وزارت ارتباطات و فناوری اطلاعات
پژوهشگاه ارتباطات و فناوری اطلاعات

هوش مصنوعی مولد: چالش‌ها، فرصت‌ها، امنیت و حریم خصوصی

(برگرفته از کارگاه هوش مصنوعی مولد اتحادیه جهانی مخابرات)

تاریخ انتشار

عنوان گزارش: هوش مصنوعی مولد: چالش‌ها، فرصت‌ها، امنیت و حریم خصوصی (برگرفته از کارگاه هوش مصنوعی مولد اتحادیه جهانی مخابرات)
کلمات کلیدی: هوش مصنوعی، چالش ها، امنیت، حریم خصوصی
تهیه کنندگان: ساسان کرمی زاده، طلا تفضلی و افسانه معدنی
ناظر علمی: امیر منصور یادگاری
گروه پژوهشی: گروه ارزیابی امنیت شبکه و سامانه‌ها
نمایندگان کارگروه ارزیابی گزارشات رصدی: نسرین تاج، مرجان بحرالعلوم
تاریخ انتشار: بهار ۱۴۰۳

چکیده

با در نظر گرفتن هوش مصنوعی به عنوان یکی از مهمترین و پیشروترین فناوری‌های حوزه فناوری اطلاعات، اتحادیه جهانی مخابرات[۱] در ۱۹ فوریه ۲۰۲۴ کارگاهی با موضوع هوش مصنوعی مولد با حضور پژوهشگرانی از کشورهای آمریکا، ژاپن، آلمان، انگلیس، فرانسه، چین، اسکاتلند و کره جنوبی برگزار نموده و در آن به طور خاص به چالش‌های امنیتی حول موضوع هوش مصنوعی مولد پرداخته شد. موضوعات مطرح شده در این نشست به طور کلی به مباحثی عمومی در مورد هوش مصنوعی از جمله چرخه حیات‌هوش مصنوعی و ذی‌نفعان آن و به طور ویژه به مخاطرات امنیتی مربوط به این فناوری مربوط می‌شود. همچنین الزامات موجود در نهاد استانداردهای اتحادیه اروپا، کمیته فنی ایمن‌سازی هوش مصنوعی اتحادیه اروپا [۲]و مرکز امنیت سایبری ملی بریتانیا[۳] نیز مورد بررسی قرار گرفته است. در انتها و در بخش نتیجه گیری، مواردی به عنوان خلاصه این نشست ارائه شده است.

[۱] International Telecommunication Union

[۲] European Union

[۳] National Cyber Security Centre is an organization of the United Kingdom Government

فهرست مطالب

۱ مقدمه

هوش مصنوعی مولد محتوا را در قالب‌های مختلف از جمله متن، تصاویر، صوت، انیمیشن، مدل‌های سه بعدی و موارد دیگر در پاسخ به درخواست‌های کاربران تولید می‌کند. چنین تطبیق پذیری با ارائه فرصت‌های امیدوارکننده و همچنین معرفی چالش‌های منحصر به فرد که نیاز به توجه دقیق دارد، تأثیری دوگانه بر امنیت دارد. از یک سو، هوش مصنوعی مولد پتانسیل افزایش قابلیت‌های امنیتی را دارد. از سوی دیگر، مهاجمان را با افزایش مقیاس-پذیری و پیچیدگی آن‌ها توانمند می‌کند. خطرات امنیتی کلیدی شامل جعل عمیق[۱]"یک تکنیک نرم‌افزاری مبتنی‌بر هوش‌مصنوعی است که در محتوای صوتی و تصویری دست می‌برد و آن را به‌دلخواه تغییر می‌دهد " و نقض حق چاپ است. علاوه بر این، هوش مصنوعی مولد تهدیدی برای حریم خصوصی داده‌ها، از جمله نقض داده‌ها، ناشناس‌سازی ناکافی، اشتراک‌گذاری غیرمجاز داده، سوگیری‌ها، عدم رضایت و شفافیت و حفظ ناکافی داده‌ها است. لذا اقدامات جامع امنیتی و حفظ حریم خصوصی برای رفع موثر این نگرانی‌ها ضروری است. در این نشست که در ژنو برگزار شد افراد از کشورهای مختلف در این نشست شرکت کرده بودند. اهداف کارگاه که توسط کمیته راهبری برای کارهای آینده در گروه مطالعاتی ۱۷ در نظر گرفته شده عبارتند از: شناسایی و ارائه یک نمای کلی جامع از مزایا و همچنین چالش‌های امنیتی و حفظ حریم خصوصی مرتبط با برنامه‌های کاربردی مبتنی بر هوش مصنوعی، شناسایی کنترل های فنی و سازمانی امنیتی و حریم خصوصی برای کاهش تهدیدات شناسایی شده، تسهیل تبادل فعالیت‌های جاری در سازمان‌های توسعه استاندارد مربوطه و سایر سازمان‌ها در رسیدگی به امنیت و حریم خصوصی برای برنامه‌های کاربردی مبتنی بر هوش مصنوعی

[۱] Deep fakes

۲ مزایا و چالش‌های مربوط به امنیت و حریم خصوصی

هوش مصنوعی مولد با ارائه فرصت‌های امیدوارکننده و همچنین معرفی چالش‌های منحصربه فردی که نیازمند توجه دقیق است، تأثیری دوگانه بر امنیت دارد. ازیک طرف، هوش مصنوعی مولد پتانسیل افزایش قابلیت‌های امنیتی را دارد. از طرف دیگر، مهاجمان را با افزایش مقیاس‌پذیری و پیچیدگی آن‌ها توانمند می‌کند. در این بخش به بررسی پویایی دوگانه هوش مصنوعی مولد پرداخته شده و چالش‌ها، خطرات و تهدیدات امنیتی و حریم‌خصوصی پیرامون هوش مصنوعی نشان داده شده است. لازم به ذکر است در این نشست پژوهشگرانی از کشورهای کره، آلمان؛ اسکاتلند، امریکا و فرانسه حضور داشتند.

۲-۱ ذی‌نفعان سیستم هوش مصنوعی

هنگام در نظر گرفتن چرخه عمر یک سیستم هوش مصنوعی، ذی‌نفعان زیر در برنامه‌ریزی، طراحی، توسعه، استقرار، بهره‌برداری و نگهداری یک سیستم هوش مصنوعی نقش دارند. که در ادامه به آن ها اشاره شده است.

- مدیریت سازمان: مدیریت سازمان یک ذینفع محسوب می‌شود زیرا استراتژی‌های سیستم هوش مصنوعی، اهداف، منابع، ارزش‌ها، روش‌های ارزیابی می‌بایست با اهداف تجاری آن سازمان هماهنگ باشند.
- معماران مدل هوش مصنوعی: زیرا در معماری مدل هوش مصنوعی نیاز است که الگوریتم‌ها، طراحی و مدل‌ها تنظیم، دقیق و مقیاس‌پذیری و تفسیرپذیری مدل اولویت‌بندی شوند.
- اپراتورهای خدمت دهنده هوش مصنوعی: زیرا این اپراتورها باید خدمات هوش مصنوعی را مدیریت و سیستم‌های هوش مصنوعی را نظارت، به روز رسانی، نگهداری و ایمن کنند.

۲-۲ تهدیدات امنیتی هوش مصنوعی مولد

در این بخش تمرکز بر انواع تهدیدات هوش مصنوعی مولد است. تهدیدات امنیتی شامل استفاده از داده‌های نادرست، سوءاستفاده از مدل‌های هوش مصنوعی، آسیب‌پذیری‌های امنیتی در سیستم‌های هوش مصنوعی، ناامنی در استقرار و استفاده، و تغییرات غیرمنتظره در عملکرد مدل‌ها می‌شود. برای مقابله با این تهدیدات، استانداردهای امنیتی مناسب از جمله رمزگذاری، احراز هویت، مانیتورینگ فعالیت‌ها و آموزش کاربران باید اجرا شود و استفاده از روش‌های بازبینی و آزمون امنیتی نیز توصیه می‌شود. تهدیدات امنیتی و خطرات مرتبط می‌توانند بر تمام مراحل

چرخه حیات هوش مصنوعی تأثیرگذار باشند. در جدول ۱ مخاطرات امنیتی مرتبط با هر یک از مراحل چرخه عمر هوش مصنوعی نشان داده شده است.

جدول ۱ تهدیدات امنیتی با توجه به مراحل چرخه عمر هوش مصنوعی

تهدیدات امنیتی	برنامه ریزی	آماده سازی داده ها	طراحی مدل	آموزش/توسعه مدل	استقرار مدل	عملیات/نظارت و نگهداری
سوگیری[۱] الگوریتمی			×	×		
آسیب کد امنیتی				×		
ناهماهنگ بودن دادهای آموزش		×		×		
مسموم کردن دادهای آموزش		×		×	×	
حمله دور زدن[۲](فرار) (دستکاری ورودی)		×		×	×	
استنتاج عضویت[۳]		×		×		
استخراج مدل[۴]				×	×	
وارونگی مدل[۵]		×		×		
ضعف الگوریتم هوش مصنوعی در ورودی پایگاه داده			×	×	×	
الگوریتم غیرقابل تفسیر[۶]				×		

[۱] Algorithm bias
[۲] Evasion attack
[۳] Membership inference
[۴] Model Extraction
[۵] Model Inversion
[۶] Algorithm unexplainability

حمله مهاجمان				×		
داده‌های آلوده یا تحریف شده				×	×	×

۲-۲-۱ الزامات امنیتی

مدیریت خطرات در طول چرخه عمر هوش مصنوعی نیازمند اجرای تدابیر امنیتی و بهره‌گیری از راهکارهای مناسب است. برخی از الزامات اصلی در این زمینه عبارتند از: توسعه و اجرای استانداردها و راه‌حل‌های امنیتی در طول تمام مراحل چرخه عمر هوش مصنوعی، آموزش و آگاهی کاربران درباره خطرات امنیتی مرتبط با هوش مصنوعی و رفتارهای امنیتی ضروری، ارزیابی مداوم و مدیریت خطرات امنیتی مرتبط با هوش مصنوعی، رصد فعالیت‌ها و پاسخ به حوادث در طول چرخه عمر هوش مصنوعی، استفاده از ابزارهای امنیتی مناسب برای شناسایی و مدیریت خطرات امنیتی، و همکاری و هماهنگی بین تیم‌های مختلف برای اجرای موثر استراتژی‌های امنیتی و مدیریت خطرات امنیتی در طول چرخه عمر هوش مصنوعی. با رعایت این الزامات و اجرای تدابیر مناسب، می‌توان خطرات امنیتی مرتبط با چرخه عمر هوش مصنوعی را کاهش داد و امنیت اطلاعات و سیستم‌های مرتبط با آن‌ها را تضمین کرد. جزئیات بیشتر مربوط به این الزامات امنیتی در جدول ۲ ارائه شده است.

جدول۲ الزامات امنیتی

مرحله چرخه عمر سیستم هوش مصنوعی	الزمات امنیتی
برنامه‌ریزی	• شناسایی و تجزیه و تحلیل تهدیدات امنیتی بالقوه ای • ارایه راهکار برای مقابله و شناسایی تهدیدات
آماده‌سازی داده‌ها	• بررسی عادی یا غیرعادی بودن داده‌ها • آمادگی در برابر تهدیدات امنیتی • حفاظت از مجموعه داده‌های آموزشی در برابر دسترسی، اصلاح یا تخریب غیرمجاز • تهیه مجموعه داده های مناسب برای آموزش و ارزیابی؛ (مثل تمیز کردن، قالب‌بندی و عادی‌سازی مجموعه داده‌ها)

طراحی مدل	• بررسی فرآیند مدیریت ریسک را با در نظر گرفتن کل چرخه حیات سیستم های هوش مصنوعی که شامل شناسایی ریسک، تجزیه و تحلیل ریسک، ارزیابی ریسک و درمان ریسک • بررسی و شناسایی تهدیدات امنیتی و آسیب پذیری های ذاتی سیستم هوش مصنوعی
آموزش/توسعه مدل	• بررسی تهدیدات امنیتی و آسیب پذیری‌های موجود در کتابخانه‌های منبع باز • آموزش تکنیک های لازم به توسعه دهندگان جهت کاهش حملات • بررسی ثبات عملیاتی کتابخانه‌های منبع باز موجود در سیستم هوش مصنوعی • استفاده از محیط‌های امن برای آموزش داده‌ها
توسعه	• اجرای فرآیند مدیریت ریسک در مرحله استقرار و عملیات سیستم هوش
عملیات / نظارت و نگهداری	• ارایه راهنما برای استفاده صحیح از خدمات هوش مصنوعی • نظارت مداوم بر مجموعه داده‌ها و مدل های یادگیری

۲-۲-۲ حفاظت از داده‌ها برای هوش مصنوعی

نحوه تعامل افراد با هوش مصنوعی، بسته به زمینه و هدف تعامل آن‌ها می‌تواند تأثیرات متفاوتی داشته باشد. به طور کلی، سه مسیر اصلی وجود دارد که از طریق آن افراد و سازمان ها به هوش مصنوعی دسترسی پیدا می کنند:

۱. دسترسی کاربران (مانند کارمندان، شهروندان و غیره) به پلتفرم‌های هوش مصنوعی در دسترس عموم همانند ChatGPT، BARD، و غیره. این موضوع، خطر سرقت داده‌ها را برای سازمان‌ها به همراه دارد.

۲. کارمندان یا دانش آموزانی که از هوش مصنوعی ارائه شده توسط سازمان ها یا مؤسسات آموزشی آن‌ها از طریق برنامه های کاربردی یا وب سایت های شرکتی استفاده می کنند. برای این موضوع در حالی که هنوز خطرات استخراج داده را به همراه دارد، می‌توان یک چارچوب امنیتی قرار داد که رعایت گردد.

۳. سازمان‌هایی که مدل‌ها یا برنامه‌های کاربردی هوش مصنوعی خود را برای استفاده داخلی توسعه می‌دهند. در این مورد نیز ممکن است داده ها توسط دیگران قابل دسترسی باشد. زیرا سوالاتی در مورد داده های مورد استفاده برای آموزش و اشتراک گذاری بالقوه آن داده ها مطرح می شود.

۲-۲-۳ خطرات امنیتی هوش مصنوعی مولد

این قسمت در مورد ملاحظات مختلف مربوط به خطرات امنیتی مرتبط با هوش مصنوعی مولد است. این موارد شامل مسائلی مانند دسترسی به خدمات، مراجع صدور مجوز، تهدیدات امنیتی احتمالی، حفظ حریم خصوصی، شفافیت، رعایت قوانین و حقوق بشر است. برخی از چالش‌های خاص ذکر شده در این خصوص شامل چالش‌های حریم خصوصی مرتبط با داده‌های ارائه شده، چارچوب‌های قانونی کنترل داده‌ها و مکانیسم‌های نظارتی است. همچنین در این قسمت به چالش‌های قانونی مرتبط با هوش مصنوعی، مانند حقوق شرکت‌های هوش مصنوعی برای استفاده از داده‌های آموزشی و اصالت محتوای تولید شده توسط هوش مصنوعی اشاره شده است. در این بخش همچنین چالش‌های نظارتی، شامل قوانین در اتحادیه اروپا و ایالات متحده در خصوص حاکمیت هوش مصنوعی نیز بحث شده است.

۲-۳ مرکز امنیت سایبری ملی بریتانیا

مرکز امنیت سایبری ملی بریتانیا، در کارگاه چالش‌ها و فرصت‌های هوش مصنوعی را چنین عنوان کرده است .چالش‌ها شامل عجله در پیاده‌سازی سیستم‌های هوش مصنوعی، ایجاد سیستم‌های خطرناک به‌طور ناخواسته، عدم حفاظت از داده‌ها در فرآیند توسعه سیستم، آسیب‌پذیری در برابر حملات هوش مصنوعی و عدم وجود کنترل‌های مورد انتظار مشابه سیستم‌های معمولی است. بنابراین، با وجود قابلیت‌های چشم‌گیر هوش مصنوعی مولد، اعتماد عمومی به آن کم است. هرچندفرصت‌هایی نیز در این حوزه وجود دارد از جمله استفاده مدافعان از هوش مصنوعی مولد برای مقابله با حملاتی که مهاجمان از آن استفاده می‌کنند شامل خودکارسازی وظایفی که در حال حاضر وقت مدافعان را می‌گیرد و تجزیه و تحلیل کارآمد داده‌های بلادرنگ بدون صرف وقت زیاد است.

۳ کنترل و حریم خصوصی

این بخش بر شناسایی کنترل‌هایی برای کاهش نگرانی‌های امنیتی و حفظ حریم خصوصی در مورد هوش مصنوعی مولد تمرکز می‌کند و دیدگاه‌های جدیدی را درباره نحوه تنظیم کنترل‌های موثر برای حفظ امنیت و حفظ حریم خصوصی به اشتراک می‌گذارد. لازم به ذکر است در این نشست پژوهشگرانی از کشورهای چین، امریکا، افریقای جنوبی، کره جنوبی، هلند تحقیقات خود را ارایه دادند.

۳-۱ بررسی در حکمرانی هوش مصنوعی مولد

این بخش به حکمرانی هوش مصنوعی مولد در طول چرخه عمر آن وکنترل‌ها و ریسک‌های مرتبط تمرکز دارد. همچنین دربرگیرنده مسائل امنیتی، خطرات مرتبط با محتوای تولید شده، خطرات ساخت مدل و مسائل حقوق مالکیت معنوی است.ضمناً به موضوعات مربوط به حقوق مالکیت فکری، که ممکن است در حفاظت از داده‌های آموزشی نقض شود، پرداخته شده است. این نشست به امنیت در طول چرخه عمر ارائه خدمات یا عملکرد سیستم، از جمله آموزش، راه‌اندازی سرویس، تولید محتوا و توزیع محتوا می‌پردازد. همچنین به جای تمرکز صرفاً بر آموزش مدل یا تولید محتوا، تأکید بر اجرای کنترل‌ها در طول کل چرخه عمر قرار دارد. این کنترل‌ها باید در طول چرخه عمر اعمال شوند، به‌ویژه برای داده‌های آموزشی که نیازمند ارزیابی منابع داده برای فیلتر کردن، برچسب‌گذاری و بازرسی هستند. استانداردهای موجود برای مقابله با این مسائل باید مورد مشاوره قرار گیرند. برای حفاظت از اطلاعات شخصی، الزامات نظارتی در مناطق مختلف باید تدوین گردد . درباره بخش مدل، باید گفت ایجاد و استفاده از استانداردها برای ارزیابی مدل امری بسیار حیاتی است. معیارها می‌توانند با استفاده از روش‌های مختلفی مانند ایجاد پرسش‌های ریسک بر اساس پایگاه دانش ریسک به صورت دستی یا استفاده از مدل دیگری برای تولید آن‌ها ساخته شوند و مدل‌ها با استفاده از پرسش‌ها ارزیابی شوند. به‌علاوه، تعبیه امنیت در مدل‌ها امری ضروری است که از طریق روش‌های مختلفی مانند معرفی عمدی نمونه‌های مجازی متنوع به داده‌های آموزشی برای تقویت آن انجام می‌شود. این تنها برای اهداف آموزشی است و در نهایت یک رویکرد مدل سنتی ممکن است برای پردازش پرسش‌های ریسک مناسب باشد.

۳-۲ اکوسیستم هوش مصنوعی و کنترل های امنیتی حیاتی

در این بخش از نشست به موضوع اکوسیستم هوش مصنوعی[۱] و کنترل های امنیتی حیاتی پرداخته شد و همچنین بیان شد که فناوری‌های هوش مصنوعی، استانداردهای فنی و شیوه‌های مورد استفاده، در حال افزایش و عملیاتی شدن هستند. با توجه به اینکه هوش مصنوعی امروزه در همه جا کاربرد پیدا کرده است دارای چالش‌هایی از قبیل مسائل حقوقی متعدد به ویژه تخلفات، حقوق مالکیت معنوی، حقوق بشر، کشف و تعارض قانون است. برای درک این چالش‌ها نیاز است تا آسیب‌پذیری‌ها، تهدیدات و حملات هوش مصنوعی و اینکه چگونه ممکن است این حملات ازیکدیگرمتفاوت باشند را باید درک نمود. به علاوه نیاز است تا آگاهی از محوریت استانداردهای امنیتی در این زمینه وجود داشته باشد.

۳-۳ انواع حملات به هوش مصنوعی مولد

در این بخش از نشست، در مورد حملاتی که به هوش مصنوعی مولد می‌شود پرداخته شد. می‌توان این حملات را بصورت زیر دسته بندی نمود.

- ✓ بدافزاری که به دلخواه، دفاع را دور می‌زند
- ✓ داده های آموزشی مسموم – دستکاری ورودی/سریع
- ✓ الگوریتم‌های هوش مصنوعی به خطر افتاده
- ✓ وارونگی مدل (زیرا عوامل تهدید خروجی مدل را برای استنتاج دستکاری می کنند)
- ✓ داده های آموزشی حساس - خطرات حفظ حریم خصوصی
- ✓ حملات اکتشافی - مدل رفتار، آسیب پذیری و حساسیت سرقت اطلاعات
- ✓ مدلDDoS
- ✓ سرویس‌های ابری نسل سوم هوش مصنوعی در معرض خطر

حملاتی که مدل‌های مولد هوش مصنوعی، یادگیری عمیق و هوش مصنوعی را به خطر می اندازد عبارتند از:

- ✓ بدافزار چند شکلی و دگرگونی در سطحی دیگر
- ✓ جعل عمیق[۲]های کاملاً قوی - عدم وجود چک کننده های یکپارچگی خوب

---

[۱] AI Ecosystem

[۲] Deepfake

- ✓ اخبار جعلی (اطلاعات دروغ و اطلاعات نادرست)
- ✓ الگوریتم‌ها/مدل‌های هوش مصنوعی به خطر افتاده
- ✓ اطمینان کامل از اطلاعات نادرست
- ✓ الگوریتم‌ها/مدل‌های هوش مصنوعی کامل آموزشی بر روی داده‌های مسموم

در ادامه در این بخش از نشست، در مورد کارهایی که برای دفاع سایبری، کنترل‌های تشخیص و پیشگیری می‌توان انجام داد موارد ذیل مطرح گردید:

- ✓ بهبود تشخیص بدافزار در ترافیک رمزگذاری شده
- ✓ بهبود احراز هویت مداوم
- ✓ آدرس راهنما بدافزار رمزگذاری شده
- ✓ بهبود تشخیص تهدید
- ✓ شناسایی و بیرون کردن مجرمان سایبری
- ✓ بهبود تشخیص و پیشگیری از نفوذ

هوش مصنوعی مولد برای تشخیص ناهنجاری/تهدید

- ✓ رفتار غیرمعمول کاربر/فرایند

تجزیه و تحلیل رفتاری امنیت سایبری

- ✓ ساخت مجموعه ای از داده‌های تهدید

توسعه مدل امن

- ✓ محفظه‌های امن، ماشین‌های مجازی و نمونه‌های مجازی، مدل‌های امن و آموزش داده‌ها

نظارت بر زمان واقعی و هوش مصنوعی حفظ حریم خصوصی

- ✓ راه حل های امنیت سایبری مبتنی بر هوش مصنوعی
- ✓ داده های مصنوعی – چالش‌های مربوط به حریم خصوصی

تسهیل تحقیقات فارنزیک دیجیتال

- ✓ کاهش زمان لازم برای حل و فصل پرونده های قضایی مربوط به جرایم سایبری
- ✓ بهبود مجازات‌های مرتبط با جرایم سایبری
- ✓ بهبود انتساب سایبری برای بازیگران تهدید دولت ملی
- ✓ تحلیلگران امنیت سایبری که می‌توانند تهدیدها را سریعتر و کارآمدتر شناسایی کرده و به آن‌ها پاسخ دهند.

۳-۴ استفاده ایمن از هوش مصنوعی مولد

اقدامات فنی برای استفاده ایمن از هوش مصنوعی مولد و همچنین جلوگیری از خطرات و مسائل امنیتی هنگام استفاده از هوش مصنوعی مولد عبارتند از:

- ✓ حریم خصوصی و محرمانه بودن داده ها
- ✓ مسائل امنیتی شخص ثالث
- ✓ بررسی آسیب پذیری رفتاری هوش مصنوعی
- ✓ مسائل حقوقی
- ✓ تکامل بازیگران تهدید (حمله کنندگان)
- ✓ مسائل مربوط به حق چاپ
- ✓ مسائل تعصب و تبعیض
- ✓ مسائل اعتماد و شهرت
- ✓ بررسی آسیب پذیری امنیتی نرم افزار
- ✓ مسائل مربوط به عملکرد، در دسترس بودن و هزینه
- ✓ مسائل اخلاقی و مقرراتی
- ✓ خطر امنیتی برتر مدل های زبان بزرگ[۱] (LLM)
- ✓ تزریق سریع[۲]
- ✓ مدیریت ناامن خروجی
- ✓ داده های آموزشی مسمومیت
- ✓ مدل انکار خدمات
- ✓ آسیب پذیری زنجیره تامین
- ✓ افشای اطلاعات حساس
- ✓ طراحی افزونه[۳] ناامن
- ✓ عاملیت بیش از حد[۴]
- ✓ اتکای بیش از حد
- ✓ مدل سرقت

---

[۱] Large language models

[۲] Prompt Injection

[۳] Plugin

[۴] Overreliance

۳-۵ کنترل‌های امنیت و حریم خصوصی هوش مصنوعی

در این بخش از نشست به انواع کنترل امنیت و حریم خصوصی هوش مصنوعی پرداخته شده که در ادامه آورده شده است:

- ✓ کنترل‌های امنیتی هوش مصنوعی برای محافظت از داده‌های شخصی
- ✓ کنترل‌های امنیتی هوش مصنوعی برای محدود کردن تأثیر رفتار مدل
- ✓ کنترل‌های اضافی برای محافظت از حقوق حریم خصوصی فردی:
  - اعتبارسنجی هدف (مثلاً تغییر هدف داده های شخصی)
  - داشتن رضایت
  - کنترل سوگیری ناخواسته
  - ارائه شفافیت/توضیح
  - دستیابی به دقت و به روز رسانی داده‌ها
  - ارایه ویژگی‌هایی برای تصحیح، دسترسی، پاک کردن

۳-۶ هوش مصنوعی مولد در مدیریت هویت و ردیابی محتوا

هوش مصنوعی بر روی همه جنبه‌های زندگی بشری تاثیر می‌گذارد. تا سال ۲۰۳۰ هوش مصنوعی زندگی بشری را مورد تاثیر قرار می‌دهد و امنیت و حریم خصوصی آن اهمیت بسزایی دارد. هوش مصنوعی کاربردهای متفاوتی دارد که در شکل ۱نشان داده شده است.

شکل ۱- کاربردهای مختلف هوش مصنوعی

دولت‌ها در سراسر دنیا قوانین و مقرراتی را برای امنیت و حریم خصوصی در هوش مصنوعی تدوین نموده‌اند. هوش مصنوعی مولد ریسک‌های امنیتی متعددی دارد. مانند ریسک تولید اطلاعات نادرست[۱] و احراز هویت. آیا محتوای تولید شده به صورت برخط توسط انسان تولید شده یا به صورت اتوماتیک توسط هوش مصنوعی. به منظور پاسخ به این سوال راهکارهای زیر پیشنهاد گردیده است:

- برچسب‌گذاری باید بر روی محتوای تولید شده انجام گیرد و امنیت به صورت برخط تامین شود تا کاربران بدانندکه محتوا توسط هوش مصنوعی تولید شده است تا تاثیر محتوای گمراه‌کننده بر روی جامعه کاهش یابد. به صورت بین‌المللی، به این مسئله به عنوان برچسب‌گذاری ارجاع می‌شود.

[۱] misinformation

- باید روش‌هایی برای ردیابی محتواهای تولید شده توسط هوش مصنوعی وجود داشته باشد تا از انتشار محتوای گمراه‌کننده و منسوخ ممانعت بعمل آید و همچنین از حق نشر[۱] و ابداع[۲] محافظت شود.

همچنین به منظور کاهش ریسک انتشار محتوای گمراه‌کننده از طریق برچسب‌گذاری باید یک اجماع جهانی صورت گیرد. برچسب‌گذاری محتوای تولید شده توسط هوش مصنوعی باید مطابق با قوانین و مقررات مرتبط با تولیدکنندگان اطلاعات برخط به کمک هوش مصنوعی، اطلاعات تولید شده توسط هوش مصنوعی و مقالات تولید شده توسط هوش مصنوعی، تصاویر و ویدیوها صورت گیرد. مثلا، خروجی هوش مصنوعی مولد باید به گونه‌ای برچسب‌گذاری شود که توسط ماشین تولید شده باشد.

در اتحادیه اروپا برای استفاده از هوش مصنوعی قوانین و مقرراتی تدوین گردیده است. برخی از آن‌ها به شرح ذیل است:

- تولیدکنندگان سیستم‌های هوش مصنوعی، صدا، تصویر و ویدئو یا متن باید مطمئن باشند که خروجی سیستم هوش مصنوعی ساختار قابل خواندن توسط ماشین دارد و مشخص است که توسط هوش مصنوعی تولید گردیده است.
- تولیدکنندگان سیستم‌های هوش مصنوعی که محتوای تصویر، ویدیو و صدا را تولید می‌کنند، مشتمل بر محتوای جعلی، باید ذکر نمایند که محتواهایشان توسط ماشین تولید شده است.

کشور کانادا نیز اصولی را برای مسئولیت‌ها، قابلیت اعتماد و فناوری‌های هوش مصنوعی مولد تدوین نموده است. از قبیل اینکه همه طرف‌ها باید اطمینان یابند که خروجی سیستم‌های هوش مصنوعی که بر روی افراد یا گروه‌ها تاثیر می‌گذارد توسط آنها به عنوان ابزار هوش مصنوعی مولدعلامت‌گذاری شده است.

استانداردسازی باید انجام گیرد تا محتوای تولید شده توسط هوش مصنوعی شناسایی شود. این استاندارد سازی با در نظر گرفتن موارد ذیل باید صورت گیرد:

---

[۱] copyright

[۲] authorship

- برچسب‌گذاری محتوای تولید شده توسط هوش مصنوعی برای اعلان به کاربران وب که محتوا توسط هوش مصنوعی تولید شده است.
  - این کار موجب کاهش ریسک استفاده نادرست و سوء استفاده می‌گردد و از استفاده نادرست کاربران وب از محتوای تولید شده توسط هوش مصنوعی ممانعت بعمل می‌آورد.
- ایجاد مکانیزم ردیابی برای محتوای تولید شده توسط هوش مصنوعی به منظور کاهش یا حذف تاثیر منفی یا نادرست انتشار محتوای تولید شده بر جامعه:
  - ریسک انتشار محتوای گمراه‌کننده: ردیابی محتوای گمراه‌کننده تولید شده توسط هوش مصنوعی.
  - ریسک امنیت مالکیت معنوی: حفظ حریم خصوصی و سایر حقوق قانونی.
- محتوای تولید شده توسط هوش مصنوعی مولد باید برچسب‌گذاری گردد.

چین به منظور تایید هویت محتوای تولیدشده، استانداردهایی را مصوب نموده است.

در سال گذشته چین اولین مستند مربوط به مقررات بر روی هوش مصنوعی مولد را منتشر نمود (راهنمای هوش مصنوعی مولد در آموزش و تحقیق).

در این مستند بیان شده که تولیدکنندگان مدل‌ها و نرم‌افزارهای هوش مصنوعی مولد باید محتواهای تولید شده مانند تصاویر و ویدئوها را مطابق با تولیدات بر روی مدیریت ترکیب‌های داده سرویس‌های اطلاعاتی مبتنی بر اینترنت برچسب‌گذاری نمایند.

استانداردسازی بر روی هویت محتواهای تولید شدهیک راهنما بر روی روش‌های شناسایی محتوای تولید شده توسط سرویسهای هوش مصنوعی مولد منتشر گردیده تا محتوای تولید شده شناسایی گردد:

- علامت گذاری برای اعلان به کاربران و ممانعت از سوء استفاده از محتوای تولید شده
- نمایش پیوسته متن اعلان
- استفاده از ته‌نقش[۱]

[۱] Watermark

- افزودن متن اعلان به تصاویر و ویدئوها
- علامت ترکیب ماشین-انسان برای ممانعت از گیج شدن در زمان استفاده از سرویس‌های مشترک انسان و هوش مصنوعی

۳-۶-۱ پیاده‌سازی معیارهای ردیابی

زمانیکه تصاویر، ویدئو و صوت توسط هوش مصنوعی تولید می‌شود، واترمارک بر روی آن باید ثبت گردد. این واترمارک باید شامل حداقل نام ارائه‌دهنده سرویس و سایر محتواها باشد.

جدول ۳- معیارهای ردیابی پیاده‌سازی شده

نوع فایل	الگوریتم	شناسایی محتوا	روش پیاده‌سازی
تصویر	Self-developed	نام مولد سرویس، شناسه محتوا	Transform domain watermark
ویدئو	Self-developed	نام مولد سرویس، شناسه محتوا	Transform domain watermark
صدا	Self-developed	نام مولد سرویس، شناسه محتوا	Frequency domain watermark

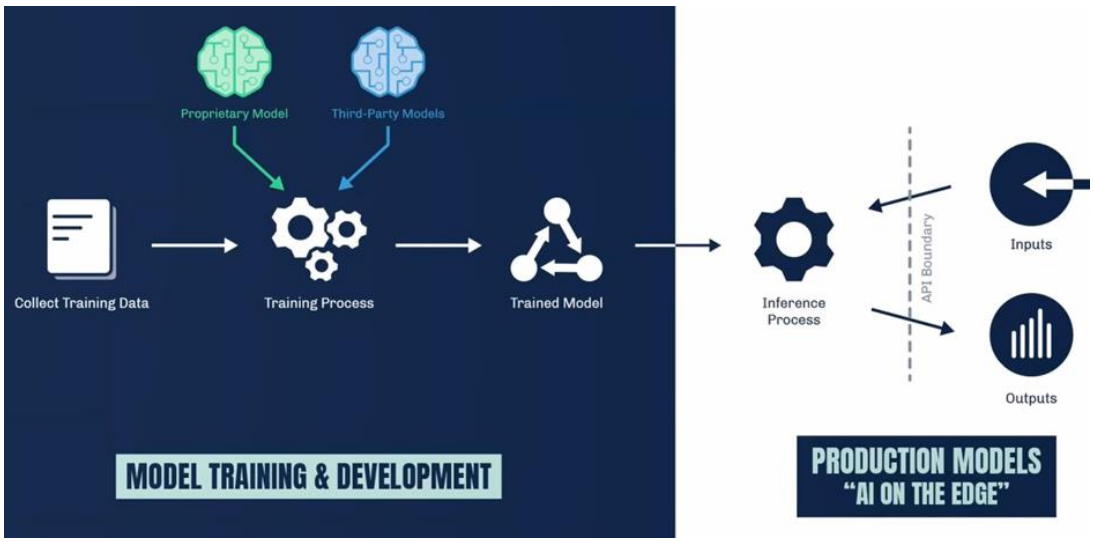
چین چهار استاندارد برای هوش مصنوعی مولدتدوین نموده است.

- نیازمندی‌های اولیه برای امنیت سرویس‌های هوش مصنوعی مولد
- روش‌های شناسایی محتوا برای سرویس‌های هوش مصنوعی
- مشخصات ایمنی برای آموزش اولیه و بهینه‌سازی داده آموزشی در هوش مصنوعی مولد
- استانداردهای ایمنی برای جداسازی دستی در هوش مصنوعی مولد

۳-۷ استفاده ایمن از LLMها و سیستم‌های هوش مصنوعی مولد

فرصت‌های ایجاد شده توسط هوش مصنوعی و ریسک آن زیاد است. هوش مصنوعی می‌تواند تا ۱۵,۷ تریلیون دلار در اقتصاد جهانی تا سال ۲۰۳۰ ایجاد نماید. امنیت تنها ریسکی است که باید در نظر گرفته شود. در طی سال ۲۰۲۲، ۳۰ درصد از حملات هوش مصنوعی موجب بالا رفتن مسمومیت داده آموزشی در هوش مصنوعی می‌شود.

حملات مهاجمین به هوش مصنوعی رو به رشد است به همین دلیل نیاز است تا مدل‌ها پویش شوند و امنیت آن‌ها تایید گردد.



شکل ۲- چرخه حیات هوش مصنوعی

مدل‌های هوش مصنوعی باید به گونه‌ای باشند که حملات به هوش مصنوعی را شناسایی نموده و به آن پاسخ دهند. همچنین عملیات امنیتی متعددی باید بر روی هوش مصنوعی انجام گیرد، مانند کشف، ایمنی و اعتماد، مانیتورینگ حمله، پاسخگویی و آگاهی‌رسانی وضعیتی. هوش مصنوعی یک بردار حمله ناامن است. هیچ راهی برای ایمن‌سازی سیستم‌های هوش مصنوعی توسط فایروال، ابر و آنتی‌ویروس وجود ندارد. مدل هوش مصنوعی امکان عبور جانبی[۱] از شبکه را فراهم می‌نماید. مدل‌های هوش مصنوعی باید پویش شوند تا امنیت و یکپارچگی آنها مورد بررسی قرار گیرد. چگونه شناسایی ریسک‌های هوش مصنوعی و روش‌های مواجهه با آن‌ها در جدول ۴ نشان داده شده است. نیاز است که ریسک‌های هر مدل شناسایی شوند و همچنین باید مدل‌ها اسکن و بر روی آن تست نفوذ انجام شود.

[۱] Lateral

جدول ۴- ریسک‌های هوش مصنوعی مولد و روش‌های مواجهه با آن

حملات به زنجیره تامین	مدیریت تقلب	هوش مصنوعی مولد	مورد مصرف هوش مصنوعی
حملات باج‌افزارها	استنتاج	استنتاج	رایج‌ترین انواع حملات
مسموم‌سازی داده	عبور	مسموم‌سازی داده	
مدل‌های در پشتی	حملات باج‌افزار	تزریق اعلان	
اسکن مدل	پویش مدل	اسکن مدل‌ها	کنترل‌های کلیدی مورد نیاز
محافظت برخط	محافظت برخط	محافظت برخط	

۴ جمع‌بندی و پیشنهادات

در بخش پایانی این نشست فرصت‌های پیش رو برای استانداردهای آینده بررسی شده و توصیه‌هایی به گروه مطالعاتی ITU-T ۱۷ برای کار آینده در این زمینه ارائه شد. همچنین در مواردی که هوش مصنوعی مولد برای امنیت استفاده می شود توضیحاتی از طرف اعضای نشست بیان شد که در ادامه به بعضی از این موارد اشاره میگردد:

- ✓ اتوماسیون دفاع سایبری
- ✓ هوش تهدید
- ✓ تولید و شناسایی کد ایمن
- ✓ شناسایی حملات سایبری
- ✓ شناسایی بدافزار
- ✓ افزایش اثربخشی فناوری های امنیت سایبری

به علاوه در مواردی که از هوش مصنوعی مولد برای حمله استفاده می شود اشاره شد.

- ✓ حملات مهندسی اجتماعی
- ✓ حملات فیشینگ
- ✓ هک خودکار
- ✓ تولید کد باج افزار
- ✓ تولید کد بدافزار

در انتها ازبین اعضای شرکت کننده در موارد مختلف مرتبط با این نشست، سوالاتی مطرح شد که در ادامه این سوالات نیز آورده شده است.

- شکاف‌های استانداردسازی:

۱. شکاف‌های فعلی استانداردهای بین المللی مربوط به امنیت و حریم خصوصی هوش مصنوعی مولد، هوش مصنوعی مولد برای امنیت چیست و چگونه می توان آن‌ها را برطرف کرد؟

۲. چگونه استانداردهای آینده می توانند قابلیت همکاری و سازگاری را در بین فناوری ها و پلت فرم های مختلف هوش مصنوعی مولد تضمین کنند؟

- ملاحظات اخلاقی:

۱. چگونگی استفاده از استانداردها و دستورالعمل‌های اخلاقی، برای هدایت توسعه و استفاده از فناوری‌های هوش مصنوعی مولد موثر است.

- استانداردهای جهانی و همکاری:

۱. همکاری بین‌المللی و تلاش‌های استانداردسازی چگونه می‌توانند باعث افزایش امنیت و حریم خصوصی فناوری‌های هوش مصنوعی مولدشود.

مواردی که در کارگاه هوش مصنوعی مولد اتحادیه جهانی مخابرات مطرح گردیده به طور خلاصه شامل موارد زیر است:

- شاغلین و کاربران هوش مصنوعی نیاز به راهنمایی دارند
- هوش مصنوعی یک فناوری نوظهور است که بر اساس سابقه طولانی کمک رایانه به صنعت، فعالیت های اجتماعی، فعالیت های تجاری و اوقات فراغت شکل گرفته است.
- بسیاری از کارهایی که هوش مصنوعی انجام می دهد تکاملی است
- بسیاری از کارهایی که هوش مصنوعی می تواند انجام دهد انقلابی است
- سرعتی که هوش مصنوعی می تواند انقلابی شود، سریعتر از چیزی است که آمادگی لازم برای آن وجود داشته باشد.
- بسیاری از نگرانی های اجتماعی ریشه در ترس از ناشناخته ها و آزمایش نشده ها دارد
- نهادهای استاندارد می توانند عقلانیت را به بحث هوش مصنوعی بیاورند
- در مورد موضوعات هوش مصنوعی باید با منطق وبه دور از احساسات روبرو شد
- استانداردها باید فقط آنچه را که برای دستیابی به اهداف لازم است مشخص کند و رویکرد خاصی را برای اجرا تحمیل نکند.
- الزام باید شامل تمام اطلاعات لازم برای درک آن الزام باشد، چه به طور مستقیم یا با ارجاع به اسناد دیگر. خواننده استاندارد نباید نیازی به فرضیاتی در مورد اجرای هر الزامی داشته باشد.
- الزامات باید واضح و دقیق، بدون جزئیات غیر ضروری که ممکن است خواننده را سردرگم کند، بیان شود.
- در عناصر فردی نیازمندی ها باید به شیوه ای مناسب و خوانا گنجانده شوند.

- هیچ تناقضی بین الزامات مختلف در استاندارد و یا سایر استانداردهای مرتبط نباید وجود داشته باشد.
- باید ابزار روشن و واضحی وجود داشته باشد که نشان دهد یک پیاده سازی با الزامات مطابقت دارد.

۵ مراجع

۱- ITU Workshop on Generative AI: Challenges and Opportunities for Security and Privacy, Geneva, Switzerland, ۱۹ February ۲۰۲۴

نشانی: تهران، انتهای کارگر شمالی، پژوهشگاه ارتباطات و فناوری اطلاعات، معاونت پژوهش و توسعه ارتباطات علمی

تلفن: ۰۲۱-۸۸۶۳۰۳۵۵

نمابر: ۰۲۱-۸۸۶۳۰۳۵۶