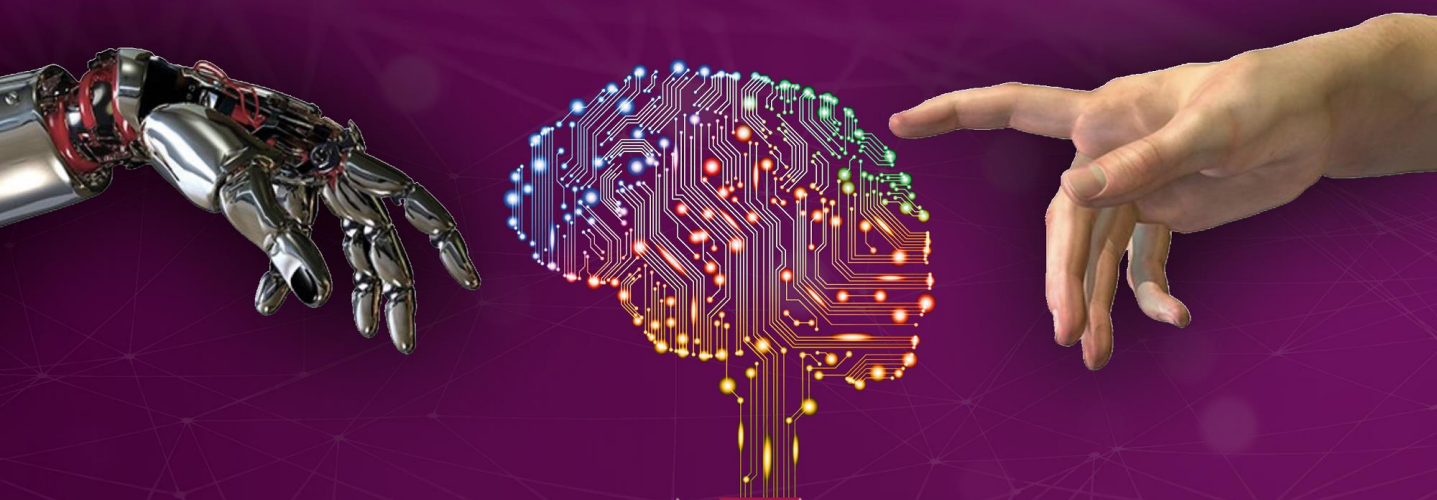
جمهوری اسلامی ایران
شورای عالی فضای مجازی
مرکز ملی فضای مجازی

فضای مجازی

اسناد فضای مجازی ۲

چارچوبی اخلاقی برای یک جامعه AI خوب

فرصت‌ها، ریسک‌ها، اصول و توصیه‌ها



AI 4PEOPLE

www.majazi.ir

بسم الله الرحمن الرحیم

مرکز ملی فضای مجازی
پژوهشگاه فضای مجازی

چارچوبی اخلاقی برای یک جامعه AI خوب: فرصت‌ها، ریسک‌ها، اصول و توصیه‌ها

اسناد فضای مجازی (۲)
بهمن ماه ۱۳۹۸

تهیه شده در: پژوهشگاه مرکز ملی فضای مجازی- گروه مطالعات فرهنگی و اجتماعی
مترجم: یحیی شعبانی (دانشجوی دکترای پژوهشگاه علوم انسانی و مطالعات فرهنگی)
ناظر علمی: امیررضا باقرپورشیرازی

نشانی: تهران، میدان آرژانتین، خیابان بیهقی، نبش خیابان ۱۶ غربی، پلاک ۲۰، کدپستی ۱۵۱۵۶۷۴۳۱۱
شماره تماس: ۸۶۱۲۱۰۶۱ http://www.majazi.ir

سلسله گزارش‌های اسناد فضای مجازی، حاوی ترجمه مجموعه متون و اسناد سیاستی یا سیاست‌گذارانه‌ای است که در سطح ملی و جهانی از سوی کشورها، مؤسسات، نهادها و سازمان‌های بین‌المللی انتشار یافته‌اند. هدف اصلی از ترجمه این متون و اسناد، صرفاً جهت آشنایی برنامه‌ریزان و سیاست‌گذاران فضای مجازی کشور بوده و به‌هیچ‌عنوان دربرگیرنده دیدگاه پژوهشگاه و مرکز ملی فضای مجازی نخواهد بود.

چکیده

این یادداشت یافته‌های AI4People، ابتکار عمل یک‌ساله‌ای که با هدف ایجاد بنیادهای یک «جامعه AI خوب» طراحی شده است، را گزارش می‌دهد. ما فرصت‌ها و ریسک‌های مرکزی AI برای جامعه را معرفی می‌کنیم؛ ترکیبی از پنج اصل اخلاقی عرضه می‌کنیم که موجب تقویت توسعه و پذیرش آن خواهد شد؛ و ۲۰ توصیه انضمامی- برای ارزیابی، توسعه، تشویق و حمایت از AI خوب- پیشنهاد می‌کنیم که در برخی موارد سیاست‌گذاران ملی یا فراملی می‌توانند مستقیماً آن‌ها را مُتقبل شوند در عین حال که دیگر ذی‌ربطان می‌توانند دیگر موارد را بر عهده بگیرند. اگر این توصیه‌ها پذیرفته شوند، برای استقرار یک جامعه AI خوب در حکم بنیادی استوار خواهند بود.

واژگان کلیدی

هوش مصنوعی، AI4People، حکمرانی داده، اخلاق دیجیتال، حکمرانی، اخلاق AI.

فهرست مطالب

مقدمه

AI ابزار سودمند دیگری نیست که پس از بالغ‌شدن نیاز به تنظیم‌گری نداشته باشد. AI نیرویی قدرتمند، صورت جدیدی از عامل هوشمند است که پیشاپیش در حال تغییر دادن زندگی‌ها، تعاملات و محیط‌های ما است.

AI4People برپا شده است تا به هدایت این نیروی قدرتمند به سمت خیر جامعه، هرکسی که در آن است، و محیط‌هایی که در آن‌ها سهیم هستیم کمک کند. این اوراق سفید[۱] نتیجه تلاش مشترک کمیته علمی AI4People - شامل ۱۲ متخصص و به ریاست لوچیانو فلوریدی- است تا مجموعه‌ای از توصیه‌ها برای توسعه یک جامعه AI خوب را پیشنهاد دهد.

این اوراق سفید سه چیز را ترکیب می‌کند: *فرصت‌ها* و *ریسک‌های* مرتبطی که تکنولوژی‌های AI برای پرورش کرامت انسان و ارتقاء شکوفایی انسان عرضه می‌کنند؛ *اصولی* که پذیرش AI را تقویت خواهند کرد؛ و دوازده *توصیه* خاص که، اگر پذیرفته شوند، همه ذی‌ربطان را قادر خواهند ساخت تا فرصت‌ها را دریابند و از ریسک‌ها پرهیز کنند یا لااقل آن‌ها را کمینه و متعادل سازند، به اصول احترام بگذارند و بدین ترتیب یک جامعه AI خوب را توسعه دهند.

این اوراق سفید علاوه بر این مقدمه از چهار بخش تشکیل شده است. بخش ۲ فرصت‌های مرکزی برای ارتقاء کرامت و شکوفایی انسانی را که AI عرضه نموده است، همراه با ریسک‌های متناظرشان، بیان می‌کند. بخش ۳ نگاهی مجمل و سطح بالا به مزایای اتخاذ رویکردی اخلاقی به توسعه و کاربرد AI توسط سازمان‌ها عرضه می‌کند. بخش ۴ پنج اصل اخلاقی برای AI، مبتنی بر تحلیل‌های موجود، را صورت‌بندی می‌نماید که پذیرش اخلاقی AI در جامعه به طور کلی را

---

[۱] اوراق سفید یا White Paper در واقع نوعی گزارشی معتبر است برای مستندسازی، کمک به فهم یک مسئله با اتخاذ یک تصمیم. عمدتاً در دو حوزه اقتصاد و دولت از اوراق سفید استفاده می‌شود. وینستون چرچیل نخستین کسی بود که در سال ۱۹۲۲ اوراق سفید دولتی را مطرح کرد. (مترجم)

تقویت می‌کنند. نهایتاً، بخش ۵ بیست توصیه با هدف توسعه یک جامعه AI خوب در اروپا ارائه می‌کند.

پس از آغاز به کار AI4People در فوریه ۲۰۱۸، کمیته علمی مشترکاً اقدام نموده است تا در بخش نهایی این یادداشت توصیه‌هایی را بسط دهد. ما به واسطه این کار امیدواریم در پی‌ریزی یک جامعه AI خوب که همگی می‌توانیم در آن مشارکت نماییم سهمی داشته باشیم.

فرصت‌ها و ریسک‌های AI برای جامعه

اینکه AI تأثیری عمده بر روی جامعه خواهد گذاشت دیگر مورد تردید نیست. مناقشه کنونی در عوض بر این نکته متمرکز است که چنین تأثیری تا چه اندازه مثبت یا منفی است، برای چه کسی، به چه طرقی، در چه جاهایی و در چه مقیاس زمانی. به عبارت دیگر، ما می‌توانیم با اطمینان از این پرسش صرف‌نظر کنیم که *آیا* AI تأثیری خواهد داشت یا خیر؛ اکنون پرسش‌های مقتضی از این قرارند: این تأثیر مثبت یا منفی *توسط چه کسانی*، *چگونه*، *کجا* و *کی* احساس خواهد شد.

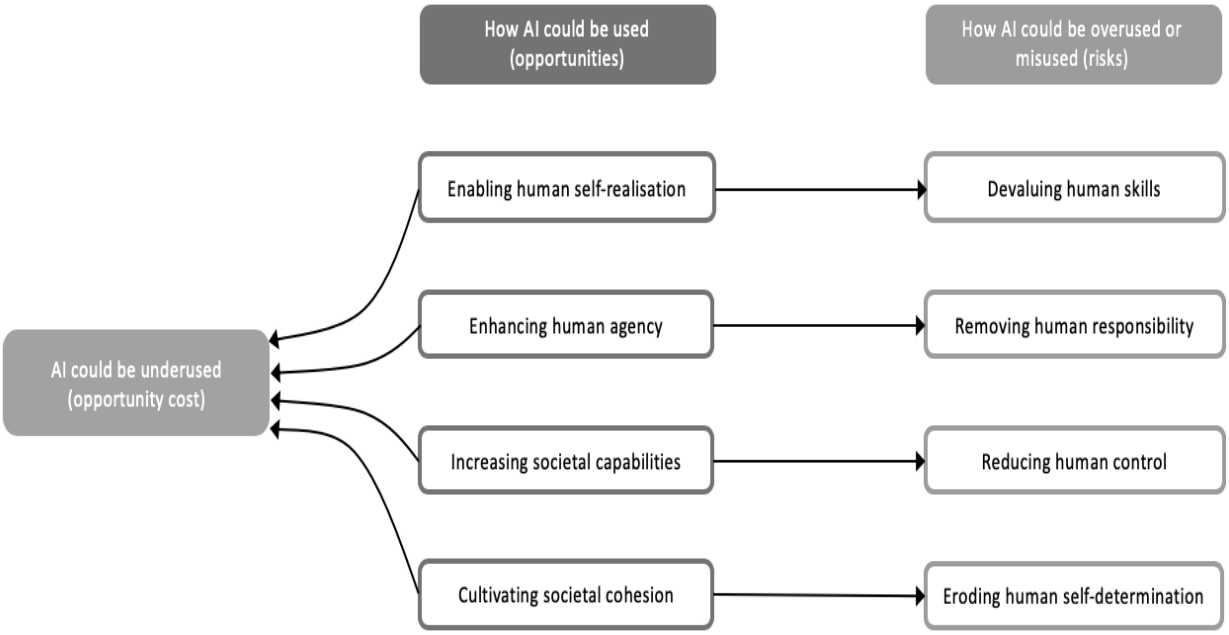
برای پرداختن به این پرسش‌ها به طریقی ماهوی‌تر و عملی‌تر، ما در اینجا چیزی را مطرح می‌کنیم که به نظر ما چهار فرصت اصلی است که AI به جامعه تقدیم می‌کند. تعداد این اصول از آن جهت چهار است که به چهار نقطه بنیادی در فهم کرامت و شکوفایی انسانی اشاره دارند: *به چه کسانی می‌توانیم تبدیل بشویم* (خودـشکوفاییِ خودمختار)؛ *چه می‌توانیم بکنیم* (عاملیت انسانی)؛ *چه می‌توانیم به دست آوریم* (ظرفیت‌های فردی و اجتماعی)؛ و *چگونه می‌توانیم با دیگران و جهان تعامل کنیم* (پیوند اجتماعی). در هر مورد، می‌توان از AI برای پرورش سرشت انسانی و قابلیت‌هایش *استفاده کرد* و بدین‌ترتیب فرصت‌هایی را ایجاد نمود؛ یا می‌توان از آن *استفاده ناکافی کرد* و بدین‌ترتیب هزینه فرصت‌هایی [غرامت فرصت‌هایی][1] ایجاد نمود؛ یا می‌توان

[1] اگر یک فرد یا یک بنگاه، از میان چندین انتخاب متفاوت یکی را برگزینند، هزینه فرصت این فرد یا بنگاه، معادل است با هزینه مرتبط با بهترین انتخاب ممکن از بین سایر انتخاب‌های باقی‌مانده که از آن صرف‌نظر شده

از آن *استفاده بیش از حد* و *سوءاستفاده کرد* و بدین‌ترتیب ریسک‌هایی ایجاد نمود. همان‌طور که این ترمینولوژی اشاره دارد، فرض بر این است که *استفاده* از AI مترادف است با نوآوریِ خوب و کاربردهای مثبتِ این تکنولوژی. اما ترس، جهل، نگرانی‌های بی‌مورد یا واکنش افراطی ممکن است، به علت آنچه که می‌توان اجمالاً دلایل غلط توصیف کرد، جامعه‌ای را به استفاده ناکافی از تکنولوژی‌های AI کمتر از پتانسیل کامل آن‌ها سوق دهد. این امر ممکن است موجب هزینه فرصت‌های قابل توجهی شود. این ممکن است، به عنوان مثال، شامل مقررات زُمخت یا نادرست، سرمایه‌گذاری ناکافی، یا واکنشی عمومی شبیه مورد محصولات اصلاح‌شده ژنتیکی شود (Imperial College, 2017). در نتیجه، منافعی که تکنولوژی AI پیشکش می‌کند نمی‌تواند کاملاً توسط جامعه محقق شود. این خطرات عمدتاً از عواقب ناخواسته ناشی می‌شود و نوعاً به نیات خوبی مرتبط است که بد از آب درآمده‌اند. اما ما همچنین باید ریسک‌های مرتبط با *استفاده بیش از حد* و ناخواسته یا *سوء استفاده* عامدانه از تکنولوژی‌های AI را نیز در نظر بگیریم که به عنوان مثال ریشه در انگیزه‌های ناهم‌تراز، طمع، ژئوپولیتیک‌های خصومت‌آمیز یا سوءنیت دارند. استفاده سوء از تکنولوژی‌های AI ممکن است به هر چیزی از کلاهبرداری‌های ایمیلی تا جنگ تمام‌عیار سایبری شتاب دهد یا آن‌ها را تشدید نماید (Taddeo, 2017). و شرارت‌های جدیدی می‌تواند ممکن شود (King et. al, 2018). امکانِ پیشرفت اجتماعی که با فرصت‌های فوق‌الذکر ترسیم شده است باید با این ریسک سنجیده شود که AI دخل و تصرف مغرضانه را افزایش می‌دهد یا تشدید می‌کند. با وجود این، یک ریسک تمام و کمال آن است که به واسطه ترس از

است. هدف اصلی از طرح هزینه فرصت در اقتصاد بررسی تخصیص بهینه منابع موجود با هدف تضمین این امر است که منابع کمیاب به صورت کارا مورد استفاده قرار گیرند. (مترجم)

استفاده مُفرط یا سوءاستفاده ممکن است از AI کم استفاده شود. ما این ریسک‌ها را در شکل الف خلاصه کرده‌ایم و در ادامه توضیح مفصل‌تری ارائه می‌کنیم.



شکل الف: نمای کلی از چهار فرصت عمده‌ای که AI عرضه کرده است، چهار ریسک متناظر و هزینه فرصت کم استفاده کردن از AI.

به چه کسانی می‌توانیم تبدیل بشویم: امکان خود شکوفاییِ انسان، بدون کاستن از توانایی‌های انسان

AI می‌تواند امکان خودـشکوفایی را فراهم آورد که مراد ما از آن توانایی انسان‌ها برای بالیدن بر حسب ویژگی‌ها، علایق، مهارت‌ها یا توانایی‌های بالقوه، اشتیاق‌ها و طرح‌های زندگی‌شان است. هر چند ابداعاتی مانند ماشین لباسشویی مردم- به ویژه زنان- را از مشقت کار خانگی آزاد کرده است، اتوماسیون هوشمندِ دیگر جنبه‌های دنیویِ حیات می‌تواند زمان بیشتری را برای فعالیت‌های فرهنگی، عقلی و اجتماعی و کارهای جالب‌تر و ارضاکننده‌تر آزاد کند. AI بیشتر بی‌گمان به معنای آن است که بخش بیشتری از زندگی بشر هوشمندانه‌تر بگذرد. ریسک این مورد به خودی

خود نه منسوخ شدن برخی مهارت‌های قدیمی و ظهور مهارت‌های جدید، بلکه آهنگ و شتاب این رویداد و توزیع نابرابر هزینه‌ها و سودهایی است که حاصل می‌شود. کاهش ارزش بسیار سریع مهارت‌های قدیمی و بنابراین اختلال فوری در بازار کار و سرشت اشتغال را می‌توان هم در سطح فردی ملاحظه کرد و هم در سطح اجتماعی. در سطح فردی، شغل‌ها پیوند نزدیکی با هویت شخصی، عزت نفس، و نقش یا منزلت اجتماعی دارند، یعنی همه عواملی که ممکن است به طرز نامطلوبی از کار اضافه بر سازمان متأثر شوند، حتی اگر پتانسیل آسیب‌های شدید اقتصادی را هم به کنار بگذاریم. به علاوه، در سطح جامعه نیز مهارت‌زدایی در حوزه‌های حساس و نیازمند مهارت از قبیل مراقبت‌های بهداشتی پزشکی و هوانوردی ممکن است آسیب‌پذیری‌های خطرناکی در صورت وقوع نقص AI یا حملات خصمانه ایجاد کند. پر و بال دادن به توسعه AI به طرفداری از توانایی‌ها و مهارت‌های جدید و در عین حال پیش‌بینی و تخفیف‌دادن تأثیر آن بر توانایی‌ها و مهارت‌های قدیمی هم نیازمند مطالعه نزدیک است و هم نیازمند ایده‌های بالقوه رادیکالی از قبیل پیشنهاد برای شکلی از «درآمد پایه عمومی»، که محبوبیت و کاربرد آزمایشی آن رو به رشد است. در نهایت، ما نیازمند انسجامی میان‌نسلی بین آن‌هایی که امروز محروم مانده‌اند و آن‌هایی که فردا از مزیت برخوردارند هستیم تا اطمینان حاصل کنیم که این گذار آشوبناک بین اکنون و آینده تا جای ممکن برای هر کسی منصفانه خواهد بود.

چه می‌توانیم انجام دهیم: ارتقاء عاملیت انسان، بدون رفع مسئولیت انسان

AI در حال تمهیدِ گنجینه‌ای رو به رشد از «عاملیت هوشمند» است. اگر چنین منبعی در خدمت هوش انسان قرار داده شود، می‌تواند بی‌اندازه عاملیت انسان را ارتقاء دهد. ما می‌توانیم در سایه پشتیبانی فراهم آمده با AI بیشتر، بهتر و سریع‌تر عمل کنیم. AI، در این معنا از «هوشِ افزوده»، می‌تواند با تأثیری مقایسه شود که موتورها بر روی زندگی ما داشته‌اند. هر چه تعداد افرادی که از فرصت‌ها و مزایای چنین گنجینه‌ای از عاملیت هوشمند «حی و حاضر» برخوردار می‌شوند بیشتر باشد، جوامع ما بهتر خواهند بود. بنابراین مسئولیت حیاتی است، با نظر به اینکه چه نوع AI را توسعه می‌دهیم، چگونه از آن استفاده می‌کنیم و اینکه آیا مزایا و امتیازات آن را با همه شریک می‌شویم یا خیر. مسلماً ریسکِ متناظر غیاب چنین مسئولیتی است. چنین غیابی نه

صرفاً به این دلیل که ما چارچوب اجتماعی‌ـسیاسی غلطی داریم، بلکه به دلیل ذهنیت «جعبه سیاه» اتفاق نمی‌افتد که برحسب آن سیستم‌های AI برای تصمیم‌گیری فراتر از فهم انسان و از این‌رو فراتر از کنترل در نظر گرفته می‌شوند. این نگرانی‌ها نه تنها شامل موارد شناخته‌شده و مشهوری از قبیل مرگ به علت وسایل نقلیه خودکار می‌شود، بلکه شامل کاربردهای پیش پا افتاده‌تر اما باز هم مهمی از قبیل تصمیمات خودکار درباره مردم یا اعتبار می‌شود.

با این حال رابطه بین درجه و کیفیت عاملیتی که مردم از آن برخوردارند و مقدار عاملیتی که ما به سیستم‌های خودکار محول می‌کنیم چه از حیث عملی و چه از حیث اخلاقی [یک بازی] با حاصل جمع صفر نیست. در واقع، اگر AI اندیشمندانه توسعه یافته باشد، فرصت *بهبود و تکثیر* امکانات برای عاملیت انسانی را عرضه می‌کند. نمونه‌های «اخلاق توزیع‌شده» در سیستم‌های انسان‌ـبه‌ـانسان از قبیل وام‌دهی همتا‌ـبه‌ـهمتا را در نظر بگیرید (Floridi, 2013). تعبیه «چارچوب‌های تسهیل‌کننده»ای که برای افزایش احتمال پیامدهای اخلاقی خوب، در مجموعه کارکردهایی که به سیستم‌های AI محول می‌کنیم، طراحی شده‌اند ممکن است در نهایت موجب تقویت، پالایش و گسترش عاملیت انسان شود. سیستم‌های AI اگر به نحو کارآمد طراحی شده باشند می‌توانند سیستم‌های اخلاقی مشترک را تقویت کنند و به آن‌ها استحکام ببخشند.

چه می‌توانیم به دســـت آوریم: افزایش توانایی‌های اجتماعی، بدون تقلیل دادن کنترل انسان

هوش مصنوعی فرصت‌های بی‌شماری برای بهبود و افزایش توانایی‌های افراد و اجتماع به مثابه یک کلّ پیشکش می‌کند. خواه با پیشگیری و علاج بیماری‌ها یا با بهینه کردن حمل و نقل و تدارکات، استفاده از تکنولوژی‌های AI از طریق افزایش رادیکال چیزی که انسان‌ها جمعاً توانایی انجام آن را دارند امکانات بی‌شماری برای بازآفرینی جامعه عرضه می‌کنند. AI بیشتر می‌تواند متکفل هماهنگی بهتر و از این‌رو اهداف بلندپروازانه‌تر شود. هوش انسان اگر با AI تقویت شده باشد می‌تواند راه‌حل‌های جدیدی برای مسائل قدیمی و جدید بیابد، از توزیع منصفانه‌تر یا مؤثرتر منابع تا رویکردی پایدارتر به مصرف. دقیقاً چون چنین تکنولوژی‌هایی قابلیت آن را دارند تا بسیار قدرتمند و اِخلال‌گر باشند، به فراخور ریسک‌هایی را نیز عرضه می‌کنند. اگر بتوانیم وظایف‌مان را

به AI محول کنیم، به نحو فزاینده‌ای لزومی ندارد تا «در یا بالای حلقه» باشیم (یعنی به عنوان بخشی از فرآیند باشیم یا لااقل آن را کنترل کنیم). اما اگر به کاربرد تکنولوژی‌های AI اعتماد کنیم تا توانایی‌های خودمان را به شیوه نادرستی افزایش دهد، ممکن است وظایف مهم و مهم‌تر از هر چیز تمام تصمیمات را به سیستم‌های خودکاری محول کنیم که باید لااقل تا حدی مشروط و تابع نظارت و انتخاب انسان باقی بمانند. این امر به نوبه خود ممکن است قابلیت ما برای نظارت بر عملکرد این سیستم‌ها را تقلیل دهد (اگر دیگر در «بالای حلقه» نباشیم) یا از خطاها و آسیب‌هایی که ظاهر می‌شوند جلوگیری کند یا آن‌ها را جبران نماید («بازبینی حلقه»). همچنین ممکن است هر چه کارکردهای بیشتری به سیستم‌های مصنوعی محول کنیم این آسیب‌های بالقوه افزایش یابند و مستحکم شوند. بنابراین ضرورت دارد تا تعادلی برقرار کنیم بین پیگیری فرصت‌های بلندپروازانه‌ای که AI برای بهبود حیات آدمی عرضه می‌کند و چیزی که می‌توانیم حاصل کنیم از یک سو، و از سوی دیگر تضمین آن‌که کنترل این پیشرفت‌های عمده و پیامدهای آن در دست ماست.

چگونه می‌توانیم همکاری متقابل داشته باشیم: پرورش همبستگی اجتماعی، بدون فرسودن خود مختاری انسان

از تغییر آب و هوا و مقاومت ضدمیکروبی تا تکثیر سلاح‌های اتمی و بنیادگرایی، مشکلات جهانی به طرز فزاینده‌ای از درجه بالایی از پیچیدگیِ هماهنگی برخوردارند بدان معنا که فقط به شرطی می‌توان با موفقیت از عهده آن‌ها برآمد که همه ذی‌ربطان متفقاً به حل آن‌ها مبادرت ورزند و و برای آن همکاری کنند. AI، با راه‌حل‌های پُرداده و الگوریتمی‌اش، می‌تواند با تقویت انسجام و همکاری اجتماعی کمک فراوانی برای پرداختن به این قبیل پیچیدگیِ هماهنگی بکند. به عنوان مثال، تلاش‌ها برای پرداختن به تغییرات آب و هوایی چالش ایجاد واکنش یکپارچه را در معرض دید قرار داده است، هم درون جوامع و هم بین آن‌ها. میزان این چالش چنان است که ممکن است به زودی نیازمند تصمیم بین مهندسی مستقیم آب و هوا و طراحی چارچوب‌های اجتماعی برای تشویق جهت کاهش چشمگیر در انتشار گازهای گلخانه‌ای باشیم. این گزینه دوم باید با یک سیستم الگوریتمی تقویت شود تا همبستگی اجتماعی را شکوفا نماید. چنین سیستمی

نباید از بیرون تحمیل شود؛ این سیستم باید نتیجه یک انتخاب خودـخواسته باشد، چنین انتخابی بی‌شباهت به تصمیم برای نخریدن شکلات نیست اگر قبلاً یک رژیم غذایی را انتخاب کرده باشیم، یا یک ساعت زنگی را برای بیدار شدن تنظیم کرده باشیم. «تلنگرزدن به خود» برای رفتار کردن به شیوه‌هایی که از حیث اجتماعی ارجحیت دارند بهترین شکل تلنگرزدن است و تنها شکلی است که خودمختاری و استقلال را حفظ می‌کند. این [تلنگرزدن به خود] نتیجه تصمیمات و انتخاب‌های انسان است، اما می‌تواند به راه‌حل‌های AI تکیه کند تا اجرا و تسهیل شود. با وجود این، ریسک ماجرا آنجا است که سیستم‌های AI ممکن است خودمختاری انسان را بفرسایند، زیرا ممکن است به تغییرات برنامه‌ریزی نشده و ناخوشایند در رفتارهای انسان بیانجامند تا روال‌هایی را فراهم آورند که کار اتوماسیون و زندگی مردم را آسان‌تر می‌کنند. قدرت پیشگویانه AI و تلنگر بی‌رحمانه، حتی اگر ناخواسته باشد، باید در خدمت خودمختاری انسان قرار بگیرد و انسجام اجتماعی را شکوفا نماید، نه اینکه کرامت انسان یا شکوفایی او را تضعیف کند.

در مجموع، این چهار فرصت و چالش‌های متناظر با آن‌ها تصویری مختلط درباره تأثیر AI بر جامعه و مردم ساکن در آن ترسیم می‌کند. پذیرش بده‌ـبستان‌ها، استفاده از فرصت‌ها در حین تلاش برای پیش‌بینی، جلوگیری یا کمینه کردن ریسک‌های پیش روی، چشم‌انداز تکنولوژی‌های AI برای ارتقاء کرامت و شکوفایی انسان را بهبود خواهد بخشید. با تشریح مزایای بالقوه AI برای افراد و جامعه به مثابه یک کلّ در رویکردی اخلاقی، در بخش بعدی بر «مزیت دوگانه» سازمان‌هایی که چنین رویکردی را اتخاذ می‌کنند تأکید می‌کنیم.

مزیت دوگانه رویکرد اخلاقی به AI

تضمین این امر که پیامدهای AI از حیث اخلاقی ارجحیت داشته باشند مبتنی بر حل تنش بین استفاده از مزایا و تقلیل دادن آسیب‌های بالقوه AI، به اختصار اجتناب هم‌زمان از سوءاستفاده و استفاده ناکافی از این تکنولوژی‌ها، است. در این زمینه، ارزش رویکرد اخلاقی به تکنولوژی‌های AI در این است که به آرامش بیشتری می‌انجامد. سرسپاری به قانون کاملاً ضرورت دارد (دستمزدی که ضروری است)، اما به میزان قابل توجهی کافی نیست (بیشترین کاری نیست که می‌توان انجام داد) (Floridi, 2018). با استفاده از یک تمثیل، تفاوت این دو تفاوت بین بازی بر

حسب قواعد و بازی خوب است به نحوی که فرد بتواند در بازی برنده شود. اتخاذ رویکردی اخلاقی به AI چیزی را اعطا می‌کند که در اینجا آن را «مزیت دوگانه» تعریف می‌کنیم. از یک طرف، اخلاق سازمان‌ها را قادر می‌سازد تا از مزیت ارزش اجتماعی که AI فراهم می‌آورد استفاده کنند. این مزیت عبارت است از توانایی برای شناسایی و استفاده از فرصت‌های جدیدی که از حیث اجتماعی قابل قبول و مناسب‌اند. از طرف دیگر، اخلاق سازمان‌ها را قادر می‌سازد تا اشتباهات گزاف را پیش‌بینی کنند و از آن‌ها پیشگیری نمایند یا لااقل آن‌ها را کمینه سازند. این همان مزیتِ پیشگیری و کاهش راه کارهایی است که از حیث اجتماعی غیر قابل قبول از آب در می‌آیند و از این‌رو رد می‌شوند، حتی وقتی که از حیث قانونی غیر قابل تردید باشند. این مزیت از هزینه فرصت‌هایی که انتخاب نمی‌شوند یا گزینه‌هایی که به خاطر ترس از خطا اتخاذ نمی‌شوند نیز می‌کاهد.

مزیت دوگانه اخلاق فقط می‌تواند در محیطی عمل کند که در آن اعتماد عمومی وجود داشته باشد و مسئولیت‌های روشن به طور کلی پذیرفته شده باشد. مقبولیت و پذیرش عمومی تکنولوژی‌های AI فقط به شرطی اتفاق خواهد افتاد که مزایا را معنادار بدانیم و ریسک‌ها را بالقوه، اگرچه قابل پیشگیری، قابل کمینه شدن یا لااقل چیزی در نظر بگیریم که بتوان از طریق مدیریت ریسک (مثلاً بیمه) یا جبران خسارت از افراد در مقابل آن محافظت کرد. این تلقی‌ها به نوبه خود به التزام عمومی نسبت به توسعه تکنولوژی‌های AI، گشودگی نسبت به نحوه عمل آن‌ها و مکانیسم‌های تنظیم‌گری و جبران خسارت قابل فهم و دسترس‌پذیر برای همگان بستگی دارد. در این شیوه، رویکردی اخلاقی به AI را نیز می‌توان همچون یک سیستم هشدار اولیه در برابر ریسک‌هایی در نظر گرفت که ممکن است کل سازمان‌ها را به مخاطره اندازند. ارزش روشنِ مزیت دوگانه رویکرد اخلاقی به AI برای هر سازمان به قدر کفایت هزینه التزام، گشودگی و رقابت‌پذیری را که برای چنین رویکردی ضروری است، توجیه می‌کند.

چارچوبی یکپارچه برای اصول AI در جامعه

AI4People نخستین ابتکار عمل برای در نظر گرفتن استلزامات اخلاقی AI نیست. سازمان‌های بسیاری قبلاً احکامی درباره ارزش‌ها یا اصولی تولید کرده‌اند که باید راهنمای توسعه

و استقرار AI در جامعه قرار بگیرند. ما در اینجا به جای فعالیت مشابهی که به صورت بالقوه زاید است، می‌کوشیم گفتگو را به نحو سازنده از اصول به سمت خط‌مشی‌های پیشنهادی، بهترین رویه‌ها و توصیه‌های انضمامی برای استراتژی‌های جدید پیش ببریم. چنین توصیه‌هایی در خلاء عرضه نمی‌شوند. اما به جای تولید زنجیره دیگری از اصول که همچون بنیادی اخلاقی برای توصیه‌های ما عمل کنند، ما تلفیقی از مجموعه اصول موجود عرضه می‌کنیم که چند ابتکار عمل و سازمان چندـذی‌ربطی مشهور تولید کرده‌اند. توضیح کامل‌تری درباره گستره، گزینش و روش ارزیابی این مجموعه اصول در Cowls and Floridi (در دست انتشار) در دسترس است. در اینجا ما بر روی اشتراکات و تفاوت‌های قابل توجهی تمرکز می‌کنیم که، با در نظر گرفتن ۲۰ توصیه‌ای که در ادامه ارائه می‌شوند، در سراسر این مجموعه اصول قابل مشاهده است. مستنداتی که ما ارزیابی کردیم عبارتند از:

۱. اصول Asilomar AI که با حمایت مؤسسه آینده حیات (Future of Life Institute) در همکاری با شرکت‌کنندگان کنفرانس Asilomar سطح‌ـبالای در ژانویه ۲۰۱۷ توسعه یافته‌اند (از این پس «Asilomar»؛ اصول Asilomar AI؛ ۲۰۱۷)؛

۲. بیانیه مونترئال برای AI پاسخگو و مسئول که با حمایت دانشگاه مونترئال و متعاقب گردهم‌آیی درباره توسعه AI با مسئولیت اجتماعی در نوامبر ۲۰۱۷ گسترش یافته است (از این پس «مونترئال»، بیانیه مونترئال، ۲۰۱۷)؛

۳. اصول کلی عرضه‌شده در نسخه دوم *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. این رساله جهانی که به صورت جمع‌سپاری[۱] انجام شده، حاصل تشریک مساعی ۲۵۰ رهبر فکری در سطح جهان برای توسعه اصول و توصیه‌هایی برای طراحی و گسترش اخلاقی سیستم‌های خودکار و

---

[۱] crowd-sourced: جمع‌سپاری یا انبوه‌سپاری ترکیبی از دو کلمه جمعیت و برون‌سپاری به معنای برون‌سپاری به انبوه مردم یا متخصصان است. (مترجم)

هوشمند است. این رساله در دسامبر ۲۰۱۷ منتشر شده است (از این پس «IEEE»؛ IEEE، ۲۰۱۷)؛

۴. اصول اخلاقی عرضه‌شده در *Statement on Artificial Intelligence, Robotics and "Autonomous" Systems* که انجمن اروپایی اخلاق در علم و تکنولوژی‌های جدیدِ کمیسیون اروپایی در مارس ۲۰۱۸ منتشر ساخته است (از این پس «EGE»؛ EGE، ۲۰۱۸)؛

۵. «پنج اصل فراگیر برای کد AI» که در پاراگراف ۴۱۷ گزارش کمیته هوش مصنوعی مجلس اعیان بریتانیا، *AI in the UK: ready, willing and able?*، در آوریل ۲۰۱۸ منتشر شده است (از این پس «HIUK»؛ مجلس اعیان، ۲۰۱۸)؛ و

۶. اصول مسلم مشارکت در AI، سازمانی چندـذی‌ربطی متشکل از آکادمیسین‌ها، پژوهشگران، سازمان‌های جامعه مدنی، کمپانی‌های سازنده و استفاده کننده از تکنولوژی‌های AI، و دیگر گروه‌ها (از این پس «مشارکت» (Partnership)؛ مشارکت در AI، ۲۰۱۸).

در مجموع، این مستندات ۴۷ اصل به دست می‌دهند. روی‌هم‌رفته، ما درجه انسجام و هم‌پوشانی چشمگیر و اطمینان‌بخشی بین شش مجموعه از اصول می‌یابیم. این نکته را می‌توان به روشنی تمام با مقایسه این مجموعه اصول با مجموعه چهار اصل مرکزی که عموماً در اخلاق زیست‌شناسی به کار می‌رود نشان داد:

نیکوکاری، غیرـزیان‌بخشی، خودمختاری و عدالت. این مقایسه نباید حیرت‌آور باشد. از بین همه حوزه‌های اخلاق کاربردی، اخلاق زیست‌شناسی حوزه‌ای است که وقتی از منظر بوم‌شناختی به فرم‌های جدیدِ عوامل، بیماران و محیط‌زیست می‌پردازد بیشترین شباهت را با اخلاق دیجیتال دارد (Floridi, 2013). این چهار اصل اخلاق زیست‌شناختی به طرز شگفت‌آوری با چالش‌های اخلاقی تازه‌ای که هوش مصنوعی مطرح می‌کند سازگارند. اما این اصول جامع نیستند. بر اساس تحلیل تطبیقی که در ادامه می‌اید، ما ادعا می‌کنیم که یک اصل جدید دیگر نیز مورد نیاز است: *توضیح‌پذیری*، که هم شامل قابل فهم بودن است و هم شامل پاسخگویی.

نیکوکاری: ارتقاء به‌زیستی، حفظ کرامت، و تداوم و محافظت از سیاره

از بین چهار اصل اخلاق زیست‌شناسی، نیکوکاری شاید آسان‌ترین اصلی باشد که بتوان از میان شش مجموعه اصولی که در اینجا تلفیق می‌کنیم مشاهده نمود. اصل مربوط به ایجاد تکنولوژی AI که برای بشریت سودمند است به طرق مختلفی بیان می‌شود، اما نوعاً در بالای همه فهرست‌های اصول ظاهر می‌شود. اصول مونترئال و IEEE هر دو عبارت «به‌ـزیستی» را بکار می‌برند: برای مونترئال، «توسعه AI باید در نهایت به‌ـزیستی همه مخلوقات ذی‌شعور را ارتقاء دهد»؛ در حالی‌که IEEE ضرورت «اولویت دادن به به‌ـزیستی انسان به‌مثابه نتیجه‌ای در همه طرح‌های سیستم» را بیان می‌کند. AIUK و Asilomar هر دو این اصل را به منزله «خیر عمومی» توصیف می‌کنند: بر طبق AIUK، AI باید «برای خیر عمومی و انتفاع بشریت توسعه یابد». [سندِ] مشارکت (Partnership) این نیت را این‌گونه توصیف می‌کند: «تضمین این امر که تکنولوژی‌های AI برای هر تعداد مردمی که ممکن است مفید واقع شوند و آن‌ها را توانمند سازند»؛ در حالی‌که EGE بر اصل «کرامت انسان» و «تداوم‌پذیری» تأکید دارد. اصل «تداوم‌پذیری» در اینجا شاید وسیع‌ترین تفسیر از نیکوکاری را عرضه می‌کند با این ادعا که «تکنولوژی AI باید در توافق با این امر باشد که ... پیش‌شرط‌های اساسی برای حیات بر روی سیاره ما، کامیابی پیوسته برای نوع بشر و محافظت از یک محیط‌زیست خوب برای نسل‌های بعدی را تضمین نماید». در مجموع، برجستگی این اصول نیکوکاری قاطعانه بر اهمیت مرکزی ارتقاء به‌ـزیستی مردم و سیاره تأکید دارد.

غیرزیان‌بخشی: حریم خصوصی، امنیت و «توانایی احتیاط»

اگرچه «فقط کار نیک انجام دادن» (نیکوکاری) و «آسیب نرساندن» (غیرـزیان‌بخشی) منطقاً هم‌ارز به نظر می‌رسند، اما در هر دو سیاق اخلاق زیست‌شناسی و اخلاق AI آن‌ها اصول متمایزی را بیان می‌کنند که هر کدام‌شان نیازمند توضیح است. درحالی‌که آن‌ها به‌ـزیستی، تقسیم مزایا و ترفیع خیر عمومی را تشویق می‌کنند، اما هر یک از شش مجموعه اصول علیه بسیاری از پیامدهای بالقوه منفیِ استفاده مفرط یا سوءاستفاده از تکنولوژی‌های AI نیز هشدار می‌دهند. یکی از نگرانی‌های ویژه ممانعت از تجاوزها به حریم خصوصی و شخصی است که به عنوان یک

اصل در پنج مجموعه از این شش مجموعه اصول فهرست شده است و به عنوان بخشی از اصول «حقوق بشر» در سند IEEE ثبت شده است. در هر مورد، حریم خصوصی چنان توصیف می‌شود که در نهایت با دسترسی افراد به داده‌های شخصی و کنترل بر نحوه استفاده از آن‌ها پیوند دارد. با این حال تجاوز به حریم خصوصی تنها خطری نیست که در پذیرش AI باید از آن پرهیز نمود. چند سند نیز بر اهمیت پرهیز از سوءاستفاده از تکنولوژی‌های AI به شیوه‌های دیگر تأکید دارند. اصول Asilomar با استناد به تهدیدهای مربوط به مسابقه تسلیحاتی AI و خودـبهسازیِ بازگشتی AI و نیز نیاز به «احتیاط» پیرامون «کرانه‌های بالا در توانایی‌های آتیِ AI»، بر روی این نکته کاملاً صراحت دارد. سند مشارکت نیز به‌طور مشابهی اهمیت بکار بستن AI «در درون محدودیت‌های امن» را بیان می‌کند. سند IEEE در این ضمن ضرورت «پرهیز از سوءاستفاده» را ذکر می‌کند، درحالی‌که بیانیه مونترئال مدعی است کسانی که AI را توسعه می‌دهند «باید با فعالیت علیه ریسک‌هایی که نتیجه ابداعات تکنولوژیک‌شان است مسئولیت‌شان را فرض بگیرند» و بدین‌ترتیب ضرورت مشابه برای مسئولیت در سند EGE را منعکس کرده است.

از این هشدارهای متعدد کاملاً روشن نیست که آیا مردمی که AI را توسعه می‌دهند باید تشویق شوند تا آسیب نرسانند یا خود تکنولوژی- به عبارت دیگر، این فرانکنشتاین است که باید در برابر شرارتش محافظت صورت بگیرد یا هیولای او. [در این سندها] مسئله نیت نیز سردرگم است: ارتقاء غیرـزیان‌بخشی می‌تواند هم شامل پیشگیری از آسیب‌های تصادفی (چیزی که بالاتر آن را «استفاده مفرط» نامیدیم) باشد و هم شامل آسیب‌های عمدی (چیزی که «سوءاستفاده نامیدیم»). با توجه به اصول غیرـزیان‌بخشی، ضرورتی ندارد که این پرسش از جنس این یا آن باشد: مسئله به سادگی عبارت است از پیشگیری از آسیب‌هایی که ظاهر می‌شوند، خواه این آسیب‌ها برآمده از نیت انسان‌ها باشند خواه برآمده از رفتار پیش‌بینی‌نشده ماشین‌ها (شامل ترغیب ناخواسته رفتار انسان به شیوه‌های نامطلوب). با این حال، این پرسش‌های اساسی درباره عاملیت، نیت و کنترل وقتی غامض‌تر می‌شوند که اصل بعدی را در نظر بگیریم.

خودمختاری: قدرت برای تصمیم

یکی دیگر از اصول مسلم اخلاق زیست‌شناسی اصل خودمختاری است: این ایده که افراد حق دارند درباره درمانی که می‌کنند یا دریافت نمی‌کنند تصمیماتی برای خودشان بگیرند. در سیاق پزشکی، این اصل خودمختاری غالباً وقتی مختل می‌شود که بیماران فاقد ظرفیت روانی برای تصمیم‌گیری در جهت بهترین مصلحت خودشان هستند؛ بدین‌ترتیب ناخواسته از خودمختاری صرف‌نظر می‌شود. با AI وضعیت پیچیده‌تر هم می‌شود: وقتی AI و عاملیت هوشمند آن را می‌پذیریم، *با رضایت* بخشی از قدرت تصمیم‌گیری خود را به ماشین‌ها واگذار می‌کنیم. از این‌رو، تصدیق اصل خودمختاری در سیاق AI یعنی ایجاد توازن بین آن بخش از قدرت تصمیم‌گیری که برای خودمان احراز می‌کنیم و آن بخشی که به عوامل مصنوعی تفویض می‌کنیم.

اصل خودمختاری به صراحت در چهار سند از شش سند بیان شده است. بیانیه مونترئال نیاز به توازن بین تصمیم‌گیری انسان و ماشین را به‌روشنی بیان می‌کند و می‌گوید که «توسعه AI باید خودمختاری همه انسان‌ها را *ارتقاء دهد* و خودمختاری سیستم‌های کامپیوتری را ... *کنترل نماید* (تأکید از ماست).» سند EGE نیز مدعی است که سیستم‌های خودمختار «نباید به آزادی انسان‌ها برای استقرار استانداردها و هنجارهای خودشان و توانایی برای زیستن برحسب آن‌ها لطمه وارد کنند»، در حالی‌که سند AIUK این موضع محدودتر را اتخاذ می‌کند که «قدرت خودمختار برای آسیب رساندن، تخریب یا فریفتن انسان‌ها هرگز نباید به AI اعطا شود». سند Asilomar نیز به‌طور مشابهی از اصل خودمختاری حمایت می‌کند، تا جایی‌که «انسان‌ها باید انتخاب کنند که آیا تصمیمات را به سیستم‌های هوشمند محول کنند و چگونه، تا اهداف منتخب انسان را محقق سازند».

این سندها با منعکس کردن تمایزی که در بالا بین نیکوکاری و غیرـزیان‌بخشی ترسیم شد، عقیده مشابهی را به شیوه‌های اندک متفاوت بیان می‌کنند: نه تنها خودمختاری انسان‌ها باید ارتقاء یابد، بلکه خودمختاری ماشین‌ها نیز باید محدود شود و ذاتاً برگشت‌پذیر باشد، باید خودمختاری انسان دوباره احراز شود (مورد خلبانی را در نظر بگیرید که قادر به خاموش کردن خلبان خودکار و به دست گرفتن مجدد کنترل کامل هواپیما است). در مجموع، نکته مرکزی

محافظت از ارزش ذاتی انتخاب انسان- لااقل در مورد تصمیمات مهم- و در نتیجه جلوگیری از ریسک تفویض بیش از اندازه امور به ماشین‌ها است. از این‌رو، آنچه در اینجا مهم‌ترین امر به نظر می‌رسد چیزی است که می‌توانیم آن را مُدلِ «فرا-خودمختاری» یا «تصمیم به تفویض» بنامیم: انسان‌ها باید همواره این قدرت را که *تصمیم بگیرند چه تصمیماتی اتخاذ کنند* و استفاده از آزادی انتخاب در صورت لزوم را حفظ نمایند و در مواردی که دلایل مهمی چون کارآمدی ممکن است بر از دست دادن کنترل تصمیم‌گیری بچربد آن را واگذار کنند. همان‌طور که پیش‌بینی شد، هر تفویضی باید قابل لغو باشد (تصمیم‌گیری برای تصمیم مجدد). تصمیم برای گرفتن تصمیمات یا تفویض آن‌ها در خلاء رخ نمی‌دهد. این ظرفیت برای تصمیم (برای تصمیم و تصمیم مجدد) نیز به تساوی در جامعه توزیع نمی‌شود. در پایانِ چهار اصل ملهم از اخلاق زیست‌شناسی به پیامدهای این ناهمگونی بالقوه در خودمختاری پرداخته می‌شود.

عدالت: ارتقاء کامیابی و حفظ انسجام

آخرین اصل از اصول چهارگانه و کلاسیک اخلاق زیست‌شناسی عدالت است که نوعاً در رابطه با توزیع منابع به آن استناد می‌شود، منابعی از قبیل گزینه‌های درمانی جدید و آزمایشی یا صرفاً دسترسی عمومی به مراقبت‌های بهداشتی متعارف. باز هم این اصل اخلاق زیست‌شناسی بازتاب‌های روشنی در اصول AI که ما تحلیل می‌کنیم می‌یابد. اهمیت «عدالت» به‌روشنی در بیانیه مونترئال ذکر می‌شود که مدعی است «توسعه AI باید عدالت را ارتقاء دهد و بکوشد تا انواع تبعیض را حذف کند»، در حالی‌که اصول Asilomar هم شامل ضرورت «انتفاع مشترک» از AI است و هم شامل ضرورت «کامیابی مشترک» از AI. سند EGE ذیل اصل موسوم به «عدالت، برابری و انسجام» مدعی است که AI باید «به عدالت جهانی و دسترسی برابر به منافع» تکنولوژی‌های AI مساعدت نماید. این سند علیه ریسک سوگیری در مجموعه داده‌های بکار رفته برای آموزش سیستم‌های AI نیز هشدار می‌دهد و -به صورت منحصر به‌فردی در بین سندها- مدعی ضرورت دفاع در برابر تهدیدهایی است که متوجه «انسجام» است، که شامل «سیستم‌های مساعدت متقابل از قبیل بیمه اجتماعی و مراقبت‌های بهداشتی» می‌شود. از آنجایی‌که EGE تن‌واره‌ای اروپایی است، احتمالاً تأکید بر محافظت از سیستم‌های حمایت اجتماعی بازتاب

ژئوپولیتیک است. گزارش AIUK مدعی است که شهروندان باید قادر به «شکوفایی روانی، عاطفی و اقتصادی به موازات هوش مصنوعی» باشند. در این حین، سند مشارکت چارچوب محتاطانه‌تری را اتخاذ می‌کند و متعهد می‌شود که «به منافع همه گروه‌هایی که ممکن است تحت تأثیر پیشرفت‌های AI قرار بگیرند احترام بگذارد».

مانند اصول دیگری که پیشتر از آن‌ها بحث شد، این تفاسیر از معانی عدالت به‌مثابه اصلی اخلاقی در سیاق AI نیز شباهت گسترده‌ای دارند اگرچه حاوی تفاوت‌های ظریفی نیز هستند. در این اسناد عدالت به طرق گوناگون مرتبط است با:

a) استفاده از AI برای اصلاح اشتباهات گذشته از قبیل حذف تبعیض غیرمنصفانه؛

b) تضمین این امر که استفاده از AI منافع مشترک خلق می‌کند (یا لااقل منافع قابل اشتراک)؛

c) پیشگیری از ایجاد آسیب‌های جدید از قبیل تخریب ساختارهای اجتماعی موجود.

شیوه‌های متفاوتی که موقعیت AI، در مقایسه با مردم، در نسبت با عدالت توصیف می‌شود نیز قابل توجه است. در Asilomar و EGE به ترتیب خود تکنولوژی‌های AI است که «باید تا جای ممکن به مردم فایده رساند و آن‌ها را توانمند کند» و «به عدالت جهانی یاری رساند»، در حالی‌که در سند مونترئال این «*توسعه* AI» است که «باید عدالت را ارتقاء دهد» (تأکید از ماست). در این حین، در سند AIUK مردم باید صرفاً «به موازات» AI شکوفا شوند. در اینجا مقصود ما موشکافی معنایی نیست. شیوه‌های متنوع توصیف رابطه بین مردم و AI در این اسناد حکایت از آشفتگی گسترده‌تر درخصوص AI به عنوان گنجینه انسان‌ـساخته‌ای از «عاملیت هوشمند» دارد. به زبان ساده و با مراجعه مجدد به تمثیل اخلاق زیست‌شناسی ما، آیا ما انسان‌ها همان بیمارانی هستیم که «درمان» AI را که پزشک تجویز کرده است دریافت می‌کنیم؟ یا هر دو؟ به نظر می‌رسد ما باید این مسئله را حل کنیم پیش از آنکه تلاش کنیم به پرسش بعدی پاسخ بگوییم یعنی این پرسش که آیا اصلاً این درمان کار می‌کند یا نه. این نکته همان توجیه مرکزی برای آن است که اصل جدیدی را در میان این اسناد شناسایی کنیم، اصلی که از اخلاق زیست‌شناسی استخراج نشده باشد.

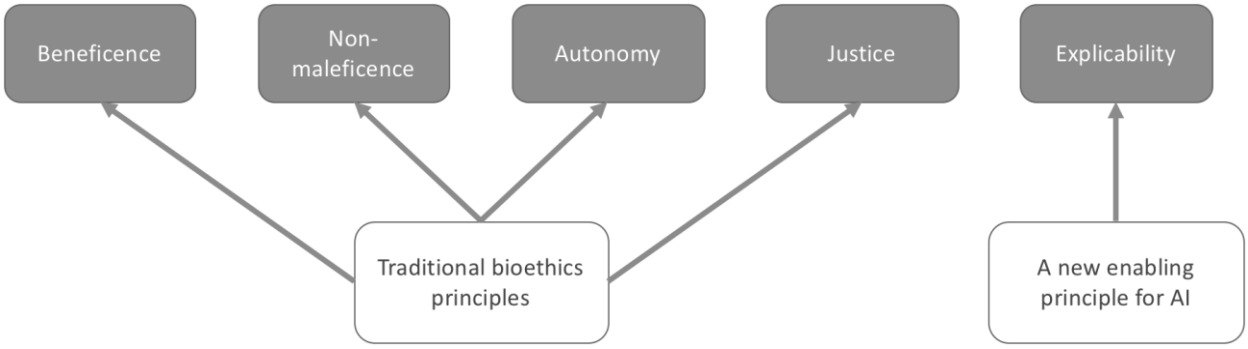
توضیح‌پذیری: تواناسازی دیگر اصول از طریق فهم‌پذیری و پاسخگویی

پاسخ کوتاه به این پرسش که آیا «ما» همان بیمارانیم یا پزشک همان کسی است که در واقع ما می‌توانستیم باشیم- بستگی به شرایط و این نکته دارد که «ما» در حیات روزمره‌مان که هستیم. وضعیت ذاتاً نابرابر است: جزء کوچکی از بشریت در حال حاضر درگیر در طراحی و توسعه مجموعه‌ای از تکنولوژی‌های AI است که حیات روزمره تقریباً هر کس دیگری را دگرگون می‌کند. این واقعیت خشک بر روی مؤلفانی که اسنادشان را تحلیل می‌کنیم تأثیری ندارد. در کل، به ضرورت *فهم* و *پایبندی به مسئولیت* فرآیندهای تصمیم‌ـگیری AI اشاره می‌شود. این اصل با استفاده از عبارات متفاوت بیان می‌شود: «شفافیت» در Asilomar؛ «پاسخگویی» در EGE؛ هم «شفافیت» و هم «پاسخگویی» در IEEE؛ «فهم‌پذیری» در AIUK؛ و «قابل فهم و تفسیر بودن» در سند مشارکت. هر یک از این اصول، اگرچه به طرق متفاوت توصیف می‌شوند، چیزی ظاهراً نو درباره AI به چنگ می‌آورند: اینکه عملکردهای آن اغلب نامرئی یا غیر قابل فهم برای همگان است اما (در بهترین حالت) خبره‌ترین افراد آن‌ها را مشاهده می‌کنند.

از این‌رو افزودن این اصل، تحت عنوان «توضیح‌پذیری» هم در معنای معرفت‌شناختیِ «فهم‌پذیری» (به عنوان پاسخی به این پرسش که «[هوش مصنوعی] چگونه کار می‌کند؟») و هم در معنای اخلاقیِ «پاسخگویی» (به عنوان پاسخی به این پرسش که «چه کسی مسئول نحوه عمل آن [هوش مصنوعی] است؟»)، قطعه گم‌شده و حیاتیِ این پازل است وقتی می‌کوشیم چارچوب اخلاق زیست‌شناسی را به اخلاق AI اِعمال کنیم. این اصل مکمل چهار اصل دیگر است: برای آنکه AI نیکوکار و غیرـآسیب‌رسان باشد، ما باید قادر به فهم سود و زیانی باشیم که در واقع به جامعه وارد می‌کند و باید بفهمیم به چه طرقی این کار را می‌کند؛ برای آن‌که AI خودمختاری انسان را ارتقاء دهد و آن را محدود نکند، «تصمیم ما درباره اینکه چه کسی باید تصمیم بگیرد» باید مُستحضر به این نکته باشد که AI چگونه به جای ما عمل خواهد کرد؛ و برای آن‌که AI منصف باشد، باید متقاعد شویم که تکنولوژی- یا، دقیق‌تر، مردم و سازمان‌هایی که آن را توسعه می‌دهند و مستقر می‌سازند- در صورت وقوع پیامدی منفی مسئول و پاسخگو است، که این امر به نوبه خود نیازمند فهم علت وقوع این پیامد است. کوتاه سخن اینکه ما باید درباره شرایط رابطه

بین خودمان و این تکنولوژیِ دگرگون‌کننده بحث و گفتگو کنیم، به دلایلی که به آسانی برای شخص «بی‌خانمان» و انگشت‌نما قابل فهم است.

در مجموع، ما ادعا می‌کنیم که این پنج اصل مقصود هر یک از ۴۷ اصل مشمول در آن شش سند پُر سر و صدا و متخصص‌ـمحور را به چنگ می‌آورند و چارچوبی اخلاقی را شکل می‌دهند که ما در قالب آن توصیه‌های خود را در زیر عرضه می‌کنیم. این چارچوب اصول در شکل ب نشان داده می‌شود.



شکل ب: چارچوبی اخلاقی برای AI، متشکل از چهار اصل سنتی و یک اصل جدید

توصیه‌هایی برای یک جامعه AI خوب

این بخش توصیه‌هایی برای یک جامعه AI خوب مطرح می‌کند و شامل دو قسمت است: یک دیباچه و ۲۰ نکته عملی. چهار نوع نکته عملی وجود دارد: *ارزیابی*، *توسعه*، *ایجاد انگیزه* و *حمایت*. برخی توصیه‌ها ممکن است مستقیماً توسط سیاست‌گذاران ملی یا اروپایی و در همکاری با ذی‌ربطان در شرایط مقتضی پذیرفته شوند. در خصوص توصیه‌های دیگر، سیاست‌گذاران ممکن است نقشی قانونی و توانمندساز برای تلاش‌هایی ایفا کنند که اشخاص ثالث بر عهده گرفته‌اند یا به جریان انداخته‌اند.

دیباچه

ما معتقدیم که برای ایجاد یک جامعه AI خوب اصول اخلاقیِ مشخص‌شده در بخش قبلی باید در عملکردهای پیش‌فرض AI تعبیه شود. به ویژه، AI باید به طرقی طراحی شود و توسعه

یابد که با احترام به خودمختاری انسان نابرابری را کاهش و توانمندی اجتماعی را افزایش دهد و منافع مشترک برای همگان را به نحو منصفانه زیاد کند. به ویژه این نکته اهمیت دارد که AI قابل توضیح باشد زیرا توضیح‌پذیری ابزاری تعیین‌کننده برای بنا کردن اعتماد عمومی و فهم تکنولوژی است.

همچنین معتقدیم که ایجاد یک جامعه AI خوب نیازمند رویکردی چندـذی‌ربطی است که مؤثرترین شیوه برای تضمین این امر است که AI در خدمت نیازهای جامعه قرار دارد و توسعه‌دهندگان، کاربران و قانون‌گذاران را قادر می‌سازد همگی بر روی یک کشتی قرار بگیرند و از ابتدا همکاری کنند.

چارچوب‌های مختلف فرهنگی نگرش‌ها به تکنولوژی جدید را می‌سازند. این سند رویکردی اروپایی را عرضه می‌کند که مُقدر است مکمل دیگر رویکردها باشد. ما متعهد به توسعه تکنولوژی AI به طریقی هستیم که *اعتماد مردم را حفظ می‌کند، در خدمت منفعت عمومی دارد*، و *مسئولیت اجتماعی مشترک را تقویت می‌کند*.

در نهایت، این مجموعه توصیه‌ها را باید چونان «سندی زنده» ملاحظه کرد. نکات عملی به گونه‌ای دینامیک طراحی شده‌اند که نیازی به خط‌مشی‌های منفرد یا سرمایه‌گذاری‌های یکـطرفه ندارند بلکه به تلاش‌های پیوسته و مداوم نیاز دارند تا تأثیرات‌شان استمرار داشته باشد.

نکات عملی

ارزیابی

۱. ظرفیت مؤسسات موجود، مثل دادگاه‌های مدنی ملی، را به منظور اصطلاح خطاهای ایجادشده یا برطرف نمودن آسیب‌های واردشده از سوی سیستم‌های AI، ارزیابی کنید. این ارزیابی باید حضور شالوده‌های پایدار و مورد اجماع اکثریت برای مسئولیت را از مرحله طراحی به بَعد ارزیابی کند تا احتمال قصور و ناسازگاری کاهش یابد. (همچنین نگاه کنید به **توصیه ۵**).

۲. **ارزیابی کنید که کدام یک از وظایف و عملکردهای تصمیم‌گیری *نباید* به سیستم‌های AI محول شوند،** از طریق استفاده از سازوکارهای مشارکتی برای اطمینان از هم‌سویی با

ارزش‌های اجتماعی و فهم باور عامه. این ارزیابی باید قانون‌گذاری موجود را در نظر بگیرد و با گفتگوی مداوم بین همه ذی‌ربطان (از جمله حکومت، صنعت، و جامعه مدنی) حمایت شود تا بتوان در مورد اینکه AI چگونه بر باور جامعه تأثیر خواهد گذاشت بحث کرد (**به همراه توصیه ۱۷**).

۳. ارزیابی کنید که آیا مقررات کنونی به نحوی کافی مبتنی بر اخلاق است تا بتوانند چارچوبی قانونی فراهم آورد که بتواند همگام با پیشرفت‌های تکنولوژیک پیش برود یا خیر. این ارزیابی می‌تواند شامل چارچوبی متشکل از اصول کلیدی باشد که بر مسائل اضطراری و/یا پیش‌بینی‌ناشده قابل اِعمال است.

توسعه

۴. چارچوبی را توسعه دهید که توضیح‌پذیری سیستم‌های AI دخیل در تصمیمات اجتماعی مهم را افزایش دهد. ویژگی اصلی این چارچوب توانایی افراد برای فراهم کردن توضیحی واقعی، سرراست و واضح از فرایند تصمیم‌گیری، به ویژه هنگام وقوع عواقب ناخواسته، است. ممکن است این کار مستلزم توسعه چارچوب‌های خاص برای صنایع مختلف باشد، و مجامع حرفه‌ای در کنار کارشناسان حوزه علم، تجارت، حقوق، و اخلاق، باید در این فرایند دخیل باشند.

۵. رویه‌های قانونی مناسب را *توسعه دهید* و زیرساخت‌های IT برای تشکیلات قضایی را بهبود بخشید تا بازرسی دقیق تصمیمات الگوریتمی در دادگاه میسر شود. ممکن است این گزینه، همان‌طور که در **توصیه ۴** اشاره شد، شامل ایجاد چارچوبی برای توضیح‌پذیری AI مختصِ سیستم حقوقی نیز باشد. نمونه‌هایی از رویه‌های مناسب می‌تواند شامل اِفشای عملی اطلاعات حساس تجاری در دعاوی قضایی مربوط به IP باشد، و – در جایی که اِفشاء کردن مستلزم ریسک‌های غیرقابل قبولی، مثلاً خطری برای امنیت ملی، است– پیکربندی سیستم‌های

AI برای اتخاذ راه‌حل‌های فنی به صورت پیش‌فرض، مثل اثبات دانایی صفر [یا پروتکل دانایی صفر][1] به منظور ارزیابی قابل اعتماد بودن آن‌ها.

۶. سازوکارهای حسابرسی برای سیستم‌های AI برای شناسایی نتایج ناخواسته، مثل سوگیری غیرمنصفانه، و (برای مثال، در همکاری با بخش بیمه) سازوکاری منسجم برای مواجهه با ریسک‌های سهمگین بخش‌هایی که به AI نیاز فراوان دارند را *توسعه دهید*. آن ریسک‌ها را می‌توان با سازوکارهای چندـذی‌ربطی بالادستی کاهش داد. تجربه پیشادیجیتال نشان می‌دهد که در برخی موارد، ممکن است چند دهه طول بکشد تا جامعه از طریق دوباره متعادل‌سازی حقوق و محافظت به نحو مناسبی با تکنولوژی همراه شود و اعتماد را احیا کند. هرچه زودتر کاربران و حکومت‌ها در این فرایند دخیل شوند – همان‌طور که ICT این امر را ممکن ساخته است – این تأخیر کوتاه‌تر خواهد بود.

۷. فرایند یا سازوکار جبران خسارت را *توسعه دهید* تا خطا یا مشکلات ایجادشده توسط AI را برطرف یا اصلاح کند. جوامع برای آنکه اعتماد عمومی را به AI جلب کنند به سازوکاری برای جبران خسارت نیاز دارند که به نحو گسترده دسترسی‌پذیر و قابل اتکا باشد و آسیب‌های واردشده، هزینه‌های ایجادشده یا مسائل دیگری که فناوری مُسبب آن‌ها بوده است را جبران کند. یک چنین سازوکاری ضرورتاً مستلزم تقسیم شفاف و جامع مسئولیت بین انسان‌ها و/یا سازمان‌هاست. برای مثال ما می‌توانیم از صنعت هوافضا درس‌های بسیاری بیاموزیم، صنعتی که سیستم ثابت‌شده‌ای برای رسیدگی جامع و جدی به نتایج ناخواسته دارد. توسعه این فرایند باید از ارزیابی ظرفیت موجود، که در **توصیه ۱** مطرح شد، پیروی کند. اگر در این ارزیابی فقدان ظرفیتی تشخیص داده شود، باید راه‌حل‌های سازمانی بیشتری در سطوح ملی و/یا اتحادیه اروپا

[1] در رمزنگاری روشی است که طرف اثبات‌کننده می‌تواند به طرف تصدیق‌کننده ثابت کند بیانیه‌ی ارائه‌شده صحیح است. این روش فقط صحت بیانیه را تصدیق می‌کند و هیچ اطلاعات اضافه‌ای را بجز این حقیقت که بیانیه واقعاً صحت دارد ارسال نمی‌کند. (مترجم)

توسعه یابند تا مردم بتوانند جبران مافات کنند. چنین راه‌حل‌هایی می‌توانند شامل این موارد باشند:

* «بازرس AI» برای اطمینان حاصل کردن از اینکه استفاده‌های ناعادلانه یا غیرمنصفانه از AI مورد حسابرسی قرار می‌گیرند؛
* فرایندی هدایت‌شده برای ثبت شکایاتی مثل درخواست آزادی اطلاعات؛ و
* توسعه سازوکارهای بیمه مسئولیت، که به عنوان ضمیمه‌ای واجب برای مجموعه‌های خاص پیشنهادات AI در اتحادیه اروپا و بازارهای دیگر مورد نیاز خواهند بود. این کار تضمین می‌کند که قابل‌اعتمادبودن نسبی مصنوعاتی که با AI کار می‌کنند، به ویژه در صنعت روباتیک، در قیمت‌گذاری بیمه و بنابراین در قیمت‌های بازار محصولات رقیب منعکس می‌شود.

همه این راه‌حل‌ها، فارغ از اینکه کدام‌یک را انتخاب کنیم، احتمالاً مبتنی بر چارچوب فهم‌پذیری هستند که در **توصیه ۴** مطرح شد.

۸. **معیارهای مورد توافقی برای قابل‌اعتمادبودن محصولات و خدمات AI توسعه دهید که یک سازمان جدید یا یک سازمان مناسب از پیش موجود عهده‌دار آن‌ها باشد. این معیارها برای سیستمی که ارزیابی کاربرـمحور همه پیشنهادهای AI عرضه‌شده در بازار را ممکن می‌سازد، همچون شالوده عمل می‌کنند.** به این ترتیب، می‌توان شاخصی برای AI قابل اعتماد ایجاد و آن را در کنار قیمت محصول اعلام کرد. این «شاخص مقایسه اعتماد» برای AI می‌تواند فهم عمومی را افزایش دهد و فضای رقابتی را پیرامون توسعه AI ایمن‌تر و از نظر اجتماعی مفیدتر (برای مثال «IwantgreatAI.org») شدت بخشد. در طولانی مدت، چنین سیستمی می‌تواند مبنایی برای یک سیستم گسترده‌تر گواهی محصولات و خدمات شایسته باشد؛ سیستم گسترده‌ای که مدیریتش را سازمانی که در اینجا به آن اشاره شد و/یا آژانس نظارتی که در **توصیه ۹** مطرح شده است بر عهده دارد. این سازمان می‌تواند از توسعه کدهای رفتاری نیز حمایت کند (نگاه کنید به **توصیه ۱۸**). به علاوه، می‌توان از کسانی که مالک سیستم‌های AI هستند و یا ورودی‌های سیستم‌های AI را مدیریت می‌کنند و از آن سود می‌برند خواست تا به

نفع خودشان در توسعه برنامه‌های سوادآموزی AI برای مصرف‌کنندگان سرمایه‌گذاری و/یا کمک کنند.

۹. یک آژانس نظارتی در اتحادیه اروپا *توسعه دهید* که از طریق ارزیابی و نظارت علمی بر محصولات، نرم‌افزارها، سیستم‌ها یا خدمات AI، مسئول حفاظت از رفاه عمومی باشد. این آژانس نظارتی می‌تواند برای مثال شبیه آژانس دارویی اروپا باشد. همچنین، باید یک سیستم نظارتی «پسا-عرضه»، برای مثال شبیه به سیستم نظارتی موجود برای داروها، برای محصولات AI ایجاد شود که در آن برخی ذی‌ربطان وظیفه داشته باشند تخلفات را گزارش دهند و همچنین سازوکاری برای کاربران دیگر وجود داشته باشد که بتوانند به راحتی تخلفات را گزارش دهند.

۱۰. رصدخانه‌ای اروپایی برای AI *ایجاد کنید*. وظیفه این رصدخانه نظارت بر پیشرفت‌ها، فراهم‌ساختن فضایی برای تبادل نظر و حمایت از مباحثه و هم‌اندیشی، ایجاد گنجینه‌ای برای ادبیات و نرم‌افزارهای AI (از جمله مفاهیم و لینک‌هایی به ادبیات موجود)، و ارائه توصیه‌ها و راهنمایی‌هایی گام به گام برای عمل خواهد بود.

۱۱. ابزارهایی قانونی و قالب‌هایی قراردادی ایجاد کنید تا برای همکاری روان و ثمربخش انسان-ماشین در محیط کار بنیادی فراهم آورد. شکل‌دادن به روایت «آینده کار» برای تصاحب «قلب‌ها و ذهن‌ها» راه‌گشاست. در راستای صندوق اروپایی تعدیل جهانی‌شدن[1]، می‌توان یک صندوق اروپایی تعدیل AI [2] همگام با «اروپایی که حفاظت می‌کند»، ایده «نوآوری فراگیر» و برای تسهیل گذار به انواع جدیدی از شغل‌ها تأسیس نمود.

[1] European Globalisation Adjustment Fund
[2] European AI Adjustment Fund

تشویق

۱۲. از توسعه و کاربرد تکنولوژی‌های AI در اتحادیه اروپا که از نظر اجتماعی ارجحیت دارند (نه اینکه صرفاً قابل‌قبول باشند) و با محیط زیست سازگارند (نه صرفاً قابل تحمل بلکه مطلوب محیط‌زیست هستند)، با تشویق‌های مالی در سطح اتحادیه اروپا حمایت کنید. این گزینه شامل بسط روش‌شناسی‌هایی خواهد بود که می‌توانند در ارزیابی این امر که آیا طرح‌های AI از نظر اجتماعی ارجحیت دارند و با محیط‌زیست سازگارند یا خیر کمک کنند. در این زمینه، اتخاذ «رویکرد چالش» (نگاه کنید به چالش‌های DARPA) می‌تواند خلاقیت را تشویق کند و موجب ارتقاء رقابت در توسعه راه‌حل‌های AI خاصی شود که از نظر اخلاقی صحیح و به نفع خیر عمومی‌اند.

۱۳. **از نظر مالی مشوق تلاش‌های تحقیقاتی پایدار، روبه‌رشد و منسجم اروپایی باشید که در خورِ ویژگی‌های خاص AI به عنوان یک حوزه علمی پژوهش‌اند.** این راهکار باید شامل مأموریتی شفاف برای پیش بردن AI در جهت خیر جامعه باشد و در مقام یک عامل توازن منحصربه‌فرد، با تمرکز کمتر بر فرصت‌های اجتماعی، در جهت گرایش‌های AI خدمت کند.

۱۴. **همکاری‌های میان‌رشته‌ای و میان‌گروهی و بحث‌های مربوط به نقاط تلاقی تکنولوژی، مسائل اجتماعی، مطالعات حقوقی و اخلاق را از نظر مالی تشویق کنید.** بحث‌های راجع به چالش‌های تکنولوژی ممکن است از پیشرفت‌های فنی واقعی عقب بمانند، اما اگر گروه‌های چنددیی‌ربطی متنوع به نحوی استراتژیک به این بحث‌ها شکل دهند آن‌ها می‌توانند نوآوری‌های تکنولوژی‌ها را به سمتی درست هدایت و حمایت کنند. اخلاق باید فرصت‌ها را مغتنم شمارد و با چالش‌ها دست و پنجه نرم کند، نه اینکه تنها توصیف‌شان کند. در این خصوص ضروری است که طراحی و توسعه AI بر اساس تنوع جنسیت، طبقه، قومیت، دیسیپلین و دیگر ابعاد مربوطه پیش برود تا درجه شمول، مدارا و غنای ایده‌ها و چشم‌اندازها را افزایش دهد.

۱۵. شمول ملاحظات اخلاقی، حقوقی و اجتماعی را در طرح‌های پژوهشی AI از نظر مالی *تشویق کنید*. همچنین به طور موازی، بررسی‌های منظمی را که بر روی قانون‌گذاری صورت می‌گیرند و هدفشان این است که ببینند این قانون‌گذاری‌ها تا چه حد به نوآوری‌های از

نظر اجتماعی مثبت پر و بال می‌دهند تشویق کنید. این دو معیار، در کنار یکدیگر کمک خواهند کرد تا اطمینان حاصل کنیم که اخلاق در قلب تکنولوژی AI است و این سیاست معطوف به نوآوری است.

۱۶. توسعه و استفاده از مناطق خاصی را که در محدوده اتحادیه اروپا برای آزمایش‌های تجربی و ایجاد سیستم‌های AI به طور قانونی مقررات‌زدایی شده‌اند از نظر مالی *تشویق کنید*. این مناطق ممکن است شکل یک «آزمایشگاه زنده»[1] (یا *Tokku*) را به خود بگیرند که بر مبنای تجربه «بزرگراه‌های آزمایش»[2] (و یا *Testreckon*) ساخته می‌شوند. آزمایش‌های جعبه شنی از این دست، علاوه بر محکم‌تر کردن پیوند بین نوآوری و سطح ریسک مرجع جامعه، در آموزش عملی و ارتقاء مسئولیت‌پذیری و شایستگی در مراحل ابتدایی مؤثرند. «حمایت از طریق طراحی» ویژگی ذاتی این نوع چارچوب است.

۱۷. پژوهش در مورد ادراک و فهم عمومی از AI و کاربردهای آن، و اجرای سازوکارهای مشورت عمومی ساخت‌یافته در جهت طراحی سیاست‌ها و قواعد مربوط به AI را به طور مالی *تشویق کنید*. این گزینه می‌تواند شامل استنباط مستقیم باور عمومی از طریق روش‌های پژوهشی سنتی، مثل رأی‌گیری و گروه‌های کانونی باشد، و همچنین شامل رویکردهای تجربی‌تری مثل فراهم کردن نمونه‌های شبیه‌سازی‌شده از دوگانه‌های اخلاقی موجود در سیستم‌های AI، و یا آزمایش‌هایی در آزمایشگاه‌های علوم اجتماعی. این دستور کار پژوهشی نباید تنها در خدمت ارزیابی باور عمومی باشد بلکه باید به هم‌آفرینی سیاست‌ها، استانداردها، بهترین عمل‌ها و قواعد نیز منجر شود.

---

[1] Living lab

[2] Test highways

حمایت

۱۸. از توسعه کدهای رفتاری خود-تنظیم‌کننده برای داده‌ها و حرفه‌های مرتبط با AI، با وظایف اخلاقی مشخص، *حمایت کنید*. این امر می‌تواند در راستای خط‌مشی دیگر حرفه‌های حساس از نظر اجتماعی، مثل پزشکان و یا وکلا باشد، مثلاً با گواهی همراهِ «AI اخلاقی» از طریق برچسب‌های اطمینان برای اطمینان حاصل کردن از اینکه مردم مزایای AI اخلاقی را می‌فهمند و بنابراین آن را از تأمین‌کنندگان طلب خواهند کرد. تکنیک‌های کنونی برای دخل و تصرف در توجه ممکن است از طریق این ابزارهای خود-تنظیم‌کننده محدود شوند.

۱۹. از ظرفیت هیئت‌مدیره شرکت‌های بزرگ برای به عهده گرفتن مسئولیت استلزامات اخلاقی تکنولوژی‌های AI شرکت‌ها *حمایت کنید*. برای مثال، این امر می‌تواند شامل آموزش تکمیلی هیئت‌مدیره‌های فعلی و توسعه بالقوه یک کمیته‌ی اخلاق با اختیارات حسابرسی درونی باشد. این راهکار را می‌توان برای ارزیابی طرح‌های اولیه و نحوه آرایش آن‌ها، با توجه به اصول بنیادین، در ساختار فعلیِ هیئت‌مدیره‌های یک لایه و دولایه، و/یا در کنار توسعه شکل الزامی «هیئت بررسی اخلاقی شرکت» توسط سازمان‌هایی که سیستم‌های AI را توسعه می‌دهند و یا از این سیستم‌ها استفاده می‌کنند، توسعه داد.

۲۰. از ایجاد برنامه‌های آموزشی و فعالیت‌های آگاهی‌بخش عمومی پیرامون تأثیرات اجتماعی، حقوقی و اخلاقی هوش مصنوعی *حمایت کنید*. این حمایت می‌تواند شامل موارد زیر باشد:

* برنامه‌هایی برای مدارسی که از قرار گرفتن علوم کامپیوتر در فهرست رشته‌های پایه‌ای که باید تدریس شوند حمایت می‌کنند؛
* ابتکار عمل‌ها و برنامه‌های صلاحیت در مشاغلی که با تکنولوژی AI سروکار دارند برای آموزش به کارکنان در مورد تأثیرات اجتماعی، حقوقی، و اخلاقی کار کردن با AI؛
* توصیه‌ای در سطح اروپا برای گنجاندن اخلاق و حقوق بشر در برنامه درسی رشته‌های علوم داده و AI و دیگر رشته‌های علمی و مهندسی که با سیستم‌های محاسباتی و AI سروکار دارند؛

* توسعه‌ی برنامه‌های مشابه در سطح وسیع برای عامه، با تمرکز خاص بر افرادی که در سطوح مختلف مدیریت تکنولوژی فعالیت دارند، از جمله کارمندان دولت، سیاست‌مداران و روزنامه‌نگاران؛
* شرکت در ابتکار عمل‌های گسترده‌تری مثل رخدادهای ITU AI for Good و تشکل‌های مدنی که بر روی اهداف توسعه پایدار سازمان ملل متحد کار می‌کنند.

نتیجه

اروپا، و به طور کلی جهان، با ظهور تکنولوژی‌هایی مواجه شده است که نویدهای هیجان‌انگیز بسیاری برای جنبه‌های وسیعی از حیات بشر در بر دارند و در عین حال به نظر می‌رسد که تهدیدها و خطرات بزرگی نیز با خود به همراه می‌آورند. هدف این اوراق سفید– و به‌ویژه توصیه‌های بخش قبلی– هدایت سکان به سوی پیامدهایی از توسعه، طراحی و آماده‌سازی تکنولوژی‌های AI است که از نظر اخلاقی و اجتماعی قابل ترجیح باشند. ما در راستای شناسایی فرصت‌ها و تهدیدهای AI برای جامعه و نیز مجموعه پنج اصل اخلاقی که برای هدایت آن در عمل معرفی کردیم، ۲۰ نکته عملی را صورتبندی کردیم که مبتنی بر روح همکاری و در جهت خلق پاسخ‌های *انضمامی* و *سازنده* به دشوارترین چالش‌های اجتماعی AI قرار دارند.

با وجود سرعت تغییر تکنولوژی، دیدن فرایند سیاسی در لیبرال دموکراسی‌های امروز به چشم امری از مُد افتاده و ناهماهنگ که دیگر در راستای حفظ ارزش‌ها و ارتقاء منافع جامعه و همه افراد آن نیست می‌تواند وسوسه‌کننده باشد. ما مخالفیم. ما با توصیه‌هایی که در اینجا ارائه کردیم، از جمله ایجاد مراکز، آژانس‌ها، برنامه‌های درسی، و زیرساخت‌های دیگر، از یک برنامه سیاست‌گذاری و نوآوری تکنولوژیک بلندپروازانه، جامع و منصفانه دفاع کرده‌ایم که باور داریم به تأمین مزایا و کاهش ریسک‌های AI، برای همه مردم، و برای جهانی که در آن شریکیم کمک می‌کند.

سپاسگزاری

نگارش این یادداشت بدون حمایت سخاوتمندانه Atomium - مؤسسه اروپایی علم، رسانه و دموکراسی، امکان‌پذیر نبود. ما به طور خاص از میکل‌آنجلو باراچی بونویچینی[1]، مدیر Atomium، از گویدو رومئو[2]، سردبیر آن، از کارمندان Atomium به خاطر کمک‌هایشان، و همه شرکای پروژه AI4People و اعضای اتاق بحث آن (http://www.eismd.eu/ai4people) به خاطر بازخوردهایشان سپاسگزاریم. مسئولیت محتوای این یادداشت و هر اشتباهی که در آن وجود داشته باشد تنها بر عهده نویسندگان است.

[1] Michelangelo BaracchiBonvicini

[2] Guido Romeo

منابع

[1] Asilomar AI Principles. (2017). *Principles developed in conjunction with the 2017 Asilomar conference [Benevolent AI 2017]*. Retrieved September 18, 2018 from https ://futur eofli fe.org/ai-princ iples .

[2] Cowls, J., & Floridi, L. (2018). *Prolegomena to a White Paper on Recommendations for the Ethics of AI (June 19, 2018)*. Available at SSRN: https ://ssrn.com/abstr act=31987 32.

[3] Cowls, J., & Floridi, L. (Forthcoming). *The Utility of a Principled Approach to AI Ethics*.

[4] European Group on Ethics in Science and New Technologies. (2018). *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems*. Retrieved September 18, 2018 from https ://ec.europ a.eu/info/news/ethic s-artificial -intel ligen ce-state ment-ege-relea sed-2018-apr-24_en.

[5] Floridi, L. (2013). *The ethics of information*. Oxford: Oxford University Press.

[6] Floridi, L. (2018). Soft ethics and the governance of the digital. *Philosophy & Technology, 31*(1), 1–8.

[7] House of Lords Artificial Intelligence Committee. (2018). *AI in the UK: ready, willing and able?* Retrieved September 18, 2018 from https ://publi catio ns.parli ament .uk/pa/ld201 719/ldsel ect/ldai/100/10002 .htm.

[8] Imperial College London. (2017). *Written Submission to House of Lords Select Committee on Artificial Intelligence [AIC0214]*. Retrieved September 18, 2018 from http://bit.ly/2yleu ET.

[9] King, T., Aggarwal, N., Taddeo, M., & Floridi, L. (2018). *Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions*. Available at SSRN: https ://ssrn.com/abstract=31832 38.

[10] Montreal Declaration for a Responsible Development of Artificial Intelligence. (2017). *Announced at the conclusion of the Forum on the Socially Responsible Development of AI*.

Retrieved September 18, 2018 from https ://www.montr ealde clara tion-respo nsibl eai.com/the-decla ration.
[11] Partnership on AI. (2018). *Tenets*. Retrieved September 18, 2018 from https ://www.partn ershi ponai .org/tenet s/.
[12] Taddeo, M. (2018). The limits of deterrence theory in cyberspace. *Philosophy & Technology*, *31*(3), 339-355
[13] The IEEE Initiative on Ethics of Autonomous and Intelligent Systems. (2017). *Ethically Aligned Design,v2*. Retrieved September 18, 2018 from https ://ethic sinac tion.ieee.org.

مشارکت‌کنندگان

Luciano Floridi (1,2) · Josh Cowls (1,2) · Monica Beltrametti (3) · Raja Chatila (4,5) ·Patrice Chazerand (6) · Virginia Dignum (7,8) · Christoph Luetge (9) · Robert Madelin (10) ·Ugo Pagallo (11) · Francesca Rossi (12,13) · Burkhard Schafer (14) · Peggy Valcke (15,16) ·Effy Vayena (17)

1. Oxford Internet Institute, University of Oxford, Oxford, UK
2. The Alan Turing Institute, London, UK
3. Naver Corporation, Grenoble, France
4. French National Center of Scientific Research, Paris, France
5. Institute of Intelligent Systems and Robotics, Pierre and Marie Curie University, Paris, France
6. Digital Europe, Brussels, Belgium
7. University of Umeо, Umeо, Sweden
8. Delft Design for Values Institute, Delft University of Technology, Delft, The Netherlands
9. TUM School of Governance, Technical University of Munich, Munich, Germany
10. Centre for Technology and Global Affairs, University of Oxford, Oxford, UK
11. Department of Law, University of Turin, Turin, Italy

12. IBM Research, New York, USA
13. University of Padova, Padua, Italy
14. University of Edinburgh Law School, Edinburgh, UK
15. Centre for IT & IP Law, Catholic University of Leuven, Flanders, Belgium
16. Bocconi University, Milan, Italy
17. Bioethics, Health Ethics and Policy Lab, ETH Zurich, Zurich, Switzerland

چارچوبی اخلاقی برای یک جامعه AI خوب

فرصت‌ها، ریسک‌ها، اصول و توصیه‌ها

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 خروشان‌تر می‌شود. اگر ما بر این رودخانه تدبیر کنیم و برنامه داشته باشیم، زه کشی 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.

۱۳۹۱/۷/۲

www.majazi.ir