Uii vessity Thhaa rr sss

ttttt tttt
Home Page: https://jolt.ut.ac.ir

Online ISSN: 2423-6209

A Legal Reflection on the Interpretability and Explainability of Artificial Intelligence: From Conceptualization to Regulatory Frameworks¯

Mahdieh Latifzadeh

Assistant Professor of Private Law, Department (The Research Group) of Islamic Jurisprudence and Law, Institute for Islamic Studies in Humanities, Ferdowsi University of Mashhad, Mashhad, Iran. E-mail: latifzadeh@um.ac.ir

RRCCIIII OOO

iiii ty:::
Research Article

iiii ll Hrrrrr:
Received: June 20, 2024
Revised: April 21, 2025
Accepted: May 17, 2025
Published online: 23 September 2025

wwww....
Rule of Law,
Fair Trial,
Algorithmic Transparency.

RRRRRRRRR

The intersection of artificial intelligence with legal requirements related to the principles of fair trial and the necessity for transparency in legal and judicial decision-making has created a fundamental challenge. The "black box" nature of deep learning systems is at odds with the principle of "reasoned justification," raising the need to reconsider the relationship between technical limitations and legal requirements. Within this framework, the present study aims to assess the feasibility of achieving explainability and interpretability in intelligent systems to meet the legal requirements of fair proceedings, using a descriptive-analytical method and a documentary approach. In this process, the concepts and boundaries of explainability and interpretability are first examined, followed by an analysis of regulatory approaches such as the General Data Protection Regulation (GDPR) and the European Union Artificial Intelligence Act. Furthermore, the technical limitations of these technologies and the resulting legal challenges are analyzed to evaluate the effectiveness of these technologies in meeting legal requirements. Ultimately, the findings indicate that reducing the issue to a binary of "Complete Transparency" or "Absolute Black Box" is an oversimplification; the optimal solution is to design a risk-based regulatory system that enables the realization of algorithmic justice in the age of artificial intelligence.

eeee iii cle: Latifzadeh, M. (2025) A Legal Reflection on the Interpretability and Explainability of Artificial Intelligence: From Conceptualization to Regulatory Frameworks. .22 (1), 99-112. http//doi.org/10.22059/jolt.2025.396027.1007387

©ublisher: University of Tehran Press.
DOI: http//doi.org/10.22059/jolt.2025.396027.1007387

¯ This work is based upon research funded by Iran National Science Foundation (INSF) under project No 4031462

انتشارات دانشگاه تهران

حقوق خصوصی

سایت نشریه: https://jolt.ut.ac.ir

شاپا الکترونیکی: ۶۲۰۹-۲۴۲۳

تاملی حقوقی بر تفسیرپذیری و توضیح‌پذیری هوش مصنوعی: از مفهوم شناسی تا چارچوب تنظیم‌گری[*]

مهدیه لطیف‌زاده

استادیار حقوق خصوصی، گروه پژوهشی فقه و حقوق اسلامی، پژوهشکده مطالعات اسلامی در علوم انسانی دانشگاه فردوسی مشهد، مشهد، ایران. رایانامه: latifzadeh@um.ac.ir

اطلاعات مقاله

نوع مقاله:
پژوهشی

تاریخ‌های مقاله:
تاریخ دریافت: ۱۴۰۳/۰۳/۳۱
تاریخ بازنگری: ۱۴۰۴/۰۲/۰۱
تاریخ پذیرش: ۱۴۰۴/۰۲/۲۷
تاریخ انتشار: ۱۴۰۴/۰۷/۱۷

کلیدواژه:
حاکمیت قانون،
دادرسی منصفانه،
شفافیت الگوریتمی،
عدالت الگوریتمی،
قانون هوش مصنوعی اتحادیه اروپا.

چکیده

تلاقی هوش مصنوعی با الزامات حقوقی مرتبط با اصول دادرسی منصفانه و ضرورت شفافیت در تصمیم‌گیری‌های حقوقی و قضایی، چالشی اساسی ایجاد کرده است. ماهیت «جعبه سیاهِ» سیستم‌های یادگیری عمیق با اصل «استدلال مستند» در تعارض قرار دارد و این مسئله ضرورت بازاندیشی در تبیین رابطه میان محدودیت‌های فنی و الزامات حقوقی را مطرح می‌کند. در این چارچوب، پژوهش حاضر با هدف امکان‌سنجی تحقق توضیح‌پذیری و تفسیرپذیری در سیستم‌های هوشمند به منظور تأمین الزامات حقوقی مربوط به دادرسی منصفانه با اتکا بر روش توصیفی-تحلیلی و رویکرد اسنادی انجام شده است. در این مسیر ابتدا مفاهیم و مرزهای توضیح‌پذیری و تفسیرپذیری بررسی و سپس رویکردهای تنظیم‌گری مانند مقررات عمومی حفاظت از داده و قانون هوش مصنوعی اتحادیه اروپا مورد واکاوی قرار گرفته است. همچنین محدودیت‌های فنی این فناوری و چالش‌های حقوقی ناشی از آن‌ها نیز تحلیل شده است تا میزان کارآمدی این فناوری در پاسخگویی به الزامات حقوقی سنجیده شود. در نهایت، یافته ها حاکی از آن است که تقلیل مسئله به دوگانه «شفافیت کامل» یا «جعبه سیاه مطلق» رویکردی ساده‌انگارانه است؛ راهکار مطلوب، طراحی نظام تنظیم‌گری مبتنی بر ارزیابی سطح خطر است، به‌گونه‌ای که امکان تحقق عدالت الگوریتمی در عصر هوش مصنوعی فراهم آید.

استناد: لطیف‌زاده، مهدیه (۱۴۰۴). تاملی حقوقی بر تفسیرپذیری و توضیح‌پذیری هوش مصنوعی: از مفهوم شناسی تا چارچوب تنظیم‌گری. *حقوق خصوصی*، ۲۲ (۱)، ۹۹-۱۱۲.

http//doi.org/10.22059/jolt.2025.396027.1007387

ناشر: مؤسسه انتشارات دانشگاه تهران.
DOI: http//doi.org/10.22059/jolt.2025.396027.1007387

[*] این اثر تحت حمایت مادی بنیاد ملی علم ایران (INSF) برگرفته شده از طرح شماره «۴۰۳۱۴۶۲» انجام شده است.

مقدمه

گسترش هوش مصنوعی در حوزهٔ حقوق مسائل مهمی را در زمینهٔ شفافیت و پاسخگویی در مورد تصمیم‌گیری هوشمند ایجاد کرده است. در واقع استفادهٔ روزافزون از این فناوری در فرایندهای حقوقی ضرورت توضیح و تفسیر تصمیمات مبتنی بر هوش مصنوعی را برجسته می‌سازد و پرسش‌هایی اساسی دربارهٔ سازگاری این سیستم‌ها با اصول بنیادین حقوقی مانند شفافیت، موجه بودن تصمیمات، و حق دادرسی منصفانه مطرح می‌کند. در این زمینه، چالش اصلی ناشی از ماهیت «جعبهٔ سیاه»[1] بسیاری از سیستم‌های هوش مصنوعی پیشرفته، به‌ویژه مدل‌های مبتنی بر یادگیری عمیق، است که با اصل «استدلال مستند (موجه)»[2] در تعارض قرار دارد (Gurrapu et al., 2023). این سیستم‌ها، به‌رغم کارآمدی، فاقد تفسیر و توضیح کافی هستند و این فقدان مانعی جدّی برای استقرار، نظارت، و کنترل مؤثر نسبت به آن‌ها به شمار می‌رود (Bordt et al., 2022). به دیگر سخن الگوریتم‌های پیچیده همچون شبکه‌های عصبی عمیق، گرچه قادر به شناسایی الگوهای پیچیده در داده‌ها هستند، فرایند تصمیم‌گیری آن‌ها برای انسان نامفهوم است (Mumford et al., 2022). با وجود این وضعیت، اهمیت دو مفهوم کلیدی «تفسیرپذیری»[3] و «توضیح‌پذیری»[4] در این میان برجسته می‌شود. «تفسیرپذیری» به امکان شناخت و فهم سازکار درونی سیستم توسط انسان اشاره دارد، درحالی‌که «توضیح‌پذیری» بر قابلیت ارائهٔ توضیحات شفاف و قابل فهم دربارهٔ نتایج یا خروجی‌های سیستم برای کاربران تأکید دارد (Rudin, 2018).

این دو مفهوم در حقیقت پاسخی به نیاز روزافزون نظام‌های حقوقی برای اطمینان از عملکرد منصفانه و قابل اعتماد سیستم‌های هوشمند هستند و در بسیاری از چارچوب‌های تنظیم‌گری فناوری‌های نوین در سراسر جهان جایگاه ویژه‌ای یافته‌اند. بدین جهت این قابلیت‌ها نه‌تنها یک مزیت فنی بلکه یک الزام حقوقی محسوب می‌شوند. مثلاً مقررات عمومی حفاظت از دادهٔ اتحادیهٔ اروپا[5] جهت تنظیم‌گری این حوزه به طور ویژه تأثیرگذار بوده است، طوری که برخی از محققین این مقررات را ایجادکنندهٔ «حق توضیح»[6] برای افراد متأثر از تصمیم‌گیری الگوریتمی تفسیر می‌کنند (Goodman & Flaxman, 2017; Vermeire & Martens, 2020). به موجب حق توضیح، اشخاص موضوع داده[7] می‌توانند توضیحاتی دربارهٔ تصمیم اتخاذشده پس از ارزیابیِ خودکار دریافت کنند و چنین تصمیمی را به چالش بکشند (Berry, 2023). با تکیه بر این مقررات، ابتکارات دیگری در مورد تنظیم‌گری این حوزه در قانون هوش مصنوعی اتحادیهٔ اروپا[8] نسبت به توضیح‌پذیری مورد توجه قرار گرفته است؛ هرچند الزامات حقوقی دقیق همچنان در حال توسعه هستند (Bordt et al., 2022). به غیر از اتحادیهٔ اروپا، برخی از چارچوب‌های حقوقی بین‌المللی نیز مانند اصول سازمان همکاری و توسعهٔ اقتصادی[9] تأکید دارند افرادی که از سیستم‌های هوش مصنوعی آسیب دیده‌اند باید بتوانند با دریافت توضیحات ساده و قابل فهم از نحوهٔ عملکرد این سیستم‌ها نتایج آن را به چالش بکشند (Leofante et al., 2024). بدین ترتیب تحولات قانونی مبین این است که نهادهای قانونگذار به اهمیت این امر پی برده‌اند. این حساسیت به‌ویژه در جایی است که مخاطراتِ فناوری‌های نوین بر حقوق اساسی افراد اثرگذار هستند. این مخاطرات از جمله «تبعیض الگوریتمی»[10] و پیامدهای ناشی از آن است که نظام‌های حقوقی در مواجهه مناسب با آن چالش‌های بسیاری دارند و توضیح‌پذیری سیستم‌ها می‌تواند امکان شناسایی، اثبات، و اصلاح این نوع تبعیض را فراهم آورد (Mollas et al., 2020; Park & Chai, 2021).

از سویی دیگر، به‌رغم اهمیت بنیادین موضوع، چالش اساسی در این حوزه رویکرد تقلیل‌گرایانهٔ نظام‌های حقوقی به فناوری‌های نوظهور است. اغلب نظام‌های حقوقی عموماً ابزارهای فناورانه را به عنوان عناصری خنثی و فاقد پیامدهای ذاتی

1. black box
2. reasoned justification
3. interpretability
4. explainability
5. general data protection regulation (GDPR)
6. right to explanation
7. data subject
8. eu artificial intelligence act (AI ACT)
9. organization for economic co-operation and development (OECD)
10. algorithmic discrimination

قلمداد می‌کنند، بدون آنکه الزامات فنی ذاتی و پیامدهای احتمالی آن‌ها را به صورت نظام‌مند مورد بررسی و تحلیل قرار دهند (Lin et al., 2023; Zhu, 2021). با توجه به این نگرش محدود نظام‌های حقوقی به فناوری‌های نوین و پیامدهای آن، ضرورت مطالعهٔ دقیق‌تر این موضوع بیش از پیش احساس می‌شود. بر همین اساس، این پژوهش با هدف امکان‌سنجی تحقق توضیح‌پذیری و تفسیرپذیری در سیستم‌های هوشمند به منظور تأمین الزامات حقوقی مربوط به دادرسی منصفانه انجام شده است. در این مسیر ابتدا مفاهیم تفسیرپذیری و توضیح‌پذیری و تمایز میان آن‌ها تبیین می‌شود، سپس با بررسی چارچوب‌های حقوقی موجود تصویری روشن از الزامات حقوقی مرتبط ترسیم می‌شود. در ادامه، محدودیت‌های فنی و چالش‌های حقوقی پیش رو واکاوی و در نهایت با جمع‌بندی یافته‌ها و سنجش موازنهٔ میان الزامات حقوقی و محدودیت‌های فناورانه امکان تحقق این دو شاخص در سیستم‌های هوشمند ارزیابی می‌شود.

پیشینهٔ پژوهش

پژوهش حاضر با هدف پیش‌گفته، در مقایسه با آثار مهم این حوزه، واجد تمایزات ماهوی است که در ادامه مورد توجه قرار خواهد گرفت. مثلاً گیدوتی[1] و همکارانش (۲۰۱۸) در اثری با عنوان «بررسی روش‌های توضیح مدل‌های جعبهٔ سیاه» صرفاً به طبقه‌بندی فنی روش‌های توضیح مدل‌های هوش مصنوعی پرداخته‌اند؛ درحالی‌که پژوهش حاضر فراتر رفته و با تحلیل حقوقی این روش‌ها الزامات قانونی متناسب با هر روش را تبیین کرده است. این تحلیل حقوقی امکان ارزیابی کفایت روش‌های مختلف توضیح‌دهی در زمینه‌های متفاوت کاربردی را فراهم می‌سازد. همچنین مین[2] و همکارانش (۲۰۲۲) در مقاله‌ای با عنوان «هوش مصنوعی توضیح‌پذیر: بررسی جامع» مروری کلی بر مفاهیم مرتبط ارائه کرده‌اند. اما فاقد تحلیل حقوقی هستند. در مقابل پژوهش حاضر با تلفیق دانش فنی و حقوقی چارچوب تنظیم‌گری مبتنی بر این تلفیق را ارائه داده و شکاف موجود در مباحث میان‌رشته‌ای را مرتفع ساخته است. این تلفیق دانش امکان درک جامع‌تر از چالش‌های تنظیم‌گری هوش مصنوعی را فراهم می‌سازد. به‌علاوه گودمن[3] و فلکسمن (۲۰۱۷) در اثری با عنوان «مقررات اتحادیهٔ اروپا دربارهٔ تصمیم‌گیری الگوریتمی و حق توضیح» تنها به بررسی حق توضیح در مقررات اروپایی پرداخته‌اند. در مقابل پژوهش حاضر تحلیلی جامع‌تر از مفاهیم شفافیت الگوریتمی ارائه داده و با شناسایی سطوح مختلف شفافیت چارچوبی منسجم‌تر تدوین کرده است. این چارچوب منسجم امکان تنظیم‌گری متناسب با هر زمینهٔ کاربردی را فراهم می‌سازد. اثر دیگر از رودین[4] (۲۰۱۸) با عنوان «توقف توضیح در مدل‌های جعبهٔ سیاه یادگیری ماشینی برای تصمیمات مهم و استفاده از مدل‌های تفسیرپذیر به جای آن» صرفاً بر ضرورت استفاده از مدل‌های تفسیرپذیر تأکید کرده است. در مقابل پژوهش حاضر رویکردی متوازن ارائه داده است که هم مزایای مدل‌های تفسیرپذیر و هم کاربردهای مشروع مدل‌های جعبهٔ سیاه را به رسمیت می‌شناسد. این رویکرد متوازن با واقعیت‌های عملی توسعه و کاربرد هوش مصنوعی همخوانی بیشتری دارد. همچنین روزن[5] و همکارانش (۲۰۲۳) در اثری با عنوان «استدلالی علیه توضیح‌پذیری» با رویکردی یک‌سویه استدلال کرده‌اند که توضیح‌پذیری همواره مطلوب نیست. این در حالی است که پژوهش حاضر با رویکردی متوازن و مبتنی بر اصل تناسب شرایط متنوع را در نظر گرفته و چارچوبی انعطاف‌پذیر ارائه داده است. این انعطاف‌پذیری امکان تطبیق با شرایط متغیر فناوری و نیازهای متنوع حقوقی را فراهم می‌سازد. اثر دیگر از کیرات[6] و همکارانش (۲۰۲۲) است که با عنوان «عدالت و توضیح‌پذیری در سیستم‌های تصمیم‌گیری خودکار: چالشی برای علوم رایانه‌ای و حقوق» منتشر شده است. در این اثر تنها به طرح چالش‌ها اکتفا شده است؛ درحالی‌که پژوهش حاضر گامی فراتر برداشته و راه‌حل‌های مبتنی بر اصول حقوقی ارائه کرده است. در نهایت اثر کانالی[7] (۲۰۲۴) با عنوان «مدل‌های تفسیرپذیر هوش مصنوعی برای تصمیم‌گیری قضایی: فراتر از توضیح‌پذیری به سوی دادرسی عادلانه» محدود به کاربرد در تصمیم‌گیری قضایی است. در مقابل

1. Guidotti
2. Minh
3. Goodman
4. Rudin
5. Rozen
6. Kirat
7. Canalli

پژوهش حاضر با رویکردی فراگیر طیف وسیعی از کاربردهای هوش مصنوعی در حوزه‌های مختلف حقوقی را پوشش داده است. این جامعیت امکان بهره‌برداری از یافته‌های پژوهش در زمینه‌های متنوع را فراهم می‌سازد.

به طور کلی، پژوهش حاضر با اهتمام نسبت به تبیین دقیق ماهیت تفسیرپذیری و توضیح‌پذیری هوش مصنوعی در بستر حقوقی سعی دارد زبان مشترکی میان متخصصان فراهم آورد تا پیش‌نیاز تنظیم‌گری مؤثر در این حوزه به دست آید. در این زمینه، چارچوب تنظیم‌گری متوازن این پژوهش، با توجه به سطوح مختلف خطر و اهمیت تصمیمات، استانداردهایی متناسب با هر کاربرد ارائه می‌کند و از رویکردهای مطلق‌گرا فاصله می‌گیرد. همچنین، با بومی‌سازی مفاهیم و ارائهٔ راهکارهای متناسب با نظام حقوقی ایران زمینهٔ اجرای عملی یافته‌های پژوهش فراهم خواهد شد.

مبانی نظری و تبیین مفهومی تفسیرپذیری و توضیح‌پذیریِ هوش مصنوعی (هوش مصنوعی تفسیرپذیر و توضیح‌پذیر[1])

در مطالعات هوش مصنوعی «تفسیرپذیری» و «توضیح‌پذیری» اغلب مترادف تلقی می‌شوند. لیکن این دو از نظر ماهیت و کارکرد تفاوت‌های اساسی دارند که در حوزهٔ حقوقی نیز نقشی تعیین‌کننده در تدوین چارچوب‌های نظارتی و تعیین الزامات مربوط به شفافیت ایفا می‌کنند (Mansi & Riedl, 2024). بدین جهت تبیین دقیق این مفاهیم، علاوه بر کمک به شناخت بهتر سیستم‌های هوشمند، زمینه‌ساز تنظیم‌گری کارآمد و متناسب در فرایندهای حقوقی است (Arsenault et al., 2024).

ماهیت‌شناسی تفسیرپذیری

تفسیرپذیری در حوزهٔ هوش مصنوعی به عنوان خصیصه‌ای ذاتی و درون‌ساختاری مطرح می‌شود که قابلیت درک مستقیم منطق استنتاجی سیستم را بدون نیاز به تشریحات تکمیلی فراهم می‌آورد. این ویژگی بنیادین، متمایز از سایر مؤلفه‌های شفافیت الگوریتمی، عمدتاً در سیستم‌های با پیچیدگی محاسباتی محدود تجلی می‌یابد که ساختار تصمیم‌گیری آن‌ها برای متخصصین حقوقی مستقیماً قابل ادراک و تحلیل است. تفسیرپذیری با الگوهای سادهٔ هوش مصنوعی مانند «رگرسیون خطی»[2] سازگاری دارد که روش استدلالی آن‌ها شباهت قابل توجهی با روش‌های سنتی استدلال در علم حقوق دارد. به بیان ساده‌تر، همان‌طور که یک حقوقدان می‌تواند مسیر استدلالی قاضی را در رأی دادگاه درک کند، متخصصان مربوطه نیز می‌توانند منطق تصمیم‌گیری این نوع سیستم‌های هوش مصنوعی را به‌روشنی درک کنند. این قابلیت با اصل «شفافیت قضایی»[3] هم‌سو است که یکی از پایه‌های اساسی نظام عدالت است. به‌علاوه در نظام‌های حقوقی مبتنی بر رویهٔ قضایی (کامن‌لا) که اصل «استدلال مستند» از مبانی اساسی آن محسوب می‌شود تفسیرپذیری اهمیتی مضاعف می‌یابد. در این نظام‌ها، قضات ملزم به ارائهٔ استدلال‌های مبسوط و مشروح هستند که مبین سوابق قضایی مرتبط، اصول حقوقی حاکم، و نحوهٔ انطباق آن‌ها با وقایع موضوعی پرونده باشد. سیستم‌های هوش مصنوعی تفسیرپذیر، با قابلیت بازنمایی این الگوی استدلالی، انطباق بیشتری با اقتضائات نظام قضایی یادشده دارند. تفسیرپذیری همچنین ارتباط وثیقی با اصل «علنی بودن دادرسی» به عنوان یکی از مؤلفه‌های اساسی حاکمیت قانون برقرار می‌سازد. این اصل مستلزم آن است که نه‌تنها نتیجهٔ نهایی بلکه فرایند استدلالی منتهی به اتخاذ تصمیم نیز برای مراجع نظارتی، متخصصان حقوقی، و در مواردی عموم شهروندان قابل دسترسی، مشاهده، و ارزیابی انتقادی باشد. سیستم‌های تفسیرپذیر این امکان را فراهم می‌آورند که سازکارهای استنتاجی آن‌ها تحت بررسی و مداقهٔ دقیق حقوقی قرار گیرد و از این رهگذر مطابقت آن‌ها با موازین دادرسی منصفانه احراز شود. در نهایت در پرتو دکترین مسئولیت‌پذیری قضایی نیز تفسیرپذیری سیستم‌های هوش مصنوعی به عنوان پیش‌شرط اساسی پاسخگویی و نظارت‌پذیری در فرایندهای تصمیم‌گیری قضایی مطرح می‌شود که امکان تشخیص خطاهای استدلالی و انحرافات از اصول حقوقی را میسر می‌سازد. در واقع، تفسیرپذیری به افراد این امکان را می‌دهد که درک کنند سیستم چگونه از داده‌های ورودی به نتیجهٔ نهایی رسیده است. تفسیرپذیری را می‌توان به «شفافیت استدلال» تشبیه کرد؛ همان‌طور که در نظام حقوقی قاضی باید دلایل رأی خود را بیان کند، سیستم هوش مصنوعی

1. Interpretable Artificial Intelligence/ Explainable Artificial Intelligence
2. linear regression
3. judicial transparency

این اصل، فارغ از سایر بسترهای حقوقی، در مادهٔ ۵ سند امنیت قضایی ایران (مصوب ۱۳۹۹) مورد تصریح قرار گرفته است.

نیز باید قادر باشد مسیر استدلالی خود را نشان دهد. این خصیصه دارای دو جنبهٔ اساسی است: نخست، قابلیت نمایش مراحل تصمیم‌گیری و دوم، قابلیت درک این مراحل توسط انسان. بدین جهت تفسیرپذیری از منظر حقوقی اهمیت ویژه‌ای دارد. زیرا تضمین می‌کند تصمیمات سیستم‌های هوشمند قضایی، همانند تصمیمات قضات انسانی، بر پایهٔ استدلال‌های قابل بررسی و نقد استوار هستند (See. Gyevnar et al., 2023).

ماهیت‌شناسی توضیح‌پذیری

با توجه به تفاوت بیان‌شده در بند سابق، حال می‌توان توضیح‌پذیری را دقیق‌تر تعریف کرد. توضیح‌پذیری در بستر هوش مصنوعی به توانایی سیستم در ارائهٔ توجیهات روشن و قابل درک دربارهٔ چرایی و چگونگی رسیدن به یک تصمیم خاص اشاره دارد، به نحوی که نه‌تنها برای متخصصان فنی، بلکه برای عموم شهروندان و اشخاص ذی‌نفع نیز قابل فهم باشد. این ویژگی، متمایز از تفسیرپذیری، عمدتاً به سازکارهای پس‌نگر[1] اشاره دارد که پس از اتخاذ تصمیم توضیحاتی را در خصوص منطق استنتاجی سیستم ارائه می‌کند؛ هرچند این توضیحات لزوماً بازنمای دقیق و کاملی از فرایند محاسباتی و الگوریتمیک زیربنایی نیستند. از منظر دکترین حقوقی، توضیح‌پذیری با اصل «بیان دلایل تصمیمات قضایی (دلیل‌مندی تصمیم قضایی)» هم‌سویی مفهومی دارد که در بسیاری از نظام‌های حقوقی به عنوان یکی از مؤلفه‌های اساسی دادرسی منصفانه مورد شناسایی قرار گرفته است. فارغ از آنچه بیان شد، توضیح‌پذیری به‌تنهایی نمی‌تواند متضمن امکان نظارت قضایی مؤثر باشد. زیرا توضیحات ارائه‌شده ممکن است فاقد عمق تحلیلی لازم برای ارزیابی انتقادی باشند. به بیان ساده‌تر، توضیح‌پذیری صرفاً «چرایی» تصمیم را به زبان ساده بیان می‌کند. اما نظارت قضایی مؤثر نیازمند درک «چگونگی» دقیق فرایند تصمیم‌گیری و امکان بررسی انتقادی آن در چارچوب اصول و موازین حقوقی است. بنابراین، در کنار توضیح‌پذیری، مؤلفه‌های دیگری همچون تفسیرپذیری- که توضیح آن بیان شد- برای تحقق نظارت قضایی کارآمد ضروری است. با وجود این توضیح‌پذیری پیوند مستحکمی با حق بنیادین دسترسی به عدالت برقرار می‌سازد. چون ارائهٔ توضیحات قابل فهم برای افراد فاقد تخصص فنی امکان بهره‌مندی معنادار از نظام قضایی را برای عموم شهروندان تسهیل می‌کند. این اصل مبتنی بر مبانی حقوقی است که هر شخص، فارغ از میزان آشنایی با مفاهیم تخصصی، باید قادر به درک دلایل تصمیماتی باشد که بر حقوق و تکالیف او تأثیرگذار است. چنین رویکردی با غایت اصلی توضیح‌پذیری، یعنی قابل فهم ساختن منطق تصمیم‌گیری برای عموم، انطباق کامل دارد (See. Dyoub et al., 2020).

از منظر کارکردی، توضیح‌پذیری به دو گونهٔ متمایز قابل تفکیک است. توضیح‌پذیری جهانی[2] که به تبیین کلیت الگوریتم و منطق زیربنایی آن می‌پردازد و توضیح‌پذیری محلی[3] که معطوف به تشریح تصمیمات خاص و منفرد است (Guidotti et al., 2018). این دوگانه انعکاس‌دهندهٔ تقابل دیرینه میان «قواعد عام» و «موارد خاص» در حقوق است. از دیدگاه دکترین حقوقی، توضیح‌پذیری محلی با اصل «فردی بودن تصمیمات قضایی»[4] همخوانی دارد که به موجب آن هر حکم قضایی باید متناسب با اوضاع و احوال خاص پرونده صادر شود. این اصل مستلزم توجه به شرایط منحصربه‌فرد متهم، بزه‌دیده، و زمینه‌های ارتکاب جرم است. در واقع توضیح‌پذیری محلی می‌تواند سازکاری برای تضمین رعایت این اصل در سیستم‌های هوش مصنوعی قضایی فراهم آورد، به نحوی که نشان دهد چگونه عوامل خاص هر پرونده در فرایند تصمیم‌گیری الگوریتمی لحاظ شده‌اند. در مقابل، توضیح‌پذیری جهانی با اصول کلی حقوقی و قواعد عام قضایی که شالودهٔ نظام حقوقی را تشکیل می‌دهند قرابت مفهومی دارد. این نوع توضیح‌پذیری امکان ارزیابی انطباق سیستم هوش مصنوعی با هنجارهای بنیادین حقوقی را فراهم می‌آورد. با وجود این، باید به این محدودیت ذاتی توجه داشت که روش‌های توضیح پس‌نگر همواره نمی‌توانند توصیف‌های جامع و تضمین‌شدهٔ دقیقی از رفتار واقعی مدل ارائه دهند. این محدودیت از منظر حقوق دادرسی بسیار حائز اهمیت است. زیرا می‌تواند منجر به ارائهٔ توضیحاتی شود که به‌ظاهر با موازین قانونی منطبق به نظر می‌رسند، اما در واقع بازنمای صادقانه‌ای از منطق زیربنایی الگوریتم

1. post-hoc
2. universal explainability
3. local explainability
4. individualization of judicial decisions

نیستند. چنین وضعیتی می‌تواند «تقلب نسبت به قانون»[1] تلقی شود (See. Bechler-Speicher et al., 2024). خلاصه اینکه تفسیرپذیری ویژگی ذاتی و ساختاری الگوریتم است که امکان مشاهده و درک مستقیم منطق تصمیم‌گیری را فراهم می‌کند؛ درحالی‌که توضیح‌پذیری قابلیتی بیرونی است که با سازکارهای الحاقی تصمیمات سیستم را برای انسان تبیین می‌کند. از منظر حقوقی، تفسیرپذیری تضمین شفافیت بیشتری ایجاد می‌کند. زیرا منطق تصمیم‌گیری را بدون واسطه و خطر تحریف آشکار می‌سازد (See. Minh et al., 2022).

الزام به توضیح‌پذیری و تفسیرپذیریِ هوش مصنوعی در نظام‌های حقوقی

با توجه به اهمیت موضوع، نظام‌های حقوقی مختلف در حال تدوین و اجرای قوانین و مقرراتی هستند که الزامات تفسیرپذیری و توضیح‌پذیری را برای سیستم‌های هوش مصنوعی تعیین و اجرایی کنند؛ هرچند سطح پیشرفت و رویکردهای حقوقی در این زمینه یکسان نیست. در این بخش، این الزامات به ترتیب ابتدا در حقوق اتحادیهٔ اروپا، سپس در نظام حقوقی ایران، آن‌گاه در نظام بین‌الملل، و در نهایت با مروری بر رویهٔ قضایی به عنوان ابعاد عملی مورد بررسی قرار خواهند گرفت.

تبیین الزامات در حقوق اتحادیهٔ اروپا

اتحادیهٔ اروپا، با وجود نقش برجسته در ایجاد چارچوب‌های تقنینی و نظارتی نسبت به فناوری، توضیح‌پذیری در سیستم‌های هوش مصنوعی را به موجب قوانین و مقررات مرتبط الزامی می‌داند. قانون هوش مصنوعی اتحادیهٔ اروپا، مصوب ۲۰۲۴[2]، به‌ویژه در بخش‌های مربوط به سیستم‌های پرخطر، الزاماتی را برای شفافیت و توضیح‌پذیری تعیین کرده است.[3] الزامات این مقررات بر اساس اینکه چه کسی از هوش مصنوعی استفاده می‌کند و سطوح مختلف خودکارسازی در فرایند تصمیم‌گیری متفاوت است. مثلاً مادهٔ ۱۳ این مقررات در بندهای مختلف به لزوم وجود توضیح‌پذیری هوش مصنوعی اشاره دارد. قانونگذار اروپایی، در بند نخست، اصل شفافیت کافی را به عنوان تکلیف قانونی برای طراحی و توسعهٔ سیستم‌های هوش مصنوعی پرخطر مقرر داشته است. این الزام حقوقی با هدف تضمین قابلیت تفسیر خروجی‌های سیستم توسط استقراردهندگان (به‌کارگیرندگان حرفه‌ای)[4] وضع شده و پیش‌نیاز اساسی برای تحقق توضیح‌پذیری محسوب می‌شود. سپس به موجب بند ۳ (ب) این ماده توضیح‌پذیری هوش مصنوعی یک الزام حقوقی برای سیستم‌های پرخطر محسوب می‌شود. این مفاد صراحتاً تکلیف به تعبیهٔ قابلیت‌های فنی برای ارائهٔ اطلاعات توضیحی مرتبط با خروجی سیستم را مقرر می‌دارد. این حکم قانونی سه تکلیف محوری را ایجاب می‌کند. این موارد وجود قابلیت‌های فنی توضیح‌دهنده در سیستم، ارتباط مستقیم توضیحات با خروجی‌های سیستم، و تصریح این قابلیت‌ها در دستورالعمل‌های مربوط به استفاده است. در مقام استناد به این مقررهٔ قانونی می‌توان اظهار داشت که قانونگذار با تصریح به لزوم قابلیت‌های فنی توضیح‌دهنده عملاً اصل منع جعبهٔ سیاه بودن سیستم‌های هوش مصنوعی پرخطر را به عنوان یک قاعدهٔ آمره وضع کرده و توضیح‌پذیری را به مثابهٔ یک الزام قانونی برای این سیستم‌ها تثبیت کرده است (European Parliament and of The Council, 2024: 59 & 60).

با توجه به آنچه بیان شد، رویکرد اتحادیهٔ اروپا به توضیح‌پذیری هوش مصنوعی مبین تلاشی نظام‌مند برای ایجاد تعادل میان نوآوری فناورانه و حمایت از حقوق بنیادین افراد است. این مقررات، با اتخاذ رویکردی مبتنی بر خطر، تعهدات حقوقی متناسب با سطح خطر سیستم‌های هوش مصنوعی تعیین می‌کند که این امر اصل تناسب[5] در قانون‌گذاری را منعکس می‌سازد. توضیح‌پذیری در این چارچوب حقوقی به عنوان امتدادی از حقوق بنیادین افراد تلقی می‌شود و با اصول مندرج در منشور حقوق بنیادین اتحادیهٔ اروپا[6] و مقررات عمومی حفاظت از داده[7] همخوانی دارد. این مقررات همچنین مسئولیت‌های حقوقی مشخصی را برای ارائه‌دهندگان

1. fraus legis
2. artificial intelligence Act (AI Act)
۳. مثلاً در مقدمات توجیهی –Recital– شمارهٔ ۲۷ مقرر شده است که سیستم‌ها باید قابل ردیابی و توضیح‌پذیر باشند و کاربران باید از تعامل با هوش مصنوعی آگاه شوند.
4. deployers
5. principle of proportionality
6. charter of fundamental rights of the european union
7. general data protection regulation (GDPR)

(تأمین‌کنندگان)[1] و استقراردهندگان (به‌کارگیرندگان حرفه‌ای)[2] سیستم‌های هوش مصنوعی تعیین می‌کند که زنجیرهٔ مسئولیت[3] را در صورت نقض الزاماتِ توضیح‌پذیری شکل می‌دهد. با این حال، تعریف دقیق «توضیح کافی» از دیدگاه حقوقی همچنان چالش‌برانگیز است و مستلزم تدوین استانداردهایی برای ارزیابی کفایت توضیحات ارائه‌شده توسط سیستم‌های هوش مصنوعی است. ناگفته نماند که تأثیرات حقوقی این مقررات به دلیل «اثر بروکسل»[4] می‌تواند فراتر از مرزهای اتحادیهٔ اروپا گسترش یابد. همچنین الگویی را برای سایر نظام‌های حقوقی در سراسر جهان فراهم سازد (See. Sovrano & Vitali, 2021). به‌علاوه، همان‌طور که در مقدمه بیان شد، مقررات عمومی حفاظت از دادهٔ اتحادیهٔ اروپا[5] نیز جهت تنظیم‌گری این حوزه به طور ویژه تأثیرگذار بوده است. به موجب حق توضیح، اشخاص موضوع داده[6] می‌توانند توضیحاتی دربارهٔ تصمیم اتخاذشده پس از ارزیابیِ خودکار دریافت کنند و چنین تصمیمی را به چالش بکشند (Berry, 2023). البته نظراتِ متفاوتِ دیگری دربارهٔ اینکه آیا این مقررات به‌صراحت «حق توضیح» ایجاد می‌کند نیز وجود دارند. لیکن در خصوص اینکه مقررات پیش‌گفته بیانِ منطق دخیل در تصمیم‌گیری خودکار را الزامی می‌داند توافق وجود دارد. چنین ضرورتی از مواد ۱۳ تا ۱۵ قابل استنباط است (Rozen et al., 2023).

تبیین الزامات در نظام حقوقی ایران

در این زمینه، نظام حقوقی ایران گرچه قانون خاصی برای تنظیم‌گری هوش مصنوعی تصویب نکرده است،[7] بسترهای حقوقی موجود مفادی برای الزامات مربوط به تفسیرپذیری و توضیح‌پذیری هوش مصنوعی ارائه کرده‌اند. در این خصوص می‌توان به اصل ۱۶۶ قانون اساسی ایران اشاره کرد که قضات را ملزم می‌کند احکام خود را مستند و مستدل به مواد قانونی و اصولی صادر کنند که بر اساس آن‌ها رأی داده‌اند. بنابراین، سیستم‌های هوش مصنوعی که در فرایندهای قضایی به کار گرفته می‌شوند باید به گونه‌ای طراحی شوند که تصمیمات آن‌ها قابل تفسیر و استناد به منابع حقوقی باشد. توجه شود که توضیح‌پذیری فراتر از شفافیت صرف سیستم یا داده‌ها است. این امر مستلزم درک عمیق‌تر از سازکارهای درونی مدل تصمیم‌گیری است که هم دیدگاه‌های هستی‌شناختی (شفافیت اطلاعات) هم معرفت‌شناختی (شفافیت مدل تصمیم‌گیری) را در بر می‌گیرد (Gorski et al., 2020). همچنین مادهٔ ۲۹۶ قانون آیین دادرسی مدنی مقرر می‌دارد رأی دادگاه باید موجه و مدلل و مستند به مواد قانون و اصولی باشد که بر اساس آن صادر شده است. به علاوه طبق مادهٔ ۳ قانون آیین دادرسی مدنی قضات موظف‌اند با توجه به دلایل و مستندات حکم مقتضی صادر کنند و در صورت فقدان قانون به استناد منابع معتبر اسلامی یا اصول حقوقی رأی دهند. این مواد نیز مؤید این امر هستند که هر سیستم هوش مصنوعی مورد استفاده باید قادر به ارائهٔ استدلال‌های حقوقی شفاف و مستند باشد. از منظر چارچوب‌های حقوقی مربوط به حقوق شهروندی نیز این مهم قابل بیان است. مثلاً مادهٔ ۴۵ منشور حقوق شهروندی بر حق دسترسی به اطلاعات تأکید می‌کند و مادهٔ ۵۷ آن حق برخورداری از دادرسی عادلانه را به رسمیت می‌شناسد. این حقوق ایجاب می‌کند شهروندان بتوانند دلایل تصمیمات اتخاذشده دربارهٔ خود را بفهمند، حتی اگر این تصمیمات توسط سیستم‌های هوش مصنوعی پشتیبانی یا اتخاذ شده باشند. در حوزهٔ حفاظت از داده‌ها نیز با آنکه قانونی در مورد حمایت از داده به تصویب نرسیده است، در اسناد پیشنهادی[8] مفادی مشابه با قانون اروپایی حمایت از داده (GDPR)، که زمینه‌ساز الزام به توضیح‌پذیری تصمیمات خودکار است، قابل مشاهده است. در نهایت نیز به موجب مادهٔ ۱ قانون مسئولیت مدنی مسئولیتِ جبران خسارت برای کسی که بدون مجوز

1. providers
2. deployers
3. chain of responsibility
4. brussels effect
5. general data protection regulation (GDPR)
6. data subject

۷. گفتنی است اقدامات صورت‌گرفته در حوزهٔ تنظیم‌گری هوش مصنوعی در ایران شامل تصویب سند ملی هوش مصنوعی و تأسیس سازمان ملی هوش مصنوعی (به همراه شورای ملی راهبردی هوش مصنوعی) در سال ۱۴۰۳ است. این موارد گرچه گام‌های مثبتی محسوب می‌شوند کفایت لازم برای مواجهه با چالش‌های این حوزه را ندارند.

۸. نظام حقوقی ایران در حوزهٔ حفاظت از داده‌های شخصی با خلأهای قانونی عمیقی مواجه است. البته ایران با ارائهٔ پیش‌نویس‌های قانونی- طرح حمایت و حفاظت از داده و اطلاعات شخصی مرکز پژوهش‌های مجلس شورای اسلامی (۱۴۰۰) و طرح حفاظت از دادهٔ شخصی مرکز پژوهش‌های مجلس شورای اسلامی (۱۴۰۳)- در این خصوص گام برداشته است. در عین حال این امر در خصوص حمایت از داده‌های شخصی کافی نیست.

قانونی به جان، سلامت، مال، آزادی، حیثیت، یا شهرت تجاری دیگری لطمه وارد سازد می‌تواند به مسئولیت طراحان و توسعه‌دهندگان حتی کاربران سیستم‌های هوش مصنوعی تعمیم یابد، به‌ویژه در مواردی که عدم توضیح‌پذیری این سیستم‌ها منجر به تصمیمات نادرست و زیان‌بار شود. ناگفته نماند در حقوق ایران «توضیح‌پذیری» با الزام به مستدل بودن آرا همخوانی دارد؛ در مقابل «تفسیرپذیری» با ضرورت موجه بودن تصمیمات و امکان درک منطق استدلال قضایی از سوی مخاطبان مطابقت دارد.

بررسی الزامات در نظام بین‌الملل

فارغ از چارچوب‌های ملی (ایران) و منطقه‌ای (اتحادیهٔ اروپا) بیان‌شده، ضرورت وجود توضیح‌پذیری و تفسیرپذیریِ هوش مصنوعی از اسناد و مبانی حقوقی بین‌الملل نیز قابل استنباط است. بدین توضیح که در چارچوب اصول دادرسی منصفانه[1] «حق بر اطلاع از مبانی تصمیم»[2] به عنوان یکی از اصول بنیادین شناخته می‌شود که به موجب آن هر شخص محق به آگاهی از دلایل تصمیماتی است که بر وضعیت حقوقی او تأثیر می‌گذارند (L. Canalli, 2024). این حق در اسناد بین‌المللی متعددی از جمله مادهٔ ۶ کنوانسیون اروپایی حقوق بشر[3] و مادهٔ ۱۴ میثاق بین‌المللی حقوق مدنی و سیاسی[4] مورد شناسایی و حمایت قرار گرفته است. در نظام‌های قضایی سنتی (غیر هوشمند) این حق از طریق اصل مستدل بودن آرا و امکان تجدیدنظرخواهی تأمین می‌شود. این اصل مستقیماً از مفهوم کلان‌تر «حاکمیت قانون»[5] نشئت می‌گیرد که مطابق آن اعمال قدرت حاکمیتی باید مبتنی بر قوانین موضوعه و نه ارادهٔ خودسرانهٔ مراجع مربوطه باشد. تصمیمات فاقد مبانی استدلالی روشن با این اصل بنیادین در تعارض قرار می‌گیرند. زیرا امکان نظارت قضایی[6] و پاسخگویی را محدود می‌سازند (Kirat et al., 2022). همچنین اصل «استماع عادلانه»[7] ایجاب می‌کند که اشخاص متأثر از تصمیمات از فرصت کافی برای ارائهٔ دیدگاه‌ها و استدلال‌های خود برخوردار باشند. این اصل مستلزم آن است که افراد بتوانند منطق تصمیم‌گیری را درک کنند تا قادر به ارائهٔ استدلال‌های مؤثر و متناسب باشند. سیستم‌های هوش مصنوعی غیر قابل توضیح این امکان را از ایشان سلب می‌کنند. به‌علاوه اصل «برابری در (سلاح‌ها) امکانات دفاع»[8] در دادرسی نیز مقتضی آن است که طرفین دعوا در موقعیتی متوازن برای دفاع از مواضع خود قرار داشته باشند. هنگامی که یک طرف (نظیر دادستان یا خواهان) از ابزارهای هوش مصنوعی پیشرفته بهره می‌گیرد، عدم توضیح‌پذیری این سیستم‌ها می‌تواند به نقض این اصل بنیادین منجر شود. برای تحقق برابری واقعی طرف مقابل باید قادر به درک و نقد استدلال‌های زیربنایی تصمیمات هوش مصنوعی باشد، حال آنکه سیستم‌های غیر قابل توضیح چنین امکانی را فراهم نمی‌آورند (See. Kirat et al., 2022; See. L. Canalli, 2024).

البته تفاوت‌های ساختاری در نظام‌های حقوقی مختلف نیز بر الزامات توضیح‌پذیری و تفسیرپذیری تأثیرگذار است. مثلاً در نظام حقوق نوشته، تأکید بر استدلال قیاسی[9]، ارجاع به مواد قانونی وجود دارد؛ درحالی‌که در نظام کامن‌لا استدلال استقرایی[10]، تحلیل سوابق قضایی، و تمایز میان پرونده‌ها اهمیت دارد. بنابراین، سیستم‌های هوش مصنوعی باید شیوهٔ استدلال و ارائهٔ توضیحات خود را متناسب با ویژگی‌های هر نظام تنظیم کنند (Nigam et al., 2024; Steging et al., 2021a). به دیگر سخن آنچه بیان شد ـ الزامات متفاوت پیش‌گفته ـ ضرورت توسعهٔ الگوریتم‌های تفسیرپذیر[11] و توضیح‌پذیر را برجسته می‌سازد. البته الگوریتم‌های یادشده باید قابلیت انطباق با الگوهای استدلالی متفاوت در نظام‌های حقوقی مختلف را داشته باشند (Y. Zhang et al., 2020).

1. due process principles
2. right to know reasons
3. convention for the protection of human rights and fundamental freedoms
4. international covenant on civil and political rights (ICCPR)
5. rule of law
6. judicial review
7. fair hearing
8. equality of arms
9. deductive reasoning
10. inductive reasoning
11. interpretable algorithms

نگاهی به رویهٔ قضایی

فارغ از اسناد حقوقی که مفاد آن‌ها مشعر بر الزام توضیح‌پذیری و تفسیرپذیری است، نگاهی به رویهٔ قضایی نیز در این موضوع خالی از فایده نیست. در این زمینه سه رأی کلیدی دیوان اروپایی حقوق بشر قابل اشاره است. نخست دیوان اروپایی حقوق بشر در پروندهٔ «تاکسکه علیه بلژیک» (شمارهٔ دعوای ۰۵/۹۲۶، مورخ ۱۶ نوامبر ۲۰۱۰)[۱] به اتفاق آرا نقض مادهٔ ۶(۱) کنوانسیون اروپایی حقوق بشر را احراز کرد. در این پرونده، دیوان تصریح کرد سیستم قضایی بلژیک در رسیدگی به اتهام قتل یک وزیر و تلاش برای قتل همسر وی علیه آقای تاکسکه تعهد خود مبنی بر ارائهٔ دلایل کافی و مستدل برای تصمیم قضایی را ایفا نکرده است. به موجب استدلال دیوان، صرف اشاره به وجود شواهد و ادله، بدون تبیین نحوهٔ ارزیابی آن‌ها و پاسخگویی به استدلال‌های دفاعی متهم، مغایر با اصل دادرسی منصفانه است. دیوان تأکید کرد که تکلیف به ارائهٔ دلایل کافی از ملزومات حق بر دادرسی منصفانه محسوب می‌شود و نقش اساسی در تضمین امکان تجدیدنظرخواهی مؤثر و نظارت عمومی بر فرایند قضایی ایفا می‌کند. مورد دوم رأی دیوان در پروندهٔ «اس و مارپر علیه بریتانیا» (۲۰۰۸) است.[۲] دیوان اروپایی حقوق بشر در این پرونده (شمارهٔ دعواهای ۰۴/۳۰۵۶۲ و ۰۴/۳۰۵۶۶، مورخ ۴ دسامبر ۲۰۰۸) به اتفاق آرا نقض مادهٔ ۸ کنوانسیون اروپایی حقوق بشر را محرز دانست. در این پرونده، دیوان استدلال کرد نگهداری نامحدود نمونه‌های دی‌ان‌ای، اثر انگشت، و سایر اطلاعات زیست‌سنجی افرادی که محکومیتی نداشته‌اند یا پروندهٔ آن‌ها مختومه شده است مداخلهٔ نامتناسب در حق بر حریم خصوصی محسوب می‌شود. دیوان با اشاره به اصل تناسب تصریح کرد به‌رغم مشروعیت هدف (پیشگیری از جرم)، فقدان محدودیت زمانی و عدم امکان حذف داده‌ها فراتر از آنچه در یک جامعهٔ دموکراتیک ضروری است تلقی می‌شود. همچنین دیوان بر لزوم تمایز میان افراد محکوم و غیر محکوم و همچنین اهمیت شفافیت در پردازش داده‌های حساس تأکید کرد. در پرونده سوم رأی دیوان در پروندهٔ «دی.اچ. و دیگران علیه جمهوری چک» (۲۰۰۷)[۳] دیوان اروپایی حقوق بشر (شمارهٔ دعوای ۰۰/۵۷۳۲۵، مورخ ۱۳ نوامبر ۲۰۰۷) با اکثریت ۱۳ رأی در برابر ۴ رأی نقض مادهٔ ۱۴ کنوانسیون را احراز کرد. در این پرونده، دیوان مفهوم تبعیض غیر مستقیم را به طور مشخص مورد شناسایی قرار داد و اعلام کرد که سیاست‌های آموزشی جمهوری چک، به‌رغم ظاهر خنثی، در عمل منجر به قرار گرفتن نامتناسب کودکان اقلیت روما در مدارس ویژهٔ کودکان با ناتوانی ذهنی شده است. دیوان با استناد به شواهد آماری و گزارش‌های نهادهای بین‌المللی تأکید کرد تأثیر نامتناسب یک سیاست بر گروه خاصی از افراد، حتی در غیاب قصد تبعیض‌آمیز، می‌تواند موجد مسئولیت دولت باشد. همچنین دیوان تصریح کرد که در موارد ادعای تبعیض غیر مستقیم بار اثبات به دولت خوانده منتقل می‌شود.

این سه رأی دیوان اروپایی حقوق بشر اصول بنیادینی را تبیین می‌کنند که مستقیماً به الزامات توضیح‌پذیری و تفسیرپذیری سیستم‌های هوش مصنوعی مرتبط هستند. این موارد «ارائهٔ استدلال کافی» است؛ یعنی سیستم‌های هوش مصنوعی باید قادر به ارائهٔ استدلال‌های روشن و کافی برای تصمیمات خود باشند (ساتاخانوف علیه روسیه). همچنین «شفافیت در استفاده از داده‌ها» است؛ یعنی باید مشخص باشد که داده‌ها چگونه جمع‌آوری، ذخیره، و پردازش می‌شوند (اس و مارپر علیه بریتانیا). و در نهایت «شناسایی و جلوگیری از تبعیض غیر مستقیم» است؛ یعنی باید سیستم‌ها قادر به توضیح این امر باشند که تصمیمات آن‌ها منجر به تبعیض غیر مستقیم علیه گروه‌های خاص نمی‌شود (دی.اچ. و دیگران علیه جمهوری چک). در مجموع، این آرای قضایی مبین این است که توضیح‌پذیری و تفسیرپذیری ضرورتی حقوقی مبتنی بر حقوق بشر است، به‌ویژه زمانی که تصمیمات هوش مصنوعی مستقیماً بر حقوق اساسی افراد اثر می‌گذارد.

---

۱. برای اطلاعات تکمیلی مراجعه شود به:
https://hudoc.echr.coe.int/eng#{%22itemid%22:[%22001-101739%22]}

۲. برای اطلاعات تکمیلی مراجعه شود به:
https://hudoc.echr.coe.int/eng#{%22itemid%22:[%22001-90051%22]}

۳ برای اطلاعات تکمیلی مراجعه شود به:
https://hudoc.echr.coe.int/eng#{%22itemid%22:[%22001-83256%22]}

محدودیت‌های فنی تفسیرپذیری و توضیح‌پذیری هوش مصنوعی و چالش‌های حقوقی ناشی از آن

از چالش‌های اساسی در کاربرد هوش مصنوعی در حوزهٔ حقوق محدودیت‌های ذاتی این فناوری در زمینهٔ تفسیرپذیری و توضیح‌پذیری است که پیامدهای حقوقی قابل توجهی به دنبال دارد. این محدودیت‌ها عمدتاً به دلیل تعارض میان پیچیدگی الگوریتمی و ضرورت شفافیت در تصمیم‌گیری‌های حقوقی ایجاد می‌شوند و پذیرش و استقرار فناوری در نظام‌های حقوقی را با دشواری‌های جدّی مواجه ساخته‌اند. یکی از مهم‌ترین چالش‌ها برقراری توازن میان دقت با قابلیت تفسیرپذیری یا توضیح‌پذیری است؛ موازنه‌ای که در مباحث تخصصی به «موازنه دقت- تفسیرپذیری»[1] معروف است. توضیح اینکه افزایش دقت و کارایی معمولاً با پیچیده‌تر شدن مدل‌ها مانند شبکه‌های عصبی عمیق همراه است که به کاهش تفسیرپذیری منجر می‌شود. این امر در حوزهٔ حقوق که شفافیت فرایند تصمیم‌گیری اهمیت اساسی دارد مشکل‌ساز است. زیرا هم دقت- برای تحقق عدالت ماهوی- هم شفافیت- برای تحقق عدالت رویه‌ای- ضرورت دارد. پژوهش‌ها نشان می‌دهد رضایت افراد از تصمیمات حقوقی بیش از نتیجه به ادراک منصفانه بودن فرایند بستگی دارد و فرایندهای شفاف حتی با نتایج نامطلوب مشروعیت بیشتری ایجاد می‌کنند (Steging et al., 2021b). چالش دیگر ضعف کیفی توضیحات و عمق استدلال حقوقی ارائه‌شده توسط سیستم‌های هوش مصنوعی است. با وجود پیشرفت فناوری، همچنان خروجی بسیاری از سامانه‌های توضیح‌پذیر و تفسیرپذیر محدود به «فهرست عوامل تأثیرگذار»[2] است و این رویکرد نمی‌تواند الزامات پیچیدهٔ حقوقی را پاسخ دهد. این در حالی است که در پرونده‌های حقوقی صرف ارائهٔ لیست متغیرها بدون تشریح نحوهٔ تعامل آن‌ها و منطق حقوقی حاکم بر تصمیم کافی نیست. این محدودیت با «حق بر استدلال کافی»[3] که در رویهٔ قضایی بسیاری از کشورها و دادگاه‌های بین‌المللی به رسمیت شناخته شده است ارتباط مستقیم دارد. مثلاً، دیوان اروپایی حقوق بشر در پروندهٔ تاکسکه علیه بلژیک (شمارهٔ دعوای ۰۵/۹۲۶، مورخ ۱۶ نوامبر ۲۰۱۰)[4]، که پیش‌تر بیان شد، تصریح کرد که پاسخ‌های یک‌کلمه‌ای به سؤالات حقوقی کلیدی الزامات دادرسی منصفانه را برآورده نمی‌کند. توضیحات ساده‌شده حاصل از برخی روش‌های توضیح‌پذیر هوش مصنوعی ممکن است نتوانند منطق پیچیده و چندبعدی استدلال حقوقی را منعکس کنند. از جنبهٔ نظری نیز هربرت هارت، فیلسوف و حقوقدان برجستهٔ قرن بیستم، با طرح مفهوم «بافت باز بودن»[5] قانون در کتاب *مفهوم قانون* بر وجود ابهام ذاتی در قوانین تأکید دارد و معتقد است تفسیر این ابهامات نیازمند قضاوت انسانی است؛ قابلیتی که حتی تفسیرپذیرترین مدل‌های هوش مصنوعی فعلی فاقد آن هستند (Licato et al., 2022). از سویی دیگر اجرای سیستم‌های هوش مصنوعی توضیح‌پذیر و تفسیرپذیر در عرصهٔ حقوق با موانع فنی، زیرساختی، و اقتصادی نیز روبه‌روست. طراحی سیستم‌های تفسیرپذیر و توضیح‌پذیر برای حل مسائل پیچیدهٔ حقوقی نیازمند تخصص و منابع مالی قابل توجهی است که می‌تواند نابرابری در دسترسی به عدالت را نیز به دنبال داشته باشد (See. Gorski et al., 2020). در نهایت، قانونگذاران با چالش انتخاب میان مدل‌های ذاتاً تفسیرپذیر با شفافیت بیشتر و مدل‌های پیچیده‌تر اما دارای توضیح‌پذیری پس‌نگر مواجه‌اند. این انتخاب پیامدهای حقوقی مهمی دارد و باید با توجه به ماهیت موضوع، حساسیت آن، میزان تأثیر بر حقوق اساسی افراد، و اصل تناسب انجام شود. شاید بتوان گفت در حوزه‌هایی مانند عدالت کیفری استفاده از مدل‌های تفسیرپذیر الزامی باشد؛ حال آنکه در موضوعات با حساسیت کمتر می‌توان از مدل‌های پیچیده‌تر، مشروط به ارائهٔ توضیح کافی، بهره برد (See. Zhang & Zhang, 2023).

نتیجه

تفسیرپذیری و توضیح‌پذیری هوش مصنوعی در تقاطع فناوری و حقوق قرار دارد. شفافیت الگوریتمی نه‌تنها یک چالش فنی بلکه یک ضرورت حقوقی است که با اصول بنیادین دادرسی منصفانه، حاکمیت قانون، و حقوق بشر ارتباط مستقیم دارد. تمایز میان

1. accuracy-interpretability trade-off
2. feature importance lists
3. right to adequate reasoning

۴. برای اطلاعات تکمیلی مراجعه شود به:

https://hudoc.echr.coe.int/eng#{%22itemid%22:[%22001-101739%22]}

5. open texture

توضیح‌پذیری و تفسیرپذیری بیانگر رویکردهای متفاوت به شفافیت است که هر یک کاربردهای خاص خود را در زمینه‌های حقوقی دارد. درحالی‌که مدل‌های ذاتاً تفسیرپذیر با استانداردهای بالای استدلال قضایی همخوانی بیشتری دارند، روش‌های توضیحِ پس‌نگر انعطاف‌پذیری بیشتری را برای استفاده از مدل‌های دقیق‌تر فراهم می‌کنند. با این حال، بررسی دقیق این دو رویکرد در بستر کاربردهای عملی نشان می‌دهد که انتخاب میان آن‌ها مستلزم موازنه‌ای دقیق میان دقت الگوریتمی و شفافیت تصمیم‌گیری است. از سویی دیگر تحلیل نظام‌مند رابطهٔ میان تفسیرپذیری و توضیح‌پذیری هوش مصنوعی با چارچوب‌های حقوقی آشکارکنندهٔ تعارض میان الزامات حقوقی و محدودیت‌های فنی است. این امر یعنی به‌رغم تصریح چارچوب‌های حقوقی بر لزوم شفافیت، قابلیت تفسیر و قابلیت توضیح سیستم‌های هوش مصنوعی موانع فنی موجود منجر به ظهور چالش‌های حقوقی‌ـ تنظیمی شده که تحقق کامل این الزامات را دشوار ساخته است. این تعارض به‌ویژه در مورد سیستم‌های مبتنی بر یادگیری عمیق که به‌رغم دقت بالا از جعبهٔ سیاه غیر شفاف برخوردارند مشهودتر است.

با توجه به یافته‌های پیش‌گفته در خصوص تعارض میان الزامات حقوقی و محدودیت‌های فنی، ضرورت اتخاذ رویکردی تدریجی در تنظیم‌گری حوزهٔ هوش مصنوعی آشکار می‌شود. به دیگر سخن به‌رغم محدودیت‌های موجود در تحقق کامل اصول تفسیرپذیری و توضیح‌پذیری در سیستم‌های هوش مصنوعی، اتخاذ رویکردی متوازن و منطبق با اقتضائات عملی می‌تواند به تأمین سطح مطلوبی از شفافیت الگوریتمی منجر شود. این رویکرد تدریجی مستلزم طراحی نظام تنظیم‌گری متناسب با سطح مخاطرات است که در آن الزامات قانونی متفاوتی برای کاربردهای مختلف هوش مصنوعی تعریف می‌شود. در این نظام برای کاربردهای پرخطرـ نظیر سیستم‌های تصمیم‌گیر در حوزهٔ سلامت، امنیت ملی، یا فرایندهای قضاییـ سطوح بالاتری از توضیح‌پذیری و تفسیرپذیری الزامی است؛ درحالی‌که برای کاربردهای کم‌خطرـ مانند سیستم‌های توصیه‌گر سرگرمی‌ـ انعطاف‌پذیری بیشتری در نظر گرفته می‌شود. برای تضمین اثربخشی این نظام، ضروری است استانداردهای فنی مشخصی جهت سنجش میزان تفسیرپذیری و توضیح‌پذیری سیستم‌های هوش مصنوعی تدوین شود. ناگفته نماند که موفقیت این امر منوط به توجه هم‌زمان به سه عامل کلیدی است. این موارد نخست تضمین حقوق بنیادین اشخاص در تعامل با سیستم‌های هوش مصنوعی، دوم برقراری توازن میان الزامات شفافیت و ضرورت حمایت از اسرار تجاری، و سوم توجه به پیامدها و تبعات اجتماعی گسترده‌تر نسبت به کاربرد این فناوری است. بنابراین آنچه اهمیت دارد این است که این نظام باید به گونه‌ای طراحی شود که ضمن تضمین حقوق شهروندان مانعی بر سر راه نوآوری و پیشرفتِ فناوری نیز ایجاد نکند.

در این میان، نظام حقوقی ایران با وجود برخورداری از الزامات حقوقی عام همچون الزام به مستند و مستدل بودن همچنین موجه و مدلل بودن آرا به موجب قانون اساسی و قوانین عادی همچنان فاقد الزامات خاص و به طور کلی قانون ویژه در حوزهٔ هوش مصنوعی است. همین خلأ ضرورت تدوین «قانون هوش مصنوعی» را برجسته می‌سازد؛ قانونی که یکی از محورهای اساسی آن باید توجه صریح به الزامات توضیح‌پذیری و تفسیرپذیری باشد. چنین قانونی می‌تواند ضمن ایجاد چارچوب مشخص برای طبقه‌بندی کاربردهای هوش مصنوعی بر اساس سطح خطر الزامات شفافیت را برای حوزه‌های حساس به صورت الزام‌آور پیش‌بینی کند. همچنین معیارهای فنی و حقوقی لازم را برای سنجش و ارزیابی خروجی‌ها تعیین کند و در عین حال تعادل میان شفافیت و حمایت از اسرار تجاری را حفظ کند.

منابع

سند امنیت قضایی مصوب ۱۳۹۹. بازیابی‌شده در ۲ اردیبهشت ۱۴۰۴ از: https://rc.majlis.ir/fa/law/show/1623986

سند ملی هوش مصنوعی جمهوری اسلامی، مصوبهٔ شورای عالی انقلاب، مورخ تیرماه ۱۴۰۳، بازیابی در ۲۵ فروردین ۱۴۰۴ از: https://rc.majlis.ir/fa/law/show/1811432

قانون اساسی جمهوری اسلامی ایران، مصوب ۱۳۵۸ با اصلاحات ۱۳۶۸.

قانون آیین دادرسی مدنی، مصوب ۱۳۷۹.

قانون مسئولیت مدنی، مصوب ۱۳۳۹.

مرکز پژوهش‌های مجلس شورای اسلامی (۱۴۰۳). *طرح حفاظت از داده‌های شخصی*. تهران: مرکز پژوهش‌های مجلس شورای اسلامی، بازیابی در ۲۰ فروردین ۱۴۰۴ از: https://rc.majlis.ir/fa/legal_draft/show/1816729

منشور حقوق شهروندی، ابلاغی رئیس‌جمهور، ۱۳۹۵.

Arsenault, P.-D., Wang, S., & Patenande, J.-M. (2024). A Survey of Explainable Artificial Intelligence (XAI) in Financial Time Series Forecasting**.** *arXiv preprint*. https://doi.org/10.48550/ARXIV.2407.15909

Bechler-Speicher, M., Globerson, A., & Gilad-Bachrach, R. (2024). The Intelligible and Effective Graph Neural Additive Networks**.** *arXiv preprint*. https://doi.org/10.48550/ARXIV.2406.01317

Berry, D. M. (2023). Explanatory Publics: Explainability and Democratic Thought**.** *arXiv preprint*. https://doi.org/10.48550/ARXIV.2304.02108

Bordt, S., Finck, M., Raidl, E., & Von Luxburg, U. (2022). Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts**.** In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT) (pp. 891–905). ACM. https://doi.org/10.1145/3531146.3533153

Dyoub, A., Costantini, S., & Lisi, F. A. (2020). Logic Programming and Machine Ethics**.** *Electronic Proceedings in Theoretical Computer Science*, 325, 6–17. https://doi.org/10.4204/EPTCS.325.6

European Parliament and Of The Council (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) Text with EEA Relevance. *Official Journal of the European Union*, L 1689, 1–144.

Goodman, B. & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision Making and a "Right to Explanation". *AI Magazine*, 38(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Gorski, L., Ramakrishna, S., & Nowosielski, J. M. (2020). Towards Grad-CAM Based Explainability in a Legal Text Processing Pipeline. *arXiv preprint*. https://doi.org/10.48550/ARXIV.2012.09603

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A Survey of Methods for Explaining Black Box Models. *arXiv preprint*. https://doi.org/10.48550/ARXIV.1802.01933

Gurrapu, S., Huang, L., & Batarseh, F. A. (2023). ExClaim: Explainable Neural Claim Verification Using Rationalization. *arXiv preprint*. https://doi.org/10.48550/ARXIV.2301.08914

Gyevnar, B., Ferguson, N., & Schafer, B. (2023). Bridging the Transparency Gap: What Can Explainable AI Learn from the AI Act? In K. Gal, A. Nowé, G. J. Nalepa, R. Fairstein, & R. Rădulescu (Eds.), *Frontiers in Artificial Intelligence and Applications* (Vol. 365, pp. 27–39). IOS Press. https://doi.org/10.3233/FAIA230367

Kirat, T., Tambou, O., Do, V., & Tsoukiàs, A. (2022). Fairness and Explainability in Automatic Decision-Making Systems: A Challenge for Computer Science and Law. *arXiv preprint*. https://doi.org/10.48550/ARXIV.2206.03226

L. Canalli, R. (2024). Interpretable AI Models for Judicial Decisionmaking: Beyond Explicability Towards Legal Due Process. *E-Publica*, 11(1), 45–67. https://doi.org/10.47345/v11n1art6

Leofante, F., Ayoobi, H., Dejl, A., Freedman, G., Gorur, D., Jiang, J., et al. (2024). Contestable AI Needs Computational Argumentation. *arXiv preprint*. https://doi.org/10.48550/ARXIV.2405.10729

Licato, J., Fields, L., & Marji, Z. (2022). Resolving Open-Textured Rules with Templated Interpretive Arguments. *arXiv preprint*. https://doi.org/10.48550/ARXIV.2212.09700

Lin, N., Liu, H., Fang, J., Zhou, D., & Yang, A. (2023). An Interpretability Framework for Similar Case Matching. *arXiv preprint*. https://doi.org/10.48550/ARXIV.2304.01622

Mansi, G. & Riedl, M. (2024). Recognizing Lawyers as AI Creators and Intermediaries in Contestability. *arXiv preprint*. https://doi.org/10.48550/ARXIV.2409.17626

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable Artificial Intelligence: A Comprehensive Review. *Artificial Intelligence Review*, *55*(5), 3503–3568. https://doi.org/10.1007/s10462-021-10088-y

Mollas, I., Bassiliades, N., & Tsoumakas, G. (2020). Altruist: Argumentative Explanations through Local Interpretations of Predictive Models. *arXiv preprint*. https://doi.org/10.48550/ARXIV.2010.07650

Mumford, J., Atkinson, K., & Bench-Capon, T. (2022). Reasoning with Legal Cases: A Hybrid ADF-ML Approach. In E. Francesconi, G. Borges, & C. Sorge (Eds.), *Legal Knowledge and Information Systems: JURIX 2022* (Vol. 359, pp. 97–106). IOS Press. https://doi.org/10.3233/FAIA220452

Nigam, S. K., Deroy, A., Maity, S., & Bhattacharya, A. (2024). Rethinking Legal Judgement Prediction in a Realistic Scenario in the Era of Large Language Models. In *Proceedings of the Natural Legal Language Processing Workshop 2024* (pp. 61–80). ACL. https://doi.org/10.18653/v1/2024.nllp-1.6

Park, M. & Chai, S. (2021). AI Model for Predicting Legal Judgments to Improve Accuracy and Explainability of Online Privacy Invasion Cases. *Applied Sciences*, *11*(23), 11080. https://doi.org/10.3390/app112311080

Rozen, H. W., Elkin-Koren, N., & Gilad-Bachrach, R. (2023). The Case Against Explainability. *arXiv preprint*. https://doi.org/10.48550/ARXIV.2305.12167

Rudin, C. (2018). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *arXiv preprint*. https://doi.org/10.48550/ARXIV.1811.10154

Sovrano, F. & Vitali, F. (2021). An Objective Metric for Explainable AI: How and Why to Estimate the Degree of Explainability. *arXiv preprint*. https://doi.org/10.48550/ARXIV.2109.05327

Steging, C., Renooij, S., & Verheij, B. (2021a). Discovering the Rationale of Decisions: Experiments on Aligning Learning and Reasoning. *arXiv preprint*. https://doi.org/10.48550/ARXIV.2105.06758

Steging, C., Renooij, S., & Verheij, B. (2021b). Rationale Discovery and Explainable AI. In E. Schweighofer (Ed.), *Frontiers in Artificial Intelligence and Applications* (Vol. 341, pp. 137–150). IOS Press. https://doi.org/10.3233/FAIA210341

Vermeire, T. & Martens, D. (2020). Explainable Image Classification with Evidence Counterfactual. *arXiv preprint*. https://doi.org/10.48550/ARXIV.2004.07511

Zhang, N. & Zhang, Z. (2023). The Application of Cognitive Neuroscience to Judicial Models: Recent Progress and Trends. *Frontiers in Neuroscience*, *17*, 1257004. https://doi.org/10.3389/fnins.2023.1257004

Zhang, Y., Tiňo, P., Leonardis, A., & Tang, K. (2020). A Survey on Neural Network Interpretability. *arXiv preprint*. https://doi.org/10.48550/ARXIV.2012.14261

Zhu, Z. (2021). Legal Regulation of Algorithmic Discrimination. *Advances in Social Behavior Research*, 1(1), 65–72. https://doi.org/10.54254/2753-7102/1/2021009