OECD publishing

نقشه‌برداری مرتبط جمع آوری‌داده ها مکانیسم هایی‌برای

آموزش‌هوش مصنوعی

مصنوعیOECD
اسناداطلاعاتی
اکتبر۲۰۲۵شماره ۴۸

OECD
BETTER POLICIES FOR BETTER LIVES

مقالات‌هوش مصنوعی OECD

نقشه برداری‌از مجموعه داده های مرتبط
مکانیسم های‌آموزش هوش مصنوعی

پیشگفتار

این‌مقاله توسط سرگی گالوز دوران، به سرپرستی کلاریس ژیروت، با پشتیبانی کریستین ریمزباخ-کوناتزه نوشته شده است. بررسی سرمقاله و کمک برای انتشار توسط آندریا فورتادوارائه شده است. این مقاله شامل بازخورد نمایندگان کمیته سیاست دیجیتال)DPC(OECD و همچنین نمایندگان گروه کاری آن در زمینه حاکمیت داده و حریم خصوصی(DGP) است. مشارکت جهانی در زمینه هوش مصنوعی (GPAI) نیز در طول جلسه عمومی پاییزی 2024 خود در مورد این کار بحث کرد. نویسنده با کمال تشکر از نظرات‌ارزشمند لیمور شمرلینگ ماگازانیک، سارا بروبه، آنتونیا فون بورن-فالویس، سلین کایرا و کاسومی سوگیموتو قدردانی می کند. این مقاله به طور قابل توجهی از مشارکت های‌خردمندانه اعضای گروه متخصص OECD در زمینه هوش مصنوعی، داده و حریم خصوصی بهره برده است. این مقاله در 18 ژوئیه 2025 توسط کمیته سیاست‌دیجیتال (DPC) تصویب و از طبقه بندی خارج شد و توسط دبیرخانه OECD برای انتشار آماده شد.

یادداشت‌برای هیئت های نمایندگی:

این‌سند همچنین در ONE Members & Partners با کد مرجع زیر موجود است:

DSTI/DPC/DGP)2025(4/نهایی

چکیده

هنگام‌توسعه سیستم های هوش مصنوعی، متخصصان اغلب بر ساخت مدل تمرکز می کنند، درحالی که گاهی اوقات اهمیت تجزیه و تحلیل مکانیسم های متنوع جمع آوری داده ها را دست‌کم می گیرند. با این حال، تنوع مکانیسم های مورد استفاده برای جمع آوری داده ها شایسته‌توجه دقیق تری است زیرا هر یک از آنها پیامدهای متفاوتی برای توسعه دهندگان هوش‌مصنوعی، صاحبان داده ها و سایر دارندگان حقوقی که داده های آنها جمع آوری شده است،دارد. این مقاله سیاستی، مکانیسم های اصلی مورد استفاده در حال حاضر برای تهیه داده هابرای آموزش سیستم های هوش مصنوعی را ترسیم می کند و یک طبقه بندی برای پشتیبانی‌از مباحث سیاستی پیرامون حریم خصوصی، مدیریت داده ها و توسعه مسئولانه هوش‌مصنوعی پیشنهاد می دهد.

فهرست‌مطالب

ارقام

جعبه ها

خلاصه‌اجرایی

عملکردو قابلیت اطمینان مدل های هوش مصنوعی (AI) ارتباط نزدیکی با کیفیت و تنوع داده های مورد استفاده در آموزش آنها دارد.در حالی که جنبه های فنی توسعه مدل اغلب در مرکز توجه قرار می گیرند،**روش های اساسی برای تهیه و جمع آوری داده های‌آموزشی نیز به همان اندازه مرتبط هستند.**مکانیسم های مختلف جمع آوری داده ها، مزایا و چالش های متمایزی رانه تنها برای توسعه دهندگان هوش مصنوعی که به دنبال مدل های قوی و نماینده تر هستند، بلکه برای افرادی که داده هایشان ممکن‌است شامل شود، به همراه دارد. در عمل، توسعه دهندگان هوش مصنوعی اغلب از چندین مکانیسم جمع آوری داده به طورهمزمان برای ساخت مجموعه داده های آموزشی جامع استفاده می کنند.**درک این مکانیسم ها برای پیشرفت سیستم های‌هوش مصنوعی قابل اعتماد و برای پرداختن به ملاحظات مربوط به حریم خصوصی و مدیریت داده ها در فرآیندتوسعه ضروری است.**

براین اساس، این مقاله**نقشه برداری و طبقه بندی از سازوکارهای اصلی که در حال حاضر برای تهیه داده ها برای آموزش سیستم های‌هوش مصنوعی استفاده می شوند را ارائه می دهد. هدف این طبقه بندی، ارائه مبنایی برای تجزیه و تحلیل های‌آینده در مورد پیامدهای حریم خصوصی و مدیریت داده ها در هر سازوکار است.**

این‌طبقه بندی، این مکانیسم های کلیدی جمع آوری داده ها را در ساختار زیر سازماندهی می کند:

۱.داده های جمع آوری شده به طور مستقیم از افراد و سازمان ها

- **داده های‌ارائه شده و مشاهده شده**حجم فزاینده ای از داده های آموزشی از داده هایی که توسط افراد ارسال می شود یا به صورت‌غیرفعال در طول تعاملات آنها با سیستم های هوش مصنوعی، به ویژه در محیط های کسب و کار به مصرف کننده (B2C) مانندچت بات ها، دستیاران مجازی و میزهای کمک خودکار، جمع آوری می شود، سرچشمه می گیرد. علاوه بر این، برخی از توسعه دهندگان‌هوش مصنوعی، مانند پلتفرم های رسانه های اجتماعی، ممکن است از داده های ارائه شده یا مشاهده شده از افراددر سراسر مجموعه گسترده تر خود برای پشتیبانی از آموزش مدل هوش مصنوعی استفاده کنند.
- **کمک های‌داوطلبانه ی داده ها:**اگرچه هنوز در حال ظهور است، اما مشارکت های داوطلبانه افراد یا سازمان ها در جمع آوری‌داده ها، پتانسیل غنی سازی مجموعه داده های آموزشی با اطلاعات متنوع و واقعی را فراهم می کند که در غیر این صورت‌دسترسی به آنها دشوار خواهد بود.

۲.داده های جمع آوری شده از ارائه دهندگان شخص ثالث

- **صدورمجوز داده های تجاری:**توافق نامه های صدور مجوز داده ها با سازمان ها، راه دیگری را برای توسعه دهندگان هوش مصنوعی‌جهت دسترسی به مجموعه داده ها فراهم می کند. بازارهای داده و دلالان داده نقش مهمی را به عنوان واسطه های داده دراین اکوسیستم ایفا می کنند و دسترسی به طیف گسترده ای از داده های شخص ثالث را فراهم می کنند.
- **رویه های‌غیرتجاری:**توسعه دهندگان هوش مصنوعی همچنین ممکن است مجموعه داده ها را از طریق روش های غیرتجاری‌به دست آورند. ابتکارات داده باز، شامل داده های بخش دولتی و خصوصی که تحت مجوزهای باز منتشر می شوند،منابع کلیدی برای توسعه مدل های هوش مصنوعی هستند. مشارکت کنندگان مهم در این زمینه، ناشران مجموعه داده‌هستند که مجموعه داده ها را از منابع مختلف گردآوری و سازماندهی می کنند و آنها را به صورت رایگان و آشکار در دسترس‌قرار می دهند. با توجه به نیاز به مجموعه داده های بزرگ و متنوع برای پشتیبانی از فرآیندهای آموزش هوش مصنوعی،جمع آوری داده ها به عنوان یک مکانیسم جمع آوری داده که به طور گسترده برای رسیدگی به این نیازها پذیرفته شده‌است، ظهور کرده است.

باتوسعه این طبقه بندی،**این مقاله به سیاست گذاران و ذینفعان رویکردی ساختاریافته برای بحث های سیاسی در مورد حریم‌خصوصی، مدیریت داده ها و توسعه قابل اعتماد هوش مصنوعی ارائه می دهد.**این خروجی، پیچیدگی و تنوع مکانیسم های‌جمع آوری داده ای را که توسعه دهندگان هوش مصنوعی به آنها متکی هستند، برجسته می کند و خاطرنشان می سازدکه رویکردهای نوظهور شامل محیط های پردازش امن و ابزارهایی مانند ابزارهای افزایش حریم خصوصی

-۶

فناوری ها(PETs) روش هایی را برای بهبود قابلیت استفاده از این سازوکارهای جمع آوری داده ها ارائه می دهند، ضمن اینکه از حریم‌خصوصی و سایر حقوق و منافع مانند مالکیت معنوی محافظت می کنند. این طبقه بندی، زمینه را برای تجزیه و تحلیل بیشتردر مورد چگونگی ایجاد تعادل بین تقاضای رو به رشد برای داده های آموزشی هوش مصنوعی (از نظر حجم و تنوع) فراهم می کند،در حالی که جنبه های حریم خصوصی و مدیریت داده ها مانند کیفیت داده ها و قابلیت ردیابی را نیز در نظر می گیرد.

۱ مقدمه

ساخت‌مدل های یادگیری ماشینی هوش مصنوعی اغلب (البته نه به صورت سیستماتیک) به حجم زیادی از داده ها نیاز دارد که ممکن‌است عمداً یا سهواً شامل داده های شخصی باشد. در حالی که نیازهای خاص داده ها بسته به نوع و هدف مدل متفاوت است،دسترسی بیشتر به داده ها عموماً مدل های هوش مصنوعی را قادر می سازد تا عملکرد بهتری داشته باشند، زیرا توانایی یادگیری‌از نقاط داده و الگوهای استنباط شده در یک فرآیند تکراری را دارند (OECD، 2022)[1]داشتن داده های متنوع و باکیفیت (مثلاً دقت، کامل بودن، ثبات، قابلیت اطمینان، اعتبار، به موقع بودن) به همان اندازه در پشتیبانی از توسعه سیستم های هوش مصنوعی‌قوی و قابل اعتماد اهمیت دارد، زیرا داده های بهتر می توانند به رفع سوگیری ها، کاهش خطاها و محدود کردن نتایج ناخواسته‌کمک کنند.

هنگام‌توسعه سیستم های هوش مصنوعی، متخصصان اغلب بر ساخت مدل - از جمله وزن ها و پارامترها - تمرکز می کنند، در حالی‌که گاهی اوقات اهمیت تجزیه و تحلیل مکانیسم های متنوع جمع آوری داده ها را دست کم می گیرند. با این حال، تنوع مکانیسم های‌مورد استفاده برای جمع آوری داده ها شایسته توجه دقیق تری است زیرا هر یک از آنها پیامدهای متفاوتی برای توسعه دهندگان‌هوش مصنوعی، صاحبان داده ها و سایر دارندگان حق که داده های آنها جمع آوری شده است، دارد (OECD ، 2022)[1]در همین راستا، توصیه نامه OECD در مورد هوش مصنوعی (از این پس، «اصول هوش مصنوعی OECD») (OECD ، ۲۰۱۹)[2] بر اهمیت احترام به حریم خصوصی و اولویت بندی سازوکارهای خاص جمع آوری داده ها که مجموعه داده های با کیفیت‌بهتری ارائه می دهند و در مقایسه با سایر سازوکارها، به صاحبان داده ها کنترل بیشتری بر داده هایشان می دهند، تأکید دارند.همانطور که در ادامه بررسی می شود، روش هایی مانند جمع آوری داده ها اغلب بدون اطلاع افراد رخ می دهد و اعمال حقوق‌قابل اجرا، مانند حقوق قانونی برای اطلاع رسانی، دسترسی، حذف و اصلاح داده هایشان را دشوار می کند.

کارهای‌قبلی OECD، به ویژه «حفاظت از حریم خصوصی در اقتصاد داده محور» (OECD، ۲۰۱۴)[3] و «افزایش دسترسی و اشتراک گذاری‌داده ها» (OECD، ۲۰۱۹)[4]، داده ها را بر اساس نحوه جمع آوری آنها طبقه بندی می کند و بین داده های داوطلبانه، داده های‌مشاهده شده، داده های مشتق شده و داده های اکتسابی تمایز قائل می شود. طبقه بندی منشأ داده ها نیز در چارچوبOECD برای طبقه بندی سیستم های هوش مصنوعی گنجانده شده است (OECD، 2022).[1]. با پیشرفت سیستم های هوش‌مصنوعی، آنها نه تنها به مقادیر زیادی داده، بلکه به مجموعه داده های متنوع و با کیفیت بالا نیز نیاز دارند. این نیاز دوگانه یک‌سوال کلیدی را مطرح می کند: چگونه می توانیم دسترسی و اشتراک گذاری داده ها را برای آموزش هوش مصنوعی - هم از نظرحجم و هم از نظر کیفیت - تسهیل کنیم و در عین حال از منافع داده های افراد محافظت کنیم؟

دراین زمینه، سازمان همکاری و توسعه اقتصادی (OECD) در سال ۲۰۲۴ تصمیم گرفت یک تحلیل ساختاری در مورد مکانیسم های مختلف مورد استفاده‌برای جمع آوری داده های شخصی برای آموزش سیستم های هوش مصنوعی انجام دهد. چنین تحلیلی به ویژه امروزه اهمیت دارد، زیرا چارچوب های‌حفظ حریم خصوصی و مدیریت داده ها به طور فزاینده ای روش های جمع آوری داده ها برای ساخت سیستم های هوش مصنوعی توسط متخصصان‌هوش مصنوعی را شکل می دهند. این امر همچنین در آینده نیز مهم خواهد بود، زیرا با وجود عدم قطعیت پیرامون اشکال دقیق هوش مصنوعی‌آینده، آنها همچنان به شدت به داده ها متکی خواهند بود.

این‌مقاله با تکیه بر کارهای قبلی OECD و بینش های کلیدی چندرشته ای از گروه متخصصان OECD.AI در زمینه هوش مصنوعی،داده و حریم خصوصی، گزیده ای از سازوکارهای رایج جمع آوری داده برای آموزش هوش مصنوعی را ارائه می دهد. این‌مقاله دو هدف را دنبال می کند. اول، سازوکارهای کلیدی جمع آوری داده مورد استفاده در آموزش هوش مصنوعی را تشریح می کندو ویژگی های اصلی آنها را برجسته می سازد. هدف این است که پیشینه لازم را برای درک بهتر مزایا و چالش های بالقوه هرسازوکار در اختیار سیاست گذاران، تنظیم کنندگان و متخصصان قرار دهد. دوم، این مقاله پایه و اساس تحقیقات آینده در موردپیامدهای حفظ حریم خصوصی و مدیریت داده ها در این سازوکارها، و همچنین ابتکارات سیاستی را که به دنبال ایجاد تعادل‌بین نیازهای حفظ حریم خصوصی و مدیریت داده ها با تقاضای رو به رشد برای داده ها در آموزش هوش مصنوعی هستند، بنامی نهد.

۲ مکانیسم های‌جمع آوری داده ها برای هوش مصنوعی آموزش

دولت هادر سراسر جهان در حال تدوین سیاست های متنوعی برای حمایت و هدایت توسعه سیستم های هوش مصنوعی قابل اعتمادو انسان محور هستند. در ژوئن 2024، اتحادیه اروپا قانون هوش مصنوعی را تصویب کرد (اتحادیه اروپا، 2024)[5] و کمیسیون‌اروپا اخیراً پیش نویس دوم آیین نامه اجرایی هوش مصنوعی همه منظوره (کمیسیون اروپا، ۲۰۲۴) را منتشر کرده است.[6]. مجلس ملی کره همچنین «قانون توسعه هوش مصنوعی و ایجاد اعتماد» را تصویب کرده است (مجلس ملی کره، ۲۰۲۴)[7] که‌قرار است از ژانویه ۲۰۲۶ لازم الاجرا شود. در دسامبر ۲۰۲۴، سنای فدرال برزیل لایحه شماره ۲۳۳۸/۲۰۲۳ را تصویب کرد (مجلس سنای‌برزیل، ۲۰۲۴)[8] که هدف آن تنظیم توسعه و استفاده از هوش مصنوعی نیز هست. در ماه مه 2025، ژاپن قانون ارتقای تحقیق،توسعه و استفاده از فناوری های مرتبط با هوش مصنوعی را تصویب کرد (مجلس ملی ژاپن، 2025)[9] که بر ترویج نوآوری‌هوش مصنوعی در عین مدیریت ریسک ها از طریق هماهنگی دولت و انطباق داوطلبانه صنعت تأکید دارد. یک نکته مشترک‌در این چارچوب های سیاستی، به رسمیت شناختن حاکمیت داده ها به عنوان امری مرتبط برای تضمین توسعه هوش مصنوعی‌قابل اعتماد است. در این زمینه، ابتکارات سیاستی بین المللی مانند «آیین نامه رفتار داوطلبانه گروه ۷» (G7، ۲۰۲۳)[10] براهمیت اجرای اقداماتی توسط توسعه دهندگان هوش مصنوعی تأکید می کنند که (۱) قابلیت ردیابی مجموعه داده ها را فراهم می کند؛(۲) داده های آموزشی با کیفیت بالا را حفظ می کند؛ (۳) از داده های شخصی و حقوق مالکیت معنوی محافظت می کند؛و (۴) با انتشار ارزیابی هایی از تأثیر مدل هوش مصنوعی بر حریم خصوصی و حفاظت از داده های شخصی، شفافیت و پاسخگویی‌را افزایش می دهد. OECD نقش محوری در پیشبرد آیین نامه رفتاری G7 و چارچوب گزارش دهی ایفا کرد که مورد دوم‌به عنوان یک پلتفرم آنلاین برای نظارت بر پذیرش داوطلبانه آیین نامه رفتاری توسط سازمان ها در فوریه ۲۰۲۵ راه اندازی شد (OECD، ۲۰۲۵)[11].

یکی‌از جنبه های کلیدی مدیریت داده های هوش مصنوعی، فرآیند جمع آوری داده ها برای ساخت مجموعه داده های آموزشی هوش‌مصنوعی است. اندازه و پیچیدگی روزافزون مدل های مدرن هوش مصنوعی در مقیاس بزرگ، به ویژه مدل های زبان بزرگ (LLM)،[1]برای آموزش مؤثر به مجموعه داده های گسترده ای نیاز است. از همه مهم تر، حوزه عمومی[2]یا مجموعه داده های تحقیقاتی‌دارای مجوز ممکن است همیشه برای پشتیبانی از مقیاس مورد نیاز برای آموزش اکثر LLM های امروزی کافی نباشند (لی،2023)[12]در نتیجه، اندازه LLM ها چالش هایی در زمینه منبع یابی داده ها ایجاد می کند. به عنوان مثال، نگرانی عمده این است‌که LLM های مدرن در درجه اول بر اساس داده های استخراج شده از وب آموزش می بینند (فدر کوپر، 2023)[13]. دعاوی و تحقیقات‌متعددی در سراسر جهان آغاز شده است که ادعا می کنند این مجموعه داده ها با نقض قوانین مربوط به حق چاپ، حریم‌خصوصی و حفاظت از داده ها جمع آوری شده اند که هنوز حل نشده اند.

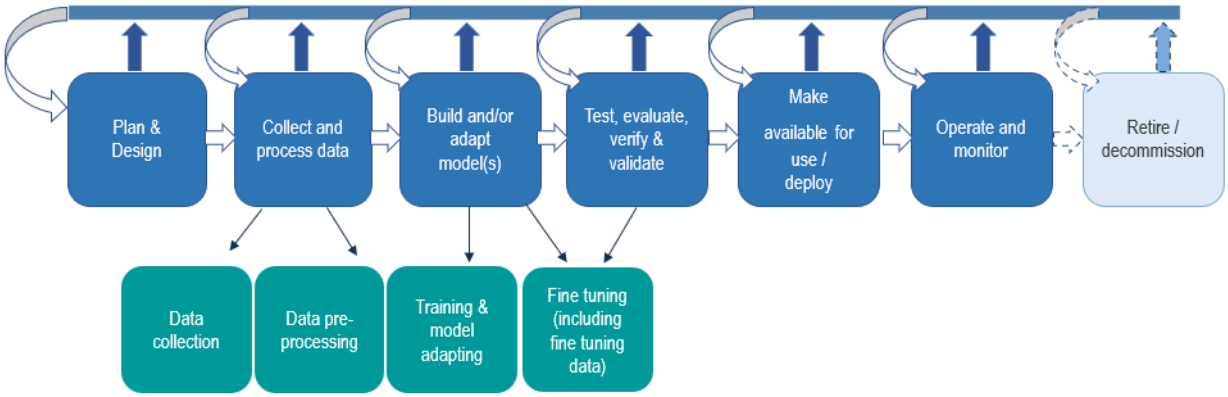
درحالی که توجه سیاست گذاران تا حد زیادی بر شیوه های جمع آوری داده ها برای LLMها متمرکز شده است، امروزه جمع آوری داده هاتنها راه جمع آوری حجم زیادی از داده ها نیست. از آنجایی که هیچ بازار یا مخزن متمرکزی برای به دست آوردن داده های آموزشی‌وجود ندارد، شرکت هایی که مدل های هوش مصنوعی را توسعه می دهند به منابع داده متعددی متکی هستند. مجموعه‌داده های آموزشی دارای مجوز آزاد، یکی دیگر از مکانیسم های محبوب برای به دست آوردن داده های آموزشی هوش مصنوعی‌است. به عنوان مثال، پروژه گوتنبرگ، یک کتابخانه دیجیتال است که بیش از 70،000 کتاب الکترونیکی رایگان ارائه می دهدکه عمدتاً شامل آثار ادبی قدیمی تر است که در حوزه عمومی قرار دارند (Gerlach, 2018)[14]. هر مکانیسم جمع آوری داده هابا سطوح مختلفی از دسترسی، قابلیت همکاری، قابلیت اطمینان و کیفیت همراه است - ویژگی هایی که لزوماً منحصر به‌فرد نیستند. به عنوان مثال، دسترسی بالای مجموعه داده های دارای مجوز آزاد تضمین نمی کند که منشأ داده های زیربنایی به خوبی‌مستند شده باشد. جنبه دیگری از این وضعیت

این‌است که اغلب چندین بازیگر در جمع آوری مجموعه داده های آموزشی هوش مصنوعی مشارکت می کنند. به عنوان مثال، مجموعه‌داده های ImageNet در ابتدا توسط محققان دانشگاه استنفورد ایجاد شد که میلیون ها تصویر متنوع را از سراسر جهان جمع آوری‌کردند (دنگ، ۲۰۰۹)[15]. این محققان همچنین تصاویر را حاشیه نویسی کردند، فرآیندی که می تواند توسط سازمان هایی که‌در برچسب گذاری و حاشیه نویسی داده ها تخصص دارند نیز انجام شود. پس از حاشیه نویسی، مجموعه داده ها توسط سازمان‌غیرانتفاعی)ImageNet, nd(ImageNet گردآوری و در دسترس عموم قرار گرفت.[16]) به عنوان منبعی برای آموزش مدل های‌مختلف هوش مصنوعی در کارهایی مانند طبقه بندی تصویر، تشخیص اشیا و تشخیص چهره عمل می کند. یک توسعه دهنده‌ممکن است از مجموعه داده های ارائه شده توسط ImageNet استفاده کند یا در طول مرحله آموزش، تغییرات بیشتری‌در آن ایجاد کند.

اغلب،دید محدودی به مجموعه داده های مورد استفاده برای ساخت مدل های هوش مصنوعی وجود دارد که ارزیابی کیفیت آنها رادشوار می کند. به عنوان مثال، شاخص شفافیت مدل بنیاد ۲۰۲۴، ۲۳ زیردامنه را مشخص می کند.۳که جنبه های مختلف توسعه‌و استقرار مدل های پایه، از جمله دسترسی به داده ها را ارزیابی می کنند (بوماسانی، 2024)[17]). از بین تمام توسعه دهندگان‌مدل پایه که در این شاخص ارزیابی شده اند، تنها یکی از آنها داده های آموزشی خود را آشکارا به اشتراک گذاشته است(بوماسانی، 2024)[17]). نکته قابل توجه این است که بین انتشار اولین گزارش در اکتبر 2023 و پیگیری آن در مه 2024، توسعه دهندگان‌امتیاز شفافیت خود را در تمام زیردامنه ها به جز افشای داده ها بهبود بخشیدند (بوماسانی، 2024)[17]). پیشرفت‌کند در نمرات شفافیت پیرامون داده ها را می توان تا حدودی با خطرات قانونی که کسب وکارها هنگام افشای مجموعه داده های‌مورد استفاده برای ساخت مدل های هوش مصنوعی متحمل می شوند، و همچنین نگرانی ها در مورد اطلاعات اختصاصی‌توضیح داد، زیرا کیفیت یک مدل اغلب با ارزش داده هایی که بر اساس آن آموزش داده می شود، ارتباط نزدیکی دارد. فراتراز پیچیدگی ها و عدم قطعیت های قانونی مربوط به مقررات حفظ حریم خصوصی و حقوق مالکیت معنوی، نگرانی هایی نیز درمورد موانع ورود به بازار برای جمع آوری داده ها وجود دارد، زیرا برخی از شرکت ها ممکن است دسترسی ممتازی به مجموعه داده های‌به دست آمده از سایر فعالیت ها در بازارهای دیجیتال داشته باشند (Competition & Markets Authority، 2023) [18]). همچنین بحث های فزاینده ای در مورد محدودیت های موجودی داده های عمومی که از طریق اینترنت به عنوان منبعی برای‌توسعه هوش مصنوعی قابل دسترسی هستند، وجود دارد. مطالعات نشان می دهد که اگر روند فعلی گسترش اندازه مجموعه‌داده های آموزشی برای LLMها به طور خاص ادامه یابد، مدل های هوش مصنوعی ممکن است بین سال های 2026 تا 2032عرضه داده های متنی عمومی تولید شده توسط انسان را به پایان برسانند (پابلو ویلالوبوس، 2022[19]). این کمبود بالقوه داده ها،سوالاتی را در مورد قابلیت اتکای بلندمدت عمدتاً به داده های عمومی موجود از اینترنت برای آموزش هوش مصنوعی مطرح‌می کند و اهمیت در نظر گرفتن طیف گسترده ای از مکانیسم های جمع آوری داده ها را افزایش می دهد. با این حال، روند روبه رشدی در صنعت به سمت توسعه مدل های بنیادی کوچک تر و کارآمدتر وجود دارد که به داده های کمتر یا مجموعه داده های‌دقیق تر متکی هستند. در نتیجه، اگرچه پتانسیل «تمام شدن داده ها» یک نگرانی قابل توجه است، اما ممکن است یک مانع‌غیرقابل مدیریت نباشد.

تصمیمات‌مربوط به داده های مورد استفاده برای ساخت مجموعه داده های آموزشی در مرحله اولیه توسعه مدل گرفته می شود. توسعه‌مدل های یادگیری ماشین هوش مصنوعی معمولاً شامل چهار مرحله است: جمع آوری داده ها، پیش پردازش، آموزش و تنظیم‌دقیق مدل (OECD، 2024)[20]). در طول فرآیند جمع آوری داده ها، توسعه دهندگان هوش مصنوعی در مورد کمیت و نوع داده هایی‌که باید در مجموعه داده های آموزشی گنجانده شوند و همچنین محل جمع آوری داده ها تصمیم می گیرند. انتخاب های انجام‌شده در طول توسعه مجموعه داده ها بر خروجی مدل هوش مصنوعی تأثیر خواهد گذاشت. در مرحله بعد، مرحله پیش پردازش‌داده ها، هر نقطه داده را به یک قالب قابل استفاده برای مدل یادگیری ماشین تبدیل می کند. این نقاط داده با هم ترکیب‌شده و در یک مجموعه داده برای آموزش جمع آوری می شوند. مرحله آموزش (یا پیش آموزش) مرحله اولیه ای است که درآن یک «مدل پایه» با استفاده از یک مجموعه داده گسترده ساخته می شود. این مدل پایه به عنوان پایه ای عمل می کند که تنظیمات‌و بهینه سازی های بیشتر در طول مرحله تنظیم دقیق، که معمولاً نیاز به استفاده از مجموعه داده های خاص دامنه دارد،بر روی آن انجام می شود.۴

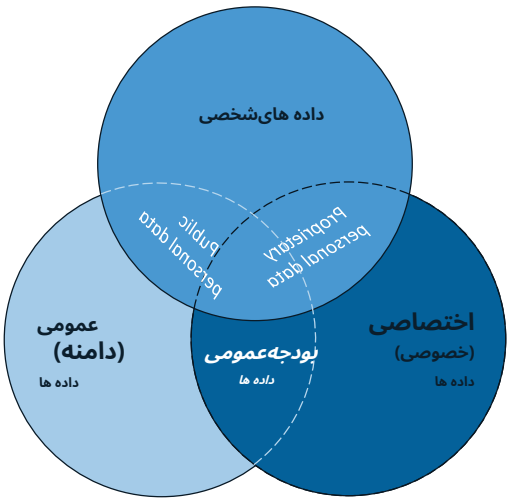
شکل۲.۱. چرخه حیات توسعه مدل هوش مصنوعی



منبع:تصویر OECD بر اساس چرخه عمر سیستم هوش مصنوعی (OECD، 2024)[20] (سازمان همکاری و توسعه اقتصادی، ۲۰۲۵)[21]

هردو مجموعه داده های آموزشی و تنظیم دقیق ممکن است شامل داده هایی باشند که دامنه های همپوشانی را در بر می گیرند. (OECD، ۲۰۱۹)[4] بین سه حوزه داده زیر تمایز قائل می شود: i)*دامنه شخصی*که شامل تمام داده های شخصی «مربوط به یک فردمشخص یا قابل شناسایی» است که صاحبان داده ها در مورد آنها منافع حریم خصوصی دارند؛ ii)*دامنه اختصاصی*که شامل تمامداده های اختصاصی می شود که معمولاً منافع اقتصادی برای حذف دیگران وجود دارد؛ و iii)*دامنه عمومی*که شامل تمام داده هایی‌است که دسترسی و استفاده مجدد از آنها رایگان است.۵این سه حوزه متقابلاً منحصر به فرد نیستند و اغلب همانطور که‌در شکل 2.2 نشان داده شده است، همپوشانی دارند که منعکس کننده منافع مختلف ذینفعان و چارچوب های حاکمیت داده قابل‌اجرا است (OECD، 2019)[22]۲۰۲۲[23]). برای مثال، داده های عمومی می توانند شامل داده هایی باشند که توسط حقوق مالکیت‌فکری یا هرگونه حقوق دیگری با اثرات مشابه محافظت نمی شوند، و همچنین مجموعه داده های اختصاصی با بودجه عمومی‌که به طور مجاز مجوز گرفته اند.

شکل۲.۲. دامنه های شخصی، اختصاصی و عمومی داده ها



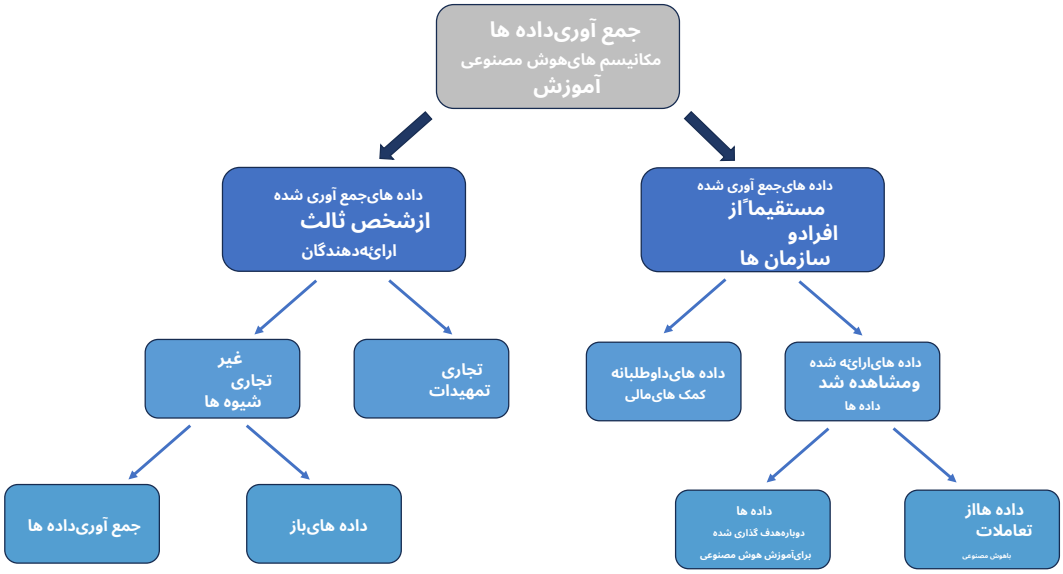
منبع:(OECD، ۲۰۱۹)[4] افزایش دسترسی و اشتراک گذاری داده ها: تطبیق خطرات و مزایای استفاده مجدد از داده ها در جوامع، انتشارات OECD، پاریس،https://doi.org/10.1787/b4d546a9-en

مرحله‌آموزش بر اندازه و تنوع مجموعه داده های آموزشی تمرکز دارد زیرا مدل های یادگیری ماشین با قابلیت های عمومی به داده های‌بزرگ و متنوع نیاز دارند. مجموعه داده های در مقیاس بزرگ، مدل هایی را با

پایه‌محکمی از دانش عمومی، که به آنها اجازه می دهد طیف متنوعی از خروجی ها را تولید کنند و بسیار مقیاس پذیر باشند. مدل های‌از پیش آموزش دیده را می توان به صورت اختیاری با استفاده از مجموعه داده های تخصصی برای به دست آوردن قابلیت های‌خاص، تنظیم دقیق کرد. در طول مرحله تنظیم دقیق، حجم داده های مورد نیاز معمولاً کمتر از مرحله آموزش است، زیراهدف آن تطبیق یک مدل از پیش آموزش دیده با وظایف و دامنه های خاص است.

بخش های‌بعدی بر مکانیسم های جمع آوری داده مورد استفاده برای هر دو مرحله آموزش و تنظیم دقیق توسعه مدل یادگیری ماشین‌تمرکز دارند. آن ها مکانیسم های جمع آوری داده مرتبط برای آموزش هوش مصنوعی را با تمرکز ویژه بر ویژگی های هر روش،تشریح می کنند. این طبقه بندی ابتدا بر اساس منابع، موارد زیر را متمایز می کند:؛توسعه دهندگان هوش مصنوعی از طریق‌کدام منابع داده، داده ها را دریافت می کنند: ۱) مستقیماً از افراد و سازمان ها؛ و ۲) از ارائه دهندگان شخص ثالث. در این دومنبع داده اصلی، مکانیسم های کلیدی و مختلف جمع آوری داده ها را شناسایی می کنیم که در شکل زیر مشخص شده اند و دربخش های بعدی با جزئیات بیشتری مورد بحث قرار می گیرند:

شکل۲.۳. مکانیسم های کلیدی جمع آوری داده ها برای آموزش هوش مصنوعی



توجه:این شکل، مروری بر سازوکارهای کلیدی جمع آوری داده ها برای آموزش هوش مصنوعی که در سند به آنها اشاره شده است، ارائه می دهد. این شکل لزوماً تمام روش های مختلف جمع آوری‌داده ها برای آموزش هوش مصنوعی را در بر نمی گیرد - اکوسیستم هوش مصنوعی پویا است و احتمالاً با ورود بازیگران و مدل های جدید کسب وکار به حوزه هوش مصنوعی، سازوکارهای‌جدید به تکامل خود ادامه خواهند داد.

لازم‌به ذکر است که PETها نقش مهمی در فراهم کردن امکان جمع آوری داده ها برای توسعه هوش مصنوعی، کاهش خطرات مربوط‌به مدیریت داده ها و حریم خصوصی ایفا می کنند. PETها با افزایش محرمانگی جمع آوری و استفاده از داده ها، از عملکرد مؤثرسازوکارهای ذکر شده در این طبقه بندی پشتیبانی می کنند. به عنوان مثال، داده های مصنوعی به طور فزاینده ای به عنوان وسیله ای‌برای جمع آوری محرمانه داده ها در نظر گرفته می شوند (OECD، 2025)[24].این طبقه بندی به طور خاص بر مکانیسم های‌جمع آوری داده های دنیای واقعی تمرکز دارد که اغلب مبنای تولید مجموعه داده های مصنوعی را تشکیل می دهند.

۳ داده های‌جمع آوری شده به طور مستقیم‌از افراد و سازمان ها

۳.۱داده های ارائه شده و داده های مشاهده شده

*داده های‌ارائه شده*به اطلاعاتی اشاره دارد که از اقدامات مستقیم انجام شده توسط یک فرد ناشی می شود، به طوری که فرد کاملاً از اقداماتی که منجر به تولید داده ها می شود، آگاه است (OECD، ۲۰۱۴)[3]این داده ها معمولاً از طریق اقدامات مستقیم کاربرمانند پر کردن فرم ها، ثبت نام در خدمات، ارائه بازخورد یا آپلود محتوا به صورت آنلاین جمع آوری می شوند. این داده ها ممکن‌است شامل انواع مختلفی از داده ها، از اطلاعات شخصی اولیه گرفته تا ورودی های پیچیده تر مانند تنظیمات برگزیده، رتبه بندی هایا محتوای تولید شده توسط کاربر (مثلاً عکس، متن و ویدیو) باشد. در حالی که افراد مربوطه ممکن است از پیامدهای‌ارائه این داده ها بی اطلاع باشند، این واقعیت که این داده ها ایجاد می شوند باید واضح - یا حداقل شهودی - باشد (OECD، ۲۰۱۴).[3].

*داده های‌مشاهده شده*داده هایی هستند که توسط دیگران مشاهده شده و در قالب دیجیتال ثبت شده اند. این داده ها می توانندیا در لحظه ایجادشان ثبت شوند یا پس از مشاهده به یک حامل دیجیتال منتقل شوند (OECD، ۲۰۱۴)[3]. این نوع جمع آوری‌داده ها معمولاً از طریق ردیابی دیجیتال (مثلاً کوکی های آنلاین، برنامه های ردیابی موقعیت مکانی) یا نظارت فیزیکی (مثلاً دوربین های مداربسته) انجام می شود. در حالی که ممکن است افراد از ایجاد داده های مشاهده شده آگاه شوند (مثلاً به دلیل‌تعامل فعال)، بخش زیادی از ایجاد داده های مشاهده شده ممکن است مورد توجه قرار نگیرد.

درزمینه سیستم های هوش مصنوعی، تمایز بین داده های ارائه شده و مشاهده شده اغلب مبهم است. هنگامی که افراد با سیستم های‌هوش مصنوعی تعامل می کنند، معمولاً هر دو نوع داده به طور همزمان جمع آوری و پردازش می شوند. به عنوان مثال،هنگامی که کاربر در یک چت بات (داده های ارائه شده) درخواستی را وارد می کند، سیستم ممکن است در همان جلسه، فراداده هایی‌مانند مهرهای زمانی، مدت زمان تعامل، اطلاعات دستگاه یا الگوهای تعامل کاربر (داده های مشاهده شده) را نیز ثبت‌کند. این همپوشانی عملی به این معنی است که در برنامه های دنیای واقعی، ترسیم مرز مشخصی بین این دو دسته می تواندشوار باشد. بنابراین، اغلب تجزیه و تحلیل دقیق تر در سطح زیرگروه ضروری است.

براین اساس، در بخش بعدی بین دو سناریو تمایز قائل می شویم: اول، داده هایی که توسط افراد در طول تعامل مستقیم با خدمات‌یا محصولات هوش مصنوعی ارائه یا مشاهده می شوند؛ و دوم، داده هایی که در ابتدا در زمینه سایر فعالیت های دیجیتال‌جمع آوری شده اند اما متعاقباً برای آموزش هوش مصنوعی استفاده می شوند.

۳.۱.۱داده های ارائه شده توسط یا مشاهده شده از افراد هنگام تعامل مستقیم با سیستم های هوش مصنوعی

یک‌مکانیسم فزاینده مرتبط برای آموزش سیستم های هوش مصنوعی، استفاده از اطلاعات تولید شده از پیام ها در ابزارهای هوش‌مصنوعی کسب و کار به مصرف کننده (B2C) مانند چت بات ها، دستیاران مجازی و سیستم های خدمات مشتری خودکار است.این چت بات های هوش مصنوعی به طور فزاینده ای "داده های تعاملی" تولید می کنند - محتوایی که هنگام تعامل یک سیستم‌هوش مصنوعی با یک کاربر تولید می شود. با ادغام بیشتر سیستم های هوش مصنوعی در محیط های مکالمه، حجم و اهمیت‌این داده های تعاملی همچنان رو به افزایش است. داده های تعاملی همچنین می توانند از طریق ادغام های کسب و کار به‌کسب و کار (B2B) با سرویس های میزبانی شده، که در آن قابلیت های هوش مصنوعی

ازطریق مشارکت مستقیم یا دسترسی به API در محصولات تعبیه شده اند. به عنوان مثال، ChatGPT از طریق یک مشارکت استراتژیک‌بین مایکروسافت و OpenAI در جستجوی مایکروسافت بینگ ادغام شده است.

دسترسی‌به داده های تعاملی محدود به شرکت هایی که تحت این مدل های کسب وکار B2C و B2B فعالیت می کنند، نیست. این نوع داده ها،که به ویژه برای تنظیم دقیق مدل های هوش مصنوعی مفید هستند، از طریق بازارهایی مانند promptbase.com در دسترس توسعه دهندگان‌هوش مصنوعی قرار دارند.

علاوه‌بر داده های تعاملی، شرکت هایی که خدمات و محصولات هوش مصنوعی ارائه می دهند، به داده های ارائه شده توسط کاربریا قابل مشاهده، مانند اطلاعات سطح حساب یا فراداده نیز دسترسی دارند. این شرکت ها ممکن است استفاده از این نوع اطلاعات‌را برای آموزش مدل های هوش مصنوعی خود مفید بیابند.

۳.۱.۲داده های (ارائه شده و/یا مشاهده شده) که برای آموزش هوش مصنوعی مورد استفاده مجدد قرار گرفته اند

علاوه‌بر داده های جمع آوری شده در طول استفاده از ابزارهای هوش مصنوعی، برخی از توسعه دهندگان هوش مصنوعی ممکن است به داده هایی (چه توسط‌کاربران ارائه شده باشد و چه از تعاملات آنها مشاهده شده باشد) از سایر سرویس های دیجیتال خود نیز دسترسی داشته باشند و ممکن است به دنبال‌استفاده از این داده ها برای آموزش مدل های هوش مصنوعی باشند. این مورد در مورد صاحبان پلتفرم های بزرگی که میزبان محتوای تولید شده توسط‌کاربر هستند، صدق می کند (FTC، 2024).[25]این پلتفرم ها انگیزه های تجاری بسیار قدرتمندی برای استفاده از محتوایی مانند پست های رسانه های‌اجتماعی، نظرات، بررسی ها، سابقه خرید یا بازخورد کاربر برای بهبود مدل های هوش مصنوعی خود دارند.

بااعمال نفوذ*داده های ارائه شده توسط کاربر*توسعه دهندگان هوش مصنوعی می توانند سیستم ها و فرآیندهایی را طراحی کنند که‌احتمالاً به نیازها و ترجیحات دنیای واقعی پاسخ می دهند. این داده ها به بهبود عملکرد برنامه های هوش مصنوعی در صنایعی مانندمراقبت های بهداشتی، آموزش، امور مالی یا سفر، که در آن ها شخصی سازی و پاسخگویی به نیازهای فردی مرتبط است، کمک‌می کند. به عنوان مثال، ویژگی های مبتنی بر هوش مصنوعی که توسط Booking.com برای شخصی سازی برنامه ریزی سفرارائه می شوند، با استفاده از داده های ناشناس از رزرو هتل و جزئیات رزرو آموزش داده می شوند (گلدنبرگ، 2021)[26].

*داده های‌مشاهده شده*داده های جمع آوری شده، به ویژه هنگامی که از طریق فناوری های ردیابی آنلاین جمع آوری می شوند، نقش مهمی‌در آموزش مدل های هوش مصنوعی طراحی شده برای تجربیات شخصی سازی شده ایفا می کنند. سیستم های هوش مصنوعی ازاین داده ها یاد می گیرند تا ترجیحات فردی را درک کرده و رفتار آینده را پیش بینی کنند. به عنوان مثال، پلتفرم های تجارت الکترونیک مانندآمازون، داده های خرید مشتری - مانند ترجیحات، جستجوها و رفتار مرور - را به LLM های خود ارائه می دهند تا توصیه های محصول‌شخصی سازی شده تری ارائه دهند (لوین، 2024)[27].

دربرخی موارد، این اطلاعات ممکن است در ابتدا برای سایر خدمات دیجیتال جمع آوری شده باشد، نه آموزش مدل هوش مصنوعی.اینکه آیا چنین استفاده ثانویه ای مجاز است یا خیر، به قوانین حفظ حریم خصوصی و حفاظت از داده ها در حوزه قضایی‌مربوطه بستگی دارد و معمولاً نیاز به ارزیابی مبنای قانونی مربوطه برای پردازش دارد. به عنوان مثال، برزیل*Autoridade Nacional de Proteção de Dados Pessoais*(در تاریخ 2 ژوئیه 2024 یک اقدام پیشگیرانه صادر کرد و از متا خواست که‌فوراً استفاده از داده های شخصی برزیلی ها را از پلتفرم رسانه اجتماعی خود برای آموزش مدل هوش مصنوعی مولد خود متوقف‌کند و نگرانی هایی را در مورد مبنای قانونی چنین پردازشی تحت قانون حفاظت از داده های برزیل مطرح کرد (بادیلو، 2024))ANPD[28]). مقاله آتی OECD تحلیل بیشتری در مورد این ملاحظات ارائه خواهد داد.

۳.۲اهدای داوطلبانه داده ها

داده های‌شخصی یا غیرشخصی ممکن است به صورت داوطلبانه توسط صاحبان داده ها یا دارندگان داده ها برای اهداف آموزشی‌هوش مصنوعی، بدون دریافت غرامت، و به نفع جامعه ارائه شوند. این عمل گاهی اوقات به عنوان «اهدای داده» یا «نوع دوستی‌داده» شناخته می شود، همانطور که در قوانینی مانند قانون حاکمیت داده اتحادیه اروپا (DGA) (اتحادیه اروپا، 2022) آمده‌است.[29]).۷این شامل افراد (یا سازمان ها) می شود که برای به اشتراک گذاشتن داده های خود با محققان، سازمان های غیرانتفاعی‌و سایر نهادها برای اهدافی مانند بهبود مراقبت های بهداشتی یا افزایش خدمات عمومی، رضایت می دهند (Kirstein, 2023)[30].

اگرچه‌این رویه ها ممکن است سوالاتی را ایجاد کنند،۸اهدای داده ها همچنین می تواند با فراهم کردن دسترسی به داده های متنوع و واقعی‌که در غیر این صورت دشوار یا پرهزینه خواهد بود، نقش مهمی در آموزش سیستم های هوش مصنوعی ایفا کند.

به‌عنوان مثال، در زمینه مراقبت های بهداشتی، نظارت بر محیط زیست و آموزش، اهدای داوطلبانه داده ها می تواند به ایجاد مجموعه‌داده های غنی تر و جامع تر کمک کند (Hirsch MC، 2020)[31]. این با اصول هوش مصنوعی OECD همسو است که دولت هارا تشویق می کند تا اعتماد داده ها و سایر سازوکارها را برای پشتیبانی از اشتراک گذاری ایمن، منصفانه، قانونی و اخلاقی‌داده ها ترویج دهند. همچنین با توصیه OECD در مورد افزایش دسترسی و اشتراک گذاری داده ها (OECD، 2021 همسواست.[32]) که از انواع مختلف همکاری های داده ای و بهره برداری از منابع داده ای جدید و موجود برای تقویت اکتشافات و نوآوری های‌علمی مبتنی بر داده در بخش های خصوصی و دولتی پشتیبانی می کند.

درزمینه هایی مانند مراقبت های بهداشتی و ژنومیک، دسترسی به مجموعه داده های بزرگ برای توسعه راه حل های مبتنی بر هوش‌مصنوعی بسیار مهم است. در این زمینه، اهدای داده ها می تواند برای کمک به افراد جهت ارائه داده های شخصی خود (مثلاً اطلاعات ژنتیکی یا داده های سلامت) به طرح های تحقیقاتی با منافع عمومی، تسریع اکتشافات در توسعه دارو، پزشکی شخصی سازی شده‌و اپیدمیولوژی مورد استفاده قرار گیرد. به عنوان مثال، می توان به UK Biobank، یک پایگاه داده و منبع تحقیقاتی‌زیست پزشکی در مقیاس بزرگ حاوی اطلاعات ژنتیکی، سبک زندگی و سلامت و نمونه های بیولوژیکی از نیم میلیون شرکت کننده‌در بریتانیا (UK Biobank, nd) اشاره کرد.[33]جالب توجه است که مطالعه ای که تمایل به اشتراک گذاری داده های سلامت‌دیجیتال برای تحقیقات در آلمان و اسرائیل را اندازه گیری می کرد، نشان داد که اکثریت قابل توجهی از شهروندان در هر دوکشور (۸۲٪ از آلمانی ها و ۸۱٪ از اسرائیلی ها) نگرش مثبتی نسبت به ایجاد یک پایگاه داده متمرکز برای تحقیقات پزشکی ابراز کرده اند،البته تا زمانی که داده ها ناشناس باشند (Weisband، ۲۰۲۳)[34]) (سازمان همکاری و توسعه اقتصادی، ۲۰۲۵)[35]در این زمینه،شایان ذکر است که قانون استفاده و دسترسی به داده ها (DUAA) که در 19 ژوئن 2025 به تصویب دارایی سلطنتی رسید،مورد توجه قرار گیرد. DUAA توسط دولت بریتانیا برای ارتقای استراتژی دیجیتال بریتانیا و آزادسازی ارزش داده ها، بهره برداری‌از قابلیت های آن برای تقویت خدمات عمومی و کمک به اقتصاد بریتانیا در نظر گرفته شده است. در میان مفاد آن، DUAA تصریح می کند که افراد می توانند «رضایت گسترده» خود را به یک حوزه تحقیقات علمی بدهند (ICO، 2025).[36].

به‌همین ترتیب، دولت ها و مؤسسات عمومی اغلب فاقد داده های لازم برای بهبود خدمات عمومی، توسعه شهرهای هوشمندتریا مقابله با چالش های اجتماعی هستند. استدلال می شود که نوع دوستی داده ها می تواند با ارائه داده های لازم برای‌آموزش مدل های هوش مصنوعی که مسائلی مانند زیرساخت های عمومی، حمل و نقل یا آموزش را بهبود می بخشند، به حمایت‌از این تلاش ها کمک کند. به عنوان مثال، شهروندان داوطلبانه داده های مربوط به مصرف انرژی یا عادات حمل و نقل خودرا به اشتراک می گذارند، می توانند به آموزش سیستم های هوش مصنوعی مورد استفاده برای بهینه سازی برنامه ریزی شهری،بهبود مدیریت ترافیک یا کاهش اتلاف انرژی در شهرهای هوشمند کمک کنند. یک مورد قابل توجه از نوع دوستی داده ها، مربوطبه پروژه DECODE است. در این ابتکار، شهروندان بارسلونا با استفاده از حسگرهای محیطی که در داخل و خارج از خانه هایشان‌قرار داده شده بودند، داده هایی در مورد سر و صدا، آلودگی هوا، دما و رطوبت جمع آوری کردند. این داده های ارزشمندمتعاقباً از طریق آنچه به عنوان "داده های مشترک" شناخته می شود، در دسترس عموم قرار گرفت (ساگارا، ۲۰۱۹).[37]).

۴ داده های‌جمع آوری شده از ارائه دهندگان شخص‌ثالث

وقتی‌داده ها مستقیماً از شخص مربوط به داده ها تهیه نمی شوند، توسعه دهندگان هوش مصنوعی که به دنبال داده ها هستند، معمولاً آن ها را از طریق شیوه های تجاری و غیرتجاری به دست می آورند. این بخش این دو مکانیسم را به عنوان وسیله ای برای تأمین داده هابرای اقتصاد هوش مصنوعی بررسی می کند.

۴.۱داده های جمع آوری شده از اشخاص ثالث بر اساس توافقات تجاری

صدورمجوز داده های تجاری یکی از راه هایی است که توسعه دهندگان هوش مصنوعی به داده ها دسترسی پیدا می کنند. به عنوان‌مثال، OpenAI قراردادهای صدور مجوز را با رسانه های خبری، از جمله آسوشیتدپرس (O'Brien، 2023) امضا کرده است. [38]) و اکسل اشپرینگر (OpenAI، ۲۰۲۳)[39]) و همچنین با پلتفرم های رسانه های اجتماعی مانند ردیت (OpenAI، 2024)[40] بسیاری‌از توسعه دهندگان هوش مصنوعی، از جمله متا، قراردادهای مجوز داده را نیز با پلتفرم توزیع تصویر، شاتر استوک، امضا کرده اند(شاتر استوک، ۲۰۲۳)[41]).

توسعه دهندگان‌هوش مصنوعی همچنین ممکن است برای دسترسی به مجموعه داده های شخص ثالث به واسطه هایی مانند بازارهای‌داده متکی باشند. این بازارهای داده، تبادل داده ها بین تأمین کنندگان و توسعه دهندگان هوش مصنوعی را تسهیل می کنند.برای روشن شدن موضوع، بازارهای داده از این نوع شامل پلتفرم هایی مانندتعریف شده.ai، appen.com، یا تبادل دادهAWS.

قراردادهای‌تجاری داده برای پشتیبانی از توسعه هوش مصنوعی همچنین می توانند شامل مدل های تجاری باشند که از دسترسی‌و اشتراک گذاری داده ها، مانند دلالان داده، درآمد کسب می کنند. دلالان داده، داده ها را از منابع مختلف جمع آوری و تجمیع‌می کنند و این داده ها را به اشخاص ثالث می فروشند. یکی از عوامل مهم تمایز بین دلالان داده و ارائه دهندگان بازار داده‌این است که دلالان داده به طور فعال در جمع آوری داده های اضافی و تجمیع آنها مشارکت دارند، در حالی که ارائه دهندگان‌بازار داده واسطه های غیرفعالی هستند که از طریق آنها کنترل کنندگان داده، از جمله دلالان، می توانند مجموعه داده های‌خود را ارائه دهند (OECD، 2019)[4]). دلالان داده مدت هاست که بخشی از اقتصاد داده محور بوده اند و در برخی حوزه های‌قضایی، شیوه های آنها توسط سازمان های نظارتی مورد تجزیه و تحلیل قرار گرفته است. یکی از فعالیت های فعلی آنهاتهیه مجموعه داده های آماده هوش مصنوعی برای آموزش مدل هوش مصنوعی است (PR Newswire، 2024)[42]).

مجموعه‌داده های توسعه هوش مصنوعی همچنین ممکن است توسط کسب وکارها به عنوان بخشی از فعالیت های تجاری ثانویه آنها به‌صورت تجاری ارائه شود. به عنوان مثال، یک تولیدکننده تجهیزات کشاورزی ممکن است از داده های حسگر جمع آوری شده از تراکتورهااستفاده نکند، اما می تواند آن را به خریداران علاقه مند برای توسعه دهندگان هوش مصنوعی که روی راه حل های کشاورزی دقیق‌کار می کنند، بفروشد. اخیراً، بایر با مایکروسافت برای توسعه یک مدل هوش مصنوعی که با داده های بایر تنظیم شده است، همکاری‌امضا کرده است. این مدل برای ارائه بینش در مورد زراعت و حفاظت از محصولات در نظر گرفته شده است و برای توزیع کنندگان‌بایر، استارتاپ های فناوری کشاورزی و احتمالاً رقبا قابل مجوز خواهد بود. از جمله موارد دیگر، مدل هوش مصنوعی قادر به‌پاسخگویی به سؤالاتی در مورد موضوعاتی مانند اجزای حشره کش و مناسب بودن محصول برای پنبه خواهد بود (Bousquette ، 2024)[43]).

۴.۲داده های جمع آوری شده از اشخاص ثالث بر اساس رویه های غیرتجاری

۴.۲.۱ترتیبات داده باز

داده های‌باز؟به سرعت به یک رویکرد برجسته برای افزایش دسترسی به داده ها تبدیل شده است (OECD، 2019[4]). همچنین، دردسترس قرار دادن داده ها از بخش های دولتی و خصوصی تحت مجوزهای داده باز، برای رشد اقتصاد هوش مصنوعی به طور فزاینده ای‌اهمیت پیدا می کند. در بخش دولتی، نمونه هایی از آن شامل دستورالعمل داده باز اتحادیه اروپا (اتحادیه اروپا، ۲۰۱۹) است.[44]) یا ابتکار ملت هوشمند سنگاپور (Smart Nation Singapore, nd[45])دولت کره همچنین نقش فعالی در ایجاد و انتشارمجموعه داده های آموزشی هوش مصنوعی برای تقویت نوآوری داخلی داشته است، با هدف ایجاد پایه ای قوی برای توسعه‌مدل های هوش مصنوعی در بخش خصوصی. به عنوان مثال، این کشور پلتفرم AI Hub را راه اندازی کرده است که دسترسی‌آزاد به طیف گسترده ای از مجموعه داده ها را فراهم می کند (AI Hub، 2025[46]). یک نمونه قابل توجه از بخش خصوصی،مجموعه داده COCO (اشیاء مشترک در متن) مایکروسافت است که شامل بیش از ۳۳۰۰۰۰ تصویر با حاشیه نویسی های‌دقیق برای ۸۰ دسته شیء است که تحت مجوز CC-BY 4.0 برای حاشیه نویسی ها قابل دسترسی است (اشیاءمشترک در متن، ۲۰۲۵[47]).

پیشرفت هادر توسعه هوش مصنوعی همچنین توسط ناشران مجموعه داده تسهیل می شود که داده ها را از منابع مختلف گردآوری‌و سازماندهی می کنند و دسترسی رایگان و آزاد به آنها را فراهم می کنند. در این راستا، ImageNet، یک مجموعه داده با مجوزآزاد از تصاویر در مقیاس بزرگ (J. Deng، ۲۰۰۹[48]) برای آموزش بسیاری از مدل های بینایی کامپیوتر اساسی بوده است. به همین‌ترتیب، پیشرفت های اخیر در تاخوردگی پروتئین که توسط سیستم یادگیری عمیق AlphaFold حاصل شده است، تنها به‌لطف دسترسی آزاد به بانک داده های پروتئین، که تقریباً نیم قرن پیش تأسیس شد، امکان پذیر شد (لیزنفلد، 2023[49]). مخازن‌مبتنی بر متن مانند (Common Crawl, nd[50]Common Crawl) همچنین مقادیر عظیمی از داده های وب را برای آموزش‌مدل های پردازش زبان طبیعی (NLP) ارائه می دهند. همچنین مخازنی از مجموعه داده های تنظیم دقیق وجود دارد که به‌صورت آزاد در دسترس هستند، مانند کتابخانه مجموعه داده های (Hugging Face، nd[51]Hugging Face).

اگرچه‌داده های باز (بدون محدودیت) ممکن است مطلوب باشند، شرایط حاکم بر اشتراک گذاری داده ها منجر به درجات مختلفی از باز بودن‌می شود، همانطور که در شکل ۴.۱ نشان داده شده و در کادر ۴.۱ در زیر توضیح داده شده است.

شکل۴.۱. درجات باز بودن داده ها

منبع:بر اساس (OECD 2015[52])، *نوآوری مبتنی بر داده: کلان داده برای رشد و رفاه*، https://doi.org/10.1787/9789264229358-en.

کادر۱.۴. درجات باز بودن داده ها

بازبودن داده ها یک «مفهوم دوتایی» نیست. بلکه پیوستاری از درجات مختلف باز بودن است که از دسترسی بسته و استفاده‌فقط توسط دارنده داده (سطح ۰ در شکل ۳) تا موارد زیر را شامل می شود:

- سطح۱: ترتیبات دسترسی و اشتراک گذاری مشروط که در آن دسترسی و اشتراک گذاری داده ها تابع «شرایطی است‌که شامل محدودیت هایی برای کاربران مجاز به دسترسی به داده ها (ترتیبات تبعیض آمیز)، شرایط استفاده ازداده ها از جمله اهدافی که داده ها می توانند برای آنها استفاده شوند و الزامات مربوط به مکانیسم های کنترل دسترسی‌به داده ها که از طریق آنها دسترسی به داده ها اعطا می شود» می شود. به عنوان مثال، توافق نامه های دوجانبه‌صدور مجوز داده ها.
- سطح۲: ترتیبات دسترسی و اشتراک گذاری بدون تبعیض داده ها، که در آن می توان به داده ها دسترسی پیدا کرد وبا پرداخت هزینه به اشتراک گذاشت، اما «بر اساس شرایطی که مستقل از هویت کاربران داده ها هستند». به عنوان‌مثال، محتوای دارای حق اشتراک، مانند مقالاتی که در وب سایت های خبری با حق اشتراک منتشر می شوندیا مجموعه داده هایی که از طریق بازارهای داده آنلاین فروخته می شوند، معمولاً در این دسته قرار می گیرند.هر کسی که مایل به پرداخت هزینه باشد، می تواند صرف نظر از اینکه چه کسی است، به داده ها دسترسی‌پیدا کند. ابتکاراتی مانند فضای داده های سلامت اروپا و مراکز ملی داده های سلامت، به عنوان مثال دیگر،به نهادهای دولتی اجازه می دهند تا دسترسی به داده ها را تحت شرایط تعریف شده و در محیط های پردازش امن‌فراهم کنند. این ترتیبات همچنین ممکن است شامل هزینه هایی برای پوشش هزینه های عملیاتی باشد.
- سطح۳: داده های باز (ترتیبات) به عنوان شکل افراطی باز بودن داده ها، که عبارتند از «ترتیبات دسترسی و اشتراک گذاری‌بدون تبعیض داده ها، که در آن داده ها قابل خواندن توسط ماشین هستند و می توانند به صورت رایگان‌به آنها دسترسی پیدا کرده و به اشتراک گذاشته شوند و توسط هر کسی برای هر هدفی مورد استفاده قرار گیرند،حداکثر با رعایت الزاماتی که یکپارچگی، منشأ، انتساب و باز بودن را حفظ می کنند». (OECD، ۲۰۲۵[53]) برای مثال،مجموعه داده های علمی منتشر شده تحت مجوزهای Creative Commons که به صورت رایگان در دسترس‌همه قرار دارند.

منبع:بر اساس (OECD، 2025[53]) ، افزایش دسترسی و اشتراک گذاری داده ها در عصر هوش مصنوعی، و (OECD، ۲۰۱۵[52])نوآوری مبتنی برداده: کلان داده برای رشد و رفاه https://doi.org/10.1787/9789264229358-en

میزان‌باز بودن داده ها به عوامل مختلفی بستگی دارد: محدودیت های صدور مجوز و استفاده، نگرانی های قانونی در مورد حریم خصوصی،حفاظت از داده ها، امنیت ملی و حقوق مالکیت معنوی. علاوه بر این، فرآیند آماده سازی، گردآوری و میزبانی مجموعه داده های‌باز در مقیاس بزرگ پرهزینه است. حفظ داده های باز با کیفیت بالا نیازمند سرمایه گذاری مداوم در تخصص انسانی و زیرساخت های‌فناوری است و همین امر، پشتیبانی مالی بلندمدت را به یک چالش تبدیل می کند (باک، 2025[54]).

لازم‌به ذکر است که چارچوب باز بودن داده ها از یک سو، برای منابع داده اصلی و از سوی دیگر، برای مجموعه داده ایجاد شده از ترکیبی‌از این منابع داده اعمال می شود. این بدان معناست که یک مجموعه داده آموزشی ممکن است شامل داده هایی با سطوح‌مختلف باز بودن باشد، اما مجموعه داده همچنان بسته باقی می ماند. بسیاری از LLM های محبوب بر روی داده هایی با درجات‌مختلف باز بودن (مانند داده های وب اسکرپینگ و دارای مجوز) آموزش دیده اند، اما مجموعه داده های آموزشی آنها همچنان‌اختصاصی است. شرح مجموعه داده ممکن است در دسترس باشد، اما دسترسی به مجموعه داده کامل یا استفاده از آن‌امکان پذیر نیست. به عنوان مثال، گزارش فنی OpenAI GPT-4 بیان می کند که GPT-4 از قبل آموزش دیده است.*با استفاده ازداده های عمومی (مانند داده های اینترنتی) و داده های دارای مجوز از ارائه دهندگان شخص ثالث*)OpenAI، 2023[55]). با این‌حال، خود مجموعه داده های آموزشی بسته باقی مانده است. این ممکن است به دلیل ترکیبی از عوامل، مانند منافع تجاری درحفظ مالکیت معنوی و اختصاصی مجموعه داده ها یا ملاحظات صدور مجوز مربوط به نحوه تهیه، سازماندهی و اشتراک گذاری‌داده ها باشد. پاسخ های سیاستی بالقوه برای پشتیبانی از در دسترس بودن داده ها برای توسعه هوش مصنوعی درمقاله ای که به زودی توسط OECD منتشر خواهد شد، بررسی خواهد شد.

۴.۲.۲.استخراج داده های اینترنتی عمومی

استخراج‌داده ها، که به آن استخراج وب نیز گفته می شود، استخراج خودکار داده های وب قابل دسترس عموم با استفاده از یک عامل نرم افزاری("ربات") است. استفاده از عامل های نرم افزاری برای مرور خودکار وب چیز جدیدی نیست و سال هاست که در بسیاری از زمینه ها به‌کار گرفته می شود. به عنوان مثال، سایت های مقایسه پروازهای هواپیمایی از "اسکرپینگ صفحه" برای اسکن وب سایت های هواپیمایی برای‌قیمت ها استفاده می کنند (ویتاکر، 2024)[56]به طور مشابه، «خزنده های وب» توسط موتورهای جستجو مانند گوگل برای فهرست بندی محتوای‌آنلاین و پیوند دادن کاربران به صفحات مرتبط استفاده می شوند (گوگل، و[57]).

ازمنظر نحوه ی ثبت داده ها، می توان خراشیدن صفحه نمایش را از خزیدن در وب متمایز کرد. خراشیدن صفحه نمایش شامل دانلوددداده های نمایش داده شده در وب سایت ها است، در حالی که خزیدن در وب بر پیوند دادن وب سایت ها از طریق تجزیه وتحلیل کلمات کلیدی و فراداده (بدون نیاز به دانلود محتوای آن وب سایت ها) تمرکز دارد. تفاوت های تحلیلی بیشتر بین این فرآیندهادر مقاله (OECD، 2025) مورد بحث قرار گرفته است.[21]) «مسائل مربوط به مالکیت معنوی در هوش مصنوعی آموزش دیده‌بر روی داده های خراشیده».

جذابیت‌جمع آوری داده ها به عنوان ابزاری برای جمع آوری داده ها، منجر به استفاده گسترده از ربات های جمع آوری داده ها برای اهداف تجاری وغیرتجاری، از جمله آموزش مدل هوش مصنوعی، شده است. جالب توجه است که مدل های سنتی یادگیری ماشینی در درجه اول با استفاده ازمجموعه داده های ساختاریافته ای که برای وظایف خاص انتخاب شده بودند، آموزش داده می شدند. با این حال، LLMها برای درک پیچیدگی های‌زبان طبیعی به مجموعه داده های بزرگتر و متنوع تری نیاز دارند. بنابراین، آنها معمولاً بر اساس داده های جمع آوری شده از وب آموزش‌داده می شوند (لی، 2023)[12]).

شرکت های‌فناوری و اپراتورهای پلتفرم، مانند پلتفرم های رسانه های اجتماعی، موتورهای جستجو و سایت های تجارت الکترونیک، هم منبع‌داده برای استخراج داده ها و هم مشارکت کنندگان فعال در اکوسیستم استخراج داده ها هستند (OECD، 2025)[21]). تقاضای روزافزون‌برای داده های وب اسکریپ شده، باعث ایجاد سازمان هایی مانند Common Crawl یا LAION، به نام تجمیع کنندگان داده های‌هوش مصنوعی، شده است که داده ها را جمع آوری کرده و با اشخاص ثالث به اشتراک می گذارند (OECD، 2025)[21]در واقع، تجمیع کننده های‌هوش مصنوعی نقش قابل توجهی در تسهیل داده های جمع آوری شده برای توسعه هوش مصنوعی ایفا می کنند؛ مطالعه ای‌که LLM ۴۷ منتشر شده بین سال های ۲۰۱۹ تا اکتبر ۲۰۲۳ را تجزیه و تحلیل کرد، نشان داد که حداقل ۶۴٪ از آنها با استفاده از داده هایCommon Crawl آموزش دیده اند (Baack، ۲۰۲۴)[58]).

دررابطه با نوع داده های به دست آمده از طریق جمع آوری اطلاعات، لازم به ذکر است که محتوای جمع آوری شده ممکن است نه تنها شامل‌داده های شخصی باشد که مستقیماً از افراد گرفته شده یا در مورد آنها مشاهده شده است، بلکه شامل داده های شخصی نیز باشدکه به طور عمومی به صورت آنلاین توسط اشخاص ثالث به اشتراک گذاشته می شود. به عنوان مثال، تصاویر، نام ها یا نظرات ممکن‌است توسط کاربران رسانه های اجتماعی در مورد افراد دیگر ارسال شود - مانند تگ کردن کسی در یک عکس یا ذکر نام او در یک نظر.در این موارد، داده های جمع آوری شده به افرادی اشاره دارد که خودشان محتوا را آپلود نکرده اند، که نشان می دهد مجموعه داده های‌جمع آوری شده ممکن است شامل داده های شخصی باشد که هم از شخص موضوع داده و هم از اشخاص ثالث سرچشمه می گیرد.

اگرچه‌داده ها ممکن است در یک وب سایت قابل دسترسی باشند، اما این به طور خودکار به این معنی نیست که واجد شرایط داده های بازهستند که می توانند آزادانه دوباره مورد استفاده قرار گیرند (مجمع جهانی حریم خصوصی، ۲۰۲۴)[59]). قانونی بودن جمع آوری داده ها برای‌آموزش مدل های هوش مصنوعی در حال حاضر در حوزه های قضایی مختلف، با تمرکز ویژه بر مالکیت معنوی، امنیت سایبری، حریم‌خصوصی و مسائل مربوط به مدیریت داده ها، مورد بحث است. در همین حال، شرکت های میزبانی وب به طور فزاینده ای در حال‌اجرای اقدامات فنی و قانونی، مانند پروتکل های حذف ربات (robots.txt) و ممنوعیت های صریح در شرایط استفاده وب سایت های‌خود، برای کاهش فعالیت های جمع آوری داده ها از وب هستند.

۵ نتیجه گیری

این‌مقاله، سازوکارهای کلیدی جمع آوری داده ها که توسط شرکت ها در توسعه مدل های هوش مصنوعی استفاده می شوند را برجسته‌می کند و بر منابعی که داده ها معمولاً از آنها به دست می آیند تمرکز دارد: ۱) مستقیماً از افراد و سازمان ها، و ۲) از ارائه دهندگان‌شخص ثالث. در این دو دسته کلی، سازوکارهای متمایزی را شناسایی می کنیم که نقش مهمی در ایجاد مجموعه داده های‌آموزشی هوش مصنوعی ایفا می کنند.

یکی‌از سازوکارهای مهم و رو به رشد برای آموزش هوش مصنوعی، استفاده از داده های تولید شده از تعاملات با ابزارهای هوش مصنوعیB2C، مانند چت بات ها، دستیاران مجازی و سیستم های خدمات مشتری خودکار است. داده های تعاملی همچنین از طریق ادغام هایB2B جمع آوری می شوند، جایی که قابلیت های هوش مصنوعی از طریق مشارکت مستقیم یا دسترسی به API در محصولات‌تعبیه می شوند.

علاوه‌بر داده های جمع آوری شده در طول استفاده از ابزارهای هوش مصنوعی، برخی از توسعه دهندگان ممکن است به داده های سایر سرویس های دیجیتال‌خود نیز دسترسی داشته باشند و ممکن است به دنبال بهره برداری از این داده ها برای آموزش مدل هوش مصنوعی باشند.

صدورمجوز داده های تجاری یکی دیگر از رویه های رایج است که در آن توسعه دهندگان هوش مصنوعی از طریق توافق نامه های صدور مجوز با سازمان هایی‌که مجموعه داده های شخص ثالث را ارائه می دهند، به داده ها دسترسی پیدا می کنند. بازارهای داده و دلالان داده نیز منابع مهمی‌برای توسعه دهندگان هوش مصنوعی هستند که دسترسی به طیف گسترده ای از داده های شخص ثالث را ارائه می دهند.

داده های‌باز برای آموزش مدل های هوش مصنوعی به طور فزاینده ای اهمیت پیدا می کنند. مفهوم باز بودن داده ها در دو سطح عمل‌می کند: منابع داده اصلی و مجموعه داده های ایجاد شده از این منابع. این بدان معناست که در حالی که یک مجموعه داده آموزشی‌ممکن است شامل داده هایی با سطوح مختلف باز بودن باشد، خود مجموعه داده حاصل می تواند بسته و اختصاصی باقی‌بماند.

محبوبیت‌روزافزون جمع آوری داده ها به عنوان ابزاری برای جمع آوری داده ها، منجر به استفاده گسترده از آن برای آموزش مدل هوش مصنوعی‌شده است. با این حال، قانونی بودن جمع آوری داده ها در حال حاضر موضوع بحث در حوزه های قضایی مختلف است و نگرانی های‌اصلی در مورد مالکیت معنوی، امنیت سایبری، حریم خصوصی و مدیریت داده ها وجود دارد. به موازات آن، شرکت های میزبانی‌وب به طور فزاینده ای اقدامات فنی و قانونی را برای محدود کردن یا جلوگیری از فعالیت های جمع آوری داده ها اجرا می کنند.

اگرچه‌اهدای داوطلبانه داده ها هنوز منبع اصلی برای آموزش مدل هوش مصنوعی نیستند و ماهیت داوطلبانه واقعی آنها می تواند در برخی‌شرایط به چالش کشیده شود، اما در نهایت می توانند نقش مهمی در آموزش مدل هوش مصنوعی ایفا کنند و دسترسی به مجموعه‌داده های متنوع و واقعی را که در غیر این صورت دستیابی به آنها دشوار و پرهزینه است، فراهم کنند.

همچنین‌لازم به ذکر است که ابزارهای نوظهوری مانند PETها نقش مهمی در فراهم کردن امکان جمع آوری داده ها برای توسعه هوش مصنوعی‌ایفا می کنند. PETها با ایمن تر کردن جمع آوری و استفاده از داده ها، خطرات مربوط به حریم خصوصی و حاکمیتی را کاهش داده‌و از عملکرد مؤثر سازوکارهای ذکر شده در این طبقه بندی پشتیبانی می کنند.

این‌یافته ها، پیچیدگی و تنوع مکانیسم های جمع آوری داده ها که توسعه دهندگان هوش مصنوعی به آن ها متکی هستند، و همچنین‌چالش های قانونی در حال تکاملی که ممکن است با آن ها همراه باشد را برجسته می کند. چشم انداز دائماً در حال تغییراست و بررسی مداوم این مکانیسم ها برای شکل دهی سیاست هایی که توسعه هوش مصنوعی را با نیازهای حفظ حریم خصوصی‌و مدیریت داده ها متعادل می کنند، ضروری است.

منابع

[46] مرکزهوش مصنوعی (۲۰۲۵)، *https://aihub.or.kr/*.

[54] باک،ب. (2025)، «به سوی بهترین شیوه ها برای مجموعه داده های باز برای آموزش LLM»، *arXiv:2501.08365*، https://arxiv.org/abs/2501.08365.

[58] باک،س. (۲۰۲۴)، *داده های آموزشی برای قیمت یک ساندویچ*، https://foundation.mozilla.org/es/research/library/generative-ai-training-data/common-crawl/ (دسترسی‌در ۱۹ آگوست ۲۰۲۵).

[28] بدیلو،م. (۲۰۲۴)، *پردازش داده های شخصی برای آموزش هوش مصنوعی در برزیل: نکات کلیدی از تصمیمات‌اولیه ANPD در پرونده متا*، https://fpf.org/blog/processing-of-personaldata-for-ai-training-in-brazil-takeaways-from-anpds-preliminary-decisions-in-the-meta-case/ (دسترسی در۱۹ آگوست ۲۰۲۵).

[17] بوماسانی،ک. (2024)، *شاخص شفافیت مدل بنیاد نسخه ۱.۱*.

[43] بوسکت،آی. (۲۰۲۴)، *این یک شرکت کشاورزی قدیمی است—و جدیدترین فروشنده هوش مصنوعی شماست*، https://www.wsj.com/articles/bayer-microsoft-generative-ai-90754f54 (دسترسی در ۱۹آگوست ۲۰۲۵).

[8] مجلس‌سنای فدرال برزیل (۲۰۲۴)، *چارچوب قانونی هوش مصنوعی برزیل*.

[50] خزیدن‌مشترک (دوم)، *https://commoncrawl.org/*.

[47] اشیاءرایج در متن (۲۰۲۵)، *https://cocodataset.org/*.

[18] سازمان‌رقابت و بازارها (۲۰۲۳)، *مدل های بنیاد هوش مصنوعی: گزارش اولیه*.

[15] دنگ،جی. (2009)، «ImageNet: یک پایگاه داده تصویر سلسله مراتبی در مقیاس بزرگ. در سال IEEE 2009 کنفرانس‌بینایی کامپیوتر و تشخیص الگو*کنفرانس IEEE در زمینه بینایی کامپیوتر و تشخیص الگو*، صفحات ۲۴۸-۲۵۵.

[6] کمیسیون‌اروپا (2024)، *پیش نویس دوم هوش مصنوعی همه منظوره*، https://digital-strategy.ec.europa.eu/en/library/second-draft-general-purpose-ai-code-practice-publishedwritten-independent-experts.

[5] اتحادیه‌اروپا (2024)، *آیین نامه (EU) 2024/1689 پارلمان اروپا و شورای۱۳ ژوئن ۲۰۲۴، وضع قوانین هماهنگ در مورد هوش مصنوعی و اصلاح مقررات (EC) شماره ۳۰۰/۲۰۰۸، (EU) شماره ۱۶۷/۲۰۱۳، (EU) شماره ۱۶۸/۲۰۱۳، (EU) ۲۰۱۸/۱۱۳۹)EU(۲۰۱۸/۸۵۸، و*.

[29] اتحادیه‌اروپا (۲۰۲۲)، *آیین نامه (EU) 2022/868 پارلمان اروپا و شورای30 مه 2022 در مورد حاکمیت داده اروپا و اصلاح مقررات (EU) 2018/1724 (قانون حاکمیت داده)*، https://eur-lex.europa.eu/legalcontent/EN/TXT/PDF/?uri=CELEX:32022R0868.

[44] اتحادیه‌اروپا (۲۰۱۹)،*دستورالعمل (EU) 2019/1024 پارلمان اروپا و شورای20 ژوئن 2019 در مورد داده های باز و استفاده مجدد از اطلاعات بخش عمومی (بازنگری)*، https://eur-lex.europa.eu/eli/dir/2019/1024/oj/eng (دسترسی در ۱۹ آگوست ۲۰۲۵).

[13] فدرکوپر، ل. (۲۰۲۳)، «گزارش اولین کارگاه آموزشی هوش مصنوعی مولد و قانون»،*حقوق و قانون ییل مقاله‌پژوهشی اقتصاد*،https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4634513.

[25] کمیسیون‌تجارت فدرال (2024)،*شرکت های هوش مصنوعی (و سایر): تغییر بی سروصدای شرایط خدمات شما می تواند ... ناعادلانه‌یا فریبنده*،https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/02/ai-companies-quietly-changing-your-terms-service-could-be-unfair-or-deceptive-other.

[10] گروه۷ (۲۰۲۳)،*آیین نامه بین المللی رفتار سازمانی فرآیند هیروشیما*.

[14] گرلاخ،ف. (۲۰۱۸)، «یک پیکره استاندارد پروژه گوتنبرگ برای تحلیل آماری داده های طبیعی» زبان‌و زبان شناسی کمی"،*arXiv:1812.08092*،https://arxiv.org/abs/1812.08092v1.

[59] مجمع‌جهانی حریم خصوصی (۲۰۲۴)،*بیانیه مشترک پایانی در مورد جمع آوری داده ها و حفاظت از آنها ازحریم خصوصی*،https://www.priv.gc.ca/en/opc-news/speeches-and-statements/2024/jsdc_20241028/ (دسترسی در ۱۵ مه ۲۰۲۵).

[26] گلدنبرگ،ل. (۲۰۲۱)، «مجموعه داده های سفرهای چند مقصدی Booking.com»،*مجموعه مقالات SIGIR '21: چهل‌و چهارمین کنفرانس بین المللی ACM SIGIR در زمینه تحقیق و توسعه در بازیابی اطلاعات*، صفحات ۲۴۵۷ -۲۴۶۲،https://doi.org/10.1145/3404835.3463240.

[62] گوگل(دوم)، .

[57] گوگل(دوم)،*جستجوی گوگل مرکزی*،https://developers.google.com/search/docs/crawling-indexing/خزنده های گوگل-کلی-مرور/ایندکس‌کردن (دسترسی در ۱۹ آگوست ۲۰۲۵).

[31] هیرش‌ام سی، آر. (۲۰۲۰)، «بیماری های نادر ۲۰۳۰: چگونه هوش مصنوعی تقویت شده از تشخیص و درمان‌بیماری های نادر در آینده»، صفحات ۷۴۰-۷۴۳،https://doi.org/10.1136/annrheumdis-2020-217125.

[51] چهره‌در آغوش گرفته (دوم)،.

[36] ICO(2025)،*قانون استفاده و دسترسی به داده ها ۲۰۲۵ (DUAA) - برای سازمان ها چه معنایی دارد؟*، https://ico.org.uk/about-the-ico/what-we-do/legislation-we-cover/data-use-and-access-act-2025/the-data-use-and-access-act-2025-what-does-it-mean-for-organisations/#innovate (دسترسی‌در ۲۶ ژوئن ۲۰۲۵).

[16] ایمیج نت(دوم)،*http://www.image-net.org/about*.

[48] جی.دنگ، دبلیو. (۲۰۰۹)، «ImageNet: یک پایگاه داده تصویر سلسله مراتبی در مقیاس بزرگ»،*کنفرانس IEEE بینایی‌کامپیوتر و تشخیص الگو*، جلد میامی، فلوریدا، ایالات متحده آمریکا، صفحات ۲۴۸-۲۵۵، https://doi.org/10.1109/CVPR.2009.5206848.

[9] رژیم‌ملی ژاپن (۲۰۲۵)،*قانون ترویج تحقیق و توسعه و بهره برداری فناوری های‌مرتبط با هوش مصنوعی*.

[30] کیرستین،ل. (۲۰۲۳)،*داده های تحرک به عنوان یک دارایی مشترک - به سوی یک زیرساخت داده های تحرک مشترک*، https://fsr.eui.eu/mobility/#_ftnref6-data-as-a-commons-towards-a-common-mobility-datainfrastructure (دسترسی در ۱۹ آگوست ۲۰۲۵).

[7] مجلس‌ملی کره (۲۰۲۴)،*قانون توسعه هوش مصنوعی و ایجاداعتماد*.

[12] لی،سی. (۲۰۲۳)، «هوش مصنوعی و قانون: نسل بعدی»،https://blog.genlaw.org/explainers/.

[27] لوین(۲۰۲۴)،*چگونه آمازون از هوش مصنوعی مولد برای بهبود توصیه های محصول استفاده می کند و توضیحات*،https://www.aboutamazon.com/news/retail/amazon-generative-ai-productsearch-results-and-descriptions (دسترسی در ۱۹ آگوست ۲۰۲۵).

[60] لیبرمن،ب. (۲۰۲۰)،*ردیابی انتشار کربن جهانی در لحظه با استفاده از Climate TRACE*، https://yaleclimateconnections.org/2020/08/climate-trace-to-track-real-time-global-carbonemissions/.

[49] لیزنفلد،آ. (2023)، «باز کردن ChatGPT: ردیابی باز بودن، شفافیت و پاسخگویی‌در مولدهای متن تنظیم شده با دستورالعمل»،*مجموعه مقالات پنجمین کنفرانس بین المللی رابط های‌کاربری محاوره ای، CUI 2023*، https://dl.acm.org/doi/10.1145/3571884.3604316.

[66] لورنز،پ. (۲۰۲۳)،*ملاحظات اولیه سیاست گذاری برای هوش مصنوعی مولد*، https://doi.org/10.1787/fae2d1e6-en.

[38] ابُرایان،م. (۲۰۲۳)،*شرکت OpenAI، سازنده ی ChatGPT، برای صدور مجوز انتشار اخبار با آسوشیتدپرس قرارداد امضا کرد.*، https://apnews.com/article/openai-chatgpt-associated-press-apf86f84c5bcc2f3b98074b38521f5f75a (دسترسی در ۱۹ آگوست ۲۰۲۵).

[53] سازمان‌همکاری و توسعه اقتصادی (2025)،*افزایش دسترسی و اشتراک گذاری داده ها در عصر هوش مصنوعی*، انتشاراتOECD، پاریس.

[35] سازمان‌همکاری و توسعه اقتصادی (2025)،*تسهیل استفاده ثانویه از داده های سلامت برای اهداف عمومی در سراسر مرزها*انتشارات OECD، پاریس،https://doi.org/10.1787/d7b90d15-en.

[21] سازمان‌همکاری و توسعه اقتصادی (2025)،*مسائل مربوط به مالکیت معنوی در هوش مصنوعی آموزش دیده بر اساس داده های خراشیده شده*، سازمان همکاری و توسعه اقتصادی انتشارات،پاریس،https://doi.org/10.1787/d5241a23-en.

[11] سازمان‌همکاری و توسعه اقتصادی (2025)،*راه اندازی چارچوب گزارش دهی فرآیند هوش مصنوعی هیروشیما (HAIP)*، https://www.oecd.org/en/events/2025/02/launch-of-the-hiroshima-ai-process-reportingframework.html (دسترسی در ۷ فوریه).

[24] سازمان‌همکاری و توسعه اقتصادی (2025)،*اشتراک گذاری مدل های هوش مصنوعی قابل اعتماد با فناوری های تقویت کننده حریم خصوصی*، سازمان همکاری و توسعه اقتصادی انتشارات،پاریس،https://doi.org/10.1787/a266160b-en.

[20] سازمان‌همکاری و توسعه اقتصادی (2024)،*یادداشت توضیحی در مورد تعریف به روز شده OECD از سیستم هوش مصنوعی*، انتشاراتOECD،https://doi.org/10.1787/623da898-en.

[65] سازمان‌همکاری و توسعه اقتصادی (2023)،*مدل های زبانی هوش مصنوعی: ملاحظات فناوری، اجتماعی-اقتصادی و سیاست گذاری*، انتشاراتOECD، پاریس،https://doi.org/10.1787/13d38f92-en..

[23] سازمان‌همکاری و توسعه اقتصادی (2022)،*راهنمای سیاست گذاری در حوزه حاکمیت داده*انتشارات OECD، پاریس، https://doi.org/10.1787/40d53904-en.

[1] »برای‌طبقه بندی سیستم های هوش مصنوعی OECD چارچوب« ،)2022(OECD*اقتصاد دیجیتال OECD مقالات*، شماره ۳۲۳، انتشارات OECD، پاریس،https://doi.org/10.1787/cb6d9eca-en.

[64] سازمان‌همکاری و توسعه اقتصادی (2021)،*ترسیم ابتکارات، فرصت ها و چالش ها در زمینه قابلیت انتقال داده ها*، سازمان همکاری و توسعه اقتصادی انتشارات،پاریس،https://doi.org/10.1787/a6edfab2-en..

[32] سازمان‌همکاری و توسعه اقتصادی (2021)،*توصیه شورا در مورد افزایش دسترسی و اشتراک گذاری داده ها*، https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0463 (دسترسی در ۱۹آگوست ۲۰۲۵).

[4] سازمان‌همکاری و توسعه اقتصادی (۲۰۱۹)،*افزایش دسترسی و اشتراک گذاری داده ها: تطبیق خطرات و مزایا برای استفاده‌مجدد از داده ها در جوامع مختلف*انتشارات OECD، پاریس،https://doi.org/10.1787/b4d546a9-en.

[22] سازمان‌همکاری و توسعه اقتصادی (۲۰۱۹)،*افزایش دسترسی و اشتراک گذاری داده ها: تطبیق خطرات و مزایا برای استفاده‌مجدد از داده ها در جوامع مختلف*انتشارات OECD، پاریس،https://doi.org/10.1787/b4d546a9-en.

[2] سازمان‌همکاری و توسعه اقتصادی (۲۰۱۹)،*توصیه شورای هوش مصنوعی*، https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (دسترسی در ۱۹ آگوست۲۰۲۵).

[52] سازمان‌همکاری و توسعه اقتصادی (۲۰۱۵)،*نوآوری مبتنی بر داده: کلان داده برای رشد و رفاه*انتشارات OECD پاریس،https://doi.org/10.1787/9789264229358-en.

[3] سازمان‌همکاری و توسعه اقتصادی (۲۰۱۴)،*حفاظت از حریم خصوصی در اقتصاد داده محور: بررسی تفکرات فعلی*، https://one.oecd.org/document/DSTI/ICCP/REG)2014(3/en/pdf.

[40] اوپن ای آی(۲۰۲۴)،*همکاری OpenAI و Reddit*، https://openai.com/index/openai-and-reddit-مشارکت/ (دسترسی در ۱۹ آگوست ۲۰۲۵).

[39] اوپن ای آی(۲۰۲۳)،*اکسل اسپرینگر استفاده مفید از هوش مصنوعی در روزنامه نگاری را تعمیق می بخشد*، https://openai.com/index/axel-springer-partnership/ (دسترسی در ۱۹ آگوست ۲۰۲۵).

[55] اوپن ای آی(۲۰۲۳)،*گزارش فنی GPT-4*، https://arxiv.org/pdf/2303.08774.

[19] پابلوویلالوبوس، آ. (2022)، «آیا داده ها تمام خواهند شد؟ محدودیت های مقیاس پذیری LLM بر اساس داده های انسانی» داده های‌تولید شده"،https://arxiv.org/abs/2211.04325.

[42])2024(PR Newswire،*نکسیس نکسیس از نکسیس دیتا پلاس، یک پلتفرم تکAPI برای کمک های مالی رونمایی کرد دسترسی‌بی سابقه سازمان ها به محتوای خبری دارای مجوز و داده های باکیفیت شرکتی که توسط هوش مصنوعی عمومی‌تأیید شده اند*، https://www.prnewswire.com/news-releases/lexisnexis-unveils-nexisdata-a-single-api-platform-giving-organizations-unprecedented-access-to-gen-ai-approvedlicensed-news-content-and-high-quality-company-data-302325319.html.

[37] ساگارا،اچ. (۲۰۱۹)،*گزارش نهایی در مورد طرح های آزمایشی بارسلونا، ارزیابی های BarcelonaNow و طرح های‌پایداری*.

[41] شاتراستوک (۲۰۲۳)،*شاتر استوک رابطه طولانی مدت خود با متا را گسترش می دهد*، https://www.prnewswire.com/news-releases/shutterstock-expands-long-standing-relationshipwith-meta-301719769.html (دسترسی در ۱۹ آگوست ۲۰۲۵).

[45] ملت‌هوشمند سنگاپور (دوم)،.

[33] بانک‌زیستی بریتانیا (دوم)،*https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank*.

[61] وزارت‌بازرگانی ایالات متحده (۲۰۲۴)،*آماده سازی داده های باز برای عصر هوش مصنوعی*، https://www.commerce.gov/news/blog/2024/01/preparing-open-data-age-ai.

[63] مجلس‌نمایندگان ایالات متحده (۲۰۲۴)،*گزارش کارگروه دو حزبی مجلس نمایندگان در مورد هوش مصنوعی: راهنما اصول،توصیه های آینده نگر و پیشنهادهای سیاستی برای اطمینان از اینکه آمریکا همچنان رهبری جهان رادر نوآوری مسئولانه هوش مصنوعی بر عهده دارد*..

وایزبند،س. (۲۰۲۳)،*مطالعه GIHF-AI 2023، اعتماد به استفاده از داده های سلامت - یک مقایسه روابطآلمان و اسرائیل.* [34]

ویتاکر،جی. (۲۰۲۴)،*رایان ایر در پرونده ی «اسکرین اسکرین» Booking.com پیروز شد*، https://www.aviationbusinessnews.com/industry-news/ryanair-wins-booking-com-screenscraper-case/ (دسترسی در ۱۹ آگوست ۲۰۲۵). [56]

یادداشت ها

۱)OECD، 2023(ها الگوریتم های پیشرفته یادگیری ماشین هستند که از طریق فرآیندی که اطلاعات را از داده های ورودی به پارامترهای‌مدل خود کدگذاری می کند، توسعه یافته و اصلاح می شوند. این امر به آنها امکان یادگیری و انجام وظایفی مانند ترجمه،تولید محتوا یا تشخیص تصویر را می دهدLLM[65]) (لورنز، 2023[66]).

۲دامنه عمومی شامل تمام داده هایی می شود که اشتراک گذاری آنها توسط حقوق مالکیت فکری یا هرگونه حقوق دیگری با اثرات‌مشابه محدود نشده است. بنابراین، در اینجا به طور گسترده تری از صرفاً عاری بودن از حمایت حق چاپ درک می شود (OECD، 2019)[4]. حوزه عمومی می تواند با حوزه های شخصی و اختصاصی همپوشانی داشته باشد (شکل ۲ را ببینید).

۳شاخص شفافیت مدل بنیادی ۲۰۲۴ شامل ۲۳ زیردامنه است که جنبه های مختلف مدل های بنیادی را از توسعه تا استقرار ارزیابی‌می کند: داده ها، تأثیر نیروی کار، دسترسی به داده ها، محاسبات، کد، تأثیر زیست محیطی، ویژگی های مدل، قابلیت ها، محدودیت ها،خطرات، قابلیت اعتماد، کاهش اثرات مدل، ارزیابی ها، به روزرسانی های مدل، فرآیند انتشار، کانال های توزیع، رابطکاربری، حفاظت از داده ها، مکانیسم های بازخورد، مستندسازی، سیاست های استفاده، جغرافیای تحت تأثیر و تأثیر (بوماسانی،۲۰۲۴)[17]

۴در برخی موارد، مانند مدل های استدلال، تنظیم دقیق ممکن است شامل یادگیری تقویتی از بازخورد انسانی باشد، نه اینکه صرفاً بر اساس مجموعه داده های خاص دامنه آموزش داده شود.

۵به یادداشت ۲ مراجعه کنید.

۶اگرچه این طبقه بندی بر منابع جمع آوری داده ها تمرکز دارد، اما توجه به این نکته ضروری است که روش های دسترسی به داده هامی تواند در این مکانیسم ها به طور قابل توجهی متفاوت باشد. در برخی موارد، داده ها به طور کامل به دست می آیند - به عنوان مثال،از طریق قراردادهای مجوز تجاری، که در آن دارنده مجوز ممکن است مالکیت یا حقوق استفاده مجدد گسترده ای رابه دست آورد. در موارد دیگر، داده ها فقط تحت شرایط تعریف شده ای مانند محیط های پردازش امن قابل دسترسی هستند، جایی‌که می توان آن ها را تجزیه وتحلیل کرد اما نمی توان آن ها را دانلود یا نگهداری کرد.

۷طبق قانون حاکمیت داده، نهادها می توانند به عنوان «سازمان های نوع دوستی داده» ثبت شوند که به اشتراک گذاری داده ها رابرای منافع عمومی مدیریت و تسهیل می کنند و تضمین می کنند که داده ها به صورت مسئولانه و مطابق با مقررات حفظ حریم‌خصوصی استفاده می شوند.

۸ماهیت داوطلبانه ی اهدای داده ها ممکن است تحت تأثیر عواملی مانند سواد دیجیتال، آگاهی از حقوق داده ها و درک خطرات مرتبط‌باشد. الگوهای مشارکت ممکن است در گروه های جمعیتی مختلف متفاوت باشد و زمینه ی درخواست های اهدای داده هامی تواند بر ماهیت واقعاً داوطلبانه ی رضایت تأثیر بگذارد.

۹«ترتیبات داده های باز» به ترتیبات دسترسی و اشتراک گذاری غیرتبعیض آمیز داده ها اشاره دارد، که در آن داده ها قابل خواندن‌توسط ماشین هستند و می توانند به صورت رایگان به آنها دسترسی پیدا کرده و به اشتراک گذاشته شوند و توسط هر کسی‌برای هر هدفی مورد استفاده قرار گیرند، حداکثر با رعایت الزاماتی که یکپارچگی، منشأ، انتساب و باز بودن را حفظ می کنند (OECD، 2021)[32].