



عدالت الگوریتمی و حکمرانی هوش مصنوعی

محمد مهریار^۱، شهروز دربندی^۲، حسام الدین ترکمن^۳

تاریخ ارسال: ۱۴۰۴/۶/۶ تاریخ پذیرش: ۱۴۰۵/۲/۱۶

چکیده

گسترش روزافزون استفاده از الگوریتم‌ها و سامانه‌های هوش مصنوعی در فرایندهای تصمیم‌گیری حقوقی، اقتصادی و اجتماعی، پرسش‌های جدی درباره عدالت، شفافیت و مسئولیت‌پذیری ایجاد کرده است. عدالت الگوریتمی به معنای تضمین تصمیم‌گیری‌های عادلانه و بدون تبعیض در سیستم‌های هوش مصنوعی می‌باشد. با گسترش استفاده از الگوریتم‌ها در تصمیم‌گیری‌های حساس، نگرانی‌ها درخصوص تعصبات موجود در داده‌ها و طراحی این سیستم‌ها افزایش یافته است. سؤال اصلی مقاله این است که چگونه می‌توان عدالت الگوریتمی را در چارچوب حکمرانی هوش مصنوعی از طریق سازوکارهای حقوقی و فنی شفافیت و پاسخ‌گویی محقق ساخت؟ برای پاسخ به این پرسش، روش تحقیق توصیفی-تحلیلی به‌کار گرفته شده و داده‌ها از طریق مطالعه کتابخانه‌ای و تحلیل تطبیقی اسناد بین‌المللی، مقررات منطقه‌ای (مانند مقررات اتحادیه اروپا) و ادبیات فنی حوزه یادگیری ماشین و هوش مصنوعی گردآوری شده‌اند. یافته‌های پژوهش نشان می‌دهد که تحقق عدالت الگوریتمی مستلزم سه رکن اساسی است: نخست، توسعه فناوری‌های توضیح‌پذیر و مقاوم در برابر سوگیری داده‌ها که امکان نظارت و ارزیابی تصمیمات الگوریتمی را فراهم سازند؛ دوم، ایجاد چارچوب‌های قانونی الزام‌آور و دستورالعمل‌های نرم که بر بی‌طرفی، شفافیت و حق اعتراض کاربران تأکید دارند و سوم، تقویت نهادهای نظارتی و همکاری‌های



دوفصلنامه
بهار و تابستان ۱۴۰۵
دوره هفتم، شماره سیزدهم

۱- استادیار گروه حقوق خصوصی، دانشکده حقوق، دانشگاه ادیان و مذاهب، قم، ایران، رایانامه: m.mehryar@urd.ac.ir (نویسنده مسئول).

۲- دانشجوی دکتری حقوق بین‌الملل عمومی، دانشگاه آزاد اسلامی، قشم، ایران، رایانامه: Shahrooz_d@yahoo.com.

۳- دانشجوی دکتری حقوق خصوصی، دانشگاه ادیان و مذاهب، قم، ایران، رایانامه: T1locean@gmail.com.

بین‌المللی برای تدوین استانداردهای مشترک حکمرانی داده و الگوریتم. همچنین نتایج نشان می‌دهد که توجه به تنوع فرهنگی و اجتماعی و آموزش بازیگران حقوقی و فناورانه، از پیش شرط‌های پذیرش عمومی و اثربخشی عدالت الگوریتمی است.

واژه‌های اصلی: عدالت الگوریتمی، الگوریتم‌های توضیح‌پذیر، تعصب و تبعیض الگوریتمی، شفافیت و پاسخ‌گویی الگوریتمی، هوش مصنوعی.

مقدمه

در جهان معاصر، نفوذ روزافزون فناوری‌های نوین، به‌ویژه الگوریتم‌ها و سامانه‌های هوش مصنوعی، تحولی بنیادین در عرصه‌های مختلف زندگی بشر به‌وجود آورده است. الگوریتم‌ها، که در اصل مجموعه‌ای از دستورالعمل‌های ازپیش‌تعریف‌شده برای پردازش اطلاعات و رسیدن به نتایج مشخص‌اند (انصاری، ۱۴۰۱: ۱۴۷)، امروزه نقشی کلیدی در ساختارهای تصمیم‌گیری در نهادهای عمومی و خصوصی ایفا می‌کنند. از سیستم‌های جست‌وجو در اینترنت گرفته تا نظام‌های پیشرفته تصمیم‌یار در حوزه‌هایی نظیر خدمات درمانی، بانکی، استخدامی و حتی نهادهای قضایی، شاهد حضور پررنگ این فناوری‌ها هستیم (Russell & Norvig, 2021: 8).

در پاسخ به افزایش نگرانی‌ها نسبت به آثار حقوقی و اجتماعی این سامانه‌ها، قانون‌گذار اتحادیه اروپا با تصویب قانون هوش مصنوعی رویکردی مبتنی بر طبقه‌بندی ریسک را برای تنظیم‌گری هوش مصنوعی اتخاذ کرده است. در این چارچوب، سامانه‌های هوش مصنوعی براساس شدت تهدید بالقوه آنها برای حقوق بنیادین، ایمنی و کرامت انسانی، از کاربردهای کم‌ریسک تا سامانه‌های با خطر غیرقابل قبول و به‌ویژه هوش مصنوعی پرخطر تفکیک می‌شوند. سامانه‌های پرخطر شامل کاربردهایی هستند که در حوزه‌های حساس نظیر استخدام، آموزش، ارزیابی اعتباری، ارائه خدمات عمومی و نظام عدالت کیفری به‌کار می‌روند و به دلیل تأثیر مستقیم و عمیق بر وضعیت حقوقی و اجتماعی افراد، مجاز شمرده می‌شوند اما مشمول الزامات سخت‌گیرانه‌ای در زمینه کیفیت داده، مدیریت ریسک، شفافیت، نظارت انسانی و پاسخ‌گویی حقوقی هستند. طرح چنین طبقه‌بندی‌ای نشان‌دهنده آن است که مسئله اصلی در حکمرانی هوش مصنوعی نه صرف استفاده از الگوریتم‌ها، بلکه نحوه به‌کارگیری آنها در بسترهای پرخطر و اثرگذار بر حقوق انسان‌ها است. دقیقاً در همین نقطه، بحث عدالت الگوریتمی به‌عنوان ضرورتی هنجاری و حقوقی مطرح می‌شود. عدالت، در بنیان خود، مفهومی هنجاری است که به توزیع منصفانه حقوق، فرصت‌ها و نتایج میان افراد و گروه‌های اجتماعی اشاره دارد (Rawls, 1971: 11). عدالت الگوریتمی، به‌مثابه مفهومی نوین در تقاطع حقوق و فناوری، ناظر بر تضمین تصمیم‌هایی است که به‌واسطه سامانه‌های الگوریتمی اتخاذ می‌شوند و باید عاری از تبعیض ناعادلانه، سازگار با اصول حقوق بشر و قابل ارزیابی و اعتراض باشند.



عدالت الگوریتمی به معنای طراحی، توسعه و استقرار سامانه‌های هوش مصنوعی به گونه‌ای است که از بازتولید تعصبات ساختاری جلوگیری کرده، فرصت‌های برابر را فراهم سازد و فرایندهای تصمیم‌گیری را شفاف و پاسخ‌گو کند. این مفهوم، صرفاً بر خروجی‌های عادلانه تمرکز ندارد، بلکه بر منصفانه بودن فرایند تصمیم‌گیری، امکان نظارت مؤثر و مسئولیت‌پذیری کنشگران انسانی و نهادی نیز تأکید می‌کند. تحقق عدالت الگوریتمی مستلزم آن است که هیچ گروه اجتماعی به‌طور سیستماتیک و پایدار، تحت تأثیر ویژگی‌هایی چون جنسیت، نژاد یا طبقه اجتماعی، در معرض تبعیض الگوریتمی قرار نگیرد (Binns, 2018: 149)؛ (Mehrabi et al., 2021: 16).

اهمیت این مسئله به‌ویژه در مورد سامانه‌های هوش مصنوعی پرخطر برجسته می‌شود؛ جایی که داده‌های آموزشی سوگیرانه یا طراحی‌های غیرشفاف می‌توانند به تصمیم‌هایی منجر شوند که دسترسی افراد به شغل، آموزش، خدمات مالی یا حمایت‌های قضایی را به‌صورت نابرابر محدود کنند (انصاری، ۱۴۰۱: ۱۵۲). چنین پیامدهایی نه تنها اعتماد عمومی به سامانه‌های هوش مصنوعی را تضعیف می‌کنند، بلکه می‌توانند اصول بنیادین حقوق بشر، به‌ویژه اصل منع تبعیض و حق دادرسی منصفانه، را نقض نمایند.

بر این اساس، هدف پژوهش حاضر تحلیل چالش‌های حقوقی و فنی عدالت الگوریتمی در چارچوب حکمرانی هوش مصنوعی است و می‌کوشد نشان دهد که چگونه می‌توان، به‌ویژه در حوزه‌های مشمول هوش مصنوعی پرخطر، از طریق سازوکارهای حقوقی و فنی مبتنی بر شفافیت، توضیح‌پذیری و پاسخ‌گویی، به تحقق عدالت الگوریتمی دست یافت.

۱- چالش‌های حقوقی و فنی اجرای عدالت الگوریتمی

عدالت، به‌عنوان یکی از اصول بنیادین نظم اجتماعی، همواره در ارزیابی عملکرد نهادهای دولتی و عمومی مورد استفاده قرار گرفته است. تحقق عدالت مستلزم سازوکارهایی است که امکان اقامه دعوی، حمایت از حقوق اشخاص و اصلاح وضعیت‌های ناعادلانه را فراهم سازند. در این میان، دولت‌ها و سازمان‌ها برای ایفای مسئولیت‌های خود در جهت تضمین این اصول، ناگزیر به بهره‌گیری از ابزارهایی هستند که تحلیل دقیق و منسجم از موقعیت‌ها و مسائل گوناگون را میسر سازند. در گذشته، نبود ابزارهای فناورانه موجب می‌شد که تحقق عدالت در سطح کلان با موانع جدی روبه‌رو باشد. با این حال، ظهور فناوری‌های نوین، به‌ویژه در عرصه هوش مصنوعی و یادگیری ماشینی، افق‌های جدیدی در تحلیل داده‌ها و پشتیبانی از تصمیم‌گیری گشوده است. این فناوری‌ها با توانمندی در پردازش انبوهی از اطلاعات، امکان پیش‌بینی، ارزیابی و مداخله در فرایندهای تصمیم‌گیری را فراهم کرده‌اند؛ فرایندهایی که در گذشته صرفاً بر عهده قاضی یا تصمیم‌گیر انسانی بود.

۱-۱- تعریف و اندازه‌گیری عدالت در چارچوب الگوریتم‌ها

تعریف و تبیین عدالت در بستر سامانه‌های الگوریتمی، یکی از بنیادی‌ترین و در عین حال پیچیده‌ترین چالش‌های نظری و عملی در حوزه عدالت فناورانه است. عدالت، برخلاف یک مفهوم ایستا و یک‌بعدی، دارای لایه‌ها و ابعاد متکثری است که در ادبیات فلسفی و حقوقی، به گونه‌هایی چون عدالت توزیعی^۱، عدالت رویه‌ای^۲ و عدالت ترمیمی^۳ طبقه‌بندی می‌شود (Kamiran & Calders, 2012:9). در سطح نخست، عدالت توزیعی ناظر بر تخصیص منصفانه منابع، فرصت‌ها و مزایا است. در بستر الگوریتم‌ها، این عدالت زمانی محقق می‌شود که توزیع خروجی‌های سیستم، میان گروه‌های مختلف اجتماعی عادلانه و بدون تبعیض باشد. به عنوان نمونه، در فرایندهای استخدامی مبتنی بر یادگیری ماشین، الگوریتم‌ها باید به گونه‌ای تنظیم شوند که دسترسی به فرصت‌های شغلی برای اقلیت‌های قومی یا جنسیتی، تحت الشعاع سوگیری‌های داده‌ای قرار نگیرد (Mehrabani et al., 2021:29). همچنین، در پلتفرم‌های خدمات‌رسان آنلاین، نباید معیارهایی چون موقعیت جغرافیایی یا سابقه جست‌وجوی کاربر، منجر به ارائه ناعادلانه محصولات یا خدمات خاص فقط برای برخی کاربران گردد؛ امری که عملاً سطح آگاهی و دسترسی اقشار دیگر را محدود می‌سازد.

سطح دوم، عدالت رویه‌ای است که بر شفافیت، قابلیت پیگیری و امکان اعتراض در فرایند تصمیم‌گیری تأکید دارد. این نوع عدالت اقتضا می‌کند که منطق و ساختار تصمیمات الگوریتمی برای کاربران و ناظران قابل فهم باشد و در صورت بروز اشتباه یا تبعیض، امکان بازبینی یا مداخله وجود داشته باشد (Kamiran & Calders, 2012:12؛ Goodman & Flaxman, 2017:55). برای نمونه، در سیستم‌های ارزیابی ریسک اعتباری^۴، باید اطلاعاتی در مورد نحوه محاسبه نمرات اعتباری ارائه شود تا کاربران بتوانند عملکرد سیستم را درک کنند و در صورت نیاز به آن اعتراض کنند.

سطح سوم عدالت، عدالت تعاملی است که به کیفیت و چگونگی تعاملات میان کاربران و سیستم‌های الگوریتمی اشاره دارد و بر تضمین رفتار بی‌طرفانه، محترمانه و عاری از تبعیض این سیستم‌ها تأکید می‌کند. این مفهوم ایجاب می‌کند که سیستم‌های هوش مصنوعی در تعامل با کاربران، استانداردهای انصاف را رعایت کنند. برای مثال، در پلتفرم‌های خدمات مشتری، پاسخ‌گویی سیستم‌های مبتنی بر هوش مصنوعی باید به گونه‌ای طراحی شود که از تبعیض براساس ویژگی‌هایی نظیر



۱- برقراری عدالت از طریق رفع تبعیض و ایجاد امکانات برابر برای همه.

۲- عدالت رویه‌ای به بررسی آثار مثبت و منفی تصمیم‌های مدیران بر کارمندان آنها از نظر میزان شفاف و منصفانه بودن اجرای این تصمیمات می‌پردازد.

۳- عدالت ترمیمی فرایندی است برای درگیر کردن تا حد امکان کسانی که دارای سهمی در یک جرم خاص هستند تا به طریق جمعی نسبت به تعیین و توجه به صدمات و زیان‌ها، نیازها و تعهدات جهت التیام و بهبود بخشیدن و اصلاح امور اقدام کنند.

۴- به‌طور مثال در بانک‌ها سیستم‌های ارزیابی ریسک اعتباری ابزارهایی هستند برای ارزیابی خطر عدم بازپرداخت وام یا تسهیلات مالی توسط وام‌گیرنده یا مشتری.

سن یا وضعیت تأهل جلوگیری کند. اما تحقق این ابعاد از عدالت، مستلزم آن است که مفاهیم کلی عدالت به شاخص‌های عینی و قابل اندازه‌گیری در طراحی الگوریتم‌ها بازتعریف شود. این بازتعریف مفهومی نیازمند چارچوب‌بندی نظری و تکنیکی دقیق است. به عنوان نمونه، عدالت توزیعی می‌تواند از طریق شاخص‌هایی نظیر «سطح تساوی در دسترسی» یا «ضریب توزیع متوازن نتایج» مورد سنجش قرار گیرد، حال آنکه عدالت رویه‌ای را با معیارهایی چون «درصد تصمیمات قابل توضیح» یا «میزان رضایت کاربران از شفافیت عملکرد سیستم» می‌توان ارزیابی نمود (Kamiran & Calders, 2012:13).

از این رو می‌توان عدالت الگوریتمی را صورت‌بندی نوین مفاهیم کلاسیک عدالت در بستر فناوری‌های هوش مصنوعی و داده‌محور دانست؛ مفاهیمی که اکنون در مواجهه با الگوهای تصمیم‌گیری غیرانسانی، نیازمند بازخوانی و بازتعریف‌اند. این‌گونه از عدالت، در پیوندی مفهومی و کارکردی با سه سنخ اصلی عدالت یعنی عدالت توزیعی، عدالت رویه‌ای و عدالت ترمیمی معنا می‌یابد. بر این بنیاد، عدالت الگوریتمی نه صرفاً صورتی جدید از عدالت، بلکه بستری برای بازتعریف مفاهیم عدالت در عصر حاکمیت فناوری‌های تصمیم‌ساز است. دستیابی به چنین عدالتی، تنها در پرتو رویکردی ترکیبی ممکن است؛ رویکردی که فناوری، حقوق و اخلاق را در فرایندی هم‌افزا و نظام‌مند با یکدیگر پیوند دهد. اهمیت حقوقی این ابعاد از عدالت، به‌ویژه در پرتو قانون هوش مصنوعی اتحادیه اروپا برجسته‌تر می‌شود. مطابق بند ۳۱ مقدمه این قانون، سامانه‌های هوش مصنوعی که به امتیازدهی اجتماعی^۱ اشخاص حقیقی می‌پردازند، به سبب خطر بالای ایجاد تبعیض، طرد اجتماعی و نقض کرامت انسانی، در زمره مصادیق ریسک غیرقابل قبول قرار گرفته و به‌طور کامل ممنوع شده‌اند. این مقرر نشان می‌دهد که قانون‌گذار اروپایی، بی‌عدالتی الگوریتمی را صرفاً یک نقص فنی یا اخلاقی تلقی نمی‌کند، بلکه آن را تهدیدی مستقیم علیه ارزش‌های بنیادینی چون برابری، کرامت انسانی و منع تبعیض می‌داند. به‌ویژه، ارزیابی یا طبقه‌بندی افراد براساس داده‌هایی که خارج از بستر اولیه تولید خود به‌کار گرفته می‌شوند و منجر به رفتار نامتناسب یا ناموجه نسبت به اشخاص یا گروه‌ها می‌گردند، نمونه روشنی از نقض عدالت توزیعی و رویه‌ای در مقیاس ساختاری است.



۲-۱- نقش تعصب و تبعیض در داده‌ها و طراحی الگوریتم‌ها

یکی از مهم‌ترین چالش‌هایی که عدالت الگوریتمی با آن روبه‌رو است، وجود تعصب و تبعیض در داده‌هایی است که الگوریتم‌ها براساس آنها آموزش می‌بینند. این تعصبات، که به صورت ناخودآگاه در داده‌ها و الگوریتم‌ها بازتولید می‌شوند، اغلب بازتاب‌دهنده الگوهای ناعادلانه‌ای هستند که در طول زمان در جامعه تثبیت شده‌اند (Angwin et al., 2016:4). این بدان معنا است که الگوریتم‌ها بدون اینکه عامدانه بخواهند، نتایجی تولید می‌کنند که می‌تواند به ضرر برخی گروه‌ها تمام شود. برای مثال،

در برخی الگوریتم‌های پیش‌بینی خطر ارتکاب جرم، مشاهده شده است که افراد سیاه‌پوست بیشتر از سفیدپوستان در دسته "پرخطر" قرار می‌گیرند، حتی وقتی شرایط مشابهی دارند (O'Neil, 2016: 178). استفاده از داده‌هایی که به‌طور تاریخی ناعادلانه بوده‌اند، می‌تواند به تداوم یا تشدید تبعیضات موجود منجر شود (Weller, 2019: 25). این موضوع، نمونه‌ای روشن از تعصب در داده‌هایی است که سیستم از آنها یاد گرفته است. از این موارد می‌توان به تعصبات پنهان در داده‌های آموزشی یاد کرد (An- gwin et al., 2016: 2). داده‌هایی که بازتاب نابرابری‌های اجتماعی و تاریخی‌اند، اغلب بی‌آنکه صراحتاً تبعیض‌آمیز باشند، الگوهایی را به الگوریتم‌ها منتقل می‌کنند که ناعادلانه‌اند (Barocas & O'Neil, 2016: 146)؛ Selbst, 2016: 683.

تعصبات فقط از طریق داده‌ها منتقل نمی‌شوند، بلکه نحوه طراحی الگوریتم هم می‌تواند باعث تبعیض شود. گاهی طراحان برای افزایش سرعت یا دقت الگوریتم، معیارهایی انتخاب می‌کنند که عدالت را نادیده می‌گیرد. مثلاً اگر الگوریتم استخدامی فقط افرادی را براساس مدرک دانشگاهی خاص انتخاب کند، ممکن است به شکلی ناآگاهانه باعث تبعیض طبقاتی یا نژادی شود (Crawford et al., 2018: 35). افزون بر این، برخی از الگوریتم‌ها به قدری پیچیده‌اند که حتی متخصصان هم نمی‌دانند دقیقاً چگونه تصمیم می‌گیرند. این سیستم‌ها که به "جعبه سیاه" معروف‌اند، تحلیل عملکردشان را سخت کرده‌اند و این مسئله نظارت و کنترل بر تبعیض را دشوارتر می‌سازد (Angwin et al., 2016: 6). در مجموع، تعصب و تبعیض در داده‌ها و طراحی الگوریتم‌ها از مهم‌ترین موانع تحقق عدالت الگوریتمی محسوب می‌شوند. این پدیده‌ها، افزون بر تداوم بخشیدن به نابرابری‌های اجتماعی، می‌توانند به تضعیف اعتماد عمومی به نظام‌های فناورانه منجر شوند.

اهمیت حقوقی این چالش‌ها در قانون هوش مصنوعی اتحادیه اروپا به‌روشنی مورد توجه قرار گرفته است. مطابق ماده ۲۷ مقدمه این قانون، رویکرد مبتنی بر ریسک باید با اصول اخلاقی بنیادین هوش مصنوعی قابل اعتماد که در دستورالعمل‌های اخلاقی ۲۰۱۹ گروه عالی‌رتبه هوش مصنوعی اتحادیه اروپا تدوین شده‌اند هم‌راستا باشد. از جمله این اصول می‌توان به تنوع، عدم تبعیض و انصاف، حکمرانی داده‌ها، نظارت انسانی، شفافیت و پاسخ‌گویی اشاره کرد. این مقررۀ تأکید می‌کند که داده‌ها باید از استانداردهای بالای کیفیت و تمامیت برخوردار باشند و طراحی سامانه‌ها به‌گونه‌ای صورت گیرد که از ایجاد سوگیری‌های ناعادلانه و آثار تبعیض‌آمیز جلوگیری شود.

از منظر ماده ۲۷ مقدمه این قانون، عدالت الگوریتمی بخشی از الزامات طراحی هوش مصنوعی انسان‌محور^۲ به شمار می‌رود. به‌ویژه، اصل «تنوع، عدم تبعیض و انصاف» اقتضا می‌کند که سامانه‌های هوش مصنوعی به‌گونه‌ای توسعه یابند که مشارکت بازیگران متنوع را در نظر گرفته، دسترسی برابر را تقویت کرده و از هرگونه اثر تبعیض‌آمیز ممنوع در حقوق اتحادیه یا حقوق ملی اجتناب کنند. این الزام، پیوند



مستقیمی میان کیفیت داده‌ها، منطق طراحی الگوریتم و تحقق عدالت توزیعی و رویه‌ای برقرار می‌سازد. این ملاحظات در حوزه‌هایی مانند عدالت کیفری از اهمیت مضاعف برخوردار است. مطابق ماده ۶۱ مقدمه این قانون، سامانه‌های هوش مصنوعی که در فرایندهای قضایی ازجمله پژوهش حقوقی، تفسیر قانون و اعمال آن بر مصادیق مشخص به‌کار گرفته می‌شوند، به دلیل تأثیر بالقوه جدی بر حاکمیت قانون، آزادی‌های فردی و حق دادرسی عادلانه، در زمره سامانه‌های پرریسک^۱ طبقه‌بندی می‌شوند. این مقرر صراحتاً بر خطر تعصبات، خطاها و عدم شفافیت در چنین سامانه‌هایی تأکید می‌کند و مقرر می‌دارد که ابزارهای هوش مصنوعی صرفاً باید تصمیم‌گیری انسانی قضات را پشتیبانی کنند، نه جایگزین آن شوند.

۱-۳- شفافیت و پیچیدگی فنی در الگوریتم‌های جعبه سیاه

الگوریتم‌های پیشرفته یادگیری ماشین، به‌ویژه مدل‌های یادگیری عمیق، به دلیل پیچیدگی ساختاری، اغلب به‌عنوان «جعبه سیاه» شناخته می‌شوند. این اصطلاح به عدم امکان درک فرایند تصمیم‌گیری توسط کاربران و حتی متخصصان اشاره دارد (Weller, 2019:29). در این مدل‌ها، تصمیم‌گیری‌ها براساس روابط پیچیده میان متغیرها انجام می‌شود که به‌سختی قابل توضیح و تحلیل هستند. این ویژگی موجب می‌شود که فرایند تصمیم‌گیری برای افراد غیرمتخصص غیرشفاف و نامفهوم باشد. عدم شفافیت در این الگوریتم‌ها پیامدهایی همچون کاهش اعتماد عمومی و افزایش خطرات قانونی به همراه دارد. در سیستم‌هایی که تصمیم‌گیری‌های حیاتی مانند تخصیص منابع یا نرخ‌گذاری بیمه به الگوریتم‌های جعبه سیاه سپرده شده است، عدم شفافیت می‌تواند اعتماد عمومی را به شدت کاهش دهد (Floridi et al., 2018:5). برای مثال، اگر کاربران ندانند که چرا نرخ بیمه‌شان افزایش یافته است، این موضوع می‌تواند منجر به احساس بی‌عدالتی شود. در واکنش به گسترش استفاده از این فناوری‌ها، بخشی از ادبیات حقوقی و سیاست‌گذاری بر لزوم شفافیت و توضیح‌پذیری الگوریتمی تأکید کرده است. مدافعان این رویکرد معتقدند که الگوریتم‌ها باید نه تنها به نتایج دقیق نزدیک شوند، بلکه ملزم‌اند فرایند تصمیم‌گیری خود را نیز به‌طور روشن و کامل قابل تبیین کنند؛ مطالبه‌ای که تحت عنوان «شفافیت الگوریتمی» شناخته می‌شود (Woods, 2023:81). در همین چارچوب، برخی تحلیل‌ها مقررات عمومی حفاظت از داده‌های اتحادیه اروپا (GDPR) را حاوی نوعی «حق بر توضیح» تفسیر کرده‌اند که براساس آن، افراد حق دارند از منطق تصمیم‌گیری خودکار مطلع شوند (Goodman & Flaxman, 2017:56).

با این حال، این انتظارات هنجاری از الگوریتم‌ها اغلب فراتر از استانداردهایی است که در عمل از تصمیم‌گیران انسانی مطالبه می‌شود. انسان‌ها به‌طور روزمره تصمیم‌هایی با آثار عمیق حقوقی و اجتماعی مانند رد ویزا، صدور رأی قضایی یا عدم اعطای تسهیلات مالی اتخاذ می‌کنند، بی‌آنکه همواره ملزم به



ارائه توضیحی شفاف و قابل راستی‌آزمایی باشند. افزون بر این، توضیحات انسانی لزوماً تضمین‌کننده صداقت یا بی‌طرفی نیستند و حتی می‌توانند ابزاری برای پوشاندن تعصبات، ترجیحات شخصی یا پیش‌داوری‌های پنهان باشند (Woods, 2023: 81). برای نمونه، قضات انسانی در بسیاری از نظام‌های حقوقی مجاز به صدور آرای اجمالی بدون بیان استدلال تفصیلی هستند؛ وضعیتی که علی‌رغم انتقادات مبتنی بر اصل علنی بودن و شفافیت دموکراتیک، تاکنون به الزام عام برای ارائه استدلال مکتوب منجر نشده است.

در پاسخ به این چالش، پژوهش‌های متعددی در حوزه هوش مصنوعی توضیح‌پذیر شکل گرفته‌اند که هدف آنها افزایش قابلیت تفسیر مدل‌های جعبه سیاه است. این رویکردها شامل استفاده از مدل‌های ساده‌شده، روش‌های توضیح محلی و تکنیک‌های بصری‌سازی جهت ارائه تصویری قابل فهم از اهمیت ویژگی‌ها و منطقی تصمیم‌گیری هستند (Ribeiro, Singh, & Guestrin, 2016: 11). با این حال، این تلاش‌ها نیز با محدودیت‌هایی مواجه‌اند. مهم‌ترین چالش، تعارض میان دقت و شفافیت است؛ به‌گونه‌ای که افزایش توضیح‌پذیری اغلب مستلزم ساده‌سازی مدل و در نتیجه کاهش کارایی پیش‌بینی آن است. افزون بر این، در مدل‌های بسیار پیچیده با میلیون‌ها پارامتر، حتی پیشرفته‌ترین ابزارهای توضیح‌پذیری نیز قادر به ارائه تصویری کامل و بی‌نقص از فرایند تصمیم‌گیری نیستند (Weller, 2019: 30).

۴-۱- چالش‌های فراخوزه‌ای: تفاوت‌های قانونی و فرهنگی

یکی از موانع مهم در مسیر تحقق عدالت الگوریتمی، تفاوت‌های جدی در نظام‌های حقوقی و فرهنگی کشورهای مختلف است. این تفاوت‌ها نه تنها بر نحوه تنظیم مقررات مربوط به الگوریتم‌ها اثر می‌گذارد، بلکه در شیوه اجرا، اولویت‌ها و میزان نظارت نیز نمود می‌یابد. در ایالات متحده، چارچوب‌های حقوقی مرتبط با الگوریتم‌ها و داده‌ها عمدتاً مبتنی بر رویکرد خودتنظیمی و انعطاف‌پذیرتر هستند. اگرچه قوانینی مانند قانون عدالت الگوریتمی در آمریکا^۲ با هدف افزایش شفافیت در تصمیم‌گیری‌های خودکار طراحی شده‌اند، اما این مقررات هنوز به اندازه‌ی مقررات سخت‌گیرانه‌ای همچون مقررات عمومی حفاظت از داده‌ها در اتحادیه اروپا^۳ قدرت اجرایی و نظارت‌پذیری ندارند (Barocas, Hardt, & Narayanan, 2019: 123). در نقطه مقابل، در جمهوری خلق چین، دولت نقش مسلط و مستقیمی در کنترل، مدیریت و بهره‌برداری از داده‌ها ایفا می‌کند (Narayanan, 2019: 123; Barocas, Hardt, & Narayanan, 2019). اولویت اصلی، نه شفافیت فردی، بلکه حفظ نظم اجتماعی و امنیت عمومی است (Chen, 2020: 250). چین با تصویب مجموعه‌ای از مقررات تخصصی، چارچوبی سلسله‌مراتبی و الزام‌آور برای تضمین عدالت، امنیت و کنترل الگوریتم‌ها ایجاد کرده است. «مقررات مدیریت خدمات توصیه‌ای الگوریتمی»^۴



1- Explainable AI

2- Algorithmic Accountability Act

3- GDPR

4- Regulations on the Administration of Algorithmic Recommendation Services

(۲۰۲۲) و «مقررات مدیریت فناوری های سنتز عمیق»^۱ (۲۰۲۳) از مهم ترین اسناد این حوزه به شمار می روند.

یکی از ویژگی های شاخص رویکرد چین، تأکید بر شفافیت کارکردی و تفسیرپذیری محدود اما هدفمند است؛ بدین معنا که الزام به توضیح تصمیمات الگوریتمی ناظر بر نتایج و منطق عملیاتی است، نه افشای کامل کد منبع یا اسرار تجاری. در همین راستا، الزام به برچسب گذاری محتوای تولید شده توسط هوش مصنوعی و شناسایی محتوای ترکیب شده، ناظر بر صیانت از منافع عمومی، اعتماد اجتماعی و سلامت فضای اطلاعاتی است. در مجموع، مدل چینی عدالت الگوریتمی بر ترکیبی از مقررات الزام آور دولتی، نظارت متمرکز و تقدم منافع عمومی استوار است (Stern et al., 2021:520).

در مقابل، اتحادیه اروپا رویکردی مبتنی بر تنظیم گری سخت گیرانه، حقوق محور و مبتنی بر ریسک را در پیش گرفته است. قانون هوش مصنوعی اتحادیه اروپا با شناسایی دسته ای از کاربردهای غیر قابل قبول کوشیده است خطوط قرمز روشنی برای صیانت از کرامت انسانی، برابری و عدم تبعیض ترسیم کند. در این چارچوب، ماده ۴۴ مقدمه این قانون، به طور خاص به سامانه های هوش مصنوعی معطوف به شناسایی یا استنباط احساسات و حالات روانی افراد بر مبنای داده های زیستی^۲ می پردازد. براساس این ماده، چنین سامانه هایی به دلیل فقدان پشتوانه علمی کافی، قابلیت اعتماد محدود، فقدان تعمیم پذیری و حساسیت شدید به تفاوت های فرهنگی و موقعیتی، می توانند به نتایج تبعیض آمیز و تعرض شدید به حقوق و آزادی های بنیادین اشخاص منجر شوند. این مقرر به ویژه تأکید می کند که بیان و تجربه احساسات نه تنها در میان فرهنگ های مختلف، بلکه حتی در میان افراد یک جامعه یا یک فرد در موقعیت های متفاوت، به شدت متغیر است. از این رو، اتکای تصمیمات مهم مانند ارزیابی عملکرد، استخدام، یا سنجش صلاحیت آموزشی به چنین سامانه هایی، از منظر عدالت الگوریتمی غیر قابل توجیه است.

۲- راهکارهای بهبود عملکرد عدالت الگوریتمی

عدالت الگوریتمی در سال های اخیر به یکی از کانون های اصلی بحث در حوزه هوش مصنوعی و حقوق فناوری تبدیل شده است. همزمان با گسترش کاربرد سامانه های تصمیم گیری مبتنی بر هوش مصنوعی در حوزه های مختلف از جمله بهداشت، آموزش، حقوق و بازار کار، دغدغه های مرتبط با انصاف، تبعیض و شفافیت بیش از پیش اهمیت یافته اند. در این راستا، سه حوزه کلیدی به عنوان راهکار برای بهبود عدالت الگوریتمی قابل توجه است: بهره گیری از تکنولوژی های پیشرفته برای کاهش تبعیض، نوآوری در قوانین و مقررات برای شفافیت و پاسخ گویی بیشتر و همکاری های بین المللی و آگاهی بخشی



1- Regulations on the Administration of Deep Synthesis Internet Information Services

2- Biometric Data

برای تضمین اجرای عدالت در سطح جهانی. این سه محور، در عین ارتباط نزدیک، از نظر اهداف و ابزارها تمایز ماهوی دارند و هر یک به طور جداگانه بررسی می‌شوند.

۲-۱- تکنولوژی‌های پیشرفته برای تحقق عدالت الگوریتمی

پیشرفت فناوری‌های نوین نقش محوری در ارتقای عدالت الگوریتمی ایفا می‌کند. هرچند سیستم‌های یادگیری ماشین و تصمیم‌گیری خودکار قادر به شناسایی الگوهای پیچیده در داده‌ها هستند، اما ماهیت پیچیده برخی مدل‌ها و وجود سوگیری‌های ساختاری در داده‌های آموزشی می‌تواند به نتایج ناعادلانه بینجامد (Barocas & Selbst, 2016: 703). به‌ویژه در مدل‌های یادگیری عمیق که به دلیل ساختار پیچیده اغلب به عنوان «جعبه سیاه» عمل می‌کنند، درک منطق تصمیم‌گیری به چالشی اساسی تبدیل شده است.

این مسئله شفافیت الگوریتمی را به یکی از کلیدی‌ترین مباحث در حکمرانی هوش مصنوعی مبدل ساخته که علاوه بر ابعاد فنی، دارای پیامدهای حقوقی و اخلاقی برای توسعه‌دهندگان و کاربران این سیستم‌ها است. از سوی دیگر، پیاده‌سازی مدل‌های تفسیرپذیر با چالش‌های عملی مواجه است، چراکه نیازمند منابع محاسباتی گسترده‌ای است که ممکن است در اختیار بسیاری از سازمان‌ها و کشورها نباشد (Floridi et al., 2018: 11). این محدودیت‌ها، لزوم توسعه راهکارهای کارآمدتر برای تضمین عدالت در سیستم‌های هوشمند را بیش از پیش آشکار می‌سازد.

یکی از راه‌حل‌های ارائه شده برای غلبه بر این چالش، تصفیه یا تعدیل تعصبات در داده‌های ورودی است؛ ازجمله از طریق وزن‌دهی مجدد داده‌ها یا بهره‌گیری از داده‌های متوازن. این روش به اصلاح داده‌های ورودی به الگوریتم می‌انجامد؛ بدین معنا که قبل از اینکه الگوریتم آموزش ببیند، تعصبات موجود در داده‌ها شناسایی و حذف شوند (Kamiran & Calders, 2012: 18). راهکار دیگر ارائه شده توسعه مدل‌هایی است که به صورت ذاتی قابلیت توضیح فرایند تصمیم‌گیری خود را داشته باشند و منطق انتخاب‌هایشان را در قالبی شفاف به کاربران و ناظران ارائه دهند (Ribeiro, Singh, & Guestrin, 2016: 9). استفاده از الگوریتم‌هایی که بتوانند به صورت شفاف توضیح دهند چرا یک تصمیم خاص را گرفته‌اند به کاربران و ناظران اجازه می‌دهد منطق پشت تصمیم را درک کرده و در صورت لزوم، آنها را به چالش بکشند. در این زمینه، تکنولوژی‌هایی نظیر الگوریتم‌های توضیح‌پذیر و یادگیری ماشین قابل تفسیر، به عنوان ابزارهایی نویدبخش برای کاهش تعصبات و افزایش شفافیت مطرح شده‌اند. توضیح‌پذیری الگوریتم‌ها به معنای توانایی مدل‌ها در ارائه دلایل و فرایندهای تصمیم‌گیری به زبانی قابل فهم برای انسان‌ها است. این قابلیت، امکان شناسایی و حذف تعصبات موجود در تصمیم‌گیری‌های الگوریتمی را فراهم می‌آورد که می‌توانند نقش هر متغیر در تصمیم‌گیری الگوریتم‌ها را توضیح دهند و به کاربران اجازه دهند عملکرد مدل‌ها را بهتر درک کنند. این مدل‌ها مانند درخت‌های تصمیم‌گیری از ساختارهایی بهره می‌برند که فرایند تصمیم‌گیری آنها به راحتی قابل توضیح است. این



تکنولوژی‌ها به‌ویژه در حوزه‌هایی مانند پزشکی و قضایی که تصمیمات تأثیر قابل توجهی بر زندگی افراد دارند، می‌توانند به افزایش شفافیت و اعتماد عمومی کمک کنند (Binns, 2018:150). راهکار دیگر ارائه شده توسط تکنولوژی‌های پیشرفته، استفاده از الگوریتم‌های یادگیری فدرال^۱ است، که با استفاده از داده‌های غیرمتمرکز حریم خصوصی را حفظ می‌کند (Kearns & Roth, 2020: 188). یادگیری فدرال یک پارادایم پیشرفته یادگیری ماشین است که در آن مدل‌های هوش مصنوعی بدون انتقال داده‌های خام کاربران به سرور مرکزی امکان آموزش را دارند. امروزه این روش به‌عنوان یکی از راهکارهای کلیدی در هوش مصنوعی مسئولیت‌پذیر شناخته می‌شود. در این روش با آموزش مدل‌ها براساس داده‌های توزیع شده، بدون نیاز به انتقال داده‌های خام، امکان تحلیل غیرمتمرکز و آموزش آنها وجود دارد.

۲-۲- نوآوری در قوانین و مقررات

در کنار توسعه تکنولوژی‌های پیشرفته، نوآوری در قوانین و مقررات یکی دیگر از راهکارهای کلیدی برای بهبود عدالت الگوریتمی است. مقررات حقوقی در این زمینه نقش ایجابی ایفا می‌کنند. قوانین و مقررات می‌توانند چارچوبی برای نظارت، شفافیت و کاهش تبعیض در استفاده از الگوریتم‌ها فراهم کنند. یکی از برجسته‌ترین مثال‌ها در این زمینه، مقررات عمومی حفاظت از داده‌ها در اتحادیه اروپا است. این مقررات، سازمان‌ها را ملزم می‌کند که فرایندهای تصمیم‌گیری خودکار را شفاف‌سازی کرده و دلایل تصمیمات الگوریتمی را به کاربران توضیح دهند. ایجاد چنین چارچوب‌های نظارتی شفاف می‌تواند نقش مهمی در کاهش تبعیض‌های الگوریتمی ایفا کند (Goodman & Flaxman, 2017:56). این نوع مقررات، الگویی برای دیگر کشورها فراهم کرده است که با الزام به پاسخگویی و شفافیت، توسعه‌دهندگان را ملزم می‌کنند که تأثیرات اجتماعی سیستم‌های خود را به‌طور جدی مورد توجه قرار دهند. همچنین از دستورالعمل‌های اخلاقی اتحادیه اروپا در زمینه هوش مصنوعی^۲ می‌توان به‌عنوان نمونه‌ای از «حقوق نرم»^۳ یاد کرد که نقش پیشرو در زمینه هوش مصنوعی دارد.

با این حال، تحول اساسی در این حوزه با تصویب قانون هوش مصنوعی اتحادیه اروپا^۴ رقم خورده است؛ قانونی که نخستین چارچوب الزام‌آور و جامع جهانی برای حکمرانی هوش مصنوعی محسوب می‌شود. ویژگی متمایز این قانون، اتخاذ رویکرد مبتنی بر ریسک است که سامانه‌های هوش مصنوعی را براساس شدت تهدید آنها برای حقوق بنیادین و منافع عمومی طبقه‌بندی می‌کند. در این چارچوب، برخی کاربردها به‌عنوان دارای «خطر غیرقابل قبول» به‌طور کلی ممنوع شده‌اند (مانند سامانه‌های امتیازدهی اجتماعی)، در حالی که سامانه‌های پرخطر ازجمله کاربردهای هوش مصنوعی در استخدام، آموزش، اعتبارسنجی، خدمات عمومی و نظام قضایی – مجاز شمرده شده‌اند، اما تحت الزامات



سخت‌گیرانه‌ای در زمینه کیفیت داده، مدیریت ریسک، نظارت انسانی، شفافیت و ثبت‌پذیری قرار گرفته‌اند. این رویکرد تنظیمی نشان می‌دهد که قانون‌گذار اروپایی، عدالت الگوریتمی را نه در قالب یک اصل انتزاعی، بلکه به عنوان معیاری عملی برای محدودسازی یا مشروط‌سازی استفاده از فناوری در حوزه‌های حساس تلقی کرده است.

در تکمیل راهکارهای فنی، سیاست‌گذاری و مقررات‌گذاری هوشمند نقشی تعیین‌کننده در تبدیل نوآوری‌های فناورانه به نتایج عادلانه و قابل اجرا ایفا می‌کند. توسعه الگوریتم‌های توضیح‌پذیر یا تکنیک‌های تعدیل سوگیری، بدون وجود سازوکارهای حاکمیتی و مشوق‌های حقوقی-اقتصادی مناسب، برای تحقق عدالت الگوریتمی کفایت نمی‌کند. آنچه تضمین‌کننده اجرای مؤثر این فناوری‌هاست، پیوند نظام‌مند میان نوآوری فنی و ابزارهای الزام‌آور حقوقی است.

در این چارچوب، خریده‌های دولتی^۱ به عنوان یکی از ابزارهای بالفعل حکمرانی، ظرفیت ویژه‌ای برای ارتقای عدالت و شفافیت الگوریتمی دارد. دولت‌ها، در مقام بزرگ‌ترین خریداران محصولات و خدمات فناورانه، می‌توانند از قدرت بازار خود برای الزام عرضه‌کنندگان به رعایت معیارهای عدالت، شفافیت و پاسخ‌گویی استفاده کنند. درج شروط مشخص در قراردادهای تأمین سامانه‌های هوش مصنوعی، امکان تنظیم سطح الزامات را متناسب با نوع فناوری، میزان ریسک و شدت تأثیر آن بر حقوق بنیادین افراد فراهم می‌سازد.

مزیت اصلی این رویکرد، انعطاف‌پذیری و پرهیز از یکسان‌سازی نامتناسب است؛ به گونه‌ای که سامانه‌های پرریسک، مانند الگوریتم‌های مورد استفاده در نظام قضایی یا امنیتی، مشمول الزامات سخت‌گیرانه‌تر افشا و حسابرسی شوند، درحالی‌که در کاربردهای کم‌خطرتر، سطح شفافیت به صورت متناسب تنظیم گردد. افزون بر این، این سازوکار با تنوع مدل‌های کسب‌وکار و ضرورت حفاظت از اسرار تجاری نیز سازگار است و امکان افشای کنترل‌شده اطلاعات در چارچوب‌های قراردادی امن را فراهم می‌آورد (Rowe & Prior, 2022:309).

در سطح ملی، نمونه‌هایی از نوآوری تقنینی نیز قابل مشاهده است. در ایالات متحده، قانون عدالت الگوریتمی^۲ نمونه‌ای از نوآوری‌های قانونی است که به طور خاص بر شناسایی و کاهش تبعیض در سیستم‌های الگوریتمی تمرکز دارد.^۳ این قانون شرکت‌ها را ملزم می‌کند که تأثیرات اجتماعی، اخلاقی و قانونی الگوریتم‌های خود را ارزیابی کنند و برای کاهش هرگونه تبعیض تلاش کنند. سنگاپور نیز قانون حفاظت از داده‌های شخصی (PDPA) را برای تضمین حفاظت از داده‌ها و حریم خصوصی افراد به تصویب رسانده است. این قانون تأکید دارد که جمع‌آوری و استفاده از داده‌های شخصی باید به گونه‌ای باشد که به حقوق و آزادی‌های افراد احترام گذاشته شود. قانون حفاظت از داده‌های شخصی همچنین



1- Public Procurement

2- Algorithmic Accountability Act

3- US Congress. (2019). Algorithmic Accountability Act of 2019. H.R. 2231.

شامل مقرراتی برای شفافیت و مسئولیت پذیری در استفاده از الگوریتم ها و هوش مصنوعی است. در سطح بین المللی، تلاش هایی برای هماهنگ سازی مقررات صورت گرفته است. علاوه بر مقررات الزام آور، دستورالعمل های اخلاقی غیرالزام آور نیز نقش مهمی در پیشبرد عدالت الگوریتمی داشته اند. برای مثال، اصول اخلاقی هوش مصنوعی سازمان همکاری و توسعه اقتصادی^۱ که در سال ۲۰۱۹ به تصویب رسید توانست چارچوبی جامع و بین المللی برای توسعه و به کارگیری مسئولانه فناوری هوش مصنوعی ارائه کند. این اصول با ایجاد زبان مشترکی میان کشورهای عضو و غیرعضو سازمان همکاری و توسعه اقتصادی، نقش به سزایی در کاهش ناهماهنگی های قانونی و تنظیم گری در سطح جهانی ایفا نموده است. جالب توجه اینکه تاکنون ۴۲ کشور - فراتر از اعضای سازمان همکاری و توسعه اقتصادی - این چارچوب را به عنوان مبنای سیاست گذاری های ملی خود در حوزه هوش مصنوعی برگزیده اند که نشان دهنده پذیرش گسترده و تأثیرگذاری عملی این استانداردهای اخلاقی است.

تحلیل محتوایی این اصول نشان می دهد که رویکرد سازمان همکاری و توسعه اقتصادی ترکیبی متوازن از ملاحظات فنی، حقوقی و اخلاقی را در برمی گیرد. در سطح فنی، تأکید بر شفافیت و قابلیت تفسیر سیستم های هوش مصنوعی؛ در بعد حقوقی، توجه به مسئولیت پذیری و پاسخ گویی و در سطح اخلاقی، التزام به ارزش های انسانی و عدالت، از ویژگی های بارز این چارچوب محسوب می شوند. این جامعیت نظری، این اصول را به یکی از تأثیرگذارترین اسناد در حکمرانی جهانی هوش مصنوعی تبدیل کرده است. همچنین دستورالعمل های مؤسسه مهندسان برق و الکترونیک^۲ برای سیستم های هوش مصنوعی قابل اعتماد، بر طراحی انسان محور و ارزیابی تأثیرات اجتماعی تأکید دارند.

در هر صورت به نظر می رسد ارائه راهکار برای اجرای عدالت در چارچوب الگوریتم ها نیازمند رویکردهای چندجانبه است. تحقق عدالت الگوریتمی نیازمند تلفیق شاخص های دقیق حقوقی، ابزارهای فنی پیشرفته و همکاری بین رشته ای است تا ضمن رعایت ملاحظات اخلاقی، از پیامدهای منفی الگوریتم ها بر جامعه جلوگیری شود (Kamiran & Calders, 2012:17). در مجموع، حکمرانی جهانی هوش مصنوعی تنها زمانی مؤثر خواهد بود که نقش یک جانبه دولت ها یا شرکت ها کنار گذاشته شود و مطالبات جنبش های اجتماعی، کارگران فناوری و جامعه مدنی به مثابه عناصر حیاتی پاسخ گویی دموکراتیک به رسمیت شناخته شوند (Bloch-Wehba, 2022:75). برای ارائه راهکار دقیق تدوین چارچوب های منعطف و قابل تطبیق با نیازهای محلی ضروری است.

۲-۳- همکاری های بین المللی و توسعه آموزش و آگاهی بخشی به متخصصان و عموم مردم

یکی از مؤثرترین و در عین حال مغفول مانده ترین ابزارها برای ارتقای عدالت الگوریتمی، سرمایه گذاری هدفمند در آموزش تخصصی و گسترش آگاهی عمومی است. عدالت الگوریتمی صرفاً از رهگذر نوآوری



فنی یا تصویب قواعد حقوقی الزام آور محقق نمی شود، بلکه مستلزم افزایش سطح دانش، حساسیت و توان انتقادی نسبت به تبعیض ها و بی عدالتی های پنهان در سامانه های هوشمند است؛ امری که هم توسعه دهندگان و بهره برداران فنی و هم کاربران عادی را دربرمی گیرد.

در این راستا، قانون هوش مصنوعی اتحادیه اروپا رویکردی معنادار اتخاذ کرده و آموزش و شایستگی انسانی را به عنوان یکی از پیش شرط های استفاده مشروع از سامانه های هوش مصنوعی، به ویژه سامانه های پرخطر به رسمیت شناخته است. مطابق این قانون، بهره برداران سامانه های پرخطر مکلف اند تدابیر فنی و سازمانی مقتضی اتخاذ کنند تا این سامانه ها مطابق دستورالعمل های استفاده طراحی شده از سوی ارائه دهندگان به کار گرفته شوند. علاوه بر این، الزام به نظارت انسانی مؤثر مستلزم آن است که اشخاص ناظر، از صلاحیت، آموزش و اختیار کافی برخوردار باشند؛ موضوعی که نشان می دهد عدالت الگوریتمی بدون سرمایه گذاری در ظرفیت انسانی قابل تحقق نخواهد بود. آموزش تخصصی توسعه دهندگان هوش مصنوعی نقشی کلیدی در پیشگیری از تعصبات الگوریتمی دارد. چنانچه مهندسان و طراحان سامانه های هوشمند، از مرحله گردآوری داده تا طراحی مدل و استقرار سیستم، با اصول عدالت، شفافیت و عدم تبعیض آشنا باشند، احتمال تولید سامانه های منصفانه تر به طور معناداری افزایش می یابد (Floridi et al., 2018:4). این آموزش ها باید فراتر از مهارت های صرفاً فنی بوده و شامل شناسایی تعصب در داده های آموزشی، استفاده از الگوریتم های توضیح پذیر و ارزیابی پیامدهای اجتماعی و حقوقی تصمیم گیری های الگوریتمی نیز باشد.

در همین چارچوب، ماده ۴ قانون هوش مصنوعی اتحادیه اروپا مفهوم سواد هوش مصنوعی^۱ را به عنوان یک تعهد صریح برعهده ارائه دهندگان و بهره برداران سامانه های هوش مصنوعی قرار داده است. براساس این رویکرد، نهادهای مسئول موظف اند، تا حد امکان، سطح مناسبی از سواد هوش مصنوعی را در میان کارکنان و سایر اشخاصی که با بهره برداری یا کاربری این سامانه ها در ارتباط اند تضمین کنند. این الزام با توجه به سطح دانش فنی، تجربه، آموزش و زمینه کاربرد سامانه تنظیم می شود و همچنین آثار احتمالی سامانه ها بر افراد یا گروه های هدف را مد نظر قرار می دهد.

در کنار آموزش تخصصی برای توسعه دهندگان، ارتقای سواد الگوریتمی در سطح عمومی از ضرورت های اساسی تحقق عدالت الگوریتمی محسوب می شود. مطالعات نشان می دهد که بخش قابل توجهی از کاربران سیستم های هوشمند، درکی ناکافی از مکانیسم های تصمیم گیری الگوریتمی و پیامدهای اجتماعی آن دارند (al., 2018:34; Crawford et). این شکاف ممکن است به پذیرش غیر واقع گرایانه نتایج الگوریتمی یا مقاومت بی اساس در برابر فناوری های نوین منجر شود.

تجربه های موفق بین المللی نشان می دهد که ترکیب این راهکارها نه تنها موجب افزایش آگاهی عمومی می شود، بلکه بستری برای شکل گیری انتظارات اجتماعی از توسعه دهندگان و سیاست گذاران



ایجاد می‌کند (al., 2018: 15). در این زمینه، استفاده از مدل‌های توضیح‌پذیر^۱ می‌تواند به عنوان پلی میان پیچیدگی‌های فنی و درک عمومی عمل کند، چراکه امکان ارائه توضیحات قابل فهم برای کاربران غیرمتخصص را فراهم می‌سازد.

یکی از اهداف اصلی آگاهی بخشی عمومی، ارتقای توان نظارت اجتماعی بر فناوری‌های نوین است. زمانی که کاربران بدانند چگونه عملکرد یک الگوریتم می‌تواند منجر به تبعیض شود، می‌توانند مطالبه شفافیت و پاسخ‌گویی بیشتری از نهادهای توسعه‌دهنده و بهره‌بردار داشته باشند. این مطالبه‌گری عمومی، در کنار نظارت سازمانی، یک ضامن مهم برای حفظ عدالت در فرایندهای خودکار است (Pasquale, 2015: 292).

۳- ملاحظات اخلاقی و پیامدهای حقوقی در عدالت الگوریتمی

هوش مصنوعی و الگوریتم‌های یادگیری ماشین با تمام مزایایی که برای بهبود کارایی و دقت تصمیم‌گیری فراهم کرده‌اند، چالش‌های اخلاقی و پیامدهای حقوقی بسیاری به همراه داشته‌اند. در این میان، سه محور اصلی در ملاحظات اخلاقی و حقوقی مرتبط با عدالت الگوریتمی شامل تعادل میان عدالت و کارایی، حفظ حقوق بشر و حریم خصوصی و پیامدهای بلندمدت به عنوان موضوعات کلیدی مورد بحث قرار می‌گیرند. این مسائل، با تأثیرگذاری بر تصمیم‌گیری‌ها در حوزه‌های قضایی، پزشکی، اقتصادی و اجتماعی، ابعاد مختلفی از زندگی انسانی را تحت تأثیر قرار می‌دهند.

دوفصلنامه
بهار و تابستان ۱۴۰۵
دوره هفتم، شماره سیزدهم



۳-۱- تعادل میان عدالت و کارایی

یکی از مهم‌ترین مسائل در عدالت الگوریتمی، ایجاد تعادل میان عدالت و کارایی است. الگوریتم‌های هوش مصنوعی اغلب با هدف بهینه‌سازی کارایی طراحی می‌شوند، اما این بهینه‌سازی می‌تواند به قیمت کاهش عدالت یا افزایش تبعیض تمام شود. کارایی به معنای بهبود سرعت و دقت تصمیم‌گیری‌ها، اغلب اولویت سازمان‌ها است، اما این رویکرد ممکن است معیارهای عدالت توزیعی و رویه‌ای را نادیده بگیرد (Binns, 2018: 151). برای نمونه، در طراحی الگوریتم‌های استخدامی، معمولاً از داده‌های تاریخی استفاده می‌شود که اگر حاوی سوگیری‌های نهادینه شده باشند، الگوریتم ممکن است در جهت بهینه‌سازی گروه‌هایی از متقاضیان را به طور ناعادلانه نادیده بگیرد. مثلاً تمرکز هوش مصنوعی بر روی فارغ‌التحصیلان سال‌های اخیر می‌تواند موجب حذف فرصت‌های شغلی برای افراد با سابقه تحصیلی قدیمی‌تر شود. این تصمیم هرچند موجب افزایش سرعت پردازش و کارایی می‌شود، اما به قیمت نقض عدالت فرصت‌ها تمام خواهد شد. این وضعیت، نمونه‌ای از تعارض میان اهداف کارایی و عدالت است. برای کاهش این تعارض، بهره‌گیری از الگوریتم‌های چندمعیاری که به طور

هم‌زمان اهداف کارایی و عدالت را مدنظر دارند، توصیه می‌شود (Selbst, 2016: 709) (Barocas & این الگوریتم‌ها با استفاده از تکنیک‌هایی چون بهینه‌سازی ترکیبی تلاش می‌کنند تعادلی پویا میان اهداف فنی و ملاحظات اخلاقی برقرار کنند. این روش‌ها، الگوریتم‌ها را قادر می‌سازند تا بدون قربانی کردن یکی از معیارها، بهبودهای قابل توجهی در هر دو زمینه ایجاد کنند. پژوهش‌ها نشان داده‌اند که در سیستم‌های قضایی، بهره‌گیری از الگوریتم‌های شفاف، قابل توضیح و عادلانه می‌تواند هم دقت تصمیم‌گیری را افزایش دهد و هم از تبعیض در پیش‌بینی خطر یا تعیین مجازات جلوگیری کند (Kearns & Roth, 2020: 192).

نکته اساسی دیگر، ضرورت نظارت مستمر بر عملکرد الگوریتم‌ها است. حتی الگوریتم‌هایی که در مرحله طراحی متعادل به نظر می‌رسند، ممکن است با گذشت زمان و تغییر داده‌های ورودی دچار سوگیری مجدد شوند. از این رو، باید به‌طور منظم عملکرد آنها ارزیابی، تعصبات احتمالی شناسایی و با داده‌های جدید بازآموزی شوند (Floridi et al., 2018: 11).

در حوزه‌های حساسی چون عدالت کیفری، این مسئله اهمیت مضاعف دارد. یکی از چالش‌های رایج در فرایند دادرسی، مدیریت حجم گسترده‌ای از داده‌ها و اسناد حقوقی است که بررسی آنها به‌صورت دستی مستعد خطا، تأخیر و تصمیمات ناعادلانه است. در چنین شرایطی، طراحی و به‌کارگیری الگوریتم‌های عادلانه می‌تواند در تحلیل دقیق شواهد، دسته‌بندی اطلاعات و پیشنهاد تصمیم‌های اولیه، کمک مؤثری به قضات و وکلای مدافع ارائه دهد، مشروط بر آنکه این ابزارها تحت نظارت حقوقی و اخلاقی دقیق توسعه یابند.

در مجموع، دستیابی به الگوریتم‌هایی که در عین برخورداری از کارایی فنی بالا، به اصول بنیادین عدالت نیز پایبند باشند، هدفی ضروری و قابل تحقق است؛ مشروط بر آنکه طراحی، ارزیابی و استفاده از این سامانه‌ها با مشارکت میان‌رشته‌ای حقوق دانان، متخصصان علوم کامپیوتر و نهادهای ناظر صورت گیرد.

۳-۲- حفظ حقوق بشر و حریم خصوصی

حفظ حقوق بشر و حریم خصوصی یکی از ارکان بنیادین در تحلیل‌های اخلاقی و حقوقی مرتبط با عدالت الگوریتمی به‌شمار می‌آید. الگوریتم‌هایی که برای تحلیل و تصمیم‌گیری در حوزه‌های حساس طراحی می‌شوند، غالباً نیازمند حجم عظیمی از داده‌های فردی‌اند؛ داده‌هایی که در صورت سوءاستفاده یا استفاده بدون رضایت، می‌توانند منجر به نقض آشکار حقوق بنیادین انسان‌ها شوند.

در بسیاری از کاربردهای الگوریتمی، به‌ویژه در حوزه‌های بازاریابی دیجیتال، سلامت، آموزش و نظام قضایی، داده‌های حساس افراد جمع‌آوری، ذخیره و تحلیل می‌شوند. به‌عنوان نمونه، الگوریتم‌های پیش‌بینی رفتار مصرف‌کننده از ردپاهای دیجیتال کاربران، شامل تاریخچه مرور، خرید، جست‌وجو و حتی موقعیت جغرافیایی برای ارائه پیشنهادات شخصی‌سازی شده استفاده می‌کنند. این نوع استفاده



از داده‌ها، در صورت عدم شفافیت یا فقدان رضایت آگاهانه، با اصول بنیادین حریم خصوصی در تعارض قرار می‌گیرد (Zuboff, 2019: 634).

در حوزه پزشکی نیز نمونه‌های روشنی از این چالش وجود دارد. استفاده از داده‌های بیماران برای آموزش الگوریتم‌های تشخیص بیماری، اگرچه می‌تواند به بهبود دقت و سرعت خدمات درمانی منجر شود، اما در صورتی که بدون رضایت صریح یا بدون اطلاع‌رسانی کافی انجام گیرد، نقض صریح کرامت انسانی و حقوق فردی محسوب می‌شود (Floridi et al., 2018: 17).

قانون هوش مصنوعی اتحادیه اروپا این دغدغه را به صورت نظام‌مند در چارچوب رویکرد مبتنی بر ریسک خود منعکس کرده است. ماده ۲۷ مقدمه این قانون، با یادآوری دستورالعمل‌های اخلاقی غیرالزام‌آور «هوش مصنوعی قابل اعتماد» مصوب ۲۰۱۹، هفت اصل محوری از جمله حفظ حریم خصوصی و حکمرانی داده‌ها، شفافیت، نظارت انسانی، تنوع، عدم تبعیض و پاسخ‌گویی را به عنوان مبانی هنجاری مشترک برای تنظیم‌گری هوش مصنوعی برجسته می‌کند. این پیوند میان اخلاق و تنظیم‌گری الزام‌آور، نشان‌دهنده آن است که حفاظت از حقوق بشر در حوزه هوش مصنوعی، صرفاً توصیه‌ای اخلاقی نیست، بلکه مبنای شکل‌گیری قواعد حقوقی سخت نیز قرار گرفته است. در همین راستا ماده ۶۹ مقدمه این قانون به صراحت تأکید می‌کند که حق حریم خصوصی و حفاظت از داده‌های شخصی باید در سراسر چرخه عمر سامانه‌های هوش مصنوعی تضمین شود. قانون‌گذار اروپایی در این زمینه، نه تنها به ابزارهایی مانند ناشناس‌سازی و رمزنگاری اشاره می‌کند، بلکه استفاده از فناوری‌هایی نظیر یادگیری فدرال را نیز به عنوان راهکاری مشروع برای آموزش الگوریتم‌ها بدون انتقال یا کپی‌برداری از داده‌های خام مورد تأکید قرار می‌دهد.

از منظر فنی همان‌طور که سابقاً بیان شده توسعه روش‌هایی مانند یادگیری فدرال^۱ و رمزنگاری پیشرفته^۲ گام‌های مؤثری در جهت حفظ حریم خصوصی در فرایند آموزش الگوریتم‌ها بوده‌اند. یادگیری فدرال این امکان را فراهم می‌سازد که داده‌های کاربران به جای ارسال به سرور مرکزی، در محل تولید داده مورد استفاده قرار گیرند و تنها به روزرسانی مدل‌ها منتقل شود. این رویکرد علاوه بر ارتقای حریم خصوصی، به کاهش خطر نشت داده و سوءاستفاده‌های ثانویه نیز کمک می‌کند.

در کنار حفاظت از اطلاعات شخصی، مفهوم حقوق بشر در بستر عدالت الگوریتمی، دربرگیرنده الزامات دیگری نیز هست. از جمله آنها می‌توان به لزوم جلوگیری از تبعیض در نتایج الگوریتم‌ها اشاره



۱- یادگیری فدرالی به شبکه‌های توزیع شده و غیرمتمرکز کمک می‌کند تا به جای ارسال داده‌های محلی و سری خود، ماشین یادگیرنده را به صورت محلی آموزش دهند و پارامترهای مدل را با یکدیگر به اشتراک بگذارند. این روش، یادگیری مشارکتی نیز نامیده می‌شود. این روش برخلاف تکنیک‌های سنتی یادگیری ماشین، متمرکز است که همه مجموعه داده‌های محلی در یک سرور بارگذاری می‌شوند.

۲- این تکنیک به طور گسترده‌ای برای رمزنگاری و افزایش امنیت اطلاعات در انتقال داده یا در رمزنگاری اطلاعات طبقه‌بندی شده توسط سازمان‌ها به کار گرفته می‌شود.

کرد. اگر الگوریتمی، بر پایه داده‌های تاریخی حاوی تعصبات نژادی، جنسیتی یا طبقاتی آموزش ببیند، احتمال بازتولید این تبعیض‌ها در خروجی‌های آن وجود دارد. به‌ویژه در سیستم‌های قضایی، استخدامی یا اعطای وام، چنین وضعیتی می‌تواند مصداق آشکار نقض حق برابری، حق دسترسی به فرصت‌ها و کرامت انسانی باشد. از این رو، طراحی عادلانه الگوریتم‌ها باید متضمن رعایت اصول عدالت توزیعی (برابری در نتایج) و عدالت رویه‌ای (شفافیت و امکان اعتراض به فرایند) باشد. این الزام، نه تنها یک دغدغه اخلاقی بلکه یک ضرورت حقوقی در چارچوب اسناد بین‌المللی حقوق بشر و استانداردهای جهانی داده‌محور به‌شمار می‌آید.

۳-۳- پیامدهای بلندمدت

کاربرد روزافزون الگوریتم‌های هوش مصنوعی در ساختارهای تصمیم‌گیری نه تنها آثار کوتاه‌مدت عملیاتی دارد، بلکه پیامدهای بلندمدتی نیز بر پیکره نظم اجتماعی، اقتصادی و حقوقی برجای می‌گذارد. این پیامدها از جنس تحول در نقش انسان، جابه‌جایی قدرت تصمیم‌گیری، گسترش مسئولیت‌های حقوقی و تغییرات عمیق فرهنگی و نهادی هستند؛ تحولات بنیادینی که بی‌توجهی به آنها می‌تواند بنیان‌های عدالت را در جوامع دستخوش بحران کند (Binns, 2018: 151).

در بعد اقتصادی، اتکای روزافزون به الگوریتم‌های مالی برای تخصیص منابع، اعطای اعتبار، یا تعیین نرخ بهره می‌تواند نابرابری‌های ساختاری را تشدید کند. اگر این سامانه‌ها بر داده‌هایی آموزش ببینند که بازتاب تبعیض‌های تاریخی یا طبقاتی‌اند، آنگاه خروجی آنها نیز واجد همان سوگیری‌ها خواهد بود. به عنوان نمونه، ممکن است افراد با سابقه مالی محدود یا اقلیت‌های قومی، به‌طور سیستماتیک از دریافت وام یا نرخ بهره مناسب محروم شوند؛ امری که می‌تواند به حاشیه‌راندگی بیشتر گروه‌های آسیب‌پذیر منجر شود (Selbst, 2016: 701 & Barocas).

در سطح اجتماعی، الگوریتم‌ها به‌ویژه از طریق پلتفرم‌های رسانه‌ای، نقش فعالی در شکل‌دهی به ترجیحات، نگرش‌ها و تعاملات انسانی یافته‌اند. الگوریتم‌های توصیه‌گر در شبکه‌های اجتماعی با تقویت اطلاعات هم‌سو با باورهای پیشینی کاربران، پدیده اتاق پژواک^۱ را تشدید کرده به گونه‌ای که افراد عمدتاً در معرض دیدگاه‌ها، اطلاعات و باورهایی قرار می‌گیرند که با دیدگاه‌های قبلی خودشان هماهنگ است و دیدگاه‌های مخالف یا متنوع به ندرت وارد آن فضا می‌شوند یا نادیده گرفته می‌شوند و از این رهگذر زمینه‌ساز قطبی‌سازی افکار عمومی می‌شوند (Crawford et al., 2018: 23). این وضعیت، علاوه بر تهدید انسجام اجتماعی، می‌تواند تهدیدی برای گفت‌وگو، عدالت اجتماعی و مشارکت برابر شهروندان باشد.

در حوزه حقوقی، استفاده از الگوریتم‌ها پرسش‌های بنیادینی درباره مسئولیت حقوقی، پاسخ‌گویی



و جبران خسارت ایجاد می‌کند. در شرایطی که یک تصمیم ناعادلانه یا تبعیض‌آمیز توسط یک الگوریتم اتخاذ شده باشد، تعیین اینکه مسئولیت متوجه توسعه‌دهنده، نهاد بهره‌بردار یا تأمین‌کننده داده است، نیازمند تبیین دقیق حقوقی است. این امر مستلزم بازنگری در نظریه‌های سنتی مسئولیت مدنی و کیفری و نیز تدوین قوانین جدید در زمینه مسئولیت‌پذیری الگوریتم‌ها است (Goodman & Flaxman, 2017:55). درنهایت، پیامدهای بلندمدت استفاده گسترده از الگوریتم‌ها با توجه به سرعت گسترش آنها بسیار فراتر از مسائلی است که اکنون بشر با آن مواجه است. این فناوری‌ها می‌توانند به تدریج بر عدالت اجتماعی، آزادی‌های فردی و نحوه اداره جوامع اثر بگذارند.

نتیجه‌گیری

عدالت الگوریتمی، به مثابه یکی از گونه‌های نوپدید عدالت در عصر داده‌ها، به مجموعه اصول و سازوکارهایی اطلاق می‌شود که با هدف تضمین انصاف، بی‌طرفی، شفافیت و پاسخ‌گویی در تصمیم‌گیری‌های الگوریتمی طراحی می‌شود. این مفهوم در پاسخ به نگرانی‌های فزاینده نسبت به تأثیرات تبعیض‌آمیز الگوریتم‌ها بر حقوق افراد، به ویژه در حوزه‌هایی مانند قضا، سلامت، اشتغال و خدمات عمومی شکل گرفته است. عدالت الگوریتمی درواقع برآمده از بازخوانی مفاهیم کلاسیک عدالت در بستر فناوری است و با انواع شناخته‌شده‌تری چون عدالت توزیعی، عدالت رویه‌ای و عدالت ترمیمی نسبت مفهومی و کاربردی دارد. عدالت توزیعی، که بر توزیع منصفانه منابع و فرصت‌ها تأکید دارد، در قالب الگوریتمی به معنای توزیع عادلانه نتایج سیستم‌های خودکار میان گروه‌های مختلف اجتماعی است. عدالت رویه‌ای که بر انصاف در فرایند تصمیم‌گیری متمرکز است، در حوزه الگوریتم‌ها ایجاب می‌کند که فرایندها شفاف، قابل توضیح و قابل اعتراض باشند. همچنین، عدالت ترمیمی که هدف آن جبران خسارات ناشی از بی‌عدالتی یا تبعیض است، در عرصه عدالت الگوریتمی می‌تواند در قالب اصلاح داده‌های تبعیض‌آمیز، بازنگری مدل‌های تصمیم‌گیر و جبران آثار ناعادلانه تصمیمات گذشته جلوه‌گر شود. بنابراین، عدالت الگوریتمی نه تنها نوعی جدید از عدالت است، بلکه بستری برای بازتعریف عدالت کلاسیک در عصر هوش مصنوعی به‌شمار می‌رود. این مقاله بر آن است که نشان دهد تحقق چنین عدالتی، نیازمند رویکردی جامع و میان‌رشته‌ای است که فناوری را با حقوق، اخلاق و سیاست‌گذاری به‌گونه‌ای یکپارچه پیوند دهد.

منشأ اصلی بی‌عدالتی در سیستم‌های الگوریتمی را می‌توان در دو سطح بررسی کرد: داده‌ها و نحوه طراحی الگوریتم. در سطح داده، مشکل اصلی آنجاست که تعصب‌ها و نابرابری‌های تاریخی که در داده‌های آموزشی وجود دارد، بدون اینکه لزوماً آگاهانه باشد، به الگوریتم منتقل می‌شود و تصمیم‌هایی ناعادلانه را تولید می‌کند. به بیان دیگر، اگر داده‌ها ناعادلانه باشند، خروجی الگوریتم هم نمی‌تواند عادلانه باشد. در سطح طراحی، مسئله از آنجا ناشی می‌شود که بسیاری از الگوریتم‌ها براساس معیارهایی طراحی می‌شوند که تنها به کارایی و دقت توجه دارند، بدون آنکه عدالت و انصاف در آنها لحاظ شود.

نتیجه چنین طراحی‌هایی این است که سیستم، حتی اگر دقیق باشد، ممکن است خروجی‌هایی تولید کند که ناعادلانه یا تبعیض‌آمیز است. برای رفع این مشکلات، لازم است عدالت نه به عنوان یک ویژگی اختیاری، بلکه به عنوان یک شرط الزام‌آور در طراحی الگوریتم‌ها وارد شود. به این معنا که معیارهای مربوط به عدالت باید در همان ابتدا و در دل فرمول‌های اصلی الگوریتم‌ها (توابع هدف) گنجانده شوند، تا سیستم از ابتدا ملزم به رعایت آنها باشد. تنها با این رویکرد می‌توان از بروز تبعیض‌های سیستماتیک جلوگیری کرد و الگوریتم‌هایی ساخت که هم دقیق و کارآمد باشند و هم عادلانه.

از حیث حقوقی، عدالت الگوریتمی نه تنها مسئله‌ای اخلاقی، بلکه به عنوان یک حق بنیادین کاربران باید در ساختار هنجاری نظام‌های حقوقی نهادینه گردد. حق اعتراض به تصمیمات خودکار، حق اطلاع از منطق تصمیم و حق درخواست بازبینی انسانی، باید بخشی از حقوق شهروندی دیجیتال محسوب شوند. تحقق این امر مستلزم آن است که نظام‌های مسئولیت مدنی، کیفری و اداری نیز بازنگری شده و قابلیت انطباق با الگوهای تصمیم‌گیری غیرانسانی را پیدا کنند. مفهوم «خطای انسانی» در این نظام‌ها، باید به گونه‌ای توسعه یابد که شامل «خطای الگوریتمی» نیز گردد و مسئولیت‌ها در زنجیره توسعه، استقرار و بهره‌برداری به درستی بازتوزیع شوند. به بیان دیگر، هنگامی که تصمیم نادرست یا زیان‌بار ناشی از عملکرد الگوریتم است، باید روشن شود که مسئولیت آن متوجه چه کسی یا نهادی است: طراح، برنامه‌نویس، بهره‌بردار، یا ناظر سیستم. تنها در صورت بازتوزیع دقیق این مسئولیت‌ها در زنجیره تولید و استفاده از سامانه‌های هوشمند است که می‌توان از بروز خلأهای حقوقی جلوگیری کرد و عدالت را در حوزه فناوری نیز نهادینه ساخت.



از منظر فنی، یکی از چالش‌های اساسی عدالت الگوریتمی، تعارض ذاتی میان شفافیت و پیچیدگی در ساختار الگوریتم‌هاست. به ویژه در الگوریتم‌های یادگیری عمیق که به دلیل معماری چندلایه خود از قدرت پیش‌بینی بالایی برخوردارند، فرایند تصمیم‌گیری به گونه‌ای پیچیده و اغلب غیرقابل تفسیر باقی می‌ماند. این در حالی است که تحقق عدالت، به ویژه در بُعد رویه‌ای آن، مستلزم شفافیت، تبیین‌پذیری و امکان اعتراض آگاهانه به تصمیمات است. برای غلبه بر این تعارض، توسعه ابزارها و روش‌های فنی ویژه‌ای ضرورت دارد؛ از جمله طراحی الگوریتم‌های توضیح‌پذیر، بهره‌گیری از مدل‌های ترکیبی که دقت را با تفسیرپذیری تلفیق کنند و استقرار سازوکارهای نظارت الگوریتمی که امکان ارزیابی بیرونی بر فرایند تصمیم‌گیری آنها را فراهم سازند. این ابزارها باید قادر باشند تا منطق درونی الگوریتم را به زبان قابل فهم برای کاربران، نهادهای ناظر و مراجع قضایی تبدیل کنند.

در غیاب این قابلیت‌های فنی، نه تنها نظارت مؤثر اداری یا قضایی بر تصمیمات خودکار ناممکن خواهد بود، بلکه اصل پاسخ‌گویی نیز به چالش کشیده می‌شود و خطر شکل‌گیری فضایی غیرشفاف و خارج از دسترس حقوقی افزایش می‌یابد. از این رو، تفسیرپذیری و توضیح‌پذیری باید به عنوان معیارهای اساسی در طراحی و ارزیابی سیستم‌های هوش مصنوعی در نظر گرفته شوند، نه صرفاً مزایای اختیاری. در سطح کلان، تحقق عدالت الگوریتمی مستلزم شکل‌گیری سازوکارهایی منسجم، چندلایه

و فراملی است. از آنجا که الگوریتم‌ها اغلب در بسترهای جهانی عمل می‌کنند و آثار آنها فراتر از مرزهای جغرافیایی گسترش می‌یابد، لازم است نظام‌های حقوقی کشورها با یکدیگر همگرا شده و به چارچوب‌های تنظیم‌گری هماهنگ و تطبیق‌پذیر دست یابند. این هماهنگی، در عین حال باید به تنوع فرهنگی، ساختارهای حقوقی بومی، و حساسیت‌های اجتماعی هر کشور احترام بگذارد؛ به گونه‌ای که تعادل میان استانداردهای جهانی و اقتضائات محلی حفظ شود. فراتر از قوانین الزام‌آور، توسعه‌ی سازوکارهای مکمل همچون اصول اخلاقی مشترک، دستورالعمل‌های بین‌المللی و استانداردهای فنی قابل قبول برای طراحی و ارزیابی الگوریتم‌ها، نقشی حیاتی در تحقق عدالت ایفا می‌کند. چنین چارچوبی تنها در صورتی اثربخش خواهد بود که مبتنی بر هم‌افزایی مستمر میان دولت‌ها، نهادهای بین‌المللی، جامعه مدنی، بخش خصوصی و مراکز علمی باشد. به بیان دیگر، عدالت الگوریتمی در سطح جهانی، بدون نهادسازی مؤثر، هماهنگی میان فرهنگی، و اجماع اخلاقی-حقوقی، در حد یک ایده باقی خواهد ماند و امکان تحقق عملی نخواهد یافت.

در نهایت، بدون گنجاندن عدالت الگوریتمی در نظام آموزشی حقوقی، هیچ‌یک از اقدامات یادشده پایدار نخواهد بود. تربیت متخصصانی که هم با فناوری‌های یادگیری ماشین آشنا باشند و هم درک عمیقی از مفاهیم عدالت، تبعیض و مسئولیت حقوقی داشته باشند، کلید گذار از سامانه‌های ناقص و ناعادلانه به سیستم‌هایی مسئولانه، شفاف و عادلانه است. این تربیت علمی، اگرچه زمان‌بر و دشوار است، اما در بلندمدت مهم‌ترین سرمایه برای تضمین عدالت در عصر هوش مصنوعی به شمار می‌آید.

بر این اساس، عدالت الگوریتمی نباید به مثابه مفهومی آرمانی، اختیاری یا زینتی نگریسته شود، بلکه باید آن را به عنوان سنگ بنای نظم حقوقی نوین تلقی کرد؛ نظامی که اگرچه ابزار تصمیم‌سازی آن به طور فزاینده‌ای از سوی سیستم‌های غیرانسانی اداره می‌شود، اما نتایج آن ناگزیر باید همسو با ارزش‌های انسانی، اصول برابری و کرامت ذاتی افراد باقی بماند. در چنین نظامی، معیار مشروعیت نه صرفاً کارایی فنی، بلکه میزان انطباق الگوریتم‌ها با موازین عدالت انسانی خواهد بود.

منابع

۱. انصاری، باقر (۱۴۰۱). مطالعه حقوقی تبعیض الگوریتمی، فصلنامه دانش حقوق عمومی، سال یازدهم، زمستان، شماره ۳۸، صفحات: ۱۷۸-۱۴۷.
2. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. *ProPublica*. P:1-11.
3. Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT* '18)*, 149-159. <https://doi.org/10.1145/3287560.3287598>
4. Barocas, S., & Hardt, M. (2016). Fairness in Machine Learning. In *Proceedings of the 2016 Conference on Fairness, Accountability, and Transparency (FAT)**. P:1-167.
5. Barocas, S., & Nissenbaum, H. (2014). Big Data's End Run around Anonymity and Consent. In J. Lane, V. Stodden, S. Bender, & H. Nissenbaum (Eds.), *Privacy, Big Data, and the Public Good: Frameworks for Engagement* (pp. 44-75). Cambridge University Press.
6. Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, 104(3), 671-732.
7. Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*. P:149-159.
8. Bloch-Wehba, Hannah. (2022). Algorithmic Governance from the Bottom Up. *BYU Law Review*, Vol. 48, p. 69.
9. Chen, L. (2020). Algorithmic Fairness in Asian Legal Contexts. *Asian Journal of Law and Society*, 7(2), 233-255.
10. Crawford, K., Whittaker, M., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., ... & Schwartz, O. (2018). AI Now Report 2018. AI Now Institute. P:1-44.
11. European Commission. (2019). Ethics Guidelines for Trustworthy AI.
12. Floridi, L., Cows, J., King, T. C., & Taddeo, M. (2018). How to Design AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics*, 24(3), 1-13.
13. Friedman, B., & Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems*, 14(3), 330-347.
14. Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3), 50-57.
15. IEEE. (2020). IEEE 7000-2020 - Standard for Ethically Aligned Design.
16. ISO. (2020). ISO/IEC 27001:2013 Information technology — Security techniques — Information security management systems — Requirements.
17. Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1), 1-33.
18. Kearns, M., & Roth, A. (2020). The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press. P:1-232.
19. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), Article 115, 1-35. <https://doi.org/10.1145/3457607>
20. Mulligan, D. K., Zalnieriute, M., Daly, A., & Gellert, R. (2019). Articulating the values of data protection: progress and problems. *International Data Privacy Law*, 9(2), 135-148.
21. OECD. (2019). Recommendation of the Council on Artificial Intelligence. *OECD Publishing*.
22. O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens



- Democracy. Crown Publishing Group. P:1-272.
23. Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press. P:1-320.
24. PDPA. (2012). *Personal Data Protection Act*. Singapore.
25. Rawls, J. (1971). *A theory of justice* (p. 11). Harvard University Press.
26. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. P:1-10.
27. Rowe, E. A., & Prior, N. (2022). Procuring Algorithmic Transparency. *Alabama Law Review*. P: 303–364.
28. Russell, S. J., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson. P:1-11.
29. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59-68). ACM.
30. Stern, R. E., Liebman, B. L., Roberts, M., & Wang, A. Z. (2021). Automating Fairness? Artificial Intelligence in the Chinese Court. *Columbia Law School Scholarship Archive*. P:515-553.
31. US Congress. (2019). *Algorithmic Accountability Act of 2019*. H.R. 2231.
32. Woods, Andrew Keane. (2022/2023). Robophobia. *University of Colorado Law Review*, Vol. 93, No. 1, pp. 51–114.
33. Weller, A. (2019). Transparency: Motivations and Challenges. In T. K. Seppälä, S. Poutanen, & A. R. Harju (Eds.), *Digital Transformation and Ethics* (pp. 23-39). Springer International Publishing.
34. Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.P:1-704.