«به نام خدا»

ارزیابی هوشمند قابلیت اعتماد در سامانه ها و الگوریتم های هوش مصنوعی

Intelligent assessment of trustworthiness in artificial intelligence systems and algorithms

چکیده

هوش مصنوعی در سال های اخیر به خصوص رشد خیره کننده ای داشته است. هوش مصنوعی می تواند مزایای قابل توجهی را بر جنبه های مختلفی از زندگی ما به ارمغان بیاورد. در عین حال، نگرانی هایی نیز مطرح شده است؛ قابل اعتماد بودن پیش نیاز مردم و جوامع برای توسعه، استقرار و استفاده از سیستم های هوش مصنوعی است. اهمیت این مساله به اندازه ای است که به تازگی اکثر کشورهای توسعه یافته قوانینی را برای قابلیت اعتماد در سامانه های هوش مصنوعی ارائه کرده اند. پیاده سازی و تحقق هوش مصنوعی قابل اعتماد، مستلزم هفت معیار است که باید به نحو شایسته ای برآورده شوند: عامل انسانی و نظارت، استحکام فنی و ایمنی، حریم خصوصی و حاکمیت داده، شفافیت، تنوع، عدم تبعیض و انصاف، رفاه اجتماعی و محیطی و مسئولیت پذیری. برای ارزیابی هر یک از این معیارها تحقیقات گسترده و متنوعی در حال انجام است. هدف این پژوهش آن است که با کمی سازی هر یک این معیارها و در نهایت تجمیع آن ها به صورت هوشمند، بتواند روشی خودکار و یکپارچه برای سنجش سامانه ها و الگوریتم های هوش مصنوعی ارائه دهد. هدف غایی آن است که بتوانیم با استفاده از این ابزار هوشمند، قابلیت اعتماد را در سیستم های مبتنی بر هوش مصنوعی بررسی و تایید کنیم و همچنین روش ها و الگوریتم ها از این جنبه بسیار مهم با هم مقایسه کنیم.

کلید واژه ها: هوش مصنوعی، قابلیت اعتماد، استحکام، شفافیت، عدم تبعیض

Abstract

Artificial intelligence (AI) has seen particularly impressive growth in recent years. It can bring significant benefits to various aspects of our lives. At the same time, some concerns have also been raised; Trustworthiness is a prerequisite for people and societies to develop, deploy and use AI-based systems. The importance of this issue is so much that recently most of the developed countries have provided laws for trustworthiness assessment in AI-based systems. The implementation and realization of trustworthy AI requires seven criteria that must be properly met: Human agency and oversight, Technical robustness and safety, Privacy and data governance, Transparency, Diversity, non-discrimination and fairness, Societal and environmental wellbeing; and Accountability. Extensive and diverse research is being done to evaluate each of these criteria. The aim of this research is to provide an automatic and integrated method for measuring AI-based systems and algorithms trustworthiness by quantifying each of these criteria and finally aggregating them intelligently. The ultimate goal is to be able to check and confirm the reliability of AI-based systems using this smart tool and also compare methods and algorithms in this very important aspect.

Keywords: artificial intelligence, trustworthy AI, robustness, transparency, fairness

مقدمه

در عصر حاضر، هوش مصنوعی به پیشرفت های قابل توجهی رسیده است و هسته ی بسیاری از فعالیت هایی است که از تکنولوژی های نوین استفاده می کنند. با وجود اینکه ریشه های هوش مصنوعی به دهه های گذشته باز می گردد، در سال های اخیر توانایی هوش مصنوعی صرف یادگیری، استدلال و سازگاری مدل ها شده است. به موجب همین توانایی ها، متدهای هوش مصنوعی به سطحی بی سابقه از عملکرد در یادگیری و حل مسائل بسیار پیچیده محاسباتی رسیده اند. کاربردهای یادگیری ماشین و یادگیری عمیق در زمینه های مختلفی مانند بینایی کامپیوتر و پردازش زبان های طبیعی و پردازش صوت و تولید محتواهای فراواقعی در حال گسترش است. این کاربرد ها در اموری حیاتی مانند پزشکی، قانون و ماشین های خودران باعث ظهور موقعیت های جدید اتوماسیون می شوند.

هوش مصنوعی در سال های اخیر به خصوص با معرفی مدل زبانی در مقیاس بالا مانند Chat-GPT، رشد خیره کننده ای داشته است. هنگامی که فناوری‌های جدیدی مانند هوش مصنوعی (AI) را در نظر می‌گیریم، پیش‌بینی می‌شود چه خوب یا بد تأثیر زیادی بر آینده بشریت داشته باشد و تصورش کمی گیج کننده است. به عنوان مثال، هوش مصنوعی می تواند مزایای قابل توجهی را بر جنبه های مختلفی از زندگی ما، از پیش بینی آب و هوا گرفته تا تشخیص سرطان، به ارمغان بیاورد. در عین حال، نگرانی هایی مطرح شده است؛ مثلا این که می تواند بسیاری از مشاغل را تهدید کند و فرآیندهای مهم تصمیم گیری را بدون شفافیت در اختیار بگیرد.

از سویی دیگر هوش مصنوعی با پیشرفت زیادی که داشته، در زمینه جنبه ها و تاثیرات آن در زندگی انسان ها مورد غفلت و نگرانی قرار گرفته است. در حقیقت روش های موجود به دقت خوبی رسیده اند، اما قابلیت اطمینان آنها نیاز به بررسی دارد. امیدها و ترس های ما در مورد هوش مصنوعی فقط مربوط به آینده های دور نبوده بلکه اغلب در مورد هوش مصنوعی امروزی است که در حال حاضر تأثیر قابل توجهی بر زندگی ما دارد. برای مثال، هوش مصنوعی هم بخشی از مشکل و هم راه حل وجود اخبار جعلی است. همچنین الگوریتم‌های هوش مصنوعی برای حمایت از بی‌طرفی مورد استفاده قرار گرفته‌اند، اما گاهی به تعصب نژادی متهم می‌شوند.

چهره های شناخته شده در این حوزه، به هر دو طرف این بحث پیوسته اند. به عنوان مثال، ایلان ماسک نگرانی های خود را مبنی بر اینکه هوش مصنوعی یک تهدید وجودی برای نژاد بشر است ابراز می کند. در حالی که بیل گیتس معتقد است که این فناوری ما را سازنده و خلاق تر می کند. با این حال، هر دو بر این بارور هستند که هوش مصنوعی طیف گسترده‌ای از فرصت‌ها و چالش‌ها را ارائه می‌کند و هر دو خواستار تأمل بر آن هستندکه چگونه می‌توانیم توسعه آن را به گونه‌ای مدیریت کنیم که مزایای آن را بدون اینکه ما را در معرض خطر قرار دهد به دست آوریم.

هوش مصنوعی قابل اعتماد[1]

[1] Trustworthy AI

قابل اعتماد بودن پیش نیاز مردم و جوامع برای توسعه، استقرار و استفاده از سیستم های هوش مصنوعی است. اگر سیستم‌های هوش مصنوعی به وضوح شایسته اعتماد نباشند، ممکن است عواقب ناخواسته‌ای به وجود بیاید و مانع از جذب آنها شود و از تحقق منافع اجتماعی و اقتصادی بالقوه گسترده‌ای که می‌توانند به همراه داشته باشند، جلوگیری کند.

اعتماد در توسعه، استقرار و استفاده از سیستم‌های هوش مصنوعی نه تنها به ویژگی‌های ذاتی فناوری است، بلکه به کیفیت سیستم‌های هوش مصنوعی نیز مربوط می‌شود. تلاش برای دستیابی به هوش مصنوعی قابل اعتماد علاوه بر این که به قابلیت اعتماد خود سیستم هوش مصنوعی مربوط می‌شود، نیازمند رویکردی جامع و سیستمی است که قابل اعتماد بودن همه فرآیندهایی را که بخشی از بافت اجتماعی-فنی سیستم در کل چرخه حیات آن هستند، در بر می‌گیرد.

هوش مصنوعی قابل اعتماد دارای سه جزء است که باید در کل چرخه عمر سیستم رعایت شود:

۱. باید قانونی[۲] و مطابق با کلیه قوانین و مقررات قابل اجرا باشد.

۲. باید اخلاقی[۳] باشد و از پایبندی به اصول و ارزش های اخلاقی اطمینان حاصل کند.

۳. باید از منظر فنی و اجتماعی مستحکم [۴] باشد، زیرا، حتی با هدف خوب، سیستم‌های هوش مصنوعی می‌توانند آسیب‌های غیرعمدی ایجاد کنند.

هر یک از این سه جزء برای دستیابی به هوش مصنوعی قابل اعتماد لازم است اما به خودی خود کافی نیست. در حالت ایده‌آل، هر سه در عملکرد خود هماهنگ و همپوشانی دارند. با این حال، در عمل ممکن است تنش هایی بین این عناصر وجود داشته باشد. یک رویکرد قابل اعتماد برای فعال کردن «رقابت‌پذیری مسئولانه»، بستری را فراهم می کند که بر اساس آن همه کسانی که تحت تأثیر سیستم‌های هوش مصنوعی قرار می‌گیرند، می‌توانند اعتماد کنند که طراحی، توسعه و استفاده از آنها قانونی، اخلاقی و مستحکم هستند. این امر کشورمان ایران را قادر می‌سازد تا خود را به عنوان یک پیشرو در زمینه هوش مصنوعی پیشرفته و شایسته اعتماد فردی و جمعی قرار دهد و تنها با اطمینان از قابل اعتماد بودن، می توانیم به طور کامل از مزایای سیستم‌های هوش مصنوعی بهره‌مند شویم. همانطور که استفاده از سیستم‌های هوش مصنوعی در مرزهای ملی متوقف نمی‌شود، تاثیر آنها نیز متوقف نمی‌گردد. بنابراین برای فرصت ها و چالش های جهانی که سیستم های هوش مصنوعی به وجود می آورند، راه حل های جهانی لازم است. ما همه ذینفعان را تشویق می‌کنیم تا در راستای چارچوبی جهانی برای هوش مصنوعی قابل اعتماد، ایجاد توافق بین‌المللی و در عین حال ترویج و حمایت از رویکرد ملی ما کار کنند.

مبانی هوش مصنوعی قابل اعتماد

اخلاق هوش مصنوعی بر مسائل اخلاقی ناشی از توسعه، استقرار و استفاده از هوش مصنوعی تمرکز دارد. دغدغه اصلی آن شناسایی این است که چگونه هوش مصنوعی می تواند زندگی خوب افراد را ارتقا دهد یا نگرانی هایی را در مورد آن چه از نظر کیفیت زندگی، چه از نظر استقلال انسانی و آزادی لازم برای یک جامعه، ایجاد کند.

---

[2] Lawful
[3] Ethical
[4] Robust

هوش مصنوعی قابل اعتماد می تواند با ایجاد رفاه و خلق ارزش، شکوفایی فردی و رفاه جمعی را به حداکثر برساند. همچنین می تواند با کمک به افزایش سلامت و رفاه شهروندان، برابری در توزیع فرصت های اقتصادی، اجتماعی و سیاسی را تقویت کند و به دستیابی به یک جامعه عادلانه کمک کند. بنابراین ضروری است که بدانیم چگونه از توسعه، استقرار و استفاده از هوش مصنوعی به بهترین شکل حمایت کنیم تا اطمینان حاصل کنیم که همه می توانند در دنیای مبتنی بر هوش مصنوعی پیشرفت کنند و آینده بهتری بسازیم و در عین حال در بتوانیم در سطح جهانی رقابت کنیم.

در این مقاله قصد داریم با معرفی مفهوم هوش مصنوعی قابل اعتماد، که به اعتقاد ما راه درستی برای ساختن آینده با هوش مصنوعی است، به این تلاش کمک کنیم. آینده ای که در آن دموکراسی، حاکمیت قانون و حقوق اساسی زیربنای سیستم های هوش مصنوعی باشد. چنین سیستم هایی به طور مداوم فرهنگ دموکراتیک را بهبود می بخشند و از آن دفاع می کنند، محیطی را ایجاد می کنند که در آن نوآوری و رقابت مسئولانه می تواند رشد کند.

اصول اخلاقی در سیستم های هوش مصنوعی

سیستم های هوش مصنوعی باید رفاه فردی و جمعی را بهبود بخشند. چهار اصل اخلاقی را فهرست می‌کنیم که ریشه در حقوق اساسی دارد، که باید به منظور اطمینان از توسعه، استقرار و استفاده از سیستم‌های هوش مصنوعی به شیوه‌ای قابل اعتماد، رعایت شوند. این اصول عبارتند از:

۱- احترام به استقلال انسان[5]

۲- پیشگیری از آسیب[6]

۳- انصاف[7]

۴- توضیح پذیری[8]

اصل احترام به استقلال انسان

سیستم‌های هوش مصنوعی نباید انسان‌ها را به‌طور غیرقابل توجیهی تابع، اجبار، فریب، دستکاری، شرطی‌سازی یا گله‌دار کنند. در عوض، آنها باید برای تقویت، تکمیل و توانمندسازی مهارت های شناختی، اجتماعی و فرهنگی انسان طراحی شوند. تخصیص کارها بین انسان ها و سیستم های هوش مصنوعی باید از اصول طراحی انسان محور پیروی کند و فرصت معناداری برای انتخاب انسان ایجاد کند. سیستم های هوش مصنوعی همچنین ممکن است به طور اساسی محیط کاری را تغییر دهند. باید از انسان در محیط کار حمایت کنند و هدف آنها ایجاد کار معنادار باشد.

[5] Respect for human autonomy
[6] Prevention of harm
[7] Fairness
[8] Explicability

اصل پیشگیری از آسیب

سیستم‌های هوش مصنوعی نباید باعث ایجاد آسیب یا تشدید آسیب شوند یا در غیر این صورت بر انسان‌ها تأثیر منفی بگذارند. سیستم های هوش مصنوعی و محیط هایی که در آن کار می کنند باید امن و ایمن باشند. پیشگیری از آسیب نیز مستلزم توجه به محیط طبیعی و همه موجودات زنده است.

اصل انصاف

توسعه، استقرار و استفاده از سیستم‌های هوش مصنوعی باید منصفانه باشد. در حالی که ما اذعان داریم که تعابیر مختلفی از انصاف وجود دارد، معتقدیم که انصاف هم بعد ماهوی و هم بعد رویه ای دارد. بعد ماهوی متضمن تعهد به موارد زیر است: تضمین توزیع برابر و عادلانه منافع و هزینه‌ها، و اطمینان از اینکه افراد و گروه‌ها عاری از تعصب، تبعیض و انگ ناعادلانه هستند. اگر بتوان از تعصبات ناعادلانه اجتناب کرد، سیستم‌های هوش مصنوعی حتی می‌توانند عدالت اجتماعی را افزایش دهند. علاوه بر این، استفاده از سیستم‌های هوش مصنوعی هرگز نباید منجر به فریب افراد یا آسیب غیرقابل توجیهی در آزادی انتخاب آنها شود. بعد رویه ای انصاف مستلزم توانایی رقابت تصمیمات اتخاذ شده توسط سیستم های هوش مصنوعی و توسط انسان هایی است که آنها را اداره می کنند.

اصل توضیح پذیری

امروزه با توجه به کاربردهای روز افزون سیستم‌های هوشمند در جنبه‌های حیاتی زندگی بشر و حل مسایل پیچیده‌ی دنیای واقعی، ضرورت بسط و توسعه ی شاخه‌ی جدیدی از هوش مصنوعی به نام هوش مصنوعی شرح‌پذیر۹ بیش از پیش مورد تمرکز قرار گرفته است. هوش مصنوعی شرح‌پذیر به دنبال ارایه‌ی روش‌ها و راه‌حل‌های هوشمند متناسب با فضای درک و شناخت انسان می‌باشد. به بیانی ساده، ابزارها و روش های شرح پذیری کمک می‌کنند تصمیمات و عملکرد سیستم‌های هوش مصنوعی به صورت قابل فهم برای کاربران انسانی توضیح داده شود و در نتیجه به کاربران کمک می‌کند تا بهتر با سیستم‌های هوش مصنوعی همکاری کنند و از آن‌ها یاد بگیرند.

در ابتدا شرح و تفسیر سیستم‌های پایه و روش‌های خطی هوش مصنوعی کار ساده ای بود. اما با توجه به اینکه در سال‌های اخیر شاهد توسعه‌ی مدل‌های پیچیده و غیر خطی تصمیم گیری مانند شبکه‌های عصبی عمیق هستیم، به ابزارها و روش‌های پیچیده و دقیق تری برای فهم انسانی این مدل‌ها نیازمندیم. .این مدل‌ها اگرچه از قدرت و دقت پیش‌بینی بالاتری برخوردارند، اما در فرآیند استنتاج و تصمیم گیری به صورت جعبه سیاه[10] عمل می‌کنند و نتایج آنها به اندازه‌ی کافی برای کاربران غیر فنی در حوزه‌هایی مانند سلامت، صنعت، حمل‌ونقل و مسایل مالی قابل توجیه و اتکا نمی‌باشند. به عنوان مثال، خروجی شبکه‌های عصبی عمیق با استفاده از روابط پیچیده ریاضی، که به عواملی از جمله توابع فعال‌ساز، نوع و اندازه‌ی داده‌ی ورودی، تعداد لایه‌ها، و الگوی ارتباط آنها بستگی دارد، بدست می‌آید. تعدد و پیچیدگی این عوامل باعث می‌شود که درک خروجی این مدل‌ها و باطبع اعتماد به تصمیم آنها یا به طور کلی مقدور نباشد و یا به دشواری انجام گیرد. به عبارتی کاربران به سختی می‌توانند پاسخ‌های قانع کننده ای برای چنین سوالاتی پیدا کنند:

- چرا مدل به ازای یک ورودی مشخص چنین خروجی مشخصی را تولید می‌کند؟
- در چه موارد مشابهی و به ازای چه ورودی‌های دیگری چنین خروجی‌هایی تولید می‌شود؟

---

[9] eXplainable Artificial Intelligence(XAI)

[10] Black-Box

- چرا مدل به ازای یک ورودی مشخص خروجی اشتباه تولید می‌کند؟
- در چه موارد مشابه دیگری امکان چنین پیش بینی‌های غلطی وجود دارد؟

ابزارها و روش‌های هوش مصنوعی شرح‌پذیر به دنبال پاسخ به چنین سوالاتی می‌باشند و تلاش می‌نمایند تا با ایجاد تغییرات موثر در فرآیند آموزش مدل‌ها و یا با استفاده از سیستم‌های کمکی ذاتا شرح پذیر (مانند مدل‌های خطی، درخت‌های تصمیم[11]، سیستم‌های مبتنی بر قوانین و غیره) ماهیت درونی مدل‌های پیچیده را آشکار و درک بهتری از این چرایی‌ها هم برای کاربران غیر فنی و هم برای توسعه دهندگان سیستم‌ها فراهم نمایند. در این راستا، کاربران غیرمتخصص به دلیل درک قابل فهم از عملکردِ درونی سیستم همکاری و تعامل بهتری با روش‌های هوشمند برقرار می‌کنند و در نتیجه کارایی و اثربخشی استفاده از سیستم‌های هوشمند افزایش می‌یابد. از طرف دیگر، ماهیت آشکار مدل‌ها می‌تواند به توسعه‌دهندگان کمک کند تا به صورت موثرتری نقاط ضعف و قوت سیستم‌های خود را شناسایی کنند و در جهت بهبود آنها اقدام نمایند. افزایش سطح اطمینان، رضایت و بهره‌وری و سهولت تعمیم و تفسیر موارد مشابه و کاهش سوگیری‌های تبعیض آمیز در تصمیم گیری‌ها از دیگر مزایای استفاده از ابزارها و روش‌های شرح پذیری می‌باشد.

الزامات هوش مصنوعی قابل اعتماد

این بخش در مورد پیاده سازی و تحقق هوش مصنوعی قابل اعتماد، فهرستی از هفت الزامی که باید برآورده شوند، ارائه می دهد.

برای دستیابی به هوش مصنوعی قابل اعتماد، اصول ذکر شده در بخش اول باید به الزامات ملموس تبدیل شود.

گروه های مختلف ذینفعان نقش های متفاوتی در حصول اطمینان از برآورده شدن الزامات دارند:

آ. توسعه دهندگان[12] باید الزامات را برای فرآیندهای طراحی و توسعه پیاده سازی و اعمال کنند.

ب. سفارش دهندگان[13] باید اطمینان حاصل کنند که سیستم هایی که استفاده می کنند و محصولات و خدماتی که ارائه می دهند الزامات را برآورده می کنند.

ج. کاربران نهایی [14] و جامعه گسترده تر باید در مورد این الزامات مطلع شوند و بتوانند درخواست کنند که مورد حمایت قرار گیرند.

فهرست الزامات زیر شامل جنبه های سیستمی، فردی و اجتماعی است:

۱. عامل انسانی و نظارت [15] از جمله حقوق اساسی، عاملیت انسانی و نظارت انسانی.

---

[11] Decision tree
[12] Developers
[13] Deployers
[14] End-users
[15] Human agency and oversight

۲. استحکام فنی و ایمنی[۱۶] از جمله انعطاف پذیری در برابر حمله و امنیت، طرح عقب نشینی و ایمنی عمومی، دقت، قابلیت اطمینان و تکرارپذیری.
۳. حریم خصوصی و حاکمیت داده[۱۷] از جمله احترام به حریم خصوصی، کیفیت و یکپارچگی داده ها و دسترسی به داده ها.
۴. شفافیت[۱۸] از جمله قابلیت ردیابی، توضیح پذیری و ارتباط.
۵. تنوع، عدم تبعیض و انصاف[۱۹] از جمله اجتناب از سوگیری ناعادلانه، دسترسی و طراحی جهانی و مشارکت ذینفعان.
۶. رفاه اجتماعی و محیطی[۲۰] از جمله پایداری و سازگاری با محیط زیست، تأثیر اجتماعی، جامعه و دموکراسی.
۷. مسئولیت پذیری[۲۱] از جمله قابلیت حسابرسی، به حداقل رساندن گزارش اثرات منفی، مبادلات و جبران خسارت.

در ادامه، هر یک از الزامات با جزئیات بیشتری توضیح داده شده است.

عامل انسانی و نظارت

سیستم‌های هوش مصنوعی باید از استقلال و تصمیم‌گیری انسان پشتیبانی کنند، همانطور که در اصل احترام به استقلال انسان آمده است. این امر مستلزم آن است که سیستم‌های هوش مصنوعی باید با حمایت از عامل کاربر و حمایت از حقوق اساسی، به عنوان توانمند برای یک جامعه دموکراتیک، شکوفا و عادلانه عمل کنند و اجازه نظارت انسانی را بدهند.

عامل انسانی: کاربران باید بتوانند در مورد سیستم‌های هوش مصنوعی تصمیم‌گیری مستقل و آگاهانه بگیرند. باید دانش و ابزار لازم برای درک و تعامل با سیستم‌های هوش مصنوعی به میزان رضایت‌بخشی به آن‌ها داده شود و در صورت امکان، بتوانند سیستم را به‌طور منطقی ارزیابی کنند یا به چالش بکشند. سیستم‌های هوش مصنوعی باید از افراد در انتخاب‌های بهتر و آگاهانه‌تر مطابق با اهدافشان حمایت کنند. سیستم‌های هوش مصنوعی گاهی اوقات می‌توانند برای شکل‌دهی و تأثیرگذاری بر رفتار انسان از طریق مکانیسم‌هایی به کار گرفته شوند که تشخیص آن‌ها دشوار باشد، زیرا ممکن است فرآیندهای ناخودآگاه، از جمله اشکال مختلف دستکاری ناعادلانه، فریب، گله‌داری و شرطی‌سازی را مهار کنند، که همگی ممکن است استقلال فردی را تهدید کنند.

نظارت انسانی: نظارت انسانی کمک می‌کند تا اطمینان حاصل شود که یک سیستم هوش مصنوعی استقلال انسان را تضعیف نمی‌کند یا اثرات نامطلوب دیگری ایجاد نمی‌کند. نظارت ممکن است از طریق مکانیسم های حاکمیتی مانند رویکرد انسان در حلقه[۲۲] (HITL)، انسان روی حلقه [۲۳] (HOTL) یا انسان در فرمان[۲۴] (HIC) به دست آید. HITL به قابلیت دخالت انسان در هر چرخه تصمیم گیری سیستم اشاره دارد که در بسیاری از موارد نه ممکن است و نه مطلوب. HOTL به قابلیت مداخله انسانی در طول چرخه طراحی سیستم و نظارت بر عملکرد سیستم اشاره دارد. HIC به توانایی نظارت بر فعالیت کلی سیستم هوش مصنوعی (شامل تأثیر اقتصادی، اجتماعی، قانونی و اخلاقی گسترده تر آن) و توانایی تصمیم گیری در مورد زمان و نحوه استفاده از سیستم در هر موقعیت خاص اشاره داردکه می تواند شامل تصمیم برای عدم استفاده از یک سیستم هوش مصنوعی در یک موقعیت خاص، ایجاد سطوحی از

16 Technical robustness and safety
17 Privacy and data governance
18 Transparency
19 Diversity, non-discrimination and fairness
20 Societal and environmental wellbeing
21 Accountability
22 human-in-the-loop
23 human-on-the-loop
24 human-in-command

اختیارات انسانی در طول استفاده از سیستم، یا اطمینان از توانایی نادیده گرفتن تصمیمی که توسط یک سیستم گرفته می شود، باشد.

استحکام فنی و ایمنی

یکی از اجزای حیاتی دستیابی به هوش مصنوعی قابل اعتماد، استحکام فنی است که ارتباط نزدیکی با اصل پیشگیری از آسیب دارد. استحکام فنی مستلزم آن است که سیستم‌های هوش مصنوعی با رویکردی پیشگیرانه نسبت به ریسک‌ها رفتار کنند و به گونه‌ای توسعه داده شوند که قابل اعتماد باشند و در عین حال آسیب‌های غیرعمدی و غیرمنتظره را به حداقل برسانند و از آسیب‌های غیرقابل قبول جلوگیری کنند.

مقاومت در برابر حمله و امنیت[۲۵]: سیستم‌های هوش مصنوعی، مانند همه سیستم‌های نرم‌افزاری، باید در برابر آسیب‌پذیری‌هایی که می‌توانند توسط دشمنان مورد سوء استفاده قرار گیرند، محافظت شوند، به عنوان مثال. هک کردن. حملات ممکن است داده ها (مسمومیت داده ها[۲۶])، مدل (نشت مدل[۲۷]) یا زیرساخت های اساسی، نرم افزار و سخت افزار را هدف قرار دهند. اگر یک سیستم هوش مصنوعی مورد حمله قرار گیرد، به عنوان مثال. در حملات خصمانه[۲۸]، داده‌ها و همچنین رفتار سیستم را می‌توان تغییر داد و سیستم را به تصمیم‌گیری‌های متفاوتی سوق داد و یا باعث از بین رفتن کلی آن شد.

طرح بازگشتی و ایمنی عمومی[۲۹]: سیستم‌های هوش مصنوعی باید محافظ‌هایی داشته باشند که در صورت بروز مشکل، یک طرح بازگشتی را ممکن می‌سازد. می تواند به این معنی باشد که سیستم های هوش مصنوعی از رویه آماری به رویه مبتنی بر قوانین تغییر می کنند، یا اینکه قبل از ادامه فعالیت خود از اپراتور انسانی درخواست می کنند. باید اطمینان حاصل شود که سیستم آنچه را که قرار است انجام دهد بدون آسیب رساندن به موجودات زنده یا محیط زیست انجام خواهد داد که شامل به حداقل رساندن عواقب و خطاهای ناخواسته است.

دقت[۳۰]: دقت به توانایی یک سیستم هوش مصنوعی برای قضاوت صحیح گفته می شود، به عنوان مثال برای طبقه بندی صحیح اطلاعات در دسته بندی های مناسب، یا توانایی آن برای پیش بینی، توصیه یا تصمیم گیری صحیح بر اساس داده ها یا مدل ها. یک فرآیند توسعه و ارزیابی صریح و به خوبی شکل گرفته می تواند خطرات ناخواسته ناشی از پیش بینی های نادرست را پشتیبانی، کاهش و اصلاح کند. هنگامی که نمی توان از پیش بینی های نادرست گاه به گاه اجتناب کرد، مهم است که سیستم بتواند میزان احتمال این خطاها را مشخص کند. سطح بالایی از دقت به ویژه در شرایطی که سیستم هوش مصنوعی مستقیماً بر زندگی انسان تأثیر می گذارد بسیار مهم است.

---

25 Resilience to attack and security
26 data poisoning
27 model leakage
28 adversarial attacks
29 Fallback plan and general safety
30 Accuracy

قابلیت اطمینان و تکرارپذیری[۳۱]: بسیار مهم است که نتایج سیستم های هوش مصنوعی قابل تکرار و همچنین قابل اعتماد باشند. یک سیستم هوش مصنوعی قابل اطمینان سیستمی است که با طیف وسیعی از ورودی ها و در موقعیت های مختلف به درستی کار می کند. این برای بررسی دقیق یک سیستم هوش مصنوعی و جلوگیری از آسیب‌های ناخواسته لازم است. تکرارپذیری توصیف می‌کند که آیا آزمایش هوش مصنوعی در شرایط مشابه، رفتار یکسانی را نشان می‌دهد یا خیر. این به دانشمندان و سیاست گذاران این امکان را می دهد تا به طور دقیق آنچه را که سیستم های هوش مصنوعی انجام می دهند، توصیف کنند.

حریم خصوصی و حاکمیت داده

حریم خصوصی ارتباط نزدیکی با اصل پیشگیری از آسیب دارد، که یک حق اساسی است و به صورت ویژه توسط سیستم های هوش مصنوعی تاثیر می پذیرد.

حریم خصوصی و حفاظت از داده ها[۳۲]: سیستم‌های هوش مصنوعی باید حریم خصوصی و حفاظت از داده‌ها را در کل چرخه حیات سیستم تضمین کنند. این شامل اطلاعاتی است که در ابتدا توسط کاربر ارائه می‌شود، و همچنین اطلاعاتی که در مورد کاربر در طول تعامل با سیستم ایجاد می‌شود (به عنوان مثال خروجی‌هایی که سیستم هوش مصنوعی برای کاربران خاص ایجاد کرده است یا نحوه پاسخ کاربران به توصیه‌های خاص). سوابق دیجیتالی رفتار انسان ممکن است به سیستم‌های هوش مصنوعی اجازه دهد که نه تنها ترجیحات افراد، بلکه گرایش جنسی، سن، جنسیت، دیدگاه‌های مذهبی یا سیاسی آنها را نیز استنباط کنند. برای اینکه افراد بتوانند به فرآیند جمع‌آوری داده‌ها اعتماد کنند، باید اطمینان حاصل شود که داده‌های جمع‌آوری‌شده درباره آنها برای تبعیض غیرقانونی یا غیرمنصفانه علیه آنها استفاده نمی‌شود.

کیفیت و یکپارچگی داده ها[۳۳]: کیفیت مجموعه داده های مورد استفاده برای عملکرد سیستم های هوش مصنوعی بسیار مهم است. هنگامی که داده ها جمع آوری می شوند، ممکن است دارای سوگیری های اجتماعی، نادرستی، خطاها و اشتباهات باشند. این باید قبل از آموزش با هر مجموعه داده ای مشخص شود. علاوه بر این، یکپارچگی داده ها باید تضمین شود. تغذیه داده های مخرب به یک سیستم هوش مصنوعی ممکن است رفتار آن را تغییر دهد، به ویژه در مورد سیستم های خودآموز. فرآیندها و مجموعه داده های مورد استفاده باید در هر مرحله مانند برنامه ریزی، آموزش، آزمایش و استقرار آزمایش و مستند شوند. این مساله باید در مورد سیستم‌های هوش مصنوعی که در داخل توسعه نیافته‌اند اما در جای دیگری خریداری شده‌اند نیز اعمال شود.

دسترسی به داده ها[۳۴]: در هر سازمانی که داده‌های افراد را مدیریت می‌کند (چه شخصی کاربر سیستم باشد یا نباشد)، باید پروتکل‌های داده‌ای حاکم بر دسترسی به داده‌ها ایجاد شود. این پروتکل ها باید مشخص کنند که چه کسانی و تحت چه شرایطی می توانند به داده ها دسترسی داشته باشند. فقط پرسنل واجد شرایط که صلاحیت و نیاز به دسترسی به داده های افراد را دارند باید مجاز به انجام این کار باشند.

---

31 Reliability and Reproducibility
32 Privacy and data protection
33 Quality and integrity of data
34 Access to data.

شفافیت[۳۵]

این الزام ارتباط نزدیکی با اصل توضیح پذیری دارد و شامل شفافیت عناصر مرتبط با یک سیستم هوش مصنوعی است: داده ها، سیستم و مدل های تجاری.

قابلیت ردیابی [۳۶] : مجموعه داده‌ها و فرآیندهایی که تصمیم سیستم هوش مصنوعی را می سازند، از جمله موارد جمع‌آوری داده‌ها و برچسب‌گذاری داده‌ها و همچنین الگوریتم‌های مورد استفاده، باید به بهترین استاندارد ممکن مستند شوند تا امکان ردیابی و افزایش شفافیت فراهم شود. این امر در مورد تصمیمات اتخاذ شده توسط سیستم هوش مصنوعی نیز صدق می کند. این امکان شناسایی دلایل اشتباه بودن یک تصمیم هوش مصنوعی را فراهم می کند که به نوبه خود می تواند به جلوگیری از اشتباهات آینده کمک کند. قابلیت ردیابی، توضیح پذیری را تسهیل می کند.

توضیح پذیری [۳۷]: توضیح پذیری به توانایی توضیح فرآیندهای فنی یک سیستم هوش مصنوعی و تصمیمات انسانی مرتبط (مانند حوزه های کاربردی یک سیستم) مربوط می شود. توضیح فنی مستلزم آن است که تصمیمات اتخاذ شده توسط یک سیستم هوش مصنوعی توسط انسان قابل درک و ردیابی باشد. علاوه بر این، ممکن است بین افزایش قابلیت توضیح یک سیستم (که ممکن است دقت آن را کاهش دهد) یا افزایش دقت آن (به قیمت توضیح پذیری) معاوضه هایی انجام شود. هر زمان که یک سیستم هوش مصنوعی تأثیر قابل توجهی بر زندگی افراد داشته باشد، باید بتوان توضیح مناسبی از فرآیند تصمیم گیری سیستم هوش مصنوعی خواست. چنین توضیحی باید به موقع و با تخصص ذینفعان مربوطه (مثلاً افراد عادی، تنظیم کننده یا محقق) تطبیق داده شود.

ارتباط[۳۸]: سیستم های هوش مصنوعی نباید خود را به عنوان انسان برای کاربران نشان دهند. انسان ها حق دارند از تعامل با یک سیستم هوش مصنوعی مطلع شوند. این مستلزم آن است که سیستم‌های هوش مصنوعی باید به این صورت قابل شناسایی باشند. فراتر از این، قابلیت‌ها و محدودیت‌های سیستم هوش مصنوعی باید به شیوه‌ای متناسب با کاربرد مورد نظر به متخصصان هوش مصنوعی یا کاربران نهایی اطلاع داده شود.

تنوع، عدم تبعیض و انصاف

برای دستیابی به هوش مصنوعی قابل اعتماد، باید شمول و تنوع را در کل چرخه حیات سیستم هوش مصنوعی فعال کنیم که شامل در نظر گرفتن و مشارکت همه ذینفعان تحت تأثیر در طول فرآیند، تضمین دسترسی برابر از طریق فرآیندهای طراحی فراگیر و رفتار برابر است. این الزام ارتباط تنگاتنگی با اصل انصاف دارد.

پرهیز از تعصب ناعادلانه[۳۹]: مجموعه داده‌های مورد استفاده توسط سیستم‌های هوش مصنوعی (هم برای آموزش و هم برای عملیات) ممکن است از داشتن مدل‌های قدیمی، ناقص بودن و مدیریت بد رنج ببرند. ادامه چنین سوگیری ها می تواند منجر به تعصب و تبعیض ناخواسته (غیر) مستقیم علیه گروه ها یا افراد خاص شود و به طور بالقوه تعصب و به حاشیه راندن را تشدید کند.

---

35 Transparency
36 Traceability
37 Explainability
38 Communication
39 Avoidance of unfair bias

آسیب همچنین می‌تواند ناشی از بهره‌برداری عمدی از سوگیری‌ها یا درگیر شدن در رقابت ناعادلانه، مانند همگن‌سازی قیمت‌ها از طریق تبانی یا بازار غیرشفاف باشد. روشی که در آن سیستم‌های هوش مصنوعی توسعه می‌یابند (مثلاً برنامه‌نویسی الگوریتم‌ها) ممکن است از سوگیری ناعادلانه رنج ببرند. با ایجاد فرآیندهای نظارتی برای تحلیل و رسیدگی به اهداف، محدودیت‌ها، الزامات و تصمیمات سیستم به شیوه‌ای شفاف و شفاف، می‌توان با این امر مقابله کرد.

دسترسی و طراحی جهانی[40]: به‌ویژه در حوزه‌های کسب‌وکار، سیستم‌ها باید کاربر محور باشند و به گونه‌ای طراحی شوند که به همه افراد امکان استفاده از محصولات یا خدمات هوش مصنوعی را بدون در نظر گرفتن سن، جنسیت، توانایی‌ها یا ویژگی‌هایشان بدهد. اگر دسترسی به این فناوری برای افراد دارای معلولیت که در تمامی گروه های اجتماعی وجود دارد، از اهمیت ویژه ای برخوردار است. سیستم‌های هوش مصنوعی نباید یک رویکرد یک‌اندازه برای همه داشته باشند و باید اصول طراحی جهانی را در نظر بگیرند که به وسیع‌ترین طیف ممکن از کاربران، پیروی از استانداردهای دسترسی مرتبط را بپردازند.

مشارکت ذینفعان[41]: به منظور توسعه سیستم های هوش مصنوعی که قابل اعتماد هستند، توصیه می شود با سهامدارانی که ممکن است به طور مستقیم یا غیرمستقیم تحت تأثیر سیستم در طول چرخه عمر آن قرار گیرند، مشورت کنید. درخواست بازخورد منظم حتی پس از استقرار و راه‌اندازی مکانیسم‌های بلندمدت برای مشارکت ذینفعان سودمند است، به‌عنوان مثال با اطمینان از اطلاعات، مشاوره و مشارکت کارگران در کل فرآیند اجرای سیستم‌های هوش مصنوعی در سازمان‌ها.

رفاه اجتماعی و محیطی

در راستای اصول انصاف و پیشگیری از آسیب، جامعه وسیع‌تر، سایر موجودات ذی‌شعور و محیط نیز باید به‌عنوان ذینفعان در طول چرخه حیات سیستم هوش مصنوعی در نظر گرفته شوند. پایداری و مسئولیت زیست‌محیطی سیستم‌های هوش مصنوعی باید تشویق شود و تحقیقات باید در زمینه راه‌حل‌های هوش مصنوعی با توجه به حوزه‌های نگرانی جهانی، مانند اهداف توسعه پایدار، تقویت شود. در حالت ایده آل، سیستم های هوش مصنوعی باید به نفع همه انسان ها، از جمله نسل های آینده، استفاده شوند.

هوش مصنوعی پایدار و سازگار با محیط زیست[42]: سیستم‌های هوش مصنوعی قول می‌دهند که به رفع برخی از مبرم‌ترین نگرانی‌های اجتماعی کمک کنند، اما باید اطمینان حاصل شود که این امر به بهترین روش ممکن برای محیط‌زیست رخ می‌دهد. توسعه، استقرار و فرآیند استفاده سیستم، و همچنین کل زنجیره تامین آن، باید در این رابطه ارزیابی شود، به عنوان مثال. از طریق بررسی انتقادی استفاده از منابع و مصرف انرژی در طول تمرین و انتخاب گزینه های کم ضررتر. برای تضمین سازگاری با محیط زیست،کل زنجیره تامین سیستم های هوش مصنوعی باید تشویق شوند.

تاثیر اجتماعی [43] : قرار گرفتن همه جا در معرض سیستم های هوش مصنوعی اجتماعی در همه زمینه های زندگی ما (چه در آموزش، کار، مراقبت یا سرگرمی) ممکن است تصور ما از عاملیت اجتماعی را تغییر دهد یا بر روابط اجتماعی و دلبستگی ما تأثیر بگذارد. در حالی که سیستم‌های هوش مصنوعی می‌توانند برای تقویت مهارت‌های اجتماعی مورد استفاده قرار گیرند، اما می‌توانند

[40] Accessibility and universal design
[41] Stakeholder Participation
[42] Sustainable and environmentally friendly AI
[43] Social impact.

به همان اندازه در بدتر شدن آن‌ها نقش داشته باشند. این می تواند بر سلامت جسمی و روانی افراد نیز تأثیر بگذارد. بنابراین اثرات این سیستم ها باید به دقت بررسی شوند.

جامعه و دموکراسی: فراتر از ارزیابی تأثیر توسعه، استقرار و استفاده یک سیستم هوش مصنوعی بر افراد، این تأثیر باید از دیدگاه اجتماعی نیز با در نظر گرفتن تأثیر آن بر نهادها، دموکراسی و جامعه به طور کلی ارزیابی شود.

مسئولیت پذیری

الزام مسئولیت پذیری، مکمل الزامات فوق است و ارتباط نزدیکی با اصل انصاف دارد. این امر مستلزم ایجاد مکانیسم هایی برای اطمینان از مسئولیت و پاسخگویی سیستم های هوش مصنوعی و نتایج آنها، قبل و بعد از توسعه، استقرار و استفاده از آنها است.

قابلیت حسابرسی[۴۴]: قابلیت حسابرسی مستلزم امکان ارزیابی الگوریتم ها، داده ها و فرآیندهای طراحی است. این لزوماً به این معنی نیست که اطلاعات مربوط به مدل‌های کسب‌وکار و مالکیت معنوی مرتبط با سیستم هوش مصنوعی همیشه باید آشکارا در دسترس باشد. ارزیابی توسط حسابرسان داخلی و خارجی، و در دسترس بودن چنین گزارش های ارزیابی، می تواند به قابل اعتماد بودن فناوری کمک کند. در برنامه‌هایی که بر حقوق اساسی تأثیر می‌گذارند، از جمله برنامه‌های کاربردی حیاتی ایمنی، سیستم‌های هوش مصنوعی باید بتوانند به طور مستقل حسابرسی شوند.

به حداقل رساندن و گزارش اثرات منفی[۴۵]: توانایی گزارش دادن در مورد اقدامات یا تصمیماتی که به نتیجه سیستم کمک می کنند و پاسخ به پیامدهای این نتیجه باید تضمین شود. شناسایی، ارزیابی، مستندسازی و به حداقل رساندن تأثیرات منفی احتمالی سیستم‌های هوش مصنوعی به ویژه برای کسانی که مستقیماً تحت تأثیر قرار می‌گیرند، حیاتی است.

مصالحات[۴۶]: هنگام اجرای الزامات فوق، ممکن است تنش هایی بین آنها ایجاد شود که ممکن است به مصالحات اجتناب ناپذیر منجر شود. چنین مبادلاتی باید به شیوه‌ای منطقی و روش‌شناختی در چارچوب مدرن مورد بررسی قرار گیرد. این مستلزم آن است که منافع و ارزش‌های مرتبط با سیستم هوش مصنوعی شناسایی شوند و در صورت بروز تعارض، مبادلات تجاری باید به صراحت مورد تایید قرار گرفته و از نظر خطراتی که برای اصول اخلاقی، از جمله حقوق اساسی دارند، ارزیابی شوند.

جبران خسارت[۴۷]: هنگامی که اشتباه رخ می دهد، مکانیسم های قابل دسترس باید پیش بینی شود که جبران خسارت کافی را تضمین کند. برای اطمینان از اعتماد، دانستن اینکه زمانی که اشتباه رخ می دهد جبران خسارت می شود لازم است. باید به افراد یا گروه های آسیب پذیر توجه ویژه ای شود.

---

[44] Auditability
[45] Minimization and reporting of negative impacts
[46] Trade-offs.
[47] Redress

هوش مصنوعی قابل اعتماد در جهان

در سال های اخیر قابلیت اعتماد به هوش مصنوعی در سراسر جهان مورد نگرانی واقع شده است. در نتیجه دولت ها و موسسات تحقیقاتی و تجاری در سراسر جهان برآنند که با تصویب قوانین برای سیستم های مبتنی بر هوش مصنوعی از مخاطرات آن بکاهند. در ادامه خلاصه ای از این موارد را ذکر می کنیم.

۱- بیانیه اتحادیه اروپا با عنوان «دستورالعمل های اخلاقی برای هوش مصنوعی قابل اعتماد» در سال ۲۰۱۹
۲- بیانیه سازمان بهداشت جهانی WHO با عنوان «اخلاق و حکمرانی هوش مصنوعی برای سلامت» در سال ۲۰۲۱
۳- بیانیه یونسکو با عنوان «توصیه هایی برای اخلاق هوش مصنوعی» در سال ۲۰۲۱
۴- دستورالعمل وزارت بهداشت و خدمات انسانی کشور امریکا با عنوان «کتاب جامع استراتژی های هوش مصنوعی قابل اعتماد» در سال ۲۰۲۱
۵- بیانیه مرکز امنیت و فناوری های نوظهور کشور چین با عنوان « هوش مصنوعی قابل اعتماد» در سال ۲۰۲۱
۶- بیانیه دولت امارات متحده عربی با عنوان «اصول و رهنمودهای اخلاق هوش مصنوعی» در سال ۲۰۲۲
۷- بیانیه شورای تحقیقات پزشکی کشور هند با عنوان «دستورالعمل های اخلاقی برای کاربرد هوش مصنوعی در تحقیقات زیست پزشکی» در سال ۲۰۲۳

به تازگی خبرگزاری رویترز در گزارشی از اجلاس سران گروه هفت (G7) (شامل کشورهای کانادا، فرانسه، آلمان، ایتالیا، ژاپن، بریتانیا و ایالات متحده آمریکا به اضافه اتحادیه اروپا) در روز بیستم ماه می ۲۰۲۳ در توکیو نوشت: «رهبران کشورهای گروه هفت روز شنبه خواستار توسعه و پذیرش استانداردهای فنی برای "قابل اعتماد" نگه داشتن هوش مصنوعی شدند و بیان کردند که مدیریت این فناوری با رشد آن همگام نبوده است.»

هوش مصنوعی در ایران

بنابر فرمایش مقام معظم رهبری در سال ۱۳۹۹: « هوش مصنوعی را دریابید. هوش مصنوعی مساله ای مهم و آینده ساز است و در اداره آینده دنیا نقش دارد. باید به گونه ای عمل کنیم که ایران جزو ده کشور برتر هوش مصنوعی در دنیا قرار بگیرد.»، محققان و دانشمندان و همچنین سیاست گذاران ایرانی برآنند تا به این دستاورد در حوزه تحقیقات و صنعت دست یابند.

در سال جاری، جمهوری اسلامی ایران در حوزه تولید علم در هوش مصنوعی در جهان رتبه ۱۵ و در استفاده از هوش مصنوعی در زندگی رتبه ۷۷ را دارد.

بحث و نتیجه گیری

هوش مصنوعی در عصر جدیدی قرار دارد که بشر را با یک خیزش علمی رو به رو کرده است. این مساله می تواند نتایج مثبتی را در همه حوزه ها، خصوصا حوزه سلامت و علوم شناختی به همراه بیاورد. البته این مساله با چالش های زیادی رو به رو است که جامعه جهانی را نیز با نگرانی مواجه کرده است. با پیشرفت بی سابقه هوش مصنوعی در سال های اخیر، اکثر کشورهای توسعه یافته درصددند تا به صورت همزمان با مدیریت صحیح هوش مصنوعی با عنوان *هوش مصنوعی قابل اعتماد* از خطرات پیش بینی شده آن بکاهند. البته این حوزه، خود در لبه دانش قرار دارد و پروژه های تحقیقاتی زیادی در جهان در این حوزه در حال اجراست.

به منظور ارزیابی قابلیت اعتماد در سامانه های هوشمند لازم است هر یک از معیارهایی که برشمردیم کمی سازی گردد که هر یک شامل چالش های زیادی است. در نهایت لازم است به صورت هوشمند این معیارها با هم تجمیع گردد تا بتوانیم قابلیت اعتماد را در سیستم ها بررسی و تایید کنیم و همچنین روش ها و الگوریتم ها از این جنبه بسیار مهم مقایسه کنیم.

و من الله التوفیق

زهرا رضوانی کاشانی

زمستان ۱۴۰۲