مدل سه‌خط دفاعی در برابر ریسک‌های ناشی از هوش مصنوعی

یوناس شوت - ترجمه: مرتضی اسدی و آرشینا منتظری

خلاصه

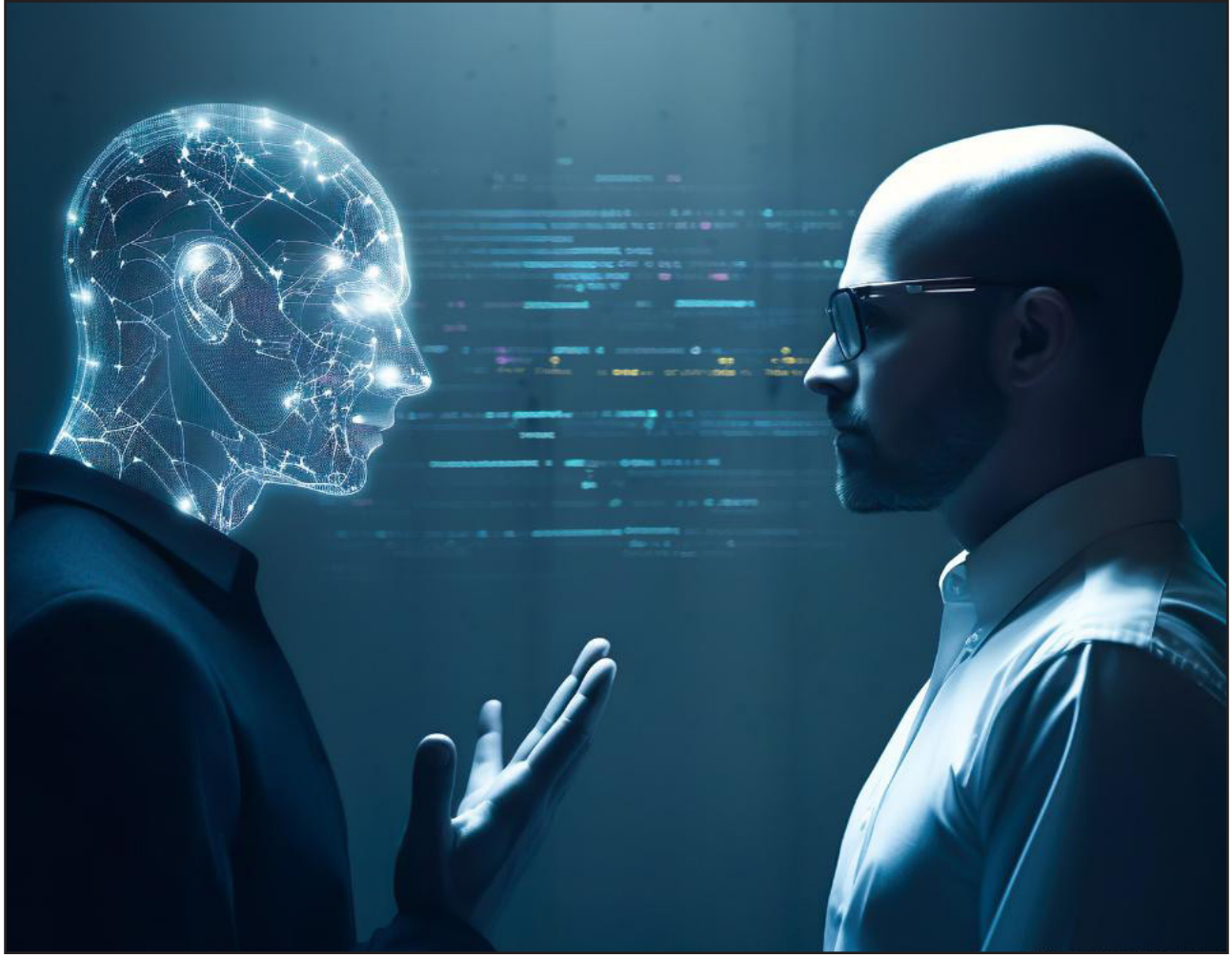
سازمان‌هایی که سیستم‌های هوش مصنوعی(AI) را توسعه می‌دهند و استقرار می‌بخشند، به دلایل اقتصادی، قانونی و اخلاقی، باید ریسک‌های مرتبط را مدیریت کنند. با این حال، همیشه مشخص نیست که چه کسی مسئول مدیریت ریسک‌های ناشی از هوش مصنوعی است. مدل سه‌خط دفاعی[1](3LoD) که بهترین روش برای بسیاری از صنایع در نظر گرفته می‌شود، می‌تواند راه‌حلی در این خصوص ارائه دهد. این مدل یک چارچوب مدیریت ریسک برای کمک به سازمان‌ها است تا نقش‌ها و مسئولیت‌های مدیریت ریسک را تعیین و هماهنگ کنند. این مقاله، روش‌هایی را پیشنهاد می‌کند که شرکت‌های هوش مصنوعی می‌توانند این مدل را پیاده‌سازی کنند. همچنین در مورد این‌که چگونه این مدل می‌تواند به کاهش ریسک‌های ناشی از هوش مصنوعی کمک کند، بحث می‌شود. به‌طوری‌که شکاف‌های پوشش ریسک را مشخص کرده و بتواند اثربخشی شیوه‌های مدیریت ریسک را افزایش دهد و هیأت مدیره را قادر سازد تا به‌طور مؤثرتری بر مدیریت نظارت کند. هدف این مقاله آگاهی‌بخشی به تصمیم‌گیرندگان در شرکت‌های پیشرو هوش مصنوعی، قانون‌گذاران[2] و نهادهای استانداردگذار[3] است.

مقدمه

سازمان‌هایی که سیستم‌های هوش مصنوعی (AI) را توسعه می‌دهند و به کار می‌گیرند، به دلایل اقتصادی، باید ریسک‌های مرتبط را مدیریت کنند، زیرا رویدادها و موارد سوءاستفاده می‌توانند تهدیدی برای عملکرد کسب‌وکار باشند. به دلایل قانونی ممکن است مقررات آتی هوش مصنوعی، آن‌ها را ملزم به اجرای یک سیستم مدیریت ریسک کند و به دلایل اخلاقی، آن‌ها موظف به جلوگیری از آسیب هستند.

بااین‌حال، همیشه مشخص نیست که چه کسی مسئول مدیریت ریسک هوش مصنوعی است: محققان و مهندسان؟ بخش حقوقی و تطبیق[4]؟ تیم راهبری؟ مدل سه‌خط می‌تواند راه‌حلی ارائه دهد. این یک چارچوب مدیریت ریسک است که به منظور بهبود راهبری ریسک سازمان با تخصیص و هماهنگ کردن نقش‌ها و مسئولیت‌های مدیریت ریسک طراحی شده است. این بهترین مدل در بسیاری از صنایع مانند مالی و هوانوردی در نظر گرفته می‌شود. این مقاله، مدل سه‌خط را در زمینه‌ی هوش مصنوعی اعمال می‌کند.

تا به امروز، کارهای آکادمیک زیادی روی اشتراک هوش مصنوعی و مدل سه‌خط انجام نشده است. استفاده از مدل برای کاهش ریسک‌های متمایز[5] ناشی از هوش مصنوعی پیشنهادهایی ارایه می‌دهد، اما این متن مختصر و مفید است. همچنین برخی ادبیات در مورد این‌که چگونه شرکت‌ها می‌توانند از هوش مصنوعی برای حمایت از مدل سه‌خط استفاده کنند وجود دارد، اما عمدتاً به نحوه‌ی اداره‌ی شرکت‌های هوش مصنوعی علاقه‌مند هستم، نه نحوه‌ی استفاده از هوش مصنوعی



برای اداره‌ی شرکت‌های غیر هوش مصنوعی. همچنین پیشنهاد شده است که دولت‌ها می‌توانند از مدل سه‌خط برای مدیریت ریسک‌های شدید ناشی از هوش مصنوعی استفاده کنند، اما در این‌جا روی چالش‌های شرکت‌ها تمرکز می‌شود، نه دولت.

انجمن حسابرسان داخلی[۶] یک مجموعه‌ی سه قسمتی را که در آن چارچوب حسابرسی هوش مصنوعی[۷] را پیشنهاد می‌کند منتشر کرده است، هر چند مطالب آن‌ها حاوی اشاره‌ای به مدل سه‌خط است، اما نقش کلیدی ایفا نمی‌کند. در نهایت، مدل سه‌خط در کتابچه‌ای که توسط مؤسسه‌ی ملی استانداردها و فناوری[۸] (NIST) در کنار چارچوب مدیریت ریسک هوش مصنوعی منتشر کرده، ذکر شده است. با این حال، این مقاله فقط اجرای مدل سه‌خط (یا مکانیسم مشابه) را پیشنهاد می‌کند، نه نحوه‌ی انجام این کار را. روی‌هم‌رفته، حداقل دو شکاف در ادبیات کنونی وجود دارد. مورد اول کاربردی است: به نظر نمی‌رسد برای این‌که چگونه سازمان‌هایی که سیستم‌های هوش مصنوعی را توسعه و استقرار می‌دهند می‌توانند مدل سه‌خط را پیاده‌سازی کنند، پیشنهاد مشخصی وجود داشته باشد و معدود پیشنهادهای موجود آن‌قدر دقیق نیستند تا راهنمایی‌های معناداری را ارائه کنند. مورد دوم دستوری است: به نظر نمی‌رسد بحث کاملی در مورد این‌که آیا اجرای مدل مطلوب است یا نه وجود داشته باشد. با توجه به این‌که این مدل مورد انتقاد قرار گرفته و شواهد تجربی زیادی برای اثربخشی آن وجود ندارد، پاسخ به این سؤال واضح نیست. با توجه به این موضوع، مقاله به دنبال پاسخ به دو سؤال زیر است:

سؤال ۱: سازمان‌هایی که سیستم‌های هوش مصنوعی را توسعه و استقرار می‌دهند چگونه می‌توانند مدل سه‌خط را پیاده‌سازی کنند؟

سؤال۲: اجرای مدل سه‌خط تا چه اندازه به کاهش ریسک‌های ناشی از هوش مصنوعی کمک می‌کند؟

مقاله بر سه حوزه متمرکز است. اول، بر سازمان‌هایی تمرکز می‌کند که سیستم‌های هوش مصنوعی پیشرفته را توسعه می‌دهند و به کار می‌گیرند، به‌ویژه آزمایشگاه‌های تحقیقاتی متوسط (مثلاً[۹] Google DeepMind و OpenAI[۱۰]) و شرکت‌های فناوری بزرگ (مانند مایکروسافت و متا)، اگرچه مرزهای بین این دو دسته مبهم است (به‌عنوان مثال Google DeepMind یکی از شرکت‌های تابعه‌ی Alphabet است و OpenAI با مایکروسافت شراکت استراتژیک دارد). در ادامه از عبارت «شرکت‌های هوش مصنوعی» برای اشاره به همه‌ی آن‌ها استفاده می‌شود. این مقاله انواع دیگر شرکت‌ها (مانند شرکت‌های سخت‌افزار)، غیرانتفاعی یا مؤسسات دانشگاهی را پوشش نمی‌دهد، اما ممکن است آن‌ها نیز از تحلیل‌ها سود ببرند. دوم، این مقاله بر بعد سازمانی مدیریت ریسک هوش مصنوعی تمرکز دارد. موضوع این نیست که چگونه شرکت‌های هوش مصنوعی باید ریسک‌های ناشی از هوش مصنوعی را تشخیص دهند، ارزیابی کنند و به

آن پاسخ دهند در عوض، موضوع برسر چگونگی تعیین و هماهنگی نقش‌ها و مسئولیت‌های مدیریت ریسک است. سوم، مقاله بر توانایی مدل برای جلوگیری از آسیب‌های اجتماعی تمرکز دارد. به ریسک‌هایی که برای خود شرکت‌ها وجود دارد کم‌تر پرداخته شده است (مثلاً ریسک‌های دعاوی حقوقی یا شهرت (اعتبار)[11]، اگرچه گاهی اوقات منافع خصوصی و عمومی همسو می‌شوند (مثلاً یکی از راه‌های کاهش ریسک دعاوی حقوقی ، جلوگیری از اتفاقات است).

ادامه‌ی این مقاله بدین شرح است. بخش دو نمایی کلی از ساختار اساسی مدل، تاریخچه، انتقادها و پایه‌ی شواهد ارائه می‌دهد. بخش ۳ راه‌هایی را پیشنهاد می‌کند که شرکت‌های هوش مصنوعی می‌توانند مدل را پیاده‌سازی کنند. بخش ۴ چگونگی کمک به کاهش ریسک‌های ناشی از هوش مصنوعی را توضیح می‌دهد. بخش ۵ نتیجه‌گیری و سؤالاتی را برای تحقیقات بیشتر پیشنهاد می‌کند.

۲. مدل سه‌خط

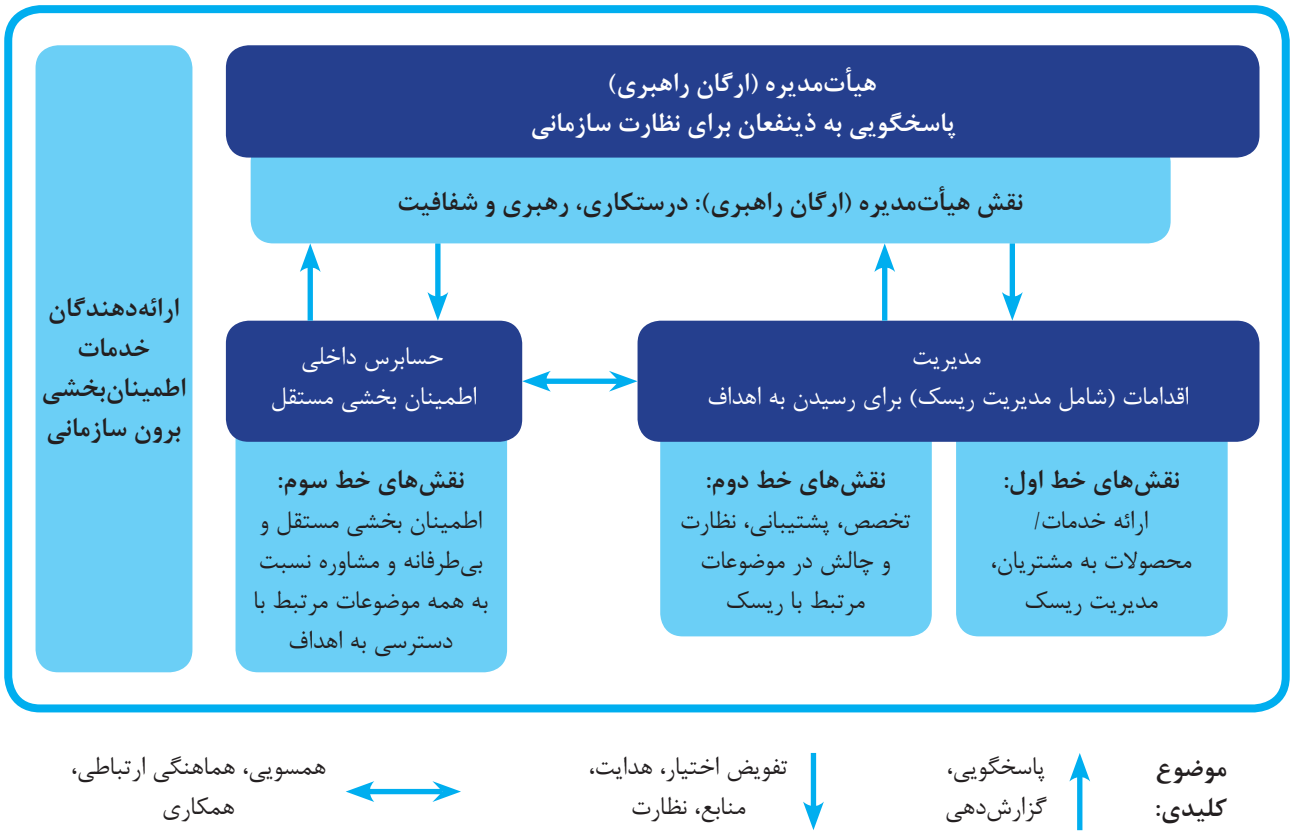
در این بخش، نمایی کلی از ساختار اصلی (بخش ۲,۱) و تاریخچه ی مدل سه‌خط (بخش ۲,۲) ارائه و برخی از انتقادهای اصلی بیان می‌شود، به‌طور خلاصه درباره مدل‌های جایگزین بحث می‌شود (بخش ۲,۳)، و شواهد تجربی را برای اثربخشی آن بررسی می‌کنیم (بخش ۲,۴).

۲,۱ ساختار اصلی

نسخه‌های مختلفی از مدل سه‌خط وجود دارد. اکثر متخصصان و محققان با نسخه منتشر شده توسط IIA2013 آشنا هستند. پس از یک فرآیند بررسی، آن‌ها نسخه‌ی به‌روز شده(IIA 2020 a) را منتشر کردند که به‌طور فزاینده‌ای جایگزین نسخه‌ی اصلی می‌شود. این مقاله عمدتاً از نسخه‌ی به‌روز شده استفاده می‌کند.

همان‌طور که در شکل ۱ نشان داده شده است. مدل به‌روز شده دارای سه نوع آیتم است: بازیگران، نقش‌ها، و روابط.

این مدل بین چهار بازیگر که به‌عنوان جعبه‌ی آبی نشان داده می‌شوند تمایز قائل می‌شود: هیأت مدیره که برای نظارت سازمانی به ذی‌نفعان پاسخ‌گو است. مدیریت که برای دستیابی به اهداف سازمان اقداماتی را انجام می‌دهد. حسابرسی داخلی، که خدمات اطمینان‌بخشی مستقلی را به هیأت مدیره ارائه می‌دهد، و همچنین



شکل۱: مدل سه‌خط توصیف شده توسط IIA2020a

ارائه‌دهندگان خدمات اطمینان‌بخشی برون‌سازمانی.

این مدل بیشتر بین چهار نقش که به صورت جعبه‌های خاکستری نشان داده شده‌اند تمایز قائل می‌شود. نقش هیأت مدیره نشان دادن درستکاری، رهبری و شفافیت است. علاوه بر آن، این مدل شامل سه نقش است که آن‌ها را «خطوط دفاع» می‌نامد. خط اول محصولات و خدمات را به مشتریان ارائه می‌دهد و ریسک های مرتبط را مدیریت می‌کند. خط دوم به خط اول در رابطه با مدیریت ریسک کمک می‌کند. این تخصص و پشتیبانی تکمیلی را ارائه می‌دهد، همچنین شیوه‌های مدیریت ریسک را نظارت و به چالش می‌کشد. خط سوم اطمینان‌بخشی و مشاوره مستقل و عینی را در مورد کلیه‌ی موارد مربوط به دستیابی به اهداف ریسک ارائه می‌دهد. دو خط اول بخشی از مدیریت هستند، در حالی که خط سوم مترادف با حسابرسی داخلی است.

در نهایت، سه نوع رابطه بین بازیگران مختلف وجود دارد که به صورت فلش نشان داده می‌شوند. روابط از بالا به پایین وجود دارد: هیأت مدیره مسئولیت را به مدیریت محول می‌کند و بر حسابرسی داخلی نظارت می‌کند. برعکس، روابط پایین به بالا وجود دارد: مدیریت و حسابرسی داخلی پاسخ‌گو هستند و به بدنه‌ی راهبری گزارش می‌دهند. و در نهایت، یک رابطه‌ی افقی بین بازیگرانی وجود دارد که کارشان باید همسو باشد، یعنی بین مدیریت و حسابرسی داخلی.

۲,۲ تاریخچه‌ی مختصر

خاستگاه این مدل مبهم است. تئوری‌هایی وجود دارد که ریشه‌های نظامی، ورزشی یا کنترل کیفیت را نشان می‌دهد احتمالاً در اواخر دهه‌ی ۱۹۹۰ یا اوایل دهه‌ی ۲۰۰۰ این مدل توسعه یافته است. در سال ۱۹۹۹، کمیته‌ی بازل[۱۲] در مورد نظارت بانکی (BCBS) رویکرد مشابهی را برای نظارت بر ریسک پیشنهاد کرد اما اولین اشاره‌ی صریح به این مدل احتمالاً در گزارشی توسط سازمان خدمات مالی بریتانیا[۱۳](۲۰۰۳) یا مقاله ای از این مدل بود.

پس از بحران مالی ۲۰۰۷-۲۰۰۸، که تا حدی ناشی از شکست‌های گسترده مدیریت ریسک بود، محبوبیت این مدل به‌شدت افزایش یافت. در پاسخ به بحران، قانون‌گذاران و مقامات نظارتی توجه فزاینده‌ای به مسئول ارشد ریسک (CRO)[۱۴] و کمیته‌ی ریسک هیأت‌مدیره[۱۵] معطوف داشتند و شروع به توصیه‌ی مدل سه‌خط کردند. بیشتر کارهای آکادمیک روی این مدل نیز پس

از بحران انجام شد و قبل از آن بسیاری از متخصصان مدیریت ریسک فقط نام مدل جدید را شنیده بودند. امروزه اکثر شرکت‌های بورسی مدل سه‌خط را پیاده‌سازی کرده‌اند. در یک نظرسنجی در سال ۲۰۱۵ از متخصصان حسابرسی داخلی در ۱۶۶ کشور (n =۱۴۵۱۸) اکثر پاسخ‌دهندگان (۷۵٪) گزارش دادند که سازمان آن‌ها از مدل سه‌خط دفاعی که توسط انجمن حسابرسان داخلی ارائه شده است پیروی می‌کند. در میان مسئولان ارشد حسابرسی (CAEs)[۱۶] در اتریش، آلمان و سوئیس (n=۴۱۵)، از یافته‌های آن‌ها پشتیبانی می‌کند. اکثر پاسخ‌دهندگان (۸۸٪) گزارش دادند که آن‌ها این مدل را پیاده‌سازی کرده‌اند، به‌ویژه نرخ پذیرش بالا در بین مؤسسات مالی تا ۹۶٪ است.

در مقابل، به نظر نمی‌رسد شرکت‌های بزرگ فناوری مدل سه‌خط را پیاده‌سازی کنند. در هیچ یک از پرونده‌های آن‌ها به کمیسیون بورس و اوراق بهادار ایالات متحده (SEC)[۱۷] یا سایر نشریات ذکر نشده است. این مدل همچنین به‌صراحت در الزامات راهبری شرکتی منتشرشده توسط نزدک ذکر نشده است، جایی که همه‌ی شرکت‌های بزرگ فناوری فهرست شده‌اند. با این حال، شایان ذکر است که شیوه‌های نظارت بر ریسک در شرکت‌های بزرگ فناوری شباهت‌هایی با مدل سه‌خط دارد. به‌عنوان مثال، به نظر می‌رسد همه‌ی آن‌ها یک عملکرد حسابرسی داخلی را دارند (به‌عنوان مثال مایکروسافت، آلفابت). بر اساس اطلاعات عمومی، آزمایشگاه‌های تحقیقاتی هوش مصنوعی با اندازه‌ی متوسط نیز به نظر نمی‌رسد این مدل را پیاده‌سازی کنند.

۲,۳ انتقادها و مدل‌های جایگزین

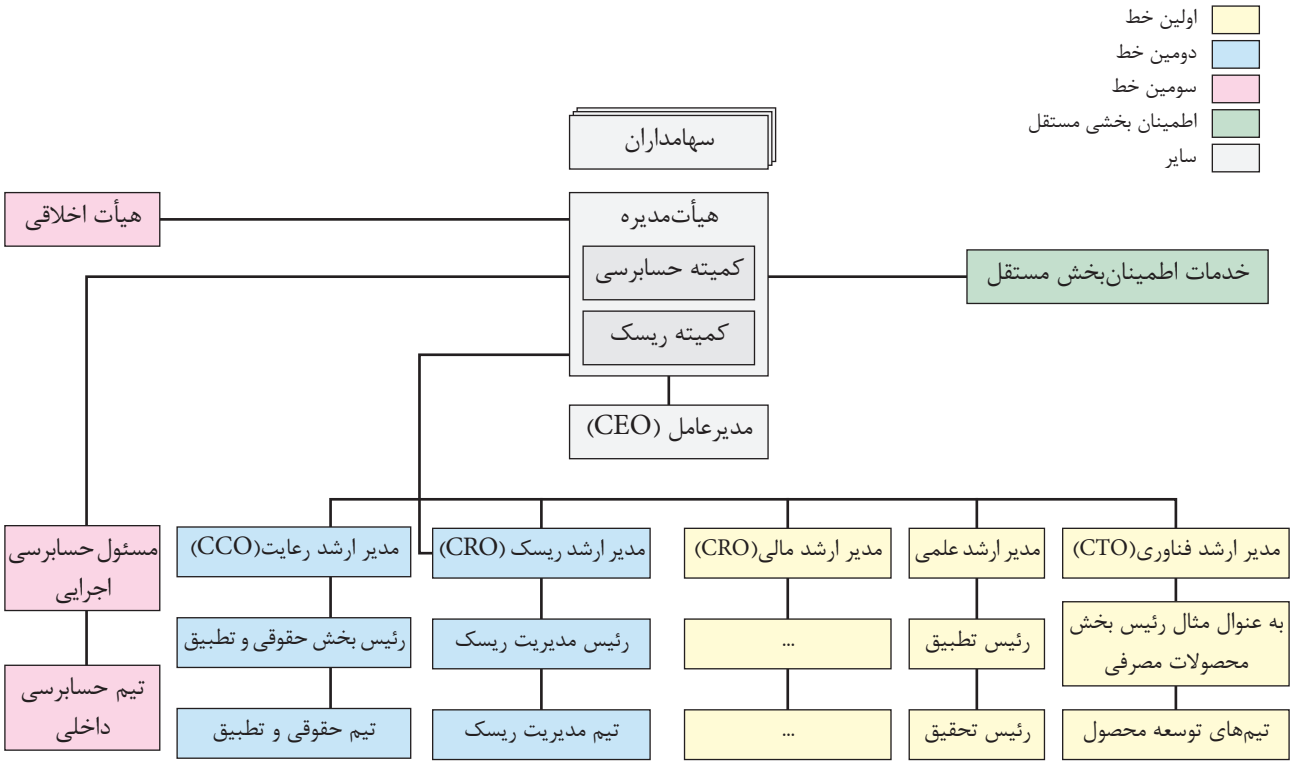
با وجود محبوبیت این مدل در بسیاری از صنایع، این مدل نیز مورد انتقاد قرار گرفته است چهار نقطه ضعف و شکست در مدل سه‌خط شناسایی شده است. اول، استدلال می‌شود که انگیزه‌های ریسک‌پذیری در خط اول اغلب نادرست است. زمانی که با جایگزینی بین ایجاد سود و کاهش ریسک مواجه می‌شوند، از لحاظ تاریخی انگیزه‌ای برای اولویت‌بندی ریسک‌های قبلی داشته‌اند. دوم، اغلب عدم استقلال سازمانی برای وظایف خط دوم وجود دارد. آن‌ها بیش از حد به افراد منفعت‌طلب (سودجو) نزدیک هستند که می‌توانند منجر به اتخاذ نگرش‌های ریسک‌پذیرتر شود. سوم، عملکردهای خط دوم اغلب فاقد مهارت و تخصص لازم برای به چالش کشیدن شیوه‌ها و کنترل‌ها در خط اول است. و چهارم، اثربخشی حسابرسی داخلی به دانش، مهارت و تجربه افراد بستگی دارد که ممکن است ناکافی باشد. یکی دیگر از انتقادهای رایج

این اســت که این مدل احساس امنیت کاذبــی را ارائه می‌دهد. به زبان ســاده، «وقتی چندین نفر مســئول هستند در واقع هیچ کس مسئول نیست»[18] انتقاد دیگر این است که این مدل بیش از حد بوروکراتیک و پرهزینه اســت. لایه‌های نظارتی اضافی ممکن اســت ریســک را کاهش دهد، اما به قیمت کارآمدی مدل تمام می‌شــود. آخرین انتقاد این مدل به جریان اطلاعات بین خطوط بستگی دارد، امــا موانع زیادی برای این موضوع وجود دارد. به‌عنوان مثال، ممکن اســت خط دوم تشــخیص ندهد که آن‌ها فقط آنچــه را که خط اول برای نشــان دادن آن‌ها انتخاب می‌کند، می‌بینند در حالی که این انتقادها کاســتی‌های مربوط را مشــخص می‌کند و باید جــدی گرفته شــود، اما مدل را به‌عنوان یک کل زیر سؤال نمی‌برد. علاوه بر این، مدل سه‌خط در طول سالیان بهبود یافته است. امروزه تمرکز بر افزایش اثربخشی مدل و پاسخ به انتقادات است.

با توجه به این انتقادات، چندین مدل جایگزین پیشنهاد شده‌اند. به‌عنوان مثال، مدل چهــار خطی را برای پاســخ‌گویی بهتر به نیازهای مؤسسات مالی پیشنهاد کردنــد. خط چهارم متشــکل از مراجع نظارتــی و حسابرســی برون‌ســازمانی اســت که قرار است با حسابرسی داخلی همکاری نزدیک داشــته باشــند. مثال دیگر مدل پنج خطی است که به‌تدریج توســط چندین محقق و سازمان توسعه یافت با این حال، تغییرات پیشــنهادی لزوماً مدل را بهبود نمی‌بخشد. استدلال شــده اســت که افزودن خطوط بیشتر مدل را بیش از حــد پیچیده می‌کند و شــرکت‌ها و قانونگذاران در حال حاضر خواهان تغییرات ســاختاری نیســتند. همچنین شایان ذکر است که مدل‌های جایگزین به‌مراتــب کمتر از مدل اصلی محبــوب هســتند. در مقایســه با این مدل‌هــای جایگزیــن، مدل ســه‌خط همچنان «بادقت‌ترین سیستم مدیریت ریســک که تاکنون توسعه یافته است» باقی می‌ماند. اما چه شواهد تجربی برای اثربخشی آن داریم؟

۲,۴ شواهد تجربی

منظور از «اثربخشی» تعیین درجه‌ای اســت که مدل مورد نظر به سازمان‌ها برای دســتیابی به اهداف‌شــان کمک می‌کند.[19] با توجه به هــدف این مقاله، مــن بیشــتر به دســتیابی بــه اهداف ریســک علاقه‌مند هســتم. این مسأله می‌تواند شــامل موارد زیر باشــد: (۱) کاهش ریسک‌های مربوط به سطح قابل قبول، (۲) اطمینــان از آگاهی مدیریت و هیــأت مدیــره از ماهیــت و مقیاس ریســک‌های کلیــدی، و (۳) تطبیق با مقــررات مربوط به ریســک. من کمتر به اهداف دیگر علاقه‌مند هســتم (مثلاً بهبود عملکــرد مالــی)، اگرچه ممکن است همپوشانی‌هایی وجود داشته باشد (مثلاً کاهش ریســک آســیب به افراد ممکن اســت خطر زیان مالی ناشــی از پرونده‌های دعاوی را کاهش دهد).



شکل ۲ – نمونه‌ی نمودار سازمانی یک شرکت هوش مصنوعی با مسئولیت‌های معادل برای هر یک از سه‌خط

به نظر نمی‌رسد هیچ مطالعه‌ای (با کیفیت بالا) در مورد اثربخشی مدل سه‌خط به معنای فوق الذکر وجود داشته باشد. فقط به نظر می‌رسد که شواهدی برای اثربخشی حسابرسی داخلی وجود دارد به‌عنوان مثال، یک نظرسنجی از مسئولان اجرایی حسابرسی (CAEs) در شرکت‌های چند ملیتی در آلمان (n =۳۷) واحدهای تجاری حسابرسی شده و حسابرسی نشده را مقایسه کرد. آن‌ها دریافتند که مدیران واحدهای حسابرسی شده نسبت به مدیران واحدهای حسابرسی نشده کاهش ریسک بیشتری را تجربه می‌کنند. مطالعات دیگر نشان می‌دهد که حسابرسی داخلی به تقویت سیستم‌های کنترل داخلی کمک می‌کند و تأثیر مثبتی بر پیشگیری و شناسایی تقلب دارد به نظر می‌رسد این واقعیت که مدل سه‌خط قادر به جلوگیری از رسوایی‌ها و بحران‌های گذشته نبود، شواهد ضعیفی علیه اثربخشی آن ارائه می‌کند (اگرچه توضیح دیگر می‌تواند شامل ضعیف بودن مدل در اجرا باشد)، در حالی که به نظر می‌رسد محبوبیت مداوم مدل شواهد ضعیفی را ارائه می‌کند. به نفع اثربخشی آن (اگرچه محبوبیت مدل همچنان می‌تواند بسته به مسیر تشریح شود). در نهایت، شواهدی در هر دو جهت وجود دارد به‌طور کلی، با وجود محبوبیت این مدل، «اثربخشی آن آزموده نشده باقی می‌ماند» و «بر اساس هیچ مدرک روشنی نیست» که ما شواهد محکمی مبنی بر ناکارآمدی مدل داشته باشیم و هنوز هم بسیار قابل‌قبول است که این مدل می‌تواند مؤثر باشد، اما مطالعات (با کیفیت بالا) شواهد تجربی برای اثربخشی آن به معنای یادشده ارائه نداده است.

این نقص (کمبود) شگفت‌انگیز شواهد، به‌طور بالقوه می‌تواند با دلایل زیر توضیح داده شود. اول، از آن‌جایی که اجرای کارآزمایی‌های تصادفی‌سازی و کنترل‌شده در مورد مداخلات سازمانی امکان‌پذیر نیست، جمع‌آوری شواهد قوی ذاتاً دشوار است. دوم، این مدل به گونه‌ای طراحی شده است که منعطف و سازگار باشد، به این معنی که یک نقشه‌ی راه واحد و استاندارد برای پیاده‌سازی آن وجود ندارد. این عدم استانداردسازی می‌تواند مقایسه‌ی پیاده‌سازی‌های مختلف مدل و ارزیابی اثربخشی آن‌ها را دشوار کند. سوم، از آن‌جایی که بیشتر شاغلین عمدتاً به عملکرد مالی اهمیت می‌دهند، ممکن است محققان تشویق شوند تا بر معیارهای اقتصادی اثربخشی تمرکز کنند تا ارتباط کار خود را توجیه کنند (اگرچه شواهد زیادی در مورد آن وجود ندارد).

حتی اگر شواهد تجربی بیشتری از سایر صنایع داشته باشیم، ممکن است ارزش اطلاعاتی آن همچنان محدود باشد. یک دلیل این است که یافته‌ها ممکن است به هوش مصنوعی تعمیم نیابد.

شرکت‌های هوش مصنوعی از نظر ساختاری با سایر شرکت‌ها متفاوت هستند، زیرا تمرکز ویژه‌ای بر تحقیق دارند، و از آن‌جایی که هوش مصنوعی یک فناوری همه منظوره است، ریسک‌های ناشی از هوش مصنوعی گسترده‌تر از ریسک‌های سایر محصولات و خدمات است. دلیل دیگر این است که بزرگ‌ترین محرک توانایی مدل برای کاهش ریسک‌ها، احتمالاً روش مشخصی است که در آن اجرا می‌شود. بنابراین، شرکت‌های هوش مصنوعی به جای این‌که بپرسند «آیا مدل سه‌خط مؤثر است؟» باید بپرسند «چگونه می‌توانیم مدل را به روشی مؤثر پیاده‌سازی کنیم.»

۳- استفاده از مدل سه‌خط در زمینه‌ی هوش مصنوعی

این بخش راه‌هایی را پیشنهاد می‌دهد که شرکت‌های هوش مصنوعی می‌توانند مدل سه‌خط را پیاده‌سازی کنند. برای هر یک از سه‌خط، نقش‌ها و مسئولیت‌های معادل را پیشنهاد می‌دهد. ابتدا، محتوای مسئولیت‌های آن‌ها را شرح می‌دهد، سپس در مورد این‌که کدام تیم یا فردی مسئول است، همان‌طور که در شکل ۲ نشان داده شده است، بحث می‌شود.

۳,۱ خط اول

خط اول دو مسئولیت اصلی دارد: ارائه‌ی محصولات و خدمات به مشتریان، که مربوط به تحقیق و توسعه‌ی محصول هوش مصنوعی است، و مدیریت ریسک‌های مرتبط. در ادامه، روی موضوع دوم تمرکز شده است.

خط اول مسئول ایجاد و حفظ ساختارها و فرآیندهای مناسب برای مدیریت ریسک است که شامل اقداماتی در تمام مراحل فرآیند مدیریت ریسک است. برای تشخیص ریسک‌ها، خط اول می‌تواند از تجزیه و تحلیل رویداد طبقه‌بندی ریسک استفاده کند. برای تخمین احتمال و تأثیر ریسک‌های مشخص‌شده، برای ارزیابی‌های احتمالی ریسک ممکن است از مطالعات دلفی[۲۰] یا از ماتریس‌های ریسک[۲۱] استفاده کنند. این تخمین‌ها معمولاً با ارزیابی مدل، به‌طور بالقوه با تمرکز بر قابلیت‌های مدل پرریسک، و ارزیابی پادمان‌های شرکت اطلاع‌رسانی می‌شوند. برای کاهش ریسک‌ها، خط اول می‌تواند مدل را بر روی یک مجموعه داده انتخاب شده از طریق یادگیری تقویتی از بازخورد انسانی (RLHF)[۲۲] یا یادگیری تقویتی از بازخورد هوش مصنوعی (RLAIF)[۲۳]، که بیشتر به‌عنوان «هوش مصنوعی قانونی» شناخته می‌شود، تنظیم کند. برای جلوگیری از فاش شدن یا سرقت و ضریب اطمینان مدل، خط اول ممکن است اقداماتی را برای تقویت امنیت اطلاعات شرکت انجام دهد و برای جلوگیری از سوءاستفاده، سیاستی را برای انتشار تحقیقات بالقوه مضر معرفی کنند همچنین ممکن است منطقی باشد که رویکردی جامع‌تر داشته باشیم و چارچوب مدیریت ریسک ویژه‌ی هوش مصنوعی را پیاده‌سازی و یا یک چارچوب کلی‌تر مدیریت ریسک سازمانی (ERM)[۲۴] را سفارشی‌سازی کنیم. چندین سازمان راهنمایی در مورد نحوه‌ی اعمال این چارچوب‌ها برای نیازهای خاص توسعه‌دهندگان هوش مصنوعی مرزی[۲۵] ارائه می‌دهند. و یا فقط اجازه‌ی دسترسی به مدل‌ها را از طریق یک رابط برنامه‌نویسی کاربردی

(API)[۲۶] می‌دهند. در ماه‌های اخیر، ایجاد سیاست‌های خاص برای توسعه و استقرار مسئولانه‌ی سیستم‌های هوش مصنوعی مرزی، که به‌عنوان «سیاست‌های مقیاس‌پذیری مسئول» یا «سیاست‌های استقرار مبتنی بر ریسک» شناخته می‌شوند، رایج شده است. برای اکثر اقدامات ذکر شده در بالا، خط اول نیاز به حمایت خط دوم دارد.

خط اول همچنین مسئول اطمینان از تطبیق با انتظارات قانونی، مقرراتی و اخلاقی است. تعهدات قانونی ممکن است ناشی از قانون ضد تبعیض[۲۷] باشد.

یک مثال قابل توجه از مقررات هوش مصنوعی، قانون پیشنهادی هوش مصنوعی اتحادیه اروپا (کمیسیون اروپایی ۲۰۲۱)[۲۸] است که ارائه‌دهندگان سیستم‌های هوش مصنوعی پر ریسک را ملزم به پیاده‌سازی یک سیستم مدیریت ریسک می‌کند. انتظارات اخلاقی ممکن است ناشی از اصول اخلاقی هوش مصنوعی باشد که سازمان‌ها به صورت داوطلبانه اتخاذ کرده‌اند. برای اطمینان از تطبیق، خط اول به پشتیبانی خط دوم متکی است.

در نهایت، خط اول مسئول اطلاع‌رسانی به ارکان راهبری در مورد نتایج اقدامات ذکر شده در بالا، میزان برآورده شدن اهداف ریسک و سطح کلی ریسک است. این باید به شکل یک گفت‌وگوی مستمر، از جمله گزارش در مورد نتایج مورد انتظار و واقعی باشد. گزارش‌ها معمولاً شامل سوابق ریسک و ماتریس‌های ریسک می‌شوند، اما می‌توانند شامل اطلاعاتی در مورد مدل‌های خاص، به شکل کارت‌های مدل، برگه‌های داده باشند و باید توجه داشته باشید که یک خط گزارش از مسئول ارشد ریسک (CRO) به مسئول ارشد اجرایی (CEO)[۲۹] و کمیته‌ی ریسک هیأت مدیره نیز وجود داشته باشد.

مدیران عملیاتی، اغلب در یک ساختار مسئولیت آبشاری، مسئول هستند. در شرکت‌های بزرگ فناوری، کمترین سطح مسئولیت متوجه مدیرانی است که مسئولیت توسعه‌ی تک‌تک محصولات هوش مصنوعی را برعهده دارند. اگر محصول مستقل هوش مصنوعی وجود نداشته باشد و سیستم‌های هوش مصنوعی تنها بخشی از یک محصول را تشکیل می‌دهند مثلاً WaveNet[۳۰] به‌عنوان بخشی از Google Assistant، کمترین سطح مسئولیت بر عهده‌ی مدیرانی است که توسعه‌ی بخش هوش مصنوعی را رهبری می‌کنند. در آزمایشگاه‌های تحقیقاتی متوسط، پایین‌ترین سطح مسئولیت مدیریت ریسک بر عهده‌ی رهبران تحقیقاتی است، یعنی محققان ارشدی که مسئول پروژه‌های تحقیقاتی فردی هستند.

معمولاً یک یا چند سطح متوسط از مسئولیت وجود خواهد داشت که ممکن است شامل تعدادی از مدیران سطح متوسط باشد که مسئول حوزه‌های گسترده‌تر محصول (مانند بازی) یا حوزه‌های تحقیقاتی (مانند یادگیری تقویتی) هستند، اگرچه جزئیات این

موضوع به ساختارهای سازمانی خاص بستگی دارد. مسئولیت نهایی مدیریت ریسک هوش مصنوعی بر عهده‌ی آن دسته از مدیران C-suite[۳۱] است که مسئولیت توسعه‌ی محصول به‌عنوان مثال مدیر ارشد فناوری[۳۲](CTO) یا تحقیق مثلاً مدیر ارشد علمی[۳۳](CSO) را بر عهده دارند. در حالی که امکان تقسیم مسئولیت‌ها بین دو یا چند مدیر اجرایی وجود دارد اغلب این موضوع توصیه نمی‌شود، زیرا می‌تواند باعث کاهش مسئولیت‌ها شود.

۳,۲ خط دوم

با توجه به مدیریت ریسک خط دوم تخصص و پشتیبانی تکمیلی را ارائه می‌دهد، اما همچنین شیوه‌های مدیریت ریسک را نیز نظارت و به چالش می‌کشد.

برخی از فعالیت‌های مدیریت ریسک به تخصص خاصی نیاز دارند که خط اول آن را ندارد. ممکن است شامل تخصص حقوقی باشد به‌عنوان مثال نحوه‌ی انطباق با الزامات مدیریت ریسک مندرج در قانون پیشنهادی هوش مصنوعی اتحادیه‌ی اروپا. تخصص فنی؛ به‌عنوان مثال، نحوه‌ی ارزیابی قابلیت‌های مدل پرریسک یا توسعه‌ی مدل‌های زبانی[۳۴] واقعی‌تر. یا تخصص اخلاقی به‌عنوان مثال، چگونگی تعریف آستانه‌های هنجاری برای مطلوبیت ممکن است شامل تخصص ویژه‌ی ریسک باشد. به‌عنوان مثال، مدل‌های زبانی چه ریسک‌هایی دارند. یا تخصص ویژه‌ی مدیریت ریسک، به‌عنوان مثال، بهترین شیوه‌ها برای فیلترهای ایمنی تیم قرمز. خط دوم می‌تواند خط اول را با تهیه‌ی پیش‌نویس خط مشی‌ها، فرآیندها و رویه‌ها و همچنین چارچوب‌ها، قالب‌ها و طبقه‌بندی‌ها پشتیبانی کند. همچنین ممکن است در مورد مسائل خاص توصیه کند به‌عنوان مثال: چگونه یک چارچوب مدیریت ریسک را سفارشی‌سازی کنیم تا نیازهای خاص شرکت را بهتر برآورده کنیم. راهنمایی‌های کلی ارائه کنیم به‌عنوان مثال نحوه‌ی اطمینان از انطباق با سیاست‌های مربوط به ایمنی در بین محققان و مهندسان. یا ارائه‌ی آموزش به‌عنوان مثال، نحوه‌ی پردازش داده‌های آموزشی به روشی مطابق با مقررات عمومی حفاظت از داده‌ها GDPR[۳۵].

خط دوم همچنین مسئول نظارت و به چالش کشیدن کفایت و اثربخشی شیوه‌های مدیریت ریسک است. اگر اهداف ریسک برآورده نشود (مثلاً شرکت قوانین و مقررات مربوط را رعایت نکند، یا نتواند ریسک‌ها را تا حد قابل قبولی کاهش دهد) شیوه‌های مدیریت ریسک غیراثربخش است یا اگر می‌توانست با منابع کمتر به نتایج مشابهی دست یابد، آن‌ها ناکافی هستند. خط دوم معمولاً از تعدادی شاخص کلیدی عملکرد[۳۶](KPI) برای ارزیابی ابعاد مختلف کفایت و اثربخشی مدیریت ریسک (مثلاً تعداد ریسک‌های شناسایی شده، تعداد اتفاقات یا درصد پرسنل آموزش دیده در مورد موضوعات خاص) استفاده می‌کند.

مسئولیت‌های خط دوم در چندین تیم تقسیم می‌شود که معمولاً شامل تیم مدیریت ریسک و همچنین تیم حقوقی و انطباق می‌شود. اگرچه اکثر شرکت‌های بزرگ فناوری در حال حاضر یک تیم مدیریت ریسک دارند، اما این تیم‌ها بیشتر درگیر ریسک‌های تجاری هستند (مانند دعوی قضایی یا ریسک شهرت). ریسک‌های ناشی از هوش مصنوعی، به‌ویژه ریسک‌های اجتماعی، معمولاً یک نگرانی عمده نیستند. اگر شرکت‌های بزرگ فناوری بخواهند این را تغییر دهند، می‌توانند مسئولیت‌های تیم‌های موجود را گسترش دهند. راه‌اندازی یک تیم جدید مدیریت ریسک ویژه‌ی هوش مصنوعی چندان مطلوب به نظر نمی‌رسد، زیرا می‌تواند منجر به پراکندگی مسئولیت‌ها شود. احتمالاً یک ساختار مسئولیت آبشاری وجود دارد که در آن مسئول ارشد ریسک (CRO) به‌عنوان واحد پاسخ‌گویی برای فرآیند مدیریت ریسک عمل می‌کند. آزمایشگاه‌های تحقیقاتی متوسط معمولاً تیم مدیریت ریسک اختصاصی ندارند. یک استثنای قابل توجه مربوط به تیم جدید آمادگی OpenAI است. آن‌ها می‌توانند یک تیم جدید راه‌اندازی کنند یا یک یا چند نفر را در تیم‌های دیگر با عملکردهای پشتیبانی مرتبط با مدیریت ریسک مأمور کنند.

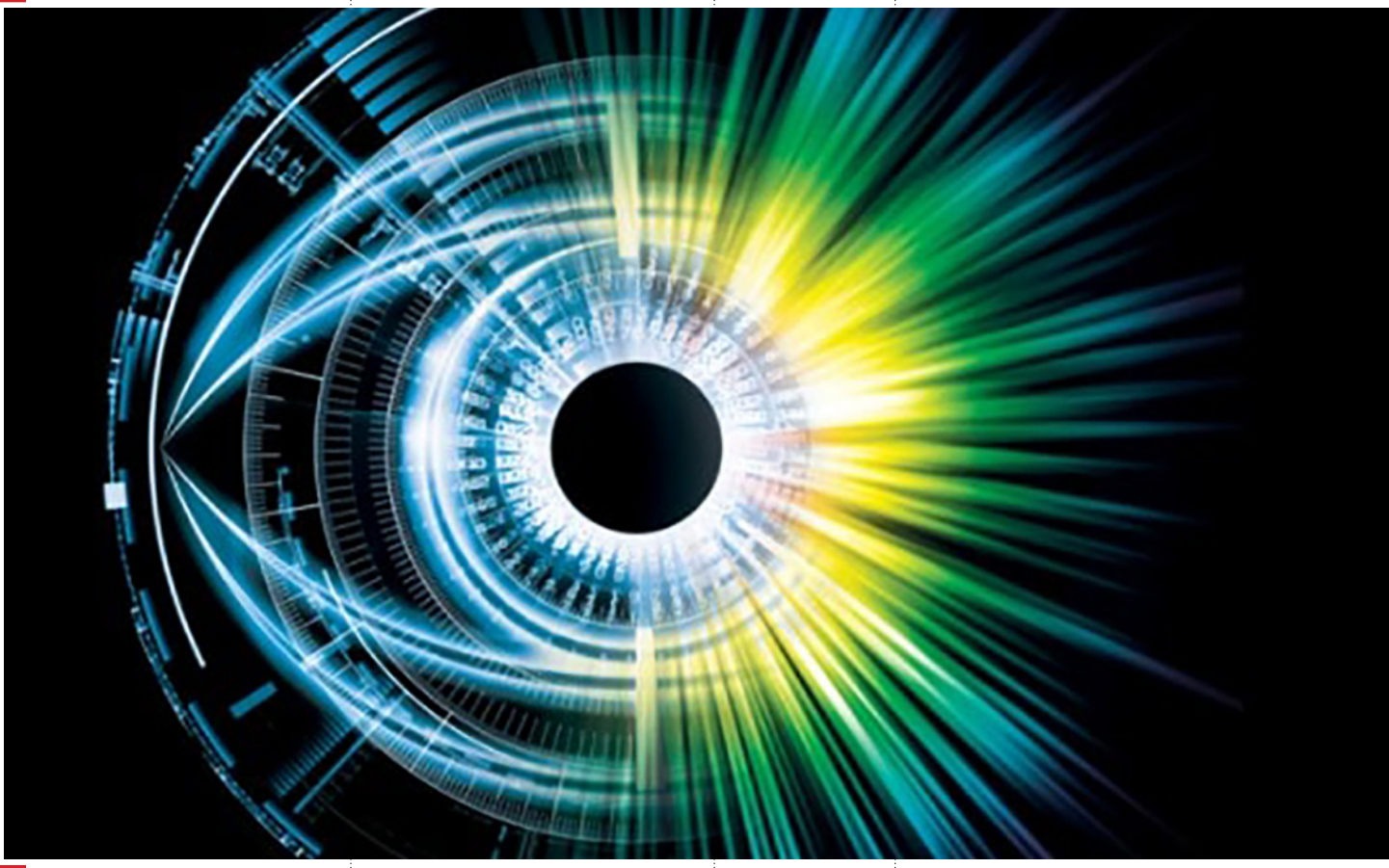
همه‌ی شرکت‌های هوش مصنوعی فراتر از مرحله‌ی راه‌اندازی اولیه، یک تیم حقوقی و انطباق دارند. سرپرست تیم، و در نهایت مسئول ارشد انطباق[۳۷](CCO) یا مسئول ارشد حقوقی[۳۸](CLO)، مسئول پشتیبانی حقوقی و انطباق مرتبط با ریسک خواهد بود. شایان ذکر است که تیم حقوقی و انطباق نیز در صورتی که واقعاً مسئولیت اطمینان از انطباق را بر عهده داشته باشند، می‌توانند جزء خط اول باشند. اگر قدرت تصمیم‌گیری نداشته باشند و فقط از خط اول حمایت کنند (مثلاً با نوشتن نظرات حقوقی) جزء خط دوم هستند. تیم حقوقی و انطباق نیز می‌توانند از شرکت‌های حقوقی برون‌سازمانی پشتیبانی بگیرند.

بسیاری از سازمان‌هایی که سیستم‌های هوش مصنوعی را توسعه و استقرار می‌دهند، تیم‌های دیگری دارند که می‌توانند مسئولیت‌های خط دوم را بر عهده بگیرند که ممکن است شامل تیم‌های ایمنی فنی، اخلاقی، خط مشی یا راهبری باشد. با این حال، در عمل، این تیم‌ها به ندرت خود را مسئول مدیریت ریسک می‌دانند. این موضوع باید هنگام اجرای مدل سه‌خط در نظر گرفته شود (به‌عنوان مثال با راه‌اندازی کارگاه‌ها برای حساس کردن آن‌ها به مسئولیت گسترده خود). به‌طور کلی، شرکت‌های هوش مصنوعی باید از واگذاری مسئولیت‌های خط دوم به آن‌ها اجتناب کنند.

۳,۳ خط سوم

خط سوم مسئول ارائه‌ی اطمینان‌بخشی مستقل است. کار دو خط اول را ارزیابی می‌کند و هر گونه کاستی را به ارکان راهبری گزارش می‌دهد.

در حالی که خط دوم در حال حاضر کفایت و اثربخشی شیوه‌های مدیریت ریسک را نظارت می‌کند و به چالش می‌کشد، خط سوم به‌طور مستقل کار



آن‌ها را ارزیابی می‌کند و به اصطلاح بر سرپرستان نظارت می‌کنند. آن‌ها می‌توانند این کار را با انجام مصاحبه (مثلاً با رهبران تحقیقاتی) و شرکت در جلسات (به‌عنوان مثال جلسات منظم تیم‌های توسعه) انجام دهند. آن‌ها همچنین می‌توانند حسابرسی داخلی انجام دهند یا حسابرسی برون‌سازمانی را سفارش دهند، چنین حسابرسی می‌تواند اهداف و دامنه‌های متفاوتی

داشته باشد. آن‌ها می‌توانند انطباق با قوانین، استانداردها یا اصول اخلاقی («حسابرسی رعایت») را ارزیابی کنند یا به دنبال شناسایی ریسک‌های جدید به روشی بازتر باشند («حسابرسی ریسک»). آن‌ها همچنین می‌توانند خود مدل را ارزیابی کنند، از جمله مجموعه داده‌ای که بر روی آن آموزش داده شده است («حسابرسی مدل»)، تأثیر مدل («حسابرسی تأثیر»)، یا راهبری شرکتی («حسابرسی راهبری»). به‌طور مشابه، خط سوم می‌تواند یک تیم قرمز را قبل یا بعد از استقرار یک مدل درگیر کند تا ارزیابی کند که آیا دو خط اول قادر به شناسایی کلیه ریسک‌های مربوط هستند یا خیر. علاوه بر آن، خط سوم می‌تواند خط‌مشی‌ها و فرآیندهای کلیدی را برای یافتن نقص‌ها و آسیب‌پذیری‌ها بررسی کند مثلاً خط‌مشی مقیاس‌پذیری مسئولانه یک شرکت یا پروتکل استقرار

آن‌ها. (توجه داشته باشید که این باید شامل یک فراارزیابی از اجرای خود مدل سه‌خط نیز باشد).

خط سوم همچنین با ارائه‌ی اطلاعات مستقل و بی‌طرفانه در مورد شیوه‌های مدیریت ریسک شرکت، از ارکان راهبری، معمولاً هیأت مدیره، پشتیبانی می‌کند. مخاطبان اصلی آن‌ها معمولاً کمیته حسابرسی است که عمدتاً از مدیران غیرموظف تشکیل شده است. اما از آن‌جایی که مدیران غیرموظف فقط به صورت پاره‌وقت کار می‌کنند و به‌شدت به اطلاعاتی که توسط مدیران به آن‌ها ارائه می‌شود وابسته هستند، آن‌ها به یک متحد[۳۹] مستقل در شرکت برای نظارت مؤثر بر مدیران نیاز دارند. خط سوم با حفظ درجه بالایی از استقلال از مدیریت و گزارش مستقیم به ارکان راهبری با پیروی از بهترین شیوه‌ها، این کار را انجام می‌دهد. اغلب به‌عنوان

«چشم و گوش»[40] ارکان راهبری توصیف می‌شوند.

خط سوم یک جایگاه سازمانی کاملاً مشخص دارد: حسابرسی داخلی. توجه داشته باشید که در این زمینه، حسابرسی داخلی به یک واحد سازمانی خاص اشاره دارد . این صرفاً به معنای حسابرسی نیست که به صورت داخلی انجام شود در عوض، این به معنای «آن دسته از افرادی است که به‌طور

مستقل از مدیریت برای ارائه‌ی اطمینان و بینش، در مورد کفایت و اثربخشی راهبری و مدیریت ریسک (از جمله کنترل داخلی) فعالیت می‌کنند.

به‌طور معمول، شرکت‌ها یک تیم حسابرسی داخلی اختصاصی دارند که توسط CAE[41] یا رئیس حسابرسی داخلی رهبری می‌شود. اکثر شرکت‌های بزرگ فناوری چنین تیمی دارند، اما مشابه تیم مدیریت ریسک، اغلب از ریسک‌های اجتماعی ناشی از هوش مصنوعی غفلت می‌کنند. به جای ایجاد یک تیم حسابرسی داخلی جداگانه مخصوص هوش مصنوعی، آن‌ها باید یک تیم فرعی در تیم حسابرسی داخلی موجود خود ایجاد کنند یا به‌سادگی یک یا چند عضو تیم را موظف کنند تا بر فعالیت‌های مدیریت ریسک خاص هوش مصنوعی تمرکز کنند. آزمایشگاه‌های تحقیقاتی متوسط معمولاً تیم حسابرسی داخلی ندارند. آن‌ها باید یک تیم یا وظیفه جدید حداقل یک نفر با مسئولیت‌های خط سوم ایجاد کنند. به‌طور خلاصه، شرکت‌های بزرگ فناوری باید «هوش مصنوعی را به حسابرسی داخلی بیاورند»، در حالی که آزمایشگاه‌های تحقیقاتی باید «حسابرسی داخلی را به هوش مصنوعی بیاورند».[42] شایان ذکر است که اگرچه پیشرفت‌های امیدوارکننده‌ای وجود دارد اما حرفه‌ی حسابرسان داخلی خاص هوش مصنوعی هنوز در مراحل ابتدایی خود است.

برخی از شرکت‌های هوش مصنوعی دارای یک هیأت اخلاقی هستند.مانند کمیته‌ی «ایتر» (کمیته‌ی اخلاق و حسابرسی داخلی) مایکروسافت و هیأت نظارت متا. این کمیته‌ها معمولاً علاوه بر حسابرسی داخلی می‌تواند مسئولیت‌های خط سوم را نیز بر عهده بگیرد. باید از نظر سازمانی مستقل از مدیریت باشد، اما همچنان بخشی از سازمان باشد (برخلاف ارائه‌دهندگان اطمینان‌بخشی برون‌سازمانی). اگر سازمان‌ها قبلاً یک هیأت مستقل اخلاقی داشته باشند (مثلاً متشکل از نمایندگان دانشگاه و جامعه‌ی مدنی)، می‌توانند یک گروه کاری تشکیل دهند که مسئولیت‌های خط سوم را بر عهده می‌گیرد.

۴ - چگونه مدل سه‌خط می‌تواند به کاهش ریسک‌های ناشی از هوش مصنوعی کمک کند

در حالی که دلایل زیادی وجود دارد که چرا شرکت‌های هوش مصنوعی ممکن است بخواهند مدل سه‌خط دفاعی را پیاده‌سازی کنند، اما این بخش فقط بر سه استدلال در مورد توانایی این مدل برای جلوگیری از آسیب‌های فردی، جمعی و اجتماعی تمرکز می‌کند. این مدل می‌تواند با شناسایی و بستن شکاف‌ها به کاهش ریسک‌های ناشی از هوش مصنوعی کمک کند. در پوشش ریسک (بخش ۴,۱)، افزایش اثربخشی شیوه‌های مدیریت ریسک (بخش ۴,۲)، و قادر ساختن ارکان راهبری برای نظارت مؤثرتر بر مدیریت (بخش ۴,۳) همچنین یک نمای کلی از مزایای دیگر ارائه می‌شود (بخش ۴,۴). شایان ذکر است که در غیاب شواهد تجربی قوی، بحث زیر نظری باقی می‌ماند و اغلب بر ملاحظات قابل قبول انتزاعی تکیه می‌کند.

۴,۱ تشخیص و بستن شکاف‌ها در پوشش ریسک

مدیریت ریسک هوش مصنوعی شامل افراد مختلف از تیم‌های مختلف

با مسئولیت‌های متفاوت است. اگر این مسئولیت‌ها به اندازه‌ی کافی هماهنگ نباشند، شکاف‌هایی در پوشش ریسک ممکن است رخ دهد. چنین شکاف‌هایی ممکن است دلایل مختلفی داشته باشند. به‌عنوان مثال، ممکن است هیچ کس مسئول مدیریت یک ریسک خاص نباشد (مثلاً ممکن است نقطه‌ی کوری[۴۳] برای ریسک‌های پراکنده[۴۴] وجود داشته باشد)، یا ممکن است مشخص نباشد که چه کسی مسئول است (مثلاً ممکن است دو تیم به اشتباه تصور کنند که تیم دیگر در حال حاضر از یک ریسک مراقبت می‌کند). همچنین ممکن است فرد مسئول نتواند ریسک را به‌طور موثر مدیریت کند (به‌عنوان مثال به دلیل نداشتن تخصص، اطلاعات یا زمان لازم). اگر یک ریسک خاص به اندازه‌ی کافی توسط سیستم مدیریت ریسک پوشش داده نشود، نمی‌توان آن را تشخیص داد و ممکن است منجر به ارزیابی ریسک نادرست شود مثلاً ریسک کل یک سیستم هوش مصنوعی ناایمن[۴۵]، قابل قبول ارزیابی شود و یک پاسخ ریسک ناکافی (مثلاً ناایمن) دریافت و سیستم هوش مصنوعی بدون اقدامات احتیاطی کافی مستقر شده باشد.

مدل سه‌خط می‌تواند با تشخیص و بستن شکاف‌ها در پوشش ریسک از این امر جلوگیری کند. می‌تواند این کار را با ارائه‌ی روشی نظام‌مند برای تخصیص و هماهنگ کردن نقش‌ها و مسئولیت‌های مرتبط با مدیریت ریسک انجام دهد. این موضوع اطمینان‌بخشی می‌کند که افرادی که نزدیک به ریسک هستند، مسئولیت مدیریت ریسک (خط اول) را بر عهده دارند و حمایت مورد نیاز خود را دریافت می‌کنند (خط دوم). راه دیگری که مدل سه‌خط می‌تواند به تشخیص نقاط کور کمک کند، از طریق عملکرد حسابرسی داخلی (خط سوم) است. آن‌ها مسئول ارزیابی کفایت و اثربخشی کل الگوی[۴۶] مدیریت ریسک هستند که شامل شکاف‌های بالقوه در پوشش ریسک است.

ممکن است کسی اعتراض کند که در عمل، وجود شکاف در پوشش ریسک نادر است، و حتی اگر رخ دهد، فقط به ریسک‌های جزئی مربوط می‌شود (مثلاً به این دلیل که شرکت‌های هوش مصنوعی راه‌های دیگری برای رسیدگی به بزرگ‌ترین ریسک‌ها پیدا کرده‌اند). با این حال، پایگاه داده رویدادهای هوش مصنوعی[۴۷] حاوی ورودی‌های متعددی است، از جمله موارد متعددی که به‌عنوان «متوسط» یا «شدید»[۴۸] طبقه‌بندی شده‌اند، که نشان می‌دهد رویدادها چندان غیر معمول نیستند. در حالی که این رویدادها دلایل مختلفی داشتند، به نظر می‌رسد که حداقل برخی از آن‌ها به شکاف‌هایی در پوشش ریسک مرتبط باشند. اما از آن‌جایی که به نظر نمی‌رسد اطلاعات عمومی در این مورد وجود داشته باشد، این موضوع همچنان در معرض حدس و گمان است.

حتی اگر کسی فکر کند که شکاف در پوشش ریسک یک مشکل رایج در میان شرکت‌های هوش مصنوعی است، ممکن است توانایی مدل برای تشخیص و بستن آن‌ها را زیر سؤال ببریم. ممکن است کسی مشکوک شود که افراد درگیر و توانایی و تمایل آن‌ها برای تشخیص شکاف‌ها نقش بسیار بیشتری ایفا می‌کنند. در حالی که مطمئناً درست است که اجرای مدل به تنهایی کافی نیست، داشتن پرسنل توانا و مشتاق نیز کافی نیست. هر دو ضروری هستند و تنها با هم می‌توانند کافی باشند (اگرچه عوامل دیگری مانند اشتراک اطلاعات بین واحدهای سازمانی مختلف نیز ممکن است نقش داشته باشند).

به‌طور کلی، به نظر می‌رسد که اجرای مدل سه‌خط به کشف برخی از شکاف‌ها در پوشش ریسک کمک کند که در غیر این صورت مورد توجه قرار نمی‌گیرند.

۴,۲ افزایش اثربخشی شیوه‌های مدیریت ریسک

برخی از شیوه‌های مدیریت ریسک بی‌اثر هستند، ممکن است روی کاغذ خوب به نظر برسند، اما در عمل کارایی ندارند. شرکت‌های هوش مصنوعی ممکن است در تشخیص ریسک‌های مربوط شکست بخورند، احتمال یا تأثیر آن‌ها را اشتباه ارزیابی کنند یا نتوانند آن‌ها را به سطح قابل‌قبولی کاهش دهند. شیوه‌های ناکارآمد مدیریت ریسک می‌توانند دلایل مختلفی داشته باشند، مانند تکیه بر یک معیار واحد (مثلاً استفاده از یک طبقه‌بندی واحد برای تشخیص طیف وسیعی از ریسک‌ها)، عدم پیش‌بینی تلاش‌های عمدی برای دور زدن اقدامات (مانند سرقت یک مدل منتشر نشده)، عدم پیش‌بینی تغییرات مرتبط در چشم‌انداز ریسک (به‌عنوان مثال: ظهور ریسک‌های سیستماتیک به دلیل اتکای فزاینده سوگیری‌های شناختی مدیران ریسک بر روی مدل‌های پایه به‌عنوان مثال: سوگیری در دسترس بودن، یعنی تمایل به «ارزیابی فراوانی یک طبقه یا احتمال یک رویداد با سهولت یادآوری موارد یا رخدادها»، و سایر خطاهای انسانی (مثلاً فردی که یک ثبت ریسک را پر می‌کند و اشتباه بیرون می‌زند).

مدل سه‌خط می‌تواند اثربخشی شیوه‌های مدیریت ریسک را با تشخیص چنین کاستی‌هایی[۴۹] افزایش دهد. همان‌طور که در بالا ذکر شد، حسابرسان داخلی اثربخشی رویه‌های مدیریت ریسک را ارزیابی می‌کنند و هرگونه کاستی را به ارکان راهبری گزارش می‌دهند که می‌تواند با مدیریت برای بهبود این شیوه‌ها تعامل داشته باشد.

ممکن است کسی اعتراض کند که بیشتر کاستی‌ها فقط در موقعیت‌های کم‌ریسک رخ می‌دهند. در موقعیت‌های پرریسک، شیوه‌های مدیریت ریسک موجود مؤثرتر هستند. برای مثال، شرکت‌های هوش مصنوعی اغلب ارزیابی‌های گسترده‌ای از ریسک، قبل

از استقرار مدل‌های پیشرفته انجام می‌دهند که این ممکن است در موارد آشکار، صادق باشد ولی در موارد کمتر آشکار به اندازه‌ی مورد نظر مؤثر نباشد مثلاً به این دلیل که نسبت به خطاهای انسانی یا تلاش‌های عمدی برای دور زدن آن‌ها حساس نیستند. به‌عنوان مثال، اخیراً یک پست وبلاگی منتشر شده است که در آن برخی از چالش‌هایی را که در هنگام ارزیابی مدل‌ها با آن مواجه شده‌اند، بیان می‌کند. در برابر این موضوع، من مطمئناً نمی‌خواهم به این استدلال متقابل تکیه کنم که اثربخشی شیوه‌های مدیریت ریسک در حال حاضر به اندازه‌ی کافی با ریسک‌های موجود افزایش می‌یابد.

برخی از شرکت‌های هوش مصنوعی ممکن است اعتراض کنند که از قبل معادل یک عملکرد حسابرسی داخلی داشتند، بنابراین پیاده‌سازی مدل سه‌خط تنها یک پیشرفت حاشیه‌ای خواهد بود. اگرچه ممکن است درست باشد که برخی از افراد در برخی از شرکت‌ها وظایفی مشابه آنچه حسابرسان داخلی انجام می‌دهند انجام می‌دهند، تا آن‌جا که من می‌دانم، ارزیابی اثربخشی رویه‌های مدیریت ریسک مسئولیت اصلی آن‌ها نیست و بهترین شیوه‌ها را دنبال نمی‌کنند و مانند حرفه‌ی حسابرسی داخلی، مستقل بودن سازمانی از مدیریت را ندارند، که می‌تواند به تفاوت‌های قابل توجهی منجر شود.

به‌طور کلی، من فکر می‌کنم این یکی از بهترین استدلال‌ها برای پیاده‌سازی مدل سه‌خط است. بدون تلاش جدی برای تشخیص شیوه‌های ناکارآمد مدیریت ریسک ، انتظار می‌رود حداقل برخی از کاستی‌ها مورد توجه قرار نگیرد. میزان صحت این موضوع عمدتاً به توانایی و تمایل حسابرسی داخلی برای انجام این وظیفه بستگی دارد.

۴،۳ توانمندسازی ارکان راهبری برای نظارت موثرتر بر مدیریت

ارکان راهبری، به‌طور معمول هیأت مدیره، مسئول نظارت بر مدیریت است. برای انجام این کار، آن‌ها به اطلاعات مستقل و عینی در مورد شیوه‌های مدیریت ریسک شرکت نیاز دارند. با این حال، آن‌ها به‌شدت به اطلاعاتی که مدیران اجرایی در اختیار آن‌ها قرار می‌دهند، متکی هستند. برای نظارت موثر بر مدیران، آن‌ها به یک متحد (هم‌پیمان) مستقل در شرکت نیاز دارند.

حسابرسی داخلی این وظیفه را با حفظ درجه‌ی بالایی از استقلال از مدیریت و گزارش مستقیم به کمیته‌ی حسابرسی هیأت مدیره انجام می‌دهد. این موضوع می‌تواند مهم باشد زیرا در مقایسه با سایر بازیگران، هیأت مدیره تأثیر قابل‌توجهی بر مدیریت دارد. برای مثال، آن‌ها می‌توانند مدیر عامل را عوض کنند (مثلاً اگر مکرراً سود را بر مصونیت اولویت می‌دهد)، تصمیم‌های راهبردی بگیرند (مثلاً جلوگیری از مشارکت راهبردی با ارتش)،[۵۰] و تغییراتی در راهبری ریسک شرکت ایجاد کنند (مثلاً ایجاد یک هیأت اخلاقی). توجه داشته باشید که یک خط گزارش تکمیلی از

مسئول ارشد ریسک (CRO) به کمیته‌ی ریسک هیأت مدیره وجود دارد.

ممکن است کسی اعتراض کند که این عملکرد می‌تواند توسط بازیگران دیگر نیز انجام شود. برای مثال، حسابرسان شخص ثالث[۵۱] نیز می‌توانند اطلاعات مستقل و عینی را در اختیار هیأت مدیره قرار دهند. در حالی که حسابرسی‌های برون‌سازمانی قطعاً می‌توانند نقش مهمی ایفا کنند، در مقایسه با حسابرسی داخلی، دارای معایبی هستند: ممکن است فاقد زمینه‌های مهم شناخت باشند،[۵۲] شرکت‌ها ممکن است نخواهند اطلاعات حساسی را با آن‌ها به اشتراک بگذارند (مثلاً در مورد پروژه‌های تحقیقاتی در حال انجام)، و حسابرسی‌ها معمولاً فقط گزارش لحظه‌ای[۵۳] در یک زمان هستند. بنابراین، شرکت‌های هوش مصنوعی باید حسابرسی برون‌سازمانی را مکمل حسابرسی داخلی بدانند، نه یک جایگزین. به همین دلیل است که مدل سه‌خط، بین حسابرسی داخلی و ارائه‌دهندگان اطمینان‌بخشی برون‌سازمانی تمایز قائل می‌شود.

می‌توان به این نکته اشاره کرد که در صنایع دیگر، حسابرسی داخلی اغلب دیر مداخله می‌کند و به جای نظارت بر آن‌ها، با مدیریت همکاری می‌کند و این واقعاً مشکل‌ساز خواهد بود. با این حال، همان‌طور که در بالا مورد بحث قرار گرفت، به نظر نمی‌رسد که این ویژگی ذاتی حسابرسی داخلی باشد. در عوض، به نظر می‌رسد که عمدتاً به روشی خاص راه‌اندازی می‌شود و افراد درگیر هدایت می‌شوند. با این حال، شرکت‌های هوش مصنوعی باید این نگرانی را جدی بگیرند و اقداماتی را برای رفع آن انجام دهند.

به‌طور کلی، من فکر می‌کنم که پیاده‌سازی مدل سه‌خط می‌تواند به‌طور قابل‌توجهی پایگاه اطلاع‌رسانی هیأت مدیره را افزایش دهد. این تأثیر در آزمایشگاه‌های تحقیقاتی متوسط قابل توجه‌تر خواهد بود، زیرا اکثر شرکت‌های فناوری بزرگ در حال حاضر دارای یک عملکرد حسابرسی داخلی هستند، البته نه مختص هوش مصنوعی.

۴,۴ سایر مزایا

پیاده‌سازی مدل سه‌خط مزایای زیادی به جز کاهش ریسک برای افراد، گروه‌ها یا جامعه دارد. اگرچه این مزایا فراتر از محدوده این مقاله هستند، به نظر می‌رسد حداقل ارائه‌ی یک نمای کلی ضروری است. در زیر به‌طور خلاصه به چهار مورد از آن‌ها می‌پردازیم.

اول، اجرای مدل سه‌خط می‌تواند از تکرارهای غیر ضروری پوشش ریسک جلوگیری کند. افراد مختلف در تیم‌های مختلف می‌توانند کار مدیریت ریسک یکسان یا بسیار مشابهی را انجام دهند و این موضوع اغلب مطلوب است زیرا می‌تواند از شکاف در پوشش ریسک جلوگیری کند . اما اگر چنین تکراری ضروری نباشد، می‌تواند منابعی مانند نیروی کار را که می‌توانند در جاهای دیگر به نحو مؤثرتری مورد استفاده قرار گیرد، هدر دهد. بنابراین شرکت‌های هوش مصنوعی با یک مبادله اثربخشی -کارایی- مواجه هستند. این‌که چگونه

ایـــن مبادلـــه باید حل شـــود، به زمینه خاص آن‌ها بستگی دارد. به‌عنوان مثال، هنگام برخورد با ریسک‌های فاجعه بار، اثربخشی (جلوگیری از شکاف در پوشش ریســـک) مهم‌تر از کارایی (جلوگیری از تکرارهای غیر ضروری پوشـــش) به نظر می‌رسد. در این مورد، شرکت‌های هوش مصنوعی باید به‌جای ریســـک شکاف‌ها در حوزه‌های مهم، در مورد پوشش بیش از حد توجه کنند.

به‌طور کلی، به نظر می‌رسد اگر به‌طور عمده به کاهش ریســـک توجه شود این مزیت اغراق آمیز و کم‌تر مرتبط باشد.

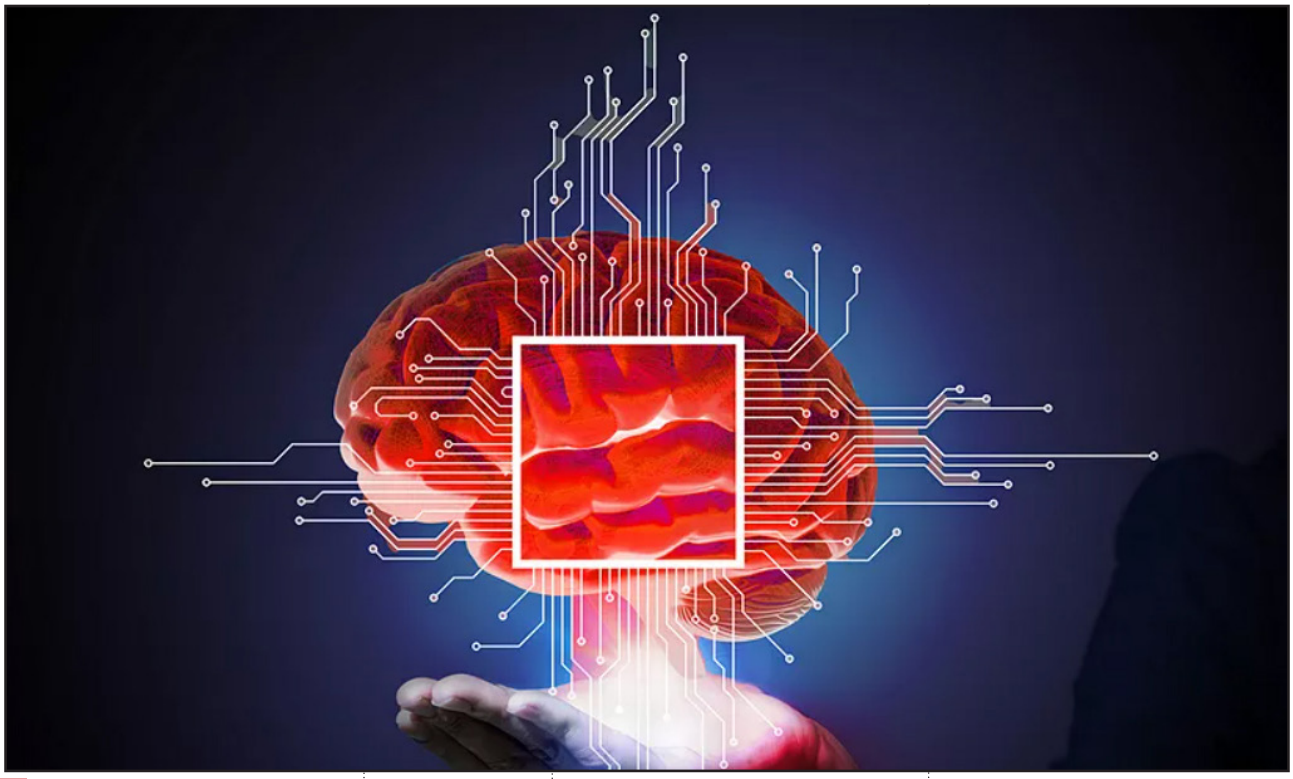
دوم، شـــرکت‌های هوش مصنوعی که مدل ســـه‌خط را پیاده‌ســـازی کرده‌اند، ممکن است مسئول‌تر تلقی شوند. به‌طور کلی، شـــیوه‌های مدیریت ریســـک در شرکت‌های هوش مصنوعی در مقایسه با بسیاری از صنایع دیگر (مانند حمل‌ونقل هوایی یا بانکی) کمتر پیشـــرفته به نظر می‌رســـد. با تطبیق بهترین شـــیوه‌های موجـــود از ســـایر صنایع، آن‌ها نشـــان می‌دهنـــد که قصد دارند تا شـــیوه‌های مدیریت ریسک خود را حرفه‌ای‌تر کنند، که می‌تواند به‌عنوان مســـئولیت‌پذیرتر تلقی شود. این تصور ممکن است فواید زیادی داشته باشد. به‌عنوان مثال، جذب و حفظ اســـتعدادهایی کـــه به اخلاق و ایمنی(صحت و سقم) اهمیت می‌دهند را آسان‌تر می‌کند. همچنین می‌تواند به جلوگیری از اقدامات بیش از حد سنگین از سوی ناظران کمک کند. حتی ممکن است در پرونده‌های دعاوی قضایی برای این ســـؤال که آیا یک ســـازمان وظیفه مراقبت خود را انجام داده اســـت یا خیر مفید باشـــد. با این حال، به نظر می‌رسد که آیا پیاده‌ســـازی مدل سه‌خط تا این حد بر ادراک تأثیر می‌گذارد، به‌ویژه در مقایسه با ســـایر اقدامات راهبری (مثلاً انتشـــار اصول اخلاقی هـــوش مصنوعی یـــا راه‌اندازی یک هیـــأت اخلاق هوش مصنوعی)، عمدتـــاً به این دلیل که اکثر ذی‌نفعـــان، از جمله بیشـــتر کارمندان، مدل را نمی‌دانند و نمی‌توانند ارتباط آن را ارزیابی کنند. یک استثنا ممکن است ناظران و دادگاه‌هایی باشـــند که بیشتر به جزئیات شـــیوه‌های مدیریت ریسک اهمیـــت می‌دهند. بهتریـــن حدس من این اســـت که پیاده‌سازی مدل، تأثیرات قابل‌توجهی بر درک چند ذی‌نفع خواهد داشـــت، در حالی که بیشـــتر ذی‌نفعان دیگر اهمیتی نمی‌دهند.

سوم، پیاده‌سازی مدل سه‌خط می‌تواند استخدام استعدادهای مدیریت ریسک را آســـان‌تر کند. حرفه‌ی مدیریت ریسک هوش مصنوعـــی در مراحل ابتدایی خود اســـت. من فرض می‌کنم که شرکت‌های هـــوش مصنوعی اســـتخدام افـــرادی با مهارت‌های هـــوش مصنوعی و مدیریت ریســـک را چالش برانگیـــز می‌دانند. در بیشتر موارد، آن‌ها می‌توانند کارشناسان هوش مصنوعی را استخدام کرده و آن‌ها را در زمینه‌ی مدیریت ریســـک آموزش دهند، یا کارشناســـان مدیریت ریســـک را از ســـایر صنایع استخدام کرده و آن‌ها را در زمینه‌ی هـــوش مصنوعی آموزش دهند. پیاده‌سازی مدل سه‌خط می‌تواند استخدام کارشناسان مدیریت ریسک از سایر صنایع را آسان‌تر کند، زیرا آن‌ها قبلاً با این مدل آشنا هستند. اگر فرض کنیم که شرکت‌های هوش مصنوعی می‌خواهند استعدادهای مدیریت ریسک بیشتری را استخدام کنند، زیرا سیستم‌ها توانمندتر می‌شـــوند و در موقعیت‌های حســـاس‌تر ایمنی اســـتفاده می‌شوند به‌عنوان مثال، ممکن اســـت اهمیت بیشتری پیدا کند. با این حال، من این اســـتدلال را چندان قانع کننده نمی‌دانم. من شـــک دارم که اجرای مدل ســـه‌خط تفاوت معنی داری در تصمیم‌گیری‌های مربوط به استخدام (به‌عنوان مثـــال در تصمیم یک داوطلب برای درخواست یا پذیرش یک پیشنهاد) ایجاد کند. از آن‌جایی که مدل مربوط به بعد سازمانی مدیریت ریسک است، تأثیر قابل‌توجهی بر کار روزمره مدیریت ریسک ندارد. با این اوصاف، ممکن است مزایای کوچکتری وجود داشـــته باشد (به‌عنوان مثال آســـان کردن فرآیند ورود). بهترین حدس من این است که تأثیر خلاف واقع اجرای مدل براستخدام کم است.

چهارم، اجرای مدل ســـه‌خط ممکن اســـت هزینه‌های تأمین مالی را کاهش دهد. آژانس‌های رتبه‌بندی تمایل دارند به شـــرکت‌هایی که چارچوب ERM را پیاده‌ســـازی کرده‌اند رتبه‌بندی بهتری بدهند (زیرا انجام این کار بهترین عمل در نظر گرفته می‌شود)، و شرکت‌هایی با رتبه‌بندی بهتـــر تمایل دارند هزینه‌های تأمین مالی کمتری داشـــته باشند زیرا شـــرایط اعتباری بهتـــری دارند. ممکن است اثر مشابهی با توجه به اجرای مدل سه‌خط وجود داشـــته باشد. هزینه‌های تأمین مالی کمتر به‌ویژه اگر فرض کنیم که هزینه‌های توسعه سیستم‌های هوش مصنوعی پیشـــرفته به دلیـــل افزایش تقاضا برای محاســـبات افزایش می‌یابد اهمیـــت خاصی دارد. در ســـناریوهایی که فشـــار تجاری بســـیار بالاتر از امروز است، هزینه‌های تأمین مالی پایین‌تر نیز می‌تواند برای ادامه تحقیقات ایمنی که به توسعه محصول کمکی نمی‌کند، مهم باشد. با این حال، من مطمئن نیستم که تا چه حد یافته‌های چارچوب‌های ERM به مدل سه‌خط تعمیم می‌یابند. بهترین حدس من این است که پیاده‌سازی مدل ســـه‌خط تأثیر معناداری بـــر هزینه‌های مالی آزمایشـــگاه‌های تحقیقاتی متوسط امروزی نخواهد داشـــت. اما انتظار دارم که با ســـودآورتر شـــدن آزمایشگاه‌ها و اســـتفاده فزاینده از ســـایر منابع مالی (مانند اعتبـــارات یـــا اوراق قرضه) این موضوع تغییر کند.

۵ نتیجه‌گیری

این مقاله، مدل سه‌خط را در زمینه‌ی هوش مصنوعی اعمال کرده است. راه‌های مشـــخصی را پیشنهاد کرده است که در آن توســـعه‌دهندگان هـــوش مصنوعی مانند OpenAI، Google DeepMind و Anthropic می‌تواننـــد ایـــن مدل را



برای کاهش ریسک‌های ناشی از هوش مصنوعی پیاده‌سازی کنند. استدلال می‌کند که اجرای این مدل می‌تواند از آسیب‌های فردی، جمعی یا اجتماعی با تشخیص و بستن شکاف‌ها در پوشش ریسک، افزایش اثربخشی شیوه‌های مدیریت ریسک، و توانمند ساختن ارکان راهبری برای نظارت مؤثرتر بر مدیریت، جلوگیری کند. به این نتیجه رسید که، در حالی که محدودیت‌هایی وجود دارد و نباید اغراق‌آمیز درباره‌ی آثار صحبت کرد، اما این مدل می‌تواند به‌طور قابل قبولی به کاهش ریسک‌های ناشی از هوش مصنوعی کمک کند.

بر اساس یافته‌های این مقاله، سؤالات زیر برای تحقیقات بیشتر پیشنهاد می‌شود. اول، بحث در مورد توانایی مدل برای کاهش ریسک‌های ناشی از هوش مصنوعی عمدتاً نظری بود و بر ملاحظات قابل قبول انتزاعی تکیه داشت. سایر محققان را تشویق می‌کنم که این ادعاها را به صورت تجربی ارزیابی کنند. یک مطالعه‌ی موردی صنعتی مشابه آنچه که موکاندر و فلوریدی (۲۰۲۲) برای حسابرسی مبتنی بر اخلاق انجام دادند، می‌تواند اولین گام باشد. دوم، اگرچه به نظر نمی‌رسد شرکت‌های هوش مصنوعی مدل سه‌خط را پیاده‌سازی کنند، اما بسیاری از فعالیت‌های ذکر شده در بالا را قبلا انجام داده اند. برای هدف‌گیری بهتر کار در آینده، بازنگری شیوه‌های مدیریت ریسک موجود در این شرکت‌ها و انجام تجزیه و تحلیل شکاف، مفید خواهد بود. از آن‌جایی که داده‌های عمومی کمیاب است، محققان باید مصاحبه یا نظرسنجی انجام دهند (مثلاً «نظرسنجی معیار مدیریت ریسک هوش مصنوعی»)، اگرچه انتظار دارم محرمانه بودن یک مانع بزرگ باشد مهم است که بدانیم آیا مقررات موجود یا آینده ممکن است حتی شرکت‌های هوش مصنوعی را ملزم به اجرای این مدل کند. به‌عنوان مثال، در حالی که ماده ۹ قانون پیشنهادی هوش مصنوعی اتحادیه اروپا به مدل سه‌خط اشاره نمی‌کند اما پیشنهاد شده است که استانداردهای هماهنگ آینده یا مشخصات مشترک باید شامل این مدل باشند. در نهایت، مقاله مدل سه‌خط را به صورت مجزا بررسی کرده است. عوامل زمینه‌ای مانند فرهنگ ریسک در شرکت‌های هوش مصنوعی را که ممکن است بر اثربخشی مدل نیز تأثیر بگذارد، حذف کرده است. درک بهتر این عوامل، پایگاه اطلاعاتی را برای تصمیم گیرندگان در شرکت‌های هوش مصنوعی و فراتر از آن بهبود می‌بخشد.

همان‌طور که جرج باکس (۱۹۷۶) گفته است، «همه‌ی مدل‌ها اشتباه هستند، اما برخی از آن‌ها مفید هستند».[۵۴] با همین روحیه، می‌توان گفت که مدل سه‌خط راه‌حل کاملی برای ریسک‌های ناشی از هوش مصنوعی نیست، اما همچنان می‌تواند نقش مهمی ایفا کند. شرکت‌های هوش مصنوعی باید آن را به‌عنوان ابزار راهبری مفید ببینند که می‌توانند برای مقابله با تهدیدات امروز و فردا از هوش مصنوعی استفاده کنند.

منبع:

Jonas Schuett, Three lines of defense against risks from AI, Springer, 2023

پی‌نویس:

1. The three lines of defense
2. regulators
3. standard-setting bodies
4. compliance
5. discrimination risks
6. Institute of Internal Auditors
7. AI auditing framework (IIA)
8. the National Institute of Standards and Technology

۹. Deep Mind در سال ۲۰۱۰ با رویکردی بین رشته‌ای برای ساخت سیستم‌های هوش مصنوعی عمومی شروع به کار کرد. این آزمایشگاه تحقیقاتی ایده‌ها و پیشرفت‌های جدید در یادگیری ماشین، علوم اعصاب، مهندسی، ریاضیات، شبیه‌سازی و زیرساخت‌های محاسباتی را همراه با روش‌های جدید سازمان‌دهی تلاش‌های علمی گرد هم آورد.

۱۰. Open AI یک آزمایشگاه تحقیقاتی خصوصی است که هدف آن توسعه و هدایت هوش مصنوعی (AI) به روش‌هایی است که به نفع بشریت به‌عنوان یک کل نگر باشد. این شرکت توسط ایلان ماسک، سام آلتمن و دیگران در سال ۲۰۱۵ تاسیس شد و دفتر مرکزی آن در سانفرانسیسکو قرار دارد.

11. litigation or reputation risks
12. Basel Committee
13. UK Financial Services Authority
14. chief risk officer
15. risk committee of the board
16. chief audit executives
17. Securities and Exchange Commission
18. when there are several people in charge—no one really is”
19. I mean the degree to which the model helps organizations to achieve their objectives.

۲۰. در سناریوهای ارزیابی ریسک، روش دلفی به شناسایی ریسک‌های بالقوه و ارزیابی احتمالات و اثرات آن‌ها کمک می‌کند. کارشناسان دید جامعی از ریسک‌ها ارائه می‌دهند و سازمان‌ها را قادر می‌سازند تا استراتژی‌های مدیریت ریسک موثری را طراحی کنند.

۲۱. ماتریس ریسک ماتریسی است که در هنگام ارزیابی ریسک برای تعریف سطح ریسک با در نظر گرفتن مقوله احتمال در مقابل دسته شدت پیامد استفاده می‌شود. این یک مکانیسم ساده برای افزایش دید ریسک‌ها و کمک به تصمیم‌گیری مدیریت است.

22. Reinforcement learning from human feedback
23. Reinforcement learning from AI feedback
24. enterprise risk management

۲۵. frontier AI هوش مصنوعی مرزی واژه‌ای است که برای توصیف مدل‌های هوش مصنوعی استفاده می‌شود که از نظر قابلیت‌ها یا وظایف مختلف با مدل‌های پیشرفته‌ی هوش مصنوعی موجود مطابقت دارند یا بهتر عمل می‌کنند. در حال حاضر، "هوش مصنوعی مرزی" به معنای مدل‌های پایه یا هوش مصنوعی همه منظوره (GPAI) است. این اصطلاح به مدل‌های پایه بسیار توانمندی که می‌توانند قابلیت‌های خطرناکی داشته باشند اشاره می‌کند به عبارت دیگر، "هوش مصنوعی مرزی" حدس و گمان است و حتی هنوز وجود هم ندارد، اما می‌تواند در گوشه و کنار وجود داشته باشد.

26. Application programming interface
27. Anti-discrimination
28. EU AI Act (European Commission)
29. Chief executive officer

۳۰. WaveNet یک شبکه عصبی عمیق برای تولید صدای خام است که توسط محققان شرکت هوش مصنوعی مستقر در لندن Deep Mind ایجاد شده است.

۳۱. گروهی از مدیران اجرایی است که مسئولیت اداره یک سازمان را بر عهده دارند.

32. chief technology officer
33. e chief scientific officer

۳۴. مدل‌های زبان هوش مصنوعی جزء کلیدی پردازش زبان طبیعی (NLP) هستند، حوزه‌ای از هوش مصنوعی (AI) که بر توانمندسازی رایانه‌ها برای درک و تولید زبان انسان متمرکز است.

35. General Data Protection Legislation
36. key performance indicators
37. chief compliance officer
38. chief legal officer
39. ally
40. eyes and ears
41. Head of Internal Audit
42. bring AI to internal audit
43. blind spot
44. diffuse risks
45. total risk of an unsafe AI system
46. regime
47. AI Incident Database
48. “moderate” or “severe”
49. shortcomings
50. blocking a strategic partnership with the military
51. third-party auditors
52. they might lack important context
53. snapshots
54. all models are wrong, but some are useful