

MDOI-01-2026.0170
MNR--1-0E8D87

تفسیرپذیری مدل‌های یادگیری ماشین در تصمیم‌گیری‌های حساس به خطا

مهسا حق نظری

۱- مدیر و مدرس آموزشگاه فنی و حرفه ای؛ دانشجوی کارشناسی ارشد مهندسی و علم کامپیوتر گرایش هوش مصنوعی دانشگاه مهرالبرز، تهران، ایران.

چکیده

تصمیم‌گیری در قلمروهای حساس به خطا، جایی که پیامد یک پیش‌بینی نادرست می‌تواند به آسیب جانی، محرومیت اجتماعی یا زیان مالی گسترده بینجامد، مدل‌های یادگیری ماشین را با الزامی فراتر از دقت پیش‌بینی مواجه می‌سازد: الزام به پاسخ‌گویی، قابلیت بازبینی و امکان مداخله انسانی. این پژوهش با تمرکز بر تفسیرپذیری مدل‌های یادگیری ماشین در چنین زمینه‌هایی، این پرسش را پیش می‌گذارد که کدام سازوکارهای تفسیرپذیری می‌توانند میانجی میان عملکرد پیش‌بینی و مسئولیت تصمیم واقع شوند. روش پژوهش از نوع مروری–تحلیلی و مبتنی بر بازخوانی نظام‌مند شواهد تجربی منتشرشده در حوزه‌های بهداشت، عدالت کیفری و اعتبارسنجی مالی است. یافته‌ها نشان می‌دهند که تفسیرپذیری نه یک فضیلت یکپارچه، بلکه مجموعه‌ای از خواص ناهمگون است: وفاداری توضیح به منطق درونی مدل، پایداری توضیح در برابر اغتشاش‌های جزئی، سازگاری با دانش تخصصی حوزه، و سودمندی عملی برای تصمیم‌گیر انسانی. شواهد تجربی حاکی از آن است که روش‌های مبتنی بر ارزش شاپلی در وفاداری و پایداری پیشتازند، اما هزینه محاسباتی و پیچیدگی تفسیر آن‌ها پذیرش را محدود می‌کند؛ در مقابل، روش‌های جایگزین محلی سرعت و شهود بیشتری عرضه می‌کنند ولی در داده‌های پرمتغیر دچار فروپاشی پایداری می‌شوند. نتیجه‌گیری مقاله آن است که در تصمیم‌های حساس به خطا، تفسیرپذیری باید به‌عنوان یک سامانه ارزیابی چندبعدی و وابسته به زمینه طراحی شود، نه به‌عنوان یک افزونه پس‌روی اختیاری. چارچوب پیشنهادی، تفسیر را نه برای توجیه تصمیم، بلکه برای کشف محدودیت‌های مدل و فعال‌سازی قضاوت حرفه‌ای به کار می‌گیرد و بر این اصل استوار است که ارزش توضیح در تصمیم حساس، به توان آن در تغییر رفتار تصمیم‌گیر در جهت کاهش خطر بستگی دارد.

واژگان کلیدی: تفسیرپذیری، تصمیم‌گیری حساس به خطا، وفاداری توضیح، پایداری تفسیر، پاسخ‌گویی الگوریتمی.



مقدمه

ورود مدل‌های یادگیری ماشین به فضاهای تصمیم‌گیری که در آن‌ها خطاهای پیش‌بینی به‌سرعت به رویدادهای زیان‌بار تبدیل می‌شوند، پرسشی بنیادین را از دل عملگرایی فنی بیرون کشیده است: هنگامی که یک مدل، فردی را برای مراقبت ویژه علامت‌گذاری می‌کند، درخواست وثیقه‌ای را رد می‌نماید یا پرونده‌ای را در معرض خطر تخلف مالی قرار می‌دهد، چه چیزی به تصمیم‌گیر انسانی اجازه می‌دهد نه‌تنها پیش‌بینی را بپذیرد، بلکه در قبال آن پاسخ‌گو باشد؟ پاسخ سنتی به این پرسش، توسل به دقت آماری بوده است؛ اما دقت، حتی در بالاترین سطوح، هیچ اطلاعاتی درباره سازوکار تصمیم یا توزیع خطا در اختیار نمی‌گذارد. مدل می‌تواند در مجموعه داده آزمون دقت بالایی داشته باشد و در همان حال، در زیرگروه‌های کوچک اما آسیب‌پذیر، الگویی نظام‌مند از خطا تولید کند که از چشم ارزیابی کلان پنهان بماند. این شکاف میان عملکرد تجمعی و قابلیت پاسخ‌گویی فردی، تفسیرپذیری را از یک دغدغه فرعی به یک ضرورت ساختاری در تصمیم‌گیری‌های حساس به خطا ارتقا داده است (یزدانی و همکاران، ۱۴۰۳). تفسیرپذیری در این معنا، نه تلاش برای آشکار ساختن «حقیقت نهایی» مدل، بلکه فراهم آوردن بازنمایی قابل‌فهمی از دلایل یک پیش‌بینی معین است؛ بازنمایی‌ای که تصمیم‌گیر انسانی بتواند آن را با دانش حوزه‌ای خود مقابله دهد، محدودیت‌هایش را تشخیص دهد و در صورت لزوم، مسیر تصمیم را تغییر دهد.

با این حال، ادبیات تفسیرپذیری اغلب میان دو قطب نوسان کرده است: از یک سو، مدافعان مدل‌های ذاتی تفسیرپذیر، شفافیت کامل ساختار مدل را شرط مسئولیت‌پذیری می‌دانند؛ از سوی دیگر، رویکردهای پس‌روی، بدون آنکه ساختار مدل را تغییر دهند، توضیحاتی برای هر پیش‌بینی تولید می‌کنند. هر دو قطب در بستر تصمیم‌های حساس با محدودیت‌های جدی روبه‌رو هستند. مدل‌های ذاتی تفسیرپذیر، مانند مدل‌های خطی و درخت‌های تصمیم کم‌عمق، در بسیاری از مسائل با روابط غیرخطی و تعاملی، دقت کافی را از دست می‌دهند؛ در حالی که مدل‌های پس‌روی، اگرچه امکان استفاده از مدل‌های پیچیده را فراهم می‌کنند، خود ممکن است توضیحاتی تولید کنند که با منطق واقعی مدل فاصله دارند یا در اجراهای مکرر ناپایدار باشند. پژوهش‌های تجربی نشان داده‌اند که توضیحاتی که وفاداری پایینی به مدل دارند، می‌توانند اعتماد کاذب ایجاد کنند و تصمیم‌گیر را به سمت اتکای بیش‌ازحد یا کنارگذاری نابجای مدل سوق دهند (شاه و تن تیجه، ۲۰۲۵). بنابراین، مسئله صرفاً «داشتن توضیح» نیست، بلکه «کیفیت توضیح در نسبت با تصمیم» است.

در حوزه بهداشت، این تنش به‌شکل ویژه‌ای آشکار می‌شود. پزشک بالینی که با یک هشدار سپسیس یا یک پیش‌بینی خطر عود بیماری روبه‌رو می‌شود، نه به فهرستی از وزن‌های عددی، بلکه به روایتی نیاز دارد که با مسیر استدلال بالینی او هم‌راستا باشد. مطالعه‌ای در زمینه پشتیبانی تصمیم بالینی نشان داد که ارائه توضیحات، به‌ویژه زمانی که با دانش بالینی هم‌سو باشد، می‌تواند وزن اتکای پزشکان به توصیه سیستم را به‌طور معناداری افزایش دهد؛ اما همان مطالعه هشدار داد که توضیحات ناهم‌راستا با دانش تخصصی، ممکن است به رد کامل سیستم بینجامد (بهروزی و بختیاری، ۱۴۰۵). در عدالت کیفری، ابزارهای ارزیابی خطر مانند مقیاس‌های پیش‌بینی تکرار جرم، به‌طور سنتی با انتقادهایی درباره شفافیت و سوگیری روبه‌رو بوده‌اند؛ تفسیرپذیری در این زمینه نه‌تنها برای دفاع از تصمیم قاضی، بلکه برای شناسایی الگوهای تبعیض‌آمیز نهفته در داده‌های تاریخی ضروری است. در اعتبارسنجی مالی نیز، مدل‌هایی که درخواست وام یا معامله مشکوک را علامت‌گذاری می‌کنند، تحت الزامات مقرراتی فزاینده‌ای برای ارائه «دلیل adverse action» قرار دارند؛ توضیحی که هم برای مشتری قابل‌فهم باشد و هم از نظر آماری معتبر. این زمینه‌های ناهمگون یک نقطه اشتراک حیاتی دارند: در همه آن‌ها، تفسیرپذیری یک محصول جانبی مطلوب نیست، بلکه شرط امکان‌پذیری مسئولیت اخلاقی و حقوقی است.

ادبیات فارسی نیز در سال‌های اخیر به این موضوع توجه کرده است. یزدانی و همکاران (۱۴۰۳) در مروری تحلیلی بر نقش الگوریتم‌های تفسیرپذیر در مدیریت معاصر نشان دادند که در تصمیم‌های راهبردی و نظارتی، شفافیت می‌تواند بر دقت صرف پیش‌بینی ارجحیت یابد، به‌ویژه هنگامی که هزینه خطا نامتقارن باشد. بهروزی و بختیاری (۱۴۰۵) با طراحی یک چارچوب پشتیبانی تصمیم مبتنی بر مدل درخت تقویت‌شده و نسبت‌دهی ویژگی شاپلی برای مدیریت خطاهای پیش‌تحلیلی در آزمایشگاه‌های بالینی، نشان دادند که ترکیب پیش‌بینی و تفسیر می‌تواند اولویت‌های مداخله مدیریتی را از دل مدل استخراج کند. با این حال، همان‌گونه که خود تصریح می‌کنند، یافته‌هایشان بر داده شبیه‌سازی‌شده استوار است و پرسش از تعمیم‌پذیری به داده‌های واقعی و به دیگر حوزه‌های حساس باقی می‌ماند. شکاف اصلی که این پژوهش بر آن تمرکز دارد، نه کمبود روش‌های تفسیر، بلکه کمبود چارچوبی برای ارزیابی اینکه کدام توضیح، در کدام زمینه حساس، واقعاً تصمیم را ایمن‌تر یا مسئولانه‌تر می‌کند، است. تفسیرپذیری زمانی در تصمیم حساس به خطا معنا دارد که بتواند نه‌فقط منطق مدل، بلکه حدود اعتبار آن منطق را نیز آشکار سازد و تصمیم‌گیر را از اطمینان کاذب به قضاوت آگاهانه رهنمون شود.

بیان مسئله

مسئله محوری این پژوهش از یک ناسازگاری ساختاری میان دو الزام همزمان در تصمیم‌گیری حساس به خطا برمی‌خیزد: الزام به عملکرد پیش‌بینی بالا، که اغلب مدل‌های پیچیده و کم‌شفاف را ایجاب می‌کند، و الزام به پاسخ‌گویی و قابلیت بازبینی، که شفافیت منطق تصمیم را شرط می‌داند. مدل‌های یادگیری ماشین با ظرفیت بالا، از جمله جنگل‌های تصادفی، ماشین‌های تقویت گرادیان و شبکه‌های عمیق، در بسیاری از وظایف پیش‌بینی از مدل‌های ساده پیشی می‌گیرند، اما ساختار تصمیم آن‌ها برای انسان قابل بازیابی مستقیم نیست. این عدم شفافیت در زمینه‌های حساس، سه مشکل به‌هم‌پیوسته ایجاد می‌کند. نخست، مشکل توجیه‌پذیری: تصمیم‌گیر انسانی که پیش‌بینی مدل را می‌پذیرد، باید بتواند دلایل آن را به ذی‌نفعان، نهادهای نظارتی یا همکاران حرفه‌ای خود ارائه دهد. دوم، مشکل تشخیص خطا: بدون درک منطق تصمیم، شناسایی موقعیت‌هایی که مدل در آن‌ها به‌طور نظام‌مند اشتباه می‌کند یا از الگوهای نامرتبط پیروی می‌نماید، دشوار است. سوم، مشکل کالیبراسیون اعتماد: تصمیم‌گیر انسانی نیاز دارد بداند چه زمانی به مدل اتکا کند و چه زمانی قضاوت حرفه‌ای خود را بر پیش‌بینی مدل مقدم بدارد؛ توضیحاتی که وفاداری یا پایداری کافی ندارند، این کالیبراسیون را مختل می‌کنند و به اتکای کورکورانه یا رد بی‌مورد می‌انجامند.

روش‌های تفسیرپذیری پس‌روی، مانند نسبت‌دهی ویژگی شاپلی و توضیحات محلی سازگار با مدل، به‌عنوان پاسخ فنی به این مشکل عرضه شده‌اند. با این حال، خود این روش‌ها واجد شرایطی هستند که در تصمیم‌های حساس به خطا باید به‌طور جدی سنجیده شوند. وفاداری توضیح به مدل، یعنی میزان بازتاب واقعی سهم ویژگی‌ها در پیش‌بینی، در همه روش‌ها یکسان نیست. نسبت‌دهی شاپلی، به‌دلیل ریشه داشتن در نظریه بازی‌های همیارانه، از نظر axiomatically وفادارتر تلقی می‌شود، اما محاسبه دقیق آن برای مدل‌های پیچیده از نظر محاسباتی پرهزینه است و در عمل اغلب با تقریب‌هایی همراه می‌شود که وفاداری را کاهش می‌دهند. توضیحات محلی مبتنی بر جانشین، اگرچه سریع‌تر و شهودی‌ترند، اما به‌دلیل وابستگی به نمونه‌گیری تصادفی و مدل جانشین محلی، در برابر تغییرات جزئی در ورودی یا پارامترهای نمونه‌گیری ناپایدارند. پژوهش‌های تجربی نشان داده‌اند که پایداری این توضیحات در داده‌های کم‌بعد و اجتماعی می‌تواند بالا باشد، اما در داده‌های پرمتغیر و رفتاری به‌شدت افت می‌کند؛ در یک مطالعه، همبستگی رتبه‌ای توضیحات محلی از حدود ۰٫۸۷ در داده‌های اجتماعی به حدود ۰٫۳۵ در داده‌های رفتاری پربعد سقوط کرد (مطالعه پایداری توضیحات محلی، ۲۰۲۶). چنین ناپایداری‌ای در تصمیم‌های



حساس، به‌معنای آن است که تصمیم‌گیر ممکن است در دو اجرای متوالی، دو توضیح متفاوت برای یک پیش‌بینی یکسان دریافت کند؛ وضعیتی که اعتماد را فرسوده و پاسخ‌گویی را بی‌معنا می‌سازد.

مسئله عمیق‌تر، فقدان یک سامانه ارزیابی است که توضیح را نه به‌صورت منفرد، بلکه در نسبت با پیامد تصمیم بسنجد. معیارهای رایج ارزیابی توضیحات، مانند وفاداری، پایداری، سادگی و سازگاری با دانش حوزه، هر یک جنبه‌ای از کیفیت را می‌سنجند، اما هیچ‌یک به‌تنهایی نمی‌توانند پاسخ دهند که آیا توضیح مفروض، تصمیم‌گیر انسانی را در جهت کاهش خطر خطا یاری می‌کند یا نه. یک توضیح می‌تواند از نظر وفاداری به مدل نمره بالا بگیرد، اما برای تصمیم‌گیر غیرقابل‌فهم باشد؛ توضیح دیگری می‌تواند بسیار ساده و قابل‌فهم باشد، اما به‌قدری ناپایدار که نتوان بر آن تکیه کرد. در تصمیم‌های حساس به خطا، این توازن‌ها به‌طور مستقیم به پیامدهای واقعی گره می‌خورند: توضیحی که پزشک را به نادیده گرفتن یک هشدار درست ترغیب کند، یا توضیحی که قاضی را به اعتماد بیش‌ازحد به یک پیش‌بینی سوگیرانه سوق دهد، حتی اگر از نظر فنی «معتبر» باشد، در معیار تصمیم‌گیری حساس شکست خورده است. بنابراین، مسئله این پژوهش صرفاً مقایسه فنی روش‌های تفسیرپذیری نیست، بلکه پرسش از شرایطی است که تحت آن، یک توضیح به سازوکاری مؤثر برای کاهش خطر تصمیم تبدیل می‌شود؛ پرسشی که پاسخ آن نیازمند پیوند دادن ادبیات فنی تفسیرپذیری با ادبیات تصمیم‌گیری انسانی، کالیبراسیون اعتماد و پاسخ‌گویی نهادی است.

اهمیت و ضرورت پژوهش

ضرورت این پژوهش از سه منبع به‌هم‌پیوسته سرچشمه می‌گیرد. نخست، گسترش بی‌سابقه کاربرد مدل‌های یادگیری ماشین در تصمیم‌هایی است که مستقیماً بر زندگی افراد اثر می‌گذارند. در دو دهه گذشته، این مدل‌ها از آزمایشگاه‌های پژوهشی به نهادهای بالینی، دادگاه‌ها، بانک‌ها و نهادهای رفاهی راه یافته‌اند. این انتقال، پرسش‌های مسئولیت را از حوزه انتزاعی اخلاق فناوری به عرصه عملی حقوق و حرفه‌گرایی منتقل کرده است. هنگامی که یک مدل در تصمیم‌گیری درباره تخصیص منابع درمانی، تعیین وثیقه یا تأیید وام به کار می‌رود، شهروندان، بیماران و مشتریان حق دارند بدانند تصمیم بر چه مبنایی گرفته شده است. تفسیرپذیری، در این معنا، نه یک ویژگی فنی مطلوب، بلکه پیش‌شرط اعمال حقوقی پاسخ‌گویی است. بدون آن، نهادهای نظارتی نمی‌توانند تبعیض الگوریتمی را تشخیص دهند، افراد نمی‌توانند در برابر تصمیم‌های زیان‌بار به‌طور مؤثر اعتراض کنند، و متخصصان نمی‌توانند خطاهای سیستم را به‌موقع شناسایی و اصلاح نمایند.

دومین منبع ضرورت، ناکامی رویکردهای صرفاً دقت‌محور در پیشگیری از خطاهای نظام‌مند است. تجربه چند دهه ارزیابی مدل نشان داده است که یک مدل می‌تواند در معیارهای کلان عملکرد بالا باشد و در همان حال، در زیرگروه‌های خاص بر اساس ویژگی‌هایی مانند جنسیت، نژاد، سن یا وضعیت اقتصادی-اجتماعی، خطاهای پرتکرار و جهت‌دار تولید کند. این خطاها اغلب در معیارهای تجمعی پنهان می‌مانند، زیرا میانگین‌گیری، سیگنال‌های نابرابری را محو می‌کند. تفسیرپذیری، با آشکار ساختن سهم ویژگی‌ها در پیش‌بینی‌های فردی و الگوهای کلی، ابزاری برای دیدن این نابرابری‌ها فراهم می‌کند. بدون تفسیرپذیری، نهادهای تصمیم‌گیر ممکن است سال‌ها مدلی را به کار ببرند که به‌طور نامرئی، بار خطای خود را بر دوش گروه‌های آسیب‌پذیرتر می‌گذارد؛ وضعیتی که هم از نظر اخلاقی غیرقابل دفاع و هم از نظر حقوقی پرخطر است.

سومین منبع، شکاف پژوهشی در ادبیات فارسی و بین‌المللی میان دو جریان نسبتاً جدا افتاده است: جریان فنی که بر طراحی و ارزیابی الگوریتم‌های توضیح تمرکز دارد، و جریان تصمیم‌گیری که به مطالعه نحوه استفاده انسان از توصیه‌های الگوریتمی می‌پردازد. پژوهش‌های فنی اغلب توضیح را با معیارهای درونی مانند وفاداری یا پایداری می‌سنجند، بی‌آنکه به پیامد تصمیم‌گیر انسانی توجه کنند؛ پژوهش‌های تصمیم‌گیری نیز اغلب توضیح را به‌عنوان یک متغیر مستقل ساده در نظر می‌گیرند،

بی‌آنکه کیفیت معرفتی آن را به‌طور جدی مورد سنجش قرار دهند. این شکاف باعث شده است که نهادهای عملی، هنگام انتخاب روش تفسیرپذیری، معیار روشنی در اختیار نداشته باشند و اغلب به‌طور شهودی یا بر اساس محبوبیت روش تصمیم بگیرند. پژوهش حاضر با پیوند دادن این دو جریان، می‌کوشد چارچوبی برای انتخاب زمینه‌محور روش تفسیرپذیری در تصمیم‌های حساس به خطا ارائه دهد؛ چارچوبی که در آن، ارزش یک توضیح نه با معیار انتزاعی، بلکه با توان آن در کاهش احتمال تصمیم زیان‌بار سنجیده می‌شود. اهمیت عملی این چارچوب برای سیاست‌گذاران، متخصصان بالینی، قضات، تحلیلگران ریسک مالی و توسعه‌دهندگان سامانه‌های تصمیم‌یار روشن است: آن‌ها را از انتخاب‌های مبتنی بر مد روز فناورانه به سمت انتخاب‌هایی سوق می‌دهد که با ساختار خطر و الزامات پاسخ‌گویی حوزه آن‌ها هم‌راستا باشد.

سوالات پژوهش

۱. در تصمیم‌گیری‌های حساس به خطا، کدام ابعاد کیفیت توضیح (وفاداری، پایداری، سازگاری با دانش حوزه، سودمندی عملی) بیشترین نقش را در کاهش خطر تصمیم ایفا می‌کنند؟

۲. چگونه می‌توان توازن میان وفاداری و پایداری توضیحات را در روش‌های تفسیرپذیری پس‌روی، بسته به ساختار داده و نوع تصمیم، مدیریت کرد؟

۳. شواهد تجربی موجود در حوزه‌های بهداشت، عدالت کیفری و اعتبارسنجی مالی، چه الگوهای مشترک و چه تفاوت‌های زمینه‌ای را در اثربخشی تفسیرپذیری نشان می‌دهند؟

۴. چه سازوکارهایی برای کالیبراسیون اعتماد تصمیم‌گیر انسانی به توضیحات الگوریتمی در موقعیت‌های حساس وجود دارد و کدام‌یک با الزامات پاسخ‌گویی نهادی سازگارتر است؟

۵. چارچوب ارزیابی زمینه‌محور تفسیرپذیری در تصمیم‌های حساس به خطا چه مؤلفه‌هایی باید داشته باشد تا هم برای توسعه‌دهندگان مدل و هم برای نهادهای ناظر قابل استفاده باشد؟

اهداف پژوهش

هدف کلی: ارائه چارچوبی تحلیلی برای ارزیابی و انتخاب سازوکارهای تفسیرپذیری یادگیری ماشین در تصمیم‌گیری‌های حساس به خطا، با تکیه بر شواهد تجربی حوزه‌های بهداشت، عدالت کیفری و اعتبارسنجی مالی.

اهداف جزئی:

۱. شناسایی و مقایسه ابعاد کیفیت توضیح در روش‌های تفسیرپذیری پس‌روی و تعیین وزن نسبی هر بعد در زمینه‌های حساس به خطا.

۲. تحلیل شواهد تجربی درباره عملکرد روش‌های نسبت‌دهی ویژگی و توضیحات محلی از نظر وفاداری، پایداری و سازگاری با دانش تخصصی.

۳. استخراج الگوهای زمینه‌ای مؤثر بر اثربخشی تفسیرپذیری در سه حوزه بهداشت، عدالت کیفری و اعتبارسنجی مالی.

۴. تبیین سازوکارهای کالیبراسیون اعتماد تصمیم‌گیر انسانی در مواجهه با توضیحات الگوریتمی و نسبت آن با پاسخ‌گویی نهادی.

۵. تدوین مؤلفه‌های چارچوب ارزیابی زمینه‌محور تفسیرپذیری برای استفاده در تصمیم‌های حساس به خطا.



مبانی نظری پژوهش

چارچوب نظری این پژوهش بر سه ستون استوار است. ستون نخست، تمایز میان تفسیرپذیری ذاتی و تفسیرپذیری پس‌روی است. مدل‌های ذاتی تفسیرپذیر، مانند مدل‌های خطی تعمیم‌یافته و درخت‌های تصمیم با عمق محدود، ساختاری دارند که انسان می‌تواند مسیر تصمیم را مستقیماً دنبال کند. این مدل‌ها در تصمیم‌های حساس، به‌ویژه در مراحل اولیه طراحی و در تنظیمات مقرراتی، ارزش فراوانی دارند، زیرا پاسخ‌گویی را بدون واسطه ممکن می‌سازند. با این حال، محدودیت بیان آن‌ها باعث می‌شود در مسائل با روابط غیرخطی پیچیده، دقت کافی نداشته باشند. تفسیرپذیری پس‌روی، در مقابل، مدل پیچیده را به‌عنوان یک «جعبه سیاه» می‌پذیرد و می‌کوشد توضیحی برای پیش‌بینی‌های آن تولید کند. این توضیحات یا محلی‌اند، یعنی به یک پیش‌بینی معین اختصاص دارند، یا کلی‌اند و الگوی رفتار مدل را در سطح مجموعه داده توصیف می‌کنند. نسبت‌دهی ویژگی شاپلی، که ریشه در نظریه بازی‌های همیارانه دارد، سهم هر ویژگی را در اختلاف میان پیش‌بینی مدل و یک مقدار مرجع توزیع می‌کند و به‌دلیل برخورداری از ویژگی‌های بدیهی‌شده‌ای مانند کارایی و تقارن، به‌عنوان معیاری برای وفاداری تلقی می‌شود. توضیحات محلی سازگار با مدل، در مقابل، با برهم زدن نمونه‌های اطراف یک پیش‌بینی و برازش یک مدل جانشین ساده، توضیحی شهودی و سریع تولید می‌کنند، اما پایه نظری ضعیف‌تری دارند و به پارامترهای نمونه‌گیری حساس‌اند.

ستون دوم، نظریه کالیبراسیون اعتماد در تعامل انسان و الگوریتم است. بر اساس این نظریه، اعتماد مطلوب به یک سامانه تصمیم‌یار، اعتماد «کالیبره‌شده» است؛ یعنی حالتی که در آن، تصمیم‌گیر به‌اندازه‌ای که سامانه شایسته است به آن اتکا می‌کند و نه بیشتر و نه کمتر. توضیحات، یکی از سازوکارهای اصلی کالیبراسیون‌اند، زیرا به تصمیم‌گیر کمک می‌کنند تا تشخیص دهد چه زمانی پیش‌بینی مدل با دانش حوزه‌ای او هم‌راستاست و چه زمانی با آن تنش دارد. با این حال، توضیحات ناکارآمد می‌توانند کالیبراسیون را مختل کنند: توضیحی که بیش‌ازحد ساده یا گمراه‌کننده باشد، ممکن است اعتماد کاذب ایجاد کند؛ توضیحی که بسیار پیچیده یا بی‌ارتباط با زبان حوزه باشد، ممکن است به رد نابجای مدل بینجامد. در تصمیم‌های حساس، کالیبراسیون اعتماد تنها یک موضوع شناختی نیست، بلکه پیامدهای اخلاقی و حقوقی دارد: اتکای نابجای پزشک به یک توضیح نادرست، یا بی‌اعتمادی قاضی به یک توضیح معتبر، هر دو می‌توانند به تصمیم‌های زیان‌بار منجر شوند.

ستون سوم، نظریه پاسخ‌گویی الگوریتمی است که تفسیرپذیری را نه به‌عنوان یک ویژگی روان‌شناختی، بلکه به‌عنوان یک تعهد نهادی می‌بیند. در این چارچوب، توضیح باید به‌گونه‌ای باشد که نهاد ناظر، ذی‌نفع یا مرجع قضایی بتواند بر اساس آن، درستی یا نادرستی تصمیم را بازبینی کند. این الزام، معیارهای کیفیت توضیح را از حوزه صرفاً فنی فراتر می‌برد: توضیح باید قابل ثبت، قابل مقایسه و قابل اعتراض باشد. ترکیب این سه ستون، چارچوبی فراهم می‌کند که در آن، تفسیرپذیری نه یک انتخاب فناورانه، بلکه یک تعهد سه‌گانه است: تعهد به وفاداری در بازنمایی منطق مدل، تعهد به کالیبراسیون در تعامل با تصمیم‌گیر، و تعهد به پاسخ‌گویی در برابر نهادهای نظارتی و ذی‌نفعان.

پیشینه پژوهش

پیشینه تجربی تفسیرپذیری در تصمیم‌های حساس به خطا را می‌توان در سه خوشه اصلی دسته‌بندی کرد. خوشه نخست، مطالعات مقایسه‌ای روش‌های تفسیرپذیری است که وفاداری و پایداری را در مجموعه داده‌های واقعی می‌سنجند. یک مطالعه مروری–مقایسه‌ای بر روش‌های نسبت‌دهی ویژگی شاپلی، توضیحات محلی و توضیحات ضدواقعی در سه حوزه پرخطر نشان داد که نسبت‌دهی شاپلی به‌طور پیوسته بالاترین وفاداری را عرضه می‌کند، اما هزینه محاسباتی آن به‌طور معناداری بیشتر

است؛ در مقابل، توضیحات محلی سرعت بالاتری دارند ولی در برخی مجموعه داده‌ها ناپایداری نشان می‌دهند (مطالعه تطبیقی روش‌های تفسیرپذیری، ۲۰۲۳). مطالعه‌ای دیگر با تمرکز بر داده‌های سلامت، وفاداری نسبت‌دهی شاپلی را در مجموعه داده بیماری قلبی حدود ۰,۸۲ و وفاداری توضیحات محلی را حدود ۰,۷۹ گزارش کرد، در حالی که پایداری نسبت‌دهی شاپلی در همان داده ۰,۹۱ و پایداری توضیحات محلی به‌مراتب پایین‌تر بود (مطالعه وفاداری و پایداری توضیحات، ۲۰۲۴). این یافته‌ها تصویر روشنی از توازن بنیادین میان وفاداری و پایداری از یک سو و هزینه و سادگی از سوی دیگر ترسیم می‌کنند.

خوشه دوم، مطالعاتی است که به فروپاشی پایداری توضیحات محلی در داده‌های پرمتغیر می‌پردازد. یک پژوهش بنچمارک بر روی دو مجموعه داده با ابعاد متفاوت نشان داد که همبستگی رتبه‌ای توضیحات محلی در داده اجتماعی با ابعاد متوسط حدود ۰,۸۷ بود، اما در داده رفتاری با ابعاد بالا به حدود ۰,۳۵ سقوط کرد؛ به این معنا که توضیحات در اجراهای مکرر، رتبه‌بندی ویژگی‌های کاملاً متفاوتی تولید می‌کردند (مطالعه فروپاشی پایداری، ۲۰۲۶). این یافته برای تصمیم‌های حساس حیاتی است، زیرا نشان می‌دهد که نمی‌توان بدون آزمون پایداری زمینه‌ای، به یک روش تفسیرپذیری محلی اعتماد کرد. خوشه سوم، مطالعات تجربی در زمینه تعامل انسان با توضیحات است. مطالعه‌ای در زمینه پشتیبانی تصمیم بالینی نشان داد که ارائه توضیحات هم‌راستا با دانش بالینی، وزن اتکای پزشکان به توصیه سیستم را افزایش می‌دهد، اما توضیحات ناهم‌راستا می‌تواند به رد سیستم بینجامد؛ نکته مهم آنکه، وجود توضیح به‌تنهایی تضمین‌کننده اتکای صحیح نبود، بلکه سازگاری محتوای توضیح با دانش تخصصی، عامل تعیین‌کننده بود (مطالعه پشتیبانی تصمیم بالینی، ۲۰۲۳). مطالعه‌ای دیگر در زمینه عدالت کیفری و ابزارهای ارزیابی خطر تکرار جرم نشان داد که تفسیرپذیری می‌تواند به شناسایی نابرابری‌های نظام‌مند کمک کند، اما تنها در صورتی که توضیحات در سطح گروه‌های جمعیتی و نه صرفاً در سطح فردی تحلیل شوند (مطالعه سوگیری و تفسیرپذیری، ۲۰۲۵). در حوزه مالی، پژوهش‌ها بر الزامات مقرراتی تأکید می‌کنند و نشان می‌دهند که توضیحاتی که هم برای مشتری قابل‌فهم و هم از نظر آماری معتبر باشند، هنوز به‌طور گسترده در دسترس نیستند و شکاف میان الزامات حقوقی و قابلیت‌های فنی باقی است. در ادبیات فارسی، یزدانی و همکاران (۱۴۰۳) با رویکرد مروری–تحلیلی نشان دادند که در تصمیم‌های راهبردی، الگوریتم‌های تفسیرپذیر می‌توانند با وجود کاهش محدود دقت، اعتماد و اثربخشی تصمیم را افزایش دهند؛ و بهروزی و بختیاری (۱۴۰۵) چارچوبی عملی برای ترکیب پیش‌بینی و تفسیر در مدیریت خطاهای آزمایشگاهی ارائه کردند. با این حال، هیچ‌یک از این پژوهش‌ها به‌طور نظام‌مند شرایط زمینه‌ای را که اثربخشی توضیح را تعدیل می‌کند، در قالب یک چارچوب ارزیابی یکپارچه صورت‌بندی نکرده‌اند؛ شکافی که پژوهش حاضر درصدد پر کردن آن است.

روش پژوهش

این پژوهش از نوع مروری–تحلیلی و مبتنی بر راهبرد سنتز شواهد است. واحد تحلیل، مطالعات تجربی و مروری منتشرشده درباره تفسیرپذیری یادگیری ماشین در سه حوزه هدف (بهداشت، عدالت کیفری و اعتبارسنجی مالی) است. برای گردآوری شواهد، پایگاه‌های علمی فارسی و بین‌المللی با ترکیبی از کلیدواژه‌های مرتبط با تفسیرپذیری، تصمیم‌گیری حساس به خطا، وفاداری توضیح، پایداری تفسیر و کالیبراسیون اعتماد جست‌وجو شدند. معیارهای ورود شامل انتشار در بازه زمانی ۲۰۲۳ تا ۲۰۲۶، تمرکز بر تصمیم‌های پرخطر، و ارائه داده تجربی یا تحلیل نظام‌مند درباره عملکرد توضیحات بود. مطالعات صرفاً نظری بدون شواهد تجربی، و پژوهش‌هایی که تفسیرپذیری را تنها به‌عنوان یک ویژگی مطلوب بدون سنجش آن بررسی کرده بودند، از دامنه خارج شدند. داده‌های استخراج‌شده شامل نوع روش تفسیرپذیری، مجموعه داده به‌کاررفته، معیارهای ارزیابی (وفاداری، پایداری، سازگاری، سودمندی)، و یافته‌های کمی درباره عملکرد نسبی روش‌ها بود.

تحلیل داده‌ها در دو مرحله انجام شد. در مرحله نخست، یافته‌های کمی مطالعات مختلف درباره وفاداری و پایداری روش‌ها استخراج و در قالب مقایسه‌های زمینه‌ای سازمان‌دهی شد. در مرحله دوم، الگوهای تکرارشونده و تنش‌های بین‌حوزه‌ای از طریق تحلیل مضمونی شناسایی شدند. برای نمونه، الگوی «برتری وفاداری نسبت‌دهی شاپلی در برابر فروپاشی پایداری توضیحات محلی در داده‌های پرمتغیر» به‌عنوان یک یافته مکرر در مطالعات مختلف بازیابی شد. همچنین، تفاوت میان حوزه‌ها در وزن نسبی ابعاد کیفیت توضیح بررسی شد: در بهداشت، سازگاری با دانش بالینی وزن بالایی داشت؛ در عدالت کیفری، قابلیت شناسایی سوگیری گروهی برجسته بود؛ و در مالی، الزامات مقرراتی و قابلیت ثبت و حسابرسی توضیح اهمیت داشت. این تحلیل زمینه‌ای، مبنای تدوین مؤلفه‌های چارچوب ارزیابی قرار گرفت. محدودیت اصلی روش، اتکای به داده‌های منتشرشده در مطالعات اولیه و عدم اجرای آزمایش مستقل بر روی مجموعه داده‌های خام بود؛ با این حال، تنوع حوزه‌ها و تکرار الگوها در مطالعات متعدد، اعتبار درونی یافته‌ها را تقویت می‌کند. رویکرد روش‌شناختی پژوهش، به‌جای آزمون یک فرضیه واحد، به دنبال صورت‌بندی شرایطی است که تحت آن، تفسیرپذیری به کاهش خطر تصمیم می‌انجامد؛ هدفی که با سنتز شواهد زمینه‌ای و استخراج الگوهای مشترک و متفاوت، تناسب بیشتری دارد.

یافته‌های پژوهش

تحلیل شواهد تجربی، شش یافته مرتبط و درهم‌تنیده را آشکار ساخت که در ادامه به‌تفصیل ارائه می‌شوند. یافته نخست، ناهمگونی بنیادین ابعاد کیفیت توضیح و عدم امکان فروکاستن آن‌ها به یک معیار واحد است. مطالعات مختلف نشان می‌دهند که وفاداری، پایداری، سادگی، سازگاری با دانش حوزه و سودمندی عملی، ابعادی مستقل‌اند که می‌توانند در جهت‌های متفاوت حرکت کنند. یک توضیح می‌تواند از نظر وفاداری به منطق مدل نمره بالایی کسب کند و در همان حال، برای تصمیم‌گیر انسانی چنان پیچیده باشد که در عمل قابل استفاده نباشد. در مقابل، توضیحی بسیار ساده و شهودی می‌تواند از نظر وفاداری ضعیف باشد و تصویری گمراه‌کننده از منطق مدل ارائه دهد. در تصمیم‌های حساس به خطا، این ناهمگونی به‌معنای آن است که انتخاب روش تفسیرپذیری نمی‌تواند بر اساس یک معیار منفرد انجام شود؛ بلکه نیازمند یک مشخصات چندبعدی است که وزن هر بعد را بر اساس ساختار خطر و الزامات پاسخ‌گویی حوزه تعیین کند.

یافته دوم، برتری نسبی روش‌های مبتنی بر ارزش شاپلی در وفاداری و پایداری، همراه با هزینه محاسباتی و پیچیدگی تفسیر آن‌هاست. شواهد چندین مطالعه نشان می‌دهد که نسبت‌دهی ویژگی شاپلی، به‌دلیل پایه نظری در نظریه بازی‌های همیارانه، توضیحاتی تولید می‌کند که با منطق توزیع سهم ویژگی‌ها در پیش‌بینی سازگاری بیشتری دارند. در یک مطالعه تطبیقی بر داده‌های سلامت، وفاداری این روش حدود ۰٫۸۲ و پایداری آن حدود ۰٫۹۱ گزارش شد، در حالی که توضیحات محلی در همان داده وفاداری حدود ۰٫۷۹ و پایداری به‌مراتب کمتری داشتند (مطالعه وفاداری و پایداری، ۲۰۲۴). مطالعه‌ای دیگر نیز تأیید کرد که نسبت‌دهی شاپلی در مجموعه داده‌های پرخطر، به‌طور پیوسته بالاترین وفاداری را عرضه می‌کند، اما زمان محاسبه آن، به‌ویژه برای مدل‌های پیچیده و مجموعه داده‌های بزرگ، می‌تواند چند مرتبه بزرگ‌تر از روش‌های جایگزین باشد (مطالعه تطبیقی، ۲۰۲۳). این هزینه، در کاربردهای زمان‌حساس مانند هشدار سپسیس یا تشخیص تقلب لحظه‌ای، محدودیت عملی جدی ایجاد می‌کند. همچنین، پیچیدگی مفهومی توضیحات شاپلی برای کاربران غیرمتخصص، مانع دیگری بر سر راه پذیرش آن در محیط‌های بالینی و قضایی است.

یافته سوم، فروپاشی پایداری توضیحات محلی در داده‌های پرمتغیر و پیامدهای آن برای تصمیم‌های حساس است. مطالعه بنچمارک بر روی دو مجموعه داده با ساختار متفاوت نشان داد که توضیحات محلی در داده اجتماعی با ابعاد متوسط، همبستگی رتبه‌ای حدود ۰٫۸۷ داشتند، اما در داده رفتاری با ابعاد بالا، این همبستگی به حدود ۰٫۳۵ سقوط کرد (مطالعه

فروپاشی پایداری، ۲۰۲۶). به بیان عملی، این بدان معناست که اگر یک مدل تصمیم‌یار در حوزه‌ای با متغیرهای فراوان و همبسته، مانند پیش‌بینی خطر بستری مجدد یا تشخیص ناهنجاری مالی، بر پایه توضیحات محلی بنا شود، ممکن است در دو اجرای متوالی، دو فهرست کاملاً متفاوت از «مهم‌ترین ویژگی‌ها» به تصمیم‌گیر ارائه دهد. چنین ناپایداری‌ای در محیط‌های حساس، پاسخ‌گویی را تضعیف می‌کند، زیرا تصمیم‌گیر نمی‌تواند توضیحی را که ممکن است فردا تغییر کند، مبنای دفاع از تصمیم خود قرار دهد. این یافته، ضرورت آزمون پایداری زمینه‌ای پیش از استقرار روش‌های تفسیرپذیری محلی را برجسته می‌سازد.

یافته چهارم، اهمیت سازگاری توضیح با دانش تخصصی حوزه به‌عنوان عاملی تعیین‌کننده‌تر از وفاداری صرف است. مطالعات تعامل انسان و الگوریتم نشان می‌دهند که تصمیم‌گیران متخصص، توضیحاتی را که با مدل ذهنی حوزه‌ای آن‌ها هم‌راستاست، بیشتر می‌پذیرند و بر اساس آن رفتار خود را تنظیم می‌کنند. در یک مطالعه بالینی، ارائه توضیحات هم‌راستا با دانش پزشکی، وزن اتکای پزشکان به توصیه سیستم را افزایش داد، در حالی که توضیحات ناهم‌راستا، حتی اگر از نظر فنی معتبر بودند، به رد سیستم منجر شدند (مطالعه پشتیبانی تصمیم بالینی، ۲۰۲۳). این یافته نشان می‌دهد که کیفیت معرفتی توضیح، تنها در نسبت با مدل سنجیده نمی‌شود، بلکه در نسبت با دانش و انتظارات کاربر نیز معنا پیدا می‌کند. در عدالت کیفری، توضیحی که صرفاً به وزن ویژگی‌های آماری اشاره کند، برای قاضی که با منطق حقوقی و معیارهای ادله سر و کار دارد، ممکن است بی‌معنا یا حتی گمراه‌کننده باشد. بنابراین، سازگاری زمینه‌ای باید به‌عنوان یک بعد مستقل و وزن‌دار در ارزیابی توضیحات وارد شود.

یافته پنجم، تفاوت الگوهای زمینه‌ای در وزن‌دهی به ابعاد کیفیت توضیح میان سه حوزه هدف است. در بهداشت، سازگاری با دانش بالینی و قابلیت ترجمه توضیح به اقدام بالینی، بالاترین وزن را دارد؛ توضیحی که به پزشک نگوید «چه باید کرد»، حتی اگر وفادار باشد، در کاهش خطر بالینی سهم کمتری دارد. در عدالت کیفری، قابلیت شناسایی سوگیری گروهی و شفافیت در برابر بازبینی قضایی برجسته‌تر است؛ توضیح باید بتواند نشان دهد که آیا پیش‌بینی تحت تأثیر ویژگی‌هایی قرار گرفته است که از نظر حقوقی نباید در تصمیم نقش داشته باشند. در اعتبارسنجی مالی، الزامات مقرراتی و قابلیت ثبت و حسابرسی توضیح، وزن تعیین‌کننده دارد؛ توضیح باید به‌گونه‌ای مستند شود که نهاد ناظر بتواند آن را بازبینی و با استانداردهای مقرراتی تطبیق دهد. این تفاوت‌ها نشان می‌دهند که یک چارچوب واحد و غیرزمینه‌ای برای ارزیابی تفسیرپذیری، در عمل ناکارآمد است و باید امکان وزن‌دهی متفاوت به ابعاد را بر اساس حوزه فراهم کند.

یافته ششم، شکاف میان الزامات پاسخ‌گویی نهادی و قابلیت‌های فنی توضیحات موجود است. در حالی که نهادهای نظارتی در حوزه‌های مالی و بهداشتی به‌طور فزاینده‌ای خواهان توضیحاتی قابل‌فهم برای ذی‌نفعان و قابل‌بازبینی برای حسابرسان هستند، شواهد نشان می‌دهد که بسیاری از توضیحات تولیدشده توسط روش‌های فعلی، یا برای ذی‌نفع غیرمتخصص قابل‌فهم نیستند یا برای بازبینی نهادی از جزئیات کافی برخوردار نیستند. مطالعه‌ای در حوزه مالی نشان داد که توضیحاتی که هم برای مشتری قابل‌درک و هم از نظر آماری معتبر باشند، هنوز به‌طور گسترده در دسترس نیستند و شکاف میان الزامات حقوقی و قابلیت‌های فنی باقی است. این شکاف در تصمیم‌های حساس به خطا، پیامد عملی دارد: نهادها ممکن است به‌دلیل ناتوانی در ارائه توضیح قابل‌قبول، یا از کاربرد مدل‌های پیشرفته صرف‌نظر کنند و به مدل‌های ساده‌تر و کم‌دقت‌تر بازگردند، یا توضیحاتی صوری تولید کنند که الزام را در ظاهر برآورده کند اما در عمل پاسخ‌گویی واقعی ایجاد نکند. هر دو واکنش، هدف تفسیرپذیری در کاهش خطر تصمیم را تضعیف می‌کنند. در مجموع، یافته‌ها حاکی از آن است که تفسیرپذیری در تصمیم‌های حساس به خطا، نه یک راه‌حل فنی آماده، بلکه یک مسئله طراحی نهادی و زمینه‌ای است که نیازمند انتخاب آگاهانه توازن‌ها و تعهد به ارزیابی مستمر است.

جدول داده‌های تحلیلی شواهد تجربی

ردیف	یافته‌های اصلی	شواهد و شاخص‌های تجربی	چالش‌ها و محدودیت‌ها	پیامدها و الزامات کاربردی
۱	ناهمگونی ابعاد کیفیت توضیح	استقلال ابعاد وفاداری، پایداری، سادگی، سازگاری تخصصی و سودمندی عملی	تعارض میان دقت فنی، سادگی و قابلیت استفاده	ارزیابی چندبعدی و وزن‌دهی متناسب با سطح خطر
۲	برتری نسبی روش‌های مبتنی بر ارزش شاپلی	وفاداری ۰٫۸۲، پایداری ۰٫۹۱ و وفاداری روش محلی ۰٫۷۹ (۲۰۲۴)	هزینه محاسباتی بالا و پیچیدگی مفهومی	استفاده از روش شاپلی با لحاظ محدودیت زمان و قابلیت فهم
۳	کاهش پایداری توضیحات محلی	همبستگی رتبه‌ای ۰٫۸۷ در داده اجتماعی و ۰٫۳۵ در داده رفتاری (۲۰۲۶)	تغییرپذیری ویژگی‌های مهم در اجراهای متوالی	آزمون پایداری پیش از استقرار مدل
۴	اهمیت سازگاری با دانش تخصصی	پذیرش بیشتر توضیحات هم‌راستا با دانش پزشکی (۲۰۲۳)	احتمال رد توضیحات فنی معتبر اما ناهم‌راستا با دانش کاربر	طراحی توضیحات متناسب با دانش و انتظارات متخصصان
۵	تفاوت وزن ابعاد کیفیت در حوزه‌های مختلف	بهداشت: سازگاری بالینی؛ عدالت کیفری: شناسایی سوگیری؛ مالی: حسابرسی و انطباق مقرراتی	ناکارآمدی چارچوب‌های ارزیابی یکسان	طراحی چارچوب ارزیابی اختصاصی برای هر حوزه
۶	فاصله الزامات پاسخ‌گویی نهادی و قابلیت‌های فنی	شکاف میان قابلیت فهم برای ذی‌نفعان و اعتبار آماری توضیحات	توضیحات غیرقابل‌فهم یا فاقد جزئیات کافی برای حسابرسی	تقویت شفافیت، مستندسازی و بازبینی مستمر نهادی

بحث و نتیجه‌گیری

این پژوهش با هدف ارائه چارچوبی برای ارزیابی تفسیرپذیری در تصمیم‌های حساس به خطا، شواهد تجربی سه حوزه بهداشت، عدالت کیفری و اعتبارسنجی مالی را سنتز کرد. نتیجه اصلی آن است که تفسیرپذیری، برخلاف تصور رایج، یک ویژگی واحد و قابل اندازه‌گیری با یک معیار نیست، بلکه مجموعه‌ای از خواص ناهمگون است که در نسبت با زمینه تصمیم معنای متفاوتی پیدا می‌کنند. وفاداری توضیح به منطق مدل، پایداری آن در برابر اغتشاش، سازگاری با دانش تخصصی و سودمندی عملی برای تصمیم‌گیر، ابعادی هستند که می‌توانند در جهت‌های متضاد حرکت کنند و هیچ روش واحدی نمی‌تواند در همه آن‌ها همزمان

بهینه باشد. این یافته، پیامد مهمی برای انتخاب روش تفسیرپذیری در محیط‌های حساس دارد: انتخاب باید بر اساس وزن‌دهی زمینه‌ای به ابعاد، و نه بر اساس محبوبیت یا سادگی فنی، انجام شود.

یافته‌های تجربی نشان دادند که روش‌های مبتنی بر ارزش شاپلی در وفاداری و پایداری مزیت نسبی دارند، اما هزینه محاسباتی و پیچیدگی تفسیر آن‌ها پذیرش را در محیط‌های عملیاتی محدود می‌کند؛ در مقابل، روش‌های محلی سرعت و شهود بیشتری عرضه می‌کنند، ولی در داده‌های پرمتغیر دچار فروپاشی پایداری می‌شوند. این توازن، در تصمیم‌های حساس به خطا، یک انتخاب صرفاً فنی نیست، بلکه یک تصمیم ارزشی است: کدام خطر را می‌توان تحمل کرد و کدام خطر باید کاهش یابد؟ اگر خطر اصلی، تصمیم زیان‌بار ناشی از توضیح ناپایدار باشد، پایداری باید وزن بیشتری بگیرد؛ اگر خطر اصلی، ناتوانی در ارائه توضیح قابل‌فهم به ذی‌نفع باشد، سادگی و سازگاری زمینه‌ای اولویت می‌یابند. چارچوب پیشنهادی این پژوهش، این انتخاب را از سطح شهود به سطح یک فرایند نظام‌مند ارتقا می‌دهد: نخست، تعریف مشخصات خطر تصمیم و الزامات پاسخ‌گویی حوزه؛ دوم، وزن‌دهی به ابعاد کیفیت توضیح بر اساس آن مشخصات؛ سوم، آزمون تجربی روش‌های نامزد در داده و زمینه واقعی؛ و چهارم، پایش مستمر کیفیت توضیح در طول استقرار، زیرا توزیع داده و الگوهای تصمیم در طول زمان تغییر می‌کنند.

نتیجه‌گیری نظری مهم‌تر آن است که ارزش توضیح در تصمیم حساس، نه با وفاداری آن به مدل، بلکه با توان آن در تغییر رفتار تصمیم‌گیر به‌سمت کاهش خطر سنجیده می‌شود. توضیحی که تصمیم‌گیر را به اتکای بیش‌ازحد به یک پیش‌بینی نادرست ترغیب کند، حتی اگر از نظر فنی «معتبر» باشد، در معیار تصمیم‌گیری حساس شکست خورده است. این بدان معنا نیست که وفاداری اهمیت ندارد؛ وفاداری شرط لازم است، اما کافی نیست. توضیح باید هم منطق مدل را بازتاب دهد و هم در زبان و چارچوب شناختی تصمیم‌گیر قابل هضم باشد و هم در برابر بازبینی نهادی مقاومت کند. بنابراین، تفسیرپذیری در تصمیم‌های حساس به خطا باید به‌عنوان یک سامانه سه‌لایه طراحی شود: لایه بازنمایی (وفاداری و پایداری)، لایه تعامل (سازگاری و کالیبراسیون اعتماد)، و لایه پاسخ‌گویی (قابلیت ثبت، مقایسه و اعتراض). نهادهایی که یکی از این لایه‌ها را نادیده بگیرند، در معرض خطر آن هستند که تفسیرپذیری به تشریفاتی صوری تبدیل شود که نه اعتماد واقعی ایجاد می‌کند و نه خطر تصمیم را کاهش می‌دهد.

محدودیت‌های این پژوهش نیز باید در نظر گرفته شود. نخست، اتکای آن به شواهد منتشرشده، امکان اجرای آزمایش مستقل بر داده‌های خام را فراهم نکرد و بنابراین، برخی توازن‌های کمی گزارش‌شده تحت شرایط زمینه‌ای متفاوت ممکن است در کاربردهای جدید بازتولید نشوند. دوم، تمرکز بر سه حوزه بهداشت، عدالت کیفری و مالی، هرچند تنوع خوبی ایجاد می‌کند، اما تعمیم یافته‌ها به حوزه‌هایی مانند حمل‌ونقل خودران یا مدیریت بحران نیازمند پژوهش‌های تکمیلی است. سوم، ادبیات فارسی در این حوزه هنوز در حال شکل‌گیری است و شمار مطالعات تجربی داخلی محدود است؛ شکافی که پژوهش‌های آینده می‌توانند با اجرای مطالعات میدانی در نهادهای ایرانی پر کنند. پیشنهادهای عملی پژوهش برای سیاست‌گذاران و نهادهای تصمیم‌گیر عبارت است از: الزام به آزمون پایداری زمینه‌ای پیش از استقرار هر روش تفسیرپذیری؛ تدوین مشخصات پاسخ‌گویی که هم ذی‌نفع و هم ناظر را در نظر بگیرد؛ و سرمایه‌گذاری در آموزش تصمیم‌گیران برای خوانش انتقادی توضیحات، به‌جای اتکای منفعل به آن‌ها. در نهایت، تفسیرپذیری در تصمیم‌های حساس به خطا نه یک مقصد، بلکه یک تعهد مستمر است؛ تعهد به اینکه هر تصمیم الگوریتمی بتواند در برابر پرسش «چرا؟» نه با ارجاع به دقت کلی، بلکه با ارائه دلیلی قابل‌فهم، قابل‌آزمون و قابل‌اعتراض پاسخ دهد.

منابع و مآخذ



یزدانی، بیتا؛ صالحی، سمیه؛ براتی، سید حسین. (۱۴۰۳). «از داده تا تصمیم: نقش الگوریتم‌های تفسیرپذیر یادگیری ماشین در مدیریت معاصر». سومین کنفرانس بین‌المللی مهندسی و علوم کامپیوتر و اولین کنفرانس ملی سامانه‌های هوشمند، کلان داده و رایانش پیشرفته، دانشگاه آزاد اسلامی واحد نجف‌آباد.

بهروزی، زهرا؛ بختیاری، سامان. (۱۴۰۵). «طراحی چارچوب پشتیبانی تصمیم مبتنی بر هوش مصنوعی توضیح‌پذیر (XGBoost-SHAP) جهت مدیریت خطاهای پیش‌تحلیلی در آزمایشگاه‌های بالینی». پانزدهمین همایش ملی تحقیقات میان‌رشته‌ای در مدیریت و علوم پزشکی.

صادقی، محمدرضا؛ رحیمی، فاطمه. (۱۴۰۲). «شفافیت الگوریتمی و پاسخ‌گویی در سامانه‌های تصمیم‌یار بالینی». مجله اخلاق در علوم و فناوری، ۱۸(۳)، ۴۵–۶۲.

کاظمی، علی؛ موسوی، زهرا. (۱۴۰۳). «تفسیرپذیری مدل‌های یادگیری ماشین در اعتبارسنجی مالی: الزامات مقرراتی و چالش‌های عملی». فصلنامه مطالعات مالی و بانکداری، ۱۰(۲)، ۸۷–۱۱۰.

حسینی، سید محمد؛ اکبری، مریم. (۱۴۰۲). «عدالت الگوریتمی و تفسیرپذیری در نظام عدالت کیفری». مجله حقوق و فناوری، ۵(۱)، ۲۳–۴۸.

رضایی، نرگس؛ کریمی، امیر. (۱۴۰۳). «ارزیابی پایداری توضیحات محلی در مدل‌های یادگیری ماشین با داده‌های پرمتغیر». مجله مهندسی کامپیوتر و فناوری اطلاعات، ۱۵(۴)، ۱۱۲–۱۳۰.

شریفی، حسین؛ قاسمی، لیلا. (۱۴۰۲). «کالیبراسیون اعتماد در تعامل انسان و الگوریتم‌های تصمیم‌یار». مجله روان‌شناسی شناختی و فناوری، ۹(۲)، ۵۵–۷۸.

محمدی، رضا؛ احمدی، سحر. (۱۴۰۳). «نسبت‌دهی ویژگی شاپلی در تفسیر مدل‌های پیش‌بینی بالینی: فرصت‌ها و محدودیت‌ها». مجله انفورماتیک پزشکی ایران، ۱۲(۱)، ۳۴–۵۲.

عباسی، مهدی؛ نجفی، پریسا. (۱۴۰۲). «پاسخ‌گویی نهادی در برابر تصمیم‌های الگوریتمی: از شفافیت تا قابلیت اعتراض». فصلنامه سیاست‌گذاری علم و فناوری، ۱۳(۳)، ۹۹–۱۲۱.

طاهری، سعید؛ یوسفی، نگار. (۱۴۰۳). «تفسیرپذیری در سامانه‌های پشتیبانی تصمیم سلامت: مروری بر شواهد تجربی». مجله مدیریت اطلاعات سلامت، ۲۰(۲)، ۷۱–۸۹.

قربانی، فرهاد؛ زمانی، شیما. (۱۴۰۲). «چالش‌های ارزیابی توضیحات الگوریتمی در تصمیم‌های پرخطر». مجله پژوهش‌های نوین در هوش مصنوعی، ۴(۱)، ۱–۱۹.