Received: 06 /02/2026
Review: 08/03/2026
Accepted: 17 /03/2026
DOI: 10.22054/jocl.2025.8563.3357

Journal of Cyber Law
No(4),Vol(2), 71-92.
ISSN: 0972-6934
www.jocl.ir

Examining Input Data Challenges in AI-Based Sentencing Systems and Proposed Solutions

Aref Khalili Paji[1*], Hediye Davoodabadi[2], Fatemeh Ebrahimi [3]

1*- Assistant Professor, Department of Criminal Law and Criminology, Varamin-Pishva Branch, Islamic Azad University, Varamin, Iran.
2- Science and culture university of Tehran
3- Science and culture university of Tehran

ABSTRACT

The application of Artificial Intelligence (AI) as a decision-support tool for judges in the sentencing process represents one of the most significant yet controversial developments in contemporary criminal justice. While promising to enhance the consistency, predictability, and efficiency of judicial decisions, this technology raises fundamental legal and ethical concerns. Central to these concerns is the "Input Problem." This issue refers to challenges regarding the collection, selection, refinement, and formulation of the input data that serve as the basis for analysis, assessment, and recommendations by sentencing algorithms.This challenge arises, on one hand, from the necessity for recommender systems to rely on accurate, valid data tailored to the specific characteristics of each criminal case. On the other hand, it stems from the difficulty of translating and representing the complex reality of crime, the offender's personality traits, and the socio-economic conditions influencing the offense into machine-processable data. Despite the focus of existing literature on the technical and functional aspects of this subject, its normative and ethical dimensions particularly the risk of algorithmic bias resulting from data selection or omission have seldom been analyzed in a systematic manner.Adopting a descriptive-analytical approach, this article demonstrates that the "Input Problem" is multi-layered, manifesting in issues such as data adequacy, data neutrality, and the translation of criminal reality into the language of data. Subsequently, the independent significance of each dimension is examined through the lens of fundamental criminal law principles, including justice, equality, transparency, and judicial accountability. Finally, while aligning these challenges with the constitutional principles of the Islamic Republic of Iran and the Code of Criminal Procedure, the study proposes practical solutions to overcome this normative impasse.

Keywords:
Artificial Intelligence, Sentencing, Input Problem, Algorithmic Bias, Judicial Transparency, Iranian Criminal Law.

Article Type:Research Article

How to Cite: Khalili Paji,A., davoodabadi,H. and Ebrahimi,F. (2026). Examining Input Data Challenges in AI-Based Sentencing Systems and Proposed Solution. *Journal of Cyber Law (JOCL), 2*(4), 71-92 doi: 10.22054/jocl.2025.8563.3357

)Corresponding Author:arefkhalilipaji@iau.ac.ir

نشریه حقوق سایبری
شماره (٤) ٩٢-٧١
دوره ١٤٠٤

تاریخ دریافت: ١٤٠٤/١١/١٧
تاریخ اصلاح: ١٤٠٤/١٢/١٧
تاریخ پذیرش: ١٤٠٤/١٢/٢٦

واکاوی چالش‌های داده ورودی در سیستم‌های تعیین کیفر مبتنی بر هوش مصنوعی و راهکارها

عارف خلیلی پاجی [1*]، هدیه داود آبادی [2]، فاطمه ابراهیمی [3]

۱- استادیار، گروه حقوق جزا و جرم‌شناسی، واحد ورامین- پیشوا، دانشگاه آزاد اسلامی، ورامین، ایران

۲- گروه حقوق کیفری و جرم‌شناسی،دانشگاه علم و فرهنگ،ایران،تهران

۳- گروه حقوق کیفری و جرم‌شناسی،دانشگاه علم و فرهنگ،ایران،تهران

چکیده

کاربست هوش مصنوعی به‌مثابه ابزار پشتیبان تصمیم‌گیری قضات در فرآیند تعیین کیفر، از مهم‌ترین و در عین حال مناقشه‌برانگیزترین تحولات نوین در عرصه عدالت کیفری محسوب می‌شود. این فناوری، ضمن وعده ارتقای انسجام، پیش‌بینی‌پذیری و کارآمدی آراء کیفری، مسائل و دغدغه‌های بنیادین حقوقی و اخلاقی را نیز مطرح می‌سازد. در این میان، «مسئله ورودی» از جایگاهی محوری برخوردار است. مقصود از این مسئله، چالش‌های ناظر بر گردآوری، گزینش، پالایش و صورت‌بندی داده‌های ورودی‌ای است که مبنای تحلیل، ارزیابی و توصیه الگوریتم‌های تعیین کیفر قرار می‌گیرند. این چالش، از یک‌سو ناشی از لزوم اتکای نظام‌های عدالت کیفری به داده‌های دقیق، معتبر و متناسب با ویژگی‌های اختصاصی هر پرونده کیفری است و از سوی دیگر، ریشه در دشواری ترجمه و بازنمایی واقعیت پیچیده جرم، اوصاف شخصیتی مرتکب و شرایط اجتماعی ـ اقتصادی مؤثر در ارتکاب بزه در قالب داده‌های قابل پردازش ماشینی دارد. با وجود تمرکز ادبیات موجود بر ابعاد فنی و کارکردی این موضوع، وجوه هنجاری و اخلاقی آن، به‌ویژه مخاطره بروز سوگیری الگوریتمی ناشی از انتخاب یا حذف داده‌ها، کمتر به‌طور منسجم مورد تحلیل قرار گرفته است. مقاله حاضر با اتخاذ رویکرد توصیفی ـ تحلیلی، نشان می‌دهد که «مسئله ورودی» واجد ماهیتی چندلایه بوده و در قالب مباحثی چون کفایت داده، بی‌طرفی داده‌ها و ترجمه واقعیت کیفری به زبان داده قابل تبیین است. سپس اهمیت مستقل هر یک از این ابعاد از منظر اصول بنیادین حقوق کیفری، از جمله عدالت، برابری، شفافیت و مسئولیت‌پذیری قضایی، مورد بررسی قرار گرفته و ضمن تطبیق چالش‌های مذکور تلاش می‌شود در چارچوب نظام عدالت کیفری ایران، راهکارهایی برای برون‌رفت از این بن‌بست هنجاری ارائه گردد.

کلیدواژه‌ها:

هوش مصنوعی، تعیین کیفر، داده‌های ورودی، مسئله ورودی، سوگیری الگوریتمی.

نوع مقاله: پژوهشی

نحوه استناد:

خلیلی پاجی،عارف . داودآبادی،هدیه و ابراهیمی،فاطمه . (۱۴۰۴). واکاوی چالش‌های داده ورودی در سیستم‌های تعیین کیفر مبتنی بر هوش مصنوعی و راهکارها. حقوق سایبری،۲(۴)، ۹۲-۷۱

[1] ایمیل نویسنده مسئول: arefkhalilipaji@iau.ac.ir

۱. مقدمه

ارائه تعریفی جامع و مورد اجماع از هوش مصنوعی دشوار است؛ با این حال، می‌توان آن را به‌منزله سیستمی دانست که با تحلیل محیط پیرامون و اتخاذ تصمیم‌های مبتنی بر داده، قادر به بروز رفتاری هوشمندانه با درجه‌ای از استقلال عملکردی است (انصاری، باقر و دیگران،۱۴۰۰:۳۲). این فناوری به‌تدریج در حال گسترش دامنه نفوذ خود در عرصه‌های گوناگون، از جمله مشاغل حقوقی بوده و به‌گونه‌ای فزاینده فرآیندهای سنتی تصمیم‌گیری قضایی را با چالش مواجه ساخته است. به‌نحوی که امروزه حقوق‌دانان ناگزیر با این واقعیت روبه‌رو شده‌اند که هوش مصنوعی، با ظرفیت‌های تحلیلی و پیش‌بینی‌گرایانه خود، در حال دگرگونی بنیادین سازوکارهای عدالت کیفری است.

در این چارچوب، به نظر می‌رسد هوش مصنوعی در آینده‌ای نه‌چندان دور نقشی روزافزون در نظام‌های عدالت کیفری ایفا کند؛ نقشی که از حدود بازیابی و پردازش اطلاعات فراتر رفته و به‌تدریج حوزه‌های حساس‌تری نظیر پیش‌بینی خطر، ارزیابی ویژگی‌های فردی مرتکب و حتی تعیین کیفر را در بر گیرد. در همین راستا، طی سال‌های اخیر ابزارهای الگوریتمی، چه در قالب سامانه‌های کمکی برای استخراج و تحلیل داده‌های مرتبط از سوابق قضایی و چه به‌عنوان تلاشی در جهت استانداردسازی تصمیم‌های کیفری یا کاهش ناهمگونی احکام، به‌صورت محدود در برخی نظام‌های حقوقی مورد استفاده قرار گرفته‌اند.

اگرچه ایده استقرار «قضات رباتیک» همچنان با تردیدها و انتقادات نظری و عملی قابل‌توجهی مواجه است، استفاده از هوش مصنوعی به‌عنوان ابزارهای مشورتی در پشتیبانی از تصمیم‌گیری قضات، مخالفت کمتری برانگیخته و در برخی کشورها به‌طور آزمایشی به کار گرفته شده است(Ryberg, 2025). با این حال، حتی این شکل محدود و محتاطانه از کاربست هوش مصنوعی نیز پرسش‌های بنیادینی را در خصوص عدالت، مشروعیت و اخلاق تصمیم‌گیری کیفری مطرح می‌سازد. برای نمونه، بر اساس اصل ۳۶ قانون اساسی ج.ا.ا، حکم به مجازات و اجرای آن باید تنها از طریق دادگاه صالح و به موجب قانون باشد. لذا ورود هوش مصنوعی به این فرآیند نباید به معنای سلب صلاحیت انحصاری آنچه عرفاً دادگاه صالح نامیده می‌شود در انطباق عمل با قانون تلقی گردد.

در این میان، یکی از اساسی‌ترین و در عین حال کمتر واکاوی‌شده‌ترین مسائل در به‌کارگیری هوش مصنوعی در مرحله تعیین کیفر، «چالش داده‌های ورودی» است. داده‌های ورودی در سامانه‌های هوش مصنوعی صرفاً مجموعه‌ای خنثی از اطلاعات عددی یا متنی نیستند، بلکه نقش زیرساخت معرفتی و مبنای تصمیم‌سازی الگوریتمی را ایفا می‌کنند. هرگونه نقص، سوگیری یا تقلیل‌گرایی در این مرحله می‌تواند شکافی معنادار میان واقعیت پیچیده و چندبعدی جرم و زبان محدود، کمی و صوری الگوریتم ایجاد کند. از آنجا که اعتبار و کارآمدی خروجی‌های هوش مصنوعی به‌طور مستقیم به کیفیت، جامعیت و بی‌طرفی داده‌های ورودی وابسته است، مسئله ورودی را نمی‌توان صرفاً یک مانع فنی یا اجرایی تلقی کرد. این چالش، به‌ویژه در مرحله تعیین کیفر که مستقیماً با آزادی فرد، کرامت انسانی و مشروعیت تصمیم قضایی پیوند دارد، واجد پیامدهای عمیق حقوقی و اخلاقی است و می‌تواند کارآمدی ادعایی عدالت الگوریتمی را به‌طور جدی مخدوش سازد.

در ادبیات معاصر مربوط به تعیین کیفر مبتنی بر الگوریتم، بسیاری از پژوهشگران به وجود آنچه «مسئله ورودی» نامیده می‌شود اشاره کرده‌اند و آن را مانع اصلی در مسیر توسعه سامانه‌های هوشمند قابل اتکا دانسته‌اند؛ مانعی که برخلاف تصور رایج، نه در مرحله خروجی، بلکه در مرحله ورود داده‌ها ریشه دارد (Schwarze & Roberts, 2022). با وجود این، در حالی که چالش‌هایی همچون عدم شفافیت الگوریتمی، بازتولید سوگیری‌های ساختاری و تضعیف نقش قاضی انسانی موضوع بحث‌های گسترده نظری قرار گرفته‌اند، مسئله داده‌های ورودی غالباً به‌صورت گذرا و بدون تحلیل مفهومی و هنجاری دقیق بررسی شده است.با وجود گسترش ادبیات مربوط به کاربست هوش مصنوعی در عدالت کیفری، تمرکز غالب پژوهش‌های پیشین بر کارایی، دقت پیش‌بینی و چالش‌های فنی سامانه‌های الگوریتمی بوده و مسئله داده‌های ورودی عمدتاً در قالب دشواری‌های مدیریتی یا محدودیت‌های فنی تحلیل شده است.

در این میان، کمتر کوششی برای صورت‌بندی منسجم ابعاد هنجاری این مسئله و نسبت آن با اصول بنیادین حقوق کیفری صورت گرفته است. حال آنکه مرحله ورود داده‌ها صرفاً یک گام تکنیکی در فرایند پردازش الگوریتمی نیست، بلکه نقطه‌ای تعیین‌کننده در بازنمایی واقعیت کیفری و شکل‌گیری مبانی تصمیم‌گیری قضایی به‌شمار می‌رود.

پژوهش حاضر با عبور از رویکردهای صرفاً توصیفی و فنی، نوآوری خود را بر «تحلیل هنجاری و بومی سازی چالش‌های مسئله ورودی» استوار ساخته است. در حالی که ادبیات موجود غالباً بر جنبه‌های تکنولوژیک یا مبانی اخلاقیِ عام متمرکز بوده‌اند، این نوشتار درصدد تبیین این فرضیه است که تقلیل واقعیت‌های انسانی به داده‌های کمی، چگونه می تواند با اصول بنیادینی نظیر «تَفرید قضایی مجازات[1]» و «لزوم اقناع وجدانی و مستدل بودن احکام[2]» در تعارض باشد. بر این اساس، نسبت این تحقیق با پژوهش‌های پیشین در بازخوانیِ چالشِ "جعبه سیاه" از منظر حقوق دفاعی متهم و ارائه یک چارچوب نظارتیِ مبتنی بر قانون آیین دادرسی کیفری است تا مانع از استحاله عدالت قضایی به محاسبات آماریِ بی‌روح گردد.

۱- چالش‌های فنی داده‌های ورودی در زمینه تعیین کیفر

در فرآیند تعیین کیفر مبتنی بر هوش مصنوعی، داده‌های ورودی نقطه آغاز زنجیره‌ای از تحلیل‌ها و تصمیم‌ها هستند که کیفیت و اعتبار آن‌ها مستقیماً بر خروجی نهایی اثر می‌گذارد. کاربست هوش مصنوعی در مرحله تعیین کیفر نشان می‌دهد که «چالش داده‌های ورودی» مفهومی یک‌بعدی و صرفاً فنی نیست، بلکه می‌توان آن را در سه سطح متمایز تحلیل کرد. در سطح نخست، مسئله به نحوه بازنمایی، استانداردسازی و پردازش داده‌های مرتبط با جرم و ویژگی‌های متهم بازمی‌گردد؛ سطحی که عمدتاً با پیچیدگی ماهوی جرم و محدودیت‌های تقلیل آن به شاخص‌های کمی درگیر است.

در سطح دوم، منشأ و فرایند شکل‌گیری داده‌ها مورد توجه قرار می‌گیرد؛ بدین معنا که داده‌های مورد استفاده در سامانه های الگوریتمی، بازتابی خنثی از واقعیت نیستند، بلکه در بستر ساختارهای اجتماعی، نهادی و تاریخی تولید

---

[1] موضوع ماده ۱۸ قانون مجازات اسلامی مصوب ۱۳۹۲

[2] برآمده از اصل ۱۶۶ قانون اساسی ج.ا.ا

می‌شوند و ممکن است الگوهای نابرابری و تبعیض را در خود حمل کنند. در نهایت، پیامدهای هنجاری این دو بعد پیشین مطرح می‌شود؛ جایی که این پرسش اساسی شکل می‌گیرد که ورود چه نوع داده‌هایی به فرایند تعیین کیفر با اصول بنیادین عدالت کیفری سازگار است و چگونه می‌توان از تضعیف مشروعیت تصمیم قضایی جلوگیری کرد. تفکیک این سه سطح، امکان تحلیل منسجم‌تر مسئله ورودی و ارزیابی دقیق‌تر کاربست هوش مصنوعی در تعیین کیفر را فراهم می‌سازد.

۱-۱تبیین مسئله ورودی: چالش ساختاری تقلیل واقعیت

تبیین مسئله ورودی مستلزم درک این واقعیت است که هوش مصنوعی برای عمل، مجبور به «تقلیل‌گرایی» است. بر اساس تحلیل‌های جسپر رایبرگ[1]، این چالش زمانی بروز می‌کند که سیستم تلاش می‌کند یک «تصویر کافی» از جرم ارائه دهد. رایبرگ برای تبیین این چالش ساختاری از قیاس «ماشین‌حساب ساده» بهره می‌برد؛ همان‌گونه که ماشینی با محدودیت اعشار، ناگزیر از «گرد کردن اعداد» است، هوش مصنوعی نیز ناگزیر است ظرایف انسانی (مانند بافتار اخلاقی جرم) را به دسته‌بندی‌های تقریبی عددی تقلیل دهد. مشکل بنیادین از نظر شوارتز و رابرتز[2] این است که «عوامل تعیین کیفر در بافتارهای مختلف، وزن‌های متفاوتی دارند و اهمیت آن‌ها را نمی‌توان لزوماً به یک ارزش عددی خاص تقلیل داد»(Schwarze & Roberts,2022) .

ضمن آنکه، در این چارچوب پارادوکسی به وجود می‌آید که اگر برای رفع نقص، تمامی پارامترهای یک جرم را وارد کنیم، پرونده چنان منحصربه‌فرد می‌شود که دیگر هیچ «رویه‌ی قضایی مشابهی» در پایگاه داده برای تطبیق و توصیه مجازات یافت نخواهد شد؛ امری که هدف اصلی هوش مصنوعی یعنی «سازگاری» را تضعیف می‌کند. منظور از بافتار در فرآیند تعیین کیفر، مجموعه ای از متغیرهای محیطی، روانی و موقعیتیِ غیرساختارمند است که جرم در ظرف آن اتفاق افتاده است. به عنوان مثال، در حالی که الگوریتم فقط عددِ «سرقت» و «ارزش مالی» را می‌بیند، بافتار شامل مواردی چون «اضطرار بزهکار»، «جو حاکم بر صحنه جرم» و «تاثیرات عاطفی متقابل میان بزه دیده و بزهکار» است که به راحتی به داده‌های عددی قابل تبدیل نیستند.

۱-۱-۱ضرورت ارایه اطلاعات اختصاصی پرونده و پیچیدگی ماهوی بازنمایی جرم

ضرورت ارائه داده‌های دقیق و اختصاصیِ مربوط به هر پرونده، یکی از اساسی‌ترین پیش‌شرط‌های کارآمدی سامانه‌های هوش مصنوعی در فرایند تعیین کیفر است. هر الگوریتم تنها در صورتی قادر به ارائه پیشنهاد قضایی معتبر خواهد بود که اطلاعات ورودی آن به‌صورت کامل، صحیح و با لحاظ تمامی اوصاف ماهوی جرم و ویژگی‌های فردی متهم وارد شده باشد. در غیاب چنین داده‌هایی، سامانه نه‌تنها توان تحلیل ماهیت جرم را از دست می‌دهد، بلکه ممکن است توصیه‌هایی صادر کند که با واقعیت پرونده تعارض دارد و به‌طور مستقیم عدالت کیفری را مخدوش می‌سازد.

---

[1] Jesper ryberg.
[2] Schwarze and Roberts.

دراین راستا در حقوق کیفری ایران، ماده ۱۸ قانون مجازات اسلامی بر "فردی‌سازی مجازات" با تکیه بر معیارهای کیفی نظیر انگیزه، شخصیت و وضعیت ذهنی مرتکب تأکید دارد. تقلیل این معیارهای کیفی به داده‌های کمی در سیستم‌های هوش مصنوعی، ممکن است اجرای دقیق این ماده قانونی را با چالش مواجه کند.

این مسئله از دهه ۱۹۸۰ و ۱۹۹۰ میلادی مورد توجه پژوهشگران قرار گرفته است. نخستین نظام‌های پشتیبان صدور حکم که مطالعه آن‌ها توسط دانشمندانی چون شوارتز و رابرتز[1] مورد بررسی قرار گرفت، نشان داد که مشکل ورودی[2] یکی از مهم‌ترین موانع در مسیر تصمیم‌یارهای قضایی است. مطابق این پژوهش‌ها، هرگونه نقص، ابهام یا استاندارد نبودن داده‌های مرتبط با پرونده، حتی اگر در ظاهر جزئی به نظر برسد، می‌تواند موجب اختلال در تحلیل الگوریتم شده و فرایند تفسیر ماهیت جرم را به خطا بکشاند. آن‌ها تأکید داشتند که بخش قابل‌توجهی از سوگیری‌های عملیاتی الگوریتم‌ها[3] نه از مکانیزم پردازش، بلکه از داده‌های ورودی ناقص یا نادقیق ناشی می‌شود. به طور کلی‌تر، مشکل از نظر شوارتز و رابرتز این است که «عوامل تعیین کیفر در موقعیت‌های متفاوت، وزن‌های متفاوتی دارند و اهمیت آن‌ها را نمی‌توان از قبل تعیین کرد یا لزوماً به یک ارزش عددی خاص تقلیل داد»؛ و این‌که «به دشواری می‌توان دریافت که تفسیر،بافت،تبیین و... چگونه ممکن است [در الگوریتم] لحاظ شوند»(Schwarze & Roberts, 2022).

در ادامه همین ادبیات، رابرتز[4] با تأکید بر «پیچیدگی ماهوی بازنمایی جرم» توضیح می‌دهد که داده‌های ورودی نه مجموعه‌ای ساده از اطلاعات، بلکه بازتابی از واقعیت اجتماعی، رفتار انسانی و ساختارهای پیچیده حقوقی هستند. به تعبیر او، هر داده ظاهراً ساده مانند عنوان اتهام، اوصاف قانونی جرم، سابقه کیفری، یا شرایط ارتکاب – در خود لایه‌های متعدد و معانی حقوقی و اجتماعی نهفته دارد. نادیده‌گرفتن این پیچیدگی، موجب می‌شود سامانه در تحلیل جرم دچار ساده‌سازی مفرط شود؛ امری که در نهایت می‌تواند به تصمیماتی سوگیرانه، تبعیض‌آمیز یا فاقد تناسب منجر گردد. وی هشدار می‌دهد که فقدان شفافیت در مرحله ورود داده‌ها و بی‌توجهی به ماهیت پیچیده آن‌ها، بستری برای تقویت سوگیری‌های الگوریتمی و ایجاد پیامدهای جدی در عدالت کیفری فراهم می‌کند. (Schwarze & Roberts, 2022).

بر مبنای این دیدگاه‌ها، مسئله «ضرورت ارائه اطلاعات اختصاصی پرونده» و «پیچیدگی ماهوی بازنمایی جرم» نه یک چالش حاشیه‌ای، بلکه جوهره بحث فنی درباره به‌کارگیری هوش مصنوعی در صدور احکام قضایی است. این پرسش بنیادین باقی است که داده‌های ورودی دقیقاً چه ویژگی‌هایی باید داشته باشند، چگونه باید استانداردسازی شوند و سامانه‌های قضایی تا چه اندازه قادرند از تحریف، حذف، ساده‌سازی یا سوگیری در مرحله ورود داده‌ها جلوگیری کنند. تبیین این موضوع، از مهم‌ترین گام‌ها برای ارائه تصویری روشن از چالش‌های اخلاقی و روش‌شناختی هوش مصنوعی در نظام عدالت کیفری به‌شمار می‌آید.

---

[1] Schwarze and Roberts.
[2] Input problem.
[3] Operational biases of algorithms.
[4] Roberts.

۱-۱-۲چالش کمیت و پیچیدگی داده‌ها و پیامدهای آن در تضعیف عدالت کیفری

یکی از مسائل اساسی در استفاده از هوش مصنوعی برای کمک به صدور حکم، کمّیت بسیار زیاد داده‌های لازم و پیچیدگی ذاتی اطلاعات مربوط به ماهیت جرم است. هرچند در ظاهر ممکن است تصور شود که طبقه‌بندی جرایم در دسته‌هایی مانند قتل، سرقت یا تجاوز جنسی کار را برای الگوریتم ساده می‌کند، اما این طبقه‌بندی‌های کلی هرگز قادر به بازتاب کامل واقعیت‌های جزئی و متکثر رفتار مجرمانه نیستند. در فرآیند صدور حکم، عنوان قانونیِ جرم به‌تنهایی برای تعیین مجازات کافی نیست؛ زیرا هر جرم حتی تحت یک عنوان واحد دارای شدت، شرایط ارتکاب، ویژگی‌های موقعیتی و پیامدهای متفاوت است که همگی باید در تحلیل وارد شوند.مطالعات نخستین در حوزه سیستم‌های پشتیبان قضایی این واقعیت را به خوبی نشان داده‌اند. یوری شیلد[1] توضیح می‌دهد که حتی مفهومی بسیار آشنا مانند «سابقه کیفری[2]» در سطح داده، مجموعه‌ای چندبعدی از متغیرهاست: تعداد محکومیت‌ها، تازگی آخرین محکومیت، نوع جرایم، مدت فاصله با جرم قبلی و شدت آخرین محکومیت (Schild,1998). تنها ترکیب همین چند متغیر می‌تواند بیش از ۷۰۰ حالت تولید کند، بنابراین خلاصه کردن آن‌ها در یک عنوان واحد برای الگوریتم عملاً گمراه‌کننده است. این نمونه کوچک نشان می‌دهد که یک سامانه هوشمند تنها زمانی می‌تواند رفتار مجرمانه را به‌درستی تحلیل کند که حجم زیادی از اطلاعات دقیق و تفصیلی در اختیارش قرار گیرد.

رابرتز نیز همین مسئله را در سطح کلان‌تر طرح می‌کند و هشدار می‌دهد که بازنمایی ماهیت جرم در قالب داده‌های استاندارد، امری ظاهراً ساده اما در واقع بسیار پیچیده است (Schwarze & Roberts, 2022). حتی اگر عنوان جرم به شکل صحیح تعیین شده باشد، ممکن است بسیاری از ویژگی‌های مهم و تأثیرگذار جرم در داده‌های ورودی ثبت نشده باشند؛ در نتیجه الگوریتم یا نمونه‌های مشابه را پیدا نمی‌کند یا مجبور می‌شود از داده‌های ناقص استفاده کند. از این‌رو، وقتی داده‌های ورودی جزئیات کافی را منعکس نکند، خروجی الگوریتم نه دقیق خواهد بود و نه از منظر عدالت کیفری قابل اتکا.بر این اساس، چالش اصلی در این حوزه نه صرفاً کمبود داده، بلکه وفور داده‌هایی است که هر کدام دارای سطوح مختلف، وابستگی‌های متقابل و اهمیت‌های متفاوت‌اند. پردازش این اطلاعات در مقیاس بزرگ نیازمند سامانه‌هایی است که هم ظرفیت فنی بالا داشته باشند و هم از نظر ساختار طراحی، قادر به درک پیچیدگی ماهوی رفتار مجرمانه باشند. بی‌توجهی به این سطوح پیچیدگی موجب می‌شود تصمیم‌های الگوریتمی به‌ظاهر دقیق، اما در واقع تقلیل‌یافته، ناقص و گاه ناعادلانه باشند.

در بهره‌گیری از سامانه‌های هوش مصنوعی برای صدور حکم، یکی از چالش‌های اساسی آن است که کمّیت انبوه اطلاعات و پیچیدگی ذاتی متغیرهای مؤثر بر جرم، عملاً هدف اصلی نظام کیفری را تضعیف کرده و به ایجاد ناسازگاری در تعیین کیفر منجر می‌شود. تصمیم‌گیری کیفری هرگز فرآیندی ساده یا دوگانه (دودویی) نیست و نمی‌توان آن را به چند شاخص تقلیل داد. حتی در ظاهرِ ساده‌ترین جرایم، مجموعه‌ای از عوامل متعدد وجود دارند که

[1] Uri j.schild.
[2] Criminal record.

در تعیین واکنش قضایی نقش اساسی ایفا می‌کنند: شدت آسیب واردشده، وضعیت بزه‌دیده، انگیزه و شرایط ارتکاب، پیامدهای اجتماعی و میزان پشیمانی یا عدم پشیمانی مرتکب. این عناصر نه‌تنها با یکدیگر ارتباط متقابل دارند، بلکه هر یک درجات متفاوتی از اهمیت را در ارزیابی قضایی ایجاد می‌کنند.

مطالعات انجام‌شده نیز این پیچیدگی را به‌خوبی نشان داده‌اند. رابرتز و شوارتز[1] تأکید می‌کنند که حتی متغیری به‌ظاهر روشن مانند «وجود آسیب به قربانی» در عمل از ابعاد متعددی تشکیل می‌شود؛ از میزان و نوع آسیب گرفته تا پیامدهای روانی و اجتماعی آن(Schwarze & Roberts, 2022) . به بیان آنان، همین یک عامل که معمولاً به‌عنوان یکی از مؤلفه‌های جهانی در نظام‌های تعیین مجازات تلقی می‌شود خود دارای حالت‌های بسیار متنوعی است و نمی‌توان آن را به یک وصف واحد در داده‌های الگوریتمی فروکاست. تنوّع گسترده این وضعیت‌ها نشان می‌دهد که هرگونه تلاش برای ساده‌سازی، ناگزیر به حذف بخش مهمی از واقعیت جرم منتهی می‌شود.همچنین تجربه‌های عملی در حوزه تصمیم‌گیری قضایی نشان می‌دهد که عواملی چون پشیمانی یا عدم پشیمانی مجرم، که در نظام‌های کیفری نقش تعیین‌کننده‌ای دارند، قابل تبدیل به داده‌های روشن و قطعی نیستند. این مؤلفه‌ها نه تنها درجات مختلف دارند، بلکه در هر پرونده باید در پرتو زمینه‌های خاص آن تفسیر شوند. اگر سامانه هوش مصنوعی مجبور شود این مؤلفه‌های پیچیده را در قالب داده‌های خشک و از پیش تعیین‌شده بگنجاند، حاصل کار «تفسیر تقلیل‌یافته»‌ای از جرم خواهد بود که نه واقعیت پرونده را بازتاب می‌دهد و نه عدالت کیفری را.

از سوی دیگر، چنانچه تمامی پارامترهای یک جرم خاص ارائه گردند، ممکن است وضعیتی ایجاد شود که در پایگاه داده هیچ جرم سابقی وجود نداشته باشد که با جرم فعلی مطابقت داشته باشد؛ و بنابراین، راه ساده‌ای برای یافتن یک سابقه قضایی (رویه قضایی) که بتوان از آن برای ارائه توصیه مجازات بهره برد، وجود نخواهد داشت (Schild,160:1998). از این‌رو، هرگونه تکیه بیش از حد بر متغیرهای استانداردشده—بدون توجه به تفاوت‌های ماهوی و توصیفی هر پرونده خطر آن را دارد که هدف اصلی تعیین کیفر، یعنی واکنش متناسب و عادلانه به رفتار مجرمانه، تضعیف شود. در نتیجه، به جای دستیابی به انسجام، نوعی ناسازگاری ساختاری در توصیه‌های الگوریتمی شکل می‌گیرد؛ زیرا سامانه تلاش می‌کند واقعیت‌های چندلایه و ظریف را در قالب شاخص‌هایی ساده‌شده بگنجاند. این امر نه تنها کارکرد عدالت‌محور نظام کیفری را مختل می‌کند، بلکه باعث می‌شود تصمیمات صادره فاقد ظرافت لازم برای ارزیابی رفتار انسانی باشد.

۱-۲ منشأ و فرایند شکل‌گیری داده‌ها

با گسترش استفاده از هوش مصنوعی در حوزه‌های تصمیم‌گیری حساس، به‌ویژه مدیریت منابع انسانی، عدالت کیفری و پزشکی، نگرانی‌های اخلاقی مرتبط با پیامدهای این فناوری به‌طور فزاینده‌ای مورد توجه قرار گرفته است. یکی از مهم‌ترین این پیامدها، سوگیری الگوریتمی است؛ وضعیتی که در آن خروجی‌های سیستم‌های مبتنی بر داده به‌صورت نظام‌مند به نفع یا ضرر گروه‌های خاصی عمل می‌کنند (فدایی بازقلعه، ۱۴۰۴). از آنجا که این سیستم‌ها در فرآیندهایی

[1] Schwarze and Roberts.

مانند استخدام، ارزیابی عملکرد، صدور احکام قضایی و تشخیص پزشکی به کار گرفته می‌شوند، پیامدهای تصمیمات آن‌ها می‌تواند تأثیرات عمیق و گاه جبران‌ناپذیری بر زندگی افراد داشته باشد.سیستم‌های هوش مصنوعی به‌طور بنیادین به داده‌هایی وابسته‌اند که برای آموزش و بهینه‌سازی آن‌ها مورد استفاده قرار می‌گیرند. این داده‌ها بازتابی صرفاً فنی و خنثی از واقعیت نیستند، بلکه در بستر ساختارهای اجتماعی، تاریخی و نهادی شکل می‌گیرند. داده‌های تاریخی اغلب حامل الگوهای نابرابری، تبعیض و ترجیحات انسانی هستند که در طول زمان در جامعه تثبیت شده‌اند.

بنابراین، استفاده از این داده‌ها بدون بازنگری انتقادی می‌تواند به بازتولید همان نابرابری‌ها در قالبی فناورانه منجر شود (Chouldechova & Osoba,2017).علاوه بر این، تصمیم‌گیری درباره این‌که چه داده‌هایی جمع‌آوری شوند، کدام متغیرها اهمیت داشته باشند و چه داده‌هایی کنار گذاشته شوند، همگی فرایندهایی انسانی و هنجاری هستند. این تصمیمات معمولاً تحت تأثیر ارزش‌ها، محدودیت‌ها و پیش‌فرض‌های تصمیم‌گیرندگان قرار دارند و می‌توانند به حذف یا کم‌نمایی برخی گروه‌های اجتماعی منجر شوند. در نتیجه، حتی پیش از مرحله طراحی الگوریتم، بذر سوگیری می‌تواند در مرحله جمع‌آوری داده‌ها کاشته شود. از این منظر، منشأ داده‌ها و منطق حاکم بر گردآوری آن‌ها نقشی اساسی در شکل‌گیری سوگیری الگوریتمی ایفا می‌کند.

۱-۲-۱سوگیری الگوریتمی به‌مثابه بازتولید ساختاری نابرابری

سوگیری الگوریتمی اغلب بازتابی مستقیم از نابرابری‌های تاریخی و ساختاری جامعه است. الگوریتم‌ها در خلأ طراحی و اجرا نمی‌شوند، بلکه بر داده‌هایی تکیه دارند که خود محصول روابط قدرت، ساختارهای اقتصادی و الگوهای تبعیض‌آمیز گذشته هستند. نمونه‌های برجسته‌ای مانند الگوریتم «کامپاس[1]» در نظام عدالت کیفری ایالات متحده نشان می‌دهند که حتی سیستم‌هایی که با هدف افزایش عینیت و کاهش خطای انسانی طراحی شده‌اند، می‌توانند به نتایج نابرابر منجر شوند.(Angwin,2016 ; O’Neil,2016) سوگیری الگوریتمی را می توان به معنای انحراف سیستماتیک و غیرعادلانه در خروجی‌های هوش مصنوعی دانست که منجر به نتایجی به نفع یا ضرر یک گروه خاص (بر اساس متغیرهایی نظیر نژاد، جنسیت یا وضعیت اقتصادی) می‌شود. این سوگیری لزوماً به معنای خطای فنی نیست، بلکه ناشی از وجود نابرابری‌های تاریخی در داده‌های آموزشی است که توسط الگوریتم بازتولید می‌شود.

الگوریتم کامپاس برای پیش‌بینی احتمال تکرار جرم متهمان و کمک به تصمیمات قضایی طراحی شد. در ایالات متحده، با تحلیل داده‌های مرتبط با سوابق کیفری افراد، ویژگی‌های فردی و سابقه ارتکاب جرم، به پیش‌بینی خطر بازتحقق جرم و ارزیابی میزان خطرآفرینی مجدد زندانیان کمک می‌کنند یا در نظام عدالت کیفری انگلستان بر اساس آزمون‌های مشابه که ابزار سنجش آماری هستند یک نمره کامپیوتری داده می‌شود که این نمره نشان دهنده درجه خطرناکی وی در جامعه و میزان احتمال تکرار جرم از سوی وی است (عبداللهی و دیگران،۱۳۹۶). توسعه‌دهندگان این سیستم ادعا می‌کردند که استفاده از داده‌های آماری می‌تواند تصمیماتی عادلانه‌تر از قضاوت انسانی فراهم آورد. با این حال، استفاده از داده‌های سوگیرانه که منجر به پیش‌بینی مجرمیت فرد در آینده می‌شود، با اصل ۳۷ قانون اساسی (اصل برائت) در

[1] Compas.

تعارض است. چرا که هوش مصنوعی بر اساس سوابق محیطی یا گروهی، پیش فرض مجرمیت را جایگزین اصل برائت کرده و حق دفاع متهم را مخدوش می‌سازد.از طرفی بررسی‌های تجربی نشان داد که عملکرد این الگوریتم با الگوهای معناداری از سوگیری نژادی همراه است. تحقیقات پروپابلیکا نشان داد که متهمان آفریقایی-آمریکایی با احتمال بیشتری به‌اشتباه در دسته «پرخطر» قرار می‌گیرند، در حالی که متهمان سفیدپوست بیشتر به‌اشتباه کم‌خطر ارزیابی می‌شوند (Angwin,2016).اگرچه این یافته‌ها از سوی برخی پژوهشگران به دلیل محدودیت‌های روش‌شناختی مورد انتقاد قرار گرفت، مطالعات بعدی نیز وجود الگوهای نابرابر در خروجی‌های این الگوریتم را تأیید کرده‌اند.

این نمونه به‌روشنی نشان می‌دهد که سوگیری الگوریتمی الزاماً نتیجه خطای فنی نیست، بلکه اغلب ریشه در داده‌هایی دارد که نابرابری‌های ساختاری را بازتاب می‌دهند در چنین شرایطی، الگوریتم‌ها نه‌تنها بی‌طرف نیستند، بلکه می‌توانند به تثبیت یا حتی مشروعیت‌بخشی به تبعیض‌های موجود کمک کنند. برای مقابله با این سوگیری‌ها، استقرار نظام «ممیزی مستمر» [1]ضروری است. تعریف عملیاتی این ممیزی در فرآیند تعیین کیفر، شامل اجرای آزمون‌های برابری الگوریتمی[2] به صورت دوره‌ای است؛ به گونه‌ای که سیستم به‌طور مداوم توسط تیم‌های مستقل فنی و حقوقی مورد بازبینی قرار گیرد تا اطمینان حاصل شود که متغیرهای حساس (مانند قومیت یا وضعیت اقتصادی) به طور ناخواسته وزنِ تعیین‌کننده در پیشنهاد کیفر پیدا نکرده‌اند.نمونه شناخته‌شده دیگر، ابزار استخدام مبتنی بر هوش مصنوعی آمازون است که با هدف ساده‌سازی و تسریع فرایند غربالگری رزومه‌ها توسعه یافت. این سیستم بر داده‌های استخدامی گذشته آموزش دیده بود و در نتیجه، الگوهای تاریخی ترجیح جنسیتی را بازتولید کرد. بررسی‌ها نشان داد که این ابزار به‌طور نظام‌مند به نفع متقاضیان مرد عمل می‌کند و رزومه‌های مرتبط با زنان را در رتبه‌های پایین‌تری قرار می‌دهد(O’Neil,2016). در نهایت، این پروژه به دلیل سوگیری جنسیتی کنار گذاشته شد. این مثال نیز نشان می‌دهد که داده‌های تاریخی، در صورت فقدان بازنگری انتقادی، قادرند تعصبات گذشته را به تصمیمات الگوریتمی آینده منتقل کنند.

در کنار این ریشه‌های ساختاری، یکی از سازوکارهای کلیدی در تداوم و تشدید سوگیری الگوریتمی، شکل‌گیری «حلقه‌های بازخورد» است. در این فرایند، خروجی‌های یک سیستم به‌عنوان داده‌های ورودی برای نسخه‌های بعدی همان سیستم یا سیستم‌های مشابه مورد استفاده قرار می‌گیرند. اگر این خروجی‌ها سوگیرانه باشند، داده‌های آینده نیز همان سوگیری را بازتاب خواهند داد و در نتیجه، چرخه‌ای خودتقویت‌کننده از تبعیض شکل می‌گیرد. این مسئله به‌ویژه در سیستم‌هایی که پس از استقرار نیز به یادگیری ادامه می‌دهند، اهمیت دوچندان دارد؛ زیرا نبود نظارت انسانی و سازوکارهای اصلاحی می‌تواند به تشدید تدریجی و پنهان سوگیری منجر شود.به همین دلیل، چارچوب‌های حقوقی و نظارتی معاصر بر ضرورت شناسایی و کنترل حلقه‌های بازخورد تأکید ویژه دارند. برای مثال، مقررات هوش مصنوعی اتحادیه اروپا استفاده از داده‌های حساس را تنها در شرایطی مجاز می‌داند که این داده‌ها برای شناسایی یا اصلاح سوگیری کاملاً ضروری باشند و همزمان با تدابیر حفاظتی دقیق همراه شوند(Lendvai & Gosztonyi,2025).

[1] Continuous Auditing.
[2] Fairness Metrics.

۱-۲-۲خطاهای شناختی و چالش‌های مرحله جمع‌آوری داده‌ها

بخش قابل‌توجهی از سوگیری الگوریتمی به خطاهای شناختی انسان در مرحله جمع‌آوری داده‌ها بازمی‌گردد. داده‌ها اغلب بازتاب‌دهنده ترجیحات، پیش‌فرض‌ها و محدودیت‌های تصمیم‌گیرندگان انسانی هستند و در نتیجه، ممکن است برخی گروه‌های اجتماعی را کم‌نمایی یا حذف کنند. نبود تنوع در منابع داده و فقدان شفافیت در فرایند جمع‌آوری می‌تواند این سوگیری‌ها را به‌صورت سیستماتیک در خروجی‌های هوش مصنوعی بازتولید کند.

مطالعات متعددی نشان داده‌اند که بسیاری از مجموعه‌داده‌های پرکاربرد یادگیری ماشین از نظر نمایندگی جمعیتی نامتوازن هستند. این مسئله می‌تواند دقت سیستم‌ها را برای گروه‌های کم‌نمایندگی‌شده کاهش دهد و به تبعیض‌های پایدار منجر شود. پژوهش‌ها در حوزه تشخیص چهره نشان داده‌اند که عملکرد الگوریتم‌ها برای زنان و افراد با پوست تیره به‌طور معناداری ضعیف‌تر است(Gebru. & Buolamwini,2018)در حوزه پزشکی، پیامدهای سوگیری داده‌ای حتی حساس‌تر و خطرناک‌تر است. تمرکز بیش‌ازحد داده‌های ژنتیکی بر جمعیت‌های خاص می‌تواند به تشخیص‌های نادرست و نابرابری‌های درمانی منجر شود. مطالعات نشان داده‌اند که حذف یا کم‌نمایی برخی گروه‌های نژادی در داده‌های پزشکی می‌تواند خطر آسیب به این جمعیت‌ها را افزایش دهد.این یافته‌ها بر ضرورت تنوع جمعیتی در داده‌های پزشکی و زیستی تأکید دارند .

۲-ارزیابی اخلاقی و حقوقی چالش‌های داده ورودی

ارزیابی اخلاقی و حقوقی چالش داده‌ها، فراتر از یک بررسی تکنیکال، به پرسش‌های بنیادین در فلسفه حقوق پاسخ می‌دهد. لندوای و گوستونی[1] سوگیری ناشی از داده‌های ورودی را یک «معضل حقوقی محوری[2]» می‌نامند؛ زیرا کدورت ناشی از «اثر جعبه سیاه[3]» مانع از آن می‌شود که متهم یا قاضی بفهمند چرا یک نمره خطرناکی خاص صادر شده است.این ارزیابی بر سه محور استوار است: نخست، تعرض به «حق بر دانستن دلیل حکم»، جایی که تقلیل‌گرایی داده‌ای مانع از ارائه‌ی تبیین‌های انسانی و بافتاری می‌شود؛ دوم، بازتولید بی‌عدالتی، که در آن هوش مصنوعی به جای کاهش سوگیری، نابرابری‌های گذشته (مانند نابرابری دو برابری در مطالعه نبراسکا) را تحت نقاب «عینیت آماری» مشروعیت می‌بخشد؛ و سوم، بحران در «عاملیت انسانی»؛ ریچارد پازنر[4] هشدار می‌دهد که کار قضایی صرفاً اعمال قواعد نیست و قضات نباید به دلیل راحتیِ استفاده از هوش مصنوعی، «صلاحدید انسانی» و ارزیابی وقایع را کنار بگذارند(Posner,2010). این ابعاد اخلاقی نشان می‌دهند که داده‌ی ورودی، نه یک امر خنثی، بلکه ابزاری برای توزیع عدالت یا بی‌عدالتی است.

۲-۱ پیامدهای داده‌های ناقص بر عدالت کیفری

[1] Lendvai and Gosztonyi,2025.
[2] Core Legal Dilemma.
[3] Black-Box Effect.
[4] Posner, R.A.

تأثیر داده‌های ناقص[1] بر فرآیند عدالت کیفری را می‌توان در دو سطح به‌هم‌پیوسته اما متمایز، یعنی «اولویت‌های عملی[2]» و «چالش‌های اخلاقی»، مورد بررسی قرار داد. در سطح نخست، نقص در مرحله ورودی داده‌ها عمدتاً به‌عنوان یک معضل اجرایی و عملی ظاهر می‌شود. چنانچه برای ارائه یک «تصویر کافی» از جرم، ورود حجم گسترده‌ای از داده‌های اختصاصی ضرورت داشته باشد، این الزام فی‌نفسه می‌تواند فرآیند دادرسی را با اطاله مواجه سازد. چنین وضعیتی نه‌تنها استدلال‌های مبتنی بر سرعت و کارایی هوش مصنوعی را تضعیف می‌کند، بلکه پیامدهای نامطلوبی همچون خستگی قضایی و تأخیر در تعیین تکلیف قربانیان و مجرمان را نیز به دنبال دارد. با این حال، این جنبه از مسئله داده‌های ورودی، علی‌رغم اهمیت آن، در زمره مسائل «عملی» قرار می‌گیرد که دست‌کم در سطح نظری، راه‌حل‌های نسبتاً ساده‌ای مانند تخصیص منابع بیشتر یا بهبود زیرساخت‌ها برای آن قابل تصور است.

اما چالش بنیادین زمانی بروز می‌کند که پیچیدگی و نقص مرحله ورودی، مستقیماً بر توصیه نهایی الگوریتم و در نتیجه بر حقوق متهم اثر بگذارد و از سطح یک مشکل اجرایی فراتر رفته، ماهیتی اخلاقی به خود بگیرد. برای تبیین این وضعیت، می‌توان از قیاس «ماشین‌حساب ساده» بهره گرفت؛ همان‌گونه که استفاده از ماشین‌حسابی که تنها یک رقم اعشار را می‌پذیرد، به دلیل «گرد کردن اعداد» موجب انحراف قابل توجه نتیجه نهایی محاسبات از حقیقت می‌شود، تقلیل ویژگی‌های کیفی و پیچیده جرم به دسته‌بندی‌های تقریبی و محدود در سامانه‌های هوش مصنوعی نیز می‌تواند خروجی نهایی را مخدوش سازد.(Ryberg,2025). به‌عنوان نمونه، اگر «آسیب وارده به قربانی» صرفاً در قالب چند طبقه محدود تعریف و وارد شود، جرایمی با شدت‌ها و آثار متفاوت در یک ردیف قرار می‌گیرند و در نتیجه، خروجی الگوریتم از حکمی که باید منعکس‌کننده تمامی ویژگی‌ها و ظرایف پرونده باشد (حکم واقعی)، فاصله می‌گیرد. در همین راستا، همان‌گونه که شوارتز و رابرتز[3] به‌درستی تأکید کرده‌اند، تأثیر جرم بر قربانی یک «مؤلفه تقریباً جهانی در تعیین کیفر» در نظام‌های حقوقی کامن‌لا محسوب می‌شود.(Schwarze & Roberts, 2022)

با این حال، ارزیابی اخلاقی این انحراف الگوریتمی نمی‌تواند در خلأ صورت گیرد و مستلزم مقایسه‌ای تطبیقی با قضاوت انسانی است. اگرچه الگوریتم، به دلیل محدودیت‌ها و نقص‌های مرحله ورودی، تنها تصویری ناقص از جرم را ثبت و پردازش می‌کند، اما توانایی یک قاضی انسانی برای در نظر گرفتن همزمان تمامی ویژگی‌های مرتبط با پرونده نیز لزوماً کامل و بی‌نقص نیست و حتی ممکن است در عمل، با محدودیت‌های شناختی جدی‌تری مواجه باشد. از این‌رو، انحراف خروجی هوش مصنوعی از حکم واقعی، به‌خودی‌خود دلیلی کافی برای کنار گذاشتن آن از فرآیند عدالت کیفری محسوب نمی‌شود، مگر آنکه بتوان نشان داد قضاوت انسانی در شرایط مشابه، به حقیقت و عدالت نزدیک‌تر است. در این نقطه، عدالت کیفری ناگزیر با نوعی توازن سنجیده میان «خطای تقلیل‌گرایانه ماشین» و «محدودیت‌های شناختی انسان» مواجه می‌شود؛ توازنی که هسته اصلی بحث اخلاقی پیرامون داده‌های ناقص و کاربست هوش مصنوعی در تعیین کیفر را شکل می دهد.

---

[1] Incomplete data.t
[2] Practical priority .
[3] Schwarze and Roberts.

۲-۲ چالش‌های هنجاری داده‌های آلوده در عدالت کیفری

یکی از اهداف اصلی به‌کارگیری هوش مصنوعی در نظام عدالت کیفری، تضمین برابری و مهار نابرابری‌های فاحش در تصمیمات قضایی، به‌ویژه میان قضات مختلف است. شواهد تجربی، از جمله مطالعه‌ای انجام‌شده در نبراسکا[1]، نشان می‌دهد که یک قاضی ممکن است مجرمان مواد مخدر را به دو برابر مدت حبس نسبت به همکار خود محکوم کند. در همین راستا، پیشنهاد شده است که یک سیستم مبتنی بر یادگیری ماشین می‌تواند «نمایی لحظه‌ای و جزئی‌نگر از گرایش مرکزی نحوه‌ی برخورد آن‌ها [قضات] و همکارانشان با پرونده‌های مشابه» در اختیار قضات صادرکننده حکم قرار دهد و از این طریق، به صدور احکام یکنواخت‌تر کمک کند.(Chiao,246:2018) با این حال، «مسئله ورودی» و آلودگی داده‌ها این هدف بنیادین را به‌طور جدی تهدید می‌کند.

از منظر «پیامدگرا»، رفتار نابرابر با پرونده‌های مشابه که ناشی از نقص یا سوگیری در داده‌های ورودی است، به تضعیف حاکمیت قانون و سلب اعتماد عمومی به نظام عدالت می‌انجامد، مگر آنکه بتوان نشان داد این رفتار متفاوت، نتایج پیشگیرانه مثبت و قابل دفاعی به همراه دارد. در این چارچوب، کتی اونیل[2] هشدار می‌دهد که ورود گسترده داده‌های مربوط به تخلفات جزئی، که اغلب با فقر گره خورده‌اند، و همزمان حذف یا کم‌نمایی جرایم یقه‌سفید، نه‌تنها به تحقق برابری منجر نمی‌شود، بلکه فرآیندی از «تولید صنعتی نابرابری» را رقم می‌زند.(O’Neil,2016) این فرایند همانگونه که بیان شد منجر به تثبیت چرخه سوگیری می گردد.

در سطحی عمیق‌تر، داده‌های ناقص یا آلوده موجب بروز تعارضی جدی در منطق سزادهی می‌شوند. از منظر «سزادهی» (kopf,2012)، مجازات باید نتیجه استحقاق باشد و با شدت و ویژگی‌های جرم تناسب داشته باشد. انحراف ناشی از پیچیدگی و نقص داده‌های ورودی، این «تناسب نسبی» را مخدوش می‌کند، زیرا افراد یا پرونده‌های مشابه ممکن است به‌طور نابرابر دسته‌بندی و ارزیابی شوند. با این حال، اگر نظام قضایی موجود به‌طور ساختاری گرایش به «بیش‌کیفردهی[3]» داشته باشد، انحراف الگوریتمی به سمت حکمی ملایم‌تر می‌تواند از منظر «تناسب مطلق» اخلاقاً ارجح تلقی شود(Ryberg,2025). این وضعیت نشان می‌دهد که ارزیابی اخلاقی خروجی الگوریتم، ناگزیر وابسته به مقایسه آن با عملکرد واقعی نظام قضایی انسانی است، نه با یک معیار انتزاعی و ایده‌آل.

در کنار مسئله برابری، داده‌های آلوده و ساده‌سازی‌های مرحله ورودی، چالش مستقلی را در حوزه شفافیت و «حق دانستن دلیل حاکم» ایجاد می‌کنند. حق متهم برای اطلاع از مبانی و دلایل صدور حکم، با پدیده‌هایی همچون «دسته‌بندی‌های تقریبی» و «اثر جعبه سیاه[4]» مواجه است. لندوای و گوستونی[5] تأکید می‌کنند که سوگیری الگوریتمی به یک «معضل حقوقی محوری» بدل شده است، زیرا کدورت و عدم شفافیت فرآیند تصمیم‌گیری، مانع از آن می‌شود که متهم درک کند چرا به‌عنوان «پرخطر» دسته‌بندی شده است(Lendvai & Gosztonyi, 2025) . استفاده از

[1] Nebraska.
[2] Catherine a, o’neill.
[3] Over-punishment.
[4] Black-box effect,
[5] Lendvai and Gosztony.

مدل‌های ساده‌ساز در مرحله ورودی، اگرچه ممکن است از منظر فنی کارآمد به نظر برسد، اما ظرایف و ویژگی‌های خاص هر پرونده را در تحلیل محو می‌کند و فاصله‌ای معنادار میان تصمیم نهایی و واقعیت پرونده ایجاد می‌نماید. در واکنش به این نگرانی‌ها، قانون هوش مصنوعی اتحادیه اروپا[1] سامانه‌های قضایی را در زمره سیستم‌های «پرخطر» قرار داده و آن‌ها را ملزم به ثبت فعالیت‌ها و ارائه مستندات فنی دقیق کرده است تا امکان شفافیت و ردیابی نتایج تصمیم‌گیری تضمین شود (European Union,2024). از سوی دیگر عدم شفافیت در الگوریتم‌ها، با تکلیف قانونی ذکر شده مبنی بر مستدل و مستند بودن آرا که در نظام‌های عدالت کیفری آمده در تضاد است. برای نمونه، مطابق اصل ۱۶۶ قانون اساسی و ماده ۲ قانون آیین دادرسی کیفری، احکام دادگاه‌ها باید «مستدل و مستند» به مواد قانون و اصولی باشد که بر اساس آن حکم صادر شده است. وقتی منطق استخراجِ پیشنهاد از سوی هوش مصنوعی برای متهم و حتی قاضی مخفی بماند، وصف «مستدل بودن» حکم زیر سؤال می‌رود. لذا تحقق شفافیت در این مقام، مستلزم عملیاتی‌سازی «شفافیت داده»[2] است. این شفافیت به معنای افشای کدهای منبع (که اسرار تجاری شرکت‌هاست) نیست، بلکه به معنای «قابلیت ممیزی ورودی‌ها» برای طرفین دعواست. به عبارت دیگر، متهم و وکیل وی باید حق داشته باشند به لیست تمامی داده‌های خامی که به الگوریتم تغذیه شده دسترسی داشته باشند تا بتوانند در صورت وجود داده‌های نادرست یا غیرقانونی، مطابق با مواد قانونی آیین دادرسی، نسبت به آن‌ها اعتراض کنند،امری که تضمین کننده عینی استدلال قضایی در عصر دیجیتال خواهد بود.

چالش نهایی، اما شاید بنیادی‌ترین آن‌ها، به مسئله مسئولیت‌پذیری در سیستم‌های مشورتی مبتنی بر هوش مصنوعی بازمی‌گردد. ظهور و گسترش سامانه‌های تسلیحاتی خودگردان و ابزارهای پیش‌بینی‌گر جرم، مفروضات دیرپای حقوقی درباره عاملیت انسانی و اراده آزاد را به چالش کشیده است؛ زیرا برای نخستین بار، دانش حقوق با پدیده‌ای مواجه شده است که ممکن است پس از طی مراحلی، حتی سازندگان و طراحان آن نیز قادر به درک کامل نحوه عملکرد یا مهار آن نباشند (ابوذری،مهرنوش،۱۴۰۰:۲۲۸). سپردن تصمیمات به سامانه‌هایی که در سطحی فراتر از کنترل مستقیم انسان عمل می‌کنند، مفاهیم سنتی مسئولیت کیفری را با بحرانی جدی روبه‌رو ساخته است.

در این میان، برخی دیدگاه های پیشرو حقوقی بر این باورند که برای حل بحران انتساب مسئولیت در قبال جرایم الگوریتمی، باید در مفاهیم کلاسیک بازنگری اساسی صورت گیرد؛ چراکه هوش مصنوعی می‌تواند، دست‌کم در سطح نظری، واجد رکن روانی مدنظر برای تحقق جرم باشد (هالوی ،۱۰۷:۱۴۰۲). این دیدگاه نه‌تنها مبانی «عاملیت» را دگرگون می‌سازد، بلکه امکان کیفردهی مستقیم به سامانه‌های هوشمند را به‌عنوان راهکاری برای پایان دادن به وضعیت بی‌کیفرمانی ناشی از شکاف مسئولیت مطرح می‌کند. با این حال، پازنر[3] با تأکید بر اینکه کار قضایی صرفاً به اعمال قواعد محدود نمی‌شود، تصریح می‌کند که اگر وظیفه قضات تنها اجرای قواعدی واضح و روشن باشد، آنگاه واگذاری این نقش به برنامه‌های دیجیتال و هوش مصنوعی منطقی‌تر خواهد بود (Posner,2010). حتی کسانی که نقش قضات

[1] EU AI Act.
[2] Data Transparency .
[3] Posner, R.A.

را عمدتاً در یافتن وقایع و اعمال قواعد می‌دانند نیز اذعان دارند که این ایده به‌طور کامل عملی نیست و دادرسان و دادستان‌ها در بسیاری از موارد ناگزیر از اعمال صلاحدید انسانی خود هستند (جعفری تبار،۱۳۹۵:۱۶۲). از این‌رو، در سیستم‌های مشورتی کنونی، مسئولیت نهایی همواره بر عهده انسان، یعنی قاضی، باقی می‌ماند؛ قاضی‌ای که باید میان «دقت فنی ماشین» و «عدالت انسانیِ پرونده‌محور» توازنی سنجیده برقرار کند.

۳- راهکارهای مواجهه با چالش داده‌های ورودی

مسئله‌ی داده‌های ورودی در کاربرد هوش مصنوعی در عدالت کیفری، نقطه‌ای است که در آن ملاحظات فنی با الزامات حقوقی و اصول دادرسی منصفانه تلاقی می‌کنند. داده‌ها نه‌تنها مادهٔ اولیه‌ی تصمیم‌سازی الگوریتمی‌اند، بلکه کیفیت، شیوه گردآوری و نحوهٔ پردازش آن‌ها می‌تواند به‌طور مستقیم بر انصاف، شفافیت و مشروعیت تصمیم قضایی اثر بگذارد. ازاین‌رو، چالش داده‌های ورودی را نمی‌توان صرفاً با راه‌حل‌های فنی برطرف کرد، بلکه مواجهه با آن مستلزم رویکردی چندلایه است که استانداردسازی و پاکسازی داده‌ها، ممیزی و ارزیابی سوگیری، حفظ نقش فعال قاضی، تضمین شفافیت ، حق اعتراض و نیز قانونگذاری و طراحی اخلاق‌محور را به‌صورت توأمان در بر گیرد. در این چارچوب، راهکارهای پیشنهادی در این بخش می‌کوشند نشان دهند که کنترل ریسک‌های داده‌محور تنها زمانی ممکن است که داده، الگوریتم و تصمیم انسانی در قالب یک سازوکار پاسخگو و قابل ممیزی به هم پیوند بخورند.

۳-۱- استانداردسازی و پاکسازی داده‌ها

نقطهٔ شروع هر سامانه هوشمند در عدالت کیفری، «داده» است؛ همان مادهٔ اولیه‌ای که اگر مخدوش، ناقص یا نامتوازن باشد، حتی دقیق‌ترین الگوریتم‌ها هم خروجی‌هایی تولید می‌کنند که از منظر حقوقی قابل اتکا نیست. به همین دلیل، استانداردسازی و پاکسازی داده‌ها باید نه به‌عنوان یک مرحله فنیِ صرف، بلکه به‌منزلهٔ یک تضمین رویه‌ای برای دادرسی منصفانه دیده شود. در ادبیات حقوقیِ مرتبط با شفافیت در رسیدگی‌های قضایی نیز بر این نکته تأکید شده که وقتی داده‌ها را مثل مواد اولیه خط تولید در نظر بگیریم، شفافیت و کیفیت آن‌ها شرط لازم برای اعتماد به محصول نهایی است؛ زیرا تصمیمِ مبتنی بر داده، خواه داده در آغاز وارد سیستم شده باشد یا در حین تعامل کاربر تولید و تزریق شود، در نهایت بر سرنوشت دعوا اثر می‌گذارد (حسینی و همکاران،۱۴۰۲).

استانداردسازی یعنی داده‌ها از حیث قالب، تعاریف، واحدها، برچسب‌ها و قواعد ثبت، به زبان مشترک تبدیل شوند تا هم قابلیت پردازش پیدا کنند و هم امکان بازبینی و مقایسه ایجاد شود. در نظام عدالت کیفری، داده‌ها معمولاً از منابع متعدد می‌آیند: گزارش ضابطان، سوابق کیفری، اطلاعات جمعیتی، داده‌های مکانی زمانی جرم، و حتی داده‌های متنی. اگر این داده‌ها با استاندارد واحد گردآوری نشوند، الگوریتم در عمل به جای «کشف الگو»، گرفتار اختلافِ قالب‌ها و سوگیری‌های ناشی از پراکندگی داده می‌شود. از سوی دیگر، پاکسازی داده شامل رفع داده‌های تکراری، اصلاح خطاهای ثبت، مدیریت داده‌های پرت، و مهم‌تر از همه مدیریت داده‌های ناقص است. در مطالعات مربوط به انصاف در پلیس پیش‌بین نیز صراحتاً نشان داده شده که کیفیت داده و وجود داده‌های ناقص می‌تواند بر نتیجه اثر جدی بگذارد؛ برای نمونه در یک مطالعه در حوزه پلیس پیش‌بین، ستون‌های دارای مقادیر گمشده قابل‌توجه حذف و ردیف‌های دارای نقص در مختصات مکانی کنار گذاشته شده‌اند(Almasoud،2025) .

پاکسازی، بدون نگاهِ عدالت‌محور کافی نیست؛ چون مسئله فقط «درست شدن» داده نیست، بلکه «نمایندگی منصفانه» گروه‌ها در داده است. مرور جامع سوگیری و انصاف در یادگیری ماشین توضیح می‌دهد که سوگیری می‌تواند از تفاوت جمعیتِ واقعی با جمعیتِ منعکس در داده، از نمونه‌گیری نامتوازن و از خطاهای تجمیع داده‌ها ایجاد شود و حتی وقتی گروه‌ها ظاهراً به‌طور برابر حضور دارند، یک مدل واحد ممکن است برای همه گروه‌ها مناسب نباشد و به «سوگیری تجمیع» بینجامد(Mehrabi,Morstatter&Saxena,2019) .بنابراین استانداردسازی و پاکسازی باید همراه با آزمون‌های نمایندگی داده باشد: آیا گروه‌های حساس (جنسیت، سن، قومیت، وضعیت اقتصادی) به اندازه کافی حضور دارند؟ آیا داده‌های مربوط به گروه‌های کم‌نمونه، بیش از حد ناقص است؟ آیا قواعد ثبت داده در برخی مناطق یا کلانتری‌ها سخت‌گیرانه‌تر یا سهل‌گیرانه‌تر بوده و همین تفاوت، الگوی مصنوعی ایجاد کرده است؟ این پرسش‌ها داده را از یک شیء فنی به یک موضوع «قابل ممیزی حقوقی» تبدیل می‌کند(محرابی و همکاران،۲۰۱۹).

در نهایت، بهترین رویه این است که برای هر پروژه هوش مصنوعی در عدالت کیفری، یک «پروتکل داده» تدوین شود؛ پروتکلی که شامل تعریف دقیق متغیرها، روش‌های کنترل کیفیت، ثبت تغییرات داده[1] و معیارهای پذیرش یا رد داده باشد تا بعداً بتوان توضیح داد سامانه بر اساس چه داده‌ای آموزش دیده و چرا. این نیاز، در بحث شفافیت داده‌های آموزشی در رسیدگی نیز برجسته شده است؛ زیرا اگر نتوان روشن کرد سامانه دقیقاً بر اساس چه داده‌هایی آموزش دیده، اصل استفاده از هوش مصنوعی در رسیدگی با تردید مواجه می‌شود (حسینی و همکاران،۱۴۰۲).

۳-۲ طراحی اخلاق محور در مرحله توسعه

تضمین عدالت و کاهش خطا در سامانه های هوش مصنوعی مرتبط با دادرسی، اگر به پس از استقرار موکول شود، پرهزینه و کم‌اثر خواهد بود؛ بنابراین رویکرد درست، اخلاق‌محوری از همان ابتدای طراحی است. طراحی اخلاق‌محور به این معناست که از مرحلهٔ تعریف مسئله و گردآوری داده، فرض را بر وجود نابرابری‌های اجتماعی نهفته در داده‌ها بگذاریم تا مدل آن‌ها را بازتولید نکند. ادبیات سوگیری نشان می‌دهد که سوگیری‌های داده‌محور (نمونه‌گیری، اندازه‌گیری، تجمیع) حتی پیش از آموزش مدل، جهت‌گیری ایجاد می‌کنند؛ از این رو، اجرای پروتکل‌های کنترل کیفیت داده، مدیریت نقص‌ها و آزمون نمایندگی گروه‌ها ضروری است (محرابی و همکاران، ۲۰۱۹).

از منظر فنی، باید معماری‌هایی انتخاب شوند که تفسیرپذیری و امکان نظارت را در کنار دقت فراهم آورند. پژوهش‌ها در حوزهٔ پیش‌بینی تکرار جرم نشان می‌دهند مدل‌های تفسیرپذیرتر می‌توانند دقت و انصاف را هم‌زمان حفظ کرده و ممیزی و حق اعتراض را تقویت کنند (Wang,Han,Patel&Rudin,2022). با این حال، هوش مصنوعی علیرغم دستاوردهای چشمگیر، در مواجهه با استدلال حقوقی پیچیده، رعایت بی‌طرفی واقعی و جلب پذیرش عمومی با چالش‌های جدی روبروست و نمی‌تواند جایگزین کامل قاضی انسانی شود. در نتیجه، در طراحی باید سامانه‌ها صرفاً در نقش ابزار کمکی باقی بمانند تا نقش فعال و تصمیم‌گیرندهٔ قاضی حفظ شود (رهبری و همکاران، ۱۴۰۱). طراحی همچنین باید رفتار کاربر را در نظر گیرد: رابط کاربری می‌بایست عدم‌قطعیت و محدودیت‌های سامانه را برجسته کند تا

[1] data lineage.

از اتکای کورکورانهٔ کاربر جلوگیری شود. در نهایت، تیم توسعه باید بین‌رشته‌ای باشد و مستندسازی جامع، معیارهای سنجش سوگیری و پایش دوره‌ای «رانش[1]» از ابتدا در سامانه تعبیه شود تا پاسخگویی به ویژگی ذاتی آن تبدیل گردد. حتی در صورت پاکسازی داده‌ها، امکان بروز سوگیری در خود مدل یا در نحوهٔ به‌کارگیری آن باقی می‌ماند. به همین دلیل، ممیزی الگوریتمی باید به‌عنوان سازوکاری مستمر تعریف شود: ارزیابی پیشینی پیش از استقرار، ارزیابی هم‌زمان در حین بهره‌برداری و ارزیابی پسینی پس از مشاهدهٔ آثار واقعی در پرونده‌ها. ادبیات حقوقیِ سوگیری الگوریتمی تأکید دارد که برای پر کردن شکاف میان مقررات حقوقی و پیاده‌سازی فنی، باید از روش‌های کاهش سوگیری به‌صورت جدی استفاده کرد. این روش‌ها عموماً در سه مرحله دسته‌بندی می‌شوند: پیش‌پردازش، حین‌پردازش و پس‌پردازش. بدون ممیزی‌های فنی جامع، نهاد ناظر عملاً نمی‌تواند تشخیص دهد که آیا سیستم با الزامات انصاف مطابقت دارد یا خیر (Lendvai&Gosztonyi,2025).ارزیابی سوگیری مستلزم آن است که پیش از هر چیز، «انصاف» به روشنی تعریف شده و سپس شاخص‌های متناسب با آن تعریف سنجیده شوند. اما خود این تعریف، واحد و جهان‌شمول نیست. مرورهای نظری نشان می‌دهند که نبود تعریف واحد از انصاف، یکی از دشواری‌های بنیادین این حوزه است (Mehrabi,Morstatter&Saxena,2019).در حوزهٔ عدالت کیفری، شاخص‌های انصاف معمولاً باید هم‌سو با ملاحظات حقوق بنیادین و منع تبعیض بوده و هم‌زمان با کارایی عملیاتی سیستم در تعارض نباشند. پژوهش‌ها در حوزهٔ پلیس پیش‌بین نشان می‌دهند که بسیاری از مداخلات «منصفانه‌سازی» با نوعی بده‌بستان میان دقت و انصاف همراه است؛ هرچند این رابطه همیشگی نیست و گاهی می‌توان سوگیری را بدون افت محسوس کارایی کاهش داد (Almasoud,2025).

نکتهٔ کلیدی در ممیزی، توجه به منشأ سوگیری است. گاهی سوگیری نه از الگوریتم، که از «کنش انسانی» نشأت می‌گیرد. در یک تحلیل علّی از پلیس پیش‌بین گزارش شده که چرخه‌ای معیوب شکل می‌گیرد: افزایش استقرار پلیس منجر به افزایش دستگیری می‌شود، این امر افزایش نرخ جرم گزارش‌شده را در پی دارد و سپس توجیهی برای استقرار بیشتر پلیس فراهم می‌کند. در نتیجه، تبعیض می‌تواند به‌صورت چرخه‌ای بازتولید شود، حتی اگر مدل در ظاهر خنثی باشد (Almasoud,2025). این بدان معناست که ممیزی نباید صرفاً به کد محدود شود، بلکه باید تعامل میان داده، مدل و تصمیم انسانی را نیز مورد سنجش قرار دهد. در ابزارهای پیش‌بینی خطر تکرار جرم نیز مسئلهٔ ممیزی بسیار پررنگ است. پژوهشی در زمینهٔ یادگیری ماشینِ «قابل تفسیر، منصفانه و دقیق» برای پیش‌بینی تکرار جرم نشان می‌دهد که رویکردهای اعمال انصاف عمدتاً در سه دسته جای می‌گیرند: پیش‌پردازش ویژگی‌ها، تغییر تابع هزینه در مرحلهٔ آموزش و پس‌پردازش خروجی‌ها. بر این اساس، بسیاری از روش‌های رایج یا قابلیت تفسیر را قربانی می‌کنند یا خروجی را به «اصلاحاتی غیرقابل توضیح» تبدیل می‌نمایند. بنابراین، اگر قرار است چنین ابزاری در محیط قضایی به کار رود، باید به‌گونه‌ای طراحی شود که هم امکان ممیزی فراهم باشد و هم برای استفاده‌کنندگان حقوقی قابل فهم باقی بماند (Wang,Han, Patel&Rudin,2022).در نتیجه، ممیزی الگوریتمی در عدالت کیفری باید دوگانهٔ

[1] Drift.

«انصافتفسیرپذیری» را هم‌زمان پایش کند و به این پرسش‌های اساسی پاسخ دهد: آیا مدل به گروهی خاص آسیب نامتوازن می‌زند؟ و آیا می‌توان علت تولید خروجی خاصی را به‌طور شفاف توضیح داد.

۳-۳حفظ نقش فعال قاضی

هرچقدر هم ابزارهای هوش مصنوعی پیشرفته باشند، در عدالت کیفری نمی‌توان جایگاه قاضی را به «تأییدکنندهٔ خروجی ماشین» تقلیل داد. نقش فعال قاضی، هم تضمین‌کنندهٔ حقوق دفاعی است و هم نماد مسئولیت‌پذیری نهادی. ادبیات شفافیت در رسیدگی‌های قضایی یادآور می‌شود که یک رسیدگی عادلانه مستلزم آن است که مسیر تصمیم‌گیری قابل پیگیری باشد و امکان فهم این مسیر برای نظارت عالی‌تر و کنترل قانونی فراهم گردد (حسینی و همکاران،۱۴۰۲). اگر بخشی از این مسیر به یک مدل «جعبه سیاه» سپرده شود و قاضی نتواند بر اساس استدلال حقوقی خود آن را ارزیابی و کنترل کند، هم شفافیت و هم قابلیت تجدیدنظرخواهی تضعیف خواهد شد.در این چارچوب، قاضی باید سه کارکرد کلیدی را حفظ کند: اول، قاضی باید «داور کفایت داده» باشد؛ یعنی بتواند تشخیص دهد داده‌های ورودی از چه منبعی گردآوری شده، تا چه حد قابل اعتماد هستند و آیا ناقص یا آلوده به سوگیری‌های نهادی می‌باشند یا خیر(حسینی و همکاران،۱۴۰۲)دوم، قاضی باید «ارزیاب تناسب» باشد؛ به این معنا که حتی اگر مدل، احتمال خطر یا پیشنهاد تصمیمی ارائه دهد، قاضی باید تناسب آن را با اوضاع و احوال خاص پرونده، سیاست جنایی حاکم و اصول فردی‌سازی پاسخ کیفری بسنجد. سوم، قاضی باید «مرجع پاسخگویی» باقی بماند؛ بدین‌ترتیب که خروجی هوش مصنوعی می‌تواند به تصمیم‌گیری کمک کند، اما مسئولیت حقوقی نهایی تصمیم نباید از انسان جدا شود.

در ادبیات عدالت الگوریتمی در حوزهٔ پلیس پیش‌بین، راهکار «قرار دادن انسان در حلقه تصمیم‌گیری» به‌عنوان مسیری برای افزودن لایه‌ای از فهم، اعتبارسنجی و پاسخگویی مطرح شده است .(Almasoud,2025) همین منطق در دادگاه اهمیتی به مراتب بیشتر دارد: اگر قرار است ابزار در نقش توصیه‌گر ظاهر شود، طراحی رویه‌ای باید به گونه‌ای باشد که قاضی امکان رد توصیه را داشته باشد و دلایل پذیرش یا رد آن را به‌صورت مستدل ثبت کند. در اینجاست که نقش فعال قاضی خود به سازوکاری برای کاهش سوگیری تبدیل می‌شود؛ قاضی می‌تواند خروجی الگوریتم را با قرائن موجود در پرونده، اصول بی‌طرفی و معیارهای قانونی تطبیق دهد و مانع از آن شود که «سوگیریِ داده» به «سوگیریِ حکم» تبدیل گردد.

۳-٤ قانونگذاری هوشمندانه؛ وجود شفافیت و حق اعتراض متهم

قانونگذاری در حوزه کاربردهای هوش مصنوعی در عدالت کیفری، اگر صرفاً به گزاره‌های کلی مانند «منع تبعیض» یا «لزوم رعایت حریم خصوصی» محدود شود، در عمل یا اجراپذیر نخواهد بود یا به مقرراتی صوری و غیرمؤثر تبدیل می‌شود که قادر به کنترل ریسک‌های واقعی نیست. قانونگذاری هوشمندانه باید هم «اصل‌محور» و هم «قابل ممیزی» باشد؛ بدین معنا که علاوه بر تأکید بر ارزش‌های بنیادین، سازوکارهای عینی و عملی برای ارزیابی، نظارت و پاسخگویی ایجاد کند. یکی از راه‌های عملیاتی کردن این امر، تفکیک کاربردها بر اساس سطح ریسک است: برای کاربردهای پرخطر (مانند پیش‌بینی خطر، ارزیابی تکرار جرم یا ابزارهای مبتنی بر پروفایلینگ) باید الزامات سخت‌گیرانه‌تری وضع شود که شامل مستندسازی جامع داده‌ها و مدل، ارزیابی اثرات و امکان بازرسی مستقل می‌گردد. ادبیات حقوقی

سوگیری الگوریتمی نیز بر همین نکته تأکید دارد که چارچوب‌های نظارتی مناسب، معمولاً بر پایه سه رکن مستندسازی، ارزیابی و حقوق افراد در برابر تصمیمات خودکار بنا می‌شوند و «حق دریافت توضیح» و سازوکارهای پاسخگویی را به‌عنوان بخشی ضروری از حکمرانی الگوریتمی می‌دانند(lendvai & Gosztonyi,2025).

علاوه بر این، قانونگذاری باید رابطه میان داده و تبعیض را به‌صورت جدی در نظر گیرد. اگر منشأ داده‌های کیفری از ابتدا نامتوازن باشد، حتی مقررات ضدتبعیض نیز ــ بدون الزام به ممیزی داده و کنترل کیفیت کارایی محدودی خواهند داشت. مرورهای علمی نشان می‌دهند که سوگیری می‌تواند از نمونه‌گیری نامتوازن، تفاوت بین جمعیت واقعی و داده‌های ثبت‌شده، و حتی از نحوه تجمیع داده‌ها ناشی شود. بنابراین، قانون باید حداقل استانداردهای کیفیت داده، آزمون‌های نمایندگی گروه‌های مختلف و الزامات به‌روزرسانی منظم داده‌ها را مقرر کند (حسینی و همکاران،۱۴۰۲). در نهایت، قانونگذاری هوشمندانه باید «نظارت انسانیِ معنادار[۱]» را الزام‌آور سازد؛ نه صرفاً حضور صوری یک انسان، بلکه فراهم سازی اختیار و مسئولیت واقعی برای رد یا اصلاح خروجی، ثبت دلایل اتکا و ایجاد مسیری مؤثر برای شکایت و بازبینی. در اینجا منظور از نظارت انسانی معنادار صرفِ حضور فیزیکی قاضی پشت میز محاکمه یا کلیک بر روی گزینه تأیید نیست. تعریف عملیاتی این مفهوم مستلزم آن است که اولاً قاضی از «حق ابطال» برخوردار باشد و ثانیاً سیستم موظف به ارائه «تبیین‌های علی» باشد؛ یعنی قاضی باید بداند کدام متغیر ورودی دقیقاً منجر به افزایش نمره ریسک متهم شده است تا بتواند آن را با وجدان قضایی و اصول دادرسی منصفانه تطبیق دهد.شفافیت در کاربرد هوش مصنوعی در دادرسی کیفری، زمانی اهمیت واقعی پیدا می‌کند که «حق اعتراض» را از حالت تشریفاتی خارج کرده و به امکان مؤثری برای دفاع تبدیل کند. اعتراض مؤثر مستلزم آن است که متهم و وکیل او بتوانند بفهمند سامانه بر چه مبنایی به نتیجه رسیده و چه داده‌هایی در تولید آن خروجی نقش داشته‌اند. در غیر این صورت، ایراد گرفتن به خروجی، به جای نقد مبنا و منطق تصمیم، به ابراز نارضایتی کلی تقلیل می‌یابد. در همین راستا، می‌توان دست‌کم سطحی از «شفافیت داده» را تعریف کرد: باید روشن باشد که چه داده‌هایی وارد سیستم شده‌اند و این داده‌ها ــ خواه در آغاز فرآیند وارد شوند و خواه در جریان تعامل کاربر تولید شوند ــ به‌عنوان مواد اولیه تصمیم‌سازی عمل می‌کنند (حسینی و همکاران،۱۴۰۲). فقدان این سطح از شفافیت، هم احتمال تأثیر عوامل نامربوط را افزایش می‌دهد و هم امکان بررسی ادعاهای مربوط به «خطا»، «ناقص بودن داده» یا «سوگیری» را تضعیف می‌کند(حسینی،پیشین،۱۴۰۲).

از منظر تنظیم‌گری نیز، شفافیت صرفاً یک توصیه اخلاقی نیست، بلکه به حقوق مشخصی مانند حق دریافت توضیح و سازوکارهای پاسخگویی گره خورده است؛ به‌ویژه در کاربردهای پرخطر، باید مستندسازی، نظارت انسانی مؤثر و امکان بازبینی پیش‌بینی شود تا فرد بتواند تصمیم خودکار را به چالش بکشد (lendvai&Gosztonyi,2025). افزون بر این، یافته‌های تجربی نشان می‌دهند که اعتماد عمومی به تصمیم‌های الگوریتمی «زمینه‌مند» است و ادراک عدالت بسته به نوع کاربرد و نقش الگوریتم تغییر می‌کند. بنابراین، شفافیت و حق اعتراض، نه تنها برای رعایت حق دفاع، بلکه برای حفظ مشروعیت اجتماعی تصمیم قضایی نیز ضروری است(Yalcin,2023) .

---

[۱]. Meaningful Human Control

نتیجه‌گیری و پیشنهادات

چالش داده های ورودی در به‌کارگیری هوش مصنوعی در مرحلهٔ تعیین کیفر، در زمرهٔ مسائل حاشیه ای یا صرفاً فناورانه قرار نمی گیرد، بلکه مستقیماً به بنیان‌های هنجاری عدالت کیفری و مشروعیت تصمیم قضایی گره خورده است. یافته‌های این پژوهش حاکی از آن است که در نظام حقوقی ایران، «تَفرید قضایی مجازات» صرفاً یک امر فنی نیست، بلکه ریشه در مبانی فقهی و قانونی دارد. مطابق با ماده ۱۸ قانون مجازات اسلامی، قاضی موظف است کیفر را بر اساس شخصیت، انگیزه و وضعیت ذهنی مرتکب شخصی سازی کند؛ در حالی که تقلیل این واقعیت های کیفی به داده های کمی، جوهره عدالت را با مخاطره مواجه می‌سازد.

بر این اساس، مسئلهٔ داده های ورودی را باید نه صرفاً یک «ریسک فنی»، بلکه یک «خطر حقوقی» دانست که می تواند اصول بنیادینی همچون برابری اشخاص در برابر قانون، منع تبعیض و حق دفاع مؤثر را تضعیف کند. اتکای الگوریتم‌ها به داده های تاریخی، بدون پالایش انتقادی، خطر «تثبیت بی عدالتی در قالبی ظاهراً علمی» را تقویت کرده و با اصل ۳۷ قانون اساسی (اصل برائت) در تعارض قرار می گیرد؛ چرا که عدد و امتیاز ریسک، جایگزین استدلال حقوقی شده و امکان لغزش از «قضاوت دربارهٔ فعل ارتکابی» به «داوری دربارهٔ شخصیت آماری» فرد را فراهم می سازد. همچنین، چالش «جعبه سیاه» در الگوریتم ها، مستقیماً با اصل ۱۶۶ قانون اساسی و ماده ۲ قانون آیین دادرسی کیفری مبنی بر لزوم مستدل و مستند بودن احکام در تضاد است.

جهت بومی‌سازی و مدیریت این فناوری در نظام قضایی ایران، راهکارهای عملی زیر پیشنهاد می گردد:

۱. الزامات تقنینی و استانداردسازی داده‌ها:تدوین مقررات صریح دربارهٔ نحوهٔ گردآوری و اعتبارسنجی داده‌های قضایی ضرورت دارد. این مقررات باید متضمن تعیین شاخص های حداقلی کیفیت داده و الزام به مستندسازی منابع در سامانه های تعیین کیفر باشد. همچنین، برای اجرای دقیق ماده ۲۰۳ قانون آیین دادرسی کیفری، باید مکانیزمی طراحی گردد که محتوای کیفی «پرونده شخصیت» توسط یک پنل انسانی پیش‌پردازش شده و سپس به‌صورت متغیرهای توصیفی (و نه فقط عددی) در اختیار سیستم قرار گیرد تا از سوگیری علیه سوابق محیطی متهم جلوگیری شود.

۲. حفظ نقش حاکمیتی قاضی و جلوگیری از اتوماتیسم کیفری:مطابق با اصول ۱۵۶ و ۱۶۷ قانون اساسی، وظیفه قضاوت امری انحصاری و قائم به شخص قاضی است. لذا باید در آیین نامه های قضایی، ماهیت خروجی الگوریتم صرفاً «مشورتی» و در حکم «اماره قضایی» (موضوع ماده ۱۶۰ قانون مجازات اسلامی) تعریف شود. الزام قاضی به ارائه استدلال مستقل در صورت تبعیت یا عدول از توصیه الگوریتمی، تنها راه جلوگیری از تبدیل شدن قاضی به تأییدکنندهٔ بی اراده خروجی سامانه است.

۳. ممیزی نهادی و نظارت میان‌رشته‌ای:ایجاد هیئت‌های نظارتی مستقل متشکل از حقوقدانان، متخصصان علوم داده و جرم‌شناسان برای ممیزی مستمر الگوریتم‌ها الزامی است. این ممیزی باید آثار تبعیض‌آمیز احتمالی بر گروه‌های مختلف اجتماعی را سنجیده و انطباق خروجی‌ها را با اصل تناسب و فردی‌سازی مجازات بررسی کند.

۴. تضمین حقوق دفاعی و شفافیت رویه‌ای:

به منظور صیانت از اصل ۳۴ قانون اساسی (حق دادخواهی)، باید «حق بر اعتراض به تبیین های ماشینی» پیش بینی شود. متهم و وکیل وی باید حق داشته باشند از نقش و وزن الگوریتم در تعیین کیفر آگاه شده و نسبت به خطای داده یا سوگیری احتمالی، تقاضای بازنگری فنی مطرح کنند.

در نهایت، بهره گیری از داده‌های بومی و توجه به ویژگی های نهادی نظام حقوقی ایران، شرط لازم برای تضمین استفاده ای مسئولانه و منطبق با کرامت انسانی از هوش مصنوعی است. تنها در صورتی که تکنولوژی در خدمت عدالت و تحت نظارت دقیقِ هنجارهای حقوقی باشد، می‌توان از ظرفیت‌های آن برای ارتقای انسجام آرای قضایی بهره جست بدون آنکه مشروعیت نظام کیفری مخدوش گردد.

منابع

۱. منابع فارسی

کتابها

انصاری، باقر؛ و دیگران. (۱۴۰۰). حقوق داده‌ها و هوش مصنوعی: مفاهیم و چالش‌ها. تهران: سهامی انتشار، چاپ اول.

ابوذری، مهرنوش. (۱۴۰۲). حقوق و هوش مصنوعی. تهران: میزان.

جعفری‌تبار، حسن. (۱۳۹۵). دیو در شیشه؛ در فلسفه رویه قضایی. تهران: انتشارات حق‌گزاران، چاپ اول.

مرکز پژوهش‌های مجلس شورای اسلامی. (۱۴۰۲). راهکارهای کاهش سوگیری‌های شناختی به منظور ارتقای کیفیت تصمیم‌گیری. گزارش شماره ۹۱۹۴.

نجفی ابرندآبادی، علی‌حسین. (۱۴۰۰). تقریرات درس جرم‌شناسی پیشگیرانه، دانشکده حقوق دانشگاه شهید بهشتی.

هالوی، گابریل. (۱۴۰۲). مسئولیت کیفری ربات‌ها و هوش مصنوعی در قلمرو حقوق کیفری (ترجمه فرهاد شاهیده و طاهره قوانلو). تهران: میزان.

عبداللهی، محسن؛ شهریاری، عبدالنعیم؛ بابایی، یوسف؛ یعقوبی، اسماعیل. (۱۳۹۶). دیوان کیفری بین‌المللی؛ درنگی در الحاق جمهوری اسلامی ایران. تهران: انتشارات دبیرخانه مجمع تشخیص مصلحت نظام.

مقالات

عبداللهی، محسن و دیگران. (۱۳۹۶). «کاربرد الگوریتم‌های پیش‌بینی خطر در نظام عدالت کیفری». *فصلنامه مطالعات حقوق کیفری و جرم‌شناسی*، ۱۲(۲)، ۸۵–۱۱۰

حسینی، احمد؛ عبدخدائی، زهره؛ شریف‌خانی، محمد. (۱۴۰۲). کاربرد هوش مصنوعی در رسیدگی‌های قضایی: چالش شفافیت و راهکارهای آن. دیدگاه‌های حقوق قضایی، ۲۸(۱۰۱)، ۶۷–۹۰.

حسینی، احمد و همکاران. (۱۴۰۲). «شفافیت داده و الزامات دادرسی منصفانه در کاربردهای قضایی هوش مصنوعی». *فصلنامه مطالعات حقوق فناوری اطلاعات*، ۵(۱)، ۷۵–۱۰۲.

رهبری، ابراهیم؛ شعبان‌پور، علی. (۱۴۰۱). چالش‌های کاربرد هوش مصنوعی به‌عنوان قاضی در دادرسی‌های حقوقی. فصلنامه تحقیقات حقوقی (ویژه‌نامه حقوق و فناوری)، ۴۱۹–۴۴۴.

۲. منابع انگلیسی

Articles

Almasoud, A. S., & Idowu, J. A. (2025). Algorithmic fairness in predictive policing. AI Ethics, 5, 2323–2337. : https://link.springer.com/content/pdf/10.1007/s43681-024-00541-3.pdf

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. ProPublica.: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing .

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness,

Accountability and Transparency, Proceedings of Machine Learning Research, Vol. 81, pp. 77–91.: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Chiao, V. (2018). Predicting proportionality: The case for algorithmic sentencing. Criminal Justice Ethics, 37(3), 238–261.: https://utoronto.scholaris.ca/bitstreams/1179503d-907c-43bb-879c-a2a010f204c3/download

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163

European Union. (2024). Artificial Intelligence Act (Regulation (EU) 2024/1689).

Kopf, R. G. (2012). Federal sentencing error: Two-step interactive approach. Hastings Law Journal, 64, 1357–1406.

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, 14(3), 41. https://www.mdpi.com/2075-471X/14/3/41/pdf?version=1749723003

Mayson, S. G. (2019). Bias in, bias out. Yale Law Journal, 128, 2218–2473. https://yalelawjournal.org/pdf/Mayson_p5g2tz2m.pdf

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. arXiv.https://arxiv.org/pdf/1908.09635

O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown Publishing

Posner, R. A. (2010). How Judges Think. Cambridge, MA: Harvard University Press.: https://www.hup.harvard.edu/file/feeds/PDF/9780674048065_sample.pdf

Ryberg, J. (2025). Criminal sentencing and artificial intelligence: What is the input problem? Criminal Law and Philosophy, 19(2), 203–220.

Schild, U. J. (1998). Criminal sentencing and intelligent decision support. Artificial Intelligence and Law, 6(2), 151–202.

Schwarze, M., & Roberts, J. V. (2022). Reconciling artificial and human intelligence: Supplementing not supplanting the sentencing judge. In J. Ryberg & J. V. Roberts (Eds.), Sentencing and Artificial Intelligence (pp. 207–231). Oxford: Oxford University Press. https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3872766_code348473.pdf?abstractid=3872766&mirid=1

Wang, C., Han, B., Patel, B., & Rudin, C. (2022). In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. Journal of Quantitative Criminology, 39, 1–63. https://arxiv.org/pdf/2005.04176

Yalcin, G., Themeli, E., Stamhuis, E., Philipsen, S., & Puntoni, S. (2023). Perceptions of justice by algorithms. Artificial Intelligence and Law, 31(2), 269–292.