مرکز ملی فضای مجازی
پژوهشگاه فضای مجازی

عصر فضای مجازی
صد و ششم

دلالت های اخلاقی و اجتماعی الگوریتم‌ها، داده‌ها و هوش مصنوعی: یک نقشه راه برای پژوهش

Ethical and societal implications of algorithms, data, and artifcial intelligence: a roadmap for research

عصر فضای مجازی

گزارش شماره ۱۰۶

مرداد ۱۴۰۱

مرکز ملی فضای مجازی
پژوهشگاه فضای مجازی

دلالت های اخلاقی واجتماعیِ الگوریتم ها، داده ها و هوش مصنوعی: یک نقشه راه برای پژوهش

تهیه شده در پژوهشگاه فضای مجازی
(گروه مطالعات فرهنگی و اجتماعی)

ناظر علمي: یحیی شعبانی (پژوهشگر گروه مطالعات فرهنگی و اجتماعی)، امیررضا باقرپور شیرازی (مدیر گروه مطالعات فرهنگی و اجتماعی)

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

فهرست

بخش چهارم

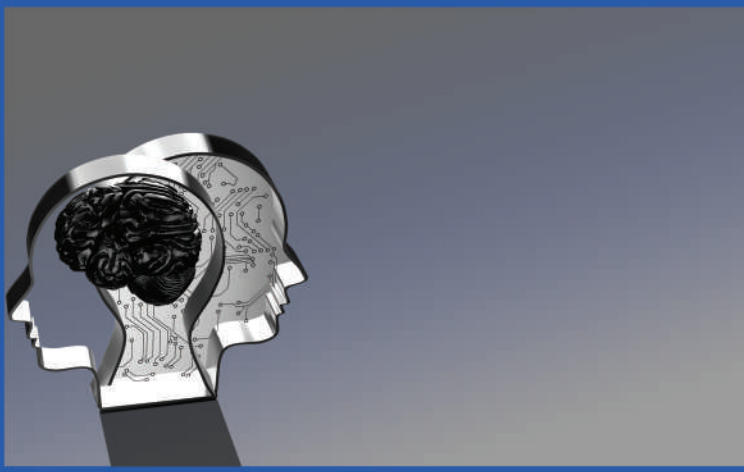
بخش پنجم

بخش ششم

بخش هفتم

بخش هشتم

سخن نخست



سخن نخست

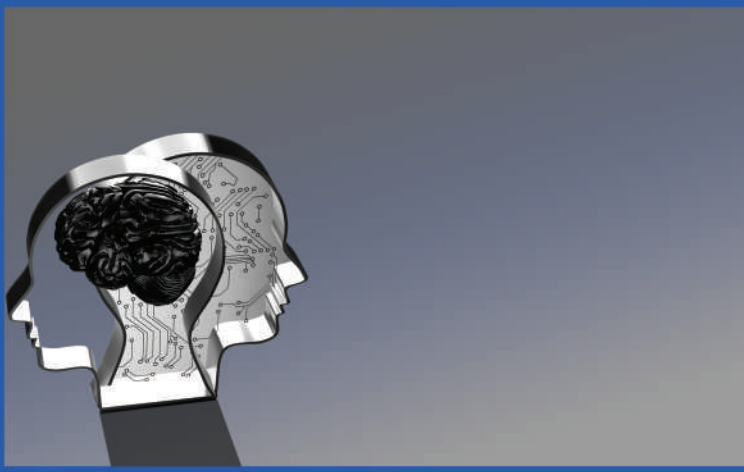
فضای مجازی با شتاب شگرف و رو به تزایدی که در حال بسط و گسترش است تمام ساحات اجتماعی، اقتصادی، سیاسی و فرهنگی زندگی بشر را درنوردیده و هر روز بخش بزرگی از زندگی واقعی را در خود فرو برده و حیات متفاوت و جدیدی به آن می‌دهد. لذا به نظر می‌رسد دو نگاه کلان به فضای مجازی وجود دارد: نگاه اول که بالاخص در ابتدای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

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

سیدابوالحسن فیروزآبادی

دبیر شورای عالی و رئیس مرکز ملی فضای مجازی

پیش گفتار



پیش‌گفتار

این گزارش[1] یک نقشه راه گسترده برای پژوهش در مورد پیامدهای اخلاقی و اجتماعی الگوریتم‌ها، داده‌ها و هوش مصنوعی (ADA) را ترسیم می‌کند. تأثیر الگوریتم‌ها، داده‌ها و هوش مصنوعی بر مردم و جامعه تقریباً همه پرسش‌ها در حوزه سیاست‌های عمومی را شکل می‌دهد، اما مباحث لزوماً بر درک مشترک از مسائل اخلاقی اصلی یا چارچوب توافق‌شده‌ای شکل نگرفته است که بتواند یک رویکرد اخلاقی برای توسعه و استقرار فناوری‌های مبتنی بر ADA باشد. حتی در مواردی که درباره مسائل اصلی نظیر سوگیری، شفافیت، مالکیت و رضایت اتفاق‌نظر وجود داشته باشد، ممکن است در زمینه‌های مختلف معانی گوناگونی به خود بگیرند- به‌عنوان مثال تفسیر این کلمات در کاربردهای فنی با تفسیر آن‌ها در سیستم قضایی متفاوت است. به همین ترتیب، ارزش‌های اخلاقی مانند انصاف می‌توانند در زبان‌ها، فرهنگ‌ها و سیستم‌های سیاسی مختلف موضوعِ تعاریف گوناگونی قرار گیرند.

اگر بخواهیم فناوری‌های مبتنی بر ADA توسعه یافته و به نفع جامعه استفاده شوند، ایضاح این مفاهیم و حل تنش‌ها و بده-بستان‌ها

۱.این نوشتار ترجمه‌ای است از منبع زیر:
Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019) Ethical and societal implications of algorithms, data, and artifcial intelligence: a roadmap for research. London: Nuffeld Foundation

بیـن اصـول و ارزش‌هـای اساسـی در ایـن زمینـه بسـیار مهـم اسـت. ایـن نقشـه راه، بـرای اولیـن بـار، جهـت پژوهش‌هایـی را مشـخص می‌کنـد کـه بایسـتی، بـرای ایجـاد پایـگاه دانـش و گفتمـان مشـترکی کـه می‌توانـد یـک رویکـرد اخلاقـی را پایه‌ریـزی کنـد، اولویت‌بنـدی شـود. بـرای هـر یـک از وظایـف اصلـی مشخص‌شـده، نویسـندگان پرسـش‌های مُفصلـی را ارائـه می‌دهنـد کـه در صـورت پرداختـن بـه آن‌هـا، توانایـی جمعـی بـرای آگاهی‌بخشـی و بهبـود اسـتانداردها، مقـررات و سیسـتم‌های نظـارت بـر فناوری‌هـای مبتنـی بـر ADA، افزایـش خواهـد یافـت.

اخیـراً بنیـاد نوفیلـد[1] - بـا مشـارکت دیگـران - مؤسسـه آدا لاولایـس (Ada)[2] را، به‌عنـوان یـک نهـاد تحقیقاتـی و مشـورتی مسـتقل بـا هـدف اطمینـان از کاربـرد داده‌هـا و هـوش مصنوعـی بـرای افـراد و جامعـه، ایجـاد کـرده اسـت. هـدف مـا از ترسـیم ایـن نقشـه راه، آگاهی‌بخشـی بـه برنامـه کاری آدا و همچنیـن کمـک بـه شـکل‌گیری برنامـه پژوهشـی پیرامـون ایـن مسـئله مهـم بـود کـه چگونـه جامعـه می‌توانـد قـدرت تحول‌آفریـن و مزایـای داده‌هـا و هـوش مصنوعـی را توزیـع کنـد و هم‌زمـان آسـیب‌های ناشـی از آن‌هـا را کاهـش دهـد.

پیـام برآمـده از نقشـه راه ایـن اسـت کـه پژوهـش دربـاره سـؤالات مطرح بایسـتی چندگانـه[3] و میان‌رشـته‌ای باشـد و منافـع متکثـر تحقیقـات علمـی، سیاسـت‌های عمومـی، بخـش خصوصـی و جامعـه مدنـی را بـه هـم متصـل کنـد. ایـن موضـوع اسـاسِ مأموریـت مؤسسـه Ada Lovelace اسـت. یکـی از اهـداف اصلـی آدا، جمع‌آوری نظـرات واگـرا، بـرای ایجـاد درک مشـترک از مباحـث اخلاقـی مطـرح پیرامـون داده‌هـا و هـوش مصنوعـی اسـت و بـرای اجرایی‌کـردنِ ایـن منظـور یک رویکرد میان‌رشـته‌ای و مشـارکتی لازم اسـت.

1. Nuffield Foundation (مترجم)
2. Ada Lovelace Institute(مترجم)
3. plural(مترجم)

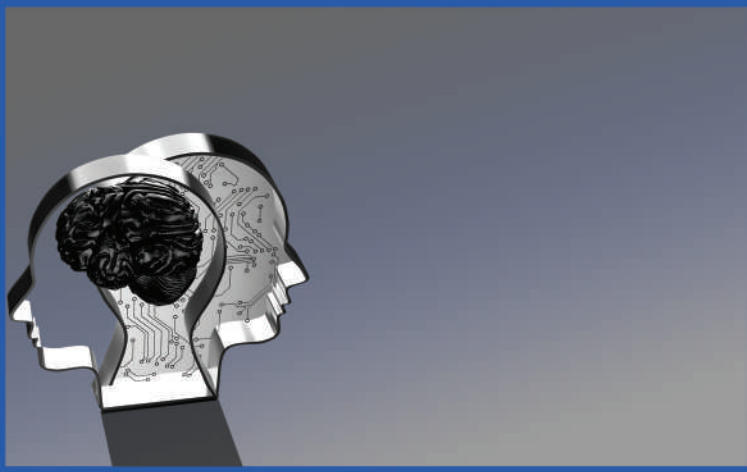
بنیاد نوفیلد به‌عنوان یک سرمایه‌گذار مستقل با مأموریت پیشبُرد رفاه اجتماعی، مشتاق تأمین بودجه تحقیقات بیشتر در این زمینه است. این سؤال که چگونه فناوری‌های دیجیتال و تأثیرات توزیعی آن‌ها می‌توانند آسیب‌پذیری را کاهش، تشدید و تغییر دهند و مفاهیم اعتماد، شواهد و اقتدار را تحت تأثیر قرار دهند، یکی از موضوعات اولویت‌بندی‌شده در استراتژی ما است. ما امیدواریم که این نقشه راه به تدوین پیشنهادات و پروپوزال‌های تحقیقاتی مرتبط کمک نماید.

من از نویسندگان این گزارش برای ارائه مشوق‌های فکری و درعین‌حال کاربردی در این زمینه مهم، تشکر می‌کنم.

تیم گاردام[1]
مدیر عامل

1. Tim Gardam(مترجم)

خلاصه اجرایی



خلاصه اجرایی

هدف این گزارش ارائه یک نقشه راه گسترده برای پژوهش درباره پیامدهای اخلاقی و اجتماعی الگوریتم‌ها، داده‌ها و هوش مصنوعی (ADA) در سال‌های آینده است. افراد درگیر در برنامه‌ریزی، تأمین مالی، پژوهش و سیاست‌گذاری‌های مربوط به این فناوری‌ها، هدف این گزارش هستند. ما از اصطلاح «فناوری‌های مبتنی بر ADA» برای گرفتن طیف گسترده‌ای از فناوری‌های اخلاقی و اجتماعی مبتنی بر الگوریتم‌ها، داده‌ها و هوش مصنوعی (AI) استفاده می‌کنیم، چراکه می‌دانیم این سه مفهوم کاملاً از یکدیگر قابل تفکیک نیستند و اغلب با هم هم‌پوشانی دارند.

مجموعه مشترکی از مفاهیم و نگرانی‌های اساسی در حال پدیدآمدن است که در آن توافق گسترده‌ای در مورد برخی از مسائلی اصلی (مانند سوگیری[1]) و ارزش‌هایی (مانند انصاف) که اخلاقیات الگوریتم‌ها، داده‌ها و هوش مصنوعی باید بر روی آن‌ها تمرکز کند، وجود دارد. طی دو سال گذشته، این موارد در کدها و مجموعه‌های مختلف «اصول» تدوین شده است. توافق در مورد این مسائل، ارزش‌ها و اصولِ سطح‌بالا گام مهمی برای اطمینان از توسعه و استفاده از

1. bias(مترجم)

فناوری‌های مبتنی بر ADA، در جهت منافع جامعه است.

با این وجود، سه خلأ اصلی در پژوهش‌های موجود مشاهده می‌کنیم: (۱) عدم وضوح یا اجماع در مورد معنای مفاهیم اخلاقی مرکزی و نحوه کاربرد آن‌ها در شرایط خاص؛ (۲) عدم توجه کافی به تنش‌های موجود بین آرمان‌ها و ارزش‌ها؛ (۳) شواهد ناکافی در موارد (الف) قابلیت‌ها و تأثیرات کلیدی فناوری، و (ب) دیدگاه‌های گروه‌های مختلف اجتماعی.

به‌منظور حل این مشکلات، توصیه می‌کنیم که پژوهش‌های آینده بایستی جهات گسترده زیر را در اولویت قرار دهند (توصیه‌های دقیق‌تر در بخش ۶ همین گزارش آمده است):

۱. کشف و حل ابهامات موجود در اصطلاحات معمول و کاربردی در ادبیات (مانند حریم خصوصی، سوگیری و توضیح‌پذیری)، از طریق:

الف. تحلیل تفسیرهای مختلف آن‌ها.

ب. شناسایی نحوه استفاده از آن‌ها در عمل در رشته‌ها، بخش‌ها، اجتماعات و فرهنگ‌های مختلف.

پ. ایجاد اجماع و توافق در مورد استفاده از آن‌ها در مواردی که از نظر فرهنگی و اخلاقی حساس هستند.

ث. تشخیص صریح تفاوت‌های اساسی در مواردی که به‌راحتی قابل توافق نیستند، و توسعه اصطلاحات برای جلوگیری از ایجاد سوءتفاهم[1] افراد از رشته‌ها، بخش‌ها، مردم و فرهنگ‌های مختلف هنگام گفتگو با یکدیگر.

۲. شناسایی و حل تنش‌ها بین شیوه‌هایی که فناوری

1. talking past one another: شرایطی را توصیف می‌کند که دو یا چند نفر در مورد موضوعات مختلف صحبت می‌کنند، درحالی‌که معتقدند آن‌ها در مورد یک چیز صحبت می‌کنند. (مترجم)

هم‌زمان ممکن است هم ارزش‌های مختلف را تهدید و هم از آن‌ها پشتیبانی کند، از طریق:

الف) بررسی موارد مشخص تنش‌های ذیل که در کاربردهای فعلی ADA نقش اساسی دارند:

۱. استفاده از الگوریتم‌ها برای دقیق‌ترکردنِ تصمیم‌گیری‌ها و پیش‌بینی‌ها **در مقابل** اطمینان از برخورد عادلانه و برابر.

۲. بهره‌مندی از مزایای افزایش شخصی‌سازی در فضای دیجیتال **در مقابل** افزایش همبستگی و شهروندی.

۳. استفاده از داده‌ها برای بهبود کیفیت و کارایی خدمات **در مقابل** رعایت حریم خصوصی و استقلال اطلاعاتی[1] افراد.

۴. استفاده از اتوماسیون برای راحت‌ترکردنِ زندگی مردم **در مقابل** ترویج خودشکوفایی[2] و عزت نفس.

ب) شناسایی تنش‌های بیشتر با در نظر گرفتن مواقعی که:

۱. هزینه‌ها و مزایای فناوری‌های مبتنی بر ADA ممکن است به‌طور نابرابر بین گروه‌های ایجادشده براساس جنسیت، طبقه، توانایی (عدم توانایی) یا قومیت توزیع شود.

۲. مزایای کوتاه‌مدت فناوری ممکن است به قیمت از دست دادن ارزش‌های درازمدت تمام شود.

۳. فناوری‌های مبتنی بر ADA ممکن است به نفع افراد یا گروه‌هایی باشد اما مشکلاتی را در سطح جمعی و کلان ایجاد می‌کند.

ج) بررسی روش‌های مختلف برای حل انواع گوناگون تنش‌ها، به‌ویژه تمایز قائل‌شدن بین آن تنش‌هایی که منعکس‌کننده

1. informational autonomy (مترجم)
2. self-actualisation (مترجم)

تعارض اساسی بین ارزش‌ها هستند و آن‌هایی که یا توهم‌آمیز هستند یا امکان ارائه راه‌حل‌های عملی دارند.

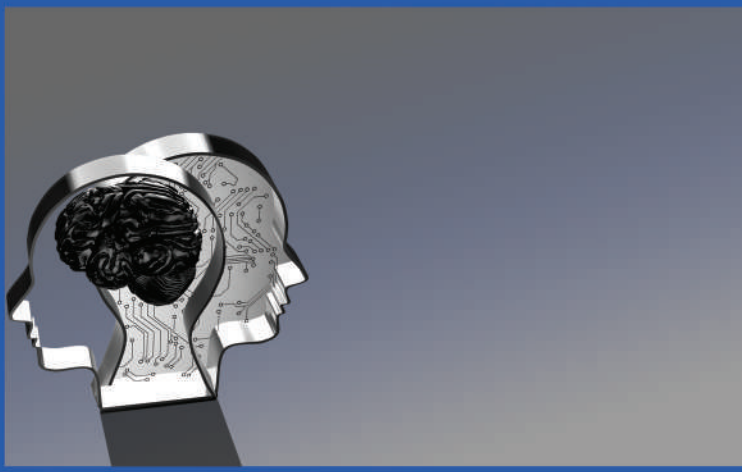
۳. ایجاد یک فضای مبتنی بر شواهدِ دقیق‌تر برای بحث در مورد مسائل اخلاقی و اجتماعی، با استفاده از:

الف) درک عمیق‌تر از آنچه از نظر فنی امکان‌پذیر است، برای ارزیابی خطرات و فرصت‌های ADA برای جامعه و تفکر واضح‌تر در مورد بده-بستان‌ها مابین ارزش‌ها

ب) یک فضای مبتنی بر شواهد قوی‌تر در مورد کاربرد کنونی و تأثیرات فناوری‌های مبتنی بر ADA در بخش‌های مختلف و بر گروه‌های گوناگون - به‌ویژه مواردی که ممکن است در گروه‌های محروم و کمتردیده‌شده (مانند زنان و افراد رنگین‌پوست) یا گروه‌های آسیب‌پذیر باشند (مانند کودکان یا افراد مسن‌تر) - ایجاد کنیم و به‌طور دقیق‌تر در مورد این موضوعات فکر کنیم که تنش‌ها بین ارزش‌ها کجا و چگونه به‌وجود می‌آیند و چگونه می‌توانیم آن‌ها را حل نماییم.

ج) بر اساس مشارکت عمومی موجود، یک کار پژوهشی برای درک چشم‌اندازهای مختلف مردم جامعه، به‌ویژه دیدگاه‌های گروه‌های حاشیه‌ای، در مورد موضوعات مهم، به‌منظور ایجاد اجماع در صورت امکان، تعریف کنیم.

مقدمه



مقدمه

۱،۱. اهداف، رویکرد و طرح کلی

هدف از این گزارش ارائه یک نقشه راه برای کار روی دلالت‌های اخلاقی و اجتماعی الگوریتم‌ها، داده و هوش مصنوعی[1] (ADA) در سال‌های آینده می‌باشد. پیشرفت‌های انجام‌شده در فهم این مسائل در محافل دانشگاهی، سیاست‌گذاری و صنعت را مرور کرده و شکاف‌های موجود در چشم‌انداز پژوهشی کنونی را شناسایی خواهیم کرد. نهایتاً اقدام به ارزیابی نقاط قوت و محدودیت‌های کارهای موجود خواهیم نمود. بر همین اساس ما سه حوزه پژوهشی وسیع را پیشنهاد کرده و پرسش‌ها و مسائل اولویت‌دار خاصِ هر یک از آن‌ها را گوشزد می‌نماییم. این پیشنهادات و گزارش به‌طور کلی، اشخاص و سازمان‌هایی را هدف قرار داده است که دست‌اندکار طرح‌ریزی، سرمایه‌گذاری و پیگیری پژوهش و کارهای سیاست‌گذاران در ارتباط با چالش‌های در حال ظهور اخلاقی و اجتماعی ADA هستند. تمرکز روی مسائل کوتاه و میان مدتی است که پیش‌تر ظاهر شده یا در فرایند ظهور و پیدایش هستند، ما روی راه‌حل‌هایی که نیازمند تحولات بنیادین سیاسی یا فناورانه هستند تمرکز نخواهیم کرد[2].

1. algorithms, data, and AI (ADA) (مترجم)

۲. سایر گروه‌ها بر اولویت پژوهش دربارهٔ چالش‌های مربوطه و بلندمدت سیستم‌های پیشرفتهٔ هوش مصنوعی تأکید کرده‌اند، که قابل‌توجه‌ترین آن‌ها مؤسسهٔ Future of Humanity Institute در دانشگاه آکسفورد است که اخیراً یک برنامهٔ پژوهشی برای حکمرانی هوش مصنوعی در بلندمدت منتشر کرده است. بررسی روابط متقابل بین اولویت‌های پژوهشی کوتاه‌مدت و بلندمدت‌تر، و اینکه چگونه این دو می‌توانند از یکدیگر بیاموزند، برای کارهای آتی مفید خواهد بود.

ما همچنین ابتدائاً بر اولویت‌های پژوهشی تمرکز خواهیم کرد نه نقش سیاست‌گذاری یا تنظیم مقررات. با این وجود تصریح می‌کنیم که این گزینه‌ها نیز مورد پژوهش واقع شده‌اند و در این گزارش چگونگی‌اش را نشان می‌دهیم.

برای تولید این پیشنهادات از مرور ادبیات گسترده کارهایی که به زبان انگلیسی انجام شده است آغاز کردیم: بیش از ۱۰۰ مقالهٔ نظری و تجربی دانشگاهی را از رشته‌هایی مثل علوم رایانه، اخلاق، تعامل انسان-رایانه، حقوق و فلسفه (و سایر رشته‌ها) پوشش دادیم. ما همچنین اسناد سیاست‌گذارانهٔ کلیدی را در قاره‌های مختلف و نیز برخی از پُرارجاع‌ترین اخبار و مقاله‌های ژورنالی در سال‌های اخیر را مرور کردیم[۱]. ما سه کارگاه (که هر یک از آن‌ها دست کم بیست متخصص را از حوزه‌های مربوطه کنار یکدیگر جمع کرد) و یک سلسله جلسات کوچک بحث و تعاطی افکار در گروه‌های ۵ الی ۱۰ نفره برگزار کردیم.

گزارش حاضر این‌گونه سازماندهی شده است:

- بخش ۲، بر اساس یک مرور ادبیات جزئی (که در پیوست اول قابل دسترسی است) یک خلاصهٔ سطح‌بالا از چشم‌انداز کنونی ارائه می‌دهد. ما برخی از دستاوردهای پژوهشی و برخی از شکاف‌هایی که هنوز وجود دارند را برجسته کرده‌ایم. ما چنین نتیجه گرفتیم که مسیر پیشِ‌رو به‌ویژه نیازمند سه نوع کار است: مفهوم‌سازی[۲]، شناسایی و حل تنش‌ها[۳] و بده-بستان‌ها[۴]، و ایجاد یک مبنای مبتنی بر شواهدِ محکم‌تر در مورد این تنش‌ها.

۱. ما آن دسته از مقالات رسانه‌ای را برگزیدیم که در آثار آکادمیک بررسی‌شده به‌کرات به آن‌ها ارجاع شده بود، یا آن مقالات رسانه‌ای را انتخاب نمودیم که طی سال گذشته در جاهای پُربازدیدی مثل نیویورک تایمز، گاردین یا تک‌کرانچ (TechCrunch) نوشته شده بودند. مرور نظام‌مندتر و جامع‌تر درگاه‌های رسانه‌ای فراتر از قلمرو این بررسی ابتدایی است، اما تحلیلی وسیع‌تر در کارهای آتی می‌تواند موجب تقویت ارزیابی ما از این فضا شود.

2. Concept building(مترجم)
3. tensions(مترجم)
4. tradeoffs(مترجم)

• بخش‌های ۳ تا ۵ به‌نوبه‌خود روی این حوزه‌های پیشنهادشده تمرکز می‌کنند: با جزئیات بیشتری توضیح می‌دهند که چرا این یک اولویت است، به‌طور کلی در این حوزه‌ها چه پژوهش‌هایی را باید در نظر بگیریم، اینکه چه پرسش‌ها یا حوزه‌های مشخصی با اهمیت به نظر می‌رسند. بخش ۶ نتایج بخش‌های گذشته را استخراج کرده و یک نقشه راه ارائه می‌دهد: مجموعه‌ای از مسیرهای پژوهشی پیشنهادشده سطح‌بالا.

۱،۲. تعاریف و واژگان کلیدی

دامنه این تحقیق گسترده بود: درنظرگرفتنِ دلالت‌های اخلاقی و اجتماعی ADA.

دلالت‌های اخلاقی و اجتماعی

ما یک تعریف گسترده‌تر از دلالت‌های اخلاقی و اجتماعی اتخاذ می‌کنیم: ملاحظه انحاء اثرگذاری ADA بر بخش‌های مختلف جامعه، و اینکه چگونه این دلالت‌ها می‌توانند ارزش‌های پذیرفته‌شده را تقویت یا تهدید کنند. ما با قصد و آگاهی از واژه دلالت‌ها استفاده می‌کنیم تا نشان بدهیم که نه تنها به تأثیرات منفی این فناوری‌ها علاقه‌مندیم (این معنا در واژگانی مثل مسائل، خطرات یا چالش‌ها نیز نهفته است) بلکه همچنین علاقمند به تأثیرات مثبت احتمالی آن‌ها نیز می‌باشیم. از آنجا که بعداً بر اهمیت درنظرگرفتنِ تنش‌های برخاسته از فرصت‌ها و تهدیدهای فناوری‌های مبتنی بر ADA تأکید خواهیم کرد، این بسیار حیاتی است.

ارزش‌ها

وقتی که صحبت از شناسایی تعارضات بین ارزش‌های مختلفی که توسط فناوری‌های جدید تقویت یا تهدید شده‌اند می‌کنیم، از واژه ارزش‌ها برای اشاره به تعهداتی استفاده می‌کنیم که به‌نحو معقول و گسترده‌ای پذیرفته شده و انسان‌ها عمیقاً به آن‌ها وفادارند. ارزش‌ها صرفاً میل‌هایی که آشکارکننده ترجیحات[1] یا لذت‌ها باشند، نیستند. آن‌ها اهداف و آرمان‌هایی هستند که مردم از روی فکر و اندیشه تأییدشان می‌کنند، انسان‌ها از طریق ارزش‌هاست که زندگی مشترکشان را سازماندهی می‌کنند[2]. در اینجا ما روی آن ارزش‌هایی تمرکز خواهیم کرد که مکرراً در بحث‌های جهانِ انگلیسی‌زبان در مورد هوش مصنوعی و فناوری‌های در حال ظهور مبتنی بر داده مطرح شده‌اند، اما تلاش خواهیم کرد تا ارزش‌هایی را که به‌نحو گسترده‌تر به‌صورت میان‌فرهنگی تکرار شده‌اند را نیز شناسایی کنیم.

الگوریتم‌ها

در ریاضیات و علوم رایانه واژه الگوریتم به‌معنای فرایندی نامبهم و روشن برای حل مجموعه‌ای از مسائل است. در این گزارش ابتدا از واژه الگوریتم، برداشتی شبیه به «الگوریتم خودکار[3]» خواهیم داشت: فرایندی که به‌منظور استنتاج / استدلال یا تصمیم‌گیری خودکار، غالباً توسط یک رایانه دیجیتالی انجام می‌شود. ما از واژه الگوریتم، اغلب به‌مثابه نوعی مختصرنویسی برای اشاره به نرم‌افزاری که این فرآیند را اجرا می‌کند استفاده می‌کنیم، و واژه‌هایی مثل «تصمیم‌گیری الگوریتمیک» کم‌وبیش معادل تصمیم‌گیری کامپیوتری‌شده هستند. از

۱. یعنی ما تصور نمی‌کنیم که صرفاً با مشاهدهٔ چگونگی رفتار آن‌ها در یک مکان بازاری بتوان ارزش‌هایی را استنباط نمود.
۲. تیبریوس (Tiberius) (۲۰۱۸) تعریف کامل‌تری به دست می‌دهد.
3. automated algorithm (مترجم)

این منظر وجه کلیدی الگوریتم‌ها این است که می‌توانند خودکار شوند و می‌توان آن‌ها را به‌گونه‌ای نظام‌مند با سرعتی بسیار بیشتر از انسان اجرا کرد که در نتیجه، فرآیندهای بسیار دیگری به‌صورت خودکار تولید می‌شود.

داده

ما داده را به‌مثابه «اطلاعات کدگذاری‌شده» در مورد یک یا چند پدیده هدف (به‌عنوان مثال ابژه‌ها، رویدادها، فرآیندها یا اشخاص) تعریف می‌کنیم. امروزه داده‌ها معمولاً به‌نحو دیجیتالی کدگذاری می‌شوند تا به‌نحو آنالوگ. داده‌ها به سه دلیل از منظر اخلاقی و اجتماعی اهمیت دارند. (۱) خودِ فرآیند جمع‌آوری و سازماندهی داده‌ها مستلزم اتخاذ مفروضاتی در مورد این است که چه چیزی مهم و سودمند بوده و ارزش توجه‌کردن دارد. از آنجا که این مفروضات در همه بافتارها برقرار نیستند، هیچ مجموعه داده‌ای کامل، دقیق و بی‌طرف نیست. (۲) داده‌های کدگذاری‌شده دیجیتالی امکان تکثیر، انتقال و تغییر اطلاعات را با سرعت و بازدهیِ بیش از همیشه می‌دهند. (۳) اَشکال جدید تحلیل، امکان استخراج بینش‌های بی‌سابقه‌ای را به کسانی که حجم عظیمی از داده‌ها را پردازش می‌کنند، می‌دهد.

هوش مصنوعی

احتمالاً هوش مصنوعی - یکی از سه واژه کلیدیِ مورد نظر این گزارش- سخت‌ترین و مناقشه‌آمیزترین واژه برای تعریف باشد. واژه هوش به انحاء گوناگون در گفتمان عمومی و در حوزه‌های دانشگاهی

مختلف با دلالت‌های ضمنی گوناگون مورد استفاده قرار گرفته است[1].
از منظر این گزارش، هوش مصنوعی اشاره به هرگونه فناوری دارد که وظایفی را اجرا می‌کند که می‌توان آن [وظایف] را هوشمندانه در نظر گرفت - و این را هم فراموش نمی‌کنیم که ممکن است باورهای ما در مورد اینکه چه چیزی هوشمندانه است در طول زمان تغییر کند. برای مثال، احتمالاً ما به‌طور شهودی، ادراک بصری یا راه‌رفتن را وظایف هوشمندانه در نظر نمی‌گیریم، زیرا آن‌ها را با کمترین تلاش آگاهانه انجام می‌دهیم: اما تلاش برای پیاده‌سازی این قابلیت‌ها در ماشین‌ها نشان داده است که آن‌ها مبتنی بر فرایندهای بسیار پیچیده هستند. ما همچنین گمان می‌کنیم که ویژگی‌های کلیدی هوش مصنوعی بیشترین اهمیت را برای اخلاق و جامعه دارند: این واقعیت که هوش مصنوعی اغلب برای بهینه‌کردنِ فرایندها مورد استفاده قرار می‌گیرد و می‌تواند طوری توسعه پیدا کند که به‌طور خودکار عمل کند، رفتارهای پیچیده‌ای را خلق می‌کند که فراتر از چیزی می‌روند که صراحتاً برنامه‌ریزی شده‌اند.

عموم[2]

واژه عموم، اغلب به‌عنوان واژه‌ای برای اشاره به تمام اشخاص جامعه مورد استفاده قرار می‌گیرد. از سوی دیگر ما از این واژه برای اشاره به گروه‌های ذینفع متفاوت و مختلف (دانشمندان، واسطه‌ها[3]، تصمیم‌گیران، فعالان و ...) که دیدگاه‌های مستقل خودشان را دارند استفاده می‌کنیم[4].

این به ما امکان پرهیز از تمرکزِ صرف بر دیدگاه‌ها و رویکردهای

۱. نگاه کنید به: Cave (۲۰۱۷)

2. Publics(مترجم)

3. mediators(مترجم)

۴. در اینجا ما از Burns et al. (۲۰۰۳) پیروی می‌کنیم، که عموم مختلف را به‌مثابه موارد مرتبط با زمینه‌های مختلف شناسایی می‌کند.

غالب- که به قیمت نادیده انگاشتن دیدگاه‌های فرعی تمام می‌شود- را می‌دهد. ما عبارت «عموم مردم» را در مقابل «متخصصان» به‌کار نمی‌بریم، چراکه در واقع، هر گروهی، تخصص پُراهمیت خاص خود را دارد.

با درنظرگرفتن این تعریف‌ها می‌توانیم روشن کنیم که چرا دلالت‌های اخلاقی و اجتماعی فناوری‌های مبتنی بر ADA می‌تواند موجب نگرانی شود. فناوری‌های مبتنی بر ADA ذاتاً دارای استفاده دوگانه هستند: هدف و منظوری که ابتدائاً برای آن طراحی شده است می‌تواند به سادگی تغییر و تحول پیدا کند و اغلب منجر به ظرفیت‌های اخلاقی کاملاً متفاوتی در آن‌ها شود. برای مثال تکنیک‌های شناسایی تصویر که کاربردهای کاملاً مثبت دارند - مثلاً در تشخیص تومورهای بدخیم - می‌توانند در خدمت اهداف زیان‌آوری مثل نظارت توده‌ای نیز قرار بگیرند (Bloomberg News, ۲۰۱۸). در همین رابطه می‌توان گفت که فناوری‌های مبتنی بر ADA، دارای قابلیت‌های کاملاً مستقل از دامنه کاربرد هستند، مثل پردازش اطلاعات، به‌دست‌آوردنِ دانش و تصمیم‌سازی. بنابراین تکنیک‌های مشابهی را می‌توان تقریباً در مورد همه وظایف به کار برد و آن‌ها را در بخش‌های مختلف جامعه نفوذ داد و فراگیر کرد. یک فناوری یکسان، در حوزه‌های کاربردی مختلف، برای گروه‌های گوناگون می‌تواند دارای خطرات و فواید کاملاً متفاوتی باشد و ارزش‌های متفاوتی را تحت تأثیر قرار دهد.

این ویژگی‌ها به همراه سرعت چشمگیر شرکت‌های قدرتمند خصوصی در ابداع کاربردهای جدید برای فناوری‌های مبتنی بر

ADA در دهه‌های اخیر، افزایش تمرکز برای تنظیم و هدایت اخلاق ADA در جهت صحیح را توضیح می‌دهد.

بخش اول

چشم‌انداز کنونی

بخش اول

چشم‌انداز کنونی

بحث دلالت‌های اخلاقی و اجتماعی ADA، به‌طور کامل ذیلِ هیچ تک رشتهٔ دانشگاهی یا تک بخش [صنعتی] قرار نمی‌گیرد. بنابراین برای درک کامل طیف مباحثی که در سال‌های اخیر پوشش داده شده‌اند، باید نگاه وسیع داشته باشیم: از انتشارات دانشگاهی مختلف از فلسفه و علوم سیاسی گرفته تا یادگیری ماشین را باید در نظر بگیریم تا صنعت و گزارش‌های رسانه‌ای. مروری که در اینجا خواهیم داشت در پی فهم دو نکته است: (۱) چه مسائل و دغدغه‌های خاصی مد نظر انواع و اقسام [اسناد] منتشرشده بوده است، (۲) چه تلاش‌هایی برای ترکیب مسائل در رشته‌ها و حوزه‌های مختلف انجام شده است.

ما دو نتیجه اصلی از این مطالعه مروری گرفتیم. (۱) مجموعه مشترکی از مفاهیم و دغدغه‌ها در حال ظهور است، اما واژگان مورد استفاده غالباً مبهم بوده و بدون تأمل دقیق استفاده شده‌اند. (۲) تلاش‌های متعدد و مختلفی برای ترکیب موضوعات و مسائل در چارچوب‌ها و مجموعه اصولِ موجود انجام شده است[۱] اما این تلاش‌ها غالباً یا غیرنظام‌مند بوده‌اند یا آن‌قدر سطح‌بالا هستند که نمی‌توانند کنش‌های عملی را هدایت کنند[۲].

۱. به‌عنوان مثال، برای یک رویکرد چارچوب‌محور نگاه کنید به: Cowls and Floridi (۲۰۱۸) و برای یک رویکرد اصول‌محور نگاه کنید به: House of Lords Select Committee on AI's (2018)

۲. این قسمت محدود به ارزیابی سطح‌بالا از چشم‌انداز فعلی اخلاقیات ADA و تأثیرات اجتماعی است. برای توصیفات و ارزیابی‌های جزئی‌تر به ضمایم ۱ تا ۴ مراجعه کنید.

۱-۱- شناسایی مفاهیم و دغدغه‌های کلیدی

اگرچه ادبیات موجود طیف وسیعی از مسائل را پوشش می‌دهد، با این همه مجموعه مشترکی از مفاهیم و دغدغه‌های مشترک در حال ظهور است. برای مثال دغدغه‌هایی در مورد سوگیری‌های الگوریتمیک و حصول اطمینان از اینکه یادگیری ماشین که [فرآیندهای مختلف] تصمیم‌گیری در مورد افراد را پشتیبانی می‌کند، به‌طور منصفانه[1] مورد استفاده قرار بگیرد، تبدیل به یکی از اصلی‌ترین موضوعات شده است، چراکه بر اهمیت شفاف‌سازی و توضیح‌پذیری[2] سامانه‌های جعبه سیاه گونهٔ[3] [هوش مصنوعی]، تأکید می‌کند. مسئله حریم خصوصی داده‌های شخصی نیز مکرراً مطرح شده است، نیز پرسش از اینکه با خودکارشدگی روزافزون تصمیمات، چگونه می‌توانیم پاسخ‌گویی[4] و مسئولیت‌پذیری[5] را [کمافی‌سابق] حفظ کنیم. تأثیر فناوری‌های مبتنی بر ADA بر اقتصاد و دلالت‌های آن برای آینده کار، تِم‌های دیگری هستند که مکرراً طرح شده‌اند. برای مشاهده پُرتکرارترین واژگان مورد استفاده در مسائل اخیر برخاسته از ADA به تصویر شماره ۱ بنگرید. این ابرِ کلمات بر اساس بسامد واژگان در چارچوب‌ها و دسته‌بندی‌های مختلفی که ما مرورشان کرده‌ایم به دست آمده است که در آن کلمات بزرگ‌تر بیشتر تکرار شده‌اند.

1. fairly (مترجم)
2. explainable (مترجم)
3. 'black box (مترجم)
4. accountability (مترجم)
5. responsibility (مترجم)

تصویر ۱: ابر کلماتِ مفاهیمِ مشترکِ در حال ظهور، براساس فراوانی تکرار آن‌ها در چارچوب‌ها و گزارش‌های مختلف

با این همه باید دقت کنیم که هرچه کلمات بیشتر استفاده می‌شوند، امکان استفادهٔ نامتأملانه یا مبهم آن‌ها نیز افزایش می‌یابد. برای مثال، مفسران اغلب بر اهمیت شفافیت تأکید کرده‌اند بی‌آنکه روشن کنند این واژه دقیقاً به چه معناست و چرا اهمیت دارد. همچنین در مورد معانیِ نسبت داده شده به این واژه‌ها در سیاق‌های مختلف ناسازگاری نیز به چشم می‌خورد: برای مثال، ممکن است واژهٔ سوگیری در یک مقاله تکنیکی یک معنای کاملاً دقیق داشته باشد، اما در یک گزارش سیاستی، مبهم‌تر باشد. ما در بخش ۳ بحث خواهیم کرد که چگونه استفاده‌ها و تفسیرهای مختلف از یک واژه یکسان می‌تواند دردسرآفرین باشد.

اگرچه شاهد ظهور نوعی اجماع در مورد مسائل کلیدی هستیم، با این همه محلّ تأکید رشته‌های مختلف، متفاوت است. تعجبی ندارد که پژوهش‌های علوم رایانه و یادگیری ماشین عمدتاً بر آن دسته از

مسائل اخلاقی تمرکز دارند که آن‌ها را به‌راحتی می‌توان با واژگان تکنیکی چارچوب‌بندی کرد، از جمله اینکه: چگونه می‌توان سامانه‌های یادگیری ماشین را تفسیرپذیرتر و قابل اعتمادتر کرد، و نیز مسائل حریم خصوصی و حفاظت از داده‌ها. مقاله‌های فلسفه و اخلاق، غالباً بر پرسش‌هایی در مورد اهمیت اخلاقی سامانه‌های پیشرفته‌تر هوش مصنوعی که وجود دارند یا در آینده وجود خواهند داشت، متمرکز هستند، و توجه کمتری به چالش‌های اخلاقی فناوری‌های رایج دارند - اگرچه مجموعه‌ای از ادبیات در مورد این مسائل کوتاه‌مدت، در رشته‌هایی مثل اخلاق اطلاعات و فناوری در حال ظهور است. ادبیات حقوقی دانشگاهی، نسبت به سایر حوزه‌هایی که ما مرور کردیم، تلاش بیشتری کرده است تا تفسیرهای مختلف از واژگانی مثل حریم خصوصی و انصاف را از یکدیگر متمایز کند و دلالت‌های هر یک از این معانی مختلف را مورد بحث قرار دهد.

وقتی که به فراسوی مقالات پژوهشی با موضوعات خاص، یعنی به تلاش‌های سطح‌بالایی نگاه می‌کنیم که برای ترکیب طیف متنوع مسائل شده است، بسیاری را می‌بینیم که رویکردهای مشابهی را برای گروه‌بندی یا دسته‌بندی این مسائل اتخاذ کرده‌اند[1]. برای مثال، شباهت‌های زیادی بین مقولاتی دیده می‌شود که DMES و PAI برای تعریف حوزه‌های پژوهشی‌شان استفاده کرده‌اند:

۱. برای فهرست‌های اخیراً پیشنهادشده از موضوعات کلیدی به ضمیمهٔ ۲ مراجعه کنید.

مضامینِ پژوهشیِ DMES[1]	ستون‌هایِ موضوعیِ PAI[2]
حریم خصوصی، شفافیت و انصاف	هوش مصنوعی منصف، شفاف و پاسخ‌گو
تأثیر اقتصادی، شمول[3]، و برابری	هوش مصنوعی، نیروی کار و اقتصاد
حکمرانی و پاسخ‌گویی	نفوذ اجتماعی هوش مصنوعی
اخلاق و ارزش‌های هوش مصنوعی	هوش مصنوعی و خیر اجتماعی
مدیریت تهدیدها، سوءاستفاده‌ها و پیامدهای ناخواستهٔ هوش مصنوعی	هوش مصنوعی حساس به ایمنی
هوش مصنوعی و چالش‌های پیچیدهٔ جهان	همکاری بین انسان‌ها و سامانه‌های هوش مصنوعی

اگر چه این گروه‌بندی‌ها به برخی ساختارهای بنیادین اشاره دارند، با این همه در وضعیت کنونی‌شان نسبتاً غیرنظام‌مند هستند. این را می‌توان بر اساس اختلافات ظریفِ گروه‌های مختلف، در مرزبندی مقوله‌هایشان‌نشان داد. آیا پاسخ‌گویی به همان مقوله‌ای تعلق دارد که انصاف و شفافیت به آن تعلق دارند (چنان‌که PAI می‌گوید) یا باید آن را همراه با حکمرانی و تنظیم مقررات، ذیل یک مقوله جدا قرار داد (چنان‌که Deep Mind می‌گوید)؟ آیا اعتماد را باید با شفافیت در یک مقوله قرار داد یا با انصاف یا با حریم خصوصی، یا باید همه آن‌ها را در یک مقوله قرار داد؟ آیا «هوش مصنوعیِ معطوف به خیر اجتماعی» باید یک مقوله مستقل باشد یا در تمام مقوله‌های دیگر مشترک باشد؟ چه موضوعات و مسائلی ممکن است به‌طور کامل در هیچ یک از این مقوله‌ها قرار نگیرد (همچون مفهوم «حقوق و آزادی‌ها»یِ AI Now)[4]؟

1. https://deepmind.com/applied/deepmind-ethics-society/research/
2. www.partnershiponai.org/about/#our-work
3. inclusion (مترجم)
4. https://ainowinstitute.org/research.html

بدون درک اینکه چرا این، و نه مسائل و مقوله‌های دیگر برگزیده شده‌اند، اطمینان از اینکه همه مسائل مهم مورد پوشش قرار داده شده‌اند دشوار است. روشن نیست که ارزش‌ها و اولویت‌های مد نظرِ چه کسی ارتقا یافته و اینکه آیا دغدغه‌های سایر اعضای جامعه - از جمله گروه‌های اقلیت- نیز پوشش داده شده است یا خیر. برخی گروه‌ها و مقاله‌ها شروع به اتخاذ رویکردی کرده‌اند که از نقشهٔ بنیادین‌تری از چشم‌انداز اخلاقی آغاز می‌کند: برای مثال، گزارش سال SDPS ۲۰۱۸ «به سوی یک اخلاق دیجیتال»، به‌گونه‌ای نظام‌مند هر یک از ارزش‌های اروپایی و این را که چگونه [این ارزش‌ها] ممکن است مورد تهدید جهانی واقع شوند که روزبه‌روز در حال دیجیتالی‌شدن است، در نظر می‌گیرد. این امر پرسش‌هایی را به ذهن متبادر می‌کند که تاکنون توجه زیادی به آن‌ها نشده است: چگونه رُخ‌سازیِ شخصی‌سازی‌شده می‌تواند همبستگی جامعه را تهدید کند، یا اینکه چگونه دردسترس‌بودگیِ داده‌ها می‌تواند عدم توازن قدرت بین حاکمیت‌ها و شرکت‌ها از یک‌سو و افراد را از سوی دیگر وخیم‌تر کند.

تلاش‌هایی مثل این درصدد تولید یک تک چارچوب نظری هستند که می‌توانند یک تک فهرست از اصول و ارزش‌ها را بازنمایانند (به‌عنوان مثال برای هماهنگی و پاسخ‌گویی عمومی)، همچنین می‌توانند توجهات را محدود کنند: برخی موضوعات را مهم‌تر از سایر موضوعات جلوه دهد. امروزه روشن است که این فضا را به شیوه‌های مختلفی می‌توان پوشش داد که هر کدام محاسن و معایب خود را دارد و برخی ارزش‌ها را بر ارزش‌های دیگر اولویت می‌دهد[1].

۱. برای ارزیابی جزئی‌تر نقاط قوت و ضعف رویکردهای مختلف به سازماندهی موضوعات به ضمیمهٔ ۲ مراجعه کنید. ضمیمهٔ ۳ حاوی چند چشم‌انداز است که می‌توان از آن‌ها برای محدودکردن بحث به مجموعهٔ محدودی از موضوعات استفاده کرد.

۱-۲- صورت‌بندی اصول اخلاقی

علاوه بر جستجوی مفاهیم کلیدی، گروه‌های مختلف شروع به تنظیم اصول تجویزی یا کدهایی برای هدایت توسعه و استفاده از فناوری‌های مبتنی بر ADA نموده‌اند. این اصول اغلب شامل مفاهیمی که در بخش‌های قبلی اشاره کردیم می‌شوند و با آن‌ها هم‌پوشانی دارند، اما تمرکز کمتری بر توضیح اینکه موضوعات چه هستند دارند و در عوض بر توضیح برخی اهدافِ استفاده و توسعه فناوری متمرکز هستند.

برای مثال اصول هوش مصنوعی آسیلومار[1] که در سال ۲۰۱۷ و همراه با کنفرانس آسیلومار برای هوش مصنوعی سودمند[2] توسعه داده شد، رهنمودهایی کلی در مورد چگونگی انجام پژوهش‌ها، اخلاق و ارزش‌هایی که هوش مصنوعی باید به آن‌ها احترام بگذارد، و ملاحظات مهم در مورد مسائل بلندمدت فراهم می‌کند. این اصول توسط صدها پژوهشگر هوش مصنوعی و پژوهشگران دیگر مثل متخصصان اخلاق و دانشمندان اجتماعی دانشگاهی امضا شد. PAI نیز مجموعه‌ای از اصول را برای هدایت توسعه و استفاده از فناوری‌های هوش مصنوعی تأسیس کرده است که تمام اعضا - از جمله بسیاری از تأثیرگذارترین شرکت‌های فناوری- سعی در اتخاذ آن دارند[3].

به‌علاوه، حاکمیت‌ها و نهادهای بین‌المللی در حال توسعه اصول خودشان هستند: گزارش اخیر کمیتهٔ منتخب لُردها برای هوش مصنوعی[4] به‌عنوان «هوش مصنوعی در انگلستان: آماده، مشتاق و توانا؟»[5] پنج اصل را برای یک دستورالعمل مقطعی هوش مصنوعی که

1. Asilomar(مترجم)
2. https://futureoflife.org/ai-principles/
برخی از نویسندگان این گزارش در آن کنفرانس حاضر و در توسعهٔ آن اصول دخیل بودند.
3. www.partnershiponai.org/tenets/
مرکز آیندهٔ هوش لورهولم (The Leverhulme Centre for the Future of Intelligence)، که نویسندگان این گزارش در آن مستقرند، یکی از اعضای ائتلاف شراکت دربارهٔ هوش مصنوعی (Partnership on AI) است.
4. Lords Select Committee on Artifcial Intelligence (مترجم)
۵. برخی از نویسندگان این گزارش شواهدی را دراختیار این کمیته قرار دادند.

قابلیت اتخاذ بین‌المللی را دارد پیشنهاد می‌دهد. همچنین IEEE نیز یک «ابتکار جهانی در مورد اخلاق سامانه‌های هوشمند و خودکار»[1] راه‌اندازی کرده و مجموعه‌ای از اصول کلی برای هدایت حکمرانی اخلاقی این فناوری‌ها توسعه داده است. حوزه صنعت نیز درگیر شده است: چشمگیرترین نمونه «اصول اخلاق هوش مصنوعی» است که در ماه ژوئن امسال توسط گوگل منتشر شد[2]. تصویر شماره ۲ واژگان کلیدی را نشان می‌دهد که در تمام مجموعه اصولِ مورد مرور ما ظاهر شده‌اند.

تصویر ۲: ابر کلماتِ مفاهیم پُربسامد در اصول و کدها بر اساس منابع پیوست ۲

هم‌پوشانی چشمگیری بین این مجموعه اصول گوناگون وجود دارد. برای مثال، یک توافق گسترده وجود دارد که فناوری‌های مبتنی بر ADA باید برای خیر عمومی مفید باشند، نباید برای آسیب‌زدن به انسان‌ها یا خدشه‌دارکردنِ حقوق آن‌ها مورد استفاده قرار بگیرند و باید به برخی از مقبول‌ترین ارزش‌هایی که پیش‌تر مورد اشاره قرار

1. https://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

برخی از نویسندگان این گزارش در این ابتکار سهیم بوده‌اند.

2. https://ai.google/principles/

گرفتند- مثل انصاف، حریم خصوصی و خودمختاری- احترام بگذارند. همچنین تلاش‌هایی برای ترکیب این اصول در یک فهرست کوتاه از اصول کلیدی (برای مثال سودرسانی، عدم زیان‌رسانی، خودمختاری، عدالت و توضیح‌پذیری)[1] انجام‌گرفته که بر اساس یک سنت غالب در اخلاق زیست‌پزشکی مُدل‌سازی شده است.

اصول، بخش ارزشمندی از یک اخلاق کاربردی می‌باشد: آن‌ها کمک می‌کنند تا مسائل پیچیده اخلاقی در چند مؤلفه معدودِ محوری فشرده شوند و به این ترتیب امکان ایجاد یک تعهد گسترده در قبال مجموعه مشترکی از ارزش‌ها را فراهم می‌کنند. آن‌ها همچنین می‌توانند ابزاری غیررسمی باشند برای مسئولیت‌پذیرکردنِ مردم و سازمان‌ها در قبال مسائل عمومی. برای مثال، جامعه یادگیری ماشین، سال گذشته در قبال مسئلهٔ سلاح‌های خودکار بسیج شد که طی آن بسیاری از گروه‌ها و پژوهشگران فردی، این تعهد عمومی را دادند که درگیر توسعه [این تسلیحات] نشوند. این مورد نشان می‌دهد که تعهد مشترک به یک اصل خاص و راهنمای عمل، می‌تواند تأثیر واقعی بر دلالت‌های اخلاقی فناوری داشته باشد.

با این وجود اغلب اصولی که برای اخلاق هوش مصنوعی پیشنهاد شده است، به اندازه کافی دقیق نیست که بتواند راهنمای عمل واقع شود. اگرچه این اصول نشان از یک توافق دارند، توافق در مورد اینکه در جریان توسعه و استفاده از فناوری‌های مبتنی بر ADA، کدام اهداف مهم و مطلوب هستند، اما آن‌ها قادر به فراهم‌کردن یک راهنمای عملی برای تفکر در موقعیت‌های جدید و چالش‌برانگیز نیستند. چالش اصلی شناسایی و پیمایش تنش‌هایی است که در عمل به این اصول

1. Cowls and Floridi (2018)

به وجود می‌آید. برای مثال ممکن است یکی از کاربردهای واقعاً سودمند هوش مصنوعی که می‌تواند زندگی انسان‌ها را نجات دهد، مستلزم چنان استفاده‌ای از داده‌های شخصی باشد که ارزش همگانی حریم خصوصی را تهدید کند، یا اینکه مستلزم چنان استفاده‌ای از الگوریتم‌ها باشد که به‌طور کامل توضیح‌پذیر نیست. بحث در مورد اصول، باید این تنش‌ها را بپذیرد و رهنمودهایی برای پیمایش بده-بستان‌های خاص این تنش‌ها داشته باشد[1].

۱-۳- مفروضات بنیادین و شکاف‌های معرفتی

همچنین قابل ذکر است که مفروضات متعددی در بحث‌های کنونی مُستتر است که شکاف‌های دانش موجود دربارهٔ اینکه به‌لحاظ تکنیکی چه چیزی امکان‌پذیر است و امکان‌پذیر خواهد بود، نیز شکاف ارزش‌های بین گروه‌های مختلف در جامعه را آشکار می‌کنند. به‌عنوان مثال، اغلبِ نگرانی‌های معطوف به سوگیری‌های الگوریتمیک، این را مفروض می‌گیرند که الگوریتم‌هایی که جایگزین تصمیم‌گیران انسانی می‌شوند به همان اندازه یا حتی بیشتر سوگیری نداشته باشند. اگرچه گاهی اوقات این مقایسه انجام می‌شود، اما به‌ندرت به‌گونه‌ای نظام‌مند بررسی شده است که چه زمانی می‌توانیم انتظار داشته باشیم که سامانه‌های الگوریتمیک بهتر از انسان‌ها عمل کنند، یا اینکه صرفاً سوگیری‌های انسانی را تکرار یا تقویت نمایند. تأکید بر شفافیت الگوریتمیک، فرض می‌کند که برخی از اقسام توضیح‌پذیری، برای همه انسان‌ها مهم است، درحالی‌که تلاش بسیار اندکی صورت گرفته تا مشخص شود که کدام نوع توضیح مطلوب چه کسانی در

۱. نمونهٔ کاملی از این مورد را در اینجا ارائه کرده‌ایم: Whittlestone et al (2019)

چـه بافتـاری اسـت. بحث‌هایـی کـه در مـورد آینـده کار شـده اسـت، غالبـاً شـامل مفروضاتـی در مـورد فوایـد و آسـیب‌های انـواع و اقسـام خودکارسـازی هسـتند، امـا تاکنـون فاقـد شـواهد خاصـی در مـورد فوایـد و آسـیب‌های عینـی خودکارسـازی یـا نظـر عامـه مـردم در مـورد ایـن مسـائل بوده‌انـد.

عملی‌کـردنِ اصـول و حـل تنش‌هـا مـا را ملـزم بـه شناسـایی ایـن قبیـل مفروضـات و پُرکـردن شـکاف‌های معرفتـی در مـورد قابلیت‌هـای فنـاوری، تأثیـر فنـاوری بـر جامعـه و نظـر عمـوم مـردم خواهـد کـرد. بـدون درک کاربردهـای کنونـی فناوری‌هـای مبتنـی بـر ADA و تأثیـرات آن‌هـا بـر جامعـه، مـا نمی‌توانیـم به‌روشـنی مسـائل و تنش‌هـای پُراهمیـت را شناسـایی کنیـم. بـدون درک اینکـه از منظـر تکنولوژیـک چه چیـزی قابل تحقق اسـت، دشـوار خواهـد بـود کـه بحثـی معنـادار داشـته باشـیم در مـورد اینکـه چـه بده-بسـتان‌هایی وجـود دارد و چگونـه می‌تـوان آن‌هـا را پیمایـش کـرد. و بـدون درک دیـدگاه گروه‌هـای مختلـف در جامعـه، مـا در معـرض خطـر انجام بده-بسـتان‌هایی قـرار خواهیـم گرفـت کـه ارزش‌هـا و نیازهـای اکثریـت را بـر اقلیـت ترجیـح می‌دهـد. صـرفِ توافـق در مـورد اینکـه بایـد خودمختـاری انسـان را حفـظ کنیـم کافـی نیسـت، بـرای مثـال: مـا بایـد درکـی عمیـق داشـته باشـیم از انحـاء خـاصِ کنونـی و آینـدۀ تضعیـف خودمختـاری توسـط فنـاوری، و بفهمیـم کـه افـراد مختلـف در کـدام سـیاق‌ها حاضـر بـه فداکردنِ بخشـی از خودمختـاریِ خـود در قبـال سـایر خیرهـا هسـتند.

۱-۴- جمع‌بندی و پیشنهادات

خلاصه اینکه

- یـک مجموعـه مفیـد از مفاهیـمِ مشـترک در حـال ظهـور اسـت، امـا در

حال حاضر بر اساس واژگان مبهمی بنا شده است که اغلب بدون تأمل به کار می‌روند. در بسیاری از واژگانی که استفاده می‌شود، ابهام‌های مهمی وجود دارد که می‌تواند تفاوت‌های مهمی که بین درکِ رشته‌ها، بخش‌ها، مردمان و فرهنگ‌های مختلف از این مفاهیم وجود دارد را پنهان کند.

• دستورالعمل‌ها و اصول مهمی تأسیس شده است، اما در مورد تنش‌هایی که لاجرم در عمل به وجود خواهند آمد، دانش کمی وجود دارد: چه زمانی ارزش‌ها با یکدیگر ناسازگار خواهند شد، چه زمانی بین نیازهای گروه‌های مختلف ناسازگاری وجود خواهد داشت، یا اینکه چه زمانی با محدودیت منابع مواجه خواهیم بود.

• اغلب بحث‌های کنونی در مورد مسائل و اصول، مبتنی بر مفروضاتی ضمنی هستند در مورد اینکه چه چیزی به‌لحاظ تکنیکی ممکن است، فناوری چگونه جامعه را تحت تأثیر قرار می‌دهد و جامعه برای کدام ارزش‌ها باید اولویت قائل شود. کاملاً حیاتی است که برای عملیاتی‌کردنِ اصول و حل این تنش‌ها، این مفروضات را شناسایی کنیم و به چالش بکشیم و به این ترتیب یک مبنای مبتنی بر شواهد عینی‌تر و قوی‌تر برای درک قابلیت‌های بنیادین تکنولوژیک، تأثیرات و نیازهای اجتماعی فراهم آوریم.

در سال‌های گذشته، در راستای درک دلالت‌های اخلاقی و اجتماعی ADA، چالش‌ها و پرسش‌های برخاسته از آن و چگونگی پرداختن به آن‌ها پیشرفت‌های چشمگیری صورت گرفته است. در ادامه باید بر این مسائل تمرکز کرد:

• ایجاد یک فهم مشترک از مفاهیم کلیدی که ابهام‌ها را تصدیق و حل می‌کند، و بین رشته‌ها، بخش‌ها، عامه مردم و فرهنگ‌ها ارتباط برقرار می‌کند. در بخش ۳، برخی از هم‌پوشانی‌های واژگانی، استفاده‌ها و تفسیرهای متفاوت، و پیچیدگی‌های مفهومی که منجر به آشفتگی و عدم توافق می‌شوند را آشکار می‌کنیم.

• شناسایی و بررسی تنش‌هایی که آن‌گاه به وجود می‌آیند که در تلاشیم تا اصول مورد توافق را در عمل به کار ببندیم. در بخشی از این گزارش دقیقاً همین کار را خواهیم کرد: شناسایی و آشکارسازی جزئی بسیاری از تنش‌هایی که نشان‌دهندهٔ ناسازگاری‌ای هستند که در این فضا در حال ظهور است، و طرح‌ریزیِ کلی برخی دستورالعمل‌ها برای حل این تنش‌ها.

• تعمیقِ درک موجود از قابلیت‌های فناورانه، تأثیرات اجتماعی و دیدگاه‌های گروه‌های مختلف به‌منظور درک بهتر مسائل در حال ظهور و نحوه حل این مسائل. در بخش ۵ توضیح می‌دهیم که چرا درک‌کردن و به‌چالش‌کشیدنِ مفروضات معطوف به فناوری و جامعه، برای حل تنش‌ها اهمیت حیاتی دارند، و سپس حوزه‌های اولویت‌دار پژوهشی را برجسته می‌کنیم.

همچنین در هر یک از بخش‌های پیش‌رو، اولویت‌ها و پیشنهادهای پژوهشی برای کارهای آینده را برجسته می‌نماییم.

بخش دوم

مفهوم‌سازی

بخش دوم

مفهوم‌سازی

یکی از موانع مهم پیشرفت در [مطالعه] مسائل اخلاقی و اجتماعی برخاسته از ADA، مُبهم‌بودن بسیاری از مفاهیم محوری است که در حال حاضر برای شناسایی مسائل مهم مورد استفاده قرار می‌گیرند. چنان‌که در بخش دوم اشاره شد، مفاهیمی مثل «انصاف»، «شفافیت» و «حریم خصوصی» نقش بسیار مهمی در ادبیات موجود دارند. علی‌رغم اینکه این مفاهیم، تِم‌های مشترک در حال ظهور از موردکاوی‌ها را برجسته می‌کنند، با این وجود بسیاری از آن‌ها دارای هم‌پوشانی و ابهام هستند. بخشی از این مسئله از این واقعیت نشئت می‌گیرد که زمینه‌ها، رشته‌ها، بخش‌ها و فرهنگ‌های مختلف می‌توانند استفاده‌های کاملاً متفاوتی از این مفاهیم داشته باشند، بخش دیگر آن ناشی از پیچیدگی‌های ذاتی خودِ این مفاهیم است. در نتیجه، بحث‌های معطوف به تأثیرات اخلاقی و اجتماعی ADA باعث شده که مردم به اشتباه گمان کنند که در حال صحبت‌کردن در مورد مفاهیم مشترک هستند.

پیشرفت سازنده در این فضا مستلزم وضوح مفهومی، تمرکز بیشتر بر ارزش‌ها و منافع مورد بحث است. ما در این بخش، با جزئیات،

چالش‌های مختلفِ سر راه دستیابی به این وضوح مفهومی را ذکر می‌کنیم.

۲-۱- هم‌پوشانی‌های واژگانی

یکی از چالش‌ها این است که واژه‌های مختلف غالباً برای بیان پدیده‌های هم‌پوشان (و نه ضرورتاً یکتا) مورد استفاده قرار می‌گیرند.

برای مثال واژه‌های شفافیت، توضیح‌پذیری، تفسیرپذیری و قابل‌فهم‌بودن[1] غالباً به جای یکدیگر و برای اشاره به این به کار می‌روند که الگوریتم‌های جعبه سیاه طور، باعث ازدست‌رفتنِ چه چیزی می‌شوند. مفسران ذکر کرده‌اند که این واژه‌ها می‌توانند به تعدادی از مسائل مجزا اشاره کنند[2]. آیا مسئله این است که شرکت‌ها یا عوامل حاکمیتی از به اشتراک گذاشتن الگوریتم‌های‌شان سرباز می‌زنند؟ یا اینکه خود مُدل‌ها بیش از حد فهم انسانی پیچیده‌اند؟ و یا اینکه [ما در بحث‌های‌مان] در مورد همه مردم صحبت می‌کنیم یا صرفاً افرادی را که دارای دانش یا تخصص علمی مربوطه هستند را مدنظر داریم؟ اگرچه به یک معنای ضعیف می‌توان گفت که همگی این پرسش‌ها شامل مسائل مرتبط با شفافیت می‌شوند، اما چالش‌های مختلفی را به وجود آورده و مستلزم راه‌حل‌های گوناگونی نیز هستند.

به‌طور مشابه واژه‌هایی مثل سوگیری، انصاف و تبعیض غالباً برای اشاره به مسائل مربوط به مجموعه داده‌ها یا الگوریتم‌هایی که به‌نحوی برخی اشخاص یا گروه‌ها را نادیده می‌انگارند استفاده می‌شوند.

1. intelligibility (مترجم)
2. Burrell 2016; Lipton (2016); Weller (2017); Selbst & Barocas (2018)

بـاز هـم روشـن نیسـت کـه آیـا همـهٔ مـوارد مـورد اشـاره توسـط ایـن واژه‌هـا، در واقـع مسـئله یکسـانی می‌باشـد یـا خیـر[1].

برخـی پژوهش‌هـا شـروع بـه گره‌گشـایی از ایـن هم‌پوشـانی‌ها کردنـد[2]. بـرای مثـال بـاروکاس (۲۰۱۴) بـر اسـاس داده‌کاوی، سـه نـوع نگرانـی در مـورد الگوریتم‌هـا را از یکدیگـر متمایـز می‌کنـد کـه همگـی ذیـل عنـوان «تبعیـض» مطـرح شـده‌اند:

۱. مـواردی کـه در آن‌هـا به‌کارگیرنـدگان یـک الگوریتـم به‌نحـو قصدمندانـه تـلاش می‌کننـد تـا برخـی کاربـران خـاص را نادیـده گرفتـه و ایـن کار را چنـان انجـام می‌دهنـد کـه شناسـایی دشـوار باشـد (بـرای مثـال از طریـق پنهان‌کـردنِ کدهـای مهـم در یـک الگوریتـم پیچیـده).

۲. مـواردی کـه در آن‌هـا تکنیک‌هـای داده‌کاوی خطاهایـی تولیـد می‌کننـد کـه منجـر بـه نادیـده انگاشـتن برخـی کاربـران می‌شـود (بـرای مثـال بـه سـبب داده‌هـای ورودی اعتمادناپذیـر یـا اسـتنتاجات ناقـص کاربـران از خروجی‌هـای الگوریتـم).

۳. مـواردی کـه در آن‌هـا یـک الگوریتـم توانایـی تصمیم‌سـازان در تمایـز قائل‌شـدن بیـن مـردم را افزایـش می‌دهـد (بـرای مثـال بـه آن‌هـا ایـن امـکان را می‌دهـد کـه بـا دقـت بیشـتری اشـخاص آسـیب‌پذیر اقتصـادی را - بـرای تحقیقـات بیشـتر - شناسـایی و هدف‌گـذاری کننـد).

از آنجـا کـه انـواع و اقسـام موضوعـات، مسـتلزم راه‌حل‌هـای مختلفـی هسـتند، ایـن گونـهٔ خـاصِ متمایزکـردنِ موضوعـات مختلفـی کـه ذیـل یـک واژهٔ خـاص درهم‌آمیختـه شـده‌اند گام اول و مهمـی اسـت بـه سـوی وضـوح مفهومـی.

1. Barocas (2014); Binns (2017)
2. E.g. Barocas (2014); Burrell (2016); Weller (2017); Zarsky (2016); Mittelstadt et al (2016)

۲-۲- تفاوت بین رشته‌ها

چالش دیگر از این واقعیت ناشی می‌شود که برخی از پُراستفاده‌ترین واژه‌ها دلالت‌ها و معانی مختلفی در سیاق‌های مختلف دارند.

برای مثال در آمار، یک «نمونهٔ دارای سوگیری» به معنای نمونه‌ای است که توزیع ویژگی‌ها در یک جمعیت مرجع را به‌نحو مناسب بازنمایی نمی‌کند (مثلاً شاملِ درصد بیشتری از مردان جوان در جمعیت کلی است). در مقابل در حقوق و روانشناسی اجتماعی، واژه سوگیری غالباً به معنای رویکردهای منفی یا پیش‌داوری در مورد یک گروه خاص می‌باشد. از این منظر یک مجموعه دادهٔ عاری از سوگیری (به معنای آماری کلمه) کماکان ممکن است سوگیری‌های رایج (به معنای اجتماعی کلمه) نسبت به افراد یا گروه‌های اجتماعی خاص را کدگذاری کرده باشد. متمایزکردنِ این استفاده‌های مختلف از یک واژه برای پرهیز از تداخل صداها اهمیت دارد[1].

جدای از این مسائل واژگانی، رشته‌های مختلف نیز شامل فرهنگ‌های پژوهشی مختلفی می‌شوند که می‌توانند وضوح و بازتعریف مفاهیم مبهم را تحت تأثیر قرار دهد. به‌عنوان مثال بسیاری از پژوهشگران یادگیری ماشین، به‌طور طبیعی در پیِ ساخت یک تعریف کاملاً دقیقِ ریاضیاتی از مثلاً انصاف[2] هستند، درحالی‌که دانشمندان کیفیِ علوم اجتماعی اغلب به دنبال برجسته‌کردنِ تفاوت‌های عمده در فهم ذینفعان گوناگون می‌باشند. به‌طور مشابه متخصصان اخلاق فلسفی نیز اغلب به دنبال تأکید بر دوراهی‌ها و مسائل بنیادینِ تعاریف مختلف یک مفهوم هستند، حال آنکه بسیاری از وکلا و پژوهشگرانِ سایر رشته‌های معطوف به سیاست‌گذاری، به

1. Barocas & Selbst (2016); London and Danks (2017)
۲. برای مثال: Zafar et al. (2017); Kusner et al. (2017); Kearns et al. (2017).

دنبال تعاریفی عملیاتی هستند که به اندازه کافی برای حل مسائل عملی خوب و مناسب باشد.

این تفاوتِ در رویکرد، تا حدودی ناشی از این است که روش‌های در دسترس رشته‌های مختلف، مناسب حل کدام مسائل هستند، و اینکه از منظر رشته‌های مختلف چه نوع پژوهشی ارزش پیگیری دارد. به علاوه اینکه راهبردهای مختلفِ مفهوم‌سازی، هم‌راستای راهبردهای مختلف حل مسائل اخلاقی و اجتماعی هستند. برای مثال درکِ فنیِ محض از این مسائل، که در آن قضاوت‌های ارزشی صرفاً در راستای مشخص‌کردن / تدقیق مسئله مورد استفاده قرار می‌گیرند، را مقایسه کنید با فهم آن‌ها به‌مثابه مسائل سیاسی که مستلزم مذاکره و مصالحه ذینفعان می‌باشد.

تلاش برای روشن‌سازی مفاهیم کلیدی مربوط به چالش‌های اخلاقی و اجتماعی ADA باید به این تفاوت‌های رشته‌ای عنایت داشته و ناخواسته، به‌صورت پیش فرض، برخی رویکردهای پژوهشی یا سیاست‌گذارانهٔ خاص را به رویکردهای دیگر اولویت ندهد.

۲-۳- تفاوت بین فرهنگ‌ها و عموم مردم

علاوه بر تفاوت‌های رشته‌ای، مفاهیم کلیدی ممکن است در فرهنگ‌های مختلف و در میان عموم مختلفِ مردم، فهم‌ها و دلالت‌های مختلف داشته باشد. مثلاً مفهوم حریم خصوصی را در نظر بگیرید. سنت‌های اخلاقی غرب مدرن (مثلاً کانت‌گرایی) حریم خصوصی شخصی را به‌مثابه یک خیر ذاتی در نظر می‌گیرند، حال آنکه سنت‌های شرقی این‌گونه نیستند. در آیین کنفوسیوس، که در

آن بیش از خیر شخصی، به خیر جمعی تأکید می‌شود، به‌طورسنتی مفهوم حریم خصوصی شخصی (در مقابل خیر جمعیِ مثلاً خانواده) کمتر مورد توجه واقع شده است (و حتی ممکن است دلالت‌های منفی داشته باشد، مثل داشتنِ اسرار شرم‌آور). بودیسم سنتی نیز به‌گونه‌ای متفاوت، باور به نفسِ خودمختار را یک توهم مخرب می‌داند، تا آنجاکه برخی سنت‌های بودیستی، توصیه کردند که افراد به‌نحو فعالانه رازهایشان را به اشتراک بگذارند تا به فقدان نفس / بی‌نفسی دست بیابند[1].

قطعاً مهم است از این فرض نیز اجتناب کنیم که همه افرادِ یک سنت فرهنگی، دارای مفاهیم مشترک هستند و کل یک فرهنگ را می‌توان تقلیل داد به آنچه که در سنت‌های فلسفی یا دینی نیرومند بیان شده است. تا وقتی که همه این‌ها را به‌مثابه «تمایلات» شناسایی کنیم، تحقیق در مورد این تفاوت‌ها برای درک دلالت‌های گوناگونی که مفاهیمِ به کار رفته در مباحث ADA برای گروه‌های مختلف دارند، اهمیت خواهد داشت. با این وجود نیازمند کارهای تجربی بیشتری (مثلاً پیمایش، مصاحبه و مطالعات انسان‌شناختی) روی نوسانات مفهومیِ درون کشوری / درون فرهنگی و بینا کشوری / بینا فرهنگی هستیم.

این نکات نه تنها در مورد فرهنگ‌های مختلف اهمیت دارند - اگر فرهنگ را بر اساس یک ملیت، زبان یا مذهب تعریف کنیم- بلکه مفاهیم اخلاقی و سیاسی کلیدی، ممکن است حتی در بین گروه‌ها یا عمومِ مردمی که با یکدیگر اشتراکاتی دارند - مثل اشتراکات

---

۱. برای اطلاعات بیشتر دربارهٔ تفاوت‌های بین فهم شرقی و غربی از حریم خصوصی به موارد زیر رجوع کنید:
Ess (2006).
The IEEE's Ethically Aligned Design, v.2, pp. 216–193.
که مورد دوم از دلالت‌های ADA چند سنت اخلاقی اعم از سنت‌های سکولار (برای مثال مکتب اصالت سود، اخلاق فضیلت، فریضه‌شناسی اخلاقی) و هم سنت‌های دینی/فرهنگی مثل بودیسم، کنفسیوئیسم، اوبونتوی آفریقا و شینتوی ژاپنی بحث می‌کند.

جنسی، طبقه‌ای، قومیتی و غیره - نیز متفاوت باشد. در این مورد، استدلالِ فمینیست‌های موج دوم که در شعار «امر شخصی، سیاسی است» خلاصه شده، بسیار روشن‌کننده است[1]. فمینیست‌های موج دوم مفهوم سنتی فضای خصوصی به‌مثابه یک فضای شخصی و غیرسیاسی، در برابر فضای عمومی سیاسی، تفکیکی که ریشه‌هایش در تاریخ اندیشه غرب به یونان باستان بازمی‌گردد، را مورد نقد قرار داده‌اند (Burch 2012, ch. 8). این مفهوم‌سازی سنتی، بیش و پیش از هر چیز منجر به این شده است که خانه‌داری غیربازاری و نگهداری از کودکان، در مباحث نیروی کار و اقتصاد چندان مهم در نظر گرفته نشوند (یا خیلی راحت نادیده گرفته شوند؛ برای مثال تأثیری در GDP ندارند)[2]. همچنین منجر به این شده که برخی پدیده‌ها، به حاشیه رانده شود (مثل اذیت و آزار جنسی).

این مثال نشان می‌دهد که چگونه به‌طور خاص بحث در مورد آینده کار و به‌طور کلی، فناوری، باید طیف وسیعی از دیدگاه‌ها را در نظر بگیرد، اگر مایل است که ارزش‌ها و مؤلفه‌های دخیل در مفاهیمی مثل «نیروی کار»، «اوقات فراغت» یا «وقت آزاد» را به‌خوبی درک کند. نکته کلی‌تر اینکه درک و ارزیابی گروه‌های عمومی مختلف در جامعه، از مفاهیم کلیدی مباحث مربوط به ADA متفاوت است. درک این تفاوت‌ها و اطمینان از اینکه ارزش‌های مدنظر تمام اعضای جامعه بازنمایی شده‌اند، برای پیمایش در این مباحث اهمیت کلیدی خواهد داشت.

۱. برای مثال نگاه کنید به: منبع (۱۹۶۹، ۲۰۰۶) Hanisch. استدلال‌ها و شعارهای مشابهی در تعدادی از جنبش‌های سیاسی دهۀ ۱۹۶۰ و ۱۹۷۰ استفاده شد. Crenshaw (1995).

2. GPI Atlantic (1999)

۲-۴- پیچیدگی مفهومی

با این همه، صرفاً متمایزکردنِ استفاده‌ها و تفاسیر مختلف، به خودی خود احتمالاً نتواند این آشفتگی‌های مفهومی را حل کند. اگرچه بسیاری از مفاهیم مهم اخلاقی، به‌طور شهودی واضح و بدون مشکل به نظر می‌رسند، غالباً تحلیل فلسفی، پیچیدگی‌های مفهومی عمیق‌تری را آشکار می‌کند.

مفهوم انصاف را دوباره در نظر بیاورید. این مفهوم غالباً به‌مثابه ارزشی که در سوگیری‌های الگوریتمیک نقش کلیدی دارد در نظر گرفته شده است. واژه‌های «سوگیری» و « انصاف» معمولاً در یکدیگر ادغام شده‌اند، با کمی بحث در مورد مواردی که در آن‌ها سوگیری، به‌مثابه تبعیض نامنصفانه تعریف شده است[1]. با این همه هنوز هیچ اجماع یکپارچه‌ای در فلسفه در مورد تعریف دقیق انصاف وجود ندارد. فیلسوفان سیاسی از تعاریف متعددی دفاع کرده‌اند که مبتنی بر شهودهای متفاوتی از این مفهوم بوده است.

برخی از نظریه‌ها روی توزیع منصفانه دستاوردها بین گروه‌های مختلف تمرکز کردند. البته کماکان باید مشخص کنیم که توزیع منصفانه دستاوردها چه معنایی دارد: زیرنظریه‌های مختلف، استدلال می‌کنند که منصفانه‌ترین توزیع، توزیعی است که نفع کلی را به حداکثر می‌رساند(فایده‌گرایی[2])، یا توزیعی است که تا جای ممکن تساوی‌انگارانه است (برابری‌طلبی[3])، یا توزیعی است که بیشترین منفعت را برای ضعیف‌ترین‌ها دارد (مینی مکس[4]). سایر نظریه‌های انصاف تمرکز کمتری بر یک توزیع خاصِ دستاوردها دارند و در عوض بر نحوه تعریف آن دستاوردها تمرکز می‌کنند: آیا منافع و

1. Friedman and Nissenbaum (1996)
2. utilitarianism (مترجم)
3. egalitarianism (مترجم)
4. minimax (مترجم)

ضررهایی که به یک شخص می‌رسد، محصول انتخاب‌های آزاد خود آن شخص است یا محصول بدشانسی‌هایی مثل ناعدالتی‌های تاریخی معطوف به برخی از اشخاص یا گروه‌های خاص است که کنترل آن‌ها در اختیار فرد نیست[1].

این تفاوت‌ها، در نحوه تفکر ما در مورد تأثیرات ADA بر انصاف، مؤثر است. برای مثال فرض کنید که در حال تحقیق در این مورد هستیم که آیا الگوریتم‌های مورد استفاده در تصمیم‌گیری‌های حوزه سلامت، نسبت به همه بیماران منصفانه هستند یا خیر. بر اساس یک مفهوم برابری‌طلبانه از انصاف، باید این را ارزیابی کنیم که آیا این الگوریتم‌ها پیامدهای یکسانی برای همه کاربران (یا همه زیرگروه‌های مرتبط - که در اینجا باید زیرگروه‌های مرتبط را مشخص کنیم) دارند یا خیر. اما بر اساس رویکرد مینیمکس (حداکثرسازی منافع برای ضعیف‌ترین‌ها) باید مطمئن شویم که این الگوریتم‌ها منجر به بهترین نتایج برای ضعیف‌ترین گروه‌های کاربری می‌شوند، حتی اگر این منجر به تفاوت‌های بیشتر در توزیع نتایج بین گروه‌های مختلف یا بدترشدنِ میانگین توزیع پیامدها بشود. اتخاذ یک مفهوم مبتنی بر انتخاب آزاد از انصاف، ما را ملزم به تعیین / تمایز دقیق شرایط انتخاب آزاد از شرایط شانسی می‌کند. برای مثال آیا سیگار کشیدن یا چاقی یک انتخاب آزاد است؟ صرفِ گفتن اینکه الگوریتم باید منصفانه باشد، نمی‌تواند بین این معانی مختلف از این مفهوم تمایز ایجاد کند[2].

پیچیدگی‌های مفهومی مشابهی را می‌توان در بسیاری از واژگان کلیدیِ

۱. برخی از سیستم‌های قانونی، علیه تبعیض یا برخوردهای ناعادلانه، حمایت‌های خاصی از گروه‌هایی که با «ویژگی‌های موردحمایت» معینی تعریف می‌شوند، انجام می‌دهند؛ ویژگی‌هایی مثل جنسیت، قومیت یا دین. این امر بعضاً بر این اساس توجیه می‌شود که این گروه‌ها به‌لحاظ تاریخی در معرض تبعیض غیرمنصفانه بوده‌اند. بااین‌حال، آنچه باعث می‌شود تبعیض علیه این گروه‌ها به‌ویژه غلط باشد، موردمناقشه است. برای مثال نگاه کنید به:
Altman (2015)

۲. برای بررسی کامل‌تری از نظریه‌های مختلف انصاف و ارتباط آن‌ها با یادگیری ماشینی نگاه کنید به: Binns (۲۰۱۷)

بحث‌های مربوط به تأثیرات ADA شناسایی کرد. اینکه یک سامانه تصمیم‌گیری الگوریتمیک باید قابل فهم باشد، چه معنایی می‌تواند داشته باشد و چرا اهمیت دارد؟ چه چیزی دادهٔ شخصی محسوب می‌شود و چرا محافظت از حریم خصوصی این داده‌ها (در مقایسه با داده‌های غیرشخصی) اهمیت دارد؟ رضایت معنادار چیست؟

درحالی‌که ادبیات فلسفیِ غنی‌ای در مورد اغلب این مفاهیم وجود دارد، کارهای نسبتاً کمی در مورد کاربست آن‌ها در چگونگی بحث‌ها و تأملات ما در مورد دلالت‌های اخلاقی ADA انجام شده است.

اگرچه وضوح و روشن‌سازی گامی مهم در راستای پیشرفت‌های سازنده است، اما همیشه گزینه‌ای سرراست و در دسترس نیست. گاهی اوقات تفاوت در درک مفاهیم کلیدی، نشان‌دهنده عدم توافق‌هایی عمیق‌تر و جدی‌تر بین گروه‌هایی است که به ارزش‌هایی به‌طور بنیادی متفاوت متعهد هستند یا تعارض منافع دارند. برای مثال اگرچه لیبرالیست‌ها یک مفهوم مبتنی بر انتخاب از انصاف را ترجیح می‌دهند، دموکرات‌های اجتماعی یک مفهوم مبتنی بر توزیع را ترجیح می‌دهند، و این صرفاً یک بحث واژگانی نیست. آن‌ها در مورد اولویت‌ها، به‌نحوی بنیادینی با یکدیگر اختلاف دارند.

تحلیل و برجسته‌سازی این اختلافات به تنهایی نمی‌تواند منجر به راه‌حل‌های غیرجنجالی در مورد این عدم توافق‌ها شود. پیمایش چنین عدم توافق‌هایی غالباً مستلزم راه‌حل‌های سیاسی است و نه تحلیل‌های مفهومیِ صرف. برای مثال با طراحی فرایندها یا نهادهای سیاسی که گروه‌های عمومی مختلف، حتی وقتی که در مورد تصمیمات شخصی اختلاف دارند، آن فرآیندها را مشروع در نظر

می‌گیرند. روشن‌سازی و تحلیل مفاهیم کلیدی، می‌تواند مواردی را که در آن‌ها بحث‌ها صرفاً واژگانی هستند متمایز کند و نشان دهد که چه کارهایی برای حل این عدم توافق‌های بنیادین باید انجام شود.

۲-۵- خلاصه و توصیه‌ها

پیشرفت در مباحث مربوط به تأثیرات اخلاقی و اجتماعی ADA، مستلزم جداسازی معانی مختلف واژگان کلیدی مورد استفاده در چارچوب‌بندی این مباحث است. برای پیشرفت در این وظیفه، سه نوع کار باید انجام گیرد:

۱. نگاشت و روشن‌سازی ابهامات

اولین کار، درک تفاوت‌ها و ابهاماتِ استفاده از مفاهیم کلیدی در مباحث ADA است.

یک گام مهم در این جهت، تمرین‌های نگاشت است که در بخش ۳.۱ اشاره شد، که هدف از آن‌ها جداسازی و طبقه‌بندی انواع و اقسام مسائل یا مواردی است که در حال حاضر ذیل یک واژه‌شناسی، در هم آمیخته‌اند. این تمرین‌های نگاشت مستلزم روشن‌سازی (a) تفاسیر و استفاده‌هایِ احتمالی از یک مفهوم مثل شفافیت، (b) اهمیتی که گروه‌ها و جوامع مختلف در عمل به این مفاهیم می‌دهند؛ خواهد بود. برای دستیابی به a، گاهی اوقات برای آشکارکردنِ پیچیدگی‌های مفهومیِ مفاهیم رایجِ مورد استفاده نیازمند تحلیل‌های فلسفی عمیق خواهیم بود. برای دستیابی به b، و به‌منظور استخراج درک‌های متفاوت از مفاهیم مشابه در رشته‌ها

و فرهنگ‌های مختلف، باید با پژوهش‌های تکنیکی مربوطه - برای مثال کارهایی که در مورد تعاریف ریاضیاتی متفاوت و محتملِ انصاف - و نیز پژوهش‌های علوم اجتماعی تجربی آشنا باشیم. این کارهای تجربی را در فصل ۵ مورد بررسی قرار خواهیم داد.

بسیاری از این قبیل کارها بر خوشه‌های مفهومی سوگیری/ انصاف/ تبعیض و شفافیت/ توزیع‌پذیری، فهم‌پذیری / تفسیرپذیری و تا حدودی حریم خصوصی و مسئولیت‌پذیری / پاسخ‌گویی متمرکز بوده‌اند. مطالعات بیشتری از این دست لازم است و باید انجام شود تا به‌گونه‌ای نظام‌مندتر، سایر مفاهیم کلیدی این حوزه (از جمله مفاهیمی که در بخش ۴ بحث می‌کنیم: کرامت انسانی، یکپارچگی، شهروندی، آسایش و خودشکوفایی) را نیز در بر بگیرد.

۲. ایجاد ارتباط بین رشته‌ها، بخش‌ها، گروه‌های عمومی و فرهنگ‌ها

تحلیل و تمرکز بر این پیچیدگی‌ها و واگرایی‌ها خطر تداخل صداها[1] را کاهش می‌دهد. با این همه نیازمند کارهایی برای ایجاد ارتباط بین این تفاوت‌ها هستیم، مثل درگیرکردنِ کاربردها و ذینفعان مرتبط، به‌منظور آگاه‌کردنِ آن‌ها از تفاوت‌های مربوطه و ایجاد فعالانه امکان مراوده بین این شاخه‌ها.

در چارچوبِ ایجاد ارتباط بین رشته‌های مختلف، نگاشت و مراوده تفاوت‌های عملی به شناسایی موقعیت‌هایی که در آن پژوهشگران یا کارورزان نسبت به یکدیگر دچار سوءتفاهم می‌شوند کمک خواهد کرد. ایجاد فعالانه امکان مراوده علاوه بر این مستلزم همکاری‌های

1. cross-talking (مترجم)

بین‌رشته‌ای خواهد بود که در آن پژوهشگران می‌توانند در انتقال یافته‌هایشان به مخاطبین هدف متفاوت، به یکدیگر کمک کنند. در حال حاضر مواردی از این همکاری‌ها در حال تحقق است، ازجمله مقالاتی که به‌طور مشترک بین وکلا و پژوهشگران فنی نوشته می‌شود. این موارد را می‌توان به‌عنوان قالبی برای همکاری‌های بیشتر در نظر گرفت. کارگاه‌هایی که رشته‌های مختلف را برای بحث در مورد مفاهیم کلیدی دور هم جمع می‌کند، می‌تواند مُدل دیگری باشد برای ایجاد ارتباط بین این تفاوت‌های واژه‌شناسانه و زبانی.

بخش عمده‌ای از بحث‌های بین‌المللی کنونی در مورد اخلاق ADA، از کشورهای غربی سرچشمه گرفته و بر اساس سنت‌های فکری غربی صورت‌بندی شده است[1]. با این وجود، در فضاهای فرهنگی دیگری نیز کارها در حال پیشروی هستند، به‌ویژه آسیای شرقی که در خط مقدم پژوهش‌های ADA است. یک گام مهم برای ادغام کامل‌تر طیف دیدگاه‌های مختلف در مباحث بین‌المللی ترجمه اسناد سیاستی و ادبیات پژوهشی مهم از زبان‌های دیگر به زبان انگلیسی و برعکس است. اطمینان از اینکه نمایندگان کشورهای مختلف با پس‌زمینه‌های گوناگون در کنفرانس‌ها و نشست‌های عمده شرکت کنند نیز اهمیت دارد. به علاوه اینکه برای شناسایی پژوهش‌هایی که در سایر کشورها انجام شده است، به‌ویژه کشورهای در حال توسعه که دیدگاه‌هایشان در حال حاضر به خوبی منعکس نشده است، کارهایی باید انجام بشود. از جمله می‌توان به ایجاد همکاری با پژوهشگران و سیاست‌گذاران در این کشورها اشاره کرد.

۱. برای اظهارنظر دربارهٔ اینکه چگونه بحث از این موضوعات در کشورهای توسعه‌یافته متفاوت از کشورهای درحال‌توسعه است، به ضمیمهٔ شمارهٔ ۱ مراجعه کنید.

۳. ایجاد اجماع و مدیریت اختلافات

در نهایت اینکه باید کارهایی برای ایجاد اجماع در مورد بهترین روش‌های مفهوم‌سازی چالش‌های اخلاقی و اجتماعی ناشی از ADA انجام شود. باید در پیِ یافتن درک مشترک و مفاهیم مشترک باشیم. این به معنای تعویض چارچوب‌های موجود در رشته‌ها یا فرهنگ‌ها نیست، بلکه به معنای انتخاب چارچوبی است که ذینفعان مختلف می‌توانند روی آن توافق کنند و اقدام سازنده مشترک شکل بدهند. اگرچه ممکن است همیشه روی یک تعریف دقیق واحد از واژه‌هایِ مهم اخلاق هوش مصنوعی توافق نداشته باشیم، اما می‌توانیم اختلافات معنادار را روشن و از تداخل صداها جلوگیری کنیم.

بسیاری از پیشنهادهایی که در رابطه با نگاشت و روشن‌سازی ابهامات، و ایجاد ارتباط بین رشته‌ها و فرهنگ‌ها مورد بحث قرار گرفتند در این مورد کمک‌کننده خواهند بود. با این وجود اگرچه تحلیل مفهومی سنتی نقطه آغاز بسیار مهمی برای حل ابهامات محسوب می‌شود، ایجاد توافق در مورد تعریف‌ها مستلزم مشارکت و درگیریِ تمام ذینفعانی است که تحت تأثیر فناوری قرار گرفته‌اند، از جمله عموم مردم. ما در بخش ۴.۴.۲ روش‌های ایجاد ارتباط با عموم مردم را مورد بحث قرار می‌دهیم.

باید تأکید کنیم که همه اختلافات اخلاقی مهم را نمی‌توان با ابزارهای مفهومی محض حل کرد. برخی اختلافات مفهومی نشان‌دهنده اختلافات عمیق‌تر رشته‌ای، فرهنگی یا سیاسی هستند. برای بررسی این موارد، لازم است که نحوه مدیریت تنش‌ها، بده-بستان‌ها و دوراهی‌های ناشی از این اختلافات را بدانیم. این‌ها را در

بخش بعدی مورد بررسی قرار می‌دهیم.

بخش سوم

کاوش و بررسی تنش‌ها

بخش سوم

کاوش و بررسی تنش‌ها

هدف از کارِ مفهومیِ توصیف‌شده در بخش ۳، ایجاد وضوح و اجماع در مورد مفاهیم و اصول کلیدیِ اخلاق ADA بود. این یک نقطه آغاز بسیار مهم است اما کافی نخواهد بود اگر نتوان این اصول را در عمل به کار گرفت، و در حال حاضر معلوم نیست که اصول بسیار سطح‌بالایِ پیشنهادشده برای اخلاق ADA بتوانند در موارد انضمامی، اقدامات را هدایت کنند یا خیر. ضمن اینکه کاربست اصول در موارد انضمامی، غالباً آشکارکنندهٔ موانع پیاده‌سازی آن‌هاست: آن‌ها ممکن است از منظر تکنیکی غیرقابل‌اعتماد باشند، بیش از حد دشوار باشند یا اینکه پیاده‌سازی‌شان ارزش‌های دیگر ما را به خطر بیندازد. به‌عنوان مثال تلاش‌های اخیر برای ارائه تعریفی از انصاف که به‌لحاظ ریاضیاتی به اندازه کافی دقیق باشد که بتوان آن را در سامانه‌های یادگیری ماشین اجرا کرد، نشان داده است که غالباً از منظر ریاضیاتی، ارائه تعریفی که دربردارنده همه وجوه شهودیِ مختلف انصاف باشد، ناممکن است[1]. بنابراین معنای اینکه ADA باید در عمل بر اساس اصول انصاف[2] عمل کند، اصلاً مشخص نیست.

برای آنکه بتوانیم درباره فناوری‌های مبتنی بر ADA و آثار این

۱. نگاه کنید به: Friedler et al. (2016); Kleinberg et al. (2016); Chouldechova (2018); Binns (2017)

۲. طبق اصل دوم از کد هوش مصنوعی میان‌بخشی پیشنهادشده توسط کمیتهٔ گزینش لُردها دربارهٔ هوش مصنوعی

فناوری‌ها به‌روشنی بیاندیشیم، باید بر پیگیری و بررسی تنش‌های ناشی از اصول و ارزش‌های مختلف - آنگاه که در عمل قصد به‌کارگیری‌شان را داریم- متمرکز شویم. درحالی‌که برخی مباحث موجود اهمیت مواجهه با این ناسازگاری‌ها را تشخیص داده‌اند، اماهیچ کدام به‌نحو نظام‌مند به آن نپرداخته است. برای مثال بیانیه مونترئال در مورد هوش مصنوعی مسئولیت‌پذیر (۲۰۱۸) بیان می‌کند که اصول این بیانیه «باید به‌گونه‌ای منسجم تفسیر شود تا از ایجاد هرگونه ناسازگاری‌ای که می‌تواند مانع عملی‌شدن آن‌ها شود، پرهیز کرد»، اما معلوم نیست که در عمل چگونه باید از ناسازگاری‌ها پرهیز نمود. به‌طور مشابه کاولز و فلوریدی (۲۰۱۸) تشخیص داده‌اند که استفاده از هوش مصنوعی برای خیر اجتماعی مستلزم «حل تنش‌های ناشی از تلفیق مزایا / منافع و کاهش آسیب‌های بالقوه هوش مصنوعی» است، اما در مورد تنش‌های خاص به‌طور جزئی و نحوه حل آن‌ها چیزی نمی‌گویند.

۳-۱- ارزش‌ها و تنش‌ها

چنان‌که در بخش دوم نیز اشاره شد، مجموعه اصول موجود به تعدادی ارزش متوسل می‌شوند که می‌توانند در کاربست‌های ADA در معرض خطر قرار بگیرند. این ارزش‌ها بیانگر انواع و اقسام اهدافی هستند که ایجاد انگیزه می‌کنند در راستای استفاده یا نگهداری از فناوری‌های مبتنی بر ADA. نکته مهم‌تر اینکه این اهداف چندگانه هستند به جای آنکه یک هدف کلی مثل کاربرد، خیر یا شکوفایی انسانی وجود داشته باشد.

این ارزش‌ها ایده‌آل‌هایی جذاب هستند، اما در عمل می‌توانند ناسازگار باشد به این معنا که اولویت‌بخشی به یک ارزش مستلزم قربانی‌کردنِ ارزش دیگر است. توسعه الگوریتم‌های پیچیده‌تر که توانایی ما را در انجام پیش‌بینی‌های دقیق‌تر در مورد سؤالات مهم بهبود می‌بخشند، می‌تواند منجر به کاهش درک ما از نحوهٔ کار این الگوریتم‌ها شود. استفاده از فناوری‌های مبتنی بر داده نیز می‌تواند دستیابی به سطح مطلوب حریم خصوصی داده‌ها را ناممکن کند. اما اگر دستاوردهای بالقوه این فناوری‌ها به اندازه کافی چشمگیر باشد - مثلاً درمان‌های جدید و بسیار مؤثر سرطان - ممکن است جوامعِ مختلف به این نتیجه برسند که می‌ارزد که هزینه‌ای مثل فداکردنِ حریم خصوصی را بپردازند.

ما از واژه چتریِ «تنش» برای اشاره به انواع و اقسام ناسازگاری بین ارزش‌ها که برخی بنیادی‌تر از برخی دیگر هستند، استفاده می‌کنیم. توجه داشته باشید که وقتی از تنش بین ارزش‌ها صحبت می‌کنیم، منظورمان تنشی است بین ارزش‌های مختلف که در کاربست‌های فناورانه پیش می‌آید نه یک تنشِ انتزاعی بین خودِ ارزش‌ها. برای مثال هدف‌هایی مثل [افزایش] بازدهی و [حفظ] حریم خصوصی، در هیچ یک از سناریوها [ی محتمل] به‌نحو بنیادین با یکدیگر ناسازگار نیستند، اما در سیاقِ برخی فناوری‌های خاصِ مبتنی بر داده، با یکدیگر ناسازگار می‌شوند. اگر همه عوامل بافتاری دست به دست هم بدهند، فناوری‌های مبتنی بر ADA ممکن است بین هر دو (یا چند) ارزش از میان این ارزش‌ها، ناسازگاری ایجاد کنند - یا یک ارزش را، هم‌زمان، به انحاء مختلف، هم تهدید کنند و هم تقویت.

برخـی از ایـن تنش‌هـا آشـکارتر و حائـز اهمیـت بیشـتری هسـتند. در ادامـه برخـی از کلیدی‌تریـن تنش‌هـایِ برخاسـته از کاربسـت‌های رایـجِ فناوری‌هـای مبتنـی بـر ADA را بیـن ارزش‌هـای مختلـف بیـان می‌کنیـم:

• کیفیـت ارائـه خدمـات در برابـر حفاظـت از حریـم خصوصـی: اسـتفاده از داده‌هـای شـخصی، به‌واسـطه متناسـب‌کردنِ خدمـات عمومـی بـر اسـاس ویژگی‌هـای شـخصی یـا جمعیت‌شـناختیِ [کاربـر]، ممکـن اسـت خدمـات عمومـی را بهبـود ببخشـد، امـا ممکـن اسـت بـه دلیـل اسـتفاده زیـاد از داده‌هـا، حریـم خصوصـی شـخصی را بـه خطـر بینـدازد.

• شخصی‌سـازی در برابـر همبسـتگی: افزایـش شخصی‌شـدگی خدمـات و اطلاعـات می‌توانـد منجـر بـه فوایـد اقتصـادی و شـخصی شـود، امـا خطـر مُنشـعب‌کردنِ جامعـه یـا افزایـش انشـعابات آن و متعاقبـاً تضعیف همبسـتگی آن را در پـی دارد.

• راحتـی در برابـر کرامـت: افزایـش خودکارسـازی و کمی‌سـازی می‌توانـد زندگی‌هـا را راحت‌تـر کنـد، امـا درعین‌حـال خطـر تضعیـف آن دسـته از ارزش‌هـا و مهارت‌هـای غیرقابـل کمی‌سـازی را درپـی دارد کـه کرامـت و فردیـت بشـری را تقویـم می‌کننـد.

• حریـم خصوصـی در برابـر شـفافیت: ضـرورتِ در نظـر گرفتـن حریـم خصوصـی یـا مالکیـت معنـوی، می‌توانـد فراهم‌کـردنِ اطلاعـات کامـلاً رضایت‌بخـش دربـاره یـک الگوریتـم یـا داده‌هایـی کـه بـا آن آمـوزش دیـده اسـت را بـا دشـواری مواجـه کنـد.

• دقـت در برابـر توضیح‌پذیـری: دقیق‌تریـن الگوریتم‌هـا ممکـن اسـت بـر اسـاس روش‌هـای پیچیـده‌ای (مثـل یادگیـری عمیـق) ایجـاد شـده باشـند کـه توسـعه‌دهندگان یـا کاربـران آن‌هـا منطـق درونی‌شـان را

به‌طور کامل درک نمی‌کنند.

• دقت در برابر انصاف: ممکن است دقیق‌ترین الگوریتم، به‌گونه‌ای نظام‌مند در قبال برخی از اقلیت‌های خاص، تبعیض قائل شود.

• رضایت از تبعیض‌ها در برابر مساوات: خودکارسازی و هوش مصنوعی می‌توانند صنایع را تقویت کرده و پیشگام فناوری‌های نو باشد، اما در این حال می‌توانند محرومیت و فقر را افزایش دهند.

• بازدهی در برابر ایمنی و پایداری: اشتیاق به پیشرفت فناوری با حداکثر سرعت ممکن است فرصت تأمل در مورد این که آیا این پیشرفت ایمن، نیرومند و قابل اعتماد است را باقی نگذارد.

با توجه به طیف گسترده کاربردهای محتمل فناوری‌های مبتنی بر ADA، و انواع و اقسام ارزش‌هایی که ممکن است تحت تأثیر (مثبت یا منفی) این کاربردها قرار بگیرند، احتمالاً هیچ فهرست ساده و کاملی از تمام تنش‌های احتمالیِ ADA در همه سیاق‌ها وجود نداشته باشد. نگاشت نظام‌مند همه این تنش‌ها فراتر از هدف‌گذاری این گزارش می‌باشد. بنابراین ما بحث خود را به چهار تنش اصلی که در جدول ۱ خلاصه شده‌اند محدود خواهیم کرد.

جدول ۱. تنش‌های کلیدیِ به وجود آمده از کاربردهای ADA

محصولات فناوری‌های ADA	ارزش‌های اصلی که با این محصولات در تنش‌اند	
دقت	انصاف	ارزش‌های اجتماعی
شخصی‌سازی	همبستگی	
کیفیت و بازدهی	خودمختاریِ اطلاعاتی	ارزش‌های فردی
راحتی	خود-شکوفایی	

تنش‌های دو ردیف اول، نشان‌دهندهٔ این هستند که چگونه محصولات به دست آمده از فناوری‌های ADA ممکن است با ایده‌آل‌های اجتماعی انصاف و همبستگی ناسازگار باشند - به همین دلیل ما این تنش‌ها را اجتماعی می‌نامیم:

۱. استفاده از الگوریتم‌ها برای تصمیم‌گیری‌ها و پیش‌بینی‌های دقیق در برابر اطمینان از مواجهه منصفانه و برابر.

۲. توزیع عواید شخصی‌سازی فزاینده در فضای دیجیتالی در برابر تقویت همبستگی و شهروندی.

اولین تنش اجتماعی بین دقت و انصاف، به‌نحو گسترده‌ای در مشاجرات و موردکاوی‌های فناوری‌های مبتنی بر ADA در سال‌های اخیر مورد بحث قرار گرفته است. تنش دوم بین شخصی‌سازی و همبستگی، توجه آشکار کمتری را را جلب کرده است - اما به عقیده ما شأنی بنیادین در اخلاق معطوف به کاربردهای فناوری‌های مبتنی بر ADA در جامعه دارد.

دو ردیف بعدی، معطوف به ایده‌آل‌های زندگی شخصی هستند بنابراین آن‌ها را تنش‌های شخصی می‌نامیم:

۳. استفاده از داده‌ها برای بهبود کیفیت و بازدهی خدمات در برابر احترام به حریم خصوصی و خودمختاری اطلاعاتی افراد.

۴. استفاده از خودکارسازی برای راحت‌ترکردنِ زندگی مردم در برابر ارتقاء خودشکوفایی و کرامت انسانی.

باز هم ما یک تنش را که در حال حاضر به‌طور گسترده پذیرفته

شده است -بین کیفیت و بازدهی خدمات و خودمختاری اطلاعاتی افراد - و یک تنش را که کمتر مورد بحث واقع شده است - بین راحتیِ ناشی از خودکارسازی از یک‌سو و تهدید خودشکوفایی از سوی دیگر - برجسته کرده‌ایم.

البته همه این چهار تنش، شباهت‌های حیاتیِ زیر را دارند: آن‌ها در بخش‌های زیادی به وجود می‌آیند و به عمیق‌ترین ایده‌آل‌های اخلاقی و سیاسی مدرنیته اشاره دارند. در این میان آن‌ها طیف گسترده‌ای از مسائل دیگر را نیز پوشش می‌دهند که ارزش دارد در مورد آن‌ها پژوهش‌های بیشتری برای مواجهه با آثار کاربردهای کنونی و قابل پیش‌بینی ADA انجام شود.

۳-۲- بازخوانی چهار تنش مهم

- تنش ۱. استفاده از الگوریتم‌ها برای تصمیم‌گیری‌ها و پیش‌بینی‌های دقیق‌تر در برابر اطمینان از مواجههٔ منصفانه و برابر

این تنش وقتی به وجود می‌آید که بخش‌های مختلف عمومی یا خصوصی، تصمیمات خود را مبتنی بر پیش‌بینی در مورد آینده رفتار افراد می‌کند (برای مثال وقتی که افسران ناظر، خطر ارتکابِ جرم دوباره را تخمین می‌زنند یا هیئت مدیره مدرسه معلمان را ارزیابی می‌کنند)[1] و از فناوری‌های مبتنی بر ADA برای بهبود پیش‌بینی‌های خود استفاده می‌نمایند. استفاده از ابزارهای کاملاً کمی برای ارزیابی چیزی پیچیده مثل رفتار انسانی یا کیفیت تدریس می‌تواند گمراه‌کننده باشد چراکه این الگوریتم‌ها صرفاً قادر به شناسایی مؤلفه‌های به‌راحتی قابل‌اندازه‌گیری‌اند[2]. با این وجود

1. Angwin, J., et al. (2016)

۲. برای مطالعهٔ بیشتر در این باره، برای مثال مراجعه کنید به کار کتی اونیل (Cathy O'Neil):
www.bloomberg.com/view/articles/2018-06-27/here-s-how-not-to-improve-public-schools

ایـن الگوریتم‌هـا گاهـی اوقـات می‌تواننـد در سـنجش برخـی از معیارهـا بسـیار دقیق‌تـر از روش‌هـای جایگزیـن باشـند، به‌ویـژه اینکـه قضاوت‌هـای انسـانی نیـز گرفتـار سـوگیری‌های نظام‌منـد هسـتند. ایـن مسـئله منجـر بـه ایـن پرسـش می‌شـود کـه آیـا منصفانـه اسـت کـه تصمیمـی را کـه زندگـی یـک فـرد را تحـت تأثیـر قـرار می‌دهـد، مبتنـی کنیـم بـر الگوریتمـی کـه لاجـرم تعمیم‌هایـی انجـام می‌دهـد، تعمیم‌هایـی کـه می‌تواننـد اطلاعـات مهمـی را از دسـت داده و به‌گونـه‌ای نظام‌منـد برخـی گروه‌هـا را نسـبت بـه گروه‌هـای دیگـر نادیـده بگیرنـد. راه دیگـری کـه الگوریتم‌هـا می‌تواننـد انصـاف و برابـری را خدشـه‌دار کنـد ایـن اسـت کـه معمـولاً توضیـح اینکـه ایـن الگوریتم‌هـا چـرا کار می‌کننـد دشـوار اسـت - یـا بـه ایـن دلیـل کـه آن‌هـا مبتنـی بـر روش‌هـای جعبـه سـیاه هسـتند، یـا بـه ایـن دلیـل کـه از نرم‌افزارهـای اختصاصـی اسـتفاده می‌کننـد- بنابرایـن توانایـی افـراد بـرای بـه چالش‌کشـیدنِ ایـن تصمیمـات حیاتـی را تضعیـف می‌نماینـد.

توضیـح فرضـی: یـک دادگاه بـرای کمک‌گرفتـن در تصمیم‌گیـری در مـورد آزادی متهمـان بـا وثیقـه یـا آزادی مشـروط آن‌هـا الگوریتمـی را بـه کار گرفـت کـه خطـر تکـرار جـرم متهمـان جنایـی را تخمیـن می‌زنـد، یعنـی احتمـال ارتـکاب دوبـاره جرم‌شـان را. اگرچـه ایـن الگوریتـم به‌طـور میانگیـن بسـیار دقیـق اسـت، به‌نحـو نظام‌منـد نسـبت بـه مجرمـان سیاه‌پوسـت تبعیـض قائـل می‌شـود، زیـرا مثبت‌هـای غلـط - یعنـی تعـداد افـرادی کـه تحـت عنـوان دارای خطـر بـالا دسـته‌بندی شـده‌اند امـا مرتکـب جـرم دوبـاره نشـده‌اند- بـرای سیاه‌پوسـت‌ها تقریبـاً دو برابـر مجرمـان سفیدپوسـت اسـت[1]. از آنجـا کـه نحـوه عملکـرد درونـی الگوریتـم،

1. Angwin, J., et al. (2016)

رازِ تجـاریِ شـرکت تولیدکننـده می‌باشـد (و در هـر حـال پیچیده‌تـر از آن اسـت کـه هیـچ فـردی بتوانـد از آن سـر در بیـاورد) متهمـان شـانس بسـیار کمـی بـرای اعـاده حیثیـت و بـه چالش‌کشـیدنِ حکمـی دارنـد کـه پیامدهـای عظیمـی روی زندگی‌شـان دارد.

- تنـش ۲. توزیـع عوایـد شخصی‌سـازی روبـه افزایـش در فضـای دیجیتالـی در برابـر تقویـت همبسـتگی و شـهروندی

شـرکت‌ها و حکومت‌هـا حـالا می‌تواننـد از داده‌هـای شـخصی افـراد بـرای اسـتنتاج در مـورد ویژگی‌هـا و اولویت‌هـای [آن افـراد] اسـتفاده کننـد، متعاقبـاً پیام‌هـا، گزینه‌هـا و خدمـاتِ در اختیـار آن‌هـا را متناسـب [بـا ایـن ویژگی‌هـا و اولویت‌هـا] نماینـد. ایـن شخصی‌سـازی پایانـی اسـت بـر راه‌حل‌هـای خـام «یکـی بـرای همـه»، و افـراد را قـادر می‌سـازد تـا محصـولات و خدمـات مناسب‌شـان را پیـدا کننـد، چیـزی کـه فوایـد بالقـوه بسـیاری بـرای بخـش سـلامت و رفـاه می‌توانـد داشـته باشـد. امـا ایـن، ایده‌آل‌هـای هدایت‌گـر دموکراسـی و رفـاه، یعنـی شـهروندی و همبسـتگی را تهدیـد می‌کنـد[۱]. ایـن ایده‌آل‌هـا مـا را دعـوت بـه تأمـل در مـورد خودمـان به‌مثابـه شـهروندان و نـه فقـط مصرف‌کننـدگان فـردی می‌کننـد، و بـه مجهزکـردنِ یکدیگـر در برابـر ضربـاتِ خـارج از کنتـرل فـردی و پیش‌بینی‌ناپذیـر سرنوشـت. تعهـدات عمومـی کـه بـر اسـاس آن‌هـا برخـی محصـولات، فـارغ از تـوان پرداخـت، بایـد در اختیـار همـهٔ شـهروندان قـرار بگیـرد (تحصیـلات، مراقبت‌هـای بهداشـتی، امنیـت، خانـه، معـاش اولیـه، اطلاعـات عمومـی) واجـد نوعـی عـدم قطعیـتِ اصیـل در مـورد ایـن اسـت کـه کـدام یـک از مـا بیمـار خواهـد شـد، شـغلش را از دسـت خواهـد داد یـا بـه انحـاء دیگـر مُتحمـل رنـج خواهـد شـد. ایـن

۱. برای شخصی‌سازی و انسجام در مراقبت‌های سلامت و بیوپزشکی مراجعه کنید به: Prainsack and Buyx (2017)

عدم قطعیت زیربنایِ تعهدات برای جمع‌آوری ریسک است و بدون آن شاهد یک تنش رو به افزایش بین ارتقاء منافع شخصی و کالاهای جمعی خواهیم بود[1].

توضیح فرضی: یک شرکت در حال بازاریابی برای یک طرح بیمه شخصی‌سازی‌شده جدید براساس الگوریتمی است که بر اساس یک مجموعه داده بزرگ آموزش دیده است. این الگوریتم، به‌منظور پیش‌بینی مؤثر آینده پزشکی، آموزشی و الزامات مراقبتی افراد، می‌تواند با دقت بسیار زیاد آن‌ها را از یکدیگر متمایز کند. به این ترتیب شرکت می‌تواند یک مواجهه کاملاً شخصی‌شده ارائه بدهد که تناسب بیشتری با احتیاجات و ترجیحات شخصی افراد دارد. موفقیت این طرح منجر به تضعیف خدماتِ مبتنی بر کمک‌های عمومی می‌شود، زیرا افرادِ برخوردار، دیگر دلیلی برای حمایت از افرادِ با نیازهای بیشتر نمی‌بینند.

- تنش ۳: استفاده از داده‌ها برای بهبود کیفیت و بازدهی خدمات در برابر احترام به حریم خصوصی و خودمختاری اطلاعاتی افراد.

این تنش وقتی به وجود می‌آید که برای بهبود طیفی از خدمات مختلف از یادگیری ماشین و کلان‌داده‌ها استفاده می‌شود: خدمات عمومی مثل مراقبت‌های بهداشتی، آموزش، مراقبت‌های اجتماعی، پلیس یا هر نوع خدمتی که به‌صورت خصوصی ارائه می‌شود. این فناوری‌ها ارائه‌دهندگان خدمت را می‌توانند قادر به متناسب‌سازی دقیقِ خدمات با نیازهای مشتری‌ها کنند، و به این ترتیب کیفیت

۱. یک نمونهٔ اخیر از این تنش مورد زیر است:
DeepMind and the Royal Free hospital: www.theguardian.com/technology/2017/jul/03/google-deepmind-16m-patient-royal-free-deal-data-protection-act

خدمات را افزایش داده و همچنین استفاده پُربازده‌تری از پول پرداخت‌کنندگان مالیات بکنند. با این وجود نیاز به استفاده بسیار زیاد از داده‌های شخصی افراد منجر به نگرانی در مورد ازدست‌رفتنِ حریم خصوصی و خودمختاری افراد بر اطلاعات خودشان می‌شود (ما از عبارت خودمختاری اطلاعاتی برای اشاره به این ارزش استفاده خواهیم کرد).

توضیح فرضی: یک بیمارستان عمومی در ازایِ راه‌اندازی یک الگوریتم یادگیری ماشین که می‌تواند توانایی پزشکان را در درمان سریع و ایمن وضعیت‌های خطرناک به شدت بهبود ببخشد، اجازهٔ دسترسی به داده‌های بیماران را به یک شرکت خصوصی می‌دهد (اسکن‌ها، رفتارها و سوابق پزشکی). این الگوریتم فقط وقتی موفق است که داده‌ها زیاد و قابل انتقال باشند، و این باعث می‌شود که پیش‌بینی اینکه این داده‌ها در کجا مورد استفاده قرار خواهند گرفت دشوار شده و به این ترتیب تضمین حریم خصوصی و اطمینان از رضایت معنادار برای بیماران نیز دشوار می‌شود.

- تنش ۴: استفاده از خودکارسازی برای راحت‌ترکردنِ زندگی مردم در برابر ارتقاء خودشکوفایی و کرامت انسانی.

در حال حاضر بسیاری از فناوری‌های مبتنی بر ADA توسط نهادهای تجاری خصوصی توسعه داده شده‌اند تا جایگزین عملکردهای فعلی شده و راه‌حل‌های مؤثرتر و ساده‌تری را در اختیار بیشترین تعداد مشتری‌ها قرار بدهند. این راه‌حل‌ها، به‌واسطه کاهش مدت

زمـان لازم و نیـز توانمندکـردنِ افـرادی کـه پیش‌تـر از بسـیاری از فعالیت‌هـا محـروم بوده‌انـد، می‌تواننـد زندگـی افـراد را واقعـاً بهبـود ببخشـند. امـا راه‌حل‌هـای خـودکار یکـی از مهم‌تریـن مؤلفه‌هـای انسـانیِ مـا را بـه خطـر می‌اندازنـد[1]. ادبیـات و هنـر مدت‌هـا اضطـراب بیـش از انـدازۀ آدمـی را در تکیـه بـر فنـاوری - تـا آنجـا کـه ظرفیت‌هـای خلاقانـه، فکـری و عاطفـی خـود را از دسـت بدهـد - کاویده‌انـد[2]. ایـن ظرفیت‌هـا، بـرای آن کـه اشـخاص بتواننـد طرح‌هـای زندگـی خـود را به‌نحـو خودمختارانـه و از روی تأمـل محقـق کننـد - ایده‌آلـی کـه غالبـاً تحـت عنـوان خودشـکوفایی و کرامـت بـه آن اشـاره می‌شـود- ضـروری هسـتند. ظهـور بسـیار سـریع سـامانه‌های هـوش مصنوعـیِ بیـش از همیشـه پُربـازده و کامـل، امـکان زوال و فرسـودگی بشـر را- و هراس‌هـای مرتبـط بـا آن: مهارت‌زدایـی، تضعیـف بنیـه، همگن‌سـازی و ازبین‌رفتـنِ تنـوعِ فرهنگـی- آشـکارتر و واقعی‌تـر کـرده اسـت. ایـن هراس‌هـا همچنیـن در جابجایـی نیـروی کار انسـانی و اسـتخدام هـوش مصنوعـی و ربات‌هـا خـودش را نشـان می‌دهـد، در واقـع کارکـردن عـلاوه بـر منبـع معـاش، منبـع معنـاداری و هویـت نیـز می‌باشـد.

توضیـــح فرضـــی: هـوش مصنوعـی سـاخت یـک دسـتیار شـخصیِ همه‌منظـوره را کـه می‌توانـد زبان‌هـای مختلـف را بـه یکدیگـر ترجمـه کنـد، پاسـخ همـه سـؤالات علمـی را در چنـد ثانیـه پیـدا کنـد و بـرای لـذت کاربـر اثـر هنـری یـا اثـر ادبـی تولیـد کنـد امکان‌پذیـر کـرده اسـت. کاربـران آن می‌تواننـد دسترسـی بی‌سـابقه‌ای بـه ثمـره تمـدن بشـری داشـته باشـند امـا آن‌هـا دیگـر نیـازی بـه کسـب و بهبـود ایـن مهارت‌هـا از طریـق تمریـن و آزمایـش منظـم نخواهنـد داشـت. ایـن تمرین‌هـا

۱. تورکل (Turkle (2016 and 2017)) این روندها را عمیقاً مورد بررسی قرار می‌دهد.
۲. داستان کوتاه دیستوپیایی E.M. Forster با عنوان "ماشین متوقف می‌شود "(1909، نگاه کنید به Forster 1947) و فیلم متحرک Wall-E 2008 که این نگرانی را شرح می‌دهند.

رفته‌رفته، بیش از پیش، همگن و غیرقابل‌تغییر می‌شوند آنچنان‌که تنوع گذشته آن‌ها صرفاً در چند فهرست و چند گزینه که بر اساس راحتی و محبوبیت فهرست شدند، بازنمایی می‌شود.

۳-۳- شناسایی تنش‌های بیشتر

چهار تنشی که در بالا طرح شدند برای تفکر در مورد پیامدهای اخلاقی و اجتماعی ADA به‌طور کلی و در وضعیت امروزی اهمیت حیاتی دارند. با این همه تنش‌های دیگری نیز وجود دارند که باید شناسایی شوند، به‌ویژه وقتی که روی وجوه خاصی از اخلاق ADA تمرکز کرده‌ایم، و این واقعیت که تأثیرات فناوری بر جامعه در طول زمان تغییر می‌کند.

رویکرد ما برای شناسایی تنش‌ها با یک فهرست از ارزش‌ها و اصول مهم آغاز می‌شود که تمایل داریم استفاده ما از فناوری‌های مبتنی بر ADA مقید به آن‌ها باشد. سپس در نظر می‌گیریم که در تحقق عملیِ این ارزش‌ها چه مشکلاتی می‌تواند به وجود بیاید، و اینکه استفاده از فناوری چگونه می‌تواند یک ارزش را تقویت یا تهدید کند.

از آنجا که نظرگاه‌های متفاوت، لاجرم تنش‌های متفاوتی را می‌توانند شناسایی کنند، این رویکرد، به‌نحو سودمندی می‌تواند توسط دیگران توسعه یافته و تکرار شود و تنش‌های دیگری که در بالا به آن‌ها اشاره شد را نیز شناسایی نماید. پیوست ۳ نمایانگر برخی شیوه‌های متفاوتِ متمایزساختنِ مسائل، گروه‌های عمومی و بخش‌های مختلف است که می‌تواند برای شناسایی انواع و اقسام

تنش‌ها مورد استفاده واقع شود. از آنجا که انحاء مختلف تهدید یا تضعیف ارزش‌های کلیدی توسط فناوری، در طول زمان تغییر می‌کند و حتی خودِ ارزش‌هایی که ما به‌مثابه جامعه به آن‌ها اولویت داده‌ایم نیز می‌توانند تغییر کنند؛ تکرار این فرآیند در طول زمان اهمیت دارد.

همچنین تأمل در مورد تنش‌ها را می‌توان از طریق ملاحظه نظام‌مند انواع و اقسام روش‌های به وجود آمدنِ این تنش‌ها توسعه داد. ما در اینجا برخی از ابزارهای مفهومی لازم برای این کار را معرفی می‌کنیم:

• **برنده‌ها در مقابل بازنده‌ها**. گاهی اوقات تنش‌ها به این دلیل ایجاد می‌شوند که هزینه‌ها و فواید فناوری‌های مبتنی بر ADA به‌نحو منصفانه‌ای بین گروه‌ها و جوامع مختلف توزیع نشده‌اند.

- فناوری‌ای که اکثریت را منتفع می‌کند، ممکن است به‌گونه‌ای نظام‌مند در قبال یک اقلیت خاص، تبعیض قائل شود: الگوریتم‌های پیش‌بینی‌کننده در یک ساختار بهداشتی- درمانی ممکن است که نتایج را در مجموع بهبود ببخشند، اما برای گروه اقلیت بدتر کنند، برای مثال گروهی که داده‌های بازنمایاننده برای آن‌ها به‌راحتی در دسترس نیست.

- خودکارسازی ممکن است که زندگی برخوردارترین انسان‌ها را غنی‌تر کند، آن‌ها را آزادتر کند تا بتوانند فعالیت‌های ارزشمندتری انجام بدهند، درحالی‌که معیشت کسانی را که شغل‌شان (توسط دستگاه‌های خودکار) از دست رفته و هیچ گزینه دیگری ندارند تهدید کند. علاوه بر تغییر در توزیع منابع مادی، حیثیت و اعتبار، قدرت و نفوذ سیاسی نیز تحت تأثیر قرار می‌گیرند.

• **بلندمـــدت در برابـــر کوتاه‌مـــدت.** تنـش بـه وجـود می‌آیـد، چراکـه ارزش‌هـا یـا فرصت‌هایـی کـه می‌تواننـد در کوتاه‌مـدت توسـط فناوری‌هـای مبتنـی بـر ADA تقویـت بیابنـد، احتمـالاً در بلندمـدت، سـایر ارزش‌هـا را تحـت تأثیـر قـرار دهنـد. بـرای مثـال:

- فنـاوری‌ای کـه زندگـی مـا را در کوتاه‌مـدت بهتـر و راحت‌تـر می‌کنـد، می‌توانـد در بلندمـدت تأثیـرات پیش‌بینی‌ناپذیـری بـر ارزش‌هـای اجتماعـی داشـته باشـد: بـرای مثـال، چنان‌کـه در بـالا هـم اشـاره شـد، افزایـش شخصی‌سـازی می‌توانـد زندگـی مـا را سـاده‌تر و راحت‌تـر کنـد، امـا از طـرف دیگـر ممکـن اسـت خودمختـاری، برابـری و همبسـتگی را در بلندمـدت تضعیـف نمایـد.
- سرعت‌بخشـیدن بـه نـوآوری می‌توانـد منافـع زیـادی را بـرای کسـانی کـه امـروزه زندگـی می‌کننـد خلـق کنـد، درحالی‌کـه تهدیدهـای بزرگ‌تـری را در بلندمـدت بـه وجـود مـی‌آورد. نوعـی بده-بسـتان وجـود دارد. بیـن به‌دسـت‌آوردنِ هـر چـه سـریع‌تر منافـع هـوش مصنوعـی، و جدی‌گرفتـنِ ایمنـی و ثبـات سـامانه‌های پیشـرفته.

• **محلـی در برابـری جهانـی.** تنش‌هـا ممکـن اسـت وقتـی بـه وجـود بیاینـد کـه کاربردهایـی کـه از منظـر محـدود یـا فـردی قابـل دفـاع هسـتند، پیامدهـای نامطلـوب ایجـاد کننـد، مسـائل جمعـی موجـود را تشـدید یـا مسـائل جدیـد خلـق کننـد. بـرای مثـال:

- فنـاوری‌ای کـه بـرای بـرآوردن نیازهـای فـردی بهینـه شـده اسـت، می‌توانـد در سـطح جمعـی تهدیدهـای پیش‌بینی‌نشـده ایجـاد کنـد: یـک الگوریتـم مراقبـت بهداشـتی ممکـن اسـت پیشـنهادی

علیه واکسیناسیون افراد دهد که می‌تواند تأثیرات نامطلوب عظیمی بر سلامت جهانی داشته باشد.

۳-۴- حل تنش‌ها

ما تا اینجا از تک واژه «تنش» برای اشاره به انواع و اقسام ناسازگاری بین ارزش‌ها استفاده کرده‌ایم که برخی بنیادی‌تر و برخی صرفاً عملی هستند. از آنجا که این تفاوت‌ها در حل این تنش‌ها تأثیر دارند، پیش از بحث در مورد راه‌حل‌ها، آن‌ها را برمی‌شماریم.

انواع تنش‌ها

ناسازگاری اخلاقی اصلی یک دوراهی واقعی است. دوراهی واقعی، یک ناسازگاری بین دو یا چند وظیفه، التزام یا ارزش است که یک عامل برای پیگیری هر دو دلایلی دارد اما نمی‌تواند (هر دو را هم‌زمان پیگیری کند). در این دوراهی‌ها هیچ راه حل اصیلی وجود ندارد زیرا خودِ آن کنشی که یک ارزش را تقویت می‌کند (برای مثال وظیفه آنتیگونه برای دفن برادر مرده خویش) ارزش دیگر را تضعیف می‌نماید (وظیفه‌اش در پیروی از شاه). ما این موارد را دوراهی‌های واقعی می‌نامیم، زیرا ناسازگاری در ذاتِ ارزش‌های مورد بررسی وجود دارد و بنابراین از طریق راه‌حل‌های عملی هوشمندانه‌تر نمی‌توان از آن‌ها اجتناب کرد. گاهی اوقات تنش‌هایی که در بالا در مورد آن‌ها بحث کردیم، شکل نوعی دوراهی را به خود می‌گیرند که در آن به‌کارگیری یک فناوری خودکار جدید بدون تضعیف و بی‌ارزش‌کردنِ برخی مهارت‌ها و قابلیت‌های خاص انسان اساساً ناممکن است. در

دوراهی‌هـای واقعـی بـرای اولویت‌بخشـی بـه یـک مجموعـه از ارزش‌هـا مثـلاً سـرعت، بازدهـی و راحتـی، در برابـر ارزش‌هـای دیگـر، مثـل موفقیـت یـا حریـم خصوصـی، ناگریـز از انتخـاب هسـتیم.

گاهـی اوقـات امـا آنچـه کـه در ابتـدا یـک انتخـاب سـخت بـه نظـر می‌رسـد، انتخابـی کـه فداکـردنِ ارزش‌هـای مهـم را ناگریـز می‌کنـد، در واقـع انتخـاب نیسـت. ادعـای وجـود یـک دوراهـی می‌توانـد اغراق‌شـده یـا بررسی‌نشـده باشـد، ماننـد زمانـی کـه یـک شـرکت ادعـا می‌کنـد کـه بایـد حریـم خصوصـی را فـدا کـرد، بی‌آنکـه مطالعـه مناسـبی در مـورد شـیوه‌ها و راه‌حل‌هـای بدیـل کـرده باشـد. در بسـیاری از مـوارد تنشـی کـه بـا آن مواجـه هسـتیم، یـک دوراهـی عملـی اسـت کـه در آن دوراهـی بنیادینـی وجـود نـدارد.

بـا توجـه بـه قابلیت‌هـا و قیـود فناورانـه کنونـی، از جملـه زمـان و منابعی کـه بـرای یافتـن یـک راه‌حـل در اختیـار داریـم، تنش بیـن شـفافیت و دقت یـک مثـال روشـن‌کننده اسـت. ایـن دو ایـده‌آل علی‌الاصـول بـا یکدیگـر ناسـازگار نیسـتند (چنان‌کـه برخـی تعاریـف ناسـازگار انصـاف [ناسـازگار] هسـتند). در اینجـا ناسـازگاری، بیشـتر عملـی اسـت؛ به‌طـور کلـی، تولیـد دقیق‌تریـن الگوریتـم ممکـن منجـر بـه اسـتدلال‌های پیچیده‌تـر خواهـد شـد کـه فهـم کامل‌شـان بـرای انسـان‌ها دشـوار اسـت. بـا ایـن وجـود ایـن پرسـش گشـوده تجربـی اسـت کـه تـا چـه حـد مجبـور بـه انجـام بده-بسـتان بیـن ایـن دو ایـده‌آل هسـتیم، و هـم اکنـون درحـال توسـعه روش‌هایـی بـرای افزایـش شـفافیت بـدون فداکـردنِ دقـت می‌باشـیم[1].

ایـن به‌نوبه‌خـود روشـن می‌کنـد کـه برخـی تنش‌هـای ظاهـری ممکـن اسـت درواقـع دوراهی‌هـای نادرسـت باشـند. این‌هـا موقعیت‌هایـی هسـتند

۱. برای مثال نگاه کنید به: Adel et al. (2018)

کـه در آن‌هـا فـرای انتخـاب بیـن دو ارزش مهـم، مجموعـه سـومی از گزینه‌هـا وجـود دارد. مـا می‌توانیـم زمـان و منابـع بیشـتری بـرای توسـعه راه‌حلـی اختصـاص دهیـم کـه هیـچ یـک از ارزش‌هـا را فـدا نمی‌کنـد یـا اینکـه به‌کارگیـری فنـاوری جدیـد را بـه تعویـق بیندازیـم، تـا وقتـی کـه پژوهش‌هـای بیشـتر، فناوری‌هـای بهتـری را در اختیارمـان قـرار دهنـد. دوراهی‌هـای غیـر واقعـی وقتـی پیـش می‌آینـد کـه مـا یـا از شناسـایی ایـن نکتـه بـاز می‌مانیـم کـه قابلیت‌هـای فناورانـه کنونی‌مـان در واقـع تـا چـه انـدازه قـادر بـه حـل یـک تنـش هسـتند، یـا هیـچ اضطـرار اجبارکننـده‌ای بـرای به‌کارگیـری بلافاصلـه یـک فنـاوری خـاص وجـود نـدارد.

بهتریـن رویکـرد بـرای حـل یـک تنـش، بسـتگی بـه اهمیـت آن تنـش خواهـد داشـت.

بده-بستان‌ها و دوراهی‌های واقعی

تـا وقتـی کـه بـا یـک دوراهـی واقعـی بیـن دو ارزش مواجـه باشـیم هـر راه‌حلـی مسـتلزم انجـام یـک بده-بسـتان بیـن آن ارزش‌هـا خواهـد بـود: انتخـاب و اولویت‌بخشـی بـه یـک ارزش، بـه قیمـت یـک ارزش دیگـر. بـرای مثـال، اگـر مشـخص کردیـم کـه هیـچ یـک از تنش‌هایـی کـه در بـالا اشـاره کردیـم بـا ابزارهـای عملـی قابـل حـل نیسـتند، آن‌گاه نیازمنـد انجام بده-بسـتان‌های ذیـل خواهیـم بـود:

- بده-بسـتان ۱: قضـاوت در ایـن مـورد کـه اسـتفاده از الگوریتمـی کـه وضعیـت را بـرای یـک زیرگـروه خـاص بدتـر می‌کنـد در چـه صورتـی قابـل قبـول اسـت، بـا ایـن فـرض کـه آن الگوریتـم به‌طـور میانگیـن بـرای کل جمعیـت دقیق‌تـر اسـت.

• بده-بستان ۲: قضاوت در این مورد که شخصی‌سازی تبلیغات و خدمات عمومی را، به نفع حفظ ایده‌آل‌های شهروندی و همبستگی، تا چه حد باید محدود و مقید کنیم؟
• بده-بستان ۳: قضاوت در این مورد که به‌خطرافتادنِ حریم خصوصی به نفع پایش بهتر بیماری یا سلامتی عمومی بیشتر، تا چه اندازه قابل قبول است؟
• بده-بستان ۴: قضاوت در این مورد که چه نوع مهارت‌هایی همواره باید در اختیار انسان‌ها باقی بماند، و متعاقباً کدام نوآوری‌ها در فناوری‌های خودکارسازی باید رد شود.

پرسش دشوار این است که چگونه باید این قضاوت‌های مبتنی بر بده-بستان را انجام دهیم. در تجارت و اقتصاد، راه‌حل بده-بستان‌ها به‌طور سنتی از تحلیل هزینه - فایده استخراج می‌شود. که بر اساس آن همه هزینه‌ها و فایده‌های یک سیاست خاص، به یک واحد ثابت در مقیاس ثابت تبدیل می‌شوند (مثل پول یا هر مقیاس کاربردی دیگر مثل رفاه) و بر این اساس که آیا فایده‌ها بر هزینه‌ها می‌چربند یا خیر، توصیه‌هایی ارائه می‌شود. این روش‌ها تقریباً در سراسر حکمرانی صنعت و تجارت مورد استفاده‌اند، زیرا رویه‌هایی روشن و اهدافی مشخص را فراهم می‌آورند. وسوسه‌کننده است که همین روش‌ها را به دوراهی‌های بالا نیز منتقل کنیم.

ما در برابر این وسوسه هشدار می‌دهیم. تحلیل هزینه - فایده، می‌تواند بخشی از فرایند تحقیق در مورد بده-بستان‌ها باشد. این فرایند، شفاف و مکانیکی است و می‌تواند داده‌های سودمند برای تصمیم‌گیری تولید کند. اما CBA به تنهایی نمی‌تواند جوابگو باشد:

تکنوکراتیک است، این واقعیت را که ارزش‌ها مبهم و غیرقابل‌کمی‌سازی هستند، و اینکه خود اعداد می‌توانند قضاوت‌های ارزشی مناقشه‌برانگیز را پنهان کنند و نهایتاً این را که خود ارزش‌گذاری اقتصادی یک خیر، می-تواند رویکرد انسان‌ها نسبت به آن را تغییر دهد (این توضیح می‌دهد که چرا کاربرد CBA در مسائل زیست‌محیطی یا سایر کالاهای عمومی و پیچیده، محل بحث‌های فراوان بوده است.) در نظر نمی‌گیرد[1].

بسته به آرایش‌های سیاسی مختلف، حل این دوراهی‌ها می‌تواند شکل‌های مختلفی به خود بگیرد. یکی از رویکردهایی که مایلیم برجسته‌اش کنیم (رویکردی که به موارد بحث در بخش ۳ نیز مرتبط است) این است که مشروعیت هر راه‌حلِ در حال ظهوری، از طریق مشورت و تأمل عمومی قابل دستیابی است. روش‌هایی برای پیاده‌سازی چنین تأملاتی - درحالی‌که گروه‌های کوچکی از شهروندان به‌واسطه بحث‌های متخصصان و ناظران هدایت و راهنمایی می‌شوند - در علوم سیاسی و در پژوهش‌های محیط‌زیستی و پزشکی که در آن‌ها مشارکت عمومی اهمیت دارد، در حال ظهور است[2]. در مورد فناوری‌های مبتنی بر ADA، چنین مشورت‌هایی هنوز جا نیفتاده‌اند اما بسیار مورد نیازند[3]. اهداف این مشورت‌ها می‌توانند این‌طور باشد:

I. دادنِ فرصت به همه ذینفعان و تعیین علایق آن‌ها با جدیت و احترام (داده‌های مربوط به هزینه‌ها و فایده‌های فناوری‌ها، در این مورد می‌تواند مفید باشد).

II. شناسایی بده-بستان‌های قابل قبول و مشروع که با حقوق و

تکالیف افرادی که تحت تأثیر این فناوری‌ها قرار گرفته‌اند سازگار باشد.
III. دستیابی به راه‌حل‌هایی که اگرچه بی‌نقص نیستند، اما دست‌کم از منظر عموم، قابل دفاع‌اند.

چنین رویکردی وقتی که با یک انتخاب تراژدیک بین ایده‌آل‌های متفاوت فضیلت و زندگی خوب مواجه می‌شود، می‌پذیرد که داوری، اعتراض، رقابت و دستیابی به اجماع، همه و همه غیرقابل‌اجتناب بوده و هیچ فرایند تکنوکراتیکی نمی‌تواند جایگزین آن‌ها شود[1].

دوراهی‌ها در عمل

از سوی دیگر، تا آنجا که در عمل با یک دوراهی مواجه هستیم، فاقد دانش و ابزارهای لازم برای تقویت ارزش‌های ناسازگاریم، بی‌آنکه یکی از آن‌ها را فدا نکنیم. در این مورد بسته به این که یک سیاست یا یک فناوری را با چه سرعت یا با استفاده از کدام منابع می‌خواهیم پیاده کنیم، بده-بستان‌ها می‌توانند قابل‌اجتناب یا غیرقابل‌اجتناب باشند. برای مثال ممکن است روش‌های مبتنی بر داده برای بهبود بازدهی خدمات عمومی و تضمین سطح‌بالای حریم خصوصی اطلاعاتی، علی‌الاصول امکان‌پذیر باشند، ولی در حال حاضر در دسترس نباشند. در مورد هر یک از چهار تنشی که مورد اشاره قرار گرفتند، ممکن است که با دانش بیشتر یا روش‌های بهتر آن‌ها را حل یا دست‌کم تضعیف‌شان کرد. در این موارد ما با یک انتخاب مواجه هستیم:

- استفاده از فناوری در وضعیت کنونی. در این مورد، لازم است که برخی بده-بستان‌های مشروع را که یک ارزش را فدای سایر ارزش‌ها می‌کند تعیین و پیاده‌سازی نماییم.

۱. دربارهٔ نقش حیاتی مشورت و رقابت تخصص‌ها در دموکراسی‌ها نگاه کنید به: Moore (2017) and Alexandrova (2018)

• متوقف‌کردنِ استفاده از این فناوری و سرمایه‌گذاری در پژوهش در مورد اینکه چگونه می‌توان کاری کرد که فناوری مزبور در خدمت همه ارزش‌هایی باشد که ما به تساوی و قوت تأییدشان می‌کنیم.

این انتخاب، تنش خاص خود را دارد، تنشی از نوع کوتاه‌مدت در برابر بلندمدت که در بخش ۳-۴ مورد بحث واقع شد: تا کجا باید منافع فناوری‌های جدید را نادیده بگیریم و به جای آن وقت و منابع را صرف حل بهتر این تنش‌ها کنیم؟ این البته یک انتخاب دوراهی نیست. بلکه ما باید در پی ایجاد تعادل باشیم: تلاش برای پیمایش بده-بستان‌های لازم برای تصمیم‌گیری در مورد چگونگی استفاده از فناوری‌های روز، و هم‌زمان سرمایه‌گذاری در پژوهش برای پی بردن به اینکه آیا این تنش در آینده به‌طور کامل حل خواهد شد.

درک بهتر تنش‌ها

چنان‌که این بحث تأکید می‌کند، به‌منظور پیشرفت در حل تنش‌های برخاسته از فناوری‌های مبتنی بر ADA لازم است که درک روشنی از ماهیت تنش‌ها داشته باشیم. آیا آن‌ها دوراهی‌های واقعی هستند یا دوراهی‌های عملی و یا دوراهی‌های کاذب؟[۱]

بنابراین در کنار توسعه روش‌هایی برای ایجاد تعادل در بده-بستان‌ها و سرمایه‌گذاری روی فناوری بهتر، ما باید به‌منظور درک بهتر ماهیت تنش‌های مهم، روی تحقیق و پژوهش نیز سرمایه‌گذاری کنیم. این کار را می‌توانیم با پرسیدن سؤالات زیر انجام دهیم:

• آیا می‌توان طوری از دقیق‌ترین الگوریتم‌های پیش‌بینی‌کننده

۱.موارد عینی می‌تواند ترکیبی از هر سه نوع دو راهی را شامل شود، هنگامی که ما در یک سطح دقیق‌تر، بین ارزش‌های مختلفی که توسط ذینفعان مختلف در یک مورد مشخص وجود دارد، تمایز قائل شویم.

استفاده کرد که انصاف و مساوات را نادیده نگیرند؟ الگوریتم‌های پیش‌بینی‌کننده‌ای را در نظر بگیرید که در حال حاضر مورد استفاده هستند (برای مثال در حوزه سلامت، جرم و استخدام) این الگوریتم‌ها تا چه اندازه، در مورد اقلیت‌های خاص تبعیض قائل می‌شوند؟

• آیا می‌توان از منافع شخصی‌سازی بهره‌برداری کرد، بدون آنکه شهروندی و همبستگی را تضعیف نمود؟ انواع مختلف شخصی‌سازی، چگونه در آینده این ایده‌آل‌های مهم را تضعیف خواهند کرد؟ چگونه می‌توان این را بررسی یا از آن اجتناب کرد؟

• آیا می‌توان از داده‌های شخصی برای بهبود کیفیت و بازدهی خدمات عمومی استفاده کرد، بی‌آنکه به خودمختاری اطلاعاتی لطمه‌ای وارد ساخت؟ روش‌های دقیق، تا چه اندازه امکان استفاده از داده‌های شخصی را به نفع منافع اجتماعی فراهم می‌آورند، درحالی‌که از حریم خصوصی اشخاص نیز محافظت می‌کنند؟

• آیا خودکارسازی می‌تواند بدون تهدید خودشکوفایی افراد، زندگی‌ها را راحت‌تر کند؟ آیا می‌توانیم یک خط فارق روشن بین رفتارهایی بکشیم که در آن‌ها خودکارسازی سودمند است یا کمترین آسیب را خواهد داشت، و قابلیت‌هایی که اساساً نباید خودکار شوند؟

پاسخ به این پرسش‌ها تا حدودی مستلزم پژوهش‌های مفهومی از سنخ آنچه در بخش ۳ بحث شد خواهد بود. برای مثال، مشخص‌کردن مهم‌ترین نوعِ انصافِ الگوریتمیک یکی از گام‌های اولیه بسیار مهم به سوی این است که آیا این مسئله را با ابزارهای فنی می‌توان

حل کرد یا خیر. به‌علاوه، از آنجا که پرسش‌های عمدتاً تجربی در مورد اینکه در واقع چه چیزی امکان‌پذیر است وجود دارد، پاسخ به آن‌ها چنان‌که در بخش بعدی مورد بحث واقع خواهد شد مستلزم تکیه بر شواهدی است در مورد اینکه چه چیزی از منظر فنی امکان‌پذیر است. در برخی موارد، فناوری کنونی یا آینده می‌تواند تنش مورد بحث را رفع یا کم‌رنگ کند.

دست آخر اینکه این تنش‌ها فقط وقتی در عمل حل می‌شوند که نهادها، قوانین و ساختارهای حکمرانی مناسبی برای حمایت از این تلاش‌ها وجود داشته باشد. استانداردها، مقررات و نظام‌های حکمرانی معطوف به فناوری‌های ADA، در حال حاضر گرفتار عدم قطعیت‌های فراوانی در مورد آینده‌شان هستند[1]. ما تأکید می‌کنیم که رویکردهای جدید به حکمرانی و قانون‌گذاری باید به درستی نسبت به تنش‌هایی که در بالا اشاره شد حساس باشند و نهادهای مشروعی را برای پیمایش همه تنش‌های ممکن در تمام سطوح به وجود بیاورند.

۳-۵- خلاصه و پیشنهادات

در این بخش، ایده و اهمیت تفکر در مورد تنش بین ارزش‌ها را در اصول حاکم بر اخلاق ADA معرفی کردیم، تا بتوانیم از اثرگذاری عملی این اصول، مطمئن باشیم. پیشنهادات سطح‌بالای ذیل را می‌توان ارائه داد:

- تمرکز اخلاق ADA را به سمت شناسایی تنش‌های برخاسته از پیاده‌سازی اخلاقی ADA ببرید.

۱. برای مطالعهٔ عدم قطعیت پیرامون دلالت‌ها و اجرای عملی GDPR نگاه کنید به: Wachter and Mittelstadt (2018)

چهـار تنشـی کـه به‌عنـوان اولویـت در بخـش ۲-۴ مطـرح شـد، دربردارنـده مشـاجرات و موردکاوی‌هایـی هسـتند کـه مفسـران در حوزه‌هـا و بخش‌هـای گوناگـون شـروع بـه بررسی‌شـان کرده‌انـد. بنابرایـن، مجموعـه پیشـنهادات بعـدی مـا عبارتنـد از:

- مصادیـق چهـار تنشـی را کـه در ایـن گـزارش مـورد تأکیـد قـرار گرفته‌انـد در بخش‌هـای مختلـف جامعـه بررسـی کنیـد و مـوارد خاصـی را جسـتجو کنیـد کـه در آن‌هـا ایـن تنش‌هـا بـه وجـود می‌آینـد:
  - اسـتفاده از الگوریتم‌هـا بـرای تصمیم‌گیـری و پیش‌بینـی دقیق‌تـر در برابـر اطمینـان از رفتـار منصفانـه و برابر.
  - توزیـع منافـع شخصی‌سـازی روزافـزون در فضـای دیجیتـال در برابـر توسـعه یکپارچگـی و شـهروندی.
  - اسـتفاده از داده‌هـا بـرای بهبـود کیفیـت و بازدهـی خدمـات در برابر احتـرام بـه حریـم خصوصـی و خودمختـاری اطلاعاتی اشـخاص.
  - اسـتفاده از خودکارسـازی بـرای راحت‌ترکـردنِ زندگـی افـراد در برابـر ارتقـای خودشـکوفایی و کرامـت.
- بـا کمـک پرسـش‌های زیـر، تنش‌هـای بعـدی و علـل بنیادی‌شـان را بـر اسـاس ناسـازگاری‌های ارزشـی دیگـر شناسـایی کنیـد:
  - در کجـا ممکـن اسـت هزینه‌هـا و فایده‌هـای فناوری‌هـای مبتنـی بـر ADA، بـه تسـاوی توزیـع نشـده باشـند؟
  - در کجـا ممکـن اسـت منافـع کوتاه‌مـدت، بـه قیمـت ارزش‌هـای بلندمـدت تمـام شـود؟
  - در کجـا ممکـن اسـت فناوری‌هـای مبتنـی بـر ADA، اشـخاص یـا گروه‌هـا را منتفـع کننـد ولـی در سـطح جمعـی، مشکل‌سـاز باشـند؟

شرح تنش‌هایی که می‌توانند به یک مورد خاص اِعمال شود اولین گام در جهت به‌کارگیری فناوری‌های اخلاقی است، گام بعدی اما، حل این تنش‌هاست. اینکه چگونه این کار را انجام می‌دهیم بستگی به ماهیت تنش مورد نظر دارد. بنابراین پیشنهاد می‌دهیم که تحقیقات بعدی معطوف باشند به:

- شناسایی اینکه تنش‌های کلیدی، تا چه حد شامل دوراهی‌های واقعی، دوراهی‌های عملی یا دوراهی-های کاذب می‌شوند. این، اغلب مستلزم بررسی مصادیق خاص تنش و ملاحظه شیوه‌های حل آن است بدون فداکردنِ هیچ یک از ارزش‌های کلیدی.
- آنجایی که در عمل با دوراهی مواجه شدیم، در این مورد تحقیق کنیم که چگونه می‌توان این دوراهی‌ها را رفع کرد. مثلاً با توسعه مرزهای امرِ امکانپذیرِ تکنیکی چنان‌که بتوانیم ارزش‌های بیشتری را در نظر بگیریم.
- آنجایی که با دوراهی‌های واقعی بین ارزش‌ها، یا دوراهی‌های عملی که اقدام عاجل می‌طلبند مواجه شدیم، درمورد حل دوراهی از طریق مشروع‌سازی بده-بستان در انظار عموم تحقیق کنیم و نیز در مورد نهادهای تنظیمی تحقیق کنیم که به‌ویژه با فناوری‌های ADA منطبق هستند.

بخش چهارم

توسعه پایگاه شواهد

بخش چهارم

توسعه پایگاه شواهد

بحث‌های کنونی در مورد دلالت‌های اخلاقی و اجتماعی ADA تحت تأثیر نامطلوب شکاف‌های فهمِ ما هستند: اینکه به‌لحاظ تکنولوژیک چه چیزی امکان‌پذیر است، اینکه فناوری‌های مختلف چگونه جامعه را تحت تأثیر قرار خواهند داد، و این که بخش‌های مختلف جامعه چه چیزی را خواهند خواست و به چه چیزی نیاز خواهند داشت. برای پیشرفت در استفاده از فناوری‌های مبتنی ADA به نفع خیر جامعه، لازم است که یک پایگاه شواهد قوی‌تر در همه این حوزه‌ها ایجاد کنیم. به‌ویژه ایجاد این پایگاه شواهد قوی‌تر، برای کسانی مثل نهادهای حاکمیتی، قانون‌گذاران و نهادهای استانداردسازی اهمیت خواهد داشت که در حال توسعه چارچوب‌ها و دستورالعمل‌های عملی برای اخلاق هوش مصنوعی هستند.

برای مثال تنش بین استفاده از داده‌ها برای بهبود خدمات عمومی و ضرورت حفاظت از حریم خصوصی شخصی، تا حدودی به این دلیل دشوار است که این مبحث فاقد شواهد مناسب در زمینه‌های زیر است:

- یادگیری ماشین و کلان‌داده تا چه حد می‌توانند خدمات عمومی را

بهبود ببخشند - و با این کار حریم خصوصی شخصی تا چه اندازه و به چه شیوه‌هایی ممکن است آسیب ببینند.

• گروه‌های عمومی مختلف تا چه اندازه، مراقبت‌های بهداشتی بهتر را نسبت به حریم خصوصی داده، ارزشمندتر می‌دانند و در کدام بافتارها، از استفاده از داده‌های‌شان خوشحال خواهند شد.

• پیامدهای بلندمدتِ استفاده روزافزون از داده‌های شخصی توسط مراجع قدرت چه می‌تواند باشد.

حصول اطمینان از اینکه از الگوریتم‌ها، داده‌ها و هوش مصنوعی در راستای منافع جامعه استفاده می‌شود، وظیفه‌ای نیست که یک‌بار و برای همیشه انجام شود، بلکه یک فرآیند دائمی و پیوسته است. این یعنی همان‌قدر که باید قابلیت‌های فناورانه و استلزامات اجتماعیِ امروزین را درک کنیم، نیز باید در مورد این بیاندیشیم که این مسائل چه تحولاتی در آینده خواهند داشت تا به این ترتیب بتوانیم راهبردهایی انطباقی برای درنظرگرفتنِ عدم قطعیت‌های آینده توسعه دهیم.

در این بخش، یک طرح کلی در مورد برخی از حوزه‌های پژوهشی عمومیِ مورد نیاز برای توسعه یک پایگاه شواهد قوی‌تر پیشنهاد خواهد شد، و بر اساس تنش‌هایی که در بخش چهار مورد بحث قرار گرفتند، برخی از پرسش‌های اولویت‌دار برجسته خواهند شد. تمرکز ما بر این است که به چه نوع پرسش‌هایی باید پاسخ داده شود و اینکه جهاتِ عمومی پژوهشی کدامند. اگرچه ما برخی روش‌های امیدبخش را برجسته خواهیم کرد، اما این همه آن چیزی نیست که

می‌تواند وجود داشته باشد. ما در این فصل تلاش نکرده‌ایم تا تمام روش‌های موجود یا ممکن برای مطالعه این پرسش‌ها را پیمایش کنیم، زیرا برای پاسخ به برخی پرسش‌ها ممکن است نیازمند راهبردهای جدید و نوآورانه باشیم. به‌طورکلی، مجموعهٔ متکثری از دیدگاه‌های انضباطی و تفکر روش‌شناختی نوآورانه است که می‌تواند بهترین پایگاه شواهدِ ممکن را فراهم کند.

۴-۱- درک قابلیت‌ها و تأثیرات فناورانه

قابلیت‌های فناورانه- چه چیزی امکان‌پذیر است؟

درک قابلیت‌های فناورانه، برای درک فرصت‌ها و تهدیدات واقعی فناوری‌ها کاملاً حیاتی است. مثلاً برای ارزیابی اینکه مهم‌ترین فرصت‌ها و تهدیدهای نشانه‌رویِ مبتنی بر داده کدامند، لازم است بدانیم که جمع‌آوری و استفاده از داده‌های شخصی‌سازی‌شده برای نشانه‌روی و مداخله شامل چه گام‌های تکنیکی‌ای می‌شود، و اینکه رویکردهای موجود با چه محدودیت‌هایی مواجه هستند[۱]. به‌منظور ارزیابیِ خطر بیکاری ناشی از پیشرفت‌های فناورانه و متعاقباً طراحی سیاست‌های تأثیرگذار برای مواجهه با آن، لازم است بدانیم که ماشین‌ها در حال حاضر در انجام چه نوع وظایفی می‌توانند از انسان‌ها پیشی بگیرند، و اینکه این وضعیت در سال‌های آینده چگونه تغییر خواهد کرد.

دانستن قابلیت‌های فناورانه به ما کمک می‌کند تا در مورد تنش‌های اخلاقی که در بخش چهار به آن‌ها اشاره شد، به شیوه‌های روشن‌تری بیاندیشیم: از طریق نشان‌دادنِ اینکه آیا این

۱. مشخص نیست که بسیاری از دعاوی دربارهٔ استفادهٔ کمبریج آنالیتیکا (Cambridge Analytica) از «psychographic microtargeting» یارای بررسی تکنیکی دقیق داشته باشند، برای مثال نگاه کنید به: Resnick (2018)

تنش‌ها دوراهی‌های واقعی‌اند یا دوراهی‌های عملی، از طریق کمک به تخمین هزینه‌ها و فایده‌های خاص یک فناوری در یک سیاق مشخص، از طریق فراهم‌آوردنِ زمینه‌های لازم برای فهم بده-بستان‌های محتمل بین ارزش‌هایی که یک فناوری آن‌ها را تهدید یا تقویت می‌کند. این نوع از شواهد برای سیاست‌گذاران و قانون‌گذارانی که روی حکمرانی فناوری‌های مبتنی بر هوش مصنوعی کار می‌کنند اهمیت حیاتی خواهد داشت، نیز به پژوهشگران کمک خواهد کرد تا خلأها و اولویت‌ها را برای پژوهش‌های آینده شناسایی کنند.

ما همچنین نیازمند پژوهش در مورد پیش‌بینی قابلیت‌های آینده هستیم که صرفاً به سنجش فناوری‌های موجود بسنده نمی‌کند، به این ترتیب بتوانیم چالش‌های جدید را پیش‌بینی و خودمان را با آن‌ها منطبق کنیم.

در مورد چهار تنش اصلی، پرسش‌های کلیدی که باید پاسخ گفته شوند عبارتند از:

- **دقت در برابر رفتار منصفانه و برابر**

= دقت تا چه اندازه با تعاریف گوناگون انصاف وارد بده-بستان می‌شود؟

= آخرین مدل‌ها چه نوع تفسیرپذیری را مطلوب دانسته و آن‌را تضمین می‌کند؟

= تا چه حد می‌توان تفسیرپذیری مناسب را بدون فداکردنِ دقت (یا سایر ارزش‌ها مثل حریم خصوصی) تضمین کرد؟

- **شخصی‌سازی در برابر همبستگی و شهروندی**

= محدودیت‌های بنیادی و عملی شخصی‌سازی دقیق (بر اساس فناوری‌های کنونی یا قابل پیش‌بینی) کدامند؟

= شخصی‌سازی تا چه حد می‌تواند به‌نحو معنادار، خروجی‌های مهم (مثل اقناع کاربر، رفتار مصرف‌کننده و الگوهای رأی‌دهی) را تحت تأثیر قرار دهد؟

• کیفیت و بازدهی خدمات در برابر حریم خصوصی و خودمختاری اطلاعاتی

= یادگیری ماشین و کلان‌داده تا چه حد می‌توانند خدمات عمومی گوناگون را بهبود ببخشند؟ آیا منافع بالقوه را می‌توان کمی‌سازی کرد؟

= بهترین روش‌های کنونی برای حفاظت از حریم خصوصی کدامند و قیود تکنیکی آن چه هستند؟

• راحتی در برابر خودشکوفایی و کرامت

= چه نوع وظایفی را ممکن است که بتوان با استفاده از فناوری‌های کنونی یا قابل پیش‌بینی، خودکار کرد؟

= هزینه‌های خودکارسازی در سطحِ وسیعِ یک وظیفه مشخص (برای مثال ملزومات زیرساختی و انرژی) چه می‌تواند باشد؟

• همچنین سؤالاتی فراگیر نیز برای بررسی وجود دارد که به همه این چهار تنش ربط پیدا کرده و به سایر تنش‌ها نیز قابل اِعمال است:

= برای درک قابلیت‌ها و محدودیت‌های فناوری، به‌منظور ارزیابی فرصت‌ها و تهدیدهای آن در بافتارهای اخلاقی و اجتماعی گوناگون، نیازمند دانستن چه چیزهایی هستیم؟

= پیشرفت قابلیت‌های فناورانه چگونه می‌تواند در کاربردهای مختلف ADA به حل تنشِ بین ارزش‌ها کمک کند، و محدودیت‌های فناوری برای انجام چنین کاری کدامند؟

این پرسش‌ها در یک سطح کلی صورت‌بندی شدند. برای کمک به حل تنش‌ها در عمل، این پرسش‌ها را باید با دامنه مسئله مشخصی که با آن مواجه هستیم متناسب کنیم، چنان‌که در سناریوی زیر مشخص شده است:

سناریوی فرضی: تصور کنید بخش مراقبت‌های بهداشتی و اجتماعی در حال توسعه دستورالعمل‌هایی است برای تعیین سطح تفسیرپذیری لازم برای الگوریتم‌های مورد استفاده در کاربردهای مختلف مراقبتی و بهداشتی، و نیز نحوه متوازن‌کردنِ آن‌ها در مقابل هزینه‌های بالقوه تأمین دقت. برای اینکه این کار را به خوبی انجام دهد باید دو نکته زیر را بداند:

- چه گزینه‌هایی برای یک کاربرد خاص وجود دارد. به‌عنوان مثال برای تحلیل تصاویر رادیولوژی، از چه مدل‌های مختلفی می‌توان استفاده کرد، و هر یک از این مدل‌ها تا چه اندازه و چگونه تفسیرپذیرند و دقیق‌کردن‌شان چه هزینه‌هایی دارد؟
- هزینه‌ها و فایده‌های مختلف در یک بافتار مشخص. در بعضی موارد، کاهش دقت ممکن است هزینه‌های بسیار زیادی داشته باشد، مثلاً تشخیص غلط می‌تواند زندگی افراد را به خطر بیندازد. و همچنین، بسته به موقعیت (اینکه آیا راه دیگری برای سنجش اعتمادپذیریِ یک الگوریتم داریم یا خیر، یا اینکه تصمیم‌های اتخاذ شده را باید به بیماران توضیح بدهیم) اهمیت انواع و اقسام تفسیرپذیری تفاوت خواهد داشت.

بدون آگاهی از این جزئیات فنی، DHSC در معرض خطر تولید

دستورالعمل‌های بسیار کلی است که در بهترین حالت، پیاده‌سازی‌شان در عمل دشوار یا ناممکن است و در بدترین حالت، آسیب‌زا (برای مثال، توصیه به عدم استفاده مطلق از مدلی که به‌طورکامل تبیین‌پذیر نیست ممکن است ما را از برخی کاربردهای سودمند آن محروم کند).

این پرسش‌ها تا حدودی معطوف به خودِ فناوری هستند و بنابراین مستلزم این است که بدانیم از منظر علوم رایانه، یادگیری ماشین و سایر حوزه‌های پژوهشی فنی چه چیزی امکان‌پذیر است. بسیاری از این حوزه‌ها سریعاً در حال پیشرفت هستند و بنابراین حیاتی است که از وضعیت و پیشرفت‌های فنی رایج کاملاً آگاه باشیم. یکی از راه‌های جمع‌آوری شواهد در مورد قابلیت‌های فناورانه، صحبت‌کردن با متخصصان این حوزه‌ها و یا پیمایش نظرات آن‌هاست[1]. از آنجا که نظرات یک تک پژوهشگر در مورد آخرین پیشرفت‌های یک حوزه نمی‌تواند نماینده باشد، بهتر است که از تعداد زیادی از متخصصان فنی نظرخواهی کنیم. یکی از چالش‌های کلیدی در اینجا، توصیف وضعیت فنی پژوهش‌ها با دقت و جزئیات کافی است، به‌گونه‌ای که برای افراد غیرتکنیکی که در حوزه اخلاق و سیاست‌گذاری این فناوری‌ها مشغولند قابل فهم و سودمند باشد.

این پرسش‌ها اما فراتر از خودِ فناوری هستند، چراکه شامل تأثیرات و پیامدهای این فناوری‌ها بر انسان‌ها نیز می‌شوند. پاسخ کامل به آن‌ها مستلزم پژوهش‌هایی از جنس روان‌شناسی و جامعه‌شناسی است. رشته تعامل انسان-رایانه، با استفاده از روش‌های روانشناسی و علوم اجتماعی بسیاری از پرسش‌های معطوف به تأثیرات فناوری‌های مبتنی بر ADA بر انسان‌ها را مورد مطالعه قرار می‌دهد.

۱. برای مثال نگاه کنید به: Grace et al. (2018)

دست آخر اینکه برخی از این پرسش‌ها نه تنها در مورد قابلیت‌های کنونی فناوری هستند، بلکه معطوف به چگونگی تکامل این قابلیت‌ها در آینده نیز می‌باشند. در حال حاضر، کارهای بسیار خوبی در مورد سنجش و پیش‌بینی قابلیت‌های فناوری وجود دارد[1]. با این همه پژوهش در مورد چالش‌های اخلاقی و اجتماعی این فناوری‌ها می‌تواند استفاده بسیار بیشتری از این کارها کرده و به این ترتیب تضمین کند که درک ما از این چالش‌های گسترده‌تر، از فهمی متقن درباره آنچه به‌لحاظ فنی امکان‌پذیر است- و آنچه می‌تواند امکان‌پذیر باشد- نشأت می‌گیرد.

بنابراین پژوهش در مورد قابلیت‌ها و تأثیرات فناورانه، احتمالاً مستلزم همکاری بین متخصصانی از حوزه‌های فنی ADA، روانشناسی / علوم اجتماعی، آینده‌پژوهی، سیاست‌گذاری و اخلاق و همچنین افرادی که قادر به ترجمه این حوزه‌های مطالعاتی مختلف به یکدیگر هستند، خواهد بود.

۴-۱-۱- استفاده‌ها و تأثیرات کنونی - چه اتفاقی دارد می‌افتد؟

علاوه بر درک امر به‌لحاظ فنی امکان‌پذیر، همچنین نیاز به درک بهتر این موارد داریم: (۱) چگونه فناوری‌های مختلف مورد استفاده قرار می‌گیرند و در واقع پیامدهایی دارند؛ (۲) زیربنای این پیامدها چه نوع علل، سازوکارها یا [عوامل] تأثیرگذارِ دیگری است؟

در مورد نکته اول، در حال حاضر بسیاری از بحث‌های معطوف به اخلاق ADA، یا بر اساس موردکاوی‌ها پیش می‌رود، مثل موردکاوی‌هایی که توسط روزنامه‌نگاران و مفسران اجتماعی آشکار

۱. برای مثال نگاه کنید به: The AI Index, and the Electronic Frontier Foundation's work on AI Progress Measurement

می‌شوند، یا بر اساس سناریوهای فرضی در مورد اینکه فناوری‌ها چگونه می‌توانند مورد استفاده قرار بگیرند. اگرچه این‌ها برای روشن‌کردنِ تأثیرات اخلاقی و اجتماعی بالقوه فناوری‌های مبتنی بر ADA حیاتی هستند اما مشخص نیست که تا چه حد نمایانگرِ توسعه‌های کنونی یا آینده می‌باشند. خطر اغراق‌کردن در مورد برخی از کاربردها و تأثیرات و دست‌کم‌گرفتنِ برخی دیگر وجود دارد.

یک نوع پژوهش مهم می‌تواند نگاشت و کمی‌سازی این باشد که فناوری‌های ADA گوناگون چگونه در بخش‌های مختلف مثل سرمایه‌گذاری، انرژی، بهداشت و سلامت و ... مورد استفاده قرار می‌گیرند[1]. پژوهش دیگر می‌تواند شناسایی این باشد که در عمل و در بخش‌های مختلف، تا چه حد تأثیرات مثبت یا منفی این فناوری‌ها مورد بحث قرار می‌گیرد. یک چالش بالقوه که باید مورد بررسی قرار بگیرد این است که نهادهای عمومی یا خصوصی تا چه اندازه مایل به افشای این اطلاعات هستند.

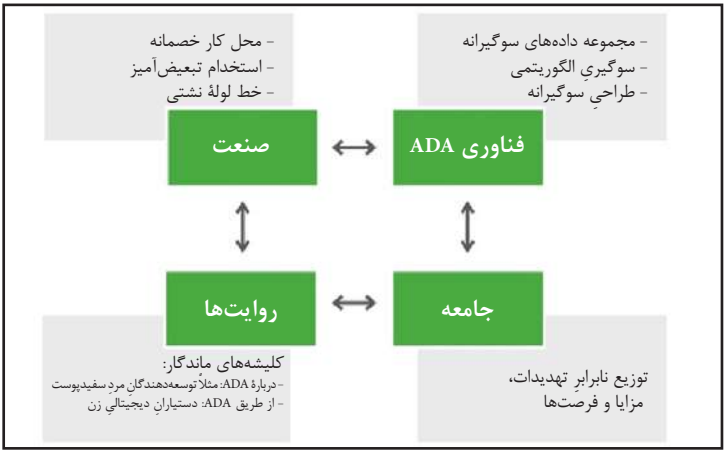
در مورد نکته دوم، فهم اینکه تأثیرات بالقوه چگونه به وجود می‌آیند، در تعیین نوع مداخلهٔ کاهش‌گرانه اهمیت حیاتی دارد، چنان‌که در موردکاوی زیر توضیح داده شده است.

موردکاوی

چرخه‌های بی‌عدالتی: نژاد و جنسیت

مواجهه با سوگیری و تبعیض الگوریتمیک مستلزم درک بهتر این است که چگونه می‌توان آن‌ها را در چرخه وسیع‌تر بی‌عدالتی جای داد، چرخه‌ای که در آن مسائل مختلف یکدیگر را تقویت می‌کنند (تصویر ۳).

۱. برای بررسی شیوه‌های دیگر تجزیهٔ فضای تأثیرات اخلاقی و اجتماعی ADA به ضمیمهٔ شمارهٔ سه نگاه کنید.



تصویر ۳: ترسیم اینکه چرخه‌ها و ناعدالتی چگونه به‌واسطهٔ نحوهٔ توسعه، به‌کارگیری و فهمِ اعضای جامعه می‌توانند تقویت شوند.

برای مثال خروجی‌های تبعیض‌آمیز یا سوگیرانه صنعت هوش مصنوعی، هم ناشی از فقدان تنوع بین پژوهشگران و توسعه‌دهنده‌ها می‌باشد و هم سوگیری‌های ازپیش‌موجود اجتماعی که در بسیاری از مجموعه داده‌ها منعکس شده‌اند (مثل همبستگی‌های مبتنی بر کلیشه‌های جنسیتی، بین واژه‌ها در منابع زبانی). به‌کارگیری این سامانه‌های دارای سوگیری منجر به پُررنگ‌تر شدنِ بی‌عدالتی‌های اجتماعی موجود می‌شود (برای مثال سامانه‌هایی که در مورد مرخصی زندانی‌ها مشاوره می‌دهند، از داده‌های دارای سوگیری استفاده می‌کنند که منجر به این می‌شود که افراد رنگین‌پوست مدت بیشتری را در زندان بمانند).

این بی‌عدالتی‌ها کسانی را که قادر به شکل‌دهی به روایت‌های فناوری هستند تحت تأثیر قرار می‌دهد، و این به‌نوبه‌خود کسانی را که قادر به ورود به صنعت هستند و نیز بی‌عدالتی‌های اجتماعی

اصلی را متأثر می‌کند. برای مثال خالقان هوش مصنوعی همواره به‌گونه‌ای مردانه تصویر شده‌اند و این باعث شده است که زن‌ها در این حوزه نادیده گرفته شوند. درک این همبستگی‌ها نقشی کلیدی در بررسی هرچه بهتر مسائل و مشکلات دارد.

برای درک بهتر انحاء مختلف به‌کارگیری فناوری‌ها، تأثیرات‌شان بر جامعه و سازوکارهای زیربنایی این تأثیرها، می‌توانیم پرسش‌های زیر را در مورد چهار تنش اصلی مطرح کنیم:

دقت در برابر رفتار منصفانه و برابر

- در کدام بخش‌ها از ADA برای تصمیم‌گیری‌هایی استفاده می‌شود که در زندگی انسان‌ها تأثیر دارد؟
- آیا می‌توان تعیین کرد که این فناوری‌ها تا چه اندازه منجر به رفتارهای مختلف با گروه‌های اجتماعی گوناگون می‌شوند؟
- تفسیر این الگوریتم‌ها چقدر راحت است و اشخاص، چه فرصت‌هایی برای به چالش کشیدن این تصمیم‌ها دارند؟

شخصی‌سازی در برابر همبستگی و شهروندی

- چه نوع پیام‌ها، مداخلات و خدمات، و در کدام بخش‌ها در حال حاضر با استفاده از یادگیری ماشین شخصی‌سازی شده‌اند؟
- این شخصی‌سازی تا چه اندازه دقیق بوده و مبتنی بر چه مقولاتی است؟
- چه شواهدی وجود دارد که این شخصی‌سازی‌ها می‌توانند رویکردها یا رفتارها را به‌نحو اساسی متأثر کنند؟

کیفیت و بازدهی خدمات در برابر حریم خصوصی و خودمختاری اطلاعاتی

- در چه بخش‌ها و کاربردهایی از ADA برای بهبود بازدهی خدمات عمومی استفاده می‌شود؟
- این کاربردهای خاص چه تأثیراتی بر خودمختاری و حریم خصوصی دارند؟

راحتی در برابر خودشکوفایی و کرامت

- کدام وظایف و شغل‌ها در سال‌های اخیر خودکار شده‌اند و در آینده نزدیک خودکار خواهند شد؟
- خودکارشدگی در حال حاضر چه تأثیراتی بر زندگی افراد گذاشته است؟

علاوه بر این‌ها پرسش‌های عمومی و کلی نیز برای تحقیق وجود دارد، پرسش‌هایی که به همه چهار تنش ربط پیدا می‌کنند:

- در حال حاضر کدام نوع از فناوری‌های مبتنی بر ADA، در بخش‌های مختلف (انرژی، سلامت، حقوق و ...) و تا چه اندازه‌ای مورد استفاده قرار می‌گیرند؟
- پیامدهای اجتماعی این کاربردهای خاص، به‌ویژه روی آن‌هایی که در بخش‌های مربوطه نادیده گرفته شده‌اند (مثل زن‌ها و افراد رنگین‌پوست) و قشر آسیب‌پذیر (مثل کودکان یا افراد سالمند) کدامند؟

۴-۲- فهم احتیاجات و ارزش‌های جوامع متأثر

به‌منظور پیشرفت در دلالت‌های اخلاقی و اجتماعی فناوری‌های ADA، دیدگاه‌های کسانی را که تحت تأثیر این فناوری‌ها قرار

گرفته‌اند یا قرار خواهند گرفت درک کنیم. به‌ویژه مذاکره در مورد بده-بستان بین ارزش‌ها فقط وقتی می‌تواند محقق شود که این ارزش‌ها و آرزوها و نگرانی‌های مربوطهٔ همه افرادی که تحت تأثیر این فناوری‌ها قرار گرفته‌اند شناسایی و ملاحظه شوند. شناسایی این دیدگاه‌ها مستلزم مشاوره با این کاربران نهایی یا دست‌کم گروه‌های نمایندهٔ آنان است[1].

باید توجه کنیم که صرفِ تقویت درک عمومی از فناوری به‌هیچ‌وجه کافی نمی‌باشد. در واقع برخی از متخصصان ارتباط‌های علمی استدلال کرده‌اند که اصلاً اهمیتی ندارد که افراد غیردانشمند، چیز کمی در مورد علم بدانند[2]. برای آنکه کاربران نهایی، تأثیرات فناوری بر زندگی‌شان را درک کنند نیازی به درک کامل چگونگی عملکرد فناوری ندارند. تعهد عموم مردم که شامل تأمل، رأی‌گیری و گفتگوی عمومی می‌باشد بسیار اهمیت بیشتری دارد: این یعنی تقویت درک متقابل بین محققان، توسعه‌دهندگان، سیاست‌گذاران و کاربران نهایی. این شامل تعامل متقابل بین این گروه‌ها به‌منظور درک علم و فناوری و همچنین تأثیرات اجتماعی، محدودیت‌ها، بده-بستان‌ها و ضعف‌های [علم و فناوری] می‌باشد.

با توجه به اهداف فعلی، تعهد عمومی برای حلِ بده-بستان‌ها و دوراهی‌ها، به‌طوری‌که از نظر همه اعضای جامعه قابل دفاع باشد اهمیت حیاتی دارد، زیرا بین علایق گروه‌های مختلف ناسازگاری وجود دارد. در مورد هر مسئله مشخصی، به‌ندرت پیش می‌آید که شهروندان دارای ارزش‌ها و دیدگاه‌های مشترک باشند. با این همه شواهدی وجود دارد که نشان می‌دهد وقتی گروه‌های مختلف

۱. برای اطلاعات ارزشمند در این قسمت، ممنون سارا کاستل Sarah Castell (Ipsos Mori) هستیم.

2. Hallman in Jamieson et al (2017)

قادر به تأمل و شرح مسائل مورد اهمیت‌شان هستند امکان کاهش ناسازگاری و دستیابی به توافق وجود دارد[1]. البته مهم است که به یاد داشته باشیم، اگرچه درک ارزش‌های عمومی مربوطه برای حل بده-بستان‌ها اهمیت دارد، اما این به‌خودی‌خود راه‌حل محسوب نمی‌شود بلکه فقط بخشی از یک فرآیند پیچیده‌تر سیاسی است[2].

طیف گسترده‌ای از روش‌ها برای افزایش مشارکت عموم مردم وجود دارد[3]. از این روش‌ها می‌توان برای استخراج طیف وسیعی از دیدگاه‌های آگاهانه و ناآگاهانه بهره جست. هدف از نظرسنجی ناآگاهانه جمع‌آوری نظراتِ کنونی گروه‌های پیمایش شده است، اما دیدگاه‌های آگاهانه را می‌توان از طریق راهبردهایی که هدف اولیه‌شان افزایش دانش گروه‌های پیمایش شده است استخراج کرد.

مشارکت عمومی که هدف آن حل بده-بستان‌ها است، می‌تواند به شکل‌های زیر باشد:

- پیمایش‌های کمی. از این پیمایش‌ها اغلب برای درک فهم عمومی استفاده می‌شود، یعنی برای درک این که گروه‌های پیمایش‌شده در حال حاضر چه اطلاعاتی در مورد موضوع دارند و این اطلاعات چه تأثیری بر عقاید و رویکردهایشان نسبت به یک فناوری می‌گذارد.
- مشاوره آنلاین مشارکتی. به‌عنوان یک مثال می‌توان به مشاوره اخیرِ مرکز انگلستان برای اخلاق داده‌ها و نوآوری اشاره کرد[4]. با استفاده از بده-بستان‌ها و تحلیل مشترک، احتمالاً به‌گونه‌ای بازی‌سازی‌شده، با این روش می‌توان دیدگاه‌های هزاران نفر از

1. Royal Society for the encouragement of Arts, Manufactures and Commerce (RSA, 2018).
۲. که می‌تواند شامل تحلیل هزینه-فایده، به‌کارگیری چشم‌اندازهای تخصصی و شواهدی دربارهٔ تأثیرات عینی تکنولوژی بر جامعه باشد.
3. International Association for Public Participation
www.dvrpc.org/GetInvolved/PublicParticipation/pdf/IAP2_public_participationToolbox.pdf
4. mwww.gov.uk/government/consultations/consultation-on-the-centre-for-data-ethics-and-innovation

شهروندان را به دست آورد و به این ترتیب یک منظر برتر در این مورد کسب کرد که هوش مصنوعی چه نقشی در جامعه بازی می‌کند و انسان‌ها چه واکنشی به تصمیمات اخلاقی مختلف نشان می‌دهند.

• پیمایش‌ها و مصاحبه‌های کیفی. این روش‌های کیفی به‌ویژه برای بررسی انگیزه‌های کنونی افراد و معنایی که آن‌ها برای تعامل‌شان با فناوری قائلند، به‌عنوان مکملی برای کارهای کمی بسیار سودمند هستند. همچنین به‌منظور دستیابی به دیدگاه‌های آگاهانه‌تر، از این روش‌ها می‌توان در ترکیب با مؤلفه‌های آموزشی نیز استفاده کرد.

• گفتگوی عمومی با طرح‌ریزی سناریو. این غالباً شامل ورودی‌هایی از یک گروه از متخصصان از جمله پیش‌بینی‌کنندگان و تحلیل‌گران فنی است که به‌نحو نظام‌مند عدم قطعیت‌های کلیدی را در یک چارچوب زمانی مشخص می‌نگارند. به این ترتیب وظیفه عموم مردم ساده‌تر می‌شود - به جای درگیری با تهدیدات و فایده‌های انتزاعیِ وجوه مختلف فناوری‌های پیچیده، خیلی ساده فقط باید به خروجی‌های برون‌یابی‌شده مختلف واکنش نشان داده و در مورد اینکه اشخاص و جامعه باید چه رفتاری در سناریوهای محتملِ مختلف نشان دهند صحبت کنند.

• مجامع شهروندی. RSA بر مجامع شهروندی به‌مثابه یکی از اَشکال مهم گفتگوی عمومی تأکید می‌کند. این‌ها فقط فرآیندهای یک‌طرفه کسب اطلاعات از عموم مردم نیستند، بلکه بر گفتگویی تمرکز دارند که در آن ذینفعانِ متخصص و شهروندان به‌منظور

تولید توصیه‌هایی برای سیاست‌گذاران با یکدیگر همکاری می‌کنند. از این روش اغلب جایی استفاده می‌شود که حل یک مسئله مستلزم پیمایش بده-بستان‌ها و ملاحظه راه‌حل‌های محتملِ مختلف باشد. این نوع از گفتگوی عمومی به‌ویژه بسیار مناسب برای بررسی و حل برخی از تنش‌هایی است که در بخش ۴ بحث کردیم.

در همه این مشارکت‌های عمومی، دیدگاه‌های حاصله صرفاً نمایانگر یک تصویر لحظه‌ای در یک تک لحظه هستند: دنبال‌کردنِ چگونگی تغییر ارزش‌ها و در طول زمان نیز اهمیت زیادی خواهد داشت. برای مثال ممکن است در چند سال آینده نگرانی در مورد حریم خصوصی داده افزایش یابد - و یا ممکن است به کلی ناپدید شود.

ما در پیوست ۱ کارهایی موجود در مشارکت عمومی را با جزئیات بیشتر به‌عنوان بخشی از مرور ادبیات بخش ۴ معرفی کرده‌ایم. بر اساس کارهایی که تاکنون انجام شده است می‌توانیم مصادیقی از پرسش‌های خاص برای تعهد عمومی حول محور چهار تنش اصلی شناسایی کنیم.

دقت در برابر رفتار منصفانه و برابر

- افراد موقعیت‌هایی را که در آن تصمیمات اصلی در مورد آن‌ها با کمک فناوری‌های ADA گرفته می‌شود، چگونه تجربه می‌کنند؟
- افراد تحت چه شرایطی مایل به پذیرش تأثیرات مختلف یک فناوری برای گروه‌های مختلف هستند؟
- انسان‌ها در بافتارهای مختلف چه چیزی را رفتار منصفانه و برابر محسوب می‌کنند[1]؟

۱. مبتنی بر آثاری مثل Grgić-Hlačа et al. (2018)، که ادراک‌های انسانی از انصاف در تصمیم‌سازی الگوریتمی در زمینهٔ پیشگیری از مخاطرات جنایی را مطالعه می‌کند و چارچوبی برای فهم این امر پیشنهاد می‌دهد که چرا افراد برخی ویژگی‌ها را به‌منزلهٔ منصفانه یا غیرمنصفانه در تصمیم‌های الگوریتمیک تلقی می‌کند.

شخصی‌سازی در برابر همبستگی و شهروندی

- انسان‌ها در چه بافتاری به دنبال اطلاعات شخصی‌شده یا گزینه‌هایی هستند که کاملاً متناسب با پروفایل آن‌هاست؟
- این، چه اندازه بستگی به نفع شخصی افراد دارد؟
- این، تا چه اندازه بستگی به حوزه مربوطه دارد (مثلاً سلامت، سرگرمی، تبلیغات سیاسی)؟
- انسان‌ها تغییراتی را که در اثر خودکارسازی در فضای عمومی پدید آمده است چگونه تجربه می‌کنند؟

کیفیت و بازدهی خدمات در برابر حریم خصوصی و خودمختاری اطلاعاتی

- افراد چه زمانی استفاده از داده‌های شخصی‌شان در جهت مؤثرتر کردنِ خدمات را تأیید می‌کنند؟
- این دیدگاه‌ها دقیقاً چه ربطی به نوع داده مورد استفاده و اینکه چه کسی از آن و به چه منظوری استفاده می‌کند، دارند؟
- این دیدگاه‌ها در بین گروه‌های مختلف چه تفاوتی می‌کنند؟

راحتی در برابر خودشکوفایی و کرامت

- افراد واگذاری شغل‌ها و وظایف مختلف به خودکارسازی را چگونه تجربه می‌کنند؟
- مؤلفه‌های جمعیت‌شناختی چه تأثیری بر پاسخِ پرسش بالا دارند؟
- الگوهای کاری ایده‌آل انسان‌ها، در پرتو خودکارسازی چه

می‌تواند باشد؟

• افراد دوست دارند چه نوع تعاملی با فناوری‌های ADA در محل کارشان داشته باشند؟ آن‌ها ترجیح می‌دهند این فناوری‌ها چه وظایفی را به عهده بگیرند؟

علاوه بر این پرسش‌ها، پرسش‌های کلی متعددی نیز برای بررسی وجود دارد:

• درک یک فناوری خاص (از جمله سازوکارها، اهداف، مالکان و خالقان آن) چرا و تا چه حدی برای عموم مردم اهمیت دارد؟

• اگر از الگوریتم‌ها به‌مثابه بخشی از فرآیند تصمیم‌گیری استفاده شود، تصمیمی که تأثیر چشمگیری بر زندگی انسان‌ها دارد، چه نوع تبیینی از تصمیمات آن الگوریتم ضروری و مناسب است؟ آیا این تبیین بسته به نوع تصمیم یا اینکه نهایتاً چه کسی مسئول آن است، تفاوت می‌کند؟

• از نظر عموم مردم بزرگ‌ترین فرصت‌ها و تهدیدهای فناوری‌های مختلف چیست و آن‌ها در مورد بده-بستان این دو چگونه می‌اندیشند؟ عوامل جمعیت‌شناختی چه تفاوتی ایجاد می‌کند؟ تجربه شخصی افراد با فناوری‌های مختلف، چه تفاوتی ایجاد می‌کند؟

۴-۳- به‌کارگیری شواهد برای حل تنش‌ها

پس از برجسته‌سازی برخی از پرسش‌ها در بخش‌های قبلی، حالا می‌خواهیم آن‌ها را در کنار یکدیگر بگذاریم تا نشان دهیم که چگونه

یـک پایـگاه شـواهد قوی‌تـر می‌توانـد بـه آشکارسـازی و حـل چهـار تنـش اصلـی مـا کمـک کنـد.

دقت در برابر رفتار منصفانه و برابر

ایـن تنـش وقتـی بـه وجـود می‌آیـد کـه کاربـران از یـک الگوریتـم به‌مثابـه بخشـی از یـک فراینـد تصمیم‌گیـری اسـتفاده می‌کننـد، امـا یـک بده-بسـتان بیـن منافـع الگوریتـم (بـرای مثـال افزایـش دقـت) و هزینه‌هـای بالقـوه آن (سـوگیری‌های بالقـوه‌ای کـه منجـر بـه نتایـج نامنصفانـه می‌شـود) وجـود دارد. در اینجـا نـه تنهـا لازم اسـت کـه نقـاط قـوت و محدودیت‌هـای آن الگوریتـم را در یـک بافتـار خـاص بدانیـم، بلکـه همچنیـن بایـد آن‌هـا را بـا نقـاط قـوت و محدودیت‌هـای نسـبی تصمیم‌گیـرانِ انسـانی نیـز مقایسـه کنیـم. در مـواردی کـه رفتـار منصفانـه و دقـت واقعـاً بـا یکدیگـر ناسـازگار هسـتند، پژوهش‌هـای بیشـتر بـرای مقایسـه دقـت پیش‌بینـی و سـوگیری‌های یـک الگوریتـم در مقایسـه بـا انسـان‌ها می‌توانـد مفیـد باشـد و بـه مـا کمـک کنـد کـه چـه زمانـی اسـتفاده از یـک الگوریتـم مناسـب اسـت.

درک نظرگاه‌هـای اجتماعـیِ مختلـف نیـز یکـی از بخش‌هـای حیاتـی پیمایـش بده-بسـتان‌های بـه وجـود آمـده در اسـتفاده از الگوریتم‌هـا در فراینـدهـای تصمیم‌گیـری اسـت. آیـا خودکارسـازیِ تصمیمـات حیاتـی، منجـر بـه افزایـش / کاهـش اعتمـاد بـه نهادهـای عمومـی می‌شـود؟ گروه‌هـای مختلـف بـرای اینکـه بـه الگوریتم‌هـا اعتمـاد کننـد، چـه سـطح و چـه نوعـی از تبیین‌پذیـری را طلـب می‌کننـد؟ اسـتفاده از چـه نـوع اطلاعـات بـا چـه ویژگی‌هایـی توسـط یـک الگوریتـم در تصمیم‌گیری‌هـای

مختلف قابل قبول است و کدام‌شان ممکن است نامنصفانه محسوب شود؟

شخصی‌سازی در برابر شهروندی و همبستگی

در اینجا به این دلیل تنش به وجود می‌آید که از داده‌ها و یادگیری ماشین می‌توان به‌منظور شخصی‌کردنِ خدمات و اطلاعات استفاده کرد، درحالی‌که دارای دلالت‌های مثبت و منفی برای خیر عمومی دموکراسی‌ها می‌باشد. به‌منظور فهم بده-بستان‌ها در اینجا نیازمند شواهد بهتری از شواهد رایج و غالباً هیجانی در مورد امرِ در حال حاضر امکان‌پذیرِ تکنیکی هستیم: از داده‌های عمومی و خصوصی در دسترس، چه نوع استنتاجاتی در مورد گروه‌ها و افراد می‌توان استخراج کرد؟ با استفاده از آن‌ها چه تأثیراتی بر رویکردها می‌توان گذاشت؟

ما همچنین باید شواهد بهتری را در مورد رویکردهای مختلف به شخصی‌سازی روزافزون جمع‌آوری کنیم: افراد چه زمانی آن را برای زندگی‌های‌شان مفید می‌دانند، آن را مضر می‌دانند، آیا می‌توان تمایز روشنی بین این‌دو قائل شد؟ شخصی‌سازی گاهی اوقات پذیرفته شده و گاهی اوقات غیرقابل‌پذیرش است، و ما باید بدانیم چرا و چه موقع. نگرانی‌های بنیادینی که منجر به عدم پذیرش شخصی‌سازی از سوی افراد در یک حوزه خاص می‌شوند کدامند - برای مثال آیا آن‌ها نگران این هستند که دیگر در یک فضای اطلاعاتی مشترک با سایر اعضای جامعه نباشند، یا آن‌ها بیشتر نگران این هستند که اطلاعات دقیق به سازمان‌ها قدرت بیش از حد برای دستکاری افراد می‌دهد؟ درست مثل حریم خصوصی ممکن است

انتظار داشته باشیم که دیدگاه‌ها در مورد شخصی‌سازی و همبستگی در طی زمان تغییر کند: باید بدانیم که این تغییرها چه هستند و چگونه می‌توانند تنش‌های موجود را تغییر بدهند آنچنان‌که نیازمند پژوهش‌های بیشتر در مورد دلالت‌های اخلاقی و سیاسی شخصی‌سازی برای دموکراسی، رفاه و مشارکت سیاسی باشیم.

کیفیت و بازدهی خدمات در برابر حریم خصوصی و خودمختاری اطلاعاتی

چنان‌که اشاره شد، این تنش به وجود می‌آید زیرا می‌توان از داده‌های شخصی برای بهبود خدمات استفاده کرد، اما چنین کاری چالش‌هایی را برای حریم خصوصی و خودمختاری اطلاعاتی افراد به وجود می‌آورد. با این همه روش‌های فنی‌ای مثل حریم خصوصی اختلافی[1] برای استنتاج از داده‌های انباشتی، در عین حفظ حریم خصوصی افراد وجود دارد[2]. هرچقدر که این روش‌ها موفقیت‌آمیز باشند و هرچقدر که ما بتوانیم مدل‌های جدیدِ رضایت را به‌کار بگیریم، تنش کمتری بین کاربردهای نوآورانه داده و حریم خصوصی وجود خواهد داشت. بنابراین درک وضعیت کنونی پژوهش‌های تکنیکی در این حوزه و اینکه چه کاری می‌توان و چه کاری نمی‌توان انجام داد برای درک این تنش حائز اهمیت خواهد بود.

اگر بده-بستان بین کیفیت خدمات و حریم خصوصی [حل نشده] باقی بماند، آن‌گاه درکِ افکارِ آگاهانه و ناآگاهانهٔ عمومی برای حل آن ضروری خواهد بود. احتمال دارد که عموم مردم استفاده از داده‌های

1. differential privacy (مترجم)

۲. هدف از روش‌های حریم‌خصوصی اختلافی به حداکثررساندن دقت استنباط‌های برگرفته از پایگاه داده‌ها و به حداقل‌رساندن شانس شناسایی سوابق افراد، از طریق تضمین این امر است که افزودن یا حذف یک نقطه دادهٔ منفرد تغییر بنیادینی در خروجی ندارد. گرچه حریم‌خصوصی اختلافی به‌معنای تضمین قطعی حریم‌خصوصی نیست، اما اطمینان می‌بخشد که خطر ماندن بخشی از اطلاعات در پایگاه داده‌ها محدود می‌شود. برای مثال نگاه کنید به:

Hilton and Dwork (2008)

شخصی‌شان را در برخی موارد تأیید کنند- برای مثال کاربردهای نجات‌بخش پزشکی- و در سایر موارد تأیید نکنند. مفهوم حریم خصوصی و اهمیت آن همچنین می‌تواند در طول زمان تکامل پیدا کند و وجوه مهم و غیرمهم آن‌را تغییر دهد. قضاوت متخصصان در مورد دلالت‌های وسیع طرح اجتماعی و اخلاقی نقض و گسترش حریم خصوصی، باید به‌عنوان مکمل این مطالعات مورد استفاده قرار بگیرد.

راحتی در برابر خودشکوفایی و کرامت

قلبِ این تنش این واقعیت است که خودکارسازی دارای منافع واضحی است: صرفه‌جویی در زمان و تلاش مردم در وظایف کم اهمیت، افزایش راحتی و دسترسی، اما درعین‌حال خودکارسازی بیش از حد می‌تواند حس موفقیت، خودشکوفایی و کرامت ما را به‌مثابه انسان تهدید کند. به‌منظور کاوش در این تنش باید از تأمل روشن در این مورد شروع کنیم که خودکارسازی عمدتاً در کجا سودمند محسوب می‌شود (شاید به این دلیل که وظایف مورد نظر بی‌روح و از خود بیگانه‌ساز هستند)، و در کجا از نظر اخلاقی آسیب‌زا و نامناسب است (مثلاً خودکارسازی وظایف پیچیده در آموزش و پرورش، جنگ، مهاجرت، عدالت و روابط می‌تواند آسیب‌زا باشد). درک نظرگاه‌های طیف وسیعی از گروه‌های مختلف در مورد این مسئله می‌تواند اهمیت ویژه داشته باشد، زیرا فعالیت‌هایی که در یک گروه یا فرهنگ بسیار با ارزش محسوب می‌شوند، در گروه یا فرهنگ دیگر می‌تواند تفاوت کنند.

اگر بتوانیم در مورد وظایفی که خودکارکردنشان می‌تواند سودمند باشد توافق کنیم، آن‌گاه می‌توانیم شروع به جمع‌آوری شواهد در مورد قابلیت‌های فناوری‌های کنونی در این حوزه‌ها و ارزیابی اینکه برای پیشرفت نیازمند چه چیزی هستیم، کنیم. و برعکس اگر بتوانیم قابلیت‌هایی را که در شکوفایی انسان اهمیت اساسی دارند و ما نمی‌خواهیم خودکارشان نماییم، به‌روشنی شناسایی کنیم، قابلیت‌های کنونی در این حوزه‌ها می‌توانند به ارزیابی بهتر این تهدیدها و تفکر در مورد واکنش‌های بالقوه کمک کنند.

علاوه بر تحقیق و پژوهش در مورد تنش‌هایی که می‌توانیم در اینجا شناسایی کنیم، ما همچنین از مطالعات بین‌رشته‌ای این تنش‌ها که تحولات ریشه‌ای‌تر سیاسی و فناورانه را کاوش می‌کنند استقبال می‌کنیم: برای مثال فناوری‌های مبتنی بر ADA چگونه خواهند بود اگر می‌شد که جذابیت اصلی‌شان سود یا برتری ژئوپلیتیک نباشد، و اینکه این فناوری‌ها چه آرایش اقتصادی-اجتماعیِ بدیلی برای سرمایه‌داری می‌توانند ایجاد کنند.

۴-۴- خلاصه و پیشنهادات

ما توصیه می‌کنیم که کارهای پژوهشی و سیاستی در مورد اخلاق فناوری‌های مبتنی بر ADA باید روی توسعه پایگاه شواهد قوی‌تر در مورد (الف) قابلیت‌های کنونی و بالقوه فناورانه؛ و (ب) رویکردها و نیازهای اجتماعی سرمایه‌گذاری کند و بسیاری از مفروضات را به چالش بکشد. به‌ویژه:

- **تعمیق درک قابلیت‌ها و محدودیت‌های فناورانه** در حوزه‌هایی

که دارای مسائل کلیدی اخلاقی و اجتماعی هستند.

غالباً بحث در مورد مسائل اخلاقی و اجتماعی بر اساس مفروضات تأمل‌نشده در مورد امر در حال حاضر امکان‌پذیرِ فناورانه بنا شده‌اند. برای ارزیابی قابل اعتماد تهدیدها و فرصت‌های ADA برای جامعه، و نیز برای تأملِ روشن‌تر در مورد بده-بستان‌های بین ارزش‌ها، باید این مفروضات را مورد بازرسی انتقادی‌تر قرار بدهیم.

• **ساخت یک پایگاه شواهد قوی‌تر در مورد کاربردها و تأثیرات کنونی فناوری‌های مبتنی بر ADA**، به‌ویژه حول محور تنش‌های کلیدی که گروه‌های به حاشیه رانده شده و نادیده گرفته شده را متأثر می‌کنند.

درک کاربردهای خاص فناوری‌های مبتنی بر ADA به ما در تفکر انضمامی‌تر در مورد اینکه احتمال به وجود آمدنِ تنش‌ها در بین ارزش‌ها در کجا و چگونه بیشتر است، و چگونگی حل آن‌ها کمک خواهد کرد. شواهدی که در مورد تأثیرات اجتماعیِ کنونیِ فناوری‌ها وجود دارد، یک مبنای نیرومندتر برای ارزیابی تهدیدها و پیش‌بینی تأثیرات محتمل آینده فراهم می‌کند.

• استفاده از تعهد عمومی موجود برای **درک بهتر نظرگاه‌های مختلف اعضای جامعه در مورد مسائل و بده-بستان‌های مهم.**

چنان‌که پیش‌تر نیز تأکید کردیم، پیمایش دلالت‌های اخلاقی و اجتماعی ADA ما را ملزم به این می‌کند که تنش بین ارزش‌ها را که استفاده از این فناوری‌ها تقویت‌شان می‌کنند به رسمیت

بشناسیم. از آنجا که گروه‌های عمومی مختلف تأثیرات مختلفی را از این فناوری‌ها خواهند گرفت، و به ارزش‌های مختلفی وفادار خواهند بود، این تنش‌ها ما را ملزم به درک عقاید متغیر عمومی در مورد مسائل مربوط به تنش‌ها و بده-بستان‌ها خواهد کرد.

بخش پنجم

نتیجه‌گیری

بخش پنجم

نتیجه‌گیری

ما در این گزارش، وضعیت کنونی پژوهش و بحث در مورد تأثیرات اخلاقی و اجتماعی الگوریتم‌ها، داده و هوش مصنوعی را به‌منظورشناسایی یافته‌ها و گام‌های بعدی مورد بررسی قرار دادیم.

در بخش ۲، چند مفهوم کلیدی را برای دسته‌بندی مسائل برخاسته از فناوری‌های مبتنی بر ADA و نیز چند اصل و ارزش اخلاقی را که بیشتر کنشگران بر اهمیت‌شان تأکید دارند، شناسایی کردیم. ما همچنین ۳ وظیفه کلیدی زیر را شناسایی کردیم که باور داریم به‌منظور پیشبردِ این مباحث باید در اولویت قرار بگیرند:

- وظیفه ۱: مفهوم‌سازی: بررسی ابهام‌های موجود در مفاهیم محوریِ مورد استفاده در بحث‌های ADA، شناسایی تفاوت‌های مهم در چگونگی استفاده و درک این واژه‌ها در رشته‌ها، بخش‌ها، گروه‌های عمومی و فرهنگی مختلف، و تلاش برای پُل‌زدن و ایجاد اجماع تا جای ممکن.
- وظیفه ۲: حل تنش‌ها و بده-بستان‌ها: شناسایی و شرح تنش‌های بین اصول و ارزش‌های مختلف در بحث‌های مربوط به ADA،

تعیین اینکه کدام یک از این تنش‌ها را براساس فناوری‌های بهتر یا دیگر راه‌حل‌های عملی می‌توان حل کرد، و توسعه روش‌های مشروع برای حل هر بده-بستانی که باید انجام شود.

• وظیفه ۳: توسعه یک پایگاه شواهد: ایجاد یک پایگاه شواهد قوی‌تر بر اساس قابلیت‌ها، کاربردها و الزامات اجتماعی مربوط به ADA و استفاده از آن برای حل تنش‌ها و بده-بستان‌ها.

در خلال این گزارش، ما توصیه‌ها و پرسش‌هایی پیشنهادی برای پژوهش در مورد این وظایف را طرح کردیم. این پرسش‌ها را در زیر خلاصه کرده‌ایم. این پرسش‌ها به‌هیچ‌وجه جامع و کامل نیستند. با این همه، حوزه‌هایی را برجسته می‌کنند که به باور ما در آن‌ها ظرفیت زیادی برای پژوهش‌های آتی وجود دارد.

ما، تحقیق در مورد تأثیرات اخلاقی و اجتماعی ADA را یک تلاش متکثرِ بین‌رشته‌ای و بین‌بخشی در نظر می‌گیریم، تلاش بر اساس بهترین روش‌های در دسترس علوم انسانی، علوم اجتماعی و رشته‌های فنی و نیز تخصص کارورزان. این توصیه‌ها، روی‌هم‌رفته یک نقشه راه برای تحقیق شکل می‌دهند، نقشه‌ای که مستلزم توازن بین احترام و یادگیری تفاوت‌های بین ذینفعان و رشته‌های مختلف است، و تشویق می‌کند به نقد مداوم و سازنده‌ای که می‌تواند دانشی مهم و عملی فراهم آورد. هدف این پایگاهِ دانشی بهبود استانداردها، مقررات و نظام‌های حکمرانیِ فناوری‌های ADA است که در حال حاضر نامطمئن و آشفته هستند. ما تأکید می‌کنیم که رویکردهای جدید به حکمرانی و مقررات، باید به‌درستی به تنش‌هایی که در بالا

توصیف شدند حساس باشند و نهادهایی مشروع و اختصاصی ایجاد نمایند که می‌توانند به جوامع در شناسایی، شرح و پیمایش این تنش‌ها، و نیز تنش‌هایی که در بافتار وسیع‌ترِ حیات این جوامع پیش می‌آیند کمک کنند.

۵-۱- پرسش‌هایی برای تحقیق

وظیفه ۱: مفهوم‌سازی

به‌منظور روشن‌سازی و حل ابهامات و اختلافات در استفاده از واژه‌های کلیدی:

- معانی مختلف واژه‌های کلیدی در بحث‌های مربوط به ADA کدامند؟ این واژه‌ها عبارت‌اند از (و نه محدود به همین‌ها): انصاف، سوگیری، تبعیض، شفافیت، توضیح‌پذیری، تفسیرپذیری، حریم خصوصی، مسئولیت‌پذیری، کرامت، همبستگی، آسودگی، توانمندسازی و خودشکوفایی.
- چگونه این واژه‌ها به جای یکدیگر استفاده می‌شوند یا باهم هم‌پوشانی معنایی دارند؟
- در کجا انواع و اقسام موضوعات، ذیل یک واژه‌شناسی مشابه در یکدیگر ادغام می‌شوند؟
- چگونه از واژه‌های کلیدی در رشته‌ها، بخش‌ها، گروه‌های عمومی و فرهنگ‌های مختلف، به‌گونه‌ای واگرا استفاده می‌شود؟

ایجاد پُل‌های مفهومی بین رشته‌ها و فرهنگ‌ها:

- چه دیدگاه‌های فرهنگی، به‌ویژه دیدگاه‌های برخاسته از کشورهای

درحال‌توسعه و گروه‌های حاشیه‌ای، هنوز در کارهای تحقیقاتی و سیاست‌گذاری در مورد اخلاق ADA، به‌خوبی منعکس نشده است؟ چگونه می‌توان این دیدگاه‌ها را هم در نظر گرفت، برای مثال از طریق ترجمه ادبیات سیاست‌گذاری و پژوهشی مربوطه یا ایجاد همکاری پیرامون موضوعات خاص؟

• در حال حاضر در پژوهش‌های اخلاق ADA، به کدام رشته‌های دانشگاهی توجه نشده است و برای رفع این مشکل، از چه نوع همکاری‌های پژوهشی بین‌رشته‌ای می‌توان استفاده کرد؟

دستیابی به اجماع و مدیریت اختلافات:

• چگونه می‌توان در جایی که در کاربرد کلمات کلیدی، ابهام و تفاوت وجود دارد، به اجماع و درک مشترک دست یافت؟

• اگر به‌راحتی نتوان به اجماع دست یافت، چگونه می‌توان آن را پذیرفت و در عین اختلافات مهم، به‌گونه‌ای مولد به کار ادامه داد؟

وظیفه ۲: تنش‌ها و بده-بستان‌ها

برای درک بهتر چهار تنش اصلی:

• تا چه اندازه با دوراهی‌های واقعی، دوراهی‌های عملی یا دوراهی‌های کاذب مواجه هستیم؟

• چگونه می‌توان طوری از دقیق‌ترین الگوریتم‌های پیش‌بینی استفاده کرد که به انصاف و مساوات لطمه‌ای وارد نشود؟

• چگونه می‌توانیم در عین برخورداری از منافع شخصی‌سازی به ایده‌آل‌های همبستگی و شهروندی نیز احترام بگذاریم؟

• چگونه می‌توانیم از داده‌های شخصی برای بهبود خدمات عمومی و حفظ یا توسعه حریم خصوصی و خودمختاری اطلاعاتی استفاده کنیم؟
• چگونه می‌توانیم از خودکارسازی در جهت راحت‌ترکردنِ زندگی‌هایمان استفاده کنیم و هم‌زمان خودشکوفایی و کرامت را نیز ارتقا دهیم؟

برای مشروعیت‌بخشی به بده-بستان‌ها:

• چگونه به بهترین نحو می‌توانیم صدای همه ذینفعان متأثر از ADA را شنیده و منافع آن‌ها را با قوت و احترام تشریح کنیم؟
• بده-بستان‌های قابل قبول و مشروعِ سازگار با حقوق و تکالیف افراد متأثر از این فناوری‌ها کدامند؟
• کدام سازوکارهای حل مسئله بیشترین احتمال پذیرفته‌شدن را دارند؟
• درمورد چهار تنش اصلی، باید بپرسیم که:
- استفاده از الگوریتم‌هایی که به ضرر یک زیرگروه خاص هستند اما به‌طورکلی دقیق‌اند، در چه مواردی قابل قبول است، اگر اصلاً قابل قبول باشد؟
- شخصی‌سازی تبلیغات و خدمات عمومی را تا چه حد باید محدود کنیم، به نفع حفظ ایده‌آل‌های شهروندی و همبستگی؟
- چه حدی از تهدید خودمختاری اطلاعاتی و حریم خصوصی به‌منظور پایش بهتر بیماری‌ها یا سلامت عمومی بیشتر قابل قبول است؟
- چه نوع مهارت‌هایی تا ابد باید در اختیار انسان‌ها باقی بماند و متعاقباً در چه مواردی باید نوآوری‌های مربوط به فناوری‌های خودکارسازی را رد کنیم؟

برای شناسایی تنش‌های جدید فراتر آن‌هایی که در این گزارش برجسته شده‌اند:

- در چه مواردی احتمال دارد که منافع و آسیب‌های فناوری‌های مبتنی بر ADA به تساوی در بین گروه‌های مختلف توزیع نشده باشد؟
- در چه مواردی ممکن است که کاربردهای فناوری‌های مبتنی بر ADA منجر به فرصت‌هایی کوتاه مدت و تهدیدهای بلندمدت شوند؟
- در چه مواردی ممکن است که درمورد اثرات فناوری بسیار محدود بیاندیشیم؟
- در چه مواردی ممکن است کاربردهایی که از منظر محدود فردی سودمند هستند، پیامدهای بیرونی نامطلوب ایجاد نمایند؟

وظیفه ۳: توسعه یک پایگاه شواهد

برای تعمیق درک‌مان از قابلیت‌ها و محدودیت‌های فناورانه:

پرسش‌های کلی:

- به‌منظور ارزیابی معنادار تهدیدها و فرصت‌هایی که فناوری‌ها در بافتارهای اخلاقی و اجتماعی گوناگون به وجود می‌آورند نیاز به درک چه چیزهایی درمورد قابلیت‌ها و محدودیت‌های فناورانه داریم؟
- چگونه ممکن است پیشرفت‌های فناورانه به حل تنش بین ارزش‌ها در کاربردهای ADA کمک کنند، و محدودیت‌های فناوری برای این منظور کدامند؟

اِعمال این پرسش‌های کلی بر ۴ تنش اصلی:

• دقت در برابر رفتار منصفانه و برابر

- دقت تا چه حد با تعاریف مختلف انصاف در بده-بستان قرار می‌گیرد؟
- چه نوع تفسیرپذیری از منظر ذینفعان مختلف، مطلوب است؟
- آخرین مدل‌ها چه نوع تفسیرپذیری را تضمین می‌کنند؟
- حفظ سطح مناسبی از تفسیرپذیری، بدون فدا کردنِ دقت، تا چه اندازه امکان‌پذیر است؟

• شخصی‌سازی در برابر همبستگی و شهروندی

- آیا جزئی‌بودنِ شخصی‌سازی، محدودیت بنیادین یا عملی دارد؟
- شخصی‌سازی، تا چه اندازه، تأثیری معنادار بر سایر خروجی‌ها (مثل رضایت کاربر، رفتار مصرف‌کننده، الگوهای رأی‌دادن) دارد؟

• کیفیت و بازدهی خدمات در برابر حریم خصوصی و خودمختاری اطلاعاتی

- یادگیری ماشین و کلان‌داده تا چه اندازه می‌توانند خدمات عمومی گوناگون را بهبود بخشند؟ آیا می‌توان منافع بالقوه را کمّی‌سازی کرد؟
- روش‌های کنونی تا چه اندازه امکان استفاده از داده‌های انباشتی را با حفظ حریم خصوصی داده‌های فردی فراهم می‌آورند؟
- بهترین روش برای اطمینان از رضایت معنادار چیست؟

• راحتی دربرابر خودشکوفایی و کرامت
- چه نوع وظایفی را می‌توان با استفاده از فناوری‌های کنونی و قابل پیش‌بینی، خودکار کرد؟
- هزینه‌های خودکارسازی گسترده (برای مثال انرژی یا زیرساخت) چه خواهد بود؟

به‌منظور ساخت یک پایگاه شواهد قوی‌تر درمورد استفاده‌ها و اثرات کنونی فناوری:

پرسش‌های کلی

• چه نوع فناوری‌های مبتنی بر ADA در حال حاضر در بخش‌های مختلف (انرژی، سلامت، حقوق و غیره) مورد استفاده است، و تا چه حد؟
• تأثیرات اجتماعی این کاربرهای خاص، به‌ویژه بر روی گروه‌های محروم (مثل رنگین‌پوستان) و نادیده گرفته شده (مثل زنان) یا آسیب‌پذیر (مثل کودکان یا سالمندان) چیست؟

اِعمال این پرسش‌های کلی بر چهار تنش خاص:

• دقت در برابر رفتار منصفانه و برابر
- در چه بخش‌ها یا کاربردهایی، از ADA برای تصمیم‌گیری/پیش‌بینی‌هایی استفاده می‌شود که پیامدهایی برای زندگی افراد دارند؟
- آیا تعیین اینکه این کاربردها در چه مواقعی منجر به رفتارهای متفاوت با گروه‌های اجتماعی مختلف می‌شود

امکان‌پذیر است؟

- تفسیر الگوریتم‌های مورد استفاده در تصمیم‌گیری‌هایی که بر زندگی افراد تأثیر دارند، چقدر ساده است؟ و افراد چه فرصت‌هایی برای به چالش کشیدن این تصمیم‌ها دارند؟

• شخصی‌سازی در برابر همبستگی و شهروندی

- چه نوع پیام‌ها، مداخلات و خدماتی و در کدام بخش‌ها در حال حاضر با استفاده از یادگیری ماشین شخصی‌سازی شده است؟

- شخصی‌سازی چقدر جزئی است و مبتنی بر چه نوع مقولاتی می‌باشد؟

- چه شواهدی وجود دارد دال بر اینکه شخصی‌سازی می‌تواند تأثیر مهمی بر رویکردها یا رفتارها داشته باشد؟

• کیفیت و بازدهی خدمات در برابر حریم خصوصی و خودمختاری اطلاعاتی

- از ADA در چه بخش‌ها و کاربردهای خاصی برای بهبود بازدهی خدمات عمومی استفاده می‌شود؟

- این کاربردهای مشخص چه تأثیراتی بر خودمختاری و حریم خصوصی دارند؟

• راحتی در برابر خودشکوفایی و کرامت

- خودکارسازی در حال حاضر چه تأثیراتی بر زندگی روزانه گروه‌های عمومی مختلف داشته است؟

به‌منظور درک بهتر دیدگاه‌های گروه‌های مختلف:

پرسش‌های کلی

- ترجیحات عموم مردم درباره درک یک فناوری مشخص (از جمله سازوکارها، اهداف، مالکان، خالقان و غیره) چیست؟
- از نظر گروه‌های عمومی مختلف، بزرگ‌ترین تهدیدها و فرصت‌های فناوری‌های مختلف چیست و آن‌ها درمورد بده-بستان این دو چگونه فکر می‌کنند؟ این دیدگاه‌ها چه تفاوتی بر اساس ویژگی‌های جمعیت‌شناختی مختلف می‌کنند؟ براساس تجربه مختلف افراد از فناورهای مختلف چه‌طور؟

اِعمال این پرسش‌های کلی به چهار تنش اصلی:

- دقت در برابر رفتار منصفانه و برابر
  - گروه‌های عمومی مختلف، تأثیرگذاری‌های مختلف یک فناوری را چگونه تجربه می‌کنند؟
  - افراد در سیاق‌های مختلف چه چیزی را به‌عنوان رفتار منصفانه و برابر در نظر می‌گیرند؟

- شخصی‌سازی در برابر همبستگی و شهروندی
  - افراد در کدام بافتارها استفاده از داده‌های شخصی‌شان را در جهت پُربازده‌کردنِ خدمات عمومی تأیید می‌کنند؟
  - این رویکردها تا چه اندازه مبتنی بر این هستند که دقیقاً از چه داده‌هایی استفاده می‌شود، چه کسی و به چه منظوری از آن‌ها استفاده می‌کند؟

- این رویکردها در میان گروه‌های مختلف چه تفاوتی می‌کند؟

• کیفیت و بازدهی خدمات در برابر حریم خصوصی و خودمختاری اطلاعاتی

- افراد در چه مواقعی استفاده از داده‌های شخصیشان به‌منظور بهبود خدمات عمومی را تأیید می‌کنند؟
- این رویکردها چه ارتباطی به این دارند که چه داده‌هایی، توسط چه کسی و به چه منظوری استفاده شده است؟
- این رویکردها، گروه به گروه چه تفاوتی می‌کنند؟

• راحتی در برابر خودشکوفایی و کرامت

- افراد بیشتر نگران کدام وظایف و شغل‌ها در برابر خودکارسازی هستند؟ پاسخ به این پرسش چه ارتباطی با عوامل جمعیت‌شناختی دارد؟
- الگوی ایده‌آل کار کردن، در پرتو خودکارسازی چه خواهد بود؟
- افراد تمایل به ایجاد چه نوع تعاملی با ADA در محل کار خود دارند؟
- کدام وظایف را به‌گونه‌ای اخلاقی می‌توان به عهده فناوری‌ها گذاشت؟

بخش ششم

پیوست ۱: خلاصه مرور ادبیات

بخش ششم

پیوست ۱: خلاصه مرور ادبیات

این گزارش بر اساس یک سلسله مرور ادبیات در مورد چگونگی بحث‌های انجام‌شده در باب دلالت‌های اخلاقی و اجتماعی ADA در مجموعه‌ای از رشته‌ها، در گزارش‌های سیاستی و جامعه مدنی و در علم و رسانه عامه-پسند تهیه شده است. مرورهای انجام‌شده بیش از ۱۰۰ مقاله دانشگاهی در رشته‌های مختلف از جمله علوم رایانه، اخلاق، تعامل انسان - رایانه، حقوق و فلسفه، ۲۰ سند سیاستی، بیش از ۲۵ کتاب پُرطرفدار و پُرارجاع، مقاله‌های رسانه‌ای و خبری و اسناد متعدد مربوط به مشارکت عمومی را پوشش می‌دهند[۱].

بخش ۳-۱ در این پیوست، اصلی‌ترین مشاهدات را از هر یک از این مرورهای ادبیات خلاصه کرده و بخش ۴ تِم‌های مشترک آن‌ها را ارائه می‌دهد.

۶-۱- ادبیات دانشگاهی

۱آ. علوم رایانه و یادگیری ماشین

ما مقالات منتشرشده در سال ۲۰۱۷ در مهم‌ترین کنفرانس‌ها و ژورنال‌های حوزه هوش مصنوعی، یادگیری ماشین، علوم داده و

۱. همهٔ منابع در قسمت مأخذشناسی آمده است و برای هر یک از حوزه‌های ادبیات موضوعی، ارجاعات مهمِ استفاده‌شده را برجسته کرده‌ایم.

داده‌کاوی را پوشش دادیم[1]. در بسیاری از موارد، کمتر از ۱ ٪ مقالات مستقیماً به تأثیرات اخلاقی یا اجتماعی فناوری ربط داشتند.

به‌طورکلی به نظر می‌رسد که یک فرهنگ اختلاف‌نظر در بین محققان فنی و مهندسانی وجود دارد که مسئولیت مسائل اخلاقی و اجتماعی برخاسته از فناوری را متوجه خودشان نمی‌دانند[2]. با این همه در دوسه سال اخیر علاقه روزافزونی به این مسائل وجود داشته است، چنان‌که کارگاه‌ها و گردهمایی‌های مربوط در مهم‌ترین کنفرانس‌ها مانند FAT/ML و ابتکار جهانیِ IEEE در سامانه‌های خودمختار و هوشمند نشان می‌دهد.

تعجبی ندارد که آن بخش از تحقیقات فنی که مستقیماً مسائل اخلاقی و اجتماعی را مورد بررسی قرار داده‌اند، بر آن مسائلی تمرکز داشته‌اند که می‌توان آن‌ها را بر اساس واژگان فنی صورت‌بندی یا ساده‌سازی کرد: چگونه می‌توانیم تصمیم‌های یک سامانه یادگیری ماشین جعبه سیاه را تفسیر یا تبیین کنیم، چگونه می‌توانیم اعتمادپذیری سامانه‌های گوناگون را در مسائل حریم خصوصی و حفاظت از داده‌ها ارزیابی کنیم و چگونه می‌توانیم ارزش‌های مهمی همچون انصاف را وارد الگوریتم‌ها و سامانه‌های هوش مصنوعی کنیم. ما همچنین مقاله‌های پیمایشی متعددی را در مورد نحوه آموزش اخلاقی متخصصان هوش مصنوعی و علوم داده پوشش دادیم. اگرچه کدهای اخلاقی برای دانشمندان علوم داده توسعه داده شده است، اما اغلب متخصصان هوش مصنوعی هیچ آموزشی در مورد مسائل اخلاقی دریافت نمی‌کنند. از آنجاکه این فضا، به‌سرعت در حال تغییر است و هیچ مسیر استانداردی برای تبدیل‌شدن به یک متخصص

۱. شامل مقالاتی از کنفرانس‌های زیر:
IJCAI, AAAI, NIPS, ICML, KDD, ICDM
و نیز ژورنال‌های زیر:
AIJ, JAIR, AIMag, AIRev, MLJ, JMLR, TKDD, TPAMI, TIST, IDA
2. Cech, (2014)

هوش مصنوعی وجود ندارد، هنوز معلوم نیست که در آموزش‌های اخلاقی، دقیقاً چه مطالبی باید وجود داشته باشد و چه کسی باید این آموزش‌ها را دریافت کند.

اب. فلسفه و اخلاق

ما مجموعه‌ای از مقالات نمایه‌شده در Philpapers.org را که یکی از نمایه‌های اصلی در فلسفه جهانِ انگلیسی‌زبان در بسیاری از موضوعات اخلاق هوش مصنوعی محسوب می‌شود مرور کردیم.

اغلب این مقاله‌ها بر اهمیت اخلاقی انواع و اقسام فناوری‌های هوش مصنوعی پیشرفته‌ای که ممکن است در آینده به وجود بیایند، تمرکز دارند[1]. پرسش‌های مورد بررسی عبارت‌اند از: آیا و چگونه عامل‌های مصنوعی ممکن است تصمیمات اخلاقی اتخاذ کنند؟ ماشین‌های هوشمند، در چه لحظه‌ای (اگر اصلاً چنین لحظه‌ای وجود داشته باشد) ممکن است دارای همان شأن اخلاقی شوند که معمولاً به انسان‌ها نسبت می‌دهیم[2]، و اینکه دلالت‌های ماشین‌های اَبَرهوشمند برای مفاهیمی مثل خودمختاری و چیستی انسان بودن، چیست؟[3]

کار روی فناوری‌های کنونی در ادبیات اخلاق هوش مصنوعی چندان مورد بحث واقع نشده است. در عوض این موضوعات غالباً تحت عنوان «اخلاق اطلاعات» یا «اخلاق رایانه» پوشش داده شده‌اند[4]. مقالات اولیه در این حوزه مسائلی مانند مسئولیت‌پذیری، سوگیری و ارزش‌باری داده‌ها و تصمیم‌گیری الگوریتمی را برجسته کرده است. اخیراً کارهای بیشتری

1. https://philpapers.org/browse/ethics-of-artificial-intelligence

۲. نگاه کنید به:
Boddington et al. (2017); Gunkel et al. (2014); Muller (2014); Wallach and Allen (2009); Allen, Varner and Zinser (2010); Anderson and Anderson (2007)

3. Bostrom (2003)

۴. برای مثال نگاه کنید به: Dignum (2018) and Brynum (2015)

برای تجزیه و تحلیلِ نحوۀ کاربست مفاهیم کلیدی مثل شفافیت، سوگیری، انصاف یا مسئولیت‌پذیری در فناوری‌های مبتنی بر ADA انجام شده است. این ادبیات، اغلب در ژورنال‌های ویژه (*Ethics and Information Technology, Philosophy & Technology, Science and Engineering Ethics*) یا در مقالات حوزه‌های فنی ADA منتشر شده‌اند.

با این وجود به نظر می‌رسد که کارهای نظام‌مندی که دلالت‌های اخلاقی فناوری‌های کنونی ADA را از یک منظر فلسفی تحلیل می‌کند، با یک فقدان نسبی مواجه هستند.

ا‌ج: حقوق

ادبیات دانشگاهی حقوق که ما پوشش دادیم این پرسش‌ها را بررسی می‌کرد: به‌ویژه با توجه به پیشرفت‌های سریع هوش مصنوعی و یادگیری ماشین، چگونه می‌توان از حقوق (موجود و بالقوه) برای کاهش تهدیدات فناوری‌های مبتنی بر ADA استفاده کرد؟ برخی از پرسش‌های کلیدی که بررسی شدند عبارت‌اند از: مقررات کنونی، در عمل، چه معنایی برای استفاده از داده و الگوریتم‌ها دارد (برای مثال: GDPR تا چه اندازه «حق تبیین و چه نوع حق تبیینی» را اجبار می‌کند؟)[1] یا آیا چنین مقرراتی می‌توانند در عمل، مسائل را حل کنند (برای مثال: حق تبیین، تا چه اندازه به مسائلی همچون حریم خصوصی، خودمختاری اطلاعاتی و اعتماد، کمک می‌کند؟)[2] و اینکه مجموعه قوانین کنونی، چگونه می‌توانند با مسئله مسئولیت‌پذیری و پاسخ‌گویی برخاسته از به‌کارگیری روزافزون هوش مصنوعی، مواجه شود[3].

۱. برای مثال نگاه کنید به:
Goodman and Flaxman (2016); Wachter, Mittelstadt, and Floridi (2017); Selbst and Powles (2017)
2. Edwards and Veale (2017)
3. Kuner et al. (2017)

عـلاوه بـر پرسـش از قوانیـن ضـروری معطـوف بـه اسـتفاده از هـوش مصنوعـی، داده و الگوریتم‌هـا، همچنیـن پرسـش‌هایی در ایـن مـورد وجـود دارد کـه ایـن فناوری‌هـا چگونـه می‌تواننـد خـودِ فرآیندهـای قانونـی را متأثـر کننـد - بـرای مثـال، نحـوه ارائـه شـهادت و شـواهد را تغییـر دهنـد. ادبیـات حقوقـی، بیـش از سـایر حوزه‌هـا تـلاش می‌کنـد تـا تفاسـیر مختلـف از واژه‌هـای مبهـم -مثـل حریـم خصوصـی یـا انصـاف- را از یکدیگر متمایـز کنـد و دلالت‌هـای آن‌هـا را بررسـی نمایـد.

۱د. تعامل انسان- ماشین[1]

HMI یـک حـوزه بین‌رشـته‌ای اسـت کـه در آن فلسـفه، روان‌شناسـی، علوم‌شـناختی، علـوم رایانـه، مهندسـی و طراحـی بـا یکدیگـر ترکیـب شـده‌اند. مـا مقـالات ایـن حـوزه را در ژورنال‌هـای مهـم مثـل *Human-Computer Interaction, Neuroethics, and AI and Society* بررسـی کردیـم[2].

مسـائل اخلاقـی و اجتماعـی تکرارشـونده‌ای کـه در ادبیـات مربـوط مـورد بحـث قرارگرفته‌انـد عبارت‌انـد از:

تأثیـرات روان‌شـناختی و عاطفـی تعامل‌هـای مختلـف انسـان-رایانه و همچنیـن تأثیـرات آن‌هـا بـر بخش‌هـای دیگـر جامعـه مثـل اقتصـاد و نگرانی‌هایـی در مـورد مسـئولیت، خودمختـاری، کرامـت، حریـم خصوصـی و مسـئولیت‌پذیری.

یکـی از وجـوه جالـب توجـه ایـن ادبیـات، توجـه به این اسـت کـه قوانین کنونـی چگونـه می‌تواننـد بـا مسـئله مسـئولیت‌پذیری و پاسـخ‌گویی ناشـی از به‌کارگیـری روزافـزون هـوش مصنوعـی مواجـه شـوند.

1. Human Machine Interaction (HMI) (مترجم)

۲. مقالات مهم شامل:
Becker (2006); Clark (2008); Jotterand and Dubljevic (2016); Menary (2007); Parens (1998); Schermer (2013); Sharkey (2014)

۱ه. علوم سیاسی و اجتماعی

ما ادبیات گسترده و روبه‌رشد علوم سیاسی، اجتماعی و اقتصادی را که از نحوه تأثیر ADA بر جامعه بحث کرده‌اند، پوشش دادیم.

اصلی‌ترین مسائلی که در این ادبیات پوشش داده شده‌اند، عبارت‌اند از: ADA چگونه رشد اقتصادی را تحت تأثیر قرار خواهد داد و اقتصاد را به‌طورکلی و شغل و بازار کار را به‌طورخاص[1] آشفته خواهد نمود، و تلاش متخصصان برای پیش‌بینی اینکه خودکارسازی چگونه و با چه سرعتی شغل را تحت تأثیر قرار خواهد داد و اینکه برای واکنش در برابر بیکاری ناشی از این فناوری‌ها از چه سیاست‌هایی می‌توان بهره جست (از جمله آموزش و طرح‌های توزیعی مثل درآمد اولیه همگانی[2]). مسئله کلیدی دیگر، تأثیر ADA بر نابرابری و رفاه جهانی است با نگرانی از اینکه فناوری شکاف بین کشورهای توسعه‌یافته و درحال‌توسعه را احتمالاً افزایش خواهد داد[3].

و نهایتاً بحث‌هایی وجود دارد در این مورد که ADA چگونه سیاست‌های ملی و بین‌المللی را متأثر خواهد کرد: سیاست، قدرت و دموکراسی درجامعه‌ای که به‌گونه‌ای روزافزون توسط ADA کنترل می‌شود[4] چگونه خواهند بود؟ و استفاده از تسلیحات خودمختار و خطر یک مسابقه تسلیحات هوش مصنوعی چگونه ثبات بین‌المللی را تهدید خواهد کرد.

۱و. سایر ادبیات‌های میان‌رشته‌ای

در نهایت ما این را بررسی کردیم که مسائل اخلاقی پیرامون ADA چگونه در سطح میان‌رشته‌ای فلسفه، حقوق و یادگیری ماشین بحث شده‌اند[5].

1. Frey (2017); Kaplan (2015); Marcus (2012); Marien (2014)
2. universal basic income(مترجم)
3. Eubanks (2018)
۴. برای مثال نگاه کنید به: Monbiot (2017) and Helbing et al. (2017)
۵. برای مثال نگاه کنید به:
Lipton (2017), Weller (2017), Binns (2018), Tene and Polonetsky (2017), and Selbst and Barocas (2016)

ارجاعـات بین‌رشـته‌ای چشـمگیری در میـان رشـته‌های مختلـف وجـود دارد، و نیـز همگـی بـر مجموعـه مفاهیـم کلیـدی مشـترکی تمرکـز دارنـد از جملـه: انصـاف، پاسـخ‌گویی، شـفافیت، سـوگیری، تبعیـض، تبیین‌پذیـری، حریـم خصوصـی و امنیـت.

امـا ایـن واژگان کلیـدی اغلـب بـدون تأمـل، بـه انحـاء گوناگـون و ناسـازگار و بـدون تحلیـل بیشـتر اسـتفاده شـده‌اند-برای مثـال پیشـنهاد دادن روش‌هایـی بـرای افزایـش شـفافیت، بـدون روشـن‌کردنِ معنـا و چرایـی اهمیـت آن.

۲. علم و رسانه‌های عامه‌پسند

مـا چگونگـی مـورد بحـث قرارگرفتـنِ مسـائل اخلاقـی و اجتماعـی ADA در رسـانه‌ها و علـم عامه‌پسـند را مـورد پیمایـش قـرار دادیـم، چراکـه توجهـات بسـیاری را بـه خـود جلـب کرده‌انـد.

۲آ. کتاب‌های علمی عام‌پسند

بـا بررسـی تعـداد زیـادی کتاب‌هـای علمـی عامه‌پسـند در مـورد هـوش مصنوعـی، پـی بـه دو موضـوع غالـب بردیم:

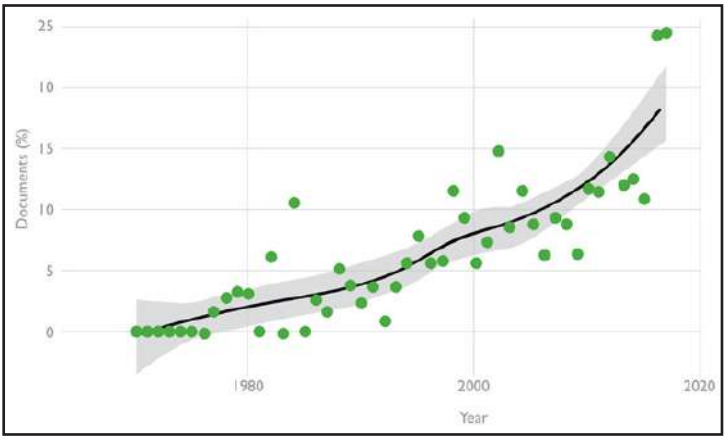
الـف) تهدیـدات ناشـی از اَبَرهـوش مصنوعـی و به‌ویـژه چالـش هم‌راسـتاکردنِ هـوش مصنوعـی پیشـرفته بـا ارزش‌هـای انسـانی و ب) آینـده کار و تأثیـرات بالقـوه خودکارسـازی. سـایر مسـائلی که کمتـر اهمیت داشـتند عبارت‌انـد از: آیـا یـک ماشـین می‌توانـد مسـئول اَعمالـش باشـد یـا حـق و حقوقـی داشـته باشـند؟ و اینکـه چگونـه می‌تـوان جلـوی سـوء اسـتفاده از کلان‌داده‌هـا را گرفـت؟ [۱]

۱. مهم‌ترین کتاب‌های پوشش داده شده عبارتند از:
Barrat (2013); Bostrom (2014); Brynjolfsson and McAfee (2014); Chace (2015); Collins (2108); Eubanks (2018); Harari (2015); Kurzweil (2012); McFarland (2008); Moravec (1988); Noble (2018); O'Connell (2017); O'Neil (2016); Shanahan (2015); Tegmark (2017); Wachter-Boettcher (2017); Wallach and Allen (2009); Walsh (2017); Zarkadakis (2015).

۲ب. رسانه‌های عامه‌پسند

مقاله‌های منتشرشده در رسانه‌های عامه‌پسند اخیراً توجه بیشتری داشته‌اند به خطرات ناشی از کاربردهای کنونی ADA به‌ویژه خطرات بالقوه آن برای سوگیری علیه جوامع نادیده گرفته شده و تهدید حریم خصوصی به‌واسطه استفاده از حجم عظیمی از داده‌های شخصی[۱].

سنجش اینکه این نگرانی‌ها از منظر عامه، چقدر اهمیت دارند ساده نیست، اما با استفاده از برخی شاخص‌ها می‌توان یک تصور کلی و ابتدایی از رشد آن‌ها بدست آورد. برای مثال، تصویر ۴، درصد اسناد مربوط به اخلاق یا تأثیرات هوش مصنوعی را از مجموع



تصویر ۴: درصد مقاله‌های منتشرشده (نقاط سبز رنگ) در موضوعات هوش مصنوعی که دست‌کم یکی از کلیدواژه‌های‌شان به اخلاق و تأثیرات هوش مصنوعی ربط داشته است. همچنین گرایش‌های کلی (خط سیاه) و خطاهای استاندارد (باند خاکستری) نیز نمایش داده شده‌اند.
روش‌شناسی [این مطالعه] به‌طور کامل در اینجا توضیح داده شده است: Martínez-Plumed et al. "The 2018 Facets of AI", IJCAI

۱. برای مثال نگاه کنید به
Crawford (2016); Li (2018); Shariff, Rahwan and Bonnefon (2016); Mukherjee (2017); Kleinberg, Ludwig and Mullainathan (2016).

نظرگاهی مشابه را از مقاله نیویورک تایمز می‌توان بدست آورد چنان‌که فَست و هُرویتز (۲۰۱۷) تحلیل کرده‌اند و به این نتیجه رسیده‌اند که بحث‌های هوش مصنوعی از سال ۲۰۰۹ به این سو رشد انفجاری داشته‌اند اما سطح بدبینی و خوش‌بینی متعادل باقی مانده است.

۳- گرایش‌های سیاستی و چشم‌انداز بین‌المللی گسترده‌تر

سازمان‌ها و نهادهای حاکمیتی مختلفی در انگلستان، آمریکا و سراسر جهان شروع به نگاشت پرسش‌های اخلاقی اجتماعی برخاسته از ADA کرده‌اند، با تمرکز بر دلالت‌های عملی آن‌ها برای سیاست‌گذاری و با تمرکز ویژه بر هر نوع تفاوت بین‌المللی، ما بررسی کردیم که این گزارش‌های سیاستی چگونه مسائل را در این فضا مورد بحث قرار داده‌اند[۱].

این گزارش‌های سیاستی طبیعتاً به طیف گسترده‌ای از مسائل می‌پردازند که ما بسیاری از آن‌ها را در بخش‌های مختلف ادبیات دانشگاهی پوشش دادیم. به‌ویژه روی این مسائل تمرکز وجود داشت: مدیریت و کاربرد داده، نمونه‌های آماری انصاف و سوگیری، شفافیت، تفسیرپذیری، مسئولیت‌پذیری، پاسخ‌گویی، آینده کار و تأثیر اقتصادی. گزارش‌های بخش‌های گوناگون جهان بر مسائل مختلف و متفاوتی متمرکز بودند. در کشورهای توسعه‌یافته تمرکز عمده بر به‌کارگیری ایمن و خطرات بالقوه فناوری بود. در کشورهای درحال‌توسعه تمرکز بحث‌ها بیشتر بر ظرفیت مساوی و زیست‌بوم فناوری و پژوهش بود.

۱. از جمله
EU EDPS Advisory Group (2018); Future Advocacy and the Wellcome Trust (2018); Government Office for Science (2016); IEEE (2018); Omidyar Network and Upturn (2018); National Science and Technology Council (2016); Royal Society (2017); Royal Society and the British Academy (2017); Select Committee on Artificial Intelligence (2018).

۴- پژوهش‌های مشارکت عمومی

مشارکت عمومی تلاشی است برای درگیرکردنِ اعضای گروه‌های عمومی مختلف در فرآیند تصمیم‌گیری در مورد مسائل مختلف سیاستی، علم، پزشکی و فناوری. این کار شامل روش‌های مختلف مانند رأی‌گیری، پیمایش، مشاوره و مجامع شهروندی می‌شود. تاکنون پیمایش‌های متعددی برای نگاشت وجوه مختلف درک عمومی از ADA انجام شده است، از جمله: Ipsos MORI and the Royal Society(2016/2017) اولین گفتگوی عمومی در مورد یادگیری ماشین را در انگلستان اجرا کردند، و رویکردهای عمومی به سلامت، مراقبت عمومی، بازاریابی، حمل‌ونقل، تجارت، پلیس، جرم، آموزش و هنر را مورد بررسی قرار دادند. این ابتکار از بحث‌های دو طرفه و از پیمایش استفاده می‌کرد و مشخص کرد که فقط ۹٪ افراد چیزهایی از یادگیری ماشین به گوش‌شان خورده است[1].

RSA (۲۰۱۷) شیوه‌های ایجاد مشارکت در شهروندان در مورد بحثِ به‌کارگیری اخلاقی فناوری‌های هوش مصنوعی را بررسی کرد و پی بُرد که بخش بسیار کمی از افراد اطلاع دارند که تصمیم‌گیری خودکار چه تأثیری روی زندگی‌شان دارد[2].

دفتر کابینه (۲۰۱۶) با استفاده از پیمایش و یک بازی داده‌ای این را مورد بررسی قرار داد که عموم مردم چه وزنی به خطرات استفاده از یادگیری ماشین و هوش مصنوعی در تصمیمات کابینه می‌دهند. آن‌ها نیز متوجه شدند که آگاهی عمومی از علوم داده محدود است[3].

Wellcome Trust (۲۰۱۶) با استفاده از یک گفتگوی وسیع عمومی، قابل‌پذیرش‌بودنِ عمومی دسترسی تجاری به داده‌های بیماران را

1. Ipsos MORI and the Royal Society, (2016); Ipsos MORI and the Royal Society. (2017).
2. Royal Society for the encouragement of Arts, Manufactures and Commerce (RSA), (2018).
3. Cabinet Office. Public dialogue on the ethics of data science in government (2016). www.ipsos.com/sites/default/files/2017-05/data-science-ethics-in-government.pdf

مـورد بررسـی قـرار داد. ایـن گفتگـو بـا یـک پیمایـش کمّـی همـراه بـود. آن‌هـا نتیجـه گرفتنـد کـه بـدون وجـود یـک نفـع عمومـی مشـخص، عموم مـردم در مـورد دسترسـی تجـاری بـه داده‌هـای مراقبـت بهداشـتی بسـیار نگـران هسـتند[1].

مرجـع تحقیقـات بهداشـتی و سـازمان بافـت انسـانی سـه گفتگـوی محلـی[2] را بـا عمـوم مـردم و ذینفعـانِ دانشـمند در مـورد رضایـت نسـبت به اشـتراک‌گذاری داده‌هـا در ژنتیـک تسـهیل کردنـد. شـرکت‌کنندگان ابـراز نگرانـی کردنـد کـه اگـر اکنـون نسـبت بـه اسـتفاده از داده‌هـای خـود در آینـده رضایـت دهنـد، در صـورت تغییـر قوانیـن ممکـن اسـت ناخواسـته به یـک جامعـهٔ دو لایـه کمـک کننـد کـه در آن افـراد می‌تواننـد بـر اسـاس ژنـوم خـود مـورد تبعیـض قـرار گیرنـد[3].

فرهنگسـتان علـوم پزشـکی در مـورد نقـش هـوش مصنوعـی در مراقبت‌هـای بهداشـتی وارد یـک گفتگـوی عمومـی و گفتگـو بـا ذینفعـان شـد.

Deep Mind (۲۰۱۸) یـک ابتـکار مشـارکتی عمومـی و ذینفعـی را بـرای توسـعه اصـول و ارزش‌هـای رفتـار جهانـی سـازماندهی کـرد[4].

۶-۲- خلاصه

در یک سطح کلی، ما مسائل مشترک زیر را شناسایی کردیم:

- حصـول اطمینـان از شـفاف/ توضیح‌پذیـر/ تفسـیرپذیر بـودنِ الگوریتم‌هـا و سـامانه‌های هـوش مصنوعـیِ جعبـه سـیاه.

1. Wellcome Trust. Public attitudes to commercial access to patient data. (2016). www.ipsos.com/sites/default/files/publication/5200-03/sri-wellcome-trustcommercial-access-to-health-data.pdf
2. Location dialogues(مترجم)
3. www.hra.nhs.uk/about-us/what-we-do/how-involve-public-our-work/what-patients-and-public-think-about- health-research/
4. https://deepmind.com/applied/deepmind-health/transparency-independent-reviewers/developing-our- values/#image-27248

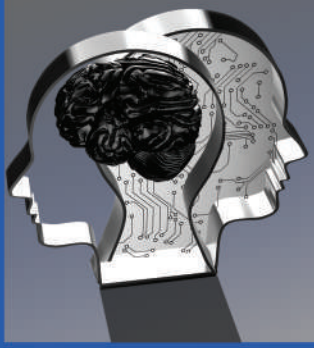
- حصول اطمینان از قابل اعتماد و قدرتمند بودنِ استفاده از ADA.
- حفظ حریم خصوصی و داده‌های شخصی.
- حصول اطمینان از اینکه الگوریتم‌ها و سامانه‌های هوش مصنوعی به‌نحو منصفانه به‌کار می‌روند و حاوی سوگیری‌های تاریخی نبوده یا منجر به سوگیری‌ها و تبعیض‌های جدید نمی‌شوند.
- حصول اطمینان از بازتاب ارزش‌های انسانی در الگوریتم‌ها.
- پرسش از اینکه آیا سامانه‌های هوش مصنوعی، هرگز قادر به تصمیم‌گیری‌های اخلاقی خواهند بود یا خیر.
- پرسش از اینکه آیا سامانه‌های هوش مصنوعی، هرگز شأن اخلاقی خواهند داشت یا خیر.
- مسائل پاسخ‌گویی و مسئولیت‌پذیری در کاربرد ADA.
- نقش قانون و آموزش‌های اخلاقی در تضمین استفاده مسئولانه از ADA.
- نقش قانون و آموزش در کاهش خطرات و افزایش سودمندی هوش مصنوعی.
- ایجاد سطح مناسبی از اعتماد در بین انسان‌ها و الگوریتم‌های ماشینی.
- دلالت‌های ADA برای عاملیت انسانی، خودمختاری و کرامت.
- تأثیرات ADA بر اقتصاد و رشد اقتصادی.
- تأثیرات ADA بر شغل و بازار کار، توسعه سیاست‌گذاری در مورد بیکاری فناورانه.
- تأثیر ADA بر نابرابری جهانی.
- تاثیر ADA بر سیاست‌های ملی، عقیده عموم و دموکراسی - از

جملـه اینکـه بیـکاری چـه تأثیـرات مُخربـی بـر عقیـده عمـوم خواهـد داشـت.

- ADA چگونه قدرت را در یک جامعه تغییر می‌دهد.
- چگونـه ADA ممکـن اسـت بـرای جهت‌دهـی بـه توجـه یـا دسـتکاری در عقایـد مـا بـه‌کار بـرود (بـرای اغـراض سیاسـی یا تجـاری).
- تأثیـر ADA بـر روابـط، ناسـازگاری و امنیـت بین‌الملـل - از جملـه تأثیـر سـلاح‌های خودمختـار و خطـر یـک مسـابقه تسـلیحاتی جهانـی.
- بـرای مواجهـه بـا چالش‌هـای ناشـی از فناوری‌هـای در حـال قدرتمندتـر شـدن، نیـاز بـه چـه شـیوه‌های حکمرانـی جهانـی اسـت.

بخش هفتم

پیوست ۲: گروه‌بندی‌ها و اصول



بخش هفتم

پیوست ۲: گروه‌بندی‌ها و اصول

در زیر برخی از شیوه‌های رایج دسته‌بندی و ساختاردهی به مسائل مربوط به ADA که سازمان‌ها از آن‌ها استفاده کرده‌اند را به همراه مجموعه اصول پیشنهادشده فهرست کرده‌ایم.

یک خط فارق بین آن‌دسته از رویکردهایی است که فهرست بلندی از اصول را ارائه داده‌اند (Asilomar, Partnership on AI) و آن‌هایی که فقط از چهار مقوله استفاده کرده‌اند (مانند گزارش سال 2017AI Now). هر یک از این رویکردها محاسن و معایب خاص خود را دارد: فهرست‌های کوتاه‌تر برای فراهم‌کردنِ یک منظر کلی از یک حوزه پیچیده مفیدند، اما خطر حذف تِم‌های مهم و تأثیرگذار را به همراه دارند. این فهرست‌ها فقط از طریق وسیع‌تر و نادقیق‌کردنِ مقوله‌هاست که می‌توانند کامل و جامع باشند. از سوی دیگر فهرست‌های بلندتر به‌منظور جامع و کامل‌بودن تهیه شده‌اند، اما خطر ازدست‌دادنِ یک منظر کلی روشن را به همراه دارند و نیز ممکن است شامل مقوله‌هایی شوند که به‌نحو ناسودمندی با یکدیگر هم‌پوشانی دارند.

یک راهبرد برای متوازن‌کردنِ این دو رویکرد، تکیه بر یک چارچوب تحلیلی وسیع‌تر است. برای مثال گروه مشاوره اخلاقی EDPS، پیشنهاد استخراج مهم‌ترین مسائل را از هشت ارزش اروپایی داده است، حال آنکه کاوِلز و فلوریدی (۲۰۱۸) پیشنهاد می‌دهند که مسائل مهم را می‌توان نتیجه این دانست که فناوری‌ها یا بیش از حد استفاده شده‌اند، یا سوء استفاده شده‌اند یا به اندازه کافی استفاده نشده‌اند. در نسبتِ با چهار نکته مهم در فهم کرامت یا شکوفایی بشری: ما می‌توانیم تبدیل به چه کسی بشویم (خودشکوفایی خودمختارانه)، ما چه می‌توانیم بکنیم (عاملیت انسانی)، ما به چه چیزی می‌توانیم دست بیابیم (توانایی‌های اجتماعی)، و ما چگونه می‌توانیم با یکدیگر و با جهان تعامل کنیم (همبستگی اجتماعی).

اگر چه این چارچوب‌ها می‌توانند نوعی توازن بین پیچیدگی و نظام‌مندی ایجاد کنند، اما همچنان با خطر نادیده انگاشتن برخی مسائل مواجه‌اند، برای مثال، مشخص نیست که مسائل سوگیری و تبعیض را در کجای لیست کاول و فلوریدی باید جای داد.

به‌علاوه اینکه، این نوع چارچوب‌های نظام‌مند عموماً یک قضاوت در مورد این را که ارزش‌های بنیادین کدامند که باید در ساختاربندی چارچوب به‌کار بروند، پیشاپیش مفروض می‌انگارند. این می‌تواند در بافتارهایی که یک تعهد پیشینی به آن ارزش‌ها وجود دارد، مفید باشد (چنان‌که می‌توان در مورد ارزش‌های اروپایی در ساختار اتحادیه اروپا این کار را کرد). اما نمی‌توان یک توافق همه‌جانبه در مورد آن ارزش‌ها را مفروض انگاشت.

انحاء مختلف تقسیم‌بندی فضای تأثیرات اخلاقی و اجتماعی ADA،

می‌توانـد اهـداف مختلفـی را دنبـال کننـد - بـرای مثـال فراهم‌کـردنِ یـک منظـر کلـی، فراچنـگ‌آوردنِ همـه مسـائل مهـم یـا فراهـم‌آوردنِ توصیه‌هـای به‌لحـاظ عملـی مهـم و مرتبـط. چارچوب‌هـای گوناگونـی کـه در پاییـن پایـش شـده‌اند می‌تواننـد بـرای همـه ایـن اهـداف سـودمند باشـند. محـل تردیـد اسـت کـه یـک چارچـوب به‌تنهایـی بتوانـد کل ایـن حـوزه را فراچنـگ بیـاورد و همـه‌ی اهـداف را بـرآورده کنـد و ایـن البتـه بـرای پیشـرفت سـازنده در ایـن مسـائل نـه لازم اسـت و نـه کافـی. در عـوض، تـلاش بـرای نگاشـت و سـازماندهی‌کردنِ مسـائل مهـم را بایـد به‌مثابـه ابزارهـای سـودمند بافتـاری بـرای اهـداف خـاص درک کـرد.

۷-۱- شیوه‌های رایج سازماندهی مسائل

گزارش Now AI ۲۰۱۷، چهار محل تحرک را شناسایی می‌کند:

۱. نیروی کار و خودکارسازی
۲. سوگیری و حذف
۳. حقوق و آزادی‌ها
۴. اخلاق و جامعه

گـروه اخـلاق و جامعـه Deep Mind، کارهایشـان را بـه شـش طـرح پژوهشـی تقسـیم کرده‌انـد:

۱. حریم خصوصی، شفافیت و انصاف
۲. تأثیر اقتصادی، شمول و برابری
۳. حکمرانی و پاسخ‌گویی
۴. اخلاق و ارزش‌های هوش مصنوعی

۵. مدیریت خطرات، سوء استفاده و پیامدهای ناخواسته هوش مصنوعی
۶. هوش مصنوعی و چالش‌های پیچیده جهان

The Partnership on AI از شش طرح زیر استفاده می‌کند:
۱. هوش‌مصنوعیِ ایمنی - انتقادی
۲. هوش مصنوعیِ منصفانه، شفاف و پاسخ‌گو
۳. همکاری بین انسان و سامانه‌های هوش مصنوعی
۴. هوش مصنوعی، نیروی کار و اقتصاد
۵. تأثیرات اجتماعی هوش مصنوعی
۶. هوش مصنوعی و خیر اجتماعی

گروه مشاوره اخلاقی EPDS، هفت چرخش عمده اجتماعی-فرهنگی را در عصر دیجیتالی برجسته می‌کند:
۱. از سوژه فردی به سوژه دیجیتالی
۲. از حیات آنالوگ به حیات دیجیتالی
۳. از حکمرانی توسط نهادها به قابلیت حکمرانی از طریق داده‌ها
۴. از یک جامعه در خطر به یک جامعه ممتاز
۵. از خودمختاری انسان به همگرایی انسان و ماشین
۶. از مسئولیت فردی به مسئولیت توزیع‌شده
۷. از عدالت کیفری به عدالت بازدارنده[1]

و تأثیرات فناوری‌های دیجیتال را بر ارزش‌های زیر در نظر بگیرید:
۱. کرامت

1. pre-emptive justice (مترجم)

۲. آزادی
۳. خودمختاری
۴. همبستگی
۵. برابری
۶. دموکراسی
۷. عدالت
۸. اعتماد

گزارش کاخ سفید اوباما: «آماده‌شدن برای آینده هوش مصنوعی» به بخش‌های زیر تقسیم می‌شود:

۱. کاربردهای هوش مصنوعی به نفع خیر عمومی
۲. هوش مصنوعی در حکمرانی
۳. هوش مصنوعی و مقررات
۴. پژوهش و نیروی کار
۵. هوش مصنوعی، خودکارسازی و اقتصاد
۶. انصاف، ایمنی و حکمرانی
۷. ملاحظات جهانی و امنیت

گزارش مشترک The Royal Society and British Academy (2017) از مقوله‌های زیر بهره می‌برد:

۱. ایمنی، امنیت، پرهیز از آسیب
۲. مسئولیت اخلاقی انسان
۳. حکمرانی، مقررات، پایش، سنجش

۴. تصمیم‌گیری دموکراتیک
۵. تبیین‌پذیری و شفافیت

گروه اروپایی اخلاق در علم و فناوری‌های جدید تقسیم‌بندی زیر را ارائه می‌دهد:

۱. حریم خصوصی و رضایت
۲. انصاف و کلیشه‌های آماری
۳. تفسیرپذیری و شفافیت
۴. شخصی‌سازی، حباب‌ها و دستکاری
۵. عدم‌توازن‌های قدرت و نابرابری
۶. آینده کار و اقتصاد
۷. تعامل انسان - ماشین

۷-۲- اصول و کدها[۱]:

اصول هوش مصنوعی Asilomar شامل اصول «اخلاقی و ارزشی» زیر می‌شود[۲]:

- ایمنی: سامانه‌های هوش مصنوعی بایستی در طول دوران عملیاتی خود ایمن و امن باشد.
- شفافیت قضایی: هر نوع دخالت سامانه‌های خودمختار در تصمیم‌گیری‌های قضایی بایستی به‌نحو رضایت‌بخشی برای یک مرجع ذی‌صلاح انسانی قابل تبیین باشد.
- مسئولیت‌پذیری: طراحان و سازندگان سامانه‌های هوش مصنوعی پیشرفته با مسئولیت و فرصت شکل‌دهی به این پیامدها ذینفعان

۱. این‌ها خلاصه‌ای از جنبه‌های مهم کدها و اصول مختلف است و لزوماً همهٔ اصول را منعکس نمی‌کند - برای مثال، برخی اوقات فقط شامل استفاده از عنوان و نه توضیح کامل هر اصل است.
۲. اصول کامل Asilomar شامل ده اصل دیگر دربارهٔ «موضوعات پژوهشی» و «موضوعات بلندمدت‌تر» است که در اینجا اشاره نکرده‌ایم.

پیامدهای اخلاقی استفاده، سوء استفاده و کنش‌ها محسوب می‌شوند.

- هم‌راستایی ارزشی: سامانه‌های هوش مصنوعیِ بسیار خودمختار باید طوری طراحی شوند که اهدافشان با ارزش‌های انسانی هم‌راستا باشد.

- ارزش‌های انسانی: سامانه‌های هوش مصنوعی باید طوری طراحی شوند و گونه‌ای عمل کنند که با ایده‌آل‌های کرامت بشری، حقوق، آزادی و تنوع فرهنگی سازگار باشند.

- حریم خصوصی: با توجه به قدرت سامانه‌های هوش مصنوعی در تحلیل و به‌کارگیری آن داده‌ها، افراد باید حق دسترسی، مدیریت و کنترل داده‌هایی که تولید کرده‌اند را داشته باشند.

- آزادی و حریم خصوصی: کاربست هوش مصنوعی در داده‌های شخصی نباید به‌گونه‌ای نامعقول، آزادی افراد را از میان ببرد.

- منافع مشترک: فناوری‌های هوش مصنوعی بایستی تا جای ممکن، افراد بیشتری را توانمند و منتفع کنند.

- کامیابی مشترک: کامیابی اقتصادیِ خلق‌شده توسط هوش مصنوعی باید در سطحی وسیع و برای منتفع‌کردنِ همه بشریت به اشتراک گذارده شود.

- کنترل انسان: انسان‌ها باید تصمیم بگیرند که به‌منظور تکمیل اهداف انتخاب‌شده توسط انسان‌ها کی و چگونه تصمیماتشان را به سامانه‌های هوش مصنوعی محول کنند.

- عدم تخریب: قدرت ناشی از کنترل سامانه‌های بسیار پیشرفته هوش مصنوعی بایستی فرآیندهای مدنی را که سلامت جامعه به

آن‌ها وابسته است، در نظر بگیرند.
- مسابقه تسلیحاتی هوش مصنوعی: باید از یک مسابقه تسلیحاتی در مورد سلاح‌های خودمختارِ کشنده پرهیز کرد.

Partnership on AI به بخش‌های زیر باور دارد:

۱. ما در پی حصول اطمینان از این هستیم که فناوری‌های هوش مصنوعی تا جای ممکن افراد بیشتری را توانمند و منتفع کنند.
۲. ما به عموم مردم آموزش و گوش خواهیم داد و به‌نحو فعالانه‌ای ذینفعان را درگیر کرده و آن‌ها را در جریان کارهای خود قرار داده و سؤالاتشان را بررسی خواهیم کرد.
۳. ما متعهد به برقراری گفتگو در مورد دلالت‌های اخلاقی، اجتماعی، اقتصادی و حقوقی هوش مصنوعی هستیم.
۴. ما باور داریم که تحقیق و توسعه هوش مصنوعی نیازمند مشارکت فعالانه طیف وسیعی از ذینفعان است.
۵. ما با ذینفعان در جامعه تجاری وارد تعامل خواهیم شد تا از نگرانی‌های خاص آن‌ها مطلع شده و فرصت‌های آن‌ها را درک و بررسی کنیم.
۶. ما تلاش خواهیم کرد تا منافع هوش مصنوعی را حداکثر و چالش‌های بالقوه آن‌را بررسی کنیم از طریق:
a. تلاش برای محافظت از حریم خصوصی و امنیت افراد.
b. تلاش برای درک و ملاحظه منافع همه گروه‌های متأثر از پیشرفت‌های هوش مصنوعی.
c. تلاش برای حصول اطمینان از اینکه پژوهش‌های هوش مصنوعی و

جوامع مهندسی کماکان به‌لحاظ اجتماعی مسئولیت‌پذیر باقی می‌مانند و مستقیماً روی تأثیرات بالقوه هوش مصنوعی بر جامعه کار می‌کنند.

d. حصول اطمینان از اینکه تحقیق و فناوری هوش مصنوعی نیرومند، قابل اعتماد و ایمن است.

e. مقابله با توسعه و استفاده از هوش مصنوعی که ممکن است معاهدات بین‌المللی یا حقوق بشر را نقض کنند و ارتقای فناوری‌هایی که آسیب به همراه ندارند.

۷. به عقیده ما قابل فهم و تفسیر بودنِ کارکرد هوش مصنوعی توسط انسان‌ها اهمیت دارد.

۸. ما تلاش می‌کنیم تا یک فرهنگ مبتنی بر همکاری، اعتماد و گشودگی در میان دانشمندان و مهندسان هوش مصنوعی ایجاد کنیم که به همه برای دستیابی بهتر به این اهداف کمک می‌کند.

گزارشِ کمیتهٔ منتخبِ لُردها در هوش مصنوعی پنج اصل زیر را برای کدهای هوش مصنوعیِ میان‌بخشی پیشنهاد می‌دهد:

۱. هوش مصنوعی باید به نفع خیر عمومی و انتقاع بشریت توسعه یابد.

۲. هوش مصنوعی باید بر اساس اصول انصاف و قابل‌فهم‌بودن عمل کند.

۳. نباید از هوش مصنوعی به‌منظور محوکردنِ حقوق داده‌ای یا حریم خصوصی افراد، خانواده‌ها یا جوامع استفاده کرد.

۴. همه شهروندان باید حق برخورداری از آموزش و پرورش و شکوفایی ذهنی و اقتصادی را در راستای هوش مصنوعی داشته باشند.
۵. نباید روی توسعه قدرت آسیب‌رسان، مُخرب یا فریب‌دهنده هوش مصنوعی سرمایه‌گذاری کرد.

IEEE نیز مجموعه‌ای از اصول کلی را برای هدایت حکمرانی اخلاقی سامانه‌های خودمختار و هوشمند توسعه داده است:

۱. حقوق بشر
۲. اولویت رفاه
۳. پاسخ‌گویی
۴. شفافیت
۵. سوء استفاده از فناوری و آگاهی از آن

اصولِ **انجمن ماشین‌آلات رایانشی**[1] برای شفافیت و پاسخ‌گویی الگوی الگوریتمیک[2]:

۱. آگاهی
۲. توضیح
۳. تبیین‌پذیری
۴. اعتبارسنجی و سنجش

دستورالعمل‌های اخلاقی **جامعهٔ هوش مصنوعیِ ژاپن**[3][4] :

۱. کمک به بشریت

1. The Association for Computing Machinery (ACM) (مترجم)
۲. نگاه کنید به:
ASM US Public Policy Council's 'Statement on Algorithmic Transparency and Accountability' (2017)
3. The Japanese Society for Artifcial Intelligence (JSAI) (مترجم)
4. http://ai-elsi.org/wp-content/uploads/2017/05/JSAI-Ethical-Guidelines-1.pdf

۲. پیروی از قوانین و مقررات
۳. احترام به حریم خصوصی دیگران
۴. انصاف
۵. امنیت
۶. عمل بر اساس همبستگی
۷. پاسخ‌گویی و مسئولیت‌پذیری اجتماعی
۸. ارتباط با جامعه و خودشکوفایی
۹. پیروی از دستورالعمل‌های اخلاقی هوش مصنوعی

اصول حکمرانی هوش مصنوعیِ[1] **ابتکارِ جامعهٔ آیندهٔ علم، حقوق و جامعه**[2]:

۱. هوش مصنوعی نباید ضربه‌ای وارد کند و اگر بتواند باید برابری حقوق، کرامت، آزادی و شکوفایی را برای همه انسان‌ها تقویت کند.
۲. هوش مصنوعی باید شفاف باشد.
۳. تولیدکنندگان و به‌کارگیرندگان هوش مصنوعی باید پاسخ‌گو باشند.
۴. تأثیرگذاری هوش مصنوعی در کاربردهای آن در جهان واقع باید قابل سنجش باشد.
۵. کاربران سامانه‌های هوش مصنوعی باید صلاحیت و تخصص مناسب را داشته باشند.
۶. هنجارهای مربوط به واگذاری تصمیم‌ها به سامانه‌های هوش مصنوعی باید طی یک فرایند گفتگوی متأملانه و همه‌گیر با جامعه مدنی تنظیم شوند.

1. www.thefuturesociety.org/science-law-society-sls-initiative/#1516790384127-3ea0ef44-2aae
2. The Future Society's Science, Law and Society Initiative(مترجم)

ده اصل برتر **اتحادیه جهانی UNI** برای اخلاق هوش مصنوعی[1]:

۱. سامانه‌های هوش مصنوعی باید شفاف باشند.
۲. سامانه‌های هوش مصنوعی را مجهز به جعبه سیاه اخلاقی کنید.
۳. هوش مصنوعی را در خدمت انسان و کره زمین قرار دهید.
۴. رویکرد انسان -در- فرماندهی[2] را اتخاذ کنید.
۵. هوش مصنوعی را بدون جنسیت و بدون سوگیری توسعه دهید.
۶. منافع هوش مصنوعی را توزیع کنید.
۷. انتقال عادلانه را تضمین کنید و از حقوق و آزادی‌های بنیادین حمایت کنید.
۸. سازوکارهای حکمرانی جهانی را تأسیس کنید.
۹. مانع انتساب مسئولیت به ربات‌ها شوید.
۱۰. مانع مسابقه تسلیحاتی هوش مصنوعی شوید.

بیانیهٔ مونترئال دربارهٔ هوش مصنوعی مسئول[3] شامل اصول زیر می‌شود[4]:

۱. رفاه
۲. احترام به خودمختاری
۳. حفاظت از حریم خصوصی
۴. همبستگی
۵. مشارکت دموکراتیک
۶. برابری
۷. شمولِ گوناگونی

1. www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf
2. human-in-command(مترجم)
3. The Montréal Declaration for Responsible AI(مترجم)
4. www.montrealdeclaration-responsibleai.com/the-declaration

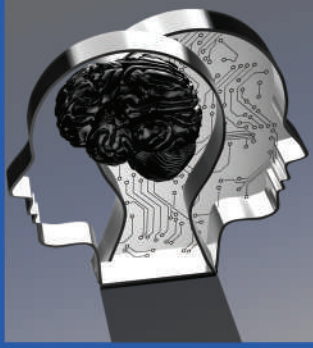
۸. احتیاط

۹. مسئولیت‌پذیری

۱۰. توسعه پایدار

بخش هشتم

پیوست ۳: دیدگاه‌های گوناگون



بخش هشتم

پیوست ۳: دیدگاه‌های گوناگون

کل فضای مباحث آثار اخلاقی و اجتماعی ADA بسیار وسیع و پیچیده است، آنچنان‌که نمی‌توان آن را در یک تک چارچوب، فراچنگ آورد. این باعث می‌شود تا فهم مسائل جزئی، بدون تمرکز بر آن و حذف برخی اطلاعات، ناممکن شود. برعکس، وجوه مهم یک مسئله خاص به‌سادگی ممکن است نادیده گرفته یا در پیچیدگی این فضا گم شود.

این پیوست، برخی از نظرگاه‌های پُراهمیت را که می‌توان در فضای اهمیت اخلاقی و اجتماعی ADA اتخاذ کرد طرح می‌کند، همراه با مثال‌هایی از زیر مجموعه‌های آن‌ها. می‌توان آن را همچون یک محور در نظر گرفت که می‌توان به‌منظور فیلترکردنِ اطلاعات روی بخش‌های مختلف آن تمرکز کرد.

از این نظرگاه‌ها می‌توان به‌صورت منفرد یا به‌صورت ترکیبی استفاده کرد تا به این ترتیب طیف مسائل را محدود یا یک مسئله را از نظرگاه‌های مختلف بررسی نمود تا مطمئن شد که حداکثر وجوه پُراهمیت ممکن، مورد ملاحظه قرار گرفته‌اند.

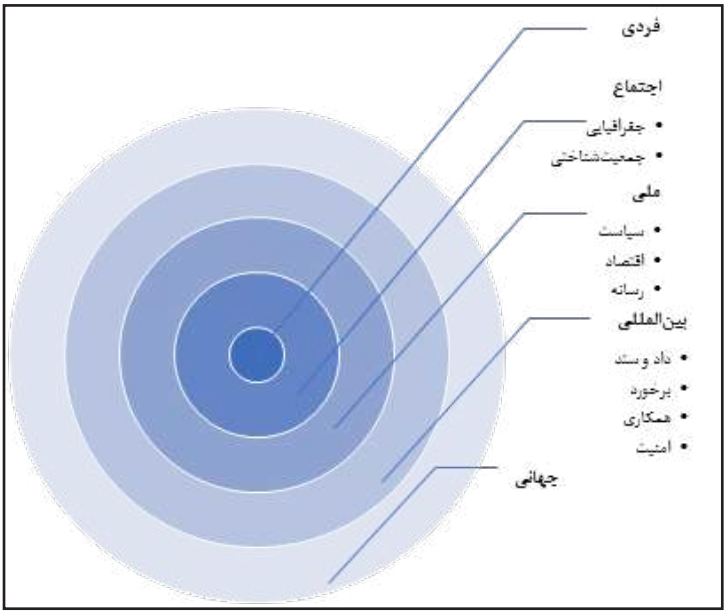
۸-۱- کدام قسمت‌ها یا بخش‌های جامعه؟

جوامع شامل بخش‌ها و قسمت‌های مختلف می‌شوند، چنان‌که حاکمیت‌ها نیز مدیریت خود را به چندین دپارتمان تقسیم می‌کنند. برای مثال می‌توان از دپارتمان‌های حکومت‌های انگلستان و آمریکا برای تمرکز بر تأثیرات ADA استفاده کرد.

کشاورزی
تجارت
فرهنگ، رسانه و ورزش
انرژی
آموزش و پرورش
محیط‌زیست
توسعه
داد و ستد
روابط بین‌المللی و نهادی
حمل و نقل
کار و نیروی کار
سلامت و مراقبت‌های اجتماعی
سرمایه‌گذاری و اقتصاد
امنیت و دفاع
اجتماع و مسکن
جرم و عدالت

۸-۲- چه سطحی از سازماندهی اجتماعی؟

روابط انسانی در سطوح مختلفِ سازمانِ اجتماعی ساختاربندی شده‌اند، از جوامع محلی گرفته تا روابط بین‌المللی. علاوه بر بررسی مسائل بر اساس بخش‌های حاکمیتی، می‌توان بر مسائلی که در سطوح مختلف سازمان اجتماعی به وجود می‌آیند نیز تمرکز کرد (تصویر ۵).



تصویر ۵: سطوح و بخش‌های مختلف جامعه که ممکن است تحت تأثیر فناوری ADA قرار گیرد.

۸-۳- کدام چارچوب زمانی؟

مسائل مرتبط با ADA ممکن است در مقیاس‌های زمانی مختلف ظهور کنند. به‌عنوان مثال ما می‌توانیم آن را به شیوه‌های زیر متمایز کنیم:

۱. چالش‌های زمان حال: چالش‌هایی که در حال حاضر از آن‌ها آگاهیم و با آن‌ها مواجهیم، کدامند؟
۲. چالش‌های آینده نزدیک: با توجه به فناوری کنونی، در آینده نزدیک، ممکن است با چه چالش‌هایی مواجه شویم؟
۳. چالش‌های بلندمدت: با پیشرفته‌ترشدنِ فناوری در طولانی‌مدت ممکن است با چه چالش‌هایی مواجه شویم؟

فکرکردن راجع به چالش‌های دسته اول بسیار راحت است، چراکه این چالش‌ها در حال حاضر جلوی ما بوده و مورد بحث هستند: برای مثال حفاظت از حریم خصوصی. چالش‌های دسته دوم و تصور اینکه چگونه فناوری-های کنونی چالش‌های جدیدی را در آینده نزدیک ایجاد خواهند کرد، نیازمند تأمل بیشتری هستند. یک مثال می‌تواند این باشد که چگونه می‌توان از تکنیک‌های ترکیب تصاویر سوء استفاده کرد.

چالش‌های دسته سوم، دشوارترین چالش‌ها برای پیش‌بینی هستند، زیرا مستلزم تأمل درباره تأثیر قابلیت‌های فناوری‌های آینده هستند. بحث در مورد تأثیرات اَبَرهوش مصنوعی می‌تواند مصداقی از آن باشد.

۸-۴- کدام گروه‌های عمومی؟

گروه‌های عمومی مختلف، نگران مسائل مختلفی هستند و نظرگاه متفاوتی نسبت به یک مسئله واحد دارند. تمایزهای ذیل میان گروه‌های عمومی مختلف و پرسش‌هایشان در مورد فناوری‌های

ADA، می‌تواند برای سازماندهی تحقیق در مورد اهمیت اخلاقی این فناوری‌ها مفید باشد:

- **طراحان و مهندسان:** من در قبال حصول اطمینان از اخلاقی‌بودنِ فناوری‌هایی که توسعه می‌دهم چه مسئولیتی دارم؟ چه استانداردهای اخلاقی را باید رعایت کنم؟ چگونه می‌توانم مطالباتی مثل حریم خصوصی، انصاف و شفافیت را از منظر فنی، دقیق کنم؟

- **کاربران/گروه‌های عمومی کلی:** یک فناوری چگونه زندگی روزانه من را متأثر می‌کند؟

چه بده-بستان‌های جدیدی را برای من پیش می‌آورد؟

- **گروه‌های به‌حاشیه‌رانده‌شده:** آیا این فناوری با توجه به وضعیت ما یک تهدید است یا یک فرصت؟ چگونه می‌توانیم از آن برای مقابله با پیش‌داوری‌ها استفاده کنیم؟

- **سازمان‌ها و نهادهای همکاری:** هزینه‌ها و فایده‌های خودکارسازی یک وظیفه/ خدمت خاص چیست؟ برای اطمینان از اخلاقی/ قابل‌اعتماد بودنِ استفاده‌مان از یک فناوری به چه چیزی نیاز داریم؟

- **سیاست‌گذاران و تنظیم‌کنندگان مقررات:** در کجا نیازمند سیاست‌گذاری یا مقررات هستیم؟ کجا نیازمند فشار افکار عمومی هستیم؟

- **NGOها و جامعه مدنی:** چگونه می‌توانیم از مشارکت عمومی گسترده مطمئن باشیم؟ علایق چه کسانی ممکن است که نادیده گرفته شده باشد؟

- **پژوهشگران:** استفاده از فناوری‌های جدید چه پرسش‌های جالب توجهی را مطرح می‌کند؟ چه مسائلی مستلزم تأملات عمیق‌تری هستند؟

- **روزنامه‌نگاران و مفسران:** کدام وجوه فناوری و تأثیراتشان بر جامعه را باید به اطلاع گروه‌های عمومی مختلف رساند؟ چگونه می‌توان این کار را به مؤثرترین شیوه ممکن انجام داد؟

۸-۵- کدام هوش؟

یک فناوری به دلایل مختلفی می‌تواند اشتباه کند. ما باید انواع و اقسام چالش‌های ممکن را در نظر داشته باشیم:

- فناوری چگونه طراحی شده یا توسعه یافته است (برای مثال چه سوگیری‌هایی ممکن است در داده‌های مورد استفاده در آموزش یک الگوریتم وجود داشته باشد: ناتوانی در بررسی اینکه یک الگوریتم دقیقاً چگونه از ویژگی‌های مختلف برای یک تصمیم‌گیری استفاده می‌کند).
- پیامدهای بیرونی یا ناخواسته چگونگی کاربرد فناوری در جامعه (برای مثال تأثیر خودکارسازی بر بازار کار، مسئله قابلیت اعتماد به الگوریتم‌هایی که در تصمیم‌گیری‌های مراقبتی-بهداشتی و سایر حوزه‌های مهم به‌کار می‌روند).
- پیامدهای عدم عملکرد صحیح و حوادث اتفاقی (تصادفِ خودروهای بی‌راننده).
- سوء استفاده از فناوری (مثلاً برای تجسس، دستکاری یا جرم).

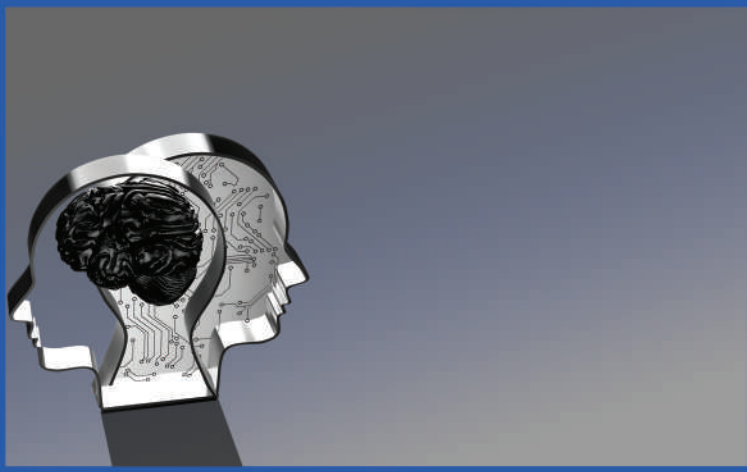
۸-۶- چه نوع راه‌حل‌هایی؟

ما سازوکارها و روش‌های متعدد و مختلفی برای بررسی چالش‌های اخلاقی و اجتماعی گوناگون برخاسته از ADA در اختیار داریم. برای

انواع مختلف مسائل نیازمند راه‌حل‌های مختلفی خواهیم بود و برای حل اغلب چالش‌ها نیازمند رویکردهای چندگانه مختلفی خواهیم بود. تفکر نظام‌مندتر درباره روش‌های گوناگون در دسترس برای مواجهه با این مسائل می‌تواند به شناسایی رویکردها و زوایای جدید کمک کند. برای مثال، راه‌حل‌های مختلف را می‌توان این‌گونه تقسیم‌بندی کرد:

- قوانین و مقررات: ملی و بین‌المللی
- سیاست‌های حاکمیتی
- مشارکت و آموزش عمومی
- عمل‌گرایی
- انواع و اقسام پژوهش‌ها:
  - پژوهش فنی
  - پژوهش فلسفی اخلاقی/ علوم انسانی
  - پژوهش علوم اجتماعی

منابع



- Acs, G., Melis, L., Castelluccia, C., & De Cristofaro, E. (2018). Differentially private mixture of generative neural networks. IEEE Transactions on Knowledge and Data Engineering.

- Adams, F. and Aizawa, K. (2001). The Bounds of Cognition. Philosophical Psychology, 14(1): 43–64.

- Adel, T., Ghahramani, Z., & Weller, A. (2018). Discovering Interpretable Representations for Both Deep Generative and Discriminative Models. In International Conference on Machine Learning: 50–59.

- Aditya, S. (2017). Explainable Image Understanding Using Vision and Reasoning. Paper presented at the AAAI.

- Aha, D. W., & Coman, A. (2017). The AI Rebellion: Changing the Narrative. Paper presented at the AAAI.

- AI Now Institute. (2017). AI Now Symposium 2017 Report.

- Alekseev, A. (2017). Artificial intelligence and ethics: Russian theory and communication practices. Russian Journal of Communication, 9(3): 294–296.

- Alexandrova A. and D. Haybron (2011) High fidelity economics, in Elgar Companion to Recent Economic Methodology (edited by John Davis and Wade Hands), Edward Elgar, 94–117.

- Alexandrova A. (2017) A Philosophy for the Science of Well-being, New York: Oxford University Press.

- Alexandrova, A. (2018) Can the Science of Well-Being Be Objective?, The British Journal for the Philosophy of Science, 69(2): 421–445. https://doi.org/10.1093/ bjps/axw027

- Altman, A. 2015. Discrimination, The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu

/ entries/discrimination/

- Alkoby, S., & Sarne, D. (2017). The Benefit in Free Information Disclosure When Selling Information to People. Paper presented at the AAAI.

- Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. Journal of Experimental &Theoretical Artificial Intelligence, 12(3): 251–261

- American Honda Motor Co. (2017). ASIMO: The World's Most Advanced Humanoid Robot. Available online: http://asimo.honda.com

- Anderson, B., & Horvath, B. (2017). The rise of the weaponized ai propaganda machine. Scout, February, 12.

- Anderson, M., & Anderson, S. L. (2007). The status of machine ethics: a report from the AAAI Symposium. Minds and Machines, 17(1): 1–10.

- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. ProPublica, May, 23.

- Ovanessoff, A. and Plastino, E. (2017). How Can AI Drive South America's Growth? Accenture Research Report.

- Arney, C. (2016). Our Final Invention: Artificial Intelligence and the End of the Human Era. Mathematics and Computer Education, 50(3): 227.

- ASI Data Science and Slaughter & May. (2017). Superhuman Resources: Responsible Deployment of AI in Business.

- Australian Computing Society. (2017). Australia's Digital Pulse in 2017.

- Barocas, S. (2014). Data mining and the discourse on discrimination. Proceedings of the Data Ethics Workshop, Conference on Knowledge Discovery and Data Mining (KDD).https://dataethics.github.io/proceedings/DataMiningandtheDiscourseOnDiscrimination.pdf

- Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104: 671.

- Becker, B. (2006). Social robots-emotional agents: Some remarks on naturalizing man-machine interaction. International Review of Information Ethics 6: 37–45.

- Bei, X., Chen, N., Huzhang, G., Tao, B., & Wu, J. (2017). Cake cutting: envy and truth. Paper presented at the Proceedings of the 26th International Joint Conference on Artificial Intelligence.

- Bei, X., Qiao, Y., & Zhang, S. (2017). Networked fairness in cake cutting. arXiv preprint arXiv:1707.02033.

- Belle, V. (2017). Logic meets probability: towards explainable AI systems for uncertain worlds. Paper presented at the Proceedings of the TwentySixth International Joint Conference on Artificial Intelligence, IJCAI.

- Bess, M. (2010). Enhanced Humans versus "Normal People": Elusive Definitions, The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine, 35(6): 641–655. https://doi.org/10.1093/jmp/jhq053

- Binns, R. (2017). Fairness in Machine Learning: Lessons from Political Philosophy. arXiv preprint arXiv:1712.03586.

- Biran, O., & McKeown, K. R. (2017). Human-Centric Justification of Machine Learning Predictions. Paper presented at the IJCAI.

- Bloomberg News. (2018). China Now Has the Most Valuable AI Startup in the World.

- Boddington, P., Millican, P., & Wooldridge, M. (2017). Minds and Machines Special Issue: Ethics and Artificial Intelligence. Minds and Machines, 27(4): 569–574.

- Bogaerts, B., Vennekens, J., & Denecker, M. (2017). Safe inductions: An algebraic study. Paper presented at the Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI).

- Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. Science Fiction and Philosophy: From Time Travel to Superintelligence, 277–284.

- Bostrom, N. (2014). Superintelligence. Oxford University Press.

- Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Filar, B. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.

- Burch, K. T. (2012). Democratic transformations: Eight conflicts in the negotiation of American identity. A&C Black.

- Burns, T.W., O'Connor, D.J., & Stocklmayer, S.M. (2003) Science communication: a contemporary definition. Public Understanding of Science, 12: 183–202.

- Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), 1–12.

- Burton, E., Goldsmith, J., Koenig, S., Kuipers, B., Mattei, N., & Walsh, T. (2017). Ethical considerations in artificial intelligence courses. arXiv preprint arXiv:1701.07769.

- Bygrave, L. A. (2001). Automated profiling: minding the machine: article 15 of the ec data protection directive and automated profiling. Computer Law & Security Review, 17(1): 17–24.

- Calders, T., &Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery,

21(2): 277–292.

- Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Paper presented at the Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

- Cave, S. (2017) Intelligence: A History. Aeon. Cave, S., & Dihal, K. (2018) Ancient dreams of intelligent machines: 3,000 years of robots. Nature, 559: 473–475.

- Cech, E. A. (2014). Culture of disengagement in engineering education? Science, Technology, & Human Values, 39(1): 42–72.

- Chace, C. (2015). Surviving AI: The promise and peril of artificial intelligence. Bradford: Three Cs Publishing.

- Chakraborti, T., Sreedharan, S., Zhang, Y., & Kambhampati, S. (2017). Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. arXiv preprint arXiv:1701.08317.

- Chierichetti, F., Kumar, R., Lattanzi, S., &Vassilvitskii, S. (2017). Fair clustering through fairlets. Paper presented at the Advances in Neural Information Processing Systems.

- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2): 153–163.

- Clark, A. (1996). Being There: Putting Brain, Body, and World Together Again. Cambridge: MIT Press.

- Clark, A. (2008). Supersizing the Mind: Embodiment, Action, and Cognitive Extension. New York: Oxford University Press.

- Clark, A. and D. Chalmers. (1998). The Extended Mind. Analysis, 58: 7–19.

- Clifford, D., Graef, I., &Valcke, P. (2018). Pre-Formulated Declarations of Data Subject Consent–Citizen-Consumer Empowerment and the Alignment of Data, Consumer and Competition Law Protections.

- Coeckelbergh, M., Pop, C., Simut, R., Peca, A., Pintea, S., David, D. &Vanderborght, B. (2016). A Survey of Expectations About the Role of Robots in RobotAssisted Therapy for Children with ASD: Ethical Acceptability, Trust, Sociability, Appearance, and Attachment. Science and Engineering Ethics 22 (1): 47–65.

- Coggon, J., and J. Miola. (2011). Autonomy, Liberty, and Medical Decision-Making. The Cambridge Law Journal, 70(3): 523–547.

- Collins, S. and A. Ruina. (2005). A bipedal walking robot with efficient and human-like gait. Proceedings IEEE International Conference on Robotics and Automation, Barcelona, Spain.

- Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., & Kramer, M. (2017). Moral Decision Making Frameworks for Artificial Intelligence. Paper presented at the AAAI.

- Cowls, J., & Floridi, L. (2018). Prolegomena to a White Paper on an Ethical Framework for a Good AI Society. SSRN Electronic Journal.

- Crawford, K. (2016). Artificial intelligence's white guy problem. The New York Times, 25.

- Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. Nature, 538(7625): 311.

- Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Color. Stanford Law Review, 43(6): 1241–1299.

- Dafoe, A. (2018). AI Governance: A Research Agenda. University of Oxford.
- Daly, A. (2016). Private power, online information flows and EU law: Mind the gap. Bloomsbury Publishing.
- Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. Paper presented at the Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence.
- Davies, J. (2016). Program good ethics into artificial intelligence. Nature News.
- Dawkins, R. (1982). The Extended Phenotype. NewYork: Oxford Press.
- Devlin, H. (2017). AI programs exhibit racial and gender biases, research reveals. The Guardian, 13.
- Dietterich, T. G. (2017). Steps toward robust artificial intelligence. AI Magazine, 38(3): 3–24.
- Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue.
- Ding, J. (2018). Deciphering China's AI Dream. University of Oxford.
- Dunbar, M. (2017). To Be a Machine: Adventures Among Cyborgs, Utopians, Hackers, and the Futurists Solving the Modest Problem of Death. The Humanist, 77(3): 42.
- Dwork, C. (2008). Differential privacy: A survey of results. Paper presented at the International Conference on Theory and Applications of Models of Computation.
- Dwork, C. (2017). What's Fair? Paper presented at the Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. Paper presented at the Proceedings of the 3rd innovations in theoretical computer science conference.
- Edwards, L., &Veale, M. (2017). Slave to the Algorithm: Why a Right to an Explanation Is Probably Not the Remedy You Are Looking for. Duke Law and Technology Review 16(1): 18.
- Ess, C. (2006). Ethical pluralism and global information ethics. Ethics and Information Technology, 8(4): 215–226.
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., &Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639): 115.
- EU EDPS Ethics Advisory Group. (2018). Towards a digital ethics.
- Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.
- European Group on Ethics in Science and New Technologies. (2018). Statement on Artificial Intelligence, Robotics and "Autonomous" Systems.
- Fast, E., & Horvitz, E. (2017). Long-Term Trends in the Public Perception of Artificial Intelligence. Paper presented at the AAAI.
- Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., &Venkatasubramanian, S. (2015). Certifying and removing disparate impact. Paper presented at the Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- Fish, B., Kun, J., & Lelkes, Á. D. (2016). A confidencebased approach for balancing fairness and accuracy. Paper presented at the Proceedings of the 2016 SIAM International Conference on Data Mining.

- Fisher, D. H. (2017). A Selected Summary of AI for Computational Sustainability. Paper presented at the AAAI.
- Forster, E. M. (1947). Collected short stories of EM Forster. Sidgwick and Jackson.
- Frank, R. H. (2000). Why is cost-benefit analysis so controversial? The Journal of Legal Studies, 29(S2): 913–930.
- Freuder, E. C. (2017). Explaining Ourselves: Human-Aware Constraint Reasoning. Paper presented at the AAAI.
- Frey, C. B., & Osborne, M. A. (2017). The future of employment: how susceptible are jobs to computerisation? Technological forecasting and social change, 114: 254–280.
- Friedler, S. A., Scheidegger, C., &Venkatasubramanian, S. (2016). On the (im) possibility of fairness. arXiv preprint arXiv:1609.07236.
- Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on Information Systems (TOIS), 14(3): 330–347.
- Future Advocacy and The Wellcome Trust. (2018). Ethical, social and political challenges of artificial intelligence in health.
- Garrett, R. K., E. C. Nisbet, and E. K. Lynch. (2013). Undermining the corrective effects of media-based political fact checking? The role of contextual cues and naïve theory. Journal of Communication 63(4): 617–637.
- Garrett, R. K., E. C. Weeks, and R. L. Neo. (2016). Driving a wedge between evidence and beliefs: How online ideological news exposure promotes political misperceptions. Journal of Computer-Mediated Communication 21(5): 331–348.
- Gellert, R. (2015). Data protection: a risk regulation? Between the risk

management of everything and the precautionary alternative. International Data Privacy Law, 5(1): 3.

- Gellert, R. (2018). Understanding the notion of risk in the General Data Protection Regulation. Computer Law & Security Review, 34(2): 279–288.

- Goh, G., Cotter, A., Gupta, M., & Friedlander, M. P. (2016). Satisfying real-world goals with dataset constraints. Paper presented at the Advances in Neural Information Processing Systems.

- Goldsmith, J., & Burton, E. (2017). Why Teaching Ethics to AI Practitioners Is Important. Paper presented at the AAAI.

- Goodman, B., & Flaxman, S. (2016). European Union regulations on algorithmic decision-making and a "right to explanation". arXiv preprint arXiv:1606.08813.

- Government Office for Science. (2016). Artificial intelligence: opportunities and implications for the future of decision making.

- Government Office for Science. (2017). The Futures Toolkit: Tools for Futures Thinking and Foresight Across UK Government.

- GPI Atlantic. (1999). Gender Equality in the Genuine Progress Index. Made to Measure Symposium Synthesis Paper, Halifax, October 3–6.

- Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When Will AI Exceed Human Performance? Evidence from AI Experts. Journal of Artificial Intelligence Research, 62: 729–754.

- Graef, I. (2016). EU Competition Law, Data Protection and Online Platforms: Data as Essential Facility: Kluwer Law International.

- Grafman, J. and I. Litvan. (1999). Evidence for Four Forms of Neuroplasticity. In Neuronal Plasticity: Building a Bridge from the Laboratory to the

Clinic. J. Grafman and Y. Christen (eds.). Springer-Verlag Publishers.

- Grbovic, M., Radosavljevic, V., Djuric, N., Bhamidipati, N., & Nagarajan, A. (2015). Gender and interest targeting for sponsored post advertising at tumblr. Paper presented at the Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

- Grgić-Hlača, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018). Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. arXiv preprint arXiv:1802.09548.

- Greenwald, A. G. (2017). An AI stereotype catcher. Science, 356(6334): 133–134.

- Gribbin, J. (2013). Computing with quantum cats: From Colossus to Qubits. Random House.

- Gunkel, D. J., & Bryson, J. (2014). Introduction to the special issue on machine morality: The machine as moral agent and patient. Philosophy &Technology, 27(1): 5–8.

- Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). The off-switch game. arXiv preprint arXiv:1611.08219.

- Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. Paper presented at the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.

- Hajian, S., & Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. IEEE transactions on knowledge and data engineering, 25(7): 1445–1459. Hajian, S., Domingo-Ferrer, J., Monreale, A., Pedreschi, D., & Giannotti, F. (2015).

Discrimination-and privacyaware patterns. Data Mining and Knowledge Discovery, 29(6): 1733–1782.

- Hanisch, C. (1969). The personal is political. Available at www.carolhanisch.org/CHwritings/PIP.html

- Hanisch, C. (2006). The personal is political: The women's liberation movement classic with a new explanatory introduction. Women of the World, Unite.

- Harari, Y. N. (2016). Homo Deus: A brief history of tomorrow. Random House.

- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Paper presented at the Advances in neural information processing systems.

- Harel, Y., Gal, I. B., & Elovici, Y. (2017). Cyber Security and the Role of Intelligent Systems in Addressing its Challenges. ACM Transactions on Intelligent Systems and Technology (TIST), 8(4): 49.

- Haybron, D. M., & Alexandrova, A. (2013). Paternalism in economics. Paternalism: Theory and practice, (eds Christian Coons and Michael Weber), Cambridge University Press, 157–177.

- Helberger, N., Zuiderveen Borgesius, F. J., & Reyna, A. (2017). The perfect match? A closer look at the relationship between EU consumer law and data protection law.

- Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., Zwitter, A. (2017). Will democracy survive big data and artificial intelligence? Scientific American, 25.

- Hilton, M. Differential privacy: a historical survey. Cal Poly State University.

- House of Commons Science and Technology Committee, The Big Data Dilemma. 12 February 2016, HC 468 2015–16.

- Hurley, S. L. (1998). Vehicles, Contents, Conceptual Structure and Externalism. Analysis 58: 1–6.

- IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2018) Ethically aligned design: a vision for prioritizing human wellbeing with artificial intelligence and autonomous systems.

- Imberman, S. P., McManus, J., & Otts, G. (2017). Creating Serious Robots That Improve Society. Paper presented at the AAAI.

- Institute of Technology and Society in Rio. (2017). Big Data in the Global South: Report on the Brazilian Case Studies.

- Ipsos MORI and the Royal Society. (2017). Public views of Machine Learning.

- Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., & Roth, A. (2016). Fairness in reinforcement learning. arXiv preprint arXiv:1611.03071.

- Johndrow, J. E., & Lum, K. (2017). An algorithm for removing sensitive information: application to raceindependent recidivism prediction. arXiv preprint arXiv:1703.04957.

- Jotterand, F. and V. Dubljevic (Eds). (2016). Cognitive Enhancement: Ethical and Policy Implications in International Perspectives. Oxford University Press.

- Kamarinou, D., Millard, C., & Singh, J. (2016). Machine Learning with Personal Data. Queen Mary School of Law Legal Studies Research Paper, 247.

- Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems,

33(1), 1–33.

- Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. Paper presented at the Data Mining (ICDM), 2010 IEEE 10th International Conference on Computer and Information Technology.

- Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. Paper presented at the Data Mining (ICDM), 2012 IEEE 12th International Conference on Dating Mining.

- Kamiran, F., Žliobaitė, I., & Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowledge and Information systems, 35(3): 613–644.

- Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. Paper presented at the Joint European Conference on Machine Learning and Knowledge Discovery in Databases.

- Kaplan, J. (2015). Humans need not apply: A guide to wealth and work in the age of artificial intelligence. Yale University Press.

- Kaplan, J. (2016). Artificial Intelligence: What everyone needs to know. Oxford University Press.

- Kearns, M., Roth, A., & Wu, Z. S. (2017). Meritocratic fairness for cross-population selection. Paper presented at the International Conference on Machine Learning.

- Kleinberg, J., Ludwig, J., Mullainathan, S. (2016). A Guide to Solving Social Problems with Machine Learning. Harvard Business Review.

- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807.

- Koops, B. (2013). On decision transparency, or how to enhance data protection after the computational turn. Privacy, due process and the computational turn: the philosophy of law meets the philosophy of technology, 189–213.

- Kraemer, F., Van Overveld, K., & Peterson, M. (2011). Is there an ethics of algorithms? Ethics and Information Technology, 13(3): 251–260.

- Kristoffersson, A., Coradeschi, S., Loutfi, A., & Severinson-Eklundh, K. (2014). Assessment of interaction quality in mobile robotic telepresence: An elderly perspective. Interaction Studies 15(2): 343–357.

- Kuner, C., Svantesson, D. J. B., Cate, F. H., Lynskey, O., & Millard, C. (2017). Machine learning with personal data: is data protection law smart enough to meet the challenge? International Data Privacy Law, 7(1), 1–2.

- Kurzweil, R. (2013). How to create a mind: The secret of human thought revealed. Penguin.

- Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. Paper presented at the Advances in Neural Information Processing Systems.

- Kuzelka, O., Davis, J., & Schockaert, S. (2017). Induction of interpretable possibilistic logic theories from relational data. arXiv preprint arXiv:1705.07095.

- Kökciyan, N., &Yolum, P. (2017). Context-Based Reasoning on Privacy in Internet of Things. Paper presented at the IJCAI.

- Langley, P., Meadows, B., Sridharan, M., & Choi, D. (2017). Explainable Agency for Intelligent Autonomous Systems. Paper presented at the AAAI.

- Lehmann, H., Iacono, I., Dautenhahn, K., Marti, P. and Robins, B. (2014).

Robot companions for children with down syndrome: A case study. Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems 15(1), pp. 99–112.

- Levy, N. Rethinking Neuroethics in the Light of the Extended Mind Thesis, The American Journal of Bioethics, 7(9): 3–11.

- Lewis-Kraus, G. (2016). The great AI awakening. The New York Times Magazine, 14.

- Li, Fei-Fei. (2018). How to Make A.I. That's Good for People. The New York Times.

- Lipton, Z. C. (2016). The Mythos of Model Interpretability. ICML 2016 Workshop on Human Interpretability in Machine Learning.

- Luong, B. T., Ruggieri, S., &Turini, F. (2011). k-NN as an implementation of situation testing for discrimination discovery and prevention. Paper presented at the Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining.

- Lyons, J. B., Clark, M. A., Wagner, A. R., & Schuelke, M. J. (2017). Certifiable Trust in Autonomous Systems: Making the Intractable Tangible. AI Magazine, 38(3).

- Marcus, G. (2012). Will a Robot Take Your Job? The New Yorker.

- Marcus, G. (2013). Why we should think about the threat of artificial intelligence. The New Yorker.

- Marien, M. (2014). The second machine age: Work, progress, and prosperity in a time of brilliant technologies. Cadmus, 2(2): 174.

- Mattu, S. and Hill, K. (2018) The House That Spied on Me. Gizmodo.

- McAllister, R., Gal, Y., Kendall, A., Van Der Wilk, M., Shah, A., Cipolla, R.,

& Weller, A. V. (2017). Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning.

- McFarland, D. (2009). Guilty robots, happy dogs: the question of alien minds. Oxford University Press.

- Mei, J.-P., Yu, H., Shen, Z., & Miao, C. (2017). A social influence based trust model for recommender systems. Intelligent Data Analysis, 21(2): 263–277.

- Menary, R. (2007). Cognitive Integration: Mind and Cognition Unbounded. Palgrave Macmillan.

- Mendoza, I., & Bygrave, L. A. (2017). The Right not to be Subject to Automated Decisions based on Profiling. In EU Internet Law: 77–98.

- Milli, S., Hadfield-Menell, D., Dragan, A., & Russell, S. (2017). Should robots be obedient? arXiv preprint arXiv:1705.09990.

- Mindell, D. (2002). Between human and machine: feedback, control, and computing before cybernetics. Baltimore: Johns Hopkins University Press.

- Minsky, M. (1982). Semantic information processing: MIT Press.

- Minton, S. N. (2017). The Value of AI Tools: Some Lessons Learned. AI Magazine, 38(3).

- Monbiot, G. (2017). Big data's power is terrifying. That could be good news for democracy. The Guardian.

- Montréal Declaration on Responsible AI. (2018). Montréal Declaration for a Responsible Development of Artificial Intelligence. Available at www.montrealdeclaration-responsibleai.com/ the-declaration

- Moore, A. (2017). Critical elitism: Deliberation, democracy, and the problem of expertise. Cambridge University Press.

- Moravec, H. (1988). Mind children: The future of robot and human intelligence. Harvard University Press.
- Mukherjee, S. (2017). A.I. Versus M.D.: What happens when a diagnosis is automated? The New Yorker.
- Müller, V. C. (2014). Risks of artificial general intelligence. Journal of Experimental and Theoretical Artificial Intelligence, 26(3): 297-301.
- Noble, S. U. (2018). Algorithms of Oppression: How search engines reinforce racism. NYU Press.
- Noë, A. (2009). Out of our heads. Hill and Wang.
- Novitske, L (2018). The AI Invasion is Coming to Africa and It's a Good Thing. Stanford Social Innovation Review.
- Nushi, B., Kamar, E., Horvitz, E., & Kossmann, D. (2017). On Human Intellect and Machine Failures: Troubleshooting Integrative Machine Learning Systems. Paper presented at the AAAI.
- Omidyar Network and Upturn. (2018). Public scrutiny of automated decisions: early lessons and emerging methods. Available online at www.omidyar.com/ insights/public-scrutiny-automated-decisions-earlylessons-and-emerging-methods
- O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.
- Open Data Institute. (2017). Helping organisations navigate concerns in their data practices. Available online at https://theodi.org/article/data-ethics-canvas/
- Pagallo, U. (2017). From automation to autonomous systems: a legal phenomenology with problems of accountability. Paper presented at the

Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17).

- Parens, E. (1998). Enhancing human traits: Ethical and social implications (Hastings Center studies in ethics). Washington, D.C.: Georgetown University Press.

- Parens, E. (2015). Shaping ourselves: On technology, flourishing, and a habit of thinking. Oxford University Press.

- Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.

- Pedreschi, D., Ruggieri, S., &Turini, F. (2009). Measuring discrimination in socially-sensitive decision records. Paper presented at the Proceedings of the 2009 SIAM International Conference on Data Mining.

- Phan, N., Wu, X., Hu, H., & Dou, D. (2017). Adaptive laplace mechanism: differential privacy preservation in deep learning. Paper presented at the 2017 IEEE International Conference on Data Mining (ICDM).

- Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. Paper presented at the Advances in Neural Information Processing Systems.

- Powles, J., & Hodson, H. (2017). Google DeepMind and healthcare in an age of algorithms. Health and technology, 7(4): 351–367.

- Prainsack, B., & Buyx, A. (2017). Solidarity in biomedicine and beyond (Vol. 33). Cambridge University Press.

- National Science and Technology Council, Obama White House. (2016). Preparing for the Future of Artificial Intelligence.

- Purtova, N. (2018). The law of everything. Broad concept of personal

data and future of EU data protection law. Law, Innovation and Technology, 10(1): 40–81.

- Quadrianto, N., & Sharmanska, V. (2017). Recycling privileged learning and distribution matching for fairness. Paper presented at the Advances in Neural Information Processing Systems.

- Reed, C., Kennedy, E., & Silva, S. (2016). Responsibility, Autonomy and Accountability: legal liability for machine learning. Queen Mary School of Law Legal Studies Research Paper No. 243/2016. Available at SSRN: https://ssrn.com/abstract=2853462

- Resnick, B. (2018). Cambridge Analytica's "psychographic microtargeting": what's bullshit and what's legit. Vox.

- Richert, A., Müller, S., Schröder, S., and Jeschke, S. (2018). Anthropomorphism in social robotics: empirical results on human–robot interaction in hybrid production workplaces. AI and Society 33(3): 413–424.

- Robins, B., Dautenhahn, K., and Dubowski, J., (2006). Does appearance matter in the interaction of children with autism with a humanoid robot? Interaction

- Studies. Social Behaviour and Communication in Biological and Artificial Systems 7(3): 509–542.

- Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review, 29(5): 582–638.

- Ross, A. S., Hughes, M. C., & Doshi-Velez, F. (2017). Right for the right reasons: Training differentiable models by constraining their explanations. arXiv preprint arXiv:1703.03717.

- Royal Society. (2017). Machine learning: the power and promise of

computers that learn by example.

- Royal Society and The British Academy. (2017). Data management and use: Governance in the 21st century.

- Royal Society for the encouragement of Arts, Manufactures and Commerce (RSA). (2018). Artificial Intelligence: Real Public Engagement.

- Russell, C., Kusner, M. J., Loftus, J., & Silva, R. (2017). When worlds collide: integrating different counterfactual assumptions in fairness. Paper presented at the Advances in Neural Information Processing Systems.

- Schermer, M. (2013). Health, Happiness and Human Enhancement – Dealing with Unexpected Effects of Deep Brain Stimulation. Neuroethics, 6(3): 435–445.

- Scheutz, M. (2017). The case for explicit ethical agents. AI Magazine, 38(4): 57–64.

- Searle, J. R. (1980). Minds, Brains, and Programs. Behavioral and Brain Sciences, 3: 417–457.

- Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. Fordham Law Review.

- Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. International Data Privacy Law, 7(4): 233–242.

- Select Committee on Artificial Intelligence. (2018). AI in the UK: Ready, Willing, and Able? HL 100 2017–19. London: House of Lords.

- Shapiro, L. A. (2004). The Mind Incarnate. MIT Press.

- Sharkey, A. (2014). Robots and human dignity: a consideration of the effects of robot care on the dignity of older people. Ethics and Information Technology 16(1): 63–75.

- Sharkey, A., and Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. Ethics and Information Technology 14(1): 27–40.

- Shirk, J. L., H. L. Ballard, C. C. Wilderman, T. Phillips, A. Wiggins, R. Jordan, E. McCallie, M. Minarchek, B. V. Lewenstein, M. E. Krasny, and R. Bonney. (2012). Public participation in scientific research: a framework for deliberate design. Ecology and Society 17(2): 29. http://dx.doi.org/10.5751/ES-04705–170229

- Shariff, A., Rahwan, I., and Bonnefon, J. (2016). Whose Life Should Your Car Save? New York Times.

- Shell International BV. (2008). Scenarios: An Explorer's Guide.

- Shirk, J. L., Ballard, H. L., Wilderman, C. C., Phillips, T., Wiggins, A., Jordan, R., Krasny, M. E. (2012). Public participation in scientific research: a framework for deliberate design. Ecology and Society, 17(2).

- Simon, H. (1969). The Sciences of the Artificial. MIT Press.

- Sintov, N., Kar, D., Nguyen, T., Fang, F., Hoffman, K., Lyet, A., &Tambe, M. (2017). Keeping It Real: Using RealWorld Problems to Teach AI to Diverse Audiences. AI Magazine, 38(2).

- Such, J. M. (2017). Privacy and autonomous systems. Paper presented at the Proceedings of the 26th International Joint Conference on Artificial Intelligence.

- Sweeney, L. (2013). Discrimination in online ad delivery. Queue, 11(3): 10.

- Tapus, A., Peca, A., Aly, A., Pop, C., Jisa, L., Pintea, S., Rusu, A. S. and David, D. O. (2012). Children with autism social engagement in interaction with

Nao, an imitative robot: A series of single case experiments. Interaction Studies 13(3): 315–347.

- Tegmark, M. (2017). Life 3.0: Being human in the age of artificial intelligence. Knopf.

- Tene, O., & Polonetsky, J. (2017). Taming the Golem: Challenges of Ethical Algorithmic Decision-Making. NC Journal of Law and Technology, 19(1): 125.

- Thelisson, E. (2017). Towards trust, transparency, and liability in AI/AS systems. Paper presented at the Proceedings of the 26th International Joint Conference on Artificial Intelligence.

- Tiberius, V. (2018). Well-Being As Value Fulfillment: How We Can Help Each Other to Live Well. Oxford University Press.

- Tolomei, G., Silvestri, F., Haines, A., & Lalmas, M. (2017). Interpretable predictions of tree-based ensembles via actionable feature tweaking. Paper presented at the Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

- Turkle, S. (2016). Reclaiming conversation: The power of talk in a digital age. Penguin.

- Turkle, S. (2017). Alone together: Why we expect more from technology and less from each other. Hachette UK.

- Vanderborght, B., Simut, R, Saldien, J., Pop, C., Rusu, A. S., Pintea, S., Lefeber, D. and David, D.O. (2012). Using the social robot probo as a social story telling agent for children with ASD. Interaction Studies 13(3): 348–372.

- Varakin, D.A., Levin, D. T. and Fidler, R. (2004). Unseen and unaware: Implications of recent research on failures of visual awareness for human

-computer interface design. Human-Computer Interaction 19(4): 389–422.

- Vempati, S. S. (2016). India and the Artificial Intelligence Revolution. Carnegie Endowment for International Peace.

- Vold, K. (2015). The Parity Argument for Extended Consciousness. Journal of Consciousness Studies, 22(3–4): 16–33.

- Wachter, S. and Mittelstadt, B.D. (2018) A right to reasonable inferences: re-thinking data protection in the age of big data and AI, Columbia Business Law Review.

- Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decisionmaking does not exist in the general data protection regulation. International Data Privacy Law, 7(2): 76–99.

- Wachter-Boettcher, S. (2017). Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech: WW Norton & Company.

- Walden, J., Jung, E., Sundar, S., and Johnson, A. (2015). Mental models of robots among senior citizens: An interview study of interaction expectations and design implications. Interaction Studies 16(1): 68–88.

- Wallach, W., & Allen, C. (2008). Moral machines: Teaching robots right from wrong. Oxford University Press.

- Walsh, T. (2016). The singularity may never be near. arXiv preprint arXiv:1602.06462.

- Walsh, T. (2017). Android Dreams: The Past, Present and Future of Artificial Intelligence. Oxford University Press.

- Weller, A. (2017). Challenges for transparency. arXiv preprint arXiv:1708.01870.

- Weiskopf, D. (2008). Patrolling the mind's boundaries. Erkenntnis, 68(2): 265–76.
- Whitfield, C. (2018). The Ethics of Artificial Intelligence. PwC Australia.
- Whittlestone, J., Nyrup, R., Alexandrova, A., and Cave, S. (2019). The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. Forthcoming in Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics and Society.
- Wheeler, M. (2010). Minds, Things, and Materiality. In L. Malafouris and C. Renfrew. (eds.), The Cognitive Life of Things: Recasting the Boundaries of the Mind. Cambridge: McDonald Institute Monographs. (Reprinted in J. Schulkin (ed.), Action, Perception and the Brain: Adaptation and Cephalic Expression. Basingstoke: Palgrave Macmillan.)
- Wilson, R. A. (1994). Wide Computationalism. Mind, 103(411): 351–72.
- Wilson, R. A. and A. Clark. (2009). How to situate cognition: Letting nature take its course. In Murat Aydede and P. Robbins (eds.), The Cambridge Handbook of Situated Cognition. Cambridge: Cambridge University Press, 55–77.
- The Wilson Centre. (2017). Artificial Intelligence: A Policy-Oriented Introduction.
- Wood, L., Lehmann, H., Dautenhahn, K., Robins, B., Rainer, A., and Syrdal, D. (2016). Robot-mediated interviews with children. Interaction Studies 17(3): 438–460.
- World Wide Web Foundation. (2017). Artificial Intelligence: Starting the Policy Dialogue in Africa.
- Yao, S., & Huang, B. (2017). Beyond parity: Fairness objectives for

collaborative filtering. Paper presented at the Advances in Neural Information Processing Systems.

- Yuste, R. et al. (2017). Four Ethical Priorities for Neurotechnologies and AI. Nature News, Nature Publishing Group. www.nature.com/news/four-ethicalpriorities-for-neurotechnologies-and-ai-1.22960

- Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K., & Weller, A. (2017). From parity to preference-based notions of fairness in classification. Paper presented at Advances in Neural Information Processing Systems.

- Zarkadakis, G. (2015). In Our Own Image: Will artificial intelligence save or destroy us? Random House.

- Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. Paper presented at the International Conference on Machine Learning.

- Žliobaite, I. (2015). A survey on measuring indirect discrimination in machine learning. arXiv preprint arXiv:1511.00148.

- Žliobaite, I., Kamiran, F., & Calders, T. (2011). Handling conditional discrimination. Paper presented at the Data Mining (ICDM), 2011 IEEE 11th International Conference on data mining.

csri.majazi.ir

حوزه فضای مجازی به اندازه انقلاب اسلامی اهمیت دارد. این فضا مثل یک رودخانه پر از آب و خروشان است که می آید و دائماً هم بر آب آن افزوده و خروشان تر می شود. اگر ما بر این رودخانه تدبیر کنیم و برنامه داشته باشیم، زه‌کشی کنیم و هدایت کنیم این رودخانه را تا به سد بریزد، می‌شود فرصت. اگر رهایش کنیم و برنامه‌ای برای آن نداشته باشیم می‌شود یک تهدید.

csri.majazi.ir