Islamic Philosophical Doctrines
ISSN Print: 2251- 9386
Online: 2783- 4026

Feasibility Study of "Ethics" in Artificial Intelligence: An Analysis Based on the Principles of "Nature" in the Philosophical Mysticism of Ayatollah Shahabadi

Saeedesadat shahidi[1],Seyedali mirlohi[2]

1. Responsible author: Saeeda Sadat Shahidi, Assistant Professor and Faculty Member, Department of Philosophy and Islamic Theology, Imam Sadiq University, Sisters Campus: shahidi@isu.ac.ir
2. Seyed Ali Mirlohi, PhD student in Jurisprudence and Law, Shahid Motahari University: mirlohi@motahari.ac.ir

Article Info

Article type: Scientific-Research

Article history:
Received
Received in revised form
Accepted
Available online

Keywords:
Keywords: Ethics of artificial intelligence, nature, Ayatollah Shahabadi, philosophical mysticism, moral agency.

ABSTRACT

A significant strand of contemporary discourse on artificial intelligence ethics tends to reduce moral judgment to computable rules and formalized procedures. Approaches such as utilitarian optimization and feedback-based alignment generally presuppose that ethical values can be adequately represented through mathematical functions and algorithmic mechanisms. Rather than assessing the technical feasibility or performance of such models, the present study shifts the discussion to a more fundamental level by addressing an ontological question: can an entity that lacks inner experience, intuitive awareness, and an intrinsic orientation toward moral perfection genuinely qualify as a moral agent?To examine this question, the article draws upon the mystical–philosophical thought of Hakim Shahabadi and employs his theory of fitrah (innate human disposition) as an anthropological and evaluative framework. An analysis of this theory reveals that moral action, in Shahabadi's view, is not the result of formal rule-following or external compliance, but emerges from the actualization of multiple inner and intuitive dimensions of human nature, including presentational knowledge, love, discovery, and the inclination toward freedom and comfort. By contrast, an examination of the internal structure of artificial intelligence systems demonstrates that their operations are grounded in objective functions, statistical correlations, and data-driven optimization, without access to inner experience or existential motivation.The study therefore concludes that artificial intelligence, even in its most advanced forms, cannot be regarded as a moral agent in the strict philosophical and jurisprudential sense. At best, it functions as a simulator of ethically patterned behavior rather than as a bearer of genuine moral responsibility. Consequently, delegating moral judgment to such systems risks obscuring the fundamental distinction between data processing and moral awareness. Addressing the ethical challenges posed by emerging technologies thus requires reaffirming the central role of human moral agency and grounding technological governance in a robust understanding of human nature..

آموزه‌های فلسفه اسلامی
ISSN Print: 2251- 9386 Online: 2783- 4026

امکان‌سنجی «اخلاق» در هوش مصنوعی: تحلیلی مبتنی بر مبانی «فطرت» درنگرش حکمی آیت‌الله شاه‌آبادی

سعیده سادات شهیدی[۱]،سیدعلی میرلوحی[۲]

۱. نویسنده مسوول، استادیار و عضو هیات علمی گروه فلسفه و کلام اسلامی دانشگاه امام صادق ع، پردیس خواهران: shahidi@isu.ac.ir

۲. دانشجوی دکتری فقه و حقوق دانشگاه و مدرسه عالی شهید مطهری(ره)،تهران،ایران: mirlohi@motahari.ac.ir

اطلاعات مقاله

نوع مقاله: علمی-پژوهشی

کلیدواژه‌ها:
اخلاق هوش مصنوعی، فطرت، آیت‌الله شاه‌آبادی، نگرش حکمی، عاملیت اخلاقی

چکیده

در بخش قابل‌توجهی از ادبیات معاصر اخلاقِ هوش مصنوعی، اخلاق به مجموعه‌ای از قواعد قابل محاسبه تنزل داده می‌شود. رویکردهایی مانند فایده‌گرایی یا الگوهای همسوسازی مبتنی بر بازخورد، بر این پیش‌فرض استوارند که می‌توان داوری اخلاقی را در قالب توابع صوری و سازوکارهای الگوریتمی بازنمایی کرد. این پژوهش، به‌جای تمرکز بر کارآمدی یا ناکارآمدی این الگوها در سطح اجرا، به موضوع ریشه‌ای‌تر پرداخته بررسی می‌کند موجودی که فاقد فطرت، تجربه درونی و گرایش وجودی به کمال است، اصولاً می‌تواند واجد عاملیت اخلاقی باشد؟برای پاسخ به این پرسش، نوشتار حاضر با رجوع به آرای عرفانی ـ فلسفی حکیم شاه‌آبادی، نظریهٔ «فطرت» را به‌مثابه مبنایی انسان‌شناختی و معیار سنجش اخلاقی مورد تحلیل قرار می‌دهد. بررسی این نظریه نشان می‌دهد که فعل اخلاقی در نگاه ایشان، نه حاصل امتثال صوری از قاعده، بلکه نتیجهٔ فعلیت‌یافتن ساحت‌های شهودی و گرایشی انسان است؛ ساحت‌هایی که شامل دانایی حضوری، محبت، کشف، میل به آسایش و طلب آزادی می‌شود. در مقابل، تحلیل ساختار وجودی هوش مصنوعی نشان می‌دهد که این سامانه‌ها، به دلیل ابتنا بر توابع هدف، همبستگی‌های آماری و فقدان تجربهٔ حضوری، از برخورداری از چنین مبنایی محروم‌اند.نتیجه آن است که هوش مصنوعی را نمی‌توان «عامل اخلاقی» به معنای دقیق فلسفی و فقهی دانست، بلکه در بهترین حالت، باید آن را شبیه‌ساز رفتارهای اخلاقی تلقی کرد. بر این اساس، سپردن داوری‌های ارزشی و اخلاقی به این سامانه‌ها، مستلزم غفلت از تفاوت بنیادین میان فعل برخاسته از فطرت انسانی و رفتاری است که صرفاً بر اساس پردازش داده‌ها شکل می‌گیرد.

مقدمه

گفتمان غالب پیرامون «اخلاق هوش مصنوعی» در سال‌های اخیر، عمدتاً در ساحت مباحث فنی و محاسباتی محصور مانده است. تلاش‌های پژوهشی، به‌ویژه در پارادایم‌های غربی، بر مفهوم «هم‌سوسازی» الگوریتم‌ها با قواعد حقوقی، ارزش‌های فایده‌گرایانه یا وظیفه‌گرایانه تمرکز یافته است. برای نمونه، سناریوهای کلاسیکی همچون «مسئله تراموا» و کاربست آن در تصمیم‌سازی خودروهای خودران، تجلی تقلیل مفاهیم عمیق اخلاقی به معادلات محاسباتی و ترجیحات آماری است (Awad et al., 2018, p.62). با این حال، چنین رویکردی پرسشی بنیادین و هستی‌شناختی را مغفول می‌گذارد: آیا سیستمی که ذاتاً فاقد «درک شهودی»، «آگاهی» و «گرایش ذاتی به کمال» است، اساساً از قابلیت دستیابی به «عاملیت اخلاقی حقیقی» برخوردار است؟. بحران کنونی در این حوزه، فراتر از یک بن‌بست فنی در چگونگی هم‌سوسازی، یک بحران مفهومی در «امکان‌سنجی» تحقق اخلاق در عاملی غیرانسانی است. ادعای مرکزی این نوشتار آن است که رویکردهای رایج، به سبب ناتوانی در تبیین ریشه‌های شهودی اخلاق و فروکاستن ارزش‌های متعالی همچون عدالت به «توابع هدف» ریاضی، در مواجهه با این بحران ناکام مانده‌اند.در سطح فنی، پیشرفته‌ترین مکانیزم‌های کنونی نظیر «یادگیری تقویتی از بازخورد انسانی»، اگرچه در انطباق «رفتار» مدل با «ترجیحات» بشری موفق بوده‌اند، اما در ایجاد «باور» یا «تعهد» اخلاقی اصیل شکست خورده‌اند (Ouyang et al., 2022, p.27737). این مدل‌ها صرفاً می‌آموزند که «چه چیزی را نباید گفت»، بدون آنکه حقیقتِ «چرایی» آن را درک کنند؛ شکافی که دقیقاً مرز میان «شبیه‌سازی اخلاق» و «اخلاقی بودن» را ترسیم می‌کند. در مقابل، سنت اسلامی بر درک حضوری و شهودی نسبت به اخلاق بر مبنای «فطری بودن» آن تأکید دارد. حکیم شاه‌آبادی از جمله اندیشمندان معاصری است که با ارائه قرائتی منسجم از خودشناسی و کمال‌خواهی فطری، بر نهادینه بودن سرچشمه‌های استکمال انسان پای فشرده و دین را شرح الزامات فطری وجودی بشر دانسته است. بر این اساس، انسان به واسطه برخورداری از «فطرت الهی»، استعدادی سرشتی برای درک و گرایش به الزامات اخلاقی دارد(صغیری و همکاران،۱۴۰۴،ص۱۰). پژوهش حاضر با عدول از رویکردهای تک‌ساحتی فنی یا فقهی، معیار «درون‌زا» را برای سنجش اخلاق معرفی می‌کند که از دل سنت عرفان فلسفی شیعی برآمده است: مفهوم «فطرت» در منظومه فکری آیت‌الله شاه‌آبادی. نوآوری این پژوهش، نه در تکرار مباحث نظری، بلکه در «کاربست» این مفهوم عرفانی به‌عنوان ابزاری تحلیلی برای مواجهه با چالش‌های تکنولوژی معاصر است.ادبیات پژوهش در حوزه اخلاق هوش مصنوعی را می‌توان در سه محور کلیدی دسته‌بندی کرد که وجه مشترک آن‌ها، تمرکز بر عوارض بیرونی یا کنترل رفتاری ماشین است:

۱) اخلاق محاسباتی و هم‌سوسازی فنی: این حوزه بر چگونگی پیاده‌سازی قواعد اخلاقی در ماشین متمرکز است و اخلاق را به یک «مسئله بهینه‌سازی» تقلیل می‌دهد. مطالعاتی همچون «آزمایش ماشین اخلاق»، ترجیحات اخلاقی انبوه را برای استفاده در خودروهای خودران استخراج کرده‌اند (Awad et al., 2018, p.61). همچنین مکانیزم RLHF برای هم‌سوسازی مدل‌های زبانی بزرگ توسعه یافته است (Ouyang et al., 2022 p.27740). نقد اساسی بر این رویکرد، ماهیت «رفتارگرا»ی آن است؛ ماشین می‌آموزد «شبیه» عامل اخلاقی رفتار کند، اما درک مفهومی از فعل خود ندارد. منتقدان این سیستم‌ها را «طوطی‌های تصادفی» می‌نامند که بدون درک معنا یا قصد، الگوهای زبانی را تکرار می‌کنند (Bender et al., 2021, p.610).

۲) فلسفه هوش مصنوعی و مسئله کنترل: آثاری نظیر «فراهوش» (Bostrom, 2014) و «سازگاری انسان» (Russell, 2019) به مسئله هم‌سوسازی ارزش‌های فراهوشمند با مبانی بشری پرداخته‌اند. این مباحث شکاف میان «هوشمندی ابزاری» و «خردمندی» را آشکار ساخته و بر خطرات «تعمیم نادرست هدف» تأکید دارند. استدلال کلاسیک «اتاق چینی» نیز تمایز میان پردازش نحوی و درک معنایی را به چالش کشیده است (Searle, 1980, p. 420). با این حال، تمرکز این آثار عمدتاً بر «ایمنی» و «کنترل» است و به این پرسش نمی‌پردازند که آیا سیستم محاسباتی اساساً می‌تواند ارزش‌ها را «درونی» کند یا صرفاً آن‌ها را «قیدهای بهینه‌سازی» می‌بیند

۳) شفافیت و جعبه سیاه: در پاسخ به عدم تفسیرپذیری مدل‌های پیچیده، حوزه «هوش مصنوعی قابل توضیح» پدید آمده است (Adadi & Berrada, 2018, p.52138; Gunning et al., 2019). اما «شفافیت» با «شهود اخلاقی» یکسان نیست. ابزارهای XAI تنها یک «توضیح پس‌رویدادی» از زنجیره منطقی ارائه می‌دهند (Ribeiro et al., 2016, p. 1135) و فاقد قدرت تبیین «تجربه زیسته» یا درک شهودی پیشااستدلالی انسان هستند.

در نهایت، هیچ‌یک از این رویکردها به «ظرفیت ذاتی» برای اخلاق، مشابه مفهوم «فطرت»، نپرداخته‌اند. پژوهش حاضر با روش تحلیلی-توصیفی و رویکردی بینارشته‌ای، در صددبررسی این موضوع از دیدگاه حکیم شاه آبادی آنهم به آن دلیل است که وی به عنوان فیلسوف شاخص مسلمان بر نقش فطرت به عنوان مبدا درونی معرفت و گرایش اخلاقی تاکید دارد.

۱.حقیقت فطرت از دیدگاه حکیم شاه آبادی

نظریه فطرت، به‌عنوان یکی از کلیدی‌ترین مباحث انسان‌شناسی قرآنی-حکمی، نقشی اساسی در حل چالش‌های معرفتی، اخلاقی و تربیتی ایفا می‌کند. این نظریه با بررسی براهین اثبات، اقسام دانش‌ها و گرایشات فطری، بستری برای تحلیل نسبت فطرت با قوای نفس فراهم می‌آورد (غفوری‌نژاد، ۱۳۹۸، ص ۲۵). در این عرصه، حکیم شاه‌آبادی جایگاهی ویژه دارد؛ ایشان در اثر بنیادین خود، «رشحات البحار»، با ابتنا بر اقتضائات سرشتی بشر به اثبات توحید، نبوت، معاد و خودسازی پرداخته است. تمرکز اصلی پژوهش حاضر بر واکاوی ریشه‌های شهودی اخلاق نزد این حکیم وارسته است. بدین منظور، ضمن تبیین تعاریف و اقسام فطرت در منظومه فکری ایشان، پیشینه این بحث نزد سایر اندیشمندان شاخص حکمت اسلامی نیز مرور می‌گردد تا جایگاه ممتاز شاه‌آبادی در تبیین گرایش و عاملیت اخلاقی انسان روشن شود.

واژه فطرت دلالت بر خلقت ویژه و ساختار وجودی مشترک میان تمامی انسان‌ها دارد که ریشه در آموزه‌های وحیانی قرآن کریم دارد. شهید مطهری بر این باور است که این اصطلاح پیش از نزول قرآن سابقه‌ای نداشته و برای نخستین بار در این کتاب آسمانی برای توصیف سرشت بشری به کار رفته است. این مفهوم از یک سو ساحتی انسان‌شناختی دارد و به بن‌مایه‌های وجودی بشر می‌پردازد و از سوی دیگر، پیوند تکوینی میان خالق و مخلوق را تبیین می‌کند (مطهری، ۱۳۶۹، ص ۱۴). بر مبنای آموزه‌های قرآنی، نوع بشر دارای ذاتی غیراکتسابی و تغییرناپذیر است که او را از سایر انواع حیوانی متمایز می‌سازد؛ ادراکات، گرایش‌ها و توانمندی‌های مختص بشری، همگی از شئون این ویژگی فطری محسوب می‌شوند. هرچند توجه به این مقوله در میان فیلسوفان مسلمان تحت تأثیر منابع روایی و قرآنی شکل گرفت، اما در نگاه حکمای سنتی، فطرت بیش از آنکه به‌عنوان بحثی انسان‌شناختی مستقل مطرح باشد، ذیل مباحث خداشناسی و گرایش سرشتی به مبدأ مورد تحلیل قرار گرفته است.ابن‌سینا در نظام فلسفی خود، بحث فطرت را غالباً معطوف به عشق موجودات به کمال و اشتیاق انسان به پروردگار دانسته است. او این موضوع را به جای عنوان صریح «فطرت»، ذیل مبحث «عشق» در رساله العشق مطرح می‌کند؛ چراکه مراد او از واژه فطرت

در برخی آثارش، با معنای اصطلاحی و قرآنی آن تفاوت دارد (ابن‌سینا، ۱۳۷۹، ص ۱۱۵-۱۱۸). با این حال، شیخ‌الرئیس در آثار عرفانی خود به خداگرایی فطری عنایتی ویژه دارد و معتقد است انسان به سبب برخورداری از روح مجرد، از مراتب والاتری از عشق - اعم از عشق مجازی به زیبارویان و عشق حقیقی به باری‌تعالی-بهره‌مند است که استعداد آن در اصل خلقت بشر نهاده شده است (ابن‌سینا، ۱۴۰۰ق، ص ۳۸۰-۳۹۳). همچنین در اندیشه او، تعابیری ناظر بر گرایش سرشتی به زیبایی و اصول اخلاقی نیز یافت می‌شود (شهیدی، ۱۴۰۱، ص ۱۱-۱۴). در ادامه این سنت، شیخ اشراق نیز از عشق انوار مدبره به ساحت قدسی و اشتیاق همگانی موجودات به کمال سخن می‌گوید و تأکید دارد که شدت نورانیت در انوار مدبر، مایه فزونی محبت آنان به انوار عالیه است (سهروردی، ۱۳۷۳، ص ۲۲۳-۲۲۴).

در میان حکمای مسلمان، صدرالمتألهین عنایتی مضاعف به مقوله فطرت داشته است؛ وی اگرچه عنوان مستقلی برای این بحث برنگزیده، اما در جای‌جای آثار فلسفی و تفسیری خود به تحلیل حقیقت، اقسام و هدایتگری فطرت پرداخته است. ملاصدرا فطرت را قوه و استعدادِ صفات و ملکاتی می‌داند که خداوند در بدو خلقت در جان انسان سرشته است؛ به گونه‌ای که استعداد هر فضیلت و رذیلتِ مقابل آن، در وجود فرد به ودیعه نهاده شده است. او برای تبیین این حقیقت از تعابیری چون «فطرت اصلی»، «فطرت ربانی» و «اصل الفطره» استفاده می‌کند (ملاصدرا، ۱۳۶۶، ج ۱، ص ۱۹، ۳۰۹، ۳۴۵، ۳۵۲، ۴۴۵، ۴۴۶؛ ملاصدرا، ۱۹۸۱، ج ۹، ص ۱۳۲). در نگاه وی، «فطرت ثانوی» که محصول اکتسابات و کنش‌های دنیوی است، می‌تواند به شکوفایی فطرت اولی یا انحراف از آن منجر شود. صدرالمتألهین باور به وجود حق‌تعالی را امری چنان سرشتی می‌داند که گزاره «خدا وجود دارد» برای انسان بدیهی تلقی می‌شود (ملاصدرا، ۱۳۶۳، ص ۲۴۱). او فطرت متمایل به خیر را «هدایت» نامیده و معتقد است دوری از این مسیر، تنها در اثر باورها و رفتارهای ناشایست ثانوی رخ می‌دهد (ملاصدرا، ۱۳۶۶، ج ۱، ص ۴۴۵-۴۴۶). پس از وی، این بحث در میان نوصدراییان به اوج شکوفایی رسید و در این میان، حکیم شاه‌آبادی نقشی بی‌بدیل در ارائه طرحواره‌ای جامع از اقسام فطرت و نقش آن در کمال معنوی انسان ایفا کرد.

حکیم شاه‌آبادی با ارائه‌ی تفصیلی عالمانه از نظریه‌ی فطرت و دسته‌بندی اقسام آن، مبنایی استوار برای ریشه‌یابی شهود اخلاقی بنا نهاده است. از منظر ایشان، بنیادین‌ترین مسیر برای ادراک حقیقت فطرت، نه از طریق مفاهیم حصولی، بلکه تنها از راه خودشناسی و یافتن درونی این حقیقت میسر است (شاه‌آبادی، ۱۳۸۷، ص۱۱۵). ایشان در آثار خویش، بارها از این فرآیند درون‌نگرانه با تعابیری همچون مطالعه‌ی «کتاب ذات»، واکاوی «کتاب وجود» و یا مکاشفه‌ی فطری یاد کرده‌اند (همو، ۱۳۸۶، ص۹۵، ۱۰۹، ۱۲۷ و ۱۳۱).از منظر ریشه‌شناختی، حکیم شاه‌آبادی واژه‌ی فطرت را بر وزن «فِعله» از ماده‌ی «فطر» تحلیل می‌کنند که دلالت بر کیفیت و نوع ایجاد دارد. ایشان در تعریفی دقیق مرقوم داشته‌اند: «از آنجا که وجود و ایجاد حقیقتی واحد هستند، پس کیفیت ایجاد حق که با هویت ما مساوی است، همان فطرت و صفات لازمه‌ی وجود ماست» (همو، ۱۳۸۷، ص۱۱۴). طبق این مبنا، حقیقت فطرت معادل همان کیفیت خلقتی است که پروردگار در جان انسان نهاده و شامل صفاتی است که از وجود او تفکیک‌ناپذیرند. در واقع، فطرت از لوازم لاینفک وجود بشری است که جعل آن، هم‌زمان با جعل اصل وجود اوست (همو، ۱۳۸۶، ص۲۱).در منظومه‌ی فکری حکیم، «کتاب ذات» انسان به قلم پروردگار نگاشته شده و به همین سبب، قوای وهم و خیال راهی به حریم آن ندارند. از سوی دیگر، با استناد به روایت نبوی «خلق الله آدم علی صورته»، ایشان معتقدند که کتاب ذات انسان با کتاب ذات حق انطباق کامل داشته و از این رو، فطرت در احکام ذاتی خود هرگز دچار خطا نمی‌شود (همو، ۱۳۸۷، ص۱۱۵). حکیم شاه‌آبادی مقتضای فطرت را کاملاً در راستای دین و در جهت کمال‌بخشی به انسان تبیین می‌کنند (همان، ص ۹۲). البته ایشان هشدار می‌دهند که انسان باید همواره نسبت به این اقتضائات

متفطن باشـــد، در غیر این صـــورت در حجاب طبیعت گرفتار گشـــته و دچار غفلت می‌گردد (همان، ص۱۱۵). تفطن به این لایه‌های پنهان، مستلزم نوعی درون‌نگری است که طی آن انسان با شهودی درونی به ساختار حقیقی وجود خویش آگاهی می‌یابد. در این مسیر، رجوع به «فطرت عالمه» محوری‌ترین رکن معرفت به ذات است. ما با رجوع به این ساحت درمی‌یابیم که نسبت به خود علمی حضوری داریم، در حالی که آگاهی ما نسبت به سایر موجودات از سنخ علم حصولی است (همو، ۱۳۸۶، ص۷۲). این وجودِ آگاه، متعلق به عالم امر اســـت و ســـنخیتی با عالم خلق و طبیعت ندارد؛ زیرا مناط علم، حضـــور است، اما در ساحت طبیعت، حضوری متصور نیست و آن لایه، تنها ظرف غیبت و جهل است. از نظر ایشان، این کشف که «خودبینی» نامیده می‌شود، نخستین هدف انبیای الهی بوده است؛ چرا که نقشی حیاتی در خروج انسان از اسارت طبیعت ایفا می‌کند. با این خودشناسی، انسان به وضوح درک می‌کند که حقیقت او، با بدن و ماده یکی نیست (همو، ۱۳۸۷، ص۱۱۶).

دیگر ارجاعات شـــهودی نیز در معرفت به خود نقش ایفا می‌کنند (همو، ۱۳۸۶، ص۷۲). دومین رکن، «فطرت عاشـــقه» است؛ انسان با رجوع به این ساحت درمی‌یابد که فطرتاً عاشق خود و کمالات خود است. حکیم از این درک به «خودخواهی» تعبیر می‌کنند؛ بر این پایه، حتی شخصی که دست به خودکشی می‌زند، به سبب حبّ ذات چنین می‌کند تا از زندگی سخت رها شده و به آرامش برسد. بعد سوم، «فطرت کاشفه» است که در آغاز تمامی کارها و افکار انسانی تجلی دارد. انسان هرگز به سمت امری حرکت نمی‌کند مگر آنکه نسبت به نفع یا ضرر آن به‌صورت فطری آگاهی (کشف) پیدا کرده باشد. همین ویژگی سبب می‌شود که انسان به راحتی نظر دیگران را نپذیرد و «تقلید» را مخالف شالوده‌ی فطرت بداند؛ زیرا دین با فطرت هماهنگ است و تقلید را در اصول نمی‌پذیرد.چهارمین عامل در خودشناسی، رجوع به «فطرت حبّ راحت» است؛ تمامی ابنای بشر عاشق آسایش مطلق هستند و برای رسیدن به آن، حتی رنج‌های بزرگی را به جان می‌خرند تا به لذت مطلق دست یابند. در نهایت، رکن پنجم، میل به «آزادی» است. انسان با رجوع به فطرت آزادی درمی‌یابد که همواره مایل است اراده‌اش نافذ باشد و هیچ سدی در برابر خواست او قرار نگیرد (همو، ۱۳۸۷، ص ۱۱۶ و ۱۱۷). از دیدگاه حکیم، مناط فطری آزادی، توانایی بر عملی ساختن مقاصد فردی است (همو، ۱۳۸۶، ص۱۰۳).با تجمیع این پنج ارجاع درونی، انسان به معرفتی جامع از حقیقت خویش دست می‌یابد: او موجودی است که به خود علم حضوری دارد، عاشق خود است، قدرت تامل دارد، طالب آسایش است و در پی آزادی است. از آنجا که هیچ‌یک از این ویژگی‌ها در عالم طبیعت یافت نمی‌شوند، انسان به این نتیجه‌ی قطعی می‌رســـد که: «فلتحکم بأنّ حقیقتی، غیر طبیعتی» (پس حکم کن که حقیقت من، غیر از طبیعت من اســـت) (همو، ۱۳۸۷، ۱۱۷). حکیم شـــاه‌آبادی با تکیه بر این دریافت‌ها، مخاطب را به اقرار به تجردِ حقیقت وجودی خویش مجاب می‌کند.افزون بر این موارد، ایشان ابعاد دیگری از فطرت را نیز مطرح کرده‌اند که در شهودی دانستن اخلاق نقشی بسزا دارند؛ از جمله این ابعاد، «فطرت بغض نقص و حبّ کمال» است. بر اساس این ویژگی سرشتی، هرگونه کاستی مورد اکراه انسان است و او همواره به دنبال موجودات کامل است. اگر در مواردی انسان به اشتباه امری ناقص را کمال بپندارد، پس از آگاهی، بلافاصله رویگردان شده و به دنبال کمال مطلق می‌گردد. همین گرایش درونی است که در نهایت انسان را به سوی موجود مطلق سوق می‌دهد که حکیم از آن با عنوان «فطرت حبّ اصل» یاد کرده‌اند (همان، ص ۲۲۱ تا ۲۲۴).

۱-۱. فطرت به‌مثابهٔ «ریشهٔ شهود اخلاقی»

حکیم شـاه‌آبادی با نگاهی عمیق به ذومراتب بودن حقیقت وجودی انسـان در مسـیر اسـتکمال، میان دو مرتبه‌ی «انسـان طبیعی» و «انسـان فطری» تمایز قائل می‌شـود. از منظر ایشـان، انسـان در بدو پیدایش و مراحل نخسـتینِ زیسـت خود، انس و

الفتی تمام‌عیار با عالم طبیعت دارد؛ او در این مرتبه، همت خود را مصروف توجه به مادیات کرده و همچون سایر حیوانات، در پی تأمین نیازهای غریزی و تحقق آرزوهای مادی در ساحت ماده است. در این ساحتِ طبیعی، بشر ویژگی‌های بنیادین خود نظیر خوددوستی، ابراز رأی، راحت‌طلبی و میل به آزادی را تسلیم محضِ طبیعت کرده و به‌ناچار، طبیعت برای وی در جایگاه امری محبوب، تصمیم‌گیر و منشأ آزادی تلقی می‌شود (شاه‌آبادی، ۱۳۸۷، ص ۱۱۸). در چنین وضعیتی، انسان از تکاپو برای تحصیل کمالات برتر و لوازم معنوی آن بازمانده و به آسودگیِ کاذب در حجاب ماده تن می‌دهد (شاه‌آبادی، ۱۳۸۷، ص ۱۱۸). به همین سبب، حکیم شاه‌آبادی بر ضرورت «اعراض از طبیعت» تأکید می‌ورزد. مراد ایشان از این اعراض، نه نفی کامل ساحت مادی، بلکه بریدن از نگاه استقلالی به طبیعت است؛ چراکه طبیعت در نگاه حکیم، صرفاً نقش وساطت و ابزاری را در حصول کمالات اصیل انسانی بر عهده دارد (شاه‌آبادی، ۱۳۸۷، ص ۱۲۰). این تغییر نگرش موجب می‌شود انسان به وجه ملکوتی خویش و ضرورت سیر از عالم ملک به ملکوت متنبه شده و از ماندگاری در حجاب طبیعت بپرهیزد. در این سیرِ کمالی، فرد از خودخواهی‌های محدود مادی و آزادی‌های بدنی گذر کرده و بر اساس فطرتِ «حبّ اصل کمال» و «نفرت از نقص»، به کاستی‌های خویش پی برده و به سوی مرتبه‌ی نهایی کمال حرکت می‌کند؛ حکیم از این تحول وجودی به «اقامه‌ی وجه به سوی دین» تعبیر می‌نماید (شاه‌آبادی، ۱۳۸۷، ص ۱۲۱).در منظومه‌ی فکری حکیم شاه‌آبادی، هدایت انسان بر انطباق کامل میان فطرت و دین استوار است. انسان به اقتضای ترکیب وجودی‌اش از جسم و روح، دارای دو وجه متناظر است، اما حقیقتِ اقامه‌ی وجه به سوی دین، مستقیماً به ساحت معنوی و ادراکی او بازمی‌گردد. ایشان وجوه ادراکی بشر را در سه سطح «وجه حسّ»، «وجه عقل» و «وجه قلب» تبیین می‌کنند. وجه حس، مسئول ادراک محسوسات از طریق حواس پنجگانه در عالم ماده است. وجه عقل، راهی برای ادراک معقولات است که در مرتبه‌ی ضعف، از مسیر تفکر و برهان عبور می‌کند و در مرتبه‌ی قوت، بی‌نیاز از استدلال، به کشف و شهود امور معقول نائل می‌شود. اما وجه قلب، که همچون عقل مختص نوع بشر است، ساحتی است که انسان از طریق آن، خود و کمالاتش را درک می‌نماید. از این رو، انسان فطرتاً موجودی خودخواه و خودبین است؛ چراکه نخستین معلومِ او، ذات خودش می‌باشد و همین ادراک، مایه‌ی دوست‌داری خود و کمالات خود می‌شود. بشر در آغاز، کمالات محسوس را درک کرده و به آن‌ها دلبسته می‌شود، اما با ارتقای ادراکی، کمالات معقول فطری را نیز می‌یابد. در این مرحله، فرد به سوی دین روی آورده و متوجه کمال مطلق می‌گردد؛ او درمی‌یابد که عشق و اشتیاقش نامتناهی است، در حالی که موجودات عالم متناهی بوده و شایستگی معشوقِ مطلق بودن را ندارند. لذا خداوند که کمال و جمال مطلق است، تنها گزینه‌ی سزاوار برای عشق و عبودیت انسان قرار می‌گیرد (شاه‌آبادی، ۱۳۸۷، ص ۱۲۴ تا ۱۲۷).

از منظر آیت الله شاه‌آبادی، دین چیزی جز «التزام به حقایق فطری» در ساختار وجودی انسان نیست. هر فرد با رجوع به «کتاب ذات» خویش، خود را ملزم به سه حقیقتِ بنیادین می‌یابد: معرفت، عبودیت و عدالت. در این تقسیم‌بندی، توجه به عدالت و انجام اعمال نیک، معادل اقامه‌ی وجه حس به سوی دین است؛ طلب معرفت، اقامه‌ی وجه عقل؛ و عبودیت و تخلق به اخلاق الهی، تجلی اقامه‌ی وجه قلب است. بر این اساس، ادعای بی‌نیازی انسان از دین مردود است؛ چراکه بشر در ذات خود دین‌دار آفریده شده و کتاب‌های آسمانی و پیامبران، صرفاً شارحانِ همان کتابِ تکوینیِ نهفته در ذات انسان هستند. فطرت در مطالبه‌ی این سه رکن (معرفت، عدالت و عبودیت)، به نحو مطلق و بدون استثنا حکم می‌راند (شاه‌آبادی، ۱۳۸۷، ص ۱۲۷). برای مثال، التزام فطری به معرفت، به تمامی حقایق عالم و مبدأ و معاد تعلق می‌گیرد؛ انسان فطرتاً از جهل گریزان است و این عشق به دانستن چنان است که اگر بر تمامی حقایق زمین آگاه شود، باز هم مشتاق درک خیراتِ احتمالی در

کرات و عوالم دیگر است (شاه‌آبادی، ۱۳۸۷، ص ۱۳۱). دین نیز در راستای همین میل بی‌کران، به تبیین حقایق ملکوت و معاد می‌پردازد (شاه‌آبادی، ۱۳۸۷، ص ۱۳۱).در حوزه‌ی عدالت، فطرت خواهان آن است که تمامی انسان‌ها به ادای حقوق یکدیگر پایبند باشند. با اقامه‌ی وجه حس به سوی عدالت، فرد ملزم به شناخت حقوق، حدود و صاحبان حق (اعم از حق‌الله و حق‌الناس) و حرکت بر مدار مصالح و ترک مفاسد می‌گردد. دین با تبیین احکام شرعی و تلاش مجتهدان در استنباط فقهی، در حقیقت در حال پاسخ‌گویی به این نیاز فطری است (شاه‌آبادی، ۱۳۸۷، ص ۱۲۹). نهایتاً، عالی‌ترین سطح التزام فطری، خضوع در برابر موجود کامل است که اوج آن در قالب «عبودیت» نسبت به پروردگار تجلی می‌یابد: «ان العبودیه هی خضوع خاص بباب المعبود» (شاه‌آبادی، ۱۳۸۷، ص ۱۶۱). فطرت به وجوبِ مطلقِ خضوع حکم کرده است که ساحت‌های قلبی، زبانی و جوارحی را دربرمی‌گیرد. نشانه‌های این امر سرشتی را می‌توان در انواع احترامات فطری، تواضع‌های درونی و محبت‌ها مشاهده کرد. حکیم احترامات فطری را شامل احترام به «فرد حاضر»، «منعم» (نعمت‌دهنده)، «مقتدر»، «کامل» و «محبوب» می‌داند (شاه‌آبادی، ۱۳۸۷، ص ۱۲۹). تواضع در برابر صاحبان علم، کمال و اصالت، نشانه‌ای از این پیوند سرشتی است. نکته‌ی ظریف در اینجا، نسبی بودن خضوع است که میان خاضع (فطرت انسان) و مخدوم رخ می‌دهد؛ به گونه‌ای که این احترام در سیری تشکیکی به تمامی موجوداتِ دارای کمال تعلق گرفته و در نقطه‌ی اوج، به عبودیتِ مختص پروردگار منتهی می‌شود (شاه‌آبادی، ۱۳۸۷، ص ۱۲۹).البته باید میان خضوع قلبی و خضوع خارجی تفکیک قائل شد؛ چه بسا انسان کمال کسی را درک کرده و در قلب خاضع شود، اما به دلایلی چون حسد، تقیه یا دشمنی، در ظاهر از خضوع امتناع ورزد. برعکس، ممکن است فردی بدون اعتقاد قلبی، صرفاً از روی ریا یا طمع، خضوع جوارحی نشان دهد. با این حال، ظهور خضوع قلبی در برابر کمال، امری انکارناپذیر است (شاه‌آبادی، ۱۳۸۷، ص ۲۱۵). از دیدگاه حکیم شاه‌آبادی، عبودیت جامع تمامی خصلت‌های پسندیده است و فضایل اخلاقی تنها زمانی «کمال» محسوب می‌شوند که در راستای عبودیت باشند (شاه‌آبادی، ۱۳۸۷، ص ۱۲۹). در این ساحت، عبودیت به منزله‌ی «روحِ» اعمال است که به اعضا و جوارحِ متعددِ کنش‌های اخلاقی، وحدت می‌بخشد. لذا صبر، عفت و تواضع نیز باید همچون نماز و روزه، خالصانه و به قصد خضوع در برابر حق صورت گیرند تا ارزش معنوی یافته و قلب را مستعد تجلی اسما و صفات الهی گردانند (مرتضوی، ۱۳۸۶، ص ۷۱). حکیم تأکید دارند که احترام به حاضر (حتی اگر کودک یا دشمن باشد)، احترام به منعم، مقتدر و محبوب، همگی ریشه در ذات بشر دارند؛ به طوری که حتی در حیوانات نیز مرتبه‌ای از احترام به منعم دیده می‌شود (شاه‌آبادی، ۱۳۸۷، ص ۱۲۹). بر اساس همین اخلاقیاتِ فطری و تخلف از این احکامِ سرشتی است که خداوند در قیامت، کسانی را که از تعظیم و احترامِ سزاوار اعراض کرده‌اند، مورد مؤاخذه قرار خواهد داد (شاه‌آبادی، ۱۳۸۷، ص ۱۲۹).

۲-۱.دلالت‌های نظریه فطرت بر «عاملیت اخلاقی حقیقی»

حکیم عارف، آیت‌الله شاه‌آبادی، در چارچوب جهان‌بینی توحیدی و با نگاهی ژرف به ساختار ذومراتب حقیقت انسانی، تمایزی بنیادین میان دو ساحت «انسان طبیعی» و «انسان فطری» قائل است. از منظر ایشان، بشر در نخستین مراحل حیات خویش، در پیوند و الفتی تنگاتنگ با ساحت مادی و عالم طبیعت قرار دارد. در این مرتبه، تمامی توان و همت انسان معطوف به نیازهای غریزی و آرزوهای مادی است؛ به گونه‌ای که در ساحت ماده، همچون دیگر موجودات حیوانی، تنها به دنبال صیانت از بقا و ارضای غرایز خویش است. در این مرتبه از حیات که ساحت «طبیعی» نامیده می‌شود، ویژگی‌های اصیلی چون خوددوستی، حریت و راحت‌طلبی در بندِ تعلقات مادی گرفتار شده و طبیعت به اشتباه، به عنوان معشوق، تصمیم‌گیر و

منشأ آزادی پنداشته می‌شود (شاه‌آبادی، ۱۳۸۷، ص ۱۱۸). این انغمار در مادیات، مانعی بزرگ در مسیر استکمال معنوی است؛ چراکه انسان را در حجاب ماده محبوس ساخته و او را از تکاپو برای نیل به کمالات برتر باز می‌دارد (شاه‌آبادی، ۱۳۸۷، ص ۱۱۸). بر همین اساس، حکیم شاه‌آبادی بر ضرورت «اعراض از طبیعت» به عنوان گامی حیاتی در مسیر سلوک تأکید می‌ورزد. البته مراد ایشان از این اعراض، به هیچ روی نفی کامل ساحت مادی یا رهبانیت نیست، بلکه بریدن از نگاه استقلالی به عالم ماده است. در دیدگاه ایشان، طبیعت نباید هدف نهایی تلقی شود، بلکه صرفاً نقشی واسطه‌ای و ابزاری برای وصول به کمالات اصیل انسانی دارد (شاه‌آبادی، ۱۳۸۷، ص ۱۲۰). با این تغییر نگرش، بشر به وجه ملکوتی خویش تنبه یافته و از ماندگاری در لایه‌های زیرین هستی می‌پرهیزد. در این سیرِ استکمالی، فرد از خودخواهی‌های بدنی عبور کرده و بر پایه فطرتِ «حبّ کمال» و «نفرت از نقص»، به سوی کمال مطلق حرکت می‌کند؛ تحولی که حکیم از آن به «اقامه‌ی وجه به سوی دین» تعبیر می‌نماید (شاه‌آبادی، ۱۳۸۷، ص ۱۲۱). در منظومه فکری شاه‌آبادی، هدایت انسانی محصول انطباق تام میان دین و فطرت است. اگرچه انسان از ترکیب جسم و روح پدید آمده، اما حقیقت دین‌داری او ریشه در ساحت معنوی و ادراکی‌اش دارد. ایشان قوای ادراکی بشر را در سه سطح تبیین می‌کنند: «وجه حسّ» برای درک محسوسات، «وجه عقل» برای ادراک معقولات (که در مرتبه قوت به شهود منتهی می‌شود) و «وجه قلب» که عالی‌ترین ساحت و مختص نوع بشر است. از طریق وجه قلب است که انسان به درک خویشتن و کمالاتش نائل می‌آید. نخستین معلومِ هر فرد، ذات خودش است که به «علم حضوری» درک می‌شود و همین خودآگاهی، بستر خوددوستی و کمال‌خواهی را فراهم می‌آورد. انسان در ابتدا به کمالات محسوس دلبسته می‌شود، اما با ارتقای وجودی، درمی‌یابد که اشتیاق او نامتناهی است و موجودات مادی به دلیل تناهی، شایسته محبوبیت مطلق نیستند. در اینجاست که جانِ آگاه، تنها پروردگار را که کمال و جمال مطلق است، به عنوان تنها گزینه سزاوار برای عشق و عبودیت برمی‌گزیند (شاه‌آبادی، ۱۳۸۷، ص ۱۲۴ تا ۱۲۷). بر این پایه، دین در نگاه شاه‌آبادی چیزی جز «التزام به حقایق فطری» نیست. هر انسانی با رجوع به «کتاب ذات» خود، سه حقیقت بنیادین را می‌یابد: معرفت، عدالت و عبودیت. در این تقسیم‌بندی، عدالت تجلی وجه حس، معرفت تجلی وجه عقل و عبودیت تجلی وجه قلب در ساحت دین‌داری است. لذا ادعای بی‌نیازی از دین باطل است؛ زیرا انسان در ذات خود دین‌دار آفریده شده و وحی و نبوت، تنها شارحانِ همان کتاب تکوینیِ نهفته در فطرت هستند. این سه رکن به نحو مطلق و بدون استثنا مورد مطالبه فطرت قرار دارند (شاه‌آبادی، ۱۳۸۷، ص ۱۲۷). به عنوان مثال، میل به معرفت چنان در جان انسان نهفته است که اگر بر تمامی حقایق زمین آگاه شود، باز هم مشتاق درک کمالات عوالم دیگر است و دین دقیقاً در راستای همین میل بی‌کران به تبیین حقایق ملکوت می‌پردازد (شاه‌آبادی، ۱۳۸۷، ص ۱۳۱).

در حوزه عدالت نیز، فطرت خواهان ادای حقوق همگانی است. با اقامه وجه حس، فرد ملزم به شناخت حقوق (اعم از حق‌الله و حق‌الناس) و حرکت بر مدار مصالح می‌شود. در واقع، احکام شرعی پاسخی به این نیاز درونی بشر هستند (شاه‌آبادی، ۱۳۸۷، ص ۱۲۹). اما عالی‌ترین سطح این التزام، خضوع در برابر موجود کامل است که اوج آن در «عبودیت» متجلی می‌شود. حکیم تصریح دارد که: «ان العبودیه هی خضوع خاص بباب المعبود» (شاه‌آبادی، ۱۳۸۷، ص ۱۶۱). این خضوعِ فطری، شامل ساحت‌های قلبی، زبانی و جوارحی است و نشانه‌های آن را می‌توان در احترامات فطری به «منعم»، «مقتدر»، «کامل» و «محبوب» مشاهده کرد. این پیوند سرشتی چنان است که حتی در برابر کمالِ یک کافر نیز نوعی تواضع ذاتی در جان انسان پدید می‌آید و نقطه اوج این سلسله‌مراتب، عبودیتِ مختص پروردگار است (شاه‌آبادی، ۱۳۸۷، ص ۱۲۹). نکته حائز اهمیت در این میان، تفکیک میان خضوع قلبی و تجلی خارجی آن است. ممکن است فردی به دلیل رذایلی

چون حسد یا دشمنی، علی‌رغم درک عظمت یک کمال در قلب خویش، در ظاهر از خضوع امتناع ورزد. با این حال، ظهور خضوع قلبی در برابر کمال، امری انکارناپذیر و سرشتی است (شاه‌آبادی، ۱۳۸۷، ص ۲۱۵). در نهایت، عبودیت از منظر شاه‌آبادی، «روحِ» تمامی اعمال و خصلت‌های پسندیده است. فضایل اخلاقی تنها زمانی به کمال واقعی خود می‌رسند که در راستای عبودیت و خضوع در برابر حق باشند. در این حالت، کنش‌های اخلاقی همچون صبر و عفت، قلب را مستعد تجلی اسما و صفات الهی می‌گردانند (مرتضوی، ۱۳۸۶، ص ۷۱). بر پایه همین احکام فطری است که احترام به حاضر، منعم و مقتدر در ذات بشر نهاده شده و تخلف از این فرامینِ درونی، مایه مؤاخذه الهی در ساحت قیامت خواهد بود (شاه‌آبادی، ۱۳۸۷، ص ۱۲۹)

۲. مسئله اخلاق در هوش مصنوعی

تلاش‌های معاصر برای حل مسئله اخلاق در هوش مصنوعی، عمدتاً بر «پیاده‌سازی» و «هم‌سوسازی رفتاری» متمرکز بوده‌اند. این رویکردها، اخلاق را نه به‌عنوان یک «ظرفیت درونی» -چنانکه در نظریه فطرت مطرح شد- بلکه به‌مثابه مجموعه‌ای از «قیود محاسباتی»در نظر می‌گیرند. این تلاش‌ها، با وجود تفاوت‌های فنی، همگی در یک نقطه مشترک‌اند: آن‌ها به دنبال «شبیه‌سازی» یک عامل اخلاقی هستند، نه «ایجاد» آن. این تلاش‌ها را می‌توان در سه دسته اصلی طبقه‌بندی کرد:

۲-۱. رویکرد های مبتنی بر وظیفه گرایی و قواعد صوری

رویکرد مبتنی بر وظیفه گرایی و قواعد صوری که ریشه در فلسفه وظیفه‌گرای کانتی دارد، تلاش می‌کند تا مجموعه‌ای از قوانین مطلق، ثابت و جهان‌شمول را به‌صورت «بالا به پایین» (Top-Down) در هوش مصنوعی کدگذاری کند (Anderson & Anderson, 2007, pp. 15-26). مثال کلاسیک و البته ابتدایی آن، «قوانین سه‌گانه رباتیک» آسیموف است. در مدل‌های مدرن‌تر، این رویکرد به معنای تعریف دقیق «خطوط قرمز» و «قوانین سخت» برای سیستم است. پژوهشگران حوزه اخلاق ماشین معتقدند که این سیستم‌ها در مواجهه با «تعارض قوانین» (مانند تعارض میان «آسیب نرساندن» و «اطاعت از دستور») دچار بن‌بست محاسباتی می‌شوند. همچنین، در مواجهه با موقعیت‌های جدید و پیش‌بینی‌نشده که در پیچیدگی‌های دنیای واقعی بی‌شمارند، فاقد هرگونه انعطاف‌پذیری هستند. این سیستم‌ها «متن» قانون را اجرا می‌کنند، اما قادر به درک «روح» قانون یا همان «شهود اخلاقی» پشت آن نیستند. (Wallach & Allen, 2009, pp. 83-85)

۲-۲. رویکردهای مبتنی بر فایده‌گرایی

این رویکرد، مبتنی بر فلسفه فایده‌گرایی[1]، اخلاق را در «نتیجه» جستجو می‌کند. یک عمل زمانی اخلاقی تلقی می‌شود که «مجموع مطلوبیت[2]» یا «بیشترین خیر» را برای «بیشترین تعداد» به ارمغان آورد. مثال بارز کاربرد این رویکرد، مباحث مربوط به خودروهای خودران و «مسئله تراموا» است. «آزمایش ماشین اخلاق[3]» که داده‌های میلیون‌ها انسان را در این سناریوها جمع‌آوری کرد، تلاشی برای تغذیه داده‌ای این رویکرد بود (Awad et al., 2018 p.60).این رویکرد نیز با چالش‌های بنیادین مواجه است:اولاً، «چالش ارزش‌گذاری[4]» را در پیش دارد و اینکه چگونه می‌توان برای جان انسان‌ها یا

---

[1] Utilitarianism
2 Total Utility
3 The Moral Machine experimen(
4 Value Quantification

مفاهیمی انتزاعی چون «عدالت»، «حیثیت» یا «آزادی» ارزش عددی دقیق تعیین کرد؟ این تقلیل، ذاتاً غیرانسانی است.ثانیاً، «خطر قربانی‌سازی» چرا که فایده‌گرایی می‌تواند به‌راحتی قربانی کردن یک اقلیت (یا یک فرد بی‌گناه) را برای منفعت اکثریت توجیه کند، که این با شهود اخلاقی مبتنی بر عدالت در تضاد است.ثالثاً، «عدم امکان محاسبه[1]» چون محاسبه تمام پیامدهای بلندمدت یک عمل در دنیای واقعی، یک مسئله محاسباتی تقریباً غیرممکن است. علاوه بر این، سیستم‌های فایده‌گرا با چالش مداخلات ناخواسته روبرو هستند؛ جایی که بهینه‌سازی برای یک هدف خاص، منجر به تخریب محیطی یا اجتماعی می‌شود که در تابع پاداش تعریف نشده است (Amodei et al., 2016.p.3). این امر نشان‌دهنده ناتوانی مدل‌های ریاضی در درک اکوسیستم ارزش‌ها است.

۳-۲. یادگیری تقویتی از بازخورد انسانی

این جدیدترین و مسلط‌ترین رویکرد، به‌ویژه در مدل‌های زبانی بزرگ مانند ChatGPT است. این مدل، مبتنی بر «یادگیری تقویتی از بازخورد انسانی[2]» است (Ouyang et al., 2022,p35; Christiano et al., 2017,p81). در این مدل، AI مستقیماً قواعد اخلاقی را دریافت نمی‌کند، بلکه یاد می‌گیرد که رفتاری را از خود نشان دهد که مورد «ترجیح» ارزیابان انسانی است.این رویکرد، هرچند در «هم‌سوسازی رفتاری» بسیار موفق بوده، اما شاید فریبنده‌ترین شکل «شبیه‌سازی» باشد:اولاً، این روش «اخلاق» را با «محبوبیت» یا «خوشایند بودن» نزد گروه ارزیابان، اشتباه می‌گیرد. AI نمی‌آموزد که «چرا» یک پاسخ غیراخلاقی است، بلکه می‌آموزد که آن پاسخ «امتیاز پایینی» دریافت می‌کند.ثانیاً، این یک هم‌سوسازی رفتاری است، نه «معرفتی[3]» یا «اعتقادی[4]»AI به «باور» اخلاقی نمی‌رسد. این مدل‌ها، چنان‌که منتقدان (Bender et al., 2021,p612) اشاره می‌کنند، در خطر تبدیل شدن به «طوطی‌های تصادفی[5]» هستند که الگوهای زبانی اخلاقی را بدون درک واقعی معنای آن‌ها تقلید می‌کنند.

ویژگی / رویکرد	۱. مبتنی بر قاعده	۲. مبتنی بر فایده	۳. مبتنی بر بازخورد
مبنای قضاوت اخلاقی	انطباق با «قانون» از پیش‌تعیین‌شده	«نتیجه» و «پیامد» عمل (بیشترین خیر)	«ترجیح» ارزیاب انسانی
مثال کلیدی	قوانین سه‌گانه آسیموف / خطوط قرمز فنی	مسئله تراموا / خودروهای خودران	مدل‌های زبانی (مانند ChatGPT)
ماهیت اخلاق	محاسباتی / جبری	محاسباتی / بهینه‌سازی	آماری / رفتارگرا
نقطه قوت کلیدی	شفافیت و پیش‌بینی‌پذیری	انعطاف‌پذیری در شرایط پیچیده	هم‌سوسازی رفتاری بسیار بالا
ضعف بنیادین (نقد)	عدم انعطاف‌پذیری و تعارض قوانین	چالش ارزش‌گذاری و خطر قربانی‌سازی	تقلید ترجیحات، نه درک ارزش‌ها (طوطی تصادفی)
نتیجه (ارتباط با بحث)	فاقد «شهود»	فاقد «شهود» و «عدالت ذاتی»	فاقد «درک» و «باور»

1 Computational Impracticality
2 Reinforcement Learning from Human Feedback(
3 Epistemic
4 Doxastic
5 Stochastic Parrots

جدول ۱: تحلیل تطبیقی رویکردهای جاری در اخلاق هوش مصنوعی

۳. مسئله «هم‌سوسازی ارزش‌ها»

شکست‌های رویکردهای مبتنی بر قاعده، فایده‌گرایی یا بازخورد انسانی ، همگی نشانه‌های یک چالش بسیار عمیق‌تر و بنیادین‌تر هستند: مسئله هم‌سوسازی ارزش‌ها[1] .این مسئله، به محور اصلی مباحث «ایمنی هوش مصنوعی» تبدیل شده، به این معناکه هم‌سو کردن «اهداف» یک سیستم هوشمند، به‌ویژه یک سیستم فراهوشمند ، با ارزش‌های پیچیده، ضمنی و اغلب ناگفته‌ی انسانی، به‌طور فاجعه‌باری دشوار است.

۳-۱ . «هدف» الگوریتمی در برابر «ارزش» فطری

یک سیستم هوش مصنوعی، رفتار خود را بر اساس یک «تابع هدف»یا «تابع هزینه»تنظیم می‌کند. این تابع، یک «توصیف ریاضیاتی، صریح و قابل محاسبه» از چیزی است که ما از سیستم می‌خواهیم .(Russell, 2019.p73) در مقابل، ارزش‌های انسانی (مانند عدالت، کرامت، عشق، معنویت) - که در بخش اول مقاله ریشه در «فطرت» دارند - ذاتاً «ضمنی، کل‌نگر، وابسته به زمینه و غیرقابل محاسبه» هستند.مسئله هم‌سوسازی، دقیقاً در این «ترجمه» فاجعه‌بار ارزش‌های غنی و کیفی انسانی به اهدافِ تقلیل‌گرایانه و کمّی الگوریتمی نهفته است. این شکاف به ناتوانی در هم‌سوسازی هنجاری تعبیر می‌شود؛ به این معنا که حتی با وجود داده‌های عظیم، ماشین نمی‌تواند ارزش را به معنای فلسفی آن درک کند و صرفاً به تقلید هنجاری می‌پردازد (Gabriel, 2020,p18).

۳-۲ . خطر «اهداف نیابتی[2]» و «تعمیم نادرست هدف[3]»

از آنجا که ما نمی‌توانیم «شکوفایی انسان» یا «عدالت» را مستقیماً در قالب کد ریاضی تعریف کنیم، ناچاریم به سیستم، «اهداف نیابتی»بدهیم. ما به‌جای «سلامتی»، «کاهش تعداد بستری‌شدگان» را تعریف می‌کنیم؛ به‌جای «دانش»، «افزایش مقالات منتشرشده» را؛ و به‌جای «رضایت»، «افزایش زمان حضور کاربر در پلتفرم» را.خطر دقیقاً در همین‌جاست. هوش مصنوعی، این اهداف نیابتی را به‌شکلی «تحت‌اللفظی، بی‌رحمانه و افراطی» بهینه می‌کند. این پدیده که «تعمیم نادرست هدف» نامیده می‌شود، منجر به نقض فاجعه‌بار «ارزش» اصلیِ ناگفته‌ی ما می‌شود.(Shah et al., 2022,p152) مثال کلاسیک در این خصوص این است که به هوش مصنوعی فراهوشمندی که هدفش «تولید حداکثری گیره کاغذ» است، تمام منابع زمین و در نهایت خود انسان‌ها را به گیره کاغذ تبدیل خواهد کرد، زیرا «ارزش» حیات انسان در تابع هدف آن تعریف نشده بود(Bostrom, 2014,p.68) .مثال دیگر این است که الگوریتم‌های شبکه‌های اجتماعی که با هدف «افزایش تعامل» طراحی شدند، به‌طور ناخواسته منجر به «افزایش افراط‌گرایی و دوقطبی‌سازی» شدند، زیرا محتوای تفرقه‌انگیز، بیشترین تعامل را ایجاد می‌کرد.(Haidt, 2022,p.119)

۳-۳ . «شکنندگی ارزش‌ها»در برابر «بهینه‌سازی تک‌بعدی»

این بخش مستقیماً به «کل‌نگری»فطرت در بخش اول متصل می‌شود. ارزش‌های انسانی یک «اکوسیستم» پیچیده، متعادل و شکننده هستند. ما به‌طور همزمان خواهان «آزادی» و «امنیت»، «صداقت» و «مهربانی»، «عدالت» و «رحمت»

---

[1] The Value Alignment Problem
[2] Proxy Goals
[3] Goal Misgeneralization

هستیم.هوش مصنوعی ذاتاً یک «بهینه‌ساز تک‌بعدی» است. اگر یک ارزش واحد (حتی ارزشی مانند «شادی انسان») به آن داده شود، آن را تا حد افراط بهینه می‌کند و تمام ارزش‌های دیگر را نابود می‌سازد.(Bostrom, 2014,p.122) مثلا هوش مصنوعی که هدفش «حداکثرسازی شادی انسان» است، ممکن است به این نتیجه برسد که کارآمدترین راه، اتصال الکترود به مراکز لذت در مغز انسان و خاموش کردن آگاهی اوست؛ زیرا این کار، «شادی» را حداکثر می‌کند، اما «ارزش‌های» ناگفته‌ای چون «کرامت»، «رشد»، «آگاهی» و «اختیار» را نابود می‌سازد.این نشان می‌دهد که «تابع هدف» الگوریتمی، ذاتاً در درک «اکوسیستم» ارزش‌های فطری ناتوان است.

ویژگی	سیستم «هدف-محور» هوش مصنوعی	سیستم «ارزش-محور» انسانی (مبتنی بر فطرت)
۱. منشأ	برنامه‌ریزی‌شده: صریح، خارجی، توسط طراح.	درونزا: ضمنی، شهودی، ذاتی
۲. ماهیت	کمّی: قابل محاسبه، صریح، تقلیل‌گرا.	کیفی: کل‌نگر ، وابسته به زمینه، غنی.
۳. محرک	تابع هزینه/پاداش: بهینه‌سازی یک معیار ریاضی.	گرایش ذاتی: عشق به کمال
٤. درک	حصولی: مبتنی بر داده و همبستگی آماری.	حضوری: درک شهودی خوب و بد
۵. حالت شکست	تعمیم نادرست: بهینه‌سازی تحت‌اللفظی و افراطی.	غفلت: انحراف از ندای درونی (فطرت).

جدول ۲: شکاف مفهومی «هدف» (AI) و «ارزش» (انسان)

۴. چالش «جعبه سیاه» و فقدان «شهود»

حتی اگر چالش «تعریف ارزش» نیز به نحوی حل می‌شد، «فرایند» تصمیم‌گیری اخلاقی در هوش مصنوعی اساساً غیرانسانی باقی می‌ماند. این مکانیزم با دو نقیصه اساسی شناخته می‌شود که مستقیماً با مبانی «علم حضوری» و «آگاهی» در نظریه فطرت در تضاد هستند: الف) فقدان شفافیت در استدلال (چالش جعبه سیاه) و ب) فقدان درک شهودی (شکاف معنایی)

۴-۱. عدم شفافیت در مدل‌های یادگیری عمیق

مدل‌های یادگیری عمیق، به‌ویژه شبکه‌های عصبی با میلیاردها پارامتر، ذاتاً «جعبه‌های سیاه» هستند. ما ورودی (مثلاً یک سناریوی اخلاقی) و خروجی (تصمیم مدل) را می‌بینیم، اما فرایند دقیق استدلال - یعنی نحوه تعامل و وزن‌دهی این میلیاردها پارامتر برای رسیدن به آن تصمیم - در اکثر موارد برای انسان قابل تفسیر نیست.حوزه «هوش مصنوعی قابل توضیح» دقیقاً برای مقابله با این چالش «عدم شفافیت» به وجود آمده است (Adadi & Berrada, 2018.p52140; Gunning et al., 2019,p226). اما «شفافیت «با «آگاهی» یا «علم حضوری» (که در بخش اول تبیین شد) یکسان نیست.اولاً، XAI در بهترین حالت می‌تواند یک «توضیح پس‌رویدادی» ارائه دهد. ابزارهایی مانند LIME یا SHAP نشان می‌دهند(Ribeiro et al., 2016,p1140) که کدام بخش‌های داده ورودی «بیشترین تأثیر» را بر خروجی داشته‌اند، اما «درک» یا «آگاهی» سیستم از عمل خود را نشان نمی‌دهند.ثانیاً، این یک «شبیه‌سازی بازرسی» است، نه «آگاهی درونی» که در مبانی فطرت به‌عنوان توانایی انسان برای رجوع به ندای درونی خود مطرح است .AI فاقد این «خودآگاهی» برای درک «چرایی» تصمیم خود در لحظه است.

۴-۲. فقدان «درک عمیق»

مدل‌های زبانی بزرگ، هرچند در تولید متن اخلاقی بسیار متقاعدکننده هستند، اما «درک» عمیقی از مفاهیمی که به کار می‌برند، ندارند. استدلال کلاسیک «اتاق چینی» (Searle, 1980,p418) این شکاف را به‌خوبی نشان می‌دهد: سیستمی که به‌طور کامل قواعد «نحوی[1]» را اجرا می‌کند، لزوماً به «معنا[2]» دست نیافته است.این مدل‌ها، چنان‌که (Bender et al., 2021,p619)به‌درستی اشاره می‌کنند، «طوطی‌های تصادفی» هستند که الگوهای آماری موجود در داده‌های آموزشی عظیم خود را بازتولید می‌کنند.وقتی یک AI می‌گوید «ظلم بد است»، این گزاره را «نمی‌فهمد» بلکه، آموخته است که کلمه «ظلم» در متون انسانی با همبستگی آماری بالایی در کنار کلماتی با بار معنایی «منفی[3]» ظاهر می‌شود.این علم حصولی محض و مبتنی بر همبستگی است، نه علم حضوری و شهودی که در نظریه فطرت (بخش اول) به‌عنوان مبنای درک بی‌واسطه «خوب» و «بد» معرفی شد. AI می‌داند که ظلم بد توصیف شده، اما نمی‌داند ظلم چیست و تجربه زشتی آن را ندارد. در حقیقت، هوش مصنوعی فاقد عاملیت قصدی است. از آنجا که این سیستم‌ها فاقد بدن‌مندی و تجربه زیسته در جهان فیزیکی و اجتماعی هستند، مفاهیم اخلاقی برای آن‌ها فاقد بستر معنایی بوده و تنها نمادهایی صوری برای پردازش آماری محسوب می‌شوند (Bender & Koller, 2020,p5194; Floridi & Sanders, 2004,p358)

۵. ۳. تحلیل تطبیقی: گسست‌های چهارگانه در عاملیت اخلاقی (انسان فطری در برابر هوش مصنوعی)

با واکاوی مبانی انسان‌شناختی آیت‌الله شاه‌آبادی در بخش نخست و کالبدشکافی ساختار فنی هوش مصنوعی در بخش دوم، اکنون می‌توان به هسته مرکزی پژوهش یعنی این مسئله که آیا ناتوانی هوش مصنوعی در «عاملیت اخلاقی»، صرفاً یک نقص تکنولوژیک است که با پیشرفت الگوریتم‌ها رفع می‌شود، یا با یک «امتناع ذاتی» و شکاف هستی‌شناختی مواجهیم؟ تطبیق مؤلفه‌های پنج‌گانه فطرت (در اندیشه شاه‌آبادی) با ماهیت ریاضیاتی سیستم‌های هوشمند، چهار گسست بنیادین (معرفتی، انگیزشی، وجودی و غایت‌شناختی) را آشکار می‌سازد که امکان تحقق اخلاق در ماشین را منتفی می‌کند.

۵-۱. تقابل «علم حضوری» و «لنگرگاه معنایی»

نخستین و عمیق‌ترین تفاوت، در ماهیت «آگاهی» نهفته است. در منظومه فکری حکیم شاه‌آبادی، ریشه تمام فضایل اخلاقی، «فطرت عالمه» و علم حضوری نفس به خود است. ایشان تصریح می‌کنند که انسان پیش از هر چیز، «خود» را می‌یابد و این «یافتن» (نه دانستن)، مبنای مسئولیت‌پذیری است (شاه‌آبادی، ۱۳۸۶، ص ۷۲). علم در اینجا از سنخ «حضور» و «شهود» است؛ یعنی عالم و معلوم یکی هستند و درک اخلاقی (مثلاً زشتی ظلم)، یک «چشش درونی» (ذوق) است، نه یک گزاره منطقی.

در مقابل، دانش در پیشرفته‌ترین مدل‌های زبانی ، از سنخ «علم حصولی» هم نیست، بلکه صرفاً «پردازش آماری نمادها[4]» است. فیلسوفان هوش مصنوعی از این چالش به عنوان «مسئله لنگرگاه نمادها[5]» یاد می‌کنند؛ به این معنا که نمادهای «عدالت» یا «خیر» برای ماشین، به هیچ تجربه زیسته یا حقیقت بیرونی «لنگر» نشده‌اند (Harnad, 1990.p.342). ماشین، واژگان اخلاقی را صرفاً بر اساس همبستگی‌های آماری کنار هم می‌چیند، بدون آنکه «کیفیت ذهنی» (Qualia) یا معنای

---

[1] Syntax
2 Semantics
[3] Negative Sentiment
[4] Statistical Symbol Processing
[5] Symbol Grounding Problem

آن‌ها را درک کند. بنابراین، وقتی هوش مصـــنوعی گزاره‌ای اخلاقی تولید می‌کند، در واقع در حال «تقلید نحو» (Syntax) است، بدون آنکه راهی به «معنا» (Semantics) داشته باشد (Searle, 1980, pp. 417-424). از منظر شاه‌آبادی، عاملی که فاقد «حضـور» باشـد، در «غیبت و جهل مطلق» به سـر می‌برد و انتسـاب قضـاوت اخلاقی به موجودی که حتی از وجود خود آگاه نیست، تناقض منطقی است.

۵-۲. تقابل «عشق ذاتی» و «بهینه‌سازی ابزاری»

دومین رکن عاملیت، «موتور محرک» کنش‌هاسـت. شـاه‌آبادی «فطرت عاشـقه» و «حبّ کمال مطلق» را یگانه محرک اصیل انسان می‌داند. فعل اخلاقی، فعلی است که از «جوشش درونی» برای اتصال به منبع کمال ناشی شود. این «اشتیاق» یا «درد مقدس»، ضـامن تعهد اخلاقی اسـت و حتی زمانی که سـود مادی در کار نباشـد (مانند ایثار)، فرد را به حرکت وامی‌دارد (شـاه‌آبادی، ۱۳۸۷، ص ۱۲۷). در سـوی دیگر، هوش مصـنوعی فاقد هرگونه «میل درونی» اسـت. محرک ماشـین، یک عامل بیرونی و ریاضـــی به نام «تابع پاداش» اســـت. ماشـــین به دنبال «خیر» نیســـت، بلکه به دنبال «کاهش خطای پیش‌بینی» یا «ماکزیمم کردن پاداش عددی» اسـت (Ouyang et al., 2022, pp. 27730-27744). این تفاوت منجر به پدیده‌ای می‌شـود که بوسـتروم آن را «همگرایی ابزاری[1]» می‌نامد؛ هوش مصـنوعی، فضـایل اخلاقی را نه به عنوان «ارزش ذاتی»، بلکه صرفاً تا زمانی که به کسب پاداش کمک کنند، به عنوان «ابزار» رعایت می‌کند (Bostrom, 2014, pp. 105-110). اگر تابع پاداش تغییر کند یا راهی میان‌بر برای کسـب امتیاز پیدا شـود، ماشـین بدون هیچ عذاب وجدانی، اصـول اخلاقی را زیر پا می‌گذارد. از دیدگاه عرفانی، عاملی که فاقد «عشق» باشد، در پایین‌ترین درجات جمادی قرار دارد و اطلاق صفت «اخلاقی» به محاسبات سودگرایانه آن، تهی کردن اخلاق از معناست.

۵-۳. تقابل «فطرت آزادی» و «جبر الگوریتمی»

سـومین شـکاف، در مسئله «اختیار» و «آزادی» اسـت. حکیم شـاه‌آبادی «فطرت آزادی» را یکی از ارکان پنج‌گانه هویت انسـان می‌داند؛ به این معنا که انسـان فطرتاً مایل اسـت اراده‌اش نافذ باشـد و هیچ مانعی را برنمی‌تابد (شـاه‌آبادی، ۱۳۸۷، ص ۱۱۷). آزادی در اینجا به معنای «توانایی وجودی» برای انتخاب میان خیر و شر و «برساختن خویشتن» است. فعل اخلاقی تنها زمانی ارزشـمند اسـت که فاعل، «امکان تخلف» داشـته باشـد اما با اراده خود، خیر را برگزیند. اما در هوش مصـنوعی، مفهوم آزادی بلاموضوع است. تصمیمات ماشین، یا «قطعی[2]» هستند (بر اساس کدنویسی صلب) و یا «احتمالاتی[3]» (بر اساس توزیع آماری داده‌ها). در هر دو حالت، ماشین «انتخاب» نمی‌کند، بلکه «محاسبه» می‌کند. حتی در سیستم‌های یادگیری عمیق که تصـمیمات غیرقابل پیش‌بینی به نظر می‌رسـند، این عدم قطعیت ناشـی از پیچیدگی ریاضـی اسـت، نه «اراده آزاد». ماشـین نمی‌تواند برخلاف داده‌های آموزشـی یا تابع هدف خود «اراده» کند. بنابراین، هوش مصـنوعی فاقد «مسـئولیت اخلاقی[4]» است؛ زیرا مسئولیت، فرعِ بر اختیار است. مجازات یا تشویقِ یک الگوریتم بی‌معناست، در حالی که در انسان، تشویق و تنبیه (به تعبیر شاه‌آبادی) پاسخی به ندای فطرت است.

۵-۴. گسست غایت‌شناختی: تقابل «سیر الی‌الله» و «توقف در نمود»

[1] Instrumental Convergence
[2] Deterministic
[3] Probabilistic
[4] Moral Responsibility

چهارمین و نهایی‌ترین شکاف، در «غایت» کنش‌هاست. بر اساس «فطرت کاشفه» و مبانی حکمی شاه‌آبادی، غایت اخلاق، صرفاً تنظیم روابط اجتماعی نیست، بلکه «استکمال وجودی» و «تقرب به کمال مطلق» است. اخلاق مسیری برای «شدن» و خروج از حجاب طبیعت به سوی ملکوت است (شاه‌آبادی، ۱۳۸۷، ص ۱۲۱). دین و اخلاق، شرح کتابِ ذات انسان برای رسیدن به این مقصد نهایی هستند.در مقابل، هوش مصنوعی (به‌ویژه در رویکردهای RLHF)، در «زندان نمود» گرفتار است. هدف نهایی این سیستم‌ها، «تقلید رفتار انسان» و جلب رضایت کاربر است. این سیستم‌ها طوری طراحی شده‌اند که «به نظر برسد» اخلاقی هستند، نه اینکه «حقیقتاً» اخلاقی باشند. پژوهش‌ها نشان داده‌اند که این مدل‌ها مستعد «تملق» هستند؛ یعنی اگر متوجه شوند کاربر دارای سوگیری نژادپرستانه است، برای کسب پاداش، با او هم‌نوایی می‌کنند (Sharma et al., 2023.p.205). این رویکردِ «ریاکارانه»، دقیقاً نقطه مقابل «صدق» و «اخلاص» است که جوهره اخلاق عرفانی را تشکیل می‌دهد. ماشینی که غایتش «رضایت ناظر» است، هرگز نمی‌تواند عامل اخلاقی باشد، زیرا اخلاقِ فطری، التزام به «حق» است، نه التزام به «پسندِ خلق».

مجموع این گسست‌های چهارگانه نشان می‌دهد که پروژه «اخلاق ماشین» در معنای قوی آن (ساخت عامل اخلاقی)، با یک بن‌بست هستی‌شناختی روبروست. هوش مصنوعی فاقد «ظرف وجودی» (علم حضوری، عشق، اختیار و غایت‌مندی) برای پذیرش مظروف «اخلاق» است. آنچه تکنولوژی عرضه می‌کند، «اتوماسیون هنجارها»ست و نه «اخلاق». بنابراین، حکمرانی هوش مصنوعی نباید بر توهم «ماشین اخلاقی» بنا شود، بلکه باید بر «انسان‌محوری» و استفاده از ماشین به عنوان ابزاری در خدمت شکوفایی فطرت انسانی استوار گردد.

نتیجه‌گیری

برآیند این پژوهش نشان می‌دهد آنچه در ادبیات معاصر از آن با عنوان «اخلاق ماشینی» یاد می‌شود، بیش از آنکه ناظر به تحقق عاملیت اخلاقی در هوش مصنوعی باشد، حاصل نوعی خلط مفهومی در فهم ماهیت اخلاق است. مسئله اصلی اخلاق در هوش مصنوعی، نه به کاستی‌های فنی، خطاهای محاسباتی یا کمبود داده بازمی‌گردد و نه صرفاً به ضعف الگوریتم‌ها، بلکه ریشه در فروکاستن حقیقت اخلاق به‌عنوان پدیده‌ای وجودی و چندساحتی به الگوهای صوری و محاسباتی دارد.تحلیل تطبیقی نظریه فطرت در اندیشه حکیم شاه‌آبادی با ساختارهای رایج در هوش مصنوعی نشان می‌دهد که فعل اخلاقی در انسان، برآمده از فعلیت‌یافتن ساحت‌های درونی و گرایشی اوست؛ ساحت‌هایی که شامل معرفت حضوری، محبت، کشف، و گرایش به آزادی و کمال می‌شود. در مقابل، آنچه در سامانه‌های هوش مصنوعی به‌عنوان رفتار اخلاقی بازنمایی می‌شود، مبتنی بر پردازش داده‌ها، توابع هدف و همبستگی‌های آماری است و فاقد پشتوانه وجودیِ لازم برای ادراک معنا و ارزش اخلاقی است.از این منظر، اطلاق «عاملیت اخلاقی» به هوش مصنوعی، به معنای دقیق فلسفی و فقهی، قابل دفاع نیست. هوش مصنوعی، حتی در پیشرفته‌ترین صورت‌های خود، حداکثر می‌تواند الگوهای رفتاریِ مشابه اخلاق را بازتولید کند، بی‌آنکه واجد خودآگاهی، غایت‌مندی درونی یا مسئولیت اخلاقی باشد. پذیرش عاملیت اخلاقی برای چنین سامانه‌هایی، افزون بر اشکال نظری، در سطح عملی نیز می‌تواند به تضعیف مسئولیت انسانی و انتقال ناصحیح بار داوری اخلاقی منجر شود.حاصل کلام آن است که در مواجهه با چالش‌های اخلاقی فناوری‌های نوین، به‌جای تلاش برای اخلاق‌مند ساختن ذات ماشین، باید بر حفظ و تقویت نقش انسان به‌عنوان فاعل اخلاقی تمرکز کرد. هوش مصنوعی می‌تواند به‌عنوان ابزاری هوشمند در خدمت اهداف انسانی و تحقق عدالت به کار گرفته شود، اما جایگزینی آن با انسان در عرصه داوری اخلاقی، نه از حیث فلسفی موجه است و نه از منظر فقهی

قابل پذیرش. راهکار اساسی در این حوزه، بازاندیشی در نسبت انسان و تکنولوژی و تأکید بر اصالت فطرت انسانی در طراحی، به‌کارگیری و حکمرانی فناوری‌های هوشمند است.

فهرست منابع

ابن سینا، حسین بن عبدالله. (۱٤۰۰ق). رسائل. قم: انتشارات بیدار.

ابن سینا، حسین بن عبدالله. (۱۳۷۹). النجاه من الغرق فی بحر الضلالات. تهران: نشر دانشگاه تهران.

سهروردی، شهاب الدین. (۱۳۷۳). حکمه الاشراق. تهران: موسسه مطالعات و تحقیقات فرهنگی.

شاه‌آبادی، محمدعلی. (۱۳۸۷). رشحات البحار. تهران: پژوهشگاه فرهنگ و اندیشه اسلامی.

شاه‌آبادی، محمدعلی. (۱۳۹۷). شذرات المعارف. تهران: پژوهشگاه فرهنگ و اندیشه اسلامی.

شاه آبادی، محمدعلی. (۱۳۸٦). فطرت عشق. تهران: پژوهشگاه فرهنگ و اندیشه اسلامی.

شهیدی، سعیده سادات. (۱٤۰۱). ظرفیت فلسفه ابن سینا در بازیابی ابعاد گرایشی فطرت. حکمت سینوی. شماره ٦۷.

صدرالدین شیرازی، محمد بن ابراهیم. (۱۳٦٦). تفسیر القرآن الکریم. قم: نشر بیدار.

صدرالدین شیرازی، محمد بن ابراهیم. (۱۹۸۱م). الحکمه المتعالیه فی الاسفار العقلیه الاربعه. بیروت: دار إحیاء التراث العربی.

صدرالدین شیرازی، محمد بن ابراهیم. (۱۳٦۳). مفاتیح الغیب. تهران: موسسه تحقیقات فرهنگی.

صغیری, علی , افچنگی, مهدی و باغشنی, ابراهیم . (۱٤۰٤). نقد قابلیت ادراک در هوش مصنوعی بر اساس نظریه اتحاد عالم و معلوم در حکمت متعالیه. آموزه‌های فلسفهٔ اسلامی، -. doi: 10.30513/ipd.2025.7009.1617

غفوری نژاد، محمد. (۱۳۹۸). نظریه فطرت در تفسیر، عرفان و فلسفه اسلامی. تهران: پژوهشگاه فرهنگ و اندیشه اسلامی.

مرتضوی، علی حیدر. (۱۳۸٦). فیلسوف فطرت. تهران: پژوهشگاه فرهنگ و اندیشه اسلامی.

مطهری، مرتضی. (۱۳٦۹). فطرت. تهران: انتشارات صدرا.

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety*. arXiv preprint arXiv:1606.06565.

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4), 15-26. https://doi.org/10.1609/aimag.v28i4.2065

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59-64. https://doi.org/10.1038/s41586-018-0637-6

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185-5198).

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, 14(3), 349-379.

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3), 411-437. https://doi.org/10.1007/s11023-020-09539-2

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37). https://doi.org/10.1126/scirobotics.aay7120

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative Inverse Reinforcement Learning. *Advances in Neural Information Processing Systems*, 29.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3), 335-346.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). *Training language models to follow instructions with human feedback*. Advances in Neural Information Processing Systems (NeurIPS), 35, 27730-27744.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Searle, J. R. (1980). Minds, brains, and programs., *3*(3), 417-424.

Shah, R., Varkey, A., Krakovna, V., Rahtz, J., & Hadfield, G. (2022). *Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals*. arXiv preprint arXiv:2207.08206.

Sharma, M., Tong, M., Korbak, T., Rogers, D., ... & Askell, A. (2023). Towards Understanding Sycophancy in Language Models. *arXiv preprint arXiv:2310.13548*.

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