Ethics By Design and Ethics of Use Approaches for Artificial Intelligence

Version 1.0
25 November 2021

اخلاق در طراحی

و

اخلاقیات در رویکردهای استفاده

برای

هوش مصنوعی

سلب مسئولیت

این سند به درخواست کمیسیون اروپا (مدیریت کل پژوهش و نوآوری) توسط گروهی از کارشناسان تهیه شده است و هدف آن افزایش آگاهی در جامعه‌ی علمی، با بهره‌مندی از پروژه‌های پژوهشی و نوآورانه‌ی اتحادیه‌ی اروپا است. این مقاله، دستورالعمل رسمی اتحادیه‌ی اروپا را تشکیل نمی‌دهد. نه کمیسیون اروپا و نه هیچ فردی از طرف آنها مسئول استفاده از آن نیست.

اخطار حقوقی

این یادداشتِ راهنما هیچ تعهد جدیدی را برای کمیسیون اروپا، عوامل اجرایی یا محققانی که خود ارزیابی اخلاقی خود را تکمیل می‌کنند، ایجاد نمی‌کند.

قدردانی

تهیه‌ی این سند با همکاری آلبنا کویومجیوا (مدیریت کل پژوهش و نوآوری؛ در حال حاضر EISMEA) انجام شده‌است. این سند توسط برندت داینو و فیلیپ بری نوشته شده است و توسط جیوانی کوماندی، جما گالدون کلاول، اینیگو دی میگوئل بریاین؛ برند استال، لارنس بروکس و کارکنان بخش اخلاق و یکپارچگی تحقیقات، مدیریت کل پژوهش و نوآوری؛ لیزا دیپندل، فرانچسکو دورانتی (در حال حاضر در HADEA) ادیتا سیکورسکا، میهالیس کریتیکوس و ایو دومونت. مورد بررسی قرار گرفته‌است.

بخش اخلاق و یکپارچگی تحقیقات، مدیریت کل پژوهش و نوآوری از همه‌ی مشارکت‌کنندگان تشکر می‌کند.

کمیسیون اروپا

مدیریت کل پژوهش و نوآوری RTD.03.001-

بخش اخلاق و یکپارچگی تحقیقات

پست الکترونیکی: RTD-ETHICS-REVIEW-HELPDESK@ec.europa.eu

بروکسل/بلژ ORBN B-1049

فهرست

معرفی

این راهنما به فعالیت‌های پژوهشی مرتبط با توسعه و استفاده از **هوش مصنوعی** (مانند رباتیک) می‌پردازد. این مقاله بر اساس **دستورالعمل‌های اخلاقی برای هوش مصنوعی قابل‌اعتماد** که توسط کارگروه تخصصی و مستقل هوش مصنوعی تهیه شده است و همچنین نتایج پژوهش‌های SIENNA و SHERPA با **اتحادیهٔ اروپا** نگاشته شده است.

این راهنما بر پایه ی کار گروهی از متخصصان مستقل سطح عالی در هوش مصنوعی و «دستورالعمل‌های اخلاقی برای هوش مصنوعی قابل اعتماد» آنها و همچنین نتایج پروژه‌های SHERPA و SIENNA با بودجه اتحادیه اروپا تهیه شده است.

این سند، راهنمایی برای اتخاذ یک رویکرد اخلاقی متمرکز در حین طراحی، توسعه، و استقرار و/ یا استفاده از راه‌حل‌های مبتنی بر هوش مصنوعی ارائه می‌دهد؛ اصول اخلاقی‌ای که سیستم‌های هوش مصنوعی باید از آنها پشتیبانی کنند را توضیح داده و ویژگی‌های کلیدی‌ای را که سیستم/برنامه‌های مبتنی بر هوش مصنوعی باید برای حفظ و ارتقاء داشته باشند، مورد بحث قرار می‌دهد:

- احترام به عامل انسانی؛
- حریم خصوصی، حفاظت از داده‌های شخصی و حاکمیت داده‌ها؛
- انصاف (عدالت)؛
- رفاه فردی، اجتماعی و محیط‌زیستی؛
- شفافیت؛
- مسئولیت پذیری و نظارت.

علاوه‌براین، این سند وظایف خاصی که باید انجام شوند تا هوش مصنوعی با این مشخصات تولید شود را نیز شرح می‌دهد.

برای پژوهشگرانی که قصد دارند از سیستم‌های مبتنی بر هوش مصنوعی در تحقیقات خود استفاده کنند، این سند ویژگی‌های اخلاقی‌ای که باید در آن سیستم‌ها وجود داشته باشند تا امکان استفاده از آنها فراهم شود را توضیح می‌دهد.

رویکرد محوری به کار رفته در این راهنما، «اخلاق در طراحی» است. هدف **اخلاق در طراحی** گنجاندن اصول اخلاقی در فرایند توسعه است که اجازه می‌دهد تا مسائل اخلاقی در اسرع وقت موردتوجه قرار بگیرند و در طول فعالیت‌های پژوهشی به‌دقت پیگیری شوند. این مقاله به صراحت وظایف مشخصی را که می‌توان به عهده گرفت و برای هر روش‌شناسی توسعه (مانند AGILE، V-Method یا CRISP-DA) می‌توانند به کار روند را مشخص می‌کند. با این حال رویکرد توصیه‌شده باید متناسب با نوع پژوهش پیشنهادی باشد و باید مدنظر داشت که ریسک‌های اخلاقی می‌توانند در طول مرحله‌ی پژوهش و استقرار و به کارگیری یا مرحله‌ی اجرا متفاوت باشند. رویکرد **اخلاق در طراحی** ارائه‌شده در این راهنما یک ابزار اضافی دیگری برای رسیدگی به نگرانی‌های مربوط به اخلاق و نشان‌دادن انطباق با اخلاقیات نیز ارائه می‌دهد. با این حال اتخاذ رویکرد **اخلاق در طراحی** مانع از دیگر اقدامات برای اطمینان از پایبندی به تمام اصول اخلاقی هوش مصنوعی و انطباق با چارچوب قانونی اتحادیه‌ی اروپا با هدف تضمین انطباق کامل اخلاقی و اجرای الزامات اخلاقی نیست.

رویکرد پیشنهادی برای متقاضیان و ذینفعان Horizon Europe اجباری نیست، اما هدف آن ارائه‌ی راهنمایی‌های بیشتر برای رسیدگی به نگرانی‌های مرتبط با اخلاق و نشان دادن انطباق با اخلاقیات است.

این سند به سه بخش تقسیم می‌شود:

- **بخش ۱**: اصول و الزامات:
  این بخش اصول اخلاقی‌ای که سیستم‌های هوش مصنوعی باید به آنها پایبند باشند را تعریف کرده و الزاماتی را برای توسعه‌ی آن‌ها استخراج می‌کند؛
- **بخش ۲**: مراحل عملی برای استفاده از **اخلاق در طراحی** در توسعه‌ی هوش مصنوعی:
  این بخش مفهوم **اخلاق در طراحی** را شرح داده و آن را به یک مدل عمومی برای توسعه‌ی سیستم‌های هوش مصنوعی مرتبط می‌کند و اقداماتی را که باید در مراحل مختلف

توسعه‌ی هوش مصنوعی با هدف رعایت اصول اخلاقی و الزامات ذکرشده در بخش ۱ انجام شوند تعریف می‌کند؛

- **بخش ۳:** استقرار اخلاق و استفاده از دستورالعمل‌ها برای پیاده‌سازی یا به‌کارگیری هوش مصنوعی به شیوه‌ای که از نظر اخلاقی مسئولانه است.

بخش ۱-

اخلاق در طراحی: اصول و الزامات

این بخش اصول اخلاقی‌ای را که سیستم‌های هوش مصنوعی باید به آن‌ها پایبند باشند تعریف کرده و الزاماتی که هوش مصنوعی باید مطابق آنها باشد را استخراج می‌کند.

الزامات اخلاقی به‌عنوان مشخصه‌های یک سیستم، به اصول آن تجسم می‌بخشند. در حالی که بسیاری از الزامات اخلاقی توسط الزامات قانونی پشتیبانی می‌شوند، انطباق اخلاقی را نمی‌توان تنها با پایبندی به تعهدات قانونی به دست آورد. اخلاق مربوط به حمایت از حقوق فردی مانند آزادی و حریم خصوصی، برابری و انصاف، اجتناب از آسیب و ارتقای رفاه فردی، و ساختن جامعه‌ای بهتر و پایدارتر است که اغلب راه‌حل‌هایی را پیش‌بینی می‌کند که در نهایت به الزامات قانونی برای رعایت آنها تبدیل می‌شوند.

اصول اخلاقی و الزامات

شش اصل کلی اخلاقی وجود دارد که هر سیستم هوش مصنوعی باید بر اساس حقوق اساسی که در منشور حقوق اساسی اتحادیه‌ی اروپا (منشور اتحادیه‌ی اروپا) و در قوانین حقوق بشر بین‌المللی مربوطه آمده است، از این اصول محافظت کرده و آنها را رعایت کند:

1. **احترام به عامل انسانی**:
   انسان‌ها باید برای تصمیم‌گیری و انجام اعمال خود مورد احترام قرار گیرند. احترام به عامل انسانی شامل سه اصل ویژه است که حقوق اساسی بشر را تعریف می‌کنند: **استقلال**، **کرامت** و **آزادی**؛
2. **حریم خصوصی و حاکمیت داده‌ها**:
   مردم حق دارند از حریم خصوصی و داده‌هایشان حفاظت شود و این موارد باید همیشه رعایت شوند؛
3. **انصاف (عدالت)**:
   باید به مردم حقوق و فرصت‌های برابر داده شود و نباید بدون داشتن شایستگی به آنها منفعت یا ضرری وارد شود؛
4. **رفاه فردی، اجتماعی و زیست‌محیطی**:
   سیستم‌های هوش مصنوعی باید به رفاه فردی، اجتماعی و زیست‌محیطی کمک کنند و به آنها آسیبی وارد نکنند؛
5. **شفافیت**:
   هدف، ورودی‌ها و عملیات برنامه‌های هوش مصنوعی باید برای ذینفعان قابل‌درک و معلوم باشد؛
6. **مسئولیت‌پذیری و نظارت**:
   انسان‌ها باید بتوانند طراحی و عملکرد سیستم‌های مبتنی بر هوش مصنوعی را درک، نظارت و کنترل کنند و بازیگران دخیل در توسعه یا عملیات آنها باید مسئولیت نحوه‌ی عملکرد این برنامه‌ها و پیامدهای ناشی از آن را بر عهده بگیرند.

به‌منظور تعبیه‌ی این شش اصل اخلاقی در طراحی‌ها، سیستم‌های هوش مصنوعی پیشنهادی باید با الزامات اخلاقی کلی که در زیر آمده‌اند منطبق باشند.

1- احترام به عامل انسانی

احترام به عامل انسانی شامل سه اصل خاص است که حقوق اساسی بشر را تعریف می‌کند: **استقلال**، **کرامت** و **آزادی**.

استقلال:

احترام گذاشتن به استقلال به این معنی است که به مردم اجازه دهیم خودشان فکر کنند، خودشان تصمیم بگیرند که چه چیزی درست و نادرست است و خودشان انتخاب کنند که چگونه می‌خواهند زندگی کنند. توجه به این نکته مهم است که سیستم‌های هوش مصنوعی می‌توانند بدون انجام کاری، استقلال انسان را محدود کنند ـ صرفاً با توجه نکردن به طیف کاملی از تنوع انسان‌ها در سبک زندگی، ارزش‌ها، باورها و سایر جنبه‌های زندگی که ما را منحصربه‌فرد می‌کنند. از این رو، چنین محدودیت‌هایی ممکن است بدون هیچ‌گونه قصد بدخواهانه‌ای پیش بیایند، اما صرفاً در نتیجه‌ی عدم درک زندگی، باورها، ارزش‌ها و ترجیحات مردم ایجاد شوند. این یک مشکل خاص با خدمات شخصی‌سازی است که هنجارهای فرهنگی متفاوت را در نظر نمی‌گیرد. علاوه بر این، خدمات شخصی‌سازی، حتی زمانی که هنجارهای فرهنگی متفاوت را در نظر می‌گیرد، ممکن است اطلاعات و گزینه‌های ارائه‌شده را محدود کند. سیستم‌های مبتنی بر هوش مصنوعی باید از محدود کردن بی‌معنای زمینه‌ی تصمیم‌گیری فردی خودداری کنند (همچنین به قسمت «آزادی» زیر مراجعه کنید).

کرامت:

کرامت بیان کننده ی این است که هر انسانی دارای ارزشی ذاتی‌ست که هرگز نباید به خطر بیفتد. از این رو، افراد نباید ابزاری، عینی یا غیرانسانی شوند، بلکه باید همیشه با احترام با آنها رفتار شود، مثلا هنگام استفاده یا قرارگرفتن در معرض سیستم‌های مبتنی بر هوش مصنوعی.

آزادی:

احترام به آزادی مستلزم آن است که مردم در انجام آن دسته از اقداماتی که باید به‌عنوان افراد مستقل دنبال کنند، مانند آزادی حرکت، آزادی بیان، آزادی دسترسی به اطلاعات و آزادی تجمع، محدود نشوند. علاوه بر این، آزادی مستلزم عدم وجود محدودیت‌هایی است که استقلال مردم را تضعیف می‌کند؛ مانند اجبار، فریب، بهره‌برداری از آسیب‌پذیری‌ها و دستکاری. اما آزادی مطلق نیست بلکه توسط قانون محدود می‌شود.

الزامات عمومی اخلاقی در عامل انسانی:

- کاربران نهایی و سایر افرادی که تحت‌تأثیر سیستم هوش مصنوعی قرار می‌گیرند، *نباید* از توانایی تصمیم‌گیری اساسی در مورد زندگی خود محروم شوند یا آزادی‌های اولیه از آنها سلب شود.
- *باید* اطمینان حاصل شود که برنامه‌های کاربردی هوش مصنوعی به طور مستقل و بدون نظارت انسانی و امکان جبران، تصمیم‌گیری نمی‌کنند: در مورد مسائل شخصی اساسی (مثلاً تأثیر مستقیم بر زندگی خصوصی یا حرفه‌ای، سلامت، رفاه یا حقوق فردی) که معمولاً توسط انسان‌ها از طریق انتخاب‌های شخصی آزاد تصمیم‌گیری می‌شوند؛ یا در مورد مسائل اساسی اقتصادی، اجتماعی و سیاسی، که معمولاً توسط مذاکرات جمعی تصمیم‌گیری می‌شوند، یا به طور مشابهی تأثیر قابل‌توجهی بر افراد می‌گذارند.
- کاربران نهایی و سایرین که تحت‌تأثیر سیستم هوش مصنوعی قرار می‌گیرند، نباید به هیچ وجه تابع، مجبور، فریب‌داده‌شده، دستکاری‌شده، مورد هدف قرارگرفته یا غیرانسانی شوند.
- وابستگی یا اعتیاد به سیستم و عملیات آن نباید عمداً تحریک شود. این اتفاق نباید از طریق عملیات و اقدامات مستقیم سیستم اتفاق بیفتد. همچنین باید تا حد امکان از کاربرد سیستم‌ها برای این اهداف جلوگیری شود.
- برنامه‌های کاربردی هوش مصنوعی باید طوری طراحی شوند که به اپراتورهای سیستم و تا حد امکان به کاربران نهایی توانایی کنترل، هدایت و مداخله در عملیات اساسی سیستم را بدهند.
- – کاربران نهایی و سایر افرادی که تحت تأثیر سیستم هوش مصنوعی قرار می‌گیرند باید اطلاعات قابل‌قبولی در مورد منطق به‌کاررفته در هوش مصنوعی و همچنین اهمیت و پیامدهای پیش‌بینی‌شده برای آنها دریافت کنند.

2- حریم خصوصی و حاکمیت داده

حفاظت از حریم خصوصی و داده‌ها حقوق اساسی‌ای هستند که باید همیشه رعایت شوند. سیستم‌های هوش مصنوعی باید به‌گونه‌ای ساخته شوند که اصول کمینه‌سازی داده‌ها و حفاظت از داده‌ها را بر اساس طراحی و به‌طور پیش‌فرض مطابق با مقررات عمومی حفاظت از داده‌های اتحادیه‌ی اروپا (GDPR) در خود داشته باشند.

حقوق مربوط به حریم خصوصی باید توسط مدل‌های حاکمیت داده‌ای که تضمین‌کنندی صحت و نمایندگی داده‌ها هستند؛ از داده‌های شخصی محافظت می‌کنند و انسان ها را قادر می‌سازند تا به‌طور فعال داده‌های شخصی خود و نحوه‌ی استفاده‌ی سیستم از آنها را مدیریت کنند، محافظت شوند. حفاظت مناسب از داده‌های شخصی می‌تواند موجب اعتماد در به اشتراک‌گذاری داده‌ها و تسهیل در جذب مدل‌های اشتراک‌گذاری داده‌ها شود. به حداقل رساندن و حفاظت از داده‌ها هرگز نباید با هدف پنهان کردن سوگیری یا اجتناب از پاسخگویی به کار گرفته شود، و این موارد باید بدون آسیب رساندن به حقوق حریم خصوصی مورد بررسی قرار گیرند.

نکته‌ی مهم این است که مسائل اخلاقی نه تنها هنگام پردازش داده‌های شخصی، بلکه زمانی که سیستم هوش مصنوعی از داده‌های غیرشخصی استفاده می‌کند (به عنوان مثال تعصب نژادی) هم مطرح شوند.

آیا می‌توانیم الزاماتی را به اعتمادسازی در داده‌ها بیفزاییم که می‌توانند به شرکت‌کنندگان در تحقیق کمک کنند تا درباره‌ی استفاده از داده‌ها مذاکره کنند؟

من فکر می‌کنم این بخش اگر توصیه‌های متناسبی را براساس انواع مختلف داده‌ها و مدل‌های داده‌ی مورداستفاده در سیستم‌های هوش مصنوعی ارائه کند، برای محققانِ هوش مصنوعی ارزش افزوده خواهد داشت؛ زیرا هر دسته از داده‌ها چالش‌های مختلفی را ایجاد می‌کنند: training data, model data, production data, knowledge data or analysis–to–data, data–to–analysis and data–and–analysis–to–lake models

الزامات عمومی اخلاقی برای حفظ حریم خصوصی و حاکمیت داده:

- سیستم‌های هوش مصنوعی *باید* داده‌های شخصی را به شیوه‌ای قانونی، منصفانه و شفاف پردازش کنند.
- اصول کمینه‌سازی و حفاظت از داده‌ها توسط طراحی و به طور پیش‌فرض *باید* در مدل‌های حاکمیت داده‌ی هوش مصنوعی ادغام شوند.
- *باید* تدابیر فنی و سازمانی مناسبی برای حفاظت از حقوق و آزادی‌های افراد هدف در داده (مانند انتصاب افسر حفاظت از داده‌ها، ناشناس‌سازی، نام مستعار، رمزگذاری، تجمیع) تنظیم شوند. برای جلوگیری از نفوذ و نشت داده‌ها *باید* تدابیر امنیتی قوی‌ای در نظر گرفته شود. انطباق با قانون امنیت سایبری و استانداردهای امنیت بین‌المللی می‌تواند مسیر امنی را برای پایبندی به اصول اخلاقی ارائه دهد.
- – داده‌ها باید به‌گونه‌ای جمع‌آوری، ذخیره و استفاده شوند که توسط انسان قابل‌بررسی باشند. تمام تحقیقاتی که با بودجه‌ی اتحادیه‌ی اروپا تامین می‌شوند باید با قوانین مربوطه و بالاترین استانداردهای اخلاقی مطابقت داشته باشند. این بدان معناست که همه ذینفعان Horizon Europe باید اصول مندرج در GDPR را به کار گیرند

3- انصاف (عدالت)

انصاف مستلزم آن است که همه‌ی مردم از حقوق و فرصت‌های اساسی یکسانی برخوردار باشند. این امر به نتایج یکسانی مثل این که افراد باید دارای ثروت یا موفقیت برابری در زندگی باشند نیاز ندارد. با این حال، نباید هیچ تبعیضی بر اساس جنبه‌های اساسی هویت که سلب‌شدنی نیستند، وجود داشته باشد. در حال حاضر قوانین مختلف تعدادی از این تبعیض‌ها مانند جنسیت، نژاد، سن، گرایش جنسی، منشاء ملی، مذهب، سلامت و ناتوانی را به رسمیت می‌شناسند. انصاف در عمل مستلزم آن است که دستورالعمل ها (روندها) به گونه‌ای طراحی نشده باشند که به طور خاص به افراد یا گروه‌های منفرد آسیب وارد کنند. انصاف ذاتی مستلزم آن است که هوش مصنوعی الگوهای تبعیضی را که به طور ناروا افراد و/یا گروه‌ها را به دلیل آسیب‌پذیری خاصشان تحت‌فشار قرار می‌دهند، پرورش ندهد.

انصاف را می‌توان با سیاست‌هایی که تنوع را ترویج می‌کنند نیز حمایت کرد. این سیاست‌ها، با ارزش‌گذاری مثبت به تفاوت‌های فردی، نه فقط ویژگی‌هایی مانند جنسیت و نژاد، بلکه شخصیت‌های متنوع، تجربیات، پیشینه‌های فرهنگی، سبک‌های شناختی و سایر متغیرهایی که بر دیدگاه‌های شخصی تأثیر می‌گذارند. حمایت از تنوع به معنای سازگاری با این تفاوت‌ها و حمایت از ترکیب متنوع تیم‌ها و سازمان‌ها است.

الزامات اخلاقی عمومی برای انصاف:

- **اجتناب از سوگیری الگوریتمی**: سیستم‌های هوش مصنوعی باید طوری طراحی شوند که از سوگیری در داده‌های ورودی، مدل‌سازی و طراحی الگوریتم جلوگیری کنند. سوگیری الگوریتمی یک نگرانی خاص است که به تکنیک‌های کاهشی خاصی نیاز دارد.
  پیشنهادهای پژوهشی باید مراحلی را برای اطمینان از اینکه داده‌های افراد معرف جمعیت هدف بوده و تنوع آنها را منعکس می‌کنند یا به اندازه‌ی کافی خنثی هستند، مشخص کنند.
- به طور مشابه، پیشنهادات پژوهشی باید به صراحت مستند کنند که سوگیری در داده‌های ورودی و در طراحی الگوریتمی ـ که می‌تواند باعث شود گروه‌های خاصی از افراد به اشتباه یا ناعادلانه نمایش داده شوند ـ، چگونه شناسایی شده و روش‌های اجتناب از آن را نیز ذکر کنند. این امر مستلزم در نظر گرفتن استنباط‌های به دست آمده توسط سیستم است که پتانسیل این را دارند که به‌طور ناعادلانه گروه‌های خاصی از مردم یا افراد مجزا را محروم کرده یا به روش‌های دیگر متحمل ضرر نمایند.
- **دسترس‌پذیری همگانی**: سیستم‌های هوش مصنوعی (در صورت لزوم) باید طوری طراحی شوند که برای انواع مختلف کاربران نهایی با توانایی‌های مختلف قابل استفاده باشند. پیشنهادهای پژوهشی برای توضیح چگونگی دستیابی به این امر، مثل انطباق با دستورالعمل‌های دسترسی مرتبط، تشویق می‌شوند. تا حد امکان، سیستم‌های هوش مصنوعی باید با ارائه‌ی همان سطح از عملکرد و مزایا به کاربران نهایی با توانایی‌ها، باورها، ترجیحات و علایق متفاوت، از تعصب عملکردی اجتناب کنند؛
- **تأثیرات منصفانه**: تأثیرات اجتماعی منفی احتمالی بر گروه‌های خاص، ـ مانند تأثیراتی غیر از آن‌هایی که ناشی از سوگیری الگوریتمی یا عدم دسترسی جهانی هستند ـ ممکن است در کوتاه‌مدت، میان‌مدت و بلندمدت رخ دهند، به‌ویژه اگر هوش مصنوعی از هدف اصلی خود منحرف شود. این تاثیرات باید کاهش یابند. سیستم هوش مصنوعی باید اطمینان حاصل کند که بر منافع گروه‌های مربوطه تأثیر منفی نمی‌گذارد. روش‌های شناسایی و کاهش اثرات منفی اجتماعی در میان‌مدت و بلندمدت باید به خوبی در طرح پیشنهادی پژوهش مستند شوند.

4- رفاه فردی، اجتماعی و زیست‌محیطی

رفاه فردی به این معنی است که افراد می‌توانند زندگی کاملی که در آن قادرند نیازها و خواسته‌های خود را با احترام متقابل دنبال کنند، داشته باشند. رفاه اجتماعی به شکوفایی جوامعی اشاره دارد که نهادهای اساسی مانند مراقبت‌های بهداشتی و سیاست در آنها به خوبی عمل می‌کنند و منابع درگیری اجتماعی به حداقل می‌رسد. رفاه زیست‌محیطی به عملکرد خوب اکوسیستم‌ها، پایداری و به حداقل رساندن تخریب محیط‌زیستی اشاره دارد.

سیستم‌های هوش مصنوعی نباید به رفاه فردی، اجتماعی یا زیست‌محیطی آسیب برساند، بلکه باید تلاش کنند تا سهم مثبتی در این اشکال رفاه داشته باشند. برای تحقق این هدف، شرکت‌کنندگان احتمالی در پژوهش، کاربران نهایی، افراد و جوامع آسیب‌دیده و ذینفعان مربوطه باید در مراحل اولیه شناسایی شوند تا امکان ارزیابی واقع‌بینانه از چگونگی بهبود رفاه آنها یا آسیب رسیدن به آنها توسط سیستم هوش مصنوعی فراهم شود. برای حمایت از رفاه و جلوگیری از آسیب، باید در طول توسعه، انتخاب‌های مستندشده‌ای انجام شود.

الزامات عمومی اخلاقی برای بهزیستی:

- سیستم های هوش مصنوعی باید همه‌ی کاربران نهایی و ذینفعان را در نظر بگیرند و نباید بهزیستی روانشناختی و عاطفی آنها را بی‌جهت یا ناعادلانه کاهش دهند.
- سیستم‌های هوش مصنوعی باید منافع و رفاه هر چه بیشتر افراد را تقویت کرده و آن‌ها را پیش ببرند.
- توسعه‌ی هوش مصنوعی *باید* به اصول پایداری زیست‌محیطی، هم در مورد خود سیستم و هم در مورد زنجیره‌ی تامینی که به آن متصل است توجه داشته باشد. در صورت لزوم، باید تلاش‌های مستندی برای در نظرگرفتن تأثیرات زیست‌محیطی کلی سیستم و اهداف توسعه‌ی پایدار وجود داشته باشند و در صورت نیاز، اقداماتی برای کاهش آنها انجام شوند. در مورد هوش مصنوعی تعبیه شده، این اقدامات باید شامل مواد استفاده‌شده و رویه‌های لغو کردن و ازرده خارج کردن آن باشد.
- سیستم‌های هوش مصنوعی که می‌توانند در حوزه‌ی رسانه، ارتباطات، سیاست، تجزیه‌وتحلیل اجتماعی، انجمن‌ها و خدمات آنلاین تجزیه‌وتحلیل رفتاری به کار روند، *باید* از نظر پتانسیل تأثیر منفی بر کیفیت ارتباطات، تعامل اجتماعی، اطلاعات، فرایندهای دموکراتیک و روابط اجتماعی (مثلاً با حمایت از گفتمان غیرمدنی، حفظ یا تقویت اخبار جعلی و دیپ‌فیک، تفکیک مردم در حباب‌های فیلتر و اتاق‌های پژواک، ایجاد روابط نامتقارن قدرت و وابستگی، و امکان دستکاری سیاسی در رای‌دهندگان) ارزیابی شوند. برای کاهش خطر چنین آسیب‌هایی باید اقدامات کاهشی انجام شوند.
- سیستم های هوش مصنوعی و رباتیک *نباید* ایمنی در محل کار را کاهش دهند. برنامه باید تأثیر احتمالی بر ایمنی محل کار، یکپارچگی کارکنان و استانداردهای انطباق، مانند استاندارد IEEE P1228 (استاندارد ایمنی نرم‌افزار) را در نظر بگیرد.

5- شفافیت، توضیح پذیری و اعتراض

شفافیت مستلزم آن است که هدف، ورودی‌ها و عملیات برنامه‌های هوش مصنوعی برای ذینفعان قابل‌درک باشند. بنابراین شفافیت بر *تمام* عناصر مرتبط با یک سیستم هوش مصنوعی تأثیر می‌گذارد: داده‌ها، سیستم و فرایندهایی که توسط آن طراحی و اجرا می‌شوند، زیرا ذینفعان باید بتوانند مفاهیم اصلی پشت آن را درک کنند (چگونه و برای چه هدفی این سیستم‌ها کار می‌کنند و تصمیم می‌گیرند).

ادعاهای مربوط به حقوق IP، محرمانه بودن یا اسرار تجاری تا زمانی که به‌طور مناسب حفظ شوند، نمی‌توانند مانع از شفافیت شوند، مثلا از طریق شفافیت انتخابی (مانند محرمانه بودن به اشخاص ثالث قابل‌اعتماد)، تعهدات فناوری یا محرمانه بودن. شفافیت برای تحقق سایر اصول یعنی احترام به عامل انسانی، حریم خصوصی و حاکمیت داده‌ها، پاسخگویی و نظارت ضروری است. بدون شفافیت (اطلاعات معنادار هدف، ورودی‌ها و عملیات برنامه‌های هوش مصنوعی)، خروجی‌های هوش مصنوعی قابل‌درک و چه بسا قابل‌بحث هم نیستند. این امر اصلاح خطاها و پیامدهای غیراخلاقی را غیرممکن می‌کند.

الزامات عمومی اخلاقی برای شفافیت:

- *باید* برای کاربران نهایی روشن شود که با یک سیستم هوش مصنوعی تعامل دارند (به ویژه برای سیستم‌هایی که ارتباطات انسانی را شبیه‌سازی می‌کنند، مانند چت‌باتها).
- هدف، قابلیت‌ها، محدودیت‌ها، مزایا و خطرات سیستم هوش مصنوعی و تصمیمات منتقل‌شده توسط آن (از جمله دستورالعمل‌های نحوه‌ی استفاده‌ی صحیح از سیستم) *باید* آشکارا به کاربران نهایی و سایر ذینفعان، اطلاع داده شوند.
- هنگام ساخت یک راه‌حل هوش مصنوعی، باید در نظر داشت که چه اقداماتی قابلیت ردیابی سیستم هوش مصنوعی را در طول کل چرخه‌ی عمر آن، از طراحی اولیه تا ارزیابی و ممیزی پس از استقرار یا در صورت مخالفت با استفاده از آن، ممکن می‌سازد.
- هر زمان لازم شد، پیشنهادی پژوهشی باید جزئیاتی در مورد اینکه چگونه تصمیمات اتخاذشده توسط سیستم برای کاربران قابل‌توضیح است ارائه دهد. در صورت امکان این امر باید شامل دلایلی باشد که چرا سیستم تصمیم خاصی گرفته است. تبیین‌پذیری یک الزام مربوط به سیستم‌هایی است که تصمیم‌گیری می‌کنند، توصیه‌هایی دارند یا اقداماتی را انجام می‌دهند که می‌توانند صدمات مهمی ایجاد کنند، بر حقوق فردی تأثیر بگذارند یا به طور قابل‌توجهی بر منافع فردی یا جمعی تأثیر بگذارند.
- فرایندهای طراحی و توسعه *باید* به تمام مسائل اخلاقی مربوطه، مانند حذف سوگیری از یک مجموعه داده بپردازند. فرایندهای توسعه (روش‌ها و ابزارها) باید سوابق تمام تصمیمات مربوطه را در این زمینه نگه دارند تا امکان ردیابی چگونگی برآورده شدن الزامات اخلاقی را فراهم آورند.

6- پاسخگویی با طراحی، کنترل و نظارت

مسئولیت‌پذیری برای هوش مصنوعی مستلزم آن است که بازیگران درگیر در توسعه یا اجرا، مسئولیت نحوه‌ی عملکرد این برنامه‌ها و پیامدهای ناشی از آن را برعهده بگیرند. البته، پاسخگویی مستلزم سطوح مشخصی از شفافیت و همچنین نظارت است. توسعه‌دهندگان یا اپراتورهای سیستم‌های هوش مصنوعی باید بتوانند توضیح دهند که چگونه و چرا یک سیستم ویژگی‌های خاصی را نشان می‌دهد یا نتایج خاصی را به نمایش می‌گذارد. نظارت انسانی مستلزم این است که بازیگران انسانی قادر به درک، نظارت و کنترل طراحی و عملکرد سیستم هوش مصنوعی باشند.

پاسخگویی به نظارت بستگی دارد: توسعه‌هندگان و اپراتورهای سیستم‌های هوش مصنوعی برای اینکه بتوانند مسئولیت را برعهده بگیرند و بر اساس آن عمل کنند، باید عملکرد و نتایج سیستم را درک و کنترل کنند. از این رو، برای اطمینان از پاسخگویی، توسعه‌دهندگان باید بتوانند توضیح دهند که چگونه و چرا یک سیستم ویژگی‌های خاصی را نشان می‌دهد.

الزامات عمومی اخلاقی برای مسئولیت‌پذیری و نظارت:

- *باید* مستند شود که چگونه اثرات نامطلوب اخلاقی و اجتماعی (مانند نتایج تبعیض‌آمیز، عدم شفافیت) سیستم می‌توانند شناسایی و متوقف شده و از تکرار مجدد آنها جلوگیری شود.
- سیستم‌های هوش مصنوعی *باید* اجازه‌ی نظارت و کنترل انسانی بر چرخه‌های تصمیم‌گیری و عملیات را بدهند، مگر اینکه دلایل قانع‌کننده‌ای ارائه شود که نشان دهد چنین نظارتی لازم نیست. چنین توجیهی باید توضیح دهد که چگونه انسان‌ها قادر خواهند بود تصمیمات اتخاذشده توسط سیستم را درک کنند و چه مکانیسم‌هایی برای لغو آنها وجود خواهد داشت.
- پیشنهاد پژوهشی در صورت لزوم و تا حدی که با نوع آن مطابقت داشته باشد (مثلاً پایه یا پیش‌رقابتی)، باید شامل ارزیابی خطرات اخلاقی احتمالی مرتبط با سیستم هوش مصنوعی پیشنهادی باشد. این باید شامل رویه‌های ارزیابی ریسک و اقدامات کاهشی (پس از استقرار) نیز باشد.
- این نکته باید در نظر گرفته شود که چگونه کاربران نهایی، افرادی که موضوع داده‌ها هستند و سایر اشخاص ثالث می‌توانند شکایات، نگرانی‌های اخلاقی یا رویدادهای نامطلوب را گزارش کنند و چگونه این موارد ارزیابی، رسیدگی و به طرف‌های مربوطه ارسال می‌شوند.
- به عنوان یک اصل کلی، همه‌ی سیستم‌های هوش مصنوعی باید توسط اشخاص ثالث مستقل قابل‌بازرسی باشند (به عنوان مثال، رویه‌ها و ابزارهای موجود تحت رویکرد XAI از بهترین عملکرد در این زمینه پشتیبانی می‌کنند). این موضوع به بازرسی تصمیمات خود سیستم محدود نمی‌شود، بلکه رویه‌ها و ابزارهای مورداستفاده در طول فرایند توسعه را نیز پوشش می‌دهد. در صورت لزوم، سیستم باید گزارش‌های قابل‌دسترسی از فرایندهای داخلی سیستم هوش مصنوعی را ایجاد کند.

بخش ۲-

نحوه‌ی به‌کارگیری اصول اخلاق در طراحی در توسعه‌ی هوش مصنوعی: مراحل عملی

اخلاق در طراحی رویکردی است که می‌توان آن را برای اطمینان از اینکه الزامات اخلاقی در طول توسعه‌ی سیستم یا تکنیک هوش مصنوعی به‌درستی موردتوجه واقع می‌شوند، به کار برد. این تنها رویکرد ممکن نیست. با این حال، این رویکرد اختصاصاً طراحی شده تا آنچه را که لازم است تا حد امکان روشن کند، وظایف خاصی را که باید انجام شوند توضیح دهد و به توسعه‌دهندگان کمک کند در هنگام توسعه‌ی یک سیستم هوش مصنوعی درباره‌ی اخلاق بیندیشند.

چرا اخلاق در طراحی؟

برای بسیاری از پروژه‌های هوش مصنوعی، مسائل اخلاقی مربوطه ممکن است تنها پس از استقرار سیستم شناسایی شوند، در حالی که برای پروژه‌های دیگر ممکن است در مرحله‌ی توسعه آشکار شوند. **اخلاق در طراحی** در وهله‌ی اول برای جلوگیری از بروز مسائل اخلاقی با پرداختن به آنها در مرحله‌ی توسعه، به جای تلاش برای رفع آنها در مراحل بعدی، در نظر گرفته شده است. این امر با استفاده‌ی فعالانه از اصول به عنوان الزامات سیستم حاصل می‌شود. علاوه بر این، از آنجایی که بسیاری از الزامات تنها در صورت ساخته‌شدن سیستم به روش‌های خاصی محقق می‌شوند، الزامات اخلاقی گاهی به جای خود سیستم هوش مصنوعی، در فرایندهای توسعه اعمال می‌شوند.

اخلاق در طراحی با یک مدل پنج لایه توصیف می‌شود؛ این مدل مشابه بسیاری از مدل‌های دیگر در علوم کامپیوتر است:سطوح بالاتر انتزاعی‌تر هستند و با پایین آمدن سطوح ویژگی‌ها زیاد می‌شوند.

شکل مدل ۵ لایه‌ی **اخلاق با طراحی**

فرایند توسعه

هدف **اخلاق در طراحی** این است که مردم را در حین توسعه‌ی یک سیستم، در خصوص نگرانی‌های اخلاقی بالقوه به فکر کردن و رسیدگی وادارد.

الزامات اخلاقی اصلی برای سیستم‌های هوش مصنوعی و رباتیک در بالا را می‌توان به صورت زیر خلاصه کرد:

- سیستم‌های هوش مصنوعی نباید بر استقلال، آزادی یا کرامت انسان تأثیر منفی بگذارد.

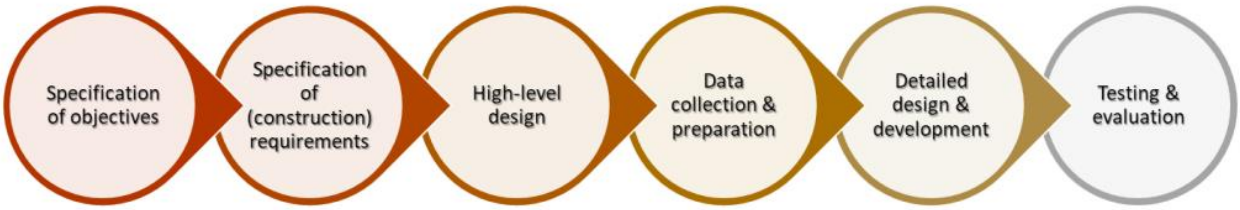
- سیستم‌های هوش مصنوعی نباید حق حفظ حریم خصوصی و حفاظت از داده‌های شخصی را نقض کنند. آنها *باید* از داده‌هایی استفاده کنند که ضروری، غیرجانبدارانه، نماینده و دقیق هستند.
- سیستم‌های هوش مصنوعی *باید* با دستور کار منصفانه و بدون تبعیض توسعه یابند.
- *باید* اقدامات لازم را انجام داد تا اطمینان حاصل شود که سیستم‌های هوش مصنوعی آسیب فردی، اجتماعی یا زیست‌محیطی ایجاد نمی‌کنند، به فناوری‌های مضر متکی نیستند و دیگران را تحت‌تأثیر قرار نمی‌دهند تا به گونه‌ای عمل کنند که باعث آسیب شده یا خودشان به صورت خزنده عمل کنند.
- سیستم‌های هوش مصنوعی باید تا حد امکان برای ذینفعان و کاربران نهایی خود شفاف باشند.
- نظارت و پاسخگویی انسانی برای اطمینان از انطباق با این اصول و رسیدگی به موارد عدم رعایت الزامی است.

اخلاق در طراحی بر این اساس استوار است که فرایندهای توسعه‌ی سیستم‌های هوش مصنوعی و روباتیک را می‌توان با استفاده از یک مدل عمومی شش مرحله‌ای توصیف کرد. این بخش مدل عمومی را تشریح می‌کند، سپس مراحل موردنیاز برای استفاده از آن را به‌منظور گنجاندن **اخلاق در طراحی** در فرایند توسعه بیان می‌کند.

با ترسیم روش توسعه‌ی خود طبق مدل عمومی مورد استفاده در اینجا، می‌توان الزامات اخلاقی مربوطه را تعیین کرد. پس از انجام این کار، **اخلاق در طراحی** به عنوان وظایف، اهداف، محدودیت ها و موارد مشابه در روش توسعه شما گنجانده می‌شود. بنابراین احتمال بروز نگرانی‌های اخلاقی به حداقل می‌رسد؛ زیرا هر مرحله در فرایند توسعه شامل اقداماتی برای جلوگیری از بروز آنها در وهله‌ی اول است.

با وجود اینکه شش مرحله از مدل عمومی در اینجا در قالب یک لیست ارائه شده‌اند، اما اینها لزوما یک فرایند متوالی نیستند. مراحل ممکن است تکراری باشند، مانند مدل V یا CRISP-DM، یا ممکن است از مدل‌های آبشاری منحرف شوند و یا اینکه افزایشی باشند، مانند Agile. در هر مورد،

تشخیص وظایف مشابه در روش توسعه‌ی موردعلاقه‌ی شخص، و سپس ترسیم وظایف مدل عمومی بر روی مراحل یا وظایف در رویکرد، نباید دشوار باشد.



شکل مدل عمومی برای توسعه‌ی هوش مصنوعی

شش وظیفه در مدل عمومی عبارتند از:

1. **مشخص کردن اهداف**: تعیین این که سیستم برای چه کاری است و باید قادر به انجام چه کاری باشد.
2. **تعیین الزامات**: توسعه‌ی الزامات فنی و غیرفنی برای ساخت سیستم، از جمله تعیین اولیه‌ی منابع مورد نیاز، ارزیابی اولیه‌ی ریسک و تجزیه‌وتحلیل هزینه و فایده، که منجر به نقشه‌ی طراحی می‌شود.
3. **طراحی سطح بالا**: توسعه‌ی یک معماری سطح بالا. این امرگاهی با توسعه‌ی یک مدل مفهومی پیش می‌رود.
4. **جمع آوری و آماده‌سازی داده**: جمع آوری، تأیید، تمیزکردن و یکپارچه‌سازی داده‌ها.
5. **طراحی و توسعه‌ی دقیق**: ساخت واقعی یک سیستم کاملاً کارآمد.
6. **آزمون و ارزیابی**: آزمودن و ارزیابی سیستم.

در ادامه هریک از این مراحل به تفصیل به همراه وظایف مربوطه که برای اطمینان از رعایت اصول اخلاقی انجام می‌شوند، بررسی می‌شوند.

ضمیمه‌ی I یک چک‌لیست نمونه، بر اساس الزامات و اصول اخلاقی فهرست‌شده ارائه می‌دهد. الزامات اخلاقی و راهنمایی‌ها باید تا حدی مطابق با نوع پژوهش پیشنهادی (از پایه تا قابل رقابت) اعمال شوند

۱. مشخص کردن اهداف

در این مرحله، اهداف سیستم براساس اصول اخلاقی و الزامات ارائه‌شده در قسمت ۱ ارزیابی می‌شوند.

نقض‌های اخلاقی احتمالی از نظر درجه‌ی جدیت متفاوت هستند. برخی از این تخلف‌ها ممکن است فقط فرضی، بسیار بعید یا با جدیت کمتری باشند. در این موارد، لزوماً نباید سیستم را رها کرد، بلکه باید برای جلوگیری از پیامدهای غیراخلاقی و کاهش زود هنگام و به‌موقع آنها، اقدامات مشخصی انجام داد.

توجه داشته باشید که برخی از مسائل اخلاقی ممکن است نگرانی‌های اخلاقی جدی‌تری را نسبت به سایرین ایجاد کنند. به عنوان مثال می‌توان این موارد را برشمرد:

- سیستم‌هایی که حقوق بشر را محدود می‌کنند،
- مردم را تحت تسلط خود می‌گیرند، فریب داده یا دستکاری می‌کنند،
- تمامیت جسمی یا ذهنی را نقض می‌کنند،
- موجب دلبستگی یا اعتیاد می‌شوند یا کنترل انسان را بر سر تصمیم‌های اصلی شخصی، اخلاقی یا سیاسی سلب می‌کنند.

برخی از این سیستم‌ها ممکن است طبق قوانین اتحادیه‌ی اروپا ممنوع تلقی شوند. از این رو، برخی از اهداف از نقطه نظر اخلاقی تحت هیچ شرایطی مجاز نیستند. اگر هدف یک سیستم اساساً با ارزش‌ها و الزامات اخلاقی ناسازگار باشد، پروژه نباید ادامه یابد. هر کاری که می‌تواند انجام شود، نباید لزوما انجام شود.

اهدافی که احتمالاً منجر به آسیب‌های جسمی، روانی یا مالی به افراد، آسیب‌های زیست‌محیطی یا آسیب به فرایندها و نهادهای اجتماعی شوند (مثلاً با حمایت از اطلاعات نادرست مردم) نیز از نظر اخلاقی قابل‌قبول نیستند. اگر پتانسیل قابل‌توجهی برای آسیب اجتماعی یا زیست‌محیطی ناشی از استفاده‌ی فناوری وجود داشته باشد، لازم است که ارزیابی اثرات اجتماعی و/ یا زیست‌محیطی انجام شود.

علاوه‌براین، در نظر بگیرید که آیا سیستم هوش مصنوعی ـ خود سیستم یا روش استفاده از آن ـ می‌تواند ناخواسته (یا عمدا) افراد را از نظر اجتماعی یا سیاسی در مضیقه قرار دهد، یا منجر به تبعیض ناعادلانه شود. در این صورت، اهداف سیستم باید اصلاح شده و اقدامات مناسبی برای کاهش این خطرات اتخاذ شوند.

توجه داشته باشید! الزامات اخلاقی در جهت شفافیت، پاسخگویی و نظارت معمولاً در مورد اهداف اعمال نمی‌شوند، بلکه در مورد معماری و طراحی دقیق سیستم به کار می‌روند. این الزامات معمولاً باید در این مرحله در نظرگرفته شوند تا مشخص شود که اهداف سیستم تا چه حد اجازه می‌دهند شفافیت و/ یا مسئولیت‌پذیری و نظارت لازم در سیستم به کار رود.

توجه:
هنگام ارزیابی اهداف، همواره احتمال سوءاستفاده‌ی عمدی یا تصادفی را در نظر داشته باشید. تا حد امکان اهداف سیستم را در جهت کاهش چنین احتمالی اصلاح کنید. چنانچه سوءاستفاده‌ی بالقوه قابل‌توجه است، ارزیابی ریسک را انجام دهید تا خطرات، عناصر طراحی مورد نیاز برای کاهش این سوءاستفاده و هر روشی که برای کاهش ریسک پس از استقرار و عملیاتی شدن لازم است،برآورد شوند.

دو دستورالعمل کلی زیر، برای مرحله‌ی مشخص کردن اهداف به کار می‌روند:

1. ارزیابی کنید که آیا اهداف پروژه‌ی طراحی با الزامات اخلاقی مربوطه مطابقت دارند یا خیر(به بخش ۱ و پیوست ۱ مراجعه کنید). توصیه می‌شود که یک متخصص اخلاق هوش مصنوعی (در صورت وجود) برای ارزیابی اهداف برای همکاری با اعضای تیم توسعه استخدام شود.
2. در هر دو مرحله‌ی تعیین اهداف و تعیین الزامات، ذینفعان خارجی را هم مدنظر قرار دهید. ممکن است ذینفعان از مسائل اخلاقی گسترده‌تری که در نتیجه‌ی استفاده از سیستم ناشی می‌شوند، آگاه باشند. در صورت امکان، باید با ذینفعان در مورد اینکه چه مسائل اخلاقی‌ای در معرض خطر هستند و چگونگی برخورد با این مسائل مشورت کرد. ذینفعان باید به‌طور مناسبی متنوع باشند (جنس، سن، قومیت، و غیره) و گروه‌های عمده‌ای را شامل شوند که به‌طور مستقیم یا غیرمستقیم تحت‌تأثیر سیستم قرار می گیرند. به این ترتیب طیف متنوعی از ایده‌ها و ترجیحات به انتخاب طراحی کمک خواهند کرد

2. تعیین الزامات (فنی و غیرفنی)

در طول این مرحله، الزامات توسعه، منابع و برنامه‌ها بر اساس الزامات اخلاقی ارزیابی می‌شوند.

کارکرد اولیه‌ی مرحله‌ی *تعیین الزامات*، رسیدن به برنامه‌ی توسعه‌ایست که شامل مشخصات طراحی سیستم، تعریف زیرساخت توسعه، تعیین منابع کارکنان مورد نیاز، تعیین نقاط عطف و سایر مهلت‌ها و غیره است.

اکثر سازمان‌ها دارای مجموعه‌ی استانداردی از ابزارهای توسعه هستند که برای همه پروژه‌ها استفاده می‌شوند. ساختارها و رویه‌های سازمانی و مدیریتی ـ به مانند روش‌شناسی توسعه ـ معمولاً با این ابزارها تنظیم می‌شوند. با این وجود، نمی‌توان فرض کرد که هر ابزار، فرایند یا عنصر سازمانی از **اخلاق در طراحی** پشتیبانی می‌کند. برخی از الزامات اخلاقی مشکلات جدیدی را در طول توسعه ایجاد می‌کنند. برای مثال، دیگر صرف اصلاح مجموعه داده‌ها برای سوگیری کافی نیست، بلکه توسعه‌دهندگان باید مستند کنند که این کار چگونه انجام شده است. در نتیجه، الزامات مربوط به نظارت و حسابرسی انسانی ممکن است نیاز به مستندسازی بسیاری از فرایندهای داخلی را به میزانی بیشتر از آنچه قبلاً وجود داشته، تحمیل کند. بنابراین باید در نظر گرفت که ابزارها و ساختارهای سازمانی مورداستفاده در پروژه‌های که قبلاً بر روش‌های توسعه متکی بودند، ممکن است نیاز به اصلاح داشته باشند.

در برخی موارد، ابزارهای توسعه‌ی مناسبی که امکان برآوردن الزامات اخلاقی را فراهم کنند، ممکن است به آسانی در دسترس نباشند. با این حال، لازم است که در جستجوی ابزار مناسب بسیار دقیق بود. الزامات تعیین‌شده در اینجا منحصر به Horizon Europe نیستند، بلکه خواسته‌های رایج بسیاری از پروژه‌های هوش مصنوعی هستند. در نتیجه، ابزارهای لازم برای برآوردن الزامات اخلاقی مربوطه به سرعت در حال توسعه هستند.

همه‌ی الزامات اخلاقی ممکن است مرتبط نباشند! درجه‌ای که ناتوانی فنی در برآوردن یک الزام اخلاقی می تواند مانع از پیشبرد پروژه شود، به الزامات اخلاقی خاص موردنظر و عملکرد سیستم بستگی دارد. به‌عنوان مثال، سیستمی که وام‌های شخصی را تأیید می‌کند باید بتواند هر تصمیم فردی را در قالبی خوانا برای انسان توضیح دهد، زیرا افراد تحت تأثیر تصمیمات آن قرار می‌گیرند. در مقابل، سیستمی که قرار ملاقات‌ها را مدیریت می‌کند ممکن است تنها تأثیر محدودی بر زندگی افراد داشته باشد، بنابراین به شفافیت کمتری نیاز دارد. در مواردی که

اعتقاد براین است که برآوردن یک الزام اخلاقی مربوطه از نظر فنی غیرممکن است، اهمیت این الزام عاملی در تعیین تأیید خواهد بود.

دستورالعمل‌های کلی زیر تا حدی که با نوع تحقیق پیشنهادی مطابقت داشته باشند (از پایه تا پیش‌رقابتی)، برای مرحله‌ی تعیین الزامات اعمال می‌شوند:

1. مشخصات طراحی پیشنهادی، محدودیت‌ها، منابع انتخاب‌شده و زیرساخت‌ها باید از نظر سازگاری با الزامات اخلاقی ارزیابی شوند. به عنوان مثال، برخی از تکنیک‌های یادگیری عمیق ممکن است برای شفافیت انتخاب خوبی نباشند. مناسب است که این مشخصات و محدودیت‌ها را در برابر الزامات اخلاقی قسمت ۱ بررسی کنید و درصورت لزوم برای اطمینان از تناسب، آنها را اصلاح کنید.
2. هنگامی که برنامه‌ی یک طراحیِ کامل تولید شد (و تا حدودی با نوع پژوهش پیشنهادی مطابقت داشت)، به‌منظور برآورد خطرات اخلاقی خاص در توسعه، استقرار و استفاده از سیستم، یک ارزیابی تاثیرات و ریسک اخلاقی توصیه می‌شود. که باید شامل ارزیابی خطرات مرتبط با استفاده‌های ناخواسته و پیامدهای سیستم باشد. برای جلوگیری از خطرات یا کاهش خطرات اجتناب‌ناپذیر باید برنامه‌ریزی‌هایی صورت گیرد. این ارزیابی ریسک باید در مراحل منظم فرایند توسعه، از طریق دسترسی بیشتر اطلاعات به‌روز شود. نیاز است که برای ارزیابی ریسک اخلاقی برنامه‌ریزی و بودجه‌بندی شود. توصیه می‌شود که در صورت امکان، این کار توسط یک متخصص اخلاق با سابقه‌ی حرفه‌ای مناسب انجام شود. ارزیابی باید بر اساس ماهیت پروژه، شدت خطرات اخلاقی و بودجه‌ی کلی برای توسعه‌ی پروژه مقیاس شود.
3. توصیه می‌شود که در این مرحله یک طرح پیاده‌سازی **اخلاق در طراحی** (EbD) تهیه شود که در آن تمام گام‌های بعدی موردنیاز برای گنجاندن EbD در فرایند توسعه، مشخص شده و بازیگران مسئول در انجام و نظارت بر وظایف تعیین شوند. در صورت لزوم، این طرح اجرایی باید ریسک اخلاقی و ارزیابی تأثیر را در صورتی که تکمیل شده باشد، در بر بگیرد.
4. علاوه‌براین، توصیه می‌شود که ارزیابی شامل یک معماری انطباق اخلاقی در زیرساخت توسعه و مجموعه‌ی ساختارها و رویه‌های سازمانی هم باشد. این معماریِ

انطباق اخلاقی باید بر ابزارها و فرایندها در سطح توسعه‌دهنده تمرکز کند، اما به مکانیسم‌هایی برای ارتباطات خارجی از سوی کاربران نهایی و سایر ذینفعان در طول آزمایش و ارزیابی هم نیاز دارد. یک راه ساده‌ی رو به جلو این است که معماری انطباق اخلاقی را با انطباق قانونی مرتبط کنیم.

5. توصیه می‌شود که ارزیابی، یک مدل حاکمیت اخلاقی که ساختارهای سازمانی برای اداره‌ی فرایند EbD (مثلاً کمیته‌های بررسی اخلاقی/ مشاور اخلاقی/ حسابرس اخلاقی) را در بر می‌گیرد، در خود داشته باشد.

3. طراحی سطح بالا

طراحی سطح بالا به توسعه‌ی معماری کلی سیستم مربوط می‌شود.

الزامات اخلاقی باید دقیقاً مانند سایر الزامات سیستم در نظر گرفته شوند. طراحی باید شامل عملکردی باشد که به وسیله آن از الزامات اخلاقی به‌صورت برنامه‌ریزی‌شده حمایت شود، مانند نگهداری گزارش‌های دستکاری داده‌های داخلی توسط سیستم. الزامات شفافیت و نظارت انسانی معمولاً به ویژگی‌هایی فراتر از آنچه برای دستیابی به هدف سیستم موردنیاز است، احتیاج دارد.

دستورالعمل‌های کلی زیر برای مرحله‌ی طراحی سطح بالا، تا حدی که با نوع پژوهش پیشنهادی (از پایه تا پیش رقابتی) مطابقت داشته باشند و هر زمان که قابل‌اجرا باشد اعمال می‌گردند:

1. بررسی کنید که آیا طراحی به یک رابط مبتنی بر اصول طراحی انسان‌محور ـ که فرصت‌های معناداری را برای انتخاب انسان ایجاد می‌کنند ـ مجالی می‌دهد یا خیر.
2. ویژگی‌ها و عملکردهایی را طراحی کنید که موجب می‌شوند قابلیت‌ها و اهداف سیستم به‌طور آشکار به کاربران و هر کسی که ممکن است تحت‌تأثیر آن قرار گیرد، منتقل شوند.
3. مکانیسم‌هایی را طراحی کنید تا مردم بدانند چه زمانی در معرض تصمیمات سیستم قرار می‌گیرند. این مکانیزم‌ها می‌توانند رویه‌های عملیاتی مورداستفاده پس از استقرار هم باشند.
4. مطمئن شوید که بعد از استقرار سیستم، هیچ جنبه‌ای از سیستم هوش مصنوعی که ممکن است بدون اطلاع قبلی، با یک انسان اشتباه گرفته شود، وجود نداشته باشد. به خاطر داشته باشید که بسیاری از افراد ممکن است درک درستی از هوش مصنوعی نداشته باشند و ناآگاهانه تصور کنند که با یک فرد در حال تعامل هستند. برای مثال، حتی زمانی که در مورد چت‌بات‌ها این مورد رعایت شود، افرادی که نمی‌دانند اصطلاح چت‌بات به چه معناست، ممکن است آن را با انسان اشتباه بگیرند.
5. بررسی کنید که طراحی از الزامات اخلاقی برای حفظ حریم خصوصی، حفاظت از داده‌های شخصی، و حاکمیت داده پشتیبانی می‌کند. اطمینان حاصل کنید که فرایندها، رویه‌ها و ابزارهای توسعه، داده‌های شخصی را به‌گونه‌ای که حقوق اساسی افراد را نقض

کند، در معرض نمایش قرار نمی‌دهند. برای مثال، گزارش‌های خطا ممکن است شامل داده‌های شخصی‌ای باشند که هنگام مواجهه با یک باگ به آنها دسترسی پیدا می‌کنند. به ویژه مهم است که اطمینان حاصل شود که توسعه‌دهندگان به اطلاعات شخصی قابل‌شناسایی دسترسی ندارند؛ مگر در موارد کاملاً ضروری و دستورالعمل‌های منطبق با GDPR.

6. یک فرایند رسمی را برای تضمین انتخاب داده‌ها برای یک سیستم منصفانه، دقیق و بی‌طرفانه وضع نمایید. برای ارزیابی اولیه‌ی منابع داده قبل از وارد شدن به سیستم، برنامه‌ریزی کنید. یک مکانیسم قابل‌بازبینی را طراحی کنید که نحوه‌ی انتخاب داده، جمع‌آوری، ذخیره و استفاده از داده‌ها را ثبت کند (هم فرایند توسعه و هم استفاده پس از عملیاتی‌شدن را پوشش دهد). نمی‌توان فرض کرد که داده‌های به دست آمده همان داده‌های مورد نظر هستند. برای مثال، مجموعه‌ی داده‌ها ممکن است ناقص باشند یا روش‌های وارد کردن داده‌ها ممکن است آن‌ها را به روش‌های غیرمنتظره‌ای تغییر دهند. این مورد شامل فرایندهای رسمی طراحی برای بررسی و تصحیح سوگیری یا خطاها پس از وارد کردن هر داده هم می‌شود.

7. ارزیابی کنید که آیا این سیستم می‌تواند هرگونه آسیب فیزیکی به افراد، حیوانات، اموال یا محیط‌زیست وارد کند یا خیر. اگر این امکان وجود دارد، ویژگی‌های طراحی را برای به حداقل رساندن خطر و/ یا شیوع و شدت آسیب احتمالی بگنجانید (به عنوان مثال اگر سیستم می‌تواند به دستورات صوتی پاسخ دهد، باید دستورات صوتی «ایست اضطراری» را در طراحی بگنجانید).

8. تمام تأثیرات منفی احتمالی بر گروه‌های مربوطه را در نظر بگیرید، از جمله تأثیراتی غیر از آن‌هایی که ناشی از سوگیری الگوریتمی یا عدم دسترسی جهانی است و اقداماتی را برای کاهش هرگونه تأثیر منفی بالقوه (مانند انگ زدن، تبعیض) طراحی کنید.

9. سیستم‌های هوش مصنوعی مرتبط با رسانه‌ها، ارتباطات، سیاست، تحلیل‌های اجتماعی و جوامع آنلاین باید از نظر پتانسیل تأثیر منفی بر کیفیت ارتباطات، تعامل اجتماعی، اطلاعات، فرایندهای دموکراتیک و روابط اجتماعی ارزیابی شوند. در صورت لزوم، پیشنهادهای پژوهشی باید اقدامات کاهشی برای کم کردن خطر چنین آسیب‌هایی را به تفصیل شرح دهند.

10. هرزمان که مرتبط باشد، باید تأثیرات زیست‌محیطی سیستم را در نظر بگیرید و در صورت نیاز اقداماتی را برای کاهش اثرات منفی انجام دهید. در مرحله‌ی تعیین اهداف، باید یک ارزیابی اولیه‌ی زیست‌محیطی انجام شود. بعد از تکمیل طراحی سطح بالای سیستم، این ارزیابی باید به تفصیل انجام شود تا نشان دهد که چگونه سیستم به روشی سازگار با محیط‌زیست ساخته خواهد شد.

11. در صورت لزوم، در نظر گرفتن تأثیر احتمالی بر ایمنی و مطابقت با استانداردهای مناسب مانند IEEE P1228 (استاندارد ایمنی نرم‌افزار) را نمایش دهید و آنها را در ویژگی‌های طراحی لحاظ کنید تا از ایمنی و انطباق، مطمئن شوید.

12. در نظر داشته باشید که چگونه تصمیمات سیستم برای کاربران و سایر ذینفعان قابل‌توضیح هستند. در صورت امکان این توضیح باید شامل دلایلی باشد که چرا سیستم تصمیم خاصی گرفته است. سیستم (یا کسانی که آن را مستقر می‌کنند) همیشه باید مکانیزمی داشته باشد که به‌وسیله‌ی آن توضیح دهد که تصمیم چه بوده و از چه داده‌ها / ویژگی‌هایی برای تصمیم‌گیری استفاده شده است. توضیح‌پذیری به‌ویژه با سیستم‌هایی که تصمیم می‌گیرند، توصیه می‌کنند، یا اقداماتی را انجام می‌دهند که ممکن است آسیب قابل‌توجهی ایجاد کنند، بر حقوق فردی تأثیر بگذارند، یا بر منافع فردی یا جمعی تأثیر زیادی بگذارند، مرتبط است.

13. روشهایی را طراحی و ابزارهایی را انتخاب و پیکربندی کنید که بتوانند فرایندهای توسعه را به‌گونه‌ای مستند کنند که افراد بتوانند تصمیمات فرایندهای طراحی و توسعه را درک و ارزیابی کنند. پیشنهادهای پژوهشی همچنین باید توضیح دهند که انسان‌ها چگونه قادر خواهند بود تصمیمات سیستم را درک کنند و چه مکانیسم‌هایی برای نادیده گرفتن آنها توسط انسان طراحی خواهد شد. برای این مستندات یک رویکرد لایه‌ای که طیف وسیعی از جزئیات فنی را ارائه می‌کند، توصیه می‌شود. این رویکرد از مرورهای اولیه، مانند خلاصه‌های اجرایی، آغاز می‌شود و طرح‌واره‌های دقیق و سایر مدل‌های فنی را دربر می‌گیرد. به این ترتیب، می‌توان اسنادی متناسب با سطح تخصص و دغدغه‌های خاص افراد را در اختیار آنها نهاد.

14. مطمئن شوید که طراحی شامل مکانیسم‌هایی است که سیستم هوش مصنوعی تصمیمات خود را به‌وسیله آن‌ها ثبت می‌کند تا بتوانند در معرض بازبینی و

پاسخگویی انسانی قرارگیرند. اگر سوژه‌های داده یا کاربران، نهایی رفتار سیستم را زیر سوال ببرند، این بررسی به‌عنوان بخشی از بررسی حاکمیت اخلاقی داخلی یا ممیزی خارجی، می‌تواند از طریق ممیزی پس از استقرار صورت گیرد.

15. تا حد امکان و متناسب، سازوکارهایی برای پاسخگویی، نظارت انسانی و ممیزی خارجی پس از استقرار طراحی کنید. این سازوکارها ممکن است به عملکرد اضافه‌ای در داخل سیستم، تنها برای گزارش فعالیت‌های داخلی که هیچ نقشی در عملکرد سیستم ندارند نیاز داشته باشند. مکانیسم‌های نظارت پس از استقرار باید به کارهای نظارتی انجام‌شده در طول توسعه دسترسی داشته باشند.

16. در طراحی یک سیستم اخلاقی برای مستندات، اطمینان حاصل کنید که برای شناسایی مسائل اخلاقی و حل آنها؛ قابل‌ردیابی و دارای توضیح کافی باشند.

17. یک رژیم آزمایشی را در طراحی بگنجانید که بتواند بررسی کند که آیا عملیات داخلی سیستم با الزامات اخلاقی مطابقت دارد یا خیر. این رژیم ممکن است به تغییراتی در نحوه‌ی دستیابی به عملکرد در سیستم نیاز داشته باشد تا امکان آزمایش و اقدامات اصلاحی مناسب را فراهم آورد. پیشنهاد پژوهشی باید جزئیاتی درباره‌ی چگونگی شناسایی، توقف و جلوگیری از تکرار مجدد اثرات نامطلوب اخلاقی و اجتماعی سیستم ارائه کند.

18. در صورت امکان و ارتباط، اطمینان حاصل کنید که آیا فرایندی وجود دارد که از طریق آن هم کارکنان داخلی و هم اشخاص ثالث بتوانند آسیب‌پذیری‌ها، خطرات یا سوگیری‌های احتمالی سیستم را، هم در طول فرایند توسعه و هم پس از استقرار گزارش کنند یا خیر.

19. مطمئن شوید که سیستم به‌گونه‌ای طراحی شده است که امکان حسابرسی اخلاقی خارجی را فراهم می‌کند تا از وجود پاسخگویی اطمینان حاصل شود. اگر مطمئن نیستید، از روش‌های حسابرسی اخلاقی موجود کمک بگیرید.

20. ویژگی‌ها و عملکردها را به‌گونه‌ای طراحی کنید که کاربران نهایی از قابلیت‌ها و محدودیت‌های سیستم هوش مصنوعی پس از استقرار نیز آگاه باشند تا از مسائل مربوط به انحراف عملکرد، اتوماسیون یا تعصبات تعبیر و ترجمه ها جلوگیری شود.

12. اطمینان حاصل کنید که توسعه‌دهندگان برای توسعه‌ی آگاهی و شیوه‌های پاسخگویی آتی (از جمله دانش مرتبط با چارچوب قانونی قابل‌اجرا در سیستم) آموزش کافی دیده‌اند.

علاوه‌براین، پیشنهادات پژوهشی برای توضیح اینکه چگونه طراحی با قابلیت دسترسی جهانی سازگار است، تشویق می‌شوند؛ (مانند نشان دادن انطباق با دستورالعمل‌های دسترسی مرتبط).

این توضیح می‌تواند شامل ارزیابی دسترسی رابط و سایر نقاط تماس باشد. بررسی طراحی رابط اولیه و سایر نقاط تماس نشان خواهد داد که آیا یک رویکرد به یک‌اندازه برای همه‌ی کاربران اعمال می‌شود یا خیر. در این صورت، ممکن است لازم باشد طرح را اصلاح کنید یا یک توجیه رسمی برای آن تهیه کنید.

۴. جمع‌آوری و آماده‌سازی داده‌ها

جمع‌آوری و آماده‌سازی داده‌ها از نظر اخلاقی یک مرحله‌ی بسیار مهم است.

باید فرض شود که هر داده‌ی جمع آوری‌شده مغرضانه، انحرافی یا ناقص است تا زمانی که خلاف آن ثابت شود. انصاف و دقت دغدغه‌های اصلی در اینجا هستند. به طور کلی، داده‌های جمع‌آوری‌شده از فعالیت‌های انسانی در هر جامعه، مانند ارتباطات نوشتاری یا الگوهای شغلی، ممکن است منعکس‌کننده‌ی سوگیری‌های آن جامعه باشند. بنابراین باید به طور فعال نشان داده شود که داده‌ها قبل از اینکه بتوان به آنها اعتماد کرد، دقیق، معرف یا خنثی هستند.

علاوه‌براین، آماده‌سازی داده‌ها ممکن است به مسائل اخلاقی نیز منجر شوند. گام‌هایی باید برداشته شود تا اطمینان حاصل شود که یادگیری آزمایشی و/ یا دستکاری الگوریتمی سوگیری‌های جدیدی ایجاد نمی‌کنند یا منجر به مسائل اخلاقی دیگر (مانند بی‌نام‌سازی) نمی‌شوند. مشکلات اغلب در مواردی ایجاد می‌شوند که آزمایش به طور دقیق، استفاده در دنیای واقعی را پس از استقرار منعکس نمی‌کنند، مانند ایجاد اتوماسیون یا سوگیری‌های ترجمه‌ای. به عنوان مثال، بسیاری از سیستم‌های تشخیص چهره به دلیل آزمایش روی جمعیت‌های قفقازی، عملکرد ضعیفی نسبت به افراد با پوست تیره‌تر دارند.

به خاطر داشته باشید که هنگام پردازش داده‌های شخصی، تمام تحقیقاتی که توسط اتحادیه‌ی اروپا تأمین مالی می‌شوند، باید با قوانین بین‌المللی، اروپایی و ملی و با بالاترین استانداردهای اخلاقی مطابقت داشته باشند. این بدان معناست که ذینفعان اتحادیه‌ی اروپا باید اصول GDPR را اعمال کنند مگر اینکه به استانداردهای بالاتر حفاظت از داده ها ملزم باشند. اگر از سازمان‌ها/ ارائه‌دهندگان خدمات خارجی برای ذخیره‌سازی/تحلیل/جمع‌آوری داده‌ها استفاده می‌کنید، مطمئن شوید که این موارد نیز با الزامات حفاظت از داده‌ها مطابقت دارند.

تا حدی که با نوع پژوهش پیشنهادی (از پایه تا پیش‌رقابتی) مطابقت داشته باشد و هر زمان که قابل اجرا باشد، دستورالعمل‌های کلی زیر برای مرحله‌ی جمع‌آوری و آماده‌سازی داده‌ها اعمال می‌شوند:

1. ارزیابی کنید که چگونه عملیات در هر فرایند ممکن است الزامات اخلاقی و/ یا حفاظت از داده‌ها را نقض کنند. تغییرات لازم را به دنبال آن ایجاد کنید. اگر تغییرات مناسب امکان پذیر نباشند، ممکن است نیاز به تغییر اهداف طراحی باشد.
2. اطمینان حاصل کنید که پردازش داده‌های شخصی با مقررات حفاظت از داده‌های عمومی (GDPR) و سایر قوانین ملی و اتحادیه‌ی اروپا مطابقت دارد. یک خط مشی حفاظت از داده‌های خاص تهیه کنید که جزئیات نحوه‌ی انطباق پروژه با الزامات حفاظت از داده‌ها را شرح دهد. اکیداً توصیه می‌شود با افسر حفاظت از داده‌های خود یا فردی با تخصص مربوط به حفاظت از داده‌ها (در صورت وجود) مشورت کنید. توجه داشته باشید که هر زمان که سیستم شما در حال پردازش داده‌های شخصی است، باید از اصل کمینه‌سازی داده‌ها پیروی کند. این بدان معناست که شما باید اطمینان حاصل کنید که فقط داده‌هایی که مرتبط، کافی و محدود به آنچه کاملاً ضروری هستند، توسط سیستم شما پردازش می‌شوند. شما همچنین باید اصول حفاظت از داده‌ها را به طور پیش‌فرض رعایت کنید و از حقوق و آزادی‌های افراد موضوع داده محافظت کنید.
3. مراحلی را مشخص کنید تا مطمئن شوید که داده‌های مربوط به افراد معرف هستند و تنوع آنها را منعکس می‌کنند یا به اندازه‌ی کافی خنثی هستند. به طور مشابه، شما باید نحوه‌ی جلوگیری از خطاها در داده‌های ورودی و طراحی الگوریتمی که می‌تواند گروه‌های خاصی از افراد را نادرست یا ناعادلانه نشان دهد، مستند کنید.
4. مطمئن شوید که داده‌های ورودی، آموزشی و خروجی همگی برای سوگیری ورودی (سوگیری در داده‌هایی که منجر به نمایش ناعادلانه، غیرنماینده یا تعصب‌آمیز افراد و گروه‌ها می‌شوند) تجزیه‌وتحلیل می‌شوند. به‌ویژه، تا حد امکان تأیید کنید که داده‌های شخصی و گروهی توضیح‌دهنده‌ی تنوع در جنسیت، نژاد، سن، گرایش جنسی، منشأ ملی، مذهب، سلامت و ناتوانی و سایر مقوله‌های اجتماعی مرتبط با این وظیفه هستند و فرضیات تعصب‌آمیز، کلیشه‌ای یا تبعیض‌آمیز دیگر در مورد افراد در این دسته‌ها قرار نمی‌گیرند. در جایی که مشخص شود سوگیری امکان پذیر است، مکانیسم‌هایی را برای اجتناب یا اصلاح آن ایجاد کنید. یعنی معیارهایی که براساس آن داده‌ها انتخاب می‌شوند را اصلاح کنید یا به‌گونه‌آی برنامه‌ریزی کنید که مجموعه‌ی

داده‌ها پس از قرارگرفتن در سیستم اصلاح شوند. الزامات شفافیت و نظارت مستلزم مستند بودن چنین اصلاحاتی هستند.

5. داده‌های آموزشی خود را تجزیه‌وتحلیل کنید و اطمینان حاصل کنید که مرتبط و نماینده هستند. ارزیابی سوگیری را به‌طور رسمی برای داده‌های واردشده به سیستم انجام دهید. فرض نکنید که داده‌های وارد شده به سیستم بی‌طرفانه هستند؛ آن‌ها را تا حد امکان با توجه به خطرات ناشی از هوش مصنوعی پیش‌بینی‌شده برای حقوق و آزادی‌های اساسی آزمایش کنید. تنوع و نمایندگی کاربران در داده‌ها، آزمایش برای جمعیت‌های خاص یا موارد استفاده‌ی مشکل‌ساز را بسنجید. مطمئن شوید که داده‌های یک گروه جمعیتی برای نشان دادن دیگری استفاده نمی‌شود، مگر اینکه به طور موجهی نماینده باشند. پتانسیل سوگیری مضری را که در مرحله‌ی آماده‌سازی داده‌ها معرفی می‌شوند، ارزیابی کنید، مانند حذف ناخواسته‌ی داده‌های مربوط به یک گروه اقلیت. اقداماتی را برای کاهش چنین خطری انجام دهید. مطمئن شوید که در صورت امکان، توانایی بازگشت به هر حالتی که سیستم در آن بوده است وجود داشته باشد تا مشخص یا پیش‌بینی شود که سیستم در زمان x چه کاری انجام می‌داد و در صورت امکان، تعیین کنید که از کدام داده‌ی آموزشی استفاده شده است.

6. اطمینان حاصل کنید که کارکنان درگیر، در مورد چارچوب قانونی قابل‌اجرا و نگرانی‌های اخلاقی آموزش مناسبی دیده‌اند.

۵. طراحی و توسعه‌ی دقیق

در مرحله‌ی طراحی و توسعه‌ی دقیق، یک سیستم کاملاً کارآمد ساخته می‌شود. اقدامات دربرگیرنده‌ی الزامات اخلاقی به وظایف مختلف طراحی تفصیلی و همچنین به زیرساخت توسعه (ابزارها، روش‌ها، رویه‌ها و هر چیز دیگری که ممکن است دقیقاً بر نحوه‌ی ساخت چیزی تأثیر بگذارد) اضافه می‌شوند.

تا حد زیادی، این مرحله شامل افزودن جزئیات بیشتر به الزامات اخلاقی سیستم و طراحی و اجرای معماری توسعه‌ی اخلاقی است. **اخلاق در طراحی** خواستار رسیدگی به مسائل اخلاقی در مرحله‌ی توسعه است، بنابراین فرایندهای توسعه‌ی موجود باید از این فعالیت پشتیبانی کنند. برای یکپارچه‌سازی الزامات اخلاقی، اطمینان حاصل کنید که دستورالعمل‌های اخلاقی به همه‌ی توسعه‌دهندگان و مهندسان ابلاغ می‌شود، و هر جا که نیاز به تصمیم‌گیری مرتبط داشته باشند، طراحی در برابر نقشه‌ی آنها برای پیاده‌سازی اخلاقی ارزیابی می‌شود. مسائلی که ممکن است به‌ویژه در این مرحله مرتبط باشد، مواردی هستند که به شفافیت، حریم خصوصی و پاسخگویی مربوط می‌شوند. آموزش مناسب و افزایش آگاهی می‌توانند کمک‌کننده باشد.

تا حدی که مطابق با نوع تحقیق پیشنهادی (از پایه تا پیش رقابتی) و در صورت لزوم، دستورالعمل‌های کلی زیر در مرحله‌ی طراحی و توسعه‌ی دقیق اعمال می‌شوند:

1. در صورت ایجاد داده‌های شخصی جدید (به‌عنوان مثال، از طریق تخمین داده‌های از دست رفته، تولید ویژگی‌های مشتق‌شده و سوابق جدید) یا جمع‌آوری یا ایجاد مجموعه‌داده‌های جدید، مطمئن شوید که همه‌ی داده‌های شخصی یا مجموعه‌داده‌های جدید ایجادشده، سطح حفاظتی یکسانی در مقایسه با سطحی که قبلاً داده‌های شخصی جمع‌آوری یا نگهداری می‌شدند، داشته باشند. داده‌های شخصی استنباط‌شده نباید قادر به تغییر خروجی مجموعه‌داده باشند، یا به‌طور قابل‌توجهی بر حقوق افراد موضوع داده تأثیر بگذارند. داده‌های شخصی تازه‌ایجادشده نباید گمراه‌کننده باشد.
2. اطمینان حاصل کنید که هیچ داده‌ی شخصی جدیدی در طول توسعه‌ی سیستم یا در طول استفاده‌ی منظم از سیستم جمع‌آوری یا ایجاد نشده است، مگر در موارد ضروری. اگر داده‌های شخصی جدید جمع‌آوری یا ایجاد شدند، مطمئن شوید که

مکانیسم‌هایی وجود دارند که محدودیت‌هایی را برای دسترسی یا استفاده از افراد ایجاد می‌کنند تا از افراد موضوع داده محافظت کنند.

3. اطمینان حاصل کنید که فرایندهایی برای محافظت از کیفیت و یکپارچگی همه داده‌های مربوطه، وجود دارند؛ از جمله فرایندهایی برای تأیید اینکه مجموعه‌داده‌ها به خطر نیفتاده یا هک نشده‌اند. اگر بر کیفیت منابع داده‌ی خارجی استفاده‌شده کنترل دارید، بسنجید که تا چه حد می‌توانید کیفیت آنها را تأیید کنید.

4. مطمئن شوید که نقش‌ها و مسئولیت‌ها برای اجرای الزامات اخلاقی و معماری مشخص هستند و همه‌ی کارکنان مربوطه الزامات را درک می‌کنند.

5. اطمینان حاصل کنید که مکانیسم‌های نظارتی برای پردازش داده‌ها (از جمله محدود کردن دسترسی تنها به پرسنل مناسب، مکانیسم‌هایی برای ثبت دسترسی به داده‌ها و ایجاد تغییرات) وجود دارند. در توسعه‌ی اخلاقیات خود با رویکرد طراحی، در نظر بگیرید که حتی زمانی که داده‌ها ناشناس هستند، ممکن است باز هم امکان ناشناس‌زدایی آنها وجود داشته باشد.

6. اگر پایگاه داده‌های تشکیل‌شده توسط داده‌های شخصی و غیرشخصی را با هم ترکیب می‌کنید، توجه داشته باشید که بدون توجه به نسبت هر نوع داده در مجموعه‌داده‌ی حاصله، باید به‌عنوان داده‌های شخصی پردازش شوند. مطمئن باشید که فرایند تعبیه‌شده‌ای وجود دارد که به افراد اجازه می‌دهد به داده‌های خود دسترسی داشته باشند و آن‌ها را از سیستم حذف کنند و/ یا خطاهای موجود در داده‌های خود را تصحیح کنند. بنابراین باید اقداماتی را انجام دهید تا تضمین کنید که افراد می‌توانند به داده‌های شخصی خود دسترسی داشته باشند، آن هم به روشی که از حریم خصوصی افراد دیگر محافظت کند.

7. اقدامات فنی و سازمانی را به‌منظور حفاظت از داده‌ها توسط طراحی و به‌طور پیش‌فرض (مانند روش‌های *حریم خصوصی در طراحی*) ایجاد نمایید؛ مانند رمزگذاری، نام مستعار، تجمیع، ناشناس‌سازی و به حداقل رساندن داده‌ها.

8. در هر سیستمی داده‌ها می‌توانند دستکاری شوند، آسیب ببینند، از دست بروند یا به‌طور نامناسبی در معرض دید قرار بگیرند. فرایندهایی را طراحی کنید تا کاهش مداوم کیفیت داده‌ها را قبل از استفاده توسط سیستم بررسی کنند. این باید شامل اقداماتی

برای جلوگیری از فساد خارجی و کاهش دادن فساد بی‌صدا و سایر اشکال فساد داده در سطح پایین باشد.

9. سوگیری‌های الگوریتمی را در مرحله‌ی توسعه به دقیق بررسی کنید. داده‌ها را می‌توان به روشی جانبدارانه پردازش کرد، بنابراین الگوریتم‌ها باید برای این مورد بررسی شوند. (به عنوان مثال با استفاده از روش‌های ارزیابی *ضد واقعیت*)

10. اطمینان حاصل کنید که طراحی رابط، اصول دسترسی جهانی را رعایت می‌کند و از معرفی سوگیری‌های عملکردی در مرحله‌ی توسعه‌ی دقیق که عملکرد سیستم را برای کاربران نهایی مختلف نابرابر می‌سازد، اجتناب کنید.

11. برای اطمینان از قابلیت ردیابی و پاسخگویی مورد نیاز، باید اقداماتی بر طبق روش‌های زیر انجام شوند:

- روش‌های مورداستفاده برای طراحی و توسعه‌ی سیستم‌ها، مانند مدل‌های ساخته‌شده، روش‌های آموزشی، اینکه کدام داده‌ها جمع‌آوری و انتخاب شده‌اند و چگونگی انجام آن.
- روش‌های مورداستفاده برای آزمایش و اعتبارسنجی سیستم‌ها، مانند سناریوها یا موارد به‌کاررفته برای آزمایش و اعتبارسنجی؛ داده‌های مورداستفاده برای آزمایش و اعتبارسنجی؛ نتایج سیستم (نتایج یا تصمیمات اتخاذشده توسط سیستم)؛ سایر تصمیمات ممکن که از موارد مختلف ناشی می‌شوند، مثلا برای سایر زیرگروه‌های کاربران.
- یک توالی از روش‌های فنی برای اطمینان از قابلیت ردیابی (مانند رمزگذاری ابرداده برای استخراج و ردیابی آن در صورت لزوم). باید راهی وجود داشته باشد که بدانیم داده‌ها از کجا آمده‌اند و توانایی ساختن نحوه‌ی ارتباط قطعات مختلف داده با یکدیگر مهیا باشد.

12. مطمئن شوید که کد به طور فعال در برنامه‌ی نرم‌افزار و در اسناد جانبی مناسب توضیح داده و مستند شده‌اند (مطابق با زبان(ها) و روش). اطمینان حاصل کنید که مستندات برای برنامه‌نویسان همکار قابل‌درک و در دسترس آنهاست. توافق‌نامه‌های محرمانه می‌توانند برنگرانی‌های مربوط به IP، محرمانه بودن یا اسرار تجاری کمک کنند.

13. مطمئن شوید که می‌دانید تصمیمات و نتایج اتخاذشده توسط سیستم تا چه حد قابل‌درک هستند، مثلا آیا به گردش کار داخلی مدل دسترسی دارید یا خیر.

14. از روش‌ها و ابزارهای رسمی برای اطمینان از توضیح‌پذیری در جاهای ممکن استفاده کنید و اگر برای سیستم خاصی طراحی‌شده، مطلوب باشد، مانند رویکردهای XAI یا *شفافیت در طراحی* و مستندات برنامه‌ریزی‌شده، مانند کارت‌های مدل.

15. اینکه سیستم می‌تواند اطلاعات نادرست یا گمراه‌کننده را به مردم ارائه دهد را در نظر داشته باشید، و در صورت لزوم، اقداماتی را برای جلوگیری از آن انجام دهید یا برای به حداقل رساندن این خطر الزامات طراحی را اضافه کنید. در برخی موارد، پس از عملیاتی شدن سیستم، این خطر افزایش می‌یابد. اگر چنین است، اسناد، عملکرد یا مکانیسم‌های دیگری را اضافه کنید تا خطر اطلاعات نادرست به حداقل برسد.

16. در نظر بگیرید که آیا سیستم به طور اجتناب‌ناپذیر داده‌ها را دستکاری می‌کند یا تصمیماتی می‌گیرد که برای انسان قابل‌ردیابی یا درک نباشند. اگر چنین است، الزامات طراحی را اضافه کنید تا عملیات داده‌ها را تا حد امکان در معرض بررسی دقیق قرار دهید، و/یا برای توضیح اینکه چرا عملیات داده‌ها نمی‌توانند و نباید حسابرسی شوند توجیه رسمی آماده کنید. توجه کنید که نگرانی‌های مربوط به مالکیت معنوی کافی نیستند. رژیم‌های آزمایش جعبه سیاه و «آزمایش ردیابی» می‌توانند برای ارزیابی خارجی عملیات داده‌های داخلی استفاده شوند.

17. در معماری توسعه، ابزارها و مکانیسم‌هایی را ایجاد کنید تا تمام اطلاعات مربوط به ارزیابی اخلاقی، مانند منبع دیتابیس و ماهیت مدل‌های مورداستفاده ثبت شوند. اطمینان حاصل کنید که کارکنان برای استفاده از ابزارها و مکانیسم‌های مربوطه آموزش دیده و تشویق می‌شوند.

18. در صورت امکان، مکانیسم‌هایی را برای ارزیابی و در صورت لزوم اقدام درباره‌ی نگرانی‌های مطرح‌شده توسط کارکنان و اشخاص ثالث ایجاد کنید. اطمینان حاصل کنید که چنین اقداماتی قبل از ادامه‌ی توسعه (یا استقرار) انجام می‌شوند.

19. ممکن است لازم باشد کنترل‌های ممیزی عمیقاً در سیستم تعبیه شوند. مطمئن شوید که کنترل‌های ممیزی برای گزارش عملکرد و ثبت تصمیمات توسط سیستم ساخته شده‌اند.

20. سند اخلاقی الزامات پروژه را اصلاح و تکمیل کنید. این احتمالاً یک فرایند تکراری است. تا آنجا که ممکن است، هرگونه تصمیم اتخاذشده در مورد نحوه‌ی مطابقت سیستم با الزامات اخلاقی خود را ثبت کنید.

21. در صورت امکان از شیوه‌های مصرف انرژی پایدار پیروی کنید. به ویژه، تصمیماتی که توسط سیستم گرفته می‌شوند و بر دنیای غیرانسانی تأثیر می‌گذارند، باید به دقت در نظر گرفته شوند.

6. آزمون و ارزشیابی

به عنوان بخشی از مرحله‌ی آزمایش و ارزیابی، یک ارزیابی اخلاقی انجام می‌شود تا ببیند آیا سیستم الزامات اخلاقی خود را برآورده می‌کند یا خیر.

ممکن است سیستم به الزامات عملکردی خود دست یابد، اما الزامات اخلاقی را برآورده نکند. در این صورت نمی‌توان سیستم را با موفقیت تکمیل کرد. با این حال، تمام هدف **اخلاق در طراحی** اجتناب از چنین نتیجه‌ای است. در صورتی که **اخلاق در طراحی** به طور جدی اعمال شود، باید از مسائل اخلاقی در این مرحله از فرایند توسعه جلوگیری کند.

به عنوان بخشی از مرحله‌ی آزمایش و ارزیابی، باید یک چک‌لیست از الزامات اخلاقی پروژه را برای طراحی یک رژیم آزمایشی داشته باشید تا بتوانید انطباق اخلاقی سیستم را بررسی کنید. بسیار بعید است که هر رژیم آزمون استانداردی را برای همه الزامات اخلاقی سیستم در نظر بگیرد، بنابراین انتخاب روش آزمون در اینجا مهم است. این آزمایش را برای تعیین اینکه آیا سیستم تمام الزامات اخلاقی خود را برآورده می‌کند، اجرا کنید. انحراف از ویژگی‌های اخلاقی مورد نظر سیستم را به همان اندازه‌ی باگ آن جدی بگیرید و کارهای اصلاحی را تا زمانی که سیستم الزامات اخلاقی خود را برآورده کند انجام دهید. اکیداً توصیه می‌شود که در این مرحله، ذینفعان را هم درگیر کنید. از ذینفعان بپرسید که آیا از اینکه سیستم به اندازه‌ی کافی ارزش‌ها و نیازهای آنها را در نظر می‌گیرد (که احتمالاً قبلاً در ابتدای پروژه مورد بحث قرار گرفته است) راضی هستند یا خیر و در صورت نیاز اقدامات لازم را انجام دهید.

فرایند آزمایش باید شامل درک رفتار سیستم توسط کاربران نهایی باشد. نمی‌توان تصور کرد که دیگران خروجی سیستم را به همان روش توسعه‌دهندگان درک می‌کنند. درک افراد آسیب‌دیده در مورد هدف سیستم، اینکه چه کسی یا چه چیزی ممکن است از آن سود ببرد و (مهمتر از همه) محدودیت‌های آن را آزمایش کنید.

علاوه بر بررسی انطباق با الزامات اخلاقی و درگیر کردن ذینفعان، و میزان تطابق با نوع پژوهش پیشنهادی (از پایه تا پیش رقابتی)، مراحل زیر را نیز در نظر بگیرید:

1. بررسی کنید که آیا کاربران درک می‌کنند که با یک عامل غیرانسانی در حال تعامل هستند و/ یا اینکه یک تصمیم، محتوا، توصیه یا خروجی نتیجه‌ی یک تصمیم الگوریتم، در موقعیت‌هایی هستند که انجام ندادن آنها برای کاربران فریبنده، گمراه‌کننده یا مضر هستند.
2. اطمینان حاصل کنید که کنترل‌های ممیزی برای بررسی عملکرد، ثبت تصمیمات اتخاذشده در مورد هدف و عملکرد سیستم (مانند گزارش در مورد تأثیرات به طور کلی، نه فقط وقوع تأثیرات منفی) در سیستم تعبیه شده‌اند. مطمئن شوید که مکانیسم‌هایی برای اطلاع‌رسانی به کاربران سازمانی و کاربران نهایی (اگر مستقیماً با آنها سروکار دارند) در مورد دلایل نتایج سیستم ایجاد شده‌اند.
3. اطمینان حاصل کنید که اطلاعات در مورد قابلیت‌ها و محدودیت‌های سیستم به طور واضح، قابل‌درک و فعالانه به ذینفعان، کاربران و سایر افراد آسیب‌دیده منتقل می‌شوند، و انتظارات واقع‌بینانه را ممکن می‌سازند.
4. در صورت امکان، مطمئن شوید که فرآیندهای عملی برای اشخاص ثالث (مانند تامین‌کنندگان، مصرف‌کنندگان، توزیع‌کنندگان/فروشندگان) یا کارگران وجود دارد تا آسیب‌پذیری‌ها، خطرات یا سوگیری‌های احتمالی در سیستم را گزارش کنند. اطمینان حاصل کنید که مکانیسم‌هایی برای بررسی و اقدام بر اساس چنین گزارش‌هایی وجود دارند.
5. فرایندهایی را برای به دست آوردن و در نظرگرفتن بازخورد کاربران و مکانیسم‌های موجود برای تطبیق سیستم در پاسخ مناسب ایجاد کنید.
6. اطمینان حاصل کنید که به کاربران و ذینفعان توضیحاتی داده می‌شود که بتوانند بفهمند چرا سیستم انتخاب خاصی را انجام داده است که منجر به نتیجه‌ی خاصی در طول آزمایش شده است تا بتوانند آن را به طور دقیق ارزیابی کنند.
7. برای کمک به توسعه‌ی شیوه‌های پاسخگویی (از جمله چارچوب قانونی قابل‌اعمال برای سیستم)، آموزش به کاربران را توسعه و ارائه دهید.
8. به طور رسمی سعی کنید پیامدها/موارد بیرونی عملیات سیستم را پیش‌بینی کنید

۳. بخش سوم: استقرار و استفاده‌ی اخلاقی

این بخش از دستورالعمل‌ها در مورد چهار روش اصلی برای استفاده از سیستم‌های هوش مصنوعی در پروژه‌های تحقیقاتی اعمال می‌شوند: مدیریت پروژه، اکتساب، اجرا و نظارت. با در نظر گرفتن تمام این فرایندها، این دستورالعمل‌های اخلاقی از افت عملکرد در هنگام استقرار هوش مصنوعی پس از پایان پروژه جلوگیری خواهند کرد.

– **مدیریت پروژه** به برنامه‌ریزی یک پروژه‌ی تحقیقاتی جدید، معمولاً در یک طرح پروژه، و مدیریت فعالیت‌های برنامه‌ریزی‌شده در طول پروژه اشاره دارد. برای این منظور، این بخش به مراحلی می‌پردازد که باید در برنامه‌ریزی و مدیریت پروژه لحاظ شوند تا اطمینان حاصل شود که مسائل اخلاقی به طور صحیح در استقرار و استفاده از یک سیستم هوش مصنوعی اعمال می‌شوند.

– **اکتساب**، به فرایند دستیابی به یک سیستم هوش مصنوعی گفته می‌شود که در خود سیستم توسعه نیافته باشد. یک سازمان در قبال وضعیت اخلاقی هر سیستم هوش مصنوعی که از آن استفاده می‌کند مسئول است؛ حتی اگر آن سیستم توسط دیگری ساخته شده باشد. اگر سیستم در خود پروژه توسعه یافته باشد، دستورالعمل‌های اخلاق در طراحی؟ مربوط به بخش ۲ این سند اعمال می‌شوند.

– **استقرار و پیاده‌سازی** به فرایند استقرار سیستم هوش مصنوعی در محیط کاربری و همچنین برنامه‌ریزی و اجرای تغییرات در سازمان اشاره دارد. نحوه‌ی استقرار یک سیستم ممکن است ویژگی‌های اخلاقی سیستم را تغییر دهد. بنابراین، پیاده‌سازی باید تضمین کند که سیستم به الزامات اخلاقی خود ادامه می‌دهد.

– منظور از **نظارت**، فرایند بررسی انطباق سیستم با الزامات و توسعه و اجرای طرح‌هایی برای بهبود عملکرد است. ویژگی‌های اخلاقی کامل یک سیستم ممکن است تا زمانی که سیستم «به‌صورت کاملا طبیعی» مستقر نشده باشد، آشکار نشوند. در نتیجه، همه سیستم‌های هوش مصنوعی نیازمند نظارت مداوم اخلاقی و در صورت لزوم، تنظیم هستند. این نظارت، معمولاً با روش ممیزی که در حال تبدیل شدن به یک الزام قانونی رایج است، انجام می‌شود.

برنامه‌ریزی و مدیریت پروژه

- برنامه‌ریزی برای وظایف اخلاقی مرتبط با کاربرد: در بودجه‌بندی و برنامه‌ریزی، مسائل اخلاقی بالقوه‌ای که ممکن است رخ دهند را در نظر بگیرید. هر زمان که مناسب باشد، انتصاب مشاور یا مشاوران اخلاقی مستقل، متخصص در زمینه‌ی اخلاق فناوری‌های جدید و نوظهور و حفاظت از داده‌های شخصی را در نظر بگیرید.
- نقش‌ها و رویه‌هایی را برای اجرای دستورالعمل‌های اخلاقی و نظارت بر اجرای آنها تعریف کنید. این می‌تواند شامل تشکیل یک افسر اخلاق هوش مصنوعی یا تیمی باشد که مسئولیت اجرای دستورالعمل‌های اخلاقی یا نظارت بر اجرای آنها را بر عهده دارند. نباید تصور کرد که هر کسی که در طول توسعه، انطباق اخلاقی را مدیریت کرد، مرجع مناسب برای این نقش است.
- اطمینان حاصل کنید که اهدافی که سیستم برای آنها استفاده می‌شوند، الزامات طراحی و انتخاب منابع با الزامات اخلاقی ارائه‌شده در مراحل اهداف و الزامات اخلاق در طراحی؟ مطابقت دارند.

اکتساب

- اگر یک سیستم هوش مصنوعی خارجی، به‌عنوان یک راه‌حل خارجی به کار رود، سیستمی را انتخاب کنید که بیشترین توانایی را برای برآورده کردن الزامات اخلاقی مشخص‌شده در قسمت ۱ دارد. اگر سیستم هوش مصنوعی به صورت سفارشی توسط یک توسعه‌دهنده‌ی خارجی ساخته شده است، اولویت را به توسعه‌دهنده‌ای بدهید که از رویکرد اخلاق در طراحی؟ استفاده می‌کند یا مایل است به الزامات اخلاقی مندرج در این راهنما پایبند باشد. تا حد ممکن، پایبندی خود به این الزامات را تأیید کنید. حداقل، فروشنده باید بتواند بسیاری از اطلاعات موردنیاز را ارائه دهد. از آنجایی که اخلاق در طراحی؟ نیازمند شفافیت و نظارت انسانی است، ممکن است در ابتدا از فروشنده بخواهیم مکانیسم‌های نظارت اخلاقی توسعه‌دهنده را توضیح داده و نمونه‌هایی از اسناد شفافیت آنها را نشان دهد. بدون شفافیت کافی، تعیین انطباق اخلاقی سیستم و یا ارائه‌ی شواهد برای پاسخگویی ممکن نخواهد بود.

- در صورت انتخاب توسعه‌ی داخلی، از روش اخلاق در طراحی؟ که در قسمت دوم این سند توضیح داده شده است، پیروی کنید و تأیید کنید که سیستم حاصل از الزامات اخلاقی ذکرشده پیروی می‌کند.
- اطمینان حاصل کنید که هر داده‌ای که قبل از استقرار برای سیستم جمع‌آوری و آماده شده است، مطابق با دستورالعمل‌های جمع‌آوری و آماده‌سازی داده‌ها است (بخش دوم، بخش ۴).
- در صورت لزوم، ارزیابی ریسک اخلاقی و ارزیابی تاثیر به منظور بررسی ریسک‌های اخلاقی خاص در استفاده از سیستم انجام شود. برای کاهش خطرات اخلاقی شناسایی‌شده، لازم است که اقدامات کاهشی انجام شوند. می‌توان این را در بالای ارزیابی اخلاقی اولیه، در طراحی اولیه‌ی پروژه ایجاد کرد. با این حال، مهم است که بدانیم که ممکن است با تکامل سیستم در طول توسعه، مسائل جدیدی به وجود بیایند که در این مورد بیشتر یاد خواهید گرفت.

استقرار و اجرا

- اگر ابزارهای هوش مصنوعی مستقرشده حاوی داده‌های شخصی هستند، آنها را حذف کنید، مگر اینکه بتوانید توجیه کنید که چرا حذف کردن احتمالاً دستیابی به اهداف خاص آن را غیرممکن کرده یا به‌طور جدی آسیب می‌رساند.
- برنامه‌ها و خط‌مشی‌هایی را ایجاد و اجرا کنید که با الزامات اخلاقی سیستم انطباق عملیاتی دارند (به بخش ۱ مراجعه کنید).
- داده‌ها، دسترسی‌ها، سیاست‌ها و رویه‌های مدیریت ریسک به‌کاررفته در سیستم را با هدف به کار بستن الزامات اخلاقی به‌روزرسانی کنید.
- در آموزش بهره‌برداری و استفاده از سیستم، سیاست‌های اخلاقی جدید را لحاظ کنید. به جنبه‌های اخلاقی درون ارتباطات در مورد راه‌اندازی سیستم توجه کنید.
- بر اجرای دستورالعمل‌های اخلاقی سیستم در طول مرحله‌ی اجرا نظارت کنید، مسائل و خطرات را شناسایی نمایید و در صورت نیاز تنظیمات لازم را اجرا کنید.

نظارت

- مکانیسم‌هایی را برای اطمینان از اینکه کاربران نهایی مطابق با خط‌مشی‌های کاربران از سیستم استفاده می‌کنند و نسبت به مسائل اخلاقی در عملیات و استفاده هوشیار هستند ایجاد نمایید و با کارکنان ارشد در مورد موضوعاتی که از نظر اخلاقی مشکل‌ساز یا مبهم هستند، مشورت کنید.
- اطمینان حاصل کنید که اهداف و معیارهای نظارتی، برای انطباق با الزامات اخلاقی وجود دارند. اگر پایش‌ها اینطور نشان دادند که انطباق کمتر از هدف است، به‌صورت دوره‌ای بر انطباق نظارت کنید و پیشنهادهایی را برای بهبود ارائه دهید.
- هنگام استقرار سیستم هوش مصنوعی در طول عمر پروژه، اطمینان حاصل کنید که ذینفعان، کاربران و افراد سیستم هوش مصنوعی، کانال‌های ارتباطی «شکایت اخلاقی» خود را دارند که از طریق آن به شما در مورد نگرانی‌های اخلاقیشان در صورت بروز هشدار دهند. به خاطر داشته باشید که مشکلات اخلاقی اغلب به این دلیل رخ می‌دهند که یک سیستم بر افرادی تأثیر می‌گذارد که هرگز انتظار نمی‌رفت تحت‌تأثیر سیستم قرار بگیرند. هوش مصنوعی منطبق با اخلاقیات برای اطمینان از حفاظت فنی و سازمانی مناسب، با هدف جلوگیری از افت عملکرد و مداخله در صورت بروز خطرات طراحی شده است.

چک لیست پیوست ۱: تعیین اهداف در مقابل الزامات اخلاقی

این چک‌لیست یک ابزار پشتیبانی است و فهرستی جامع از همه‌ی الزامات اخلاقی که می‌توانند برای توسعه‌ی هر سیستم هوش مصنوعی خاص قابل‌اجرا باشد را تشکیل نمی‌دهد. این باید همراه با دستورالعمل‌های فعلی بخش ۱ تا ۳ مورد استفاده قرار گیرد و تا حدی که مطابق با نوع سیستم هوش مصنوعی و پژوهش پیشنهادی (از پایه تا پیش‌رقابتی) هستند، اعمال شوند.

تعیین اهداف در مقابل الزامات اخلاقی	بله	خیر (چگونه خطرات احتمالی کاهش خواهند یافت؟)
احترام به عامل انسانی		
کاربران نهایی و دیگر افراد تحت‌تاثیر هوش مصنوعی از توانایی تصمیم‌گیری در مورد زندگی خود محروم نیستند، آزادی‌های اولیه از آنها سلب شده است.		
کاربران نهایی و دیگر افراد تحت‌تاثیر هوش مصنوعی، تابع، مجبور، فریب‌خورده، دستکاری‌شده، عینی‌شده یا غیرانسانی‌شده نمی‌شوند، و همچنین وابستگی یا اعتیاد نسبت به سیستم و عملیات آن تحریک نمی‌گردد.		
این سیستم به‌طور مستقل در مورد موضوعات حیاتی که معمولاً توسط انسان‌ها از طریق انتخاب‌های شخصی آزاد یا مشورت‌های جمعی تصمیم‌گیری می‌شوند، یا به طور مشابه بر افراد تأثیر می‌گذارد، تصمیم‌گیری نمی‌کنند.		
این سیستم به گونه‌ای طراحی شده است که به اپراتورهای سیستم و تا حد امکان به کاربران نهایی توانایی کنترل، هدایت و مداخله در عملیات اساسی سیستم (در صورت لزوم) را بدهد.		
حریم خصوصی و حاکمیت داده		

این سیستم داده‌ها را مطابق با الزامات قانونی، انصاف و شفافیت تنظیم‌شده در چارچوب قانونی حفاظت از داده‌های ملی و اتحادیه‌ی اروپا و انتظارات معقول سوژه‌های داده پردازش می‌کند.		
تدابیر فنی و سازمانی برای حفاظت از حقوق سوژه‌های داده (از طریق اقداماتی مانند ناشناس‌سازی، نام مستعار، رمزگذاری و تجمیع) وجود دارند.		
اقدامات امنیتی برای جلوگیری از نقض و نشت داده‌ها (مانند مکانیسم‌های ثبت دسترسی به داده‌ها و اصلاح آنها) وجود دارند.		
انصاف		
این سیستم برای جلوگیری از سوگیری الگوریتمی در داده‌های ورودی، مدل‌سازی و طراحی الگوریتم طراحی شده است. این سیستم برای جلوگیری از سوگیری تاریخی و انتخابی در جمع‌آوری داده‌ها، سوگیری نمایشی و عملی در آموزش الگوریتمی، سوگیری تجمیع و ارزیابی در مدل‌سازی و سوگیری اتوماسیون در استقرار طراحی شده است.		
این سیستم به گونه‌ای طراحی شده است که می‌توان از انواع مختلف کاربران نهایی با توانایی‌های مختلف (در صورت امکان/مرتبط) استفاده کرد.		
این سیستم تأثیرات اجتماعی منفی بر گروه‌های مربوطه ندارد، تأثیراتی به جز آنهایی که ناشی از سوگیری الگوریتمی یا عدم دسترسی جهانی هستند.		
رفاه فردی، اجتماعی و زیست‌محیطی		
سیستم هوش مصنوعی رفاه همه‌ی ذینفعان را در نظر می‌گیرد و آن را بی‌جهت یا ناعادلانه کم/تضعیف نمی‌کند.		

سیستم هوش مصنوعی به اصول پایداری زیست‌محیطی توجه دارد، هم در مورد خود سیستم و هم در مورد زنجیره‌ی تامینی که به آن متصل است (در صورت لزوم)		
این سیستم هوش مصنوعی پتانسیل تأثیر منفی بر کیفیت ارتباطات، تعامل اجتماعی، اطلاعات، فرایندهای دموکراتیک و روابط اجتماعی (در صورت لزوم) را ندارد.		
این سیستم ایمنی و یکپارچگی در محیط کار را کاهش نمی‌دهد و با مقررات مربوط به بهداشت و ایمنی و استخدام مطابقت دارد.		
شفافیت		
کاربران نهایی می‌دانند که در حال تعامل با یک سیستم هوش مصنوعی هستند.		
هدف، قابلیت‌ها، محدودیت‌ها، مزایا و خطرات سیستم هوش مصنوعی و تصمیم‌های منتقل‌شده به همراه پیامدهای احتمالی آن به طور آشکار به اطلاع کاربران نهایی و سایر ذینفعان می‌رسد.		
افراد می توانند بررسی، پرس‌وجو، اعتراض کنند و به دنبال تغییر یا اعتراض به فعالیت‌های هوش مصنوعی یا روباتیک باشند (در صورت لزوم)		
این سیستم هوش مصنوعی قابلیت ردیابی در کل چرخه‌ی عمر خود، از طراحی اولیه تا ارزیابی و ممیزی پس از استقرار را امکان‌پذیر می‌کند.		
این سیستم جزئیاتی را در مورد چگونگی گرفتن تصمیمات و دلایلی که این تصمیمات بر اساس آنها گرفته شده‌اند (در صورت امکان و مرتبط بودن) ارائه می‌دهد.		
این سیستم سوابق تصمیمات گرفته‌شده را نگه می‌دارد (در صورت لزوم)		

مسئولیت‌پذیری و نظارت		
این سیستم جزئیاتی را درباره‌ی چگونگی شناسایی، توقف و جلوگیری از تکرار مجدد اثرات نامطلوب اخلاقی و اجتماعی ارائه می‌دهد.		
این سیستم هوش مصنوعی اجازه‌ی نظارت انسانی د ر طول چرخه‌ی عمر پروژه را / با توجه به چرخه‌ی تصمیم‌گیری و عملیات آنها (د ر صورت لزوم) می‌دهد.		

پایان.

درود

اگر شما این راهنما را مطالعه کرده اید، احتمالا پژوهشگر، طراح، توسعه دهنده ی کسب و کار و در کل دغدغه مندی در حوزه فناوری هستید که تصمیم گیری درست در حوزه هوش مصنوعی چالش اوست . من **ندا صالحی پژوهشگر طراحی، مشاور و مدرس طراحی اخلاقی** مفتخرم اذعان کنم که با یاری دو فرهیخته ی گرامی راهنمای حاضر؛ ترجمه، ویراستاری و صفحه بندی شد و در اختیار اهل فن قرار گرفت. قطعا بدون ارزیابی دقیق و موشکافانه خانم مهندس **مریم رضائی** حاجیده ترجمه ی این راهنما پر نقص و از زبان مادری خوانندگان محترم دور می بود و حقیقتا بدون تلاش خالصانه و نگاه تخصصی همکار ارجمند خانم **مهتاب خواره**، این کتابچه زیبا رصفحه آرایی نمی یافت. ضمن تشکر و ابراز ارادت از همراهی دو دوست عزیز، امیدوارم نکات با ارزش برآمده از این پژوهش راهگشایی هرچند اندک در جهت توسعه پروژه های هوش مصنوعی اخلاق مدار باشد.

با احترام

ندا صالحی

موسس استودیو طراحی انسانی

راه های تماس در اینستاگرام
@human_design_studio
@m_hajideh
@mahtab_khavareh

www.humandesignstudio.ir
humandesignst@gmail.com

حامی باش

از اینکه وقت ارزشمند خود را صرف مطالعه ی این متن کردید، از شما سپاسگزاریم. امید است این راهنما بتواند درک بهتری از مفهوم «اخلاق در هوش مصنوعی» ایجاد کند.

چنانچه این محتوا برایتان مفید بوده، در صورت تمایل میتوانید با پرداخت مبلغ دلخواه به شماره کارت زیر از این طرح **حمایت** کنید. هر کمک شما، دلگرمی بزرگیست برای ما تا مطالب بیشتری را ترجمه کرده و در اختیار جامعه ی طراحی و کسب و کار قرار دهیم.

پیشاپیش از حمایت های شما متشکریم.

شماره کارت بانک اقتصاد نوین – ندا صالحی رزوه

6274-1211-9742-3536