هوش مصنوعی و امنیت سایبری

تاریخ انتشار: آذر ۱۴۰۴

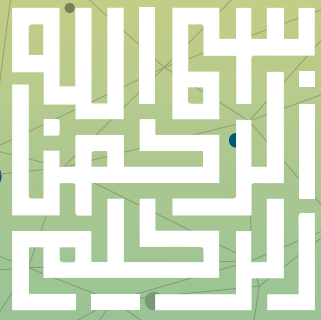
ITRC

پژوهشگاه ارتباطات
و فناوری اطلاعات
(مرکز تحقیقات مخابرات ایران)

هوش مصنوعی و امنیت سایبری

تاریخ انتشار: آذر ۱۴۰۴

ITRC
پژوهشگاه ارتباطات
و فناوری اطلاعات
(مرکز تحقیقات مخابرات ایران)



عنوان گزارش: هوش مصنوعی و امنیت سایبری

کلمات کلیدی: هوش مصنوعی، امنیت سایبری، ریسک‌های امنیت سایبری، مدیریت ریسک

تهیه کنندگان: مهدی عزیزی مهماندوست، مسعود جمشیدی‌ها

گروه پژوهشی:

تاریخ نشر: آذر ۱۴۰۴

ITRC
پژوهشگاه ارتباطات
و فناوری اطلاعات
(مرکز تحقیقات مخابرات ایران)

خلاصه مدیریتی

هوش مصنوعی (AI) در حال تبدیل‌شدن به نیرویی محوری برای پیشبرد نوآوری در بخش‌های مختلف از جمله خدمات درمانی، مالی، حمل‌ونقل و تولید است. بااین‌حال، با ادغام روزافزون هوش مصنوعی در سامانه‌ها و زیرساخت‌های حیاتی، ریسک‌های قابل‌توجهی در زمینه امنیت سایبری ایجاد می‌شود که باید به طور اثربخش مدیریت شوند. این گزارش تحلیل جامعی از این ریسک‌ها و راهنمایی راهبردی برای توسعه سیاست‌های استوار به‌منظور کاهش این ریسک‌ها در منطقه آسیا و اقیانوسیه ارائه می‌کند.

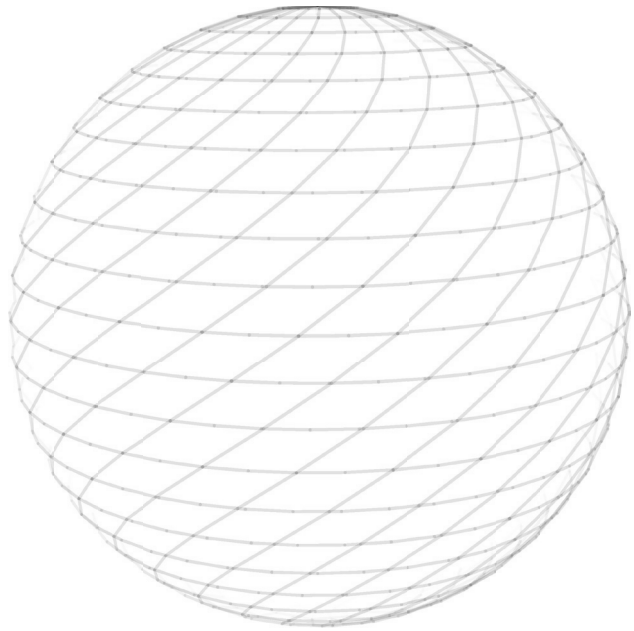
این گزارش با بررسی چشم‌انداز فعلی هوش مصنوعی آغاز می‌شود و سامانه‌های هوش مصنوعی را به دو دسته پیش‌بینی‌کننده و مولد تقسیم می‌کند و کاربردهای آن‌ها را در بخش‌های کلیدی توضیح می‌دهد. علی‌رغم اینکه هوش مصنوعی ظرفیت ایجاد پیشرفت‌های دگرگون‌کننده‌ای را دارد، اما آسیب‌پذیری‌های منحصربه‌فردی را نیز به همراه دارد؛ مانند مسمومیت داده‌ها، حملات فرار از مدل و نگرانی‌های اخلاقی که می‌تواند یکپارچگی و امنیت سامانه‌های هوش مصنوعی را به خطر بیندازد. در اینجا برای مقابله با این چالش‌ها، یک چارچوب جامع امنیت سایبری هوش مصنوعی پیشنهاد شده است. این چارچوب شامل تعدادی مؤلفه اصلی مانند نظارت، مدیریت چرخه عمر، امنیت مدل، حاکمیت داده، شفافیت و راهبردهای پاسخ به حوادث می‌شود. این چارچوب به‌گونه‌ای طراحی شده است که بدون مشکل با قوانین موجود و استانداردهای شناخته‌شده بین‌المللی یکپارچه شود و اطمینان ایجاد کند که رویکرد منسجم و مؤثری برای حاکمیت هوش مصنوعی در سراسر منطقه آسیا و اقیانوسیه حاصل می‌شود.

این توصیه‌های سیاستی بر لزوم به‌روزرسانی راهبردهای ملی امنیت سایبری، تدوین دستورالعمل‌های مختص به هوش مصنوعی، سرمایه‌گذاری در تحقیق‌وتوسعه، ترویج همکاری‌های بین‌المللی و ارتقای سواد در زمینه هوش مصنوعی تأکید دارد. این اقدامات برای اطمینان از این که استقرار سامانه‌های هوش مصنوعی هم امن و هم اخلاقی باشد، ضروری هستند و از نوآوری حمایت می‌نمایند و در عین حال به‌صورت پیش‌نگرانه به تهدیدات نوظهور رسیدگی می‌کنند.

اهمیت تلاش هماهنگ بین دولت‌ها، صنایع و دانشگاه‌ها برای توسعه و پیاده‌سازی سیاست‌های امنیت سایبری استوار برای هوش مصنوعی غیرقابل‌چشم‌پوشی است. چنین سیاست‌هایی برای

حافظت در مقابل ریسک‌های فعلی و همچنین پیش‌بینی و کاهش چالش‌های آینده ضروری هستند و اطمینان حاصل می‌کنند که هوش مصنوعی همچنان نیروی محرکه‌ای برای تغییرات مثبت در این منطقه باقی می‌ماند.

فهرست مطالب



ITRC
پژوهشگاه ارتباطات
و فناوری اطلاعات
(مرکز تحقیقات مخابرات ایران)

۱ مقدمه

هوش مصنوعی به‌عنوان یکی از دگرگون‌کننده‌ترین فناوری‌های عصر حاضر ظهوریافته است که توانایی ایجاد انقلابی در صنایع، تغییر شکل‌دادن جوامع و پیشبرد نوآوری بی‌سابقه را در سطح جهانی دارد. از خدمات درمانی و مالی گرفته تا حمل‌ونقل و انرژی، هوش مصنوعی برای حل مسائل پیچیده، بهینه‌سازی فرایندها و ایجاد فرصت‌های جدید برای رشد و توسعه به کار گرفته می‌شود. پیشرفت‌های سریع در فناوری‌های هوش مصنوعی، مانند یادگیری ماشینی، یادگیری عمیق و پردازش زبان طبیعی و اخیراً ابزارهای هوش مصنوعی مولد که برای استفاده مصرف‌کنندگان در دسترس قرار گرفته‌اند، باعث شده است توسعه سامانه‌های هوشمندی امکان‌پذیر شود که قادرند کارهایی را انجام دهند که پیش از این منحصراً در حیطه هوش انسانی بوده است.

بااین‌حال، با جاافتادن روزافزون هوش مصنوعی در زندگی روزمره و زیرساخت‌ها، ضروری است که ریسک‌های امنیت سایبری مرتبط با این فناوری شناسایی و رسیدگی شوند. همان ویژگی‌هایی که هوش مصنوعی را تا این حد دگرگون‌کننده می‌سازد (توانایی آن در یادگیری، تطبیق و اتخاذ تصمیمات به‌صورت خودمختار) در عین حال آسیب‌پذیری‌ها، سطح‌های حمله و قابلیت‌های تهاجمی جدیدی را به وجود می‌آورد که بازیگران مخرب می‌توانند از آن‌ها بهره‌برداری کنند. این ریسک‌ها با افزایش پیچیدگی و عدم شفافیت سامانه‌های هوش مصنوعی تشدید می‌شوند که می‌تواند شناسایی و کاهش نقض‌های امنیتی را دشوار کند.

پذیرش سریع هوش مصنوعی باعث برجسته‌شدن نیاز به سیاست‌های جامع و همسو برای مقابله با ریسک‌های احتمالی امنیت سایبری و اطمینان از استقرار امن، ایمن و اخلاقی سامانه‌های هوش مصنوعی شده است [۱]. دولت‌ها، صنعت و دانشگاه‌ها باید با یکدیگر همکاری کنند تا یک چارچوب حاکمیت قوی را ایجاد کنند که چالش‌های منحصربه‌فرد ناشی از هوش مصنوعی را بررسی نماید و در عین حال مروج نوآوری و ارتقادهنده توسعه و استقرار مسئولانه این فناوری دگرگون‌کننده باشد.

هدف این گزارش ارائه درک جامعی نسبت به ریسک‌های امنیت سایبری هوش مصنوعی به‌عنوان ابزاری آگاهی‌بخش برای کمک به سیاست‌گذاری مؤثر است. این گزارش نمایی کلی از وضعیت فعلی هوش مصنوعی، از جمله فناوری‌ها، کاربردها و استانداردهای کلیدی را ارائه می‌دهد و توصیه‌های

سیاستی به دولت‌ها و نهادهای تنظیم‌کننده مقررات برای تقویت و همسوسازی سیاست‌ها به‌منظور استقرار امن سامانه‌های هوش مصنوعی در منطقه آسیا و اقیانوسیه ارائه می‌دهد.

۲ چشم‌انداز فعلی هوش مصنوعی

۲-۱ شرحی کلی از فناوری‌های هوش مصنوعی و کاربردهای آن‌ها

در حالی که اصطلاح «هوش مصنوعی» به یک اصطلاح رایج تبدیل شده است، هنوز یک تعریف جهانی و مورد توافق برای آن وجود ندارد. به‌عنوان‌مثال، مرکز امنیت سایبری ملی بریتانیا (NCSC) هوش مصنوعی را «هر سامانه رایانه‌ای قادر به انجام کارهایی که معمولاً نیاز به هوش انسانی دارند، مانند ادراک بصری، تولید متن، بازشناسی گفتار یا ترجمه بین زبان‌ها» [۲] تعریف می‌کند. این تعریف نشان‌دهنده ماهیت وسیع و متنوع فناوری‌های هوش مصنوعی است. در مقابل، تعریفی که انجمن پیشبرد هوش مصنوعی (AAAI) ارائه می‌دهد، هوش مصنوعی را «درک علمی از سازوکارهای زیربنایی تفکر و رفتار هوشمند و مجسم شدن آن‌ها در ماشین‌ها» [۳] توصیف می‌کند. این تعریف بر جنبه‌های علمی و شناختی هوش مصنوعی تأکید دارد و بر درک و تکرار هوش انسانی در ماشین‌ها متمرکز است.

برای مقاصد این گزارش سیاست‌محور، از تعریف ارائه‌شده در ماده ۳-۱ قانون هوش مصنوعی اتحادیه اروپا به‌عنوان مرجع اصلی استفاده خواهد شد. طبق این تعریف، سامانه هوش مصنوعی «نرم‌افزارهایی توسعه‌یافته با استفاده از یک یا چند فن و رویکرد که قادرند برای مجموعه‌ای مشخص از اهداف تعریف‌شده توسط انسان، خروجی‌هایی مانند محتوا، پیش‌بینی، توصیه یا تصمیماتی ایجاد کنند که بر محیط‌هایی که با آن تعامل دارند تأثیر می‌گذارند» [۴]. در اینجا این تعریف پذیرفته شده است، زیرا ماهیت چندوجهی سامانه‌های هوش مصنوعی را در برمی‌گیرد و در بحث سیاست‌گذاری امنیت سایبری، اساسی را برای درک شاخه‌های متنوع هوش مصنوعی و کاربردهای مرتبط آن‌ها فراهم می‌کند. این درک، همان‌طور که در ادامه توضیح داده خواهد شد، برای مسیریابی در میان استانداردها و سیاست‌های هوش مصنوعی مفید خواهد بود.

۲-۱-۱ مراحل توسعه هوش مصنوعی

اگرچه حوزه هوش مصنوعی در سال‌های اخیر تحولات سریعی داشته است، مفهوم کلی آن در سال ۱۹۵۶ در کارگاه پروژه تحقیقاتی تابستانی هوش مصنوعی در دارتموث توسعه یافت [۵]. در طول سالیان، توسعه و کاربرد هوش مصنوعی با پیشرفت‌ها و موانع قابل‌توجهی مواجه شده است. برخی از فنون هوش مصنوعی، مانند یادگیری ماشین، در دهه ۱۹۹۰ به‌صورت عملی درآمدند و در اواخر دهه ۲۰۰۰ و اوایل دهه ۲۰۱۰ با رشد انفجاری داده و قدرت رایانشی رونق گرفتند [۶]. هوش مصنوعی

مولد در سال ۲۰۱۴ توجهات زیادی را جلب کرد و در سال ۲۰۲۰ با معرفی ChatGPT 3.0 توسط شرکت OpenAI توجه عمومی زیادی را به خود جلب کرد [۷]. با وجود این پیشرفت‌ها، بسیاری از کارشناسان معتقدند که توسعه هوش مصنوعی هنوز در مراحل اولیه قرار دارد. توانمندی هوش مصنوعی به دو مرحله توسعه تقسیم می‌شود:

- **هوش مصنوعی محدود (ANI)**[1]: سامانه‌های هوش مصنوعی محدود، وظایف خاص یا بازه محدودی از وظایف را انجام می‌دهند. از نمونه‌های هوش مصنوعی محدود، می‌توان به دستیارهای شخصی مجازی و سامانه‌های توصیه‌گر اشاره کرد. بیشتر قابلیت‌های هوش مصنوعی فعلی در این دسته قرار دارند [۸].
- **هوش جامع مصنوعی (AGI)**[2]: که به آن هوش مصنوعی قوی نیز گفته می‌شود، هوش جامع مصنوعی ظرفیت نرم‌افزارهایی را با هوش شبیه به انسان، قابلیت‌های خودآموزی و مجموعه‌ای گسترده از کارکردهای انطباق‌پذیر دارد. به‌این‌ترتیب، انتظار می‌رود که نرم‌افزارهای هوش جامع مصنوعی قادر به انجام وظایف جدیدی باشند که برای آن آموزش ندیده‌اند. البته در حال حاضر هوش جامع مصنوعی یک فناوری فرضی است که در خط مقدم تحقیقات قرار دارد.

۲-۱-۲ دو دسته سامانه‌های هوش مصنوعی: پیش‌بینی‌کننده و مولد

هوش مصنوعی، به‌ویژه در معنای محدود فعلی، به‌طورکلی به دو دسته تقسیم می‌شود: هوش مصنوعی پیش‌بینی‌کننده و مولد. درک این تمایزها برای سیاست‌گذاران جهت رسیدگی اثربخش به چالش‌ها و فرصت‌های منحصربه‌فردی که هر دسته ایجاد می‌کند، ضروری است.

سامانه‌های هوش مصنوعی پیش‌بینی‌کننده برای تحلیل داده‌ها و پیش‌بینی رویدادهای آینده طراحی شده‌اند. این سامانه‌ها در زمینه‌های مختلفی مانند پیش‌بینی بازار در دانش مالی، پیش‌بینی شیوع بیماری در خدمات درمانی، بهینه‌سازی زنجیره تأمین در آماد (لجستیک) و پیش‌بینی وضعیت اقلیمی در محیط‌زیست استفاده می‌شوند. به‌عنوان‌مثال، در بخش خدمات درمانی، سامانه سلامت واتسون[3] شرکت آی‌بی‌ام از واکاوش‌های پیشگویانه برای بهبود مراقبت از بیماران از طریق پیش‌بینی پیشرفت بیماری و پیشنهاد طرح‌های درمانی شخصی‌سازی‌شده استفاده می‌کند [۹]. فناوری شبکه پیش‌بینی‌کننده سیسکو[4] از منابع دورسنجی[5] مختلف داده جمع‌آوری می‌کند و با به‌کارگیری هوش

1. Artificial Narrow Intelligence
2. Artificial General Intelligence
3. Watson Health
4. Cisco Predictive Network
5. telemetry

مصنوعی و مدل‌ها برای آموختن الگوها، پیش‌بینی مشکلات تجربه کاربر و ارائه گزینه‌های حل مسئله استفاده می‌کند و به این شکل شبکه‌های خودترمیم کننده‌ای ایجاد می‌کند که قادر به یادگیری، پیش‌بینی و طرح‌ریزی هستند [۱۰]. در بخش محیط‌زیست، به‌ویژه در آسه‌آن (منطقه اتحادیه کشورهای جنوب شرق آسیا) که در بین سال‌های ۲۰۰۴ تا ۲۰۱۴، بیش از ۵۰٪ از مرگ‌ومیرهای جهانی ناشی از بلایای طبیعی در آن رخ داد [۱۱]، هوش مصنوعی پیش‌بینی‌کننده می‌تواند نقشی حیاتی در مدیریت بحران‌ها ایفا کند. این فناوری می‌تواند مسیر طوفان‌ها را پیش‌بینی نماید و احتمال وقوع زمین‌لرزه و سونامی را ارزیابی کند و راهبردهای پاسخ به بحران مؤثرتری را امکان‌پذیر سازد.

از سوی دیگر سامانه‌های هوش مصنوعی مولد قادرند بر اساس داده‌هایی که با استفاده از آن آموزش‌دیده‌اند، محتوای جدید ایجاد کنند، مانند متن، تصویر، صدا و ویدئو. رشد اخیر در توسعه و توجه به هوش مصنوعی مولد، ناشی از پیشرفت‌های فنونی مانند شبکه‌های مولد تخاصمی (GAN) و مدل‌های زبانی بزرگ (LLM) است. سکوی ChatGPT شرکت OpenAI از این دسته از هوش مصنوعی برای تولید متن، تصویر و کد منسجم بر اساس پرسش‌های[1] کاربر استفاده می‌کند. هوش مصنوعی مولد را همچنین می‌توان برای خلق هنر دیجیتال، ادبیات و موسیقی بازتاب‌دهنده میراث‌فرهنگی به کار گرفت و همچنین برای ساخت دستیارهای هوشمند تعاملی.

۳-۱-۲ انواع هوش مصنوعی

هوش مصنوعی اصطلاح گسترده‌ای است که فناوری‌های مختلفی را در برمی‌گیرد که هرکدام از الگوریتم‌ها و روش‌های آموزشی مختلفی استفاده می‌کنند که برای کاربردهای خاصی طراحی شده است.

یادگیری ماشین که یکی از زیرمجموعه‌های هوش مصنوعی است، بر توسعه الگوریتم‌هایی متمرکز است که رایانه‌ها را قادر می‌سازد از داده‌ها یاد بگیرند و عملکرد خود را در طول زمان بدون برنامه‌نویسی صریح بهبود دهند. یادگیری ماشین عمدتاً برای وظایف پیش‌بینی مانند پیش‌بینی وضعیت آب‌وهوا، تحلیل بازار سهام، شناسایی هرزنامه (اسپم) و بازشناسی تصاویر استفاده می‌شود.

یادگیری عمیق که یک شاخه تخصصی از یادگیری ماشین است، از شبکه‌های عصبی چندلایه برای تحلیل ساختارهای داده پیچیده استفاده می‌کند. یادگیری عمیق به دلیل توانایی استخراج خودکار ویژگی‌ها از داده‌های خام در وظایفی مانند بازشناسی تصویر و گفتار قوی است. در حالی که می‌توان از یادگیری عمیق در کارهای سنتی یادگیری ماشین نیز استفاده کرد، یادگیری عمیق قابلیت‌های پیشرفته‌تری مانند مدل‌های زبانی بزرگ (LLM)، بینایی رایانه‌ای، سامانه‌های خودمختار و تشخیص

1. prompt

پزشکی را نیز امکان‌پذیر می‌سازد.

۲-۲ کاربرد هوش مصنوعی در بخش‌های کلیدی

هوش مصنوعی از زمان سرآغاز خود در اواسط قرن بیستم به طور قابل‌توجهی تکامل‌یافته است. این فناوری از زمان پیروزی جنجالی مشهور سامانه هوش مصنوعی Deep Blue در برابر گری کاسپارف، قهرمان جهانی شطرنج، در سال ۱۹۹۷ [۱۲]، شاهد مراحل مختلفی از توسعه بوده است که هرکدام با رویکردها و دستاوردهای خاص خود همراه بوده‌اند و به کاربردهای گسترده‌تری منتهی شده‌اند. امروزه، موارد استفاده هوش مصنوعی شامل بخش‌های کلیدی مختلفی می‌شود و احتمال پیشبرد اهداف ملی را در اشکال مختلف دارد.

۲-۲-۱ دولت و بخش عمومی

هوش مصنوعی می‌تواند بخش دولتی و عمومی را با بهبود ارائه خدمات، تقویت تصمیم‌گیری و بهینه‌سازی تخصیص منابع دگرگون کند. یادگیری عمیق و پردازش زبان طبیعی بر روی مجموعه‌داده‌های عظیم، مانند سوابق بهداشت الکترونیکی و بازخورد عمومی، اعمال می‌شود تا از تصمیم‌گیری‌های بالینی پشتیبانی نماید، فعالیت‌های مشکوک را شناسایی و وظایف اداری را خودکارسازی کند. به‌عنوان‌مثال، در سنگاپور، دولت از هوش مصنوعی در ابتکار ملت هوشمند خود استفاده می‌کند که هدف آن بهبود زندگی شهری از طریق فناوری است. برنامه MyResponder که از هوش مصنوعی بهره می‌برد، به یافتن نزدیک‌ترین پاسخ‌دهندگان آموزش‌دیده در موارد اضطراری قلبی کمک می‌کند که زمان پاسخگویی و نرخ بقا را به طور قابل‌توجهی بهبود می‌بخشد [۱۳].

علاوه بر این، هوش مصنوعی می‌تواند شرایط زیست‌محیطی را نظارت کند، پیش‌بینی رویدادهای جوی نماید و مدیریت شرایط اضطراری را انجام دهد. یکی از کارکردهای بافایده روزافزون، توانایی مدل‌های هوش مصنوعی در تحلیل تصاویر ماهواره‌ای و داده‌های هواشناسی به‌منظور پیش‌بینی بلایای طبیعی مانند سیل‌ها یا زلزله‌ها است که امکان تخلیه به‌موقع و استقرار منابع را فراهم می‌آورد [۱۴]. دولت‌ها همچنین از هوش مصنوعی برای افزایش بازده عملیات اداری، اطمینان از رعایت استانداردهای مقرراتی و بهبود تحلیل‌های اقتصادی، نظارت بر تجارت و برنامه‌ریزی زیرساخت‌ها از طریق نگهداری پیشگویانه و واکاوش تدارکات استفاده می‌کنند. به‌عنوان‌مثال، واکاوش‌های پیشگویانه می‌تواند فرسایش زیرساخت‌ها را پیش‌بینی کند که امکان رسیدگی به‌موقع و کاهش هزینه‌ها را فراهم می‌کند.

۲-۲-۲ خدمات درمانی

هوش مصنوعی می‌تواند با بهبود تشخیص، درمان و مراقبت از بیماران، انقلابی را در بخش خدمات

درمانی ایجاد کند. فنون هوش مصنوعی مانند یادگیری عمیق و پردازش زبان طبیعی به داده‌های سوابق بهداشتی الکترونیکی، منابع علمی پزشکی و داده‌های ژنتیکی اعمال می‌شود تا الگوها شناسایی شود، نتایج پیش‌بینی گردد و از تصمیم‌گیری بالینی پشتیبانی شود. با تحلیل تصاویر پزشکی توسط الگوریتم‌های هوش مصنوعی برای تشخیص نشانه‌های اولیه بیماری‌هایی مانند سرطان، می‌توان جان افراد بسیاری را نجات داد. علاوه بر نجات جان‌ها، هوش مصنوعی می‌تواند کیفیت زندگی را با کمک به توسعه طرح‌های درمان شخصی‌سازی‌شده از طریق تحلیل اطلاعات ژنتیکی و تاریخچه پزشکی بیمار ارتقا دهد و اثربخش‌ترین درمان‌ها را توصیه کند.

در زمینه کشف دارو، هوش مصنوعی با پیش‌بینی نحوه تعامل ترکیبات مختلف با اهداف در بدن انسان، فرایند کشف دارو را تسریع می‌کند و می‌تواند زمان و هزینه‌های مربوط به عرضه داروهای جدید به بازار را کاهش دهد. به‌علاوه، نوآوران این حوزه به دنبال به کمال رساندن جراحی رباتیک و دستگاه‌های پزشکی کمک گرفته از هوش مصنوعی برای بهبود دقت جراحی و نتایج آن هستند که ریسک‌های ناشی از عوارض را کاهش می‌دهد و زمان‌های بهبودی را سرعت می‌بخشد.

۳-۲-۲ دانش مالی

هوش مصنوعی مزایای زیادی در صنعت مالی دارد، از شناسایی تقلب و ارزیابی ریسک گرفته تا تجارت الگوریتمی و خدمات مشاوره‌ای رباتیک. مدل‌های یادگیری ماشین مقادیر وسیعی از داده‌های مالی را تحلیل می‌کنند تا تراکنش‌های تقلبی را شناسایی نمایند، ریسک اعتبار را ارزیابی کنند و پول‌شویی را شناسایی نمایند. به‌عنوان‌مثال، هوش مصنوعی می‌تواند الگوهای غیرعادی‌ای را در تراکنش‌ها شناسایی کند که ممکن است نشان‌دهنده تقلب باشند که این امر امکان واکنش سریع‌تر و کاهش خسارات مالی را فراهم می‌کند.

در فیلیپین، مشاوران رباتیک متکی بر هوش مصنوعی مشاوره‌های سرمایه‌گذاری شخصی‌سازی‌شده ارائه می‌دهند که برنامه‌ریزی مالی را برای گروه‌های وسیع‌تری از جمعیت قابل‌دسترس‌تر می‌کند. بر اساس گزارش بانک مرکزی فیلیپین[1]، این خدمات مبتنی بر هوش مصنوعی با ارائه خدمات مشاوره‌ای کم‌هزینه به اجتماعات کم‌برخوردار، شمولیت مالی[2] را افزایش داده‌اند [۱۵].

۴-۲-۲ حمل‌ونقل

هوش مصنوعی نقشی اساسی در توسعه وسایل نقلیه خودران دارد که وعده انقلابی در حمل‌ونقل را با بهبود ایمنی، کاهش ازدحام و ارتقای قابلیت تحرک می‌دهند. فنونی مانند بینایی رایانه‌ای و یادگیری

1. Bangko Sentral ng Pilipinas
2. financial inclusion

عمیق تقویتی به وسایل نقلیه این امکان را می‌دهند که محیط اطراف خود را درک کنند، تصمیم‌گیری کنند و در سناریوهای ترافیک پیچیده مسیریابی کنند. اگرچه نظارت انسانی هنوز ضروری است، وسایل نقلیه خودران می‌توانند تصادفات ناشی از اشتباهات انسانی را کاهش دهند، بازده سوخت را بهبود بخشند و راه‌حل‌هایی برای حرکت افرادی فراهم کنند که قادر به رانندگی نیستند، مانند کسانی که دچار معلولیت هستند. در اقتصادهای پیشرفته‌تر، هوش مصنوعی برای بهینه‌سازی مدیریت ترافیک با تحلیل داده‌های بی‌درنگ از دوربین‌های ترافیک، حسگرها و دستگاه‌های سامانه موقعیت‌یاب جهانی (GPS) جهت تنظیم زمان‌بندی چراغ‌های راهنمایی و مدیریت ازدحام استفاده می‌شود. واکاوش پیشگویانه می‌تواند تقاضا برای خدمات حمل‌ونقل عمومی را پیش‌بینی نماید و امکان برنامه‌ریزی و تخصیص منابع بهتر را فراهم کند. در آماد، هوش مصنوعی برنامه‌ریزی مسیرها را بهبود می‌بخشد، زمان تحویل را کاهش می‌دهد و هزینه‌های عملیاتی را پایین می‌آورد که این باعث افزایش بازده کلی و رضایت مشتری می‌شود.

۵-۲-۲ انرژی

هوش مصنوعی تولید، توزیع و مصرف انرژی را بهینه‌سازی می‌کند. به‌عنوان‌مثال، در تایلند، مدل‌های یادگیری ماشین میزان تقاضای انرژی را پیش‌بینی می‌نمایند و به‌کارگیری انرژی‌های تجدیدپذیر را بهینه‌سازی می‌کنند. شرکت تولید برق تایلند (EGAT) از هوش مصنوعی برای تحلیل الگوهای جوی و پیش‌بینی تولید انرژی خورشیدی و بادی استفاده می‌کند که به تعادل عرضه و تقاضا در شبکه برق کمک می‌کند [۱۶]. در صنعت نفت و گاز، هوش مصنوعی با تحلیل داده‌های زمین‌شناختی جهت شناسایی سایت‌های حفاری دارای بیشترین پتانسیل، پیش‌بینی خرابی‌های تجهیزات و بهبود ایمنی، عملیات‌های حفاری را بهینه‌سازی می‌کند. مدل‌های نگهداری پیشگویانه می‌توانند زمان خرابی احتمالی تجهیزات را پیش‌بینی کنند که این کار امکان مداخلات به موقعی را فراهم می‌نماید که از وقفه‌ها و حوادث هزینه‌بر جلوگیری می‌کند.

۶-۲-۲ تولید

هوش مصنوعی امکان نگهداری پیشگویانه، کنترل کیفیت، بهینه‌سازی زنجیره تأمین و یکپارچه‌سازی رباتیک را فراهم می‌سازد.

می‌توان داده‌های حسگر تجهیزات را برای پیش‌بینی و جلوگیری از خرابی‌ها تحلیل کرد که این امر زمان خرابی و هزینه‌های نگهداری را کاهش می‌دهد. به‌عنوان‌مثال، هوش مصنوعی می‌تواند پیش‌بینی کند که یک قطعه ماشین چه زمانی احتمالاً فرسوده خواهد شد و قادر است نگهداری را پیش از خرابی برنامه‌ریزی کند تا از تأخیر در تولید جلوگیری شود. بینایی رایانه‌ای و یادگیری عمیق با شناسایی

بسیار دقیق نقص‌ها و ناهنجاری‌ها در محصولات، کنترل کیفیت را بهبود می‌بخشند و اطمینان ایجاد می‌کنند که فقط محصولات باکیفیت بالا به بازار عرضه می‌شود. پیش‌بینی تقاضا و بهینه‌سازی زنجیره تأمین با استفاده از هوش مصنوعی، با پیش‌بینی دقیق‌تر تقاضای مصرف‌کننده و تنظیم متناسب برنامه‌های تولید، مدیریت موجودی را بهینه‌سازی می‌کند که باعث کاهش ضایعات و بهبود کارایی می‌شود. یکپارچه‌سازی رباتیک در فرایندهای تولید، خودکارسازی را افزایش، خطاهای انسانی را کاهش و سرعت و یکدستی تولید را تقویت می‌دهد.

۷-۲-۲ خرده‌فروشی

رشد در بخش خرده‌فروشی و سکوهای تجارت الکترونیکی مانند آمازون از طریق ویژگی‌هایی تأمین می‌شود که توسط هوش مصنوعی امکان‌پذیر می‌شوند، مانند پیشنهادات شخصی‌سازی‌شده، قیمت‌گذاری پویا و تجربیات مشتری. مدل‌های یادگیری ماشین داده‌های مشتری را، از جمله تاریخچه مرور و خرید، تحلیل می‌کنند تا پیشنهادات محصولات شخصی‌سازی‌شده و بازاریابی هدفمند ارائه دهند. تنها شرکت‌های بزرگ چندملیتی نیستند که از هوش مصنوعی بهره برده‌اند، بلکه سکوهای تجارت الکترونیکی محلی در اندونزی مانند Tokopedia و Bukalapak و سکوی مستقر در سنگاپور Shopee نیز از هوش مصنوعی برای پیشنهاد محصولات بر اساس رفتار گذشته مشتری استفاده می‌کنند که این امر باعث افزایش فروش و رضایت مشتری می‌شود. چت‌بات و دستیارهای مجازی مبتنی بر هوش مصنوعی پشتیبانی مشتری ۲۴ ساعته در ۷ روز هفته ارائه می‌دهند و به درخواست‌های معمول پاسخ می‌دهند و زمان پاسخگویی را بهبود می‌بخشند. این سامانه‌ها می‌توانند به وظایفی مانند ردیابی سفارش، پردازش مرجوعی‌ها و اطلاعات محصولات کمک کنند و وقت کارکنان انسانی را برای تمرکز بر مسائل پیچیده‌تر آزاد کنند.

۳ هوش مصنوعی و امنیت سایبری

با پیشرفت و گسترش بیشتر فناوری‌های هوش مصنوعی، دو حوزه اصلی از نظر ریسک‌های امنیت سایبری به وجود می‌آید: حملات علیه هوش مصنوعی و حملات تسهیل شده و تقویت‌شده توسط هوش مصنوعی. هوش مصنوعی نیز مانند هر فناوری یا سامانه‌ای، در معرض ریسک نفوذ یا دست‌کاری قرار دارد. البته جنبه‌های منحصربه‌فردی در نحوه توسعه و پیاده‌سازی هوش مصنوعی وجود دارد که نیاز به توجه ویژه دارند. به‌عنوان‌مثال، یک مهاجم ممکن است داده‌های آموزشی را به گونه نامحسوسی دست‌کاری کند که منجر به تولید اطلاعات غلط یا ایجاد خروجی‌های نادرستی شود که بتواند بر سامانه‌های حیاتی و ضروری تأثیر منفی بگذارد. در مقابل، هوش مصنوعی می‌تواند ابزاری برای تقویت تاکتیک‌ها، فنون و رویه‌های موجود (TTPs) در حملات سایبری باشد. به‌عبارت‌دیگر،

هوش مصنوعی سد مقابل دسترسی مجرمان سایبری را تضعیف می‌کند و افرادی را که دارای کمترین دانش فنی هستند قادر به آغاز حملات سایبری پیچیده می‌کند. به‌عنوان‌مثال، ابزارهای مبتنی بر هوش مصنوعی می‌توانند فرایند ایجاد نرم‌افزارهای مخرب را خودکارسازی کنند و حملات مهندسی اجتماعی را تقویت کنند که این امر باعث می‌شود مهاجمان بامهارت کمتر، با سهولت بیشتری بتوانند با پیام‌های فیشینگ بسیار قانع‌کننده، از باج‌افزار یا بدافزار استفاده کنند.

تکامل فناوری‌های هوش مصنوعی همچنین مرز بین رسانه‌های مصنوعی و محتوای تولیدشده توسط انسان را محو می‌کند و شناسایی جعل‌های عمیق[1] را پیچیده‌تر می‌نماید. این جعل‌های عمیق بسیار واقع‌گرایانه تهدیدات قابل‌توجهی را به همراه دارند، از کارزارهای اطلاعات غلط گرفته تا دزدی هویت. فناوری‌های شناسایی در مقابله با پیشرفت‌های سریع محتوای تولیدشده توسط هوش مصنوعی به تکاپو افتاده‌اند که این امر جعل‌های عمیق را هدفمندتر و خطرناک‌تر از همیشه می‌کند.

با وجود این چالش‌ها، جا برای امیدواری وجود دارد. جیتو پاتل، معاون اجرایی و مدیر محصول سیسکو، تصریح کرده است: «این زمان خوبی برای متمایل کردن کفه ترازو به نفع مدافعان است». این نشان می‌دهد که اگرچه هوش مصنوعی ریسک‌های جدیدی را ایجاد می‌کند، اما فرصت‌های بی‌سابقه‌ای را نیز برای تقویت دفاع‌های امنیت سایبری فراهم می‌آورد [۱۷].

همچنین مهم است که بدانیم ریسک‌های مختص به هوش مصنوعی در انزوا وجود ندارند، بلکه در بافت تهدیدات امنیت سایبری سنتی قرار دارند. اولین گام در توسعه شیوه‌های دفاعی استوار، درک کامل نحوه وقوع حملات و ریسک‌های مرتبط با آن‌ها است. در این فصل، یک طبقه‌بندی از ریسک‌های امنیت سایبری گردآوری شده است که مستخرج از یک مرور کوتاه از چارچوب‌های موجود شورای امنیت داده هند (DSCI) [۱۸] و مؤسسه ملی استاندارد و فناوری ایالات متحده (NIST) [۱۹] است.

۳-۱ ریسک‌های امنیت داده

این دسته شامل ریسک‌های مرتبط با محرمانگی، یکپارچگی و حریم خصوصی داده‌هایی است که برای آموزش و عملیات مدل‌های هوش مصنوعی استفاده می‌شود. حملاتی که این حوزه‌ها را هدف قرار می‌دهند، قصد دست‌کاری، دزدی یا استنباط اطلاعات حساس را دارند که تهدیدات قابل‌توجهی را برای امنیت کلی و قابلیت اعتماد سامانه‌های هوش مصنوعی ایجاد می‌کند.

۳-۱-۱ مسمومیت داده

مسمومیت داده شامل دست‌کاری داده‌های آموزشی توسط مهاجمان برای ایجاد آسیب‌پذیری‌ها یا

1. deepfakes

درهای پشتی[1] به مدل هوش مصنوعی می‌شود. این فعالیت مخرب می‌تواند درستی و اطمینان‌پذیری مدل را به‌شدت تضعیف کند.

مثال: مهاجم تغییراتی را در مجموعه‌داده‌ای ایجاد می‌کند که برای آموزش یک مدل شناسایی بدافزار استفاده می‌شود. مهاجم می‌تواند با ویرایش دقیق نمونه‌های بدافزار در مجموعه‌داده، مدل را به دسته‌بندی اشتباه برخی از انواع بدافزار به‌عنوان نرم‌افزار بی‌ضرر وادارد. این دسته‌بندی اشتباه می‌تواند این امکان را ایجاد کند که نرم‌افزار مخرب شناسایی را دور بزند که باعث نفوذ امنیتی و آسیب احتمالی به سامانه‌ها و داده می‌شود.

سناریوی ممکن در دنیای واقعی: در یک سناریوی فرضی، ممکن است عملکرد یک سامانه هوش مصنوعی خدمات درمانی که برای تشخیص بیماری از طریق تصاویر پزشکی طراحی شده باشد، در صورت دست‌کاری داده‌های آموزشی توسط مهاجمین، به اختلال بیفتد. ممکن است مدل با اضافه‌شدن تصاویری با تغییرات نامحسوس ولی تأثیرگذار، طوری آموزش داده شود که برخی بیماری‌ها را به‌اشتباه تشخیص دهد که این امر ممکن است منجر به توصیه‌های درمانی نادرست شود.

۳-۱-۲ استخراج داده

حملات استخراج داده شامل استنباط یا بازسازی اطلاعات حساس درباره داده‌های آموزشی بر مبنای خروجی‌ها یا رفتارهای مدل توسط مهاجمان می‌شود. این نوع حمله می‌تواند اطلاعات محرمانه را افشا کند که این امر ریسک‌های قابل‌توجهی را در زمینه حریم خصوصی یا اطلاعات انحصاری ایجاد می‌کند.

مثال: یک مهاجم به شکلی برای یک مدل زبان پرسمان ارسال می‌کند که حاوی پرسش‌هایی است که به‌دقت طوری تنظیم شده است که اطلاعات حساس گنجانده شده در داده آموزشی مدل را استخراج کند. برای مثال، مهاجم می‌تواند پرسش‌هایی را وارد کند که مدل زبان را به فاش کردن اطلاعات خصوصی وادار کند، مانند آدرس‌های ایمیل، شماره‌تلفن‌ها یا شناسه‌های شخصی که بخشی از داده‌های آموزشی اصلی بوده‌اند.

سناریوی ممکن در دنیای واقعی: در بافت یک چت‌بات خدمات مشتری که با استفاده از مجموعه‌داده گسترده‌ای از تعاملات با مشتریان آموزش‌دیده است، مهاجم می‌تواند مدل را به‌گونه‌ای مورد سوءاستفاده قرار دهد که اطلاعات حساس مشتری را استخراج کند. ممکن است مهاجم با استفاده از پرسمان‌های خاص، اطلاعات محرمانه‌ای را از تراکنش‌های گذشته یا داده‌های شخصی به دست آورد که حریم خصوصی کاربران را نقض کند و شاید منجر به سرقت هویت یا تقلب مالی بشود.

1. backdoor

۳-۱-۳ حملات استنباطی

حملات استنباطی مهاجمان را قادر می‌سازد مشخص کنند آیا یک نقطه داده خاص بخشی از مجموعه‌داده آموزشی مدل بوده است (استنباط عضویت) یا خیر یا به آن‌ها این امکان را می‌دهد که مشخصات کلی‌ای را درباره داده‌های آموزشی استنباط کنند (استنباط مشخصات). این حملات می‌توانند محرمانگی و یکپارچگی داده‌های آموزشی را به خطر بیندازند.

مثال: یک مهاجم خروجی‌های یک مدل یادگیری ماشین را تجزیه‌وتحلیل می‌کند تا بفهمد آیا داده‌های یک فرد خاص در مجموعه‌داده‌های آموزشی گنجانده شده است یا خیر. این استنباط عضویت می‌تواند باعث مشخص‌شدن حضور در مجموعه‌داده شود و احتمال افشای اطلاعات شخصی یا حساس را دارد. همچنین، استنباط مشخصات به مهاجم این امکان را می‌دهد که مشخصه‌های کلی‌تر داده‌های آموزشی مانند حضور برخی گروه‌های جمعیت‌شناختی خاص را استنباط کند.

سناریوی ممکن در دنیای واقعی: یک مدل یادگیری ماشین را در نظر بگیرید که با استفاده از داده‌های تراکنش مالی برای شناسایی فعالیت‌های متقلبانه آموزش‌داده‌شده است. مهاجم ممکن است از این مدل برای استنباط اینکه آیا تاریخچه تراکنش‌های یک فرد خاص در داده‌های آموزشی وجود داشته است یا خیر استفاده کند. ممکن است مهاجم با تجزیه‌وتحلیل دقیق پاسخ‌های مدل به ورودی‌های مختلف، رفتارها یا الگوهای مالی حساسی را فاش کند که حریم خصوصی فرد را به خطر بیندازد و بتواند منجر به جرایم مالی هدفمند شود.

۳-۲ ریسک‌های امنیتی مدل

ریسک‌های امنیتی مدل شامل تهدیداتی می‌شود که به طور خاص معماری، پارامترها یا مرزهای تصمیم‌گیری مدل‌های هوش مصنوعی را هدف قرار می‌دهند. این حملات با هدف آسیب‌رساندن به یکپارچگی و اطمینان‌پذیری خروجی‌های مدل طراحی شده‌اند که ممکن است منجر به پیامدهای مضر شوند. درک و کاهش این ریسک‌ها برای اطمینان‌یافتن از امنیت و استواری سامانه‌های هوش مصنوعی ضروری است.

۳-۲-۱ حملات دورزنی

حملات دورزنی[1] شامل ساخت نمونه‌های تخاصمی‌ای توسط مهاجمین می‌شود که برای واداشتن مدل به دسته‌بندی اشتباه یا تولید خروجی‌های نادرست طراحی شده‌اند. این تغییرات نامحسوس می‌توانند حتی مدل‌های با درستی بالا را فریب دهند و منجر به نتایج احتمالاً خطرناکی شوند.

مثال: یک مهاجم تغییرات کوچک و تخاصمی را در یک تصویر ایجاد می‌کند که باعث می‌شود

1. evasion attack

سامانه بینایی رایانه‌ای یک وسیله نقلیه خودران یک علامت توقف را به‌اشتباه به‌عنوان یک علامت محدودیت سرعت دسته‌بندی کند. این اشتباه دسته‌بندی ممکن است باعث عدم توقف وسیله نقلیه در تقاطع‌ها و منجر به بروز تصادفات و تهدید جان انسان‌ها شود.

سناریوی ممکن در دنیای واقعی: در سال ۲۰۱۸، محققان نشان دادند که با قراردادن برچسب‌های کوچک روی علائم توقف، می‌توان سامانه شناسایی علائم ترافیکی مبتنی بر هوش مصنوعی را فریب داد تا این علائم را اشتباه دسته‌بندی کنند. این نوع حملات دورزنی آسیب‌پذیری‌های مدل‌های هوش مصنوعی‌ای را که در کاربردهای حساس مانند رانندگی خودران به کار گرفته می‌شوند برجسته می‌سازد [۲۰].

۳-۲-۲ حملات در پشتی

حملات در پشتی شامل تزریق محرک‌های مخفی توسط مهاجم به مدل در هنگام مرحله آموزش آن است. زمانی که این محرک‌ها فعال شوند، باعث اشتباه‌های هدفمند در دسته‌بندی مدل می‌شود، اغلب بدون آنکه عملکرد مدل در ورودی‌های معمولی تغییری کند.

مثال: یک مهاجم یک مدل بازشناسی چهره را در هنگام آموزش با تزریق تصاویری که یک پوشاک جانبی خاص مانند یک عینک دارد، آلوده می‌کند. وقتی افراد این پوشاک جانبی را می‌پوشند، مدل آن‌ها را اشتباه دسته‌بندی می‌کند و نمی‌تواند به‌درستی آن‌ها را شناسایی کند. این می‌تواند به افراد غیرمجاز امکان عبور از سامانه‌های امنیتی را بدهد.

سناریوی ممکن در دنیای واقعی: محققان در یک مطالعه دریافتند که با تزریق داده‌های آموزشی آلوده به سامانه‌های بازشناسی چهره، می‌توانند درهای پشتی‌ای بسازند که به آن‌ها امکان عبور از تمهیدات امنیتی را صرفاً با پوشیدن یک جفت عینک خاص بدهد [۲۱]. این فن می‌تواند برای دورزدن سامانه‌های امنیتی در محیط‌های حساس مانند فرودگاه‌ها یا تأسیسات امن استفاده شود.

۳-۲-۳ مسمومیت مدل

مسمومیت مدل شامل دست‌کاری پارامترها یا معماری مدل توسط مهاجمان به‌قصد ایجاد آسیب‌پذیری یا کاهش عملکرد مدل می‌شود. این موضوع به‌ویژه در محیط‌های یادگیری مشارکتی مانند یادگیری فدرال نگران‌کننده است.

مثال: در یک محیط یادگیری فدرال که چندین مشارکت‌کننده برای آموزش یک مدل جهانی همکاری می‌کنند، یک مشارکت‌کننده مخرب از عمد به‌روزرسانی‌های مدل آلوده به سرور مرکزی ارسال می‌کند. این به‌روزرسانی‌های مسموم باعث کاهش عملکرد مدل جهانی می‌شود و ممکن است باعث نقص آن در کاربردهای حیاتی گردد.

سناریوی ممکن در دنیای واقعی: یادگیری فدرال در خدمات درمانی برای آموزش مشارکتی مدل‌های پیشگویانه در بیمارستان‌های مختلف بدون به‌اشتراک‌گذاری داده‌های بیماران استفاده می‌شود. اگر یک هستار[1] مخرب به‌روزرسانی‌های آلوده تزریق کند، ممکن است مدل حاصل پیش‌بینی‌های پزشکی نادرست ارائه دهد و احتمال آسیب‌رسانی به بیماران وجود دارد.

۳-۲-۴ سرقت مدل

سرقت مدل زمانی رخ می‌دهد که مهاجمان معماری، پارامترها یا کارکرد مدل را از طریق دسترسی پرسمان مکرر استخراج یا شبیه‌سازی کنند. این می‌تواند مدل‌های انحصاری و اموال فکری را به خطر بیندازد و به مهاجمان امکان ایجاد نسخه‌های مشابه غیرمجاز را بدهد.

مثال: یک مهاجم به طور مکرر به یک مدل یادگیری ماشینی قابل‌دسترسی از طریق میانای برنامه‌سازی کاربردی[2] پرسمان می‌دهد و خروجی‌ها را تجزیه‌وتحلیل می‌کند تا مدلی با کارکرد معادل آن را بازتولید کند. این مدل دزدیده‌شده سپس می‌تواند بدون مجوز استفاده شود و ممکن است از محافظت‌ها عبور کند یا زیان‌های رقابتی ایجاد کند.

سناریوی ممکن در دنیای واقعی: محققان نشان داده‌اند که با ارائه پرسمان‌های مکرر به میاناهای برنامه‌سازی کاربردی یادگیری ماشینی تجاری مانند میاناهای گوگل، آمازون و مایکروسافت، می‌توان مدل‌های زیربنایی را عملاً بازتولید کرد [۲۲]. این نوع حمله می‌تواند منجر به سرقت قابل‌توجه دارایی‌های فکری شود و مزیت رقابتی ارائه‌دهندگان خدمات هوش مصنوعی را کاهش دهد.

۳-۳ ریسک‌های زیرساختی

ریسک‌های زیرساختی شامل تهدیداتی می‌شود که به زیرساخت‌ها و ابزارهای زیربنایی مورد استفاده برای توسعه، آموزش و پیاده‌سازی مدل‌های هوش مصنوعی مربوط می‌شوند. این حملات، در دسترس بودن، عملکرد و یکپارچگی زنجیره تأمین سامانه‌های هوش مصنوعی را هدف قرار می‌دهند و ممکن است منجر به اختلالات و آسیب‌پذیری‌های قابل‌توجهی شوند.

۳-۳-۱ حملات انکار خدمات

حملات انکار خدمات (DoS) شامل از کار انداختن یک سامانه هوش مصنوعی با تعداد زیادی درخواست می‌شود که سامانه را برای کاربران قانونی غیرقابل‌دسترسی یا غیرپاسخگو می‌نماید. این حملات می‌توانند خدمات حیاتی را مختل کنند و تجربه کاربری کلی را کاهش دهند.

1. entity
2. API

مثال: یک مهاجم یک خدمت بازشناسی تصویر مبتنی بر هوش مصنوعی را غرق در حجم زیادی از درخواست می‌کند. این سیلاب ترافیک باعث سرریزشدن ظرفیت این خدمت می‌شود و آن را از پردازش درخواست‌های مشروع کاربران عاجز می‌سازد که منجر به توقف خدمات و اختلالات عملیاتی می‌شود.

سناریوی ممکن در دنیای واقعی: یک حمله منع خدمت توزیع‌شده (DDoS) سکوی میزبانی مخزن گیت‌هاب (GitHub) را هدف قرار داد. اگرچه این حمله متکی بر هوش مصنوعی نبود، اما این حمله نشان داد که چگونه درخواست‌های بیش از حد می‌تواند زیرساخت‌های حیاتی را از کار بیندازد [۲۳]. تاکتیک‌های مشابه می‌تواند برای سامانه‌های هوش مصنوعی نیز به کار گرفته شود که در آن هدف، اختلال در خدمات از طریق اشباع منابع سامانه است.

۳-۳-۲ حملات اشباع منابع

حملات اشباع منابع، باعث تکمیل ظرفیت منابع رایانشی یک سامانه هوش مصنوعی می‌شود و در نتیجه عملکرد یا قابلیت دسترسی آن را کاهش می‌دهند.

مهاجمان پرسمان‌ها یا ورودی‌هایی ایجاد می‌کنند که از ضعف‌های سامانه بهره‌برداری می‌کند و منجر به مصرف بیش از اندازهٔ منابع رایانشی و افزایش تأخیر در رسیدگی به درخواست‌های مشروع می‌شود.

مثال: یک مهاجم ضعفی را در یک سامانه هوش مصنوعی شناسایی می‌کند که به وی این امکان را می‌دهد که درخواست‌های رایانشی سنگین ارسال کند. مهاجم با ارسال مکرر این درخواست‌ها، باعث مصرف بیش از حد توان رایانشی سامانه هوش مصنوعی می‌شود که منجر به پاسخ‌دهی کند و کاهش عملکرد برای کاربران مشروع می‌شود.

سناریوی ممکن در دنیای واقعی: در بافت خدمات هوش مصنوعی مبتنی بر ابر، حملات اشباع منابع می‌تواند منجر به افزایش هزینه‌های عملیاتی به دلیل مصرف بیش از اندازه منابع رایانشی شود. برای مثال، یک مدل هوش مصنوعی که روی یک سکوی ابری مستقر شده باشد، در صورت هدف قرارگرفتن توسط چنین حملاتی، می‌تواند کاهش سرعت قابل‌توجهی را تجربه کند و با افزایش هزینه مواجه شود.

۳-۳-۳ حملات زنجیره تأمین

حملات زنجیره تأمین به یکپارچگی ابزارها، کتابخانه‌ها یا سکوهای توسعه هوش مصنوعی آسیب می‌زنند و آسیب‌پذیری‌هایی را ایجاد می‌کنند که مهاجمان می‌توانند از آن‌ها بهره‌برداری کنند. این حملات می‌توانند پیامدهای گسترده‌ای داشته باشند، زیرا مؤلفه‌های آسیب‌دیده معمولاً در سامانه‌ها و کاربردهای هوش مصنوعی متعددی به طور وسیع مورد استفاده قرار می‌گیرند.

مثال: یک مهاجم یک چارچوب یادگیری ماشینی متن‌باز را که به طور گسترده استفاده می‌شود نفوذ نموده و یک آسیب‌پذیری ایجاد می‌کند که به وی این امکان را می‌دهد که به داده‌های حساس پردازش‌شده توسط سامانه‌های هوش مصنوعی ساخته‌شده با استفاده از این چارچوب دسترسی پیدا کند. این می‌تواند منجر به نقض‌های گسترده داده و سوءاستفاده بالقوه از سامانه‌های هوش مصنوعی آسیب‌پذیر شود.

سناریوی ممکن در دنیای واقعی: حمله زنجیره تأمین SolarWinds تأثیر بالقوه چنین حملاتی بر زیرساخت‌های نرم‌افزاری را برجسته ساخت. یک روزآمدسازی معیوب به نرم‌افزار SolarWinds منجر به نفوذ گسترده به شبکه‌های دولتی و شرکتی شد [۲۴]. به طور مشابه، حمله به ابزارهای توسعه هوش مصنوعی می‌تواند امنیت بسیاری از سامانه‌های هوش مصنوعی را به خطر بیندازد.

۳-۴ ریسک‌های کاربرد

ریسک‌های کاربرد تهدیدات و چالش‌هایی هستند که با کاربرد یا مورد استفاده خاص یک سامانه هوش مصنوعی مرتبط هستند. این ریسک‌ها می‌توانند ناشی از تعامل بین مدل هوش مصنوعی و کاربران آن و همچنین ناشی از پیامدهای اخلاقی و اجتماعی کلی به‌کارگیری هوش مصنوعی در بافت‌های مختلف باشند. رسیدگی به این ریسک‌ها برای اطمینان از استفاده مسئولانه و اخلاقی از فناوری‌های هوش مصنوعی ضروری است.

۳-۴-۱ حملات تزریق پرسش

حملات تزریق پرسش شامل دست‌کاری پرسش‌های ورودی ارائه‌شده به یک سامانه هوش مصنوعی توسط مهاجم می‌شود که باعث ایجاد خروجی‌های ناخواسته، مضر یا سوگیرانه می‌شود. این حملات از اتکای سامانه به داده‌های ورودی سوءاستفاده می‌کنند تا رفتار آن را به‌گونه‌ای تحت‌تأثیر دهند که می‌تواند مخرب یا گمراه‌کننده باشد.

مثال: مهاجم پرسش گمراه‌کننده‌ای طراحی می‌کند که باعث می‌شود مدل زبانی محتوایی را تولید کند که مقاصد سیاسی خاصی را در پوشش اخبار عینی اشاعه می‌دهد. این می‌تواند کاربران را فریب دهد و اطلاعات نادرست را پخش کند و بر افکار و اعتماد عمومی به منابع اطلاعاتی تأثیر بگذارد.

سناریوی ممکن در دنیای واقعی: ممکن است مهاجمان در سکوهای رسانه اجتماعی، از تزریق پرسش برای گسترش تبلیغات سیاسی (پروپاگاندا) یا اخبار جعلی استفاده کنند. به‌عنوان‌مثال، در دوره‌های انتخابات، پرسش‌های دست‌کاری‌شده می‌توانند باعث شوند سامانه‌های هوش مصنوعی اطلاعات نادرستی تولید و منتشر کنند که بر رفتار رأی‌دهندگان تأثیر می‌گذارد و فرایندهای دموکراتیک را تضعیف می‌کند.

۳-۴-۲ مشکلات یکپارچگی خروجی

مشکلات یکپارچگی خروجی زمانی رخ می‌دهند که یک سامانه هوش مصنوعی خروجی‌های نامنسجم، غیرقابل‌اعتماد یا سوگیرانه تولید می‌کند. این مشکلات می‌توانند منجر به اشتباهات در تصمیم‌گیری، آسیب به اعتبار یا درز اطلاعات خصوصی شوند و اعتماد و بازده سامانه هوش مصنوعی را به خطر بیندازند.

مثال: یک سامانه نظارت محتوای مبتنی بر هوش مصنوعی قادر به شناسایی و پرچم‌گذاری مداوم گفتار نفرت‌آمیز نیست و این باعث انتشار محتوای مضر در یک سکوی رسانه اجتماعی می‌شود. این ناهماهنگی می‌تواند منجر به انتقاد از سکو به دلیل عدم مدیریت مؤثر محتوای مضر شود و به اعتبار و اعتماد کاربران به سکو آسیب بزند.

سناریوی ممکن در دنیای واقعی: فیس‌بوک و توییتر به دلیل سامانه‌های نظارت محتوای هوش مصنوعی خود که گاهی در شناسایی و حذف محتوای مضر مانند سخنان نفرت‌آمیز و اطلاعات نادرست ناکام بوده‌اند، مورد انتقاد قرار گرفته‌اند. این شکست‌ها منجر به بازخورد منفی عمومی قابل‌توجهی شده است و نگرانی‌هایی را در مورد اطمینان‌پذیری هوش مصنوعی در نظارت بر محتوا مطرح ساخته است.

۳-۴-۳ ریسک‌های اخلاقی و اجتماعی

ریسک‌های اخلاقی و اجتماعی شامل پیامدهای گسترده‌ای می‌شود که سامانه‌های هوش مصنوعی بر افراد، گروه‌ها یا جامعه دارند. این ریسک‌ها می‌توانند ناشی از پیامدهای ناخواسته استفاده از هوش مصنوعی باشند که منجر به تأثیرات منفی مانند تحکیم سوگیری‌ها، نقض حریم خصوصی یا تشدید نابرابری‌های اجتماعی می‌شوند.

مثال: یک سامانه هوش مصنوعی که برای پلیس پیشگویانه استفاده می‌شود ممکن است سوگیری‌های موجود در داده‌های تاریخی جرایم را بیاموزد و آن‌ها را تقویت کند که این منجر به سخت‌گیری بیش از اندازه به برخی اجتماع‌های خاص و استمرار نابرابری‌های اجتماعی می‌شود. این می‌تواند منجر به افزایش نظارت و آزار گروه‌های حاشیه‌ای شود و اعتماد به نیروی انتظامی را کاهش دهد.

سناریو واقعی محتمل: استفاده از هوش مصنوعی در پلیس پیشگویانه از بعد از آنکه گزارش‌ها نشان داد این سامانه‌ها اجتماع‌های اقلیت را بدون تناسب بیشتر هدف قرار داده‌اند، مدام زیر سؤال رفته است. مطالعات نشان می‌دهد که سوگیری‌های تاریخی در داده‌های جرایم باعث شده است که سامانه‌های هوش مصنوعی تمرکز غیرمنصفانه‌ای روی محله‌های خاصی داشته باشند که نگرانی‌های اخلاقی‌ای را در مورد انصاف و تأثیر این فناوری‌ها ایجاد کرده است [۲۵].

۴ توسعه چارچوبی برای مواجهه با ریسک‌های امنیت سایبری هوش مصنوعی

۴-۱ رویکرد

۴-۱-۱ مبنا قراردادن و تکمیل قوانین و مقررات موجود

منطقه آسیا - اقیانوسیه که میزبان برخی از پویاترین و سریع‌الرشدترین اقتصادهای جهان است، شاهد افزایش چشمگیر پذیرش هوش مصنوعی است. بااین‌حال، این پیشرفت در فناوری، سازمان‌ها را در معرض چشم‌انداز پیچیده‌ای از تهدیدهای امنیت سایبری و سایر ریسک‌های مرتبط با استفاده از هوش مصنوعی نیز قرار می‌دهد.

رویکردهای حاکمیتی و نظاری هوش مصنوعی باید از قوانین و مقررات موجود برای رسیدگی به نگرانی‌های نوظهور امنیت سایبری بهره بگیرند. این راهبرد باعث اطمینان از انسجام مقرراتی، جلوگیری از سردرگمی ناشی از هم‌پوشانی قوانین و مبنا قراردادن و تکمیل رویه‌های برتر موجود می‌شود. بسیاری از قوانین موجود، در همین حال حاضر نگرانی‌های کلیدی مرتبط با هوش مصنوعی، مانند حریم خصوصی داده، عدم تبعیض و مالکیت فکری را پوشش می‌دهند. سازمان‌ها می‌توانند با همسوسازی اقدامات امنیت سایبری هوش مصنوعی با چارچوب‌های موجود، راه انطباق را هموار کنند، اجرای اثربخش را تسهیل نمایند و همکاری میان بخش‌ها را ارتقا دهند.

برای این منظور، رویکردی گام‌به‌گام در تدوین چارچوبی برای امنیت سایبری هوش مصنوعی ضروری است. این رویکرد شامل موارد زیر است:

- نگاشت مقررات موجود برای شناسایی و درک چشم‌انداز مقرراتی و یافتن دقیق نقاطی که در آن می‌توان از قوانین موجود بهره برد.
- یکپارچه‌سازی رویه‌های برتر از مقررات موجود درون چارچوب امنیت سایبری هوش مصنوعی برای بهره‌مند ساختن این چارچوب از استانداردها و روش‌شناسی‌های تثبیت‌شده.
- متناسب‌سازی مقررات موجود برای رسیدگی به ریسک‌های منحصربه‌فرد مرتبط با سامانه‌های هوش مصنوعی از طریق به‌روزرسانی الزامات مقرراتی به‌منظور پوشش آسیب‌پذیری‌های خاص هوش مصنوعی، مانند دورزدن مدل و مسمومیت داده.
- توسعه سازوکارهایی برای پایش مستمر و بررسی انطباق به‌منظور اینکه سامانه‌های هوش مصنوعی بتوانند از طریق ممیزی، ارزیابی ریسک و گزارش‌دهی منظم، از الزامات مقرراتی به‌روز پیروی کنند.

- **مقررات کلیدی در سطح منطقه**

چندین کشور در منطقه آسیا - اقیانوسیه چارچوب‌های حقوقی برای رسیدگی به ریسک‌های امنیت سایبری هوش مصنوعی تدوین کرده‌اند. از جمله این کشورها می‌توان به سنگاپور، استرالیا، ژاپن و کره جنوبی اشاره کرد.

- **چارچوب مدل حاکمیت هوش مصنوعی سنگاپور، قانون حفاظت از داده‌های شخصی**

چارچوب حاکمیت هوش مصنوعی سنگاپور، دستورالعمل‌هایی عملی جهت مدیریت استقرار هوش مصنوعی به سازمان‌ها ارائه می‌کند؛ تمرکز این چارچوب معطوف بر موضوعاتی مانند شفافیت، انصاف و پاسخگویی است. این چارچوب می‌تواند توسعه سامانه‌های هوش مصنوعی را با تدابیر امنیتی درونی، مانند ممیزی‌ها و ارزیابی‌های ریسک منظم، هدایت نماید تا از تهدیدهای امنیت سایبری پیشگیری شود و از اثرات این تهدیدات کاسته شود [۲۶]. علاوه بر این، قانون حفاظت از داده‌های شخصی (PDPA) در سنگاپور دستورالعمل‌های جامعی را درباره حفاظت از داده ارائه می‌کند که برای امن ساختن سامانه‌های هوش مصنوعی وابسته به مجموعه‌داده‌های بزرگ حیاتی است. اصول پاسخگویی و کمینه‌سازی داده در قانون حفاظت از داده‌های شخصی را می‌توان بر سامانه‌های هوش مصنوعی اعتماد کرد تا اطمینان حاصل شود که تنها داده‌های ضروری پردازش می‌شود و سازمان‌ها در قبال حفاظت از داده پاسخگو هستند [۲۷].

- **چارچوب اخلاق هوش مصنوعی استرالیا، قانون امنیت زیرساخت‌های حیاتی**

چارچوب اخلاق هوش مصنوعی استرالیا اصولی را برای حصول اطمینان از استقرار اخلاقی هوش مصنوعی تعیین می‌کند که شامل ملاحظاتی در خصوص امنیت و حریم خصوصی می‌شود. می‌توان اصل «حفاظت از حریم خصوصی و امنیت» در این چارچوب اخلاق هوش مصنوعی را برای اجرای تدابیر حفاظت از داده استوار در سامانه‌های هوش مصنوعی به کاربست و ریسک نقض داده و دسترسی غیرمجاز را کاهش داد [۲۸]. همچنین، قانون امنیت زیرساخت‌های حیاتی در استرالیا، رویه‌های سخت‌گیرانه‌ای را برای حفاظت از زیرساخت‌های حیاتی، از جمله سامانه‌های هوش مصنوعی مستقرشده در چنین بخش‌هایی، مقرر می‌کند. اعمال الزامات مدیریت ریسک و گزارش‌دهی به سامانه‌های هوش مصنوعی این قانون می‌تواند انعطاف‌پذیری برنامه‌های هوش مصنوعی را در بخش‌های حیاتی مانند انرژی و مالی تقویت کند [۲۹].

- **قانون پایه امنیت سایبری ژاپن، دستورالعمل‌های بهره‌برداری از هوش مصنوعی**

قانون پایه امنیت سایبری ژاپن مبنایی قانونی برای تلاش‌های ملی امنیت سایبری این کشور ایجاد می‌کند و بر حفاظت از زیرساخت‌های اطلاعاتی حیاتی و اهمیت امنیت سایبری در توسعه هوش

مصنوعی تأکید دارد. با بهره‌گیری از این تأکید می‌توان پیاده‌سازی تدابیر امنیت سایبری استوار در سامانه‌های هوش مصنوعی به‌کارگرفته‌شده در بخش‌های حیاتی را هدایت کرد [۳۰]. دستورالعمل‌های بهره‌برداری از هوش مصنوعی در ژاپن به ترویج استفاده ایمن و امن از هوش مصنوعی می‌پردازد و بر شفافیت، پاسخگویی و اعتماد کاربر تأکید می‌کند. می‌توان با اعمال تمرکز این دستورالعمل بر شفافیت اطمینان حاصل نمود که سامانه‌های هوش مصنوعی توضیحات مشخصی از فرایندها و تصمیمات خود ارائه می‌دهند که به تشخیص و پیشگیری از فعالیت‌های مخرب کمک می‌کند [۳۱].

- **قانون حفاظت از اطلاعات شخصی کره جنوبی، راهبرد ملی هوش مصنوعی**

قانون حفاظت از اطلاعات شخصی کره جنوبی (PIPA) حفاظت قدرتمندی را برای داده‌های شخصی ارائه می‌کند که برای سامانه‌های هوش مصنوعی که با اطلاعات حساس کار می‌کنند، ضروری است. الزامات سخت‌گیرانه حفاظت از داده در قانون حفاظت از اطلاعات شخصی را می‌توان برای اطمینان از پردازش و ذخیره‌سازی امن داده‌های شخصی در سامانه‌های هوش مصنوعی اعمال نمود و ریسک درز و نقض داده را کاهش داد [۳۲]. راهبرد ملی هوش مصنوعی کره جنوبی طرح‌هایی را برای توسعه هوش مصنوعی شرح می‌دهد و بر نیاز به تدابیر امنیت سایبری استوار برای حمایت از نوآوری در هوش مصنوعی تأکید می‌کند. تمرکز این راهبرد بر امنیت سایبری می‌تواند هدایتگر یکپارچه‌سازی اصول امنیت در طراحی[1] در توسعه سامانه‌های هوش مصنوعی باشد و اطمینان حاصل نماید که سامانه‌های هوش مصنوعی از آغاز در برابر تهدیدهای سایبری استوار باشند [۳۳].

- **راهنمای حاکمیت و اخلاق هوش مصنوعی آسه‌آن، چارچوب حاکمیت داده دیجیتال آسه‌آن، نظام جهانی قوانین حریم خصوصی فرامرزی**

ترویج همکاری میان کشورها باعث تحکیم همسویی منطقه‌ای در مقررات امنیت سایبری هوش مصنوعی می‌شود که موجب تسهیل توسعه استانداردها و دستورالعمل‌های منطقه‌ای می‌شود که از همکاری فرامرزی در مواجهه با ریسک‌های امنیت سایبری هوش مصنوعی حمایت می‌کنند. نمونه‌هایی از این چارچوب‌ها عبارت‌اند از راهنمای حاکمیت و اخلاق هوش مصنوعی آسه‌آن [۳۴]، چارچوب حاکمیت داده دیجیتال آسه‌آن و نظام جهانی قوانین حریم خصوصی فرامرزی (CBPR). چارچوب حاکمیت داده دیجیتال آسه‌آن، حفاظت از داده و حریم خصوصی را در میان کشورهای عضو آسه‌آن ترویج می‌دهد و مبنایی برای همسویی تدابیر امنیت سایبری هوش مصنوعی فراهم می‌آورد. با اجرای اصول این چارچوب در سامانه‌های هوش مصنوعی می‌توان از وجود رویه‌های حفاظت از داده یکدست در سراسر منطقه اطمینان حاصل نمود و امنیت و قابلیت اعتماد برنامه‌های هوش مصنوعی را ارتقا داد [۳۵]. نظام قوانین حریم خصوصی فرامرزی جریان‌های داده فرامرزی را تسهیل می‌کند و در

1. security-by-design

عین حال اطمینان حاصل می‌نماید که معیارهای سطح بالایی در زمینه حریم خصوصی برقرار هستند. همسو ساختن تدابیر امنیت سایبری هوش مصنوعی با نظام قوانین حریم خصوصی فرامرزی می‌تواند از اشتراک و پردازش امن داده‌ها در برنامه‌های هوش مصنوعی پشتیبانی کند و در عین حفاظت از حریم خصوصی کاربران، نوآوری را ارتقا دهد [۳۶].

۲-۱-۴ همسوسازی با استانداردهای بین‌المللی معتبر

علاوه بر بررسی چارچوب‌های منطقه‌ای، یک روش دیگر برای تقویت سیاست‌های امنیت سایبری، نگاه به استانداردهای شناخته‌شده بین‌المللی معتبر برای همسوسازی است. این استانداردها دستورالعمل‌های جامعی را ارائه می‌دهند که طیف وسیعی از جنبه‌های امنیت سایبری را از مشخصات فنی تا ملاحظات اخلاقی پوشش می‌دهند و باعث حاصل‌شدن رویکردی جامع در خصوص حاکمیت هوش مصنوعی می‌شوند.

این همسوسازی رویکرد ساختاریافته‌ای را در شناسایی، مدیریت و کاهش ریسک‌ها فراهم می‌آورد که برای حفظ یکپارچگی فناوری‌های هوش مصنوعی ضروری است. سامانه‌های هوش مصنوعی که در چارچوب این ساختار توسعه می‌یابند، می‌توانند برای رعایت یک معیار جهانی از نظر امنیت و اطمینان‌پذیری طراحی شوند و اعتماد کاربران و ذی‌نفعان را جلب کنند. برای منطقه آسیا و اقیانوسیه که مشتمل بر مجموعه متنوعی از اقتصادها و محیطهای نظارتی مختلف است، پیروی از این استانداردها می‌تواند فرایندهای انطباق را هموار نماید و بار نظارتی سازمان‌ها را کاهش دهد و همکاری منطقه‌ای و جهانی را تقویت کند.

- **استانداردهای فنی**

استانداردهای فنی بر اقدامات خاص لازم برای امن ساختن سامانه‌های هوش مصنوعی متمرکزند. این استانداردها دستورالعمل‌های دقیقی را برای اجرای کنترل‌های امنیتی فراهم می‌آورند و اطمینان حاصل می‌نمایند که سامانه‌های هوش مصنوعی به شیوه‌ای امن طراحی و اداره می‌شوند. رعایت آخرین تحقیقات و دستورالعمل‌های امنیتی؛ مانند تحقیقات و دستورالعمل‌های مؤسسه ملی فناوری و استانداردهای ایالات متحده (NIST)، سازمان بین‌المللی استاندارد/ کمیسیون بین‌المللی الکتروتکنیک (ISO/IEC)، سازمان پیشبرد استانداردهای اطلاعات ساختمند (OASIS) و پروژه باز امنیت نرم‌افزاری جهانی (OWASP)؛ برای سازمان‌ها جهت کاهش ریسک‌های خصمانه و همسوسازی با رویه‌های امنیتی پیشرفته ضروری است. در اینجا به برخی از مهم‌ترین سازمان‌های استانداردسازی فنی و چارچوب‌های مرتبط با هوش مصنوعی آن‌ها پرداخته شده است.

- **کارگروه فرعی SC 42 کارگروه فنی مشترک ۱ سازمان بین‌المللی استاندارد (ISO) و کمیسیون بین‌المللی الکتروتکنیک (IEC)**

ISO و IEC اغلب با یکدیگر همکاری می‌کنند؛ در حقیقت، IEC و ISO اولین سازمان‌های بین‌المللی استانداردی هستند که گروه خبره‌ای را برای انجام فعالیت‌های استانداردسازی در زمینه هوش مصنوعی راه‌اندازی کرده‌اند. SC 42 بخشی از کارگروه فنی مشترک (ISO/IEC JTC 1) است. SC 42 تمام بوم‌سازگانی را که سامانه‌های هوش مصنوعی در آن توسعه می‌یابند و مستقر می‌شوند در نظر می‌گیرد. این کارگروه فرعی استانداردهای افقی‌ای را توسعه می‌دهد که مبنایی جهت ایجاد راهکارهای هوش مصنوعی برای صنایع مختلف ارائه می‌دهند. SC 42 با کارگروه‌های فنی IEC و ISO همکاری نزدیکی دارد که بر استانداردهای عمودی مختص به بخش‌ها متمرکزند [۳۷].

- JTC1/SC42 استانداردهای زیر را تدوین نموده‌اند:

- ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system[۱][۳۸]
- ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management[۲][۳۹]
- ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)[۳][۴۰]

دیگر استانداردهای مرتبط با امنیت سایبری شامل ISO/IEC 27001[۴] برای مدیریت امنیت اطلاعات است که الزامات لازم برای ایجاد، اجرا، نگهداری و بهبود مستمر سامانه مدیریت امنیت اطلاعات (ISMS) را مشخص می‌کند [۴۱]. سازمان‌ها می‌توانند با همسوسازی با این استاندارد اطمینان یابند که سامانه‌های هوش مصنوعی آن‌ها از چارچوبی ساختند برای مدیریت اطلاعات حساس و کاهش مجموعه‌ای گسترده از ریسک‌های امنیتی پیروی می‌کند.

- **چارچوب مدیریت ریسک هوش مصنوعی و چارچوب امنیت سایبری مؤسسه ملی فناوری و استانداردهای ایالات متحده (NIST)**

چارچوب مدیریت ریسک هوش مصنوعی (AI RMF) مؤسسه NIST یک دستورالعمل داوطلبانه برای کمک به سازمان‌ها در توسعه و مدیریت سامانه‌های هوش مصنوعی است. این چارچوب مجموعه‌ای از فرایندها و فعالیت‌ها را تشریح می‌کند که سازمان‌ها می‌توانند برای مدیریت ریسک‌ها در سراسر طول عمر سامانه‌های هوش مصنوعی به کار گیرند و به همکاری ذی‌نفعان متنوع می‌پردازد [۴۲]. مؤسسه

---

۱. استاندارد ملی ایران/ایزو/آی‌ای‌سی ۴۲۰۰۱: فناوری اطلاعات - هوش مصنوعی - نظام مدیریت
۲. استاندارد ملی ایران ۲۳۷۳۳: فناوری اطلاعات – هوش مصنوعی – راهنمای مدیریت مخاطره (ریسک)
۳. استاندارد ملی ایران ۲۳۶۸۱: چارچوب سامانه‌های هوش مصنوعی با استفاده از یادگیری ماشین
۴. استاندارد ملی ایران ۲۷۰۰۱: امنیت اطلاعات، امنیت سایبری و حفاظت از حریم خصوصی – سامانه مدیریت امنیت اطلاعات - الزامات

NIST در کنار چارچوب مدیریت ریسک هوش مصنوعی، یک ابزار را با نام پلی‌بوک (Playbook) در بستر گیت‌هاب معرفی کرده است که توصیه‌های بیشتری در مورد اقدامات، منابع و مستنداتی که سازمان‌ها می‌توانند به کار بگیرند ارائه می‌کند [۴۳]. انتظار می‌رود که این سند زنده در واکنش به پیشرفت‌های فناوری هوش مصنوعی و تغییرات چشم‌انداز ریسک تکامل یابد [۴۴].

مؤسسه NIST در پاسخ به رشد هوش مصنوعی مولد، یک نمایه میان‌بخشی و منبع همراه برای چارچوب مدیریت ریسک هوش مصنوعی به نام «نمایه هوش مصنوعی مولد» ایجاد کرد. این نمایه سعی دارد ریسک‌های نوظهور یا ریسک‌هایی را که می‌توانند با استفاده از هوش مصنوعی مولد تشدید شوند تعریف کند. این نمایه همچنین مجموعه‌ای از اقدامات را ارائه می‌دهد که سازمان‌ها می‌توانند برای «حاکمیت، نگاشت، اندازه‌گیری و مدیریت» ریسک‌های مذکور اتخاذ کنند [۴۵].

هدف چارچوب امنیت سایبری NIST کمک به سازمان‌ها در توسعه یک راهبرد امنیت سایبری جامع است که به ریسک‌های منحصربه‌فرد سامانه‌های هوش مصنوعی بپردازد. این چارچوب، یک چارچوب سیاست برای راهنمایی امنیتی رایانه ارائه می‌کند که شامل پروتکل‌های دقیقی برای شناسایی، حفاظت، تشخیص، پاسخ‌دهی و بازیابی از تهدیدات سایبری می‌شود [۴۶]. به‌عنوان‌مثال، می‌توان از دستورالعمل‌های چارچوب NIST در مورد ارزیابی ریسک برای شناسایی آسیب‌پذیری‌های بالقوه در مدل‌های هوش مصنوعی، مانند حملات خصمانه و مسمومیت داده، استفاده کرد و اقدامات کاهشی مناسب برای این آسیب‌پذیری‌ها اجرا کرد؛ این نوع راهنمایی در اکثر قوانین محلی دیده نمی‌شود.

- **سایر استانداردهای شناخته‌شده بین‌المللی**

پروژه باز امنیت نرم‌افزاری جهانی (OWASP)

پروژه باز امنیت نرم‌افزاری جهانی (OWASP) یک بنیاد غیرانتفاعی است که هدف آن بهبود امنیت نرم‌افزاری است. این بنیاد در زمینه امنیت و حریم خصوصی هوش مصنوعی راهنمایی ارائه می‌دهد تا بینش‌هایی در خصوص «طراحی، ایجاد، آزمون و تهیه» سامانه‌های هوش مصنوعی با اولویت امنیت و حریم خصوصی ارائه دهد [۴۷].

چشم‌انداز تهدیدات خصمانه MITRE برای سامانه‌های هوش مصنوعی (ATLAS)

MITRE ATLAS یک پایگاه دانش از تاکتیک‌ها، فنون و مطالعات موردی سامانه‌های هوش مصنوعی است. هدف آن آگاه‌ساختن دولت‌ها، صنعت و دانشگاه‌ها از تهدیدات موجود علیه سامانه‌های هوش مصنوعی در دنیای واقعی، امکان‌پذیر ساختن ارزیابی تهدیدات و تشکیل تیم قرمز[1] و مستندسازی حملات خصمانه به سامانه‌های هوش مصنوعی است و چارچوبی برای ارزیابی تهدیدات و تشکیل تیم

1. red teaming

قرمز داخلی ارائه می‌کند [۴۸].

ائتلاف هوش مصنوعی امن (CoSAI)

ائتلاف هوش مصنوعی امن (CoSAI) یک کنسرسیوم بین‌المللی متن‌باز[1] است که به دنبال رسیدگی به مسائل امنیتی هوش مصنوعی از طریق ارائه «روش‌شناسی‌های متن‌باز و ابزار و چارچوب‌های استانداردسازی شده» است [۴۹]. ائتلاف CoSAI به جریان‌های کاری مختلفی مرتبط با صنعت و دانشگاه‌ها در زمینه‌هایی مانند امنیت زنجیره تأمین نرم‌افزار برای سامانه‌های هوش مصنوعی، آمادگی دفاع‌کنندگان برای یک چشم‌انداز امنیتی در حال تغییر و حاکمیت امنیت هوش مصنوعی تقسیم شده است [۵۰]. این ابتکارات بر یکپارچه‌سازی و استفاده امن از هوش مصنوعی در سراسر سازمان‌ها در تمام مراحل توسعه و استفاده متمرکزند.

همسویی با استانداردها و دستورالعمل‌های شناخته‌شده بین‌المللی منجر به یک رویکرد سیاست‌گذاری مبتنی بر ریسک خواهد شد. نهادها باید با پیاده‌سازی مدل‌های جاافتاده‌ای مانند چارچوب امنیت سایبری NIST، چارچوب مدیریت ریسک هوش مصنوعی، COSO ،ISO/IEC 27001 ERM، به‌صورت پیش‌نگرانه[2] تهدیدات امنیت سایبری مرتبط با هوش مصنوعی را شناسایی، ارزیابی و اولویت‌بندی کنند و کاهش دهند. نیاز به چارچوب‌های مدیریت ریسک جامعی که با استانداردهای شناخته‌شده بین‌المللی همسو باشند که به‌دقت بررسی شده باشند، قابل‌چشم‌پوشی نیست. بدین‌طریق اطمینان حاصل می‌شود که منابع به طور کارآمد تخصیص می‌یابد، ضروری‌ترین تهدیدات در اولویت رسیدگی قرار می‌گیرد و وضعیت امنیتی سازمان‌ها به‌طورکلی بهبود می‌یابد.

تعهدات شرکت‌ها به امنیت سایبری و هوش مصنوعی

علاوه بر مقررات منطقه‌ای و استانداردهای بین‌المللی، در چند سال اخیر، شرکت‌های جهانی مختلفی تعهدات قابل‌توجهی را برای تقویت امنیت سایبری و اطمینان از توسعه و استقرار مسئولانه فناوری‌های هوش مصنوعی متقبل شده‌اند. این تعهدات معمولاً بر اهمیت امنیت، شفافیت و ملاحظات اخلاقی در هوش مصنوعی تأکید دارند:

فراخوان رم برای اخلاق هوش مصنوعی (فوریه ۲۰۲۰)

ابتکار مشهور به «فراخوان رم برای اخلاق هوش مصنوعی» توسط واتیکان رهبری شد و شرکت‌های فناوری پیشتازی مانند سیسکو، آی‌بی‌ام و مایکروسافت را به همراه مؤسسات دانشگاهی معتبر گرد هم آورد. این پیمان به دنبال ترویج رویکردی اخلاقی برای توسعه هوش مصنوعی بود و از شش اصل اساسی پیروی می‌کرد که یکی از آن‌ها به طور خاص بر نیاز به امنیت و حریم خصوصی در سامانه‌های هوش مصنوعی تأکید داشت. فراخوان رم بر به رسمیت شناخته‌شدن جهانی چالش‌های اخلاقی و

1. open source
2. proactive

امنیتی ناشی از هوش مصنوعی و مسئولیت جمعی پرداختن به این مسائل تأکید کرد [۵۱].

تعهدات کاخ سفید ایالات متحده در زمینه هوش مصنوعی (ژوئیه و سپتامبر ۲۰۲۳)

دولت بایدن - هریس در ایالات متحده تعهدات داوطلبانه‌ای از هفت شرکت بزرگ هوش مصنوعی، از جمله آمازون، آنتروپیک، گوگل، اینفلکشن، متا، مایکروسافت و OpenAI دریافت کرد. این شرکت‌ها متعهد شدند که توسعه ایمن، امن و شفاف هوش مصنوعی را در اولویت قرار دهند [۵۲]. در سپتامبر ۲۰۲۳، تعهدات بیشتری از شرکت‌هایی مانند ادوبی، Cohere، آی‌بی‌ام، انویدیا، پالانتیر، سیلزفورس، Scale AI و Stability دریافت شد [۵۳]. این تعهدات نشان‌دهنده تلاش هماهنگ دولت و بخش خصوصی برای مدیریت ریسک‌های مرتبط با هوش مصنوعی و اطمینان‌یافتن از توسعه فناوری‌های هوش مصنوعی با تدابیر امنیتی استوار است [۵۴].

اجلاس ایمنی هوش مصنوعی در پارک بلچلی (نوامبر ۲۰۲۳)

اجلاس ایمنی هوش مصنوعی در پارک بلچلی شاهد گردهمایی شرکت‌های جهانی فناوری و دولت‌ها برای تعهد به همکاری در رسیدگی به چالش‌های ایمنی هوش مصنوعی بود. این اجلاس لزوم به‌رسمیت‌شناختن نیاز به همکاری بین‌المللی در مدیریت ریسک‌های هوش مصنوعی را برجسته ساخت و تمرکز ویژه‌ای بر همسوسازی تلاش‌ها در کشورها و صنایع مختلف برای محافظت در برابر تهدیدات بالقوه مرتبط با هوش مصنوعی داشت [۵۵].

اجلاس ایمنی هوش مصنوعی سئول (مه ۲۰۲۴)

در اجلاس ایمنی هوش مصنوعی سئول، شرکت‌های فناوری مانند آمازون، سیسکو، گوگل، متا و مایکروسافت، به همراه دیگر توسعه‌دهندگان برجسته هوش مصنوعی مانند OpenAI و Zhipu AI، پیمانی را امضا کردند تا چارچوب‌هایی را منتشر کنند که تشریح نماید چگونه ریسک‌های مرتبط با مدل‌های هوش مصنوعی «مرزی»[1] خود را اندازه‌گیری می‌کنند و کاهش می‌دهند [۵۶]. این اجلاس بر نقش حیاتی ارزیابی و مدیریت مستمر ریسک در مستقرسازی ایمن فناوری‌های پیشرفته هوش مصنوعی تأکید کرد [۵۷].

پیمان Thorn و All Tech Is Human (۲۰۲۴)

ابتکار Thorn و All Tech Is Human شرکت‌های پیشرو در هوش مصنوعی را گرد هم آورد تا پایبندی خود را به حفاظت از ایمنی برخط کودکان متعهد شوند [۵۸]. این تعهد بخشی از یک تلاش وسیع‌تر برای رسیدگی به اثرات اجتماعی هوش مصنوعی و حصول اطمینان از توسعه و استقرار فناوری‌های هوش مصنوعی به شکلی است که سلامت و رفاه جمعیت‌های آسیب‌پذیر را در اولویت قرار دهد [۵۹].

1. frontier

این تعهدات نشان‌دهنده درک فزاینده از سوی دولت‌ها و شرکت‌ها در خصوص اهمیت امنیت سایبری و ملاحظات اخلاقی در توسعه هوش مصنوعی است. این نهادها با تعهد به این اصول، گام‌های پیش‌نگرانه‌ای را برای رسیدگی به ریسک‌های بالقوه مرتبط با هوش مصنوعی و ترویج یک آینده دیجیتال ایمن‌تر و مطمئن‌تر برمی‌دارند.

۲-۴ اجزای کلیدی برای توسعه چارچوب امنیت سایبری مناسب برای هوش مصنوعی

توسعه یک چارچوب امنیت سایبری هوش مصنوعی استوار مستلزم مبنا قراردادن و تکمیل ساختارهای موجود، همسوسازی با استانداردهای جهانی و رسیدگی به ریسک‌های منحصربه‌فرد سامانه‌های هوش مصنوعی است.

در این تلاش، شش مؤلفه کلیدی وجود دارد: راهنمایی و نظارت، مدیریت چرخه عمر سامانه‌های هوش مصنوعی، حاکمیت و حفاظت از داده، امنیت و استواری مدل، شفافیت و پاسخگویی و پاسخ به حوادث و بازیابی. این مؤلفه‌ها سنگ بنای سامانه‌های هوش مصنوعی امن، اطمینان‌پذیر و قابل‌اعتماد هستند.

۱-۲-۴ راهنمایی و نظارت

نیاز به وجود یک کمیته نظارت میان تخصصی متشکل از مدیران ارشد در یک سازمان وجود دارد که در خصوص رویه‌ها و سیاست‌های هوش مصنوعی مشاوره دهد. این کمیته همچنین نقطه ارجاع و بررسی برای استفاده‌های پرریسک از هوش مصنوعی خواهد بود.	
مرحله کلیدی	**توضیح**
کمیته رهبری	ارائه نظارت رهبری از طریق یک کمیته هوش مصنوعی مسئول متشکل از مدیران ارشد تخصص‌های مختلف (مانند فروش، امنیت، حریم خصوصی، مهندسی، قانونی، حقوق بشر، امور دولتی، منابع انسانی).
نظارت رهبری	مشاوره و پایش سازمان در خصوص رویه‌های مسئولانه هوش مصنوعی و اتخاذ چارچوب حاکمیت هوش مصنوعی.
بررسی موارد استفاده و حوادث	بررسی استفاده‌های پیشنهادی حساس یا پرریسک هوش مصنوعی و مدیریت گزارش‌های حوادث در خصوص سوگیری یا تبعیض.

۲-۲-۴ مدیریت چرخه عمر سامانه هوش مصنوعی

مدیریت چرخه عمر سامانه هوش مصنوعی شامل نظارت بر کل چرخه عمر سامانه‌های هوش مصنوعی می‌شود، از توسعه و استقرار تا پایش و خارج‌سازی از فعالیت. مقررات چرخه عمر اثربخش اطمینان ایجاد می‌نماید که سامانه‌های هوش مصنوعی در طول عمر عملیاتی خود امن، اطمینان‌پذیر و همسو با استانداردهای اخلاقی باقی می‌مانند.	
مرحله کلیدی	**توضیح**
توسعه، طراحی و آموزش	ملاحظات امنیتی و اخلاقی باید از ابتدا در طراحی لحاظ شوند. در این مرحله، استفاده از رویه‌های کدگذاری امن، انجام مدل‌سازی تهدید دقیق و آزمون امنیتی گام‌هایی ضروری هستند. این موضوع شامل انتخاب، آماده‌سازی و یکپارچه‌سازی دارایی‌های آموزشی نیز می‌شود. این مراحل نیز باید همان فرایند امنیت سایبری دقیق را پشت سر بگذارند تا اطمینان حاصل شود که از حملاتی مانند مسمومیت مدل مصون هستند.

مرحله کلیدی	توضیح
استقرار	مدل‌های هوش مصنوعی باید به‌صورت امن در محیط‌های عملیاتی ادغام شوند. پیاده‌سازی کنترل‌های دسترسی قوی، حصول اطمینان از پیکربندی‌های امن و انجام ارزیابی‌های آسیب‌پذیری از اقدامات کلیدی برای حفاظت از سامانه‌های هوش مصنوعی در این مرحله‌اند.
پایش، نگهداری و تحلیل رانش مدل	پایش مداوم سامانه‌های هوش مصنوعی برای شناسایی و پاسخ‌دهی به تهدیدات جدید و اطمینان از عدم رانش مدل از اهداف طراحی آن به شکلی که بردارهای حمله جدیدی را فراهم کند، ضروری است. پیاده‌سازی پایش بی‌درنگ[1]، ثبت وقایع و تشخیص ناهنجاری می‌تواند در شناسایی و کاهش اثر سریع حوادث کمک کند. به‌روزرسانی‌ها و وصله‌های[2] منظم نیز برای رسیدگی به آسیب‌پذیری‌ها ضروری هستند.
خارج‌سازی از فعالیت	سامانه‌های هوش مصنوعی باید هنگام پایان چرخه عمر آن‌ها به طور امن از فعالیت خارج شوند. این امر شامل حذف امن داده‌ها، غیرفعال‌سازی مدل‌ها و اطمینان از عدم وجود ریسک‌های باقی‌مانده می‌شود.

۳-۲-۴ امنیت و استواری مدل

نظارت بر امنیت و استواری مدل‌های هوش مصنوعی برای حفاظت در برابر حملات خصمانه و اطمینان‌یافتن از عملکرد اطمینان‌پذیر ضروری است.

مرحله کلیدی	توضیح
اعتبارسنجی و آزمون مدل	انجام اعتبارسنجی و آزمون دقیق مدل‌های هوش مصنوعی می‌تواند به شناسایی و کاهش آسیب‌پذیری‌های بالقوه کمک کند. این امر شامل آزمون استواری، انصاف و سوگیری می‌شود.
پایش مدل	پیاده‌سازی سامانه‌های پایش و هشدار خودکار می‌تواند به حفظ یکپارچگی و عملکرد مدل‌های هوش مصنوعی کمک کند.

۴-۲-۴ حاکمیت و حفاظت از داده

حاکمیت و حفاظت از داده برای اطمینان از امنیت و حفظ حریم خصوصی داده‌های مورد استفاده در سامانه‌های هوش مصنوعی ضروری است. حاکمیت داده اثربخش شامل پیاده‌سازی سیاست‌ها و روال‌هایی برای مدیریت داده در کل چرخه عمر آن می‌شود، از گردآوری و ذخیره‌سازی تا پردازش و امحا.

مرحله کلیدی	توضیح
حریم خصوصی داده	رعایت قوانین حفاظت از داده؛ مانند مقررات عمومی حفاظت از داده اتحادیه اروپا (GDPR) و قانون حفاظت از داده‌های شخصی (PDPA) سنگاپور امری حیاتی است. پیاده‌سازی ناشناس‌سازی داده، رمزنگاری و کنترل‌های دسترسی می‌تواند به حفاظت از داده‌های حساس کمک نماید.
کیفیت و یکپارچگی داده	درستی و یکپارچگی داده برای عملکرد اطمینان‌پذیر مدل‌های هوش مصنوعی ضروری است. پیاده‌سازی فرایندهای اعتبارسنجی، پاک‌سازی و ممیزی داده می‌تواند به حفظ کیفیت داده کمک کند.
کنترل دسترسی به داده	پیاده‌سازی کنترل دسترسی بر مبنای نقش (RBAC) و اصول حق ویژه کمینه می‌تواند به جلوگیری از دسترسی غیرمجاز به داده‌های حساس کمک کند. بررسی‌ها و ممیزی‌های منظم دسترسی برای حصول اطمینان از رعایت سیاست‌های حاکمیت داده لازم است.

1. real-time
2. patches

۴-۲-۵ شفافیت و پاسخگویی

شفافیت و پاسخگویی برای اطمینان از اینکه سامانه‌های هوش مصنوعی به‌صورت اخلاقی عمل می‌کنند و کاربران می‌توانند به این سامانه‌ها اعتماد داشته باشند ضروری هستند.	
مرحله کلیدی	**توضیح**
توضیح‌پذیری	پیاده‌سازی فنون هوش مصنوعی توضیح‌پذیر (XAI) می‌تواند به ارائه توضیحات واضح و قابل‌فهم درباره تصمیم‌های مدل هوش مصنوعی کمک کند. این کار باعث ارتقای شفافیت می‌شود و به ایجاد اعتماد میان کاربران کمک می‌کند.
قابلیت ممیزی	چارچوب‌هایی که اطمینان حاصل نماید که سامانه‌های هوش مصنوعی قابل ممیزی هستند امکان تصدیق مستقل عملکرد و پیروی این سامانه‌ها از استانداردهای اخلاقی فراهم می‌کند. پیاده‌سازی ثبت وقایع و سوابق ممیزی می‌تواند به تحقق این امر کمک کند.
دستورالعمل‌های اخلاقی	اتخاذ دستورالعمل‌ها و اصول اخلاقی، مانند اصول هوش مصنوعی سازمان توسعه و همکاری اقتصادی (OECD) و ابتکار جهانی IEEE در خصوص اخلاق سامانه‌های خودمختار و هوشمند، می‌تواند به اطمینان از توسعه و استقرار سامانه‌های هوش مصنوعی به‌گونه‌ای که حقوق بشر را رعایت نماید و رفاه اجتماعی را ترویج دهد کمک کند.

۴-۲-۶ پاسخ به حوادث و بازیابی

سازوکارهای عملیاتی پاسخ به حوادث و بازیابی، مؤلفه‌های مهمی در کاهش اثرات نقض‌های امنیتی و محافظت از تداوم سامانه‌های هوش مصنوعی هستند.	
مرحله کلیدی	**توضیح**
برنامه مدیریت امنیت اطلاعات	باید گام‌هایی جهت اتخاذ رویه‌های جامع امنیت سایبری و امنیت زنجیره تأمین مطابق با استانداردهای شناخته‌شده بین‌المللی (مانند ISO و NIST) برداشته شود.
شناسایی و پاسخ به حادثه	سامانه‌های پاسخ به حوادث و پایش بی‌درنگ می‌توانند به شناسایی و پاسخ‌دهی سریع به حوادث امنیتی کمک کنند. تدوین و آزمودن منظم طرح‌های پاسخ به حوادث برای آمادگی ضروری است.
بازیابی از فاجعه	باید طرح‌های بازیابی از فاجعه برای کمک به احیای سریع عملیات هوش مصنوعی پس از وقوع نقض امنیتی یا خرابی سامانه برقرار باشد. آزمودن و به‌روزرسانی منظم این طرح‌ها برای اطمینان از اثربخشی آن‌ها لازم است.
بهبود مستمر	مقرر کردن یک فرایند بهبود مستمر شامل یادگیری از حوادث امنیتی و به‌روزرسانی سیاست‌ها، روال‌ها و فناوری‌ها برای ارتقای انعطاف‌پذیری سامانه‌های هوش مصنوعی می‌شود.

۵ توصیه‌های سیاستی - مدیریت ریسک‌های امنیت سایبری هوش مصنوعی و تقویت حاکمیت هوش مصنوعی

۵-۱ به‌روزرسانی راهبرد ملی امنیت سایبری برای رسیدگی به نگرانی‌های مرتبط با هوش مصنوعی

یکپارچه‌سازی هوش مصنوعی در بخش‌های مختلف، ریسک‌های جدیدی را در حوزه امنیت سایبری ایجاد کرده است که مستلزم پاسخ‌های هدفمند در عرصه سیاست است. مدیریت اثربخش این ریسک‌ها شامل به‌روزرسانی چارچوب‌های موجود و تدوین مقررات و رویه‌های جدید متناسب با چالش‌های منحصربه‌فرد هوش مصنوعی است. این امر موارد شامل موارد زیر می‌شود:

- **تهدیدهای منحصر به هوش مصنوعی:** شناسایی و مقابله با تهدیدهای منحصر به هوش مصنوعی مانند مسمومیت داده، سرقت مدل و حملات خصمانه در چارچوب‌های ملی امنیت سایبری.
- **یکپارچه‌سازی میان‌بخشی:** اطمینان از اینکه تدابیر امنیت سایبری هوش مصنوعی در تمامی بخش‌ها از جمله خدمات درمانی، مالی، حمل‌ونقل و انرژی یکپارچه‌سازی شود تا دفاع منسجمی در برابر تهدیدهای مرتبط با هوش مصنوعی حاصل گردد.
- **مشارکت‌های عمومی - خصوصی:** تشویق همکاری میان دستگاه‌های دولتی و سازمان‌های بخش خصوصی برای مشارکت در توسعه سیاست‌ها و اشتراک اطلاعات تهدید، رویه‌های برتر و منابع مقابله با تهدیدهای امنیت سایبری هوش مصنوعی.
- **توانمندسازی نیروی کار:** تسریع ایجاد فرصت‌های شغلی و نقش‌های جدید با شتاب گرفتن تغییرات نیروی کار جهانی به‌واسطهٔ هوش مصنوعی، فراهم‌سازی دسترسی به برنامه‌های آموزشی برای کارکنان و قادر ساختن کسب‌وکارها به ارتباط با کارکنان ماهر.

۵-۲ تدوین دستورالعمل‌های امنیت سایبری مختص به هوش مصنوعی

توسعه و اجرای دستورالعمل‌های امنیت سایبری مختص به هوش مصنوعی می‌تواند چارچوب روشنی را برای پیروی ارائه دهد و از حفاظت منسجم و اثربخش در بخش‌های مختلف اطمینان ایجاد نماید. این دستورالعمل‌ها باید شامل موارد زیر باشند:

- **نظارت مناسب:** راه‌اندازی نظارت مبتنی بر ریسک بر توسعه و استقرار هوش مصنوعی با تمرکز بر موارد استفاده پرریسک تعریف‌شده به‌صورت دقیق و آثار مرتبط با آن‌ها.
- **پروتکل‌های ارزیابی ریسک هوش مصنوعی:** تدوین پروتکل‌هایی مبتنی بر رویه‌های برتر شناخته‌شدهٔ بین‌المللی برای ارزیابی ریسک‌های امنیت سایبری سامانه‌های هوش مصنوعی

در سراسر چرخه عمر آن‌ها؛ از توسعه تا استقرار و خارج‌سازی از فعالیت.

- **رویه‌های برتر حفاظت از داده:** تعریف رویه‌های برتر حاکمیت داده، از جمله یکپارچگی داده، حریم خصوصی و کنترل دسترسی به‌منظور جلوگیری از حملات مبتنی بر داده به سامانه‌های هوش مصنوعی.
- **تدابیر مبتنی بر ریسک مرتبط با مدل:** پیشنهاد تدابیر امنیتی برای حفاظت از مدل‌های هوش مصنوعی مانند آموزش خصمانه، آزمون استواری و سازوکارهای به‌روزرسانی امن مدل.
- **همسویی با استانداردهای شناخته‌شده بین‌المللی:** چارچوب‌های امنیت سایبری هوش مصنوعی مبتنی بر استانداردهای شناخته‌شده بین‌المللی و در عین حال تطبیق‌پذیر با چالش‌های نوظهور، برای حفاظت در برابر تهدیدهای امنیت سایبری و در نتیجه ترویج نوآوری امن و مسئولانه در حوزه هوش مصنوعی.

۳-۵ سرمایه‌گذاری در تحقیق‌وتوسعه امنیت سایبری هوش مصنوعی

سرمایه‌گذاری در تحقیق‌وتوسعه متمرکز بر امنیت سایبری هوش مصنوعی برای تسلط بر تهدیدهای نوظهور و توسعه راهکارهای نوآورانه حیاتی است. حوزه‌های کلیدی برای سرمایه‌گذاری باید شامل موارد زیر باشد:

- **تشخیص پیشرفته تهدید:** تشویق توسعه روش‌های پیشرفته شناسایی و کاهش تهدیدهای مختص به هوش مصنوعی.
- **چارچوب‌های هوش مصنوعی امن:** ایجاد چارچوب‌ها و ابزارهای امن توسعه هوش مصنوعی که اصول امنیت در طراحی را به کار بسته و ساخت سامانه‌های هوش مصنوعی استوار و قابل‌اعتماد را تسهیل می‌نماید.
- **همکاری با دانشگاه‌ها:** ترویج مشارکت میان نهادهای دولتی، صنعت و مؤسسات دانشگاهی برای پیشبرد پژوهش و نوآوری در امنیت سایبری هوش مصنوعی.

۴-۵ ترویج همکاری و هماهنگی بین‌المللی

چالش‌های امنیت سایبری هوش مصنوعی ماهیت جهانی دارند و رسیدگی اثربخش به آن‌ها مستلزم تلاش‌های هماهنگ بین‌المللی است. سیاست‌گذاران باید موارد زیر را در اولویت قرار دهند:

- **همسوسازی استانداردها:** تلاش در جهت همسوسازی استانداردها و مقررات امنیت سایبری هوش مصنوعی در سراسر منطقه و با شرکای جهانی برای اطمینان از رویکردی متحد در مدیریت ریسک‌های هوش مصنوعی.
- **اشتراک اطلاعات:** ایجاد سازوکارهای فرامرزی برای تبادل اطلاعات مرتبط با تهدید، رویه‌های

برتر و درس‌آموخته‌های حوادث امنیت سایبری هوش مصنوعی.

- **ابتکارات مشترک:** مشارکت در ابتکارات مشترک و پروژه‌های مشارکتی با شرکای بین‌المللی برای توسعه و پیاده‌سازی راهکارهای امنیت سایبری هوش مصنوعی.

۵-۵ ارتقای سواد هوش مصنوعی و توسعه نیروی کار

ایجاد نیروی کار ماهر و ارتقای سواد هوش مصنوعی برای تقویت امنیت سایبری سامانه‌های هوش مصنوعی ضروری است. این امر شامل موارد زیر می‌شود:

- **برنامه‌های آموزشی:** توسعه برنامه‌ها و طرح درس‌های آموزش با تمرکز بر امنیت سایبری هوش مصنوعی که اطمینان حاصل نماید که نسل آینده متخصصان به دانش و مهارت‌های لازم مجهز می‌شوند.
- **آموزش حرفه‌ای:** ارائه برنامه‌های آموزش و گواهی برای متخصصان فعلی به‌منظور آگاه نگه‌داشتن آنان از جدیدترین تهدیدها، فنون و رویه‌های برتر امنیت سایبری هوش مصنوعی.
- **کارزارهای آگاهی‌رسانی عمومی:** اجرای کارزارهای آگاهی‌بخشی عمومی برای آموزش به شهروندان در خصوص ریسک‌های امنیت سایبری هوش مصنوعی و راه‌های حفاظت از خود در برابر تهدیدهای مرتبط با هوش مصنوعی.

۶ جمع‌بندی و پیشنهادات

این گزارش اهمیت حیاتی اطمینان‌یافتن از ایمنی و امنیت سایبری هوش مصنوعی را برجسته می‌سازد و تأکید می‌کند که با یکپارچه‌سازی بیشتر فناوری‌های هوش مصنوعی در بخش‌های مختلف، ریسک‌های همراه با آن نیز باید به طور اثربخش مدیریت شوند.

توسعه یک چارچوب جامع امنیت سایبری هوش مصنوعی مستلزم مبنا قراردادن و تکمیل سیاست‌های موجود، همسویی با استانداردها و رویه‌های شناخته‌شده بین‌المللی و رسیدگی به ریسک‌های منحصربه‌فرد مرتبط با سامانه‌های هوش مصنوعی است. شش مؤلفه کلیدی در این چارچوب نقش اساسی دارند: راهنمایی و نظارت، مدیریت چرخه عمر سامانه هوش مصنوعی، حاکمیت و حفاظت از داده، امنیت و استواری مدل، شفافیت و پاسخگویی و پاسخ به حوادث و بازیابی.

برای کاهش ریسک‌های امنیت سایبری مرتبط با هوش مصنوعی و تقویت حاکمیت هوش مصنوعی، باید راهبردهای ملی امنیت سایبری به‌روزرسانی شوند تا به تهدیدهای مختص به هوش مصنوعی پرداخته شود، تدابیر میان بخش‌ها یکپارچه شود و مشارکت‌های عمومی-خصوصی ترویج داده شود. همچنین لازم است دستورالعمل‌های مختص به هوش مصنوعی ایجاد شود، از جمله

نظارت مبتنی بر ریسک و رویه‌های برتر حفاظت از داده. سرمایه‌گذاری در تحقیق‌وتوسعه امنیت سایبری هوش مصنوعی نیز برای پیشبرد چارچوب‌های امن و تشخیص تهدید ضروری است. نیاز به همکاری بین‌المللی برای همسویی استانداردها و اشتراک اطلاعات مرتبط با تهدید وجود دارد. علاوه بر این، ارتقای سواد هوش مصنوعی و توسعه نیروی کار از طریق برنامه‌های آموزشی و تعلیم حرفه‌ای و کارزارهای آگاهی‌بخشی عمومی برای مجهز کردن افراد به توانایی‌های لازم برای کاهش ریسک‌های مرتبط با هوش مصنوعی حیاتی است.

رسیدگی به پیچیدگی‌های هوش مصنوعی نیازمند رویکردی هماهنگ میان دولت‌ها، صنایع و شرکت‌های فناوری است. ضروری است که اصول درست در طراحی و استقرار فناوری‌های هوش مصنوعی نهادینه شود تا نسبت به نوآوری مسئولانه اطمینان حاصل شود. استانداردهای مشترک و همکاری مستمر برای مسیریابی در چشم‌انداز در حال تحول هوش مصنوعی و امنیت سایبری ضروری خواهد بود، تا منافع هوش مصنوعی محقق شود و در عین حال ریسک‌های آن به طور اثربخش کاهش یابد.

۷ مراجع

1. Cisco’s AI Readiness Index (https://www.cisco.com/c/m/en_us/solutions/ ai/readiness-index. html) that surveyed over 8,000 global companies found 84% of companies think AI would have a very significant or significant impact on their business and 97% of companies say the urgency to deploy AI-powered technologies increased. There is a profound gap between accelerating pace of AI development and how prepared organizations are in adopting it.
2. https://www.ncsc.gov.uk/guidance/ai-and-cyber-security-what-you- need-to-know
3. https://aaai.org/
4. https://eur-lex.europa.eu/legal-content/EN/TXT/ HTML/?uri=CELEX:52021PC0206
5. https://theconversation.com/a-brief-history-of-ai-how-we-got-here- and-where-we-are-going-233482
6. https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history- of-machine-learning-every-manager-should-read/
7. https://news.mit.edu/2023/explained-generative-ai-1109
8. https://aisel.aisnet.org/pacis2021/44
9. https://www.researchgate.net/profile/Mani-Madhukar/ publication/316202671_IBM’s_Watson_Analytics_for_Health_Care/ links/5e2bed2d92851c3aadd7d440/IBMs-Watson-Analytics-for- HealthCare.pdf?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UG FnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19
10. https://www.cisco.com/c/m/en_us/solutions/predictive-networks/index. html
11. https://asean.org/wp-content/uploads/2021/01/fa-220416_DM2025_ email.pdf
12. https://www.kasparov.com/timeline-event/deep-blue/
13. https://www.smartnation.gov.sg/nais/
14. https://www.technologyreview.com/2023/12/29/1084699/machine- learning-earthquake-prediction-ai-artificial-intelligence/
15. https://www.bsp.gov.ph/SitePages/InclusiveFinance/InclusiveFinance. aspx
16. https://newsroom.ibm.com/2021-02-04-EGAT-Adopts-IBM-AI-to-Help- Improve-Efficiency-Across-Thailands-Major-Power-Plants
17. https://www.weforum.org/agenda/2024/02/ai-cybersecurity-how-to- navigate-the-risks-and-opportunities/
18. https://www.dsci.in/resource/content/mitigating-security-privacy- risks-guide-enterprise-use-generative-ai
19. https://csrc.nist.gov/pubs/ai/100/2/e2023/final
20. https://ieeexplore.ieee.org/document/8578273
21. https://arxiv.org/abs/1708.06733
22. https://dl.acm.org/doi/10.5555/3241094.3241142

23. https://www.a10networks.com/blog/5-most-famous-ddos-attacks/
24. https://www.techtarget.com/whatis/feature/SolarWinds-hack- explained-Everything-you-need-to-know
25. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423
26. https://www.pdpc.gov.sg/help-and-resources/2020/01/second-edition- of-model-artificial-intelligence-governance-framework
27. https://sso.agc.gov.sg/Acts-Supp/40-2020
28. https://www.industry.gov.au/publications/australias-artificial- intelligence-ethics-framework
29. https://www.cisc.gov.au/legislation-regulation-and-compliance/soci- act-2018
30. https://www.japaneselawtranslation.go.jp/en/laws/view/3677/en
31. https://www.soumu.go.jp/main_content/000658284.pdf
32. https://elaw.klri.re.kr/eng_mobile/viewer. do?hseq=62389&type=part&key=4
33. https://www.msit.go.kr/bbs/view. do?sCode=eng&nttSeqNo=9&bbsSeqNo=46&mId=10&mPid=9
34. https://asean.org/book/asean-guide-on-ai-governance-and-ethics/
35. https://asean.org/wp-content/uploads/2012/05/6B-ASEAN-Framework- on-Digital-Data-Governance_Endorsedv1.pdf
36. https://www.globalcbpr.org/wp-content/uploads/Global-CBPR- Framework-2023.pdf
37. https://www.iec.ch/ords/f?p=103%3A7%3A704219401578297%3A%3A%3 A%3AFSP_ORG_ID%3A21538
38. https://www.iso.org/standard/81230.html
39. https://www.iso.org/standard/77304.html
40. https://www.iso.org/standard/74438.html
41. ISO/IEC 27001 https://www.iso.org/standard/27001
42. https://www.nist.gov/itl/ai-risk-management-framework
43. https://www.brookings.edu/articles/nists-ai-risk-management- framework-plants-a-flag-in-the-ai-debate/
44. https://www.nist.gov/itl/ai-risk-management-framework
45. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
46. https://www.nist.gov/cyberframework
47. https://owasp.org/www-project-ai-security-and-privacy-guide/
48. https://atlas.mitre.org/pdf-files/MITRE_ATLAS_Fact_Sheet.pdf
49. https://www.oasis-open.org/2024/07/18/introducing-cosai/
50. https://www.coalitionforsecureai.org/about/

51. https://www.romecall.org/the-call/

52. https://www.whitehouse.gov/briefing-room/statements- releases/2023/07/21/ fact-sheet-biden-harris-administration-secures- voluntary-commitments-from-leading-artificial-intelligence- companies-to-manage-the-risks-posed-by-ai/

53. https://www.whitehouse.gov/briefing-room/statements- releases/2023/09/12/ fact-sheet-biden-harris-administration-secures- voluntary-commitments-from-eight-additional-artificial-intelligence- companies-to-manage-the-risks-posed-by-ai/

54. https://www.whitehouse.gov/briefing-room/statements- releases/2023/09/12/ fact-sheet-biden-harris-administration-secures- voluntary-commitments-from-eight-additional-artificial-intelligence- companies-to-manage-the-risks-posed-by-ai

55. https://www.aisafetysummit.gov.uk

56. https://apnews.com/article/south-korea-seoul-ai-summit-uk-2cc2b297 872d860edc-60545d5a5cf598

57. https://aiseoulsummit.kr/press/?mod=document&uid=43

58. https://www.thorn.org/blog/generative-ai-principles/

59. https://www.thorn.org/blog/generative-ai-principles

60. CCAPAC (Coalition for Cybersecurity in Asia-Pacific) Report: AI and Cybersecurity 2024 https://ccapac.asia

نشانی: تهران، انتهای خیابان کارگر شمالی،
پژوهشگاه ارتباطات و فناوری اطلاعات
ایمیل: info@itrc.ac.ir