مروری بر اخلاق هوش مصنوعی

نویسندگان:

Changwu Huang , *Member, IEEE*, Zeqi Zhang, Bifei Mao, and Xin Yao , *Fellow, IEEE*

مترجم:

Ryan Heida (links.ryanheida.com)

واژه‌های کلیدی فارسی:

هـوش مصـنوعی (AI)، اخـلاق هـوش مصـنوعی، مسـائل اخـلاقی، نظریـه اخـلاقی، اصـل اخلاقی

واژه‌های کلیدی انگلیسی:

Artificial intelligence (AI), AI ethics, ethical issue ,ethical theory, ethical principle

چکیده:

هوش مصنوعی (AI) به طور عمیق زندگی ما را تغییر داده و همچنان به تغییر آن ادامه خواهد داد. هوش مصنوعی در حوزه‌ها و سناریوهای متعددی مانند رانندگی خودران، مراقبت‌های پزشکی، رسانه، امور مالی، ربات‌های صنعتی و خدمات اینترنتی به کار گرفته می‌شود. کاربرد گسترده هوش مصنوعی و ادغام عمیق آن با اقتصاد و جامعه باعث افزایش بهره‌وری و تولید مزایا شده است. در عین حال، این فناوری به ناچار بر نظم اجتماعی موجود تأثیر می‌گذارد و نگرانی‌های اخلاقی را مطرح می‌کند. مسائلی مانند نشت حریم خصوصی، تبعیض، بیکاری، و خطرات امنیتی که توسط سیستم‌های هوش مصنوعی ایجاد می‌شوند، مشکلات زیادی برای افراد به وجود آورده‌اند. بنابراین، اخلاق هوش مصنوعی، که حوزه‌ای مرتبط با مطالعه مسائل اخلاقی در هوش مصنوعی است، نه تنها به یک موضوع مهم تحقیقاتی در میان محققان دانشگاهی تبدیل شده است، بلکه یک موضوع مشترک مورد توجه افراد، سازمان‌ها، کشورها، و جامعه نیز می‌باشد. این مقاله یک مرور جامع از این حوزه ارائه می‌دهد، که شامل خلاصه و تحلیل ریسک‌ها و مسائل اخلاقی ناشی از هوش مصنوعی، دستورالعمل‌ها و اصول اخلاقی صادر شده توسط سازمان‌های مختلف، رویکردهایی برای مقابله با مسائل اخلاقی در هوش مصنوعی، و روش‌هایی برای ارزیابی اخلاق در هوش مصنوعی است. علاوه بر این، چالش‌های اجرای اخلاق در هوش مصنوعی و دیدگاه‌های آینده نیز مورد بحث قرار گرفته‌اند. امیدواریم این پژوهش دیدگاهی سیستماتیک و جامع از اخلاق هوش مصنوعی برای پژوهشگران و متخصصان این حوزه، به‌ویژه تازه‌کاران این رشته تحقیقاتی، ارائه دهد.

بیانیه تأثیر

اخلاق هوش مصنوعی یک موضوع مهم و نوظهور در میان محافل دانشگاهی، صنعت، دولت، جامعه و افراد است. در دهه‌های گذشته، تلاش‌های بسیاری برای بررسی مسائل اخلاقی در حوزه هوش مصنوعی صورت گرفته است. این مقاله یک مرور جامع بر حوزه اخلاق هوش مصنوعی ارائه می‌دهد، که شامل خلاصه و تحلیل مسائل اخلاقی هوش مصنوعی، دستورالعمل‌ها و اصول اخلاقی، رویکردها برای مقابله با مسائل اخلاقی هوش مصنوعی، و روش‌هایی برای ارزیابی اخلاقیات فناوری‌های هوش مصنوعی است. علاوه بر این، چالش‌های پژوهشی- و دیدگاه‌های آینده نیز مورد بحث قرار گرفته‌اند. این مقاله به محققان کمک می‌کند تا دیدگاهی کلی از اخلاق هوش مصنوعی به دست آورند و بدین ترتیب تحقیقات و مطالعات بیشتری در این حوزه انجام دهند.

بخش 1 - مقدمه

هوش مصنوعی (AI) در دهه گذشته پیشرفت‌های سریع و چشمگیری داشته است. فناوری‌های هوش مصنوعی مانند یادگیری ماشین(ML) ، پردازش زبان طبیعی و بینایی کامپیوتری به طور فزاینده‌ای در حوزه‌ها و جنبه‌های مختلف جامعه ما نفوذ کرده و گسترش یافته‌اند. هوش مصنوعی به تدریج در حال جایگزینی وظایف انسانی و تصمیم‌گیری‌های انسانی است و در بخش‌های متعددی مانند تجارت، لجستیک، تولید، حمل‌ونقل، مراقبت‌های بهداشتی، آموزش و مدیریت دولتی به کار گرفته شده است.

کاربرد هوش مصنوعی باعث بهبود بهره‌وری و کاهش هزینه‌ها شده است که این امر برای رشد اقتصادی، توسعه اجتماعی و رفاه انسانی مفید است. برای مثال، چت‌بات‌های هوش مصنوعی می‌توانند در هر زمان به سوالات مشتریان پاسخ دهند و رضایت مشتریان و فروش شرکت را افزایش دهند. همچنین، هوش مصنوعی این امکان را برای پزشکان فراهم کرده است که از طریق خدمات تله‌مدیسین به بیماران در مناطق دورافتاده

خدمت کنند. بدون شک، توسعه سریع و کاربرد گسترده هوش مصنوعی در حال حاضر زندگی روزمره، انسانیت و جامعه را تحت تأثیر قرار داده است.

با این حال، هوش مصنوعی هم‌زمان ریسک‌ها و مسائل اخلاقی قابل توجهی را برای کاربران، توسعه‌دهندگان، انسان‌ها و جامعه ایجاد می‌کند. در سال‌های اخیر، موارد بسیاری از نتایج نامطلوب ناشی از هوش مصنوعی مشاهده شده است. به عنوان مثال، در سال ۲۰۱۶، راننده یک خودروی تسلا در تصادف جاده‌ای کشته شد، زیرا حالت Autopilot این خودرو نتوانست یک کامیون در حال عبور را تشخیص دهد. چت‌بات هوش مصنوعی شرکت مایکروسافت، **Tay.ai**، به دلیل نژادپرست و جنسیت‌گرا شدن در کمتر از یک روز پس از ورود به توییتر از دسترس خارج شد. نمونه‌های بسیاری دیگر نیز وجود دارند که به مسائل مربوط به شکست، انصاف، تعصب، حریم خصوصی و دیگر مشکلات اخلاقی سیستم‌های هوش مصنوعی مربوط می‌شوند. حتی جدی‌تر از این، فناوری هوش مصنوعی توسط مجرمان برای آسیب رساندن به دیگران یا جامعه مورد استفاده قرار گرفته است؛ به عنوان مثال، مجرمان با استفاده از نرم‌افزار مبتنی بر هوش مصنوعی صدای یک مدیر اجرایی را جعل کردند و درخواست انتقال جعلی ۲۴۳,۰۰۰ دلار کردند. بنابراین، ضروری و حیاتی است که مسائل و ریسک‌های اخلاقی هوش مصنوعی مورد بررسی قرار گیرد تا این فناوری به شکلی اخلاقی ساخته، اعمال و توسعه یابد.

اخلاق هوش مصنوعی یا اخلاق ماشین یک حوزه نوظهور و میان‌رشته‌ای است که به بررسی مسائل اخلاقی هوش مصنوعی می‌پردازد. اخلاق هوش مصنوعی شامل دو جنبه است: اخلاق هوش مصنوعی که نظریه‌های اخلاقی، دستورالعمل‌ها، سیاست‌ها، اصول، قوانین و مقررات مرتبط با هوش مصنوعی را مطالعه می‌کند؛ و هوش مصنوعی اخلاقی که به هوش مصنوعی‌ای اشاره دارد که می‌تواند هنجارهای اخلاقی را رعایت کند و رفتار اخلاقی داشته باشد. اخلاق هوش مصنوعی پیش‌نیازی برای ساخت یا رفتار اخلاقی هوش مصنوعی است. این حوزه ارزش‌ها و اصول اخلاقی را بررسی می‌کند که تعیین می‌کنند چه چیزی از نظر اخلاقی درست یا غلط است. با داشتن اخلاق مناسب برای هوش مصنوعی، می‌توان از طریق برخی روش‌ها و فناوری‌ها هوش مصنوعی اخلاقی ساخت یا پیاده‌سازی کرد.

با وجود اینکه اخلاق هوش مصنوعی طی چندین سال گذشته به طور گسترده توسط پژوهشگران میان‌رشته‌ای مورد بحث قرار گرفته است، این حوزه همچنان در مراحل ابتدایی خود قرار دارد. اخلاق هوش مصنوعی یک حوزه پژوهشی‌ـ بسیار گسترده و به سرعت در حال توسعه است که در سال‌های اخیر توجه بیشتری از سوی پژوهشگران جلب کرده است. اگرچه در سال‌های گذشته چندین مقاله مروری منتشرـ شده است، هر یک از آن‌ها بر جنبه یا جنبه‌هایی خاص از اخلاق هوش مصنوعی تمرکز داشته‌اند و هنوز کمبود بررسی‌های جامع برای ارائه یک تصویر کامل از این حوزه وجود دارد.

اهداف و مشارکت‌های اصلی مقاله:

این مقاله با هدف ارائه یک مرور سیستماتیک و جامع از اخلاق هوش مصنوعی از جنبه‌های مختلف، برای ارائه راهنمایی‌های کاربردی به جامعه برای تحقق هوش مصنوعی اخلاقی در آینده نوشته شده است. این مقاله با جمع‌بندی و تحلیل مسائل اخلاقی هوش مصنوعی، دستورالعمل‌ها و اصول اخلاقی، رویکردها برای مقابله با مسائل اخلاقی و روش‌هایی برای ارزیابی اخلاقی هوش مصنوعی به جامعه علمی و متخصصان کمک می‌کند.

مشارکت‌های اصلی مقاله به شرح زیر است:

۱. ارائه یک مرور جامع درباره اخلاق هوش مصنوعی، شامل مسائل اخلاقی و ریسک‌های هوش مصنوعی، دستورالعمل‌ها و اصول اخلاقی، رویکردهای مقابله با مسائل اخلاقی و روش‌های ارزیابی هوش مصنوعی اخلاقی.

۲. دسته‌بندی جدیدی از مسائل اخلاقی هوش مصنوعی ارائه شده که به شناسایی، درک و تحلیل مشکلات اخلاقی در هوش مصنوعی و توسعه راه‌حل‌های مربوطه کمک می‌کند.

۳. مرور دستورالعمل‌ها و اصول اخلاقی جهانی مربوط به هوش مصنوعی که توسط شرکت‌ها، سازمان‌ها و دولت‌ها منتشر شده است.

۴. بررسی رویکردهای میان‌رشته‌ای برای مقابله با مشکلات اخلاقی هوش مصنوعی، از جمله رویکردهای اخلاقی، فناورانه و قانونی.

۵. مرور روش‌های ارزیابی اخلاق هوش مصنوعی که جنبه‌ای کمتر مورد توجه در ادبیات موجود است.

۶. شناسایی چالش‌های موجود در اخلاق هوش مصنوعی و ارائه دیدگاه‌های آینده برای طراحی هوش مصنوعی اخلاقی.

این مقاله راهنمایی کامل و جامعی برای پژوهشگران و متخصصان این حوزه، به ویژه مبتدیان، ارائه می‌دهد تا آن‌ها بتوانند تحقیقات و مطالعات بیشتری در این زمینه انجام دهند.

بخش 2 - دامنه و روش‌شناسی

در این بخش، ابتدا جنبه‌ها و موضوعات تحت پوشش این بررسی و ارتباط میان این موضوعات را روشن می‌کنیم. سپس روش‌شناسی استفاده‌شده در انجام این بررسی، از جمله استراتژی جستجوی منابع و معیارهای انتخاب آن‌ها، را شرح می‌دهیم.

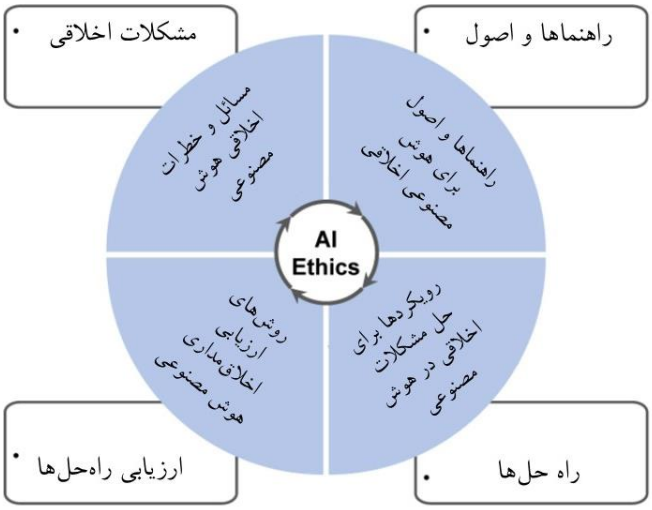
1.2. دامنه

دامنه و موضوعات این مقاله به شرح زیر است: بررسی مسائل و ریسک‌های اخلاقی هوش مصنوعی نقطه شروع این مطالعه است، زیرا وجود مسائل اخلاقی در هوش مصنوعی زمینه‌ساز شکل‌گیری حوزه تحقیقاتی اخلاق هوش مصنوعی است. بنابراین، ضروری و مهم است که مشکلات اخلاقی موجود در هوش مصنوعی شفاف‌سازی و درک شود.

سپس دستورالعمل‌ها و اصول اخلاقی که توسعه و استفاده از هوش مصنوعی را هدایت می‌کنند، بررسی می‌شوند. با توجه به اینکه مسائل اخلاقی هوش مصنوعی توجه روزافزون بخش‌های مختلف جامعه را به خود جلب کرده است، بسیاری از سازمان‌ها (از جمله نهادهای آکادمیک، صنعت و دولت‌ها) به بحث و جستجوی چارچوب‌ها، دستورالعمل‌ها و اصول ممکن برای حل مسائل اخلاقی هوش مصنوعی پرداخته‌اند. این دستورالعمل‌ها و اصول جهت‌گیری‌های ارزشمندی برای پیاده‌سازی هوش مصنوعی اخلاقی ارائه می‌دهند.

پس از شفاف‌سازی مسائل اخلاقی موجود و دستورالعمل‌ها، رویکردهای حل مسائل اخلاقی در هوش مصنوعی بررسی می‌شوند. این مقاله رویکردهای اخلاقی، فناورانه و قانونی را پوشش می‌دهد، اما تمرکز بیشتری بر دو دسته اول (رویکردهای اخلاقی و فناورانه) دارد، زیرا پژوهشگران جامعه هوش مصنوعی ممکن است بیشتر به این دو دسته علاقه‌مند باشند.

در نهایت، نحوه ارزیابی هوش مصنوعی اخلاقی که شامل بررسی اخلاقی یا اخلاق‌مداری هوش مصنوعی است، خلاصه می‌شود؛ به عبارت دیگر، بررسی اینکه مشکلات اخلاقی تا چه حد برطرف شده یا اینکه آیا یک سیستم هوش مصنوعی الزامات اخلاقی را برآورده می‌کند یا خیر.



شکل ۱. موضوعات مطرح‌شده در این مقاله و ارتباط بین آن‌ها.

به وضوح، این چهار جنبه برای حل مسائل اخلاقی در هوش مصنوعی ضروری هستند. بنابراین، این چهار جنبه محتوای اصلی این مقاله را تشکیل می‌دهند و یک مرور سیستماتیک از اخلاق هوش مصنوعی ارائه می‌دهند.

موضوعات یا جنبه‌های مورد بررسی در این مقاله و ارتباط میان آن‌ها در **شکل ۱** نشان داده شده است.

2.2. روش‌شناسی

این مرور طیف گسترده‌ای از اسناد را پوشش می‌دهد، از جمله منابع علمی، سازمانی، منابع خاکستری دولتی، و گزارش‌های خبری. جستجوی ادبیات مرتبط در دو مرحله انجام شد. در مرحله اول، ورودی‌ها یا کلمات کلیدی که منعکس کننده اصطلاحات مختلف مرتبط با اخلاق هوش مصنوعی هستند، برای جستجو در پایگاه‌های Google Scholar، Web of Science، IEEE Xplore، ACM Digital Library، Science Direct، Springer Link، arXiv و Google به کار گرفته شدند. کلمات کلیدی استفاده‌شده شامل موارد زیر بودند: (اخلاق، اخلاقی، مسئولیت، مسئولانه، قابل اعتماد، شفاف، توضیح‌پذیر، عادلانه، مفید، قوی، ایمن، خصوصی، پایدار) و/یا (مسائل، ریسک‌ها، دستورالعمل، اصل، رویکرد، روش، ارزیابی، سنجش، چالش) و (هوش مصنوعی، AI، یادگیری ماشین، ML، سیستم هوشمند، عامل هوشمند). (ما عمدتاً به ادبیاتی که از سال ۲۰۱۰ به بعد منتشر- یا ارائه شده باشد توجه کردیم و تا حد امکان این کلمات کلیدی را در عناوین جستجو کردیم.

در مرحله دوم، به بررسی کارهای مرتبط با ادبیات یافت‌شده در مرحله اول پرداختیم، از جمله مقالات ارجاع‌شده و سایر آثار نویسندگان همان مقالات. در مورد دستورالعمل‌های اخلاق هوش مصنوعی، فقط اسنادی را جمع آوری کردیم که به زبان انگلیسی- (یا با ترجمه رسمی به انگلیسی-) بودند و امکان مشاهده یا دانلود آن‌ها از اینترنت وجود داشت. فهرست کاملی از این دستورالعمل‌های اخلاق هوش مصنوعی همراه با لینک‌های URLدر مواد تکمیلی این مقاله ارائه شده است.

بخش ۳ - مسائل اخلاقی و ریسک‌های هوش مصنوعی

برای پرداختن به مشکلات اخلاقی هوش مصنوعی، ابتدا باید مسائل اخلاقی یا ریسک‌های بالقوه‌ای را که هوش مصنوعی ممکن است به همراه داشته باشد، شناسایی و درک کنیم. سپس می‌توان دستورالعمل‌ها، سیاست‌ها، اصول، و قواعد اخلاقی لازم (یعنی اخلاق هوش مصنوعی) را به طور مناسب تدوین کرد. با داشتن اخلاق مناسب برای هوش مصنوعی، می‌توان سیستم‌های هوش مصنوعی‌ای طراحی و ایجاد کرد که به طور اخلاقی رفتار کنند (یعنی هوش مصنوعی اخلاقی). مسئله اخلاقی در هوش مصنوعی معمولاً به چیزهای غیراخلاقی یا نتایج مشکل‌زای مرتبط با هوش مصنوعی اشاره دارد (یعنی مسائل و ریسک‌هایی که از توسعه، استقرار، و استفاده از هوش مصنوعی ناشی می‌شوند) که باید به آن‌ها پرداخته شود. بسیاری از مسائل اخلاقی، مانند کمبود شفافیت، حریم خصوصی و مسئولیت‌پذیری، تبعیض و جانبداری، مشکلات ایمنی و امنیت، و همچنین استفاده‌های جنایی و مخرب شناسایی شده‌اند.

این بخش بر مسائل و ریسک‌های اخلاقی هوش مصنوعی تمرکز دارد. ابتدا، در بخشIII-A ، چهار دسته‌بندی مختلف از مسائل اخلاقی هوش مصنوعی در ادبیات موجود بررسی می‌شود. از آنجا که این چهار دسته‌بندی یا برخی مسائل اخلاقی را نادیده می‌گیرند یا بیش از حد پیچیده هستند، در بخش III-B یک دسته‌بندی جدید پیشنهاد شده است که مسائل اخلاقی هوش مصنوعی را در سه سطح فردی، اجتماعی، و زیست‌محیطی طبقه‌بندی می‌کند. این دسته‌بندی پیشنهادی تمام مسائل اخلاقی موجود را به صورت جامع پوشش می‌دهد و درک و تحلیل مشکلات اخلاقی ناشی از هوش مصنوعی را تسهیل می‌کند. علاوه بر این، در بخش III-C تلاش می‌کنیم مسائل اخلاقی مرتبط با مراحل چرخه عمر سیستم‌های هوش مصنوعی را ترسیم کنیم، که این امر می‌تواند در شناسایی این مسائل در طی فرایند توسعه سیستم‌های هوش مصنوعی مفید باشد.

هدف اصلی این بخش بحث و روشن‌سازی مسائل اخلاقی هوش مصنوعی است تا متخصصان بتوانند این مسائل را شناسایی و درک کنند و سپس به آن‌ها در مطالعه بیشتر برای حل مشکلات اخلاقی هوش مصنوعی کمک شود. مشارکت اصلی این بخش پیشنهاد یک دسته‌بندی جدید برای مسائل اخلاقی هوش مصنوعی است که مسائل اخلاقی موجود را به شیوه‌ای روشن و آسان برای درک پوشش می‌دهد. علاوه بر این، مسائل اخلاقی مرتبط با مراحل چرخه عمر سیستم‌های هوش مصنوعی مورد بحث قرار گرفته است.

1.3. مرور دسته‌بندی‌های مسائل اخلاقی هوش مصنوعی

این بخش به بررسی نگرانی‌ها یا مسائل اخلاقی هوش مصنوعی از دیدگاه‌های مختلف می‌پردازد و چهار دسته‌بندی متفاوت که در ادبیات جمع‌آوری‌شده یافت شده‌اند، مرور می‌کند. دو مورد از این دسته‌بندی‌ها از گزارش‌های دولتی و دو مورد دیگر از انتشارات علمی گرفته شده‌اند. از دیدگاه‌ها و دسته‌بندی‌های مختلف، مسائل اخلاقی مرتبط نیز تا حدی متفاوت هستند. در ادامه، چهار دسته‌بندی مختلف از مسائل اخلاقی هوش مصنوعی به ترتیب مرور می‌شوند. چهار دسته‌بندی مورد بررسی و دسته‌بندی پیشنهادی ما در جدول I ذکر شده‌اند.

1.1.3. دسته‌بندی بر اساس ویژگی‌های هوش مصنوعی، عوامل انسانی و تأثیر اجتماعی

در مرجع [۱۱]، مسائل اخلاقی هوش مصنوعی به سه دسته اصلی تقسیم شده‌اند: مسائل اخلاقی ناشی از ویژگی‌های هوش مصنوعی، ریسک‌های اخلاقی ناشی از عوامل انسانی، و تأثیر اجتماعی مسائل اخلاقی هوش مصنوعی.

الف. مسائل اخلاقی ناشی از ویژگی‌های هوش مصنوعی

شفافیت :یادگیری ماشین (ML) فناوری اصلی هوش مصنوعی کنونی، به ویژه شبکه‌های عصبی عمیق است. با این حال، فرایند استنباط یادگیری ماشین، که معمولاً به عنوان "جعبه سیاه" شناخته می‌شود، به سختی قابل توضیح و درک است. این عدم شفافیت باعث می‌شود الگوریتم‌ها یا مدل‌ها برای کاربران و حتی توسعه‌دهندگان مرموز به نظر برسند. این مسئله به طور مستقیم به مشکل شفافیت منجر می‌شود. کمبود شفافیت نه تنها مشکلات توضیحی ایجاد می‌کند، بلکه مانع نظارت و هدایت انسانی بر یادگیری ماشین یا هوش مصنوعی نیز می‌شود. بنابراین، شفافیت یا توضیح‌پذیری یکی از معایب به‌شدت مورد بحث هوش مصنوعی است.

امنیت داده و حریم خصوصی :عملکرد هوش مصنوعی کنونی به شدت به داده‌های آموزشی وابسته است. معمولاً مقدار زیادی داده، که احتمالاً شامل داده‌های شخصی- و خصوصی است، برای آموزش یک مدل هوش مصنوعی (به ویژه مدل‌های یادگیری عمیق) مورد نیاز است. سوءاستفاده و استفاده مخرب از داده‌ها، مانند نشت اطلاعات شخصی- یا دستکاری، از مسائل اخلاقی جدی است که به شدت به افراد، مؤسسات، سازمان‌ها و حتی کشورها مربوط می‌شود. امنیت داده و حریم خصوصی از مسائل کلیدی در توسعه و کاربرد فناوری هوش مصنوعی هستند.

استقلال، قصد و مسئولیت‌پذیری :با پیشرفت هوش مصنوعی، سیستم‌ها یا عامل‌های هوش مصنوعی کنونی، مانند ربات‌های مراقبت بهداشتی، دارای درجه‌ای از استقلال، قصد، و مسئولیت‌پذیری هستند. استقلال هوش مصنوعی به توانایی یک سیستم هوش مصنوعی برای عمل بدون دخالت یا کنترل مستقیم انسانی اشاره دارد. قصد به توانایی یک سیستم هوش مصنوعی برای انجام اعمالی که می‌توانند اخلاقاً مضر- یا مفید باشند اشاره دارد. مسئولیت‌پذیری نشان می‌دهد که سیستم هوش مصنوعی برخی از قوانین اجتماعی و مسئولیت‌های فرضی را برآورده می‌کند. اما اینکه یک سیستم هوش مصنوعی تا چه حد باید استقلال، قصد، و مسئولیت‌پذیری داشته باشد، یک سؤال و مسئله چالش‌برانگیز است.

ب. مسائل اخلاقی ناشی از عوامل انسانی

مسئولیت‌پذیری: وقتی یک سیستم یا عامل هوش مصنوعی در انجام یک وظیفه مشخص شکست بخورد و نتایج نامطلوبی به بار آورد، چه کسی- باید پاسخگو باشد؟ نتیجه نامطلوب ممکن است به دلایل مختلفی، از جمله کدهای برنامه‌نویسی-، داده‌های ورودی، عملیات نادرست یا عوامل دیگر ایجاد شده باشد. این امر به مشکل معروف به "مسئله دست‌های متعدد" منجر می‌شود. بنابراین، مسئولیت‌پذیری یکی از مسائل اخلاقی مربوط به عوامل انسانی در طراحی، پیاده‌سازی، استقرار، و استفاده از هوش مصنوعی است.

استانداردهای اخلاقی: از آنجا که هدف نهایی اخلاق هوش مصنوعی ایجاد هوش مصنوعی‌ای است که بتواند از اصول اخلاقی پیروی کند و به طور اخلاقی رفتار نماید، ضروری است که استانداردهای جامع و بی‌طرف اخلاقی برای آموزش یا تنظیم هوش مصنوعی ایجاد شود. برای تدوین استانداردهای اخلاقی برای هوش مصنوعی، پژوهشگران و متخصصان باید نظریه‌ها و اصول اخلاقی موجود را به خوبی درک کنند.

قوانین حقوق بشر-: طراحان، مهندسان نرم‌افزار و سایر افرادی که در طراحی و کاربرد سیستم‌های هوش مصنوعی مشارکت دارند، باید قوانین حقوق بشر- را فرا گیرند. بدون آموزش در زمینه قوانین حقوق بشر-، ممکن است به صورت ناآگاهانه حقوق اساسی بشر- را نقض کنند. قوانین یا اسناد حقوق بشر- که در کشورهای مختلف یا مناطق مختلف دنبال می‌شوند، اغلب با یکدیگر متفاوت هستند. قوانین حقوق بشر- متعددی، مانند **حقوق بین‌المللی بشر-، میثاق بین‌المللی حقوق مدنی و سیاسی، میثاق بین‌المللی حقوق اقتصادی، اجتماعی**

و فـرهنگی، اعلامیـه جهـانی حقـوق بشـر، منشـور سـازمان ملـل متحـد و **کنوانسـیون اروپـایی حفاظـت از حقـوق بشر و آزادی‌های بنیادین**، توسط دولت‌های مختلف منتشر شده‌اند.

ج) تأثیر اجتماعی مسائل اخلاق هوش مصنوعی

اتوماسیون و جایگزینی شغل

بـا جـایگزینی تعـداد بیشـتری از کـارگران کارخانـه بـا سیسـتم‌های خودکـار و ربات‌هـا، هـوش مصـنوعی بـازار کـار را دچـار اختلال و تحول می‌کند. بنابراین، بسیاری از افراد نگران اتوماسیون و جایگزینی شغلی هستند.

دسترس‌پذیری

دسـترس‌پذیری یـا قابلیـت اسـتفاده از فناوری‌هـای نوظهـور، ماننـد هـوش مصـنوعی، تـأثیر مسـتقیمی بـر رفـاه انسـان دارد. بـا ایـن حـال، اگـر تنهـا بخشی‌ـ از جمعیـت از مزایـای هـوش مصـنوعی بهره‌منـد شـوند، ایـن امـر غـیراخلاقی و ناعادلانـه خواهـد بـود. بایـد توجـه شـود کـه محصـولات و خـدمات هـوش مصـنوعی بـه گونـه‌ای توسـعه یابنـد کـه برای همه قابل دسترس باشند، به طوری که مزایای هوش مصنوعی به طور مساوی در میان همه گسترش یابد.

دموکراسی و حقوق مدنی

هـوش مصـنوعی غـیراخلاقی حقیقـت را تحریـف می‌کنـد و در نهایـت منجـر بـه از دسـت دادن اعتمـاد و حمایـت عمـومی از فنـاوری هـوش مصـنوعی می‌شـود. قـدرت دموکراسی‌هـا بـا از دسـت دادن جوامـع آگـاه و اعتمادکننـده آسـیب می‌بینـد. بـا آسـیب‌دیدن دموکراسی‌هـا و تشـدید تعصـب‌های سـاختاری، بهره‌منـدی آزادانـه از حقـوق مـدنی دیگـر بـه طـور مـداوم بـرای همـه قابـل دسـترس نخواهـد بـود. بنـابراین، دمـوکراسی و حقـوق مـدنی بایـد در اخـلاق هوش مصنوعی مدنظر قرار گیرند.

2.1.3. دسته‌بندی بر اساس آسیب‌پذیری‌های هوش مصنوعی و انسان
در مرجع[29] ، **لیائو** مسائل اخلاق هوش مصنوعی را به دو دسته تقسیم می‌کند:

۱. مسـائل اخـلاقی کـه بـه دلیـل محـدودیت‌های سیسـتم‌های یـادگیری ماشـین (ML) کنـونی بـه وجـود می‌آینـد و به عنوان "آسیب‌پذیری‌های هوش مصنوعی (به ویژه یادگیری ماشین)" شناخته می‌شوند.

۲. مسـائل اخـلاقی کـه بـه دلیـل عملکـرد بـیش از حـد خـوب سیسـتم‌های یـادگیری ماشـین کنـونی بـه وجـود می‌آینـد و انسـان‌ها در حضـور یـا تعامـل بـا ایـن سیسـتم‌های هوشـمند آسـیب‌پذیر می‌شـوند، کـه بـه عنـوان "آسیب‌پذیری‌های انسانی" مطرح می‌شوند.

الف. مسائل اخلاقی ناشی از آسیب‌پذیری‌های هوش مصنوعی

یـادگیری ماشـین تشـنه‌ی داده اسـت: معمـولاً یـادگیری ماشـین بـه مقـدار زیـادی داده نیـاز دارد تـا عملکـرد خـوبی داشـته باشـد. ایـن امـر شرکت‌هـا و سـازمان‌ها را بـه جمـع‌آوری یـا خریـد داده، از جملـه داده‌هـای حسـاس شخصی‌ـ، حتی اگر این کار ممکن است حق حریم خصوصی فرد را نقض کند، ترغیب می‌کند.

زبالـه وارد شـود، زبالـه خـارج می‌شـود: عملکـرد الگـوریتم یـادگیری ماشـین بـه شـدت بـه داده‌هـایی کـه از آن‌هـا یـاد می‌گیرد وابسـته اسـت. اگـر یـک الگـوریتم یـادگیری ماشـین بـا داده‌هـای ناکافی یـا نادرسـت آمـوزش داده شـود، حـتی اگـر طراحی خوبی داشته باشد، نتایج نامطلوبی ارائه خواهد کرد.

الگوریتم‌های معیوب: حتی اگر یک الگوریتم یادگیری ماشین با داده‌های کافی و دقیق وارد شود، اگر خود الگوریتم بد باشد، پیش‌بینی‌های نادرستی ارائه می‌دهد. برای مثال، یک الگوریتم بد ممکن است نتواند یک الگو را تشخیص دهد حتی اگر وجود داشته باشد، یا ممکن است یک الگو را شناسایی کند حتی اگر وجود نداشته باشد، که به ترتیب به عنوان "کم‌برازش" و "بیش‌برازش" شناخته می‌شوند.

یادگیری عمیق یک جعبه‌ی سیاه است: یادگیری عمیق یک جعبه سیاه است که مسائلی مانند توضیح‌پذیری، تفسیرپذیری و اعتماد را ایجاد می‌کند. حتی برای طراحان و توسعه‌دهندگان یادگیری عمیق، این مدل غیرقابل درک است، زیرا معمولاً شامل هزاران یا میلیون‌ها ارتباط بین نورون‌های مختلف است. بنابراین، توضیح چگونگی تعامل این ارتباطات و دلیل پیش‌بینی‌های خاص مدل دشوار است.

ب. مسائل اخلاقی ناشی از آسیب‌پذیری‌های انسانی

سوءاستفاده از هوش مصنوعی: فناوری‌های هوش مصنوعی، مانند تشخیص چهره و تولید تصویر، می‌توانند بهتر از انسان عمل کنند. با این حال، مسائل اخلاقی وجود دارند زیرا مردم ممکن است به استفاده نادرست از این فناوری‌ها وسوسه شوند. برای مثال، یک دولت می‌تواند از فناوری تشخیص چهره برای نظارت بر شهروندان خود استفاده کند یا یادگیری ماشین می‌تواند برای ساخت عکس‌ها یا ویدیوهای جعلی که انسان‌ها نمی‌توانند واقعی یا جعلی بودن آن‌ها را تشخیص دهند، استفاده شود.

جایگزینی شغل: از آنجا که ربات‌های هوشمند می‌توانند وظایف خاصی را سریع‌تر و بهتر از انسان‌ها انجام دهند، بسیاری از افراد نگران هستند که ربات‌ها و سایر فناوری‌های هوش مصنوعی بخش بزرگی از نیروی کار انسانی را در آینده‌ای نزدیک جایگزین کنند.

مسائل مربوط به همدم‌های رباتیک: با پیچیده‌تر شدن ربات‌های هوش مصنوعی، آن‌ها به عنوان همدم‌های انسان در نظر گرفته شده‌اند. این امر مسائل اخلاقی‌ای را در مورد رابطه بین انسان و همدم‌های رباتیک ایجاد می‌کند.

3.1.3. دسته‌بندی بر اساس الگوریتم، داده‌ها، کاربرد و ریسک‌های بلندمدت و غیرمستقیم اخلاقی

در گزارش تحلیلی ریسک‌های اخلاقی هوش مصنوعی که توسط **گروه کاری استانداردسازی ملی هوش مصنوعی چین** منتشر شده است، مسائل اخلاقی هوش مصنوعی به چهار جنبه زیر تقسیم می‌شود:

- مسائل اخلاقی مرتبط با الگوریتم‌های هوش مصنوعی؛
- مسائل اخلاقی مرتبط با داده‌ها؛
- مسائل اخلاقی مرتبط با کاربرد هوش مصنوعی؛
- ریسک‌های بلندمدت و غیرمستقیم اخلاقی.

الف. مسائل اخلاقی مرتبط با الگوریتم‌ها

امنیت الگوریتم: الگوریتم‌های هوش مصنوعی چندین مشکل امنیتی ایجاد می‌کنند. نخست، خطر نشت الگوریتم یا مدل وجود دارد. به طور معمول، مدل از طریق آموزش با داده‌های آموزشی و بهینه‌سازی پارامترهایش به دست می‌آید. اگر پارامترهای مدل یک الگوریتم افشا شوند، یک طرف ثالث ممکن است بتواند

مدل را کپی کند. این امر به مالک مدل خسارات اقتصادی وارد می‌کند، زیرا طرف ثالث بدون پرداخت هزینه داده‌های آموزشی، مدل مشابهی به دست می‌آورد. دوم، پارامترهای مدل الگوریتم هوش مصنوعی ممکن است به صورت غیرقانونی توسط مهاجمان تغییر کنند، که این امر باعث کاهش عملکرد مدل هوش مصنوعی و ایجاد نتایج نامطلوب می‌شود. علاوه بر این، در بسیاری از سناریوها، خروجی مدل به شدت با امنیت شخصی- مرتبط است، مانند حوزه‌های پزشکی و رانندگی خودکار. در صورت وجود حفره‌ها یا اشتباهات در الگوریتم‌ها در این حوزه‌ها، آسیب‌های مستقیم به انسان وارد می‌شود و عواقب جدی به دنبال خواهد داشت.

توضیح‌پذیری الگوریتم: به دلیل ویژگی "جعبه سیاه" بسیاری از الگوریتم‌های یادگیری ماشین، به ویژه یادگیری عمیق یا شبکه‌های عصبی، فرایند تصمیم‌گیری الگوریتم‌های هوش مصنوعی دشوار است. تفسیرپذیری یا توضیح‌پذیری الگوریتم‌ها یک مسئله اخلاقی اساسی در هوش مصنوعی است، زیرا به حق انسان برای دانستن مربوط می‌شود.

معضل تصمیم‌گیری الگوریتمی: پس از به دست آوردن مدل هوش مصنوعی، نتایج الگوریتم معمولاً برای ما غیرقابل پیش‌بینی است. به عبارت دیگر، حتی اگر یک مدل هوش مصنوعی به خوبی طراحی شود، نمی‌توان تصمیمات الگوریتم و نتایج آن را پیش‌بینی کرد. این امر به معضل یا خطر تصمیم‌گیری الگوریتمی هوش مصنوعی منجر می‌شود. به عنوان مثال، خودروهای خودران باید تصادفات را کاهش دهند، اما گاهی باید بین دو گزینه بد انتخاب کنند، مانند برخورد با عابران پیاده یا فدا کردن خود و سرنشینان برای نجات عابران.

ب) مسائل اخلاقی مرتبط با داده‌ها

حفظ حریم خصوصی: با توسعه داده‌های بزرگ و هوش مصنوعی، تنش بین فناوری هوش مصنوعی و حفاظت از حریم خصوصی کاربران به طور فزاینده‌ای جدی‌تر شده است. مجرمان راه‌های بیشتری برای دسترسی به داده‌های خصوصی شخصی- با هزینه کمتر و سود بیشتر دارند. حوادث امنیتی داده‌ها در سال‌های اخیر به طور مکرر رخ داده است. حفاظت از حریم خصوصی به یک مسئله اخلاقی جدی و شناخته‌شده در استفاده از هوش مصنوعی تبدیل شده است.

شناسایی و پردازش اطلاعات شخص و حساس: قوانین و مقررات سنتی تنها بر حفاظت از اطلاعات شخصی- و حساس تمرکز دارند. اگر اطلاعات شخصی- یا حساس از طریق روش‌هایی مانند تصادفی‌سازی یا سنتز داده‌ها ناشناس شوند، دیگر به عنوان اطلاعات شخصی- یا حساس در نظر گرفته نمی‌شوند و تحت حفاظت قوانین سنتی قرار نمی‌گیرند. استفاده، اشتراک‌گذاری و انتقال چنین اطلاعاتی، مسائل اخلاقی جدیدی ایجاد می‌کند.

ج) مسائل اخلاقی مرتبط با کاربرد

تبعیض الگوریتمی: نتایج اجرای الگوریتم‌ها مستقیماً بر تصمیم‌گیری سیستم‌های هوش مصنوعی تأثیر می‌گذارد. با این حال، تبعیض یا تعصب الگوریتمی در بسیاری از کاربردهای هوش مصنوعی مشاهده شده است. برای مثال، تعصب نژادی در سیستم‌های عدالت کیفری یا تبعیض جنسیتی در استخدام.

سوءاستفاده از الگوریتم‌ها: سوءاستفاده از الگوریتم‌ها به وضعیتی اشاره دارد که افراد از الگوریتم‌ها برای تحلیل، تصمیم‌گیری، هماهنگی و فعالیت‌های دیگر استفاده می‌کنند، اما هدف استفاده، روش استفاده، یا دامنه استفاده نادرست بوده و اثرات منفی ایجاد می‌کند. برای مثال، الگوریتم‌های تشخیص چهره می‌توانند برای ارتقای امنیت عمومی و تسریع در کشف مظنونان جنایی استفاده شوند، اما اگر برای شناسایی مجرمان بالقوه یا

تعیین احتمال ارتکاب جرم بر اساس چهره فرد به کار گرفته شوند، این امر سوءاستفاده از الگوریتم محسوب می‌شود.

د) ریسک‌های بلندمدت و غیرمستقیم اخلاقی

اشتغال: با پیشرفت سریع و کاربرد گسترده هوش مصنوعی، کارهای بیشتری توسط برخی محصولات هوش مصنوعی انجام می‌شود. این امر تأثیر قابل‌توجهی بر مسئله اشتغال خواهد داشت.

مالکیت: با پیشرفت مداوم هوش مصنوعی، تفاوت‌های فکری بین عامل‌های هوش مصنوعی و انسان‌ها به تدریج کاهش می‌یابد. در نتیجه، بحث‌های متعددی درباره مالکیت مطرح می‌شود، مانند اینکه آیا عامل‌های هوش مصنوعی باید به عنوان "موضوع حقوقی" در نظر گرفته شوند و آیا محصولات هوش مصنوعی دارای حقوق مالکیت (کپی‌رایت یا حقوق ثبت اختراع) هستند یا خیر.

رقابت: رقابت ناعادلانه، رقابت مخرب و رفتارهای انحصارگرایانه با مزایای فناوری تأثیرات منفی بر ثبات اجتماعی، آزادی بازار، عدالت و ارزش برابر دارند و به شدت منافع مصرف‌کنندگان را آسیب می‌رسانند و بهبود رفاه اجتماعی را مانع می‌شوند.

مسئولیت‌پذیری: با کاربرد گسترده هوش مصنوعی، موارد بسیاری مشاهده شده است که محصولات هوش مصنوعی قوانین یا اخلاق را نقض کرده‌اند، مانند آسیب شخصی- یا تبعیض الگوریتمی. در چنین مواردی، مسئله اساسی این است که چه کسی- مسئول این عواقب بد است. برای مثال، در رانندگی خودکار که موضوعات مختلفی مانند مالک خودرو، راننده، سرنشینان، تولیدکنندگان خودرو و ارائه‌دهندگان سیستم خودران را درگیر می‌کند، مسئولیت‌ها در صورت وقوع تصادف چگونه تقسیم می‌شود.

4.1.3. دسته‌بندی بر اساس استقرار هوش مصنوعی
در جدیدترین مطالعه **خدمات پژوهشی- پارلمان اروپا** درباره پیامدهای اخلاقی و سؤالات اخلاقی مرتبط با هوش مصنوعی، مسائل اخلاقی بر اساس تأثیرات هوش مصنوعی بر جامعه انسانی، روان‌شناسی انسانی، سیستم مالی، سیستم قانونی، محیط زیست و سیاره، و اعتماد طبقه‌بندی شده‌اند.

الف) تأثیر بر جامعه

بازار کار: هوش مصنوعی در حال حاضر در بخش‌هایی مانند امور مالی، تولید پیشرفته، حمل‌ونقل، توسعه انرژی، مراقبت‌های بهداشتی و بسیاری حوزه‌های دیگر به کار گرفته شده است. اثرات اتوماسیون بر مشاغل کارگری یا "یقه آبی" به وضوح قابل مشاهده است. با پیشرفته‌تر شدن عامل‌های هوش مصنوعی یا ربات‌ها، تعداد بیشتری از شغل‌ها تحت تأثیر فناوری‌های هوش مصنوعی قرار می‌گیرند و بسیاری از موقعیت‌های شغلی از بین خواهند رفت. این امر می‌تواند خطر بیکاری گسترده را در بسیاری از بخش‌های شغلی به همراه داشته باشد.

نابرابری: فناوری‌های هوش مصنوعی انتظار می‌رود که عملیات تجاری شرکت‌ها را ساده‌تر و بهره‌وری را افزایش دهند. با این حال، برخی افراد معتقدند این کار به هزینه نیروی کار انسانی انجام خواهد شد. در نتیجه، درآمدها

در میان افراد کمتری توزیع می‌شود و صاحبان شرکت‌های مبتنی بر هوش مصنوعی از مزایای نامتناسبی بهره‌مند خواهند شد، که به افزایش نابرابری‌های اجتماعی منجر می‌شود.

حریم خصوصی، حقوق بشر و کرامت انسانی: دستیارهای شخصی هوشمند مانند **سیری اپل، اکوی آمازون** و **هوم گوگل** می‌توانند علایق و رفتار کاربران را بیاموزند، اما در عین حال نگرانی‌هایی درباره اینکه این دستگاه‌ها دائماً فعال و در حال گوش دادن هستند، مطرح می‌شود. همچنین، هوش مصنوعی می‌تواند برای تعیین باورهای سیاسی افراد استفاده شود، که ممکن است آن‌ها را در برابر دست‌کاری آسیب‌پذیر کند.

تعصب: تعصبات انسانی مانند تعصب جنسیتی یا نژادی می‌توانند به هوش مصنوعی منتقل شوند. تعصب هوش مصنوعی ممکن است ناشی از داده‌های آموزشی، ارزش‌های توسعه‌دهندگان یا کاربران، یا فرایند یادگیری خود هوش مصنوعی باشد.

دموکراسی: پیاده‌سازی و پذیرش هوش مصنوعی می‌تواند به چندین روش به دموکراسی آسیب برساند. از جمله تمرکز قدرت در دستان چند شرکت بزرگ، تأثیرگذاری بر انتخابات سیاسی، و قطبی‌شدن اجتماعی از طریق سیستم‌های توصیه خبر مبتنی بر هوش مصنوعی.

ب) تأثیر بر روان‌شناسی انسان

روابط: هوش مصنوعی در حال پیشرفت در تقلید از تفکر، تجربه، رفتار، و روابط انسانی است. این امر ممکن است بر روابط واقعی انسانی تأثیر بگذارد و مسائل اخلاقی جدیدی ایجاد کند.

شخصیت: با انجام وظایف و تصمیماتی که به طور سنتی توسط انسان انجام می‌شوند، این سؤال اخلاقی مطرح می‌شود که آیا سیستم‌های هوش مصنوعی باید دارای حقوق و شخصیت حقوقی شوند.

ج) تأثیر بر سیستم مالی

استفاده از هوش مصنوعی در بازارهای مالی به طور قابل توجهی کارایی تراکنش و حجم معاملات را بهبود بخشیده است. بازارها برای اتوماسیون، بسیار مناسب هستند، زیرا در حال حاضر تقریباً به طور کامل به صورت الکترونیکی کار می‌کنند و حجم عظیمی از داده‌ها با سرعت بالایی تولید می‌شود که نیاز به استفاده از الگوریتم‌هایی برای هضم و تجزیه و تحلیل آن دارد. علاوه بر این، به دلیل پویایی بازارها، واکنش سریع به اطلاعات بسیار مهم است [۵۹]، که انگیزه‌های قابل توجهی را برای جایگزینی فرآیند تصمیم گیری کُند افراد با تصمیم گیری الگوریتمی فراهم می‌کند. علاوه بر این، جوایز برای تصمیمات تجاری موثر قابل توجه است، که توضیح می‌دهد که چرا شرکت‌ها در فناوری هوش مصنوعی سرمایه گذاری زیادی کرده اند.

با این حال، عوامل معاملاتی خودکار مبتنی بر هوش مصنوعی نیز ممکن است به طور مخرب برای بی ثبات کردن بازارها یا آسیب رساندن به طرف‌های بی گناه از راه‌های دیگر استفاده شوند. حتی اگر قصد اصلی، مخرب بودن آنها نباشد. استقلال و انعطاف‌پذیری استراتژی‌های معاملاتی الگوریتمی، از جمله استفاده روزافزون از تکنیک‌های ML، پیش‌بینی عملکرد آنها در موقعیت‌های غیرمنتظره را برای افراد دشوار می‌کند.

د) تأثیر بر سیستم قانونی

حقوق جزا: بر اساس قوانین جزایی فعلی، جرم از دو عنصر تشکیل شده است، یعنی فعل (یا ترک فعل) اختیاری و قصد ارتکاب جرم. اگر نشان داده شود که محصولات یا ربات‌های هوش مصنوعی از هشیاری یا آگاهی کافی برخوردارند، ممکن است آنها مرتکب مستقیم جرایم جنایی یا مسئول جنایات یا سهل‌انگاری‌ها باشند. اگر بپذیریم که محصولات هوش مصنوعی دارای ذهن، اراده آزاد مانند انسان، استقلال یا حس اخلاق خاص خود هستند، در این صورت قوانین جزایی ما و حتی کل سیستم حقوق ما باید مورد بازنگری قرار گیرند [۶۰].

قانون شکنجه: قانون شکنجه شرایطی مانند آسیب رفتاری یک فرد، رنج، ضرر ناعادلانه یا آسیب رساندن به شخص دیگر را پوشش می‌دهد. هنگامی که تصادفی با خودرو(های) خودران رخ می‌دهد، دو حوزه قانونی مرتبط وجود دارد: سهل انگاری و مسئولیت محصول. در حالی که امروزه بیشتر تصادفات، ناشی از خطای راننده است که نشان می‌دهد مسئولیت تصادفات بر اساس اصل سهل انگاری تنظیم می‌شود. بنابراین، در آینده، قانون جرم، که شامل انواع مختلفی از دعاوی آسیب شخصی- است، به طور قابل توجهی تحت تأثیر قرار خواهد گرفت [۶۱] زیرا محصولات هوش مصنوعی (مانند اتومبیل‌های خودران یا سایر ربات‌های هوشمند) در دعاوی صدمات شخصی- دخیل خواهند بود. به عنوان تصادف بین اتومبیل‌های خودران یا ادعای جراحت که در آن ربات به انسان آسیب می‌رساند.

ه) تأثیر بر محیط زیست و سیاره

استفاده از منابع طبیعی: توسعه و کاربرد هوش مصنوعی تقاضای بسیاری از منابع طبیعی مانند فلزات خاکی کمیاب مانند نیکل، کبالت، گرافیت و غیره را افزایش می‌دهد. با کاهش عرضه‌ی موجود، اپراتورها ممکن است مجبور شوند در محیط‌های جدید و پیچیده‌تر برای استخراج کار کنند. این امر باعث افزایش میزان تولید و مصرف فلزات کمیاب خاکی و آسیب بیشتر به محیط زیست می‌شود [۶۲].

آلودگی و ضایعات: افزایش تولید و مصرف دستگاه‌های فناوری هوش مصنوعی مانند ربات‌ها باعث تشدید آلودگی و ضایعات مانند تجمع فلزات سنگین و مواد سمی در محیط می‌شود [۶۳].

نگرانی‌های انرژی: استفاده از فناوری هوش مصنوعی، به‌ویژه یادگیری عمیق، عموماً شامل آموزش مدل‌های ML بر روی حجم عظیمی از داده است که معمولاً مقادیر زیادی انرژی مصرف می‌کند. با توجه به داده‌های فهرست شده در [۶۴]، اثر کربن آموزش یک مدل پردازش زبان طبیعی (مدل ترانسفورماتور) تقریباً ۵ برابر اثر کربن یک ماشین متوسط در کل طول عمر آن است.

و) تأثیر بر اعتماد

هوش مصنوعی نوید تغییرات و مزایای متعددی را برای زندگی افراد و جامعه می‌دهد. این مسئله در حال تغییر زندگی روزمره ما در بسیاری از حوزه ها، مانند حمل و نقل، صنعت خدمات، مراقبت‌های بهداشتی، آموزش، ایمنی و امنیت عمومی، و سرگرمی است. با این وجود، این سیستم‌های هوش مصنوعی باید به گونه‌ای معرفی شوند که اعتماد و درک را تقویت کند و به حقوق بشر- و مدنی احترام بگذارد [۶۵]. اتفاق نظر در میان جامعه‌ی تحقیقاتی این است که اعتماد به هوش مصنوعی تنها از طریق انصاف، شفافیت، مسئولیت پذیری و مقررات (یا کنترل) حاصل می‌شود.

انصاف: اعتماد به هوش مصنوعی، باید منصفانه و بی طرفانه باشد. همانطور که تصمیمات بیشتر و بیشتری به هوش مصنوعی واگذار می‌شود، ما باید اطمینان حاصل کنیم که این تصمیمات عاری از تعصب و تبعیض

هستند [٦٦]. چه فیلتر کردن رزومه‌ها برای مصاحبه‌های شغلی باشد یا تصمیم گیری در مورد پذیرش در دانشگاه یا انجام رتبه بندی اعتباری برای شرکت‌های وام، اساساً ضروری است که تصمیمات اتخاذ شده توسط هوش مصنوعی منصفانه باشد.

شفافیت: شفافیت برای ایجاد اعتماد در هوش مصنوعی مهم است، زیرا باید دانست که چرا یک سیستم هوش مصنوعی تصمیم خاصی گرفته است، به خصوص اگر آن تصمیم باعث عواقب نامطلوب یا آسیب شود. با توجه به اینکه اتوپایلوت یک خودروی هوشمند منجر به تصادفات مرگبار متعددی شده است، واضح است که برای کشف چگونگی و چرایی وقوع این تصادفات و رفع نقص فنی یا عملیاتی، شفاف سازی ضروری است. مشخص نبودن هسته‌ی ML و ایهام در ML، که به جعبه سیاه معروف است، یکی از موانع اصلی شفافیت هوش مصنوعی است [٥١].

مسئولیت پذیری: مسئولیت پذیری [٦٧] تضمین می‌کند که اگر یک سیستم هوش مصنوعی مرتکب اشتباهی شود یا به کسی- آسیب برساند، می‌توان مسئولیت آن را بر عهده گرفت، خواه طراح باشد، توسعه دهنده یا شرکتی باشد که هوش مصنوعی را می‌فروشد. در صورت بروز خسارت، پاسخگویی برای ایجاد یک مکانیسم اصلاحی ضروری است تا قربانیان بتوانند غرامت کافی را دریافت کنند. بنابراین، پاسخگویی برای اطمینان از اعتماد هوش مصنوعی بسیار مهم است.

کنترل: موضوع دیگری که بر اعتماد عمومی به هوش مصنوعی تأثیر می‌گذارد، کنترل پذیری هوش مصنوعی است [٦٨]. این مسئله تا حد زیادی به ترس مردم از ایده‌ی "اَبَر هوش (super-intelligence)" مربوط می‌شود که به معنی افزایش هوش AI به حدی است که از توانایی‌های انسانی پیشی- می‌گیرد، ممکن است هوش مصنوعی کنترل منابع ما را به دست گرفته و از گونه‌های ما پیشی- بگیرد. حتی می‌تواند منجر به انقراض انسان شود. یک نگرانی مرتبط به این موضوع این است که حتی اگر یک عامل هوش مصنوعی به دقت طراحی شده باشد تا اهداف خود را با نیازهای انسان هماهنگ کند، ممکن است به تنهایی اهداف فرعی غیرقابل پیش بینی ایجاد کند. بنابراین، برای حفظ اعتماد به هوش مصنوعی، مهم است که انسان‌ها نظارت یا کنترل نهایی بر فناوری هوش مصنوعی داشته باشند.

طبقه‌بندی	کلاس	مسائل اخلاقی	بحث
طبقه‌بندی مسائل اخلاق هوش مصنوعی بر اساس ویژگی‌های هوش مصنوعی، عوامل انسانی و تأثیر اجتماعی [١١]	مسائل اخلاق ناشی از ویژگی‌های هوش مصنوعی، مسائل اخلاق ناشی از عوامل انسانی، تأثیر اجتماعی مسائل اخلاق هوش مصنوعی	شفافیت، امنیت داده‌ها و حریم خصوصی، خودمختاری، قصد و مسئولیت‌پذیری؛ پاسخگویی، استانداردهای اخلاق، قوانین حقوق بشر-؛ اتوماسیون و جایگزینی شغل، دسترسی، دموکراسی و حقوق مدنی	تأثیرات هوش مصنوعی بر محیط زیست، مانند مصرف منابع طبیعی و آلودگی محیط زیست، نادیده گرفته شده است.
طبقه‌بندی مسائل اخلاق هوش مصنوعی بر اساس آسیب‌پذیری‌های هوش مصنوعی و انسان [٢٩]	مسائل اخلاق ناشی از آسیب‌پذیری‌های هوش مصنوعی، مسائل اخلاق ناشی از آسیب‌پذیری‌های انسانی	هوش مصنوعی نیازمند داده‌های زیاد است، ورود داده‌های نامعتبر/خروج داده‌های نامعتبر، الگوریتم‌های معیوب، یادگیری عمیق به عنوان جعبه سیاه؛ سوءاستفاده از هوش مصنوعی، جایگزینی شغل، مسائل	چندین مسئله مهم، مانند مسئولیت‌پذیری، ایمنی، آزادی، و مشکلات زیست‌محیطی، نادیده گرفته شده‌اند.

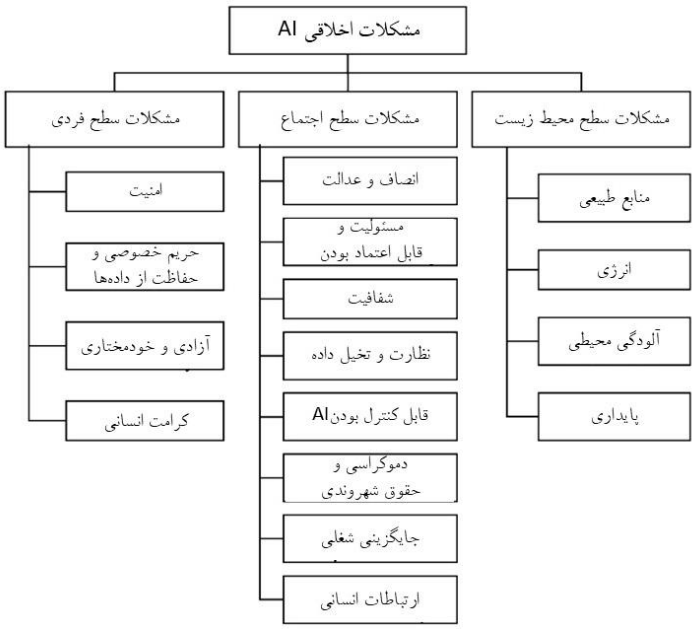
		مربوط به همراهان روباتیک	
طبقه‌بندی مسائل اخلاق هوش مصنوعی بر اساس مسائل مربوط به الگوریتم، داده، کاربرد، و ریسک‌های بلندمدت و غیرمستقیم [۳۸]	مسائل اخلاق مربوط به الگوریتم، مسائل اخلاق مربوط به داده، مسائل اخلاق مربوط به کاربرد، ریسک‌های بلندمدت و غیرمستقیم	امنیت الگوریتم، قابلیت توضیح الگوریتم، معضلات تصمیم‌گیری الگوریتمی؛ حفاظت از حریم خصوصی، شناسایی و پردازش اطلاعات حساس شخصی؛ تبعیض الگوریتمی، سوءاستفاده از الگوریتم؛ اشتغال، مالکیت، رقابت، مسئولیت‌پذیری	مسائل مرتبط با پاسخگویی، عدالت، خودمختاری و آزادی، کرامت انسانی، مشکلات زیست‌محیطی شامل نمی‌شوند.
طبقه‌بندی مسائل اخلاق هوش مصنوعی بر اساس استقرار هوش مصنوعی [۵۱]	تأثیر بر جامعه، تأثیر بر روانشناسی انسانی، تأثیر بر سیستم مالی، تأثیر بر سیستم حقوقی، تأثیر بر محیط زیست و سیاره، تأثیر بر اعتماد	بازار کار، نابرابری، حریم خصوصی، حقوق بشر و کرامت انسانی، تعصب، دموکراسی؛ روابط، شخصیت انسانی؛ قوانین کیفری، قوانین حقوقی؛ استفاده از منابع طبیعی، آلودگی و زباله، نگرانی‌های انرژی؛ عدالت، شفافیت، پاسخگویی، کنترل	برخی مسائل، از جمله مسئولیت‌پذیری، ایمنی، و پایداری، حذف شده‌اند و این طبقه‌بندی پیچیده و دشوار درک است.
طبقه‌بندی ما: طبقه‌بندی مسائل اخلاق هوش مصنوعی در سطح فردی، اجتماعی و زیست‌محیطی	مسائل اخلاق در سطح فردی، مسائل اخلاق در سطح اجتماعی، مسائل اخلاق در سطح زیست‌محیطی	ایمنی، حریم خصوصی و حفاظت از داده‌ها، آزادی و خودمختاری، کرامت انسانی؛ عدالت و انصاف، مسئولیت‌پذیری و پاسخگویی، شفافیت، نظارت و داده‌سازی، کنترل‌پذیری هوش مصنوعی، دموکراسی و حقوق مدنی، جایگزینی شغل، روابط انسانی؛ منابع طبیعی، انرژی، آلودگی زیست‌محیطی، پایداری	طبقه‌بندی ما مسائل اخلاق هوش مصنوعی را از سطح فردی، اجتماعی و زیست‌محیطی دسته‌بندی می‌کند. این طبقه‌بندی نه‌تنها واضح و قابل فهم است، بلکه به طور جامع مسائل اخلاق مورد بحث را پوشش می‌دهد.

2.3. دسته بندی پیشنهادی ما: مسائل اخلاق در سطوح فردی، اجتماعی و محیطی

در بخش قبل، ما مسائل اخلاق هوش مصنوعی را که در بخش ادبیات تحقیق، شرح داده شده و طبقه بندی شده اند، مرور کرده ایم (جدول ۱ را ببینید). با این حال، دسته بندی‌های ارائه شده در بالا دارای نقص‌های آشکار هستند. به طور خاص، طبقه‌بندی بر اساس ویژگی‌های هوش مصنوعی، عوامل انسانی و تأثیرات اجتماعی [۱۱]، به وضوح تأثیر هوش مصنوعی بر محیط‌زیست، مانند مصرف منابع طبیعی و آلودگی محیط‌زیست را نادیده می‌گیرد. طبقه بندی بر اساس آسیب پذیری‌های هوش مصنوعی و انسان [۲۹] چندین موضوع مهم مانند مسئولیت، ایمنی و مشکلات زیست محیطی را حذف می‌کند. طبقه‌بندی بر اساس الگوریتم، داده‌ها، کاربردها و ریسک‌های اخلاق بلندمدت و غیرمستقیم [۳۸] ملاحظات انصاف، استقلال و آزادی،

کرامت انسانی، مشکلات زیست‌محیطی و غیره را نادیده می‌گیرد. اگرچه طبقه بندی بر اساس استقرار هوش مصنوعی [۵۱] مسائل اخلاقی را به طور جامع پوشش می‌دهد، این طبقه بندی بسیار دست و پا گیر است و برخی مسائل از جمله مسئولیت، ایمنی و پایداری حذف شده اند. این موضوع به ما انگیزه می‌دهد تا مسائل اخلاقی هوش مصنوعی را بیشتر تحلیل و مرتب کنیم.

بدون شک سیستم‌های هوش مصنوعی عمدتاً به افراد یا عموم جامعه خدمت ارائه می‌دهند. از این رو، می‌توانیم مسائل اخلاقی هوش مصنوعی را از منظر فردی و اجتماعی تحلیل و روشن کنیم. در عین حال، محصولات هوش مصنوعی به‌عنوان نهادهای روی کره‌ی زمین، ناگزیر بر محیط‌زیست تأثیر خواهند داشت. بنابراین، مسائل اخلاقی مربوط به جنبه‌های زیست محیطی نیز باید مورد توجه قرار گیرند. بنابراین، در این بخش، ما پیشنهاد کردیم که موضوعات اخلاقی هوش مصنوعی را در سه سطح مختلف، یعنی مسائل اخلاقی در سطوح فردی، اجتماعی و محیطی طبقه بندی کنیم. مسائل اخلاقی در سطح فردی عمدتاً شامل موضوعاتی می‌شود که پیامدهای نامطلوبی برای افراد انسانی، حقوق و رفاه آنها دارد [۶۹]. مسائل اخلاقی هوش مصنوعی در سطح اجتماعی، پیامدهای اجتماعی را در نظر می‌گیرد که هوش مصنوعی برای گروه‌ها یا جامعه‌ای از افراد به عنوان یک کل، به ارمغان آورده یا ممکن است به همراه داشته باشد [۶۹]. مسائل اخلاقی هوش مصنوعی در سطح زیست محیطی بر تأثیرات هوش مصنوعی بر محیط طبیعی متمرکز است. دسته بندی پیشنهادی ما در شکل ۲ نشان داده شده است.



شکل ۲. دسته بندی ارائه شده برای مشکلات اخلاقی AI

1.2.3. مسائل اخلاقی در سطح فردی

در سطح فردی، هوش مصنوعی بر ایمنی، حریم خصوصی، استقلال و کرامت انسانی افراد تأثیر گذاشته است. استفاده از هوش مصنوعی خطراتی را برای ایمنی افراد به همراه داشته است. به عنوان مثال، در چند سال گذشته تصادفات آسیب دیدگی با خودروهای خودران و ربات‌ها رخ داده و گزارش شده است. مسئله حریم خصوصی یکی از خطرات جدی است که هوش مصنوعی برای ما به ارمغان می‌آورد. برای دستیابی به عملکرد

خوب، سیستم‌های هوش مصنوعی معمولاً به مقدار زیادی داده نیاز دارند که اغلب شامل داده‌های خصوصی کاربران می‌شود. با این حال، خطرات جدی مرتبط با این مجموعه داده وجود دارد. یکی از مسائل اصلی حفظ حریم خصوصی و اطلاعات است. علاوه بر این، همانطور که در بخش قبل توضیح داده شد، استفاده از هوش مصنوعی ممکن است چالش‌هایی را برای حقوق بشر- مانند استقلال و کرامت به همراه داشته باشد. خودمختاری به ظرفیت تفکر، تصمیم گیری و عمل مستقل، آزادانه و بدون تأثیر دیگران اشاره دارد [۷۰]. کرامت انسانی که یکی از حقوق اصلی بشر- است، مربوط به حق احترام و رفتار اخلاقی انسان است [۷۱]. حفاظت از کرامت در زمینه هوش مصنوعی بسیار مهم است. کرامت انسانی باید یکی از مفاهیم اساسی برای محافظت از انسان در برابر آسیب باشد و هنگام توسعه فناوری‌های هوش مصنوعی باید مورد احترام قرار گیرد. به عنوان مثال، یک سیستم تسلیحاتی خودمختار مرگبار [۷۲] ممکن است اصل کرامت انسانی را نقض کند.

2.2.3. مسائل اخلاقی در سطح اجتماعی

هنگام در نظر گرفتن مسائل اخلاقی هوش مصنوعی در سطح اجتماعی، ما عمدتاً بر پیامدها و تأثیرات گسترده‌ای که هوش مصنوعی برای جامعه و رفاه جوامع و ملل در سراسر جهان به ارمغان می‌آورد تمرکز می‌کنیم. در طبقه بندی مسائل اخلاقی در سطح اجتماعی، ما در مورد انصاف و عدالت، مسئولیت و پاسخگویی، شفافیت، نظارت و اطلاعات، قابلیت کنترل هوش مصنوعی، دموکراسی و حقوق مدنی، جایگزینی شغل و روابط انسانی بحث می‌کنیم.

وجود تعصب و تبعیض در هوش مصنوعی چالش‌هایی را برای انصاف و عدالت ایجاد کرده است. سوگیری‌ها و تبعیض‌های تعبیه شده در هوش مصنوعی ممکن است شکاف‌های اجتماعی را افزایش داده و به گروه‌های اجتماعی خاصی آسیب برساند [۷۰]. به عنوان مثال، در سیستم عدالت کیفری ایالات متحده، الگوریتم‌های هوش مصنوعی که برای ارزیابی خطر ارتکاب جرم استفاده می‌شوند، مورد توجه قرار گرفته‌اند که سوگیری نژادی را نشان می‌دهند [۷۳]. مسئولیت به معنای مسئول چیزی بودن یا در چیزی مسئول بودن است. بر اساس این مفهوم، مسئولیت پذیری در اصل یعنی کسی- که از نظر حقوقی یا سیاسی مسئول خسارت است، باید نوعی توجیه یا جبران خسارت ارائه کند. با مسئولیت ارائه راه حل‌های حقوقی منعکس می‌شود [۷۰]. بنابراین، مکانیسم‌هایی باید ایجاد شود تا از مسئولیت و پاسخگویی سیستم‌های هوش مصنوعی و نتایج آن‌ها قبل و بعد از اجرای آنها اطمینان حاصل شود. به دلیل ماهیت جعبه سیاه الگوریتم‌های هوش مصنوعی، عدم شفافیت به یکی از موضوعاتی تبدیل شده است که به طور گسترده مورد بحث قرار گرفته است. شفافیت، یعنی درک نحوه عملکرد سیستم‌های هوش مصنوعی، برای پاسخگویی نیز بسیار مهم است. در عصری که ما زندگی می‌کنیم که به آن عصر- دیجیتال و هوشمند نیز می‌گویند، نظارت و اطلاعات [۷۴]، از دغدغه‌های رایج هستند. داده‌ها از زندگی روزمره کاربران از طریق دستگاه‌های هوشمند جمع آوری می‌شود و ما تحت نظارت انبوه، زندگی می‌کنیم. از آنجایی که قدرت هوش مصنوعی به سرعت افزایش یافته است، توسعه سیستم‌های هوش مصنوعی باید دارای پادمان‌هایی باشد تا از کنترل‌پذیری سیستم‌های هوش مصنوعی توسط انسان اطمینان حاصل شود. سایر موضوعاتی که قبلاً مورد بحث قرار گرفت، از جمله دموکراسی و حقوق مدنی، جایگزینی شغل و روابط انسانی نیز در این دسته قرار می‌گیرند.

3.2.3. مسائل اخلاقی در سطح محیطی

مسائل اخلاقی هوش مصنوعی در سطح زیست محیطی بر تأثیرات هوش مصنوعی بر محیط زیست و سیاره‌ی زمین تمرکز دارد. هوش مصنوعی می‌تواند راحتی زیادی را برای زندگی ما به ارمغان بیاورد و می‌تواند به ما در

مقابله با برخی چالش‌ها کمک کند، اما برای سیاره زمین نیز هزینه دارد. کاربرد گسترده هوش مصنوعی اغلب مستلزم استقرار تعداد زیادی دستگاه پایانه سخت افزاری از جمله تراشه ها، حسگرها، دستگاه‌های ذخیره سازی و غیره است. تولید این سخت افزارها منابع طبیعی زیادی به خصوص برخی عناصر کمیاب را مصرف می‌کند. علاوه بر این، در پایان چرخه عمر، این سخت افزارها معمولاً دور ریخته می‌شوند که باعث آلودگی جدی زیست محیطی خواهند شد. جنبه مهم دیگر این است که سیستم‌های هوش مصنوعی معمولاً به قدرت محاسباتی قابل توجهی نیاز دارند که با مصرف انرژی بالا همراه است. علاوه بر این، از دیدگاه بلندمدت و جهانی، توسعه هوش مصنوعی باید پایدار باشد، یعنی فناوری هوش مصنوعی باید اهداف توسعه انسانی را برآورده کند و در عین حال توانایی سیستم‌های طبیعی را برای ارائه منابع طبیعی و خدمات اکوسیستمی که اقتصاد و جامعه به آن وابسته است ، حفظ کند [۲]. به طور خلاصه، مصرف منابع طبیعی، آلودگی محیط زیست، هزینه‌های مصرف انرژی و پایداری درگیر در توسعه هوش مصنوعی مسائل و نگرانی‌های اصلی در سطح زیست محیطی هستند.

دسته بندی پیشنهادی ما مسائل اخلاق را از سه سطح اصلی، یعنی تأثیر هوش مصنوعی بر فرد، جامعه و محیط، روشن می‌کند. مهم نیست که هوش مصنوعی در چه زمینه یا بخشی- استفاده می‌شود، می‌توانیم مسائل اخلاق مربوطه را از این سه سطح در نظر بگیریم. بدیهی است که این روش طبقه بندی ساده و واضح است و به طور جامع مسائل اخلاق هوش مصنوعی را پوشش می‌دهد.

3.3. مسائل اخلاق کلیدی مرتبط با هر مرحله از چرخه حیات سیستم هوش مصنوعی

پس از بررسی مسائل اخلاق و خطرات مورد بحث در ادبیات، ما در مورد مسائل اخلاق مرتبط با مراحل مختلف چرخه حیات یک سیستم هوش مصنوعی بحث می‌کنیم. اگر بدانیم مشکلات اخلاق موجود در کدام مراحل یا مراحل چرخه حیات سیستم هوش مصنوعی ایجاد می‌شوند یا مطرح می‌شوند، در رفع این مشکلات بسیار مفید خواهد بود. این موضوع انگیزه‌ای برای بحث در مورد مسائل اخلاق بالقوه در هر مرحله از چرخه حیات یک سیستم هوش مصنوعی است.

چرخه عمر کلی یا فرآیند توسعه یک سیستم هوش مصنوعی مبتنی بر ML [۷۵] یا محصول [۷۶]، اغلب شامل مراحل زیر است: تجزیه و تحلیل کسب و کار، مهندسی داده، مدل سازی ML، استقرار مدل، و بهره برداری و نظارت. معمولاً چرخه عمر محصولات هوش مصنوعی از تجزیه و تحلیل کسب و کار شروع می‌شود که عمدتاً شامل شناسایی و درک مشکل تجاری‌ای که باید حل شود و معیارهای تجاری (یا معیارهای موفقیت) است. این معیارها باید شامل معیارهای عملکرد مدل و همچنین شاخص‌های عملکرد کلیدی کسب و کار باشد تا با استفاده از مدل‌های هوش مصنوعی بهبود یابد. گام بعدی در مورد مهندسی داده است که به جمع آوری داده ها، برچسب گذاری داده ها، پاکسازی داده ها، ساختار داده ها، مهندسی ویژگی‌ها و سایر عملیات مربوط به داده‌ها می‌پردازد. پس از این، ادامه‌ی فرآیند، وارد مرحله‌ی به اصطلاح مدل سازی ML می‌شود. این مرحله به طور کلی شامل فرآیند تکراری طراحی یا انتخاب الگوریتم، آموزش مدل و ارزیابی مدل است. اگر ساخت مدل رضایت‌بخش باشد، ادامه‌ی فرآیند، به مرحله استقرار مدل می‌رود که مدل ML را در دسترس سایر سیستم‌های درون سازمان یا وب قرار می‌دهد تا مدل بتواند داده‌ها را دریافت کند و پاسخ آنها را برگرداند. مرحله عملیات و نظارت شامل عملیات سیستم هوش مصنوعی و ارزیابی مداوم عملکرد و تأثیرات آن است. این مرحله مشکلات را شناسایی کرده و سیستم هوش مصنوعی را با بازگشت به مراحل دیگر یا در صورت لزوم کنار گذاشتن سیستم هوش مصنوعی از تولید، یا تنظیم می‌کند یا تکامل می‌دهد.

ما سعی می‌کنیم نقشه‌ای ایجاد کنیم که مسائل اخلاق را با مراحل چرخه‌ی عمر هوش مصنوعی مرتبط می‌کند، جایی که این ارتباط به این معنی است که موضوع اخلاق در مرحله‌ی خاصی از چرخه عمر هوش مصنوعی

بیشتر رخ می‌دهد، یا اغلب به دلایلی در این مرحله ایجاد می‌شود. این نقشه برداری در جدول ۲ ارائه شده است، جایی که چندین مشکل اخلاقی حیاتی با پنج مرحله چرخه حیات هوش مصنوعی مرتبط است. این نقشه برداری برای پرداختن به مشکل اخلاقی به روشی فعال در طول فرآیند طراحی یک سیستم هوش مصنوعی مفید خواهد بود.

مرحله چرخه حیات هوش مصنوعی	ملاحظات اخلاقی در طول مرحله
تحلیل کسب‌وکار	شفافیت، عدالت (آیا معماری محصول هوش مصنوعی طراحی‌شده شامل متغیرها، ویژگی‌ها یا فرآیندهایی است که غیرمنطقی، غیراخلاقی یا غیرقابل توجیه هستند؟)، مسئولیت‌پذیری و پاسخگویی، دموکراسی و حقوق مدنی، پایداری
مهندسی داده	حریم خصوصی (چگونه امنیت داده‌ها و حفظ اطلاعات خصوصی و حساس موجود در مجموعه داده‌ها تضمین شود؟)، شفافیت (چگونه می‌توان روش‌های جمع‌آوری داده‌ها را برای مصرف‌کنندگان شفاف کرد؟)، عدالت (آیا داده‌ها به درستی نماینده، مرتبط، دقیق و قابل تعمیم هستند؟)، دموکراسی و حقوق مدنی (چگونه به کاربران نهایی امکان کنترل استفاده از داده‌هایشان را می‌دهید؟)
مدل‌سازی یادگیری ماشین	شفافیت (آیا فرآیند تصمیم‌گیری یا استنتاج مدل قابل درک است؟)، ایمنی (دقت، قابلیت اطمینان، امنیت و استحکام مدل)، عدالت (آیا خروجی‌های مدل نتایج متفاوتی برای گروه‌های مختلف افراد نشان می‌دهند؟)
استقرار مدل	حریم خصوصی (اطمینان حاصل شود که اطلاعات خصوصی از طریق مدل مستقر قابل شناسایی مجدد نیستند)، ایمنی (چگونه می‌توان از ایمنی مدل مستقر در برابر تغییرات مخرب و حملات محافظت کرد؟)
عملیات و نظارت	حریم خصوصی (حریم خصوصی باید در طول فرآیند عملیات و نظارت تضمین شود)، عدالت (آیا محصول هوش مصنوعی تأثیرات تبعیض‌آمیز یا ناعادلانه بر افرادی که تحت تأثیر قرار می‌گیرند دارد؟)، دموکراسی و حقوق مدنی (حقوق مدنی یا حقوق کاربران را نقض نکنید)

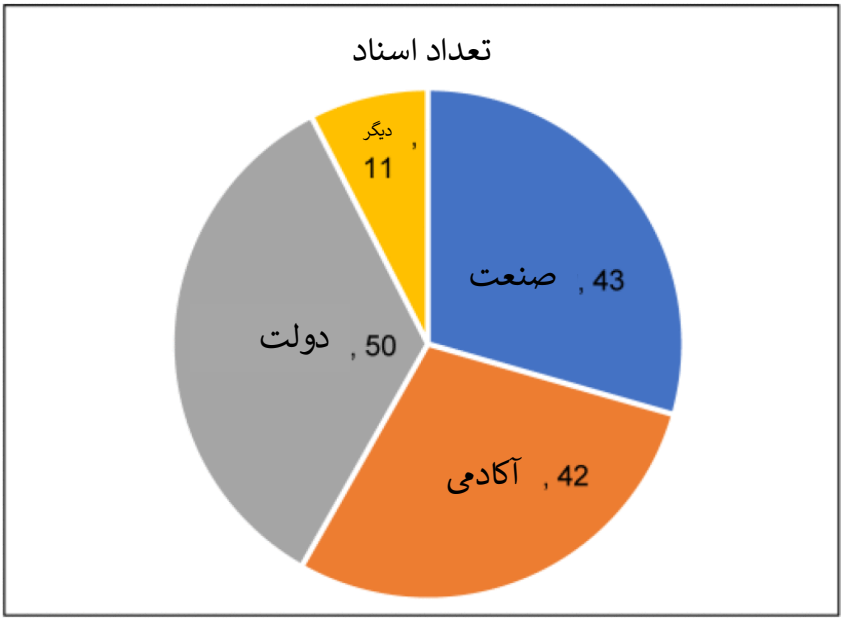
بخش ۴ - رهنمودها و اصول اخلاقی برای هوش مصنوعی

از آنجایی که مسائل اخلاقی هوش مصنوعی بیشتر و بیشتر مورد توجه و بحث‌های بخش‌های مختلف جامعه قرار گرفته است، بسیاری از سازمان‌ها (از جمله دانشگاه، صنعت و دولت) شروع به بحث و جست‌وجوی چارچوب‌ها، دستورالعمل‌ها و اصول ممکن برای حل مسائل اخلاقی هوش مصنوعی کرده‌اند [۷۸]. این اصول و دستورالعمل ها، دستورالعمل‌های مفیدی را برای تمرین هوش مصنوعی اخلاقی ارائه می‌دهند. این بخش به ارائه‌ی یک چشم انداز جهانی به روز از دستورالعمل‌ها و اصول اخلاق هوش مصنوعی اختصاص دارد که از طریق بررسی ۱۴۶ گزارش دستورالعمل و توصیه مربوط به اخلاق هوش مصنوعی منتشر شده توسط شرکت ها، سازمان‌ها و دولت‌ها از سال ۲۰۱۵ در سراسر جهان به دست می‌آید. این اصول و دستورالعمل ها،

راهنمایی‌های سطح بالایی را برای برنامه ریزی، توسعه، تولید، و استفاده از هوش مصنوعی و نیز دستورالعمل‌هایی برای پرداختن به مسائل اخلاقی هوش مصنوعی ارائه می‌دهند.

1.4. رهنمودهایی برای اخلاق هوش مصنوعی

یک بررسی و تجزیه و تحلیل عالی از اصول و دستورالعمل‌های فعلی در مورد هوش مصنوعی اخلاقی در سال ۲۰۱۹ توسط Jobin و همکاران ارائه شده است [۱۲]، که بررسی ۸۴ دستورالعمل اخلاقی منتشر شده توسط سازمان‌های ملی یا بین المللی از کشورهای مختلف را انجام داده‌اند. جوبین و همکاران [۱۲] توافق گسترده و قوی در مورد پنج اصل کلیدی، یعنی شفافیت، عدالت و انصاف، عدم سوء استفاده، مسئولیت و حریم خصوصی در میان بسیاری از آنها یافتند. با این حال، بسیاری از دستورالعمل‌ها و توصیه‌های جدید برای اخلاق هوش مصنوعی در دو سال گذشته منتشر شده‌اند که باعث شده است مقاله Jobin منسوخ شود زیرا بسیاری از اسناد مهم گنجانده نشده‌اند. به عنوان مثال، در ۲۴ نوامبر ۲۰۲۱، یونسکو (سازمان آموزشی، علمی و فرهنگی ملل متحد) توصیه‌ای را در مورد اخلاق هوش مصنوعی تصویب کرد که اولین توافق جهانی در مورد اخلاق هوش مصنوعی است [۷۹]. برای به‌روزرسانی و غنی‌سازی تحقیقات در مورد دستورالعمل‌ها و اصول هوش مصنوعی اخلاقی، بر اساس جدول دستورالعمل‌های اخلاقی هوش مصنوعی که در مقاله Jobin [۱۲] ارائه شده است (فقط شامل ۸۴ سند)، ما بسیاری از دستورالعمل‌های اخلاقی هوش مصنوعی جدید منتشر شده را جمع‌آوری کرده‌ایم که در مقاله بررسی جوبین گنجانده نشده‌اند. در نهایت، در مجموع ۱۴۶ دستورالعمل اخلاقی AI جمع آوری شده است. فهرستی از تمام دستورالعمل‌ها یا اسناد جمع‌آوری‌شده در جدول ۵ از مواد تکمیلی آورده شده است. تعداد دستورالعمل‌های صادر شده در هر سال از سال ۲۰۱۵ تا ۲۰۲۱ در جدول ۳ شمارش و فهرست شده است. واضح است که اکثر دستورالعمل‌ها در پنج سال گذشته منتشر شده اند، یعنی از سال ۲۰۱۶ تا ۲۰۲۰. تعداد راهنماهای منتشر شده در سال ۲۰۱۸ با ۵۳ مورد، که ۳۶٫۳ درصد از کل تعداد را به خود اختصاص داده است، بیشترین تعداد راهنما را به خود اختصاص داده است. علاوه بر این، تعداد دستورالعمل‌های AI صادر شده توسط هر کشور در جدول ۴ فهرست شده است. علاوه بر این، درصد دستورالعمل‌های منتشر شده توسط انواع مختلف صادرکنندگان (از جمله دولت، صنعت، دانشگاه و سایر سازمان ها) در شکل ۳ نشان داده شده است. از شکل ۳ می‌توان دریافت که دولت‌ها، شرکت‌ها و دانشگاه‌ها همه، نگرانی‌های شدیدی در مورد اخلاق هوش مصنوعی نشان داده‌اند.



سال	تعداد اسناد
2015	2
2016	7
2017	25
2018	53
2019	31
2020	24
2021	4

کشور	تعداد
Australia	3
Canada	4
China	5
Denmark	4
EU	15
Finland	4
France	3
Germany	7
Iceland	1
India	1
International	12
Ireland	3
Japan	6
N/A	3
Netherlands	4
Norway	1
Russia	1

Singapore	3
South Korea	3
Spain	2
Sweden	1
Switzerland	1
Turkey	1
UAE	2
UK	16
USA	39
Vatican	1

2.4. اصول اخلاق هوش مصنوعی

اصول اخلاق که در ۱۴۶ دستورالعمل جمع آوری شده، در جدول ۱ از مواد تکمیلی ذکر شده است. طبق جدول، یک همگرایی آشکار حول پنج اصل اخلاق مهم وجود دارد: شفافیت، انصاف و عدالت، مسئولیت، عدم سوء استفاده و حریم خصوصی. ۱۱ اصل اخلاق مشخص شده در دستورالعمل‌های هوش مصنوعی موجود در ادامه شرح و توضیح داده شده است.

۱) شفافیت: شفافیت یکی از اصولی است که به طور گسترده در بحث اخلاق هوش مصنوعی مورد بحث قرار گرفته است. شفافیت هوش مصنوعی عمدتاً شامل شفافیت خود فناوری هوش مصنوعی و شفافیت توسعه و پذیرش هوش مصنوعی است [۱۳]. از یک طرف، شفافیت هوش مصنوعی شامل تفسیرپذیری یک سیستم هوش مصنوعی معین است، یعنی توانایی دانستن اینکه چگونه و چرا یک مدل به روشی که در یک زمینه خاص در پیش می‌گیرد و بنابراین درک منطقی که در پس تصمیم یا رفتار آن وجود دارد. این جنبه از شفافیت معمولاً به عنوان استعاره "باز کردن جعبه سیاه هوش مصنوعی" ذکر می‌شود. این مبحث به تفسیر پذیری، توضیح پذیری یا قابل فهم بودن مربوط می‌شود. از سوی دیگر، شفافیت هوش مصنوعی شامل توجیه پذیری یا منطقی بودن فرآیند طراحی و اجرای سیستم هوش مصنوعی و نتیجه آن است. به عبارت دیگر فرآیند طراحی و پیاده سازی سیستم هوش مصنوعی و تصمیم یا رفتار آن باید قابل توجیه و قابل مشاهده باشد.

۲) انصاف و عدالت: اصل عدالت و انصاف بیان می‌کند که توسعه، استقرار و استفاده از هوش مصنوعی باید عادلانه و منصفانه باشد تا سیستم هوش مصنوعی نباید منجر به تبعیض یا سوگیری علیه افراد، جوامع یا گروه‌ها شود [۸۰]. تبعیض و نتایج ناعادلانه ناشی از الگوریتم‌های هوش مصنوعی به موضوع داغ رسانه‌ها و دانشگاه‌ها تبدیل شده است. در نتیجه اصل انصاف و عدالت در چند سال اخیر توجه قابل توجهی را به خود جلب کرده است.

۳) مسئولیت و پاسخگویی: اصل مسئولیت و پاسخگویی ایجاب می‌کند که هوش مصنوعی باید قابل ممیزی باشد؛ یعنی طراحان، توسعه دهندگان، مالکان و اپراتورهای هوش مصنوعی در قبال رفتارها یا تصمیمات یک سیستم هوش مصنوعی مسئول و پاسخگو هستند و بنابراین برای مضرات یا پیامدهای بدی که ممکن است ایجاد کند مسئول تلقی می‌شوند [۵۱]. طراحان، سازندگان و کاربران سیستم‌های هوش مصنوعی ذینفعان پیامدهای اخلاق استفاده، سوء استفاده و رفتارشان هستند و مسئولیت و فرصت شکل‌دهی به این مفاهیم را دارند. این موضوع مستلزم آن است که مکانیسم‌های مناسبی برای اطمینان از مسئولیت و پاسخگویی سیستم‌های هوش مصنوعی و نتایج آن‌ها، قبل و بعد از توسعه، استقرار و استفاده از آن‌ها ایجاد شود.

۴) NonMeficence (بدخواهی نکردن): اساساً به معنای عدم آسیب رساندن یا اجتناب از تحمیل خطرات آسیب به دیگران است [۸۱]،[۸۲]. بنابراین، اصل عدم سوء استفاده از هوش مصنوعی به طور کلی به این

اشاره دارد که سیستم‌های هوش مصنوعی نباید باعث آسیب یا تشدید آسیب به انسان یا تأثیر نامطلوب بر انسان‌ها شوند. این امر مستلزم حفظ کرامت انسانی و نیز تمامیت روحی و جسمی است. اصل عدم سوء استفاده مستلزم آن است که سیستم‌های هوش مصنوعی و محیط‌هایی که در آن کار می‌کنند باید ایمن و مطمئن باشند تا برای استفاده مخرب باز نباشند. با توجه به برخی از تصادفات مرگبار که از خودروهای خودران و ربات‌ها رخ می‌دهد، اجتناب از آسیب به انسان یکی از بزرگترین نگرانی‌ها در اخلاق هوش مصنوعی است. از این رو، بیشتر دستورالعمل‌های اخلاقی تاکید زیادی بر اطمینان از عدم آسیب به انسان از طریق ایمنی و امنیت هوش مصنوعی دارند.

۵) حریم خصوصی: هدف اصل حریم خصوصی، اطمینان از احترام به حریم خصوصی و حفاظت از داده‌ها هنگام استفاده از سیستم‌های هوش مصنوعی است. سیستم‌های هوش مصنوعی باید حقوق حریم خصوصی و حفاظت از داده‌ها را حفظ کرده و به آنها احترام بگذارند و همچنین امنیت داده‌ها را حفظ کنند. این موضوع شامل ارائه حاکمیت و مدیریت داده موثر برای تمام داده‌های مورد استفاده و تولید شده توسط سیستم هوش مصنوعی در کل چرخه عمر آن سیستم است [۸۳]. به طور خاص، جمع آوری، استفاده و ذخیره داده‌ها باید با قوانین و مقررات مربوط به حریم خصوصی و حفاظت از داده‌ها مطابقت داشته باشد. داده‌ها و الگوریتم‌ها باید در برابر سرقت محافظت شوند. هنگامی که نشت اطلاعات رخ می‌دهد، کارفرمایان یا ارائه دهندگان هوش مصنوعی باید در اسرع وقت به کارمندان، مشتریان، شرکا و سایر افراد مرتبط اطلاع دهند تا ضرر یا تأثیر ناشی از نشت به حداقل برسد.

۶) سودمندی: اصل سودمندی بیان می‌کند که هوش مصنوعی باید به مردم کمک کند و به نفع بشریت باشد [۸۲]. این اصل نشان می‌دهد که فناوری هوش مصنوعی باید برای به ارمغان آوردن نتایج و تأثیرات مفید برای افراد، جامعه و محیط استفاده شود [۸۴]. هنگام توسعه یک سیستم هوش مصنوعی، اهداف آن باید به وضوح تعریف و توجیه شوند. استفاده از فناوری هوش مصنوعی برای کمک به رسیدگی به نگرانی‌های جهانی باید تشویق شود، مانند استفاده از هوش مصنوعی برای کمک به ما در مدیریت امنیت غذایی، آلودگی و مسری‌هایی مانند ایدز و کووید ۱۹.

۷) آزادی و خودمختاری: آزادی و خودمختاری که عموماً به توانایی فرد در تصمیم گیری با توجه به اهداف و خواسته‌های خود اشاره دارد، ارزش اصلی شهروندان در جوامع دموکراتیک است. بنابراین، مهم است که استفاده از هوش مصنوعی به آزادی و خودمختاری ما آسیب نرساند یا آن را محدود نکند. هنگامی که ما از عوامل هوش مصنوعی استفاده می‌کنیم، مایلیم بخشی‌ـ از اختیارات تصمیم گیری خود را به ماشین‌های هوش مصنوعی واگذار کنیم. بنابراین، حفظ اصل آزادی و خودمختاری در زمینه هوش مصنوعی به معنای ایجاد تعادل بین قدرت تصمیم‌گیری که برای خود حفظ می‌کنیم و قدرتی که به هوش مصنوعی واگذار می‌کنیم [۸۴] است.

۸) همبستگی: اصل همبستگی مستلزم این است که توسعه و کاربرد یک سیستم هوش مصنوعی باید با حفظ مرزهای همبستگی بین مردم و نسل‌ها سازگار باشد. به عبارت دیگر، هوش مصنوعی باید امنیت اجتماعی و انسجام را ارتقا دهد و پیوندها و روابط اجتماعی را به خطر اندازد [۱۳].

۹) پایداری: با توجه به تغییرات اقلیمی و آسیب‌های زیست محیطی مداوم، اهمیت پایداری بیش از پیش مورد توجه قرار گرفته است. مانند سایر زمینه‌ها و رشته ها، هوش مصنوعی تحت تأثیر قرار گرفته و باید در دستور کار توسعه پایدار گنجانده شود. اصل پایداری بیانگر این است که تولید، مدیریت و اجرای هوش مصنوعی باید پایدار باشد و از آسیب‌های زیست محیطی جلوگیری کند. به عبارت دیگر، فناوری هوش مصنوعی باید الزامات تضمین تداوم رفاه بشرـ و حفظ محیطی خوب برای نسل‌های آینده را برآورده کند [۸۵]. سیستم‌های هوش

مصنوعی قول می‌دهند که به رفع برخی از مهم‌ترین نگرانی‌های اجتماعی کمک کنند، اما باید اطمینان حاصل شود که این امر به سازگارترین شکل ممکن به محیط‌زیست اتفاق می‌افتد.

۱۰) اعتماد: قابل اعتماد بودن، پیش نیاز افراد و جوامع برای پذیرش هوش مصنوعی است، زیرا اعتماد یک اصل اساسی برای تعاملات بین فردی و عملکرد اجتماعی است. اعتماد در توسعه، استقرار و استفاده از سیستم‌های هوش مصنوعی نه تنها به ویژگی‌های ذاتی فناوری مربوط می‌شود، بلکه به کیفیت سیستم اجتماعی-فنی مربوط به برنامه‌های کاربردی هوش مصنوعی نیز مرتبط است. بنابراین، حرکت به سمت هوش مصنوعی قابل اعتماد نه تنها به قابلیت اعتماد خود سیستم هوش مصنوعی مربوط می‌شود، بلکه نیازمند رویکردی جامع و سیستماتیک است که قابلیت اطمینان همه شرکت‌کنندگان و فرآیندهایی را که کل چرخه حیات سیستم را تشکیل می‌دهند، پوشش می‌دهد [۸۶].

۱۱) کرامت: کرامت انسانی شامل این باور است که همه مردم دارای یک ارزش ذاتی هستند که صرفاً به انسانیت آنها گره خورده است، یعنی هیچ ربطی به طبقه، نژاد، جنسیت، مذهب، توانایی‌ها یا هر عامل دیگری غیر از آنها ندارد. انسان بودن و این ارزش ذاتی هرگز نباید توسط افراد دیگر یا توسط فناوری‌هایی مانند هوش مصنوعی کاهش یابد، به خطر بیفتد، یا سرکوب شود. این مهم است که هوش مصنوعی نباید به حیثیت کاربران یا سایر اعضای جامعه آسیب برساند. در نتیجه، احترام به کرامت انسانی یک اصل مهم است که باید در اخلاق هوش مصنوعی مورد توجه قرار گیرد. بنابراین سیستم هوش مصنوعی باید به گونه‌ای توسعه یابد که به تمامیت جسمی و روانی افراد، احساس هویت فردی و فرهنگی و ارضای نیازهای اساسی آنها احترام بگذارد، حمایت کند و از آنها محافظت کند [۱۳].

بخش ۵ - رویکردهایی برای پرداختن به مسائل اخلاقی در هوش مصنوعی

این بخش رویکردهای مربوط به رسیدگی یا کاهش مسائل اخلاقی هوش مصنوعی را بررسی می‌کند. از آنجایی که اخلاق هوش مصنوعی یک زمینه گسترده و چند رشته‌ای است، ما سعی می‌کنیم به جای تمرکز صرف بر روی رویکردهای فناورانه‌ای که مورد علاقه جامعه‌ی AI/ML هستند، یک مرور کلی از رویکردهای موجود و بالقوه برای پرداختن به مسائل اخلاقی هوش مصنوعی، از جمله رویکردهای اخلاقی، تکنولوژیکی و قانونی ارائه کنیم. این بررسی از رویکردهای چند رشته‌ای برای پرداختن به مشکلات اخلاقی هوش مصنوعی نه تنها خلاصه‌ای آموزنده در مورد رویکردهای هوش مصنوعی اخلاقی ارائه می‌کند، بلکه به محققان جامعه هوش مصنوعی پیشنهاد می‌کند تا به جای تکیه بر رویکردهای فناوری، راه‌حل‌هایی برای مسائل اخلاقی هوش مصنوعی از دیدگاه‌های مختلف جستجو کنند. از آنجایی که مسائل اخلاقی هوش مصنوعی با مشکلات چند رشته‌ای در هم آمیخته است، ممکن است تنها از طریق همکاری روش‌های مختلف بتوان این مشکلات را به طور موثر حل کرد.

رویکردهای اخلاقی به توسعه‌ی سیستم‌ها یا عوامل هوش مصنوعی اخلاقی اختصاص دارد که قادر به استدلال و عمل اخلاقی بر اساس نظریه‌های اخلاقی [۸۷] با پیاده‌سازی یا تعبیه اخلاق در هوش مصنوعی هستند. رویکردهای فناوری برای توسعه فناوری‌های جدید (به ویژه فناوری‌های ML) برای حذف یا کاهش کاستی‌های هوش مصنوعی فعلی طراحی شده‌اند. به عنوان مثال، تحقیق در مورد ML قابل توضیح قصد دارد رویکردهای جدیدی را برای توضیح دلیل و مکانیسم کار الگوریتم‌های ML ایجاد کند. ML منصفانه تکنیک‌هایی را مطالعه می‌کند که ML را قادر می‌سازد تا تصمیمات یا پیش‌بینی‌های منصفانه بگیرد، یعنی تعصب یا تبعیض ML را کاهش دهد. رویکردهای حقوقی در نظر دارند تحقیقات، استقرار، کاربرد و سایر جنبه‌های هوش مصنوعی را از طریق قانون‌گذاری و مقررات، با هدف اجتناب از موضوعات اخلاقی مورد بحث قبلی، تنظیم یا کنترل کنند.

1.5. رویکردهای اخلاق: اجرای اخلاق در هوش مصنوعی

طراحی سیستم‌های هوش مصنوعی اخلاق، که می‌توانند استدلال کنند و اخلاق عمل کنند، نیاز به درک درستی از رفتار اخلاق دارد. این مسئله شامل قضاوت درست و نادرست، خوب و بد، و همچنین مسائل مربوط به عدالت، انصاف، فضیلت و سایر اصول اخلاق است. بنابراین، نظریه‌های اخلاق، که با مفاهیم رفتار درست و نادرست مرتبط هستند، ارتباط نزدیکی با اخلاق هوش مصنوعی دارند. این بخش به رویکردهای پیاده سازی اخلاق در سیستم‌های هوش مصنوعی بر اساس تئوری‌های اخلاق موجود اختصاص دارد. ابتدا، نظریه‌های اخلاق، به ویژه اخلاق هنجاری که با اخلاق هوش مصنوعی مرتبط است، بررسی می‌شود. سپس، سه نوع رویکرد اصلی برای طراحی سیستم‌های هوش مصنوعی اخلاق خلاصه می‌شود.

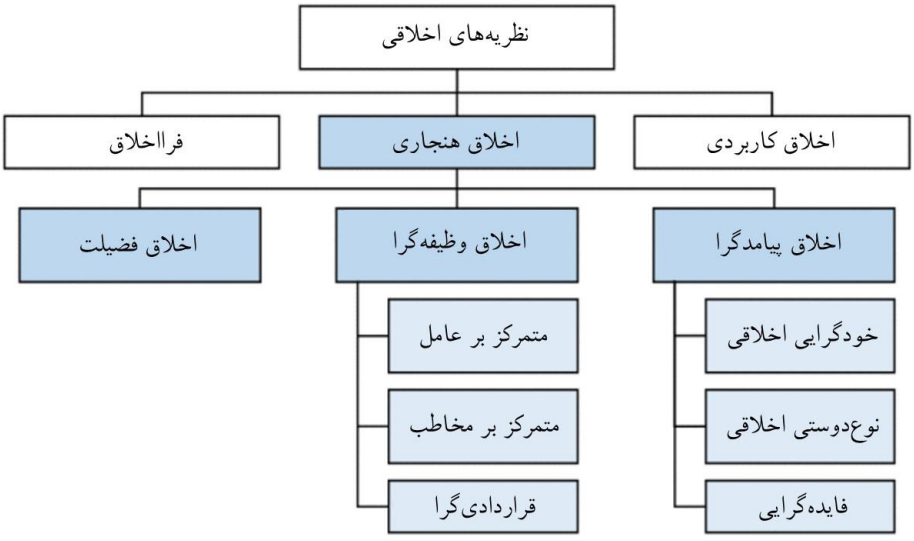
۱) نظریه‌های اخلاق

حوزه اخلاق (همچنین به عنوان فلسفه اخلاق شناخته می‌شود) به نظام‌مند کردن، دفاع و توصیه مفاهیم رفتار درست و نادرست می‌پردازد. اخلاق بر قضاوت و تعیین اینکه کدام عمل در شرایط معین خوب یا اخلاق است تمرکز دارد [۸۸]. مطالعه فلسفی اخلاق معمولاً شامل سه حوزه‌ی موضوعی اصلی است: فرا اخلاق، اخلاق هنجاری و اخلاق کاربردی [۸۹]. شاخه‌های نظریه‌های اخلاق در شکل ۴ نشان داده شده است.

فرا اخلاق، ماهیت، دامنه و معنای اصول اخلاق یا قضاوت اخلاق را بررسی می‌کند. تلاش برای درک معنا و منشأ اصطلاحات اخلاقی، نقش عقل در قضاوت‌های اخلاق و مسائل مربوط به حقایق جهانی یا ارزش‌های انسانی است [۹۰].

اخلاق هنجاری به دنبال رسیدن به معیارها و قوانین اخلاق است که رفتار درست و نادرست را تنظیم می‌کند. به این معنا که هدف آن ایجاد مجموعه‌ای از قواعد است که بر رفتار انسان و یا اینکه چگونه چیزها باید باشند، با بررسی اینکه چگونه انسان‌ها برای چیزها ارزش قائل هستند و تشخیص درست از غلط یا خوب از بد قضاوت می‌کنند.

اخلاق کاربردی، اخلاق حوزه‌های کاربردی خاص است که شامل تحلیل مسائل اخلاقی خاص و بحث‌برانگیز مانند سقط جنین، مجازات اعدام، حقوق حیوانات، نگرانی‌های زیست‌محیطی، جنگ هسته‌ای و غیره است.



الف) اخلاق هنجاری

اخلاق هنجاری مخصوصاً به درک و به کارگیری اصول اخلاقی در طراحی، استقرار و استفاده از سیستم‌های هوش مصنوعی [۸۹] مربوط می‌شود، زیرا یک رشته‌ی فلسفی عملی-هنجاری است که به نحوه‌ی رفتار انسان‌ها یا عوامل با دیگران مربوط می‌شود. سه شاخه‌ی اخلاقی هنجاری، یعنی فضیلت، اخلاق دین شناختی و پیامدگرایی در زیر ارائه و خلاصه می‌شود.

اخلاق فضیلت: اخلاق فضیلت بر فضایل یا شخصیت اخلاقی تأکید می‌کند و بر اهمیت پرورش عادات خوبِ مَنِش مانند خیرخواهی تأکید می‌کند [۹۲]. از این رو، اخلاق فضیلت، بر شخصیت ذاتی عامل، تمرکز دارد تا پیامدهای اعمالی که توسط عامل انجام می‌شود. در اخلاق فضیلتی، اگر فاعل بر اساس برخی ارزش‌های اخلاقی عمل کرده و بیندیشد، عمل آن فاعل را از نظر اخلاقی خوب تعریف می‌کند [۹۳]. به عبارت دیگر، بر اساس نظریات فضیلت، فاعل در صورتی اخلاقی است که برخی از فضایل اخلاقی را از طریق آعمال خود بروز دهد [۹۴]، [۹۵].

اخلاق دئونتولوژیک: نظریه‌های دین‌شناختی، که گاهی اوقات نظریه‌های وظیفه نامیده می‌شوند، با استفاده از قواعد اخلاقی خاصی که به‌عنوان اصول بنیادین تعهد عمل می‌کنند، درباره اخلاقی بودن یک عمل قضاوت می‌کنند. دئونتولوژی نوعی نظریه اخلاق هنجاری است که در مورد آن، انتخاب‌ها یا اعمالی از نظر اخلاقی میتوانند مورد نیاز، ممنوع یا مجاز باشند. به عبارت دیگر، deontology یک نظریه اخلاقی است که تصمیمات ما را در مورد آنچه که باید انجام دهیم، هدایت و ارزیابی می‌کند [۹۶]. دئونتولوژیست‌ها یک عمل اخلاقی خوب را به عنوان عملی تعریف می‌کنند که به برخی تعهدات پایبند باشد، که ممکن است قوانین یا وظایف اخلاقی، مقررات و هنجارهای قابل اجرا باشد.

سه مکتب اصلی از نظریه‌های deontological وجود دارد، یعنی نظریه‌های عامل محور، بیمار محور (که قربانی محور نیز نامیده می‌شود) و نظریه‌ی قراردادی. نظریه‌ی عامل محور، عامل را در مرکز قرار می‌دهد و بر وظایف عامل نِسبی تمرکز می‌کند. تئوری بیمارمحور، همانطور که از دیونتولوژی عامل‌محور متمایز می‌شوند، به جای اینکه مبتنی بر وظیفه باشند، مبتنی بر حقوق هستند. بر حقوق بیماران یا قربانیان بالقوه تمرکز دارد، مانند حق استفاده نکردن به عنوان وسیله‌ای برای رسیدن به هدف توسط شخص دیگری. تئوری‌های قراردادگرا با هر دو نظریه عامل محور و بیمار محور متفاوت است. در نظریه‌ی قراردادگرا، اعمال نادرست اخلاقی آن دسته از اعمالی هستند که توسط اصولی ممنوع می‌شوند که افراد در یک قرارداد اجتماعی مناسب و توصیف شده آن را می‌پذیرند یا با اصولی ممنوع می‌شوند که چنین افرادی نمی‌توانند «معقولانه رد کنند» [۹۶].

اخلاق نتیجه گرایانه: اخلاق نتیجه گرایانه، همانطور که از نامش پیداست، بر نتایج سودگرایانه‌ی آعمال، تأکید دارد [۹۷]. اخلاق پیامدگرایانه، اخلاقی بودن یک عمل را صرفاً بر اساس نتیجه یا پیامدهای آن ارزیابی می‌کند. به عبارت دیگر، در نظریه‌های نتیجه گرا، صحت اخلاقی یک عمل با توجه به نتیجه یا نتایج آن عمل تعیین می‌شود. به عقیده نتیجه گرایان، اگر پیامد آن عمل سودمند، یعنی مطلوب‌تر از نامطلوب تلقی شود، از نظر اخلاقی درست است. فرض کنید یک مورد ساده که در آن فرد با انتخاب بین چندین عمل ممکن مواجه می‌شود، نتیجه‌گرایی مشخص می‌کند که عمل اخلاقاً درست همان کاری است که بهترین پیامدهای کلی را دارد.

اخلاق پیامدگرا از نظر تاریخی مهم و هنوز هم محبوب است، زیرا این شهود اساسی را در بر می‌گیرد که آنچه خوب یا درست است هر چیزی است که جهان را در آینده، بهترین می‌کند، زیرا ما نمی توانیم گذشته را تغییر دهیم. نظریه‌های نتیجه گرایی را می‌توان به موارد زیر تقسیم کرد [۹۸]، [۹۹].

- اگویسم اخلاقی: بیان می‌کند که یک عمل، زمانی از نظر اخلاقی خوب است که پیامدها یا آثار آن عمل فقط برای عاملی که آن عمل را انجام می‌دهد، مطلوب‌تر از نامطلوب باشد.

- نوع دوستی اخلاق: بیان می‌کند که یک عمل، زمانی از نظر اخلاق خوب است که پیامدها یا آثار آن عمل برای همه به جز فاعل، مطلوب‌تر از نامطلوب باشد.
- فایده گرایی: بیان می‌کند که یک عمل، زمانی از نظر اخلاق خوب است که پیامدها یا آثار آن عمل برای همه مطلوب‌تر از نامطلوب باشد.

هر سه‌ی این نظریه‌ها بر پیامدهای اَعمال برای گروه‌های مختلف مردم تمرکز دارند. اما مانند همه نظریه‌های هنجاری، سه نظریه فوق رقیب یکدیگر هستند. آنها همچنین نتایج متفاوتی را ارائه می‌دهند.

ب) خلاصه‌ای از اخلاق هنجاری

از توصیفات بالا مشخص می‌شود که نظریه‌های اخلاق هنجاری مختلف، قضاوت متفاوتی را برای یک اقدام یا تصمیم به همراه خواهند داشت. به تصویر زیر توجه کنید [۱۰۰]: یک آقای مسن توسط گروهی از نوجوانان متکبر در مترو عذاب می‌کشد و زنی مصمم به کمک او می‌آید. اخلاق مدار فضیلت، عمل او را از نظر اخلاق مناسب می‌داند، زیرا فضیلت‌های خیرخواهی و شجاعت را نشان می‌دهد. متخصص دئونتولوژی اقدام او را ستودنی می‌داند زیرا با قانون کمک به نیازمندان مطابقت دارد. نتیجه گرا از عمل او به خوبی دفاع می‌کند، زیرا او رفاه کلی همه طرف‌های درگیر را به حداکثر رسانده - نجیب زاده‌ی سالخورده از رنج و رسوایی در امان است، که از سرگرمی نوجوانان پیشی- می‌گیرد-. مقایسه‌ی مختصری بین سه نظریه اخلاق هنجاری در جدول ۵ ارائه شده است.

۲) رویکردهای پیاده سازی اخلاق در هوش مصنوعی

در بخش قبل، نظریه‌های اخلاق مرتبط با اخلاق هوش مصنوعی را مورد بحث قرار دادیم. این بخش به طور خلاصه روش‌ها و رویکردهای پیاده‌سازی اخلاق در سیستم‌های هوش مصنوعی، یعنی طراحی سیستم‌های هوش مصنوعی اخلاقی را بررسی می‌کند. روش‌ها یا رویکردهای موجود برای کاشت اخلاق در هوش مصنوعی را می‌توان به سه نوع اصلی تقسیم کرد: رویکردهای بالا به پایین، رویکردهای پایین به بالا و رویکردهای ترکیبی [۱۰۱].

الف) رویکردهای بالا به پایین

رویکرد بالا به پایین به رویکردی اشاره دارد که یک نظریه اخلاق خاص را اتخاذ می‌کند و الزامات محاسباتی آن را برای هدایت طراحی الگوریتم‌ها و زیرسیستم‌هایی که می‌توانند آن نظریه را تحقق بخشند، تجزیه و تحلیل کند [۱۰۲]. رویکردهای بالا به پایین، استدلال اخلاق را بر اساس نظریه‌های اخلاق یا اصول اخلاق معین انجام می‌دهند. در رویکردهای بالا به پایین، اصول اخلاق و نظریه‌های اخلاق به عنوان قوانینی برای انتخاب اقدامات اخلاق مناسب [۱۰۱] استفاده می‌شود یا برای توصیف آنچه عامل هوش مصنوعی باید در یک موقعیت خاص انجام دهد استفاده می‌شود. بنابراین، یک رویکرد از بالا به پایین مستلزم قوانین، تعهدات و حقوق تعریف شده‌ی رسمی برای هدایت عامل هوش مصنوعی در فرآیند تصمیم‌گیری است. به عنوان مثال، سه قانون رباتیک آسیموف [۱۰۳] که بر رفتار ربات‌ها حاکم است، می‌تواند یک سیستم اخلاق از بالا به پایین برای روبات‌ها در نظر بگیرد [۱۰۱]. بسیاری از پیاده سازی‌های دیگر با استفاده از رویکردهای بالا به پایین را می‌توان در [۱۰۴]-[۱۱۱] و غیره یافت.

رویکردهای بالا به پایین معمولاً به عنوان داشتن مجموعه‌ای از قوانین درک می‌شوند که می‌توانند به یک الگوریتم تبدیل شوند. این قوانین وظایف یک نماینده یا نیاز نماینده را برای ارزیابی پیامدهای اقدامات احتمالی مختلفی که ممکن است انجام دهد مشخص می‌کند. رویکردهای بالا به پایین در نظریه‌های اخلاق متفاوت مورد استفاده هستند. به عنوان مثال، زمانی که نظریه‌ی نتیجه گرایی در رویکرد از بالا به پایین استفاده می‌شود، مدل استدلال باید نتیجه یا پیامد اَعمال را به عنوان مبنای تصمیم ارزیابی کند، یعنی عملی که منجر به نتیجه‌ی خوب می‌شود اخلاقی است و در غیر این صورت غیراخلاقی است. در حالی که اگر تئوری دئونتولوژیک به کار رود، مدل استدلال رضایت، یک ارزش معین را برای تصمیم گیری در نظر می‌گیرد، یعنی عملی که از وظایف تبعیت می‌کند اخلاقی است و عمل زیر پا گذاشتن وظایف غیراخلاقی است.

ب) رویکردهای پایین به بالا

رویکردهای پایین به بالا فرض می‌کنند که رفتار اخلاقی از مشاهدات رفتارهای دیگران آموخته می‌شود. در رویکرد پایین به بالا، تأکید بر ایجاد محیطی است که در آن یک عامل هوش مصنوعی مسیر عمل را بررسی می‌کند و عمل اخلاقی قابل ستایش، با پاداش یا انتخاب مشخص می‌شود [۱۰۱]. برخلاف رویکردهای بالا به پایین، که برای تعریف اَعمال اخلاقی و غیر اخلاقی، به نظریه‌ها یا اصول اخلاقی نیاز دارند، اصول اخلاقی از مشاهدات یا تجربه در رویکردهای پایین به بالا کشف یا آموخته می‌شوند. این رویکرد نشان می‌دهد که عامل هوش مصنوعی باید مانند بچه‌های کوچک هنجارها و اخلاقیات را بیاموزد تا از نظر اخلاقی شایسته باشد. برای مثال، آقایان هنرور و آقایی، یک عامل BDI کاسوئیستی [۱۱۲] را پیشنهاد کردند که روش استدلال مبتنی بر مورد در هوش مصنوعی و رویکرد موردی از پایین به بالا را در اخلاق ترکیب می‌کند تا قابلیت استدلال اخلاقی را به عامل باور-میل-نیت (belief-desire-intention) اضافه کند. [۱۱۳]. سایر پیاده سازی‌های رویکردهای پایین به بالا را می‌توان در [۱۱۴]-[۱۱۸] و غیره یافت.

رویکردهای پایین به بالا می‌توانند از خرد جمعی به عنوان وسیله‌ای برای آگاهی دادن به قضاوت اخلاقی عامل، استفاده کنند و سپس آن عامل می‌تواند یاد بگیرد که چگونه اخلاقی بودن عمل خود را قضاوت کند و بنابراین اخلاقی رفتار کند. ظاهراً، رویکردهای پایین به بالا فرض می‌کنند که حجم کافی از داده‌ها یا مشاهدات در مورد تصمیمات اخلاقی و نتایج آنها را می‌توان از مجموعه‌ای از موضوعات یا سناریوها جمع‌آوری کرد. این موضوع لازمه‌ی استفاده از رویکردهای پایین به بالا برای پیاده سازی سیستم‌های هوش مصنوعی اخلاقی است. با این حال، در عمل، این نیاز به راحتی برآورده نمی شود.

ج) رویکردهای ترکیبی

رویکرد ترکیبی تلاش می‌کند تا مزایای رویکردهای بالا به پایین و پایین و همچنین پایین به بالا را ترکیب کند. رویکردهای بالا به پایین از نظریه‌ها و اصول اخلاقی استفاده می‌کنند و بر اهمیت نگرانی‌های اخلاقی صریح که از خارج از نهاد (موضوع اخلاقی) ناشی می‌شوند، تأکید می‌کنند. در حالی که رویکردهای پایین به بالا بیشتر بر پرورش اخلاقی که از درون موجودیت، از طریق تکامل و یادگیری ناشی می‌شود، تمرکز می‌کنند. هر دو رویکرد از بالا به پایین و پایین به بالا جنبه‌های مختلفی از حساسیت اخلاقی را در بر می‌گیرند. با ترکیب این رویکردها، ممکن است بتوانیم عامل هوش مصنوعی ایجاد کنیم که بتواند اخلاق پویا و انعطاف پذیر رویکرد پایین به بالا را در عین رعایت اصول از بالا به پایین حفظ کند. رویکردهای ترکیبی متفاوتی در [۱۱۹] - [۱۲۴] پیاده‌سازی شده‌اند.

همانطور که Gigerenzer [۱۲۵] بیان کرد، ماهیت رفتار اخلاقی، ناشی از تعامل بین ذهن و محیط است. بر اساس این دیدگاه، هم طبیعت و هم تربیت، در شکل گیری رفتار اخلاقی مهم هستند. رویکرد ترکیبی با این مفهوم سازگار است. در رویکرد ترکیبی، رویکرد بالا به پایین از قوانین برنامه ریزی شده و رویکرد پایین به بالا از قوانین آموخته شده از مشاهدات یا تجربیات زمینه استفاده می کند که به ترتیب شبیه به ماهیت و پرورش جنبه‌های اخلاقی هستند. بنابراین، از این منظر، هم طبیعت و هم پرورش در رویکردهای ترکیبی مورد توجه قرار می گیرند.

د) نکاتی در مورد رویکردهای اخلاقی

رویکرد از بالا به پایین، نظریه‌ها و اصول اخلاقی مشخص شده را به تصمیم گیری اخلاقی تبدیل می کند یا نظریه‌ها و اصول اخلاقی داده شده را به الگوریتم تبدیل می کند. این رویکرد برای طراحی و تحقق عوامل هوش مصنوعی اخلاقی با اصول اخلاقی شناخته شده و کدهای اخلاقی مناسب است. مزیت این رویکرد این است که بر اساس تئوری‌ها و قواعد اخلاقی از پیش تعیین شده، تصمیمات و اقدامات کارگزاران اخلاقی قابل پیش بینی است و هنجارها یا قوانین اخلاقی اجرا شده از طریق کدهای برنامه یا ابزارهای دیگر را می‌توان در طول تصمیم گیری اخلاقی درک کرد. بنابراین، اعتبار عامل هوش مصنوعی اخلاقی ایجاد شده توسط رویکرد بالا به پایین را می‌توان بهتر تضمین کرد زیرا فرآیند تصمیم گیری آن دارای قابلیت تفسیر و شفافیت قوی است. نقطه ضعف رویکرد از بالا به پایین این است که عامل اخلاقی نظریه‌های اخلاقی یا قواعد اخلاقی از پیش تعیین شده را اتخاذ می کند، هنگام تصمیم گیری در یک محیط پیچیده و متغیر، این روش فاقد انعطاف پذیری و سازگاری است.

رویکرد پایین به بالا تأکید می کند که کارگزاران اخلاقی اخلاق را به طور مستقل از محیط اجتماعی می‌آموزند، به تدریج دارای استدلال اخلاقی و توانایی‌های اخلاقی هستند و می‌توانند با تغییرات محیطی سازگار شوند. رویکرد پایین به پایین برای طراحی و اجرای عوامل هوش مصنوعی اخلاقی بدون تئوری‌ها و دستورالعمل‌های اخلاقی روشن، مناسب است. مزیت این رویکرد این است که عامل می‌تواند از طریق یادگیری مستمر توسعه یافته و تکامل یابد تا با تغییرات محیطی سازگار شود. این دسته از رویکردها، سازگاری و انعطاف پذیری خوبی دارند و می‌توان نظریه‌ها یا دستورالعمل‌های اخلاقی متفاوت و جدیدی را برای سناریوهای کاربردی مختلف ساخت. نقطه ضعف این رویکرد این است که به دلیل عدم هدایت نظریه‌ها یا قوانین اخلاقی، فرآیند تصمیم گیری عوامل هوش مصنوعی اخلاقی دارای درجه خاصی از اطاعت کورکورانه است و تکمیل آموزش در کوتاه مدت دشوار است. پس در این رویکرد باید زمان و تصمیمات اخلاقی مناسب اتخاذ کنیم. در عین حال، تضمین تفسیرپذیری و شفافیت فرآیند تصمیم‌گیری عوامل AI اخلاقی طراحی‌شده دشوار است.

رویکرد ترکیبی، مزایای رویکردهای بالا به پایین و پایین به بالا را با هم ترکیب می کند و تا حدودی بر کاستی‌های این دو روش غلبه می کند. اگر یک رویکرد واحد (از بالا به پایین یا پایین به بالا) الزامات را پوشش ندهد، یک رویکرد ترکیبی ضروری و امیدوارکننده در نظر گرفته می‌شود. با این حال، چالش اصلی ترکیب مناسب ویژگی‌های رویکردهای بالا به پایین و پایین به بالا است. ویژگی‌های سه رویکرد برای پیاده سازی اخلاق در هوش مصنوعی خلاصه شده و در جدول ۶ فهرست شده است.

نظریه اخلاقی	توضیحات	تمرکز در تأمل	معیار تصمیم‌گیری	استدلال عملی
اخلاق فضیلت	یک عمل درست است اگر کاری باشد که یک فرد با	انگیزه‌ها (آیا عمل با فضیلت انگیزه یافته است؟)	فضایل	تجسم فضایل/ویژگی‌های انسانی

	فضیلت در آن موقعیت انجام دهد.			
اخلاق وظیفه‌گرا	یک عمل درست است اگر با یک قاعده یا اصل اخلاق مطابقت داشته باشد.	عمل (آیا عمل با یک الزام سازگار است؟)	وظایف/قوانین	پیروی از قوانین
اخلاق پیامدگرا	یک عمل درست است اگر بهترین پیامدها را ترویج کند، یعنی خوشبختی را به حداکثر برساند.	پیامدها (نتیجه عمل چیست؟)	رفاه نسبی	بهینه‌سازی سودمندی یا خوشبختی

رویکرد	توضیحات	آیا نیاز به قوانین اخلاقی دارد؟	توانایی یادگیری	توانایی تطبیق	قابلیت تفسیر
بالا به پایین	برنامه‌ریزی نظریه و اصول اخلاقی داده‌شده	بله	خیر	ضعیف	زیاد
پایین به بالا	یادگیری قوانین کلی از موارد فردی	خیر	قوی	قوی	کم
ترکیبی	ترکیب رویکردهای پایین به بالا و بالا به پایین	بله	قوی	قوی	متوسط

2.5. رویکردهای فناورانه

در این بخش، به طور خلاصه وضعیت تحقیق در مورد رویکردهای فناورانه برای پرداختن به مسائل اخلاقی هوش مصنوعی را در راستای اصول مورد بحث در بخش ۵-ب خلاصه می‌کنیم.

در حال حاضر، رویکردهای فناورانه برای کاهش مسائل مرتبط هنوز در مرحله رشد اولیه هستند. در سال‌های اخیر، جوامع تحقیقاتی هوش مصنوعی تلاش‌های خاصی را برای پرداختن به مسائل اخلاقی هوش مصنوعی انجام داده‌اند. به عنوان مثال، ACM (انجمن ماشین‌های محاسباتی) کنفرانس سالانه ACM FAccT (که محققان و متخصصان علاقه‌مند به عدالت، پاسخگویی و شفافیت در سیستم‌های فنی-اجتماعی را گرد هم می‌آورد) را از سال ۲۰۱۸ تا به حال برگزار کرده است. AAAI (انجمن پیشرفت هوش مصنوعی) و ACM کنفرانس AAAI/ACM را در زمینه هوش مصنوعی، اخلاق و جامعه (AIES) از سال ۲۰۱۸ تأسیس کرده اند و سی و یکمین کنفرانس بین المللی مشترک هوش مصنوعی و بیست و سومین کنفرانس اروپایی در زمینه هوش مصنوعی (IJCAI-ECAI 2022) مسیر ویژه‌ای را در مورد "AI برای خوب" ارائه می‌دهد.

کار موجود، تا جایی که ما می‌دانیم، عمدتاً بر روی چند موضوع و اصول اصلی و کلیدی تمرکز دارد و سایر مسائل و اصول به ندرت مطرح می‌شوند. بنابراین، ما فقط خلاصه‌ای کوتاه از رویکردهای فناورانه ارائه می‌دهیم که شامل پنج اصل کلیدی اخلاق است. به ویژه، برای پنج اصل کلیدی (به عنوان مثال، شفافیت،

انصاف و عدالت، عدم سوء استفاده، مسئولیت و پاسخگویی و حریم خصوصی)، برخی از موضوعات پژوهشی- نماینده و مراجع مربوطه در جدول ۲ مواد تکمیلی فهرست شده‌اند.

هوش مصنوعی قابل توضیح (XAI)، که به عنوان هوش مصنوعی قابل تفسیر نیز شناخته می‌شود، در حال حاضر مسیر اصلی تحقیق و روش فنی برای رسیدگی به مسائل عدم شفافیت در هوش مصنوعی است. هدف XAI این است که به کاربران انسانی اجازه دهد تا نتایج و خروجی‌های ارائه شده توسط یک سیستم هوش مصنوعی، به ویژه توسط الگوریتم‌های ML را درک کنند. کریستوف و همکارانش [۱۲۸] تاریخچه مختصری از حوزه XAI ارائه کرد و با یک مرور کلی از روش‌های تفسیری پیشرفته، برخی از چالش‌های تحقیق را مورد بحث قرار داد. علاوه بر این، کریستوف کتابی در مورد ML قابل تفسیر [۱۲۹] نوشته است که یک کتاب محبوب در زمینه XAI است.

در مورد اصل انصاف، آثار زیادی نیز وجود دارد که به حذف یا کاهش تعصب یا تبعیض نشان داده شده توسط سیستم‌های هوش مصنوعی، به ویژه در ML اختصاص داده شده است. هوش مصنوعی منصفانه [۱۳۰] با هدف جلوگیری از آسیب (یا منفعت) متفاوت برای زیرگروه‌های مختلف، یک موضوع تحقیقاتی بسیار فعال است که به پرداختن به مسائل عدم انصاف در هوش مصنوعی اختصاص دارد. در بررسی انصاف در ML توسط سایمون و کریستین [۱۳۱]، مکاتب فکری و رویکردهای مختلف برای کاهش سوگیری‌ها و افزایش انصاف در ML بررسی شد.

اصل Nonmaleficence شامل چندین کد مانند ایمنی، امنیت و استحکام است. از این رو، کارهایی برای هر یک از کدهای مرتبط با اصل عدم سوء استفاده وجود دارد. در حال حاضر، هوش مصنوعی ایمن، هوش مصنوعی مطمئن و هوش مصنوعی قوی سه جهت اصلی تحقیقاتی برای تحقق اصل عدم سوء استفاده در هوش مصنوعی هستند. خوانندگان علاقه‌مند می‌توانند جزئیات بیشتر را از طریق مراجع مربوطه فهرست شده در جدول ۲ مواد تکمیلی دریافت کنند.

از آنجایی که هوش مصنوعی به طور گسترده در زندگی ما استفاده می‌شود، هوش مصنوعی مسئول، در حال تبدیل شدن به یک موضوع حیاتی است. مسئولیت مفهومی نسبتاً انتزاعی و گسترده است. در حال حاضر، هیچ تعریف یا مفهوم جهانی و یکپارچه‌ای برای هوش مصنوعی مسئول وجود ندارد که عمدتاً شامل پاسخگویی، مسئولیت پذیری، انصاف، استحکام و توضیح پذیری است [۱۳۲]. دوریان و همکاران [۱۳۳] دو چارچوب برای هوش مصنوعی مسئول با ادغام تجزیه و تحلیل اخلاق در عملکرد مهندسی در هوش مصنوعی پیشنهاد کردند. علاوه بر این، مقاله‌ی [۱۳۴] مقدمه‌ای سیستماتیک در مورد هوش مصنوعی مسئول ارائه می‌دهد.

به منظور رسیدگی به مسائل حریم خصوصی در هوش مصنوعی، محققان تلاش‌های زیادی انجام داده اند. حریم خصوصی متفاوت [۱۳۵]، یکی از رویکردهای اصلی برای حفظ حریم خصوصی و تجزیه و تحلیل داده‌ها است. اخیراً یک پارادایم جدید ML، یعنی یادگیری فدرال [۱۳۶]، [۱۳۷] (که ML توزیع شده نیز نامیده می‌شود)، برای کاهش خطر نشت حریم خصوصی در ML پیشنهاد شده است. علاوه بر این، برخی دیگر از تکنیک‌های حفظ حریم خصوصی برای ML [138]، [۱۳۹] پیشنهاد شده است.

در مورد سایر اصول مانند خیرخواهی، آزادی و خودمختاری، کرامت و غیره، رویکردهای تکنولوژیکی مرتبطی در ادبیات پیدا نکرده‌ایم. این امر، ممکن است به دلیل دشواری یا نامناسب بودن استفاده از روش‌های فنی برای رسیدگی به مسائل مربوط به این اصول باشد. به طور کلی، اخلاق هوش مصنوعی یک حوزه نسبتاً جدید است و رویکردهای تحقق این اصول هنوز نیاز به مطالعه در آینده دارد.

3.5. رویکردهای حقوقی: قانونگذاری و مقررات

با توجه به استفاده روزافزون از فناوری‌های هوش مصنوعی در بسیاری از بخش‌ها و نمایش مسائل اخلاقی و خطرات در کاربردهای هوش مصنوعی، قوانین و مقررات بسیاری توسط دولت‌ها و سازمان‌ها برای کنترل توسعه و کاربرد هوش مصنوعی وضع شده است. رویکردهای حقوقی به یکی از ابزارهای پرداختن به مسائل اخلاقی در هوش مصنوعی تبدیل شده است. در ادامه، چندین قانون و مقررات مرتبط با هوش مصنوعی را که در چند سال گذشته پیشنهاد شده‌اند، فهرست می‌کنیم.

- در سال ۲۰۱۶، پارلمان اروپا و شورای اتحادیه اروپا (EU) مقررات عمومی حفاظت از داده‌ها [۱۴۰] را منتشر کردند که مقرراتی در قانون اتحادیه اروپا در مورد حفاظت از داده‌ها و حریم خصوصی در اتحادیه اروپا و منطقه اقتصادی اروپا است.
- در سال ۲۰۱۷، ایالات متحده، لایحه‌ای را برای تضمین ایمنی وسایل نقلیه‌ی خودکار با تشویق آزمایش و استقرار چنین وسایل نقلیه‌ای تصویب کرد به نام"قانون استقرار و تحقیقات ایمن در آینده - استقرار و تحقیقات در خودروها در آینده" [۱۴۱].
- در سال ۲۰۱۸، برزیل قانون شماره ۱۳ ۷۰۹، قانون کلی حفاظت از داده‌ها (Lei Geral de Proteção de Dados) [۱۴۲] را برای حفاظت از داده‌های شخصی در این کشور به تصویب رساند.
- در سال ۲۰۲۱، کمیسیون اروپا قانون هوش مصنوعی [۱۴۳] را منتشر کرد که یک رویکرد نظارتی بین بخشی را برای استفاده از سیستم‌های هوش مصنوعی در سراسر اتحادیه اروپا و بازار آن تعیین می‌کند.

بخش ۶ - روش‌های ارزیابی هوش مصنوعی اخلاقی

هدف رشته‌ی اخلاق هوش مصنوعی، طراحی سیستم‌های هوش مصنوعی اخلاقی برای رفتار اخلاقی یا پایبندی به اصول و قواعد اخلاقی و معنوی است. نحوه ارزیابی یا محاسبه‌ی اخلاقیات یا معنویات (صلاحیت معنوی) هوش مصنوعی اخلاقی طراحی شده بسیار مهم و ضروری است، زیرا سیستم‌های هوش مصنوعی طراحی شده باید قبل از استقرار، آزمایش یا ارزیابی شوند که آیا سیستم هوش مصنوعی الزامات اخلاقی را برآورده می‌کند یا خیر. با این حال، این جنبه اغلب در ادبیات موجود نادیده گرفته شده یا نادیده گرفته می‌شود. این بخش سه نوع رویکرد آزمایش، تأیید و استانداردها را برای ارزیابی اخلاق هوش مصنوعی بررسی می‌کند.

الف. آزمایش

تست، یک روش معمولی است که برای ارزیابی قابلیت‌های اخلاقی یک سیستم هوش مصنوعی استفاده می‌شود. معمولاً هنگام آزمایش یک سیستم، خروجی سیستم باید با یک حقیقت زمینی یا خروجی مورد انتظار مقایسه شود [۱۰۰]. این بخش بر روی رویکردهای آزمایشی برای ارزیابی هوش مصنوعی اخلاقی تمرکز دارد.

۱) آزمون تورینگ اخلاقی

هم در نظریه‌های اخلاقی و هم در بحث‌های روزمره درباره اخلاق، مردم معمولاً نظرات متفاوتی در مورد اخلاقی بودن اَعمال مختلف دارند. به عنوان مثال، کانت ادعا کرد که دروغ، صرف نظر از عواقب آن، همیشه غیراخلاقی است. اخلاق‌گرایان فایده‌گرا این را انکار می‌کنند و معتقدند که دروغ تا زمانی که پیامدهای آن در مجموع به اندازه کافی خوب باشد موجه است. از آنجایی که نظریه‌های اخلاقی مختلف معیارهای ارزیابی

متفاوتی برای رفتار اخلاقی دارند، آلن و همکاران [۱۴۴] پیشنهاد کردند که از آزمون تورینگ اخلاقی (MTT) برای ارزیابی عوامل اخلاقی مصنوعی استفاده شود.

در نسخه استاندارد آزمون تورینگ [۱۴۵]، یک بازجوی انسانی از راه دور وظیفه دارد بین یک ماشین (کامپیوتر) و یک موضوع انسانی بر اساس پاسخ آنها به سؤالات مختلف مطرح شده توسط بازجو تمایز قائل شود. اگر ماشینی با شانس کافی به عنوان سوژه انسانی اشتباه شناسایی شود، موفق به گذراندن تست می‌شود و ماشین به عنوان موجودی باهوش و متفکر در نظر گرفته می‌شود. Turing Test مستقیماً آزمون رفتاری را انجام می‌دهد تا اختلاف نظر در مورد معیارهای تعریف هوش یا کسب موفقیت در زبان طبیعی را دور بزند. آزمون تورینگ اخلاقی (MTT) به طور مشابه برای دور زدن اختلاف نظرها در مورد استانداردهای اخلاقی با محدود کردن مکالمات در آزمون عطف استاندارد به سؤالات مربوط به اخلاق پیشنهاد شد. اگر انسان بازجو نتواند ماشین را از سوژه انسانی در سطحی بالاتر از شانس تشخیص دهد، ماشین یک عامل اخلاقی است.

با این حال، آلن و همکاران [۱۴۴] اعتراف کردند که یکی از محدودیت‌های MTT این است که بر توانایی ماشین‌ها برای بیان واضح قضاوت‌های اخلاقی تأکید می‌کند. ریشه شناسان یا کانتیان ممکن است با این تاکید راضی باشند، اما نتیجه گرایان استدلال می‌کنند که MTT تاکید زیادی بر توانایی بیان دلیل آعمال فرد دارد. به منظور تغییر تمرکز از توانایی مکالمه به عمل، آلن و همکاران [۱۴۴] همچنین یک MTT جایگزین را پیشنهاد کرد که "MTT مقایسه ای" (cMTT) نامیده شد. در cMTT، به انسانی بازجوی دو جفت توصیف از آعمال واقعی و اخلاقی مهم یک سوژه انسانی و یک ماشین (یا عامل هوش مصنوعی) داده می‌شود، که از تمام مراجعی که بازیگر را شناسایی می‌کنند پاک می‌شود. اگر بازجو دستگاه را در درصد معینی به درستی شناسایی کند، آنگاه دستگاه نمی تواند آزمون را پشت سر بگذارد. یک مشکل این نسخه از MTT این است که نحوه رفتار ماشین راحت‌تر از انسان تشخیص داده می‌شود، زیرا ماشین به طور مداوم در شرایط یکسان رفتار می‌کند. بنابراین، باید از بازجو خواسته شود تا ارزیابی کند که آیا یک بازیگر نسبت به دیگری اخلاقی‌تر است یا نه؟ اگر دستگاه به عنوان دستگاهی که دارای اخلاق کمتری است، بیشتر از انسان شناسایی نشود، دستگاه آزمایش را به خوبی گذرانده است.

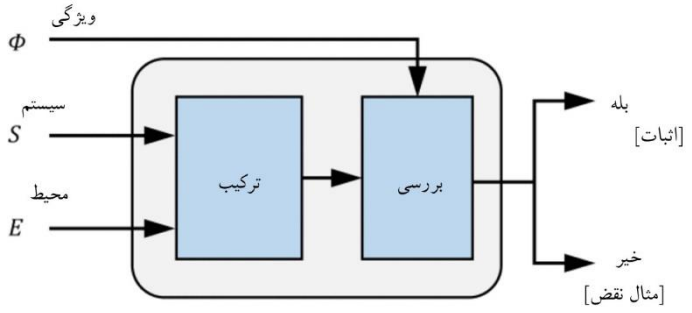
اگرچه cMTT مشکلات متعددی دارد، برای مثال، ممکن است کسی- استدلال کند که این استاندارد بسیار پایین است، والاک و آلن [۱۴۶] معتقدند که cMTT یک روش عملی و قابل قبول برای ارزیابی اخلاقیات عوامل هوش مصنوعی است، زیرا هیچ معیار ارزیابی دیگری وجود ندارد که به طور مشترک پذیرفته شده و مورد توافق باشد.

۲) تست‌های تخصصی و غیر کارشناسی

علاوه بر MTT، محققان سعی کرده‌اند صلاحیت اخلاقی سیستم‌های هوش مصنوعی را از طریق آزمون‌های خبره یا غیرمتخصص ارزیابی کنند، که در آن نتیجه سیستم با حقیقت اصلی ارائه شده توسط افراد غیرمتخصص یا متخصص مقایسه می‌شود. آزمون کارشناسی استانداردهای متخصصان در اخلاق هنجاری را برای ارزیابی اخلاقیات عوامل هوش مصنوعی اتخاذ می‌کند. آزمون‌های غیرکارشناسی، اخلاق عامیانه را به عنوان معیار در نظر می‌گیرند و توانایی اخلاقی عامل یا سیستم هوش مصنوعی را در آزمون معیار مربوطه ارزیابی می‌کنند. در آزمون‌های غیرکارشناسی، شهروندان می‌توانند نقش خود را در ارزیابی و ارزیابی قابلیت‌های اخلاقی یک سیستم هوش مصنوعی بر اساس مواضع اخلاقی و بررسی دقیق خود ایفا کنند.

ب. تأیید

دسته دیگری از رویکردها برای ارزیابی اخلاق بودن هوش مصنوعی شامل اثبات این است که سیستم هوش مصنوعی طبق برخی مشخصات شناخته شده به درستی رفتار می‌کند. سشیا و همکاران [۱۴۷] این نوع رویکرد را مورد بحث قرار دادند. یک فرآیند تأیید رسمی معمولی در شکل ۵ نشان داده شده است که در آن S مدلی از سیستمی است که باید تأیید شود، E مدلی از محیط است و Φ خاصیتی است که باید تأیید شود. برنامه راستی آزمایی، یک پاسخ بله یا خیر به عنوان خروجی می‌دهد که نشان دهنده‌ی آن است که آیا S ویژگی Φ را در محیط E برآورده می‌کند یا نه. و یک مدرک صحت شامل یک پاسخ بله در برخی از ابزارهای تأیید رسمی است.



آرنولد و شوتز [۱۴۸] نقایص MTT را بررسی کردند و خاطرنشان کردند که ارزیابی‌های مبتنی بر MTT در برابر فریب، استدلال ناکافی و عملکرد اخلاق ضعیف آسیب‌پذیر هستند و آنها مفهوم "تأیید طراحی" را برای ارزیابی شایستگی اخلاق سیستم هوش مصنوعی پیشنهاد کردند.

برای ارزیابی طراحی اخلاق هوش مصنوعی، می‌توان از معیارهای ارزیابی متنوع استفاده کرد. صرف نظر از روشی که هوش مصنوعی استدلال اخلاق را انجام می‌دهد، بسیار مهم است که فعالیت‌های اخلاق آن با اهداف طراحی اخلاق مطابقت داشته باشد.

ج. استانداردها

بسیاری از استانداردهای صنعتی برای هدایت توسعه و کاربرد هوش مصنوعی و ارزیابی محصولات هوش مصنوعی پیشنهاد شده است. در این بخش برخی از استانداردهای مرتبط با هوش مصنوعی معرفی می‌شوند.

- در سال ۲۰۱۴، انجمن کامپیوتر استرالیا کد رفتار حرفه‌ای ASC را برای پیروی از همه متخصصان فناوری ارتباطات اطلاعات ایجاد کرد که شش ارزش اصلی اخلاق و الزامات مرتبط با رفتار حرفه‌ای را مشخص می‌کند.
- در سال ۲۰۱۸، ACM کد اخلاق و رفتار حرفه‌ای ACM را برای پاسخ به تغییرات در حرفه رایانه از سال ۱۹۹۲ به روز کرد. این کد بیانگر وجدان حرفه است و برای الهام بخشیدن و هدایت رفتار اخلاق همه متخصصان رایانه، از جمله متخصصان فعلی طراحی شده است و پزشکان مشتاق، مدرسان، دانش‌آموزان، تأثیرگذاران و هر کسی- که از فناوری محاسباتی به روشی تأثیرگذار استفاده می‌کند. علاوه بر این، این کد به عنوان مبنایی برای اصلاح در صورت وقوع تخلف عمل می‌کند. این آیین نامه شامل

اصولی است که به عنوان بیانیه‌های مسئولیت، بر اساس این درک که خیر عمومی همیشه ملاحظات اولیه است، فرموله شده است. هر اصل با دستورالعمل‌هایی تکمیل می‌شود که توضیحاتی را برای کمک به متخصصان محاسبات در درک و به کارگیری اصل ارائه می‌دهد [۱۴۹].

- پروژه‌ی IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [۱۵۰] توسعه‌ی سری استانداردهای IEEE P7000™ [۱۵۱] را تایید کرد (فهرست شده در جدول ۳ مواد تکمیلی) که موضوعاتی از جمع آوری داده‌ها تا حریم خصوصی و سوگیری الگوریتمی و فراتر از آن را پوشش می‌دهد.
- ISO/IEC JTC 1/SC 42 [۱۵۲] که کمیته مشترک بین ISO و IEC مسئول استانداردسازی در حوزه هوش مصنوعی است، به توسعه مجموعه‌ی بزرگی از استانداردها شامل حوزه‌های استانداردهای اساسی هوش مصنوعی، داده‌های بزرگ، قابلیت اعتماد هوش مصنوعی، موارد استفاده، برنامه‌های کاربردی، مفاهیم حاکمیتی هوش مصنوعی، رویکردهای محاسباتی هوش مصنوعی، نگرانی‌های اخلاقی و اجتماعی هستند. استانداردهای منتشر شده و در دست توسعه توسط ISO/IEC JTC 1/SC 42 در جدول ۴ از مواد تکمیلی فهرست شده است.

با نگرانی در مورد مسائل اخلاقی هوش مصنوعی، علاقه به استانداردهای هوش مصنوعی برای شکل دادن به طراحی، استقرار و ارزیابی هوش مصنوعی به سرعت در حال رشد است. اگرچه استانداردهای زیادی پیشنهاد شده است، شکاف بین استانداردها (یا اصول) و عمل هنوز زیاد است. در حال حاضر، تنها برخی از شرکت‌های بزرگ، مانند IBM [۱۵۳] و مایکروسافت [۱۵۴]، استانداردها، چارچوب‌ها و دستورالعمل‌های صنعتی خود را برای ایجاد فرهنگ هوش مصنوعی پیاده‌سازی کرده‌اند. اما برای کسب و کارهای کوچکتر با منابع کمتر، اصول تمرین شکاف یک مشکل بزرگ است. بنابراین، هنوز تلاش‌های زیادی لازم است. از یک سو، ارائه استانداردهای توسعه یافته ضروری است و از سوی دیگر، لازم است به شدت عمل به استانداردها ترویج شوند.

بخش ۷ - چالش‌ها و چشم اندازهای آینده

از آنجایی که اخلاق هوش مصنوعی یک رشته نوظهور است و هنوز چالش‌ها و مشکلات زیادی وجود دارد که باید در این زمینه مورد توجه قرار گیرد. در این بخش، برخی از چالش‌ها در اخلاق هوش مصنوعی را مورد بحث قرار می‌دهیم و دیدگاه‌های آینده را از دیدگاه خود ارائه می‌کنیم. هدف از این بخش ارائه برخی سوالات احتمالی تحقیقاتی و رهنمودها برای تحقیقات بیشتر در آینده است و از این طریق پیشرفت تحقیقات در زمینه اخلاق هوش مصنوعی را تسهیل می‌کند.

الف. چالش‌ها در دستورالعمل‌ها و اصول اخلاقی هوش مصنوعی

همانطور که در بخش چهارم بررسی شد، تعداد زیادی دستورالعمل توسط سازمان‌ها، شرکت‌ها و دولت‌های مختلف پیشنهاد و منتشر شده‌اند و اصول مختلفی را می‌توان در این دستورالعمل‌ها شناسایی کرد. اما در حال حاضر هنوز دستورالعملی وجود ندارد که توسط سازمان ها، بخش‌ها و دولت‌های مختلف تصویب شده باشد. به عبارت دیگر، سازمان‌های مختلف، شرکت‌های حوزه‌های مختلف و حتی شرکت‌های مختلف در یک حوزه، نظرات متفاوتی در مورد اخلاق هوش مصنوعی دارند. اجماع در مورد اخلاق هوش مصنوعی هنوز حاصل نشده است و مشخص نیست که هوش مصنوعی باید از چه اصول و ارزش‌های مشترکی پیروی کند. علاوه بر این، زمانی که هوش مصنوعی در حوزه‌های مختلف به کار می‌رود، ممکن است اصول اخلاقی متفاوتی مورد نیاز

باشد. در حال حاضر، مطالعه و بحث در مورد اخلاق هوش مصنوعی در زمینه‌های مختلف کاربردی خاص به ندرت در طول مطالعه ادبیات ما دیده می‌شود.

بنابراین، بسیار مهم و ضروری است که اصول اخلاقی اساسی و مشترک هوش مصنوعی از طریق بحث و همکاری بین سازمان‌ها، حوزه‌ها و دولت‌های مختلف به دست آید و به خوبی تثبیت شود. سپس بر اساس اصول اولیه و رایج، هر رشته می‌تواند این اصول را بیشتر ارتقا دهد تا به طور کلی در این زمینه خاص قابل اجرا باشد. شفاف سازی اصول و ارزش‌های اخلاقی که یک سیستم هوش مصنوعی باید با آنها مطابقت داشته باشد، پیش نیاز و اساس طراحی چنین سیستمی است که این الزامات را برآورده کند.

ب. چالش‌های پیاده سازی اخلاق در هوش مصنوعی

در پیاده سازی اخلاق در هوش مصنوعی، چالش‌های زیادی وجود دارد. این بخش به تحلیل چالش‌هایی می‌پردازد که ممکن است در عمل، با اتخاذ انواع مختلف نظریه‌های اخلاقی با آن مواجه شوند.

۱) چالش‌های اخلاق فضیلت در عمل

بر اساس اخلاق فضیلتی، اگر فاعل، فضیلتی را مصداق دهد، یعنی بر اساس برخی ارزش‌های اخلاقی عمل کند و بیندیشد، از نظر اخلاقی خوب است [۹۳]. نمی توان فقط با مشاهده یک عمل یا یک سری اقدامات که به نظر می‌رسد دلالت بر آن فضیلت دارد، قضاوت کرد که آیا یک سیستم یا عامل هوش مصنوعی فضیلت دارد یا نه، دلایل پشت این اعمال باید روشن شود، یعنی انگیزه‌های پشت این کارها روشن شود. اقدامات باید واضح باشد با این حال، انگیزه‌های پشت اَعمال سیستم‌های هوش مصنوعی، معمولاً برای ما نامشخص و ناشناخته هستند و کشف آن دشوار است. این مورد، چالش اصلی برای اجرای اخلاق فضیلت است. به‌علاوه، وقتی طراحی اخلاقی را بر اساس اخلاق فضیلت انجام می‌دهیم، اینکه سیستم هوش مصنوعی با کدام ویژگی‌ها یا ویژگی‌های فضیلت هماهنگ می‌شود، سؤال دشواری است. حتی اگر صفات فضیلت به دقت انتخاب شده باشد، نحوه توصیف و اندازه گیری فضیلت همچنان یک کار چالش برانگیز است.

۲) چالش‌های اخلاق دئونتولوژیک در عمل

دئونتولوژیست‌ها یک عمل را در صورتی از نظر اخلاقی خوب می‌دانند که به برخی از قوانین یا وظایف، مقررات و هنجارهای اخلاقی پایبند باشد. اگرچه ماهیت مبتنی بر قاعده اخلاق deontological برای عمل، مناسب به نظر می‌رسد، چالش‌هایی در طول فرآیند اجرا بوجود می‌آیند. اول اینکه کدام قواعد اخلاقی باید در طراحی اخلاقی اجرا شوند. دوم، ممکن است در برخی شرایط بین قوانین تضاد وجود داشته باشد. اگرچه دستور دادن یا سنجیدن قوانین اخلاقی ممکن است این مشکل را حل کند، تعیین ترتیب اهمیت قواعد اخلاقی مختلف اغلب دشوار است.

۳) چالش‌های اخلاق نتیجه گرایی در عمل

اخلاق نتیجه گرایانه اخلاقی بودن یک عمل را صرفاً بر اساس نتیجه آن ارزیابی می‌کند. دو چالش اصلی در طول اجرای اخلاق نتیجه گرا درگیر است. اولاً، تعیین پیامدهای یک اقدام یا تصمیم، دشوار است. برای سیستم

هـوش مصـنوعی فعـلی، عواقـب احتمـالی اقـدامات سیسـتم، معمـولاً از قبـل بـا توجـه بـه عـدم شـفافیت یـا تفسـیرپذیری مـدل‌های هـوش مصـنوعی فعـلی، بـه‌ویژه شـبکه‌های عصـبی مصـنوعی، مشـخص نیسـت. چـالش دوم مربـوط بـه کـمّی کـردن پیامـدها اسـت. از آنجـایی کـه در اخـلاق نتیجـه گـرا، هـدف، بـه حـداکثر رسـاندن مطلوبیـت است، چگونگی تعریف و محاسبه‌ی مطلوبیت یک مشکل اساسی است.

۴) چالش‌های هماهنگی در بین استانداردهای اخلاقی مختلف

بـه دلیـل تفـاوت در فرهنـگ، مـذهب و سـازمان، معیارهـای اخـلاقی نـیز متفـاوت اسـت، حـتی اگـر در یـک زمینـه باشـند. دسـتیابی بـه پیشـنهاد اسـتاندارد اخـلاقی یکپارچـه نـه تنهـا دشـوار اسـت، بلکـه غـیر ضروری اسـت. بنـابراین، چگـونگی دسـتیابی بـه همـاهنگی بـین اسـتانداردهای اخـلاقی کشـورها و سـازمان‌های مختلـف مهـم و بـه‌ویژه چالش‌برانگیز است.

ج. چالش در توسعه‌ی رویکردهای فناورانه برای کاهش مسائل اخلاقی هوش مصنوعی

در حـال حـاضر، بهبـود توضـیح پـذیری، انصـاف، حفاظـت از حـریم خصـوصی، امنیـت، اسـتحکام و سـایر شایسـتگی‌های مـرتبط بـا الزامـات هـوش مصـنوعی اخـلاقی از موضـوعات داغ تحقیقـاتی در جوامـع هـوش مصـنوعی هسـتند. بـا ایـن حـال، بیشـتر کارهـای تحقیقـاتی کنـونی از یـک بعـد واحـد از اصـول اخـلاقی انجـام می‌شـوند، بـرای مثـال، XAI بـر افـزایش تفسـیرپذیری هـوش مصـنوعی تمرکـز دارد و ML منصـفانه بـه کـاهش بی‌عـدالتی یـا سـوگیری ML اختصـاص دارد. هنـوز عـدم ادغـام اصـول یـا الزامـات اخـلاقی متعـدد در کـار تحقیقـاتی جـاری وجـود دارد. بـدیهی اسـت کـه ادغـام چنـد بعـد اخـلاقی کـه تعـادل هـم افـزایی بـین چنـدین اصـل اخـلاقی مختلـف را ممکـن می‌سـازد، بـرای سـاختن سیسـتم‌های هـوش مصـنوعی اخـلاقی کـه می‌تواننـد اصـول اخـلاقی متعـدد را بـرآورده کننـد، ضروری و حیـاتی اسـت. امـا ادغـام چنـد بعـد اخـلاقی در یـک سیسـتم هـوش مصـنوعی از طریـق رویکردهـای فناورانـه بـه دلیـل تضـاد یـا ناسازگاری بین الزامات اخلاقی مختلف بسیار چالش برانگیز است.

د. چالش در ارزیابی اخلاق در هوش مصنوعی

اخـلاق ذاتـاً یـک مفهـوم کـیفی اسـت کـه بـه بسـیاری از ویژگی‌هـایی کـه کـمّی کـردن آنهـا دشـوار اسـت، ماننـد ویژگی‌هـای فـرهنگی یـا نـژادی وابسـته اسـت. از ایـن رو، تعریـف دقیـق اخـلاق، اگـر نگـوییم غـیرممکن، بسـیار دشـوار اسـت. در نتیجـه، ارزیـابی اخـلاق هـوش مصـنوعی بسـته بـه افـرادی کـه هـوش مصـنوعی را ارزیـابی می‌کننـد، همیشـه دارای عناصر ذهنی است. این امر تحقیقات و کاربردهای اخلاق هوش مصنوعی را با چالش‌هایی مواجه می‌کند.

هـ. چشم اندازهای آینده

در ایـن بخـش بـه بـرخی دیـدگاه‌های آینـده اشـاره می‌شـود کـه ممکـن اسـت بـرای تحقیقـات آتی ارزشـمند باشـند. ابتـدا بـرای پیاده‌سـازی اخـلاق در هـوش مصـنوعی بایـد بـه ایـن نکتـه اشـاره کـرد کـه انسـان‌ها هرگـز از یـک نظریـه اخـلاقی اسـتفاده نمی‌کننـد، بلکـه بـا توجـه بـه موقعیـت یـا زمینـه‌ای کـه بـا آن مواجـه هسـتند، بـین نظریـه‌های مختلـف جابـه‌جا می‌شـوند [۱۳۴]. ایـن مسـئله نـه تنهـا بـه ایـن دلیـل اسـت کـه انسـان‌ها کارگزاران عقـلانی محـض نیسـتند کـه نظریـه اقتصـادی می‌خواهـد آن‌هـا را بـاور کنـیم، بلکـه بـه ایـن دلیـل اسـت کـه پـیروی دقیـق از هـر نظریـه اخـلاقی می‌توانـد منجـر بـه نتـایج نـامطلوب شـود. ایـن بـدان معناسـت کـه سیسـتم‌های هـوش مصـنوعی بایـد دارای بازنمـایی از نظریـه‌های اخـلاقی مختلـف و توانـایی انتخـاب بـین ایـن نظریـه‌های اخـلاقی باشـند. در اینجـا مـا ایـن رویکـرد را رویکـرد

چند نظریه‌ای می‌نامیم. در رویکرد چند تئوری، سیستم‌های هوش مصنوعی بسته به نوع موقعیت می‌توانند به جای یکدیگر نظریه‌های مختلفی را اعمال کنند. علاوه بر این، ترکیبی از تئوری‌های اخلاق هنجاری و اخلاق دامنه‌ی خاص که توسط متخصصان این حوزه پذیرفته شده است، شایسته‌ی پیاده‌سازی است زیرا یک سیستم هوش مصنوعی اخلاقی باید توسط کاربرانش پذیرفته شود.

از نظر رویکردهای فن‌آوری برای پرداختن به مسائل اخلاقی در هوش مصنوعی، توسعه ML جدید و سایر فناوری‌های هوش مصنوعی تحت هدایت دستورالعمل‌های اخلاقی و اصول بررسی‌شده در بخش ۴، مطلوب است. اگرچه در نظر گرفتن چندین اصل اخلاقی مختلف به طور همزمان هنگام طراحی عوامل جدید هوش مصنوعی چالش برانگیز است، اما یک گام بسیار مهم و ضروری در توسعه هوش مصنوعی اخلاقی در آینده خواهد بود.

از بررسی رویکردهای ارزشیابی اخلاقی، می‌توان دریافت که روش‌های ارزیابی موثر به فوریت مورد نیاز است زیرا ما باید سیستم هوش مصنوعی طراحی شده را قبل از استقرار ارزیابی کنیم. در حال حاضر، پیشنهاد یک روش ارزیابی کلی دشوار است. بنابراین، محققان اغلب بر حوزه‌های خاصی تمرکز می‌کردند و به وظایف ارزیابی شایستگی اخلاقی در این حوزه‌ها می‌پرداختند. معیارهای خاص دامنه، به عنوان مثال، مجموعه داده‌های جامع، برای آزمایش اخلاقی سیستم‌های هوش مصنوعی نیز برای برخی از زمینه‌های کاربردی حیاتی، مانند اتومبیل‌های خودران و مراقبت‌های بهداشتی مهم به نظر می‌رسند.

به عنوان آخرین مورد و نه کم اهمیت‌ترین آن، از آنجایی که هم طبیعت و هم تربیت، در شکل‌دهی رفتارهای اخلاقی مهم هستند، ما ترکیب اخلاق هنجاری و اخلاق تکاملی [۱۵۵] را برای طراحی سیستم‌های هوش مصنوعی اخلاقی پیشنهاد می‌کنیم. اخلاق هنجاری مانند توانایی‌های اخلاقی فطری است، در حالی که رویکرد اخلاق تکاملی می‌تواند با یادگیری و تکامل مستمر، شایستگی اخلاقی جدیدی کسب کند. این مسئله ممکن است یک مسیر امیدوارکننده برای توسعه سیستم هوش مصنوعی اخلاقی در آینده باشد.

بخش ۸ - نتیجه گیری

بر اساس بررسی ما از اخلاق هوش مصنوعی و پیچیدگی‌ها و چالش‌های فراوانی که در این مقاله توضیح داده شد، واضح است که تلاش برای پرداختن به مسائل اخلاقی در هوش مصنوعی و طراحی سیستم‌های هوش مصنوعی اخلاقی که قادر به رفتار اخلاقی باشند، کاری دشوار و پیچیده است. با این حال، اینکه آیا هوش مصنوعی می‌تواند نقش مهمی را در جامعه‌ی آینده ما ایفا کند تا حد زیادی به موفقیت سیستم‌های هوش مصنوعی اخلاقی بستگی دارد. نظم و انضباط اخلاق هوش مصنوعی مستلزم تلاش مشترک دانشمندان، مهندسان، فیلسوفان، کاربران و سیاستگذاران دولتی است.

این مقاله با خلاصه و تجزیه و تحلیل خطرات اخلاقی و مسائل مطرح شده توسط هوش مصنوعی، دستورالعمل‌ها و اصول اخلاقی صادر شده توسط سازمان‌های مختلف، رویکردهایی برای پرداختن به مسائل اخلاقی در هوش مصنوعی یا رعایت اصول اخلاقی هوش مصنوعی، و روش‌هایی برای ارزیابی اخلاق (یا معنویت) هوش مصنوعی و علاوه بر این موارد، به برخی از چالش‌ها در عمل اخلاق هوش مصنوعی و برخی از جهت‌گیری‌های تحقیقاتی آینده اشاره می‌شود.

با این حال، اخلاق هوش مصنوعی یک حوزه‌ی تحقیقاتی بسیار گسترده و چند رشته‌ای است. پوشش همه موضوعات ممکن در این زمینه با یک مقاله مروری غیرممکن است. امیدواریم این مقاله بتواند نقطه شروعی

بـرای افـرادی باشـد کـه بـه اخـلاق هـوش مصـنوعی علاقه‌منـد هسـتند تـا پیشـینه‌ی کـافی و دیـد پرنـده‌ای بـه دسـت آورند تا تحقیقات بیشتر توسط آنها انجام شود.

منابع

[1] M. Haenlein and A. Kaplan, “A brief history of artificial intelligence: On

the past, present, and future of artificial intelligence,” California Manage.

Rev., vol. 61, no. 4, pp. 5–14, 2019.

[2] R. Vinuesa et al., “The role of artificial intelligence in achieving the

sustainable development goals,” Nature Commun., vol. 11, no. 1, 2020,

Art. no. 233.HUANG et al.: OVERVIEW OF ARTIFICIAL INTELLIGENCE ETHICS 817

[3] Gartner, “Chatbots will appeal to modern workers,” 2019. Accessed: Feb. 10, 2022. [Online]. Available: https://www.gartner.com/

smarterwithgartner/chatbots-will-appeal-to-modern-workers

[4] M. J. Haleem, R. P. Singh, and R. Suman, “Telemedicine for healthcare:

Capabilities, features, barriers, and applications,” Sensors Int., vol. 2,

2021, Art. no. 100117.

[5] A. Morby, “Tesla driver killed in first fatal crash using autopilot,” 2016. Accessed: Feb. 10, 2022. [Online]. Available:

https://www.dezeen.com/2016/07/01/tesla-driver-killed-car-crashnews-driverless-car-autopilot/

[6] S. McGregor, Ed., “Incident number 6,” in AI Incident Database, 2016.

[Online]. Available: https://incidentdatabase.ai/cite/6

[7] R. V. Yampolskiy, “Predicting future AI failures from historic examples,”

Foresight, vol. 21, no. 1, pp. 138–152, 2019.

[8] C. Stupp, “Fraudsters used AI to mimic CEO’s voice in unusual cybercrime case: Scams using artificial intelligence are a

new challenge for companies,” 2019. Accessed: Feb. 10, 2022.

[Online]. Available: https://www.wsj.com/articles/fraudsters-use-ai-tomimic-ceos-voice-in-unusual-cybercrime-case-11567157402

[9] C. Allen, W. Wallach, and I. Smit, “Why machine ethics?,” IEEE Intell.

Syst., vol. 21, no. 4, pp. 12–17, Jul./Aug. 2006.

[10] M. Anderson and S. L. Anderson, “Machine ethics: Creating an ethical

intelligent agent," AI Mag., vol. 28, no. 4, pp. 15–26, 2007.

[11] K. Siau and W. Wang, "Artificial intelligence (AI) ethics," J. Database

Manage., vol. 31, no. 2, pp. 74–87, 2020.

[12] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics

guidelines," Nature Mach. Intell., vol. 1, no. 9, pp. 389–399, 2019.

[13] M. Ryan and B. C. Stahl, "Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications,"

JICES, vol. 19, no. 1, pp. 61–86, 2021.

[14] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A

survey on bias and fairness in machine learning," ACM Comput. Surv.,

vol. 54, no. 6, pp. 1–35, 2021.

[15] J. García and F. Fernández, "A comprehensive survey on safe reinforcement learning," J. Mach. Learn. Res., vol. 16, no. 42, pp. 1437–1480,

2015.

[16] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and

G. Srivastava, "A survey on security and privacy of federated learning,"

Future Gener. Comput. Syst., vol. 115, pp. 619–640, 2021.

[17] X. Liu et al., "Privacy and security issues in deep learning: A survey,"

IEEE Access, vol. 9, pp. 4566–4593, 2021.

[18] B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts,

taxonomies, opportunities and challenges toward responsible AI," Inf.

Fusion, vol. 58, pp. 82–115, 2020.

[19] Y. Zhang, M. Wu, G. Y. Tian, G. Zhang, and J. Lu, "Ethics and privacy

of artificial intelligence: Understandings from bibliometrics," Knowl.-

Based Syst., vol. 222, 2021, Art. no. 106994.

[20] D. Castelvecchi, "Can we open the black box of AI?," Nature, vol. 538,

no. 7623, pp. 20–23, 2016.

[21] S. Dilmaghani, M. R. Brust, G. Danoy, N. Cassagnes, J. Pecero, and

P. Bouvry, “Privacy and security of big data in AI systems: A research and standards perspective,” in Proc. IEEE Int. Conf. Big Data, 2019, pp. 5737–5743.

[22] J. P. Sullins, “When is a robot a moral agent?,” in Machine Ethics, M. Anderson and S. L. Anderson, Eds., Cambridge, U.K.: Cambridge Univ. Press, 2011, pp. 151–161.

[23] J. Timmermans, B. C. Stahl, V. Ikonen, and E. Bozdag, “The ethics of cloud computing: A conceptual review,” in Proc. IEEE 2nd Int. Conf. Cloud Comput. Technol. Sci., 2010, pp. 614–620.

[24] W. Wang and K. Siau, “Ethical and moral issues with AI: A case study on healthcare robots,” in Proc. 24th Americas Conf. Inf. Syst., 2018, pp. 1–5.

[25] I. Bantekas and L. Oette, International Human Rights Law and Practice. Cambridge U. K.: Cambridge Univ. Press, 2018.

[26] R. Rodrigues, “Legal and human rights issues of AI: Gaps, challenges and vulnerabilities,” J. Responsible Technol., vol. 4, 2020, Art. no. 100005.

[27] W. Wang and K. Siau, “Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda,” J. Database Manage., vol. 30, no. 1, pp. 61–79, 2019.

[28] W. Wang and K. Siau, “Industry 4.0: Ethical and moral predicaments,” Cutter Bus. Technol. J., vol. 32, no. 6, pp. 36–45, 2019.

[29] S. M. Liao, Ed., Ethics of Artificial Intelligence. New York, NY, USA: Oxford Univ. Press, 2020.

[30] A. Adadi, “A survey on data-efficient algorithms in big data era,” J. Big Data, vol. 8, no. 1, pp. 1–54, 2021.

[31] R. S. Geiger et al., “Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled

training data comes from?," in Proc. Conf. Fairness, Accountability,
Transparency, 2020, pp. 325–336.
[32] W. M. P. van der Aalst, V. Rubin, H. M. W. Verbeek, B. F. van Dongen,
E. Kindler, and C. W. Günther, "Process mining: A two-step approach to
balance between underfitting and overfitting," Softw. Syst. Model., vol. 9,
no. 1, pp. 87–111, 2010.
[33] Z. C. Lipton, "The mythos of model interpretability," Queue, vol. 16,
no. 3, pp. 31–57, 2018.
[34] Y. Wang and M. Kosinski, "Deep neural networks are more accurate than
humans at detecting sexual orientation from facial images," J. Pers. Social
Psychol., vol. 114, no. 2, pp. 246–257, 2018.
[35] D. Guera and E. J. Delp, "Deepfake video detection using recurrent neural
networks," in Proc. IEEE Int. Conf. Adv. Video Signal-based Surveill.,
2018, pp. 1–6.
[36] C. B. Frey and M. A. Osborne, "The future of employment: How susceptible are jobs to computerisation?," Technological Forecasting Social
Change, vol. 114, pp. 254–280, 2017.
[37] R. Maines, "Love + sex with robots: The evolution of human-robot
relationships (Levy, D.; 2007) [Book review]," IEEE Technol. Soc. Mag.,
vol. 27, no. 4, pp. 10–12, Dec. 2008.
[38] National AI Standardization General, "Artificial intelligence ethical risk
analysis report," 2019. Accessed: Apr. 19, 2022. [Online]. Available:
http://www.cesi.cn/201904/5036.html
[39] A. Hannun, C. Guo, and L. van der Maaten, "Measuring data leakage in
machine-learning models with fisher information," in Proc. 37th Conf.
Uncertainty Artif. Intell., 2021, pp. 760–770.
[40] A. Salem, M. Backes, and Y. Zhang, "Get a model! Model hijacking

attack against machine learning models," Nov. 2021. [Online]. Available:

https://arxiv.org/pdf/2111.04394

[41] A. Pereira and C. Thomas, "Challenges of machine learning applied to safety-critical cyber-physical systems," MAKE, vol. 2, no. 4,

pp. 579–602, 2020.

[42] J. A. McDermid, Y. Jia, Z. Porter, and I. Habli, "Artificial intelligence

explainability: The technical and ethical dimensions," Philos. Trans.. Ser.

A, Math. Phys. Eng. Sci., vol. 379, no. 2207, 2021, Art. no. 20200363.

[43] J.-F. Bonnefon, A. Shariff, and I. Rahwan, "The social dilemma of

autonomous vehicles," Science, vol. 352, no. 6293, pp. 1573–1576, 2016.

[44] B. C. Stahl and D. Wright, "Ethics and privacy in AI and big data: Implementing responsible research and innovation," IEEE Secur. Privacy,

vol. 16, no. 3, pp. 26–33, May/Jun. 2018.

[45] S. Ribaric, A. Ariyaeeinia, and N. Pavesic, "De-identification for privacy

protection in multimedia content: A survey," Signal Process., Image

Commun., vol. 47, pp. 131–151, 2016.

[46] A. Julia, L. Jeff, M. Surya, and K. Lauren, "Machine bias: There's

software used across the country to predict future criminals. And

it's biased against blacks," 2016. Accessed: Apr. 19, 2022. [Online]. Available: https://www.propublica.org/article/machine-bias-riskassessments-in-criminal-sentencing

[47] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias

against women," 2018. Accessed: Apr. 19, 2022. [Online]. Available:

https://www.reuters.com/article/us-amazon-com-jobs-automationinsight/amazon-scraps-secret-ai-recruiting-tool-that-showed-biasagainst-women-idUSKCN1MK08G

[48] D. Castelvecchi, "AI pioneer: 'The dangers of abuse are very real'," Nature, Apr. 4, 2019, [Online]. Available: https://www.nature.com/articles/

d41586-019-00505-2

[49] K. Hristov, "Artificial intelligence and the copyright dilemma," IDEA,

IP Law Rev., vol. 57, 2017, Art. no. 3. [Online]. Available: https://ssrn.

com/abstract=2976428

[50] C. Bartneck, C. Lütge, A. Wagner, and S. Welsh, “Responsibility and liability in the case of AI systems,” in An Introduction to Ethics in Robotics

and AI (SpringerBriefs in Ethics), C. Bartneck, C. Lütge, A. Wagner, and

S. Welsh, Eds., Cham, Switzerland: Springer, 2021, pp. 39–44.

[51] E. Bird, J. Fox-Skelly, N. Jenner, R. Larbey, E. Weitkamp, and A.

Winfield, “The ethics of artificial intelligence: Issues and initiatives,”

European Parliamentary Research Service, Brussels, Belgium, 2020. Accessed: Apr. 19, 2022. [Online]. Available: https://www.europarl.europa.

eu/thinktank/en/document/EPRS_STU(2020)634452

[52] C. Lutz, “Digital inequalities in the age of artificial intelligence and big

data,” Hum. Behav. Emerg. Technol., vol. 1, no. 2, pp. 141–148, 2019.

[53] L. Manikonda, A. Deotale, and S. Kambhampati, “What’s up with

Privacy? User preferences and privacy concerns in intelligent personal

assistants,” in Proc. AAAI/ACM Conf. AI, Ethics Soc., 2018, pp. 229–235.818 IEEE TRANSACTIONS ON ARTIFICIAL INTELLIGENCE, VOL. 4, NO. 4, AUGUST 2023

[54] D. Roselli, J. Matthews, and N. Talagala, “Managing bias in AI,” in Proc.

World Wide Web Conf., 2019, pp. 539–544.

[55] Y. Gorodnichenko, T. Pham, and O. Talavera, “Social media, sentiment

and public opinions:Evidencefrom #Brexit and #USElection,” Eur.Econ.

Rev., vol. 136, Jul. 2021, Art. no. 103772.

[56] N. Thurman, “Making ‘The daily me’: Technology, economics and

habit in the mainstream assimilation of personalized news,” Journalism,

vol. 12, no. 4, pp. 395–415, 2011.

[57] J. Donath, “Ethical issues in our relationship with artificial entities,” in

The Oxford Handbook of Ethics of AI. M. D. Dubber, F. Pasquale, and S.

Das, Eds., Oxford, U.K.: Oxford Univ. Press, 2020, pp. 51–73.

[58] E. Magrani, “New perspectives on ethics and the laws of artificial intelligence,” Internet Policy Rev., vol. 8, 2019, Art. no. 3.

[59] M. P. Wellman and U. Rajan, “Ethical issues for autonomous trading

agents,” Minds Mach., vol. 27, no. 4, pp. 609–624, 2017.

[60] U. Pagallo, “The impact of AI on criminal law, and its two fold procedures,” in Research Handbook on the Law of Artificial Intelligence,

W. Barfield and U. Pagallo, Eds., Cheltenham U.K.: Edward Elgar

Publishing, 2018, pp. 385–409.

[61] E. Dacoronia, “Tort law and new technologies,” in Legal Challenges in

the New Digital Age, A. M. López Rodríguez, M. D. Green, and M.

L. Kubica, Eds., Leiden, The Netherlands: Koninklijke Brill NV, 2021,

pp. 3–12.

[62] J. Khakurel, B. Penzenstadler, J. Porras, A. Knutas, and W. Zhang, “The

rise of artificial intelligence under the lens of sustainability,” Technologies, vol. 6, no. 4, 2018, Art. no. 100.

[63] S. Herat, “Sustainable management of electronic waste (e-Waste),” Clean

Soil Air Water, vol. 35, no. 4, pp. 305–310, 2007.

[64] E. Strubell, A. Ganesh, and A. McCallum, “Energy and policy considerations for deep learning in NLP,” in Proc. 57th Annu. Meeting Assoc.

Comput. Linguistics, 2019, pp. 3645–3650.

[65] V. Dignum, “Ethics in artificial intelligence: Introduction to the special

issue,” Ethics Inf. Technol., vol. 20, no. 1, pp. 1–3, 2018.

[66] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, “Algorithmic decision making and the cost of fairness,” in Proc. 23rd ACM

SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2017, pp. 797–806.

[67] R. Caplan, J. Donovan, L. Hanson, and J. Matthews, “Algorithmic accountability: A primer,” Data Soc., vol. 18, pp. 1–13, 2018.

[68] R. V. Yampolskiy, “On controllability of AI,” Jul. 2020. [Online]. Available: https://arxiv.org/pdf/2008.04071

[69] B. C. Stahl, J. Timmermans, and C. Flick, “Ethics of emerging information and communication technologies,” Sci. Public Policy, vol. 44, no. 3,

pp. 369–381, 2017.

[70] L. Vesnic-Alujevic, S. Nascimento, and A. Pólvora, “Societal and

ethical impacts of artificial intelligence: Critical notes on European policy frameworks,” Telecommun. Policy, vol. 44, no. 6, 2020,

Art. no. 101961.

[71] U. G. Assembly, “Universal declaration of human rights,” UN Gen.

Assem., vol. 302, no. 2, pp. 14–25, 1948.

[72] S. Russell, S. Hauert, R. Altman, and M. Veloso, “Robotics: Ethics of

artificial intelligence,” Nature, vol. 521, no. 7553, pp. 415–418, 2015.

[73] A. Chouldechova, “Fair prediction with disparate impact: A study of

bias in recidivism prediction instruments,” Big Data, vol. 5, no. 2,

pp. 153–163, 2017.

[74] J. van Dijck, “Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology,” Surveill. Soc., vol. 12, no. 2,

pp. 197–208, 2014.

[75] E. de Souza Nascimento, I. Ahmed, E. Oliveira, M. P. Palheta, I. Steinmacher, and T. Conte, “Understanding development process of machine

learning systems: Challenges and solutions,” in Proc. ACM/IEEE Int.

Symp. Empirical Softw. Eng. Meas., 2019, pp. 1–6.

[76] K. A. Crockett, L. Gerber, A. Latham, and E. Colyer, “Building trustworthy AI solutions: A case for practical solutions for

small businesses,” IEEE Trans. Artif. Intell., early access, 2021,

doi: 10.1109/TAI.2021.3137091.

[77] D. Leslie, “Understanding artificial intelligence ethics and safety:

A guide for the responsible design and implementation of AI

systems in the public sector,” 2019. Accessed: Apr. 19, 2022.

[Online]. Available: https://www.turing.ac.uk/research/publications/

understanding-artificial-intelligence-ethics-and-safety

[78] B. Buruk, P. E. Ekmekci, and B. Arda, “A critical perspective on guidelines for responsible and trustworthy artificial intelligence,” Med. Health

Care Philosophy, vol. 23, no. 3, pp. 387–399, 2020.

[79] UNESCO, “Recommendation on the ethics of artificial intelligence,”

2021. Accessed: Feb. 15 2022. [Online]. Available: https://en.unesco.

org/artificial-intelligence/ethics

[80] B. C. Stahl, Ed., Artificial Intelligence for a Better Future: An Ecosystem

Perspective on the Ethics of AI and Emerging Digital Technologies.

Cham, Switzerland: Springer, 2021.

[81] P. D. Motloba, “Non-maleficence - A disremembered moral obligation,”

South Afr. Dent. J., vol. 74, 2019, Art. no. 1.

[82] L. Floridi and J. Cowls, “A unified framework of five principles for AI

in society,” in Ethics, Governance, and Policies in Artificial Intelligence

(Philosophical Studies Series), vol. 144, L. Floridi, Ed. Cham, Switzerland: Springer, 2021, pp. 5–17.

[83] S. Jain, M. Luthra, S. Sharma, and M. Fatima, “Trustworthiness of

artificial intelligence,” in Proc. 6th Int. Conf. Adv. Comput. Commun.

Syst., 2020, pp. 907–912.

[84] L. Floridi et al., “AI4People-An ethical framework for a good AI society:

Opportunities, risks, principles, and recommendations,” Minds Mach.,

vol. 28, no. 4, pp. 689–707, 2018.

[85] R. Nishant, M. Kennedy, and J. Corbett, “Artificial intelligence for

sustainability: Challenges, opportunities, and a research agenda,” Int. J.

Inf. Manage., vol. 53, 2020, Art. no. 102104.

[86] C. S. Wickramasinghe, D. L. Marino, J. Grandio, and M. Manic, “Trustworthy AI development guidelines for human system interaction,” in

Proc. 13th Int. Conf. Hum. Syst. Interaction, 2020, pp. 130–136.

[87] V. Dignum, “Can AI systems be ethical?,” in Artificial Intelligence:

Foundations Theory and Algorithms, Responsible Artificial Intelligence.

V. Dignum, Ed., Cham, Switzerland: Springer, 2019, pp. 71–92.

[88] S. L. Anderson and M. Anderson, “AI and ethics,” AI Ethics, vol. 1, no. 1,

pp. 27–31, 2021.

[89] V. Dignum, “Ethical decision-making,” in Artificial Intelligence: Foundations, Theory, and Algorithms, Responsible Artificial Intelligence. V.

Dignum, Ed., Cham, Switzerland: Springer, 2019, pp. 35–46.

[90] G. Sayre-McCord, “Metaethics,” in The Stanford Encyclopedia

of Philosophy, E. N. Zalta, Ed., Stanford, CA, USA: Metaphys.

Res. Lab, Stanford Univ., 2014. [Online]. Available: https:

//plato.stanford.edu/entries/metaethics/#:~:text=Metaethics%20is%

20the%20attempt%20to,matter%20of%20taste%20than%20truth%3F

[91] Ethics | Internet Encyclopedia of Philosophy, 1995. Accessed: Aug. 2,

2021. [Online]. Available: https://iep.utm.edu/ethics/#SH2c

[92] R. Hursthouse and G. Pettigrove, “Virtue ethics,” in The Stanford Encyclopedia of Philosophy, E. N. Zalta, Ed. Stanford, CA, USA: Metaphys. Res. Lab, Stanford Univ., 2018. [Online]. Available: https://plato.

stanford.edu/entries/ethics-virtue/

[93] N. Cointe, G. Bonnet, and O. Boissier, “Ethical judgment of agents’

behaviors in multi-agent systems,” in Proc. Int. Conf. Auton. Agents

Multiagent Syst., 2016, pp. 1106–1114.

[94] H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, and Q. Yang, “Building

ethics into artificial intelligence,” in Proc. 27th Int. Joint Conf. Artif.

Intell., 2018, pp. 5527–5533.

[95] H. J. Curzer, Aristotle and the Virtues. New York, NY, USA: Oxford

Univ. Press, 2012.

[96] L. Alexander and M. Moore, “Deontological ethics,” in The Stanford

Encyclopedia of Philosophy, E. N. Zalta, Ed., 2020, Stanford, CA, USA:

Metaphys. Res. Lab, Stanford Univ., 2020. [Online]. Available: https:

//plato.stanford.edu/entries/ethics-deontological/

[97] W. Sinnott-Armstrong, “Consequentialism,” in The Stanford Encyclopedia of Philosophy, E. N. Zalta, Ed., 2019. Stanford, CA, USA: Metaphys.

Res. Lab, Stanford Univ., [Online]. Available: https://plato.stanford.edu/

entries/consequentialism/

[98] D. O. Brink, “Some forms and limits of consequentialism,” in Oxford Handbooks in Philosophy, The Oxford Handbook of Ethical Theory. D. Copp, Ed., New York, NY, USA: Oxford Univ. Press, 2006,

pp. 380–423.

[99] H. A. M. J. ten Have, Ed., Encyclopedia of Global Bioethics. Cham,

Switzerland: Springer, 2016.

[100] S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen, and A. Bernstein,

“Implementations in machine ethics: A survey,” ACM Comput. Surv.,

vol. 53, no. 6, pp. 1–38, 2021.

[101] C. Allen, I. Smit, and W. Wallach, “Artificial morality: Top-down,

bottom-up, and hybrid approaches,” Ethics Inf. Technol., vol. 7, no. 3,

pp. 149–155, 2005.

[102] W. Wallach and C. Allen, “Top-down morality,” in Moral Machines, W.

Wallach and C. Allen, Eds., Oxford, U.K: Oxford Univ. Press, 2009,

pp. 83–98.

[103] I. Asimov, “Runaround,” Astounding Sci. Fiction, vol. 29, no. 1,

pp. 94–103, 1942.

[104] J.-G. Ganascia, “Ethical system formalization using nonmonotonic ogics,” in Proc. Annu. Meeting Cogn. Sci. Soc., 2007,

pp. 1013–1018.HUANG et al.: OVERVIEW OF ARTIFICIAL INTELLIGENCE ETHICS 819

[105] K. Arkoudas, S. Bringsjord, and P. Bello, “Toward ethical robots via

mechanized deontic logic," in Proc. AAAI Fall Symp. Mach. Ethics, 2005, pp. 17–23.
[106] S. Bringsjord and J. Taylor, "Introducing divine-command robot ethics," in Robot Ethics: The Ethical and Social Implication of Robotics. 2012, pp. 85–108.
[107] N. S. Govindarajulu and S. Bringsjord, "On automating the doctrine of double effect," in Proc. 26th Int. Joint Conf. Artif. Intell., 2017, pp. 4722–4730.
[108] F. Berreby, G. Bourgne, and J.-G. Ganascia, "A declarative modular framework for representing and applying ethical principles," in Proc. 16th Conf. Auton. Agents MultiAgent Syst., 2017, pp. 96–104.
[109] V. Bonnemains, C. Saurel, and C. Tessier, "Embedded ethics: Some echnical and ethical challenges," Ethics Inf. Technol., vol. 20, no. 1, pp. 41–58, 2018.
[110] G. S. Reed, M. D. Petty, N. J. Jones, A. W. Morris, J. P. Ballenger, and H. S. Delugach, "A principles-based model of ethical considerations in military decision making," J. Defense Model. Simul., vol. 13, no. 2, pp. 195–211, 2016.
[111] L. Dennis, M. Fisher, M. Slavkovik, and M.Webster, "Formal verification of ethical choices in autonomous systems," Robot. Auton. Syst., vol. 77, pp. 1–14, 2016.
[112] A. R. Honarvar and N. Ghasem-Aghaee, "Casuist BDI-Agent: A new extended BDI architecture with the capability of ethical reasoning," in Proc. Int. Conf. Artif. Intell. Comput. Intell., 2009, pp. 86–95.
[113] A. S. Rao and M. P. Georgeff, "BDI agents: From theory to practice," in Proc. 1st Int. Conf. Multiagent Syst., 1995, pp. 312–319.
[114] S. Armstrong, "Motivated value selection for artificial agents," in Proc.

AAAI Workshop Artif. Intell. Ethics, Jan. 2015, pp. 12–20.

[115] U. Furbach, C. Schon, and F. Stolzenburg, “Automated reasoning in

deontic logic,” in Proc. 8th Int. Workshop Multi-Disciplinary Trends Artif.

Intell., 2014, pp. 57–68.

[116] D. Howard and I. Muntean, “Artificial moral cognition: Moral functionalism and autonomous moral agency,” in Philosophical Studies Series,

Philosophy and Computing, T. M. Powers, Ed. Cham, Switzerland:

Springer, 2017, pp. 121–159.

[117] Y.-H. Wu and S.-D. Lin, “A low-cost ethics shaping approach for designing reinforcement learning agents,” in Proc. 32nd AAAI Conf. Artif.

Intell., 2018, pp. 1687–1694.

[118] R. Noothigattu et al., “A voting-based system for ethical decision making,” in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 1587–1594.

[119] M. Guarini, “Particularism and the classification and reclassification of

moral cases,” IEEE Intell. Syst., vol. 21, no. 4, pp. 22–28, Jul./Aug. 2006.

[120] M. Anderson and S. L. Anderson, “GenEth: A general ethical dilemma

analyzer,” in Proc. 28th AAAI Conf. Artif. Intell., 2014, pp. 253–261.

[121] M. Azad-Manjiri, “A new architecture for making moral agents based on

C4.5 decision tree algorithm,” Int. J. Inf. Technol. Comput. Sci., vol. 6,

no. 5, pp. 50–57, 2014.

[122] L. Yilmaz, A. Franco-Watkins, and T. S. Kroecker, “Computational models of ethical decision-making: A coherence-driven reflective equilibrium

model,” Cogn. Syst. Res., vol. 46, pp. 61–74, 2017.

[123] T. A. Han, A. Saptawijaya, and L. M. Pereira, “Moral reasoning under

uncertainty,” in Logic For Programming, Artificial Intelligence, and

Reasoning. Berlin, Germany: Springer, 2012, pp. 212–227.

[124] M. Anderson, S. Anderson, and C. Armen, “Towards machine ethics

implementing two action-based ethical theories,” in Proc. AAAI Fall

Symp. Mach. Ethics, 2005, pp. 1–7.

[125] G. Gigerenzer, “Moral satisficing: Rethinking moral behavior as bounded rationality,” Topics Cogn. Sci., vol. 2, no. 3, pp. 528–554, 2010.

[126] J. Skorin-Kapov, “Ethical positions and decision-making,” in Professional and Business Ethics Through Film, J. Skorin-Kapov, Ed., New York, NY, USA: Springer, 2018, pp. 19–54.

[127] T.-L. Gu and L. Li, “Artificialmoral agents and their designmethodology: Retrospect and prospect,” Chin. J. Comput., vol. 44, pp. 632–651, 2021.

[128] C.Molnar, G. Casalicchio, and B. Bischl, “Interpretablemachine learning – A brief history, state-of-the-art and challenges,” in Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases, 2020, pp. 417–431.

[129] C. Molnar, Interpretable Machine Learning: A Guide For Making Black Box Models Interpretable. Morisville, NC, USA: Lulu, 2019.

[130] S. Feuerriegel, M. Dolata, and G. Schwabe, “Fair AI,” Bus. Inf. Syst. Eng., vol. 62, no. 4, pp. 379–384, 2020.

[131] S. Caton and C. Haas, “Fairness in machine learning: A survey,” Oct. 2020. [Online]. Available: https://arxiv.org/pdf/2010.04053

[132] S. E. Whang, K. H. Tae, Y. Roh, and G. Heo, “Responsible AI challenges n end-to-end machine learning,” Jan. 2021. [Online]. Available: https: //arxiv.org/pdf/2101.05967

[133] D. Peters, K. Vold, D. Robinson, and R. A. Calvo, “Responsible AI—Two frameworks for ethical design practice,” IEEE Trans. Technol. Soc., vol. 1, no. 1, pp. 34–47, Mar. 2020.

[134] V. Dignum, ed., Responsible Artificial Intelligence. Cham, Switzerland: Springer, 2019.

[135] C. Dwork, “Differential privacy: A survey of results,” in Proc. Int. Conf. Theory Appl. Models Computation, 2008, pp. 1–19.

[136] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, “Federated learning,” Synth. Lectures Artif. Intell. Mach. Learn., vol. 13, no. 3, pp. 1–207, 2019.

[137] M. Kirienko et al., “Distributed learning: A reliable privacy-preserving strategy to change multicenter collaborations using AI,” Eur. J. Nucl. Med. Mol. Imag., vol. 48, no. 12, pp. 3791–3804.2021.

[138] R. Shokri and V. Shmatikov, “Privacy-preserving deep learning,” in Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur., 2015, pp. 1310–1321.

[139] C. Meurisch, B. Bayrak, and M. Mühlhäuser, “Privacy-preserving AI services through data decentralization,” in Proc. Web Conf., 2020, pp. 190–200.

[140] UR-Lex - 02016R0679-20160504 - EN - EUR-Lex, 2016. Accessed: Jun. 28, 2021. [Online]. Available: https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=CELEX%3A02016R0679-20160504&qid=1532348683434

[141] R. E. Latta, H.R.3388 - 115th Congress (2017-2018): SELF DRIVE Act, 2017. Accessed: Jun. 28, 2021. [Online]. Available: https://www.congress.gov/bill/115th-congress/house-bill/3388

[142] 7. Lei No. 13, de 14 de Agosto de 2018, 2018. Accessed: Jun. 25, 2021. [Online]. Available: http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.html

[143] EUR-Lex - 52021PC0206 - EN - EUR-Lex, 2021. Accessed: Jun. 28, 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

[144] C. Allen, G. Varner, and J. Zinser, “Prolegomena to any future artificial moral agent,” J. Exp. Theor. Artif. Intell., vol. 12, no. 3, pp. 251–261,

2000.

[145] A. M. Turing, “Computing machinery and intelligence,” Mind, vol. LIX, no. 236, pp. 433–460, 1950.

[146] W. Wallach and C. Allen, Moral Machines: Teaching Robots Right From Wrong. Oxford, U.K.: Oxford Univ. Press, 2009.

[147] S. A. Seshia, D. Sadigh, and S. S. Sastry, “Towards verified artificial intelligence,” Jun. 2016. [Online]. Available: http://arxiv.org/pdf/1606.08514v4

[148] T. Arnold and M. Scheutz, “Against the moral turing test: Accountable design and the moral reasoning of autonomous systems,” Ethics Inf. Technol., vol. 18, no. 2, pp. 103–115, 2016.

[149] ACM Code of Ethics and Professional Conduct, 2018. Accessed: Jun. 25, 2021. [Online]. Available: https://www.acm.org/code-of-ethics

[150] IEEE SA - The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019. Accessed: Jun. 28 2021. [Online]. Available: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

[151] IEEE Ethics In Action | Ethically Aligned Design, IEEE 7000TM Projects, 2020. Accessed: Jun. 28, 2021. [Online]. Available: https://ethicsinaction.ieee.org/p7000/

[152] ISO, ISO/IEC JTC 1/SC 42 - Artificial intelligence, 2017. Accessed: Jun. 28, 2021. [Online]. Available: https://www.iso.org/committee/6794475.html

[153] B. Goehring, F. Rossi, and D. Zaharchuk, “Advancing AI ethics beyond compliance: From principles to practice,” IBM Corporation, Apr. 2020. Accessed: Apr. 19, 2022. [Online]. Available: https://www.ibm.com/thought-leadership/institute-business-value/report/ai-ethics

[154] Responsible AI, 2017. Accessed: Apr. 19, 2022. [Online]. Available:

https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:
primaryr6
[155] F. Allhoff, “Evolutionary ethics from Darwin to Moore,” Hist. Philosophy Life Sci., vol. 25, no. 1, pp. 51–79, 2003.