تأثیر هوش مصنوعی بر تحول هنجارهای اخلاقی: رویکردی مبتنی بر روان‌شناسی اخلاق

علی اکبرپور[1]
[1]دانشجوی کارشناسی ارشد روان‌شناسی تربیتی، دانشگاه تهران، Ali.akbarpoor@ut.ac.ir

چکیده

این پژوهش به بررسی نقش هوش مصنوعی در دگرگونی هنجارهای اخلاقی انسان از منظر روان‌شناسی اخلاق می‌پردازد. هدف مطالعه، تبیین چگونگی تأثیرگذاری سامانه‌های هوش مصنوعی، به‌ویژه الگوریتم‌های توصیه‌گر و پلتفرم‌های تعاملی، بر شکل‌گیری و بازتعریف ارزش‌هایی چون عدالت و همدلی است. در این پژوهش، با رویکردی نظری و مبتنی بر مرور انتقادی ادبیات، مفاهیم اصلی روان‌شناسی اخلاق و کارکردهای شناختی ـ عاطفی در قضاوت‌های اخلاقی تحلیل شده و با ویژگی‌های فناوری‌های نوین هوش مصنوعی مقایسه می‌شوند. نتایج تحلیل نشان می‌دهد که هدایت توجه و اولویت‌بندی محتوا توسط الگوریتم‌ها می‌تواند الگوهای قضاوت اخلاقی افراد را تغییر داده و در برخی موارد به کاهش حساسیت نسبت به پیامدهای انسانی تصمیم‌ها بینجامد. در عین حال، طراحی سامانه‌های هوش مصنوعی بر پایه اصول روان‌شناختی و توجه به سازوکارهای همدلی و انصاف، این ظرفیت را دارد که به تقویت هنجارهای اخلاقی همسو با ارزش‌های انسانی کمک کند. بر این اساس، ضرورت ادغام ملاحظات روان‌شناختی در فرایند توسعه و استقرار فناوری‌های هوش مصنوعی مورد تأکید قرار می‌گیرد.

کلید واژه : ارزش‌های انسانی، روان‌شناسی اخلاق، فناوری هوش مصنوعی، هنجارهای اخلاقی

۱- مقدمه :

در دهه‌های اخیر، هوش مصنوعی به یکی از تعیین‌کننده‌ترین عوامل تحول در زندگی فردی و اجتماعی انسان بدل شده است. از الگوریتم‌های توصیه‌گر در شبکه‌های اجتماعی و پلتفرم‌های محتوایی گرفته تا دستیارهای هوشمند و سیستم‌های تصمیم‌یار، بخش قابل توجهی از تجربه‌های روزمره ما در بستری شکل می‌گیرد که به‌طور نامرئی توسط سامانه‌های هوش مصنوعی سازمان‌دهی می‌شود. این تحول صرفاً جنبه فنی یا کارکردی ندارد، بلکه به‌تدریج نحوه ادراک ما از خود، دیگران و جهان اجتماعی را نیز دگرگون می‌کند. در چنین زمینه‌ای، یکی از پرسش‌های اساسی آن است که هوش مصنوعی چگونه می‌تواند بر هنجارهای اخلاقی انسان و شیوه‌های داوری درباره «درست» و «نادرست» اثر بگذارد.

اخلاق هوش مصنوعی عمدتاً در قالب مباحثی چون شفافیت الگوریتمی، مسئولیت‌پذیری، حریم خصوصی و تبعیض الگوریتمی پیگیری شده است؛ اما کمتر به این موضوع پرداخته شده که خودِ مواجهه مداوم با سامانه‌های هوش مصنوعی چگونه می‌تواند ساختار قضاوت‌های اخلاقی انسان و حساسیت او نسبت به ارزش‌هایی مانند عدالت و همدلی را تغییر دهد. به بیان دیگر، مسئله فقط این نیست که «سیستم اخلاقی باشد یا نه»، بلکه این است که «تعامل با سیستم، اخلاق ما را چگونه بازسازی می‌کند». این پرسش در حوزه روان‌شناسی اخلاق، که به سازوکارهای شناختی و عاطفی قضاوت‌های اخلاقی می‌پردازد، اهمیت ویژه‌ای دارد.

نظریه‌های کلاسیک و معاصر در روان‌شناسی اخلاق – از الگوهای رشد اخلاقی تا رویکردهای عاطفی-شناختی – نشان می‌دهند که قضاوت‌های اخلاقی برآیند تعامل میان فرآیندهای شناختی (مانند استدلال، توجه و تفسیر موقعیت) و فرآیندهای عاطفی (مانند همدلی، احساس گناه یا انزجار اخلاقی) هستند. در فضای دیجیتالِ متکی بر هوش مصنوعی، هر دو بُعد شناختی و عاطفی در معرض شکل‌دهی و هدایت الگوریتمی قرار می‌گیرند. الگوریتم‌های توصیه‌گر با اولویت‌بندی انواع خاصی از محتوا، می‌توانند بر آنچه می‌بینیم، به چه چیز حساس می‌شویم و چه نوع روایت‌هایی از عدالت و رنج انسانی در معرض دید ما قرار می‌گیرد، اثر بگذارند. این فرایند می‌تواند به مرور، مرزهای «طبیعی» و «پذیرفته‌شده» در قضاوت‌های اخلاقی را جابه‌جا کند.

از سوی دیگر، هوش مصنوعی صرفاً تهدیدی برای ارزش‌های اخلاقی به شمار نمی‌آید، بلکه می‌تواند به‌عنوان ابزاری برای تقویت هنجارهای اخلاقی انسانی نیز به‌کار گرفته شود؛ به شرط آن‌که طراحی آن بر پایه فهم دقیق‌تری از سازوکارهای روان‌شناختی اخلاق انجام گیرد. برای مثال، می‌توان سامانه‌هایی طراحی کرد که به‌طور هدفمند تعاملات همدلانه را تقویت کنند، تنوع دیدگاه‌ها را افزایش دهند و از تشدید حباب‌های اطلاعاتی و قطبی‌سازی اخلاقی جلوگیری نمایند.

با وجود این ظرفیت‌ها، شکاف محسوسی میان ادبیات اخلاق هوش مصنوعی و یافته‌های روان‌شناسی اخلاق وجود دارد. بسیاری از بحث‌های رایج در حوزه اخلاق فناوری، بیش از آن‌که بر فرایندهای روان‌شناختی تحول هنجارهای اخلاقی تمرکز کنند، معطوف به مقررات، اصول کلی یا پیامدهای اجتماعی هستند. این مقاله می‌کوشد با اتخاذ رویکردی مبتنی بر روان‌شناسی اخلاق، به این پرسش بپردازد که تعامل مستمر با سامانه‌های هوش مصنوعی چگونه می‌تواند به بازتعریف هنجارهای اخلاقی مرتبط با عدالت و همدلی بینجامد.

در این چارچوب، ابتدا به تبیین مفاهیم بنیادین در روان‌شناسی اخلاق و جایگاه عواطف و شناخت در قضاوت اخلاقی پرداخته می‌شود. سپس نقش سازوکارهای الگوریتمی در جهت‌دهی به ادراک‌ها و داوری‌های اخلاقی تحلیل می‌گردد و در ادامه، پیامدهای این تحلیل برای طراحی اخلاق‌محور سامانه‌های هوش مصنوعی مورد بحث قرار می‌گیرد. در نهایت، با جمع‌بندی یافته‌های نظری، بر ضرورت ادغام ملاحظات روان‌شناختی در توسعه و سیاست‌گذاری مرتبط با هوش مصنوعی برای حفظ و تقویت ارزش‌های انسانی تأکید می‌شود.

۲- ۲- روش تحقیق

این پژوهش از نوع مطالعه نظری ـ تحلیلی است و با رویکردی مبتنی بر مرور نظام‌مند مفاهیم و ادبیات علمی انجام شده است. هدف، تبیین سازوکارهای روان‌شناختی مؤثر در تحول هنجارهای اخلاقی در تعامل با سامانه‌های هوش مصنوعی است؛ بنابراین روش تحقیق بر تحلیل مفهومی و تطبیقی استوار است و شامل مراحل زیر بوده است:

۲-۱- جمع‌آوری ادبیات تخصصی

منابع معتبر حوزه روان‌شناسی اخلاق، نظریه‌های عاطفی-شناختی، مطالعات اخلاق فناوری و آثار پژوهشی مرتبط با تأثیر الگوریتم‌ها بر رفتار انسان بررسی شده‌اند. برای انتخاب منابع، از پایگاه‌هایی مانند *Google Scholar*، *Scopus* و *IEEE Xplore* استفاده شده است.

تحلیل نظری چارچوب‌های روان‌شناختی

نظریه‌های کلیدی از جمله رشد اخلاقی کالبرگ، نظریه بنیادهای اخلاقی هیدت، مدل‌های همدلی، و الگوهای شناختی قضاوت اخلاقی استخراج و مبانی مفهومی آنها دسته‌بندی شد.

مقایسه تطبیقی میان داده‌های روان‌شناختی و سازوکارهای هوش مصنوعی

نحوه عملکرد الگوریتم‌های توصیه‌گر، سیستم‌های تعاملی و سازوکارهای شخصی‌سازی محتوا تحلیل شد و با یافته‌های روان‌شناختی مرتبط با توجه، پردازش عاطفی، سوگیری‌های شناختی و قضاوت اخلاقی تطبیق داده شد.

استخراج الگوی اثرگذاری هوش مصنوعی بر هنجارهای اخلاقی

بر اساس تطبیق مفاهیم، یک چارچوب نظری پیشنهادی برای نشان دادن مسیرهای احتمالی تأثیرگذاری هوش مصنوعی بر ارزش‌هایی مثل عدالت و همدلی تدوین شد.

تحلیل پیامدها و پیشنهادات اخلاقی

نتایج تحلیلی در قالب پیامدهای نظری و کاربردی برای طراحی سامانه‌های اخلاق‌محور هوش مصنوعی ارائه شده است.

به دلیل ماهیت نظری پژوهش، ابزارهای میدانی و آماری به‌کار گرفته نشده و تمرکز اصلی بر تحلیل استدلالی، انسجام مفهومی و استنتاج نظری بوده است.

۳- ۳- مبانی نظری و پیشینه پژوهش

۳-۱- روان‌شناسی اخلاق و الگوهای کلاسیک قضاوت اخلاقی

روان‌شناسی اخلاق به مطالعه سازوکارهای شناختی و عاطفی دخیل در تصمیم‌گیری اخلاقی می‌پردازد. یکی از نظریه‌های بنیادین در این حوزه، نظریه رشد اخلاقی کالبرگ [۱] است که تحول قضاوت اخلاقی را در سه سطح پیش‌قراردادی، قراردادی و پس‌قراردادی توضیح می‌دهد. این مدل نشان می‌دهد که ظرفیت انسان برای داوری اخلاقی در تعامل با تجارب اجتماعی، آموزش و تکامل شناختی شکل می‌گیرد.

در کنار این رویکرد شناختی، دیدگاه‌های نوین‌تر بر تعامل عواطف و شناخت تأکید دارند. جاناتان هِیدت [۲] با ارائه نظریه بنیادهای اخلاقی [۳] بیان می‌کند که قضاوت‌های اخلاقی محصول واکنش‌های سریع عاطفی مانند همدلی، خشم و انزجار هستند و استدلال اخلاقی معمولاً در مرحله‌ای ثانویه وارد می‌شود. در این چارچوب، ارزش‌هایی مانند عدالت و مراقبت مستقیماً تحت تأثیر محیط اجتماعی و محرک‌های بیرونی قرار دارند.

[۱] Kohlberg, 1984

همچنین جاشوا گرین [۴] با مدل «دو فرایندی» خود نشان می‌دهد که تصمیم‌های اخلاقی در تعامل دو سیستم شکل می‌گیرند:

۱) سیستم سریع، هیجانی و شهودی

۲) سیستم کند، تحلیلی و مبتنی بر استدلال

این مدل به‌طور مستقیم قابل استفاده برای تحلیل نحوه تأثیر الگوریتم‌ها بر قضاوت‌های انسانی است؛ زیرا هوش مصنوعی اغلب سیستم سریع شهودی را تحریک می‌کند و فرصت پردازش تحلیلی را کاهش می‌دهد.

۳-۲- تغییر هنجارهای اخلاقی و نقش محیط اجتماعی

هنجارهای اخلاقی مجموعه باورهای مشترکی هستند که رفتار پذیرفته‌شده را در جامعه تعریف می‌کنند. پژوهش‌های الایر تورنر[۵] و شور و وینبرگر [۶]نشان می‌دهد که هنجارهای اخلاقی در اثر سازوکارهای فرهنگی، رسانه‌ای و ارتباطی تغییر می‌کنند. با گسترش رسانه‌های دیجیتال، زمینه‌های جدیدی برای شکل‌گیری ارزش‌ها پدید آمده است؛ به‌گونه‌ای که کانال‌های اطلاعاتی نقش مهمی در گسترش سوگیری‌های شناختی، کاهش تنوع دیدگاه‌ها و تغییر حساسیت اخلاقی ایفا می‌کنند.

۳-۳- نقش هوش مصنوعی و الگوریتم‌ها در شکل‌دهی ارزش‌های اخلاقی

مطالعات معاصر در حوزه اخلاق هوش مصنوعی نشان می‌دهد که سامانه‌های هوشمند تنها ابزارهای فنی نیستند، بلکه بخشی از محیط اجتماعی-شناختی انسان محسوب می‌شوند.

لوتجیانو فلوریدی [۷]به‌عنوان یکی از برجسته‌ترین نظریه‌پردازان «اخلاق اطلاعات»، معتقد است که فناوری‌های هوشمند «محیط اخلاقی» را بازسازی می‌کنند و بر نحوه درک انسان از مسئولیت، عدالت و هنجارهای اجتماعی اثر می‌گذارند.

همچنین کیت کراوفورد [۸] همکاران (در حوزه سوگیری الگوریتمی) نشان داده‌اند که الگوریتم‌های توصیه‌گر با ایجاد «حباب‌های اطلاعاتی» و شخصی‌سازی شدید محتوا می‌توانند حساسیت اخلاقی افراد را تغییر دهند. این سامانه‌ها با هدایت توجه به انواع خاصی از محتوا، شیوه مواجهه افراد با رنج انسانی، عدالت اجتماعی و رفتار اخلاقی را دست‌کاری می‌کنند.

در حوزه طراحی سیستم‌های اخلاق‌محور، وِندل والاچ [۹]و پیتر آلن [۱۰]چارچوب‌هایی مطرح کرده‌اند که بر لزوم ادغام اصول اخلاقی و روان‌شناختی در طراحی سامانه‌ها تأکید دارد. این پژوهش‌ها نشان می‌دهند که هوش مصنوعی می‌تواند هم هنجارهای اخلاقی را تضعیف کند و هم در شرایط طراحی‌شده و نظارت‌شده، آن‌ها را تقویت نماید.

۳-۴- جمع‌بندی پیشینه علمی

ادبیات علمی نشان می‌دهد که:

کالبرگ، هیدت و گرین سازوکارهای روان‌شناختی اخلاق را توضیح داده‌اند.

فلوریدی، کراوفورد و ونکاتاسوبرامانیان نقش فناوری را در تغییر هنجارهای اخلاقی بررسی کرده‌اند.

اما ترکیب این دو حوزه و تحلیل روان‌شناختی اثر الگوریتم‌ها بر هنجارهای اخلاقی هنوز محدود است.

---

[۴] Haidt, 2001

[۵] Moral Foundations Theory

[۶] Greene, 2013

[۷] Turner, 1991

[۸] Shweder & Weinberger, 1992

[۹] Luciano Floridi

[۱۰] Kate Crawford, 2021

[۱۱] Wendell Wallach

[۱۲] Peter Allen

از این‌رو، پژوهش حاضر تلاش می‌کند با ادغام این دو سطح نظری، سازوکاری را نشان دهد که طی آن هوش مصنوعی می‌تواند به بازتعریف ارزش‌هایی مانند عدالت و همدلی منجر شود.

۴- بحث و تحلیل

تحلیل تحول هنجارهای اخلاقی در عصر هوش مصنوعی مستلزم ترکیب دو حوزه نظری است:

۱) روان‌شناسی اخلاق (کالبرگ، هیدت، گرین)

۲) اخلاق فناوری و هوش مصنوعی (فلوریدی، کراوفورد، ونکاتاسوبرامانیان، والاچ)

از این منظر، تعامل انسان با سامانه‌های هوشمند تنها یک کنش تکنیکی نیست، بلکه فرایندی است که بر ساختارهای شناختی و عاطفی قضاوت اخلاقی اثر می‌گذارد.

۴-۱- تأثیر الگوریتم‌ها بر سازوکارهای شناختی اخلاق

مطابق نظریه رشد اخلاقی کالبرگ، قضاوت اخلاقی انسان در سطوحی از استدلال مبتنی بر اصول شکل می‌گیرد. اما سامانه‌های توصیه‌گر با ایجاد محیط‌های شناختی محدود و شخصی‌سازی‌شده، ظرفیت افراد برای پردازش تحلیلی و استدلال اخلاقی را کاهش می‌دهند.

به تعبیر جاشوا گرین (۲۰۱۳)، تصمیم‌های اخلاقی نیازمند تعادل میان سیستم هیجانی سریع و سیستم کند شناختی هستند؛ اما محیط الگوریتمی رسانه‌ها با تسریع جریان محرک‌های هیجانی و کوتاه‌کردن فرصت پردازش، باعث تقویت سیستم سریع و تضعیف سیستم کند می‌شود. این وضعیت می‌تواند به قضاوت‌های سطحی‌تر، واکنش‌های احساسی شدید و کاهش عمق استدلال اخلاقی منجر شود.

۴-۲- اثر هوش مصنوعی بر حساسیت‌های عاطفی اخلاقی

بر اساس نظریه بنیادهای اخلاقی هیدت، ارزش‌هایی همچون همدلی، مراقبت و عدالت ریشه‌های عاطفی دارند. الگوریتم‌های مبتنی بر اولویت‌بندی محتوای هیجانی می‌توانند الگوهای عاطفی را بازتعریف کنند.

کیت کراوفورد (۲۰۲۱) نشان می‌دهد که پلتفرم‌ها تمایل دارند محتواهای برانگیزاننده‌ٔ خشم یا انزجار را بیشتر نمایش دهند؛ زیرا این واکنش‌ها تعامل بیشتری تولید می‌کنند.

این امر به‌تدریج حساسیت نسبت به رنج انسانی را کاهش می‌دهد و واکنش‌های همدلانه (که زیربنای اخلاق مراقبت‌اند) را تضعیف می‌کند.

همچنین تحلیل‌های سورِش ونکاتاسوبرامانیان در حوزه سوگیری الگوریتمی نشان می‌دهد که فناوری‌های توصیه‌گر با تقویت کلیشه‌های اخلاقی یا اجتماعی ممکن است «چارچوب‌های عاطفی» افراد را در جهت‌گیری‌های خاص ثابت نگه دارند. این فرایند می‌تواند در بلندمدت به کاهش انعطاف اخلاقی و کاهش توانایی انسان در درک دیدگاه‌های متفاوت منجر شود.

۴-۳- بازتعریف هنجارهای عدالت در محیط‌های هوشمند

عدالت یکی از بنیادی‌ترین ارزش‌های اخلاقی در نظریه‌های کالبرگ و هیدت است. در محیط‌های دیجیتال، نوع مواجهه افراد با مسائل عدالت‌محور تحت‌تأثیر نحوه توزیع اطلاعات و تجربه‌های شخصی‌سازی‌شده قرار می‌گیرد.

لوسانو فلوریدی اشاره می‌کند که سامانه‌های هوش مصنوعی «معماری اخلاقی محیط» را تغییر می‌دهند؛ یعنی تعیین می‌کنند که افراد چه چیزی را ببینند و چه چیزی خارج از میدان توجه آن‌ها قرار گیرد. این بازطراحی محیطی به بازطراحی هنجارهای عدالت نیز می‌انجامد.

۴-۴- ظرفیت‌های هوش مصنوعی برای تقویت هنجارهای اخلاقی

در کنار پیامدهای چالش‌برانگیز، پژوهشگران اخلاق فناوری مانند وِندل والاچ و پیتر آلن بر پتانسیل هوش مصنوعی برای تقویت ارزش‌های اخلاقی تأکید کرده‌اند.

به شرط طراحی صحیح، سامانه‌های هوشمند می‌توانند:

تنوع دیدگاه‌های اخلاقی را افزایش دهند،

محتواهای همدلانه را تقویت کنند،

سوگیری‌ها را کاهش دهند،

کاربران را در تصمیم‌گیری‌های اخلاقی یاری دهند.

برای نمونه، طراحی سامانه‌هایی که مداخلات مبتنی بر همدلی ارائه می‌کنند یا محتواهای اخلاقی را با دقت بیشتری برجسته می‌کنند، می‌تواند هنجارهای عدالت و مراقبت را تقویت نماید.

۵-۴- جمع‌بندی تحلیل

تحلیل حاضر نشان می‌دهد که هوش مصنوعی دو سطح از سازوکارهای اخلاقی انسان را درگیر می‌کند:

سطح شناختی: جهت‌دهی به توجه، کاهش استدلال اخلاقی، تقویت سیستم سریع تصمیم‌گیری

سطح عاطفی: تغییر حساسیت‌های همدلانه، تقویت واکنش‌های هیجانی، دست‌کاری قالب‌های عاطفی

نتیجه این فرایند، بازتعریف تدریجی هنجارهای اخلاقی است؛ به این معنا که عدالت، همدلی و مسئولیت‌پذیری در بستری شکل می‌گیرند که توسط الگوریتم‌ها طراحی و هدایت می‌شود.

۵- نتیجه‌گیری

یافته‌های این پژوهش نشان می‌دهد که هوش مصنوعی نه‌فقط به‌عنوان یک ابزار فناوری، بلکه به‌عنوان یک بافت *(context)* روان‌شناختی جدید عمل می‌کند که در آن، فرایندهای رشد اخلاقی، قضاوت اخلاقی و هویت اخلاقی افراد شکل می‌گیرند و دگرگون می‌شوند. از منظر روان‌شناسی اخلاق، تعامل مداوم با سامانه‌های هوشمند می‌تواند سطوح مختلف تجربه اخلاقی را تحت‌تأثیر قرار دهد:

در سطح شناختی، الگوریتم‌های توصیه‌گر و محیط‌های دیجیتال شخصی‌سازی‌شده، الگوهای توجه، تفسیر موقعیت‌های اجتماعی و دسترسی به اطلاعات را تغییر می‌دهند. این تغییرات می‌تواند مسیر طی‌شده از قضاوت‌های پیش‌قراردادی تا پس‌قراردادی را، آن‌گونه که کالبرگ توصیف می‌کند، دچار دگرگونی کند؛ زیرا فرد بخش قابل‌توجهی از مواجهه‌های اخلاقی خود را در فضایی تجربه می‌کند که توسط منطق الگوریتمی فیلتر شده است. در نتیجه، فرصت مواجهه با تعارض‌های اخلاقی پیچیده و تمرین استدلال اصول‌محور کاهش یافته و قضاوت‌های اخلاقی بیش از پیش به الگوهای تثبیت‌شده و گروه‌محور وابسته می‌شود.

در سطح عاطفی–شهودی، نظریه بنیادهای اخلاقی هیدت به ما یادآور می‌شود که ارزش‌هایی مثل مراقبت و انصاف ریشه در واکنش‌های عاطفی و شهودی دارند. محیط‌های مبتنی بر هوش مصنوعی، از طریق برجسته‌سازی مکرر برخی محرک‌های هیجانی (مانند خشم، ترس یا تحقیر) و کم‌رنگ‌سازی موقعیت‌های برانگیزاننده همدلی، به‌تدریج «پروفایل عاطفی اخلاقی» فرد را بازتنظیم می‌کنند. این فرایند می‌تواند به کاهش حساسیت همدلانه، عادی‌شدن رنج دیگران و تقویت قضاوت‌های قطبی‌شده بینجامد؛ پدیده‌ای که برای سلامت روان اجتماعی و انسجام جمعی مخاطره‌آمیز است.

در سطح هویت اخلاقی و خودآگاهی، زندگی در بافتی که واکنش‌ها، ترجیحات و قضاوت‌های فرد مدام توسط سامانه‌های هوشمند بازتاب و تقویت می‌شوند، می‌تواند تجربه فرد از «منِ اخلاقی» را تغییر دهد. فرد ممکن است کمتر در معرض بازخوردهای چالش‌برانگیز قرار گیرد، کمتر نیاز به بازنگری خودانتقادی در باورهای اخلاقی‌اش احساس کند و در نتیجه، فرایندهای رشد هویت اخلاقی کند یا منحرف شود. برای روان‌شناسان تربیتی، مشاوران و درمانگران، این نکته پیامدهای مستقیم دارد؛ چراکه مداخلات اخلاقی و آموزشی دیگر نمی‌توانند مستقل از بافت هوش مصنوعی طراحی شوند.

در عین حال، نتایج این مطالعه نشان می‌دهد که همین سازوکارها می‌توانند به‌صورت معکوس به خدمت تقویت اخلاق انسانی نیز درآیند .اگر طراحان سامانه‌های هوش مصنوعی با تکیه بر یافته‌های روان‌شناسی اخلاق عمل کنند، می‌توان محیط‌هایی ساخت که:

- مواجهه با دیدگاه‌های متنوع اخلاقی را تسهیل کند،
- موقعیت‌های برانگیزاننده همدلی را برجسته سازد،
- فرصت تمرین استدلال اخلاقی را افزایش دهد،
- و رشد هویت اخلاقی خودآگاه و مسئول را حمایت کند.

بنابراین، این پژوهش بر ضرورت شکل‌گیری گفت‌وگویی فعال بین روان‌شناسی و علوم داده تأکید می‌کند. برای حفظ و تقویت سلامت اخلاقی فرد و جامعه در عصر هوش مصنوعی، کافی نیست که صرفاً درباره «اخلاقی بودن» یا «نبودن» فناوری صحبت کنیم؛ بلکه باید به‌طور مشخص بررسی کنیم که این فناوری‌ها چگونه بر پردازش‌های شناختی، الگوهای هیجانی، هویت اخلاقی و روابط میان‌فردی اثر می‌گذارند و بر اساس این شناخت، هم در سطح طراحی سامانه‌ها و هم در سطح مداخلات آموزشی و درمانی، رویکردهای جدیدی توسعه دهیم.

مراجع سبک *(APA)*

Greene, J. D. (2013). Moral tribes: Emotion, reason, and the gap between us and them. Penguin Press.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. Psychological Review, 108(4), 814–834.

Kohlberg, L. (1984). The psychology of moral development: The nature and validity of moral stages. Harper & Row.

Crawford, K. (2021). The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

Floridi, L. (2013). The ethics of information. Oxford University Press.

Venkatasubramanian, S. (2019). The social consequences of algorithmic bias. Annual Review of Statistics and Its Application, 6, 29–43.

Wallach, W., & Allen, C. (2009). Moral machines: Teaching robots right from wrong. Oxford University Press.