

به نام خدا



وزارت علوم، تحقیقات، و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران

(ایرانداک)

خلاصه گزارش طرح پژوهشی سازمانی

عنوان طرح پژوهشی

شناسایی و تدوین اصول و رهنمودهای اخلاقی برای مواجهه با چالش‌های اخلاقی در توسعه هوش مصنوعی

تاریخ شروع طرح پژوهشی	۱۴۰۲/۰۷/۱۶	تاریخ پایان طرح پژوهشی	۱۴۰۴/۰۲/۱۶
نام و نام خانوادگی مجری طرح پژوهشی	علیرضا ثقه‌الاسلامی		
نام و نام خانوادگی همکاران طرح پژوهشی	ندارد		

درباره طرح پژوهشی

(بیشینه ۳۰۰ واژه درباره مقدمه، مسئله، و هدف‌های پژوهش و آنچه که انجام، بررسی، یا تعیین شده است.)

با گسترش و استفاده فراگیر از کامپیوتر و افزایش قدرت محاسباتی در ماشین‌های محاسبه‌گر، دهه ۱۹۵۰ تا ۱۹۷۰ را می‌توان سرآغاز «عصر طلایی هوش مصنوعی» تلقی کرد. از این رو، دولت‌ها و شرکت‌های خصوصی بزرگ سرمایه‌گذاری‌های کلانی را صرف تحقیقات نظری و کاربردی در زمینه هوش مصنوعی کردند. اما با گذر زمان، برخی دانشمندان استفاده از سیستم‌های خودمختار هوشمند و فراگیری استفاده از آن را تهدیدی جدی برای جامعه بشریت معرفی کردند. در ابتدا، خطرهایی نظیر اشتغال ربات‌ها به جای انسان‌ها، قدرتمند شدن قدرت‌های جهانی، به مخاطره افتادن ارزش‌های اخلاقی و اجتماعی با جانشینی سیستم‌های هوشمند به جای انسان‌ها و مواردی دیگر مطرح شدند. اما به مرور، طوفان احساسات افراطی و تفریطی درباره توسعه هوش مصنوعی و پیدایش چالش‌های اخلاقی اغراق‌آمیز فرو نشست، و نگاهی واقع‌بینانه و مستدل به توانایی‌ها و محدودیت‌های هوش مصنوعی آشکار گردید. امروزه این نگاه معتدل در قامت حوزه‌ای مطالعاتی، مسائل و چالش‌های اخلاقی مرتبط با هوش مصنوعی را از منظر اخلاق تکنولوژی صورتبندی می‌کند، آن‌ها را تحلیل و تبیین می‌نماید، و سرانجام به معرفی و تدوین رهنمودهایی اخلاقی برای توسعه هوش مصنوعی می‌پردازد.

هدف از این طرح پژوهشی شناسایی و تدوین اصول و رهنمودهای اخلاق هوش مصنوعی است. دانش‌افزایی این طرح پژوهشی در گردآوری موضوعات و مسائل اخلاقی، تدوین اصول اخلاقی منسجم، و در نهایت مقوله‌بندی و

پیشنهاد کدها و رهنمودهای اخلاقی در حوزه اخلاق هوش مصنوعی ذیل ذینفعان متعددی هم‌چون کاربران، طراحان، سیاست‌گزاران و دیگر کنشگران در این قلمرو است. برای دستیابی به این هدف، در چارچوب مطالعات کتابخانه‌ای و بررسی منابع متنوع، اسناد و دستورالعمل‌های بین‌المللی مرتبط با چالش‌های اخلاقی مطرح در هوش مصنوعی شناسایی و معرفی می‌شود. سپس، این اسناد و دستورالعمل‌ها با دو معیار روزآمد بودن و معتبر بودن سازمان‌های متولی گردآوری و مبتنی بر روش پژوهش‌های اسنادی بررسی و تحلیل می‌شود. و در پایان، اصول و رهنمودهایی اخلاقی برای مواجهه با چالش‌های اخلاقی هوش مصنوعی تدوین و معرفی می‌گردد.

روش پژوهش

(بیشینه ۳۰۰ واژه درباره گام‌های پژوهش و چگونگی گردآوری و تحلیل داده‌ها (ابزار، جامعه، نمونه، روش نمونه‌گیری، روش‌های تجزیه و تحلیل داده‌ها).

روش پژوهش این طرح پژوهشی در دو گام تحقق یافته است. در گام اول این پژوهش، ابتدا در چارچوب مطالعات کتابخانه‌ای، منابع متنوع، اسناد و دستورالعمل‌های بین‌المللی مرتبط با چالش‌های اخلاقی مطرح در هوش مصنوعی شناسایی و معرفی می‌شود. این اسناد و دستورالعمل‌ها با دو معیار روزآمد بودن و معتبر بودن سازمان‌های متولی گردآوری و بررسی می‌شوند.

در گام دوم و در چارچوب بررسی‌های تحلیلی در گام پیشین، مسائل و چالش‌های اخلاقی شاخص (و غیرقابل تقلیل به یکدیگر) و هم‌چنین شناسایی ذینفعان متعددی هم‌چون کاربران، طراحان، سیاست‌گزاران و دیگر کنشگران در حوزه هوش مصنوعی است. در ادامه، اصول اخلاقی حداقلی و منسجمی برای ارجاع و استناد کدها و رهنمودهای اخلاق هوش مصنوعی به این اصول تهیه و تدوین می‌گردد. و در پایان، ناظر به مسائل و چالش‌های اخلاقی در هوش مصنوعی و با استناد به اصول اخلاقی مدون، کدها و رهنمودهای اخلاقی ذیل ذینفعان و کنشگران در حوزه هوش مصنوعی مقوله‌بندی و پیشنهاد می‌شود. برای ارزیابی روایی یافته‌ها و دستاورد پژوهش، اصول و رهنمودهای اخلاقی پیشنهادی برای مواجهه با چالش‌های اخلاقی در هوش مصنوعی، توسط گروهی کانونی (Focus Group) از متخصصان هوش مصنوعی، فلسفه و اخلاق تکنولوژی و علم اطلاعات نقد و بررسی می‌شود. این بررسی‌ها تا زمان دستیابی به اجماعی نسبی میان متخصصان گروه کانونی ادامه و فرآیند اعتباربخشی به یافته‌های پژوهش تحقق می‌یابد.

یافته‌های طرح پژوهشی

(بیشینه ۵۰۰ واژه درباره یافته‌ها و دستاوردهای کلیدی پژوهش در پاسخ به پرسش‌ها یا بررسی فرضیه‌های پژوهش).

در مرحله اول گام دوم این پژوهش، رهنمودهای اخلاقی در بررسی هفت سند اخلاق هوش مصنوعی از طریق جست‌وجوی کلمات کلیدی یکسان و مشابه برای هر یک از اصول اخلاقی پنج‌گانه استخراج و گردآوری شدند. به عنوان مثال، برای اصل خیرخواهی، علاوه بر واژه کلیدی beneficence و مشتقات و مترادف‌های آن، و کلماتی

مانند dignity, wellbeing و sustainability نیز در تمامی این اسناد جست‌وجو شدند و رهنمودهایی که برخوردار از این واژه‌های کلیدی بود استخراج و گردآوری شدند. در مجموع ۱۳۰ رهنمود ذیل ۵ اصل اخلاقی از ۷ سند اخلاق هوش مصنوعی استخراج شدند. ۱۵ رهنمود برای اصل خیرخواهی، ۴۹ رهنمود برای اصل عدم آسیب‌رسانی، ۲۵ رهنمود برای اصل خودمختاری، ۲۸ رهنمود برای اصل عدالت، و ۲۸ رهنمود برای اصل توضیح‌پذیری از این اسناد گردآوری شدند.

در مرحله دوم گام دوم این پژوهش، رهنمودهای ویرایش، از نظر اصطلاحات یک‌دست‌سازی شد. هم‌چنین رهنمودهای تکراری حذف و رهنمودهای که از جامعیت کمتری نسبت به رهنمودهای دیگر برخوردار بودند حذف شدند. به این ترتیب، در مجموع ۶۷ رهنمود در قالب ۸ رهنمود برای اصل خیرخواهی، ۱۶ رهنمود برای اصل عدم آسیب‌رسانی، ۱۳ رهنمود برای اصل خودمختاری، ۱۴ رهنمود برای اصل عدالت، و ۱۶ رهنمود برای اصل توضیح‌پذیری حفظ شدند.

در مرحله سوم از گام دوم این پژوهش، رهنمودهای اخلاقی حاصل از مرحله دوم در گروهی کانونی بحث و بررسی گردید. این گروه شامل ۶ نفر متخصص در حوزه هوش مصنوعی و اخلاق حرفه‌ای (دو متخصص علم اطلاعات و دانش‌شناسی؛ یک متخصص فلسفه علم و تکنولوژی؛ یک متخصص مهندسی تکنولوژی اطلاعات؛ یک متخصص مهندسی صنایع؛ یک متخصص سیاست‌گذاری بر پایه مدل) برگزار شد. این نشست ۴ ساعت به طول انجامید و ۶۷ رهنمود اخلاقی مرحله دوم در چارچوب پنج اصل اخلاقی، بار دیگر الویت‌بخشی، اصلاح و حذف شدند. نتیجه آن‌که اصل اخلاقی خیرخواهی با ۷ رهنمود، اصل عدم آسیب‌رسانی با ۱۲ رهنمود، اصل خودمختاری با ۷ رهنمود، اصل اخلاقی عدالت با ۱۰ رهنمود، و اصل اخلاقی توضیح‌پذیری با ۱۰ رهنمود، ویرایش و صورت‌بندی مجدد گردید.

نتیجه‌گیری و پیشنهادها

(بیشینه ۳۰۰ واژه بحث و نتیجه‌گیری و بیان پیشنهادهای برگرفته از پژوهش).

در پایان این طرح پژوهشی دو نکته بسیار مهم، لازم به توضیح است: اول آن‌که گروه کانونی این طرح پژوهشی، رهنمودهای مورد بحث را با رویکرد اخلاق حرفه‌ای و تأکید بر قلمرو مخاطبان اصلی آن‌ها که شامل طراحان، متخصصان و توسعه‌دهندگان سیستم‌های هوش مصنوعی است، مبنای مشارکت در مباحث نقادانه خود قرار دادند. اما در صورت‌بندی نهایی رهنمودهای اخلاقی تلاش شده است، مخاطبان این رهنمودها، تمامی کشگران و عامل‌های انسانی که با سیستم‌های هوش مصنوعی تعامل دارند را هدف قرار می‌دهد. البته، بی‌تردید نقش طراحان، متخصصان و توسعه‌دهندگان سیستم‌های هوش مصنوعی برای توجه به رهنمودهای اخلاقی پیشنهادی به‌مراتب گسترده‌تر و مسئولانه‌تر است. نکته دوم آن‌که در بررسی‌های گروه کانونی اخلاق در هوش مصنوعی، اصل عدم

آسیب‌رسانی همچون بررسی اولیه هفت سند اخلاق هوش مصنوعی، از بیشترین رهنمودهای اخلاقی برخوردار شد. این موضوع، اهمیت عدم آسیب‌رسانیِ سیستم‌های هوش مصنوعی به جامعه انسانی و محیط زیست را نشان می‌دهد. از این رو، به نظر می‌رسد سیستم‌های هوش مصنوعی در طول چرخه عمر خود، و در فرآیند توسعه و استقرارشان در سرتاسر کره زمین، پیش و بیش از هر دستاوردی، باید عدم آسیب رساندن به محیط زیست و موجودات صاحب شعور سکونت‌یافته در این سیاره را هدف نهایی خود قرار دهند.