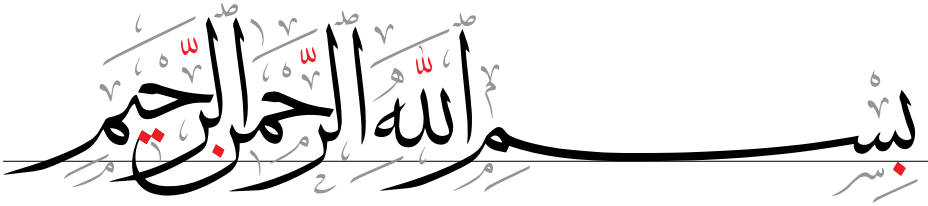
مرکز پژوهش‌های مجلس شورای اسلامی
دفتر مطالعات حکمرانی
مسلسل: ۲۰۹۶۸ ■ شهریور ۱۴۰۴

حکمرانی هوش مصنوعی (۸):
هوش مصنوعی توضیح‌پذیر و نقش آن در ورود فناوری هوش مصنوعی به بخش عمومی


بسم الله الرحمن الرحيم

تاریخ انتشار:
۱۴۰۴/۶/۲۵

مرکز پژوهش‌های مجلس شورای اسلامی

شماره مسلسل: ۲۰۹۶۸
کد موضوعی: ۳۵۰

عنوان گزارش:
حکمرانی هوش مصنوعی (۸): هوش مصنوعی توضیح‌پذیر و نقش آن در ورود فناوری هوش مصنوعی به بخش عمومی

نوع گزارش: طرح ولایحه□ ، نظارتی□ ، راهبردی■ ،پیش‌نویس قانونی□

نام دفتر:
مطالعات بنیادین حکمرانی (گروه سیاست پژوهی و آزمایشگاه حکمرانی)

تهیه و تدوین:
ایمان اکبری (گروه سیاست‌پژوهی و آزمایشگاه حکمرانی)

مدیر مطالعه:
حسین بابایی مجرد

اظهارنظرکنندگان:
یحیی مرتب (دفتر مطالعات مدیریت، گروه دولت الکترونیک و مدیریت داده)، سهیلا خردمندنیا، زهرا جعفری (دفتر مطالعات صنعت، معدن و انرژی، گروه توسعه فناوري و تولید دانش‌بنیان)

اظهارنظرکننده خارج از مرکز:
احسان دهقانی (عضو هیئت‌علمی دانشگاه علم و صنعت ایران)

ناظر علمی:
مهدی عبدالحمید

گرافیک و صفحه آرایی:
سیده فاطمه ابوطالبی
ساجده زارع مرزی

ویراستار ادبی:
سیده مرضیه موسوی راد

واژه‌های کلیدی:
۱. هوش مصنوعی توضیح‌پذیر
۲. حکمرانی هوش مصنوعی
۳. هوش مصنوعی تفسیرپذیر
۴. هوش مصنوعی قابل اعتماد
۵. هوش مصنوعی اخلاق‌مدار

تاریخ شروع مطالعه:
۱۴۰۲/۰۹/۰۱

دفتر مطالعات حکمرانی

فهرست مطالب

فهرست شکل ها

فهرست جدول

حکمرانی هوش مصنوعی(۸): هوش مصنوعی توضیح‌پذیر و نقش آن در ورود فناوری هوش مصنوعی به بخش عمومی

Doi: 10.22034/report.mrc.2025.1404.33.6.20968

چکیده

هوش مصنوعی روزبه‌روز در جنبه‌های مختلف زندگی روزمره وارد شده و بر تصمیم‌گیری‌ها در بخش‌های گوناگون تأثیر می‌گذارد. این فناوری به‌علت قابلیت هوشمندی، خودمختاری و تصمیم‌گیری در بسیاری از حوزه‌ها، نتایج خارق‌العاده‌ای ایجاد کرده و به‌شدت مورد توجه قرار گرفته است. بااین‌حال ورود این فناوری به بخش عمومی به دلایلی نیازمند دقت و احتیاط است. الزاماتی مانند اهمیت شفافیت، پاسخ‌گویی و مسئولیت‌پذیری عاملان، عمل در محدوده قوانین و مقررات، اقناع عمومی نسبت به تصمیم‌های اتخاذ شده، اطمینان بالا نسبت به نتایج تصمیم‌ها و تنوع مسائل و تخصص‌های مورد نیاز، تصمیم‌گیری و عمل در این بخش را واجد اقتضائاتی خاص کرده است. بنابراین ورود فناوری هوش مصنوعی به بخش عمومی مسیری خاص و ویژه را می‌طلبد، به‌نحوی‌که توجه به الزامات بیان شده از بخش عمومی را تضمین کند. این گزارش، هوش مصنوعی توضیح‌پذیر را که شاخه‌ای نوین در حوزه علوم هوش مصنوعی است، به‌عنوان دریچه ورود این فناوری به بخش عمومی معرفی می‌کند. هدف هوش مصنوعی توضیح‌پذیر، ایجاد مدل‌ها و سیستم‌هایی است که تصمیم‌ها و خروجی‌هایشان برای انسان‌ها قابل توضیح و فهم باشد. این قابلیت توضیح می‌تواند تضمین‌کننده استفاده صحیح این فناوری در این بخش باشد. دستیابی به تعادل بین عملکرد و قابل توضیح بودن، یک چالش اصلی است که نیاز به بررسی دقیق و رویکردهای میان‌رشته‌ای دارد. علاوه‌بر این، عوامل انسانی نقش اساسی در اجرای موفقیت‌آمیز هوش مصنوعی توضیح‌پذیر دارند. گزارش حاضر پس از بررسی این حوزه و کاربردها و چالش‌های آن، به بیان توصیه‌هایی جهت توسعه این رویکرد پرداخته است.

خلاصه مدیریتی

بیان/شرح مسئله

امروزه فناوری هوش مصنوعی برای هیچ فردی ناآشنا نیست. پس از توسعه مدل‌های زبانی بزرگ و شاخه هوش مصنوعی مولد و ارائه چت‌بات‌هایی مانند چت جی‌پی‌تی و جمینی و تولید تصاویر و فیلم و خدمات متعدد دیگر توسط سامانه‌های مبتنی‌بر هوش مصنوعی، در حال حاضر کمتر کسی است که با این فناوری مواجهه نداشته باشد. این حوزه به شکل روزافزون در جنبه‌های مختلف زندگی روزمره وارد شده و بر اقدامات و تصمیم‌گیری‌ها در بخش‌های گوناگون تأثیر گذاشته است. نقطه قوت این فناوری که موجب تحولات گسترده و روزافزون شده، قابلیت یادگیرندگی، هوشمندی، خودمختاری و تصمیم‌گیری است که در بسیاری از حوزه‌ها، نتایج جذاب و خارق‌العاده‌ای ایجاد کرده است. امروزه در بخش عمومی نیز در حوزه‌هایی مانند پزشکی و سلامت، حمل‌ونقل، حقوق، مالی و بانکی و... نمونه‌هایی از ورود فناوری هوش مصنوعی قابل مشاهده است. بااین‌حال ورود این فناوری به بخش عمومی نیازمند دقت و احتیاط ویژه است. هرگونه تصمیم‌گیری و اقدام در این بخش، مستلزم تضمین شفافیت، پاسخ‌گویی و مسئولیت‌پذیری، رعایت قوانین و مقررات، اقناع عمومی نسبت به تصمیم‌های اتخاذ شده و اطمینان از نتایج تصمیم‌ها است. علاوه‌بر این تنوع مسائل و تخصص‌های مورد نیاز در این بخش نیز بر پیچیدگی این شرایط افزوده است. بنابراین ورود فناوری هوش مصنوعی به بخش عمومی مانند بخش خصوصی و تجاری نیست، بلکه ضروری است با ملاحظاتی خاص و از مسیر و طریقی ویژه به‌نحوی‌که الزامات بیان شده را تضمین کند، پیگیری شود. این گزارش به هوش مصنوعی توضیح‌پذیر که شاخه‌ای نوین در حوزه علوم هوش مصنوعی است و بر شفافیت و قابل‌فهم بودن عملکرد مدل‌های هوش مصنوعی تمرکز دارد، خواهد پرداخت و این شاخه را به‌عنوان دریچه ورود فناوری هوش مصنوعی به بخش عمومی معرفی خواهد کرد. هدف هوش مصنوعی توضیح‌پذیر این است که به کاربران اجازه درک اینکه یک مدل هوش مصنوعی چگونه به یک نتیجه خاص رسیده است را بدهد. این امر به‌ویژه در حوزه‌هایی که تصمیمات مدل‌ها می‌تواند تأثیرات قابل‌توجهی بر زندگی انسان‌ها داشته باشد، مانند پزشکی، مالی و حقوق، بسیار مهم است.

نقطه‌نظرات/ یافته‌های کلیدی

این گزارش پس از معرفی این شاخه نوین، بیان می‌کند که برخلاف مدل‌های معمول هوش مصنوعی که عملکرد داخلی آنها نامشخص و نامفهوم بوده، هوش مصنوعی توضیح‌پذیر بر ایجاد سیستم‌هایی متمرکز است که قادر به ارائه توضیحات واضح و قابل درک برای انسان‌ها درباره نحوه رسیدن به نتایج خود هستند. هرچند مفهوم و الزامات توضیح‌پذیری ممکن است بسته به زمینه یا مخاطب خاص، قابل تغییر باشد.

توضیح‌پذیری می‌تواند دستاوردهای متعددی داشته باشد. مانند:

- توانمندسازی افراد برای مقابله با پیامد منفی تصمیم‌گیری خودکار؛
- ارتقای پذیرش نتایج تصمیمات خودکار با فهم شیوه حصول نتایج و انتخاب آگاهانه آنها؛
- آشکارسازی تهدیدات احتمالی و محافظت از آسیب‌پذیری‌های امنیتی؛
- ایجاد زمینه ادغام الگوریتم‌ها با ارزش‌های انسانی؛
- ارتقای استانداردهای صنعت برای توسعه محصولات مبتنی‌بر هوش مصنوعی؛
- شناخت نقاط ضعف الگوریتم‌ها و امکان ارتقای آنها؛
- امکان شهود ایده‌های جدید براساس عملکرد الگوریتم‌ها؛
- امکان تحقق حق توضیح برای شهروندان در هنگام استفاده از داده‌هایشان.

به‌کارگیری فناوری هوش مصنوعی در بخش عمومی اگر با الزام توضیح‌پذیری و با رعایت این ویژگی محقق شود، می‌تواند شرایط و ملاحظات تصمیم‌گیری در این بخش ازجمله شفافیت، پاسخ‌گویی و مسئولیت‌پذیری، رعایت قوانین و مقررات، اقناع عمومی نسبت به تصمیم‌های اتخاذ شده، اطمینان از نتایج تصمیم‌ها را تضمین کرده و به ارتقای کارآمدی بخش عمومی منجر شود. در غیر این صورت، عدم سازگاری ویژگی‌های تصمیمات خودکار و الگوریتمی، یا موجب آسیب رساندن به ارزش‌های عمومی مانند شفافیت و پاسخ‌گویی و حتی سوءاستفاده احتمالی افراد شده، یا بر اثر مقاومت و واپس زدن این فناوری، بخش عمومی از ظرفیت بالقوه این فناوری نوین بی‌بهره خواهد شد.

هوش مصنوعی توضیح‌پذیر می‌تواند کاربردهای متعددی در بخش‌های مختلف مانند بهداشت و درمان، مالی و اقتصادی، حقوقی، حمل‌ونقل، وسایل نقلیه خودران، آموزش، رسانه و سرگرمی، خرده‌فروشی و مدیریت منابع انسانی داشته باشد. همچنین امکان تعمیم‌پذیری، ارتقای امنیت الگوریتم‌ها و استحکام الگوریتمی از دستاوردهای جانبی توجه به توضیح‌پذیری در الگوریتم‌هاست. در ضمن توجه به این شاخه از هوش مصنوعی و توسعه آن می‌تواند در بلندمدت منجر به افزایش اعتماد به هوش مصنوعی، بهبود تصمیم‌گیری‌ها و تحقق هوش مصنوعی اخلاقی و عادلانه و رشد اقتصادی و نوآوری در سطح جامعه بشود.

شایان ذکر است که توسعه این شاخه از هوش مصنوعی با چالش‌هایی مانند چالش فنی و پیچیدگی بالا، توازن بین عملکرد و توضیح‌پذیری و عوامل انسانی و شناختی مواجه بوده که ضروری است مدنظر قرار گیرند.

پیشنهاد راهکارهای تقنینی، نظارتی یا سیاستی

نظر به اهمیت و دستاوردهای هوش مصنوعی توضیح‌پذیر و همچنین اهمیت بهره‌مندی از ظرفیت فناوری هوش مصنوعی در بخش‌های مختلف و به‌طور خاص بخش عمومی، توصیه‌های زیر پیشنهاد می‌شود:

- الزام سیستم‌های هوش مصنوعی مورد استفاده در بخش‌های مهم و عمومی به تحقق درجاتی از توضیح‌پذیری و ایجاد چارچوب‌های قانونی و اخلاقی؛
- ایجاد استانداردها و پروتکل‌هایی برای ارزیابی و تضمین کیفیت سیستم‌های هوش مصنوعی توضیح‌پذیر به همراه سازوکارهای نظارتی مناسب در دستگاه‌های مربوطه مانند سازمان ملی استاندارد؛
- تشویق به پذیرش استانداردهای توضیح‌پذیری در بخش صنعتی و ایجاد نهادهای نظارتی یا کمیته‌های تخصصی ناظر در بخش خصوصی؛
- قرار دادن تسهیلات و بودجه برای تحقیق و توسعه در زمینه توضیح‌پذیری الگوریتم‌ها و مدل‌ها بالاخص در تحقیقات بین‌رشته‌ای در حوزه‌های علوم‌شناختی، روان‌شناسی، علوم رایانه و هوش مصنوعی در بخش دانشگاهی و پژوهشگاهی؛
- تسهیل برنامه‌های آموزشی برای قانونگذاران، توسعه‌دهندگان و عموم مردم برای افزایش آگاهی و درک فناوری‌های هوش مصنوعی و پرداختن به نیازها و ترجیحات متنوع کاربران مختلف و افزایش آگاهی آنها با هدف افزایش تقاضا برای راه‌حل‌های شفاف هوش مصنوعی؛
- حمایت از نوآوری ایجاد مشوق‌ها و تسهیلات مالی برای شرکت‌های نوپا و کارآفرینان در حوزه هوش مصنوعی توضیح‌پذیر؛
- در نظر داشتن رویکردی قابل انعطاف و آینده‌نگر در حوزه اسناد و قوانین مرتبط با فناوری هوش مصنوعی جهت ارتقای ظرفیت تطبیق با توسعه روزافزون این فناوری و مسائل مربوطه.

۱. مقدمه

امروزه هوش مصنوعی فناوری‌ها و سیستم‌هایی را شامل می‌شود که امکان درک و تحلیل داده‌ها و تقلید عملکردهای انسانی را برای ماشین‌ها فراهم می‌سازد. این سیستم‌ها، قادر به انجام وظایف خاص و استخراج الگوهای پنهان هستند تا براساس آنها، پیش‌بینی‌ها و تصمیمات هوشمندانه‌ای اتخاذ کنند. این فناوری تحول‌آفرین در حوزه‌های مختلف، کارایی را افزایش داده و مشکلات پیچیده‌ای را حل می‌کند که در غیر این صورت برای ذهن انسانی بسیار چالش‌برانگیز بودند. در این زمینه می‌توان به تحلیل کلان‌داده‌ها، قابلیت تعامل با انسان و پردازش زبان انسانی و قابلیت تولید محتوای ابتکاری اشاره کرد [۱].

با توجه به اینکه مسائل بخش حاکمیتی از پیچیده‌ترین و پر حجم‌ترین مسائل هستند، خواه‌ناخواه حرکت به سمت بهره‌مندی از این فناوری تحولی، برای دولت‌ها جذابیت فراوانی خواهد داشت. در سال‌های اخیر دولت‌ها برای مدیریت عمومی، خط‌مشی‌گذاری و ارائه خدمات در بخش‌های مختلف به هوش مصنوعی رو آورده‌اند تا مؤثرتر حکمرانی کنند [۲]. این امور از تقویت بخش‌های اقتصادی و مالی تا بهبود مراقبت‌های بهداشتی و مدیریت زیرساخت‌ها را شامل شده است. ادبیات خط‌مشی‌گذاری مبتنی‌بر شواهد که در ابتدای قرن ۲۱ با هدف ارتقای کیفیت تصمیم‌گیری در بخش عمومی در انگلستان مطرح شد، اکنون با تولید انبوهی از داده‌ها و ظرفیت پردازش آنها توسط الگوریتم‌های هوش مصنوعی، جان تازه‌ای گرفته است [۳]. بااین‌حال، این روند با مانع بزرگی با عنوان قابلیت توضیح و تفسیر تصمیم‌ها مواجه است [۴].

نکته قابل‌توجه این است که روند تولید خروجی‌های این سیستم‌ها اغلب ناشناخته یا به اصطلاح «جعبه سیاه»[1] است. به این معنا که دلایل و فرایندهای پشت این تصمیمات حتی برای متخصصان، قابل توضیح نیست [۵]. این مسئله به‌ویژه در حوزه‌های حساس که تصمیمات اتخاذ شده می‌توانند تأثیر زیادی بر زندگی افراد داشته باشند، مانند مراقبت‌های بهداشتی، امور مالی و قضایی و در کل، در بخش عمومی، اهمیت فراوان دارد.

بنابراین حساسیت تصمیمات بخش عمومی و گستره آثار آنها، لزوم اطمینان از کیفیت و کارآمدی تصمیمات را ضروری ساخته است. علاوه‌بر این، بخش عمومی اقتضائات دیگری مانند پاسخ‌گویی، شفافیت، ضرورت اقناع عمومی و عمل در محدوده قوانین و مواردی از این دست را دارد. هرگونه تحول در بخش عمومی بدون درنظرگیری این اقتضائات ممکن است علاوه‌بر عدم بهبود شرایط، به عکس خود تبدیل شده و آثار مخربی بر حاکمیت بگذارد. توسعه هوش مصنوعی نیز از این امر مستثنا نیست. بنابراین ضروری است ورود این فناوری به بخش عمومی با رعایت تدابیری خاص و تحت ارزیابی و کنترل کیفیت دقیق انجام شود. این گزارش به دنبال معرفی شاخه‌ای از علوم هوش مصنوعی با عنوان هوش مصنوعی توضیح‌پذیر[2] به‌عنوان دریچه ورود سنجیده این فناوری به بخش عمومی است. در ادامه پس از بررسی اجمالی فناوری هوش مصنوعی و ظرفیت تحولی آن در ارتقای بخش‌های مختلف، اقتضائات خاص بخش عمومی و حاکمیتی بیان شده و هوش مصنوعی توضیح‌پذیر به‌عنوان الزام و متضمن این اقتضائات معرفی خواهد شد. سپس کاربردها، فواید جانبی و آینده این شاخه از هوش مصنوعی و چالش‌های پیاده‌سازی آن بررسی شده و درنهایت توصیه‌های سیاستی جهت توسعه این بستر ارائه خواهد شد.

1. Black Box
2. Explainable Artificial Intelligence (XAI)

۲. درک هوش مصنوعی و تأثیر آن

۲-۱. انواع هوش مصنوعی

هوش مصنوعی به مجموعه‌ای از فناوری‌ها و الگوریتم‌ها گفته می‌شود که توانایی شبیه‌سازی فرایندهای شناختی انسان ازجمله یادگیری، استدلال و تصمیم‌گیری را دارد. هوش مصنوعی از طریق ترکیبی از مجموعه داده‌های بزرگ، الگوریتم‌ها و فرایندهای یادگیری به دنبال تحقق این فرایندهای شناختی است.

محبوبیت فزاینده این فناوری در سال‌های اخیر عمدتاً به‌دلیل توانایی آن در خودکارسازی وظایف، مانند پردازش مقادیر زیادی از اطلاعات یا شناسایی الگوها و در دسترس بودن گسترده آن برای عموم است [۶].

فناوری‌های هوش مصنوعی را می‌توان به‌طور عمده به دو نوع تقسیم کرد:

– **هوش مصنوعی خاص منظور:** این نوع هوش مصنوعی که به‌عنوان هوش مصنوعی ضعیف هم شناخته می‌شود، در یک زمینه محدود عمل می‌کند و مختص یک کار خاص است. برای مثال می‌توان به وسایل نقلیه خودران اشاره کرد. هوش مصنوعی خاص منظور برای انجام مجموعه محدودی از وظایف برنامه‌ریزی شده و درک فراتر از بخش برنامه‌ریزی شده خود را ندارد و در این پژوهش، این نوع هوش مصنوعی مدنظر است.

– **هوش مصنوعی عمومی:** این نوع هوش مصنوعی توانایی درک، یادگیری و به‌کارگیری هوش در طیف گسترده‌ای از وظایف، مشابه توانایی‌های شناختی انسان را دارد. هوش مصنوعی عمومی هنوز وجود ندارد، اما حوزه قابل‌توجهی از تحقیقات در زمینه هوش مصنوعی را شامل می‌شود و به آن بسیار نزدیک شده‌ایم.

۲-۲. نقش هوش مصنوعی در جامعه

کاربردهای هوش مصنوعی امروز چندین بخش را دربرمی‌گیرد که تنها به چهار مورد از آنها به‌صورت مختصر در زیر اشاره می‌کنیم:

– **مراقبت‌های بهداشتی:** هوش مصنوعی در فرایندهای تشخیصی، پزشکی شخصی و مدیریت مراقبت از بیمار، ازجمله برنامه‌های کاربردی دیگر کمک می‌کند.

– **مالی:** هوش مصنوعی برای تجارت الگوریتمی، کشف تقلب و خدمات مشتری در بانکداری استفاده می‌شود.

– **حمل‌ونقل:** رانندگی و مدیریت ترافیک حوزه‌های کلیدی هستند که در آنها هوش مصنوعی به‌طور قابل‌توجهی مورد استفاده قرار می‌گیرد.

– **نظامی:** کمک در ساخت وسایل و تجهیزات نظامی خودکار و هوشمند و ارتقای توانمندی‌های دفاعی و نظامی از مهم‌ترین کاربردهای هوش مصنوعی است.

با گسترش و پیچیده‌تر شدن سیستم‌های هوش مصنوعی، فهم چگونگی تصمیم‌گیری‌های آنها اهمیت بیشتری پیدا می‌کند [۴]. به‌رغم استفاده رو به رشد از هوش مصنوعی، بسیاری از سیستم‌های مبتنی‌بر این فناوری به‌گونه‌ای عمل می‌کنند که هم برای توسعه‌دهندگان، هم برای استفاده‌کنندگان و هم برای کسانی که تحت‌تأثیر استفاده از آن هستند، غیرشفاف خواهند بود [۷]. این پدیده معمولاً به‌عنوان اثر «جعبه سیاه» شناخته می‌شود. علاوه‌بر این، سوگیری‌ها (انحرافات نظام‌مند در داده‌ها یا فرایندهای تصمیم‌گیری) می‌توانند منجر به نتایج ناعادلانه یا نادرست شوند. ازآنجاکه سیستم‌های هوش مصنوعی از داده‌ها یاد می‌گیرند، ممکن است به‌صورت ناخواسته، سوگیری‌های موجود در آن داده‌ها (شامل سوگیری‌های اجتماعی، فرهنگی یا اقتصادی) را پذیرفته و تقویت کنند. به همین دلیل، توجه به سوگیری‌ها در توسعه و پیاده‌سازی سیستم‌های هوش مصنوعی به‌ویژه در زمینه‌های حساس (مانند بهداشت و درمان، مالی و اجرای قانون) ضروری است.

این موارد تنها بخش کوچکی از تأثیرات هوش مصنوعی هستند. حال آنکه ظرفیت بالقوه تحول کل صنایع و ساختارهای اجتماعی از آینده توسعه این فناوری متصور است. درک این اصول و نقش‌های هوش مصنوعی، رویکرد آگاهانه‌تری را برای بحث و قانونگذاری فناوری هوش مصنوعی ایجاد و تضمین می‌کند که می‌توان مزایای آن را در عین کاهش خطراتش به حداکثر رساند.

۳. هوش مصنوعی و بخش عمومی: چرایی هوش مصنوعی توضیح‌پذیر

مسئله بهره‌مندی از ظرفیت داده‌ها در تصمیم‌گیری بخش عمومی امری جدید نیست. ادبیات خط‌مشی‌گذاری مبتنی‌بر شواهد که در ابتدای قرن ۲۱ در دولت انگلستان با هدف مدرن‌سازی دولت مطرح شد، بر این امر تأکید داشت. این دستورکار بر بهبود خط‌مشی‌گذاری از طریق داده‌ها، تحقیقات و بهره‌برداری از فناوری اطلاعات برای منافع عمومی تأکید می‌ورزد [۸]. همچنین سازمان همکاری و توسعه اقتصادی (OECD) خط‌مشی‌گذاری آگاه از شواهد[1] را مطرح کرده است و آن را فرایندی می‌داند که به‌موجب آن منابع متعدد اطلاعات، ازجمله آمار، داده‌ها و بهترین شواهد و ارزیابی‌های تحقیقاتی موجود، قبل از تصمیم‌گیری برای برنامه‌ریزی، اجرا و (در صورت لزوم) تغییر خط‌مشی‌ها و برنامه‌های عمومی مورد مشورت قرار می‌گیرند [۹]. بااین‌حال به‌رغم تولید انبوهی از داده‌ها در عصر کنونی، حصول شواهد عملیاتی و کاربردی برای تصمیم‌گیران در میان این حجم از داده‌ها، امری آسان نخواهد بود. فناوری هوش مصنوعی با ایجاد توانمندی‌های فراانسانی در تحلیل کلان داده‌ها، می‌تواند در این زمینه کمک‌کننده باشد.

توسعه کاربست فناوری هوش مصنوعی در بخش خصوصی در کنار توسعه روزافزون زیرساخت‌های این فناوری ازجمله زیرساخت‌های تولید، ذخیره و پردازش داده و توسعه الگوریتم‌ها زمینه‌ساز ورود این فناوری به بخش عمومی خواهد شد[۲]. در زمینه حکمرانی و بخش عمومی، هوش مصنوعی می‌تواند به بهبود کارایی، شفافیت و پاسخ‌گویی در ارائه خدمات عمومی کمک کند. کارکردهای این فناوری را در بخش عمومی می‌توان به شرح زیر برشمرد [۱۰]:

- ارتقای نظام قانونگذاری و خط‌مشی‌گذاری؛
- ارتقای کارآمدی نظام اداری؛
- ارتقای سرعت و کیفیت ارائه خدمات عمومی.

بااین‌حال این بخش اقتضائات و شرایط خاصی دارد که پیش از هرگونه اقدام برای ورود این فناوری به این بخش لازم به توجه است. جدول زیر به مقایسه شرایط بخش عمومی و خصوصی پرداخته است [۱۱]:

جدول ۱. مقایسه شرایط بخش عمومی و خصوصی

ویژگی	بخش دولتی	بخش خصوصی
منابع مورد استفاده	منابع عمومی و نیاز به پاسخ‌گویی	منابع خصوصی و پاسخ‌گویی صرفاً به سهام‌داران
مقیاس تصمیم‌گیری	کلان و گسترده در سطح جامعه و معمولاً برگشت‌ناپذیر	محدود و حوزه‌ای و قابلیت تغییر
محیط تصمیم‌گیری	محیط سیاسی و دمکراتیک	محیط رقابتی و خروجی محور
بستر اجرا تصمیم	خط‌مشی‌ها، قوانین و مقررات	مصوبات هیئت‌مدیره

1. Evidence-Informed Policymaking

ویژگی	بخش دولتی	بخش خصوصی
بازیگران و ذی‌نفعان	بازیگران و سازمان‌ها و ذی‌نفعان متعدد و سطوح مختلف	بازیگران و ذی‌نفعان محدود شامل هیئت‌مدیره، کارکنان و مشتریان
ارزش و هدف	ارزش‌های عمومی و منفعت عامه، ارزش‌های متعدد	ارزش‌های اقتصادی و منفعت شخصی
ساختار تصمیم‌گیری	بوروکراسی و لختی سازمانی	ساختارهای قابل انعطاف و چابک
مسئله‌یابی	مواجهه با نامسئله‌ها در تعریف بسیاری از مسائل	مبتنی‌بر تشخیص هیئت‌مدیره
افق زمانی	معمولاً بلندمدت	معمولاً کوتاه‌مدت
ارزیابی نتایج تصمیمات	ارزیابی چندگانه و از منظر ارزش‌های مختلف	مبتنی‌بر خروجی و منفعت اقتصادی و پولی

مأخذ: یافته‌های پژوهش.

این مسائل ضرورت توجه به موارد زیر را روشن می‌سازد:

- **پاسخ‌گویی و شفافیت:** با توجه به مصرف بودجه عمومی در کنار وجود سازوکارهای دمکراتیک و ذی‌نفعان متعدد، پاسخ‌گویی و شفافیت تصمیمات در بخش عمومی اهمیت فراوانی دارد.
- **عمل در محدوده قوانین:** به‌دلیل اینکه بستر اجرای تصمیمات در بخش عمومی، قوانین و مقررات بوده، ضروری است اعمال حاکمیت در این بخش براساس رعایت قوانین و تناسب با مقررات سنجیده شود. نهادهای ناظر در بخش عمومی این وظیفه را دارند.
- **تصمیمات مطمئن و قابل اعتماد:** با توجه به گستره تصمیمات حاکمیتی که معمولاً تمام افراد جامعه را شامل شده و آثاری بدون بازگشت می‌گذارد و با در نظرگیری افق زمانی طولانی در مشخص شدن نتایج تصمیمات این بخش، اهمیت ارزشیابی تصمیمات پیش از اجرا و تضمین قابلیت اعتماد آنها اهمیت فراوانی دارد.
- **تنوع مسائل و تخصصی بودن تصمیمات:** مسائل این بخش معمولاً درهم‌تنیده و پیچیده و دارای ابعاد مختلف است. در ضمن تنوع مسائل این حوزه، ضرورت توجه به تخصص‌های مختلف را آشکار ساخته است.
- **مسئولیت‌پذیری و سازوکار تصمیم‌گیری و سلسله‌مراتب:** غلبه ساختار سلسله‌مراتبی و بوروکراتیک در بخش عمومی، لزوم پاسخ‌گویی به مقام مافوق را در این بخش بیش‌ازپیش ساخته است. در این راستا اهمیت توجه به بحث مسئولیت‌پذیری مطرح شده است.
- **تصمیمات قابل اعتماد و مورد پذیرش عمومی:** با توجه به وجود بازیگران مختلف در جامعه که هرکدام ارزش و عقلانیت خاص خود را دارند و هدف بخش عمومی تحقق منفعت عامه از طریق تعامل با این بازیگران است، علاوه بر پاسخ‌گویی، توجه به میزان قابل اعتماد بودن تصمیمات و میزان پذیرش عمومی تصمیمات در این بخش اهمیت فراوانی می‌یابد.
- **رعایت ملاحظات اخلاقی:** تضمین رعایت ملاحظات اخلاقی در تصمیمات بخش عمومی به‌علت وکالت مردم به حاکمیت در پیشبرد امور آنها و مسئولیت حاکمیت در حفظ ارزش‌های عمومی، ضروری است.

به‌طورکلی موارد مذکور لزوم توجه و دقت پیش از ورود فناوری‌های خودمختار و خود تصمیم‌گیر را به بخش عمومی روشن می‌سازد. هوش مصنوعی ازجمله این فناوری‌هاست. در این راستا هوش مصنوعی توضیح‌پذیر مطرح شده است که می‌تواند به دغدغه‌های فوق رسیدگی کند.

۴. چیستی و اهمیت هوش مصنوعی توضیح‌پذیر

۴-۱. چیستی هوش مصنوعی توضیح‌پذیر

هوش مصنوعی توضیح‌پذیر(XAI[1]) یک مفهوم یکپارچه نیست و مشتمل بر قیودی مانند قابلیت توضیح،[2] تفسیرپذیری،[3] شفافیت،[4] پاسخ‌گویی[5] یا قابلیت فهم[6] است [۱۲]. مفاهیم این حوزه کماکان در حال تکمیل هستند و مواردی مانند اهداف، درجه، محدوده و مخاطبان توضیح‌پذیری مورد بحث است [۱]. برخی از تعاریف صرفاً و به‌طور محدود بر مفاهیم فنی متمرکز هستند، درحالی‌که برخی دیگر فقط بر ابعاد گسترده اجتماعی و سیاسی متمرکزند [۱۳]. بااین‌حال یک تعریف مورد پذیرش، توضیح‌پذیری را ایجاد دلایل و جزئیاتی تعریف کرده که عملکرد یک مدل را برای یک مخاطب خاص ساده یا آسان می‌کند [۱۴]. در ادبیات، اصطلاحات هوش مصنوعی قابل توضیح،[7] هوش مصنوعی قابل تفسیر،[8] هوش مصنوعی شفاف،[9] هوش مصنوعی قابل درک[10] و هوش مصنوعی مسئول[11] به جای یکدیگر استفاده می‌شوند [۱۴].

بااین‌حال، علی و همکاران [۱۵] بیان می‌کنند که توضیح‌پذیری، توسعه‌دهندگان را قادر می‌سازد تا در فرایند تصمیم‌گیری مدل، تحقیق کنند. همچنین اعتماد آنها را از طریق درک موقعیت‌هایی که مدل نتایج خود را دریافت می‌کند، افزایش می‌دهد. به‌جای یک پیش‌بینی ساده، توضیح‌پذیری رابطی را ارائه می‌کند که اطلاعات یا توضیحات اضافی را ارائه می‌دهد که برای تفسیر عملکرد زیربنایی یک سیستم هوش مصنوعی ضروری است و به توسعه‌دهندگان کمک می‌کند. ازسوی‌دیگر، توضیح‌پذیری بینشی از تصمیم الگوریتم هوش مصنوعی به کاربر نهایی می‌دهد تا اعتماد ایجاد کند که هوش مصنوعی تصمیم‌های درست و غیرمغرضانه براساس حقایق اتخاذ کرده است.

می‌توان گفت برخلاف مدل‌های معمول هوش مصنوعی که عملکرد داخلی آنها نامشخص و نامفهوم است، این شاخه از هوش مصنوعی بر ایجاد سیستم‌هایی متمرکز است که قادر به ارائه توضیحات واضح و قابل درک برای انسان‌ها درباره نحوه رسیدن به نتایج خود هستند. توضیح‌پذیری به این معناست که هوش مصنوعی باید بتواند پیش‌بینی‌های به دست آمده از یک مدل را از دیدگاه روش‌شناختی عمیق‌تری برای کاربران توضیح دهد [۱۶]. این شفافیت ضروری است تا بتوان به سیستم‌های هوش مصنوعی اعتماد کرد و به‌طور مؤثرتری با آنها تعامل داشت. زمانی که هوش مصنوعی بتواند دلایل خود را توضیح دهد، کاربران بیشتر تمایل دارند تا توصیه‌های آن را بپذیرند و براساس آنها عمل کنند که این موضوع خود منجر به پذیرش و استفاده گسترده‌تر از فناوری‌های هوش مصنوعی در بخش‌های مختلف خواهد شد.

آژانس پروژه‌های تحقیقاتی پیشرفته دفاعی[12] آمریکا (دارپا) که پیشران و پیشرو بسیاری از فناوری‌ها در سراسر جهان است در سال ۲۰۱۷ برنامه‌ای مرتبط با هوش مصنوعی توضیح‌پذیر با هدف ایجاد مجموعه‌ای از روش‌ها و جعبه ابزارهای توضیح‌پذیری فناوری‌های مرتبط با هوش مصنوعی ازجمله یادگیری ماشین را ایجاد کرد.[13] نتایج این برنامه در ویژه نامه‌ای از مجله هوش مصنوعی کاربردی[14]

1 .Explainable Artificial Intelligence
2 .Explainable
3 .Interpretable
4 .Transparent
5 .Accountable
6 .Understandable
7 .Explainable AI
8 .Interpretable AI
9 .Transparent AI
10 .Understandable AI
11 .Responsible AI
12 .Defense Advanced Research Projects Agency (DARPA)
13 .https://www.darpa.mil/research/programs/explainable-artificial-intelligence
14 .Applied AI Letters

در سال ۲۰۲۱ با چاپ ۲۵ مقاله منتخب منتشر شد[1]. در ضمن جعبه ابزار متن باز هوش مصنوعی توضیح‌پذیر بر این اساس در اختیار عموم قرار گرفت.[2] در سال‌های بعد تحقیقات در مورد هوش مصنوعی قابل توضیح افزایش یافته و در ارتباط با خلاصه کردن خط‌مشی، همکاری انسانی، تجسم، تأیید و غیره متمرکز شده است [۱۷].

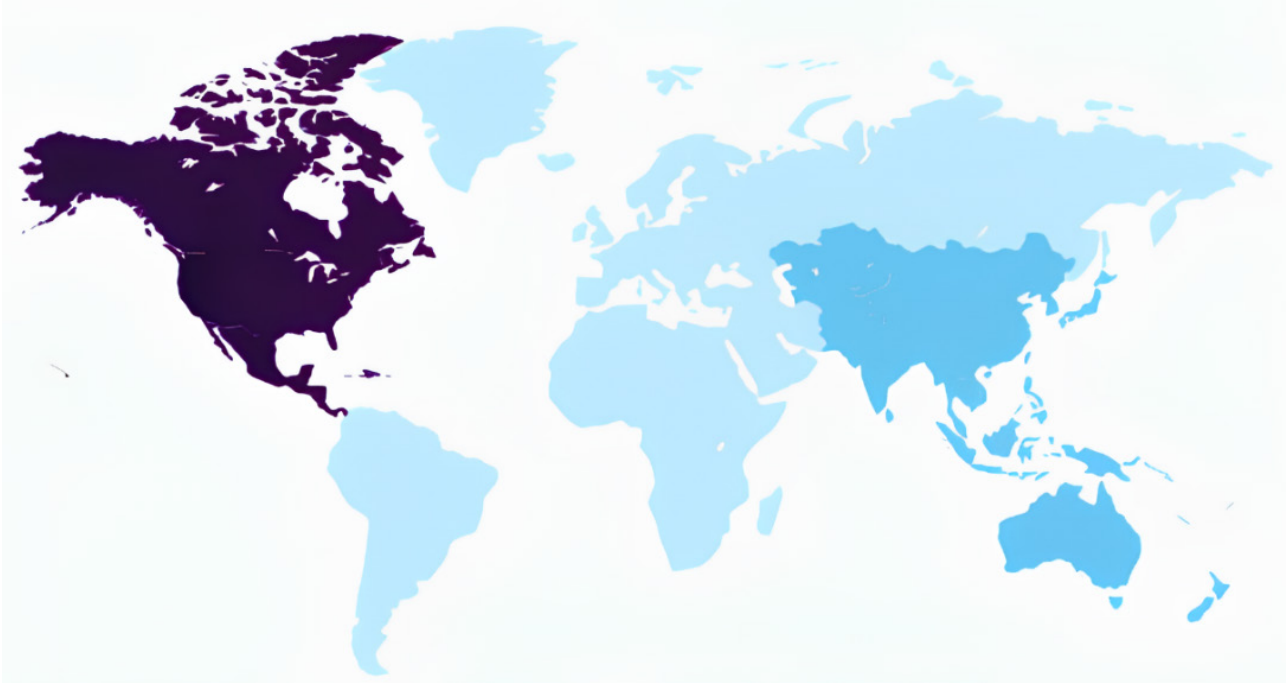
هوش مصنوعی توضیح‌پذیر به شناسایی و کاهش سوگیری‌ها کمک می‌کند. با شفاف‌سازی فرایندهای تصمیم‌گیری هوش مصنوعی، توسعه‌دهندگان و ذی‌نفعان قادر خواهند بود بهتر بفهمند که این پیش‌داوری‌ها از کجا ناشی می‌شوند و چه اقداماتی می‌تواند برای رفع آنها انجام شود. این رویکرد نه‌تنها به بهبود دقت و کارایی سیستم‌های هوش مصنوعی کمک می‌کند، بلکه اعتماد عمومی را نیز افزایش و اطمینان می‌دهد که تصمیمات اتخاذ شده عادلانه و مبتنی‌بر شواهد هستند. علاوه‌بر این، توضیح‌پذیری امکان بهبود و ارتقای مدل‌ها و تخصصی‌تر شدن آنها را نیز فراهم می‌کند.

هوش مصنوعی به سرمایه‌گذاری‌های فزاینده و توسعه‌های مهم در بازار نیاز دارد. این سرمایه‌گذاری‌ها برای پیشرفت در فناوری هوش مصنوعی ضروری هستند. بازار هوش مصنوعی توضیح‌پذیر نیز به‌عنوان یک بخش کلیدی در این توسعه، نقش مهمی ایفا می‌کند.

در گزارشی در سال ۲۰۲۲ [۱۸]، بازار استفاده از هوش مصنوعی توضیح‌پذیر براساس نوع مصرف، به بخش‌های مختلفی مانند بهداشت و درمان، خدمات مالی، هوافضا، خرده‌فروشی، بخش عمومی، فناوری اطلاعات، خودروسازی و غیره تقسیم می‌شود. بخش فناوری اطلاعات و ارتباطات با سهم ۹۹/۱۷٪ پیشرو بوده است. رتبه دوم در بخش مالی است که هوش مصنوعی توضیح‌پذیر می‌تواند بهره‌وری را افزایش و هزینه‌ها را کاهش دهد و رشد بازار را تقویت می‌کنند. در رتبه‌های بعدی نیز بخش‌های بهداشت و درمان و خرده فروشی قرار می‌گیرند.

شکل ۱، میزان توسعه هوش مصنوعی توضیح‌پذیر را براساس موقعیت جغرافیایی نشان می‌دهد:

شکل ۱. تصویر مناطقی با بیشترین سهم از بازار و بیشترین نرخ رشد [۱۸]



1. https://onlinelibrary.wiley.com/toc/26895595/2021/2/4
2 .https://xaitk.org/

همچنین در گزارش [۱۸] آمریکای شمالی با سهم ۴۰/۵۲٪ اکثریت بازار را را در اختیار دارد و پیش‌بینی می‌شود در دوره پیش‌بینی تا سال ۲۰۳۰ با نرخ رشد مرکب سالیانه۱۳/۴٪ به رشد خود ادامه دهد. زیرساخت‌های قوی فناوری اطلاعات در کشورهایی مانند آلمان، فرانسه، آمریکا، انگلستان، ژاپن و کانادا از عوامل اصلی حمایت‌کننده رشد بازار هوش مصنوعی توضیح‌پذیر در این کشورهاست. همچنین حمایت‌های گسترده دولتی برای به‌روزرسانی زیرساخت‌های فناوری اطلاعات به توسعه این بازار کمک می‌کند. بااین‌حال، انتظار می‌رود کشورهایی مانند هند و چین در دوره پیش‌بینی‌شده رشد بیشتری داشته باشند. رشد اقتصادی مطلوب در این کشورها، سرمایه‌گذاری‌های متعددی را به سمت توسعه کسب‌وکار هوش مصنوعی توضیح‌پذیر جذب می‌کند.

علاوه‌بر این، پیش‌بینی می‌شود منطقه آسیا و اقیانوسیه سریع‌ترین نرخ رشد مرکب سالیانه با ۸/۲۴ ٪ را در دوره پیش‌بینی داشته باشد. پیشرفت‌های چشمگیر فناوری در کشورهای این منطقه، رشد بازار را تقویت می‌کند.

برای دستیابی به توضیح‌پذیری، روش‌های مختلفی وجود دارد، ازجمله استفاده از مدل‌های ساده‌تر مانند درخت تصمیم یا رگرسیون خطی و یا توسعه روش‌هایی که خروجی مدل‌های پیچیده‌تر مانند شبکه‌های عصبی را توضیح‌پذیرتر می‌کنند.

روش‌های ایجاد توضیح‌پذیری می‌توانند از حیث دامنه، کاربرد و روش‌شناسی توضیح‌پذیری، متفاوت باشند.

به‌طورکلی دو نوع رویکرد برای توضیح‌پذیری مدل‌های هوش مصنوعی از حیث دامنه توضیح‌پذیری وجود دارد:

۱. توضیحات محلی: دامنه این رویکرد، ارائه توضیحات برای نتایج یک داده خاص از مدل مدنظر است. برای مثال فرض کنید یک سیستم هوش مصنوعی برای ارزیابی وام در یک بانک استفاده می‌شود. یک مشتری درخواست وام داده و سیستم پیش‌بینی کرده که این فرد نباید وام بگیرد. روش LIME[1] [۱۹] که یک روش تفسیر محلی است، به ما کمک می‌کند تا بفهمیم چرا مدل، این تصمیم را برای این شخص خاص گرفته است. مثلاً ممکن است نشان دهد که درآمد پایین و تاریخچه اعتباری ضعیف عوامل اصلی در این تصمیم‌گیری بوده‌اند. به این ترتیب، با بررسی این ویژگی‌ها، می‌توان به توضیح دقیقی برای تصمیم مدل دست یافت و اگر لازم بود، به مخاطب اطلاعات بیشتری ارائه داد. همچنین روش SHAP[2] براساس نظریه بازی‌ها، تأثیر هر ویژگی بر خروجی مدل را محاسبه می‌کند و توضیحات جامع‌تری ارائه می‌دهد.

۲. توضیحات کلی: دامنه این رویکرد، ارائه تصویری کلی از نحوه عملکرد مدل در کل داده‌هاست. برای مثال، فرض کنید همان سیستم هوش مصنوعی برای ارزیابی وام در بانک، پیش‌بینی‌هایی برای هزاران مشتری انجام داده است. روشی مانند «مدل جانشین کلی»[3] به ما کمک می‌کند تا یک مدل ساده‌تر را آموزش دهیم که رفتار مدل پیچیده را تقریب بزند. با استفاده از این مدل ساده‌تر، می‌توان به یک درک کلی از تأثیر ویژگی‌ها بر تصمیمات مدل اصلی دست یافت. برای مثال، ممکن است متوجه شویم که درآمد، تاریخچه اعتباری و بدهی فعلی، مهم‌ترین عواملی هستند که مدل برای تصمیم‌گیری از آنها استفاده می‌کند. این اطلاعات به بانک کمک می‌کند تا سیاست‌های خود را بهتر تنظیم کند و تصمیمات عادلانه‌تری بگیرد [۱۹] و [۲۰].

همچنین می‌توان تکنیک‌های توضیح‌پذیری را از حیث محدوده توضیح‌پذیری به دو دسته تقسیم کرد:

اول، روش‌هایی که عمومی هستند و می‌توانند برای توضیح‌پذیر کردن هر الگوریتم یادگیری ماشینی به‌کار گرفته شوند (مانند SHAP و LIME).

دوم، روش‌هایی که مختص مدل‌های شبکه عصبی عمیق (DNN) هستند که در آن عملیات درونی مدل، مانند توابع فعال‌سازی و وزن‌ها توضیح داده شده، در نظر گرفته می‌شود (مانند GRAD-CAM[4] و DeepLIFT[5]) [۵].

به‌علاوه از منظر کاربرد تکنیک‌های توضیح‌پذیری در الگوریتم‌ها می‌توان تکنیک‌های توضیح‌پذیری را به روش‌هایی که به‌عنوان

1 .Local Interpretable Model-Agnostic Explanations
2 .Shapley Additive Explanations
3. Global Surrogate
4 .GRADient Class Activation Mapping
5 .Deep Learning Important FeaTures

بخشی از یک مدل خاص یادگیری ماشین و به شکل درونی مورد استفاده قرار گرفته‌اند (رویکرد ذاتی (مختص مدل))[۱] و روش‌هایی که می‌توانند برای هر مدلی به‌عنوان یک عمل پسافرایندی و از بیرون اعمال شوند، تقسیم کرد (رویکرد POST HOC (مدل - آگنوستیک).[۲] در یک رویکرد ذاتی، توضیح‌پذیری به‌عنوان بخشی از فرایند آموزش در معماری مدل یادگیری ماشین ادغام می‌شود و نمی‌تواند به معماری‌های دیگر منتقل شود. در یک رویکرد POST-HOC، روش XAI به هیچ معماری وابسته یا مرتبط نیست و می‌تواند برای هر مدل یادگیری ماشین آموزش‌دیده، اعمال شود [۵].

از حیث روش‌شناسی توضیح‌پذیری می‌توان روش را به دو دسته روش مبتنی‌بر پس انتشار[۳] یا مبتنی‌بر اغتشاش[۴] تقسیم‌بندی کرد [۴] که توضیحات بیشتر از حوصله این گزارش خارج است.

روش دیگر برای طبقه‌بندی متدهای توضیح‌پذیری، تقسیم آنها به اثبات‌ها،[۵] اعتبارسنجی‌ها[۶] و مجوزها[۷] است. اثبات‌ها، توضیحاتی قابل آزمایش، قابل ردیابی و بدون ابهام هستند که از طریق پیوندهای علّی، گزاره‌های منطقی یا فرایندهای شفاف قابل اثبات هستند. معمولاً اثبات‌ها فقط برای سیستم‌های هوش مصنوعی در دسترس هستند که از تکنیک‌های «ذاتاً قابل تفسیر»[۸] مانند قوانین،[۹] درخت‌های تصمیم‌گیری[۱۰] یا رگرسیون‌های خطی[۱۱] استفاده می‌کنند. اعتبارسنجی‌ها، توضیحاتی هستند که صحت سیستم هوش مصنوعی را تأیید می‌کنند. این راستی‌آزمایی‌ها از طریق روش‌های آزمایش، تکرارپذیری، تقریب‌ها و انتزاع‌ها و توجیه‌ها انجام می‌شود. مجوزها، فرایندهایی هستند که در آن اشخاص ثالث نوعی استاندارد، تصویب، ممنوعیت یا ممیزی را ارائه می‌کنند. مجوزها ممکن است مربوط به مدل هوش مصنوعی، عملکرد آن در موارد خاص یا حتی فرایند ایجاد هوش مصنوعی باشند.

شایان ذکر است که آنچه را می‌توان قابل توضیح دانست، ممکن است بسته به زمینه یا مخاطب خاص، قابل تغییر باشد. آنچه که یک دانشمند علم داده و یک قانونگذار بدون پیشینه علم کامپیوتر «قابل توضیح» تلقی می‌کنند، بسیار متفاوت است [۳]. برای یک مدل تشخیص سرطان با استفاده از تصاویر میکروسکوپی، توضیح ممکن است به معنای نقشه‌ای از پیکسل‌های ورودی باشد که به خروجی مدل کمک می‌کنند. برای یک مدل تشخیص گفتار، توضیح ممکن است اطلاعات طیف توان در طول یک زمان خاص باشد که بیشتر به تصمیم خروجی فعلی کمک می‌کند [۴]. بنابراین پاسخ به سه سؤال پیش از انتخاب هر استراتژی توضیح‌پذیری ضروری است [۱۳]:

- توضیح خوب چیست؟
- توضیح برای کیست؟
- توضیح چگونه ارائه خواهد شد؟

توضیحات بسیار وابسته به زمینه هستند. میزان «خوب بودن» یک توضیح به نیازها و اهداف مخاطب توضیح (کاربر) و توضیح‌دهنده (یک XAI) بستگی دارد، مخاطبان مختلف اولویت‌های متفاوتی دارند. برنامه‌نویسان و توسعه‌دهندگان سیستم در درجه اول به توضیحات مرتبط با کارآمدی الگوریتم علاقه‌مند هستند. کاربران الگوریتم‌ها بر کارایی یا اثربخشی تمرکز می‌کنند. متخصصان آن حوزه خاص نگران صحت نتایج الگوریتم‌ها هستند و قانونگذاران به پیامدهای خط‌مشی علاقه‌مندند. کاربران غیرمتخصص یک سیستم نیز توضیحاتی را می‌خواهند که اعتماد ایجاد و مسئولیت‌پذیری را فراهم کند [۱۳].

1 .Model-Specific
2 .Model-Agnostic
3 .Backpropagation-Based
4 .Perturbation-Based
5 .Proofs
6 .Validations
7 .Authorizations
8 .Inherently Interpretable
9 .Rules
10 .Decisions Trees
11 .Linear Regressions

هرکدام از رویکردها و تکنیک‌های مذکور، کاربردهای خاصی دارد که ضروری است بسته به هدف از توضیح‌پذیری و مرحله استفاده از مدل یادگیری ماشین، میان آنها انتخاب صورت گیرد.

همچنین علی و همکاران این حوزه مطالعاتی را به توضیح‌پذیری داده، توضیح‌پذیری مدل، توضیح‌پذیری پسافرایندی و ارزیابی توضیح‌پذیری تقسیم کرده‌اند که هرکدام از این موارد مخاطبان خاص خود را (اعم از متخصصان هوش مصنوعی، عموم شهروندان، سیاستمداران، کاربران مدل‌ها و...) دارد [۱۵].

۲-۴. اهمیت هوش مصنوعی توضیح‌پذیر در دنیای امروز

اهمیت هوش مصنوعی توضیح‌پذیر قابل چشم‌پوشی نیست. در دنیایی که تصمیمات هوش مصنوعی می‌تواند تأثیرات قابل‌توجهی بر افراد و جامعه داشته باشد، از تأیید وام‌ها گرفته تا تشخیص بیماری‌ها، ضروری است که این تصمیمات قابل‌فهم و قابل‌توجیه باشند. این اهمیت در سالیان اخیر با دلایلی مانند نیاز مدل‌ها به تأیید، بهبود، انطباق با قانون، پرهیز از سوگیری و تشخیص آثار ویژگی‌های مرتبط، بیش‌ازپیش شده است [۳]. قابلیت توضیح و تفسیر، قابلیت اعتماد به مدل‌ها را افزایش می‌دهد. چراکه کاربران احتمالاً زمانی که با فرایند اتخاذ تصمیمات این مدل‌ها آشنا باشند، بیشتر به سیستم‌های هوش مصنوعی توجه خواهند کرد و تصمیمات آنها را خواهند پذیرفت [۲]. این مسئله به‌ویژه در زمینه‌های حساس مانند پزشکی اهمیت دارد، جایی که بیماران و ارائه‌دهندگان خدمات بهداشتی نیاز دارند تا دلایل پشت تشخیص‌ها و توصیه‌های درمانی ارائه شده توسط هوش مصنوعی را درک کنند [۱]. با پر کردن شکاف بین سیستم‌های هوش مصنوعی پیچیده و درک انسانی، توضیح‌پذیری نقش مهمی در بهره‌برداری ایمن، اخلاقی و مؤثر از فناوری‌های هوش مصنوعی در بخش‌های مختلف ایفا می‌کند [۴].

هوش مصنوعی توضیح‌پذیر با در نظر گرفتن مواردی مانند شفافیت،[۱] واقع‌نمایی (اعتماد)،[۲] عدالت،[۳] مسئولیت،[۴] قابل‌فهم بودن،[۵] قابلیت استفاده،[۶] امنیت،[۷] حریم خصوصی،[۸] اخلاق،[۹] استحکام[۱۰] و ثبات،[۱۱] دستاوردهای زیر را در یک الگوریتم هوش مصنوعی محقق خواهد کرد [۱]، [۱۵]، [۲۱] و [۲۲]:

- توانمندسازی افراد در مواجهه با پیامدهای منفی ناشی از تصمیم‌گیری خودکار؛
- کمک به افراد در انتخاب آگاهانه‌تر؛
- افشای تهدیدها و محافظت از آسیب‌پذیری‌های امنیتی؛
- ادغام و سوق الگوریتم‌ها به سمت ارزش‌های انسانی؛
- ارتقای استانداردهای صنعت برای توسعه محصولات مبتنی‌بر هوش مصنوعی و درنتیجه بهبود اعتماد مصرف‌کننده و کسب‌وکار؛
- شناخت نقاط ضعف الگوریتم‌ها و امکان ارتقای آنها؛
- شهود ایده‌های جدید براساس عملکرد الگوریتم‌ها؛
- اجرای خط‌مشی‌ها و قوانین حق توضیح برای کاربران.

برای مثال، در سیستم عدالت کیفری، هوش مصنوعی توضیح‌پذیر می‌تواند کمک کند تا در خصوص آزادی یا صدور حکم، عادلانه

1 .Transparency
2 .Trust
3 .Fairness
4 .Accountability
5 .Understandability,
6 .Usability
7 .Security
8 .Privacy
9 .Ethics
10 .Robustness
11 .Stability

و براساس معیارهای عینی و نه داده‌های متعصبانه تصمیم‌گیری شود. همچنین، چارچوب‌های قانونی در سراسر جهان، شفافیت در پیاده‌سازی هوش مصنوعی را مطالبه می‌کنند. هوش مصنوعی توضیح‌پذیر از طریق تصویب قوانین و مقرراتی برای محافظت از حق افراد برای دریافت توضیحات مرتبط با یک تصمیم اداری که براساس یک فرایند الگوریتمی اتخاذ شده، مورد توجه قرار گرفته است [۲۲]. قوانینی مانند سازمان مقررات عمومی حفاظت از داده‌ها (GDPR) در اروپا اجبار می‌کند که افراد حق دریافت توضیحات مرتبط با تصمیم‌گیری خودکار را داشته باشند [۲۳]. این «حق توضیح»، جزو مقرراتی است که از ۲۵ مه ۲۰۱۸ در سراسر اتحادیه اروپا اجرایی شده است [۲۴] و موضوع توضیح‌پذیری را نه‌تنها تبدیل به یک ویژگی مطلوب، بلکه آن را یک نیاز ضروری برای مطابقت با استانداردهای قانونی می‌کند.

البته این فقط جزئی از سیاست‌های وسیع‌تری است که در دنیا در حال پیگیری است. در قانون هوش مصنوعی (AI Act) که هدف آن حمایت از توسعه هوش مصنوعی قابل اعتماد در اروپاست، این کار به‌وسیله تضمین ایمنی و حقوق افراد و کسب‌وکارها در زمینه هوش مصنوعی و همچنین تقویت پذیرش، سرمایه‌گذاری و نوآوری در این حوزه در سراسر اتحادیه اروپا انجام می‌شود [۲۳].

۵. هوش مصنوعی توضیح‌پذیر و بخش عمومی

همان‌طور که بیان شد مهم‌ترین شرایط و ملاحظات تصمیم‌گیری در بخش عمومی عبارت‌اند از:

- پاسخ‌گویی و شفافیت؛
- عمل در محدوده قوانین؛
- تصمیمات مطمئن و قابل اعتماد؛
- تنوع مسائل و تخصصی بودن تصمیمات؛
- مسئولیت‌پذیری و سازوکار تصمیم‌گیری و سلسله‌مراتب؛
- قابلیت اعتماد و پذیرش عمومی تصمیمات؛
- رعایت ملاحظات اخلاقی.

به‌کارگیری فناوری هوش مصنوعی در این بخش اگر با الزام توضیح‌پذیری و با رعایت این ویژگی محقق شود، خواهد توانست این شرایط و ملاحظات را تضمین کرده و به ارتقای کارآمدی بخش عمومی منجر شود. در غیر این صورت عدم سازگاری ویژگی‌های تصمیمات خودکار و الگوریتمی یا موجب آسیب رساندن به ارزش‌های عمومی مانند شفافیت و پاسخ‌گویی و حتی سوءاستفاده احتمالی افراد خواهد شد، یا بر اثر مقاومت و واپس زدن این فناوری، بخش عمومی از ظرفیت بالقوه این فناوری نوین بی‌بهره خواهد شد. بنابراین توجه به الزام توضیح‌پذیری در الگوریتم‌های به‌کار گرفته شده در بخش عمومی ضروری است. این توضیح‌پذیری به‌طور خاص می‌تواند موارد زیر را محقق کند:

۵-۱. قابلیت اعتماد و شفافیت

در عصر دیجیتال امروز، سیستم‌های هوش مصنوعی، هر روز بیشتر و بیشتر تصمیماتی می‌گیرند که بر زندگی روزمره ما تأثیر می‌گذارد. برای مثال در حوزه‌های بهداشت و درمان و مالی، این تصمیمات می‌توانند پیامدهای قابل‌توجهی داشته باشند. برای اینکه هوش مصنوعی به‌طور مؤثر در جامعه ادغام شود، ضروری است که به این سیستم‌ها توسط کاربرانشان اعتماد شود. هوش مصنوعی توضیح‌پذیر، اعتماد را با ارائه نکاتی در مورد چگونگی تصمیم‌گیری‌ها ایجاد می‌کند.

شفافیت در هوش مصنوعی به این معناست که کاربران می‌توانند فرایندهای زیربنایی و منطق منجر شده به یک نتیجه خاص را درک کنند. این موضوع به‌ویژه در سناریوهایی که ممکن است سیستم‌های هوش مصنوعی خطا کنند یا تصمیمات آنها مورد ارزیابی قرار گیرد، اهمیت دارد. برای مثال، در تشخیص‌های پزشکی، اگر یک سیستم هوش مصنوعی درمان خاصی را توصیه کند، ضروری است پزشکان و بیماران دلایل پشت این توصیه را درک کنند تا با آرامش‌خاطر از آن پیروی کنند. شفافیت به عمومی‌تر کردن هوش مصنوعی کمک و آن را قابل دسترس‌تر و مقبول‌تر می‌کند که این موضوع باعث تشویق به پذیرش آن می‌شود. علاوه‌بر این، شفافیت فقط درباره ایجاد اعتماد در کاربران نهایی نیست، بلکه به توسعه‌دهندگان و پژوهشگران این امکان را می‌دهد تا مدل‌های هوش مصنوعی را با شناسایی و اصلاح نقص‌ها، تعصبات و خطاها، بهبود بخشند. سیستم‌های هوش مصنوعی شفاف اجازه شناسایی و اصلاح سوگیری یا خطاها را می‌دهند، درنتیجه از نتایج ناعادلانه یا آسیب‌زا جلوگیری می‌کنند.

توضیح‌پذیری در الگوریتم‌های هوش مصنوعی می‌تواند در تحقق شفافیت و ارتقای اعتماد نسبت به تصمیمات متخذه و افزایش پذیرش عمومی این تصمیمات که در بخش عمومی نقشی حیاتی دارد، نقش‌آفرین باشد.

۲-۵. پاسخ‌گویی، مسئولیت‌پذیری و ملاحظات اخلاقی

به هر میزان که سیستم‌های هوش مصنوعی نقش‌های مهم‌تری در جامعه برعهده می‌گیرند، نیاز به مسئولیت‌پذیری نسبت به نتایج تصمیمات بیشتر می‌شود. مسئولیت‌پذیری در هوش مصنوعی به توانایی ردیابی و فهم فرایند تصمیم‌گیری اشاره دارد تا بفهمیم که این تصمیمات به‌صورت **اخلاقی** و **مسئولانه** گرفته شده‌اند یا نه. هوش مصنوعی توضیح‌پذیر به ذی‌نفعان این امکان را می‌دهد تا انصاف و درستی تصمیمات هوش مصنوعی را بررسی و ارزیابی کنند.

ملاحظات اخلاقی در خط مقدم توسعه هوش مصنوعی قرار دارد. سیستم‌های هوش مصنوعی، اگر به دقت طراحی و نظارت نشوند، می‌توانند تعصبات و نابرابری‌های موجود را حفظ یا حتی تشدید کنند. برای مثال، اگر یک سیستم هوش مصنوعی که در فرایندهای استخدام استفاده می‌شود با داده‌های متعصبانه آموزش داده شده باشد، ممکن است گروه‌های خاصی از متقاضیان را ناعادلانه محروم کند که در بخش عمومی این امر قابل پذیرش نیست. هوش مصنوعی توضیح‌پذیر به شناسایی این تعصبات با شفاف‌سازی فرایند تصمیم‌گیری کمک می‌کند و به توسعه‌دهندگان اجازه می‌دهد تا به‌طور فعال این مسائل را حل‌وفصل کنند.

این شاخه تنها مرتبط با اجتناب از تبعیض و پیش‌داوری نبوده و علاوه‌بر آن، شامل احترام سیستم‌های هوش مصنوعی به حریم خصوصی کاربران، رعایت نفع کاربران در تصمیمات و جلوگیری از آسیب‌های غیرمنتظره نیز می‌شود. این موضوع به کاربران و ذی‌نفعان قدرت می‌دهد تا از تصمیمات هوش مصنوعی سؤال کنند و به چالش بکشند که این امر منجر به استفاده مسئولانه‌تر و اخلاقی‌تر از هوش مصنوعی می‌شود [۲۵].

ضمن اینکه فهم فرایند اخذ تصمیمات و ایجاد توضیح نسبت به آنها تناسب بیشتری با محیط بوروکراتیک و سلسله‌مراتبی که اقتضای بخش عمومی است دارد. زیرا لازمه تحقق پاسخ‌گویی سلسله‌مراتبی، وجود سطحی از پاسخ‌گویی و مسئولیت‌پذیری است که تصمیم‌گیری خودکار الگوریتمی بدون در نظر گرفتن ملاحظات توضیح‌پذیری قابل تحقق نخواهد بود.

۳-۵. رعایت مقررات

نهادهای نظارتی به اهمیت شفافیت در هوش مصنوعی پی برده‌اند. **قوانین و مقررات** متعددی در حال وضع شدن هستند تا سیستم‌های هوش مصنوعی به‌صورت منصفانه، شفاف و مسئولانه عمل کنند. رعایت این مقررات برای سازمان‌هایی که فناوری‌های هوش مصنوعی را توسعه و استقرار می‌دهند، ضروری است.

هوش مصنوعی توضیح‌پذیر به تحقق این الزامات قانونی کمک می‌کند و با ارائه بینش‌های واضح در مورد چگونگی تصمیم‌گیری

سیستم‌های هوش مصنوعی، امکان عمل به این قوانین را فراهم می‌سازد. این امر نه‌تنها یک ضرورت قانونی است، بلکه یک مزیت رقابتی نیز محسوب می‌شود. شرکت‌هایی که توضیح‌پذیری را در سیستم‌های هوش مصنوعی خود اولویت قرار می‌دهند، می‌توانند به‌راحتی اعتماد مصرف‌کنندگان و نهادهای نظارتی را جلب کنند و از چالش‌های قانونی کمتری برخوردار باشند.

علاوه‌بر این به‌دلیل وجود دستگاه‌های نظارتی متعدد در بخش عمومی و وجود دستورکار عملیاتی به نام **قوانین** که این دستگاه‌ها به دنبال ارزیابی میزان عمل به آن قوانین هستند، ضروری است تصمیمات اتخاذ شده در بخش عمومی و اجرایی، ملاحظه عمل به قوانین و مقررات حوزه مربوطه را تأمین کنند. وجود سطحی از توضیح‌پذیری جهت تأمین این ملاحظه ضروری خواهد بود.

۶. کاربردهای هوش مصنوعی توضیح‌پذیر

۶-۱. بهداشتی و درمانی

در بخش مراقبت‌های بهداشتی، استفاده از فناوری‌های هوش مصنوعی وعده دگرگونی و تحول در تشخیص‌ها، برنامه‌ریزی درمان، طراحی دارو و مراقبت از بیماران را می‌دهد. بااین‌حال، پیچیدگی تصمیمات پزشکی نیازمند شفافیت و قابلیت توضیح‌پذیری است. هوش مصنوعی توضیح‌پذیر می‌تواند به پزشکان و بیماران اطلاعات واضح و قابل درکی درباره چگونگی ارائه توصیه‌های تشخیصی یا درمانی ارائه دهد [۱].

برای مثال، الگوریتم‌های هوش مصنوعی که در رادیولوژی برای تشخیص پزشکی استفاده می‌شوند، باید توضیح دهند چرا مناطق خاصی از تصویر را به‌عنوان نگران‌کننده علامت‌گذاری کرده‌اند؟ این موضوع به رادیولوژیست‌ها اجازه می‌دهد تا نتایج هوش مصنوعی را اعتبارسنجی کنند تا هیچ اطلاعات مهمی نادیده گرفته نشود و دقت تشخیص را بهبود می‌دهد. علاوه‌بر این، هوش مصنوعی توضیح‌پذیر می‌تواند منطق پشت برنامه‌های درمان شخصی‌سازی شده را روشن کند که به پزشکان امکان می‌دهد تصمیمات آگاهانه‌تری بگیرند و اعتماد بیماران به مراقبت‌های پزشکی با کمک هوش مصنوعی را بیشتر کنند [۲۶].

یکی دیگر از کاربردهای هوش مصنوعی توضیح‌پذیر در منابع آب و فاضلاب است، به‌طوری‌که به سیاستگذاران اجازه می‌دهد اقدامات پیشگیرانه بهتری برای اطمینان از ایمنی آب آشامیدنی انجام دهند. براساس نتایج این مطالعه [۲۷]، روش‌های سنتی ارزیابی کیفیت اغلب توانایی ارائه توضیحات دقیق و قابل درک برای یافته‌های خود را ندارند، اما با استفاده از هوش مصنوعی توضیح‌پذیر، این مطالعه راهی برای پر کردن این شکاف ارائه می‌دهد تا دلیل پیش‌بینی‌های هوش مصنوعی فهمیده و اقدامات، آگاهانه انجام شود. نتایج این مطالعه نشان می‌دهد که استفاده از «هوش مصنوعی توضیح‌پذیر» نه‌تنها به بهبود دقت ارزیابی‌های ایمنی آب کمک می‌کند، بلکه می‌تواند به بهبود کلی سلامت عمومی در مناطق شهری منجر شود.

اهمیت هوش مصنوعی توضیح‌پذیر در مراقبت‌های بهداشتی به ملاحظات اخلاقی نیز گسترش می‌یابد. با شفاف کردن فرایندهای تصمیم‌گیری هوش مصنوعی، ارائه‌دهندگان مراقبت‌های بهداشتی می‌توانند هرگونه تعصب در مدل‌های هوش مصنوعی را شناسایی و رفع کنند. این موضوع باعث می‌شود که همه بیماران به‌صورت عادلانه و منصفانه، بدون در نظر گرفتن پس‌زمینه‌شان، درمان شوند. علاوه‌بر این، نهادهای نظارتی مانند سازمان غذا و دارو، خواهان توضیح برای سیستم‌های هوش مصنوعی مورد استفاده در مراقبت‌های بهداشتی هستند تا استانداردهای ایمنی و اثربخشی را برآورده کنند [۲۸].

۶-۲. مالی و اقتصادی

در بخش مالی، هوش مصنوعی برای وظایف مختلفی ازجمله امتیازدهی اعتباری، تشخیص تقلب و تجارت الگوریتمی استفاده

می‌شود. ماهیت مبهم بسیاری از سیستم‌های هوش مصنوعی می‌تواند به بی‌اعتمادی و نگرانی‌های اخلاقی منجر شود، به‌ویژه زمانی که این سیستم‌ها تصمیمات مهمی را اتخاذ می‌کنند که بر رفاه افراد تأثیر می‌گذارد. هوش مصنوعی توضیح‌پذیر به این مسائل پرداخته و شفافیت لازم در نحوه اتخاذ تصمیمات مالی را فراهم می‌کند.

سیستم‌های امتیازدهی اعتباری مانند FICO [۲۹] و Vantage [۳۰] که در بانک‌های آمریکایی مورد استفاده قرار می‌گیرند به‌راحتی قابل تفسیر هستند و در اسناد مربوط به این سیستم‌های امتیازدهی می‌توان دید که هرکدام از این دو سیستم به یک ویژگی مالی و اعتباری مانند پرداخت به‌موقع اقساط وام به چه حد اهمیت داده‌اند. به‌عبارت‌دیگر در این سیستم‌ها برای کاربر شفاف است که به چه علتی از امتیاز مالی آنها کاسته شده است.

به همین ترتیب با استفاده روزافزون هوش مصنوعی در زمینه تشخیص کلاهبرداری و همچنین تخلفات مالی، ضرورت استفاده از مدل‌های توضیح‌پذیر در این قسمت هم دیده می‌شود و تلاش شده است تا این موارد به کمک مدل‌های توضیح‌پذیر موارد بیشتری از صرف پیش‌بینی ارائه کنند و به ما بگویند که دقیقاً کدام ویژگی از پرونده و مشخصات ورودی مدنظر، ما را مشکوک به وجود مشکلات اقتصادی در پرونده کرده است.

از دیگر کاربردهای این شاخه در زمینه اقتصادی می‌توان به این مورد اشاره کرد که این روش‌ها در سایر زمینه‌های اقتصادی نیز قابل اعمال هستند و قابلیت ارتقای کارایی این بخش‌ها را ایجاد می‌کنند. برای مثال در زمینه توزیع برق می‌توان این روش‌ها شامل: مدل‌های توضیح‌پذیر برای مشکلات مختلف بخش توزیع برق ازجمله مدیریت منابع، تشخیص آسیب‌پذیری‌های شبکه و همچنین پیش‌بینی مقدار برق تولیدی از منابع تجدیدپذیر را به‌کار گرفت. تمامی موارد گفته شده قابلیت پیش‌بینی به کمک هوش مصنوعی را دارند و به کمک روش‌های توضیح‌پذیری، شفافیت و اعتمادپذیری این مدل‌ها افزایش می‌یابد [۵].

برای نمونه آژانس رگولاتوری انرژی برزیل (ANEEL) مدلی بر مبنای افزایش رضایت مشترکین طراحی کرده است و سعی می‌کند براساس این مدل، تعیین کند که مهم‌ترین ویژگی‌ها جهت کسب رضایت مشترکین چیست تا در مراحل بعد بر مبنای این ویژگی‌ها قوانینی وضع کند که هزینه‌ها را کاهش داده و همچنین شفافیت و اعتمادپذیری را ارتقا دهد [۳].

۳-۶. حقوقی

در بخش حقوقی از هوش مصنوعی برای کاربردهایی مانند پیش‌بینی نتایج پرونده، تحلیل اسناد و بررسی قراردادها استفاده می‌شود. بااین‌حال، استفاده از هوش مصنوعی در تصمیم‌گیری‌های حقوقی نگرانی‌های قابل‌توجهی درباره شفافیت و پاسخ‌گویی را مطرح می‌کند. هوش مصنوعی توضیح‌پذیر می‌تواند این نگرانی‌ها را با ارائه توضیحات واضح و قابل‌فهم برای بینش‌های حقوقی مبتنی‌بر هوش مصنوعی برطرف کند.

برای نمونه در یکی از استفاده‌هایی که هوش مصنوعی توضیح‌پذیر در این زمینه داشته، به این صورت است که مشخص می‌کند با توجه به قوانین موجود و اسناد پرونده خروجی مدل به کدام‌یک از قوانین بیشتر ارتباط دارد. برای مثال اگر مدل تشخیص داده قانونی نقض شده است، کدام‌یک از آنهاست [۳۱].

۴-۶. حمل‌ونقل

حوزه حمل‌ونقل شامل: سیستم‌های ناوبری، سیستم‌های پشتیبانی تصمیم برای خودروهای خودران و سیستم‌های تغییر مسیر پرواز برای صنعت هوانوردی است که توضیحات موردمحور[1] برای سیستم‌های پشتیبانی تصمیم‌گیری خودکار خودرو ترجیح داده می‌شوند. علاوه‌بر این، توضیحات ارائه شده به کاربر می‌تواند بصری، متنی، نشانگرهای نوری یا حالت ترکیبی باشد. برای سیستم‌های ناوبری،

1 .Case-based

نیازهای کاربر کمی متفاوت است، زیرا کاربران به توضیحات درخواست شده در مورد خاص و همچنین استدلال مناسب در پشت هر تصمیمی نیاز دارند [۲۲].

۶-۵. وسایل نقلیه خودران

توسعه وسایل نقلیه خودران به‌شدت به هوش مصنوعی برای ناوبری در محیط‌های رانندگی پیچیده متکی است. هوش مصنوعی توضیح‌پذیر در این بخش برای ایمنی و پذیرش عمومی خودروهای خودران مهم است. توضیح‌پذیری شفافیت لازم در نحوه تصمیم‌گیری‌های لحظه‌ای وسایل نقلیه خودران را فراهم می‌کند که از تشخیص علائم ترافیکی تا اجتناب از موانع را شامل می‌شود. برای مثال، اگر یک وسیله نقلیه خودران به‌طور ناگهانی توقف کند، هوش مصنوعی توضیح‌پذیر می‌تواند عواملی که منجر به این تصمیم شده‌اند، مانند شناسایی عابر پیاده یا تغییرات ناگهانی در شرایط جاده را روشن کند[۳۲]. همان‌طور که در شکل ۲ نشان داده شده، در خروجی مدل که تصویر سمت راست است ماشین جلویی که مانع ماست، به‌عنوان یک قید مهم در تصمیم‌گیری دیده می‌شود.

شکل ۲. نمونه‌ای از کاربرد هوش مصنوعی توضیح‌پذیر در ماشین‌های خودران [۳۲]

در مجموع می‌توان گفت هوش مصنوعی توضیح‌پذیر دارای تأثیرات گسترده‌ای در بخش‌های مختلف است و شفافیت، اعتماد و پاسخ‌گویی را افزایش می‌دهد. با شفاف و قابل‌فهم کردن فرایندهای تصمیم‌گیری هوش مصنوعی تضمین می‌کنیم که خطرات ناشی از به‌کارگیری آن فناوری به حداقل خواهد رسید. با ادامه تکامل و ادغام هوش مصنوعی در جنبه‌های مختلف جامعه، نقش توضیح‌پذیری مهم‌تر خواهد شد و آینده توسعه و به‌کارگیری هوش مصنوعی را شکل خواهد داد.

۶-۶. رسانه و سرگرمی

این حوزه شامل: سیستم‌های توصیه موسیقی، سیستم‌های توصیه فیلم، سیستم‌های توصیه هنری برای انجمن‌ها و وب‌سایت‌ها، سیستم‌های توصیه مقاله خبری و سیستم‌های بازی است. در این موارد کاربران به توضیحاتی نیاز دارند که حاوی جزئیات توصیه‌های شخصی‌سازی شده باشد که شامل اصول اولیه کار سیستم و جزئیات مربوط به اطلاعات شخصی استفاده شده است. کاربران همچنین نگرانی خود را در مورد میزان اطلاعات ارائه شده ابراز می‌کنند. زیرا اطلاعات بیش از حد می‌تواند بار شناختی ایجاد کند. بنابراین، سیستم نباید کاربر را با اطلاعات غیرضروری و مبهم غرق کند و باید هم توضیحات متنی و هم تصویری ارائه دهد [۲۲].

۶-۷. آموزش

حوزه آموزشی شامل: سیستم‌های تدریس خصوصی هوشمند، تصمیم‌گیری در مورد پذیرش دانشگاه و سیستم‌های تخمین نمره است که ایجاد توضیح‌پذیری در آنها، قابلیت استفاده از سیستم را بهبود می‌بخشد. توضیحات باید اطلاعاتی در مورد رفتار سیستم و روش کار و همچنین منطق پشت برخی وظایف تصمیم‌گیری، مانند تصمیم‌گیری پذیرش، ارائه دهد [۲۲].

۶-۸. خرده‌فروشی

در صنعت خرده‌فروشی، هوش مصنوعی به‌طور گسترده‌ای برای توصیه‌های شخصی‌سازی شده، مدیریت موجودی و خدمات مشتری استفاده می‌شود. هوش مصنوعی توضیح‌پذیر این کاربردها را با شفاف و قابل‌فهم کردن فرایندهای تصمیم‌گیری پشت توصیه‌های هوش مصنوعی تقویت می‌کند. برای مثال، با استفاده از روش‌های توضیح‌پذیری در سیستم‌های توصیه شخصی‌سازی شده می‌توان توضیح داد که محصول توصیه شده برای کاربر مدنظر به چه دلیلی پیشنهاد شده است که باعث افزایش اعتمادپذیری این محصولات می‌شود [۳۳].

۶-۹. مدیریت منابع انسانی

در مدیریت منابع انسانی با استفاده از نتایج الگوریتم‌ها مانند تصمیم‌گیری درخصوص استخدام یا ارتقای افراد، نگرش کارکنان نسبت به پذیرش تصمیمات مبتنی‌بر هوش مصنوعی تحت‌تأثیر عوامل مختلفی قرار دارد و کارمندان اغلب بر این باورند که تصمیم‌گیری می‌تواند مغرضانه، دست‌کاری شده و تجاوز به حریم خصوصی باشد. کاهش شکاف دانش با افزایش شفافیت و توضیح‌پذیری نتایج می‌تواند به درک و پذیرش تصمیمات کمک کند. برای نمونه ارائه دلایل مناسب در پشت تصمیم‌گیری، توضیح نمرات ارزیابی داوطلب و نشان دادن استخدام‌های مشابه در سازمان، علاوه‌بر این، همکاری با کاربران انسانی در مرحله طراحی می‌تواند شانس پذیرش سیستم را افزایش دهد الزامات استخدام‌کننده برای توضیح‌پذیری شامل است [۲۲].

۷. فواید جانبی و آینده هوش مصنوعی توضیح‌پذیر

۷-۱. تعمیم‌پذیری

یکی از چالش‌های اصلی در هوش مصنوعی، تعمیم‌پذیری مدل‌ها به داده‌های جدید و ناشناخته است. مدل‌های توضیح‌پذیر به محققان این امکان را می‌دهند تا درک بهتری از نحوه یادگیری مدل‌ها و روابط بین ویژگی‌های مختلف داشته باشند. این امر می‌تواند به بهبود تعمیم‌پذیری مدل‌ها کمک کند، زیرا توسعه‌دهندگان می‌توانند از اطلاعات تفسیر شده برای تنظیم مدل‌ها به‌گونه‌ای استفاده کنند که در شرایط و داده‌های مختلف عملکرد بهتری داشته باشند [۲۲].

۷-۲. استحکام و مقاومت

مدل‌های هوش مصنوعی توضیح‌پذیر می‌توانند به بهبود استحکام سیستم‌های هوش مصنوعی کمک کنند. استحکام به توانایی یک سیستم در مقابله با داده‌های غیرمنتظره و تغییرات محیطی اشاره دارد. وقتی که مدل‌های هوش مصنوعی به‌طور شفاف توضیح دهند که چگونه به تصمیمات خود رسیده‌اند، توسعه‌دهندگان می‌توانند بهتر نقاط ضعف و قوت سیستم را شناسایی کرده و بهبودهای لازم را اعمال کنند [۱۴] و [۱۵].

۷-۳. امنیت

در حوزه امنیت، هوش مصنوعی توضیح‌پذیر می‌تواند نقش حیاتی ایفا کند. مدل‌های هوش مصنوعی که قابل تفسیر باشند، می‌توانند به شناسایی و مقابله با حملات سایبری کمک کنند. برای مثال، در سیستم‌های تشخیص نفوذ، مدل‌های توضیح‌پذیر می‌توانند به تحلیلگران امنیتی نشان دهند که چگونه یک حمله تشخیص داده شده است و این امکان را فراهم کنند تا روش‌های مقابله مؤثرتری طراحی شود.

۷-۴. روندها و پیش‌بینی‌ها

حوزه هوش مصنوعی توضیح‌پذیر به سرعت در حال تکامل است و این امر به‌دلیل نیاز روزافزون به شفافیت، اعتماد و پاسخ‌گویی در سیستم‌های هوش مصنوعی است. چندین روند کلیدی و پیش‌بینی‌ها نشان‌دهنده جهت‌گیری آینده این حوزه هستند [۱۵]:

– **ادغام با هوش مصنوعی رایج:** تکنیک‌های توضیح‌پذیری در حال تبدیل شدن به جزء جدایی‌ناپذیر توسعه سیستم‌های هوش مصنوعی در صنایع مختلف هستند. با پیشرفت هوش مصنوعی، گنجاندن توضیح‌پذیری از مرحله طراحی اولیه به یک رویه استاندارد تبدیل خواهد شد تا اینکه پس از توسعه به آن افزوده شود. این تغییر برای اثربخشی و قابلیت فهم سیستم‌های هوش مصنوعی از ابتدا ضروری است.

پیشرفت در تکنیک‌های جدید و بهبود یافته برای توضیح‌پذیری به‌طور مداوم در حال ظهور هستند. روش‌هایی مانند توضیحات مقایسه‌ای، استدلال‌های خلاف واقع و استنتاج علّی در حال رشد هستند. برای مثال، توضیحات مقایسه‌ای بر برجسته کردن تغییراتی در ورودی تمرکز می‌کنند که منجر به نتیجه‌ای متفاوت می‌شود که به کاربران کمک می‌کند تا مرز تصمیم‌گیری مدل را درک کنند.

– **توضیحات شخصی‌سازی شده:** تأکید بیشتری بر ارائه توضیحات برای گروه‌های مختلف کاربران وجود دارد. توضیحات شخصی‌سازی شده با در نظر گرفتن پس‌زمینه، ترجیحات و نیازهای کاربر، اطمینان حاصل می‌کنند که اطلاعات ارائه شده برای هر فرد، مربوط و قابل‌فهم است. برای مثال، یک تحلیلگر مالی ممکن است به توجیهات آماری دقیق برای امتیاز اعتباری نیاز داشته باشد، درحالی‌که یک فرد عادی ممکن است فقط به نسخه ساده‌شده‌ای نیاز داشته باشد که عوامل کلیدی تأثیرگذار بر امتیاز او را شرح دهد.

– **اخلاق و عدالت در هوش مصنوعی:** توضیح‌پذیری در رسیدگی به نگرانی‌های اخلاقی و عدالت در سیستم‌های هوش مصنوعی نقش ایفا خواهد کرد. با شفاف کردن تصمیمات هوش مصنوعی، شناسایی و کاهش سوگیری‌ها آسان‌تر می‌شود که منجر به نتایج عادلانه‌تر برای گروه‌های مختلف جمعیتی می‌شود. این امر به‌ویژه در بخش‌هایی مانند استخدام و اجرای قانون اهمیت دارد، جایی که سیستم‌های هوش مصنوعی متعصب می‌توانند تأثیرات منفی چشمگیری داشته باشند.

۷-۵. تأثیر بلندمدت در جامعه

با پذیرش گسترده تکنیک‌های توضیح‌پذیری انتظار می‌رود که مزایای بلندمدت قابل‌توجه و تغییرات تحول‌آفرینی در جنبه‌های مختلف جامعه ایجاد شود[۴]، [۱۴]، [۱۵] و [۳۴]:

– **افزایش اعتماد به هوش مصنوعی:** با شفاف و قابل فهم‌تر شدن سیستم‌های هوش مصنوعی، اعتماد عمومی به فناوری‌های هوش مصنوعی افزایش خواهد یافت. این اعتماد برای پذیرش گسترده‌تر و یکپارچگی هوش مصنوعی در زندگی روزمره ضروری است. سیستم‌های هوش مصنوعی قابل اعتماد در بخش‌های حیاتی مانند مراقبت‌های بهداشتی، مالی و خدمات عمومی سریع‌تر پذیرفته می‌شوند که منجر به بهبود کارایی و نتایج خواهد شد.

– **بهبود تصمیم‌گیری:** توضیح‌پذیری منجر به تصمیم‌گیری بهتر در چندین حوزه می‌شود. متخصصین در زمینه‌هایی مانند مراقبت‌های بهداشتی، مالی و خدمات حقوقی می‌توانند هوش مصنوعی را به‌طور مؤثرتری به‌کار گیرند که نتایج بهبود یافته و کاهش ریسک‌ها را به همراه دارد.

– **هوش مصنوعی اخلاقی و عادلانه:** با فهمیدن سوگیری‌ها، توضیح‌پذیری به توسعه سیستم‌های هوش مصنوعی اخلاقی کمک خواهد کرد. این امر به جلوگیری از ترویج تبعیض کمک خواهد کرد که منجر به جامعه‌ای عادلانه‌تر خواهد شد. شیوه‌های اخلاقی هوش مصنوعی به‌ویژه در حوزه‌هایی مانند استخدام، وام‌دهی و اجرای قانون که عدالت بسیار اهمیت دارد، بسیار مهم خواهد بود.

– **رشد اقتصادی و نوآوری:** سیستم‌های هوش مصنوعی شفاف، نوآوری و رشد اقتصادی را با به وجود آوردن امکان توسعه برنامه‌ها

و خدمات جدید رشد خواهند داد. کسب‌وکارهایی که توضیح‌پذیری را به‌کار می‌گیرند، با افزایش اعتماد و رضایت مشتری، مزیت رقابتی کسب خواهند کرد.

۸. چالش‌های پیاده‌سازی هوش مصنوعی توضیح‌پذیر

پیاده‌سازی هوش مصنوعی توضیح‌پذیر شامل چندین چالش است که جنبه‌های فنی، عملکردی و انسانی را دربرمی‌گیرد. این چالش‌ها باید برطرف شوند تا سیستم‌هایی ایجاد شوند که هم مؤثر و هم قابل تفسیر باشند.

۸-۱. چالش‌های فنی

یکی از چالش‌های اصلی فنی در توضیح‌پذیری پیچیدگی ذاتی مدل‌های یادگیری ماشین مدرن است. روش‌هایی مانند یادگیری عمیق شامل معماری‌های بسیار پیچیده با لایه‌ها و پارامترهای فراوان هستند که ردیابی فرایند تصمیم‌گیری را دشوار می‌کند. ساده‌سازی این مدل‌های پیچیده برای قابل تفسیر کردن آنها اغلب به روش‌های پیچیده‌ای نیاز دارد. علاوه‌بر این، توسعه روش‌هایی که بتوانند توضیحات را برای انواع مختلف مدل‌ها ارائه دهند چالش قابل‌توجهی است. به‌عبارت‌دیگر اکثر روش‌های توسعه داده شده یا مخصوص یک مدل خاص تولید شده‌اند و یا اینکه در اجرا و پیاده‌سازی مشکل‌های خاص خودشان را دارند. ضمن اینکه فقدان یک استاندارد و تعریف مشخص و مورد قبول برای توضیح‌پذیری، امکان ارزیابی آن را با مشکل مواجه می‌کند.

۸-۲. توازن بین عملکرد و توضیح‌پذیری

یکی از معضلات اصلی در این شاخه تعادل بین دقت مدل و قابلیت تفسیر آن است. مدل‌های بسیار دقیق مانند شبکه‌های عصبی عمیق اغلب پیچیده و مبهم هستند. ازسوی‌دیگر، مدل‌های ساده‌تر توضیحات واضح‌تری ارائه می‌دهند، اما ممکن است دقت لازم برای برخی وظایف را نداشته باشند. رسیدن به تعادل میان عملکرد بالا و قابلیت تفسیر، چالش‌برانگیز است. کاهش پیچیدگی مدل برای بهبود قابلیت تفسیر می‌تواند منجر به کاهش عملکرد شود که ممکن است در کاربردهای حیاتی مانند تشخیص پزشکی یا رانندگی خودکار قابل‌قبول نباشد. برعکس، افزایش دقت مدل با افزودن لایه‌های پیچیده می‌تواند فرایند تصمیم‌گیری را مبهم کند و توضیح آنها دشوار شود. محققان فعالانه در حال بررسی روش‌های ترکیبی هستند که تلاش می‌کند نقاط قوت هر دو مدل پیچیده و ساده را ترکیب کند. روش‌هایی مانند شبکه‌های عصبی توضیح‌پذیر که عناصر طراحی شده برای افزایش شفافیت را شامل می‌شوند و روش‌های تفسیری که مدل‌ها را پس از آموزش تحلیل می‌کنند، نمونه‌هایی از تلاش‌ها برای این تعادل هستند [۷]، [۱۴] و [۲۲].

۸-۳. عوامل انسانی

عوامل انسانی نقش مهمی در پیاده‌سازی توضیح‌پذیری دارند. یکی از چالش‌های اصلی این است که توضیحات سیستم‌های هوش مصنوعی، برای مخاطبان هدف (متفاوت در تخصص فنی و توانایی‌های شناختی) قابل درک باشد. توضیحات باید با سطح درک کاربران تنظیم شوند، یعنی از اصطلاحات فنی خودداری کنند و درعین‌حال مفاهیم لازم را منتقل کنند.

علاوه‌بر این، مسئله رضایت از توضیح وجود دارد. حتی اگر توضیحی به‌صورت فنی درست و قابل‌فهم باشد، ممکن است همیشه نیاز کاربران به شفافیت را برآورده نکند. کاربران اغلب به دنبال توضیحاتی هستند که نه‌تنها چگونگی اتخاذ یک تصمیم را توصیف کنند، بلکه چرایی اتخاذ آن را نیز به طریقی که با ارزش‌ها و انتظارات آنها همخوانی داشته باشد، توضیح دهند. برآوردن این نیاز، مستلزم رویکردی ظریف است که واکنش‌های روان‌شناختی و احساسی کاربران را در نظر بگیرد.

رسیدگی به این عوامل انسانی، مستلزم پژوهش‌های بین‌رشته‌ای ازجمله به‌کارگیری بینش‌های روان‌شناختی، علوم‌شناختی، تعامل انسان و کامپیوتر با هدف ارائه توضیحاتی است که شهودی و قابل اعتماد باشند. این توضیحات می‌تواند پذیرش و اعتماد کاربران به سیستم‌های هوش مصنوعی را افزایش داده و درنهایت اثربخشی و تأثیر اجتماعی آنها را بهبود بخشد[۲۲] و [۳۵].

یک چالش مهم دیگر این است که توضیحات می‌تواند اعتماد کاربر را افزایش دهد؛ اما در بلندمدت، خروجی‌های مدل‌ها ممکن است توصیه‌های دقیقی را ارائه نکنند. چنین خطری ممکن است باعث شود کاربران اعتماد به نفس نادرست داشته باشند و به نتایج اشتباه اعتماد کنند. این امر نیز نیازمند دقت ویژه است [۷].

۴-۸. پیش‌نیازهای راهبردی

برای پیاده‌سازی موفق یک هوش مصنوعی توضیح‌پذیر، سه سؤال مهم باید پاسخ داده شود [۳۶]: یک توضیح خوب چه چیزی است؟ این توضیح برای چه کسی است؟ و درنهایت چگونه ارائه خواهد شد؟ توضیحات باید متناسب با زمینه و نیازهای کاربر و هدف سیستم باشند. تحقیقات نشان می‌دهد که افراد به دلایل مختلفی ازجمله پیش‌بینی رویدادهای مشابه، تشخیص، ارزیابی مقصر بودن یا بی‌گناهی، توجیه یا عقلانی‌سازی یک اقدام و غیره به توضیحات نیاز دارند.

کیفیت یک توضیح همچنین به نحوه ارائه آن بستگی دارد. توضیحات می‌توانند در لحظه و به‌صورت مستمر ارائه شوند یا پس از وقوع اتفاق و خلاصه شده باشند. توضیحات تعاملی معمولاً ترجیح داده می‌شوند، اما همیشه مناسب یا قابل اجرا نیستند. فرمت‌های متنی، تصویری و چندرسانه‌ای با نتایج متفاوت مورد مقایسه قرار گرفته‌اند و پاسخ‌های متنی آشنا یا توضیحات تصویری ساده برای کاربران غیرمتخصص اغلب مؤثرتر هستند.

۹. جمع‌بندی و نتیجه‌گیری

در این گزارش، هوش مصنوعی توضیح‌پذیر به‌عنوان یک زمینه پژوهشی و عملیاتی در حوزه هوش مصنوعی بررسی شد. اهمیت این موضوع در تسهیل عمومی‌سازی فناوری هوش مصنوعی در بخش‌های مختلف است.

آینده حوزه هوش مصنوعی توضیح‌پذیر به‌شدت به تلاش‌های مشترک ذی‌نفعان مختلف، ازجمله خط‌مشی‌گذاران، فعالان صنعتی، پژوهشگران و کاربران نهایی بستگی دارد.

خط‌مشی‌گذاران، قانونگذاران و نهادهای نظارتی: دولت‌ها و نهادهای نظارتی باید خط‌مشی‌هایی ایجاد و اجرا کنند که استفاده از هوش مصنوعی توضیح‌پذیر در بخش‌های حساس را الزام‌آور می‌سازد. این امر شامل: به‌روزرسانی مقررات موجود برای گنجاندن الزامات شفافیت و پاسخ‌گویی در سیستم‌های هوش مصنوعی می‌شود. خط‌مشی‌گذاران باید از ابتکارات پژوهشی حمایت کنند و بودجه‌ای برای توسعه روش‌های جدید توضیح‌پذیری فراهم کنند. ایجاد استانداردهایی مبتنی‌بر توضیح‌پذیری در ورود این فناوری به بخش عمومی می‌تواند مفید و ضروری باشد. به‌علاوه سازوکارهای تضمین اینکه استفاده از تکنیک‌های توضیح‌پذیری به‌هیچ‌وجه حریم خصوصی و امنیت داده‌های کاربران را نقض نمی‌کند، ضرورت دارد. همچنین توجه به بحث توضیح‌پذیری در نهادهای راهبری هوش مصنوعی و نهادهای نظارتی بخش عمومی مبتنی‌بر اهمیت لحاظ شیوه‌های تضمین این توضیح‌پذیری جهت ارزیابی نتایج حاصله و تصمیمات متخذه می‌تواند در توسعه این شاخه از هوش مصنوعی در بخش عمومی مؤثر واقع شود.

هم‌آفرینی بخش‌های مختلف: با توجه به پیچیدگی فناوری هوش مصنوعی که اقتضائات اجرایی مهمی دارد و ازطرف‌دیگر اهمیت پذیرش استانداردها، شرکت‌های توسعه‌دهنده فناوری هوش مصنوعی، بخش‌های دانشگاهی و دستگاه‌های نظارتی باید برای ایجاد

استانداردهای توضیح‌پذیری همکاری و هم‌آفرینی کنند. چارچوب‌ها و شیوه‌های بهینه مشترک به سازگاری و اعتماد در پیاده‌سازی‌ها کمک خواهد کرد. این امر منجر به کیفیت بهتر و اعتماد بیشتر به سیستم‌های هوش مصنوعی می‌شود.

مؤسسات دانشگاهی و پژوهشی: پژوهشگران به‌دلیل توسعه روش‌های جدید و اعتبارسنجی آنها قطعه مهمی در جورچین پیشرفت توضیح‌پذیری هستند. سرمایه‌گذاری مستمر در پژوهش هوش مصنوعی نوآوری را فعال می‌کند و کیفیت کلی سیستم‌های توضیح‌پذیر را بهبود می‌دهد. مؤسسات نیز می‌توانند پژوهش‌های بین‌رشته‌ای ازجمله علوم‌شناختی، روان‌شناسی، علوم کامپیوتر و هوش مصنوعی را برای پرداختن به چالش‌های پیچیده این حوزه شروع و نسبت به پرداخت بورسیه‌های تحصیلی مرتبط و تشویق پژوهش‌ها به این سمت اقدام کنند.

آگاه‌سازی و مشارکت عمومی: آگاه‌سازی عمومی درباره اهمیت توضیح‌پذیری و تأثیر آن بر زندگی روزمره بسیار مهم است. بسیار مهم است که عموم مردم و کسانی که نتایج الگوریتم‌ها بر آنها اعمال خواهد شد، نسبت به الگوریتمی بودن تصمیمات و سطحی از توضیح نسبت به نتایج، آگاه باشند. افزایش آگاهی علاوه‌بر ارتقای ظرفیت پذیرش فناوری، تقاضا برای راه‌حل‌های شفاف هوش مصنوعی را افزایش می‌دهد و کاربران را قادر می‌سازد تا تصمیمات آگاهانه‌ای درباره فناوری‌هایی که استفاده می‌کنند، بگیرند. مشارکت عمومی می‌تواند بازخوردهای ارزشمندی به توسعه‌دهندگان ارائه دهد و به آنها در اصلاح و بهبود سیستم‌های توضیح‌پذیر کمک کند.

ارزیابی سیستم‌های هوش مصنوعی در بخش عمومی: با توجه به اهمیت و گستره تأثیر فناوری هوش مصنوعی و پیش‌بینی فراگیری و پذیرش بالای این فناوری در سال‌های آینده، ضروری است سامانه‌های مبتنی‌بر هوش مصنوعی در این بخش به‌صورت دوره‌ای از منظر سوگیری‌های احتمالی و اعتبار نتایج، پایش شده و به سمت سطحی از توضیح‌پذیری سوق داده شوند.

درنتیجه، آینده هوش مصنوعی توضیح‌پذیر وعده‌های بزرگی برای افزایش اعتماد، پاسخ‌گویی و استانداردهای اخلاقی در سیستم‌های هوش مصنوعی دارد. با پذیرش این روندها و ترویج همکاری بین ذی‌نفعان، جامعه می‌تواند از ظرفیت کامل هوش مصنوعی بهره‌مند شود و درعین‌حال بداند که این فناوری به‌صورت شفاف و عادلانه عمل می‌کند. به‌طور خلاصه، حوزه هوش مصنوعی توضیح‌پذیر نقش قابل‌توجهی در جهت افزایش شفافیت، اعتماد و مسئولیت‌پذیری در سیستم‌های هوش مصنوعی دارد. همان‌طور که در بخش‌های قبلی بیان شد، این گزارش شامل تعریف، روندها، چالش‌های فنی و عوامل انسانی است که برای توسعه و اجرای آن بسیار مهم هستند. پیش‌بینی‌ها در این‌باره نشان‌دهنده چشم‌انداز رو به رشد، ازجمله یکپارچگی با سیستم‌های قبلی، توضیحات شخصی‌سازی شده و وارد شدن ملاحظات اخلاقی است که منعکس‌کننده نیاز هرچه بیشتر به این شاخه در زمینه‌های مختلف است. چالش‌های فنی، مانند پیچیدگی مدل‌های مدرن یادگیری ماشین، ضرورت پیدا کردن راه‌حل‌های نوآورانه برای ساده‌سازی مدل‌ها بدون فدا کردن دقت را بیان می‌کنند. دستیابی به تعادل بین عملکرد و قابل توضیح بودن یک چالش اصلی است که نیاز به بررسی دقیق و رویکردهای میان‌رشته‌ای دارد. علاوه‌بر این، عوامل انسانی نقش اساسی در اجرای موفقیت‌آمیز هوش مصنوعی توضیح‌پذیر دارند. پرداختن به نیازها و ترجیحات متنوع کاربران و اطمینان از اینکه توضیحات قابل‌فهم و قابل اعتماد هستند، از اهمیت بالایی برخوردار است.

در پایان باید گفت، هوش مصنوعی توضیح‌پذیر مسیری را برای افزایش اعتماد، شفافیت و مسئولیت‌پذیری در سیستم‌های هوش مصنوعی به ما نشان داده است و راه را برای پذیرش گسترده‌تر این فناوری هموار می‌سازد.

منابع و مآخذ

[1] A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” IEEE Access, vol. 6, pp. 52138–52160, Sep. 2018, doi: 10.1109/ACCESS.2018.2870052.

[2] C. van Noordt and G. Misuraca, “Artificial intelligence for the public sector: results of landscaping the use of AI in government across the European Union,” Gov Inf Q, vol. 39, no. 3, 2022, doi: 10.1016/j.giq.2022.101714.

[3] M. S. D. Carvalho and G. L. D. Silva, “Inside the black box: Using Explainable AI to improve Evidence-Based Policies,” in Proceedings - 2021 IEEE 23rd Conference on Business Informatics, CBI 2021 - Main Papers, 2021, pp. 57–64. doi: 10.1109/CBI52690.2021.10055.

[4] A. Das, G. Student Member, P. Rad, and S. Member, “Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey,” Jun. 2020, Accessed: Jan. 04, 2025. [Online]. Available: https://arxiv.org/abs/2006.11371v2

[5] R. Machlev, M. Perl, A. Caciularu, J. Belikov, K. Y. Levy, and Y. Levron, “Explaining the decisions of power quality disturbance classifiers using latent space features,” International Journal of Electrical Power & Energy Systems, vol. 148, p. 108949, Jun. 2023, doi: 10.1016/J.IJEPES.2023.108949.

[6] “The state of AI in 2022—and a half decade in review | McKinsey.” Accessed: Dec. 21, 2024. [Online]. Available:https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review

[7] R. Machlev et al., “Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities,” Energy and AI, vol. 9, p. 100169, Aug. 2022, doi: 10.1016/J.EGYAI.2022.100169.

[8] “Modernising Government White Paper (Hansard, 9 December 1999).” Accessed: Jan. 04, 2025. [Online].Available:https://api.parliament.uk/historic-hansard/westminster-hall/1999/dec/09/modernising-government-white-paper

[9]“BuildingCapacityforEvidence-InformedPolicy-Making,”Sep.2020,doi:10.1787/86331250-EN.

[۱۰]ایمان. اکبری و عطیه، یوسفی و محمدمهدی، مهربان‌هلان .بررسی لایحه برنامه هفتم توسعه (۸۸) : توسعه پایدار هوش مصنوعی در کشور ، ۱۹۳۹۵ ، ماهنامه گزارش‌های کارشناسی مرکز پژوهش‌های مجلس شورای اسلامی، ۳۱(۸)، ۱۹۳۹۵.
: https://report.mrc.ir/article_9802.html

[۱۱] ایمان، اکبری. طراحی سیستم پشتیبان هوشمند قانونگذاری ۱: معرفی رویکرد تحلیل داده‌محور خط‌مشی کاربست هوش مصنوعی و فناوری مبتنی‌بر داده در تحلیل خط‌مشی، ۱۹۸۱۰ ، ۳۲(۳)، ۱۹۸۱۰.
, vol. 32, no. 3, pp. 1–28, May 2024, doi: 10.22034/REPORT.2024.16913.1793.

[12] E. Tjoa and C. Guan, “A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI,” IEEE Trans Neural Netw Learn Syst, vol. 32, no. 11, pp. 4793–4813, Nov. 2021, doi: 10.1109/TNNLS.2020.3027314.

[13] M. Ridley, "Explainable Artificial Intelligence (XAI)," Information Technology and Libraries, vol. 41, no. 2, 2022, doi: 10.6017/ITAL.V41I2.14683.
[14] A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/J.INFFUS.2019.12.012.
[15] S. Ali et al., "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," Information Fusion, vol. 99, p. 101805, Nov. 2023, doi: 10.1016/J.INFFUS.2023.101805.
[16] P. Antunes, V. Herskovic, S. F. Ochoa, and J. A. Pino, "Structuring dimensions for collaborative systems evaluation," ACM Computing Surveys (CSUR), vol. 44, no. 2, Mar. 2008, doi: 10.1145/2089125.2089128.
[17] L. Wells and T. Bednarz, "Explainable AI and Reinforcement Learning—A Systematic Review of Current Approaches and Trends," Front Artif Intell, vol. 4, p. 550030, May 2021, doi: 10.3389/FRAI.2021.550030/BIBTEX.
[18] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds., "Explainable AI: Interpreting, Explaining and Visualizing Deep Learning," vol. 11700, 2019, doi: 10.1007/978-3-030-28954-6.
[19] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 13-17-August-2016, pp. 1135–1144, Aug. 2016, doi: 10.1145/2939672.2939778.
[20] C. Molnar, "Interpretable Machine Learning A Guide for Making Black Box Models Explainable", Accessed: Dec. 18, 2024. [Online]. Available: http://leanpub.com/interpretable-machine-learning
[21] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," ACM Comput Surv, vol. 51, no. 5, Sep. 2018, doi: 10.1145/3236009/SUPPL_FILE/GUIDOTTI.ZIP.
[22] A. B. Haque, A. K. M. N. Islam, and P. Mikalef, "Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research," Technol Forecast Soc Change, vol. 186, p. 122120, Jan. 2023, doi: 10.1016/J.TECHFORE.2022.122120.
[23] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation,'" AI Mag, vol. 38, no. 3, pp. 50–57, Jun. 2016, doi: 10.1609/aimag.v38i3.2741.
[24] "A right to explanation | The Alan Turing Institute." Accessed: Dec. 18, 2024. [Online]. Available: https://www.turing.ac.uk/research/impact-stories/a-right-to-explanation
[25] A. Howard, C. Zhang, and E. Horvitz, "Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems," Proceedings of IEEE Workshop on

Advanced Robotics and its Social Impacts, ARSO, Sep. 2017, doi: 10.1109/ARSO.2017.8025197.
[26] B. M. de Vries, G. J. C. Zwezerijnen, G. L. Burchell, F. H. P. van Velden, C. W. Menke-van der Houven van Oordt, and R. Boellaard, "Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review," Front Med (Lausanne), vol. 10, p. 1180773, May 2023, doi: 10.3389/FMED.2023.1180773/BIBTEX.
[27] N. Hellen and G. Marvin, "Explainable AI for Safe Water Evaluation for Public Health in Urban Settings," 2022 International Conference on Innovations in Science, Engineering and Technology, ICISET 2022, pp. 227–232, 2022, doi: 10.1109/ICISET54810.2022.9775912.
[28] "FDA Releases Artificial Intelligence/Machine Learning Action Plan | FDA." Accessed: Dec. 18, 2024. [Online]. Available: https://www.fda.gov/news-events/press-announcements/fda-releases-artificial-intelligencemachine-learning-action-plan
[29] "Understanding Your FICO® Scores | FICO." Accessed: Dec. 21, 2024. [Online]. Available: https://www.fico.com/en/latest-thinking/fact-sheet/understanding-fico-scores
[30] "User Guide VantageScore 4.0," 2022, Accessed: Dec. 18, 2024. [Online]. Available: www.VantageScore.com
[31] I. Psychoula, A. Gutmann, P. Mainali, S. H. Lee, P. Dunphy, and F. Petitcolas, "Explainable Machine Learning for Fraud Detection," Computer (Long Beach Calif), vol. 54, no. 10, pp. 49–59, Oct. 2021, doi: 10.1109/MC.2021.3081249.
[32] J. Kim and J. Canny, "Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention," Proceedings of the IEEE International Conference on Computer Vision, vol. 2017-October, pp. 2961–2969, Dec. 2017, doi: 10.1109/ICCV.2017.320.
[33] X. Wang, X. He, Y. Cao, M. Liu, and T.-S. Chua, "KGAT: Knowledge Graph Attention Network for Recommendation", doi: 10.1145/3292500.3330989.
[34] B. Leichtmann, C. Humer, A. Hinterreiter, M. Streit, and M. Mara, "Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task," Comput Human Behav, vol. 139, p. 107539, Feb. 2023, doi: 10.1016/J.CHB.2022.107539.
[35] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," Artif Intell, vol. 267, pp. 1–38, Feb. 2019, doi: 10.1016/J.ARTINT.2018.07.007.
[36] M. Ridley, "Explainable Artificial Intelligence (XAI) Adoption and Advocacy," Information Technology and Libraries, vol. 41, no. 2, 2022, doi: 10.6017/ital.v41i2.14683.

گزیده سیاستی

این گزارش به چیستی و اهمیت هوش مصنوعی توضیح پذیر به عنوان یک شاخه نوین از علوم هوش مصنوعی و نقش و ضرورت توضیح پذیری در به کارگیری فناوری هوش مصنوعی در بخش عمومی پرداخته و توصیه‌هایی در این راستا ارائه کرده است.

مرکز پژوهش های مجلس شورای اسلامی

تهران، خیابان پاسداران، روبروی پارک نیاوران (ضلع جنوبی، پلاک ۸۰۲)

تلفن: ۷۵۱۸۳۰۰۰ صندوق پستی: ۵۸۵۵-۱۵۸۷۵ پست الکترونیک: mrc@majles.ir

وب‌سایت: rc.majles.ir