فصلنامه علمی تخصصی انگاره های نو در تحقیقات آموزشی؛ سال پنجم ، شماره ۱۵،صص ۱۳-۲۵
شاپای الکترونیکی:۹۱۶۹-۲۹۸۰

رابطه بین هویت اخلاقی و سواد هوش مصنوعی: پیامدهایی برای تعامل اخلاقی در عصر هوشمند سازی

طیبه احمدی[۱]

تاریخ پذیرش: ۱۴۰۵/۰۲/۱۰ تاریخ انتشار: ۱۴۰۵/۰۳/۰۵

چکیده

مقدمه و هدف: با گسترش هوش مصنوعی، درک تعامل اخلاقی کاربران با این فناوری اهمیت یافته است. این پژوهش با تأکید بر نقش هویت اخلاقی، به بررسی رابطه آن با سواد هوش مصنوعی و تأثیر این دو بر رفتار اخلاقی در استفاده از سیستم‌های هوشمند می‌پردازد.

روش‌شناسی پژوهش: پژوهش حاضر به روش کتابخانهای با رویکرد توصیفی-تحلیلی و با استفاده از منابع روانشناسی اخلاق، اخلاق هوش مصنوعی، چارچوب های اخلاق اسلامی و رهنمودهای کاربردی جامعه اطلاعاتی انجام شده است.

یافته ها: مطالعه منابع نشان میدهد که هویت اخلاقی (شامل درونی سازی صفاتی مانند انصاف، صداقت و مسئولیتپذیری) نقش واسطه ای و تعدیلگر در نحوه تفسیر خروجیهای هوش مصنوعی، ارزیابی ریسکهای اخلاقی و تصمیمگیری دارد. در حالی که سیستمهای هوش مصنوعی میتوانند استدلال اخلاقی را شبیهسازی کنند، فاقد عاملیت اخلاقی اصیل، آگاهی، قصد (نیت) و مسئولیت پذیری هستند. مدلهای معاصر هوش اخلاقی-اخلاقی بر این نکته تأکید دارند که سواد هوش مصنوعی اثربخش باید سه بعد شایستگی فنی، استدلال اخلاقی و آگاهی معنوی-اخلاقی را تلفیق کند. همچنین، چارچوبهایی مانند توصیه نامه یونسکو، اخلاق هوش مصنوعی جامعه اطلاعاتی ایالات متحده و اصول اخلاقی اسلامی (از جمله خلافت، امانت، عدل و نیت) همگی بر ضرورت پاسخگویی انسانی، شفافیت و عدالت در کاربست هوش مصنوعی تأکید دارند.

نتیجه گیری: تحول آموزش در عصر دیجیتال نیازمند پرورش هویت اخلاقی قوی و سواد جامع هوش مصنوعی است که از طریق رویکردهای تلفیقی، ارزیابی مستمر و تقویت مسئولیت‌پذیری محقق شده و به تعامل اخلاقی و حفظ عاملیت انسانی می‌انجامد.

واژه های کلیدی: هویت اخلاقی، سواد هوش مصنوعی، پاسخگویی

[۱] کارشناسی ارشدمشاوره، آموزش و پرورش ناحیه ۱شیراز Ahmadita1396@gmail.com

مقدمه

هوش مصنوعی به نیرویی فراگیر در تصمیم‌گیری در حوزه‌هایی مانند مراقبت‌های بهداشتی، آموزش، حکمرانی و زندگی شخصی تبدیل شده است. سیستم‌های معاصر هوش مصنوعی وظایفی را انجام می‌دهند که به‌طور سنتی با شناخت انسان مرتبط است، از جمله استدلال، پیش‌بینی و تعامل اجتماعی (راسل[1] ۲۰۱۰). با ورود هوش مصنوعی به حوزه‌های حساس اخلاقی مانند قضا و پزشکی، نگرانی دربارهی درستی و اخلاقی بودن کاربرد آن به یک مسئلهی اصلی در دانشگاه و جامعه تبدیل شده است.(باستروم و یودکوسکی[2] ۲۰۱۸).

به موازات این تحول فناورانه، اهمیت سواد هوش مصنوعی رو به افزایش است. سواد هوش مصنوعی به‌عنوان توانایی درک، ارزیابی انتقادی و استفادهی مسئولانه از سیستم‌های هوش مصنوعی تعریف می‌شود. توصیه‌نامه ی یونسکو در مورد اخلاق هوش مصنوعی[3](۲۰۲۱) تأکید می‌کند که سواد هوش مصنوعی باید فراتر از شایستگی فنی رفته و آگاهی اخلاقی، ملاحظات حقوق بشر و ظرفیت پرسش‌گری از خروجی‌های الگوریتمی را در زمینه‌های اجتماعی، فرهنگی و اخلاقی گسترده‌تر در بر گیرد. به‌طور مشابه، چارچوب اخلاق هوش مصنوعی جامعهٔ اطلاعاتی ایالات متحده[4](۲۰۲۰) تأکید می‌کند که استفادهی اخلاقی از هوش مصنوعی نه تنها به آزمایش و شفافیت، بلکه به قضاوت و پاسخگویی انسانی در تمام مراحل چرخهی حیات یک سیستم هوش مصنوعی نیاز دارد.

سواد هوش مصنوعی، اگرچه برای تعامل مؤثر با سیستم‌های هوشمند ضروری است، اما به تنهایی نمی‌تواند تنوع رفتارهای اخلاقی کاربران را تبیین کند. در این میان، عامل مهم و در عین حال کمتر مورد توجه قرار گرفته، «هویت اخلاقی» است. هویت اخلاقی به درجه‌ای اشاره دارد که اخلاقی بودن برای خودپنداره و هویت یک فرد مرکزی محسوب می‌شود. افرادی که هویت اخلاقی قوی دارند، خود را دارای صفاتی مانند انصاف، صداقت و مسئولیت‌پذیری می‌دانند؛ از این رو، هنگام مواجهه با خروجی‌های مغرضانه یا ناعادلانه هوش مصنوعی، به شناسایی و نقد آن‌ها می‌پردازند. در مقابل، افراد با هویت اخلاقی ضعیف‌تر، تمایل بیشتری به پذیرش بدون چون‌وچرای خروجی‌های الگوریتمی دارند، زیرا ارزش‌های اخلاقی نقش پررنگی در خودپنداره آن‌ها ایفا نمی‌کند. بنابراین، برای دستیابی به تعامل اخلاقی اثربخش با سیستم‌های هوشمند، صرفاً تقویت سواد فنی کافی نیست، بلکه پرورش همزمان هویت اخلاقی کاربران ضروری است. (آکوینو و رید[5] ۲۰۰۲). هویت اخلاقی که ریشه در روان‌شناسی اخلاق دارد، به درونی‌سازی صفات اخلاقی مانند انصاف، شفقت، صداقت و مسئولیت‌پذیری در خودتعریف‌گری فرد اشاره دارد. افرادی با هویت اخلاقی قوی به احتمال بیشتری در موقعیت‌های پیچیده یا مبهم به‌طور پیوسته بر اساس اصول اخلاقی عمل می‌کنند

[1] Russell
[2] Bostrom & Yudkowsky
[3] Recommendation on the ethics of artificial intelligence
[4] United States Intelligence Community
[5] Aquino & Reed

(کلبرگ[1] ۱۹۸۴؛ رست[2] ۱۹۸۶). هنگام تعامل با هوش مصنوعی، این افراد به احتمال بیشتری نسبت به خروجی‌های مغرضانه نقادانه برخورد می‌کنند، در مورد ریسک‌های اخلاقی مانند نظارت یا دستکاری محتاط‌تر هستند و در مورد پیامدهای تصمیمات مبتنی بر هوش مصنوعی تأمل بیشتری دارند.

تحولات اخیر در پژوهش هوش مصنوعی این رابطه را پیچیده‌تر کرده است. مدل‌های زبانی بزرگ و دیگر سیستم‌های هوش مصنوعی مولد به‌طور فزاینده‌ای از سوی کاربران به‌عنوان دارا بودن تخصص اخلاقی تلقی می‌شوند و گاهی در اعتمادپذیری درک‌شده و انسجام استدلالی از متخصصان اخلاق انسانی پیشی می‌گیرند (دیلون و دیگران[3] ۲۰۲۵). این درک سؤالات مهمی را مطرح می‌کند: کاربران چگونه توصیه‌های اخلاقی تولیدشده توسط هوش مصنوعی را تفسیر می‌کنند؟ آیا افراد با هویت اخلاقی قوی‌تر به‌گونه‌ای متفاوت با سیستم‌های هوش مصنوعی تعامل می‌کنند؟ و سواد هوش مصنوعی که به‌عنوان شایستگی فنی و اخلاقی درک می‌شود چگونه این تعامل را شکل می‌دهد؟

مفهوم هوش اخلاقی چارچوب مفیدی برای بررسی این سؤالات فراهم می‌کند. چارچوب هوش اخلاقی شامل ابعادی مانند تخصص اخلاقی، حساسیت، انسجام، شفافیت و پاسخگویی است که می‌تواند برای عامل‌های انسانی و مصنوعی به کار رود (استوپو و دیگران[4] ۲۰۲۲؛ هابارد، کید و استوپو[5] ۲۰۲۵). یافته‌های تجربی نشان می‌دهد که اگرچه مدل‌های معاصر هوش مصنوعی می‌توانند جنبه‌هایی از استدلال اخلاقی را شبیه‌سازی کنند، اما فاقد عاملیت اخلاقی اصیل، درگیری عاطفی، هویت خودبنیانِ تأملی و از همه مهمتر ظرفیت قصد اخلاقی و آگاهی معنوی‌اند که مشخصهی آگاهی اخلاقی انسان است (محمود و دیگران[6] ۲۰۲۵). این تمایز بر اهمیت هویت اخلاقی انسان در تفسیر، هدایت و مسئولیت‌پذیری در قبال استفاده از هوش مصنوعی تأکید می‌کند.علاوه بر این، نگرانی‌هایی مانند سوگیری الگوریتمی، تصمیم‌گیری مبهم و دستکاری رفتاری، محدودیت‌های اساسی سیستم‌های هوش مصنوعی را برجسته می‌سازند. هوش مصنوعی می‌تواند تبعیض‌های اجتماعی را از طریق داده‌های آموزش مغرضانه تقویت کند، رفتار کاربر را از طریق الگوریتم‌های شخصی‌سازی هدفمند دستکاری نماید و به دلیل ماهیت پیچیده، خودمختار یا غیرشفاف خود، خطوط پاسخگویی را مبهم سازد (اونیل[7] ۲۰۱۶). این چالش‌ها کاربرانی را می‌طلبد که نه تنها از نظر فنی باسواد باشند، بلکه از نظر اخلاقی نیز ریشه‌دار باشند ـ کسانی که بتوانند تشخیص دهند چه زمانی یک خروجی هوش مصنوعی اصول عدالت، امانت یا رفاه عمومی را نقض می‌کند، همان‌طور که هم در اندیشهٔ اخلاقی اسلامی و هم در اخلاق سکولار هوش مصنوعی بیان شده است (غزالی، ۱۹۸۳؛ باستروم و یودکوسکی، ۲۰۱۴؛ چارچوب اخلاق هوش مصنوعی جامعهٔ اطلاعاتی ایالات متحده، ۲۰۲۰).

---

[1] Kohlberg
[2] Rest
[3] Dillon, et al
[4] Stupu et al
[5] Hubbard et al
[6] Mahmood
[7] O’Neil

علیرغم گسترش پژوهش در مورد اخلاق و سواد هوش مصنوعی، تقاطع بین هویت اخلاقی و سواد هوش مصنوعی هنوز به‌طور کافی بررسی نشده است. بنابراین، این مقاله با هدف پاسخ به پرسش پژوهشی زیر تدوین شده است:

هویت اخلاقی چگونه بر سواد هوش مصنوعی و تعامل اخلاقی با سیستم‌های هوش مصنوعی تأثیر می‌گذارد؟این مقاله یافته‌های روان‌شناسی اخلاق، اخلاق هوش مصنوعی، مطالعات تجربی اخیر در مورد هوش اخلاقی در سیستم‌های هوش مصنوعی و چارچوب‌های هنجاری برگرفته از اندیشهٔ اخلاقی اسلامی و همچنین حکمرانی غربی هوش مصنوعی را ترکیب می‌کند (باستروم و یودکوسکی، ۲۰۱۴؛ محمود و دیگران، ۲۰۲۵).

روش شناسی پژوهش

در این پژوهش از یک رویکرد مرور تلفیقی یعنی توصیفی- تحلیلی استفاده شد. منابع بر اساس ارتباط آن‌ها با هویت اخلاقی، سواد هوش مصنوعی و حکمرانی اخلاقی انتخاب شدند.

یافته ها

هویت اخلاقی به‌عنوان پایه‌ای برای تعامل اخلاقی با هوش مصنوعی

هویت اخلاقی به‌طور معناداری بر نحوه‌ی تفسیر و پاسخ افراد به چالش‌های اخلاقی مربوط به هوش مصنوعی تأثیر می‌گذارد. افراد با هویت اخلاقی قوی به احتمال بیشتری ملاحظات اخلاقی مانند انصاف، حریم خصوصی و عدم آسیب‌رسانی را بر راحتی، کارایی یا مرجعیت الگوریتمی ترجیح می‌دهند. در زمینه‌ی هوش مصنوعی، این امر به بررسی دقیق‌تر خروجی‌های هوش مصنوعی، آگاهی بیشتر از ریسک‌های اخلاقی و گرایش قوی‌تر به مطالبه‌ی پاسخگویی انسانی ترجمه می‌شود.پژوهش در روان‌شناسی اخلاق به‌طور مداوم نشان می‌دهد که هویت اخلاقی، رفتار نوع‌دوستانه و تصمیم‌گیری اخلاقی را در طیف وسیعی از حوزه‌ها هدایت می‌کند (آکوینو و رید[1] ۲۰۰۲؛ رست، ۱۹۸۶). اخلاق هنگامی که در مورد استفاده از هوش مصنوعی به کار گرفته شود، این بدان معناست که افراد مبتنی بر اخلاق، به احتمال کمتری توصیه‌های هوش مصنوعی را کورکورانه باور می‌کنند، به‌ویژه در موقعیت‌های اخلاقی سنگین مانند استخدام، پلیس، دسته‌بندی پزشکی یا تعدیل محتوا. برای مثال، در سناریوهای تصمیم‌گیری که شامل انصاف یا سوگیری نژادی/جنسیتی هستند، افراد با هویت اخلاقی بالا به احتمال بیشتری نتایج تبعیض‌آمیز را شناسایی و به چالش می‌کشند، حتی زمانی که آن نتایج توسط یک سیستم هوش مصنوعی به ظاهر معتبر تولید شده باشند (اونیل، ۲۰۱۶).

از دیدگاه اخلاق اسلامی، این ریشه‌دار بودن اخلاقی صرفاً روان‌شناختی نیست، بلکه هستی‌شناختی

[1] Aquino & Reed

است. انسان به‌عنوان خلیفه (جانشین زمین) توصیف شده که به امانت (مسئولیت اخلاقی) سپرده شده و به عقل و روح مجهز است (محمود و دیگران، ۲۰۲۵؛ بقره ۳۰). هویت اخلاقی در این سنت از تکلیف (مسئولیت الهی) و پرورش فضایلی مانند عدل، نیت و مصلحت جدایی‌ناپذیر است. چنین چارچوبی تقویت می‌کند که تعامل اخلاقی با هوش مصنوعی یک مهارت صرفاً فنی نیست، بلکه یک وظیفهٔ اخلاقی و معنوی است.

سواد هوش مصنوعی فراتر از شایستگی فنی

سواد هوش مصنوعی اغلب به‌طور محدودی به‌عنوان مجموعه‌ای از مهارت‌های فنی تعریف می‌شود، از جمله درک الگوریتم‌ها، ساختارهای داده، محدودیت‌های مدل و مفاهیم پایه برنامه‌نویسی. با این حال، یافته‌های چندین منبع نشان می‌دهد که سواد اخلاقی نیز به همان اندازه ضروری است. کاربران باید بتوانند تشخیص دهند که یک خروجی هوش مصنوعی چه زمانی شامل یک قضاوت اخلاقی است، چه فرض ارزش‌گذاری شده‌ای را در خود جای داده و چه پیامدهای افتراقی برای گروه‌های مختلف دارد (یونسکو، ۲۰۲۱؛ چارچوب اخلاق هوش مصنوعی جامعهٔ اطلاعاتی ایالات متحده، ۲۰۲۰).

چارچوب هوش اخلاقی برجسته می‌کند که سیستم‌های هوش مصنوعی می‌توانند تخصص اخلاقی را به معنای تولید پاسخ‌های منسجم و قاعده‌محور به معماهای اخلاقی از خود نشان دهند، اما اغلب فاقد انسجام (یکپارچه‌سازی ارزش‌های اخلاقی متعدد) و شفافیت (قابل توضیح بودن استدلال اخلاقی) هستند (استوپو و دیگران، ۲۰۲۲؛ هابارد و دیگران، ۲۰۲۵). افزون بر این، همان‌طور که باستروم و یودکوسکی (۲۰۱۴) استدلال می‌کنند، حتی سیستم‌های هوش مصنوعی پیشرفته ممکن است در زمینه‌هایی عمل کنند که توسط طراحانشان پیش‌بینی نشده است، و پیش‌بینی محلی رفتار اخلاقی را غیرممکن می‌کند. این یک ریسک قابل توجه ایجاد می‌کند: کاربران ممکن است توانایی‌های اخلاقی هوش مصنوعی را بیش از حد برآورد کرده و قضاوت اخلاقی را به‌طور نامناسب به آن واگذار کنند.

افرادی که دارای سواد فنی بالای هوش مصنوعی هستند، اما هویت اخلاقی ضعیفی دارند، همچنان ممکن است اگر فاقد پایه‌ریزی اخلاقی باشند از هوش مصنوعی به‌طور نادرست استفاده کنند. آن‌ها ممکن است به‌طور کارآمد یک الگوریتم ناعادلانه را عملیاتی کنند یا یک فرآیند ناقض حریم خصوصی را بدون تردید اخلاقی خودکار سازند. بنابراین، سواد مؤثر هوش مصنوعی باید حداقل سه بعد به هم پیوسته را یکپارچه سازد:

- شایستگی فنی (درک چگونگی کارکرد هوش مصنوعی و محدودیت‌های آن)؛
- استدلال اخلاقی (به‌کارگیری اصولی مانند عدالت، عدم آسیب‌رسانی و شفافیت)؛
- آگاهی اخلاقی (بازشناسی هویت اخلاقی و مسئولیت خود به‌عنوان یک عامل انسانی).

مرجعیت اخلاقی درک‌شده‌ی سیستم‌های هوش مصنوعی

یکی از چشمگیرترین یافته‌های پژوهش‌های اخیر این است که مردم، اغلب سیستم‌های هوش مصنوعی را مشاوران اخلاقی معتبر یا حتی مراجع اخلاقی درک می‌کنند. دیلون و دیگران[1] (۲۰۲۵) دریافتند که توصیه‌های اخلاقی تولیدشده توسط هوش مصنوعی گاهی به‌عنوان قابل اعتمادتر، متعادل‌تر و واضح‌تر از توصیه‌های متخصصان اخلاق انسانی ارزیابی می‌شود. این پدیده یک تناقض عمیق ایجاد می‌کند: هوش مصنوعی از نظر ظاهری شایسته‌ی اخلاقی به نظر می‌رسد، با این حال، بر اساس اجماع موجود در ادبیات، فاقد درک اخلاقی اصیل، آگاهی، قصد و ظرفیت تجربهٔ ذهنی است.افراد با هویت اخلاقی قوی به احتمال بیشتری این محدودیت را تشخیص می‌دهند. آن‌ها خروجی‌های هوش مصنوعی را به‌عنوان ابزاری برای تأمل تفسیر می‌کنند، نه به‌عنوان قضاوتهای اخلاقیِ قطعی. آنها سؤالات سطح دوم می‌پرسند: چرا هوش مصنوعی این توصیه را تولید کرد؟ روی چه داده‌ای آموزش دیده است؟ چه ارزش‌هایی در آن تعبیه شده؟ اگر توصیه باعث آسیب شود چه کسی پاسخگو است؟ در مقابل، افراد با هویت اخلاقی ضعیف‌تر ممکن است بدون ارزیابی انتقادی، تصمیمات هوش مصنوعی را بپذیرند و به‌طور مؤثر عاملیت اخلاقی را به یک ماشین برون‌سپاری کنند.سناریویی که مستقیماً اصول اخلاق اسلامی و انسان‌گرایانهٔ تکلیف (مسئولیت اخلاقی شخصی) را تضعیف می‌کند(باستروم و یودکوسکی، ۲۰۱۴؛ محمود و دیگران، ۲۰۲۵).

حساسیت اخلاقی و تصمیم‌گیری هوش مصنوعی

یافته‌های تجربی در مورد حساسیت اخلاقی هوش مصنوعی نشان می‌دهد که سیستم‌های معاصر هوش مصنوعی می‌توانند معماهای اخلاقی را تشخیص دهند (مثلاً شناسایی یک معاوضه بین صداقت و وفاداری) اما اغلب آن‌ها را با قطعیّتی بی‌جای یا انعطاف‌ناپذیری الگوریتمی حل می‌کنند. برای مثال، در سناریوهای معاوضه‌ی تراژیک (مانند تصمیم‌گیری خودروهای خودران)، مدل‌های هوش مصنوعی به‌طور مداوم یک گزینه را انتخاب می‌کنند، با وجود اینکه به پیچیدگی معما اذعان دارند (هابارد و دیگران، ۲۰۲۵). این منعکس‌کنندهٔ یک محدودیت ساختاری است: سیستم‌های هوش مصنوعی برای بهینه‌سازی انسجام در توزیع آموزش خود برنامه‌ریزی شده‌اند، نه برای نوعی فروتنی اخلاقی، درگیری عاطفی یا حساسیت زمینه‌ای که از عامل‌های اخلاقی انسانی انتظار می‌رود.

[1] Dillon et al

کاربران با هویت اخلاقی بالا به احتمال بیشتری این ناسازگاری را متوجه شده و خروجی‌های هوش مصنوعی را زیر سؤال می‌برند. همچنین آن‌ها با ابهام اخلاقی راحت‌تر هستند و تشخیص می‌دهند که برخی معماهای اخلاقی راه‌حل کاملاً رضایت‌بخشی ندارند (هانزلمن و تانر[1] ۲۰۰۸). در اندیشهٔ اسلامی، این شناخت با مفهوم اجتهاد (استدلال مستقل) و این درک همسو است که قضاوت اخلاقی نه تنها به قواعد، بلکه به حکمت و نیت نیاز دارد، که هیچ‌کدام در اختیار هوش مصنوعی نیست (محمود و دیگران، ۲۰۲۵).

نقش هویت اخلاقی در تفسیر سوگیری و دستکاری

سیستم‌های هوش مصنوعی می‌توانند از طریق داده‌های آموزش مغرضانه، انتخاب ویژگی، طراحی مدل و حلقه‌های بازخورد، سوگیری‌های اجتماعی را تداوم و تشدید بخشند و همچنین با به‌کارگیری الگوریتم‌های شخصی‌سازی هدفمند، اصطکاک‌زدایی‌ها و تحلیل‌های پیش‌بینی‌کننده، رفتار کاربران را دستکاری نمایند (اونیل، ۲۰۱۶). در این میان، هویت اخلاقی نقش تعیین‌کننده‌ای در نحوه پاسخ افراد به این ریسک‌ها ایفا می‌کند؛ به گونه‌ای که افراد با هویت اخلاقی بالا، مقاومت بیشتری در برابر دستکاری داشته، احتمال بالاتری در تشخیص سوگیری نشان می‌دهند، گرایش قوی‌تری به مطالبه انصاف و شفافیت دارند و در صورت نقض اصول اخلاقی توسط هوش مصنوعی، توصیه آن را لغو یا رد می‌کنند. در مقابل، افراد با هویت اخلاقی پایین، آسیب‌پذیری بیشتری در برابر دستکاری داشته و به احتمال کمتر خروجی‌های مغرضانه را به چالش می‌کشند و اغلب «الگوریتم چنین گفت» را به‌عنوان توجیهی کافی می‌پذیرند. برای مثال، در محیط‌های رسانه‌های اجتماعی، کاربرانی با ارزش‌های اخلاقی قوی مانند عدالت و صداقت، محتوای هدفمند را زیر سؤال برده، شخصی‌سازی دستکاری‌گرانه را تشخیص داده و انصاف را بر راحتی یا معیارهای تعامل ترجیح می‌دهند. همچنین در محیط‌های نهادی، چنین افرادی با احتمال بیشتری به‌عنوان «سوت‌زن» اخلاقی عمل کرده یا ممیزی الگوریتمی را مطالبه می‌کنند، همان‌طور که در چارچوب اخلاق هوش مصنوعی جامعهٔ اطلاعاتی ایالات متحده بر بازبینی و پاسخگویی مستمر تأکید شده است (چارچوب اخلاق هوش مصنوعی جامعهٔ اطلاعاتی ایالات متحده، ۲۰۲۰).

شفافیت، اعتماد و قضاوت اخلاقی

شفافیت (قابل توضیح‌بودن و قابل تفسیر بودن) سنگ‌بنای حکمرانی اخلاقی هوش مصنوعی در هر

[1] Hanselmann & Tanner

سه سند منبع است (یونسکو، ۲۰۲۱؛ باستروم و یودکوسکی، ۲۰۱۴؛ چارچوب اخلاق هوش مصنوعی جامعهٔ اطلاعاتی ایالات متحده، ۲۰۲۰). با این حال، بسیاری از سیستم‌های پیشرفته هوش مصنوعی به‌عنوان «جعبه‌های سیاه» عمل می‌کنند و درک چگونگی رسیدن به یک تصمیم خاص را برای کاربران دشوار یا غیرممکن می‌سازند. این ابهام، چالشی مستقیم برای عاملیت اخلاقی ایجاد می‌کند: اگر نتوانم توضیح دهم که چرا هوش مصنوعی یک توصیه را ارائه کرده است، آیا می‌توانم از نظر اخلاقی در قبال عمل بر اساس آن مسئول باشم؟هویت اخلاقی نقش مهمی در هدایت این تنش ایفا می‌کند. کاربران با هویت اخلاقی قوی به احتمال بیشتری:

– شفافیت را از سیستم‌های هوش مصنوعی و توسعه‌دهندگان آن‌ها مطالبه می‌کنند.

– اعتمادپذیری را بر اساس توضیحات ارزیابی می‌کنند، نه صرفاً بر اساس نتایج.

– عامل‌های انسانی را در قبال خروجی‌های مبهم یا آسیب‌رسان هوش مصنوعی پاسخگو می‌دانند.

همچنین آن‌ها به احتمال کمتری به یک سیستم هوش مصنوعی صرفاً به این دلیل که از نظر فنی «دقیق» است اعتماد می‌کنند؛ آن‌ها به انسجام با اصول اخلاقی و شفافیت در مورد محدودیت‌ها نیاز دارند. در مقابل، کاربران با هویت اخلاقی ضعیف‌تر ممکن است توصیه‌های مبهم را تا زمانی که راحت یا کارآمد هستند بپذیرند و بدین ترتیب تقسیم کاری مشکوک از نظر اخلاقی را ممکن سازند: هوش مصنوعی تصمیم می‌گیرد، انسان‌ها تمکین می‌کنند.

تعامل بین هویت اخلاقی و سواد هوش مصنوعی

رابطه بین هویت اخلاقی و سواد هوش مصنوعی تعاملی و تقویت‌کنندهٔ متقابل است، نه صرفاً افزایشی:

– سواد هوش مصنوعی مهارت‌های شناختی و فنی را برای درک، ارزیابی و استفاده از سیستم‌های هوش مصنوعی فراهم می‌کند.

– هویت اخلاقی جهت‌گیری اخلاقی، انگیزه و تعهد خودتنظیمی را برای استفاده از آن مهارت‌ها به روش‌های مسئولانه اخلاقی فراهم می‌کند.

بدون هویت اخلاقی، سواد هوش مصنوعی ممکن است به استفاده‌ی کارآمد اما غیراخلاقی منجر شود: فرد می‌داند چگونه یک سیستم هوش مصنوعی را مستقر کند، اما قطب‌نمای اخلاقی برای زیر سؤال بردن اینکه آیا باید در یک زمینهٔ خاص مستقر شود یا چگونه آسیب‌های آن را کاهش دهد را ندارد. بدون سواد هوش مصنوعی، هویت اخلاقی ممکن است برای هدایت سیستم‌های فناورانهٔ پیچیده ناکافی باشد: فرد نیت خوبی دارد اما نمی‌تواند سوگیری تعبیه‌شده را شناسایی کند، دستکاری را تشخیص دهد یا شفافیت مناسب را مطالبه کند.بنابراین، بالاترین سطح تعامل اخلاقی با هوش مصنوعی زمانی رخ می‌دهد که هویت اخلاقی قوی با سواد بالای هوش مصنوعی همراه شود. چنین افرادی نه تنها از نظر

فنی شایسته هستند، بلکه از نظر اخلاقی تأملی، نهادی مسئول و قادر به اعمال آنچه سنت اسلامی تکلیف (مسئولیت اخلاقی) در محیط‌های با واسطهٔ فناورانه می‌نامد، می‌باشند (محمود و دیگران، ۲۰۲۵).

پیامدهای آموزشی و سیاست‌گذاری

یافته‌های این تلفیق به‌شدت نشان می‌دهد که برنامه‌های آموزش و تربیت هوش مصنوعی باید فراتر از کسب مهارت فنی رفته و موارد زیر را ادغام کنند:

چارچوب‌های رشد اخلاقی بر این پیشفرض استوارند که توانایی قضاوت و استدلال اخلاقی در انسان‌ها به صورت مرحله‌ای و تدریجی توسعه می‌یابد. کلبرگ (۱۹۸۴) شش مرحله رشد اخلاقی را در سه سطح پیش‌عرفی، عرفی و پس عرفی شناسایی کرد که از جهت‌گیری مبتنی بر اجتناب از تنبیه تا جهت‌گیری مبتنی بر اصول جهانی عدالت، حقوق بشر و وجدان فردی امتداد می‌یابد. رست (۱۹۸۶) نیز مدل چهارجزئی خود را ارائه داد که شامل حساسیت اخلاقی (تشخیص وجود یک مسئله اخلاقی)، قضاوت اخلاقی (تشخیص درست از نادرست)، انگیزش اخلاقی (اولویت‌دهی به ارزش‌های اخلاقی بر سایر ارزش‌ها) و اجرای اخلاقی (شجاعت عمل بر اساس قضاوت اخلاقی) است. کاربرد این چارچوب‌ها در آموزش سواد هوش مصنوعی به این معناست که برنامه‌های آموزشی باید متناسب با سطح رشد اخلاقی یادگیرندگان طراحی شوند و فرصت‌هایی برای تأمل در معماهای اخلاقی، بحث درباره دوراهی‌های اخلاقی مرتبط با هوش مصنوعی و تمرین قضاوت اخلاقی در موقعیت‌های شبیه‌سازی‌شده فراهم آورند. این رویکرد به افراد کمک می‌کند تا از سطوح پایین‌تر رشد اخلاقی (مانند پذیرش صرفِ دستورات الگوریتم به دلیل ترس از عواقب) به سطوح بالاتر (مانند ارزیابی مستقل خروجی‌های هوش مصنوعی بر اساس اصول جهانی عدالت و کرامت انسانی) حرکت کنند.

استدلال اخلاقی در مواجهه با هوش مصنوعی نمی‌تواند صرفاً بر اساس یک سنت فکری واحد شکل گیرد، بلکه نیازمند بهره‌مندی از ظرفیت‌های موجود در سنت‌های مختلف اخلاقی است. مقاصد الشریعه (اهداف شریعت اسلامی) که توسط اندیشمندانی مانند شاطبی و غزالی تدوین شده، پنج ضرورت اساسی را برای حفظ کرامت انسانی تعریف می‌کند: حفظ دین، جان، عقل، نسل و مال (محمود و دیگران، ۲۰۲۵). در حوزه هوش مصنوعی، این چارچوب می‌تواند مبنایی برای ارزیابی سیستم‌ها قرار گیرد؛ برای مثال، آیا سیستم هوش مصنوعی به حفظ عقل (با تقویت تفکر انتقادی) کمک می‌کند یا آن را تضعیف می‌نماید؟ آیا به حفظ مال (با جلوگیری از کلاهبرداری‌های الگوریتمی) کمک می‌کند یا نه؟ اخلاق فضیلت نیز بر پرورش صفات پسندیده‌ای مانند شجاعت، انصاف، صداقت و فروتنی تأکید دارد و در مقابل، رذایلی مانند طمع، خودخواهی و بی‌تفاوتی را نفی می‌کند. از منظر اخلاق فضیلت، پرسش کلیدی این نیست که «آیا این عمل هوش مصنوعی درست است؟» بلکه این است که «انسانِ فضیلتمند در مواجهه با این خروجی هوش مصنوعی چه می‌کند؟» اصول وظیفه‌گرا نیز بر پایه احترام به کرامت انسانی

و عدم استفاده از انسان به عنوان وسیله صرف استوارند؛ از این منظر، استفاده از هوش مصنوعی برای دستکاری رفتار انسان یا نقض حریم خصوصی اساساً نادرست است، صرف نظر از اینکه پیامدهای مفیدی داشته باشد یا نه. تحلیل پیامدگرا نیز بر ارزیابی نتایج و عواقب عمل تأکید دارد و می‌پرسد: «استفاده از این سیستم هوش مصنوعی چه سود و زیان‌هایی برای بیشترین تعداد افراد دارد؟» (باستروم و یودکوسکی، ۲۰۱۴). آموزش ترکیبی این چهار چارچوب به افراد امکان می‌دهد تا از یک سو، اصول جهانی اخلاقی را بشناسند و از سوی دیگر، به تنوع فرهنگی و دینی مخاطبان احترام بگذارند و در موقعیت‌های خاص، با استفاده از استدلال چندوجهی، بهترین تصمیم اخلاقی را اتخاذ کنند.

بخش سوم و عملیاتی‌ترین مؤلفه، استفاده از سناریوهای واقعی و معماهای اخلاقی مرتبط با هوش مصنوعی در فرآیند آموزش است. این معماها باید به گونه‌ای طراحی شوند که شرکت‌کنندگان را در موقعیت‌هایی قرار دهند که نیاز به شناسایی سوگیری دارند؛ مثلاً سناریویی که در آن یک سیستم هوش مصنوعی استخدام، به طور سیستماتیک متقاضیان زن را رد می‌کند (به دلیل داده‌های تاریخی مغرضانه) و شرکت‌کنندگان باید تشخیص دهند که سوگیری در کجای زنجیره داده، مدل یا خروجی رخ داده است. همچنین این مطالعات موردی باید فرصت زیر سؤال بردن سیستم‌های مبهم (جعبه سیاه) را فراهم آورند؛ برای مثال، سناریویی که در آن یک الگوریتم تشخیص چهره در فرودگاه، فردی را به اشتباه به عنوان مظنون معرفی می‌کند، اما هیچ توضیحی درباره چگونگی این تصمیم ارائه نمی‌شود. شرکت‌کنندگان می‌آموزند که سوالاتی مانند «الگوریتم از چه ویژگی‌هایی استفاده کرده؟»، «داده‌های آموزش شامل چه کسانی بوده؟» و «چه کسی پشت این سیستم است؟» را بپرسند. علاوه بر این، این معماها باید انتساب پاسخگویی را تمرین کنند؛ یعنی در سناریویی که یک خودروی خودران دچار حادثه می‌شود، شرکت‌کنندگان باید تصمیم بگیرند که چه کسی مسئول است: برنامه‌نویس؟ سازنده خودرو؟ مالک خودرو؟ یا خود الگوریتم؟ در نهایت، این آموزش باید شامل تمرین لغو یا اصلاح توصیه‌های هوش مصنوعی در صورت نقض اصول اخلاقی باشد. برای مثال، در سناریویی که یک پزشک از هوش مصنوعی برای تشخیص بیماری استفاده می‌کند، اما هوش مصنوعی توصیه‌ای می‌کند که با قضاوت بالینی پزشک در تضاد است و پیامدهای جدی برای بیمار دارد. شرکت‌کنندگان می‌آموزند که چگونه می‌توان با حفظ احترام به هوش مصنوعی به عنوان یک ابزار، اما با تکیه بر عاملیت اخلاقی خود، توصیه نادرست را لغو کرده یا اصلاح نمود. این تمرینات عملی، پل ارتباطی بین دانش نظری اخلاقی و مهارت عملی مواجهه با هوش مصنوعی در دنیای واقعی را ایجاد می‌کنند.

بحث و نتیجه گیری

یافته ها نشان داد که هویت اخلاقی یک عامل حیاتی و در عین حال اغلب نادیده گرفته شده در درک سواد هوش مصنوعی و تعامل اخلاقی با سیستم‌های هوش مصنوعی است. در حالی که سواد

هوش مصنوعی افراد را به دانش و مهارت‌های لازم برای تعامل با فناوری‌های هوش مصنوعی مجهز می‌کند، هویت اخلاقی تعیین می‌کند که چگونه این مهارت‌ها در موقعیت‌های پیچیده، مبهم یا پُرریسک اخلاقی به کار گرفته می‌شوند.یافته‌ها، که از روان‌شناسی اخلاق، ادبیات اخلاق هوش مصنوعی، چارچوب‌های اخلاقی اسلامی و رهنمودهای کاربردی اطلاعاتی تلفیق شده‌اند، نشان می‌دهند که سیستم‌های هوش مصنوعی علیرغم پیچیدگی فزاینده‌شان فاقد عاملیت اخلاقی اصیل، آگاهی، قصد، تجربهٔ ذهنی و پاسخگویی هستند. آن‌ها می‌توانند استدلال اخلاقی را شبیه‌سازی کنند و حتی در معیارهای محدود سازگاری اخلاقی از انسان پیشی بگیرند، اما نمی‌توانند مسئولیت اخلاقی را بر عهده بگیرند، آگاهی معنوی را اعمال کنند یا دارای «نیت» باشند. در نتیجه، مسئولیت نهایی در قبال تصمیم‌گیری اخلاقی، نظارت و پاسخگویی بر عهده ی کاربران انسانی باقی می‌ماند.

افراد با هویت اخلاقی قوی، خروجی‌های هوش مصنوعی را انتقادی ارزیابی کرده، ریسک‌های اخلاقی را تشخیص می‌دهند، شفافیت و پاسخگویی انسانی را مطالبه می‌کنند و از هوش مصنوعی به‌عنوان ابزاری برای تأمل استفاده می‌نمایند نه جایگزینی برای قضاوت اخلاقی؛ در مقابل، اتکا به هوش مصنوعی بدون پایه‌ریزی اخلاقی، منجر به پذیرش غیرانتقادی، تشدید بی‌عدالتی و تضعیف کرامت انسانی می‌شود. بنابراین، این مقاله بر رویکردی چندبعدی و یکپارچه به سواد هوش مصنوعی تأکید دارد که شایستگی فنی، استدلال اخلاقی و آگاهی معنوی–اخلاقی را ترکیب می‌کند و از چارچوب‌هایی مانند یونسکو، جامعهٔ اطلاعاتی ایالات متحده و اخلاق مقاصد اسلامی بهره می‌برد. در عصری که هوش مصنوعی بر تصمیمات حساس انسانی تأثیر می‌گذارد، پژوهش‌های تجربی آینده باید به بررسی این رابطه در زمینه‌های فرهنگی و نهادی متنوع بپردازند.

ملاحظات اخلاقی

پیروی از اصول اخلاق پژوهش

درتحقیق انجام شده، از اصول اخلاق پژوهش، پیروی شده است.

حامی مالی

هزینه‌های پژوهش حاضر، توسط نویسنده تامین شد.

تعارض منافع

فاقد هرگونه تعارض منافع بوده است.

تشکر وقدردانی

بدین وسیله از بزرگوارانی که در انجام این پژوهش، بنده را یاری فرموده اند، نهایت تشکر وقدردانی می نماییم.

منابع:

Aquino, K., & Reed II, A. (2002). The self-importance of moral identity. *Journal of personality and social psychology*, *83*(6), 1423.

Bostrom, N., & Yudkowsky, E. (2018). The ethics of artificial intelligence. In *Artificial intelligence safety and security* (pp. 57-69). Chapman and Hall/CRC.

Dillion, D., Mondal, D., Tandon, N., & Gray, K. (2025). AI language model rivals expert ethicist in perceived moral expertise. *Scientific Reports*, *15*(1), 4084.

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

Hanselmann, M., & Tanner, C. (2008). Taboos and conflicts in decision making: Sacred values, decision difficulty, and emotions. *Judgment and Decision making*, *3*(1), 51-63.

Hubbard, S., Kidd, D., & Stupu, A. (2025). *Crocodile tears: Can the ethical-moral intelligence of AI models be trusted?* [Manuscript submitted for publication].

Kohlberg, L. (1987). The psychology of moral development. *Ethics*, *97*(2).

Mahmood, S., Abbasi, E., & Awan, T. A. (2025 Artificial intelligence and human identity: Ethical challenges in Islamic thought. *Contemporary Journal of Social Science Review*, *3*(4), 85-100.

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

Russell, P. N. (2010). Artificial intelligence: a modern approach by stuart. *Russell and Peter Norvig contributing writers, Ernest Davis...[et al.]*, 22.

Rest, J. R. (1986). Moral development: Advances in research and theory. Praeger.

Sternberg, R. J. (2025). A trilogy theory of moral intelligence. Review of General Psychology, 29(1), 1–18.

Stupu, A. G., & Rusu, A. S. (2022). Integrative analysis of ethical intelligence and moral intelligence: New conceptual models and developments in education. *Educatia 21*, (23), 55-68.

Unesco. (2022). *Recommendation on the ethics of artificial intelligence*. United Nations Educational, Scientific and Cultural Organization. https://digitallibrary.un.org/record/4062376?v=pdf

United States Intelligence Community. (2020). Artificial intelligence ethics framework for the intelligence community (Version 1.0). Office of the

Director of National Intelligence.

The Relationship Between Moral Identity and AI Literacy: Implications for Ethical Engagement in the Age of Intelligent Systems

Tayebeh Ahmadi[1]

Abstract

Introduction and Objective: With the expansion of artificial intelligence, understanding users' ethical interaction with this technology has become increasingly important. This study, emphasizing the role of moral identity, examines its relationship with AI literacy and the impact of both on ethical behavior in the use of intelligent systems.

Research Methodology: The present study employs a library-based method with a descriptive-analytical approach, drawing on sources from moral psychology, AI ethics, Islamic ethical frameworks, and practical guidelines from the intelligence community.

Findings: A review of the sources indicates that moral identity (including the internalization of traits such as fairness, honesty, and responsibility) plays a mediating and moderating role in how AI outputs are interpreted, how ethical risks are assessed, and how decisions are made. While AI systems can simulate moral reasoning, they lack genuine moral agency, consciousness, intention (niyyah), and accountability. Contemporary models of ethical-moral intelligence emphasize that effective AI literacy must integrate three dimensions: technical competence, ethical reasoning, and spiritual-moral awareness. Furthermore, frameworks such as the UNESCO Recommendation, the US Intelligence Community's AI Ethics Framework, and Islamic ethical principles (including khilafah, amanah, ‘adl, and niyyah) all underscore the necessity of human accountability, transparency, and justice in the application of AI.

Conclusion: Educational transformation in the digital age requires cultivating strong moral identity and comprehensive AI literacy, achieved through integrated approaches, continuous assessment, and the reinforcement of responsibility, leading to ethical engagement and the preservation of human agency.

Keywords: Moral Identity, AI Literacy, Accountability

[1] M.A in Counseling, Education Office of District 1, Shiraz, Iran.Ahmadita1396@gmail.com