بسم الله الرحمن الرحیم

عصر فضای مجازی

گزارش شماره ۷۶

شهریور ۱۴۰۰

مرکز ملی فضای مجازی
پژوهشگاه فضای مجازی

برخی چالش‌های اخلاقی عمومی و اختصاصی هوش مصنوعی

محتوای انتشار یافته در این اثر
الزاماً بیانگر دیدگاه مرکز ملی فضای مجازی نیست

تهیه شده در پژوهشگاه فضای مجازی
(گروه مطالعات اخلاقی فضای مجازی)

تهیه‌کنندگان: سعیده بابایی (دانشجوی دکتری فلسفه علم و فناوری پژوهشگاه علوم انسانی)
فائزه نوروزی (کارشناسی ارشد فلسفه علم دانشگاه صنعتی شریف)

ناظر علمی: محمدمهدی نصرهرندی

نشانی: تهران، میدان آرژانتین، خیابان بیهقی، نبش خیابان ۱۶غربی، پلاک ۲۰
تلفن: ۰۲۱-۸۶۱۵۱۰۶۱
کد پستی: ۱۵۱۵۶۷۴۳۱۱

فهرست

سخن نخست

سخن نخست

فضای مجازی با شتاب شگرف و رو به تزایدی که در حال بسط و گسترش است تمام ساحات اجتماعی، اقتصادی، سیاسی و فرهنگی زندگی بشر را درنوردیده و هر روز بخش بزرگی از زندگی واقعی را در خود فرو برده و حیات متفاوت و جدیدی به آن می‌دهد. لذا به نظر می‌رسد دو نگاه کلان به فضای مجازی وجود دارد: نگاه اول که بالاخص در ابتدای رشد و تکوین فضای مجازی مسلط شده بود، آن را همچون ابزاری کنار سایر ابزارهای بشری تصویر می‌کرد که تنها طریقیت داشت. اما نگاه دوم، در نتیجه رشد تحولات خیره‌کننده فضای مجازی و سایه گستری آن در حوزه‌ها و شئون بشر در یک دهه اخیر آن را چون سکویی می‌داند که بسیار فراتر از شأن ابزاری حیات انسان‌ها را سامان جدیدی داده و ادعای تمدن نوینی را دارد. رویکردی که از قضا از چشمان بصیر رهبر انقلاب نیز دور نمانده و انتظاری تمدنی از فضای مجازی در ایران را مطالبه داشته‌اند.

در همین راستا گزارش‌های عصرفضای مجازی تلاش می‌کند تا فهم سازمان‌ها و دستگاه‌های مرتبط با حوزه فضای مجازی را ارتقاء بخشیده و آن‌ها را برای مواجهه فعال و خردمندانه با تحولات این عرصه مهیا سازد.

سید ابوالحسن فیروزآبادی
دبیر شورای عالی و رئیس مرکز ملی فضای مجازی

چکیده

چکیده

با افزایش قابلیت‌های هوش مصنوعی در دهه اخیر و برخی از چالش‌هایی که در نتیجه استفاده از سیستم‌های هوشمند در زمینه‌های مختلف مطرح شده است، محققان بسیاری از رشته‌ها، از جمله اخلاق‌دانان، دریافته‌اند که ساختن سیستم‌های هوشمند اخلاقی و ایمن ضرورتی انکارناپذیر است. مباحثاتی که پیرامون چالش‌های اخلاقی هوش مصنوعی صورت گرفته است، هنوز نوپاست و نیاز به تأملاتی عمیق و بنیادی درباره شأن اخلاقی انسان و جایگاه هوش مصنوعی در زندگی بشر دارد. در این گزارش، به برخی از مهم‌ترین چالش‌های اخلاقی ناشی از پیشرفت‌های هوش مصنوعی و کاربرد آن‌ها در زندگی روزمره می‌پردازیم تا اهمیت تأثیرات اخلاقی این فناوری را روشن کنیم و بر ضرورت توجه متخصصان این حوزه، به تأملات اخلاقی بیشتر در این باره تأکید کنیم.

واژگان کلیدی: فضای مجازی، هوش مصنوعی، اخلاق، سوگیری، شأن اخلاقی، تکینگی، پیش‌بینی‌پذیری، عواقب ناخواسته

مقدمه

مقدمه

هوش مصنوعی، یکی از حوزه‌هایی است که به دلیل ماهیت بین‌رشته‌ای و کارایی‌های بسیاری که دارد، در دهه اخیر بسیار مورد توجه قرار گرفته است و سرعت رشد بالایی دارد. این حوزه دارای پیچیدگی‌ها و ظرافت‌های بسیاری است و تاثیرات فاحشی در ابعاد گوناگون بر جامعه، فرهنگ، اخلاق و سبک زندگی می‌گذارد. از این رو، نمی‌توان از این تاثیرات غافل بود و باید برای پیش‌بینی مخاطرات حاصل از هوش مصنوعی پیش از آنکه مسیر توسعه آن غیر قابل تغییر شود، چاره‌اندیشی نمود. فناوری، از جمله هوش مصنوعی، باید عاملی برای شکل‌گیری زندگی خوب[1] باشد و ما را تبدیل به انسان‌های بهتری کند. همانطور که استیون هاوکینگ ادعا می‌کند، ظهور هوش مصنوعی می‌تواند «بدترین یا بهترین اتفاقی باشد که برای بشریت رخ داده است» و این مسئولیت ما به عنوان طراحان، توسعه‌دهندگان، دولت‌ها، کسب و کارها و نهایتا انسان‌ها و اعضای جامعه است که از تحقق دومین حالت اطمینان حاصل کنیم.

هوش مصنوعی به نحو روزافزونی در حال نفوذ به همه ساحات جامعه از جمله بانک‌داری، سلامت، اعمال قانون، زیرساخت‌های شهری

1. Good Life

و ... است. انتظار می‌رود که بین سال‌های ۲۰۲۰ تا ۲۰۶۰، ابرکامپیوترها تقریباً در همه حوزه‌ها از قابلیت‌های انسانی هم فراتر بروند. بخشی از تأثیرات هوش مصنوعی، نامطلوب و نیازمند مداخله‌هایی از جنس سیاست‌گذاری است. غول‌های فناوری مانند آلفابت، آمازون، فیس‌بوک، آی‌بی‌ام و مایکروسافت (همچنین افرادی از جمله استیون هاوکینگ و ایلان ماسک) بر این باورند که زمان آن فرا رسیده تا درباره چشم‌انداز نسبتا بی حد و مرز هوش مصنوعی سخن بگوییم. اما به همان اندازه که فناوری‌های نوظهوری مانند هوش مصنوعی جدید هستند، مباحثات درباره چالش‌های اخلاقی و ارزیابی مخاطرات آن‌ها نیز تازگی دارد.

اخلاق هوش مصنوعی یکی از شاخه‌های اخلاق فناوری است که شامل بخش‌هایی از جمله اخلاق روباتیک و اخلاق ماشین می‌شود. امکان ساختن ماشین‌های متفکر، مسائل اخلاقی متعددی را پیش روی ما قرار می‌دهد. این پرسش‌ها شامل مواردی از این قبیل می‌شوند که آیا چنین ماشین‌هایی موجب آسیب رساندن به انسان‌ها و سایر موجوداتی که دارای شأن اخلاقی هستند، می‌شوند؟ شأن اخلاقی این ماشین‌ها چیست؟ هوش مصنوعی از چه جهاتی، با توجه به ارزیابی اخلاقی ما از آن، با انسان‌ها فرق دارد؟ و .... آیزاک آسیموف یکی از نخستین کسانی است که در باب چالش‌های اخلاقی ناشی از هوش مصنوعی سخن گفته است. او علی‌رغم اینکه در کتاب خود «من، روبات» سه قانون روباتیک را برای سیستم‌های هوشمند مصنوعی نوشت، اما معتقد بود هیچ مجموعه معینی از قوانین نمی‌توانند به قدر کافی همه شرایط ممکن را پیش‌بینی کنند. در این گزارش نیز امکان

پرداختن به همه مخاطرات محتمل پیش روی هوش مصنوعی وجود ندارد، اما سعی کردیم به عمده چالش‌های اخلاقی ناشی از هوش مصنوعی از نگاه صاحب‌نظران و متخصصان این حوزه اشاره کنیم. این چالش‌ها در دو ساخت عمومی و اختصاصی (مختص به حوزه مورد استفاده هوش مصنوعی) بررسی خواهند شد.

ما باید با جهان و هر آنچه در آن است با کرامت و احترام برخورد کنیم. این مهم است که بدانیم تصمیماتی که درباره جهان می‌گیریم و عاداتی که آن تصمیمات در ما ایجاد می کنند، ما را تبدیل به چگونه انسان‌هایی خواهد کرد؟ پرسش اصلی این است که ما می‌خواهیم چگونه انسان‌هایی بشویم؟ باید به این توجه کنیم که فناوری‌های ساخته دست ما، ما را انسان‌های بهتر یا بدتری می‌کند. پیشرفت‌های فناوری، لزوما ما را تبدیل به انسان‌های بهتری نکرده است و پیشرفته‌تر بودن ما نسبت به نسل‌های پیشین، به معنی خردمندتر بودن ما نیست. وقتی ماشین‌ها همه کارهایی که انسان انجام می‌داده را بر عهده بگیرند، کاری برای انسان باقی نخواهند گذاشت. در نتیجه، ممکن است انسان‌ها رشد ویژگی‌های انسانی خود در بسیاری از زمینه‌ها را متوقف کنند و آن را به ماشین واگذار کنند و به این ترتیب ماشین‌ها در همه زمینه‌ها عملکرد بهتری از انسان پیدا کنند. ما باید تصمیماتی آگاهانه در این باره بگیریم که کدام بخش از کارهایمان را می‌خواهیم به فناوری واگذار کنیم و کدام بخش از این کارها قابل واگذاری به فناوری هستند و حضور انسان در انجام آن‌ها ضروری نیست. تکیه بیش از حد به فناوری می‌تواند عواقب اخلاقی و وجودی سوئی برای نسل بشر داشته باشد، بنابراین، ضرورت

تأمـلات و دوراندیشـی‌هایی از ایـن دسـت، بـرای تصمیم‌گیـری دربـاره آینـده فناوری‌هـا، خصوصـا فناوری‌هـای پیچیـده و نویدبخشـی ماننـد هـوش مصنوعـی، انکارناپذیـر اسـت.

بخش اول

بخش اول

برخی چالش‌های اخلاقی عمومی هوش مصنوعی

۱-۱- عامل‌های اخلاقی هوشمند

امروزه، انواع مختلفی از روبات‌ها و ماشین‌های هوشمند طراحی شده‌اند که کاربردهای مختلفی مثلا در پزشکی، صنایع نظامی، برقراری تعاملات اجتماعی و ... دارند و در موقعیت‌هایی مورد استفاده قرار می‌گیرند که مستلزم تصمیم‌گیری‌های اخلاقی است. مثلاً روبات‌هایی که در تعامل با و مراقبت از سالمندان و معلولین به کار می‌روند و باید متناسب با شرایط حساس روحی و جسمی آن‌ها تصمیماتی را بگیرند. روبات‌هایی که معلم یا دستیار معلم یک کلاس هستند نیز باید در نسبت با دانش‌آموزان تصمیم‌گیری‌هایی را انجام دهند. روبات پرستار نیز نمونه دیگری از این روبات‌هاست که برای تعامل با کودکان به کار می‌رود و کودک را از کاری منع یا به آن تشویق می‌کند و این مستلزم تصمیم‌گیری‌های اخلاقی این روبات و تنظیم کردن نحوه تعامل او با کودکان است.

امکان ساخت ماشین‌های اخلاقی و ماشین‌هایی که بتوانند تصمیم‌گیری کنند، پرسش‌هایی از این دست را پیش روی ما می‌نهد: آیا روبات‌ها می‌توانند عامل‌های اخلاقی هوشمند[1] (AMA) باشند

1. artificial moral agents

و تصمیم‌های اخلاقی بگیرند؟ کدام چارچوب اخلاقی باید مبنای تصمیمات آن‌ها باشد؟ آیا می‌توان چارچوب اخلاقی معینی (مثلا چارچوب اخلاقی کانتی یا اخلاق سودگرایی و ...) را در یک ماشین جای‌دهی کرد؟ روبات‌هایی که به طور کامل اخلاقی نیستند، ممکن است چه آسیب‌هایی به انسان برسانند؟ عاملیت اخلاقی روبات‌ها چگونه قابل ارزیابی است؟ پرسش‌هایی از این دست موجب شد که ما تأملی دوباره درباره مفهوم عاملیت[1] بکنیم و معنای سنتی آن را بازبینی کنیم.

ماشین‌های اخلاقی، ماشین‌هایی هستند که به گونه‌ای طراحی می‌شوند که به انسان‌ها و موجودات دیگری که دارای شأن اخلاقی[2] هستند، آسیبی نمی‌رسانند. همچنان که میزان خودمختاری هوش مصنوعی بیشتر می‌شود، طراحان و مهندسان نیز باید انتخاب‌ها و کنش‌های این ماشین‌ها را در موقعیت‌های مختلف پیش‌بینی کنند و عواقب خواسته یا ناخواسته آن‌ها را از منظری اخلاقی ارزیابی و برای آن‌ها چاره‌اندیشی کنند. در پاسخ به این نیاز، حوزه‌های مطالعاتی مختلفی همچون اخلاق ماشین[3] (Anderson and Anderson ۲۰۰۶a و Wallach et al. ۲۰۰۸)، اخلاق هوش مصنوعی[4] (Danielson ۱۹۹۲)، اخلاقی محاسباتی[5] (Allen ۲۰۰۲)، و هوش مصنوعی دوستانه[6] (Yudkowsky ۲۰۰۱) ظهور یافته‌اند. وقتی از اخلاق در هوش مصنوعی صحبت می‌کنیم، عموماً سه بعد را در نظر داریم: ۱. نظام‌های اخلاقی که در هوش مصنوعی جای‌دهی[7] می‌کنیم ۲. اخلاق برای افرادی که هوش مصنوعی را طراحی و

1. agency
2. moral status
3. Machine Ethics
4. Ethics of AI
5. Computational Ethics
6. Friendly AI
7. embed

استفاده می‌کنند (طراحان، مهندسان، کاربران و ...) ۳. اخلاق برای نحوه تعامل و برخورد کاربران با هوش مصنوعی. (Asaro, ۲۰۰۶)

به طور کلی، می‌توان از چهار سطح برای عاملیت اخلاقی سخن گفت(Moor, ۲۰۰۶)؛ ۱. عامل‌های دارای تأثیر اخلاقی[1]: در پایین‌ترین سطح عاملیت، عامل اخلاقی دارای تاثیر اخلاقی است. یعنی یک فناوری با اینکه خود هیچ اخلاقیاتی ندارد، اما دارای تاثیر اخلاقی است. ۲. عامل‌های اخلاقی ضمنی[2]: سطح دوم، عاملیت اخلاقی ضمنی است که به این معناست که فناوری برخی از کارها را بدون اینکه بداند، انجام می‌دهد؛ چیزی مشابه عادت‌های روتین. در این سطح، اخلاق به نحو ضمنی در کارکرد فاعل وجود دارد، حتی اگر فاعل در آن لحظه اخلاقی بودن را انتخاب نکرده باشد. این را می‌توان ناشی از ارزش‌بار بودن فناوری دانست. ۳. عامل‌های اخلاقی صریح[3]: عامل اخلاقی واجد الگوریتم‌هایی است که موجب می‌شوند او به نحو اخلاقی عمل کند ۴. عامل‌های اخلاقی کامل[4]: در سطح چهارمِ عاملیت اخلاقی، صحبت از نسبت دادن مسئولیت اخلاقی به عامل می‌شود. عامل در این سطح، دارای اراده آزاد، آگاهی و قصدمندی است.

ما از کسی که مسئولیت اخلاقی دارد، انتظار پاسخ داریم. اگر چیزی به لحاظ اخلاقی درست نباشد، او را سرزنش و مجازات می‌کنیم. اما ممکن است فردی را مسئول بدانیم ولی او را سرزنش نکنیم، زیرای خطای او سهوی بوده است. فناوری را، دست کم در حال حاضر، نمی‌توان سرزنش کرد، چرا که آن را مسئول نمی‌دانیم. ما معمولا مسئولیت را به شکل ضمنی متوجه روبات می‌دانیم و شاید برای دفعه بعد، آن را طوری برنامه‌ریزی کنیم که درست‌تر عمل کند. هوش مصنوعی

1. ethical impact agents
2. implicit ethical agents
3. explicit ethical agents
4. full ethical agents

نیـز از بسـیاری جهـات ماننـد فناوری‌هـا و ماشـین‌های دیگـر اسـت، ولـی قابلیت‌هایـی در آن وجـود دارد کـه می‌توانـد بـر اسـاس اصـول اخلاقـی کدنویسـی شـود تـا قـادر بـه انتخـاب و انجـام دادن عمـل درسـت اخلاقـی باشـد و حتـی ممکـن اسـت بتوانـد زمانـی کـه دو قانـون یـا ارزش اخلاقـی در تضـاد بـا هـم قـرار می‌گیرنـد، موقعیـت را تحلیـل کنـد و ارزش‌هـا را اولویت‌بنـدی کنـد. هـوش مصنوعـی گاهـی می‌توانـد از انسـان‌ها اخلاقی‌تـر عمـل کنـد، زیـرا برخـی از سـوگیری‌های انسـانی (کـه شـاید خـود انسـان هـم از آن آگاه نباشـد) را نـدارد. مثـلا در انتخـاب میـان اینکـه چـه کسـی را در یـک فاجعـه نجـات دهنـد.

یـک عامـل اخلاقـی کامـل، فاعلـی نیسـت کـه هرگـز اشـتباه نمی‌کنـد. چنیـن فاعلـی اشـتباه می‌کنـد. اگـر انسـان‌ها را عامل‌هـای اخلاقـی بدانیـم، اشـتباه بخشـی از زیسـت اخلاقـی آنهاسـت. خطاهـای انسـانی اسـت کـه باعـث می‌شـود عذرخواهـی کنیـم، احسـاس گنـاه و عـذاب وجـدان داشـته باشـیم. ولـی ابزارهـا ماننـد ماشـین‌ها یـا پهپادهـا بـرای اشـتباه کـردن طراحـی نمی‌شـوند. مـا آن‌هـا را بـه نحـوی طراحـی می‌کنیـم کـه کارشـان را بی‌نقـص و بـدون خطـا انجـام دهنـد. پـس اگـر ماشـین‌ها فاعل‌هایـی باشـند کـه حـق اشـتباه کـردن ندارنـد، یعنـی آزادی انتخـاب هم ندارنـد. در نتیجـه، هرگـز نمی‌توانـد کامـلا خودمختـار باشـند و خودمختـار بـودن را یاد نمی‌گیرنـد. پـس هرگـز شـأن یـک عامـل کامـلا اخلاقـی را نخواهند داشـت.

۱-۲- نقش احساسات در تصمیم‌های اخلاقی

یکـی از چالش‌هـای مهـم در تصمیم‌گیری‌هـای اخلاقـی روبات‌هـا و ماشـین‌های هوشـمند ایـن اسـت کـه آیـا ایـن ماشـین‌ها می‌تواننـد دارای

احساسات باشند یا خیر. این موضوع از این جهت حائز اهمیت است که نسبت میان عقل و احساس در فعالیت‌های شناختی[1] و قضاوت‌های اخلاقی همواره مورد بحث بوده است. یونانیان باستان، متاثر از افلاطون، احساسات را مایه انحراف و مانع تفکرِ موثر می‌دانستند و تفکر و تصمیم‌گیری عقلانی را مستقل از احساسات تلقی می‌کردند. اما در قرن ۱۸، دیوید هیوم و آدام اسمیت و سپس به طور جدی‌تر، ویلیام جیمز بر اهمیت احساسات اخلاقی تأکید ورزیدند، تا جایی که هیوم، احساسات را مقدم بر عقل می‌دانست. مطالعات عصب‌شناختی نیز اخیراً نشان داده است که یکی از مهمترین عوامل در فرایند تصمیم‌گیری انسان، احساسات است و نمی‌توان فرایندهای عقلانی را مستقل از احساسات در نظر گرفت. حال ممکن است این پرسش مطرح شود که احساسات به چه معنا و دارای چه ماهیتی هستند؟ پاسخ‌های مطرح شده به این پرسش را می‌توان در قالب دو دیدگاه گردآوری کرد: ۱. قضاوت در مورد حال عمومی فرد، ۲. واکنش‌های جسمانی فرد (Thagard, 2005).

قائلان به دیدگاه اول (از جمله نوسبام[2]، اتلی[3]، شِرر[4] و ...) قضاوت فرد از شرایطی که برای او پیش می‌آید و موجب نزدیکی او به، یا دوری او از اهدافش می‌شود را عامل اصلی بروز انواع احساسات مانند خشم، ترس و ... می‌دانند. بنابراین، احساسات حاصل شکل‌گیری یک بازنمایی کلی از موقعیت نهایی برای حل مسئله توسط فرد است. از آنجایی که فرایند حل مسئله در انسان بسیار پیچیده و مستلزم تحقق چندین هدف متضاد است، احساسات کمک می‌کنند که فرد جوانب مهم موقعیت‌های مختلف را شناسایی کند و روی آن‌ها تمرکز کند و

1. cognitive
2. Nussbaum
3. Oatley
4. Scherer

بـا در نظـر گرفتـن حـس کلـی خـود نسـبت بـه موقعیـت پیش‌امـده، انگیـزه کافـی را بـرای تصمیم‌گیـری و عمـل پیـدا کنـد. در نتیجـه، بـرای تبییـن دلیـل اعمـال افـراد، بایـد بـه حالت‌هـای احساسـی آن‌هـا توجـه کـرد کـه ایـن حالت‌هـا قابـل تحویـل بـه بازنمایی‌هـای کلامـی نیسـتند. بـرای فهـم احساسـات فـرد، بایـد خـود را در موقعیـت او تصـور کـرد و بـا برقـراری ایـن تمثیـل، رابطـه همدلانـه‌ای ایجـاد کـرد تـا احساسـی شـبیه احسـاس او را درک کـرد.

قائـلان بـه دیـدگاه دوم، از جملـه ویلیـام جیمـز (۱۸۹۰)، احساسـات را نـه مبتنـی بـر قضاوت‌هـای شـناختی، کـه حاصـل واکنش‌هـای جسـمانی در نظـر می‌گیرنـد. بـه ایـن معنـا کـه اگـر بـا قرارگیـری فـرد در یـک موقعیـت، تغییراتـی جسـمانی در او حاصـل شـود، مغـز واکنش‌هایـی از جنـس احساسـات بـه ایـن تغییـرات جسـمانی نشـان می‌دهـد. رویکـرد میانـه‌ای نیـز وجـود دارد کـه قائـل بـه جمـع میـان دو دیـدگاه پیش‌گفتـه اسـت. بـرای مثـال، موریـس[1] (۲۰۰۲) احساسـات را متکـی بـر تعامـل میـان پیام‌هـای جسـمانی و ارزیابی‌هـای شـناختی می‌دانـد.

برخـی از مطالعـات نشـان می‌دهـد کـه کسـانی کـه از قـدرت همدلـی برخـوردار نیسـتند، نمی‌تواننـد قضـاوت اخلاقـی درسـتی داشـته باشـند. حـس گنـاه و سـرزنش، بی‌تفاوتـی نسـبت بـه مجـازات و نداشـتن همدلـی و بـه طـور کلـی نداشـتن احساسـات، در قضـاوت اخلاقـی، رشـد اخلاقـی و انگیـزش اخلاقـی نقـش دارنـد. برخـی از متخصصـان معتقدنـد کـه حـس همدلـی در روابـط انسـانی ضـروری اسـت و هـوش مصنوعـی اگـر بخواهـد جایگزیـن بخشـی از روابـط انسـانی شـود، بایـد بتوانـد بـه نحـوی همدلـی را بازنمایـی کنـد. در غیـر ایـن صـورت، ممکـن اسـت تهدیـدی بـرای کرامـت

1. Morris

انسانی باشد (Thagard, 2005).

سوال اصلی ما در نسبت احساسات با هوش مصنوعی این است که آیا می‌توان احساسات را در سیستم‌های هوشمند جای‌دهی کرد؟ این پرسش از رو مهم است که احساسات نقش مهمی در قضاوت‌های اخلاقی دارند و دستگاه‌های هوشمند برای اینکه قادر به تصمیم‌گیری و داوری اخلاقی باشند، باید بتوانند چیزی از جنس احساس را تجربه کنند. به علاوه، هوش مصنوعی برای ایجاد ارتباط احساسی با انسان، باید بتواند احساسات انسانی را تشخیص دهد و پاسخگوی آن‌ها باشد. بنابراین، دریافتن نحوه ارتباط تفکر و احساسات در فرایند تصمیم‌گیری، امری مهم محسوب می‌شود. تصمیم‌گیری فرایندی احساسی/شناختی است و با توجه به ارزیابی‌های مثبت و منفی از تصویر موقعیت‌های مختلف صورت می‌گیرد. اگرچه احساسات نقش مهمی در تصمیم‌گیری دارند، اما اذعان به این مطلب به معنای عمل کردن بی چون و چرا بر اساس احساسات نیست، بلکه صرفا باید آن‌ها را در تصمیم‌گیری‌ها مدنظر قرار داد.

۱-۳- شأن اخلاقی هوش مصنوعی

با یک انسان، نه به مثابه یک ابزار، بلکه به مثابه یک غایت باید برخورد کرد و محدودیت‌های اخلاقی جدی را در مواجهه با او باید پذیرفت؛ مانند ممنوعیت قتل او، دزدی کردن از او، یا انجام کارهای دیگری در حق او یا اموالش بدون رضایت او. انسان دارای شأن اخلاقی است و انجام این قبیل کارها در حق او نارواست. اما آیا این شأن را می‌توان برای هوش مصنوعی هم قائل بود یا فقط موجودات بیولوژیک

از چنین حقوقی برخوردارند؟ یکی از جدی‌ترین چالش‌ها در حوزه اخلاق هوش مصنوعی این است که حقوق روبات‌ها چگونه مشخص می‌شود و این حقوق، چه نسبتی با وظایف آن‌ها دارد؟ آیا می‌توان حقوقی مانند آزادی، حیات، آزادی بیان و برابری را برای روبات‌ها هم قائل بود؟ وقتی یک سیستم هوش مصنوعی در انجام وظیفه خود ناموفق عمل می‌کند، چه کسی مقصر است؟ برنامه‌نویس؟ کاربر نهایی؟

سیستم‌های هوش مصنوعی در حال حاضر نسبتا سطحی هستند و شاید این سوالات در مورد آن‌ها هنوز موضوعیتی نداشته باشند، اما به مرور پیچیده‌تر و واقع‌نماتر می‌شوند و در تصمیم‌گیری‌ها مداخله بیشتری می‌کنند. وقتی ماشین‌ها را به مثابه هستومند[1]هایی که توانایی ادراک[2]، احساس و عمل دارند در نظر بگیریم، اندیشیدن در باب شأن اخلاقی و قانونی آن‌ها عجیب نخواهد بود. آیا باید با ماشین‌های هوشمند به مثابه موجوداتی با سطح هوش مشابه حیوانات رفتار کنیم؟ آیا می‌توان به سیستمی که وقتی عملیات پاداشش ورودی منفی دریافت می‌کند، تجربه رنج کشیدن یا عذاب وجدان نسبت داد؟ آیا رنج ماشین‌های «دارای احساس» را باید در نظر بگیریم؟

پرسش درباره شأن اخلاقی، در برخی از حوزه‌های اخلاق کاربردی بسیار مهم است. توافق عمومی بر این است که سیستم‌های هوش مصنوعی فعلی دارای هیچگونه شأن اخلاقی نیستند. ما می‌توانیم برنامه‌های کامپیوتری را مطابق میلمان تغییر دهیم و آن‌ها را کپی، پاک یا استفاده کنیم. همه محدودیت‌های اخلاقی ما در مواجهه با سیستم‌های هوش مصنوعی، فعلی ریشه در مسئولیت‌های ما نسبت

1. entity
2. perception

به موجودات دیگر، مانند انسان‌ها و حیوانات دارد نه در وظایف ما نسبت به خود این سیستم‌ها. با این حال، هنوز توافقی بر سر اینکه که دقیقا چه ویژگی‌هایی شأن اخلاقی را به وجود می‌آورند نیست. اما به طور کلی می‌توان دو رویکرد را در نسبت دادن شأن اخلاقی به مصنوعات متمایز کرد: ۱. رویکرد پیوسته[1] ۲. رویکرد ناپیوسته[2].

طبق رویکرد پیوسته، عامل‌های اخلاقی هوشمند و عامل‌های اخلاقی انسانی هیچگونه تفاوت کیفی قابل اعتنایی با یکدیگر ندارند. در مقابل، رویکرد ناپیوسته، تفاوتی ذاتی میان انسان‌ها و عامل‌های اخلاقی هوشمند قائل است. به عقیده باستروم (۲۰۱۱)، معمولا دو ویژگی به عنوان ویژگی‌های مربوط به شأن اخلاقی پیشنهاد می‌شوند: ۱. ادراک: توانایی تجربه پدیداری[3] یا تجربه کیفیات[4]، مانند توانایی حس کردن درد و رنج. ۲. معرفت[5]: مجموعه‌ای از توانایی‌ها که به هوش عالی‌تر مربوط می‌شوند، مانند خودآگاهی و پاسخگویی به عقل. طبق این دیدگاه، یک سیستم هوش مصنوعی اگر ظرفیت درک کیفیاتی مانند درد را داشته باشد، دارای شأن اخلاقی خواهد بود. اگر علاوه بر ادراک، سیستم هوش مصنوعی دارای معرفتی از نوع معرفت انسان بالغ معمولی نیز باشد، آنگاه می‌تواند دارای شأن اخلاقی کاملی معادل انسان‌ها باشد.

مبناهای اخلاقی‌ای مانند مبنای اخلاقی باستروم و جیمز مور که پیش‌تر مطرح شد، این پیش‌فرض را دارند که هیچگونه تفاوت ذاتی میان انسان و عامل‌های اخلاقی هوشمند وجود ندارد و شأن اخلاقی دارای طیفی پیوسته است که همه این موجودات با درجات

1. Continuity Approach
2. Discontinuity Approach
3. phenomenal experience
4. qualia
5. knowledge

مختلف خودمختاری و فعل و انفعلات و تأثیرات اخلاقی، در نقاط مختلف آن قرار دارند. بنابراین، وظایف اخلاقی ما که برگرفته از ملاحظات شأن اخلاقی هستند، ممکن است ایجاب کند که با یک ذهن مصنوعی به همان شیوه‌ای رفتار کنیم که با ذهن انسان طبیعی با کیفیتی معادل آن در موقعیتی مشابه رفتار می‌کنیم. ما انسان‌ها معمولا به فردی که انواع متعارفی از رفتار انسانی را نشان می‌دهد و با نگاه از بیرون، کنش‌های آن‌ها تمایزی با کنش‌های ما ندارد، ادراک و تجربه آگاهی و به طور کلی، هوشمندی را نسبت می‌دهیم. این در حالی است که ذهن مصنوعی ممکن است ساختاری کاملا متفاوت با ذهن انسانی داشته باشد و در عین حال، رفتاری شبه انسانی بروز دهد. شأن اخلاقی سیستم‌های هوشمند، موضوعی است که هنوز توافق جامعی درباره آن وجود ندارد.

۱-۴- شفافیت و منبع-باز بودن هوش مصنوعی

هوش مصنوعی می‌تواند تاثیرات شگرفی روی ساحات مختلف زندگی بشر داشته باشد. از این رو، برخی از متخصصان معتقدند که مردم این حق را دارند که در مورد نحوه پیشرفت این فناوری مطلع باشند و توسعه‌دهندگان آن موظفند حداکثر شفافیت[1] را درباره هوش مصنوعی با مردم داشته باشند. برای این منظور، موسسات منبع-بازی[2] مانند Open AI و AI Now توسط افراد صاحب‌نامی در حوزه هوش مصنوعی مانند ایلان ماسک[3]، سم آلتمن[4]، مردیث ویتاکر[5] و کیت کرافورد[6] تاسیس شده است. منبع-باز بودن به این معناست که

1. Transparency
2. open source
3. Elon Musk
4. Sam Altman
5. Meredith Whittaker
6. Kate Crawford

کدهای منبع، بینش‌های علمی و الگوریتم‌ها، ایده‌ها، داده‌ها، تکنیک‌های امنیتی، قابلیت‌ها و اهداف سازمانی در حوزه هوش مصنوعی، در انحصار عده‌ای معدود نباشد و عموم مردم بتوانند به آن‌ها دسترسی داشته باشند.

شیوه‌های یادگیری عمیق و شبکه‌های عصبی مصنوعی به گونه‌ای طراحی شده‌اند که ما واقعا نمی‌فهمیم سیستم چه چیزی را یاد می‌گیرد. چنین سیستمی مانند جعبه سیاهی رفتاری می‌کند که نمی‌توانیم آن را کنترل کنیم. الگوریتم‌های پیچیده در سیستم‌های هوشمند با هدف‌های بزرگ و متنوعی طراحی شده‌اند. آن‌ها می‌توانند صورت انسان را بازشناسی کنند، دست خط را تشخیص بدهند، تقلب در کارت‌های اعتباری را کشف کنند، اسپم‌ها را مسدود کنند، زبان‌ها را به یکدیگر ترجمه کنند، تومورها را در تصاویر پزشکی کشف کنند و در بازی‌های شطرنج و گو در مقابل انسان‌ها پیروز شوند. اما با افزایش پیچیدگی این الگوریتم‌ها و نرم‌افزارها همزمان شفافیت آن‌ها نیز برای ما کمتر و کمتر می‌شود.

در حال حاضر، بسیاری از کمپانی‌های بزرگ همچون گوگل[1]، فیسبوک[2]، مایکروسافت[3] و بایدو[4] به طور منظم کارهای اخیر خود را تا حد زیادی در کنفرانس‌های تخصصی ارائه می‌کنند و آن‌ها را روی سرورهای خود قرار می‌دهند. منبع-باز بودن موجب می‌شود بسیاری از حوزه‌های نگرانی‌برانگیز از جمله حوزه کاربردهای نظامی، کنترل اجتماعی، و مخاطرات سیستمی که ناشی از اطمینان به فرایندهای خودمختار پیچیده هستند، شناسایی شود.

در مورد آثار مثبت و منفی کوتاه‌مدت، میان‌مدت و بلندمدت شفافیت

1. Google
2. Facebook
3. Microsoft
4. Baidu

هوش مصنوعی نظرات مختلفی مطرح است. یکی از مهمترین نگرانی‌هایی که وجود دارد این است که منبع-باز بودن هوش مصنوعی می‌تواند منجر به شتاب گرفتن رشد این حوزه و نفوذ و کاربرد آن شود. حوزه هوش مصنوعی هنوز به حوزه‌ای کاملا رقابتی تبدیل نشده است، اما منبع-باز کردن آن می‌تواند منجر به دسترسی آزاد و کم‌هزینه آن برای همه بشود و برای این اتفاق وجوه مثبت متعددی را در کوتاه‌مدت می‌توان برشمرد. در میان‌مدت، که زمان به قدر کافی سپری شده است تا تحقیقات جدید عملیاتی و کاربردی شوند، مسائلی از جنس مسائل اقتصادی و سرمایه‌گذاری در این حوزه مطرح می‌شود. شرکت‌ها دیگر دارای حق انحصاری برای استفاده از ایده‌ها و نوآوری‌هایی که تولید می‌کنند، نیستند و این می‌تواند زیان اقتصادی برای آن‌ها به همراه داشته باشد.

منبع-باز شدن هوش مصنوعی همچنین موجب می‌شود که شرکت‌های رقیب سعی کنند اولین گروهی باشند که هوش مصنوعی پیشرفته‌تری را عرضه می‌کنند و در نتیجه، سرعت رشد این حوزه به مراتب بالاتر می‌رود. این امر موجب می‌شود مخاطرات و پیامدهای ناخواسته[1] ناشی از این فناوری هم بیشتر و پیچیده‌تر شود و کنترل آن‌ها دشوارتر باشد. برای کاهش این مخاطرات و کنترل عواقب ناخواسته هوش مصنوعی، لازم است زیرساخت‌های مناسب برای آمادگی در مواجهه با این تغییرات ایجاد شود. از سوی دیگر، باز بودن منبع این کمپانی‌ها زمینه را برای مشارکت و مساهمت بیشتر افراد بیرونی فراهم می‌کند. این افراد داوطلبانه وقت آزاد خود را در جهت بالا بردن امنیت پروژه‌های هوش مصنوعی صرف

1. unintended consequences

می‌کنند و اهداف و برنامه‌هایی سازمان‌یافته برای نظارت این افراد بر هوش مصنوعی پیش روی آن‌ها قرار می‌گیرد (Bostrom, 2017). محققان هوش مصنوعی نیز می‌توانند با این شرکت‌ها و سازمان‌ها همکاری کنند و کارهای خود را در کنفرانس‌های فنی ارائه کنند که در نتیجه آن، پیوندی میان محققان مستعد و شرکت‌ها برقرار می‌شود و پیشرفت علمی این حوزه سرعت بیشتری پیدا می‌کند.

۱-۵- پیش‌بینی‌پذیری هوش مصنوعی

عامل‌های هوشمند پیچیده، می‌توانند رفتارهای نوظهوری[1] را از خود بروز دهند که گاهی برای سازندگان و برنامه‌نویسان آن‌ها نیز قابل پیش‌بینی نیست. این رفتارها تنها زمانی خود را نشان می‌دهند که سیستم‌های هوشمند در تعامل با جهان و عامل‌های دیگر در محیط قرار بگیرند. بنابراین، محدودیت‌های نظری بنیادینی برای ما وجود دارد برای اینکه پیش‌بینی کنیم که یک کد مشخص، خصوصیات مطلوبی که تعیین کرده‌ایم را بروز می‌دهد یا خیر و برای فهمیدن این موضوع، فقط باید کد را اجرا کنیم و رفتار سیستم را در عمل مشاهده کنیم.

پیش‌بینی دقیق و مداوم فرایند تصمیم‌گیری و کنش‌های یک سیستم هوشمند، یا سیستمی هوشمندتر از انسان، حتی در حالتی که اهداف نهایی او برای ما مشخص است، امکان‌پذیر نیست. این سیستم‌ها به مثابه جعبه سیاهی عمل می‌کنند که دستیابی به مدل‌هایی برای پیش‌بینی رفتار آن‌ها امری بسیار دشوار است و توانایی ما برای فهم تأثیرات آن‌ها محدود است. آشوبناک[2] بودن یا

1. emergent
2. chaotic

متلاطم[1] بودن دینامیک سیستم‌ها در هوش مصنوعی، موجب محدودیت ما در شناسایی مدل درست برای توضیح وقایع پیشین، ارزیابی موقعیت‌های فعلی و پیش‌بینی وقایع آینده می‌شود. حتی اگر برنامه‌نویسان همه کارها را درست انجام دهند، ممکن است رفتار موضعی و خاص هوش مصنوعی قابل پیش‌بینی نباشد و این رفتارها می‌توانند مخاطرات بالقوه‌ای را برای انسان‌ها به وجود بیاورند.

چنانکه وولفرام (۲۰۰۲) ادعا می‌کند، رفتارهای پیچیده برنامه‌ها را نمی‌توان بدون اینکه این برنامه‌ها را در واقعیت اجرا کرد، پیش‌بینی نمود. متخصصانی که در حوزه هوش مصنوعی عمومی[2] فعالیت می‌کنند نیز اگرچه ممکن است بدانند هدف نهایی سیستمی که طراح کرده‌اند چیست و بتوانند آن را پیش‌بینی کنند، اما نمی‌دانند که این سیستم هوشمند یا ابرهوشمند در واقعیت و گام به گام چه نقشه‌ای را اجرا می‌کند و چه عواقبی را ممکن است به همراه داشته باشد. همانطور که پیش‌تر هم گفتیم، به طور کلی هر فناوری‌ای که وارد عرصه استفاده عمومی می‌شود، عواقب ناخواسته‌ای را به همراه می‌آورد که ممکن است توسط طراح پیش‌بینی نشده باشد و در سیستم‌های هوشمند، به دلیل ماهیت خاصی که دارند، این عواقب ناخواسته به مراتب بیشتر و پیچیده‌تر است.

ورنر وینج[3] نیز از یک تکینگی فناورانه سخن می‌گوید که در آن، افق پیش‌بینی نامعلومی وجود دارد که ما وقایع بعد از آن را نمی‌توانیم پیش‌بینی کنیم. این در حالی است که به باور باستروم، ما تا حد معقولی می‌توانیم با اطمینان، وقایع پس از این تکینگی فناورانه را پیش‌بینی کنیم. برای مثال، با استفاده از ملاحظات عقلانیت

1. turbulent
2. Artificial General Intelligence
3. Vernor Vinge

اقتصادی، می‌توانیم رفتار عامل‌های رقیبی که از یکدیگر مستقل‌اند را تا حدی پیش‌بینی کنیم، زیرا این موجودات ساخته ما انسان‌ها هستند و ما می‌توانیم تا حد خوبی تعیین کنیم که چه ارزش‌هایی در آن‌ها جای‌دهی[1] شود. این ارزش‌ها هستند که تعیین می‌کنند جهان به چه سمتی حرکت کند. بنابراین، باستروم معتقد است که اینکه سازندگان این سیستم‌های هوشمند چه ارزش‌هایی را برای طراحی این مصنوعات انتخاب کنند و در آن‌ها جای‌دهی کنند، می‌تواند به ما قدرت پیش‌بینی رفتارهای این سیستم‌ها را تا حد نسبتا خوبی بدهد و ما را قادر به پیش‌بینی چند سناریوی احتمالی برای آینده جهانی که این مصنوعات در شکل دادن آن نقش دارند، بکند.

بنابراین، بسیار مهم است که الگوریتم‌های هوش مصنوعی که قرار است برخی از وظایف اجتماعی را بر عهده بگیرند، برای کسانی که آن‌ها را می‌سازند، پیش‌بینی‌پذیر و تحت کنترل باشند. به پیشنهاد باستروم، یکی از شیوه‌های کنترل تأثیرات سیستم‌های هوشمند بر جهان آینده، این است که چارچوبی از ارزش‌های مطلوب، موازین و سازوکارهایی مشخص برای امنیت این سیستم‌ها طراحی شود. این پیشنهاد می‌تواند به ما در طراحی سیستم‌های هوشمند به گونه‌ای که متناسب با اقتضائات بومی و دینی ما باشند، نیز کمک کند. این دستورالعمل‌ها می‌توانند معیار طراحی این مصنوعات و ارزیابی آن‌ها قرار بگیرند و مهندسان و سیاست‌گذاران، این چارچوب ارزش‌های اخلاقی را معیار فعالیت‌های خود در جهت رسیدن به هوش مصنوعی ایمن‌تر قرار دهند. اما به دلیل اینکه هوش مصنوعی ٪۱۰۰ ایمن امری غیر ممکن است، ما باید برای مواجهه با عواقب ناخواسته

1. embed

احتمالی این سیستم‌ها نیز آمادگی داشته باشیم.

۱-۶- سوگیری هوش مصنوعی

سوگیری[1] به معنای توجه بیشتر به برخی از ارزش‌ها و نادیده گرفتن برخی از ارزش‌های دیگر است. اگرچه هوش مصنوعی قادر به پردازش سریع است و کارامدی آن در برخی از موارد بسیار فراتر از توانایی انسانی است، با این حال نمی‌توان همیشه به بی‌طرفی آن مطمئن بود. نباید فراموش کنیم که سیستم‌های هوش مصنوعی توسط انسان‌هایی ساخته می‌شوند که می‌توانند دارای سوگیری و قضاوت باشند.

هوش مصنوعی امروزه به شیوه یادگیری و یادگیری عمیق و توسط داده‌های بسیار زیادی که به عنوان ورودی به آن داده می‌شود، آموزش می‌بینند. معنای هر داده‌ای وابسته به سیاق مربوط به آن داده است و وقتی این سیاق مبهم باشد، هوش مصنوعی مبتنی بر یادگیری عمیق نیز مانند مغز انسان گاهی همبستگی‌های نادرستی را می‌بینند که می‌تواند منجر به سوگیری و رفتار متعصبانه شود. دادن اطلاعات بیشتر به چنین سیستم‌هایی به منزله عینی بودن یا درست بودن اطلاعات و تصمیمات آن‌ها نیست، بلکه تصمیم هوش مصنوعی هم صرفا یک نظر است نه حقیقت محض. اگر در سوگیری تعادلی وجود نداشته باشد، جامعه فرومی‌پاشد. زیرا تصمیمات و قضاوت‌هایی علیه برخی احزاب و گروه‌ها وجود خواهد داشت.

به علاوه، امکان ویروسی شدن ماشین هوشمند (مشابه بیماری مغزی در انسان) این نگرانی را تشدید می‌کند. سیستم‌های هوش مصنوعی

1. bias

در مقابل سوگیری‌هایی که سازندگان آن‌ها خواسته یا ناخواسته در این سیستم‌ها جای می‌دهند، آسیب‌پذیرند و ممکن است این سوگیری‌ها را در تصمیمات بزرگ و کوچک خود وارد کنند. داده‌هایی که در یادگیری عمیق و یادگیری ماشینی وارد سیستم هوش مصنوعی می‌شود ممکن است دارای سوگیری‌های نژادی، جنسیتی، طبقه‌ای و ... باشد. مثلا در تصمیم‌گیری یک ماشین هوشمند در انتخاب افراد مناسب برای یک شغل، اعطای وام، تشخیص فرد مظنون، فرستادن او به زندان و عفو مشروط او سوگیری‌هایی از این دست وجود دارد. یکی از تنگناهای اخلاقی پیش روی هوش مصنوعی این است که بهتر است این سوگیری‌ها در آن حذف شود یا اینکه داده‌ها همانطور که در واقعیت دارای سوگیری هستند، به ماشین داده شوند.

بیشتر سیستم‌های هوشمند به غیر از سازندگانشان، برای دیگران غیر شفافند و ما نمی‌توانیم تعیین کنیم که چرا یک الگوریتم نتیجه، پیشنهاد یا ارزیابی خاصی را تولید می‌کند. این می‌تواند در مواردی که از ماشین‌ها به جای قاضی یا هیئت منصفه استفاده می‌کنیم مشکل‌ساز شود. زیرا ما ابتدایی‌ترین فهم را درباره نحوه کار کردن الگوریتم‌ها نداریم و تا زمانی که موجب آسیبی نشده‌اند، نمی‌توانیم بدانیم که چگونه نتایج ضعیفی تولید می‌کنند.

مشاهدات نشان می‌دهند که الگوریتم‌ها سوگیری‌هایی دارند که موجب شده احتمال ارتکاب جرم افراد سیاه‌پوست را دو برابر افراد سفیدپوست تخمین بزنند. از آنجایی که الگوریتم‌ها شفافیت ندارند، دشوار است که این سوگیری را به خطای ارزیابی نسبت دهیم یا آن را حاصل کدنویسی‌هایی بدانیم که منعکس‌کننده سوگیری‌های

سازندگانشان هستند. علاوه بر این، پیش‌بینی‌های این الگوریتم‌ها آنقدر هم دقیق نیست. پیش‌بینی الگوریتم‌ها فقط درباره ٪۲۰ افرادی که پیش‌بینی شده بود که مرتکب جرم‌های خشونت‌امیز خواهند شد، صادق بود و با ٪۸۰ دیگر به مثابه مجرم برخورد شد.

صحبت از چیزی که فاقد هرگونه سوگیری باشد، واقع بینانه نیست. برای مثال، یک جاده آسفالت هم سوگیری دارد، چرا که آسفالت برای کودکان که جمجمه نرمتری دارند، خطرناک‌تر است و آن‌ها را بیشتر تهدید می‌کند. از آنجایی که ممکن است الگوریتم‌ها سوگیری‌های انسانی را بازتولید کنند، نیاز به شفافیت و پاسخگویی در آن‌ها نیاز داریم. شرکتی آمریکایی نرم‌افزاری طراحی کرده که می‌تواند احتمال تکرار یک جرم توسط مجرم را محاسبه کند. این الگوریتم در تعداد قابل توجهی از پیش‌بینی‌هایش در مورد افراد آفریقایی-آمریکایی خطا کرده و تبعیض نژادی اعمال کرده است. بنابراین، چنین ابزارهایی باید قابل بررسی باشند و معیارهای تصمیم‌گیری آن‌ها شفاف باشد تا از بی‌عدالتی در تصمیم‌گیری جلوگیری شود.

ما باید همواره از پاسخگویی این فناوری‌ها مطمئن باشیم و از تعصب و سوگیری آن‌ها جلوگیری کنیم. این عادلانه نیست که فردی به خاطر مخاطرات احتمالی که شاید پیش بیایند یا احتمال کم ابتلای او به بیماری در آینده، از فرصت‌های شغلی فعلی خود محروم شود. هوش مصنوعی نمی‌تواند به ما مجوز عدم مسئولیت در قبال مسائل اخلاقی از این دست را بدهد. ما به عنوان انسان‌های سازنده هوش مصنوعی، موظفیم که مسئولیت طراحی الگوریتم‌ها، ارزیابی

و شفافیت معنایی آن‌ها را بر عهده بگیریم و همواره بر کار هوش مصنوعی نظارت اخلاقی داشته باشیم.

۱-۷- موقعیت‌مندی و بدن‌مندی هوش و ادراک

سیستم‌های هوش مصنوعی عموماً یک دوره یادگیری دارند که طی آن «یاد می‌گیرند» الگوهای صحیح را شناسایی کنند و متناسب با ورودی‌های خود عمل کنند. وقتی یک سیستم به طور کامل آموزش دید، می‌تواند وارد مرحله آزمون شود که طی آن در معرض موقعیت‌ها و نمونه‌های مختلف قرار می‌گیرد و عملکردش بررسی می‌شود. بدیهی است که مرحله آموزش نمی‌تواند همه موقعیت‌هایی را که یک سیستم ممکن است در جهان واقعی با آن‌ها مواجه شود، پوشش دهد. ماشین هوشمند همواره ممکن است در موقعیت‌های پیش‌بینی‌نشده و منحصربفردی قرار بگیرد که برای آن‌ها برنامه‌ریزی نشده و ممکن است رفتارهای نوظهوری از خود بروز دهد.

هر موقعیتی که پیش می‌آید، منحصربفرد است و پیچیدگی‌ها ویژگی‌های خاص خود را دارد و در نتیجه، نمی‌توان از پیش همه شرایط مربوط به آن موقعیت را دانست و پیش‌بینی کرد. این همان تذکری است که هیوبرت درایفوس، فیلسوف برجسته، تحت عنوان موقعیت‌مندی یا سیاق‌مندی ادراک از آن یاد می‌کند. پیچیدگی‌های یک سیاق و به طور کلی، همه وجوه واقعیت را نمی‌توان به زبان گزاره‌ها و به نظریه تقلیل داد و همه معرفت را نمی‌توان صوری کرد. در نتیجه، یک سیستم هوشمند ممکن است با موقعیت‌های پیش‌بینی‌نشده و جدیدی مواجه شود که برای آن‌ها آماده نشده

است و این احتمال وجود دارد که رفتاری که در این موقعیت‌ها از خود نشان می‌دهد، منجر به بروز عواقب ناخواسته‌ای شود که طراحان این سیستم‌ها برای آن‌ها برنامه‌ریزی نکرده‌اند. در حالی که ادراک انسان سیاق‌مند است و با یک ابژه (موضوع مورد شناخت)، به مثابه یک کل مواجه می‌شود. انسان، پس‌زمینه و سیاقی که ابژه در آن قرار دارد را نیز در ادراک خود از یک موقعیت وارد می‌کند و بر اساس نگاهی کل‌گرایانه به موقعیت، درباره آن تصمیم می‌گیرد.

در همین راستا، مبحثی در فلسفه ذهن مطرح می‌شود تحت عنوان مسئله چارچوب[1]. این مسئله مربوط به قدرت تشخیص امر مربوط از امر نامربوط در هوش مصنوعی است. جهان واقعی و غیر ایزوله، پویاست و بی‌نهایت امکان برای تغییر در آن وجود دارد. بی‌شمار گزاره واقعی نیز می‌توانند به تغییر رخ داده و پاسخ معقول به آن، مربوط باشند. هر برنامه هوش مصنوعی که بخواهد بر مبنای اصول منطقی در جهان واقعی کار کند، باید راهی برای حل مسئله چارچوب، یعنی تفکیک امور مرتبط از امور نامرتبط با وضعیت فعلی روبات، بیابد.

مسئله چارچوب، مسئله امکان ساخت روبات‌هایی است که قادر باشند به درستی تشخیص دهند چه اطلاعات و چه استنتاجی در یک موقعیت معین، مناسب است و باید به کار گرفته شود. برای مثال، اگر روباتی که برای ریختن چای برنامه‌ریزی شده است را در نظر بگیریم، موقعیت فعلی فنجان در قالب یک جمله در کنار سایر گزاره‌های نشان‌دهنده موقعیت محیط، در حافظه این روبات موجود است. بعد از این که روبات فنجان را برداشت، بخشی از اطلاعات

1. Frame Problem

باید به‌روز شوند که اولین داده، موقعیت فعلی فنجان است. اما چه داده‌های دیگری نیاز به تغییر دارند؟ دمای محیط، موقعیت سماور و... تغییری نکرده اند، اما موقعیت قاشقی که داخل فنجان بوده نیز باید به روز شود. مشکل معرفت شناختی که در این میان پیش می‌آید این است که روبات چطور باید داده‌هایی که احتمال تغییر دارند را تشخیص دهد؟ هرچه روبات هوشمندتر و دایره فعالیت‌هایش گسترده‌تر باشد، داده‌های درون حافظه و شرایط پیرامونی آن نیز پیچیده‌تر و تشخیص گزاره‌های مرتبط دشوارتر می شود.

مبحث بسیار مهم دیگری که امروزه در سیستم‌های هوشمند پیگیری می‌شود، رویکرد بدن‌مندانه به هوش است. ادعای رویکرد بدن‌مندی[1] این است که فرایندهای شناختی، به تجاربی که از داشتن یک بدن با ظرفیت‌های حسی حرکتی ویژه حاصل می شود، بستگی دارد؛ بدنی که به طور مداوم در کنش متقابل با جهان اطرافش است و به واسطه ویژگی‌های منحصربفردی که دارد، جهان را به نحوی خاص ادراک می‌کند.

مضمون مشترک رهیافت‌های مختلف در شناخت بدن‌مند، بازگرداندن و احیای بدن در کانون مطالعات علوم شناختی، تاکید بر تعامل میان عامل و محیط، تاکید بر اهمیت تجارب و تفاوت‌های فردی و کاربرد یک دیدگاه کل‌گرایانه در بررسی مساله شناخت است. لذا در عوض نگاه به ذهن به عنوان آیینه جهان، که در پارادایم محاسبه‌گرایی[2] پیگیری می‌شود، در این دیدگاه، ادراک حسی، صرفا تاثیری ذهنی یا روانی بر مغز انسان نیست، بلکه موقعیت هوشمندانه بدن در عالم است و تفکر و تصمیم‌گیری، فرایندی مجزا از کنش بدن‌مند نیست.

1. embodiment
2. computationalism

وجه دیگر بدن‌مندی، نسبت دادن نوعی از دربارگی[1] به سمت ابژه‌ها برای بدن است. لذا بدن، مستقل از ذهن و فکر انسان، سطوحی از هوشمندی را دارد. امتداد این رویکرد شناختی در هوش مصنوعی، ساخت روبات‌هایی است که دارای بدنی مشابه انسان هستند و با جهان تعامل می‌کنند.

بنابراین، هوش انسانی بسیار متأثر از موقعیت فرد در جهان و ساختار بدنی اوست و این پیچیدگی‌ها، در عمل به انسان کمک می‌کنند که در هر موقعیت یکتا، چگونه رفتاری اخلاقی و مناسب از خود بروز دهد و خطاهای فاحشی از او سر نزند. در حالی که سیستم‌های هوشمند ممکن است دچار اشتباهات و فریب‌هایی شوند که انسان‌ها در حالت عادی احتمالاً مرتکب آن‌ها نمی‌شوند. مثلا، یک الگوی تصادفی از نقاط ممکن است باعث شود ماشین بدون در نظر گرفتن سیاقی که این نقاط در آن قرار دارند، چیزهایی را «ببیند» که وجود ندارند و بر اساس آن، تصمیم‌های اشتباهی بگیرد. سیستم‌های هوش مصنوعی را می‌توان مانند «غولی در چراغ جادو» در نظر گرفت که اگرچه برخی از رویاها و آرزوهای ما را برآورده می‌کنند، ولی ممکن است عواقب ناخواسته سنگینی هم به همراه داشته باشند.

این مسئله به این معنا نیست که یک ماشین ممکن است قصد سوئی داشته باشد، بلکه این نکته را یادآور می‌شود که ماشین محتمل است که دچار نوعی فهم ناقص از کل بافتاری شده باشد که مقصودی در آن محقق شده است. مثلا فرض کنید از یک سیستم هوش مصنوعی خواسته شده که سرطان را در جهان ریشه‌کن کند.

1. intentionality

این سیستم پس از محاسبه بسیار، ممکن است فرمولی را ارائه کند که سرطان را با کشتن همه افراد کره زمین از بین می‌برد. این کامپیوتر به هدف «ریشه‌کن شدن سرطان» به نحوی بسیار کارامد دست یافته، اما نه به نحوی که مقصود ما بوده است (Boss-mann, 2016). بنابراین، ماشین ممکن است در تصمیم‌گیری خود، کل‌گرایانه و ناظر به سیاق عمل نکند و تصمیماتش عواقب جبران‌ناپذیری را برای ما به بار بیاورد. اگرچه ممکن است این مثال دور از ذهن به نظر برسد، اما خطاهایی از این دست امروزه هم توسط ماشین‌ها صورت می‌گیرد.

۱-۸- سوء استفاده از خطاهای هوش مصنوعی

هوش مصنوعی فناوری قدرتمندی است که برای کسانی که در جستجوی قدرت هستند، مانند مجرمان سازمان‌یافته، گروه‌های افراطی و تروریست‌ها جذابیت بالایی دارد. از آنجایی که امکان محافظت کامل از این فناوری در مقابل سوءاستفاده و هک شدن وجود ندارد، این موضوع نیز می‌تواند نگران‌کننده باشد. اگر ما قرار است برای انجام برخی از امورمان به هوش مصنوعی تکیه کنیم، باید بیشترین اطمینان را حاصل کنیم از اینکه ماشین، همانطور که برنامه‌ریزی شده است، عمل می‌کند و انسان‌ها نمی‌توانند آن را برای مقاصد سوء خود به کار ببرند.

الگوریتم‌های هوش مصنوعی باید در مقابل دستکاری و سوء استفاده‌های احتمالی بسیار مقاوم باشند. برای مثال، یک سیستم بینایی ماشینی که از آن برای اسکن چمدان‌ها برای یافتن بمب استفاده می‌شود، باید در

مقابل دشمنان انسانی مقاوم باشد و بتواند نقص‌های قابل سوء استفاده در الگوریتم‌ها را بیابد. بنابراین، سیستم هوش مصنوعی باید دارای امنیت اطلاعات باشد و در مقابل دستکاری، مقاوم و قابل پیش‌بینی باشد. این موضوع، وظیفه سنگین‌تری را بر دوش طراحان و برنامه‌نویسان می‌گذارد تا در برنامه‌ریزی‌ها و طراحی‌های خود، این سوءاستفاده‌های احتمالی را شناسایی کنند و نگاه دقیق‌تر و دوراندیشانه‌تری به آینده داشته باشند.

هر چقدر یک فناوری قوی‌تر می‌شود، در راه اهداف خیر و شر بیشتری می‌تواند به کار گرفته شود. این موضوع منحصر به سلاح‌های خودمختار یا روبات‌هایی که جایگزین سربازها در جنگ می‌شوند، نیست. بلکه در مورد همه سیستم‌های هوش مصنوعی که در صورت کاربرد شریرانه قابلیت آسیب‌رسانی دارند، نیز صادق است، چرا که این نزاع‌ها صرفا در میدان نبرد اتفاق نمی‌افتند. بنابراین، اهمیت موضوعاتی مانند امنیت سایبری روز به روز در حال افزایش است، زیرا ما با سیستم‌هایی روبرو هستیم که به مراتب سریع‌تر و تواناتر از ما هستند و وقایعی که پس از ورود آن‌ها به جوامع می‌توانند اتفاق بیفتند، بسیار متنوع و بدیع هستند.

۱-۹- هوش مصنوعی ضعیف و قوی

در پی پیشرفت‌های گسترده در علوم کامپیوتر و ساخت روبات‌های هوشمندِ مبتنی بر نظریه محاسباتی ذهن، این بحث میان اندیشمندان مطرح شد که آیا این پیشرفت‌ها به جایی می‌رسد که سیستم هوشمند در تعقل، هوشمندی و کسب علم جایگزین هوش انسانی شود؟

چالش پیرامون شبیه‌سازی ذهن انسان توسط هوش مصنوعی توسط جان سرل[1] با دوگانه‌ی هوش مصنوعی ضعیف[2] و هوش مصنوعی قوی[3] صورت‌بندی شد. قائلان به هوش مصنوعی ضعیف، معتقدند سیستم‌های هوشمند ابزارند و تنها می‌توان بعضی از کاربردهای محاسباتی ذهن انسان را توسط آن‌ها پیاده‌سازی کرد. در مقابل، طرفداران هوش مصنوعی قوی قائلند که رایانه‌ها توانسته‌اند یا خواهند توانست همه‌ی توانایی‌های ذهن انسان از قبیل آگاهی، اراده، تفکر، فهم معنا و زبان و ... را داشته باشند.

با هوش مصنوعی قوی، سیستم‌های هوشمند مشابه انسان، تنها یک ابزار برای مطالعه ذهن نیستند، بلکه سیستم‌هایی که به درستی برنامه‌ریزی شده اند، همان ذهن انسانی هستند. بدین معنا می‌توان گفت که سیستمی که برنامه درستی را اجرا می‌کند، می‌فهمد و حالت‌های شناختی[4] دارد. جان سرل برای رد امکان هوش مصنوعی قوی، آزمون اتاق چینی را طرح کرد که مناقشات بسیاری در فلسفه هوش مصنوعی در پی داشت. روشن است که هوش مصنوعی به صورت ضعیفِ آن محقق شده است و نزاعی در این باره نیست، اما امکان تحقق هوش مصنوعی قوی محل بحث است و مناقشاتی را درباره چیستی و ماهیت انسان و تفاوت او با سایر موجودات برانگیخته است. مباحثی از این دست هستند که موجب شده‌اند شأن اخلاقی سیستم‌های هوشمند برای ما مسئله‌ساز شوند.

۱-۱۰- پدیده انفجار هوش و تکینگی

عرف عام تایید می‌کند که دلیل اصلی تسلط انسان‌ها بر موجودات دیگر،

1. John Searl
2. Weak Artificial Intelligence
3. Strong Artificial Intelligence
4. Cognitive state

هوش، استعدادهای منحصربفرد، عقل و اختیار آنهاست. ما بر حیوانات بزرگ‌تر، سریع‌تر و قوی‌تر غلبه داریم، زیرا می‌توانیم ابزارهایی را برای کنترل آن‌ها بسازیم: هم ابزارهای فیزیکی مانند قفس و سلاح، و هم ابزارهای شناختی مانند تربیت و شرطی‌سازی. اما آیا روزی فرا می‌رسد که هوش مصنوعی بتواند از تسلط بشر رها شود و همین برتری را نسبت به او داشته باشد و او را تحت سلطه خود درآورد؟ اگر ماشینی به قدر کافی پیشرفته باشد، ممکن است دیگر متوقف کردن یا کنترل کردن آن از اختیار ما خارج شود، زیرا چنین ماشینی می‌تواند اینگونه رفتارهای ما را پیش‌بینی کند و در مقابل آن‌ها از خود دفاع کند. برخی از متخصصان هوش مصنوعی این پدیده را «تکینگی[1]» می‌نامند. تکنیگی یعنی نقطه‌ای در زمان که انسان‌ها دیگر تنها موجودات هوشمند روی زمین نیستند و موجودات هوشمندی به وجود می‌آیند که به اندازه انسان یا هوشمندتر از او باشند.

سیستم‌های هوش مصنوعی که به قدر کافی هوشمند شده باشند و بتوانند طراحی خود را درک و تحلیل کنند، می‌توانند خود را بازطراحی کنند یا سیستم‌های جایگزینی بسازند، و بنابراین می‌توانند کاری کنند که هوشمندتر هم بشوند، و همینطور در چرخه‌ای بازخوردی این روند می‌تواند ادامه یابد. فراهوش، یکی از چندین مخاطره جدی مطرح‌شده توسط باستروم (۲۰۰۲) است؛ مخاطره‌ای «که در آن یک خروجی ناسازگار و پیش‌بینی نشده می‌تواند یا زندگی موجود هوشمند نشأت‌گرفته از زمین (یعنی انسان) را نابود کند یا توانایی او را به طور دائمی و موثری محدود کند.» در حالی که هوش مصنوعی می‌تواند در جهتی توسعه پیدا کند که خروجی مثبت فراهوش

1. singularity

این باشد که زندگی هوشمند نشأت‌گرفته از زمین (انسان) را حفظ کند و به محقق شدن توانایی‌های او کمک کند. بنابراین، یکی از راه‌حل‌های پیشنهاد شده برای حل مسئله تکنیکی، ساختن «هوش مصنوعی دوستانه[1]» است که طبق آن، روبات‌ها باید به نحوی طراحی شوند که به طور ذاتی دوست‌دار انسان‌ها باشند و نتوانند زندگی او را تهدید کنند.

۱-۱۱- هوش مصنوعی و بحران شغلی

با توسعه روبات‌ها و فناوری‌های مبتنی بر هوش مصنوعی، این نگرانی به وجود آمده است که ممکن است مردم شغلشان را به خاطر این فناوری‌ها از دست بدهند یا تغییراتی نامتوازن در وضعیت دستمزدها به وجود بیاید. این نگرانی فقط مربوط به ظهور فناوری‌های هوش مصنوعی نیست. از بدو پیدایش فناوری، افراد همیشه به خاطر ظهور فناوری‌هایی که جایگزین کارکردهای انسانی شده است، شغلشان را از دست داده‌اند. فناوری اغلب جایگزین شغل افرادی می‌شود که کارهای نوعا انسانی را انجام می‌دهند و این کارها را با کارامدی و سرعت بالاتری انجام می‌دهند و معمولا، صرفه اقتصادی استفاده از ماشین‌ها بیش از استخدام نیروی کار انسانی است.

با ابداع شیوه‌های جدید برای اتوماتیک شدن شغل‌ها، فرصتی برای پذیرفتن نقش‌های پیچیده‌تر توسط انسان‌ها پیش آمد و کارهای فیزیکی که پیش‌تر عمدتا توسط انسان انجام می‌شد، به ماشین واگذار شد. در نتیجه این فرایندها، انسان بیشتر به سمت کارهای شناختی سوق پیدا کرد که ویژگی آن‌ها استراتژیک و

1. Friendly Artificial Intelligence

مدیریتی بودن است. یک نمونه، باربری صنعتی است که در حال حاضر برای میلیون‌ها نفر در آمریکا اشتغال‌زایی کرده است. اگر باربری اتوماتیک که شرکت تسلا (متعلق به ایلان ماسک) وعده آن را داده است، در دهه آینده در دسترس عمومی قرار گیرد چه اتفاقی برای افرادی رخ خواهد داد که به این مشاغل اشتغال دارند؟ وجه دیگر مسئله این است که اگر ریسک پایین‌تر تصادفات را معیار بگیریم و بخواهیم بر این اساس تصمیم‌گیری کنیم، باربری اتوماتیک انتخابی اخلاقی‌تر به نظر می‌رسد.

صنعت باربری صرفا یک نمونه از صدها نمونه شغل‌های دیگری است که ممکن است با سیستم‌های هوش مصنوعی جایگزین شوند و نظام اشتغال فعلی را به کلی دگرگون کنند. در ده الی بیست سال آینده، حدودا نیمی از مشاغلی که امروز وجود دارند در معرض تهدید الگوریتم‌ها خواهند بود و ۴۰٪ از پانصد شرکت برتر فعلی جهان در دهه آینده از بین خواهند رفت. (Helbing, et al., 2019) با این حال، بسیاری از متخصصان معتقدند که هوش مصنوعی لزوما منجر به بحران اشتغال نخواهد شد، بلکه آرایش و انواع مشاغل را تغییر می‌دهد و موجب به وجود آمدن مشاغل جدیدی متناسب با اقتضائات خود خواهد شد و باز هم گزینه‌های مختلف اشتغال برای افراد در دسترس خواهد بود.

باید توجه داشت که فناوری وقتی در جامعه معرفی و به کار گرفته شود، دیگر شرایط به عقب برگشت نمی‌کند، زیرا بخشی از سیستم می‌شود و نظام‌ها، نهادها و ساختارهای مادی را متناسب با خود تغییر می‌دهد، به گونه‌ای که از بین بردن این ساختارها کاری بسیار

دشوار خواهد بود. ما در ساختارها جایگاه مهمی به فناوری داده‌ایم. برای مثال، وقتی کارهایی نظیر پرداخت پول یا تشخیص مشتری و حرف زدن مودبانه با او را بر عهده ماشین گذاشتیم، افرادی که قبلا چنین کارهایی را انجام می‌دادند، شغلشان را از دست می‌دهند. مدافعان هوش مصنوعی می‌گویند که این افراد همچنان آزادند که مشاغل دیگری را انتخاب کنند؛ مثلا مشاغل جدیدی که توسط هوش مصنوعی خلق شده است، اما اینکه چنین پیشنهادی چه مقدار قابلیت عملی شدن را دارد، محل بحث است و باید به راهکارهای عملیاتی شدن آن اندیشید.

بخش دوم

بخش دوم

برخی چالش‌های اخلاقی هوش مصنوعی در حوزه‌های خاص

در این فصل، برخی از کاربردهای انضمامی هوش مصنوعی در سطح سیاسی، اجتماعی و در برخی از حرفه‌های خاص را بررسی می‌کنیم و تأثیر چالش‌های اخلاقی که در فصل پیش مطرح کردیم را در کاربرد عملی سیستم‌های هوشمند و در زندگی روزمره، به نحوی ملموس‌تر از نظر می‌گذرانیم. به طور خاص، به دو حوزه صنایع نظامی و حرفه‌های حقوقی به نحوی مفصل‌تر می‌پردازیم، چرا که کاربرد هوش مصنوعی در آن‌ها تا حد خوبی پیش رفته است و حساسیت تصمیم‌گیری و قضاوت در این حوزه‌ها بالاست، و پرسش‌های اخلاقی که استفاده از سیستم‌های هوشمند، پیشِ روی متخصصین این دو حوزه قرار داده است را مرور می‌کنیم. سپس، به چالش‌های اخلاقی دیگری که کاربرد هوش مصنوعی در حوزه‌هایی مانند عدالت اقتصادی، تعاملات اجتماعی و اشتغال ایجاد کرده است، اشاره می‌کنیم.

۲-۱- هوش مصنوعی در صنایع نظامی

سلاح‌های هوشمند نظامی، واقعیتِ آینده ما هستند. اگرچه هنوز

انسان‌ها در زنجیره تصمیم‌گیری برای اِعمال نیروی کشنده حضور دارند، ولیکن ما در حال پیشروی به سمتی هستیم که تصمیم‌ها به طور کامل به روبات‌ها واگذار شوند. بنابراین، زمان آن فرا رسیده که با چالش‌های احتمالی کاربرد این سیستم‌ها روبرو شویم. امروزه بیش از ۵۰ کشور در جهان در حال توسعه فناوری روبات‌های نظامی هستند. در همه روبات‌هایی که در حال حاضر به کار برده می‌شوند، اپراتوری وجود دارد که پرواز آن‌ها را کنترل می‌کند و نیروی کشنده آن‌ها را به کار می‌گیرد. اما این شیوه به سرعت در حال تغییر است و همه (از جمله آمریکا، چین، روسیه، اسرائیل و انگلستان) به دنبال ساختن و استفاده از روبات‌های خودمختار در میدان جنگ هستند. هدف نهایی، دستیابی به شبکه‌ای از روبات‌های زمینی، دریایی و هوایی است که بتوانند با کمک یکدیگر و به نحوی خودمختار و بدون مداخله انسانی، هدف خود را ردیابی و نابود کنند. (Sharkey, 2012)

استفاده از هوش مصنوعی در امور نظامی و جنگ‌ها اگرچه طرفدارانی جدی در عرصه نظامی و سیاسی دارد، اما چالش‌های اخلاقی ناشی از آن، خصوصا وقتی پای بحث خودمختاری آن مطرح می‌شود و روبات‌ها قرار است تصمیم بگیرند چه کسی باید کشته شود، موجب شده منتقدان بسیاری علیه ساخت و کاربرد این سلاح‌ها سخن بگویند. در حال حاضر، بسیاری از کشورها روی سلاح‌های هوشمند سرمایه‌گذاری کرده‌اند که این با مخالفت افراد سرشناسی همچون استیون هاوکینگ[1] (فیزیکدان)، جان تالین[2] (موسس اسکایپ)، نوام چامسکی[3] (زبان‌شناس) و مکس تگمارک[4] (کیهان‌شناس) و ... روبرو

1. Stephen Hawking
2. Jaan Tallinn
3. Noam Chomsky
4. Max Tegmark

شده است، زیرا این سلاح‌ها می‌توانند منجر به فجایع آنی و عظیمی شوند.

در کنفرانسی که در سال ۲۰۰۹ توسط انجمن توسعه هوش مصنوعی برگزار شد، مباحثاتی پیرامون تأثیر بالقوه روبات‌ها و کامپیوترها و امکان خودکفایی آن‌ها در تصمیم‌گیری مطرح شد که طبق آن، این سطح از خودمختاری هوش مصنوعی او را قادر می‌سازد که به نحوی مستقل، هدف‌هایی را برای حمله انتخاب کند. بنابراین، سلاح‌های هوشمند در موقعیت تصمیم‌گیری درباره اینکه در چه زمانی، چه کسی را بکشند هستند و این یکی از مهمترین و حساس‌ترین چالش‌های پیش روی هوش مصنوعی است. در موضوع «روبات‌های جنگی و سلاح‌های خودمختار»، علاوه بر آرکین که به دنبال کنترل اخلاقی الگوریتمی روبات‌هاست، افراد دیگری نیز به ارائه نظرات خود پرداخته‌اند. کلوس بر مبنای اصول سودگرایی، پیشنهاد خود را برای کنترل روبات‌ها ارائه می‌دهد. لوکاس و کامستوک[1] نیز نوع دیگری از سودگرایی را مبنای کار خود قرار می‌دهند. اندرسون و اندرسون[2] از سوی دیگر، در مخالفت با رویکردهای سودگرایانه، مفهوم «وظایف بدیهی» را به عنوان اساس کار خود انتخاب می‌کنند که توضیح این پیشنهادات را به پژوهش تفصیلی در این باره موکول می‌کنیم.

روبات‌های خودمختار در حال حاضر نمی‌توانند میان افراد نظامی، غیر نظامی، مهاجمان، تماشاچیان، افراد خطرناک و بی‌خطر تمایز قائل شوند. همچنین، تشخیص افراد زخمی و افراد تسلیم‌شده برای آن‌ها مقدور نیست. روبات‌ها برای تشخیص این تمایزها با سه نوع چالش مواجه‌اند: ۱. آن‌ها به سیستم‌های پردازش حسی و تصویری کافی

1. Lucas and Comstock
2. Anderson and Anderson

برای تمایز نظامیان از غیر نظامیان دسترسی ندارند. امکانات روبات محدود به دوربین، حسگرهای فروسرخ، ردیاب صوتی، لیزر، حسگر دما و ... است. ۲. تعریفی مشخص و قابل ترجمه به زبان کد و الگوریتم برای افراد غیر نظامی وجود ندارد. قوانین جنگ نیز نمی‌تواند اطلاعات لازم را برای یک روبات فراهم کند. پیروی از قرارداد ۱۹۴۹ ژنو مستلزم استفاده از عرف عام است. ۳. حتی اگر ماشین‌ها توانایی تشخیص افراد نظامی را داشتند، باز هم فاقد آگاهی از صحنه نبرد و توانایی استدلال عرفی درباره آن بودند تا بتوانند بر اساس آن تصمیم‌گیری کنند.

در حال حاضر، نمی‌توان به ماشین‌ها در تشخیص انسان‌ها اعتماد کرد و شاهدی هم علیه این مدعا وجود ندارد. روبات‌ها آگاهی موقعیت‌مند ندارند و نمی‌توانند تصمیماتی متناسب با موقعیت بگیرند. معیاری که برای مناسب بودن عملکرد این روبات‌ها در نظر گرفته می‌شود، کمینه کردن آسیب جانبی از طریق انتخاب مناسب‌ترین سلاح و نشانه‌گیری درست آن است. اما مسئله اصلی در عملکرد متناسب این است که تصمیم درستی درباره اینکه آیا در یک موقعیت خاص باید از نیروی کشنده استفاده کرد یا خیر گرفته شود. ماشین حس ششم ندارد و جنگ هم ماجرایی سیاه و سفید نیست که بتوان چنین تصمیماتی را فارغ از اقتضائات خاص یک موقعیت گرفت. هیچ شیوه‌ای نیز برای فهم نحوه تفسیر انسان از قوانین جنگ در موقعیت‌های خاص وجود ندارد و همچنین، روشی برای حل ابهامات قوانین متعارض در موقعیت‌های جدید در دست نیست تا بتوان آن‌ها را به شکل کد برنامه‌نویسی کرد.

نکته دیگری که در طراحی سلاح‌های هوشمند باید مورد توجه قرار بگیرد این است که طبق استدلال لاتور و وینوگراد، ساخت یک مصنوع همواره نیازمند یک فرایند ترجمه است. معماری و توانایی‌های یک ماشین با یک انسان متفاوت است و برای طراحی یک ماشین باید به این تفاوت‌ها دقت کرد. مثلا سربازی که مجروح شده، به دنبال پناهگاهی برای محافظت از خود می‌گردد. اما ماشین ترسی از جراحت ندارد و می‌تواند به کمک خود دشمن، مکان او را شناسایی کند و به او شلیک متقابل کند.

وینوگراد توجه ما را به تغییراتی جلب می‌کند که ممکن است در ترجمه کردن معنای کلمات روزمره و وابسته به سیاق انسانی به زبان الگوریتم و مستقل از سیاق به وجود بیاید. فرایند ترجمه باعث تغییری جدی در معنای کلمات و مفاهیمی می‌شود که حامل قوانین جنگ و قواعد درگیری هستند. به علاوه، این قوانین گاهی در معرض پیشنهادها و نقدهای عمومی بوده‌اند، در حالی که شکل الگوریتمی جدید آن‌ها که ترجمه وفاداری هم نیست، پشت درهای بسته و به صورت محرمانه تولید می‌شود و شفافیت ندارند. این کدها «کدهای بسته» هستند. یعنی عملکرد آن‌ها شفاف نیست و نمی‌توان حدس زد که چگونه عمل می‌کنند و همین عدم شفافیت ممکن است منجر به خسارات جانی بیشتری شود.

مسئله دیگر، پاسخگویی[1] است. در حال حاضر، اتفاق آرا بر این است که روبات دارای عاملیت اخلاقی نیست و نمی‌تواند برای اعمالش بازخواست شود. عاملیت اخلاقی می‌تواند به فرمانده یا برنامه‌نویس یا شرکت سازنده یا سیاست‌گذاری که اجازه استفاده از روبات

1. accountability

را داده است، نسبت داده شود. همانطور که گفتیم، برخی معتقدند که برای اینکه یک موجود، عاملی اخلاقی باشد، وجود احساسات لازم است. زیرا سیستم باید بتواند رفتار خود را به صورت «تابعی عاطفی» از احساساتی مانند گناه، غم و ... تغییر دهد و متناسب با موقعیت عمل کند. همچنین، باید توجه داشت که اگر عاملیت اخلاقی همه عوامل مؤثر در فرایند استفاده از نیروی کشنده مشخص نباشد، انسان‌های غیر نظامی بیشتری کشته خواهند شد. (Sharkey, 2012)

قانع کردن قانون‌گذاران، سیاست‌گذاران و نظامیان درباره افزایش کاربرد روبات‌های خودمختار به این وسیله که چنین روبات‌هایی می‌توانند حالت‌های عاطفی مانند گناه و دلسوزی داشته باشند و در نتیجه استدلال اخلاقی کنند، کار چندان دشواری نیست. زیرا می‌توان ادعا کرد که فناوری می‌تواند به گونه‌ای طراحی شود که چالش‌های اخلاقی را حل کند و از انسان‌ها هم انسانی‌ترین عمل کند. اما برخی از متخصصان اخلاق سلاح‌های هوشمند (Sharkey, 2012) معتقدند که این نوعی دور زدن مسئله است. همیشه تکیه کردن به فناوری نمی‌تواند مشکلات بشر را حل کند. اگر سربازان و نظامیان در جنگ‌ها غیر اخلاقی عمل می‌کنند، چاره کار روی آوردن به روبات و فناوری نیست. بلکه می‌توان قدرت استدلال اخلاقی را در آن‌ها به قدری تقویت کرد که بتوانند در موقعیت‌های دشوار در جنگ نیز اخلاقی عمل کنند.

اگرچه روبات‌ها خستگی و ترس ندارند و از روی عصبانیت و ناامیدی تصمیم نمی‌گیرند، اما نمی‌توان ادعا کرد که روبات در برخی از مواقع،

انسانی‌تر از انسان‌ها رفتار می‌کند. زیرا صحبت از انسانی بودن یک موجود بی‌جان معنا ندارد. بلکه بهتر است بگوییم که انسان از فناوری به نحوی انسانی استفاده می‌کند. روبات‌های فعلی، در بهترین حالت صرفاً می‌توانند حامل ارزش‌های انسانی باشند. به باور شارکی، ماشین‌های خودمختار به جای اینکه جنگ را انسانی‌تر و اخلاقی‌تر کنند، به انسان‌زدایی از آن می‌افزایند. ما باید به جای این کار، مطمئن شویم که انسان‌ها تصمیم اخلاقی می‌گیرند و روی سلاح‌های کشنده کنترل دارند.

پاسخ به پرسش‌های اخلاقی مانند اینکه «آیا باید سرباز دشمن را کشت یا خیر» مستلزم داشتن بینش‌هایی درباره امور مختلف مثل مرگ و زندگی است تا بر اساس آن بتوان تشخیص داد که آیا این سرباز سزاوار مرگ هست یا خیر. باید تأملاتی داشت درباره هدف این جنگ، درباره توجیه اخلاقی استفاده از نیروی کشنده در این جنگ، درباره درستیِ اخلاقی هدف اپراتور ماشین، و همچنین بدیل‌های دیگر برای استفاده از نیروی کشنده. هیچ ماشینی در حال حاضر امکان پاسخ به چنین پرسش‌هایی را ندارد و ما به عنوان فیلسوفان، اخلاق‌دانان و طراحان باید ابتدا این پرسش‌ها را شناسایی کنیم و پاسخی برای آن‌ها بیابیم تا بتوانیم روبات‌های هوشمند اخلاقی تولید کنیم.

۲-۲- هوش مصنوعی در حرفه‌های حقوقی

یکی از کاربردهای دیگر هوش مصنوعی که استقبال از آن روز به روز در حال افزایش است، به حرفه‌های حقوقی (Artificial Legal

Intelligence یا ALI) ارتباط پیدا می‌کند. هوش مصنوعی در این حرفه‌ها می‌تواند به مثابه سیستمی باشد که قابلیت مشاوره حقوقی تخصصی یا تصمیم‌گیری دارد. هوش مصنوعی همچنین می‌تواند در برخی از فرایندهای قضاوتی یا غیر قضاوتی موجود ادغام شود. به نظر می‌رسد این فناوری‌ها تاثیر مهمی روی بخش قضایی داشته باشند و بسیاری از آرایش‌های شغلی فعلی در حرفه‌های مربوط به حقوق، دیگر در ۲۰ سال آینده وجود نداشته باشند و با هوش مصنوعی جایگزین شوند. این کاربرد از هوش مصنوعی نیز موجب مطرح شدن پرسش‌هایی از این قبیل شده است که آیا جایگزین شدن ماشین‌های هوشمند در فرایندهای تحلیل و تصمیم‌گیری، می‌تواند موجب جایگزینی نقش وکلا یا قاضی‌ها شود؟ قضاوت به کمک هوش مصنوعی در ۳۰ سال آینده با چه تغییراتی مواجه خواهد شد؟ آیا قضاوت دارای ویژگی‌ها یا بخش‌هایی هست که تضمین کند این حرفه همیشه یک فعالیت انسانی خواهد بود؟

هوش مصنوعی امروزه در تحلیل اسناد و داده‌ها در فرایندهای اکتشاف حقوقی مورد استفاده قرار می‌گیرد، چون می‌تواند در زمانی کوتاه‌تر و با هزینه‌ای بسیار کمتر از انسان‌ها میلیون‌ها واژه را تحلیل کند. این کار می‌تواند موجب اتوماتیک شدن بسیاری از فرایندها شود. علاوه بر این، می‌تواند روی نتایج پرونده‌های واقعی نیز تاثیرگذار باشد. هوش مصنوعی برای جهانِ تصمیم‌های حقوقی نویدهایی به ارمغان می‌آورد. در کانادا، تیم پروفسور رندی گوبل[1] در دپارتمان علوم کامپیوتر دانشگاه آلبرتا به همراه گروهی از محققان ژاپنی الگوریتمی را طراحی کرده‌اند که می‌تواند شواهد حقوقی متناقض

1. Randy Goebel

را بسنجد، در پرونده‌ها حکم ارائه بدهد، و نتایج محاکمه‌های آتی را پیش‌بینی کند. هدف سازندگان این سیستم‌ها این است که با استفاده از ماشین‌ها به انسان‌ها کمک کنیم که تصمیمات حقوقی بهتری بگیرند.

نقش‌هایی که در سیستم قضاوت وجود دارند مدام در حال تغییرند و فناوری جنبه‌های مختلف این سیستم را تغییر داده است. کدگذاری پیشگویانه، تحلیل پیشگویانه و یادگیری ماشین از جمله روش‌هایی هستند که در هوش مصنوعی حقوقی به کار می‌روند. برای مثال، کدگذاری پیشگویانه به منظور بررسی احتمال بازگشت به جرم توسط مجرم و همچنین، کمک به تصمیم‌گیری قاضی درباره ارائه حکم به کار می‌رود.

به ادعای پروفسور نیکولاس آلتراس، محققان انتظار ندارند که هوش مصنوعی به طور کامل جایگزین قاضی‌ها و وکیل‌ها در دادگاه شود، اما بسیار محتمل است که ابزارهای هوشمند به آن‌ها در قانون‌گذاری کمک کنند. قاضی با استفاده از چنین برنامه‌هایی می‌توانند برای تحلیل یک پرونده، از پرونده‌های مشابه استفاده کند و مشابهت‌ها و تفاوت‌های میان آن‌ها را پیدا کند و حکم هوش مصنوعی درباره پرونده را بفهمد.

اگرچه، الگوریتم‌های ارائه حکم با سوگیری‌های پنهان بسیار مورد توجه قرار گرفته‌اند، اما عده‌ای معتقدند که استفاده از هوش مصنوعی می‌تواند تصمیم‌گیری‌های سوگیرانه قاضی‌ها را اصلاح کند. با ترکیب مجموعه کلان‌داده‌ها با هوش مصنوعی و پیدا کردن سوگیری‌های رفتاری می‌توان به قضاوت‌های عادلانه‌تری دست

یافت. تصمیمات قاضی اغلب متأثر از عواملی است که به پرونده مورد بررسی ارتباطی ندارند. مثلا اگر یک قاضی دو مرتبه پشت سر هم حکم به جنون بدهد، ممکن است فکر کند که بیش از اندازه مدارا کرده است و در پرونده بعدی چنین حکمی ندهد. این تأثیری است که پرونده‌های دیگر روی پرونده فعلی او گذاشته‌اند.

یکی از تجربیات استفاده از هوش مصنوعی در حقوق، پیش‌بینی تصمیمات قاضی در پرونده‌های مربوط به جنون به کمک یادگیری ماشینی است که تا حد زیادی موفق عمل کرده است. این سیستم با استفاده از دانش مربوط به هویت قاضی و ملیت کسی که به دنبال گرفتن حکم جنون است، پیش‌بینی خود را انجام می‌دهد. حال، این پرسش مطرح می‌شود که چرا قاضی پیش از اینکه وقایع را بررسی کند، تا این حد قابل پیش‌بینی است؟ می‌توان با استفاده از این سیستم، به قضاتی که اینگونه حکم می‌دهند هشدار داد و آن‌ها را به تأمل بیشتر روی پرونده دعوت کرد. آگاهی فرد از سوگیری‌هایش و آموزش او می‌تواند در کاهش این سوگیری‌ها موثر باشد.

بیشتر پیشرفت‌های فناورانه اخیر در جهت بهبود عملکرد فرایندهای دادگاه‌محور بوده‌اند. مثلا امروزه بسیاری از افراد می‌توانند از خدمات عدالت آنلاین بهره‌مند شوند و درباره فرایندها، گزینه‌ها و بدیل‌های عدالت، از طریق سیستم‌های اطلاعاتی وب‌محور اطلاعات کسب کنند. خدمات حقوقی «غیر همراه» نیز در چند سال اخیر توسط شرکت‌های حقوقی آنلاین ارائه شده است. برخی از اطلاعات وب‌محور (مثل ویدیوی دیجیتالی)، ویدیو کنفرانس (مثل تماس‌های ویدیویی گروهی اینترنت‌محور)، تله‌کنفرانس و ایمیل می‌توانند

جایگزین رویکردهای چهره به چهره در دادگاه شوند و جزء دسته دوم تغییرات فناورانه محسوب می‌شوند. مثلا فرایندهای دادگاهی آنلاین استفاده زیادی در برخی از انواع مجادلات و در رابطه با امور عدالت مجرمان دارند (خصوصا در درخواست وثیقه).

تکنیک‌های تحلیل داده‌های کلان می‌توانند برای پیوند دادن پایگاه‌های داده مختلف مورد استفاده قرار گیرند و تغییرات قانونی مربوطه را که می‌توانند روی موکلان خاص موثر باشند به کار گیرند و در نتیجه، فرصت‌های بالقوه را به مرتبط‌ترین تیم اطلاع دهند. برخی از انواع سیستم‌های متخصص می‌توانند به پیش‌بینی نتیجه انواع خاصی از پرونده‌های حقوقی کمک کنند. برای مثال، متخصصان در «پروژه پیش‌بینی دیوان عالی کشور[1]» در آمریکا، با پردازش داده‌های حاصل از حکم‌های قبلی دادگاه، موفق شدند سه چهارم نتایج پرونده‌های دیگر را پیش‌بینی کنند. یک نمونه مدرن از ابزارهای پیش‌بینی پرونده که به وکلا کمک می‌کند استراتژی‌های خود را بهبود ببخشند، Lex Machina است. این فناوری تحلیل حقوقی، با استخراج داده‌های مربوط به دعاوی قضایی و استفاده از پردازش زبان طبیعی و یادگیری ماشین توانست نتایج پرونده‌ها را پیش‌بینی کند.

البته، نگرانی‌هایی وجود دارد درباره دسترسی عمومی‌تر به عدالت و گسترش سیستم‌های حل منازعات آنلاین در ابعاد بزرگ که امروزه هم در برخی از دادگاه‌ها استفاده می‌شود. چنین تغییراتی منجر به دموکراتیک شدن عدالت می‌شوند و می‌توانند از پیشرفت روبات قاضی حمایت کنند. پیشرفت‌هایی که در جهت جایگزینی دادگاه‌های فیزیکی با بدیل‌های آنلاین آن‌ها هستند، در عین حال از مقام و

1. Supreme Court forecasting project

بسیاری از پیشرفت‌ها در سیستم‌های هوش مصنوعی مانند Siri، Al-exa وCortana بر مبنای روش‌هایی نظیر پردازش زبان طبیعی بوده است. این روش می‌تواند در حوزه حقوق، برای ساختن روبات‌های وکیل و همچنین، ساختن سیستم‌هایی که به فهم بیشتر قانون توسط مردم کمک می‌کنند به کار گرفته شود. حقوق، رشته‌ای است مبتنی بر ارزش‌هایی همچون عدالت و برابری. اما سابقه سوگیری‌های هوش مصنوعی در تصمیم‌گیری نگران‌کننده است. احتمال داشتن هوش مصنوعی وکیل (عاملانی خودمختار و تصمیم‌گیرنده که می‌توانند به ما مشاوره قانونی بدهند یا نماینده ما باشند) چقدر است؟ مشکلات و امکان‌های ساختن چنین سیستم‌هایی چیست؟ آیا نیاز به قواعد اخلاقی برای هوش مصنوعی وکیل داریم؟ چگونه مطمئن باشیم که هوش مصنوعی وکیل قادر است تصمیمات اخلاقی بگیرد؟ به طور ایده‌ال، یک روبات وکیل باید بتواند به سادگی با حقوق تطبیقی، ترجمه‌های حقوقی و پیچیدگی تمایز میان سیستم‌های حقوقی رسیدگی کند. سیستم‌های حقوقی مختلف، قواعد تفسیری متفاوتی را می‌طلبند و تفسیر متون حقوقی مختلف وابسته به منطقه یا نظام حقوقی است. روبات وکیل چگونه از عهده این مسئله برمی‌آید؟ آیا باید همیشه محدود به یک یا چند سیستم حقوقی مشابه باشد؟

همه هدف هوش مصنوعی این است که سیستم‌هایی بسازد که درجه‌ای از خودمختاری و توانایی تقلید رفتار هوشمندانه انسان را داشته باشند. بنابراین، تعریف قواعد اخلاقی برای چنین رفتاری ضروری است تا از کنترل خارج نشود. عدالت، بی‌طرف و عینی است و غایت آن منصف بودن است. بنابراین، باید مطمئن شویم

که سیستم‌های خودمختار تصمیم‌گیر خطای زیادی ندارند و ما می‌توانیم خطاهای آن‌ها را اصلاح کنیم. ما باید طراحی فناوری‌هایمان را مبتنی بر ارزش عدالت انجام دهیم و راه‌حل‌های الگوریتمیکی داشته باشیم که به هیچ روی موجب تبعیض نمی‌شوند.

ادعا بر این است که سیستم‌های هوش مصنوعی مانند انسان‌ها نیستند که تحت تاثیر احساسات قرار بگیرند، یا اقناع شوند یا داشتن یک روز خسته‌کننده روی تصمیماتشان تاثیرگذار باشد. مثلا اگر قاضی از وکیلی خوشش نیاید، این می‌تواند روی نظرش تاثیرگذار باشد. ماشین بسیاری از سوگیری‌ها، نامعقولیت‌ها و اشتباهات انسانی را ندارد.

علاوه بر تعارض منافع، سوگیری نژادپرستانه (آشکارا یا پنهان) نیز نظام حقوقی آمریکا را دچار مشکل کرده است. اقلیت‌ها نیز دسترسی کمتری به دادگاه‌ها دارند و نتایج نامطلوب‌تری را به علت محدودیت در کیفیت معرفی خود و سوگیری‌های آگاهانه و ناخودآگاه می‌گیرند. اگرچه، با نگاهی عمیق‌تر به تصمیمات حقوقی با کمک ماشین‌ها درمی‌یابیم که ماشین‌ها آنطور هم که به نظر می‌رسد بی‌طرف و نامتناقض نیستند.

یکی از مهمترینِ این مشکلات، ترجمه قانون به الگوریتم است. جنبه‌های زبانی این ترجمه و ابعاد اخلاقی ساختن چنین وکلایی چیست؟ مفهوم «عدالت در طراحی» استانداردی حداقلی برای رفتار اخلاقی است که در عامل‌های هوش مصنوعی جای‌دهی می‌شود. برای این منظور، باید به چیستی قانون پردازیم تا بتوانیم شیوه‌ای را برای ترجمه آن به زبان الگوریتم‌ها پیدا کنیم و روبات خودمختار وکیل بسازیم. در این کار، زبان‌شناسی و ترجمه اهمیت زیادی دارند. چیستی

قانون برای ارائه تحلیل درباره وکلای ناانسان باید بررسی شود. آیا قانون نوعی فرمول ریاضیاتی قابل تبدیل به الگوریتم است یا چیزی بیش از آن است و قابل تقلیل به الگوریتم نیست، زیرا هر پرونده خاص، شرایط و شیوه‌های تفسیر خاص خود را دارد.

علاوه بر این، تضمینی وجود ندارد که ماشین‌ها بتوانند از پس نکات ظریف حقوقی به نحوی کارآمد بربیایند. زیرا بسیاری از مسائل حقوقی نیازمند قاضی‌هایی است که بتوانند منافع مختلف را در نسبتی متعادل با یکدیگر قرار دهند. مثلا ممکن است لازم باشد قاضی حریم خصوصی قربانی را در برخی از پرونده‌ها مانند تجاوز جنسی در نظر بگیرد و این تصمیمی است که قاضی می‌گیرد و این قبیل تصمیمات تاثیرات مهمی روی پرونده دارند. در اینگونه موارد، هیچ پاسخ حقوقی مشخصی وجود ندارد که بگوید چگونه این ارزش‌ها باید در تعادل با یکدیگر قرار بگیرند و این تصمیم، وابسته به قاضی است.

وجود ملاحظات اینچنینی که وابسته به بافتار موضوع هستند، در پرونده‌ها رایج است. توجه به بافتار برای ارائه حکم بسیار ضروری است. گرین وود[1] می‌گوید که ماشین‌ها می‌توانند نتایج سازگاری را تولید کنند، اما فاقد مهارت‌های انتقادی دیگری هستند که برای تضمین تحقق عدالت لازمانند.

پالیبلنک[2] معتقد است که پیش‌بینی‌پذیر کردن قانون با استفاده از ماشین‌ها می‌تواند منجر به تولید مسائل بیشتری به نسبت مسائلی که این ماشین حل می‌کند بشود. یعنی وقتی شما سعی می‌کنید قانون را پیش‌بینی‌پذیرتر کنید، در واقع خطر قربانی شدن انصاف را بیشتر کرده‌اید.

1. Greenwood
2. Pulleyblank

پالیبلنک و گرین وود نتیجه می‌گیرند که ماشین‌ها احتمالا می‌توانند در حرفه‌های حقوقی دستیار انسان باشند که در نتیجه آن، این صنعت دچار تغییراتی خواهد شد. اما با جایگزین شدن انسان با ماشین در فرایند حقوقی، نیاز به ایجاد تغییراتی در خود قانون نیز خواهیم داشت. به علاوه، کاربرد هوش مصنوعی برای کارهای پیچیده مستلزم رشد پردازش عاطفی است که در آینده نزدیک شاید امکان‌پذیر شود.

آزمایش‌هایی انجام شده است که با استفاده از برنامه‌های کامپیوتری هوشمند، نتایج پرونده بر اساس اطلاعات زمینه‌ای (تحلیل پیشگویانه) پیش‌بینی شده است. آلتراس و همکارانش برنامه‌ای طراحی کرده‌اند که در دادگاه حقوق بشر اروپا، تصمیم‌هایی را که مربوط به نقش حقوق بشر بودند تحلیل زمینه‌ای کرده است تا الگوهایی را در قضاوت کشف کند. برنامه، این الگوها را یاد گرفت و توانست نتیجه پرونده‌هایش را با دقت ٪۷۹ پیش‌بینی کند. این برنامه، نمونه سیستم هوشمندی است که با تحلیل داده‌های قبلی، قواعد قابل تعمیمی را برای آینده پیدا کند. اگرچه، به ادعای سوردن، یادگیری ماشینی محدودیت‌هایی در توسعه هوش مصنوعی برای پیش‌بینی نتیجه دارد. تکنیک‌های یادگیری ماشینی فقط در مواردی مفیدند که اطلاعات تحلیل‌شده، مشابه اطلاعات جدیدی باشند که به هوش مصنوعی ارائه می‌شود. چنین مسائلی زمانی مطرح می‌شوند که اندازه نمونه پرونده‌های قبلی به قدر کافی بزرگ نباشد که کامپیوتر الگوها را کشف کند و تعمیم‌های مفیدی را بسازد.

یک مسئله دیگر این است که آیا یک برنامه کامپیوتری یا فرایند خودکار،

دارای اقتدار حقوقی برای تصمیم‌گیری به جای انسان هست یا خیر. چه کسی دارای اقتدار حقوقی برای تصمیم‌گیری است؟ برنامه‌نویس؟ سیاست‌گذار؟ تصمیم‌گیر انسانی؟ یا خود کامپیوتر؟

مسئله بعدی، دقت ترجمه قانون به کدهایی است که برنامه کامپیوتری بتواند آن را بفهمد. زبان حقوق بسیار دقیق و نیازمند فهمی سیاقمند است. برنامه‌نویسان معمولا تجربه یا صلاحیت حقوقی ندارند و فاقد تخصص سیاستی یا اجرایی هستند. بنابراین با این چالش مواجهیم که ظرافت‌ها و دقت‌های موجود در قانون به نحوی صحیح در فرایند خودمختار کدنویسی شوند. کدها باید به طور مداوم به‌روز شوند، زیرا دائما با اصلاحات جدید، تصمیم‌های جدید در پرونده‌ها و مقررات پیچیده مواجهیم. این چالش‌ها می‌توانند با وارد کردن وکلا و سیاست‌گذاران در فرایند ایجاد و به‌روز کردن برنامه‌های کامپیوتری تا حدی رفع شوند.

بسیاری از قضاوت‌ها در نظام حقوقی شامل عنصر بصیرت هستند. برنامه‌های کامپیوتری بر مبنای منطق ساخته شده‌اند که در آن‌ها، اطلاعات ورودی به وسیله الگوریتم‌های برنامه‌نویسی‌شده پردازش می‌شوند تا به یک خروجی ازپیش‌تعیین‌شده برسند. صلب بودن این فرایند موجب به وجود آمدن نوعی ناسازگاری با تصمیمات بصیرت‌مندانه می‌شود. تصمیمات بصیرت‌مندانه ممکن است مستلزم در نظر گرفتن ارزش‌های جامعه، ویژگی‌های خاص احزاب، و بسیاری از شرایط محیطی مرتبط دیگر باشد. قانون طی فرایندی ساده‌سازی می‌شود و کامپیوترها قادر می‌شوند آن را راحت‌تر پردازش کنند. اما این اصلاحات و ساده‌سازی‌ها ممکن است منجر به تصمیمات

ناعادلانه و بلاضابطه شود، زیرا محتمل است که ظرایف قانونی از بین بروند.

البته، شکل فعلی قضاوت وابسته به عوامل و دارای سوگیری‌هایی است که اینها در هوش مصنوعی می‌توانند حذف شوند. نتیجه قضاوت دادگاه می‌تواند متاثر از عوامل متعددی شامل نحوه ارائه شدن مسئله، منابع در دسترس برای شاکی، نحوه تصمیم‌گیری و ... باشد. این عوامل در تصمیم‌گیری دادگاه نیز تاثیرگذارند: قاضی در چه زمانی و چه چیزی را خورده است؟ زمان دادگاه چه زمانی از روز است؟ تعداد تصمیم‌های دیگری که قاضی در این روز گرفته است، چقدر بوده است؟ ارزش‌های او چه هستند؟ پیش‌فرض‌های ناخودآگاه او چه هستند؟ اتکای او به شهودش چه اندازه است؟ چه احساساتی در تصمیم‌گیری او دخیل هستند؟

قاضی حتی اگر به چنین عواملی آگاه شود، میزان تاثیر آن‌ها را ناچیز می‌شمارد. اگرچه، درباره اینکه استفاده از هوش مصنوعی موجب کاهش این سوگیری‌ها می‌شود یا خیر مناقشه وجود دارد. همانطور که در فصل پیش گفتیم، الگوریتم‌ها می‌توانند نتایج ناخواسته‌ای داشته باشند و تبعیض‌هایی را بازتولید کنند. استفاده از روبات قاضی ممکن است باعث کاهش ظرفیت فرایند قضاوت در برخورد کرامت‌مدارانه و انسانی با مردم در دادگاه‌ها شود، زیرا این نوع برخوردها مستلزم احساس و شفقت است. اگرچه، این امید وجود دارد که فناوری بتواند به احساسات انسانی پاسخ مناسب بدهد.

علی‌رغم اینکه این امکان وجود دارد که هوش مصنوعی جایگزین انسان‌ها در فرایند قضاوت شود، برخی از متخصصان معتقدند که بهتر است از این فناوری برای کامل کردن کارهای انسانی و بالا

بردن کارایی انسان‌ها استفاده شود، نه جایگزین شدن به جای انسان. سیستم‌هایی که می‌توانند بر اساس اطلاعات تصمیم بگیرند، می‌توانند قضاوت اولیه‌ای برای پرونده ارائه دهند که قاضی با استفاده از آن، و با تکیه بر بصیرت‌مندی خود به نتیجه نهایی برسد. همچنین می‌توان از هوش مصنوعی برای ساختن فراانسان‌هایی استفاده کرد که در آن‌ها قابلیت‌های فکری، فیزیکی و روانی انسان‌های معمولی ارتقاء پیدا کند.

بنابراین، با چالش‌هایی درباره نقش دادگاه‌ها و قاضی‌ها، نحوه مدیریت و دسته‌بندی داده‌ها، و محل و چگونگی انجام کارهای اجرایی و قضایی مواجهیم. به علاوه، مسائلی درباره مالکیت فکری و اینکه چه کسی مسئولیت کنترل و ورودی‌های روبات قاضی را بر عهده دارد و الگوریتم‌های شفاف چگونه هستند نیز وجود دارد. قاضی کاری بسیار بیشتر از حکم دادن و رسیدن به نتیجه درباره منازعات را بر عهده دارد. مدیریت پرونده و حل و فصل یک منازعه بر عهده قاضی است. بسیاری از قضات، نقشی آموزشی نیز بر عهده دارند که هم به طرفین نزاع و هم به وکلا درباره رهیافت‌های قابل اتخاذ آگاهی می‌دهند و هم آموزش اجتماعی را در سطحی وسیع‌تر بر عهده می‌گیرند. کسانی که از جایگزینی کامل قاضی با یک سیستم هوش مصنوعی دفاع می‌کنند، این نکته را در نظر نمی‌گیرند که قاضی مساهمت‌هایی در جامعه دارد که فراتر از قضاوت است و شامل امور مهمی درباره پذیرش و اجابت قانون می‌شود.

روشن است که انجام بسیاری از کارهای قضایی نیازمند هوش انسانی است و برنامه‌های کامپیوتری راه درازی برای کسب این قابلیت‌ها

و تعامل مشفقانه و احساسی، و پاسخگویی زیرکانه به مردم دارند. آیا پیشرفت‌های فناورانه می‌توانند روزی جایگزین قاضی در دادگاه شوند و فناوری‌های عاطفی‌تر به چه طریقی در این کار مشارکت کنند؟

بهتر است به جای اینکه بپرسیم آیا هوش مصنوعی کارکردهای قضایی را تغییر خواهند داد، درباره زمان و میزان این تغییرات پرسش کنیم. فناوری امروز هم در حال تغییر شغل‌های قضایی است و در بخش‌های مختلف از جمله فرایندهای مشاوره‌ای به کار می‌رود. در آینده نزدیک، دادگاه‌ها به ایجاد و گسترش پلتفرم‌های آنلاین و سیستم‌هایی برای بایگانی، مراجعه و فعالیت‌های دیگر ادامه خواهند داد. این تغییرات چارچوبی را فراهم می‌کنند که روبات قاضی بتواند در آن شکوفا شود. چالش‌هایی در این باره وجود دارد که روبات قاضی در چه محدوده‌ای مجاز به قضاوت است و چه پرونده‌های نیازمند بصیرت‌مندی قاضی است و روبات نمی‌تواند در آن‌ها ورود کند. در برخی از حوزه‌ها، واضح است که استفاده از هوش مصنوعی می‌تواند بر دقت، سرعت و به‌صرفه بودن اقتصادی فرایند تاثیر مثبتی بگذارد.

مسائل متعددی پیش روی ایده روبات قاضی وجود دارد از جمله اینکه بصیرت‌مندی روبات و محدوده جایگزین شدن انسان با آن چه میزان است. قاضی پرونده را مدیریت می‌کند، چارچوبی پاسخگو و انسانی ارائه می‌کند، پرونده‌ها را فیصله می‌دهد، سیستم‌ها و فرایندهای دادگاهی را مدیریت می‌کند و نقش آموزش عمومی را نیز بر عهده دارد. به علاوه، برای مشخص کردن مرزهای عمل روبات قاضی، باید ملاحظات اخلاقی را در نظر گرفت و به پرسش‌هایی درباره نوشتن الگوریتم‌ها پاسخ داد. یک آینده محتمل این است که

هوش مصنوعی مکمل قاضی انسانی باشد و قابلیت‌های او را ارتقاء دهد یا به او در تصمیم‌گیری کمک کند.

۲-۳- هوش مصنوعی در فرایندهای تصمیم‌سازی کلان

همانطور که پیش‌تر گفتیم، سیستم‌های هوشمندی که با شیوه یادگیری عمیق تربیت می‌شوند و از کلان‌داده‌ها برای تصمیمات خود استفاده می‌کنند به حقیقت مطلق دسترسی ندارد و تصمیمات آن به نوعی نظر شخصی آن محسوب می‌شوند که ممکن است آمیخته با خطا یا سوگیری باشند. اگر از ماشین‌های هوشمند یا فراهوشمند برای تصمیم‌گیری‌های کلان اجتماعی یا سیاسی بهره بگیریم، ممکن است این ماشین‌ها مرتکب اشتباه شوند و منجر به بروز مخاطراتی بسیار بزرگ و فاجعه‌بار در سطح ملی و بین‌المللی شوند. به علاوه، امکان رهایی و غیر قابل کنترل شدن این ماشین‌ها، عمل کردن آن‌ها طبق میل خودشان و حتی دروغ گفتن آن‌ها نیز دور از ذهن نیست.

سیستم‌های هوش مصنوعی امروزه در بسیاری از صنایع و سازمان‌ها، تصمیم‌گیرنده هستند. اما ما هنوز موازین مشخصی را برای تعیین اینکه چه تصمیماتی قابل پذیرش هستند نداریم. آیا روبات‌ها می‌توانند تصمیمات اخلاقی خوبی بگیرند؟ آیا ما مجازیم که به آن‌ها حق چنین تصمیم‌گیری‌هایی را بدهیم؟ ما در اینجا با دو مسئله متعارض مواجهیم: آیا باید تلاش کنیم روبات‌هایی بسازیم که قادر به گرفتن تصمیمات اخلاقی هستند؟ یا باید از قرار دادن روبات‌ها در موقعیت‌هایی که مستلزم کفایت اخلاقی و داشتن فهمی از موقعیت اجتماعی هستند اجتناب کنیم؟ به بیان دیگر،

آیا ما باید هوش مصنوعی را صرفا به عنوان ابزاری کمکی برای انجام کارهایمان طراحی کنیم؟ یا اینکه سیستم‌های هوشمندی بسازیم که بتوانند به عنوان همکارانی با توانایی‌ها و حقوق مشابه ما، درکنار ما زندگی کنند؟ هر یک از این رویکردها موافقان و مخالفانی دارد و نظرگاه‌های فلسفی متفاوتی درباره شأن وجودی و اخلاقی موجودات انسانی و غیر انسانی از هر یک از آن‌ها پشتیبانی می‌کند که بررسی آن‌ها، مستلزم پژوهشی مستقل در این باره است.

۲-۴- هوش مصنوعی و عدالت اقتصادی

نظام اقتصادی کنونی در جهان، مبتنی بر نوعی تعادل در توزیع اقتصادی است که این امر معمولاً با تعیین دستمزد ساعتی اتفاق می‌افتد. اغلب شرکت‌ها در توزیع سود حاصل از محصولات و خدمات خود، هنوز به کار ساعتی متکی‌اند. اما در آینده‌ای نزدیک، شرکت‌ها می‌توانند با استفاده از هوش مصنوعی، تا حد زیادی وابستگی به نیروی کار انسانی را کاهش دهند و این به معنای واگذاری و تقسیم سود میان نیروی انسانی کمتری است. در نتیجه، افرادی که مالکیت شرکت‌های مبتنی بر هوش مصنوعی را دارند، بخش عمده سود را خود برداشت می‌کنند.

همین امروز هم شکاف رو به افزایشی در توزیع ثروت در این کسب و کارها قابل مشاهده است. صاحبان استارتاپ‌ها بخش عمده سود اقتصادی‌ای که خلق کرده‌اند را تصاحب می‌کنند. بنابراین یکی از چالش‌های اخلاقی که هوش مصنوعی به همراه می‌آورد این است که اگر ما در حال ساختن جامعه‌ای با سازماندهی شغلی متفاوتی هستیم، چگونه

باید اقتصاد عادلانه‌ای متناسب با آن را بسازیم که در آن، میان صاحبان این فناوری‌ها با دیگران، شکاف عمیقی ایجاد نشود؟ سود حاصل از فناوری‌های چگونه می‌تواند به نحو عادلانه‌ای توزیع شود؟ آیا علوم و قوانین اقتصادی متفاوتی را باید تولید کنیم؟ به نظر می‌رسد که پاسخ این پرسش مثبت است.

۲-۵- تأثیر هوش مصنوعی بر رفتارها و تعاملات انسانی

سیستم‌های هوش مصنوعی روز به روز در حال بهتر شدن در زمینه مدل‌سازی مکالمه و روابط انسانی هستند. برای مثال، در سال ۲۰۱۵، روباتی به نام یوجین گوستمن[1] برای نخستین بار برنده چالش تورینگ شد. تورینگ در سال ۱۹۵۰ آزمونی[2] را طراحی کرد که با استفاده از آن، می‌توان یک موجود هوشمند را از موجودی غیر هوشمند تمییز داد. در این چالش، داورهایی که گروهی از انسان‌ها بودند، ورودی‌های متنی برای صحبت کردن با موجودی نامشخص (نمی‌دانستند که با یوجین صحبت می‌کنند یا با یک انسان) را به او عرضه کردند و پس از دریافت پاسخ‌های او، حدس زدند که موجود طرف صحبتشان انسان بوده یا ماشین. یوجین گوستمن بیش از نیمی از داورها را که گمان می‌کردند در حال صحبت با انسان هستند، فریب داد.

این واقعه، نقطه آغاز عصری است که ما در آن تعاملی مشابه انسان‌ها را با ماشین‌ها هم خواهیم داشت؛ چه در قالب ارائه‌دهنده خدمات به مشتری و چه در قالب یک دوست اجتماعی. در حالی که انسان‌ها در توجه و محبتی که می‌توانند به دیگری ابراز کنند

1. Eugene Goostman
2. Turing test

محدود هستند، روبات‌های مصنوعی می‌توانند از منابع نامحدود مجازی خود را در ساختن روابط استفاده کنند. فیلم Her که در سال ۲۰۱۳ ساخته شد، نمونه چنین سیستمی را به خوبی به تصویر کشیده است.

علاوه بر این سیستم‌ها، امروزه شاهد ماشین‌هایی هستیم که مراکز پاداش در مغز انسان‌ها را برمی‌انگیزند و از روش‌های متعددی برای اعتیاد کاربر به بازی‌های ویدیویی و موبایلی استفاده می‌کنند. اعتیاد به فناوری جدیدترین نوع وابستگی در انسان است که می‌تواند مخاطراتی جدی را برای سلامت روحی و جسمی او به وجود بیاورد. در حالی که اگر جهت‌دهی به توجه انسان و برانگیختن فعالیت‌های خاص در او به درستی و با اهداف نیکی رخ دهد، این ویژگی می‌تواند تبدیل به فرصتی برای حرکت جامعه به سوی رفتارهای مفیدتر شود.

۲-۶- خودروهای خودران هوشمند

ساختن خودروهای خودران[1] (یا خودروهای بدون راننده[2] یا خودروهای خودمختار[3])، یکی از چالش‌برانگیزترین مسائل در حوزه اخلاق هوش مصنوعی است. این خودروها می‌توانند بدون مداخله انسانی و برای مدتی طولانی رانندگی کنند. پیش‌بینی می‌شود که خودروهای نیمه‌اتوماتیک یا تمام‌اتوماتیک در آینده‌ای نه چندان دور مورد استفاده فراگیری قرار بگیرند. پرسش‌هایی از این دست برای مهندسان و شرکت‌های سازنده این خودروها مطرح است: این خودروها در هنگام برخورد با موانع چگونه تصمیم‌گیری می‌کنند و تصمیم‌های آن‌ها چه قدر قابل اعتماد است؟ راننده انسانی خاطی

1. self-driving car
2. driverless
3. autonomous

را می‌توان مجازات کرد و به او مسئولیت نسبت داد، اما وقتی این خودروها تصادف می‌کنند، چه کسی مسئول است؟ چه کسی باید مجازات شود؟ بسیار محتمل است که این روبات‌ها در موقعیت‌هایی قرار بگیرند که نیازمند تصمیم‌گیری درباره مرگ و زندگی است؛ در این حالت، باید از خود محافظت کنند یا از انسان‌های دیگر؟

برای مثال این موقعیت را در نظر بگیرید: خودروی خودرانی در موقعیتی قرار گرفته است که یا باید به سمت چپ منحرف شود و با دختربچه هشت ساله‌ای برخورد کند، یا به سمت راست منحرف شود و با مادربزرگ هشتاد ساله‌ای تصادف کند. سرعت این خودرو نیز به اندازه‌ای است که در صورت برخورد با هر یک از این دو نفر، آن‌ها کشته خواهند شد. تصمیم اخلاقی درست در چنین موقعیتی چیست؟ طراح چنین خودرویی چگونه باید برنامه‌نویسی کند و خودرو را برای مواجهه با چنین موقعیتی آماده کند؟ به نظر می‌رسد هر دوی این تصمیم‌ها، انتخابی غیر اخلاقی هستند. همانطور که در بحث سوگیری نیز اشاره کردیم، هوش مصنوعی اگر بخواهد عادلانه عمل کند، باید فارغ از تفاوت‌های جنسیتی، نژادی، مذهبی، سنی، ملیتی و ... تصمیم‌گیری کند. بدین ترتیب، تصمیم‌گیری در موقعیت‌های اینچنینی، مسئله‌ای است که باید در حوزه اخلاق بررسی شود و اندیشیدن به مسائلی از این دست و یافتن پاسخ‌هایی برای آن‌ها باید پیش از فراگیری کاربرد این خودروها صورت بگیرد تا برای مواجهه با مخاطرات این ماشین‌ها بیشترین آمادگی را پیدا کنیم.

جمع بندی

جمع‌بندی

در گزارش حاضر، جدی‌ترین و پربسامدترین پرسش‌ها، دغدغه‌ها و نگرانی‌هایی که متخصصان حوزه اخلاق هوش مصنوعی با آن‌ها دست و پنجه نرم می‌کنند را مطرح کردیم. سیستم‌های هوشمند هنوز در مراحل آغازین تحول خود به سر می‌برند و در همین مراحل هم چالش‌هایی جدی را پیش روی ما قرار داده‌اند. به نظر می‌رسد که ما در همین مراحل اولیه، باید موضع خود را نسبت به برخی از مسائل مشخص کنیم و نظریات اخلاقی خود را به‌روز کنیم تا بتوانیم با استفاده از آن‌ها، پاسخ پرسش‌های اخلاقی فعلی را بیابیم و علاوه بر آن، بنیان قابل اتکایی را برای مواجهه با چالش‌های اخلاقی آینده‌ای که سیستم‌های هوشمند در مسیر ما قرار می‌دهند، فراهم کنیم. چارچوبی که معیار ما برای ارزیابی سیستم‌های هوشمند و تصمیم‌گیری برای چگونگی توسعه آنهاست، باید به قدری جامعیت، پویایی و انعطاف داشته باشد که بتواند پاسخگوی سوالات متنوع و متعددی باشد که در نتیجه کاربرد هوش مصنوعی در حوزه‌های انضمامی مختلف مطرح می‌شوند. برای تهیه چنین چارچوبی، ضروری است که به مبانی فلسفی و انسان‌شناختی عمیق‌تری که

برگزیده‌ایم رجوع کنیم و در صورت لزوم، در آن‌ها بازاندیشی کنیم و این موازین و اصول را متناسب با نیازهای امروزمان، تفصیل و تفسیر کنیم .

منابع

[۱] کارمن، تیلر، مرلوپونتی ستایشگر فلسفه، ۱۳۹۴، ترجمه هانیه یاسری، نشر ققنوس

[2] Alac, M. (2015). Social robots: Things or agents?. AI & Society

[3] Aletras, N. et al, 'Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective' [2016] (October) PeerJ Computer Science

[4] Altmann, J., Asaro, P., Sharkey, N., & Sparrow, R. (2013). Armed Military Robots: Editorial. Ethics and Information Technology, 15(2), 73–76.

[5] Andreas, Matthias (2011) Algorithmic Moral Control of War Robots: Philosophical Questions, Law, Innovation and Technology, 3:2, 279-301

[6] Anderson, M. & Anderson, S. L., (Eds.), (2011). Machine ethics. Cambridge: Cambridge University Press.

[7] Arkin, R. 'Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture', Tech Report No GIT-GVU-07-11, Mobile Robot Laboratory, College of Computing, Georgia Institute of Technology, 2007.

[8] Asaro, P. M. (2006). What should we want from a robot ethic? International Review of Information Ethics, 6, 9–16.

[9] Asaro, P. (2012). On banning autonomous lethal systems: human rights, automation and the dehumanizing of lethal decision-making. Special Issue on New Technologies and Warfare, International Review of the Red Cross 94(886), 687–709.Bostrom, N., Singularity and Predictability. 1998: Available at: http://mason.gmu.edu/~rhanson/vc.html.

[10] Bostrom, N. (. (2017). Strategic Implications of Openness in AI Development. Global Policy.

[11] Bostrom, N., Yudkowsky, E., 2011, THE ETHICS OF ARTIFICIAL

INTELLIGENCE in Cambridge Handbook of Artificial Intelligence, eds. William Ramsey and Keith Frankish, Cambridge University Press.

[12] Cloos, C. 'The Utilibot Project: An Autonomous Mobile Robot Based on Utilitarianism' (2005) in Anderson et al

[13] Dreyfus H (1972) What computers can't do. MIT Press, New York2. Dreyfus H (1992) What computers still can't do. MIT Press, New York

[14] Ernst, E., Merola, R., Samaan, D., 2018. The economics of artificial intelligence: Implications for the future of work. International Labour

[15] Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. Minds and Machines, 14(3), 349–379

[16] Helbing, D., Frey, B., Gigerenzer, G., Hafen, E., Hanger, M., Hofstetter, Y., . . . Zwitter, A. (2019). Will democracy survive big data and artificial intelligence? In D. Helbing, Towards Digital Enlightenment. Springer International Publishing AG, part of Springer Nature.

[17] Kurzweil, R. (2005). The singularity is near. When Humans transcend biology. New York: Viking.

[18] Latour, L. 'A Collective of Humans and Nonhumans: Following Daedalus's Labyrinth' in David Kaplan (ed), Readings in the Philosophy of Technology (Rowman & Littlefield, 2nd edn 2009).

[19] Lin, P., Abney, K., & Bekey, G. (Eds.). (2012). Robot ethics—The Ethical and Social Implications of Robotics. Cambridge: The MIT Press.

[20] Lin, P. (2015). Why ethics matters for autonomous cars. In M. Maurer, J. C. Gerdes, B. Lenz, H. Winner (Eds.), Autonomes Fahren: Technische, rechtliche und gesellschaftliche aspekte (pp. 69–85). Berlin Heidelberg: Springer.

[21] Moor, J. H. (2006). The nature, importance, and difficulty of machine

ethics. IEEE Intelligent Systems, 21(4), 18–21.

[22] Moor, J. H. (2007). Four kinds of ethical robot. Philosophy Now, 72, 12–14.

[23] Müller V (ed), 2016. Risks of artificial intelligence. CRC Press, Boca Raton.

[24] Russell, S. (2016, June) Should we fear super smart robots? Scientific American, 314, 58 9

[25] Rutkin, A. (2014, September) Ethical trap: robot paralyzed by choice of who to save, New Scientist. Amsterdam: Elsevier.

[26] Searle, J.R. (1980), Minds, brains, and programs. Behavioral and Brain Sciences 3 (3): 417-457

[27] Sharkey, A. J. C., & Sharkey, N. E. (2012). Granny and the robots: Ethical issues in robot care for the elderly. Ethics and Information Technology, 14(1), 27–40.

[28] Sharkey, N. (2012). The Evitability of Autonomous Robot Warfare. International Review of the Red Cross, 787-799.

[29] Sharkey, A. (2016). Should we welcome robot teachers? Ethics and Information Technology, 18(4), 283–297

[30] Surden, H. 'Machine Learning and Law' (2014) 89 Washington Law Review 87

[31] Susskind, R., & Susskind, D. (2015) The Future of Professions: How technology will transform the work of human experts. Oxford: Oxford University Press Thagard, P. (2005). Mind: Introduction to cognitive science. MIT press.

[32] Thagard, P. (2005). Mind: Introduction to cognitive science. MIT press.

[33] Torrance, S. (2011). Machine ethics and the Idea of a more-than-human moral world. In M. Anderson & S. L. Anderson (Eds.), Machine ethics (pp. 115–137). Cambridge: Cambridge University Press.

[34] Vinge, V. Technological singularity. in VISION-21 Symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute. 1993.

[35] Winograd, T. 'Thinking Machines: Can There Be? Are We?' in James J Sheehan and Morton Sosna (eds), The Boundaries of Humanity: Humans, Animals, Machines (University of California Press, 1991).

[36] Wolfram, S., A new kind of science. Vol. 5. 2002: Wolfram media Champaign, IL

[37] Yampolskiy R.V. (2013) Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach. In: Müller V. (eds) Philosophy and Theory of Artificial Intelligence. Studies in Applied Philosophy, Epistemology and Rational Ethics, vol 5. Springer, Berlin, Heidelberg.

[38] Yampolskiy, Roman, Unpredictability of AI, 2019.

[39] Yapo, A. & Weiss, J. 2018. Ethical Implications of Bias in Machine Learning. 51st Hawaii International Conference on System Sciences

csri.majazi.ir