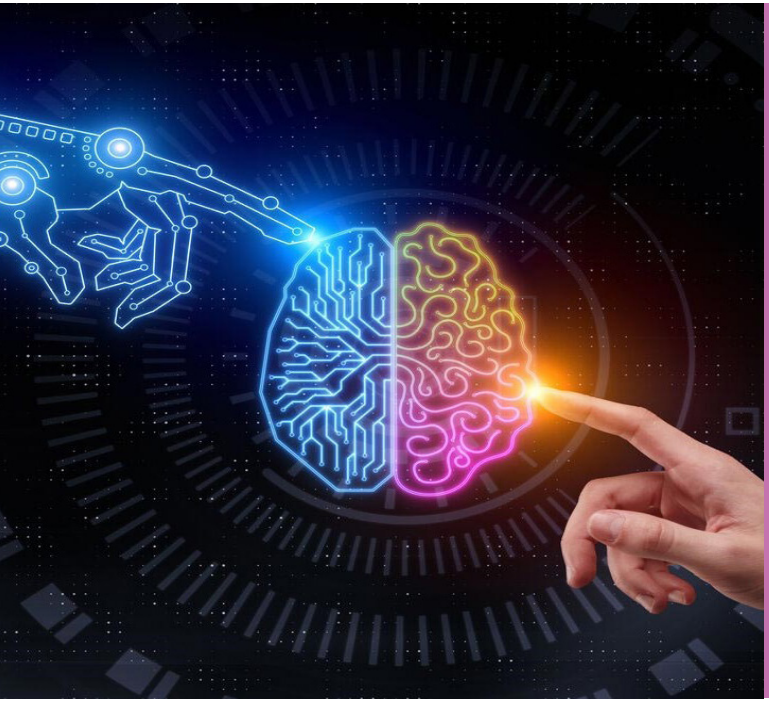
مرکز ملی فضای مجازی
پژوهشگاه فضای مجازی

عصر فضای مجازی

صدوسی



هوش مصنوعی تبیین پذیر

Explainable artificial intelligence

بسم الله الرحمن الرحیم

عصر فضای مجازی

گزارش شماره ۱۳۰

فروردین ۱۴۰۲

مرکز ملی فضای مجازی
پژوهشگاه فضای مجازی

هوش مصنوعی تبیین‌پذیر

محتوای انتشار یافته در این اثر
الزاماً بیانگر دیدگاه مرکز ملی فضای مجازی نیست

تهیه‌شده در: پژوهشگاه فضای مجازی-
گروه مطالعات بنیادین

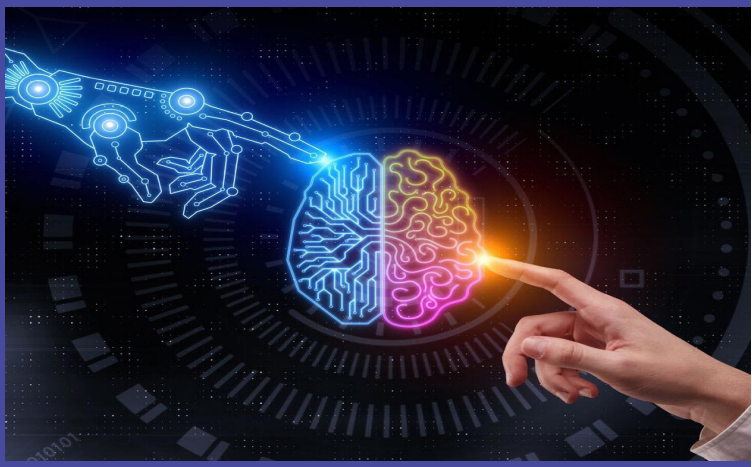
تهیه‌کننده: دکتر امیرحسین زادیوسفی
(استادیار گروه فلسفه دانشگاه تربیت مدرس)

ناظر علمی: دکتر حسین مطلبی کربکندی
(مدیرگروه مطالعات بنیادین)
دکتر مهدی خلیلی

نشانی: تهران، میدان آرژانتین، خیابان بیهقی، نبش خیابان ۱۶ غربی، پلاک ۲۰،
کدپستی ۱۵۱۵۶۷۴۳۱۱
شماره تماس: ۸۶۱۲۱۰۶۱
http://www.majazi.ir

فهرست

سخن نخست



سخن نخست

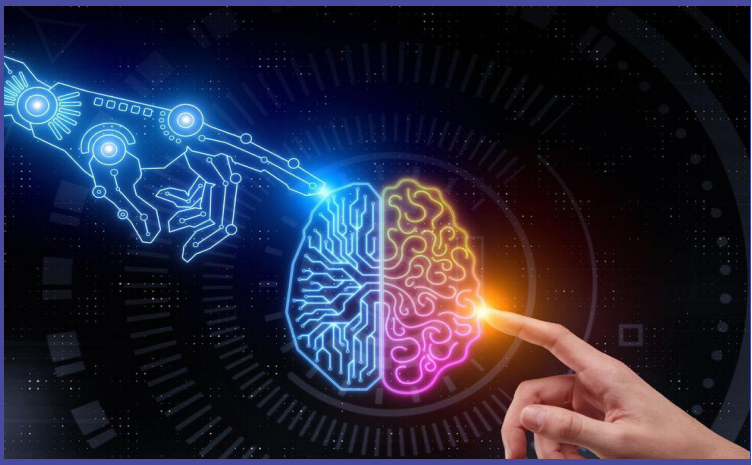
فضای مجازی با شتاب شگرف و رو به تزایدی که در حال بسط و گسترش است تمام ساحات اجتماعی، اقتصادی، سیاسی و فرهنگی زندگی بشر را درنوردیده و هر روز بخش بزرگی از زندگی واقعی را در خود فرو برده و حیات متفاوت و جدیدی به آن می‌دهد. لذا به نظر می‌رسد دو نگاه کلان به فضای مجازی وجود دارد: نگاه اول که بالاخص در ابتدای رشد و تکوین فضای مجازی مسلط شده بود، آن را همچون ابزاری کنار سایر ابزارهای بشری تصویر می‌کرد که تنها طریقیت داشت. اما نگاه دوم، در نتیجه رشد تحولات خیره‌کننده فضای مجازی و سایه گستری آن در حوزه‌ها و شئون بشر در یک دهه اخیر آن را چون سکویی می‌داند که بسیار فراتر از شأن ابزاری حیات انسان‌ها را سامان جدیدی داده و ادعای تمدن نوینی را دارد. رویکردی که از قضا از چشمان بصیر رهبر انقلاب نیز دور نمانده و انتظاری تمدنی از فضای مجازی در ایران را مطالبه داشته‌اند.

در همین راستا گزارش‌های عصرفضای مجازی تلاش می‌کند تا فهم سازمان‌ها و دستگاه‌های مرتبط با حوزه فضای مجازی را ارتقاء بخشیده و آن‌ها را برای مواجهه فعال و خردمندانه با تحولات این عرصه مهیا سازد.

سیدابوالحسن فیروزآبادی

دبیر شورای عالی و رئیس مرکز ملی فضای مجازی

چکیده



چکیده

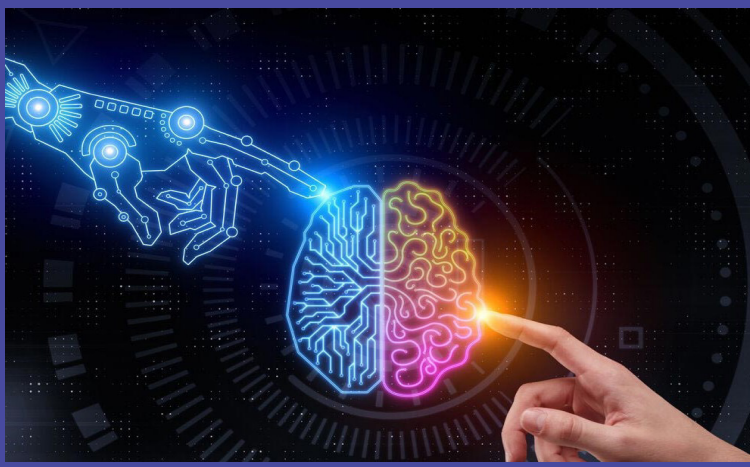
سیستم‌های هوشمند را می‌توان از یک نقطه نظر به دو بخش تقسیم کرد: سیستم‌های هوشمند خبره و سیستم‌های هوشمند مبتنی بر یادگیری عمیق. ویژگی سیستم‌های خبره این است که می‌توان فرایند اینکه سیستم چگونه به نتیجه خاصی رسیده است را بازبینی نمود. یعنی، این فرایند پوشیده نبوده و تبیین‌پذیر است. اما در مقابل، در سیستم‌های مبتنی بر یادگیری عمیق، فرایند تصمیم‌سازی توسط سیستم هوشمند حتی برای طراحان نیز معلوم نبوده و به یک بیان، فرایند تصمیم‌سازی در این سیستم‌ها مبهم است. این ابهام و تبیین‌ناپذیری به مسئله جعبه سیاه موسوم است. «هوش مصنوعی تبیین‌پذیر» واکنشی است به مسئله جعبه سیاه. هدف از این گزارش، معرفی هوش مصنوعی تبیین‌پذیر است. در این راستا، در ابتدا سیستم‌های خبره و سیستم‌های مبتنی بر یادگیری عمیق را معرفی می‌کنیم. پس از آن به مسئله جعبه سیاه اشاره خواهیم نمود. در بخش بعد، هوش مصنوعی تبیین‌پذیر را به عنوان واکنشی به مسئله جعبه سیاه بررسی خواهیم کرد. با تمایز میان این سؤال که «چگونه می‌توان ازنقطه‌نظر تکنولوژیکی، فنی و مهندسی به هوش مصنوعی تبیین‌پذیر دست یافت؟» و این سؤال که «چرا اساساً هوش مصنوعی تبیین‌پذیر امری ارزشمند است؟»،

سـؤال اول را فـرو گذاشـته و بـه بیـان پاسـخ‌های ارائـه شـده بـرای سـؤال دوم می‌پردازیـم. در ادبیـات علمـی پیرامـون هـوش مصنوعـی تبیین‌پذیـر چهـار پاسـخ بـه ایـن سـؤال وجـود دارد کـه بـه آنهـا اشـاره خواهیـم نمود. پـس از اشـاره بـه ایـن چهـار پاسـخ، ملاحظاتـی را پیرامـون ایـن چهـار دلیـل بـه پیـش می‌کشـیم. در بخـش بعـد، پیشـنهادهایی راهبـردی را بـرای اسـتفاده مدیـران و تصمیم‌سـازان عرصـه هـوش مصنوعـی جمهـوری اسـلامی ایـران ارائـه کـرده و در نهایـت گـزارش را بـا یـک جمع‌بنـدی بـه پایـان می‌بریـم.

واژگان کلیدی

یادگیری عمیق، مسئله جعبه سیاه، هوش مصنوعی تبیین‌پذیر

مقدمه



مقدمه

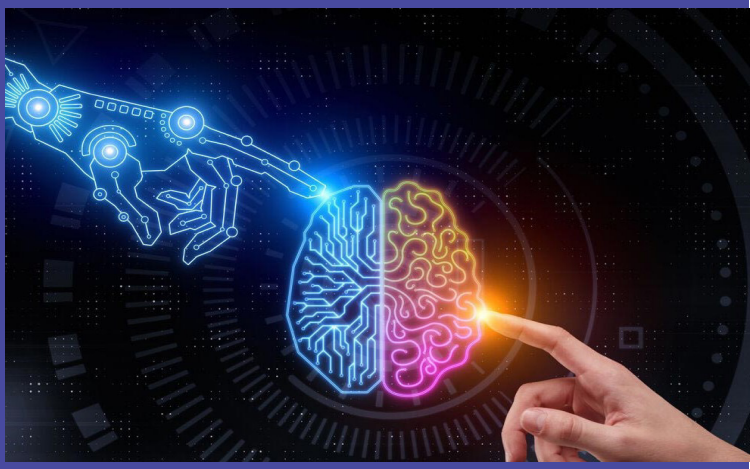
سیستم‌های هوشمند را در یک نگاه کلی می‌توان به دو بخش تقسیم نمود:

سیستم‌های هوشمند خبره[1] و سیستم‌های هوشمند مبتنی بر یادگیری عمیق[2]. از ویژگی‌های سیستم‌های خبره این است که می‌توان دریافت که سیستم هوشمند چگونه به یک خروجی دست یافته است. در مقابل، سیستم‌های مبتنی بر یادگیری عمیق به گونه‌ای هستند که حتی طراحان آنها نیز نمی‌توانند دریابند که سیستم هوشمند چگونه به خروجی مشخصی رسیده است. از آنجا که فرایند رسیدن به یک خروجی مشخص در سیستم‌های هوشمند مبتنی بر یادگیری عمیق، فرایندی نامعلوم است، به این سیستم‌های «جعبه سیاه»[3] نیز اطلاق می‌شود. «هوش مصنوعی تبیین‌پذیر» رویکردی است که قصد دارد سیستم هوشمندی را طراحی کند که برخلاف سیستم‌های هوشمند مبتنی بر یادگیری عمیق بتوان این را که سیستم هوشمند چگونه به خروجی مشخصی رسیده است را تبیین کرد. در این گزارش قصد داریم «هوش مصنوعی تبیین‌پذیر» را معرفی کنیم.

1. Expert systems
2. deep learning
3. Black box

ساختار این گزارش به این شکل است: در بخش دوم به معرفی سیستم‌های هوشمند خبره و سیستم‌های مبتنی بر یادگیری عمیق می‌پردازیم. در بخش سوم، ضمن معرفی مسئله جعبه سیاه، به معرفی رویکرد هوش مصنوعی تبیین‌پذیر برای حل این مسئله خواهیم پرداخت. در بخش چهارم، به پاسخ‌های مختلفی که به سؤال «چرا اساساً هوش مصنوعی تبیین‌پذیر امری ارزشمند است؟» اشاره خواهیم کرد. در بخش پنجم، ملاحظات خود را پیرامون مسئله هوش مصنوعی تبیین‌پذیر ارائه می‌کنیم. در نهایت و در بخش ششم، پیشنهاداهای راهبردی‌ای را برای استفاده مدیران و تصمیم‌سازان عرصه هوش مصنوعی جمهوری اسلامی بیان می‌کنیم. در نهایت، گزارش را با یک جمع بندی به پایان می‌بریم.

بخش اول



یادگیری عمیق و مسئله جعبه سیاه

همان‌طور که گفتیم سیستم‌های هوشمند، از یک نقطه نظر، به دو دسته تقسیم می‌شوند: سیستم‌های هوشمند خبره و سیستم‌های هوشمند مبتنی بر یادگیری عمیق. اجازه دهید درباره این دو گونه سیستم هوشمند توضیح دهیم. سیستم‌های خبره که می‌توان آنها را سیستم‌های «اگر – آنگاه» نامید سیستم‌های هستند که از چهار بخش اصلی تشکیل می‌شوند. بخش اصلی سیستم‌های خبره چیزی است که به «پایگاه دانش» موسوم است. پایگاه دانش در یک سیستم خبره مجموعه همه اطلاعاتی است که طراحان به سیستم خبره داده‌اند و اغلب به صورت گزاره‌های شرطی بیان می‌شوند. به بیان دقیق‌تر، در پایگاه دانش قوانین مختلفی به صورت جملات شرطی اگر- آنگاه تعبیه شده است. بخش دوم سیستم‌های خبره چیزی است که به «موتور استنتاج» معروف است. موتور استنتاج که شاید بتوان گفت مهم‌ترین بخش یک سیستم خبره است با ترکیب ورودی‌هایی که کاربر به سیستم می‌دهد با آنچه در پایگاه دانش در آن ذخیره شده است، خروجی‌هایی را تحویل می‌دهد. موتور استنتاج از دو طریق می‌تواند به نتایج دسترسی یابد. روش اول به روش مبتنی بر داده و روش دیگر روش مبتنی بر هدف نام دارد. در

روش اول سیستم با استفاده از داده‌ها و نیز گزاره‌های شرطی اگر - آنگاه، از مقدم این گزاره‌ها شروع کرده، به نتایج مناسب می‌رسد. در روش دوم، سیستم برعکس عمل می‌کند و برای نتیجه‌ای خاص به دنبال مقدماتی که می‌توانند به آن نتیجه منجر شوند می‌گردد. برای مثال، فرض کنید که سیستم خبره‌ای برای تشخیص و عیب‌یابی خوردها طراحی کرده‌ایم. هنگامی که ما به سیستم خبره این ورودی را می‌دهیم که «از گزوز این ماشین دود سیاه رنگی خارج می‌شود» آنگاه این سیستم خبره بر اساس این اطلاعات که «اگر از اگزوز ماشینی دود سیاه رنگی خارج شد، آنگاه واشرهای سیلندر و روغن دان بازبینی شود» که در پایگاه دانش آن وجود دارد نتیجه خواهد کرد که «واشرهای سیلندر و روغن دان بازبینی شود». بخش سوم یک سیستم خبره، «امکانات توضیح»[1] نام دارد که کاربر یا طراح را قادر می‌سازد که بتواند تمام مراحلی را که سیستم خبره از ورودی به سمت خروجی طی نموده است دنبال نماید. درواقع، این قابلیت این امکان را به کاربر یا طراح می‌دهد تا دریابد که سیستم خبره چگونه به نتیجه‌ای خاص رسیده است. بخش چهارم یک سیستم خبره، چیزی است که به «رابط کاربر» موسوم است. رابط کاربر این امکان را به کاربر می‌دهد تا بتواند داده‌های را به عنوان ورودی به سیستم خبره بدهد. معمولاً رابط کاربر شامل یک پردازنده زبانی است که می‌تواند زبان طبیعی را که قابل درک برای کاربر انسانی است به داده‌های قابل درک برای سیستم خبره تبدیل کند (Nath,2015). همین اندازه برای معرفی سیستم‌های خبره کافی است. آنچه درباره سیستم‌های خبره برای ما مهم است ویژگی و شاخصه دوم این سیستم‌ها، یعنی «امکانات توضیح» است. این ویژگی ما را قادر خواهد ساخت از جزئیات اینکه

1. Explanation Facilities

چگونه سیستم خبره به نتیجه‌ای خاص رسیده است آگاه شویم. حال، که دانستیم سیستم‌های هوشمند خبره چه هستند، اجازه دهید به بررسی سیستم‌های هوشمند مبتنی بر یادگیری عمیق بپردازیم.

یادگیری عمیق زیرمجموعه‌ای از یادگیری ماشین[1] است که نوعی فناوری هوش مصنوعی است. به طور خاص، یادگیری عمیق نوعی از یادگیری ماشین است که هدف آن این است که به کامپیوتر بیاموزد که چگونه از طریق مثال‌ها و تجارب بیاموزد. یادگیری ماشین مستلزم این است که داده‌های زیادی به سیستم داده شود تا سیستم بتواند از طریق آنها بیاموزد. باید به این نکته توجه کرد که یادگیری عمیق ازنقطه‌نظر نحوه

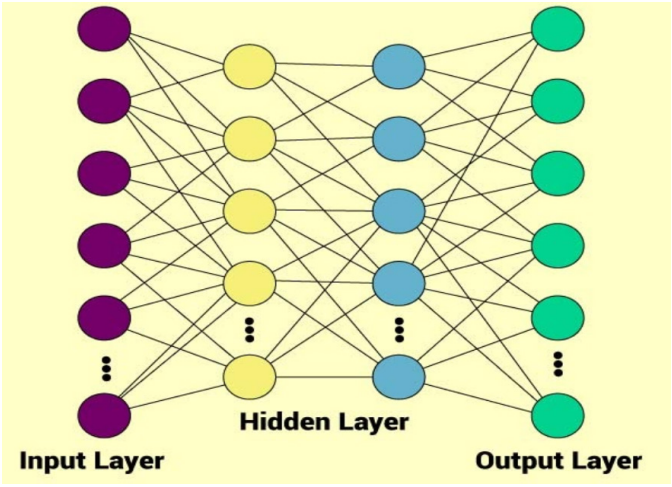
1. Machine learning

کار، با سایر انواع یادگیری ماشین متفاوت است. در یادگیری عمیق سیستم خود، به‌تنهایی، از داده‌ها یاد می‌گیرد؛ بنابراین ایده یادگیری عمیق این است که کامپیوتر به طور مستقل یاد بگیرد. (بدون اینکه انسان به او بگوید دنبال چه چیزی باشد).

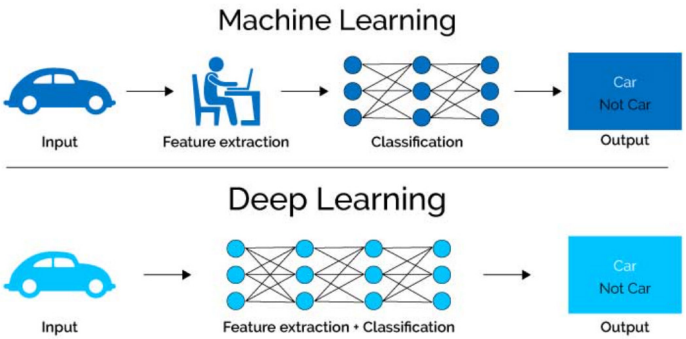
باید به این نکته توجه کرد که یادگیری عمیق از نحوه کارکرد مغز انسان الهام گرفته شده است. در یادگیری عمیق از چیزی به نام شبکه‌های عصبی مصنوعی[1] استفاده می‌شود. بنابراین، برای پاسخ به سؤال «یادگیری عمیق چیست؟»، در ابتدا باید بدانیم که شبکه‌های مصنوعی عصبی چه هستند.

شبکه عصبی مصنوعی یک سیستم رایانه‌ای است که از نحوه کار مغز انسان الهام گرفته شده است. این شبکه عصبی مصنوعی کمابیش مانند یک فیلتر عمل می‌کند. شبکه‌های عصبی مصنوعی دارای لایه‌هایی از «گره‌ها»[2] یا «نورون‌ها»[3] هستند. این گره‌ها هرکدام ورودی را تجزیه و تحلیل می‌کنند و سپس نتایج را با گره‌های موجود در لایه بعدی به اشتراک می‌گذارند. بنابراین، ما به شبکه عصبی مصنوعی داده‌هایی را برای پردازش می‌دهیم. این داده‌ها ورودی سیستم هستند. سپس، تمام گره‌های موجود در لایه اول ورودی را پردازش می‌کنند. سپس خروجی آنها به عنوان ورودی برای لایه بعدی از گره‌ها مورد استفاده قرار گرفته و دوباره پردازش می‌شوند. این فرایند برای هر لایه اتفاق می‌افتد. این لایه‌های میانی «لایه‌های پنهان»[4] نام دارند. آنچه در نهایت سیستم به عنوان خروجی به ما تحویل می‌دهد پاسخ شبکه عصبی مصنوعی به ورودی‌ای است که به آن داده بودیم (Grover,2019).

1. artificial neural networks (ANN)
2. nodes
3. Neurons
4. hidden layers



بنابراین، در پاسخ به این سؤال که «یادگیری عمیق چیست؟» باید بگوییم که «یادگیری عمیق یک شبکه عصبی مصنوعی بسیار عظیم است». به طور معمول، یک شبکه عصبی مصنوعی فقط چند لایه پنهان (حداکثر ۲ تا ۳ لایه) دارد که ورودی از طریق آن فیلتر می‌شود. اما در یادگیری عمیق تعداد این لایه‌های میانی بسیار زیاد است (مثلاً میلیون‌ها لایه میانی). واژه «عمیق» در عنوان «یادگیری عمیق» دقیقاً به‌خاطر وجود همین لایه‌های میانی بسیار زیاد است. به طور خلاصه، یادگیری عمیق یک شبکه عصبی مصنوعی عظیم است که به ماشین‌ها اجازه می‌دهد از طریق مثال یاد بگیرند. هرچه نمونه‌ها و مثال‌های بیشتری به سیستم‌های مبتنی بر یادگیری عمیق ارائه شود، سیستم عملکرد بهتری خواهد داشت.



اجـازه دهیـد بـرای درک بهتـر عملکـرد سیسـتم‌های مبتنـی بـر یادگیری عمیـق مثالـی بزنیـم. فـرض کنیـد کـودک نوپایـی اولیـن کلمـه‌ای فراگرفته اسـت «گربـه» اسـت و می‌توانـد به‌خوبـی مصادیق گربـه را از سـایر حیوانات تمایـز دهـد. اجـازه دهیـد ببینیـم ایـن کـودک چگونه توانسـته اسـت به این توانایـی دسـت یابـد. این کودک در هر بار به شـیئی اشـاره کـرده و واژه گربه را بـرای نامیـدن آن بـکار می‌بـرد. والدیـن او نیز در هـر بار به او ایـن بازخورد را می‌دهنـد کـه آیـا شـیئی کـه کـودک بـه آن اشـاره کرده اسـت یـک گربه اسـت یـا نـه. کودک به مـرور زمان و از طریـق تجارب مختلـف و بازخوردهای والدینـش در نهایـت بـه ایـن توانایـی خواهد رسـید کـه بتواند گربـه را از غیر آن متمایـز کنـد. سیسـتم‌های مبتنی بـر یادگیری عمیق نیز به همین شـکل هسـتند. آن‌هـا نیـز از طریـق تجربـه و از طریـق بازخـوردی کـه بـه آن داده می‌شـود بـه مـرور زمـان می‌تواننـد بیاموزنـد کـه خروجـی خواسـته شـده از آن‌هـا را بـه کاربر تحویـل دهند.

اجـازه دهیـد به برخـی مثال‌هـا و کاربردهای یادگیـری عمیق اشـاره کنیم. یکـی از کاربردهـای الگوریتم‌هـای یادگیـری عمیـق برای تشـخیص و دسـته

بنـدی اشـیاء اسـت. برای مثـال، می‌تـوان سیسـتمی طراحی کرد کـه چهره افـراد را تشـخیص دهـد و یا بتواند اشـیاء خاصـی را که مطلوب ماسـت، مثلاً ماشـین پراید را، از سـایر اشیاء دیگر تشـخیص دهد. یکی دیگر از کاربردهای موفـق یادگیـری عمیـق در پـردازش زبان طبیعی اسـت. برای مثـال، مترجم گـوگل کـه می‌توانـد گفتارهـا و متون‌هایـی را از زبـان مبـدأ بـه زبـان مقصد ترجمـه کنـد نمونـه‌ای اسـت از کاربردهـای یادگیـری عمیـق. یکـی دیگر از کاربردهـای یادگیـری عمیـق در خودروهـای خـودران اسـت. ایـن خودروهـا از طریـق یادگیـری عمیـق می‌آموزنـد کـه شـرایط محیطـی را، هماننـد یک راننـده انسـانی، تشـخیص دهند. برای مثـال، یاد خواهند گرفـت هنگامی که کودکـی بـه وسـط خیابان پریـد آن را به عنوان یک انسـان تشـخیص دهد و سـپس ترمـز نماید. به نظر می‌رسـد همیـن اندازه برای آشـنایی بـا یادگیری عمیـق کافـی اسـت (Grover,2019). حـال اجـازه دهیـد «مسـئله جعبه سـیاه»[1] را معرفـی کنیم.

هـر سیسـتم محاسـباتی یـک ورودی را دریافت کرده، بر روی آن پردازشـی می‌کنـد و در نهایـت خروجـی‌ای را تحویـل می‌دهـد. امـا یـک سیسـتم محاسـباتی یـا یـک برنامـه هوشـمند «جعبـه سـیاه» نامیـده می‌شـود بـه شـرطی کـه بـه مـا ایـن امـکان را بدهـد کـه ورودی و خروجـی را مشـاهده کنیـم، ولـی هیـچ فراینـدی و عملکـرد بیـن ورودی و خروجـی را مشـاهده نکنیـم. بنابرایـن جعبـه سـیاه هـوش مصنوعـی به ایـن واقعیـت اشـاره دارد کـه بـا بیشـتر ابزارهـای مبتنـی بـر هـوش مصنوعـی، مـا نمی‌دانیم کـه آنها چگونـه کارهایـی را کـه انجـام می‌دهند را انجـام می‌دهنـد. به‌عبارت‌دیگر، ما از از داده‌هایـی کـه سیسـتم هوشـمند با آنها شـروع می‌کند (یعنـی ورودی) آگاهیـم (بـه عنوان مثـال، عکس‌های حیوانـات). همچنین، ما از پاسـخی که

1. black box problem

سیستم هوشمند تولید می‌کند (یعنی خروجی) نیز آگاهیم (مثلاً، جدا کردن تصاویر گربه از تصاویر سایر حیوانات). اما به‌خاطر مسئله جعبه سیاه، ما نمی‌دانیم که چگونه توانسته است از ورودی‌هایی که به آن داده‌ایم آن خروجی‌های خاص را به ما تحویل دهد. اما این عدم آگاهی هنگامی که سیستم هوشمند خروجی‌های نادرست، ناخواسته و مشکل‌دار را به ما تحویل بدهد، خود را به صورت یک مسئله نمایان می‌کند.

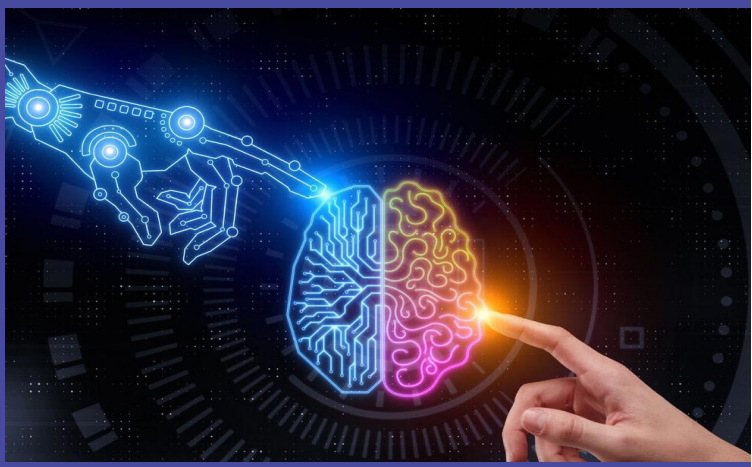
اما می‌توان پرسید «چه چیزی سبب بروز مسئله جعبه سیاه می‌شود؟» در پاسخ به این سؤال باید گفت که رایج‌ترین ابزارهایی که از مسئله جعبه سیاه رنج می‌برند، ابزارهایی هستند که از شبکه‌های عصبی مصنوعی یا همان یادگیری عمیق که در بالا درباره آن سخن گفتیم استفاده می‌کنند. همان‌طور که در بالا اشاره کردیم، شبکه‌های عصبی مصنوعی دارای بی‌شمار لایه‌های مخفی از گره‌ها (نورون‌ها) هستند. این گره‌ها هر یک ورودی داده شده را پردازش کرده و خروجی خود را به لایه بعدی گره‌ها منتقل می‌کنند. همان‌طور که بیان کردیم، یادگیری عمیق، یک شبکه عصبی مصنوعی عظیم، با بی‌شمار لایه میانی پنهان است که خود می‌تواند از طریق مثال‌ها و تجربه‌ها بیاموزد. اما نکته اینجاست که ما نمی‌توانیم دریابیم که گره‌ها چه چیزهایی را یاد گرفته‌اند. در واقع، ما از خروجی بین لایه‌ها اطلاعی نداشته و فقط نتیجه را دریافت می‌کنیم. ازاین‌رو، ما نمی‌توانیم نحوه تجزیه و تحلیل داده‌ها توسط گره‌ها را دریابیم و در واقع (بر اساس تعریف بالا) با یک جعبه سیاه روبرو خواهیم بود (Zednik,2019).

از آنجا که سیستم‌های هوشمند مبتنی بر یادگیری عمیق برگرفته و الهام گرفته از کارکرد مغز انسان هستند، مسئله جعبه سیاه درباره مغز

انسـان نیـز مطـرح می‌شـود. چرا که مـا دقیقـاً نمی‌دانیـم که مغز مـا چگونه کار می‌کنـد و چگونـه بـه نتیجـه یا یک خروجـی خاص می‌رسـد. در واقع، به هیچ وجه مشـخص نیسـت که در سـطحی پایـه‌ای، چگونـه تصمیمات خود را (مثـلاً تصمیـم به نوشـیدن یک لیـوان آب)، اخـذ می‌کنیم. در واقـع، برای اینکـه بتوانیـم چنیـن تصمیم سـاده‌ای را اتخاذ کنیـم تعداد زیـادی نورون و سـیناپس در مغزمـان فعالیـت می‌کنند.

حال که دانسـتیم مسـئله جعبه سـیاه چگونه مسـئله‌ای اسـت اجازه دهید توضیـح دهیـم چـرا جعبـه سـیاه یـک مسـئله و مشـکل قلمـداد می‌شـود و اینکـه چـه رویکردی برای حل این مسـئله وجـود دارد. در بخـش بعد به این مطلـب می‌پردازیم.

بخش دوم



هوش مصنوعی تبیین‌پذیر و اکنشی به مسئله جعبه سیاه

در بخش قبل بیان کردیم که مسئله جعبه سیاه این مسئله است که حتی طراحان سیستم هوشمند نیز نمی‌دانند که چگونه سیستم هوشمند به خروجی خاصی رسیده است. اما چرا این مسئله باید یک مشکل قلمداد شود. در ادامه به این سؤال پاسخ می‌دهیم.

گاهی اوقات سیستم‌های هوشمند مبتنی بر یادگیری عمیق به نتایجی دست پیدا می‌کنند که مخالف با اهدافی است که طراحان آن برای آن در نظر گرفته‌اند. در واقع، گاهی اوقات این سیستم‌ها از داده‌ها و تجربیات طوری می‌آموزند که نتایجی برخلاف آنچه برای آن طراحی شده‌اند را ارائه می‌دهند. برای مثال، یک سیستم مبتنی بر یادگیری عمیق که برای تشخیص تصاویر اسب طراحی شده بود، در نهایت یاد گرفته بود که تصاویری را که بر روی آن برچسب کپی‌رایت وجود دارد را از تصاویری که این برچسب را ندارد متمایز کند (چراکه بالاتفاق بر روی تصاویر اسبی که به آن نمایش داده می‌شود برچسب کپی‌رایت نیز وجود داشت).

یکی از اشتباهات و خطاهای سیستم‌های هوشمند، سیستم هوش مصنوعی ساخت شرکت IBM است که به پزشکان در حوزه تشخیص

بیماری یاری می‌رساند. در سال ۲۰۱۳ شرکت IBM اولین محصول تجاری خود را برای درمان سرطان تولید کرد. اما این در حالی بود که برخی از پزشکان بیان کردند که این سیستم هوشمند توصیه‌های اشتباهی را برای درمان سرطان ارائه می‌دهد که می‌تواند سبب مشکلات جدی و حتی نتایج مرگبار شود (Peng,2018).

مثال دیگر، کشته‌شدن یک عابر پیاده در اثر تصادف با یک اتومبیل خودران اوبر[1] بود که در ۲۸ مارس سال ۲۰۱۸ در آریزونای آمریکا اتفاق افتاد. بعد از بررسی شرکت اوبر اعلام کرد که ماشین خودران بعد از اینکه تشخیص داده بود که یک عابر پیاده در مقابلش هست ولی تصمیم گرفته بود که هیچ اقدامی را انجام ندهد (Peng,2018).

می‌توان حالت‌های فرضی مختلف دیگری را نیز برای اشتباهات اتومبیل‌های خودران در نظر گرفت. برای مثال، سیستم هوشمند اتومبیل‌های خودران می‌تواند کودکان کوتاه قامت را به‌عنوان یک عابر انسانی شناسایی نکند یا درصورتی‌که امر دائر شود میان تصادف با یک ماشین دیگر و تصادف با یک عابر پیاده این‌طور تشخیص دهد که تصادف با عابر پیاده بهتر است از تصادف با یک ماشین دیگر. حتی سیستم هوشمند به دلایلی نامعلوم ممکن است این را آموخته باشد که در برابر عابرانی که با خود یک جعبه حمل می‌کنند توقف نکند.

1. Uber self-driving

یکی دیگر از خطاهای سیستم‌های هوشمند مبتنی بر یادگیری عمیق، سیستمی است که شرکت آمازون برای استخدام افراد استفاده کرده است. شرکت آمازون در خلال سال‌های ۲۰۱۴ تا ۲۰۱۷ از سامانه هوشمندی برای بررسی بی‌شمار رزومه ارسالی و ارائه پیشنهاداتی استفاده می‌کرده است. اما این شرکت بعدها متوجه شد که سیستم هوشمند به دلایل نامشخصی متقاضیان مرد را نسبت به متقاضیان زن در اولویت قرار می‌دهد (Peng,2018). همین مقدار برای مثال‌ها کافی است.

مثال‌های بالا، همگی حاکی از یک مشکل اخلاقی است که ناشی از نادیده گرفتن مسئله جعبه سیاه است. به نظر می‌رسد درصورتی‌که سیستم‌های هوشمند بتوانند نحو نادرستی بیاموزند و درنتیجه خروجی‌هایی را ارائه کنند که مطلوب ما نبوده است آنگاه دیگر نمی‌توان به این سیستم‌ها اعتماد نمود. برای مثال، در حوزه پزشکی اگر سیستم هوشمند طوری آموخته باشد که توصیه‌های

اشـتباهی بـرای درمـان ارائـه دهـد، آنـگاه ایـن مسـئله جـان بیمـاران را بـه خطـر خواهـد انداخـت. همچنیـن، اگـر سیسـتم هوشـمند در بررسـی رزومه‌هـای متقاضیـان بـرای کار طـوری آموختـه باشـد کـه نسـبت بـه مـردان سـوگیری داشـته باشـد، آنـگاه فرصت‌هـای شـغلی بـرای زنـان لایـق از دسـت خواهـد رفـت و ایـن نتیجـه‌ای جـز تبعیـض جنسـیت و نابرابـری نخواهـد داشـت. نیـز درصورتی‌کـه ماشـین‌های خـودران بـه اشـتباه بیاموزنـد کـه در برابـر عابرانـی خـاص (مثـلاً عابرانـی کـه بـا خـود یـک جعبـه حمـل می‌کننـد) توقـف نکننـد، آنـگاه جـان افـراد بـه خطـر خواهـد افتـاد.

حـال می‌تـوان گفـت کـه مسـئله جعبـه سـیاه مسـتلزم یـک مسـئله مهم‌تـر، یعنـی مسـئله «عـدم اعتمـاد بـه هـوش مصنوعـی» اسـت. در واقـع، درصورتی‌کـه مـا ندانیـم کـه آیـا سیسـتم هوشـمندی کـه از آن اسـتفاده می‌کنیـم قابل‌اعتمـاد اسـت یـا خیـر، تمایلـی بـرای اسـتفاده از آن نخواهیـم داشـت.

در واکنـش بـه مسـئله جعبـه سـیاه و بـه دنبـال آن، مسـئله عـدم اعتماد، در چنـد سـال اخیـر جریانـی بـه راه افتـاده اسـت کـه بـه «هـوش مصنوعی تبیین‌پذیـر»[1] موسـوم اسـت. هـوش مصنوعـی تبیین‌پذیـر هـوش مصنوعـی‌ای اسـت کـه می‌تـوان فراینـد تصمیم‌سـازی‌اش را درک نمـود. در واقـع، در هـوش مصنوعـی تبیین‌پذیـر انسـان‌ها می‌تواننـد دریابنـد کـه یـک سیسـتم هوشـمند چگونـه بـه تصمیـم خاصـی رسـیده اسـت. بنابرایـن، بـا هـوش مصنوعـی تبیین‌پذیـر می‌تـوان اشـکالات و خطاهـای سیسـتم‌های هوشـمند را دریافـت و از ایـن خـلال از بـروز مشـکلات اخلاقـی، ناعدالتـی و حتـی فاجعه‌بـار جلوگیـری نمـود.

1. Explainable Artificial Intelligence (XAI)

در ادبیات علمی در باب هوش مصنوعی تبیین‌پذیر معمولاً دو سؤال از یکدیگر متمایز می‌شوند (Colaner,2021). پرسش اول این است که :

«چگونه می‌توان ازنقطه‌نظر تکنولوژیکی، فنی و مهندسی به هوش مصنوعی تبیین‌پذیر دست‌یافت؟»[1]

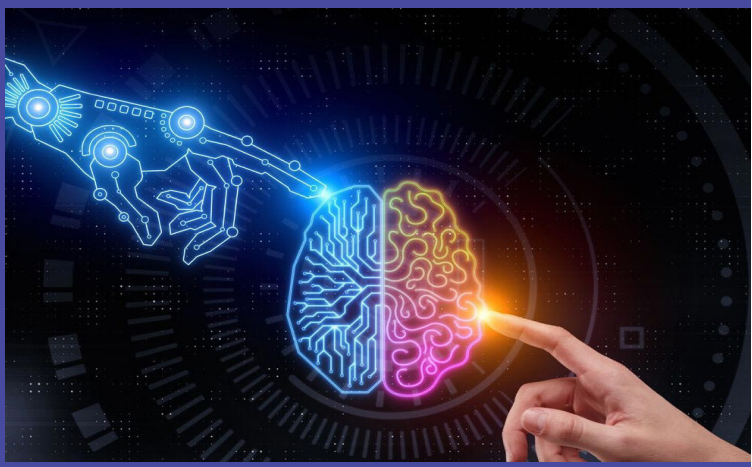
پرسش دومی که می‌توان در حوزه هوش مصنوعی تبیین‌پذیر پرسید این پرسش است که:

«چرا اساساً هوش مصنوعی تبیین‌پذیر امری ارزشمند است؟»

ما در این گزارش با پرسش اول کاری نخواهیم داشت. پرسش اول پرسشی است مربوط به مسائل فنی در حوزه هوش مصنوعی. در عوض، در بخش‌های بعد به پرسش دوم خواهیم پرداخت.

۱. برای نمونه ببینید: (Vellido et al. 2012; Kim et al. 2016; Tamagnini et al. 2017; Guidotti et al. 2018; Ayesha et al. 2020)

بخش سوم



ارزشمندی هوش مصنوعی تبیین‌پذیر

در ادبیات علمی شکل‌گرفته پیرامون هوش مصنوعی تبیین‌پذیر چهار دلیل عمده برای پاسخ به این پرسش که «چرا اساساً هوش مصنوعی تبیین‌پذیر امری ارزشمند است؟» ارائه شده است. در ادامه این چهار پاسخ را بررسی خواهیم کرد. این چهار پاسخ در دو دسته تقسیم‌بندی می‌شوند:

پاسخ‌هایی که به ارزش ابزاری هوش مصنوعی تبیین‌پذیر اشاره دارند و پاسخ‌هایی که به ارزش ذاتی هوش مصنوعی تبیین‌پذیر اشاره دارند. منظور از ارزش ابزاری این است که هوش مصنوعی تبیین‌پذیر به این دلیل خوب است که وسیله و ابزاری است برای دست‌یافتن به چیزهای دیگری که برای ما مطلوب و ارزشمند هستند. اما در مقابل، پاسخ‌هایی که سعی دارند ارزش هوش مصنوعی تبیین‌پذیر را بر اساس یک ارزش ذاتی توضیح دهند تلاش می‌کنند بیان کنند هوش مصنوعی تبیین‌پذیر به‌خاطر ویژگی‌های خوبی که دارد ارزشمند است. در این رویکرد بیان نمی‌شود که هوش مصنوعی تبیین‌پذیر خوب است چون نتایج خوبی برای ما به دنبال دارد. بلکه در عوض، خود سیستم تبیین‌پذیر به‌خودی‌خود

ویژگی‌های ارزشمندی دارد، فارغ از اینکه آیا نتایج حاصل از آن برای ما مطلوب باشد یا نه. در ادامه این پاسخ‌ها را بررسی خواهیم کرد.

رویکرد اول (ارزش ابزاری)

پاسخ اول به این پرسش که «چرا اساساً هوش مصنوعی تبیین‌پذیر امری ارزشمند است؟» (که در بخش‌های قبل مختصراً به آن اشاره شد) این است که امر مطلوبی را ذکر کنیم که بتوان از طریق هوش مصنوعی تبیین‌پذیر، به آن دست‌یافت. در واقع، در این پاسخ سعی می‌شود به نحو غیرمستقیم توضیح داده شود که چرا ارائه تبیین امر ارزشمندی است. به بیان دیگر، بر اساس این رویکرد هوش مصنوعی تبیین‌پذیر ابزاری است برای دستیابی به آن فایده خوب (Colaner,2021). یک مثال خوب برای این پاسخ، صورت‌بندی است که آژانس پروژه‌های تحقیقاتی پیشرفته دفاعی ایالات متحده آمریکا[1] که هدف آن تلاش برای فهم تکنولوژی‌های نو ظهور در حوزه امنیت ملی است، ارائه داده است. یکی از اهداف این آژانس تولید سیستم‌های هوشمندی است که برای انسان‌ها توضیح پذیر باشند. از نظر آن‌ها اگر انسان‌ها موفق شوند هوش مصنوعی تبیین‌پذیر را تولید کنند، آنگاه شرایطی برای کاربران محقق خواهد شد که بتوانند به نحو درستی به هوش مصنوعی اعتماد کرده و به نحو مؤثری آن را مدیریت کنند[2] (DARPA,2016). همان‌طور که پیداست «اعتماد» و «مدیریت» اهداف مطلوبی هستند که می‌توان از طریق تبیین به آنها دست یافت. به بیان دیگر، می‌توان گفت دو امر ارزشمند به نام‌های «اعتماد» و «مدیریت»

1. Defense Advanced Research Projects Agency (DARPA)
2. appropriately trust, and effectively manage

وجـود دارد کـه می‌تـوان از طریـق ایجـاد هـوش مصنوعـی تبیین‌پذیـر بـه آن‌هـا دسـت یافـت؛ امـوری کـه نمی‌تـوان از طریـق سیسـتم‌های هوشـمند فعلـی (کـه تبیین‌پذیـر نیسـتند) بـه آن‌هـا دسـت پیـدا کـرد. بنابرایـن، می‌تـوان گفـت کـه تبیین‌پذیـری در ایـن رویکـرد ارزشـی ابـزاری دارد. یعنـی، تبیین‌پذیـری ارزشـمند اسـت زیـرا سـبب می‌شـود بـه دو ارزش دیگـر، یعنـی اعتمـاد بـه و مدیریـتِ هـوش مصنوعـی دسـتیابیم.

ایـن رویکـرد، یعنـی ارزش ابـزاری تبیین‌پذیـری (بـرای دسـتیابی بـه دو ارزش اعتمـاد و مدیریـت) همچنیـن در «آیین‌نامـه عمومـی حفاظـت از داده‌هـا»[1] نیـز موردتوجـه قـرار داده شـده اسـت. یکـی از پایه‌هـای ایـن چارچـوب مقرراتـی، تأکیـد بـر داشـتن حـق تبییـن و توضیـح[2] اسـت. بـرای مثـال، اگـر فـردی کـه بـرای اسـتخدام در شـرکت آمـازون اقـدام کـرده و رزومـه وی توسـط بررسـی اولیه سیسـتم هوشـمند رد شـده اسـت، ایـن حـق را دارد کـه بپرسـد «چـرا درخواسـت مـن رد شـده اسـت؟». در واقـع، از نظـر ایـن چارچـوب مقرراتـی، هنگامی‌کـه یـک الگوریتم هوشـمند تصمیمـی را اتخـاذ می‌کنـد، ایـن حـق بـرای افـراد (مثـلاً کسـانی کـه از آن تصمیـم متأثـر می‌شـوند) وجـود دارد کـه بداننـد چـرا سیسـتم هوشـمند بـه آن تصمیـم رسـیده اسـت (Wachter et al,2017).

همان‌طـور کـه پیداسـت در رویکـرد نخسـت تبیین‌پذیـری امـری ارزشـمند اسـت زیـرا وسـیله و ابـزاری اسـت بـرای دسـت‌یافتن بـه ارزش‌هـای دیگـری همچـون، «اعتمـاد»، «قابلیـت مدیریـت» و «حـق تبییـن». بنابرایـن، می‌تـوان اسـتدلالی را ماننـد زیـر بـرای رویکـرد نخسـت ترتیـب داد:

۱) اگـر چیـزی ماننـد الـف ابـزار و راهـی بـرای رسـیدن بـه ارزش دیگـری

1. General Data Protection Regulation (GDPR)
2. right to an explanation

مانند ب باشد، آنگاه الف نیز خود دارای ارزش خواهد بود.

۲) هوش مصنوعی تبیین‌پذیر ابزاری است برای رسیدن به اموری همچون «اعتماد»، «قابلیت مدیریت» و «حق تبیین».

۳) «اعتماد»، «قابلیت مدیریت» و «حق تبیین» اموری ارزشمند هستند.

۴) بنابراین، هوش مصنوعی تبیین‌پذیر ارزشمند است.

رویکرد دوم (ارزش ذاتی)

برخلاف رویکرد اول در پاسخ به این سؤال که «چرا اساساً هوش مصنوعی تبیین‌پذیر امری ارزشمند است؟»، رویکرد دیگری وجود دارد که بیان می‌دارد که اساساً هوش مصنوعی تبیین‌پذیر ذاتاً امری ارزشمند است. رویکرد دوم، با این سؤال آغاز می‌کند که آیا اساساً می‌توان تبیین‌ها را اموری دانست که ذاتاً دارای ارزش هستند؟ این امکان که تبیین و توضیح امری ذاتاً ارزشمند باشد امری معنادار خواهد بود درصورتی‌که بپذیریم اگر انسان بخشی از یک فرایند غیر تبیین‌پذیر باشد، آنگاه این امر مستقیماً بر روی شرایط انسانی وی تأثیر خواهد گذاشت (با صرف‌نظر از اینکه آیا خروجی فرایند تبیین‌پذیر مطلوب وی باشد یا نه). به بیان دیگر، ازنقطه‌نظر این رویکرد تبیین‌پذیری ذاتاً ارزشمند است فارغ از اینکه تبیین‌پذیری چه نتایجی را به دنبال دارد. این استدلال گه گاه با نام «استدلال فردیت»[1] شناخته می‌شود زیرا مبتنی بر سؤالاتی در باب فردیت از قبیل خودمختاری فردی[2] و کرامت انسانی[3] است.

در ادبیات بحث سه نسخه از استدلال فردیت ارائه شده است که در

1. personhood argument
2. individual autonomy
3. human dignity

ادامـه هـر یـک را بیـان می‌کنیـم.

نسخه اول (استدلال از طریق مشارکت)

هنگامی‌کـه تصمیمـات و اقدامـات یـک سیسـتم هوشـمند غیـر قابـل تبییـن و توضیـح باشـد، آنـگاه انسـان‌ها بـه عنـوان موجوداتـی کـه قـرار اسـت از تصمیمـات و اقدامـات هـوش مصنوعـی متأثـر شـوند دیگـر نمی‌تواننـد در فراینـد تصمیم‌سـازی مشـارکتی داشـته باشـند. بـرای مثـال، همان‌طـور کـه در بخش‌هـای قبـل بیـان کردیـم هـوش مصنوعـی در حـوزه پزشـکی تصمیماتـی را بـرای بیمـار می‌گیـرد کـه نـه بیمـار و نـه پزشـک معالـج وی هیـچ مشـارکتی در اتخـاذ آن تصمیـم نداشـته‌اند. بنابرایـن، می‌تـوان ادعـا کـرد هرچقـدر کـه یـک سیسـتم هوشـمند بیشـتر غیـر قابـل تبییـن و توضیـح باشـد، مـا انسـان‌ها کمتـر می‌توانیـم در اتخـاذ یـک تصمیـم مشـارکت داشـته باشـیم. بـرای تقریـب بـه ذهـن، اجـازه دهیـد مثالـی را از حـوزه عدالـت اجتماعـی ارائـه کنیـم. محققیـن حـوزه عدالـت اجتماعـی بیـان می‌کننـد هنگامی‌کـه گروهـی از انسـان‌ها از فراینـد تصمیم‌سـازی بـرای همـه انسـان‌ها کنـار گذاشـته می‌شـوند، آنـگاه ناعدالتـی اجتماعـی رخداده اسـت. بـرای مثـال، یانـگ[1] ایـن مثـال را بیـان می‌کنـد کـه کارگـران یـک کارخانـه از ایـن اطلاعیـه کـه رئیـس یـک کارخانـه بـزرگ در شـهر، قصـد دارد کارخانـه‌اش را تعطیل کند بسـیار خشـمگین هسـتند. حـرف آن‌هـا ایـن اسـت کـه چـرا تصمیم‌سـازان ایـن کارخانـه بـدون اطـلاع و در جریـان گذاشـتن، مشـورت و هشـدار قبلـی بـه کارگـران کارخانـه، چنیـن تصمیمـی را اتخـاذ کرده‌انـد (Young, ۱۹۹۰). بایـد بـه ایـن نکتـه دقـت کـرد کـه نکتـه ایـن مثـال نتایـج ناعادلانـه

1. Young

تعطیلی کارخانه برای کارگران نیست. نکته این مثال این است که کارگران از این‌که چرا آن‌ها در فرایند تصمیم‌سازی برای مسئله به این مهمی دخالت داده نشده‌اند عصبانی هستند.

در واقع، می‌توان فرض کرد که مدیران کارخانه به کارگران این وعده را بدهند که بعد از تعطیلی کارخانه به آن‌ها مبلغ زیادی پول بلاعوض خواهند داد و حق بیمه آن‌ها را نیز پرداخت خواهند کرد. اگرچه این شرایط و این نتیجه به لحاظ اقتصادی عادلانه به نظر می‌رسد، ولی نکته این است که کارگران کماکان می‌توانند از تصمیم اتخاذ شده ناراضی باشند، تنها به این دلیل که آن‌ها در جریان این تصمیم مهم قرار داده نشده و مشارکتی برای اتخاذ این تصمیم نداشته‌اند.

در واقع، ممکن است افراد بتوانند با نتایج یک تصمیم کنار بیایند ولی صرفاً به این دلیل که آن‌ها در جریان اتخاذ تصمیم قرار نگرفته‌اند احساس ناعدالتی نمایند. همچنین، این اصل که عدالت مستلزم گوش فرادادن به سخن ذی‌نفعان در خلال یک فرایند تصمیم‌سازی است، ازنقطه‌نظر اصول مدیریت حقوق سیاسی نیز بحث شده است (Zhu et al,2019). باتوجه‌به این توضیحات می‌توان نسخه اول از استدلال فردیت را مانند زیر بیان کرد (Colaner,2019):

۱) در مواقعی که نتایج و خروجی‌های یک فرایند تصمیم‌سازی بر روی فرد یا جامعه تأثیر می‌گذارد، کرامت انسانی مستلزم فرصت مشارکت معنادار در آن فرایند است.

۲) الگوریتم‌های مبتنی بر یادگیری عمیق اساساً غیرقابل‌درک و

توضیح هستند.

۳) بنابراین، نمی‌توان در فرایند تصمیم‌سازی الگوریتم‌های مبتنی بر یادگیری عمیق مشارکت معنادار داشت.

۴) بنابراین، استفاده از الگوریتم‌های مبتنی بر یادگیری عمیق در تصمیم‌سازی برای فرد یا جامعه، مستلزم نقض کرامت انسانی خواهد شد.

همان‌طور که در استدلال بالا پیداست، ازآنجاکه سیستم‌های هوشمند مبتنی بر یادگیری عمیق تبیین‌ناپذیر هستند، بنابراین، انسان‌ها نمی‌توانند در فرایند تصمیم‌سازی آن‌ها مشارکتی داشته باشند و ازآنجاکه مشارکت در یک فرایند تصمیم‌سازی که نتایج آن متوجه فرد و جامعه است، شرط ضروری برای کرامت انسانی است، پس می‌توان نتیجه گرفت استفاده از یادگیری عمیق مستلزم نقض کرامت انسانی است. در استدلال بالا باید به این نکته دقت شود که برای ما مهم نیست که نتایج حاصل از تصمیمات مبتنی بر یادگیری عمیق درست است یا خیر، اخلاقی است یا خیر، قانونی است یا خیر، برایمان سودمند است یا خیر. حتی اگر نتایج حاصل از یادگیری عمیق درست، اخلاقی، قانونی و سودمند نیز باشد، صرفاً به این دلیل که ما انسان‌ها در فرایند تصمیم‌سازی مشارکتی نداشته‌ایم استفاده از آن منافی امری است که ارزش ذاتی دارد (یعنی کرامت انسانی).

می‌توان استدلال بالا را به نحو ایجابی نیز بیان کرد:

۱) استفاده از هوش مصنوعی تبیین‌پذیر به این معنی است که بتوان در فرایند تصمیم‌سازی مشارکت داشت.

۲) مشارکت در فرایند تصمیم‌سازی ذاتاً ارزشمند است
۳) بنابراین، استفاده از هوش مصنوعی تبیین‌پذیر ذاتاً ارزشمند است.

همان‌طور که مشخص است این استدلال قصد ندارد ارزشمندی استفاده از هوش مصنوعی تبیین‌پذیر را به‌خاطر ارزش دیگری همچون اعتماد و مدیریت پذیری بیان کند. این استدلال بیان می‌کند صرف هوش مصنوعی تبیین‌پذیر، فارغ از اینکه نتیجه تصمیمات آن چه باشد، امری ارزشمند است، تنها به این دلیل که حق مشارکت در تصمیمات برای انسان‌ها محفوظ است.

نسخه دوم (استدلال از طریق علم)

نسخه دوم استدلال فردیت، برخلاف نسخه اول که بر «مشارکت» در امر تصمیم‌سازی تأکید می‌کرد، بر «علم» تأکید دارد (Colaner,2021). این استدلال می‌توان مانند زیر بیان کرد:

۱) کرامت انسانی شامل این است که من بدانم چرا آنچه برای من اتفاق افتاده است، اتفاق افتاده است
۲) اما در صورت استفاده از سیستم‌های مبتنی بر یادگیری عمیق (تبیین‌ناپذیر) علم به چرایی اینکه چرا چیزی برای من اتفاق افتاده است غیرممکن است
۳) بنابراین، در صورت استفاده از سیستم‌های هوشمند مبتنی بر یادگیری عمیق، کرامت انسانی نقض شده است.

همان‌طور که پیداست نسخه دوم، بر این تأکید دارد که اصل اینکه فرد بداند برای او چه اتفاقی افتاده است و چرا سیستم هوشمند تصمیمی خاصی را برای وی اتخاذ کرده است امری ارزشمند و مهم

است. زیرا در این صورت است که کرامت انسانی انسان‌ها حفظ شده است. همان‌طور که پیداست، در استدلال بالا برای ما مهم نیست که نتایج حاصل از تصمیمات مبتنی بر یادگیری عمیق درست است یا خیر، اخلاقی است یا خیر، قانونی است یا خیر، برایمان سودمند است یا خیر. حتی اگر نتایج حاصل از یادگیری عمیق درست، اخلاقی، قانونی و سودمند نیز باشد، صرفاً به این دلیل که ما انسان‌ها به اینکه چرا سیستم هوشمند تصمیم خاصی را گرفته است علم نداریم، استفاده از آن دارای اشکال می‌شود. در مقابل، ازآنجاکه استفاده از هوش مصنوعی تبیین‌پذیر امکان علم را به ما می‌دهد همین امر فی ذاته، سبب ارزشمندی استفاده از هوش مصنوعی تبیین‌پذیر می‌شود.

نسخه سوم (استدلال از طریق فعلیت انسان)[1]

نسخه سوم از استدلال فردیت که نیثن کولانر[2] آن را ارائه می‌کند، بر این امر تأکید دارد که در صورتی انسان‌ها می‌توانند به خود و جامعه خود را فعلیت بخشند که برای خود و جامعه تصمیم اتخاذ نمایند. به بیان دیگر، ایده اصلی این نسخه از استدلال فردیت این است که اتخاذ تصمیم برای تحقق و فعلیت انسانیت انسان امری ضروری است (Colaner,2021). نکته مهمی که باید به آن توجه کرد است که این نسخه از استدلال با نسخه اول ارتباط دارد اما دقیقاً همان نیست.

نکته این است که نسخه اول بر روی مسئله «قدرت در مشارکت» تأکید داشت اما این نسخه بر مسئله «فعلیت انسان» تأکید دارد.

1. actualization
2. Nathan Colaner

ایـن نسـخه از اسـتدلال ریشـه در ایده‌هـای ارسـطو در بـاب ماهیـت و طبیعـت انسـان دارد. از نظـر ارسـطو مـا انسـان‌ها بیـش از موجوداتـی همچـون زنبورهـا، اجتماعـی هسـتیم.

در واقـع از نظـر ارسـطو انسـانیت مـا انسـان‌ها در گـرو زندگـی اجتماعـی بـا یکدیگـر اسـت. بنابرایـن، ایـن خصیصـه سـبب خواهـد شـد کـه مـا انسـان‌ها بـه ایـن سـمت متمایـل شـویم کـه چگونـه جامعـه خـود را سـامان دهیـم تـا بتوانیـم در سـایه عدالـت در کنـار یکدیگـر زندگـی کنیم. ارسـطو در ایـن بـاب بیـان مـی‌دارد:

> فـردی کـه قـادر بـه زندگـی در جامعـه نیسـت، یـا فـردی کـه بـه دلیـل خودکفـا بـودن هیـچ نیـازی نـدارد، یـا بایـد یـک جانـور باشـد یـا یـک خـدا. چنیـن فـردی هیـچ بخشـی از یـک جامعـه نیسـت [...] عدالـت پیونـد انسـان‌ها در یـک جامعـه اسـت (Aristotle,1984).

بـا توجـه بـه ایـن نکتـه کـه لازمـه اجتماعـی بـودن انسـان تصمیم‌سـازی بـرای خـود و جامعـه اسـت و از طـرف دیگـر فعلیـت گوهـر انسـانی نیـز مبتنـی بـر تصمیم‌سـازی وی بـرای خـود و جامعـه اسـت، می‌تـوان نسـخه سـوم از اسـتدلال فردیـت را، به‌صـورت تقریبـی، ماننـد زیـر بیـان کـرد:

۱) فعلیـت کامـل انسـانیت انسـان مسـتلزم ایـن اسـت کـه وی بـرای زندگـی خـود و جامعـه‌اش تصمیم‌سـازی کنـد (نـه سیسـتم هوشـمند).

۲) امـا هـوش مصنوعـی تبیین‌ناپذیـر ایـن فرصـت را بـه انسـان نمی‌دهـد تـا بتوانـد بـرای خـود و جامعـه تصمیـم بگیـرد.

۳) بنابرایـن، اسـتفاده از هـوش مصنوعـی تبیین‌ناپذیـر سـبب خواهـد

شـد انسـانیت انسـان فعلیـت نیابـد.

۴) اما هوش مصنوعی تبیین‌پذیر چنین مشکلی را پدید نمی‌آورد.

۵) بنابراین، هوش مصنوعی تبیین‌پذیر ارزشمند است.

بنـا بـر اسـتدلال فـوق، هـوش مصنوعـی تبیین‌پذیـر، برخـلاف هـوش مصنوعـی تبیین‌ناپذیـر، بـه ایـن دلیـل کـه سـبب می‌شـود انسـان بتوانـد از طریـق دخالـت در فراینـد تصمیم‌سـازی بـرای خـود و جامعـه حقیقـت خـود را فعلیـت بخشـد، امـری ارزشـمند اسـت. همان‌طـور کـه پیداسـت، ایـن اسـتدلال بیـان نمی‌کنـد هـوش مصنوعـی تبیین‌پذیـر بـه دلیـل اینکـه سـبب می‌شـود نتایـج مطلوبـی بـه دسـت آیـد امـری ارزشـمند اسـت. بلکـه ایـن اسـتدلال صرفـاً نظـر از نتایـج عملـی هـوش مصنوعـی تبیین‌پذیـر (از قبیـل کاهـش خطاهـا، اعتمـاد و ...) خـودِ تبیین‌پذیـری را امـری دارای ارزش می‌دانـد.

بـه بیـان دیگـر، در اسـتدلال بـالا بـرای مـا مهـم نیسـت کـه نتایـج حاصـل از تصمیمـات مبتنـی بـر یادگیـری عمیـق درسـت اسـت یـا خیـر، اخلاقـی اسـت یـا خیـر، قانونـی اسـت یـا خیـر، برایمـان سـودمند اسـت یـا خیـر. حتـی اگـر نتایـج حاصـل از یادگیـری عمیـق درسـت، اخلاقـی، قانونـی و سـودمند نیـز باشـد، صرفـاً بـه ایـن دلیـل کـه مـا انسـان‌ها از طریـق اعمـال تصمیـم نتوانسـته‌ایم گـروه خـود را فعلیـت دهیـم، اسـتفاده از آن دارای اشـکال می‌شـود.

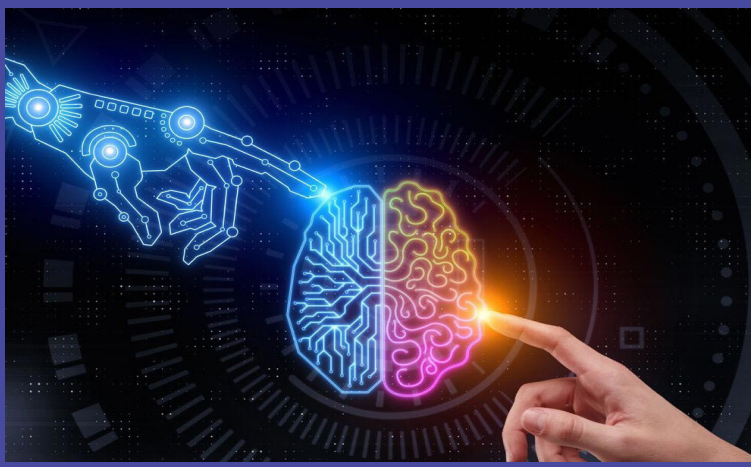
ممکـن اسـت تصـور شـود توصیـه اسـتدلال بـالا ایـن اسـت کـه اساسـاً بایـد از اسـتفاده از سیسـتم‌های هوشـمند بـرای تصمیم‌سـازی دسـت کشـید. در پاسـخ، نیشـن کولانـر بیـان می‌کنـد کـه ایـن امـر، لازمـه

استدلال بالا نیست. ما می‌توانیم آنچه ارسطو بیان کرده است را جدی تلقی کنیم ولی درعین‌حال بر این عقیده باشیم که جایی برای تصمیم‌سازی هوش مصنوعی نیز وجود دارد. در این صورت می‌توان انواع مختلف تصمیم‌سازی هوش مصنوعی را از یکدیگر متمایز کرد:

- چه تصمیماتی را می‌توان به‌صورت کامل به هوش مصنوعی سپرد؟ (برای مثال، توصیه‌هایی برای خرید)
- چه تصمیماتی می‌باید حاصل تعامل میان انسان و هوش مصنوعی باشد؟ (برای مثال، تصمیم برای استخدام افراد)
- چه تصمیماتی باید توسط هوش مصنوعی اتخاذ شود اما با این امکان که انسان بتواند آن را نپذیرد؟ (برای مثال، درخواست دادن برای کارت اعتباری یا وام)
- چه تصمیماتی باید منحصراً در حیطه اختیارات انسان باشد نه هوش مصنوعی؟ (برای مثال، انتخاب یک رئیس‌جمهور).

نیشن کولانر بیان می‌کند که در این حالت، بحث بر سر پاسخ به سؤالات فوق خواهد بود. از نظر وی تصمیماتی از نوع دوم و سوم، با فرض هوش مصنوعی غیر تبیین‌پذیر نمی‌تواند محقق شود. ازاین‌رو، استدلال بالا می‌تواند کماکان در مورد تصمیمات از نوع دوم و سوم جاری باشد و بتوان نتیجه گرفت که هوش مصنوعی تبیین‌ناپذیر مانعی خواهد بود برای تصمیمات از نوع دوم و سوم.

بخش چهارم



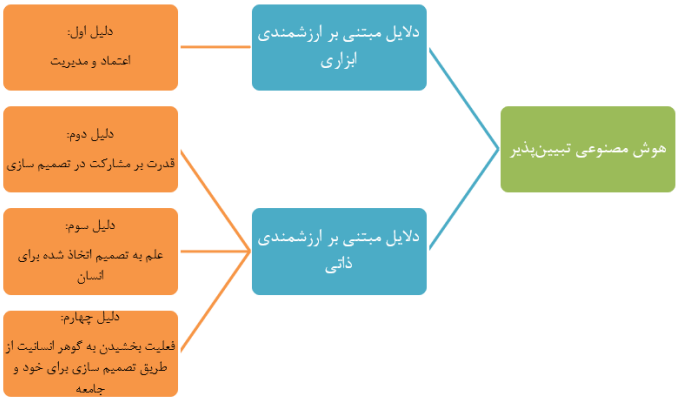
ملاحظاتی پیرامون هوش مصنوعی تبیین‌پذیر

همان‌طور که در بخش‌های قبل ملاحظه کردیم، اساساً چهار دلیل عمده برای ارزشمندی هوش مصنوعی تبیین‌پذیر ارائه شده است که خود به دو بخش دلایل ابزاری و دلایل ذاتی تقسیم می‌شوند. بر اساس دلایل ابزاری، هوش مصنوعی تبیین‌پذیر ارزشمند است زیرا وسیله‌ای است برای دستیابی به ارزش‌های دیگری همچون اعتماد. در واقع، هوش مصنوعی تبیین‌پذیر به ما کمک می‌کند که بتوانیم به نحو معقولی به هوش مصنوعی اعتماد کرده و از آن استفاده کنیم و در صورت بروز اشکال و خطر، بتوانیم دریابیم که اشکال کار کجاست. بنابراین، بر اساس این نگاه استفاده از هوش مصنوعی تبیین‌پذیر ارزشمند است اما نه به‌خودی‌خود، بلکه به این سبب که منافعی را برای ما به دنبال می‌آورد. بر اساس دلایل ذاتی، هوش مصنوعی تبیین‌پذیر، صرف‌نظر از اینکه چه منافع و فوایدی را برای ما به دنبال می‌آورد امری ارزشمند است. این دلایل خود از سه طریق این مطلب را اثبات می‌کنند.

بنابراین، دلایل دوم تا چهارم به‌قرار زیرند: دلیل دوم اینکه، هوش مصنوعی تبیین‌پذیر ارزشمند است زیرا قابلیت مشارکت

در تصمیم‌سازی را میسر می‌کند (حتی در مواردی که نتایج به‌دست‌آمده از یک سیستم هوشمند همگی اخلاقی، قانونی و منطبق بر منافع و تمایلات ما باشد)؛ دلیل سوم اینکه، هوش مصنوعی تبیین‌پذیر ارزشمند است زیرا سبب خواهد شد که ما به اینکه چرا فلان تصمیم درباره‌مان گرفته شده است علم پیدا کنیم (حتی در مواردی که نتایج به‌دست‌آمده از یک سیستم هوشمند همگی اخلاقی، قانونی و منطبق بر منافع و تمایلات ما باشد) و دلیل چهارم اینکه، هوش مصنوعی تبیین‌پذیر ارزشمند است زیرا سبب خواهد شد که ما انسان‌ها، با اتخاذ تصمیم برای خود و جامعه، خویشتن واقعی خود را فعلیت بخشیم. اما به نظر می‌رسد در میان دلایل ذکر شده برای ارزشمندی هوش مصنوعی تنها دلیل اول قابل‌قبول بوده و سه دلیل دیگر قابل دفاع نباشد.

در ادامه در این باره توضیح می‌دهیم.



همان‌طور که دیدیم، دلایل دوم تا چهارم، برخلاف دلیل اول، دلایل ذاتی هستند و نه دلایل ابزاری. اجازه دهید دلیل دوم را بررسی کنیم. بر اساس دلیل دوم هوش مصنوعی تبیین‌پذیر ارزشمند است زیرا قابلیت مشارکت در تصمیم‌سازی را میسر می‌کند.

همان‌طور که بیان کردیم این دلیل با نتایجی که یک هوش مصنوعی برای ما به بار می‌آورد کاری ندارد، نتایج هر چه می‌خواهد باشد (به سود یا به ضرر ما)، آنچه ارزشمند است این است که ما انسان‌ها بتوانیم در فرایند تصمیم‌سازی مشارکت کنیم. در واقع، نفس مشارکت امری مهم و ارزشمند است. سپس این دلیل بیان می‌کرد ازآنجاکه در یادگیری عمیق چنین چیزی میسر نبوده ولی در هوش مصنوعی تبیین‌پذیر میسر است، بنابراین هوش مصنوعی تبیین‌پذیر امری ارزشمند است. اما به نظر می‌رسد در اینجا اشکالی وجود داشته باشد. اشکال این است که همین عدم مشارکت در تصمیم‌سازی، در حوزه‌های انسانی نیز به چشم می‌خورد.

در واقع، در بسیاری از مواقع، کارکنان یک شرکت، کارگران یک کارخانه، شهروندان یک شهر و مردمان یک کشور هیچ مشارکتی در تصمیم‌سازی‌های خرد و کلان مسئولین ندارند. حال می‌توان گفت، اگر استفاده از هوش مصنوعی غیر تبیین‌پذیر صرفاً به این خاطر که نمی‌توان در امر تصمیم‌سازی مشارکت نمود مذموم و دارای اشکال است، آنگاه باید این را نیز پذیرفت که اصل تصمیم‌سازی مسئولین برای زیر دستان خود (در سطوح مختلف) بدون مشارکت آن‌ها نیز امری مذموم و دارای اشکال است.

اما به نظر می‌رسد که پذیرفتن مورد اخیر این لازمه را خواهد

داشت که همه افراد در همه تصمیمات مشارکت کنند که به لحاظ عملی امری است غیرممکن.

البته ممکن است کسی بیان کند در نظام‌های دموکراتیک این مردم هستند که مسئولین را انتخاب کرده و تصمیم‌سازی در امور کشور را به او تفویض می‌کنند. ازاین‌رو، به نحو غیرمستقیم مردم در تصمیم‌سازی‌ها مشارکت کرده‌اند.

اما به نظر می‌رسد پاسخ بالا چندان موفقیت‌آمیز نباشد. زیرا فرض کنید که درباره مسئله‌ای برگزیدگان و نمایندگان مردم تصمیم الف را اتخاذ می‌کنند اما اکثریت مردمی که به آن نمایندگان رأی داده‌اند با اتخاذ تصمیم الف مخالف‌اند. به نظر می‌رسد در این موارد می‌توان به نحو معقولی مدعی شد که مردم (اگرچه خود نمایندگان را انتخاب کرده‌اند) در تصمیم‌سازی مشارکت نکرده‌اند؛ چراکه اکثریت آن‌ها نظری خلاف مسئولین خود دارند.

حال اجازه دهید به دلیل سوم بپردازیم. بر اساس دلیل سوم، اصل علم به اینکه چگونه هوش مصنوعی تصمیمی را برای ما اتخاذ کرده است امری ارزشمند است؛ امری که هوش مصنوعی جعبه سیاه به دلیل نامعلوم بودن فرایند تصمیم‌سازی، تأمین نمی‌کند. اما به نظر می‌رسد در اینجا نیز مشکلی وجود داشته باشد.

فارغ از اینکه هوش مصنوعی تبیین‌پذیر به لحاظ تکنولوژیکی و فنی و مهندسی قابل دستیابی باشد یا نه، مسئله این است که اگر تصمیم‌سازی هوش مصنوعی غیر تبیین‌پذیر امری است مشکل‌دار (آن‌گونه که دلیل سوم بیان می‌دارد)، آنگاه تصمیم‌سازی انسان‌ها نیز امری است مشکل‌دار. اجازه دهید توضیح دهیم.

ما در زندگی خود تصمیمات مختلفی را از تصمیمات ساده تا تصمیمات پیچیده اتخاذ می‌کنیم. برای مثال، فرض کنید تصمیم می‌گیریم تا به سمت یخچال رفته و لیوان آبی را بنوشیم. باتوجه‌به اینکه مغز انسان یک مجموعه عظیم نورونی است (همان‌گونه که سیستم‌های یادگیری عمیق ملهم از مغز انسانی هستند)، این سؤال مطرح است که آیا ما واقعاً تبیینی از این مسئله داریم که چه فرایند نورونی در مغزمان رخداده است که در نتیجه آن ما تصمیم به نوشیدن آب گرفته‌ایم؟ به نظر می‌رسد جواب این سؤال خیر باشد. ما هیچ اطلاعی از اینکه چه فرایندهایی در مغزمان رخ می‌دهد نداریم. در واقع می‌توان گفت مغز ما همانند یک جعبه سیاه است که فرایند تصمیم‌سازی در آن برای ما نامعلوم است.

بنابراین، به نظر می‌رسد در تصمیمات انسانی نیز تبیین‌ناپذیری وجود دارد. حال می‌توان از مدافعان دلیل سوم پرسید چرا استفاده از تصمیمات سیستم هوشمند تبیین‌ناپذیر امری مشکل‌دار بوده ولی استفاده از تصمیمات انسان‌ها (که در آن نیز تبیین‌ناپذیری وجود دارد) بلا اشکال است. در واقع، اگر بپذیریم تصمیم‌سازی توسط سیستم هوشمند جعبه سیاه مشکل‌دار است، آنگاه باید این را نیز بپذیریم که تصمیم‌سازی مغز انسان‌ها نیز مشکل‌دار است. اما به نظر می‌رسد مطلب اخیر نپذیرفتنی نباشد.

حال اجازه دهید به بررسی دلیل چهارم بپردازیم. بر اساس دلیل چهارم هوش مصنوعی تبیین‌پذیر ارزشمند است زیرا سبب خواهد شد که ما انسان‌ها، با اتخاذ تصمیم برای خود و جامعه، خویشتن واقعی خود را فعلیت بخشیم. اما به نظر می‌رسد در اینجا

نیـز مشـکلی وجـود داشـته باشـد. ایـن دلیـل بـر ایـن تأکیـد دارد کـه اصـل تصمیم‌سـازی سـبب بالفعـل شـدن حقیقـت و گوهـر انسـان خواهـد شـد. سـپس بیـان می‌کنـد کـه ایـن امـر در هـوش مصنوعـی تبیین‌ناپذیـر محقـق نیسـت. امـا بـه نظـر می‌رسـد بـه خـلاف ادعـای دلیـل چهـارم، ایـن امـر در هـوش مصنوعـی تبیین‌ناپذیـر محقـق اسـت. درسـت اسـت کـه مـا نمی‌توانیـم در فراینـد تصمیم‌سـازی یـک سیسـتم هوشـمند تبیین‌ناپذیـر دخالتـی کنیـم ولـی در اصـل اسـتفاده یـا عـدم اسـتفاده از سیسـتم هوشـمند تصمیمـی را اتخـاذ کرده‌ایـم. در واقـع، مـا پیـش از آنکـه سیسـتم هوشـمند تصمیمـی را بـرای مـا یـا جامعـه مـا اتخـاذ کنـد، تصمیـم گرفته‌ایـم کـه از سیسـتم هوشـمند تبیین‌ناپذیـر اسـتفاده کنیـم. بنابرایـن، بـه نظـر می‌رسـد کـه حتـی در سیسـتم هوشـمند تبیین‌ناپذیـر ملاحظـه مدنظـر دلیـل چهـارم (یعنـی اعمـال تصمیم‌سـازی توسـط انسـان به‌منظـور فعلیـت بخشـیدن بـه خـود و جامعـه) تأمیـن خواهـد شـد.

از نظـر نگارنـده، شـاید بهتریـن دلیـل بـرای اسـتفاده از سیسـتم‌های هوشـمند تبیین‌پذیـر، همـان دلیـل اول باشـد. در واقـع، شـاید بهتریـن دلیـل ایـن باشـد کـه اسـتفاده از هـوش مصنوعـی تبیین‌پذیـر ارزشـمند اسـت زیـرا سـبب خواهـد شـد بتوانیـم اشـکالات موجـود در سیسـتم‌های هوشـمند را کاهـش داده و بتوانیـم اساسـاً بـرای اسـتفاده از سیسـتم‌های هوشـمند بـه آن‌هـا اعتمـاد کنیـم.

همچنیـن، حتـی ممکـن اسـت بتـوان دلایـل را ارائـه کـرد کـه نشـان دهنـد اسـتفاده از سیسـتم‌های هوشـمند تبیین‌ناپذیـر، علی‌رغـم وجـود مسـئله جعبـه سـیاه، امـری بـی مشـکل اسـت.

اجـازه دهیـد یکـی از ایـن دلایـل را ارائـه کنیـم. ایـده اصلـی ایـن دلیـل

ایـن اسـت کـه همان‌طـور کـه تصمیمـات اشـتباه و خطـا در میـان انسـان‌ها امـری پذیرفتـه شـده اسـت، اشـتباهات و خطاهـای سیسـتم‌های هوشـمند نیـز می‌بایـد امـری پذیرفتـه شـده باشـد.

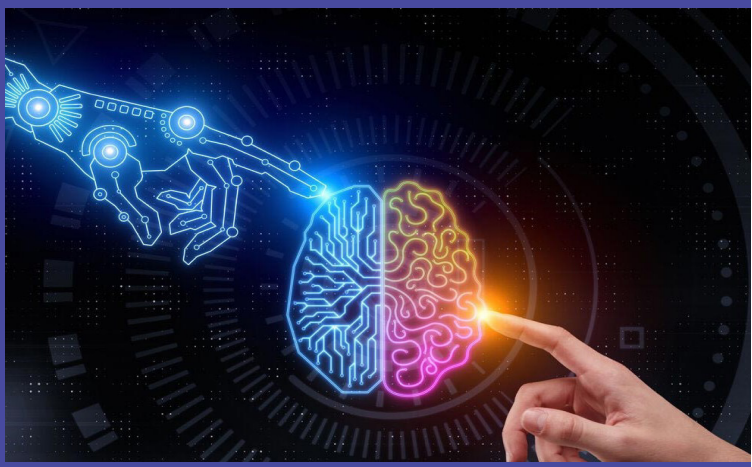
توضیـح آنکـه، انسـان‌ها به‌صـورت فـردی و یـا به‌صـورت جمعـی بعضـاً تصمیمـات اشـتباهی اتخـاذ می‌کننـد. امـا مسـئله ایـن اسـت کـه ایـن تصمیمـات اشـتباه و خطـا بعضـاً اتفـاق می‌افتـد. حـال اگـر یـک سیسـتم هوشـمند گاهـی اوقـات دچـار تصمیمـات اشـتباه و خطـا شـود، بـه نحـو معقولـی می‌تـوان آن را هماننـد خطاهـای انسـانی در نظـر گرفـت. بـه بیـان دیگـر، درصورتی‌کـه تصمیمـات اشـتباه یـک سیسـتم هوشـمند بسـیار کمتـر از تصمیمـات درسـت آن باشـد، آنـگاه می‌تـوان بـه نحـو معقولـی بـر روی آن اعتمـاد کـرده و از آن اسـتفاده کـرد. ایـن ایـده برخاسـته از نظریـه اتکاگرایـی[1] در معرفت‌شناسـی اسـت.

بـر اسـاس نظریـه اتکاگرایـی در بـاب توجیـه، اگـر باورهـای تولیـد شـده توسـط یـک قـوای شـناختی در اکثـر مواقـع درسـت باشـند، آنـگاه آن قـوه شـناختی و به‌تبـع باورهـای تولیـد شـده توسـط آن معتبـر بـوده و می‌تـوان بـه آن اتـکا کـرد (Lemos,2007,p.85). بـه نظـر می‌رسـد می‌تـوان همیـن ایـده را نیـز در مـورد سیسـتم‌های هوشـمند تبیین‌ناپذیـر نیـز بـه کار بـرد. اگـر یـک سیسـتم هوشـمند در اکثـر مواقـع (مثـلاً ۹۰ درصـد از مواقـع) تصمیمـات درسـتی اتخـاذ کنـد و تنهـا در مواقـع اندکـی (مثـلاً ۱۰ درصـد مواقـع) تصمیمـات اشـتباهی اتخـاذ کنـد، آنـگاه ایـن سیسـتم قابل‌اتـکا بـوده و می‌تـوان بـه آن اعتمـاد کـرد. ولـی اگـر براثـر آزمایشـات مشـخص شـد کـه یـک سیسـتم هوشـمند در اکثـر مواقـع تصمیمـات اشـتباهی اتخـاذ می‌کنـد سیسـتم مذکـور قابل‌اتـکا نبـوده و نمی‌تـوان از

1. Relaibilism

آن استفاده کرد. بنابراین، به نظر می‌رسد حتی بتوان استدلال کرد که استفاده از سیستم‌های غیر تبیین‌پذیر با وجود مسئله جعبه سیاه امری بی مشکل است.

بخش پنجم



پیشنهادهای راهبردی

بـر اسـاس آنچـه در بـالا بیـان شـد، نگارنـده پیشـنهادهای راهبـردی زیـر را بـرای اسـتفاده مدیـران و تصمیم‌سـازان عرصـه هـوش مصنوعـی جمهـوری اسـلامی ایـران ارائـه می‌کنـد:

همان‌طـور کـه دیدیـم، مسـئله هـوش مصنوعـی تبیین‌پذیـر مسـئله نوپایـی در عرصـه هـوش مصنوعـی اسـت کـه هنـوز ابهامـات زیـادی در آن وجـود دارد. پیشـنهاد می‌شـود بودجه‌هـای تحقیقاتـی‌ای بـرای کاوش ابعـاد مختلـف مسـئله اختصـاص داده شـود تـا بتـوان بـرای سـؤالات زیـر پاسـخ‌های قابـل قبولـی یافـت:

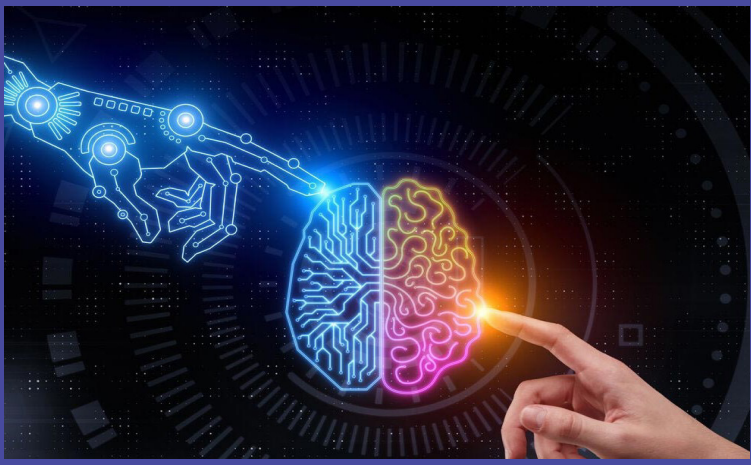
«آیـا دسـتیابی بـه هـوش مصنوعـی تبیین‌پذیـر ضـرورت دارد؟»، «آیـا دلایـل معقولـی بـرای حرکـت بـه سـمت هـوش مصنوعـی تبیین‌پذیـر وجـود دارد؟» و «آیـا دسـتیابی بـه هـوش مصنوعـی تبیین‌پذیـر بـه لحـاظ تکنولوژیکـی و فنـی و مهندسـی اساسـاً ممکـن اسـت؟».

در گام بعـد و پـس از مشخص‌شـدن پاسـخ سـؤالات فـوق نیـاز اسـت تـا بـر اسـاس پاسـخ‌های به‌دسـت‌آمده در گام قبلـی، تصمیمـات درسـتی اتخـاذ شـود.

بـرای مثـال، اگـر بـه ایـن نتیجـه رسـیدیم کـه دسـتیابی بـه هـوش

مصنوعی تبیین‌پذیر ضرورت داشته و امری ممکن است، آنگاه باید سیاست‌های راهبردی برای دستیابی به آن اتخاذ شود.

جمع‌بندی



جمع‌بندی

در این گزارش در ابتدا توضیحی درباره انواع مختلف سیستم‌های هوشمند ارائه شد: سیستم‌های هوشمند خبره و سیستم‌های هوشمند مبتنی بر یادگیری عمیق. ویژگی سیستم‌های خبره این است که می‌توان فرایند اینکه سیستم چگونه به نتیجه خاصی رسیده است را بازبینی نمود. یعنی، این فرایند پوشیده نبوده و تبیین‌پذیر است. اما در مقابل در سیستم‌های مبتنی بر یادگیری عمیق، فرایند تصمیم‌سازی توسط سیستم هوشمند حتی برای طراحان نیز معلوم نبوده و به یک بیان، فرایند تصمیم‌سازی در این سیستم‌ها مبهم است.

پس از بیان این مطلب، مسئله جعبه سیاه را شرح دادیم. مسئله جعبه سیاه به همین مسئله عدم تبیین‌پذیری و نامشخص بودن فرایند رسیدن به یک خروجی در سیستم‌های مبتنی بر یادگیری عمیق اشاره دارد. پس از توضیحاتی در باب مسئله جعبه سیاه، بیان کردیم که یکی از رویکردها برای در پاسخ به مسئله جعبه سیاه، رویکردی است که به «هوش مصنوعی تبیین‌پذیر» معروف است. پس از تمایز میان این سؤال که «چگونه می‌توان ازنقطه‌نظر تکنولوژیکی، فنی و مهندسی به هوش مصنوعی تبیین‌پذیر دست‌یافت؟» و این سؤال که «چرا اساساً هوش مصنوعی تبیین‌پذیر امری ارزشمند

است؟»، سؤال اول را فروگذاشته و به بیان پاسخ‌های ارائه شده برای سؤال دوم پرداختیم. در ادبیات بحث چهار پاسخ به این مسئله وجود دارد. این چهار دلیل خود به دو بخش دلایل ابزاری و دلایل ذاتی تقسیم می‌شوند.

بر اساس دلایل ابزاری، هوش مصنوعی تبیین‌پذیر ارزشمند است زیرا وسیله و ابزاری است برای دستیابی به ارزش‌های دیگری همچون اعتماد. در واقع، هوش مصنوعی تبیین‌پذیر به ما کمک می‌کند که بتوانیم به نحو معقولی به هوش مصنوعی اعتماد کرده و از آن استفاده کنیم و در صورت بروز اشکال و خطر، بتوانیم دریابیم که اشکال کار کجاست. بنابراین، بر اساس این نگاه استفاده از هوش مصنوعی تبیین‌پذیر ارزشمند است اما نه به‌خودی‌خود، بلکه به این سبب که منافعی را (مثلاً «اعتماد» به آن را)، برای ما به دنبال می‌آورد. بر اساس دلایل ذاتی، هوش مصنوعی تبیین‌پذیر، صرف‌نظر از اینکه چه منافع و فوایدی را برای ما به دنبال می‌آورد امری ارزشمند است.

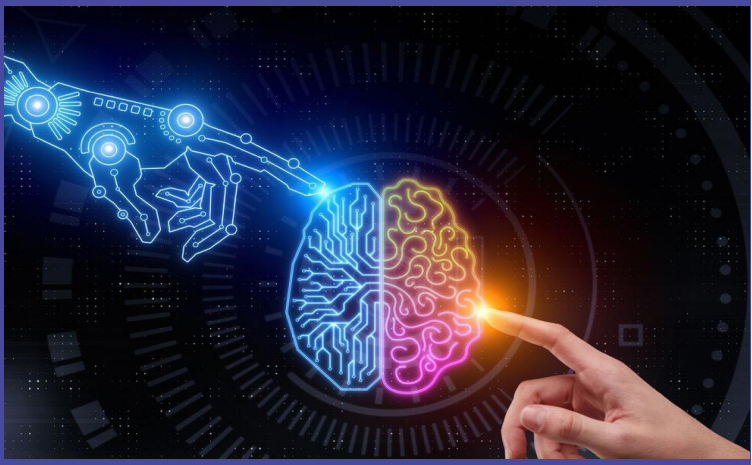
این دلایل خود از سه طریق این مطلب را اثبات می‌کنند. بنابراین، دلایل دوم تا چهارم به‌قرار زیرند: دلیل دوم اینکه، هوش مصنوعی تبیین‌پذیر ارزشمند است زیرا قابلیت مشارکت در تصمیم‌سازی را میسر می‌کند (حتی در مواردی که نتایج به‌دست‌آمده از یک سیستم هوشمند همگی اخلاقی، قانونی و منطبق بر منافع و تمایلات ما باشد)؛ دلیل سوم اینکه، هوش مصنوعی تبیین‌پذیر ارزشمند است زیرا سبب خواهد شد که ما به اینکه چرا فلان تصمیم درباره‌مان گرفته شده است علم پیدا کنیم (حتی در مواردی که نتایج به‌دست‌آمده از یک سیستم هوشمند همگی اخلاقی، قانونی و منطبق بر منافع و تمایلات ما باشد) و دلیل چهارم اینکه، هوش مصنوعی تبیین‌پذیر ارزشمند است زیرا سبب خواهد شد که ما

انسان‌ها، با اتخاذ تصمیم برای خود و جامعه، خویشتن واقعی خود را فعلیت بخشیم. در بخش بعدی به برخی ملاحظات در باب هوش مصنوعی تبیین‌پذیر اشاره کردیم.

یکی از ملاحظات بر این نکته تأکید داشت که دلایل دوم تا چهارم قابل چالش بوده و شاید بهترین دلیل برای دستیابی و استفاده از هوش مصنوعی تبیین‌پذیر، همان دلیل اول باشد. ملاحظه دیگر، این بود که حتی می‌توان استدلال کرد که استفاده از هوش مصنوعی غیر تبیین‌پذیر امری بلا مشکل است. زیرا، همان‌طور که این مسئله که انسان‌ها در تصمیم‌سازی دچار خطا می‌شوند مسئله‌ای پذیرفته شده است، خطا و اشتباهات هوش مصنوعی قابل‌اعتماد نیز می‌توان پذیرفت.

در پایان پیشنهاداتی را برای استفاده مدیران و تصمیم‌سازان عرصه هوش مصنوعی جمهوری اسلامی ایران ارائه کردیم. یکی از این پیشنهادهای راهبردی این بود که باتوجه‌به نوظهور بودن هوش مصنوعی تبیین‌پذیر و مشخص نبودن ابعاد مختلف آن، راه‌اندازی پروژه‌های تحقیق برای شناسایی ابعاد مختلف مسئله امری ضروری است.

منابع



منابع

[1] Aristotle J (1984) The complete works of Aristotle: the revised oxford translation. Edited by Jonathan Barnes. Princeton University Press, Princeton (Volume Two. Bollington Series LXXI.2)D

[2] Ayesha S, Hanif MK, Talib R (2020) Overview and comparative study of dimensionality reduction techniques for high dimensional data. Inf Fusion 58–59:44

[3] Buckner, Cameron and James Garson. (2019). «Connectionism», The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2019/entries/connectionism/>.

[4] Colaner, N. (2021). Is explainable artificial intelligence intrinsically valuable?. AI & SOCIETY, 8-1.

[5] DARPA (2016) Explainable Artificial Intelligence (XAI) Program, Broad Agency Announcement Explainable Artificial Intelligence (XAI) DARPA-BAA16-.D

[6] Grover, Richa. (2019). Deep Learning - Overview, Practical Examples, Popular Algorithms. Retrived from:

https://www.analyticssteps.com/blogs/deep-learning-overview-practical-examples-popular-algorithms

[7] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A survey of methods for explaining black box models. ACM Comput Surv 1:(5)51

[8] Kim B, Khanna R, Kovejo O (2016) Examples are not Enough, Learn to Criticize! Criticism for Interpretability. Proceedings of Advances in Neural Information Processing Systems, Barcelona, Spain, 29.

[9] Lemos, N. (2007). An introduction to the theory of knowledge. Cambridge University Press.

[10] Nath, Prasenjit. (2015). AI & Expert System in Medical Field: A study by survey method. AITHUN. Volume-I. 100.

[11] Peng, Tony. (10 .(2018 AI Failures. Retrived from: https://medium.com/syncedreview/-2018in-review-10-ai-failures-c18faadf5983

[12] Tamagnini P, Krause J, Dasgupta A, Bertini E (2017) Interpreting Black-Box Classifiers Using Instance-Level Visual Explanations. In Proceedings of HILDA'17, Chicago, IL, USA.

[13] Vellido A, Jos´e D Martin-Guerrero, J Lisboa (2012) Making machine learning models interpretable. Proceedings of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges, Belgium.

[14] Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law, 99-76 ,(2)7.

[15] Young IM (1990) Justice and the politics of difference. Princeton University Press, PrincetonD

[16] Zednik, C. (2019). Solving the black box problem: a normative framework for explainable artificial intelligence. Philosophy & Technology, 24-1.

[17] Zhu YN, Sun L, Li C (2019) Inclusive Leadership, Proactive Personality and Employee Voice: A Voice Role Identity Perspective. Proceedings of the Academy of Management Annual Meeting.


مرکز ملی فضای مجازی
پژوهشگاه فضای مجازی
csri.majazi.ir

حـوزه فضـای مجـازی به اندازه انقـلاب اسـلامی اهمیت دارد. این فضـا مثل یک رودخانـه پر از آب خروشان است که مـی‌آید و دائمـاً هم بر آب آن افزوده و خروشـان تر مـی‌شـود. اگـر مـا بر ایـن رودخانـه تدبیـر کنیـم و برنامـه داشـته باشـیم، زه‌کشـی کنیـم و هدایـت کنیـم ایـن رودخانـه را تا به سـد بریـزد، می‌شـود فرصـت. اگـر رهایـش کنیـم و برنامـه‌ای بـرای آن نداشـته باشـیم می‌شـود یـک تهدید.

csri.ac.ir