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 2016، 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 243,000 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 پژوهشی 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 منتشر 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:**

1. 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.
2. 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.
3. مرور دستورالعمل‌ها و اصول اخلاقی جهانی مربوط به هوش مصنوعی که توسط شرکت‌ها، سازمان‌ها و دولت‌ها منتشر شده است.
4. بررسی رویکردهای میان‌رشته‌ای برای مقابله با مشکلات اخلاقی هوش مصنوعی، از جمله رویکردهای اخلاقی، فناورانه و قانونی.
5. مرور روش‌های ارزیابی اخلاق هوش مصنوعی که جنبه‌ای کمتر مورد توجه در ادبیات موجود است.
6. 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.

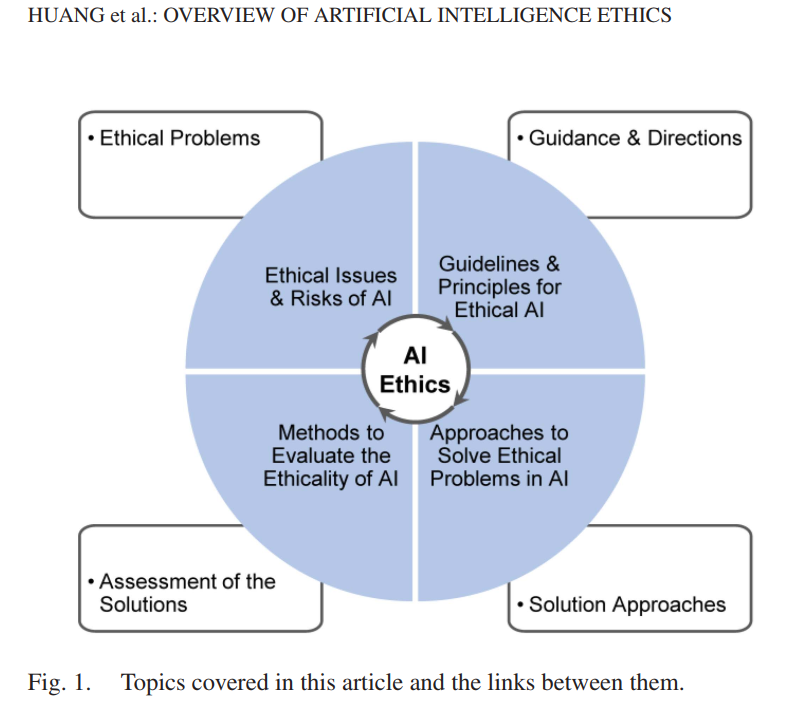
سپس دستورالعمل‌ها و اصول اخلاقی که توسعه و استفاده از هوش مصنوعی را هدایت می‌کنند، بررسی می‌شوند. با توجه به اینکه مسائل اخلاقی هوش مصنوعی توجه روزافزون بخش‌های مختلف جامعه را به خود جلب کرده است، بسیاری از سازمان‌ها (از جمله نهادهای آکادمیک، صنعت و دولت‌ها) به بحث و جستجوی چارچوب‌ها، دستورالعمل‌ها و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

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 1** 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 منعکس‌کننده 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 2010 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 اخلاقی هوش مصنوعی، فقط اسنادی را جمع‌آوری 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 اخلاقی هوش مصنوعی همراه با لینک‌های URL در مواد تکمیلی این مقاله ارائه شده است.

# بخش 3 - 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 بخش III-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 [11]، 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 فرهنگی**، **اعلامیه جهانی حقوق بشر**، **منشور سازمان ملل متحد** و **کنوانسیون اروپایی حفاظت از حقوق بشر و آزادی‌های بنیادین**، توسط دولت‌های مختلف منتشر شده‌اند.

**ج) تأثیر اجتماعی مسائل اخلاقی هوش مصنوعی**

**اتوماسیون و جایگزینی شغل**  
با جایگزینی تعداد بیشتری از کارگران کارخانه با سیستم‌های خودکار و ربات‌ها، هوش مصنوعی بازار کار را دچار اختلال و تحول می‌کند. بنابراین، بسیاری از افراد نگران اتوماسیون و جایگزینی شغلی هستند.

**دسترس‌پذیری**  
دسترس‌پذیری یا قابلیت استفاده از فناوری‌های نوظهور، مانند هوش مصنوعی، تأثیر مستقیمی بر رفاه انسان دارد. با این حال، اگر تنها بخشی از جمعیت از مزایای هوش مصنوعی بهره‌مند شوند، این امر غیراخلاقی و ناعادلانه خواهد بود. باید توجه شود که محصولات و خدمات هوش مصنوعی به گونه‌ای توسعه یابند که برای همه قابل دسترس باشند، به طوری که مزایای هوش مصنوعی به طور مساوی در میان همه گسترش یابد.

**دموکراسی و حقوق مدنی**  
هوش مصنوعی غیراخلاقی حقیقت را تحریف می‌کند و در نهایت منجر به از دست دادن اعتماد و حمایت عمومی از فناوری هوش مصنوعی می‌شود. قدرت دموکراسی‌ها با از دست دادن جوامع آگاه و اعتمادکننده آسیب می‌بیند. با آسیب‌دیدن دموکراسی‌ها و تشدید تعصب‌های ساختاری، بهره‌مندی آزادانه از حقوق مدنی دیگر به طور مداوم برای همه قابل دسترس نخواهد بود. بنابراین، دموکراسی و حقوق مدنی باید در اخلاق هوش مصنوعی مدنظر قرار گیرند.

### .2.1.3 دسته‌بندی بر اساس آسیب‌پذیری‌های هوش مصنوعی و انسان

در مرجع [29]، **لیائو** مسائل اخلاقی هوش مصنوعی را به دو دسته تقسیم می‌کند:

1. مسائل اخلاقی که به دلیل محدودیت‌های سیستم‌های یادگیری ماشین (ML) کنونی به وجود می‌آیند و به عنوان "آسیب‌پذیری‌های هوش مصنوعی (به ویژه یادگیری ماشین)" شناخته می‌شوند.
2. مسائل اخلاقی که به دلیل عملکرد بیش از حد خوب سیستم‌های یادگیری ماشین کنونی به وجود می‌آیند و انسان‌ها در حضور یا تعامل با این سیستم‌های هوشمند آسیب‌پذیر می‌شوند، که به عنوان "آسیب‌پذیری‌های انسانی" مطرح می‌شوند.

**الف. مسائل اخلاقی ناشی از آسیب‌پذیری‌های هوش مصنوعی**

**یادگیری ماشین تشنه‌ی داده است:** معمولاً یادگیری ماشین به مقدار زیادی داده نیاز دارد تا عملکرد خوبی داشته باشد. این امر شرکت‌ها و سازمان‌ها را به جمع‌آوری یا خرید داده، از جمله داده‌های حساس شخصی، حتی اگر این کار ممکن است حق حریم خصوصی فرد را نقض کند، ترغیب می‌کند.

**زباله وارد شود، زباله خارج می‌شود:** عملکرد الگوریتم یادگیری ماشین به شدت به داده‌هایی که از آن‌ها یاد می‌گیرد وابسته است. اگر یک الگوریتم یادگیری ماشین با داده‌های ناکافی یا نادرست آموزش داده شود، حتی اگر 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 می کنند و حجم عظیمی از داده ها با سرعت بالایی تولید می شود که نیاز به استفاده از الگوریتم هایی برای هضم و تجزیه و تحلیل آن دارد. علاوه بر این، به دلیل پویایی بازارها، واکنش سریع به اطلاعات بسیار مهم است [59]، که انگیزه های قابل توجهی را برای جایگزینی فرآیند تصمیم گیری کُند افراد با تصمیم گیری الگوریتمی فراهم می کند. علاوه بر این، جوایز برای تصمیمات تجاری موثر قابل توجه است، که توضیح می دهد که چرا شرکت ها 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 طرف های بی گناه از راه های 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 اخلاقی خاص خود هستند، در این صورت قوانین جزایی ما و حتی کل سیستم حقوقی ما باید مورد بازنگری قرار گیرند [60].

**قانون شکنجه:** قانون شکنجه شرایطی مانند آسیب رفتاری یک فرد، رنج، ضرر ناعادلانه یا آسیب رساندن به شخص دیگر را پوشش می دهد. هنگامی که تصادفی با خودرو(های) خودران رخ می دهد، دو حوزه قانونی مرتبط وجود دارد: سهل انگاری و مسئولیت محصول. در حالی که امروزه بیشتر تصادفات، ناشی از خطای راننده است که نشان می دهد مسئولیت تصادفات بر اساس اصل سهل انگاری تنظیم می شود. بنابراین، در آینده، قانون جرم، که شامل انواع مختلفی از دعاوی آسیب شخصی است، به طور قابل توجهی تحت تأثیر قرار خواهد گرفت [61] زیرا محصولات هوش مصنوعی (مانند اتومبیل های خودران یا سایر ربات های هوشمند) در دعاوی صدمات شخصی دخیل خواهند بود. به عنوان تصادف بین اتومبیل های خودران یا ادعای جراحت که در آن ربات به انسان آسیب می رساند.

**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 محیط های جدید و پیچیده تر برای استخراج کار کنند. این امر باعث افزایش میزان تولید و مصرف فلزات کمیاب خاکی و آسیب بیشتر به محیط زیست می شود [62].

**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 [63].

**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 داده های فهرست شده در [64]، اثر کربن آموزش یک مدل پردازش زبان طبیعی (مدل ترانسفورماتور) تقریباً 5 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 می دهد. این مسئله در حال تغییر زندگی روزمره ما در بسیاری از حوزه ها، مانند حمل و نقل، صنعت خدمات، مراقبت های 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 [65]. 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 می شود.

**انصاف:** اعتماد به هوش مصنوعی، باید منصفانه و بی طرفانه باشد. همانطور که تصمیمات بیشتر و بیشتری به هوش مصنوعی واگذار می شود، ما باید اطمینان حاصل کنیم که این تصمیمات عاری از تعصب و تبعیض هستند [66]. چه فیلتر کردن رزومه ها برای مصاحبه های شغلی باشد یا تصمیم گیری در مورد پذیرش در دانشگاه یا انجام رتبه بندی اعتباری برای شرکت های 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 [51].

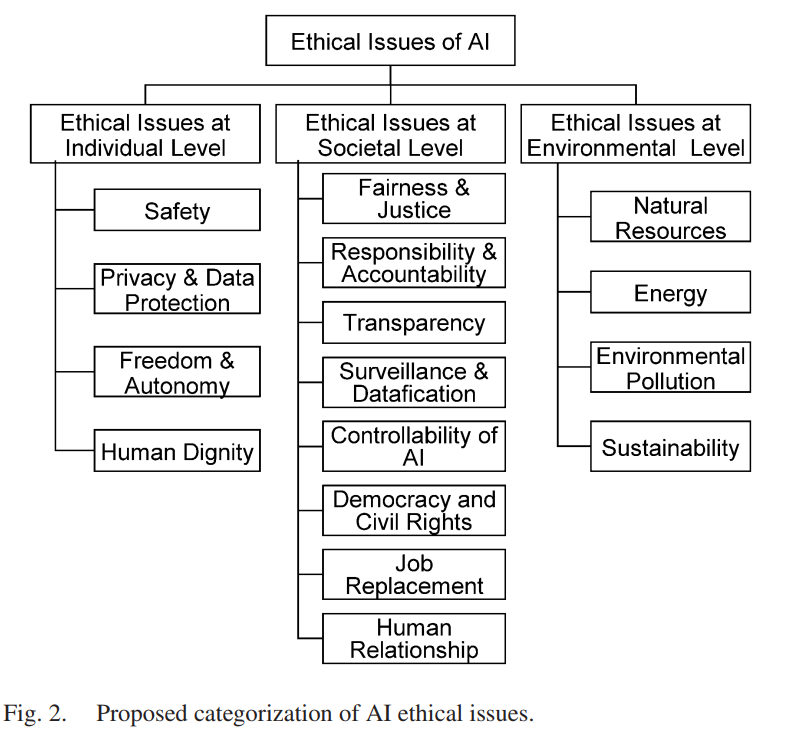
**مسئولیت پذیری:** مسئولیت پذیری [67] تضمین می کند که اگر یک سیستم هوش مصنوعی مرتکب اشتباهی شود یا به کسی آسیب برساند، می توان مسئولیت آن را بر عهده گرفت، خواه طراح باشد، توسعه دهنده یا شرکتی باشد که هوش مصنوعی را می فروشد. در صورت بروز خسارت، پاسخگویی برای ایجاد یک مکانیسم اصلاحی ضروری است تا قربانیان 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 می گذارد، کنترل پذیری هوش مصنوعی است [68]. این مسئله تا حد زیادی به ترس مردم از ایده‌ی "اَبَر هوش (super-intelligence)" مربوط می شود که به معنی افزایش هوش AI به حدی است که از توانایی های انسانی پیشی می گیرد، ممکن است هوش مصنوعی کنترل منابع ما را به دست گرفته و از گونه های 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 انسان ها نظارت یا کنترل نهایی بر فناوری هوش مصنوعی داشته باشند.

**TABLE I** List and Discussion of the Reviewed Categorization of Ethical Issues of AI and Our Proposed Categorization

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

در بخش قبل، ما مسائل اخلاقی هوش مصنوعی را که در بخش ادبیات تحقیق، شرح داده شده و طبقه بندی شده اند، مرور کرده ایم (جدول 1 را ببینید). با این حال، دسته بندی های ارائه شده در بالا دارای نقص های آشکار هستند. به طور خاص، طبقه‌بندی بر اساس ویژگی‌های هوش مصنوعی، عوامل انسانی و تأثیرات اجتماعی [11]، به وضوح تأثیر هوش مصنوعی بر محیط‌زیست، مانند مصرف منابع طبیعی و آلودگی محیط‌زیست را نادیده می‌گیرد. طبقه بندی بر اساس آسیب پذیری های هوش مصنوعی و انسان [29] چندین موضوع مهم مانند مسئولیت، ایمنی و مشکلات زیست محیطی را حذف می کند. طبقه‌بندی بر اساس الگوریتم، داده‌ها، کاربردها و ریسک‌های اخلاقی بلندمدت و غیرمستقیم [38] ملاحظات انصاف، استقلال و آزادی، کرامت انسانی، مشکلات زیست‌محیطی و غیره را نادیده می‌گیرد. اگرچه طبقه بندی بر اساس استقرار هوش مصنوعی [51] مسائل اخلاقی را به طور جامع پوشش می دهد، 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 می دهد تا مسائل اخلاقی هوش مصنوعی را بیشتر تحلیل و مرتب کنیم.

بدون شک سیستم های 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 جنبه های 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 می شود که پیامدهای نامطلوبی برای افراد انسانی، حقوق و رفاه آنها دارد [69]. مسائل اخلاقی هوش مصنوعی در سطح اجتماعی، پیامدهای اجتماعی را در نظر می گیرد که هوش مصنوعی برای گروه ها یا جامعه‌ای از افراد به عنوان یک کل، به ارمغان آورده یا ممکن است به همراه داشته باشد [69]. مسائل اخلاقی هوش مصنوعی در سطح زیست محیطی بر تأثیرات هوش مصنوعی بر محیط طبیعی متمرکز است. دسته بندی پیشنهادی ما در شکل 2 نشان داده شده است.

### .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 ربات ها رخ داده و گزارش شده است. مسئله حریم خصوصی یکی از خطرات جدی است که هوش مصنوعی برای ما به ارمغان می آورد. 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 چالش هایی 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 [70]. کرامت انسانی که یکی از حقوق اصلی بشر است، مربوط به حق احترام و رفتار اخلاقی انسان است [71]. 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 فناوری های هوش مصنوعی باید مورد احترام قرار گیرد. به عنوان مثال، یک سیستم تسلیحاتی خودمختار مرگبار [72] 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 گسترده ای که هوش مصنوعی برای جامعه و رفاه جوامع و ملل در سراسر جهان به ارمغان می آورد تمرکز می کنیم. 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 می کنیم.

وجود تعصب و تبعیض در هوش مصنوعی چالش هایی را برای انصاف و عدالت ایجاد کرده است. سوگیری ها و تبعیض های تعبیه شده در هوش مصنوعی ممکن است شکاف های اجتماعی را افزایش داده و به گروه های اجتماعی خاصی آسیب برساند [70]. 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 [73]. 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 حل های حقوقی منعکس می شود [70]. بنابراین، مکانیسم هایی باید ایجاد شود تا از مسئولیت و پاسخگویی سیستم های هوش مصنوعی و نتایج آن ها قبل و بعد از اجرای آنها اطمینان حاصل شود. به دلیل ماهیت جعبه سیاه الگوریتم های 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 سیستم های 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 [74]، از دغدغه های رایج هستند. داده ها از زندگی روزمره کاربران از طریق دستگاه های هوشمند جمع آوری می شود و ما تحت نظارت انبوه، زندگی می کنیم. 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 می گیرند.

### .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، دستگاه های ذخیره سازی و غیره است. تولید این سخت افزارها منابع طبیعی زیادی به خصوص برخی عناصر کمیاب را مصرف می کند. علاوه بر این، در پایان چرخه عمر، این سخت افزارها معمولاً دور ریخته می شوند که باعث آلودگی جدی زیست محیطی خواهند شد. جنبه مهم دیگر این است که سیستم های 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 [2]. به طور خلاصه، مصرف منابع طبیعی، آلودگی محیط زیست، هزینه های مصرف انرژی و پایداری درگیر در توسعه هوش مصنوعی مسائل و نگرانی های اصلی در سطح زیست محیطی هستند.

دسته بندی پیشنهادی ما مسائل اخلاقی را از سه سطح اصلی، یعنی تأثیر هوش مصنوعی بر فرد، جامعه و محیط، روشن می کند. مهم نیست که هوش مصنوعی در چه زمینه یا بخشی استفاده می شود، می توانیم مسائل اخلاقی 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 اخلاقی هوش مصنوعی را پوشش می دهد.

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

پس از بررسی مسائل اخلاقی و خطرات مورد بحث در ادبیات، ما در مورد مسائل اخلاقی مرتبط با مراحل مختلف چرخه حیات یک سیستم هوش مصنوعی بحث می کنیم. اگر بدانیم مشکلات اخلاقی 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 انگیزه ای برای بحث در مورد مسائل اخلاقی 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 [75] یا محصول [76]، 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 می شود 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 ویژگی ها و سایر عملیات مربوط به داده ها می پردازد. پس از این، ادامه‌ی فرآیند، وارد مرحله‌ی به اصطلاح مدل سازی ML می شود. 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 اخلاقی را با مراحل چرخه‌ی عمر هوش مصنوعی مرتبط می‌کند، جایی که این ارتباط به این معنی است که موضوع اخلاقی 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 2 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.

**TABLE II** Ethical Considerations Along Each Stage of the AI Lifecycle

# بخش 4 - 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 [78]. این اصول و دستورالعمل ها، دستورالعمل های مفیدی را برای تمرین هوش مصنوعی اخلاقی ارائه می دهند. این بخش به ارائه‎ی یک چشم انداز جهانی به روز از دستورالعمل ها و اصول اخلاق هوش مصنوعی اختصاص دارد که از طریق بررسی 146 گزارش دستورالعمل و توصیه مربوط به اخلاق هوش مصنوعی منتشر شده توسط شرکت ها، سازمان ها و دولت ها از سال 2015 در سراسر جهان به دست می آید. این اصول و دستورالعمل ها، راهنمایی های سطح بالایی را برای برنامه ریزی، توسعه، تولید، و استفاده از هوش مصنوعی و نیز دستورالعمل هایی برای پرداختن به مسائل اخلاقی هوش مصنوعی ارائه می دهند.

## .1.4 رهنمودهایی برای اخلاق هوش مصنوعی

یک بررسی و تجزیه و تحلیل عالی از اصول و دستورالعمل های فعلی در مورد هوش مصنوعی اخلاقی در سال 2019 توسط Jobin و همکاران ارائه شده است [12]، که بررسی 84 دستورالعمل اخلاقی منتشر شده توسط سازمان های ملی یا بین المللی از کشورهای مختلف را انجام داده‌اند. جوبین و همکاران [12] 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 24 نوامبر 2021، یونسکو (سازمان آموزشی، علمی و فرهنگی ملل متحد) توصیه ای را در مورد اخلاق هوش مصنوعی تصویب کرد که اولین توافق جهانی در مورد اخلاق هوش مصنوعی است [79]. برای به‌روزرسانی و غنی‌سازی تحقیقات در مورد دستورالعمل‌ها و اصول هوش مصنوعی اخلاقی، بر اساس جدول دستورالعمل‌های اخلاقی هوش مصنوعی که در مقاله Jobin [12] ارائه شده است (فقط شامل 84 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 146 دستورالعمل اخلاقی AI جمع آوری شده است. فهرستی از تمام دستورالعمل‌ها یا اسناد جمع‌آوری‌شده در جدول 5 از مواد تکمیلی آورده شده است. تعداد دستورالعمل های صادر شده در هر سال از سال 2015 تا 2021 در جدول 3 شمارش و فهرست شده است. واضح است که اکثر دستورالعمل ها در پنج سال گذشته منتشر شده اند، یعنی از سال 2016 تا 2020. تعداد راهنماهای منتشر شده در سال 2018 با 53 مورد، که 36.3 درصد از کل تعداد را به خود اختصاص داده است، بیشترین تعداد راهنما را به خود اختصاص داده است. علاوه بر این، تعداد دستورالعمل های AI صادر شده توسط هر کشور در جدول 4 فهرست شده است. علاوه بر این، درصد دستورالعمل های منتشر شده توسط انواع مختلف صادرکنندگان (از جمله دولت، صنعت، دانشگاه و سایر سازمان ها) در شکل 3 نشان داده شده است. از شکل 3 می توان دریافت که دولت‌ها، شرکت‌ها و دانشگاه‌ها همه، نگرانی‌های شدیدی در مورد اخلاق هوش مصنوعی نشان داده‌اند.

Fig. 3.Percentage of guidelines released by different types of issuers.

**TABLE III** Number of Documents Issued Each Year From 2015 to 2021

**TABLE IV** Number of Guidelines Issued By Each Country or Region

می، ها، ی، ای، های، هایی، تر، ترین،