فناوری همراه
fanavari hamrah
فصلنامه‌ی خبری تحلیلی
زمستان ۱۴۰۲
شماره‌ی دهم
رصد فناوری

مدیریت امنیت، ریسک و قابلیت اعتماد در هوش مصنوعی

امروزه از هوش مصنوعی در کاربردهای بسیار متنوعی همچون شهر هوشمند، سلامت هوشمند، تولیدات کارخانه‌ای هوشمند، دنیای مجازی و متاورس بهره‌گیری می‌شود. به‌طوری که گسترش استفاده از این ابزار نگرانی‌های جدی در زمینه ریسک‌ها، امنیت و قابلیت اطمینان در استفاده از آن ایجاد کرده است. با وجود توسعه روزافزون هوش‌مصنوعی نگرانی‌های ایجاد شده در استفاده اخلاقی از آن و نیز نگرانی‌های ناشی از عملکرد قابل فهم برای انسان مانع استفاده کامل از قابلیت‌های آن شده است. کاهش نگرانی‌ها و ایجاد اطمینان بیشتر جهت استقرار گسترده سامانه‌های مبتنی بر هوش مصنوعی نیازمند یک چارچوب قانونمند خواهد بود. یکی از سودمندترین راه‌کارهای تضمین قابلیت اطمینان و اعتماد به سیستم‌های هوش مصنوعی اتکای به چارچوب مدیریت امنیت، ریسک و اعتماد هوش مصنوعی (AI TRiSM) است. این چارچوب اگرچه به تازگی معرفی شده است، توانسته است در حوزه‌های مختلف نوآوری‌های کسب‌وکارها و جامعه بسیار موثر واقع شود.

کلمات کلیدی: هوش مصنوعی، مدیریت امنیت، قابلیت اعتماد، ریسک، چارچوب AI TRiSM، حملات متخاصم

سعید خشک‌دهان
کارشناسی ارشد
مهندسی فناوری
اطلاعات شبکه‌های
کامپیوتری از دانشگاه
صنعتی شریف

مقدمه

کاربردهای هوش مصنوعی تقریباً در تمام ابعاد زندگی ما قابل مشاهده و ردگیری هستند. استفاده از هوش مصنوعی در موتورهای پیشنهاد محصول، شهر هوشمند، آموزش، خودروهای بدون سرنشین و حتی کاربردهای آن در زمینه‌های حیاتی مانند سلامت هوشمند، اقتصاد، حمل‌ونقل و مخابرات مزایای چشم‌گیری به همراه داشته است [۱].

گسترش بی‌وقفه و همه‌جانبه استفاده از هوش مصنوعی نگرانی‌های گسترده‌ای در حوزه اعتماد، ریسک و امنیت به وجود آورده است. از سوی دیگر ایده اطمینان به هوش مصنوعی شامل تضمین استفاده صادقانه، اخلاقی و قابل اعتماد کاربران از آن نیز می‌باشد. در استفاده از هوش مصنوعی می‌بایست مراقب ریسک خروجی‌های اشتباه حاصل شده نیز باشیم. این ریسک‌ها شامل بایاس‌ها[1]، نتایج ناخواسته، تجاوز به حریم خصوصی و حتی آسیب ناشی از به‌کارگیری هوش مصنوعی است [۲] و [۳]. برای پاسخ به این چالش‌ها موسسه مشاوره‌ای گارتنر چارچوبی با عنوان مدیریت امنیت، ریسک و قابلیت اعتماد در هوش مصنوعی[2] یا به اختصار AI TRiSM معرفی نموده است. این چارچوب با هدف مدیریت چالش‌های مرتبط با به‌کارگیری هوش مصنوعی مانند استفاده منصفانه، حاکمیت بر آن، قابلیت اعتماد و حفظ حریم خصوصی ارائه شده و ارزیابی ساختاریافته قابلیت اطمینان سیستم‌های هوش مصنوعی را با ارزیابی میزان شفافیت، توضیح‌پذیری[3] و مسئولیت‌پذیری[4] آن‌ها بررسی می‌نماید. عنوان AI TRiSM اخیراً در یک مقاله منتشر شده در گارتنر به عنوان یکی از ۱۰ رویه اصلی فناوری که باید در سال ۲۰۲۴ مورد توجه قرار گیرد، معرفی شده است [۴]. در گارتنر AI TRiSM به صورت زیر تعریف شده است:

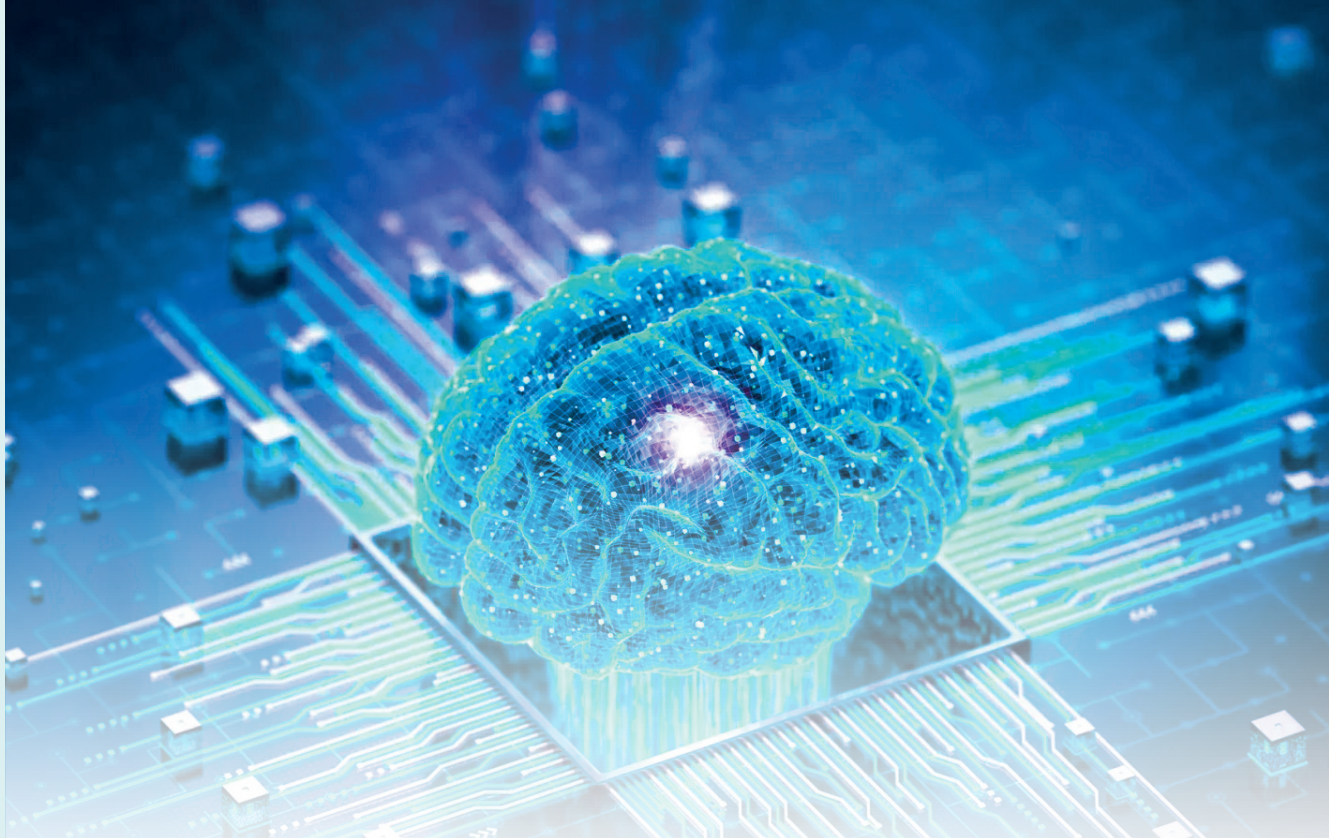
"چارچوب مدیریت اعتماد در هوش مصنوعی، ریسک و امنیت (AI TRiSM) می‌کوشد تا حفاظت از داده‌ها، بازدهی عملکرد، مقاومت، قابلیت اطمینان، انصاف، قابلیت اعتماد و حاکمیت بر مدل‌های هوش مصنوعی را تضمین نماید. این فرآیند شامل راه‌کارها و روش‌هایی برای تضمین قابلیت همکاری متقابل بین مدل‌ها و توصیف‌پذیری آن‌ها، حفاظت از داده‌های هوش مصنوعی، عملیات اختصاصی مدل هوش مصنوعی و مقاومت در برابر حملات خصمانه است" [۵].

در ادامه این گزارش ابتدا قابلیت اطمینان، ریسک‌ها و امنیت موجود در هوش مصنوعی مورد بررسی قرار گرفته و سپس چارچوب TRiSM در هوش مصنوعی و کاربردهای آن در توسعه سیستم‌های مبتنی بر هوش مصنوعی معرفی می‌گردد. چالش‌ها و جهت‌گیری‌های آتی AI TRiSM نیز در قسمت آخر مورد بررسی قرار خواهند گرفت.

امنیت، ریسک و قابلیت اطمینان در هوش مصنوعی

اعتماد و اطمینان در به‌کارگیری هوش مصنوعی: اعتماد یک المان

1- Biases
2- Artificial Intelligence Trust, Risk and Security Management
3- Explainability
4- Accountability



کلیدی برای یکپارچه‌سازی موفق هوش مصنوعی است و تاکنون تلاش‌های بسیاری برای شفاف‌سازی عملکرد، توصیف‌پذیری، انصاف و مسئولیت‌پذیری به منظور ایجاد اعتماد برای کاربران این ابزار و ذینفعان این حوزه صورت گرفته است. به صورت کلی اعتماد عبارت است از پذیرش استفاده از یک سیستم یا یک قابلیت ضمن پذیرش ریسک‌های احتمالی به منظور دریافت فایده مدنظر از یک کاربرد تعریف شده. اعتماد تاثیری چشم‌گیر در پذیرش یک ابزار مبتنی بر هوش مصنوعی خواهد داشت، زیرا بهره‌برداری بیشتر از یک پلتفرم جدید منجر به گسترش استفاده از آن و به همان نسبت افزایش بلوغ آن خواهد شد.

در منبع [۲] با عنوان ارتباط احساسی و اعتماد به هوش مصنوعی سعی شده است که نحوه برخورد کاربران هوش مصنوعی با کاربردهای هوش مصنوعی به لحاظ احساسی تحلیل شود و روش‌هایی برای افزایش میزان اعتماد کاربران به هوش مصنوعی پیشنهاد گردد. مثلاً اشاره شده است که در یک نظرسنجی تنها ۴۴% کاربران گفته‌اند که حاضر به سوار شدن به یک ماشین بدون راننده شرکت Uber هستند. یک نظرسنجی دیگر نشان داده است که ۴۲% مردم اعتماد عمومی به هوش مصنوعی ندارند و ۴۹% مردم نیز هیچ کاربرد هوش مصنوعی که بتوانند به آن اعتماد کنند در ذهن خود نداشته‌اند. دلایل مختلفی برای عدم اعتماد مردم به هوش مصنوعی وجود دارد که مهم‌ترین آن عدم فهم نحوه عملکرد و تصمیم‌گیری هوش مصنوعی است. Explainable AI تلاشی برای رفع این مشکل بوده است که کوشیده است نحوه تصمیم‌گیری هوش مصنوعی را به شکلی قابل فهم برای عموم مردم ارائه نماید. دلیل دیگر ترس از هوش مصنوعی ترس از عواقب به‌کارگیری هوش مصنوعی مانند ترس از بی‌کارشدن یا ترس از ربات‌های هوش‌مصنوعی خودکار و قاتل است. در این راستا، نمایش درست مزایا و نتایج مطلوب و باکیفیت کاربردهای هوش مصنوعی به کاربران می‌تواند ابزاری برای ایجاد احساس خوب برای ارتباط کاربر با هوش مصنوعی باشد و عدم اعتماد کاربر به هوش مصنوعی را کاهش دهد.

از نگاه کاربر، جانب‌داری[5]، تبعیض[6] و تجاوز به حریم خصوصی فاکتورهای اصلی عدم اعتماد به هوش مصنوعی هستند. در ادامه به توضیح هر یک از این موارد می‌پردازیم.

جانب‌داری و تبعیض: راه‌کارهای هوش مصنوعی الگوریتم‌های پیش‌بینی‌کننده رفتار پیچیده انسان‌ها را ارائه می‌کنند که بعضاً توانایی تحلیل داده با دقتی بیشتر از انسان نیز دارا است. همچنین هوش مصنوعی نشان داده است که الگوریتم‌های پیش‌بینی‌کننده رفتار آن قادرند از جانب‌داری که رفتاری خاص نوع بشر است دوری کرده و به عنوان مثال بی‌طرفی و عینیت در قضاوت رفتار مجرمان را فراهم نماید. در سوی دیگر ماجرا، اگر از داده‌های جانب‌دارانه و تبعیضی برای آموزش مدل‌های هوش مصنوعی استفاده شود، به‌کارگیری هوش مصنوعی در امر قضاوت می‌تواند یکپارچگی قضاوت را دچار خدشه نموده و منجر به رخدادهای ناخواسته و نامطلوب مانند جانب‌داری برنامه‌ریزی شده، تبعیض نژادی و حتی افزایش نرخ زندانیان در یک سیستم قضایی گردد [۶]. به‌عنوان نمونه در سال ۲۰۱۷ آمازون سیستم مصاحبه و استخدام نیروی مبتنی بر هوش مصنوعی خود را به دلیل جانب‌داری جنسیتی که در استخدام آقایان نسبت به خانم‌ها به جهت کمتر بودن میزان داده‌های آموزشی کسب‌شده صورت می‌داد، متوقف کرد. بنابراین، وقوع نتایج جانب‌دارانه در تصمیم‌گیری‌های مبتنی بر هوش‌مصنوعی هر چند به صورت تصادفی یا به جهت موضوعات سیستماتیک، میزان اعتماد به آن را به‌شدت کاهش خواهد داد.

تجاوز به حریم خصوصی کاربران: باتوجه به اینکه در سیستم‌های هوش مصنوعی به صورت گسترده از داده‌ها به‌منظور یادگیری و بهبود عملکرد استفاده می‌شود، در صورت استفاده

5- Bias
6- Discrimination

نادرست از داده‌های حساس کاربران حریم خصوصی ایشان آسیب خواهد دید. به‌عبارتی ممکن است بدون اینکه خود اشخاص متوجه شوند، داده‌های خصوصی ایشان برای اتخاذ تصمیماتی در مورد ایشان استفاده شود. با توجه به اینکه استفاده از هوش مصنوعی می‌تواند نتایج مثبت بسیاری در مراقبت‌های ویژه پزشکی، مدیریت انتخاب و هدایت افراد در سیستم‌های مالی و حمل‌ونقل داشته باشد و با توجه به اینکه تصمیم‌گیری در بسیاری موارد به مرور به هوش مصنوعی واگذار می‌گردد، تضمین اعتماد کاربران به هوش مصنوعی اهمیت ویژه‌ای خواهد داشت و مستلزم تنظیم قوانین رگولاتوری دقیق به‌منظور نظارت بر توسعه اخلاقی سامانه‌های هوش مصنوعی و استفاده اخلاقی از این سامانه‌ها می‌باشد.

علاوه بر چالش‌های عنوان شده جهت اعتماد به هوش مصنوعی، ریسک‌ها و چالش‌های امنیتی نیز در استفاده از هوش مصنوعی وجود دارند که می‌بایست مورد توجه قرار گرفته و بررسی شوند. در ادامه به معرفی برخی از آن‌ها خواهیم پرداخت.

تفهیم ریسک‌های هوش مصنوعی: در به‌کارگیری هوش مصنوعی با وجود مزایا، ریسک‌های مختلفی نیز وجود دارند. به‌عنوان مثال، مدیریت و هدایت افکار عمومی جامعه از ریسک‌های مهمی است که اخیراً در حوزه به‌کارگیری هوش مصنوعی مورد توجه قرار گرفته است. حضور و تاثیر رسانه‌های اجتماعی در جوامع کنونی در حوزه سرگرمی، انتشار اطلاعات و هدایت افکار عمومی و توسعه کسب‌وکارها غیرقابل انکار است. استفاده از هوش مصنوعی در رسانه‌های اجتماعی نگرانی‌های عمیق در زمینه مدیریت افکار عمومی و تجاوز به حریم خصوصی کاربران و زورگیری‌های سایبری ایجاد نموده است. موتورهای جستجو می‌کوشند با نمایش نتایج اختصاصی متناسب با علایق و نیازمندی‌های کاربران تجربه کاربری بهتری ایجاد نمایند. هر چند که این موتورها به‌جهت فیلترکردن اطلاعات نمایشی، جانب‌داری و محدود کردن تنوع اخبار نمایشی مورد انتقاد هستند.

فناوری DeepFake که به کمک آن کاربر می‌کوشد صوت، تصویر یا ویدئوهایی تولید نماید که علی‌رغم ساختگی بودن به کاربر حس اصالت و اعتبار می‌دهد، از دیگر چالش‌های بزرگ به‌کارگیری هوش مصنوعی است.

سیستم‌های تسلیحات خودکار مرگبار (LAWS)[7] ابزارهایی هستند که برای حمله به یک شخص یا گروهی از افراد توسط هوش مصنوعی بدون دخالت مستقیم انسان‌ها با استفاده از ارائه‌ای از حسگرها و الگوریتم‌های کامپیوتری استفاده می‌شوند. پیدایش مفهوم LAWS سبب شده که توسعه سیستم‌های هوش مصنوعی با چالش‌های جدی در استفاده اخلاقی از آن روبرو شود.

تفهیم امنیت هوش مصنوعی: با توسعه کاربردهای هوش مصنوعی با توجه به دسترسی آن‌ها به داده‌هایی بسیار حساس تضمین امنیت آن‌ها اهمیت ویژه‌ای می‌یابد. مدیریت امنیت هوش مصنوعی شامل توسعه تمرینات و ابزارهایی برای حفاظت سیستم‌های هوش مصنوعی و داده‌های تحت پردازش آن‌ها در برابر فعالیت‌های مخرب، شکاف‌های امنیتی و دسترسی‌های غیرقانونی است. هوش مصنوعی همان‌قدر که می‌تواند برای شناسایی و مقابله با تهدیدات امنیتی استفاده شود، می‌تواند برای کاربردهای مخرب امنیتی و نفوذ، دسترسی و آسیب به داده‌های خصوصی افراد استفاده شود. هوش مصنوعی به کمک ابزارهایی مانند DeepFake می‌تواند برای آسیب‌هایی استفاده شود که

7- Lethal Autonomous Weapons Systems

امنیت فیزیکی، دیجیتال و سیاسی را به‌خطر بیندازد. استفاده از هوش مصنوعی برای تقویت سیستم‌های امنیت سایبری گرچه از حجم عظیم حملات سنتی کاسته است، اما مهاجمان مخرب نیز می‌توانند با سوءاستفاده از نقاط ضعف الگوریتم‌های هوش مصنوعی نتایج مطلوب را تغییر دهند که این موضوع می‌تواند زندگی گروه زیادی از افراد را متاثر نماید. استفاده از اصول برنامه‌نویسی طراحی مبتنی بر حریم خصوصی و استفاده از روش‌های ناشناس‌سازی می‌تواند به حفظ داده‌های حساس کاربران کمک نماید.

چارچوب AI TRiSM

تا اینجای مقاله درخصوص اهمیت اعتماد، امنیت و ریسک در کاربردهای هوش مصنوعی صحبت شد. دیدیم که توسعه سامانه‌های مبتنی بر هوش مصنوعی مستلزم افزایش اعتماد به آن است. تضمین انطباق با قوانین حریم خصوصی نیازمند انطباق با GDPR[8] است. چارچوب‌های جاری برای مدیریت امنیت، ریسک و اعتماد به هوش مصنوعی مستقل از هم و جداگانه عمل می‌کنند و فاقد انسجام و همسویی لازم هستند. این چارچوب‌های مستقل و مجزا اغلب همکاری هماهنگ ندارند. از این‌رو استراتژی تکه‌تکه شده‌ای برای مدیریت هوش مصنوعی ایجاد می‌کنند و سبب می‌شود اعتماد به هوش مصنوعی در مواقعی بدون درک کامل ریسک‌ها و مشکلات امنیتی آن ایجاد شود. چارچوب AI TRiSM می‌کوشد پلی برای یکپارچه‌سازی چارچوب‌های مستقل مدیریت ریسک هوش مصنوعی باشد. این چارچوب با ترکیب اجزای کلیدی سایر چارچوب‌ها، هم‌افزایی و همکاری متقابل در ابعاد مختلف مدیریت و حاکمیت بر هوش مصنوعی ایجاد خواهد کرد. این چارچوب می‌کوشد اعتماد، ریسک و امنیت هوش مصنوعی را در کل زنجیره حیات سامانه‌های هوش مصنوعی شامل طراحی، توسعه، استقرار و عملیات آن‌ها مدنظر قرار دهد. این چارچوب همچنین به کسب‌وکارها در پیاده‌سازی استراتژی‌های هوش مصنوعی که با اهداف و ارزش‌های ایشان یکسو باشد کمک می‌کند. شکل ۱ چهار اصل پایه AI TRiSM را نشان می‌دهد که شامل نظارت بر مدل، عملیات مدل (ModelOps)، امنیت کاربردهای هوش مصنوعی و حریم خصوصی مدل است. در ادامه، به توضیح هر یک از این اصول می‌پردازیم.

نظارت بر مدل: کاربران هوش مصنوعی اغلب در اعتماد به آن خصوصاً زمانی که تصمیمات اتخاذی توسط آن بسیار پیچیده است و به‌سادگی قابل فهم نیست دچار مشکل هستند. به همین جهت اهمیت ویژه‌ای دارد که در کاربردهای واقعی مدل‌های هوش مصنوعی، انصاف، مسئولیت‌پذیری، شفافیت و حاکمیت

8- General Data Protection Regulation
(قانون مصوب اتحادیه اروپا که به منظور حفظ حریم خصوصی داده‌های کاربران در اتحادیه اروپا تنظیم و تصویب شده است)

شکل ۱- چهار ستون اصلی چارچوب AI TRiSM به منظور مدیریت اعتماد، ریسک و امنیت سیستم‌های هوش مصنوعی [۱]

شکل ۲- تجسم ارتباط ویژگی‌های متصل به هم در ایجاد یک زنجیره ModelOps موثر [۱]

کاربران در دستور کار قرار گرفته شود. لذا نظارت بر مدل و تلاش برای توصیف‌پذیری مدل‌ها به هوش مصنوعی کمک می‌کند که جانب‌دارانه عمل نکند. با پیاده‌سازی توصیف‌پذیری و نظارت بر مدل‌های هوش مصنوعی می‌توان اطمینان حاصل کرد که این مدل‌ها درست و بدون جانب‌داری عمل می‌کنند. نظارت بر عملکرد هوش مصنوعی مستلزم فهم کامل نحوه عملکرد مدل‌های هوش مصنوعی، نحوه اتخاذ تصمیمات قابل دفاع توسط آن‌ها و ترویج شفافیت در عملکرد آن‌ها است.

عملیات مدل‌های هوش مصنوعی: با وجود قابلیت‌های شگرف مشاهده شده در هوش مصنوعی یکپارچه‌سازی آن در کسب‌وکارها به جهت نبود ابزارهای مناسب و متدلوژی‌های تسهیل‌گر زنجیره حیات توسعه راه‌کارهای هوش مصنوعی، هنوز در مراحل ابتدایی است. وظایف مهمی مانند آماده‌سازی داده، طراحی مدل‌ها و آموزش آن‌ها، توسعه نرم‌افزار، تضمین کیفیت، استقرار، نظارت، بازخورد و تضمین قابلیت بازتولید و قابلیت ممیزی بخشی از زنجیره حیات توسعه نرم‌افزارهای هوش مصنوعی هستند. از این رو، یک بخش اساسی چارچوب AI TRiSM فرآیند ModelOps است که عملیاتی شدن مدل‌های هوش مصنوعی نظیر مدیریت زنجیره حیات، حاکمیت مدل‌ها و همچنین مسئولیت مدیریت زیرساخت‌های حیاتی و محیطی را مورد نظر دارد تا به واسطه آن بتوان عملکرد بهینه مدل‌ها را تضمین نمود. شکل ۲ یک فرآیند جامع ModelOps را که شامل مراحل کلیدی آن می‌باشد، نشان می‌دهد.

کاربردهای امنیت هوش مصنوعی: سامانه‌های امنیت مبتنی بر هوش مصنوعی با تحلیل حجم عظیم داده قادرند به کمک الگوریتم‌های پیچیده یادگیری ماشین نقاط ضعف، دسترسی‌های غیرمجاز و رفتارهای مخرب را شناسایی نمایند. در چارچوب AI-TRiSM امنیت داده اهمیتی ویژه در نظارت بر سلامت و موضوعات اقتصادی دارد. همچنین در AI-TRiSM چارچوب‌های حفاظت داده مشابه استفاده از داده‌های مصنوعی[9] [۷]، حریم خصوصی تفاضلی[10]، به همراه استفاده از پروتکل‌هایی مانند رمزنگاری هومومورفیک کامل (FHE[11]) و محاسبات چندعاملی امن (SMPC[12]) به‌کار گرفته می‌شوند که برای حفاظت از حجم عظیم داده موردنیاز هوش مصنوعی، تضمین اعتماد، امنیت و کاهش ریسک سیستم‌های هوش مصنوعی حیاتی هستند.

حریم خصوصی داده‌ها: به این معناست که سیستم‌های هوش مصنوعی داده‌های حساس و خصوصی را با رعایت الزامات حقوقی رگولاتوری پردازش می‌کنند. این کار مستلزم اخذ تاییدیه‌های لازم از صاحبان داده، استفاده از روش‌های ناشناس‌سازی و استفاده از رویه‌های مدیریت داده امن به منظور حفاظت از حریم خصوصی کاربران است. روش‌های حریم خصوصی تفاضلی با افزودن نویز یا تصادفی‌سازی در یک دیتاست، می‌توانند برای حفاظت حریم خصوصی کاربران در یک دیتاست به کار گرفته شوند. حفاظت از حریم خصوصی امری ضروری برای تضمین استفاده اخلاقی از هوش مصنوعی و تضمین استقرار قابل‌اعتماد و مسئولانه هوش مصنوعی خصوصاً در کاربردهایی مانند مراقبت از سلامت و تبادلات مالی و بانکی است.

چالش‌های AI-TRiSM

چنانکه دیدیم توسعه کاربردهای هوش مصنوعی نیازمند توسعه یک چارچوب دقیق برای ایجاد اطمینان، مدیریت ریسک و امنیت است، لیکن استقرار کامل AI-TRiSM امروزه با چالش‌هایی روبرو است. از جمله آن که AI-TRiSM خود نیازمند قابلیت اطمینان و پیشرفت ویژه در سامانه‌های هوش مصنوعی است.

نظارت بر مدل‌های هوش مصنوعی با چالش‌هایی از قبیل (و نه صرفاً محدود به) رانش داده[13] روبه‌رو است. مدل‌های هوش مصنوعی براساس داده‌های آموزشی که دریافت می‌کنند

9- Synthetic data
Synthetic data (داده‌های مصنوعی به‌واسطه مدل‌های الگوریتمی تولیدکننده داده

10- Differential privacy
11- Fully Homomorphic Encryption
12- Secure MultiParty Computation
13- Data Drift
(مدل‌های هوش‌مصنوعی وابسته به داده‌های آموزشی ورودی هستند و اگر مشخصات داده‌های ورودی در گذر زمان تغییر کند، پدیده رانش داده اتفاق می‌افتد که به جهت

عمل می‌کنند. تغییر این داده‌هـــا در گذر زمان می‌تواند منجر به رانش‌داده شود که برای رفع آن نظارت مستمر و مکرر در بازه‌های زمانی مختلف نیاز اســـت. از طرف دیگـــر دوری از جانب‌داری و رعایت انصاف در مدل‌هـــای هوش مصنوعی خـــود در معرض آسیب جانب‌داری اســـت! لذا انتخاب معیارهای مناسب جهت رعایت انصاف و ارزیابی مکرر معیارها ضروری خواهد بود. مدیریت این چالش‌ها مســـتلزم اســـتفاده ترکیبـــی از روش‌هایی مانند گردآوری داده‌های جاری، نظارت بر خطوط لوله[۱۴]، نسخه‌بندی مدل‌ها، هشدارهای خودکار و تنظیم چرخه‌های بازخورد شامل متخصصان انسانی است.

انطباق بـــا قوانیـــن رگولاتـــوری ماننـــد GDPR بـــا توجه به پیچیدگی‌های آن‌ها خصوصاً برای شرکت‌های کوچک‌تر پرهزینه و دشوار خواهد بود. مهاجمان اغلب می‌کوشـــند از نقاط ضعف سامانه‌های هوش مصنوعی سوءاستفاده کنند. تضمین استحکام مدل‌های هوش مصنوعی علیه چنین حملاتی چالش‌برانگیز است. برای مثال مهاجم ممکن است بکوشد الگوریتم‌های پیش‌بینی مدل را فریب داده یا تمهیدات امنیتی را دور بزند که منجر به نتایج مثبت یا منفی غلط در شناسایی ریسک‌ها خواهد شد. با توجه به اینکه استقرار کاربردهای هوش مصنوعی وابسته به دقت بالای داده‌های آموزشـــی ورودی است، دســـتکاری عمدی داده‌های آموزشی ورودی یا تزریق نمونه‌های مخرب که سیستم را از هدف اصلی آن به سمتی نامشخص هدایت می‌کند، از دیگر چالش‌های به‌کارگیری این چارچوب است. فاصله دانش تخصصی تیم مجری این چارچوب با پیشرفت‌های بسیار سریع حوزه هوش‌مصنوعی یک چالش بسیار بزرگ در حوزه استقرار AI-TRiSM است و این فاصله با توجه به سرعت تغییر چشم‌انداز حملات و تهدیدات هر روز بزرگ‌تر و پیچیده‌تر می‌گردد.

نتیجه‌گیری

AI-TRiSM چارچوبی است که توسط گارتنر به منظور استقرار امن و قابل‌اطمینان سامانه‌های هوش مصنوعی معرفی شده است. اهمیت سامانه‌های هوش مصنوعی در سال‌های اخیر باتوجه به سرعت بالای توسعه آن‌ها و نیاز روزافزون افراد و کسب‌وکارها به آن افزایش یافته اســـت. AI-TRiSM چارچوبی با نقشی حیاتی برای تضمین قانون‌گذاری مناسب جهت استقرار مدل‌های هوش مصنوعی و مدیریت موثر ریسک‌های احتمالی است. با به‌کارگیری AI-TRiSM ســـازمان‌ها می‌توانند فهم ارزشـــمندی از طراحی، توسعه و توزیع مدل‌های هوش مصنوعی به‌دست آورند که آن‌ها را ضمن حفظ اعتبار و قابلیت اطمینان قادر به نظارت موثر و کاهش ریسک‌های احتمالی می‌سازد. چارچوب AI-TRiSM می‌کوشد اعتماد، ریسک و امنیت هوش مصنوعی را در طول زنجیره حیات ســـامانه‌های هوش مصنوعی شامل طراحی، توســـعه، استقرار و عملیات آن‌ها مدنظر قرار دهد. در ایـــن مقاله، چهار اصل پایه AI-TRiSM شامل نظارت بر مدل، عملیات مدل (ModelOps)، امنیت کاربردهای هوش مصنوعی و حریم خصوصی مدل توصیف گردیده و اهمیت استفاده از آن‌ها مطرح شد. البته باوجود اهمیت ویژه این چارچوب استقرار آن با چالش‌هایی روبه‌رو است که حل بخشی از آن‌ها نیز خود نیازمند توســـعه بیشتر هوش مصنوعی است. در انتهای این گزارش برخی از چالش‌های توسعه چارچوب AI-TRiSM نیز معرفی گردید.

منابع:

[1] M. K. A. a. M. A. A. Adib Habbal a, "Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions," Expert Systems With Applications Elsevior, pp. 1-14, 2024.

[2] T. A. M. S. B. S. K. R. B. D. R. S. Omri Gillath, "Attachment and trust in artificial intelligence," Computers in Human Behavior En Elsevior Journal 2021.

[3] J. Bharadiya, " Artificial Intelligence in Transportation Systems A Critical Review," American Journal of Computing and Engineering, p. 34–45, 2023.

[4] B. Willemsen, "Gartner Top 10 Strategic Technology Trends 2024," Gartner Magzine , NY, 2023.

[5] I. Gartner, "Definition of AI TRiSM, Gartner Information Technology Glossary,". Available: https://www.gartner.com/en/information-technology/glossary/ai-trism2024.

[6] M. A. Malek, "Criminal courts' artificial intelligence: The way it reinforces bias and discrimination," AI and Ethics, p. 233–245, 2022.

[7] A. B. &. et., "Synthetic data protection: Towards a paradigm change in data regulation?," Big Data & Society, https://doi.org/10.1177/20539517241231277, 2024.