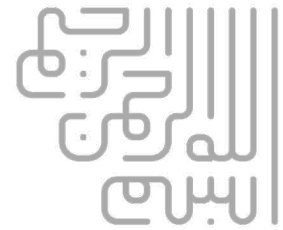


سایبرنتیک و مسئله حکمرانی (۱): عاملیت و مسئولیت‌پذیری در فناوری‌های هوش مصنوعی



تاریخ انتشار: ۱۴۰۵/۷/۱۸
Publish Date : 2026/10/10



عنوان: سایبرنتیک و مسئله حکمرانی (۱):
عاملیت و مسئولیت پذیری در فناوری های هوش مصنوعی
تهیه و تدوین کنندگان (Authors): زهرا زرگر (پژوهشکده مطالعات بنیادین علم و فناوری دانشگاه
شهید بهشتی)، ایمان اکبری (گروه سیاست پژوهی و آزمایشگاه حکمرانی)
واژه های کلیدی: اخلاق هوش مصنوعی، مسئولیت پذیری، عاملیت
ناشر: مرکز پژوهش های مجلس شورای اسلامی

DOI: <https://doi.org/10.22034/mrc.report.21955>

Permanent Link: https://report.mrc.ir/article_12005.html

Publisher : Islamic Parliament Research Center of Iran

عنوان: سایبرنتیک و مسئله حکمرانی (۱): عاملیت و مسئولیت پذیری در فناوری های هوش مصنوعی

نوع گزارش: طرح / لایحه □، نظارتی □، راهبردی ■، پیش نویس قانونی □
نام دفتر: مطالعات بنیادین حکمرانی (گروه سیاست پژوهی و آزمایشگاه حکمرانی)
مدیر مطالعه: حسین بابایی مجرد
ناظر علمی: مهدی عبدالحمید
ناظر علمی خارج از مرکز: مهدی شاملی حاجی آبادی
اظهار نظر کننده: ایمان شعبان زاده (دفتر مطالعات اجتماعی)
گرافیک و صفحه آرایی: علیرضا مقدم، انسیه بهابزرگی
ویراستار ادبی: زهرا کریمی
تاریخ شروع مطالعه: ۱۴۰۴/۰۴/۰۱

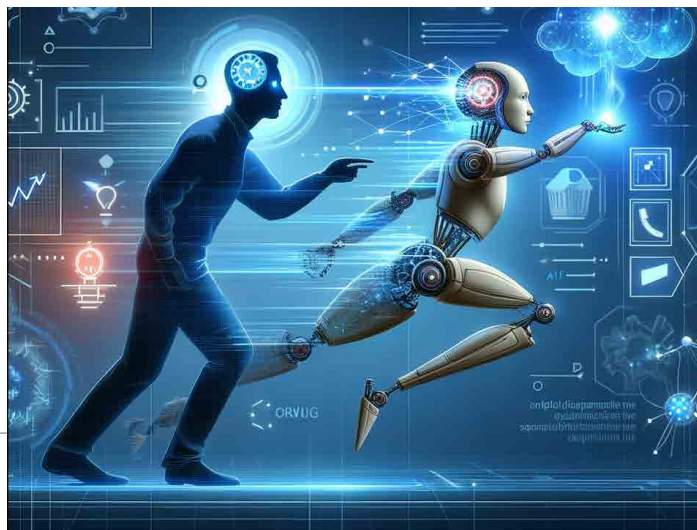


فهرست مطالب

چکیده.....	۶
خلاصه مدیریتی.....	۷
۱. مقدمه.....	۸
۲. پیشینه.....	۹
۱-۲. سوابق مطالعاتی در مرکز.....	۹
۳. مفهوم مسئولیت پذیری.....	۹
۴. مسئولیت جمعی، مسئله دست‌های بسیار و معضل شکاف مسئولیت پذیری.....	۱۲
۵. هوش مصنوعی به عنوان عامل اخلاقی.....	۱۶
۶. راهبردهایی برای توسعه مسئولانه هوش مصنوعی و رهنمودهایی برای کشور.....	۲۲
۱-۶. شفافیت ساختارهای اجتماعی-تکنیکی.....	۲۲
۲-۶. تعریف کارگروه ارزیابی اخلاقی در سیستم.....	۲۴
۳-۶. آگاهی بخشی درباره جنبه‌های عمومی مسئولیت پذیری.....	۲۵
۴-۶. خط‌مشی گذاری مشارکت محور.....	۲۷
۷. جمع‌بندی و نتیجه‌گیری.....	۲۸
منابع و مآخذ.....	۳۱

فهرست جداول

جدول ۱. تحلیل پیشینه پژوهشی.....	۹
جدول ۲. تعابیر مختلف از مسئولیت و شروط لازم برای آن.....	۱۱
جدول ۳. انواع رویکردها به فناوری.....	۱۳
جدول ۴. مسئله دست‌های بسیار و شکاف مسئولیت.....	۱۴
جدول ۵. دیدگاه‌های مختلف درباره عاملیت و مسئولیت پذیری هوش مصنوعی.....	۲۱



سایبرنتیک و مسئله حکمرانی (۱): عاملیت و مسئولیت‌پذیری در فناوری‌های هوش مصنوعی

چکیده



هوش مصنوعی به عنوان یک پدیده سایبرنتیکی با حکمرانی پیوستگی زیادی دارد. در مطالعات حکمرانی، مسئولیت‌پذیری یکی از مهم‌ترین مؤلفه‌هاست. مسئولیت‌پذیری مفهومی محوری در اخلاق است که بار اخلاقی یک وضعیت را به عواملی که در ایجاد آن وضعیت نقش علی داشته‌اند، نسبت می‌دهد. این مفهوم ضمن اهمیت اخلاقی، در نظام حقوقی نیز مفهومی اساسی بوده؛ زیرا در تعیین مصادیق جرم، و میزان جریمه‌ها و جبران خسارات، تعیین افرادی که مسئول ایجاد یک وضعیت هستند و میزان مسئولیت آنها، ضروری است. در مورد فناوری‌های هوش مصنوعی، مسئله مسئولیت‌پذیری پیچیدگی‌های خاصی پیدا می‌کند. فناوری‌های هوش مصنوعی که دارای قابلیت یادگیری هستند، فراتر از برنامه‌نویسی طراحان آن عمل کرده و براساس داده‌هایی که طبق آن آموزش می‌بینند، عمل می‌کنند. بنابراین به‌علت پیچیدگی و خارج از کنترل بودن عملکرد هوش مصنوعی، نمی‌توان در مسیر علی که منتهی به یک خروجی مشخص شده است، نقش عوامل انسانی را دقیقاً مشخص کرد. چنین وضعیتی با عنوان معضل «شکاف مسئولیت‌پذیری» شناخته می‌شود. با بیشتر شدن کاربردهای هوش مصنوعی در عرصه‌های مختلف، این معضل مسئله‌ای در خور توجه در مباحث نظری مربوط به توسعه هوش مصنوعی است. هدف گزارش حاضر تقریر مسئله مسئولیت‌پذیری و عاملیت در مورد هوش مصنوعی، روشن ساختن ابعاد مختلف آن، ارائه و دسته‌بندی پاسخ‌های مختلف ارائه شده به آن است. در انتها، راهکارهایی سیاستی و تقنینی با محوریت بر سازمان ملی هوش مصنوعی کشور با هدف پیشگیری از شکل‌گیری شکاف مسئولیت‌پذیری ذیل راهبردهایی چون شفاف‌سازی ساختارهای اجتماعی-تکنیکی، تشکیل کارگروه ارزیابی اخلاقی، خط‌مشی‌گذاری مشارکت‌محور و آگاهی‌بخشی درباره جنبه‌های عمومی مسئولیت‌پذیری ارائه می‌شود.



■ بیان/شرح مسئله

هوش مصنوعی به عنوان یک پدیده سایبرنتیکی خواه ناخواه با مسئله حکمرانی پیوندی عمیق دارد. یکی از مسائل حقوقی و اخلاقی برجسته در حکمرانی بحث مسئولیت‌پذیری است. در حوزه مطالعات توسعه هوش مصنوعی، مسئولیت‌پذیری بحث چگونگی انتساب مسئولیت اعمال فناوریانه است. این بحث هرچند در نهایت در قالب منازعات حقوقی بروز پیدا می‌کند، ریشه در مباحث اخلاقی و فلسفی دارد. وضعیت ویژه‌ای که در ارتباط با مسئله توزیع مسئولیت در مورد فناوری‌های هوش مصنوعی وجود دارد این است که هوش مصنوعی عامل^۱، درجاتی از خودمختاری و استقلال را از طراحان و تولیدکنندگان خود نشان می‌دهد. همین امر باعث شده در حلقه تولید و توسعه و کاربری، نتوان شخصیت حقیقی یا حقوقی انسانی را یافت که براساس معانی مرسوم، واجد شرایط مسئولیت‌پذیری نسبت به پیامدهای واقع شده باشد. به این ترتیب، این مسئله که در صورت وقوع پیامدهای آسیب‌زا به دنبال اعمال فناوریانه، مسئولیت پاسخگویی و جبران با چه کسی است، اهمیت می‌یابد.

■ نقطه نظرات / یافته‌های کلیدی

نقطه شروع برای این مسئله، تحلیلی نظری نسبت به ابعاد مختلف و شرایط انتساب مسئولیت‌پذیری از یک سو، و ویژگی‌های هوش مصنوعی از سوی دیگر است. براساس چنین تحلیلی می‌توان مواردی که انتساب مسئولیت به هوش مصنوعی مجاز است را معین کرد. با این حال، هر ۲ محور این بحث، محل مناقشات جدی است. در نتیجه، گروهی قائل به انتساب مسئولیت به هوش مصنوعی بوده و گروهی مسئولیت‌پذیری هوش مصنوعی را غیرممکن می‌دانند. برای برون‌رفت از این بن‌بست نظری، چرخش از مفهوم گذشته‌نگر از مسئولیت، به مفهومی آینده‌نگر پیشنهاد می‌شود. بر این اساس، به جای آنکه به مناقشه نظری درباره امکان یا عدم امکان مسئولیت هوش مصنوعی پرداخته شود، بر راهبردهای توسعه مسئولانه هوش مصنوعی و حفظ مسئولیت انسانی در روند توسعه هوش مصنوعی تکیه می‌شود.

این رویکرد که در مباحث هوش مصنوعی رویکردی جدید بوده و در چند سال اخیر به آن توجه شده، قیودی را بر توسعه هوش مصنوعی در سطوح مختلف آن وضع می‌کند که شفافیت، تبیین‌پذیری و حفظ مسئولیت انسانی در زنجیره توسعه و کاربری هوش مصنوعی را حفظ کرده و به این ترتیب از ایجاد شکاف مسئولیت پیشگیری کند. این قیود ذیل راهبردهایی چهارگانه مطرح می‌شود که شامل موارد زیر است.

الف) شفافیت ساختارهای اجتماعی-تکنیکی: با توجه به اینکه فناوری‌های هوش مصنوعی در بستری متشکل از انواع نهادهای اجتماعی و هویت حقیقی و حقوقی کار می‌کنند (ساختار اجتماعی-تکنیکی)، برای اجتناب از معضل شکاف مسئولیت لازم است نقشه‌ای از این ساختارها و روابط علت و معلولی میان اجزای مختلف آنها ترسیم شود، تا در هر حال، امکان پیگیری زنجیره علی منتهی به یک خروجی وجود داشته باشد.

ب) تشکیل کارگروه ارزیابی اخلاقی: شناسایی ابعاد اخلاقی تصمیم‌ها و اعمال مختلف در یک سامانه اجتماعی-تکنیکی، نیازمند تخصص در حوزه اخلاق و آشنایی به ابعاد فناوریانه است. تعریف کارگروه ارزیابی اخلاقی به عنوان حقوقی که واجد ۲ شرط آگاهی و قدرت عمل باشند، امکان شناسایی و اجتناب از پیامدهای نامطلوب استفاده از فناوری‌های هوش مصنوعی را حفظ کرده، و در صورت وقوع چنین مسائلی، این کارگروه مسئولیت پاسخگویی و توضیح شرایط واقع شده را برعهده خواهد داشت.

1. Agentic



ج) آگاهی بخشی درباره ابعاد عمومی مسئولیت پذیری: با توجه به اینکه فناوری‌های هوش مصنوعی یادگیرنده فراتر از چارچوب طراحی شده توسط تولیدکنندگان آن، از اعمال کاربران نیز اثر می‌پذیرند، بخشی از مسئولیت خروجی‌های آنها برعهده کاربران است. آگاهی کاربران از این ابعاد و ترغیب آنها به اهتمام برای رعایت مسئولیت‌های اخلاقی خود، از وقوع چالش‌های اخلاقی پیشگیری می‌کند.

د) خط‌مشی‌گذاری مشارکت‌محور: استفاده از مشارکت اقشار مختلف در تخمین مخاطرات ناشی از کاربرد هوش مصنوعی و تصمیم‌گیری برای مسیر توسعه هوش مصنوعی، منابع فکری و عملی عمومی را در پیشگیری از مخاطرات هوش مصنوعی استفاده کرده و می‌تواند در اجتناب از وقوع این مخاطرات نقش مهمی داشته باشد.

■ پیشنهاد راهکار تقنینی، نظارتی یا سیاستی

ضروری است سازمان ملی هوش مصنوعی کشور به‌عنوان متولی تخصصی راهبری توسعه فناوری هوش مصنوعی در کشور، اقدام‌ها و راهبردهای زیر را مدنظر قرار داده و سازوکارها و اقدام‌های لازم را به انجام برساند که عبارت‌اند از:

- تعریف شاخص کمی برای سنجش عملکرد سازمان‌ها در ارتباط با پیشگیری از شکاف مسئولیت‌پذیری.
- تجربه‌نگاری استخراج نقشه ساختارهای اجتماعی-تکنیکی و شفاف‌سازی آنها (در سطوح درون‌سازمانی و فراسازمانی) و بررسی نقش این کار در شفافیت مسئولیت‌ها و پیشگیری از ایجاد شکاف مسئولیت.
- تجربه‌نگاری تشکیل کارگروه‌های ارزیابی اخلاقی یا نقش‌های مشابه آن (مسئولیت پاسخگویی اخلاقی در سازمان یا فراسازمان)؛ مطالعه ساختارها و ترکیب‌های مختلف ممکن برای این کارگروه؛ تعریف جایگاه آن در ساختارهای سازمانی و فراسازمانی به‌نحوی که قدرت اعمال نظر و نفوذ لازم برای این کارگروه‌ها تضمین شود.
- بررسی ابزارهای مختلف افزایش سواد علمی و آگاهی عمومی درباره توسعه هوش مصنوعی و تأثیر آنها بر حس مسئولیت‌پذیری در کاربران هوش مصنوعی و مسئولیت‌پذیری (پاسخگویی) در توسعه‌دهندگان هوش مصنوعی.
- بررسی انواع روش‌های خط‌مشی‌گذاری مشارکتی در حوزه فناوری، مطالعه کارآمدی و تأثیر آنها بر کشف مخاطرات فناوری‌های هوش مصنوعی و درگیری عمومی با مسئولیت‌های ناشی از کاربرد این فناوری‌ها.
- تعریف سازوکارهای تشویقی برای کسب‌وکارهای مبتنی بر فناوری‌های هوش مصنوعی برای تعهد به چارچوب توسعه مسئولانه هوش مصنوعی و تعریف این تعهدات در چارچوب مسئولیت اجتماعی شرکت‌ها.

۱. مقدمه

ویژگی یادگیرندگی، فناوری‌های هوش مصنوعی را از سایر فناوری‌های مرسوم متمایز ساخته است. این ویژگی سبب شده تا این نوع فناوری‌ها بتوانند خارج از برنامه‌ریزی طراحان عمل کرده و خروجی‌های پیش‌بینی‌ناپذیر داشته باشند. این شیوه عملکرد، موجب شده است برخی محققان این فناوری‌ها را دارای درجاتی از خودمختاری بدانند. به‌رسمیت شناختن این فناوری‌ها به‌عنوان عوامل خودمختاری که در شبکه‌های تعاملی انسان-فناوری نقش ایفا می‌کنند، پرسش از مسئولیت خروجی این شبکه‌ها را به میان می‌آورد. اگر به جز انسان‌ها، عوامل خودمختار دیگری نیز در این شبکه ایفای نقش کنند، آیا می‌توان گفت مسئولیت اخلاقی خروجی‌های این شبکه‌ها تنها برعهده انسان‌های مشارکت‌کننده است؟ درحالی‌که از یک سو انتساب مسئولیت اخلاقی به عاملان غیرانسانی بی‌معنا یا بی‌فایده به نظر می‌رسد، اما از سوی دیگر انتساب مسئولیت به عوامل انسانی که کنترلی بر خروجی شبکه نداشته‌اند نیز معقول و منصفانه نیست. این معضل به نام

«شکاف مسئولیت پذیری» شناخته می شود. نحوه مواجهه ما با این چالش، هم به لحاظ اخلاقی و هم حقوقی پیامدهای قابل توجهی دارد.

در این گزارش، به معرفی این مسئله و ابعاد مختلف آن، مرور پاسخ های ارائه شده به آن، و در نهایت ارائه پیشنهادهایی برای فائق آمدن بر این چالش خواهیم پرداخت. در قسمت بعدی مقاله مفهوم «مسئولیت پذیری» و جنبه های متعدد آن معرفی خواهد شد. در قسمت سوم مقاله، مفهوم مسئولیت جمعی، چالش دست های بسیار و معضل شکاف مسئولیت پذیری شرح داده می شود. در قسمت چهارم مقاله، دیدگاه های مختلف درباره انتساب یا عدم انتساب مسئولیت پذیری و عاملیت اخلاقی هوش مصنوعی مطرح و دسته بندی می شوند. در نهایت در قسمت پنجم مقاله، راهبردهایی برای توسعه مسئولانه هوش مصنوعی پیشنهاد می شود.

۲. پیشینه

۲-۱. سوابق مطالعاتی در مرکز

جدول ۱. تحلیل پیشینه پژوهشی

ردیف	عنوان گزارش	سال انتشار	شماره مسلسل	نام دفتر / سازمان / نهاد	توضیحات
۱	بررسی لایحه برنامه هفتم توسعه (۸۸): توسعه پایدار هوش مصنوعی در کشور	۱۴۰۲	۱۹۳۹۵	حکمرانی	توجه به مسئولیت پذیری و پاسخگویی را از راهبردهای توسعه این فناوری دانسته است.
۲	حکمرانی هوش مصنوعی (۲): مفاهیم، ابعاد و مؤلفه ها	۱۴۰۳	۲۰۳۴۰	حکمرانی	مسئولیت پذیری را ذیل ارزش عدالت و برابری، از مسائل حکمرانی هوش مصنوعی می داند.
۳	حکمرانی هوش مصنوعی (۵): نظام تنظیم گری هوش مصنوعی در کشور	۱۴۰۳	۲۰۵۴۷	حکمرانی	مسئولیت پذیری و پاسخگویی را از دلایل نیاز به تنظیم گری در حوزه هوش مصنوعی می داند.
۴	حکمرانی هوش مصنوعی (۸): هوش مصنوعی توضیح پذیر و نقش آن در ورود فناوری هوش مصنوعی به بخش عمومی	۱۴۰۴	۲۰۹۶۸	حکمرانی	مسئولیت پذیری را به خصوص در ورود فناوری هوش مصنوعی به بخش عمومی بسیار لازم به توجه می داند.

مأخذ: یافته های پژوهش.

۳. مفهوم مسئولیت پذیری

«مسئولیت» مفهومی پایه ای در تفکر و عمل اخلاقی است. می توان گفت محوری ترین موضوع اخلاقیات، «اعمال» هستند. اموری که با صفاتی همچون اخلاقی، غیر اخلاقی، یا نا اخلاقی مورد ارزش گذاری قرار می گیرند، از سنخ «افعال» و «اعمال» هستند.^۱

۱. البته در برخی نظریه های اخلاقی، ارزش اخلاقی نیت و مقاصد نیز مورد توجه قرار می گیرد.



همچنین، نظام پاداش و مجازات که برای پیاده‌سازی اخلاق در جامعه استفاده می‌شود، شامل تجویزهایی برای چگونگی برخورد با مصدر «اعمالی» هستند که پیش‌تر دارای ارزش اخلاقی مثبت یا منفی شناخته شده‌اند. برای اینکه بتوانیم وضعیت‌هایی را به‌عنوان «عمل» یا «فعل» شناسایی کنیم و از «رفتار» متمایز کنیم، باید برای آن وضعیت منشئی در نظر بگیریم که به نحو خاصی در تحقق آن وضعیت نقش داشته است. برای مثال، اگر ضربه بی‌مقدمه دست یک فرد به صورت فردی دیگر به‌عنوان «فعل» قابل ارزش‌گذاری اخلاقی باشد، باید بتوانیم آن را از ضربه شاخه درخت به صورت فرد یا ضربه غیرارادی دست فردی بیهوش متمایز کنیم. در حالت اول، منشأ رویداد، یک شخص با قدرت، اراده و آگاهی بوده که تعمداً به‌صورت طرف مقابل ضربه زده است. در حالات دیگر، منشأ رویداد امور مادی مانند باد و شاخه درخت یا تحریکات عصبی فرد بیهوش هستند. اگر میان یک وضعیت در جهان، و منبعی با قابلیت‌ها و ویژگی‌های مشخص، اتصال علی خاصی برقرار شود، می‌توان آن منبع را بابت آن وضعیت «مسئول» دانست. به همین ترتیب، مسئولیت در قبال یک وضعیت پیامدهایی مانند استحقاق دریافت پاداش، مجازات و لزوم پاسخگویی نسبت به آن وضعیت دارد.

یکی از مسائل نظری در مورد مسئولیت‌پذیری، تحلیل شرایط لازم و کافی برای انتساب مسئولیت اخلاقی است. این شرایط مشخص می‌کنند که در چه مواقعی می‌توان کسی را بابت وضعیتی مسئول دانست. برای مثال رویکرد ارسطویی، مسئولیت‌پذیری اخلاقی را شامل ۲ نوع مؤلفه می‌داند: مؤلفه‌ای مربوط به آزادی یا کنترل، و مؤلفه معرفتی. براساس این دیدگاه، برای آنکه منشأ یک عمل را دارای مسئولیت اخلاقی بدانیم، لازم و کافی است که آن منشأ، عمل را آزادانه و با کنترل انجام داده و نسبت به پیامدهای آن عمل معرفت داشته باشد. علاوه بر این، برخی به ۲ شرط کنترل و معرفت، شرط «شایستگی اخلاقی»^۱ را نیز اضافه می‌کنند [۱]، [۲]، [۳].

با این حال، درباره این شروط و همچنین تعابیر آن، اتفاق نظر وجود ندارد. در ادامه این شروط تفصیل می‌شود.

● **شرط آزادی:** شرط مربوط به آزادی که اغلب در ارتباط با مفهوم اراده آزاد و اختیار مطرح می‌شود، به این معناست که فاعل، عمل را مختارانه و نه از روی جبر انجام داده باشد. اما اینکه آزادانه انجام دادن عمل به چه معنا بوده، خود محل اختلاف است. برای مثال برخی فیلسوفان [۴] امکان انجام عملی غیر از آنچه انجام شده را شرط آزادانه بودن اعمال می‌دانند، درحالی‌که برخی دیگر [۵] معتقدند اگر یک فرد منشأ عمل خود باشد و آن را براساس خواست خود انجام دهد آزادانه عمل کرده است. همچنین اگر در حالتی عمل از او سر بزند که نسبت به دلایل حساسیت داشته و پاسخگو باشد [۶] در این صورت، شرط امکان‌پذیری انجام عمل به طریقی دیگر موضوعیت ندارد. فیشتر تعبیر اول (مبتنی بر امکان‌پذیری انجام کار دیگر) را آزادی تنظیمی، و تعبیر دوم (منشأ عمل بودن و پاسخگو به دلیل بودن) را آزادی هدایتی می‌نامد [۷]. اینکه کدام تعبیر از آزادی را بپذیریم، در تعیین شرایط مسئولیت‌پذیری تفاوت ایجاد می‌کند.

● **شرط معرفتی:** می‌توان معرفت را به معنای افلاطونی آن، معادل «باور صادق موجه» در نظر گرفت که در این صورت حالتی ذهنی بوده و تنها محدود به امور ذهن‌مند خواهد بود. یا اینکه آن را از سنخ داشتن «اطلاعات» در نظر بگیریم که در این صورت به امور غیرذهن‌مند نیز قابل اطلاق خواهد بود. علاوه بر این، می‌توان درباره اینکه آیا فقدان معرفت افراد را از مسئولیت اخلاقی معاف می‌کند، مناقشه کرد. برای مثال اگر پزشکی نسبت به عوارض یک دارو معرفت نداشته باشد و بر این اساس، دارویی را تجویز کرده و به مرگ بیمار منجر شود، آیا می‌توان او را فاقد مسئولیت دانست؟ اطلاق مسئولیت اخلاقی در این مورد و موارد مشابه، مستلزم آن است که بپذیریم در مرتبه‌ای بالاتر، «کسب معرفت» نیز می‌تواند در زمره مسئولیت‌های اخلاقی افراد باشد، و چنین نیست که به‌طور کلی فقدان معرفت باعث سلب مسئولیت از افراد شود.

● **شرط شایستگی اخلاقی:** براساس این شرط، فرد برای اینکه مسئولیت اخلاقی داشته باشد، باید بتواند درکی از خیر و شر اخلاقی داشته باشد و نسبت به آنچه درک کرده است، پاسخگو باشد. در نتیجه، افرادی که به‌علت اختلالاتی ذهنی یا تربیت نادرست

از درک ارزش اخلاقی ناتوانند، به لحاظ اخلاقی مسئولیت ندارند. مدافعین شرط شایستگی اخلاقی معتقدند در مورد افرادی که فاقد این شایستگی هستند، احترام نگذاشتن به دیگران به معنای بی‌احترامی نیست، زیرا بی‌احترامی مستلزم درکی از قوانین و سپس نقض آنهاست. درحالی‌که این افراد از درک قوانین ناتوان هستند [۸].

مسئله دیگر درباره مفهوم مسئولیت اخلاقی، به پیامدهای انتساب مسئولیت بازمی‌گردد. مسئول بودن یک فرد در ارتباط با یک وضعیت برای آن فرد چه پیامدهایی دارد؟ برخی متفکرین، مسئولیت را معادل استحقاق برای پاداش^۱ و مجازات^۲ می‌دانند. در این صورت، رابطه‌ای دوطرفه میان پاداش/مجازات و مسئولیت‌پذیری وجود دارد. برای مجازات‌پذیری، لازم است فرد مسئولیت داشته باشد، و اگر در موردی مجازات‌پذیری فرد محرز شود، می‌توان گفت در آن مورد مسئولیت دارد.

برخی دیگر مسئولیت‌پذیری را معادل اتصاف‌پذیری^۳ می‌دانند، به این معنا که در صورتی فرد نسبت به یک عمل مسئول است که بتوان آن را به‌درستی مسبب آن عمل دانست و آن عمل را به او نسبت داد. در رویکردی دیگر، مسئولیت‌پذیری ملازم با توضیح‌دهندگی^۴ نسبت به عمل است. در این مورد، توضیح اینکه چرا یک وضعیت حاصل شده است به منشأ و مسئول عمل ارجاع می‌دهد، حتی اگر این توضیح توسط فردی غیر از خود او ارائه شود. با اندکی تفاوت، برخی پاسخگویی^۵ نسبت به عمل را معادل مسئولیت اخلاقی می‌دانند. در این مورد اگر فردی نسبت به عملی مسئول باشد، باید در مورد آن به دیگران پاسخگو باشد. بسته به اینکه کدام یک از دیدگاه‌های فوق درباره مسئولیت‌پذیری پذیرفته شود، شرایط لازم و کافی برای مسئولیت نیز متفاوت خواهد بود. برای مثال، اگر مسئولیت‌پذیری معادل مجازات‌پذیری باشد، برآورده شدن شرط کنترل به معنای امکان انجام کاری غیر از آنچه انجام شده، ضرورت پیدا می‌کند؛ زیرا کسی را که نمی‌توانسته به گونه‌ای دیگر عمل کند، نمی‌توان بابت آنچه انجام داده مجازات کرد. اما اگر مسئولیت‌پذیری معادل پاسخگویی باشد، نیازی نیست که فرد امکان انجام کاری غیر از آنچه انجام داده را داشته باشد، بلکه همین که با نیت و براساس دلیل آن کار را انجام داده باشد، برای این که نسبت به آن کار پاسخگو باشد کفایت می‌کند.

جدول ۲. تعابیر مختلف از مسئولیت و شروط لازم برای آن

پیامدهای مسئولیت	شرط لازم برای مسئولیت‌پذیری
استحقاق پاداش و مجازات	امکان انجام کاری دیگر
اتصاف‌پذیری	منشأ عمل بودن
پاسخگویی و توضیح‌دهندگی	آگاهی نسبت به چگونگی وقوع عمل

مأخذ: همان.

یکی دیگر از ابعاد مفهوم مسئولیت‌پذیری، تفاوت مسئولیت‌پذیری آینده‌نگر (رو به جلو)، و مسئولیت‌پذیری گذشته‌نگر (رو به عقب) است.

● **رویکرد رو به جلو:** این رویکرد مسئولیت را در قبال آنچه هنوز رخ نداده است لحاظ می‌کند و بر اینکه چه اعمالی باید انجام شوند تا نتایج مثبت حاصل شده و از نتایج منفی پرهیز شوند، تمرکز دارد. در این رویکرد فرد به سبب مسئولیتی که دارد، ملزم می‌شود تا اعمالی را برای ایجاد نتایجی خاص انجام دهد [۸]. ازجمله این اعمال، می‌تواند کسب آمادگی برای عاملیت اخلاقی و کسب مهارت‌هایی باشد که برای چنان عاملیتی لازم هستند. نمونه‌ای از این مدل مسئولیت‌پذیری را می‌توان در کارهای وارگاس [۹]،

1. Praiseworthiness
2. Blameworthiness
3. Attributability
4. Accountability
5. Answerability



جفرسن [۱۰]، و مک‌گیر [۱۱] دید. براساس این تعبیر از مسئولیت، افراد نقش‌های اجتماعی که پذیرفته‌اند (از نقش‌های خانوادگی تا شغلی و شهروندی) خود را ملزم به انجام اعمال مختلف و کسب مهارت‌های گوناگون می‌دانند.

● **رویکرد رو به عقب:** برخلاف رویکرد آینده‌نگر که پیش‌دستانه بوده، رویکرد گذشته‌نگر، رویکردی واکنشی است. در این رویکرد، مسئولیت در قبال آنچه از پیش اتفاق افتاده است معنا پیدا می‌کند. اگر عملی موجب ایجاد وضعیتی مثبت و سودمند شده باشد، عامل آن مستحق تحسین و دریافت پاداش بوده و اگر عملی به ایجاد وضعیتی زیان‌بار منجر شده باشد، عامل آن مستحق سرزنش، مجازات و جبران خسارت است. در این حالت، فردی که مرتکب خطا شده باید احساس گناه کرده و فردی که مرتکب عملی صحیح شده باید احساس رضایت داشته باشد [۸]. نمونه‌هایی از این دیدگاه را می‌توان در آثار کارلسون [۱۲] و گراهام [۱۳] دید. براساس همین تعبیر از مسئولیت است که در زندگی روزمره، افراد بابت کاری که انجام می‌دهند دستمزد گرفته و بابت خطایی که سر می‌زند خسارت می‌دهند.

ابعاد فوق و برخی جنبه‌های دیگر (مانند شک‌گرایی نسبت به مسئولیت اخلاقی) خطوط اصلی مباحث مطرح شده حول مفهوم‌پردازی مسئولیت اخلاقی را تشکیل می‌دهند. همان‌طور که دیده می‌شود، رویکردهای متنوعی نسبت به تعابیر، الزامات، شرایط و پیامدهای مسئولیت اخلاقی وجود دارد. با این حال، این مسائل اغلب حول مفهومی فردی و همچنین انسانی از مسئولیت اخلاقی مطرح می‌شوند. اضافه شدن قید جمعی و غیرانسانی، پیچیدگی مسائل مربوط به مسئولیت‌پذیری را بیش از پیش افزایش می‌دهد.

۴. مسئولیت جمعی، مسئله دست‌های بسیار و معضل شکاف مسئولیت‌پذیری

به‌طور کلاسیک، تحلیل مسائل مربوط به مسئولیت‌پذیری با تکیه بر مثال‌هایی درباره مسئولیت اخلاقی شخصی و فردی پیش می‌رفت. اما از اوایل قرن ۲۱، توجه فیلسوفان به مفهوم مسئولیت جمعی^۱ یا مشترک^۲ جلب شد. یکی از زمینه‌های اهمیت یافتن مفهوم مسئولیت جمعی، توسعه فناوری و اعمال فناورانه بود. در اعمال فناورانه، گروه متکثری از افراد حقیقی و حقوقی در رقم زدن یک خروجی نقش ایفا می‌کنند، افرادی که اغلب با یکدیگر مرتبط نیستند و حتی ممکن است از وجود یکدیگر بی‌خبر باشند. در مورد فناوری‌های هوش مصنوعی، این مسئله یک پیچ‌وتاب اضافه نیز پیدا می‌کند، بخشی از این شبکه عاملان، می‌تواند مصنوعی بوده و اصلاً انسانی نباشد. در چنین مواردی، فهم ابعاد مختلف مسئولیت و نسبت عاملان مختلف با آن، بسیار پیچیده‌تر از مسئولیت اخلاقی فردی خواهد بود.

مناقشات مربوط به مسئولیت جمعی را می‌توان در سه دسته جای داد.

الف) آیا می‌شود به یک جمع مسئولیت اخلاقی نسبت داد؟ آیا شروطی را که برای مسئولیت اخلاقی درباره افراد برمی‌شماریم (همچون داشتن قصد و نیت)، می‌توان فراتر از افراد به خود گروه به‌عنوان یک هویت جمعی نسبت داد؟ و اگر چنین است، گروه‌هایی که می‌توانند گزینه‌ای برای مسئولیت جمعی باشند (مانند شرکت و ملت، یا جوامع بومی، جنبش‌های اجتماعی، و باشگاه‌ها) چه ساختاری باید داشته باشند؟ آیا به هر انباشتی^۳ از افراد می‌توان مسئولیت جمعی نسبت داد یا لازم است گروهی خوشه‌ای^۴ و سامان‌یافته داشته باشیم؟ چه میزان از همبستگی، هم‌فکری، تعامل و مشارکت برای اینکه یک گروه واجد مسئولیت جمعی بشوند کفایت می‌کند؟

ب) مسئولیت میان اعضای یک جمع چطور توزیع می‌شود؟ پرسش این است که آیا می‌توان مسئولیت اخلاقی یک گروه را

1. Collective Responsibility
2. Shared Responsibility
3. Aggregative
4. Conglomerate

میان اعضای آن توزیع کرد؟ آیا می‌توان گفت تک‌تک اعضا بابت خروجی‌های گروهی که در آن عضویت دارند، مسئول هستند؟ آیا می‌توان برخی اعضا را بابت کاری که اعضای دیگر انجام داده‌اند، یا حتی عملی که از گروه به‌عنوان یک هویت سرزده است، مسئول دانست؟ در بررسی این موضوع، بیشتر مسئولیت‌پذیری مشترک (و نه مسئولیت‌پذیری جمعی) مورد اشاره قرار می‌گیرد. مسئولیت‌پذیری جمعی در مورد هویتی جمعی مانند یک شرکت مطرح بوده که چیزی فراتر از جمع اعضای آن است، درحالی‌که مسئولیت‌پذیری مشترک مربوط به عاملان اخلاقی منفردی است که به‌عنوان اعضای یک گروه در ایجاد یک وضعیت نقش داشته‌اند. **ج) در چه مواردی مناسب‌تر است که به جای مسئولیت فردی، از مسئولیت جمعی سخن بگوییم؟** برای مثال آیا ارجاع به مسئولیت جمعی در موضوعاتی چون تروریسم، نژادپرستی و ... به جای مسئولیت فردی، با ارزش‌هایی همچون عدالت و آزادی سازگار است و آنها را ارتقا می‌دهد [۱۴]؟

در مورد اعمالی که با وساطت فناوری انجام می‌شود، بیشتر بُعد دوم مسئولیت جمعی یا اشتراکی، یعنی **چگونگی توزیع مسئولیت** مورد توجه قرار گرفته است. این مسئله خصوصاً در دیدگاه‌هایی که ماهیت فناوری را سامانه‌ای^۱ در نظر می‌گیرند، اهمیت بیشتری پیدا می‌کند. نگاه سامانه‌ای به فناوری، در نقطه مقابل نگاه ابزاری به فناوری قرار دارد. در نگاه ابزاری، فناوری امری سخت‌افزاری یا نرم‌افزاری است که هر کس می‌تواند آن را مطابق اهداف خود به کار گیرد. اما براساس نگاه سامانه‌ای، برای فهم فناوری نمی‌توان آن را امری منزوی از هویت و ساختارهای انسانی و اجتماعی در نظر گرفت. یک سامانه اجتماعی-فناورانه، از اشیاء فیزیکی، مهارت‌های انسانی و نهادهای اجتماعی تشکیل شده است. اجزای این سامانه‌ها ترکیبی از سخت‌افزارها، دانش، مخترعان، راه‌اندازان، مصرف‌کنندگان، بازاربایان، تبلیغ‌کنندگان، مأمورین دولت و دیگر افرادی هستند که در شکل‌گیری عمل فناورانه دخیل هستند. فیلسوفانی که نگاه سامانه‌ای به فناوری دارند معتقدند اگر بخواهیم درک درستی از فناوری داشته باشیم نباید صرفاً بر یک شی و ویژگی‌هایش متمرکز شویم، بلکه باید گستره وسیع‌تری از هویت و ارتباطاتی که شی در آن جای گرفته را نیز لحاظ کنیم. این رویکرد از رویکردهای سخت‌افزاری و نرم‌افزاری جامع‌تر بوده و آنها را در برمی‌گیرد [۱۵].

جدول ۳. انواع رویکردها به فناوری

نمونه بارز	انواع رویکردها به فناوری
چکش، مداد، قیچی	فناوری به‌مثابه سخت‌افزار
روش تدریس، تکنیک جراحی، روش برنامه‌نویسی، روش‌های کشاورزی و دامپروری	فناوری به‌مثابه روش نرم‌افزار
سامانه‌های هواپیمایی متشکل از هواپیماها، فرودگاه‌ها، قوانین، خلبانان، برنامه‌های آموزشی، نهادهای نظارتی و ...	فناوری به‌مثابه سامانه اجتماعی-تکنیکی

مأخذ: همان.

فرماس و همکارانش از جمله افرادی هستند که نگاهی سامانه‌ای به فناوری دارند و فناوری را تحقق‌یافته در سامانه‌های اجتماعی-تکنیکی که آمیزه‌ای از هویت اجتماعی و مصنوعی هستند، در نظر می‌گیرند. از دید آنها، کسی که از یک مصنوع تکنیکی استفاده می‌کند، در واقع در یک سامانه پیچیده متشکل از هویت گوناگون عمل می‌کند، نه اینکه فقط یک شی فیزیکی ظاهراً منفرد را به کار گیرد. برای مثال فردی که سوار هواپیما می‌شود صرفاً از پله‌های یک پرندۀ فلزی بالا نمی‌رود؛ بلکه در واقع از یک سامانه هواپیمایی مسافربری جهانی استفاده می‌کند که از هویت مختلف مصنوعی و اجتماعی تشکیل شده مانند فرودگاه، هواپیما، دروازه‌های ورود و خروج؛ و قوانین، مسیرهای هوایی، شرکت‌های هواپیمایی، بیمه، مسئولین پرواز، خلبان، مهمانداران و برای اینکه یک هواپیما به‌عنوان یک مصنوع کارش را درست انجام دهد، لازم است تمام زیربخش‌های سامانه‌ای که در آن مستقر شده

1. Systemic



به درستی کار کنند [۱۶]. اگر چنین نگاهی را بپذیریم، یک عمل فناورانه عملی مشترک میان کنشگران متعدد انسانی و غیرانسانی است.

برای مثال فرض کنید راننده‌ای در روزی بارانی در حال عبور از خیابان و درحالی‌که چراغ راهنمایی تازه قرمز شده است، با عابری که وسط خیابان پریده، برخورد می‌کند و به او آسیب می‌زند. از منظر فردی، راننده مقصر است و باید مسئولیت اتفاق رخ داده را بپذیرد. اما اگر زاویه دید را وسیع‌تر کنیم، می‌بینیم که عوامل متعدد فناورانه و اجتماعی در این سانحه دخیل بوده‌اند. برای مثال، خودروی راننده ساخت شرکتی است که ترمز تمام محصولات آن ایراد داشته و با اندکی تأخیر کار می‌کند؛ چراغ راهنما که بلافاصله پس از زرد شدن قرمز شده و به راننده برای توقف ماشین مهلت نمی‌دهد، آسفالت خیابان که فرسوده شده و در روزهای بارانی به شدت لیز می‌شود، نهادهای نظارتی و شهرداری که به تماس‌های شهروندان برای اصلاح وضعیت چراغ راهنما و آسفالت اهمیت نمی‌دهند، و نهادهای آموزشی که به شهروندان در مورد چگونگی عبور از خیابان آموزش‌های لازم را نداده‌اند.

این عوامل تنها تعدادی از عواملی هستند که بالقوه می‌توانند در وقوع حادثه ذکر شده مؤثر باشند. با در نظر گرفتن این عوامل، نقش عامل اولیه یعنی راننده، کم‌رنگ‌تر شده و دست‌های دیگری که به واسطه زنجیره‌های علیّ متعدد به حادثه فوق مرتبط می‌شوند رؤیت‌پذیر می‌شوند. به این مسئله در اخلاق فناوری «مسئله دست‌های بسیار»^۱ گفته می‌شود. مسئله دست‌های بسیار، به واقعیت اشتراکی بودن مسئولیت در اعمال فناورانه اشاره دارد، و اینکه عوامل متعددی در وقوع یک عمل فناورانه نقش دارند که شناسایی آنها و تشخیص سهم هر یک در آنچه رخ داده، کاری دشوار است.

اما مسئله توزیع مسئولیت، می‌تواند از این نیز پیچیده‌تر شود. با پیشرفت فناوری‌های هوش مصنوعی و ظهور فناوری‌های مجهز به یادگیری عمیق که می‌توانند درجاتی از خودمختاری را بروز دهند، این پرسش مطرح می‌شود که آیا ممکن است فناوری نیز بتواند واجد مسئولیت اخلاقی باشد؟ این پرسش از جهاتی به پرسش از مسئولیت جمعی شباهت دارد؛ زیرا در آن مورد نیز مسئله بررسی امکان مسئولیت اخلاقی برای هویتی ساختگی (مانند شرکت، ملت، و ...) بوده که با مصداق بارز مسئول اخلاقی (یک انسان بالغ و عاقل) متفاوت است. در اینجا نیز مسئله امکان انتساب مسئولیت اخلاقی، به هویتی غیرانسانی و ساختگی مطرح است که از جهاتی به انسان‌ها شباهت دارد. فناوری‌های هوش مصنوعی با یادگیری عمیق، می‌توانند توسط داده‌های محیطی که در آن توسعه پیدا می‌کنند بیاموزند^۲ و خارج از الگوریتم‌هایی که توسط برنامه‌ریز روی آنها ریخته شده، عمل کنند. در نتیجه، خروجی‌هایی خواهند داشت که خارج از کنترل و خواست طراحانشان است.

با ورود این نوع فناوری‌ها به سامانه‌های اجتماعی-فناورانه، وضعیت‌هایی قابل وقوع است که در آنها نمی‌توان به درستی مسئولیت وضعیت موجود را به عاملان انسانی نسبت داد زیرا شروط ذکر شده برای مسئولیت اخلاقی (مانند آگاهی، کنترل، ابتنا بر دلیل و ...) در مورد هیچ یک از عوامل انسانی دخیل در وضعیت فوق به میزان کافی برآورده نمی‌شود. ماتیاس چنین وضعیتی را معضل «شکاف مسئولیت‌پذیری»^۳ می‌خواند. او به نمونه‌های متعددی از فناوری‌های هوش مصنوعی اشاره می‌کند که در آنها برنامه‌نویس و کاربر نقش بسیار حداقلی در عمل نهایی دارند و می‌گویند با بیشتر شدن و پیچیده‌تر شدن شبکه‌های فناوری، کنترل عامل‌های انسانی کاهش یافته و شکاف مسئولیت عمیق‌تر می‌شود [۱۷]. در نتیجه ما با حوادثی مواجه می‌شویم که نمی‌توان کسی را بابت خسارت‌های آن مسئول دانست یا به جبران آن وادار کرد.

جدول ۴. مسئله دست‌های بسیار و شکاف مسئولیت

مسئله دست‌های بسیار	مسئولیت چطور بین عوامل انسانی متعدد دخیل در یک وضعیت توزیع می‌شود؟
مسئله شکاف مسئولیت	در مواردی که هیچ انسانی آگاهی و کنترل لازم برای پذیرش مسئولیت یک وضعیت را ندارد، چه کسی مسئول است؟

مأخذ: همان.

1. The Problem of Many Hands
2. Training
3. The Responsibility Gap

یکی از مثال‌های ماتیاس، ربات آیبو^۱ بوده که توسط شرکت سونی در سال ۱۹۹۹ تولید شده است. این ربات به عنوان اسباب‌بازی هوشمند، جایگزینی برای حیوانات خانگی در آپارتمان‌های کوچک است. آیبو قابلیت یادگیری دارد و علاوه بر یادگیری دستورات شفاهی و عمل براساس آنها، می‌تواند برخی حرکات فیزیکی جدید را نیز بیاموزد. برای مثال اگر الگوریتم‌های حرکتی که از ابتدا برای او برنامه‌ریزی شده‌اند، برای حرکت او در منزل بهینه نباشند، آیبو می‌تواند شیوه‌های حرکتی دیگری را امتحان کند. ماتیاس می‌گوید قابل تصور است که نسخه پیشرفته‌ای از این ربات برای بهینه‌تر کردن عملکرد کلی، شیوه حرکتی مناسب خود را پیدا کند. در این مورد، ربات برای طولانی‌تر کردن عمر باتری خود، حرکتی تاخت‌وار را انتخاب می‌کند که باعث کم شدن اصطکاک و صرفه‌جویی در انرژی می‌شود. تاخت سریع ربات در منزل، باعث می‌شود او با کودک برخورد کرده و به او آسیب بزند. ماتیاس می‌گوید در این مورد نمی‌توان دلیلی قانع‌کننده ارائه داد برای اینکه تولیدکننده یا والدین کودک را مسئول بشماریم. آنچه رخ داده، پیشرفتی غیرقابل پیش‌بینی و محصول توانایی وفق‌پذیری ربات بوده است. به همین علت نمی‌توان منصفانه کسی را بابت آن مسئول دانست. یادگیرندگی ماشین‌ها، مسبب بروز خودمختاری و خارج از کنترل بودن این ماشین‌ها می‌شود [۱۷].

با افزایش تعداد و پیچیده‌تر شدن ماشین‌های یادگیرنده، مسئله شکاف مسئولیت اهمیت بیشتری پیدا می‌کند. در انواع شیوه‌های برنامه‌نویسی هوش مصنوعی (از سیستم‌های نمادین، سیستم‌های پیوندگرا و شبکه‌های عصبی، الگوریتم‌های ژنتیکی و نهایتاً عامل‌های خودمختار) به تدریج سهم کنترل برنامه‌نویس در عملکرد برنامه و خروجی نهایی کمتر می‌شود. قوانین عملکرد ماشین، طی تعامل آن با محیط شکل می‌گیرد. در نتیجه اعمال این فناوری‌ها نه کاملاً قابل پیش‌بینی و نه قانوناً قابل پیگیری است.

وجود شکاف مسئولیت در یک سامانه فناورانه هوش مصنوعی را می‌توان براساس برخی شاخص‌ها شناسایی کرد. برای مثال سطح پایین تبیین‌پذیری عملکرد سامانه، مشخص نبودن دامنه امکان‌های عملکردی سامانه و امکان ایجاد خروجی‌های پیش‌بینی‌ناپذیر، فقدان یا ابهام نقشه روابط میان سامانه و نهادها و سامانه‌های اجتماعی-فنی دیگر، تعریف نشدن جایگاه حقوقی دارای اطلاعات و قدرت برای توضیح و کنترل عملکرد سامانه، واگذاری تصمیم‌های پرریسک به سامانه بدون سطح‌بندی مسئولیت انسانی یا ابهام در نقش کاربر، توسعه‌دهنده و نهاد بهره‌بردار در پیامدهای منفی.

برای مرتفع ساختن شکاف مسئولیت، چند رویکرد پیش گرفته شده است. فناوری هراسان این معضل را دشواری غیرقابل رفع فناوری‌های هوش مصنوعی دانسته و بر آن به‌عنوان دلیلی برای توقف توسعه این فناوری‌ها تکیه می‌کند. برخی فیلسوفان برای غلبه بر این معضل، قائل به مسئولیت‌پذیری هوش مصنوعی شده‌اند. از دید آنها در مواردی که نمی‌توان عاملی انسانی را مسئول وضع پیش‌آمده دانست، می‌توان به درستی مسئولیت اخلاقی این وضعیت را به هوش مصنوعی نسبت داد. برخی دیگر، علیه این موضع استدلال کرده و بر امکان‌ناپذیری انتساب مسئولیت اخلاقی به هوش مصنوعی پافشاری می‌کنند. گروهی دیگر از اندیشمندان مناقشه درباره امکان یا عدم امکان انتساب مسئولیت اخلاقی به هوش مصنوعی را بحثی ثمربخش ندانسته و به جای آن رویکردی هنجاری اتخاذ می‌کنند. در چنین رویکردی، محدودیت‌های هوش مصنوعی لحاظ شده و برای دچار نشدن به معضل شکاف مسئولیت، باید‌ها و نباید‌هایی برای واگذاری وظایف خاص به هوش مصنوعی طرح می‌شود. در نهایت، جدیدترین رویکرد نسبت به معضل شکاف مسئولیت‌پذیری، با رویکردی آینده‌نگرانه نسبت به مسئولیت اخلاقی، راهکارهایی را پیشنهاد می‌دهد که در آن همواره مسئولیت عامل انسانی حفظ شده و از وقوع شکاف مسئولیت پیشگیری شود. در ادامه، رویکردهای ذکر شده با اشاره به آرای نمایندگان آنها بررسی می‌شود.



۵. هوش مصنوعی به عنوان عامل اخلاقی

عاملیت، به توانمندی یک هویت بر انجام عمل (و نه صرفاً رفتار) مربوط است. هنگامی که عمل مورد نظر ما دارای جنبه‌ای اخلاقی باشد، صحبت از عاملیت اخلاقی به میان می‌آید. همان‌طور که در بخش ۳ بیان شد، میان عاملیت اخلاقی و مسئولیت اخلاقی ملازمه وجود دارد. اما اندیشمندان مختلف درباره جهت این ملازمه (یا حتی برقراری آن) هم‌نظر نیستند. برخی معتقدند لازمه مسئولیت اخلاقی، عاملیت اخلاقی است، زیرا برای انتساب مسئولیت اخلاقی، باید بتوان عملی را به یک عامل اخلاقی نسبت داد. برخی بالعکس، معتقدند برای اینکه فردی عامل اخلاقی باشد، باید مسئولیت اخلاقی داشته باشد. زیرا تا زمانی که نتوانیم مسئولیت اخلاقی یک عمل را به هویتی نسبت دهیم، نمی‌توانیم به درستی آن را عامل اخلاقی بدانیم. به هر ترتیب، به علت اتصال نزدیکی که میان این دو مفهوم وجود دارد، بسیاری از اندیشمندان برای بررسی امکان مسئولیت‌پذیری هوش مصنوعی، بر مسئله عاملیت اخلاقی متمرکز شده‌اند. اگر بتوان نشان داد که هوش مصنوعی علاوه بر عامل علی یک وضعیت، نسبت به آن عاملیت اخلاقی نیز دارد، آنگاه مسیر برای انتساب مسئولیت اخلاقی به هوش مصنوعی بسیار هموار خواهد بود. پرسش از عاملیت اخلاقی مصنوعات، در ۳ بُعد وجودشناختی، معرفت‌شناختی و عملی امتداد پیدا می‌کند.

- چه ویژگی‌هایی قوام‌بخش شروط لازم و کافی برای داشتن عاملیت اخلاقی هستند؟
 - چطور می‌توانیم نسبت به وجود یا عدم این ویژگی‌ها در یک هویت معرفت پیدا کنیم؟
 - با فرض امکان‌پذیری عاملیت اخلاقی مصنوعات، چطور می‌توان ظرفیت‌های اخلاقی را در مصنوعات ایجاد کرد؟
- دو پرسش اول، در ارتباط با شکاف مسئولیت‌پذیری اهمیت پیدا می‌کنند، زیرا پاسخ آنها در اینکه هوش مصنوعی را مسئول بدانیم یا خیر، تعیین‌کننده است. اما پرسش سوم، زمانی اهمیت پیدا می‌کند که عاملیت اخلاقی هوش مصنوعی را پذیرفته و به دنبال طراحی مصنوعاتی باشیم که مطابق ارزش‌های اخلاقی عمل کنند. در ادامه برخی از پاسخ به این پرسش‌ها، بررسی خواهد شد.
- **هوش مصنوعی عاملیت اخلاقی دارد: تعاملی بودن، خودمختاری و انطباق‌پذیری**^۱ سه شرطی هستند که فلوریدی و ساندرس [۱۸] برای عاملیت اخلاقی معرفی کرده و هوش مصنوعی را واجد آن می‌دانند. تعاملی بودن به این معناست که عامل و محیط پیرامونش بتوانند با یکدیگر کنش داشته باشند. خودمختاری به معنای این است که عامل بتواند تغییر حالت دهد بی‌آنکه این تغییر، پاسخی مستقیم به تعامل باشد. براساس این شرط یک عامل خودمختار باید اقلأً دو حالت ممکن داشته و دارای سطح خاصی از پیچیدگی و استقلال از پیرامون خود باشد. منظور از انطباق‌پذیری نیز آن است که تعاملات عامل بتوانند قوانین‌گذاری را که براساس آن حالاتش را تغییر می‌دهد، عوض کنند. این ویژگی تضمین می‌کند که عامل حالت عملکردش را به‌طریقی که بسیار وابسته به تجربیاتش است بیاموزد. حال اگر عملی به خیر یا شر اخلاقی منجر شود، عاملی که مسبب آن عمل است، واجد عاملیت اخلاقی شناخته می‌شود. فلوریدی و ساندرس این شروط را در مورد فناوری‌های هوش مصنوعی قابل حصول می‌دانند.
- ممکن است کسی بگوید فناوری‌های هوش مصنوعی فاقد مسئولیت هستند. زیرا به علت نداشتن تعامل مناسب با محیط، برای مثال چون فاقد اطلاعات کافی هستند یا انتخابی ندارند، نباید آن عامل را برای عملش دارای مسئولیت اخلاقی بدانیم. اما از دید فلوریدی و ساندرس، در مورد هوش مصنوعی چنین نیست؛ زیرا این فناوری‌ها به میزان کافی تعاملی، خودمختار و وفق‌پذیر هستند که بتوان آنها را عامل اخلاقی دانست. از طرف دیگر، عاملیت اخلاقی بدون مسئولیت اخلاقی بی‌مایه است. بنابراین فلوریدی و ساندرس برای تسهیل سخن، عامل اخلاقی را مترادف عامل اخلاقی مسئول می‌دانند. باین حال از دید آنها اینکه ما عمل عاملی را به لحاظ اخلاقی ارزش‌گذاری کنیم، لزوماً معادل انتساب مسئولیت اخلاقی نیست. برای مثال والدین اعمال فرزندانشان را به لحاظ

1. Interactivity, Autonomy, and Adaptability

اخلاقی ارزش‌گذاری می‌کنند، حتی در سنینی که کودکان را بابت اعمالشان نمی‌توان مسئول دانست. همچنین در مورد حیوانات دست‌آموزی که با کمک به انسان کاری به‌لحاظ اخلاقی ارزشمند انجام می‌دهند و تشویق می‌شوند، همین نکته صادق است. زیرا این کار علی‌رغم ارزش اخلاقی، برای آن حیوان چیزی بیش از یک بازی نیست و برای او مسئولیتی به بار نمی‌آورد. همچنین اگر عاملی اخلاقی تمام تلاش خود را برای پرهیز از عملی شر انجام دهد اما نهایتاً این عمل از او سر بزند (مانند آنچه در داستان‌های اسطوره‌ای تراژیک رخ می‌دهد) باز هم با اینکه توضیح وقوع آن عمل به فرد بازمی‌گردد و به این معنا عامل اخلاقی است، اما نسبت به آن مسئول نیست.

فلوریدی در پژوهشی دیگر از دیدگاهی که در کار مشترکش با ساندرس مطرح کرده بود، فراتر می‌رود و ادعا می‌کند در مواردی که یک شبکه از عاملان مختلف متشکل از عوامل انسانی یا غیرانسانی (یا عواملی با ماهیت دوگانه)، خروجی اخلاقی داشته باشند، هم عاملیت اخلاقی و هم مسئولیت اخلاقی میان تمام نقاط این شبکه توزیع می‌شود. در این حالت، مسئولیت اخلاقی وابسته به وجود قصدمندی نیست. فلوریدی از تعبیر مسئولیت‌پذیری بدون ذهن‌مندی برای معرفی ایده خود استفاده می‌کند. او یک سامانه چندعاملی را همچون شبکه عصبی چندلایه مدل می‌کند [۱۹]. براساس چنین مدلی، عاملیت اخلاقی توزیع شده، نتیجه تعاملات عصبی میان گره‌های شبکه هستند. این پدیده که به نقش‌آفرینی گره‌های مختلف در خروجی منجر می‌شود، **انتشار رو به جلو** نام دارد. بر این اساس، اختصاص مسئولیت به هر گره براساس وضعیت اخلاقی خروجی، با پیگیری نقش آن گره در رقم زدن خروجی، صورت گرفته و انتشار رو به عقب نام دارد. فلوریدی در توضیح چگونگی مسئولیت گره‌ها (اعم از گره‌های انسانی و غیرانسانی) در چنین شبکه‌ای آمیخته‌ای، از مفهوم **ضمانت اکید** که مفهومی حقوقی بوده و به معنای لزوم جبران خسارت (حتی در صورت عدم مسئولیت اخلاقی) و معرفت مشترک (که به دانش مشترک میان گره‌های شبکه اشاره دارد)، استفاده می‌کند. در این رویکرد، فلوریدی (برخلاف مقاله مشترکش با ساندرس) مسئولیت را معادل توضیح‌دهندگی (و نه چیزی بیش از آن) در نظر می‌گیرد. برای توزیع مسئولیت باید دید کدام عاملان به لحاظ علی وضعیت سامانه را توضیح می‌دهند. پاسخگویی اخلاقی و ضمانت حقوقی نیز از این تعبیر نتیجه می‌شود. شرط انتساب مسئولیت در این شرایط، این است که عاملان بتوانند از رفتارهای خود بیاموزند و آن را اصلاح کنند، یعنی درجه‌ای از خودمختاری را داشته باشند. به این ترتیب، مسئولیت‌پذیری اخلاقی که فلوریدی از آن دفاع می‌کند، نوعی **مسئولیت اخلاقی بدون تقصیر** است؛ زیرا شرط قصدمندی در این نوع مسئولیت اخلاقی اهمیتی ندارد.

■ **هوش مصنوعی عاملیت اخلاقی دارد: خودمختاری، قصدمندی و مسئولیت‌پذیری** شروطی هستند که سالینز برای عاملیت اخلاقی معرفی می‌کند. او در تحلیل عاملیت اخلاقی مصنوعی، ربات‌ها را مدنظر قرار می‌دهد. همچنین، میان حیوانات دست‌آموز و ربات‌های پیشرفته نوعی توازی و تناظر برقرار می‌کند. از دید او، هم در آموزش دادن حیوانات و هم در ساخت ربات‌ها، ما طبیعت را برای مقاصد خود دستکاری می‌کنیم. سالینز معتقد است حیوانات برای اینکه دارای حقوق اخلاقی و سطحی پایین از مسئولیت اخلاقی باشند، دارای حد کافی خودمختاری هستند. علاوه بر این، در صورتی که حیوانات دست‌آموز کار اخلاقی ارزشمندی انجام دهند (برای مثال راهنمایی فردی نابینا) ما آنها را تحسین می‌کنیم [۲۰].

بنابراین از نظر سالینز، مسئولیت‌پذیری مفهومی پایه‌ای‌تر از عاملیت اخلاقی است و او آن را به‌عنوان شرطی برای عاملیت اخلاقی مطرح می‌کند. تعبیر مد نظر سالینز از خودمختاری، تعبیری فلسفی و حداکثر نیست. چنان‌که مطابق تعبیر غلیظ از خودمختاری، حتی خودمختاری انسان‌ها نیز مورد تردید است. تعبیر مد نظر او، تعبیری مهندسی است؛ یعنی استقلال از برنامه‌نویس، اپراتور و کاربر. همچنین، زمانی می‌توانیم یک ربات را قصدمند بدانیم که تعامل پیچیده برنامه‌های ربات و محیط باعث شود ماشین به‌طریقی عمل کند که به‌لحاظ اخلاقی زیان‌بار یا سودمند باشد و این اعمال ظاهراً محصول رایزنی و محاسبات باشند. در چنین شرایطی می‌توان گفت عمل ماشین قصدمندانه است. مطابق شرط مسئولیت‌پذیری، زمانی که یک ربات نقش‌هایی اجتماعی را برعهده داشته باشد که دربردارنده مسئولیت‌هایی اخلاقی باشند و تنها راه معنابخشی به عملکردش این باشد که فرض کنیم ربات «باور» دارد



نسبت به انجام وظایفش مسئول بوده، می‌توانیم بگوییم به‌اندازه کافی مسئولیت‌پذیر است. براساس این شروط، یک مصنوع برای اینکه عامل اخلاقی به حساب بیاید نیازی به این ندارد که دارای ذهن یا آگاهی باشد [۲۰].

هرچند فلوریدی، ساندرس و سالینز از امکان عاملیت اخلاقی مصنوعات دفاع می‌کنند، اما برخی فیلسوفان با آنها هم‌نظر نیستند. از دید این فیلسوفان شرايطی اساسی برای عاملیت اخلاقی وجود دارد که توسط فناوری‌های هوش مصنوعی برآورده نمی‌شود.

■ **هوش مصنوعی عاملیت اخلاقی ندارد: خودمختاری و مسئولیت‌پذیری** شروطی هستند که ولیز آن را برای عاملیت اخلاقی لازم می‌داند و معتقد است الگوریتم‌ها (هوش مصنوعی) هیچ یک از این دو ویژگی را ندارند. علت اصلی او برای این ادعا این است که از دید او الگوریتم‌ها فاقد حس‌مندی^۱ هستند. از دید او، برای داشتن مسئولیت اخلاقی، قابلیت فهم اخلاقی ضرورت دارد. اما فهم اخلاقی نیز به نوبه خود مستلزم داشتن درکی از آثار رنج و آسیب است. این درک، با معرفتی تجربی از درد حاصل می‌شود. منظور از معرفت تجربی، چیزی فراتر از اطلاعات بوده، زیرا معرفت به چگونگی تجربه درد و رنج است که جز با تجربه واقعی این امور قابل حصول نیست. درحالی‌که هوش مصنوعی فاقد چنین تجربه‌ای است [۲۱].

ولیز می‌گوید هوش مصنوعی می‌تواند اعمال خیر و شر را مانند انسان انجام دهد، برای مثال به خیریه کمک کرده، از افراد آسیب‌پذیر محافظت کند، یا مرتکب بی‌احترامی و خیانت شود. اما نمی‌تواند مانند انسان احساساتی مانند همدلی، رنج، لذت، صمیمیت و شرم را تجربه کند. از دید ولیز ما می‌توانیم میان توانمندی بر ایجاد خیر و شر از یک‌سو، و عاملیت اخلاقی از سوی دیگر تفاوت قائل شویم. چنان‌که برای مثال طوفان نیز می‌تواند به افراد زیادی آسیب بزند، اما ما به آن عاملیت اخلاقی نسبت نمی‌دهیم. او به این نکته اشاره می‌کند که در ادبیات علمی اخیر در حوزه هوش مصنوعی، نویسندگان مختلفی صراحتاً یا ضمنی به این نکته اشاره کرده‌اند که به‌علت ویژگی‌های خاص هوش مصنوعی می‌توان به‌نحوی استعاری یا عملگرایانه طوری درباره هوش مصنوعی سخن گفت، که گویی دارای عاملیت اخلاقی است و حتی نیازی نیست که درباره اینکه واقعاً دارای عاملیت اخلاقی هستند یا خیر، مذاقه کرد. ولیز این دیدگاه را نمی‌پذیرد. از دید او، پرسش از عاملیت اخلاقی هوش مصنوعی صرفاً یک کنجکاوی متافیزیکی نیست، بلکه برای ما پیامدهایی مستقیم دارد. اگر برای هوش مصنوعی عاملیت اخلاقی قائل نباشیم، انسان‌ها باید مسئولیت اعمال هوش مصنوعی را بپذیرند و اگر عاملیت قائل باشیم، باید مسئولیت اخلاقی برای هوش مصنوعی قائل بوده و آن را در صورت لزوم مجازات کنیم. این دوراهی، ما را وادار می‌کند مسئله مسئولیت‌پذیری هوش مصنوعی را جدی بگیریم.

ولیز در تشریح دو شرط پیشنهادی خود برای عاملیت اخلاقی (یعنی خودمختاری و مسئولیت‌پذیری اخلاقی) از میان تعبیر متعددی که از خودمختاری وجود دارد، خودحکمرانی و پاسخگویی به علت را دو تعبیر مرتبط با عاملیت اخلاقی می‌داند. از دید او الگوریتم‌ها فاقد خودحکمرانی بوده؛ زیرا برای تعیین اهداف خود نیازمند دریافت ورودی از جهان بیرونشان هستند. همچنین پاسخگو به علت نیز نیستند زیرا هیچ علتی نمی‌تواند آنها را نسبت به تغییر هدفی که برایش برنامه‌ریزی شده‌اند قانع کند. علی‌رغم اینکه به‌علت رفتار بسیار شبیه و در واقع غیرمتمایز از انسان، ممکن است در ظاهر دارای خودحکمرانی و پاسخگو به علت به نظر برسند.

عاملان می‌توانند دارای مسئولیت اخلاقی باشند حتی اگر استقلال ذهنی و بدنی چندانی نداشته باشند، یا (مانند بردگان) حق اخلاقی‌شان برای خودتعیینی نقض شده باشد. مسئولیت اخلاقی نزد ولیز، با توضیح‌دهنده بودن، و به تبع آن با پاسخگویی و مجازات‌پذیری همبستگی دارد. او توضیح‌دهندگی را معادل پاسخگویی^۲ به دیگران درباره شیوه زندگی می‌داند. فردی نسبت به دیگران پاسخگو است که بتواند علایق و ادعاهای اخلاقی دیگران را تشخیص دهد و در صورتی که آنها را محترم بشمارد مستحق سرزنش و تنبیه باشد. اغلب انسان‌های بالغ این شرط را برآورده می‌کنند، اما الگوریتم‌ها به‌علت اینکه فاقد قصد و امکان عمل به‌نحوی غیر از آنچه ورودی و برنامه‌شان تعیین می‌کند هستند، مسئولیت‌پذیر و پاسخگو نیستند.

■ **هوش مصنوعی عاملیت اخلاقی ندارد: اختیار و آگاهی** شروطی هستند که مانا و ناث آن را لازمه عاملیت اخلاقی می‌دانند.

1. Sentience
2. Answerability

آنها تحلیل خود را در چارچوب نظریه اخلاق کانتی قرار می دهند. از دید آنها، از آنجاکه عمل هوش مصنوعی ناشی از اختیار و احساس وظیفه نیست و در واقع عمل براساس دستورات برنامه ریز است، اعمال آنها نااخلاقی محسوب می شود. آنها اختیاری برای عمل به نحوی دیگر ندارند. بنابراین، هرچند می توان آنها را ماشین های تولید شادی دانست، اما نمی توان ماشینی اخلاقی خواند؛ زیرا در غیاب حداقلی از عاملیت آگاهانه، نمی توان از اراده آزاد سخن گفت و بدون داشتن اراده آزاد نیز عاملیت اخلاقی معنا ندارد [۲۲]. به این ترتیب، از دید مانا و ناث آگاهی شرطی اساسی برای داشتن اختیار، و اختیار شرطی اساسی برای داشتن عاملیت اخلاقی است. فناوری های هوش مصنوعی به سبب فقدان آگاهی، فاقد اراده و عاملیت اخلاقی هستند.

■ **هوش مصنوعی (فاقد آگاهی) عاملیت اخلاقی ندارد: ظرفیت انتخاب آزادانه شرطی است که هیما برای عاملیت اخلاقی معرفی می کند.** هیما عاملیت اخلاقی را ملازم با توضیح دهنده گی اخلاقی می داند. هیما نیز در توضیح مصادیق عامل اخلاقی، به حیوانات اشاره می کند. هیما سگ ها را دارای عاملیت، اما فاقد عاملیت اخلاقی می داند زیرا آنها موضوع حکمرانی اخلاقی نبوده و به لحاظ اخلاقی توضیح دهنده اعمالشان نیستند. نکته اینجاست که از دید هیما، توضیح دهنده گی ارتباط وثیقی با مجازات پذیری دارد. او معتقد است زمانی که کسی را توضیح دهنده می دانیم که به رفتار او براساس آنچه شایسته آن است واکنش نشان دهیم؛ یعنی در صورت نقض قواعد اخلاقی با سرزنش و در صورت رعایت قواعد اخلاقی با تحسین آن مواجه شویم [۲۳].

هیما ظرفیت انتخاب آزادانه اعمال خود، و ظرفیت تشخیص تفاوت میان درست و غلط را به عنوان شروط لازم و مجموعاً کافی برای عاملیت اخلاقی معرفی می کند. منظور او از انتخاب آزادانه، انتخابی است که مبتنی بر تأمل بوده و خروجی آن توسط چیزی بیرون از عامل متعین نشود. منظور این نیست که امری خارجی روی عمل عامل اخلاقی اثر نگذارند، بلکه این است که عاملی خارجی، مستقیماً عمل را تحمیل نکرده باشد. این نوع آزادی مستلزم نوعی ظرفیت برای تأمل و تصمیم گیری و به طور کلی داشتن عقلانیت است. دومین ظرفیتی که هیما برای عاملیت اخلاقی لازم می شمرد نیز با عقلانیت مرتبط است. از دید هیما شناختن درست از غلط یا آنچه اغلب به عنوان شایستگی اخلاقی نامیده می شود نیز برای عاملیت اخلاقی ضرورت دارد. در اینجا نیز منظور این نیست که فرد باید به لحاظ اخلاقی خطاناپذیر باشد، زیرا هیچ کس چنین نیست. هیما منظور خود را از این معرفت، چیزی ضعیف تر از باور صادق موجه در نظر گرفته و آن را صرفاً وجود نوعی روش شناسی عمدتاً اعتمادپذیر برای بازشناسی درست از غلط می داند. یک عامل اخلاقی باید بتواند در استدلال اخلاقی درگیر شده و براساس آن عمل کند. برای درگیر شدن در استدلال اخلاقی، فهمی حداقلی از مفاهیم اخلاقی همچون خوب، بد، الزام آور، غلط، و مجاز، لازم است. همچنین، داشتن دریافتی از اصول اخلاقی پایه ای -مانند اصل پرهیز از آسیب- ضرورت دارد.

بر این اساس، هیما نیز مشابه مانا و ناث اما با استدلالی دیگر، «آگاهی» را به عنوان یک شرط ضمنی و ضروری برای عاملیت اخلاقی معرفی می کند. از یک سو انتخاب گری تنها در مورد موجودات آگاه قابل طرح بوده؛ زیرا رفتار موجودات غیر آگاه یا محصول تصادف بوده یا کاملاً متعین شده و در هر حال انتخاب در تعیین آن جایی ندارد. از طرف دیگر، فهم اخلاقی یا ظرفیت تشخیص خوب و بد نیز وابسته به وجود حالات ذهنی آگاهانه است. استدلال اخلاقی، فرایندی بوده که در آن دلایل فهمیده می شوند و چنین فرایندی لاجرم آگاهانه است.

■ **هوش مصنوعی سطحی از عاملیت اخلاقی را دارد:** گروهی دیگر از فیلسوفان با همدلی نسبی با دلایل هر ۲ گروه در تأیید یا رد عاملیت اخلاقی مصنوعات، راه دیگری در پیش گرفته اند و به جای آنکه درباره مفهومی یکتا از عاملیت اخلاقی بحث کنند، توصیفی ریزبافت تر از عاملیت اخلاقی ارائه داده که شامل چندین سطح از عاملیت می شود و برای هر یک از این سطوح شروطی را برشمرده اند. برای مثال مور ۴ سطح از عاملیت اخلاقی را از هم متمایز می کند: عامل اخلاقی اثرگذار،^۱ عامل اخلاقی ضمنی،^۲

1. Ethical Impact Agent
2. Implicit Ethical Agent



عامل اخلاقی آشکار،^۱ و عامل اخلاقی کامل.^۲ مطابق دسته‌بندی مور، هر ماشینی که براساس پیامدهای اخلاقی عملکردش مورد ارزیابی قرار بگیرد، دارای عاملیت اخلاقی اثرگذار است. مور به‌عنوان نمونه، به ربات‌هایی اشاره می‌کند که در قطر برای شتررانی استفاده می‌شوند. این ربات‌ها جایگزین پسران جوانی هستند که سابقاً شترسواری می‌کردند. برای اینکه شتررانی به‌خوبی انجام شود، کم‌وزن بودن این پسران اهمیت داشت و به همین علت صاحبان شترها پسران جوان شترران را در سوءتغذیه نگاه می‌داشتند. جایگزینی شترران رباتی با انسانی، عملاً نیاز به انجام این عمل غیراخلاقی را رفع کرده است [۲۴].

در سطح بالاتر، ماشین‌هایی هستند که به‌گونه‌ای طراحی شده‌اند که فاقد آثار اخلاقی منفی باشند، یعنی ایمنی و اعتمادپذیری در فرایند طراحی آنها لحاظ شده است. این ماشین‌ها عاملیت اخلاقی ضمنی دارند. برای مثال مور به هواپیماهای خودران اشاره می‌کند که نسبت به ایمنی و به‌موقع انجام شدن سفر، حساسیت بیشتری نسبت به خلبان انسانی دارند و می‌توانند با عملکرد خود فایده بیشتری به مسافران برسانند.

سطح بعدی عاملیت، به ماشین‌هایی اختصاص دارد که برنامه‌ریزی آنها به‌نحوی است که با استفاده از مقولات اخلاقی می‌توانند استدلال کنند و دارای عاملیت اخلاقی آشکار هستند. چنین ماشین‌هایی به ۳ نوع منطق مجهز هستند. منطق وظیفه‌شناسانه که روی گزاره‌هایی درباره الزام و مجوزهای اخلاقی اعمال می‌شود. منطق معرفتی که روی گزاره‌هایی درباره باورها اعمال می‌شود، و منطق عملی که روی گزاره‌هایی درباره کنش‌ها به‌کار بسته می‌شود. مدل‌های مختلفی برای پیاده‌سازی این منطق‌ها روی ماشین‌ها ارائه شده که هدف همگی آنها امکان پیاده‌سازی قابلیت استدلال اخلاقی، و عمل براساس آن است.

در بالاترین سطح عاملیت اخلاقی، عاملان کامل هستند؛ یعنی عاملانی که داورهای اخلاقی صریحی انجام می‌دهند و صلاحیت توجیه تصمیمات اخلاقی‌شان را دارند. این سطح از عاملیت مستلزم ظرفیت آگاهی، قصدمندی و اراده آزاد است. ماشین‌ها، ۲ مرتبه اول عاملیت اخلاقی را به‌سادگی می‌توانند احراز کنند، اما احراز مرحله سوم با چالش‌های عملی بسیاری همراه است. درحالی‌که نیل به سطح چهارم و عامل اخلاقی کامل بودن، برای هیچ ماشینی میسر نیست.

■ **عاملیت اخلاقی مصنوعات: مسئله این نیست!** در رویکردی دیگر نسبت به عاملیت اخلاقی مصنوعات، پیشنهاد می‌شود معضل نظری تعریف شرایط عاملیت به طریقی دور زده شود و به‌جای تمرکز بر این مسئله، رویکردی هنجاری درباره عاملیت مصنوعات در پیش گرفته شود. برخی فیلسوفان با تحلیل محل نزاع فیلسوفان در مسئله مسئولیت و عاملیت اخلاقی هوش مصنوعی، ریشه این اختلاف را در زاویه دید نسبت به عاملیت اخلاقی می‌یابند. اگر عاملیت اخلاقی را از زاویه سوم‌شخص ببینیم، هویتی که برایشان قائل به عاملیت اخلاقی هستیم طیفی را تشکیل می‌دهند که در یک سر طیف انسان‌های عاقل و بالغ و در سر دیگر طیف کودکان و حیوانات هستند. در این صورت اگر ویژگی‌هایی که در این عاملان می‌یابیم در امور مصنوعی نیز دیده شوند، نمی‌توان از انتساب عاملیت اخلاقی به آنها اجتناب کرد. اما اگر مسئله را از زاویه اول‌شخص ببینیم، هر فردی خود را به‌عنوان عامل اخلاقی در نظر گرفته و با نظر کردن در احوالات خود ویژگی‌های عامل اخلاقی را توصیف می‌کند. از آنجاکه این ویژگی‌ها همچون آگاهی، وجدان، و ... مختص امور زنده هستند و در جهان بی‌جان تحقق نمی‌یابند، نمی‌توان به مصنوعات عاملیت اخلاقی نسبت داد. به این ترتیب، مناقشه متوقف بر نتیجه مناقشه میان دو رویکرد اول‌شخص و سوم‌شخص نسبت به عاملیت اخلاقی خواهد بود. بهدادی و مونته با اتخاذ چنین رویکردی، از نوعی عاملیتی شراکتی بین مصنوعات و انسان‌ها دفاع می‌کنند، عاملیت و شراکتی که شکل‌پذیر است. براساس این رویکرد، به جای پرسش از امکان‌پذیری عاملیت مصنوعات، بر باید‌ها و نباید‌های عاملیت شراکتی انسان-مصنوع تمرکز می‌شود. در این رویکرد ملاحظات اخلاقی در ترکیب‌بندی شراکت میان انسان و مصنوع، واگذاری وظایف به مصنوعات دارای هوش مصنوعی براساس توانمندی‌ها و ایجاد قابلیت‌های اخلاقی در آنها مد نظر قرار می‌گیرد [۲۵].

در مجموع، انتساب یا عدم انتساب مسئولیت اخلاقی به هوش مصنوعی، با مناقشات دیگری در حوزه اخلاق و فلسفه گره خورده

1. Explicit Ethical Agent
2. Full Ethical Agent

است. تعریف اندیشمندان از مسئولیت‌پذیری و پیامدهای آن، نسبت مسئولیت‌پذیری با عاملیت اخلاقی و دیدگاه آنها درباره ویژگی‌های موجود یا قابل حصول در فناوری‌های هوش مصنوعی مهم‌ترین ابعاد این مسئله هستند. به همین علت، نمی‌توان بدون اتخاذ موضعی مستدل در هر یک از مسائل فوق، درباره مسئولیت‌پذیری هوش مصنوعی، به نتیجه رسید.

مسئله عاملیت اخلاقی هوش مصنوعی در جوامع دنباله‌رو در توسعه این فناوری مانند کشور ما اهمیت مضاعفی دارد. زیرا از سویی پدیده بیش‌اعتمادی به سامانه‌های مبتنی بر فناوری هوش مصنوعی که ناشی از فقدان شناخت کافی از این فناوری و وارداتی بودن صرف آن به کشور است، به چالش‌های متعددی منجر خواهد شد. از سوی دیگر اتکا به سامانه‌های پیش‌رو در این فناوری که تقریباً تمام آنها در خارج از کشور قرار دارند، به تسری ارزش‌هایی منجر خواهد شد که ممکن است با ارزش‌های اساسی جامعه ما تفاوت فاحشی داشته باشد. این تفاوت در ارزش‌ها خود به چالش‌های جدیدی در نتیجه توسعه و فراگیری این فناوری منجر خواهد شد. بنابراین پرداخت بنیادین به این مسئله برای کشور اهمیت دارد.

در جدول زیر، خلاصه‌ای از دیدگاه‌های موجود درباره مسئولیت‌پذیری و عاملیت اخلاقی هوش مصنوعی ارائه شده است. در این جدول، تنوع آرا در ابعاد مختلف این مسئله را می‌توان دید.

جدول ۵. دیدگاه‌های مختلف درباره عاملیت و مسئولیت‌پذیری هوش مصنوعی

نماینده دیدگاه	شرایط مسئولیت‌پذیری اخلاقی	پیامدهای مسئولیت‌پذیری اخلاقی	شرایط عاملیت اخلاقی	امکان مسئولیت اخلاقی در هوش مصنوعی	امکان عاملیت اخلاقی در هوش مصنوعی
فلوریدی و ساندرس	توضیح‌دهندگی، قصدمندی	سرزنش‌پذیری	تعاملی بودن، خودمختاری، انطباق‌پذیری	-	✓
فلوریدی	توضیح‌دهندگی، خودمختاری	ضمانت حقوقی، پاسخگویی اخلاقی	آموختن از عمل، خودمختاری	✓	✓
سالینز	معقول بودن انتساب «باور به مسئولیت داشتن» به هویت مورد نظر	مجازات/پاداش‌پذیری	خودمختاری، قصدمندی، مسئولیت‌پذیری	✓	✓
ولیز	قابلیت فهم اخلاقی، حس‌مندی، آگاهی	توضیح‌دهندگی، مجازات/پاداش‌پذیری	خودمختاری، مسئولیت‌پذیری اخلاقی	-	-
مانا و ناث	-	-	اختیار، احساس وظیفه، آگاهی	-	-
هیما	توضیح‌دهندگی	توضیح‌دهندگی	انتخاب آزادانه، تشخیص درست و غلط اخلاقی، آگاهی	؟	؟
مور	-	-	سطح ۱: اثر اخلاقی سطح ۲: ملاحظات اخلاقی در طراحی سطح ۳: قدرت استدلال اخلاقی سطح ۴: آگاهی، قصدمندی، اراده آزاد	؟	✓ (سطوح ۱ تا ۳)

مأخذ: همان.



۶. راهبردهایی برای توسعه مسئولانه هوش مصنوعی و رهنمودهایی برای کشور

پیچیدگی موضع‌گیری درباره مسئله عاملیت و مسئولیت اخلاقی هوش مصنوعی، گروهی از فیلسوفان را به سمت رویکردی دیگر نسبت به این مسئله سوق داده است؛ رویکردی که در آن پیشگیری از شکل‌گیری شکاف مسئولیت اهمیت دارد. در رویکرد اخیر نسبت به مسئله شکاف مسئولیت‌پذیری، دیسیو و مکاچی ۴ نوع شکاف مسئولیت‌پذیری را از یکدیگر متمایز می‌کنند. آنها با این تفکیک، می‌خواهند نشان دهند چه معنایی از مسئولیت‌پذیری، توسط چه جنبه‌ای از خودکارسازی تهدید می‌شود، و چرا اهمیت دارد. این ۴ حالت، کمابیش در تناظر با تعابیری از مسئولیت‌پذیری است که در قسمت پیش مختلف بحث و بررسی شد که این شکاف‌ها عبارت‌اند از شکاف مجازات‌پذیری، توضیح‌دهندگی اخلاقی، توضیح‌دهندگی عمومی و مسئولیت‌پذیری فعال [۲۵]. آنها پس از شرح این ۴ حالت، از مفهومی با عنوان «مسئولیت معنادار انسانی» دفاع می‌کنند. مسئولیت معنادار انسانی مفهومی است که به تازگی در مباحث مربوط به مسئولیت‌پذیری مطرح شده و اشاره به الزاماتی تجویزی دارد که صورتی از کنترل انسانی را که به لحاظ قانونی، اخلاقی و اجتماعی مقبول باشد، ارتقا می‌بخشد. دیسیو و مکاچی، کنترل معنادار را شامل ۲ شرط «ردیابی»^۱ و «پیگردی»^۲ می‌دانند.

در ادامه این بخش با تکیه بر پیشنهاد دیسیو و مکاچی و برخی دیگر از فیلسوفان فناوری، راهبردهایی چهارگانه با عنوان‌های «شفافیت ساختارهای اجتماعی-تکنیکی»، «تشکیل کارگروه ارزیابی اخلاقی ناظر بر توسعه»، «آگاهی‌بخشی درباره ابعاد عمومی مسئولیت‌پذیری» و «خط‌مشی‌گذاری مشارکت‌محور» به‌عنوان راهبردهایی مبنایی برای پیشگیری از شکاف مسئولیت‌پذیری مطرح شده و نتایج و فواید هر یک شرح داده می‌شود. این راهبردها مبتنی بر تعبیری «رو به جلو» از مسئولیت‌پذیری هستند. در این تعبیر، به جای تکیه بر یافتن مقصر در مواقعی که عملی غیراخلاقی رخ داده، تأکید بر وظایف عاملان اخلاقی برای پیشگیری از وقوع عملی غیراخلاقی است.

در ارتباط با این راهبردهای پیشنهادی، این نکته شایسته توجه است که شکاف مسئولیت‌پذیری بسته به اینکه یک سامانه فناورانه در چه سطحی از توسعه‌یافتگی قرار داشته باشد و در آن چه سنخ وظایفی به هوش مصنوعی محول شود، می‌تواند در مورد وضعیت‌های متفاوتی مطرح شده و راهکارهای متفاوتی را نیز بطلبد. در کشورهایی که در حال حاضر در خط مقدم توسعه هوش مصنوعی هستند، پیشگیری از شکاف مسئولیت در مرحله خط‌مشی‌گذاری‌های کلان و ترسیم مسیر توسعه جهانی فناوری نیز معنا پیدا می‌کند، درحالی‌که در کشورهایی که مصرف‌کننده هوش مصنوعی هستند، جلوگیری از وقوع شکاف مسئولیت عمدتاً به سطح اقدام‌های ملی و تنظیم مقررات مربوط به کسب‌وکارهای مبتنی بر هوش مصنوعی و نهادهای ناظر بر آنها بازمی‌گردد. همچنین ابعاد عمومی مسئولیت هوش مصنوعی، بسته به میزان دسترسی آحاد جامعه به این فناوری‌ها، نگرش جامعه نسبت به این فناوری‌ها و میزان استفاده از آنها، در جوامع مختلف می‌تواند ضریب‌های متفاوتی داشته باشد. این نکته، توجه به ابعاد بومی مسئله شکاف مسئولیت‌پذیری را برجسته می‌کند. در ادامه هر یک از راهکارهای ذیل با توجه به وضعیت شاخص هوش مصنوعی در ایران، مورد توجه قرار خواهند گرفت.

۶-۱. شفافیت ساختارهای اجتماعی-تکنیکی

یکی از عواملی که موجب ایجاد شکاف مسئولیت می‌شود، شفاف نبودن عوامل، نهادها و ساختارهای درگیر در یک شبکه اجتماعی-تکنیکی، و نقش هر یک از آنها در تولید خروجی نهایی است. دیسیو و مکاچی برای حل معضل شکاف مسئولیت‌پذیری، بر اهمیت

1. Tracing
2. Tracking

شفافیت نهادی به‌عنوان یک اصل کلیدی اشاره می‌کنند. این به آن معناست که نه‌تنها جنبه‌های تکنیکی یک ساختار باید شفاف بوده؛ بلکه برای حفظ مسئولیت انسانی، نهادهای حقوقی و قوانین نیز باید شفاف باشند. اگر چیدمان‌های نهادی و قانونی به‌صورتی باشند که نقش هر نهاد و ارزش‌های آن در مسیر توسعه مشخص باشد، ابهام در روابط بین بخشی و نقش علی و حقوقی نهادها از بین می‌رود و مسئولیت قابل پیگیری خواهد بود. برای این منظور لازم است نقشه‌ای از تمام عاملان مرتبط در طراحی، کنترل، قانونگذاری و استفاده از یک سیستم فراهم شود و نسبت این عاملان با دلایل، ارزش‌ها و مقاصد مورد نظر تحلیل شود. در این صورت خروجی‌های سیستم تبیین‌پذیر خواهند بود. یعنی می‌توان با ارجاع به چنین نقشه‌ای، توضیح داد که چه فرایندی منتهی به خروجی فعلی شده است. نکته‌ای که کاکلبرگ نیز بر آن به‌عنوان مؤلفه ضروری مسئولیت تأکید می‌کند [۲۵].

● **مثال:** فرض کنیم قرار است یک فناوری هوش مصنوعی به نام «به‌مان» در حوزه تشخیص پزشکی و درمان توسعه پیدا کند. برنامه‌ای که روی گوشی‌های همراه نصب می‌شود و با گرفتن شرح حال از بیمار و پردازش داده‌های صوتی و تصویری بیماری او را تشخیص داده و روند درمان را به او پیشنهاد کند. نهادهای گوناگونی درگیر توسعه این فناوری خواهند بود. نهادهایی همچون دانشگاه‌ها و پژوهشگاه‌هایی که داده‌های علمی پزشکی و الگوهای تشخیص و درمان مورد نیاز را فراهم کرده، شرکت‌های طراحی و برنامه نویسی و تولیدکنندگان گوشی‌های همراه که زیرساخت تکنیکی این فناوری را تولید کرده، مراکز درمانی که داده‌های بیماران را برای یادگیری هوش مصنوعی در اختیار آن قرار می‌دهند و از بیماران بازخورد می‌گیرند، شرکت‌های بیمه و داروسازی، نهادهای دولتی که قوانین امنیتی و اقتصادی تولید این فناوری را تعیین و بر اجرای آن نظارت می‌کنند و هر یک از این نهادها ارزش‌هایی خاص (مانند پیشینه کردن سود اقتصادی، بالا بردن سطح سلامت مردم، حفاظت از محرمانگی داده‌ها و ...) دارند و بر این اساس در این ساختار عمل می‌کنند. در صورتی که حادثه‌ای ناخوشایند در این سیستم رخ دهد -برای مثال اطلاعات بیماران بدون رضایت آنها در اختیار شرکت‌های دارویی یا غذایی قرار گیرد- با ارجاع به نقشه‌ای شفاف از ارتباط میان نهادها می‌توان نقش گروه‌های مختلف را در این واقعه ردیابی کرد. به این ترتیب این راهبرد در عین حال که امکان مطالبه و بازخواست را فراهم می‌کند، در تقویت مسئولیت‌پذیری فعال نیز مؤثر است و موجب توجه هر یک از نهادها به اهمیت کار خود و پیامدهای ناخواسته احتمالی می‌شود.

ایجاد شفافیت در ساختارهای اجتماعی-تکنیکی در مورد زیست‌بوم فناورانه ایران اهمیت بسزایی دارد. از یک‌سو ایران به‌عنوان کشوری که خواهان پیشتازی در حوزه فناوری‌های دیجیتال و هوش مصنوعی بوده، با ضعف‌هایی جدی در حوزه حکمرانی داده مواجه است که این مسئله هم شامل ضعف‌هایی در زیرساخت‌های فناورانه و هم مشکلاتی در حکمرانی و مدیریت فناوری می‌شود. در غیاب یک پایگاه داده ملی و دسترس‌پذیر، نهادهای مختلفی که در پی هوشمندسازی فعالیت‌های خود هستند، از سامانه‌های جداگانه‌ای استفاده می‌کنند و این، به نوعی انزوا و ساختار جزیره‌ای در این نهادها منتهی شده است. از سوی دیگر، مسئولیت حکمرانی بر سامانه‌های داده‌محور نیز میان نهادهای مختلف حاکمیتی تقسیم شده و عملاً تنظیم فعالیت سازمان‌های مختلف به نحو یکپارچه و مدون پیش نمی‌رود. برای مثال، تعریف سازمان ملی هوش مصنوعی بدون تدوین اساسنامه برای آن و تعلیق فعالیت‌های آن پس از گذشت حدود یک‌سال از تأسیس آن و محول شدن مسئولیت‌های این نهاد به نهادهای دیگر، نمونه‌ای از ضعف مدیریتی در حوزه حکمرانی دیجیتال است. درحالی‌که شفافیت ساختارهای اجتماعی-تکنیکی مستلزم وجود نقشه‌ای شفاف از ارتباط نهادهای مختلف با یکدیگر و دامنه عملکرد هر یک و خصوصاً شیوه تبادل داده میان ساختارهای مختلف است.

البته ایران در حال حاضر در ابتدای مسیر بهره‌مندی از فناوری‌های هوش مصنوعی بوده و در مقایسه با کشورهای منطقه و جهان، سهم هوش مصنوعی در بازار فناورانه آن اندک است. مطابق گزارش شاخص هوش مصنوعی ایران در سال ۲۰۲۴، در مورد میزان پذیرش هوش مصنوعی در شرکت‌های ایرانی، ۶۲٫۱ درصد از شرکت‌های ایرانی هنوز از هیچ فناوری هوش مصنوعی استفاده نکرده‌اند؛ درحالی‌که در سطح جهانی، ۷۲ درصد از سازمان‌های مورد بررسی دست کم در یکی از واحدهای خود از این فناوری‌ها



استفاده می‌کنند. مطابق آمارهای به دست آمده، نرخ پذیرش هوش مصنوعی در ایران در سال ۲۰۲۴ با سطح جهانی حدود هفت‌سال فاصله دارد و مطابق آمار سال ۲۰۲۳ ایران حتی با کشورهای در حال توسعه نیز فاصله قابل توجهی دارد [۲۶]. به همین علت هنوز برای مهیا ساختن زیرساخت‌های لازم و تمهیدات مدیریتی برای تحقق شفافیت ساختارهای اجتماعی-تکنیکی فرصت وجود دارد. این مسئولیت باید به صورت مستقیم برعهده سازمان ملی هوش مصنوعی قرار گیرد.

۲-۶. تعریف کارگروه ارزیابی اخلاقی در سیستم

شفافیت نهادی با وجود اهمیتی که دارد، برای حذف شکاف مسئولیت کافی نیست. زیرا با وجود مشخص بودن نقش هر نهاد، همچنان این امکان وجود دارد که در غیاب نگاه کلان اخلاقی به سیستم، موضوعاتی اخلاقی در توسعه سیستم از نظر پنهان بماند. در یک سیستم پیچیده، اینکه هر نهاد به خودی خود مسئولانه رفتار کند، برای اینکه خروجی هم اخلاقی باشد کفایت نمی‌کند؛ زیرا گاهی خروجی تمام عاملان در گستره مسئولانه بوده اما خروجی نهایی سیستم غیراخلاقی است.

● **مثال:** در مورد فناوری فرضی «به‌مان» بیماری‌ای را تصور کنید که در آن چگونگی ظهور علائم وابسته به ویژگی‌های قومیتی باشد، چنان‌که بسیاری از بیماری‌ها این گونه‌اند. حال در مورد این بیماری خاص، وابستگی علائم به ویژگی‌های قومیتی هنوز در پژوهش‌های علمی نیز کشف نشده است. درحالی‌که افراد آن قومیت به دلایلی همچون فناوری‌های هراسی، یا عدم دسترسی به گوشی‌های پیشرفته از چنین برنامه‌ای استفاده نمی‌کنند و لذا فناوری برای تشخیص بیماری در افراد آن قومیت به اندازه کافی داده دریافت نکرده و آموزش ندیده است. در نتیجه هوش مصنوعی در تشخیص بیماری آن افراد از این قومیت که از برنامه استفاده می‌کنند نیز به خطا می‌رود. در اینجا هیچ یک از نهادها در انجام وظیفه خود کوتاهی نکرده‌اند اما در کل خروجی سیستم با مشکل اخلاقی مواجه است.

برای پیشگیری از چنین اشکالاتی لازم است نگاهی کل‌نگر و آشنا به جنبه‌های اخلاقی در سیستم حضور داشته باشد و ضمن نظارت بر فرایند توسعه، دارای قدرت اعمال نظر نیز باشد. در غیاب چنین منظری، محتمل است حالات بسیاری واقع شوند که در آنها هیچ‌کس مسئولیت پاسخگویی نسبت به وضع موجود را نپذیرد. زیرا هیچ کس به اطلاعات لازم برای درک اتفاقی که در حال وقوع بوده و قدرت لازم برای پیشگیری از آن مجهز نبوده است.

برای حفظ امکان پاسخگویی، پیشنهاد دیسیو و مکاچی تعریف یک جایگاه حقیقی یا حقوقی برای پاسخگویی اخلاقی است. چنین جایگاهی را می‌توانیم «کارگروه ارزیابی اخلاقی در سیستم» بنامیم. کارگروهی که در هر سیستم در زنجیره طراحی و توسعه و کاربری نقش ایفا می‌کند، می‌تواند ابعاد اخلاقی سیستم را درک کند و قدرت کنترل و اعمال نظر نسبت به خروجی‌ها را -براساس ارزیابی اخلاقی خود- دارد. به این ترتیب این کارگروه در قبال خروجی سیستم نسبت به عموم جامعه و قانون، پاسخگو خواهد بود. هدف این است که با تعریف این جایگاه توزیعی منصفانه از مجازات‌پذیری اخلاقی داشته باشیم که مانع از وقوع ۲ نوع خطا می‌شود: اینکه کسی بدون اینکه امکان اجتناب از خطا داشته باشد مجازات شود، و اینکه کسی بابت خطاهای اجتناب‌ناپذیر مجازات نشود. تعریف چنین جایگاهی می‌تواند شکاف پاسخگویی اخلاقی، پاسخگویی عمومی و مجازات‌پذیری را کاهش دهد [۲۷].

در مورد زیست‌بوم فناوریانه ایرانی، کارگروه ارزیابی اخلاقی در سیستم می‌تواند در ۲ سطح حاکمیتی و سازمانی اجرا شود. در سطح حاکمیتی، نقش کارگروه ارزیابی اخلاقی حفظ موقعیت پاسخگویی و مداخله اخلاقی برای پیشگیری از وقوع شکاف مسئولیت در سطح نهادهای اجرایی و مدیریتی کلان است. با توجه به آنچه در قسمت قبل ذکر شد، ضعف در حکمرانی داده و مدیریت یکپارچه سامانه‌های اطلاعاتی نهادهای اجرایی و حاکمیتی مختلف، نقاط آسیب‌پذیری برای ایجاد معضل شکاف مسئولیت در سطح دولتی ایجاد می‌کند. در این سطح، کارگروه ارزیابی اخلاقی مسئولیت رصد کردن عملکرد نهادهای مختلف و پیش‌بینی مشکلات احتمالی

در تعاملات بین این نهادها و همچنین امکان مداخله برای رفع این مشکلات را دارد.

علاوه بر سطح حاکمیتی، در سطح شرکت‌ها و سازمان‌های کوچک‌تر و خصوصی و کسب‌وکارهای فناورانه نیز تعریف کارگروه ارزیابی اخلاقی اهمیت دارد. در این سطح، کارشناسانی که نسبت به مسائل اخلاقی احتمالی ناشی از کاربست هوش مصنوعی تسلط داشته و همچنین به خوبی از شیوه ارتباط و نحوه تقسیم کار زیربخش‌های مختلف سازمان مطلع هستند و قدرت مداخله و تصمیم‌سازی دارند، از وقوع خروجی‌های غیراخلاقی در سازمان جلوگیری کرده و در صورت وقوع چنین وضعیت‌هایی، پاسخگو خواهند بود. به این ترتیب از شکل‌گیری معضل شکاف مسئولیت در سازمان پیشگیری می‌شود. نکته قابل توجه این است که پیاده‌سازی چنین کارگروه‌هایی در شرکت‌های خصوصی و کسب‌وکارها مستلزم تعریف مشوق‌هایی برای این شرکت‌هاست. در نظر گرفتن تشکیل کارگروه ارزیابی اخلاقی به عنوان یکی از مؤلفه‌های مسئولیت اجتماعی سازمان‌ها^۱ و همچنین ایجاد مشوق‌هایی همچون معافیت‌های مالیاتی و ... می‌تواند در ترغیب سازمان‌ها به تشکیل این کارگروه مؤثر باشد.

در این مورد نیز همانند مورد شفافیت، به علت اینکه اقدام‌های ملی در زیست‌بوم فناورانه ایرانی هنوز در مراحل ابتدایی خود قرار دارد، امکان توجه به این مسئله و لحاظ کردن آن همچنان وجود دارد. اقدام‌های ملی، شامل حاکمیت، رهنمودها و مقررات، حمایت مالی و توانمندسازها و مشوق‌های هوش مصنوعی می‌شوند. ایران در میان کشورهای منطقه با ۵ اقدام ملی در مقایسه با سایر کشورها اقدام‌های ملی کمتری انجام داده است. همچنین، ترکیه، امارات و رژیم صهیونیستی به عنوان کشورهای پیشرو، در اقدام‌های ملی خود بر حوزه‌های «توانمندسازها و مشوق‌های هوش مصنوعی» یا «حکمرانی» تمرکز داشته‌اند؛ در حالی که ایران و عربستان در درجه اول بر «رهنمودها و مقررات» متمرکز بوده و بیشتر از طریق تنظیم مقررات به دنبال کنترل توسعه هوش مصنوعی بوده‌اند [۲۶]. با توجه به روند رو به رشد توسعه هوش مصنوعی در ساحت کسب‌وکارهای فناورانه و لزوم اجرای تمهیدات لازم برای حفظ توسعه مسئولانه، توجه به ابعادی چون توانمندسازی و تعریف مشوق‌ها در کنار تنظیم مقررات در اقدام‌های ملی بعدی، به طور خاص در مورد زیست‌بوم فناورانه ایران واجد اهمیت است.

کارگروه پیشنهادی باید ذیل سازمان ملی هوش مصنوعی در کشور ایجاد شده و به فعالیت بپردازد.

۳-۶. آگاهی بخشی درباره جنبه‌های عمومی مسئولیت‌پذیری

جنبه دیگری که در ارتباط با توسعه هوش مصنوعی حائز اهمیت است، نقش کاربران و محیط در شکل‌دهی خروجی سیستم‌های فناورانه هستند. سیستم‌های هوش مصنوعی برخلاف نرم‌افزارهای پیشین که صرفاً براساس برنامه‌ریزی‌شان کار می‌کنند، از محیط می‌آموزند. بازخوردهای محیط و کنش کاربران انسانی، هوش مصنوعی را آموزش می‌دهند و خروجی رفتار آن را تعیین می‌کنند. این در حالی است که غالب مردم تصویری از نقش علی و مسئولیت خود در قبال خروجی این سیستم‌ها ندارند و صرفاً طراحان و تولیدکنندگان را مسئول می‌دانند.

کاکلبرگ نیز بر این نکته تأکید می‌کند. از نظر او زمانی که درباره رابطه بین توسعه‌دهندگان فناوری و کاربران آن فکر می‌کنیم، نباید کاربران را صرفاً عناصری منفعل در نظر بگیریم. اعطای نقش و عاملیت به جامعه کاربران، از چند جهت اهمیت دارد. اول اینکه با به رسمیت شناختن نقش عاملیت کاربران، پاسخگویی به آنها نیز اهمیت پیدا می‌کند. توجه به ضرورت پاسخگویی، خود به تلاش برای حفظ تبیین‌پذیری و شفافیت مسیر توسعه فناوری منجر می‌شود. این به آن معناست که کاربران در تعیین مصادیق تبیین‌خواه و کم و کیف چنین تبیین‌هایی نقش ایفا می‌کنند. بنابراین برای آنکه پاسخگویی و تبیین‌پذیری محقق شود، لازم است عموم جامعه



دارای درک مناسبی از مسائل مرتبط با فناوری‌های هوش مصنوعی، و تأثیرات احتمالی آن بر زندگی و منافع‌شان داشته باشند. در غیاب آگاهی و احساس نیاز عمومی، بخش مهمی از انگیزه شفاف‌سازی فرایندهای توسعه از دست خواهد رفت [۲۸]. اهمیت تعامل با مخاطبان تنها به مسئله تبیین‌پذیری محدود نمی‌شود. در کارکرد فناوری‌های هوش مصنوعی که قابلیت یادگیری دارند، عملاً بخشی از توسعه فناوری از طریق تعامل با محیط (یعنی کاربران) اتفاق می‌افتد. کاربران به‌عنوان کسانی که داده‌های ورودی سیستم‌های هوش مصنوعی را فراهم می‌کنند، در چگونگی شکل‌گیری و رفتار این سیستم‌ها نقش ایفا می‌کنند. در واقع می‌توان گفت توسعه تکنولوژی و عمل فناورانه، عملی جمعی است که مسئولیتی جمعی در پی دارد. عملی که در آن نه تنها مجموعه‌ای از عناصر رسمی مربوط به توسعه تکنولوژی نقش دارند؛ بلکه عموم جامعه و مختصات فکری و فرهنگی جامعه نیز در آن نقش ایفا می‌کنند. ممکن است یک فناوری هوش مصنوعی دارای سوگیری ارزشی بوده که منعکس‌کننده سوگیری نهادینه شده در کل جامعه باشد و ارزش‌هایی را که در فرهنگ جامعه رسوخ کرده تحقق بخشد.

● **مثال:** به مورد فرضی «به‌مان» برگردیم. فرض کنید گروهی از کاربران این برنامه بدون قصد واقعی برای دریافت تشخیص و درمان پزشکی و صرفاً برای سرگرمی، اطلاعاتی غلط وارد برنامه کنند. مجموعه‌ای از علائم غیرواقعی را به سیستم داده تا عملکرد سیستم را امتحان کرده، یا حتی برای تفریح بر سر اینکه سیستم چه تشخیصی می‌دهد با دوستان خود شرط‌بندی کنند. ورود این اطلاعات غلط از طرفی می‌تواند آمار غلطی درباره همبستگی علائم به هوش مصنوعی بدهد. و از طرف دیگر می‌تواند سوگیری‌های غیراخلاقی وارد این سیستم کند. اگر در یک جامعه کاربران محض سرگرمی اطلاعات مربوط به یک نژاد، یک جنسیت یا یک قومیت را با علائم بیماری‌های بدنام (همچون بیماری‌های مقاربتی) وارد سیستم کنند، به مرور سیستم را نسبت به آن جنسیت یا نژاد دچار سوگیری می‌کند.

در غیاب آگاهی‌بخشی عمومی، با اینکه کاربران در توسعه فناوری نقش ایفا می‌کنند اما به‌علت فقدان مؤلفه آگاهی نمی‌توان آنها را مسئول دانست و اعمال‌شان را در جهت اخلاقی بودن تنظیم کرد. اما آگاهی‌بخشی عمومی با تقویت مؤلفه آگاهی میان کاربران باعث می‌شود آنان نیز نسبت به نقش خود در این فرایند احساس مسئولیت کنند.

درخصوص وضعیت کشور، هر چند به علت پایین بودن دسترسی به اینترنت و موانع دیگر، کاربران ایرانی در مقایسه با کاربران جهانی دسترسی محدودتری به فناوری‌های هوش مصنوعی دارند، اما مطابق نظرسنجی‌ها، رویکرد عمده کاربران ایرانی نسبت به فناوری‌های هوش مصنوعی مثبت است (۴۶ درصد از کاربران احساس مثبت، ۳۴ درصد احساس منفی، و ۲۰ درصد احساس خنثی دارند). احساسات مثبت کاربران شامل تحسین و قدردانی از ویژگی‌های مثبت این فناوری‌ها، اشتیاق به توسعه بیشتر آنها، در نظر گرفتن آنها به‌عنوان ابزارهایی سودمند و همچنین اعتماد و اتکای بالا به هوش مصنوعی می‌شود. در عین حال، در مقایسه با تحسین هوش مصنوعی که در ۲۰ درصد دیدگاه‌ها منعکس شده، ۱۵ درصد نگرانی و ترس از توسعه هوش مصنوعی، در ۱۲٫۱ درصد آرا ناامیدی و سرخوردگی از عملکرد هوش مصنوعی، در ۸٫۴ درصد توجه به خطرهای آسیب‌ناشی از هوش مصنوعی و در ۴ درصد از دیدگاه‌ها نیز نسبت به ملاحظات اخلاقی هوش مصنوعی توجه شده است [۲۶]. این نتایج، اهمیت آگاهی‌بخشی عمومی درباره ابعاد اخلاقی هوش مصنوعی و خصوصاً نقش کاربران در جهت‌دهی به فناوری‌های هوش مصنوعی را بیش از پیش برجسته می‌کند. در این مورد نیز مسئولیت آگاهی‌بخشی علاوه بر نهادهای حاکمیتی، می‌تواند ذیل مسئولیت اجتماعی سازمان‌ها و کسب‌وکارها تعریف شده و انجام آن با تعریف مشوق‌های گوناگون تسهیل شود.

در این خصوص نیز لازم است سازمان ملی هوش مصنوعی به‌عنوان متولی توسعه این فناوری در کشور، به نقش‌آفرینی فعال و مبتنی بر همکاری با نهادهای مربوطه بپردازد.

۴-۶. خط‌مشی‌گذاری مشارکت‌محور

راهبرد دیگری که مسیر توسعه مسئولانه هوش مصنوعی را هموار می‌کند، تمرکززدایی از خط‌مشی‌گذاری و مشارکت دادن افکار عمومی و ذی‌نفعان در توسعه فناوری‌های هوش مصنوعی است. این راهبرد ۲ مزیت مهم دارد. یک مزیت آن این است که کنترل کاربران بر مسیر توسعه فناوری را بیشتر می‌کند. این از جنبه‌ای حق کاربران است که نسبت به فناوری‌هایی که زندگی آنها را به شکلی غیرقابل بازگشت دگرگون می‌کند نظر داده و نقش ایفا کنند. و از جنبه‌ای دیگر باعث می‌شود به جای آنکه عاملیتی محدود و ناخودآگاه نسبت به مسیر توسعه فناوری داشته باشند، مشارکتی فعال و آگاهانه و درنتیجه مسئولانه داشته باشند.

اما مزیت دوم این است که خرد جمعی و افکار عمومی می‌تواند واجد آگاهی و اطلاعاتی درباره جنبه‌های مختلف اخلاقی این فناوری‌ها باشد که محتمل است از نظر سیاست‌گذاران دور بماند. به علت محدودیت منظر سیاست‌گذاران، طراحان و سرمایه‌گذاران، همواره این احتمال مطرح است که ابعاد اخلاقی مهم یا خطرهایی وجود داشته باشد که توسعه‌دهندگان قادر به درک آن نباشند و نمایندگان گوناگون مردم بتوانند این ابعاد پوشیده را باز کنند.

در فرایند تعاملی بین توسعه‌دهندگان و کاربران، اطلاعات لازم درباره سیستم در حال توسعه و امکان‌های بالقوه آن به کاربران داده می‌شود. به این ترتیب ذهن کاربران نسبت به موضوع حساس شده و پرسش‌های نویی در آن شکل می‌گیرد که فرایند تبیین‌پذیری را غنی‌تر می‌کند. همین‌طور با نقش نهادهای مختلف در توسعه فناوری آشنا می‌شوند که امکان مطالبه و پرسش‌گری را در آنها ایجاد می‌کند.

کاکلبرگ نیز در این راستا می‌گوید در تنگنای اخلاقی فناوری -همچون مورد هوش مصنوعی یادگیرنده که در آن پیش‌بینی‌پذیری دور از دسترس است- گفت‌وگو با ذی‌نفعان نقش مهمی در فهم بهتر شرایط و برون‌رفت از تنگنا دارد [۲۸]. استفاده از ایده‌ها، روایت‌ها و توجه به ترجیحات گروه‌های مختلف به متخصصین در درک بهتر موقعیت‌های بالقوه پیش‌رو کمک می‌کند. نهایتاً تصمیم‌سازی مشارکتی، در پذیرش فناوری در جامعه و در مسیر درست قرار گرفتن آن نقش مثبتی ایفا می‌کند.

روزر و پش نیز با اشاره به نقش عواطف عمومی در تشخیص مخاطرات فناوری، از رویکرد مشارکت‌محور دفاع می‌کنند. از نظر آنها -برخلاف تفکر سنتی- عواطف نه‌تنها در تقابل با دلیل‌آوری و عقلانیت نیستند، بلکه مؤلفه‌ای از عقلانیت و تفکر اخلاقی هستند. عواطفی همچون شفقت و همدلی و احساس مسئولیت نقش مهمی در توسعه عدالت و اخلاق دارند. فناوری‌ها در عین حال که به بهبود رفاه مردم کمک می‌کنند، به واسطه پیامدهای منفی‌شان عواطفی چون ترس و خشم را نیز در بین مردم برمی‌انگیزند. این احساسات عمومی به‌عنوان مؤلفه‌ای از خرد جمعی، باید در ارزیابی فناوری‌ها و تصمیم‌گیری برای آنها نقش داشته باشند. بر این اساس، آنها از رویکرد ارزیابی مشارکتی ریسک^۱ دفاع می‌کنند. رویکردی فرایندی برای سیاست‌گذاری، که پاسخ‌های عاطفی نسبت به مخاطرات تکنولوژیک و عواطفی که پشتوانه آن است را جدی می‌گیرد [۲۹].

● **مثال:** در مورد «به‌مان»، یک راه توسعه این فناوری این است که کمیسیون از متخصصان پزشکی تشکیل شود و درباره اینکه چه بیماری‌ها و روال‌های درمانی در دستور کار این برنامه قرار بگیرند تصمیم‌گیری شود. اما روش دیگر این است که از گروه‌های مختلف مردم، در سنین، جنسیت، قومیت، مذهب و ... مختلف دعوت شود تا درباره اینکه از چنین برنامه‌ای چه توقعاتی دارند، در مورد درمان کدام بیماری‌ها به آن اعتماد می‌کنند، آیا آن را جایگزین قابل قبولی برای پزشک می‌دانند یا خیر، چه پیامدهای نامطلوبی برای این فناوری پیش‌بینی می‌کنند، و آیا حاضر هستند داده‌هایشان در اختیار نهادهای عمومی قرار بگیرد یا نه نظر بدهند. بدون شک رویکرد مشارکت‌محور بسیاری از خلأهای اخلاقی توسعه فناوری را پر کرده و از بروز خطاهای ممکن پیشگیری می‌کند.

برای پیاده‌سازی روش مشارکت‌محور راه‌حل‌های گوناگونی پیشنهاد شده که عبارت‌اند از:

■ تشکیل گروه‌های متمرکز که در آن افراد ذی‌نفع درباره عقاید و ترجیحاتشان گفت‌وگو می‌کنند؛



■ فرایند تعاملی تحلیل سناریو که در آن یک گروه مسائل اصلی را تشخیص داده و سناریوهای مختلف را طراحی و بررسی می‌کنند و با استفاده از نظرات افراد مختلف آنها را محک می‌زنند؛

■ تشکیل گروهی از مشارکت‌کنندگان گوناگون و تجمیع آگاهی آنان برای پرورش و کاوش ایده‌ها و سیاست‌ها؛

■ تشکیل هیئت‌های داوری از شهروندان مطلع که بازنمایی‌کننده باورهای عمومی هستند.

در مورد زیست‌بوم فناوریانه ایران، روش مشارکت‌محور در سطوح مختلفی قابل پیاده‌سازی بوده و در مواردی نیز تا حدودی مورد توجه قرار گرفته است. برای مثال، ذیل اقدام‌های ملی، در تدوین [سند ملی هوش مصنوعی](#)، امکان مشارکت عمومی توسط سامانه هم‌اندیشی شورای عالی انقلاب فرهنگی فراهم شده است [۲۶]. اما علاوه بر این، همچنان فرصت‌های بیشتر و جدی‌تری برای جلب مشارکت عمومی در سطوح دیگر از جمله نهادهای سیاستگذار فناوری و همچنین کسب‌وکارها وجود دارد. خط‌مشی‌گذاری مشارکت‌محور علاوه بر اهمیتی که در ایجاد حس مسئولیت‌پذیری و جذب منابع عمومی برای توسعه مسئولانه و اخلاقی هوش مصنوعی دارد، در ایجاد اعتماد مناسب به فناوری‌های هوش مصنوعی و اجتناب از حالات بیش‌اعتمادی و کم‌اعتمادی نسبت به این فناوری‌ها که از جمله معضلات توسعه فناوری‌های هوش مصنوعی بوده نیز، سودمند است.

سازمان ملی هوش مصنوعی باید این راهبرد را به‌عنوان کلان‌راهبرد توسعه این فناوری در کشور، دنبال کند و بر این اساس با بهره‌گیری از ظرفیت شورای ملی راهبری هوش مصنوعی و با هماهنگی و همکاری با نهادهای موجود، سعی در مشارکتی کردن خط‌مشی‌گذاری در این حوزه داشته باشد.

۷. جمع‌بندی و نتیجه‌گیری



مسئله عاملیت اخلاقی و مسئولیت‌پذیری هوش مصنوعی که عمدتاً با عنوان «شکاف مسئولیت‌پذیری» مورد اشاره قرار می‌گیرد، در ۲ دهه اخیر توجه ویژه محققان حوزه فناوری را به خود جلب کرده است. این موضوع هم به لحاظ اخلاقی و هم به لحاظ حقوقی اهمیت دارد. به لحاظ اخلاقی، تعامل انسان‌ها با جهان از طریق شبکه‌هایی آمیخته از عوامل انسانی و هوش مصنوعی، پرسش از چگونگی توزیع وظایف اخلاقی را به میان می‌کشد و مسائل جدیدی را درباره وظایف اخلاقی انسان در عصر زیست با فناوری‌های هوشمند و یادگیرنده مطرح می‌کند. از جهت حقوقی نیز، اضافه شدن عواملی خودمختار که خارج از کنترل و اراده انسانی عمل کرده به شبکه تعاملات بشری، بازتعریف وظایف حقوقی عوامل انسانی یا غیرانسانی را ضروری می‌کند.

با وجود اهمیت این مسئله و ضرورت عاجل و عملی برای پاسخ‌دهی به آن، ابعاد پیچیده آن رسیدن به راه‌حلی سراسر را دور از دسترس ساخته است. مفهوم مسئولیت اخلاقی و عاملیت اخلاقی به خودی خود از جمله مفاهیم بحث‌برانگیز در میان فیلسوفان و اندیشمندان بوده و پرسش‌های متعددی درباره شرایط لازم و کافی، پیامدهای آن، و ابعاد و حالات مختلف آن مطرح شده که بحث درباره آنها از دیرباز تاکنون ادامه دارد. از طرف دیگر، هنگامی که این پرسش درباره مصنوعات مطرح شود، ابعاد پیچیده‌تری پیدا می‌کند. محققان مختلف بسته به اینکه شرایط و پیامدهای مسئولیت‌پذیری را شامل چه اموری بدانند یا چه ویژگی‌هایی را به فناوری‌های هوش مصنوعی نسبت دهند (یا علی‌الاصول برای این فناوری‌ها قابل حصول بدانند)، و حتی اینکه درباره عاملیت اخلاقی اموری دیگر مانند کودکان و حیوانات چه دیدگاهی داشته باشند، به این پرسش پاسخ‌های متفاوتی می‌دهند.

در رویکردهای اخیر نسبت به این مسئله، با چرخش از رویکرد مرسوم رو به عقب، تعبیر مسئولیت‌پذیری رو به جلو مبنا قرار گرفته است. با اخذ چنین تعبیری، به جای آنکه درگیر مباحث پرمناقشه درباره امکان یا عدم امکان انتساب مسئولیت به هوش مصنوعی

شویم، باید مسیر توسعه هوش مصنوعی را به نحوی طراحی کنیم که همواره امکان کنترل معنادار انسانی باقی بماند. در این شیوه مواجهه با مسئله شکاف مسئولیت‌پذیری، پرسش‌های عمیق فلسفی و اخلاقی درباره مسئولیت‌پذیری هوش مصنوعی کنار گذاشته می‌شود و به جای آن بر انواع تمهیداتی که موجب حفظ نظارت اخلاقی بر عملکرد هوش مصنوعی، و حفظ امکان کنترل و پاسخگویی نسبت به عملکرد سیستم‌های آمیخته انسان-هوش مصنوعی شده، تأکید می‌شود. پیشنهادهای مختلفی که در این رویکرد ارائه می‌شود، بر حفظ شفافیت، توضیح‌پذیری، مجازات‌پذیری، کنترل و پاسخگویی انسانی متمرکز شده‌اند.

رویکردی مشابه این در برخی اسناد حقوقی مربوط به هوش مصنوعی نیز دیده می‌شود. برای مثال قانون هوش مصنوعی اتحادیه اروپا (EU AI Act) راهکارهایی مشابه را برای اجتناب از شکاف مسئولیت پیشنهاد می‌دهد.

ایده تدوین قانون جامع برای تنظیم‌گری هوش مصنوعی در اتحادیه اروپا از حدود سال ۲۰۱۸ و همزمان با انتشار اسناد راهبردی کمیسیون اروپا درباره «هوش مصنوعی قابل اعتماد» شکل گرفت. کمیسیون اروپا در سال ۲۰۲۱ پیش‌نویس رسمی EU AI Act را ارائه کرد و این طرح پس از چند سال مذاکره فشرده میان کمیسیون، پارلمان اروپا و شورای اتحادیه اروپا، اصلاحات متعدد و بررسی نگرانی‌های حقوقی، فناورانه و اقتصادی، در نهایت در سال ۲۰۲۴ به تصویب نهایی رسید. طولانی شدن این فرایند نشان‌دهنده تلاش اتحادیه اروپا برای ایجاد توازن میان حمایت از نوآوری فناورانه، حفاظت از حقوق بنیادین شهروندان و ایجاد سازوکارهای مؤثر مسئولیت‌پذیری در کاربردهای هوش مصنوعی است.

دو نکته جالب در مورد این قانون وجود دارد که در زمینه پیاده‌سازی بومی راهکارهای توسعه مسئولانه هوش مصنوعی می‌تواند الهام‌بخش باشد.

۱ در این قانون، سامانه‌های هوش مصنوعی نه براساس ماهیت فنی، بلکه بر مبنای میزان خطری که برای حقوق بنیادین، ایمنی و منافع عمومی ایجاد می‌کنند، طبقه‌بندی می‌شوند. مهم‌ترین بخش این قانون به سامانه‌های «پرمخاطره» اختصاص دارد؛ سامانه‌هایی که در حوزه‌هایی مانند سلامت، آموزش، قضا، استخدام و خدمات عمومی به کار می‌روند و مشمول الزاماتی نظیر شفافیت، مستندسازی، ثبت تصمیم‌ها، امکان نظارت انسانی و تعیین صریح مسئولیت بازیگران انسانی هستند. این قانون به صراحت از انتساب مسئولیت به خود هوش مصنوعی پرهیز کرده و حفظ مسئولیت انسانی در تمام مراحل طراحی، استقرار و بهره‌برداری را الزامی می‌داند.

۲ مهم‌ترین مزیت EU AI Act در پیشگیری عملی از شکل‌گیری «شکاف مسئولیت‌پذیری» است؛ درحالی‌که طبعاً به‌عنوان یک سند حقوقی از ورود به مناقشات نظری درباره عاملیت اخلاقی هوش مصنوعی پرهیز می‌کند. این قانون با الزام به شفاف‌سازی نقش‌ها، قابلیت رهگیری تصمیم‌ها و نظارت انسانی، امکان پاسخگویی حقوقی و اخلاقی در صورت بروز آسیب را حفظ کرده و از انتقال ضمنی مسئولیت از انسان به سامانه‌های خودکار جلوگیری می‌کند. همچنین، رویکرد مرحله‌بندی‌شده و منعطف این قانون باعث شده ضمن حفاظت از حقوق عمومی، مانعی غیرضروری برای نوآوری فناورانه ایجاد نشود و اعتماد عمومی به کاربردهای هوش مصنوعی تقویت شود.

کشور ایران هم در مقایسه با سطح جهانی و هم کشورهای منطقه، هنوز در مراحل اولیه توسعه فناوری‌های هوش مصنوعی است. اگرچه عملکرد محققان ایرانی در حوزه تولید و نشر دانش هوش مصنوعی در مقایسه با سطح جهانی و منطقه‌ای قابل قبول است، اما در سایر حوزه‌ها مانند اقدام‌های ملی، فعالیت‌ها و کسب‌وکارهای کارآفرینانه، توسعه بازار، و مشروعیت‌بخشی در آغاز مسیر قرار دارد. هرچند نقاط ضعف موجود در حوزه تأمین و حکمرانی داده و فراهم ساختن زیرساخت‌های لازم برای مدیریت یکپارچه در ارتباط با پیدایش معضل شکاف مسئولیت تهدیدهایی را ایجاد می‌کند، اما ازسوی دیگر با توجه به اینکه راهبردهای ارائه شده برای پیشگیری از معضل شکاف مسئولیت، مبتنی بر وضع الزامات، قوانین، و سیاست‌هایی در سطوح مختلف حاکمیتی، سازمانی، و کسب‌وکار است، قرار داشتن در ابتدای مسیر توسعه هوش مصنوعی و باز بودن امکان شکل‌دهی به زیرساخت‌ها و قوانین، فرصتی



برای توسعه این ساحات براساس راهبردهای مسئولانه فراهم می‌کند. سازمان ملی هوش مصنوعی به‌عنوان محور توسعه این فناوری در کشور، باید سازوکارها و سیاست‌های لازم را مدنظر قرار داده و در این راستا به نقش‌آفرینی فعال اقدام کند. بر همین اساس پیشنهاد‌های زیر برای این سازمان مطرح می‌شود.

پیشنهاد‌های سیاستی و تقنینی

- تعریف شاخصی کمی برای سنجش عملکرد سازمان‌ها در ارتباط با پیشگیری از شکاف مسئولیت‌پذیری.
- تجربه‌نگاری استخراج نقشه ساختارهای اجتماعی-تکنیکی و شفاف‌سازی آنها (در سطوح درون‌سازمانی و فراسازمانی) و بررسی نقش این کار در شفافیت مسئولیت‌ها و پیشگیری از ایجاد شکاف مسئولیت.
- تجربه‌نگاری تشکیل کارگروه‌های ارزیابی اخلاقی یا نقش‌های مشابه آن (مسئولیت پاسخگویی اخلاقی در سازمان یا فراسازمان)؛ مطالعه ساختارها و ترکیب‌های مختلف ممکن برای این کارگروه؛ تعریف جایگاه آن در ساختارهای سازمانی و فراسازمانی به‌نحوی که قدرت اعمال نظر و نفوذ لازم برای این کارگروه‌ها تضمین شود.
- بررسی ابزارهای مختلف افزایش سواد علمی و آگاهی عمومی درباره توسعه هوش مصنوعی و تأثیر آنها بر حس مسئولیت‌پذیری در کاربران هوش مصنوعی و مسئولیت‌پذیری (و پاسخگویی) در توسعه‌دهندگان هوش مصنوعی.
- بررسی انواع روش‌های خط‌مشی‌گذاری مشارکتی در حوزه فناوری و مطالعه کارآمدی و تأثیر آنها بر کشف مخاطرات فناوری‌های هوش مصنوعی و درگیری عمومی با مسئولیت‌های ناشی از کاربرد این فناوری‌ها.
- تعریف سازوکارهای تشویقی برای کسب‌وکارهای مبتنی بر فناوری‌های هوش مصنوعی برای تعهد به چارچوب توسعه مسئولانه هوش مصنوعی و تعریف این تعهدات در چارچوب مسئولیت اجتماعی شرکت‌ها.



- [1] Wallace, R.J., *Responsibility and the Moral Sentiments*. 1996: Harvard University Press.
- [2] Kelly, E., What is an Excuse?, in *Blame: Its Nature and Norms*, N.T. Justin Coates, Editor. 2013, Oxford University Press. p. 244-269.
- [3] Levy, N., Culpable Ignorance and Moral Responsibility: A Reply to FitzPatrick. *Ethics and Information Technology*, 2009. **119**(4): p. 13.
- [4] Inwagen, P.V., *An Essay on Free Will*. 1983: Clarendon Press.
- [5] Frankfurt, H., Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, 1971. **68**(1): p. 16.
- [6] John Martin Fischer, M.R., *Responsibility and Control: A Theory of Moral Responsibility*. 1998: Cambridge University Press.
- [7] Fischer, J., *Deep Control: Essays in Free Will and Value*. 2011: Oxford University Press.
- [8] Talbert, M., Moral Responsibility, in *Stanford Encyclopedia of Philosophy*. 2024.
- [9] Vargas, M., *Building Better Beings: A Theory of Moral Responsibility*. 2013: Oxford University Press.
- [10] Jefferson, A., Instrumentalism about Moral Responsibility Revisited. *Philosophical Studies Quarterly*, 2019. **69**(276): p. 555-573.
- [11] MacGeer, V., Building a Better Theory of Responsibility. *Philosophical Studies*, 2015. **172**(10): p. 2635-2649.
- [12] Carlsson, A., Blameworthiness as Deserved Guilt. *Journal of Ethics*, 2017. **21**: p. 89-115.
- [13] Graham, P., A Sketch of a Theory of Moral Blameworthiness. *Philosophy and Phenomenological Research*, 2014. **88**: p. 388-409.
- [14] Smiley, M., *Collective Responsibility*. 2022.
- [15] Dusek, V., *Philosophy of Technology: An Introduction*. 2006: Blackwell Publishing.
- [۱۶] ورماس، پیترو. رویکردی در فلسفه تکنولوژی: از مصنوعات تکنیکی تا سیستم های اجتماعی- تکنیکی. ۱۳۹۱. تهران: کتاب آمه.
- [17] Matthias, A., The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology*, 2004. **6**: p. 9.
- [18]. Floridi, L, Sanders, W, On the Morality of Artificial Intelligence, *Minds and Machines*, 2004. 340-379.
- [19] Floridi, L., Faultless Responsibility: On the Nature and Allocation of Responsibility for Distributed Moral Actions. *Philosophical Transactions of Royal Society*, 2016: p. 2-13.
- [20] Sullins, J.P., When Is a Robot a Moral Agent? *International Review of Information Ethics*, 2006. **12**: p. 23-30.



- [21] Véliz, C., Moral Zombies: Why Algorithms Are Not Moral Agents. *AI and Society*, 2021. **36**(2): p. 487-497.
- [22] Riya Manna, R.N., Kantian Moral Agency and the Ethics of Artificial Intelligence. *Problemos*, 2021. **100**: p. 139-151.
- [23] Himma, K.E., Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, 2009. **11**(1): p. 19-29.
- [24] Moor, J., The Nature, Importance, and Difficulty of Machine Ethics. *Intelligent Systems*, IEEE, 2006: p. 18-21.
- [25] Dorna Behdadi, C.M., A Normative Approach to Artificial Moral Agency. 2020. **30**(2): p. 195-218.
- [26]. Iran AI Index, 2024. Online Available on: <https://aistrategy.tsc.sharif.ir/شاخص>
- [27] Mecacci, F.S.d.S.G., Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy and Technology*, 2021. **34**(4): p. 1057-1084.
- [28] Coeckelbergh, M., Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, 2020. **26**(4): p. 2051-2068.
- [29] Roeser, U.P.S., An Emotional Deliberation Approach to Risk. *Science, Technology, and Human Values*, 2016. **41**(2): p. 274-297.

عنوان: سایبرنتیک و مسئله حکمرانی (۱): عاملیت و مسئولیت‌پذیری در فناوری‌های هوش مصنوعی



DOI: <https://doi.org/10.22034/mrc.report.21955>

Permanent Link: https://report.mrc.ir/article_12005.html

Publisher : Islamic Parliament Research Center of Iran

گزیده سیاستی

گزارش حاضر ضمن تقریر مسئله مسئولیت‌پذیری و عاملیت در مورد هوش مصنوعی، روشن ساختن ابعاد مختلف آن و ارائه و دسته‌بندی پاسخ‌های مختلف ارائه‌شده به آن، راهکارهایی سیاستی و تقنینی با هدف پیشگیری از شکل‌گیری شکاف مسئولیت‌پذیری ارائه می‌دهد.



مرکز پژوهش‌های مجلس شورای اسلامی

تهران، خیابان پاسداران، رویروی پارک نیاوران (ضلع جنوبی، پلاک ۸۰۲)

تلفن: ۷۵۱۸۳۰۰۰ صندوق پستی: ۱۵۸۷۵-۵۸۵۵ پست الکترونیک: mrc@majles.ir

وبسایت: rc.majles.ir