



Feasibility of “Ethics” in Artificial Intelligence: An Analysis Based on the Principles of *Fiṭrah* in the Philosophical-Mystical Thought of Ayatollah Shāhābādī

Sa'ideh Sādāt Shahīdī¹ | Sayyid 'Alī Mīrlohi²

1. Corresponding Author, Assistant Professor, Department of Islamic Philosophy and Theology, Imam Sadiq University (Women's Campus), Tehran, Iran. Email: shahidi@isu.ac.ir

2. Ph.D. Student in Jurisprudence and Law, Shahid Motahari University and Higher School, Tehran, Iran. Email: mirlohi@motahari.ac.ir



Article Info

Article type:

Research Article

Received: 1 January 2026

Accepted: 20 January 2026

Available Online: 21 February 2026

Keywords:

Ethics of artificial intelligence, *fiṭrah*, Ayatollah Shāhābādī, philosophical mysticism, moral agency.



Abstract

A significant strand of contemporary discourse on artificial intelligence ethics tends to reduce moral judgment to computable rules and formalised procedures. Approaches such as utilitarian optimisation 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 Ḥakīm Shāhābādī and employs his theory of *fiṭrah* (innate human disposition) as an anthropological and evaluative framework. An analysis of this theory reveals that moral action, in Shāhābādī's view, is not the result of formal rule-following or external compliance, but emerges from the actualisation of multiple inner and intuitive dimensions of human nature, including presential 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 optimisation, 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 behaviour 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.

Cite this article: Shahīdī, S., S; Mīrlohi, S. A. (2026). Feasibility of “Ethics” in Artificial Intelligence: An Analysis Based on the Principles of *Fiṭrah* in the Philosophical-Mystical Thought of Ayatollah Shāhābādī. *Islamic Philosophical Doctrines*, 21(38), 281-314. <https://doi.org/10.30513/ipd.2026.8031.1705>





Extended Abstract

Introduction

The dominant discourse surrounding “Artificial Intelligence Ethics” in recent years has remained largely confined to technical, computational, and legal frameworks. With the advent of Large Language Models (LLMs) and generative AI, which possess an astonishing ability to simulate empathy and moral reasoning, the tendency toward the “anthropomorphism” of machines has intensified. Prevailing Western paradigms attempt to resolve the ethical challenges of these intelligent systems through the “alignment” of algorithms with legal regulations, utilitarian values, or deontological principles. Scenarios such as the “Trolley Problem” in autonomous vehicles exemplify this reductionism, where profound, qualitative ethical concepts are reduced to quantitative mathematical equations and the optimisation of objective functions.

However, these approaches overlook a fundamental ontological question: can a system that inherently lacks intuitive understanding, consciousness, lived experience, and an intrinsic inclination toward perfection genuinely possess the capacity for moral agency? The current crisis in this field is far more than a technical deadlock regarding output control; it is a conceptual crisis regarding the feasibility of realising ethics in a non-human entity. State-of-the-art mechanisms such as Reinforcement Learning from Human Feedback (RLHF), while somewhat successful in aligning model behaviour with human preferences, have entirely failed to instill genuine moral belief, intention, or commitment. These models merely learn to reproduce social norms without grasping their intrinsic value.

This research aims to fill the gap in contemporary Islamic philosophical literature regarding technological challenges by turning to the theory of innate disposition (*fiṭrah*) in the mystical-philosophical thought of Ayatollah Shāhābādī. The primary objective is to move beyond the superficial “control of machine behaviour” and delve into the machine’s existential capacity to embody ethics. The central question is: based on Ḥakīm Shāhābādī’s anthropological principles, can AI achieve responsible moral agency, or will it forever remain a mechanical simulator of ethics?

Method

This research was conducted using a descriptive-analytical method with an interdisciplinary approach, intersecting Islamic philosophy and computer science. In the first step, through library and documentary research—specifically examining key foundational works of Ayatollah Shāhābādī such as *Rashāhāt al-baḥr* and *Shadharāt al-ma‘ārif*—the components of the theory of *fiṭrah*, the levels of the soul, and the pillars of moral agency were extracted and formulated into a conceptual framework. In the second step, by reviewing authoritative and recent literature in AI and the philosophy of technology



(focusing on sources from 2016 to 2024), the technical structure of intelligent systems—including deep neural networks, reinforcement learning algorithms, and rule-based approaches—was dissected. Finally, utilising comparative analysis and dialectical rational argumentation, the findings from these two contrasting domains were juxtaposed to deduce the fundamental gaps between the *fiṭrah*-endowed human and the algorithm-driven machine.

Findings

This perspective understands ethics not as a product of social conventions or cost-benefit calculations, but as something rooted in the very ontological structure of the human being. It distinguishes between the “natural human,” bound by instincts, and the “primordial (*fiṭrī*) human,” oriented toward perfection, and identifies five intuitive dimensions of human nature: immediate self-awareness as the ground of responsibility; an innate love of absolute perfection as the inner motive of moral action; a pre-discursive perception of good and evil; a desire for ultimate tranquillity through perfection; and a tendency toward freedom from the determinism of nature. Accordingly, a moral act is realised when the three human faculties—sense, intellect, and heart—are aligned with the three fundamental realities of justice, knowledge, and devotion.

The analytical section then explains the limitations of prevailing machine-ethics paradigms. Rule-based (deontological) approaches are fragile because machines cannot grasp the spirit of rules and face rule conflicts; utilitarian systems suffer from goal mis-generalisation and unintended interventions; and explainable AI provides only post-hoc accounts that cannot substitute for human intuitive understanding. Likewise, models trained with reinforcement learning from human feedback primarily reproduce linguistic patterns without genuine access to meaning. Ultimately, the comparison reveals four fundamental gaps between humans and machines: an epistemic gap (immediate self-knowledge versus statistical representation), a motivational gap (inner love versus reward functions), an ontological gap (free will versus algorithmic determinism), and a teleological gap (existential perfection versus behavioural simulation). Hence, human morality appears intrinsic and purpose-oriented, whereas machine morality remains instrumental and imitative.

Conclusion

The present study demonstrates that attributing “moral agency” to artificial intelligence stems from an epistemological error, confusing the simulation of a phenomenon with the reality of that phenomenon. Based on the criterion of *fiṭrah* in Ayatollah Shāhābādī’s school of thought, AI suffers from an intrinsic impossibility of becoming ethical due to its lack of an existential vessel (comprising presential knowledge, the agony of love, free will, and immaterial teleology). What is currently presented as machine ethics is, at best, the automation of social norms and, at worst, algorithmic deception. These findings carry profound implications for technology policy and jurisprudence: first, the illusion of constructing autonomous moral machines must be discarded, as



a machine can never bear the heavy burden of agency; second, delegating ethical and legal judgments to algorithms is rejected from a philosophical standpoint and is highly dangerous from a societal perspective. The ultimate solution in technological governance lies in adhering to the Human-in-the-Loop principle and redefining AI strictly as a smart tool serving the flourishing of human *fiṭrah*, rather than an independent, judging agent.

Author Contributions

Both authors have collaborated with each other until the final version was compiled.

Data Availability Statement

This research is based exclusively on published philosophical, mystical, and technical sources. No new empirical data were generated. All sources are fully cited in the reference list.

Acknowledgements

The authors thank the journal's reviewers and editorial board for their evaluation and constructive comments.

Ethical Considerations

In writing this article, all standards and principles of research ethics, including originality of the text, confidentiality in references, and prevention of plagiarism, have been fully observed.

Funding

This research was conducted at personal expense and received no specific financial support from public, commercial, or non-profit funding bodies.

Conflict of Interest

The authors declare that there are no conflicts of interest.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

No artificial intelligence was used in the production of scientific content, idea generation, and formulation of the main arguments of this research. However, the authors declare, based on the standards of publication ethics, that in the final stages, they used the Gemini artificial intelligence assistant exclusively for linguistic editing, integrating the text structure, and specialised translation of the detailed abstract into English. All outputs have been carefully reviewed, corrected, and finally approved by the authors.



امکان سنجی «اخلاق» در هوش مصنوعی؛ تحلیلی مبتنی بر مبانی «فطرت» در نگرش حکمی آیت الله شاه‌آبادی

✉ سعیده سادات شهیدی^۱ | سید علی میرلوحی^۲

۱. نویسنده مسئول، استادیار، گروه فلسفه و کلام اسلامی، دانشگاه امام صادق علیه السلام، پردیس خاوران، تهران، ایران. رایانامه: shahidi@isu.ac.ir

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

چکیده

در بخش قابل توجهی از ادبیات معاصر اخلاقی 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 به جای تمرکز بر کارآمدی یا ناکارآمدی این الگوها در سطح اجرا، به موضوع ریشه‌ای‌تر می‌پردازد و بررسی می‌کند آیا 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 - فلسفی حکیم شاه‌آبادی، نظریه «فطرت» را به مثابه مبانی انسان‌شناختی و معیار سنجش اخلاقی تحلیل می‌کند. بررسی این نظریه نشان می‌دهد که فعل اخلاقی در نگاه ایشان، نه حاصل امتثال صوری از قاعده، بلکه نتیجه فعلیت یافتن 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 تجربه حضوری، از چنین مبانی 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.

استناد: شهیدی، سعیده سادات؛ میرلوحی، سید علی. (۱۴۰۵). امکان سنجی «اخلاق» در هوش مصنوعی؛ تحلیلی مبتنی بر مبانی «فطرت» در نگرش حکمی آیت الله شاه‌آبادی. آموزه‌های فلسفه اسلامی، ۲۱(۳۸)، ۲۸۱-۳۱۴.

<https://doi.org/10.30513/ipd.2026.8031.1705>

© نویسندگان.

ناشر: دانشگاه علوم اسلامی رضوی.



اطلاعات مقاله

نوع مقاله: پژوهشی

تاریخ دریافت: ۱۴۰۴/۱۰/۱۱

تاریخ پذیرش: ۱۴۰۴/۱۰/۳۰

تاریخ انتشار برخط: ۱۴۰۴/۱۲/۰۲

کلیدواژه‌ها

اخلاق هوش مصنوعی، فطرت، آیت الله شاه‌آبادی، نگرش حکمی، عاملیت اخلاقی.





مقدمه

گفتمان غالب درباره «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 قابلیت دستیابی به «عاملیت اخلاقی حقیقی» را دارد؟ بحران کنونی در این حوزه فراتر از یک بن‌بست فنی در چگونگی هم‌سوسازی، یک بحران مفهومی در «امکان سنجی» 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 حقیقت «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 معاصر 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 به‌واسطه برخورداری از «فطرت الهی»، استعدادی سرشتی برای درک و گرایش به الزامات اخلاقی دارد (صغیری و همکاران، ۱۴۰۴، ص. ۱۰). 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 به عنوان



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 «شبهه» 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 می‌سازند 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 پیش‌استدلالی انسان‌اند.

در نهایت، هیچ یک از این رویکردها به «ظرفیت ذاتی» برای اخلاق، مشابه مفهوم «فطرت»، نپرداخته‌اند. پژوهش حاضر با روش تحلیلی - توصیفی و رویکردی بینارشته‌ای، درصدد بررسی



این موضوع از دیدگاه حکیم شاه آبادی، آن هم به آن دلیل است که وی به عنوان فیلسوف شاخص مسلمان، بر نقش فطرت به عنوان مبدأ درونی معرفت و گرایش اخلاقی تأکید دارد.

۱. حقیقت فطرت از دیدگاه حکیم شاه آبادی

نظریه فطرت به عنوان یکی از کلیدی ترین مباحث انسان شناسی قرآنی - حکمی، نقشی اساسی در حل چالش های معرفتی، اخلاقی و تربیتی ایفا می کند. این نظریه با بررسی براهین اثبات، اقسام دانش ها و گرایش های فطری، بستری برای تحلیل نسبت فطرت با قوای نفس فراهم می آورد (غفوری نژاد، ۱۳۹۸، ص. ۲۵). در این عرصه، حکیم شاه آبادی جایگاهی ویژه دارد. او 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 ریشه های شهودی اخلاق نزد این حکیم وارسته استوار است. بدین منظور، ضمن تبیین تعاریف و اقسام فطرت در منظومه فکری او، پیشینه این بحث نزد سایر اندیشمندان شاخص حکمت اسلامی نیز مرور می گردد تا جایگاه ممتاز شاه آبادی در تبیین گرایش و عاملیت اخلاقی انسان روشن شود.

واژه «فطرت» دلالت بر خلقت ویژه و ساختار وجودی مشترک میان تمامی انسان ها دارد. این مفهوم ریشه در آموزه های وحیانی قرآن کریم دارد. شهید مطهری بر این باور است که این اصطلاح پیش از نزول قرآن سابقه ای نداشته و برای نخستین بار در این کتاب آسمانی برای توصیف سرشت بشری به کار رفته است. این مفهوم از یک سو، ساحتی انسان شناختی دارد و به بن مایه های وجودی بشر می پردازد و از سوی دیگر، پیوند تکوینی میان خالق و مخلوق را تبیین می کند (مطهری، ۱۳۶۹، ص. ۱۴). بر مبنای آموزه های قرآنی، نوع بشر دارای ذاتی غیراقتسابی و تغییرناپذیر است که او را از سایر انواع حیوانی متمایز می سازد. ادراکات، گرایش ها و توانمندی های مختص بشری همگی از شئون این ویژگی فطری محسوب می شوند. 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 آن که به عنوان بحثی انسان شناختی مستقل مطرح باشد، ذیل مباحث خداشناسی و گرایش سرشتی به مبدأ تحلیل شده است. ابن سینا 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



رسالة العشق مطرح می‌کند، چراکه مراد او از واژه فطرت در برخی آثارش، با معنای اصطلاحی و قرآنی آن تفاوت دارد (ابن سینا، ۱۳۷۹، ص. ۱۱۵-۱۱۸). با این حال، شیخ الرئیس در آثار عرفانی‌اش 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 زیارویان و عشق حقیقی به باری تعالی، بهره‌مند است که استعداد آن در اصل خلقت بشر نهاده شده است (ابن سینا، ۱۴۰۰، 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 مدبر 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 عالی 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 جای جای آثار فلسفی و تفسیری خود، به تحلیل حقیقت، اقسام و هدایتگری فطرت پرداخته است. ملاصدرا فطرت را قوه و استعداد صفات و ملکاتی می‌داند که خداوند در بدو خلقت در جان انسان سرشته است، به گونه‌ای که استعداد هر فضیلت و رذیلت مقابل آن، در وجود فرد به ودیعه نهاده شده است. او برای تبیین این حقیقت، از تعابیری چون «فطرت اصلی»، «فطرت ربانی» و «اصل الفطرة» 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 حق تعالی 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» می‌نامد و معتقد است دوری از این مسیر، تنها بر 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 ارائه تفصیلی عالمانه از نظریه فطرت و دسته‌بندی اقسام آن، مبنایی استوار برای ریشه‌یابی شهود اخلاقی بنا نهاد. از منظر او، 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، ص. ۱۱۵). او در آثار خویش، بارها از این فرایند درون‌نگرانه با تعبیری همچون مطالعه «کتاب ذات»، واکاوی «کتاب وجود» و یا مکاشفه فطری یاد کرده است (شاه‌آبادی، ۱۳۸۶، ص. ۹۵، ۱۰۹، ۱۲۷ و ۱۳۱). از منظر ریشه‌شناختی، حکیم شاه‌آبادی واژه «فطرت» را بر وزن «فعله» از ماده «فطر» تحلیل می‌کند که دلالت بر کیفیت و نوع ایجاد دارد: «از آن جاکه وجود و ایجاد حقیقتی واحدند، پس کیفیت ایجاد حق که با هویت ما مساوی است، همان فطرت و صفات لازمه وجود ماست (شاه‌آبادی، ۱۳۸۷، ص. ۱۱۴).

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 (شاه‌آبادی، ۱۳۸۶، ص. ۲۱). در منظومه فکری حکیم، «کتاب ذات» انسان به قلم پروردگار نگاشته شده است و به همین سبب، قوای وهم و خیال راهی به حریم آن ندارند. از سوی دیگر، شاه‌آبادی با استناد به روایت نبوی «خلق الله آدم علی صورته»، معتقد است که کتاب ذات انسان با کتاب ذات حق انطباق کامل دارد و از این رو، فطرت در احکام ذاتی خود هرگز دچار خطا نمی‌شود (شاه‌آبادی، ۱۳۸۷، ص. ۱۱۵). حکیم شاه‌آبادی مقتضای فطرت را کاملاً در راستای دین و در جهت کمال‌بخشی به انسان تبیین می‌کنند (شاه‌آبادی، ۱۳۸۷، ص. ۹۲). البته او هشدار می‌دهد که انسان باید همواره نسبت به این اقتضائات متفطن باشد، و در غیر این صورت، در حجاب طبیعت گرفتار می‌شود و دچار غفلت می‌گردد (شاه‌آبادی، ۱۳۸۷، ص. ۱۱۵). تفتن به این لایه‌های پنهان، مستلزم نوعی درون‌نگری است که طی آن، انسان با شهودی درونی به ساختار حقیقی وجودش 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، درحالی‌که آگاهی ما به سایر موجودات از سنخ علم حصولی است (شاه‌آبادی، ۱۳۸۶، ص. ۷۲). این وجود آگاه متعلق به عالم امر است و سنخیتی با عالم خلق و طبیعت ندارد، زیرا مناط علم حضور است، اما در ساحت طبیعت، حضوری متصور نیست و آن لایه تنها ظرف غیبت و جهل است. از نظر او، این کشف که «خودبینی» نامیده می‌شود، نخستین هدف انبیای الهی بوده است، چراکه نقشی حیاتی در خروج انسان از اسارت طبیعت ایفا می‌کند. با این خودشناسی، انسان به وضوح درک می‌کند که حقیقتش با بدن و ماده یکی نیست (شاه‌آبادی، ۱۳۸۷، ص. ۱۱۶).



دیگر ارجاعات شهودی نیز در معرفت به خود نقش ایفا می‌کنند (شاه‌آبادی، ۱۳۸۶، ص. ۷۲). دومین رکن «فطرت عاشقه» است. انسان با رجوع به این ساحت درمی‌یابد که فطرتاً عاشق خود و کمالاتش است. حکیم از این درک به «خودخواهی» تعبیر می‌کند. بر این پایه، حتی شخصی که دست به خودکشی می‌زند، به سبب حب ذات چنین می‌کند تا از زندگی سخت رها شود و به آرامش برسد. بعد سوم «فطرت کاشفه» است که در آغاز تمام کارها و افکار انسانی تجلی دارد. انسان هرگز به سمت امری حرکت نمی‌کند مگر آن‌که نسبت به نفع یا ضرر آن به صورت فطری آگاهی (کشف) پیدا کرده باشد. همین ویژگی سبب می‌شود که انسان به راحتی نظر دیگران را نپذیرد و «تقلید» را مخالف شالوده فطرت بداند، زیرا دین با فطرت هماهنگ است و تقلید را در اصول نمی‌پذیرد. چهارمین عامل در خودشناسی، رجوع به «فطرت حب راحت» است. همه ابنای بشر عاشق آسایش مطلق‌اند و برای رسیدن به آن، حتی رنج‌های بزرگی را به جان می‌خورند. 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 (شاه‌آبادی، ۱۳۸۷، ص. ۱۱۶ و ۱۱۷). از دیدگاه حکیم، مناظ فطری آزادی، توانایی بر عملی ساختن مقاصد فردی است (شاه‌آبادی، ۱۳۸۶، ص. ۱۰۳). با تجمیع این پنج ارجاع درونی، انسان به معرفتی جامع از حقیقتش دست می‌یابد؛ او موجودی است که به خود علم حضوری دارد، عاشق خود است، قدرت تأمل دارد، طالب آسایش است و در پی آزادی است. از آن‌جا که هیچ یک از این ویژگی‌ها در عالم طبیعت یافت نمی‌شوند، انسان به این نتیجه قطعی می‌رسد که: «فلتحکم بأن حقیقتی غیر طبیعتی»؛ پس حکم کن که حقیقت من غیر از طبیعت من است (شاه‌آبادی، ۱۳۸۷، ص. ۱۱۷).

حکیم شاه‌آبادی با تکیه بر این دریافت‌ها، مخاطب را به اقرار به مجرد حقیقت وجودی خویش مجاب می‌کند. افزون بر این موارد، او ابعاد دیگری از فطرت را نیز مطرح می‌کند که در شهودی دانستن اخلاق نقشی بسزا دارند. از جمله این ابعاد، «فطرت بغض نقص و حب کمال» است. بر اساس این ویژگی سرشتی، انسان از هر گونه کاستی اکراه دارد و همواره به دنبال موجودات کامل است. اگر در مواردی، انسان به اشتباه امری ناقص را کمال بیندارد، پس از آگاهی، بلافاصله رویگردان می‌شود و کمال مطلق را می‌جوید. همین گرایش درونی است که



در نهایت، انسان را به سوی موجود مطلق سوق می دهد، که حکیم از آن با عنوان «فطرت حبّ اصل» یاد کرده است (شاه آبادی، ۱۳۸۷، ص. ۲۲۱ تا ۲۲۴).

۱-۱. فطرت به مثابه «ریشه شهود اخلاقی»

حکیم شاه آبادی با نگاهی عمیق به ذومراتب بودن حقیقت وجودی انسان در مسیر استكمال، میان دو مرتبه «انسان طبیعی» و «انسان فطری» تمایز قائل می شود. از منظر ایشان، انسان در بدو پیدایش و مراحل نخستین زیست خود، انس و الفتی تمام عیار با عالم طبیعت دارد. او در این مرتبه، همتش را مصروف توجه به مادیات می کند و همچون سایر حیوانات، در پی تأمین نیازهای غریزی و تحقق آرزوهای مادی در ساحت ماده است. در این ساحت طبیعی، بشر ویژگی های بنیادین خود نظیر خود دوستی، ابراز رأی، راحت طلبی و میل به آزادی را تسلیم محض طبیعت می کند و به ناچار، طبیعت برای وی در جایگاه امری محبوب، تصمیم گیر و منشأ آزادی تلقی می شود (شاه آبادی، ۱۳۸۷، ص. ۱۱۸). در چنین وضعیتی، انسان از تکاپو برای تحصیل کمالات برتر و لوازم معنوی آن باز می ماند و به آسودگی کاذب در حجاب ماده تن می دهد (شاه آبادی، ۱۳۸۷، ص. ۱۱۸). به همین سبب، حکیم شاه آبادی بر ضرورت «اعراض از طبیعت» تأکید می ورزد. مراد او از این اعراض، نه نفی کامل ساحت مادی، بلکه بریدن از نگاه استقلال بی طبیعت است، چراکه طبیعت در نگاه حکیم، صرفاً نقش وساطت و ابزاری را در حصول کمالات اصیل انسانی بر عهده دارد (شاه آبادی، ۱۳۸۷، ص. ۱۲۰). این تغییر نگرش موجب می شود انسان به وجه ملکوتی خویش و ضرورت سیر از عالم ملک به ملکوت متنبه شود و از ماندگاری در حجاب طبیعت بپرهیزد. در این سیر کمالی، فرد از خودخواهی های محدود مادی و آزادی های بدنی گذر می کند و بر اساس فطرت «حب اصل کمال» و «نفرت از نقص»، به کاستی هایش پی می برد و به سوی مرتبه نهایی کمال حرکت می کند. حکیم از این تحول وجودی به «اقامه وجه به سوی دین» تعبیر می نماید (شاه آبادی، ۱۳۸۷، ص. ۱۲۱). در منظومه فکری حکیم شاه آبادی، هدایت انسان بر انطباق کامل میان فطرت و دین استوار است. انسان به اقتضای ترکیب وجودی اش از جسم و روح، دارای دو وجه متناظر است، اما حقیقت اقامه وجه به سوی دین مستقیماً به ساحت معنوی و ادراکی او باز می گردد. شاه آبادی وجوه ادراکی بشر را در سه سطح «وجه حس»، «وجه عقل» و «وجه قلب» تبیین می کند. وجه حس، مسئول ادراک محسوسات از طریق حواس پنج گانه در عالم ماده است. وجه عقل راهی برای ادراک معقولات است که در مرتبه ضعیف، از مسیر تفکر و برهان عبور می کند



و در مرتبه 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 معلوم او ذات خودش است و همین ادراک مایه دوست‌داری خود و کمالاتش می‌شود. بشر در آغاز، کمالات محسوس را درک می‌کند و به آن‌ها دلبسته می‌شود، اما با ارتقای ادراکی، کمالات معقول فطری را نیز می‌یابد. در این مرحله، فرد به سوی دین روی می‌آورد و متوجه کمال مطلق می‌گردد. او درمی‌یابد که عشق و اشتیاقش نامتناهی است، درحالی‌که موجودات عالم متناهی‌اند و شایستگی معشوق مطلق بودن را ندارند. لذا خداوند که کمال و جمال مطلق است، تنها گزینه سزاوار برای عشق و عبودیت انسان قرار می‌گیرد (شاه‌آبادی، ۱۳۸۷، ص. ۱۲۴ تا ۱۲۷).

از منظر آیت‌الله 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 حقیقت بنیادین می‌یابد: معرفت، عبودیت و عدالت. در این تقسیم‌بندی، توجه به عدالت و انجام اعمال نیک معادل اقامه وجه حس به سوی دین است. طلب معرفت اقامه وجه عقل، و عبودیت و تخلق به اخلاق الهی تجلی اقامه وجه قلب است. بر این اساس، ادعای بی‌نیازی انسان از دین مردود است، زیرا بشر در ذاتش دین‌دار آفریده شده است و کتاب‌های آسمانی و پیامبران صرفاً شارحان همان کتاب تکوینی نهفته در ذات انسان‌اند. فطرت در مطالبه 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 تمام حقایق عالم و مبدأ و معاد تعلق می‌گیرد. انسان فطرتاً از جهل‌گریزان است و این عشق به دانستن چنان است که اگر بر تمام حقایق زمین آگاه شود، باز مشتاق درک خیرات 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 حوزه عدالت، فطرت خواهان آن است که همه انسان‌ها به ادای حقوق یکدیگر پایبند باشند. با اقامه 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: «إن العبودية هي خضوع خاص بباب المعبود» (شاه‌آبادی، ۱۳۸۷، ص. ۱۶۱).

فطرت به وجوبِ مطلقِ خضوع حکم کرده است که ساحت‌های قلبی، زبانی و جوارحی را در بر می‌گیرد. 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. نکته ظریف در این جا نسبی بودن خضوع است که میان خاضع (فطرت انسان) و مخدوم رخ می‌دهد، به‌گونه‌ای که این احترام در سیری تشکیکی به تمام موجودات دارای کمال تعلق می‌گیرد و در نقطه 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 کند و در قلب خاضع شود، اما به دلایلی چون حسد، تقيه 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 منزله «روح» اعمال است که به اعضا و جوارح متعدد 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 یابند و قلب را مستعد تجلی اسما و صفات الهی گردانند (مرتضوی، ۱۳۸۶، ص. ۷۱). حکیم تأکید دارد که احترام به حاضر، حتی اگر کودک یا دشمن باشد، احترام به منعم، مقتدر و محبوب، همگی ریشه در ذات بشر دارند، به‌طوری که حتی در حیوانات نیز مرتبه‌ای از احترام به منعم دیده می‌شود (شاه‌آبادی، ۱۳۸۷، ص. ۱۲۹). بر اساس همین اخلاقیات فطری و تخلف از این احکام سرشتی است که خداوند در قیامت، کسانی را که از تعظیم و احترام سزاوار اعراض کرده‌اند، مؤاخذه خواهد کرد (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 او، بشر در نخستین مراحل حیاتش در پیوند و الفتی تنگاتنگ با ساحت مادی و عالم طبیعت است. در این مرتبه، تمام 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 بند تعلقات مادی گرفتار می‌آید و طبیعت به اشتباه 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 می‌سازد 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 او از این اعراض به هیچ رو نفی کامل ساحت مادی یا رهبانیت نیست، بلکه بریدن از نگاه استقلال‌ی به عالم ماده است. در دیدگاه او، 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 می‌یابد و از ماندگاری در لایه‌های زیرین هستی می‌پرهیزد. در این سیر استکمالی، فرد از خودخواهی‌های بدنی عبور می‌کند و بر پایه فطرت «حب کمال» و «نفرت از نقص»، به سوی کمال مطلق حرکت می‌کند؛ تحولی که حکیم از آن به «اقامه 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 است، اما حقیقت دین‌داری او ریشه در ساحت معنوی و ادراکی‌اش دارد. شاه‌آبادی قوای ادراکی بشر را در سه سطح تبیین می‌کند: «وجه حس» 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 خویش‌شن و کمالاتش نائل می‌آید. نخستین معلوم هر فرد ذات اوست که به «علم حضوری» درک می‌شود و همین خودآگاهی بستر خوددوستی و کمال‌خواهی را فراهم



می آورد. انسان در ابتدا به کمالات محسوس دلبسته می شود، اما با ارتقای وجودی، درمی یابد که اشتیاق او نامتناهی است و موجودات مادی به دلیل تناهی، شایسته محبویت مطلق نیستند. در این جاست که جان آگاه تنها پروردگار را که کمال و جمال مطلق است، گزینه سزاوار برای عشق و عبودیت برمیگزیند (شاه آبادی، ۱۳۸۷، ص. ۱۲۴ تا ۱۲۷). بر این پایه، دین در نگاه شاه آبادی چیزی جز «التزام به حقایق فطری» نیست. هر انسانی با رجوع به «کتاب ذات» خود، سه حقیقت بنیادین را می یابد: معرفت، عدالت و عبودیت. در این تقسیم بندی، عدالت تجلی وجه حس، معرفت تجلی وجه عقل و عبودیت تجلی وجه قلب در ساحت دین داری است. لذا ادعای بی نیازی از دین باطل است، زیرا انسان در ذاتش دین دار آفریده شده است و وحی و نبوت تنها شارحان همان کتاب تکوینی نهفته در فطرت اند. این سه رکن به نحو مطلق و بدون استثنا مورد مطالبه فطرت قرار دارند (شاه آبادی، ۱۳۸۷، ص. ۱۲۷). برای مثال، میل به معرفت چنان در جان انسان نهفته است که اگر بر تمام حقایق زمین آگاه شود، باز مشتاق درک کمالات عوالم دیگر است و دین دقیقاً در راستای همین میل بی کران به تبیین حقایق ملکوت می پردازد (شاه آبادی، ۱۳۸۷، ص. ۱۳۱). در حوزه عدالت نیز فطرت خواهان ادای حقوق همگانی است. با اقامه وجه حس، فرد ملزم به شناخت حقوق (اعم از حق الله و حق الناس) و حرکت بر مدار مصالح می شود. در واقع، احکام شرعی پاسخی به این نیاز درونی بشرند (شاه آبادی، ۱۳۸۷، ص. ۱۲۹). اما عالی ترین سطح این التزام، خضوع در برابر موجود کامل است که اوج آن در «عبودیت» متجلی می شود. حکیم تصریح دارد که: «إِنَّ الْعِبَادِيَّةَ هِيَ خُضُوعٌ خَاصٌّ بِبَابِ الْمَعْبُود» (شاه آبادی، ۱۳۸۷، ص. ۱۶۱).

این خضوع فطری شامل ساحت های قلبی، زبانی و جوارحی است و نشانه های آن را می توان در احترامات فطری به «منعم»، «مقتدر»، «کامل» و «محبوب» مشاهده کرد. این پیوند سرشتی چنان است که حتی در برابر کمال یک کافر نیز نوعی تواضع ذاتی در جان انسان پدید می آید و نقطه اوج این سلسله مراتب، عبودیت مختص پروردگار است (شاه آبادی، 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، علی رغم درک عظمت یک کمال در قلبش، در ظاهر از خضوع امتناع ورزد. با این حال، ظهور خضوع قلبی در برابر کمال، امری انکارناپذیر و سرشتی است (شاه آبادی، ۱۳۸۷، ص. ۲۱۵). در نهایت، عبودیت از منظر شاه آبادی، «روح» تمام اعمال و خصلت های پسندیده است. فضایل اخلاقی زمانی به کمال واقعی شان می رسند 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 است و تخلف از این فرامین 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 به عنوان «ظرفیت درونی» چنان‌که در نظریه فطرت مطرح شد، بلکه به مثابه مجموعه‌ای از «قیود محاسباتی» در نظر می‌گیرند. این تلاش‌ها با وجود تفاوت‌های فنی، همگی در یک نقطه مشترک‌اند: آن‌ها به دنبال «شبیه‌سازی» یک عامل اخلاقی‌اند، نه «ایجاد» آن. این تلاش‌ها را می‌توان در سه دسته اصلی طبقه‌بندی کرد:

۱-۲. رویکردهای مبتنی بر وظیفه‌گرایی و قواعد صوری

رویکرد مبتنی بر وظیفه‌گرایی و قواعد صوری که ریشه در فلسفه وظیفه‌گرایی کانتی دارد، تلاش می‌کند تا مجموعه‌ای از قوانین مطلق، ثابت و جهان‌شمول را به صورت «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 پیش‌بینی نشده که در پیچیدگی‌های دنیای واقعی بی‌شمارند، فاقد هر گونه انعطاف‌پذیری‌اند. 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،^۱ اخلاق را در «نتیجه» جست‌وجو می‌کند. یک عمل

1. Utilitarianism



زمانی اخلاقی تلقی می‌شود که «مجموع مطلوبیت»^۲ 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^۳ 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 زیر مواجه است:

اولاً «چالش ارزش‌گذاری»^۴ را در پیش دارد و این‌که چگونه می‌توان برای جان انسان‌ها یا مفاهیمی انتزاعی چون «عدالت»، «حیثیت» یا «آزادی» ارزش عددی دقیق تعیین کرد؟ این تقلیل ذاتاً غیرانسانی است.

ثانیاً «خطر قربانی‌سازی»؛ چراکه فایده‌گرایی می‌تواند به راحتی 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»^۵ چون محاسبه تمام پیامدهای بلندمدت یک عمل در دنیای واقعی، مسئله محاسباتی‌ای تقریباً غیرممکن است. به علاوه، سیستم‌های فایده‌گرا با چالش مداخلات ناخواسته روبه‌رویند؛ 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 ارزش‌هاست.

۲-۳. 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»^۶ است (Ouyang et al., 2022, p. 35; Christiano et al., 2017, p. 81). در این مدل، AI مستقیماً قواعد اخلاقی را دریافت نمی‌کند، بلکه یاد می‌گیرد که رفتاری از خود نشان دهد که مورد «ترجیح» ارزیابان انسانی است. این رویکرد هرچند در «هم‌سوسازی رفتاری» بسیار موفق بوده است، اما شاید فریبنده‌ترین شکل «شبیه‌سازی» باشد:

2. Total Utility

3. The Moral Machine experiment

4. Value Quantification

5. Computational Impracticality

6. Reinforcement Learning from Human Feedback



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»^۷ یا «اعتقادی»^۸. AI به «باور» اخلاقی نمی‌رسد. این مدل‌ها چنان‌که منتقدان (Bender et al., 2021, p. 612) اشاره می‌کنند، در خطر تبدیل شدن به «طوطی‌های تصادفی»^۹ 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 پیش تعیین شده	«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	آماري / 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»

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 بنیادین‌ترند: مسئله هم‌سوسازی ارزش‌ها.^{۱۰} این مسئله به محور اصلی مباحث «ایمنی هوش مصنوعی» تبدیل شده است، به این معنا که هم‌سو کردن «اهداف»

7. Epistemic

8. Doxastic

9. Stochastic Parrots

10. The Value Alignment Problem



یک سیستم هوشمند، به ویژه سیستمی فراهوشمند، با ارزش های پیچیده، ضمنی و اغلب ناگفته انسانی، به طور فاجعه باری دشوار است.

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

یک سیستم هوش مصنوعی رفتار خود را بر اساس «تابع هدف» یا «تابع هزینه» تنظیم می کند. این تابع «توصیف ریاضیاتی، صریح و قابل محاسبه» از چیزی است که ما از سیستم می خواهیم (Russell, 2019, p. 73). در مقابل، ارزش های انسانی مانند عدالت، کرامت، عشق، و معنویت که ریشه در «فطرت» دارند، ذاتاً «ضمنی، کل نگر، وابسته به زمینه و غیر قابل محاسبه» هستند. مسئله هم سوسازی دقیقاً در این «ترجمه» فاجعه بار ارزش های غنی و کیفی انسانی به اهداف تقلیل گرایانه و کمی الگوریتمی نهفته است. این شکاف به ناتوانی در هم سوسازی هنجاری تعبیر می شود، به این معنا که حتی با وجود داده های عظیم، ماشین نمی تواند ارزش را به معنای فلسفی آن درک کند و صرفاً به تقلید هنجاری می پردازد (Gabriel, 2020, p. 18).

۲-۳. خطر «اهداف نیابتی» و «تعمیم نادرست هدف»^{۱۱}

از آن جاکه ما نمی توانیم «شکوفایی انسان» یا «عدالت» را مستقیماً در قالب کد ریاضی تعریف کنیم، ناچاریم به سیستم، «اهداف نیابتی» بدهیم. ما به جای «سلامتی»، «کاهش تعداد بستری شدگان» را تعریف می کنیم؛ به جای «دانش»، «افزایش مقالات منتشر شده»؛ و به جای «رضایت»، «افزایش زمان حضور کاربر در پلتفرم» را. خطر دقیقاً در همین جاست. هوش مصنوعی این اهداف نیابتی را به شکلی «تحت اللفظی، بی رحمانه و افراطی» بهینه می کند. این پدیده که «تعمیم نادرست هدف» نامیده می شود، منجر به نقض فاجعه بار «ارزش» اصلی ناگفته ما می شود (Shah et al., 2022, p. 152). 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 انسان ها 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 الگوریتم های شبکه های اجتماعی که با هدف «افزایش تعامل» طراحی شدند، به طور ناخواسته منجر به «افزایش افراط گرایی و دوقطبی سازی» شدند، زیرا محتوای تفرقه انگیز بیشترین تعامل را ایجاد می کرد (Germano, Gmez & Sobbrío, 2022, p. 21).

11. Proxy Goals

12. Goal Misgeneralization



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

این بخش مستقیماً به «کل‌نگری» فطرت در بخش اول متصل می‌شود. ارزش‌های انسانی یک «اکوسیستم» پیچیده، متعادل و شکننده‌اند. ما به‌طور هم‌زمان خواهان «آزادی» و «امنیت»، «صداقت» و «مهربانی»، «عدالت» و «رحمت» هستیم. هوش مصنوعی ذاتاً یک «بهینه‌ساز تک‌بعدی» است. اگر به آن یک ارزش، حتی ارزشی مانند «شادی انسان» داده شود، آن را تا حد افراط بهینه می‌کند و تمام ارزش‌های دیگر را نابود می‌سازد (Bostrom, 2014, p. 122). مثلاً هوش مصنوعی که هدفش «حداکثرسازی شادی انسان» است، ممکن است به این نتیجه برسد که کارآمدترین راه اتصال الکترونی 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	کمی: قابل محاسبه، صریح، تقلیل‌گرا.	کیفی: کل‌نگر، وابسته به زمینه، غنی.
۳. محرک	تابع هزینه/پاداش: بهینه‌سازی یک معیار ریاضی.	گرایش ذاتی: عشق به کمال
۴. 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 تضادند: الف) فقدان شفافیت در استدلال (چالش جعبه سیاه) و ب) فقدان درک شهودی (شکاف معنایی).



۱-۴. 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, p. 52140; Gunning et al., 2019, p. 226). 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, p. 1140) 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 به عنوان توانایی انسان برای رجوع به ندای درونی خود مطرح است. AI فاقد این «خودآگاهی» برای درک «چرایی» تصمیم خود در لحظه است.

۲-۴. فقدان «درک عمیق»

مدل‌های زبانی بزرگ، هرچند در تولید متن اخلاقی بسیار متقاعدکننده‌اند، اما «درک» عمیقی از مفاهیمی که به کار می‌برند، ندارند. استدلال کلاسیک «اتاق چینی» (Searle, 1980, p. 418) این شکاف را به خوبی نشان می‌دهد: سیستمی که به طور کامل قواعد «نحوی»^{۱۳} را اجرا می‌کند، لزوماً به «معنا»^{۱۴} دست نیافته است. این مدل‌ها چنان‌که به درستی اشاره می‌کنند (Bender et al., 2021, p. 619) «طوطی‌های تصادفی» هستند که الگوهای آماری موجود در داده‌های آموزشی عظیم خود را بازتولید می‌کنند. وقتی 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»^{۱۵} ظاهر می‌شود. این علم حصولی محض و مبتنی بر هم‌بستگی است، نه علم حضوری و شهودی که در نظریه فطرت (بخش اول) به عنوان مبنای درک بی‌واسطه «خوب» و «بد» معرفی شد. AI

13. Syntax

14. Semantics

15. Negative Sentiment



می‌داند که ظلم بد توصیف شده است، اما نمی‌داند ظلم چیست و تجربه زشتی آن را ندارد. در حقیقت، هوش مصنوعی فاقد عاملیت قصدی است. از آن جاکه این سیستم‌ها فاقد بدن مندی و تجربه زیسته در جهان فیزیکی و اجتماعی‌اند، مفاهیم اخلاقی برای آن‌ها فاقد بستر معنایی است و تنها نمادهایی صوری برای پردازش آماری محسوب می‌شوند (Bender & Koller, 2020, p. 5194; Floridi & Sanders, 2004, p. 358).

۵. 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 پرداخت، یعنی این مسئله که آیا ناتوانی هوش مصنوعی در «عاملیت اخلاقی» صرفاً نقصی تکنولوژیک است که با پیشرفت الگوریتم‌ها رفع می‌شود یا با یک «امتناع ذاتی» و شکاف هستی‌شناختی مواجه‌ایم؟ 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. او تصریح می‌کند که انسان پیش از هر چیز، «خود» را می‌یابد و این «یافتن» (نه دانستن)، مبنای مسئولیت‌پذیری است (شاه‌آبادی، ۱۳۸۶، ص. ۷۲). علم در این جا از سنخ «حضور» و «شهود» است، یعنی عالم و معلوم یکی‌اند و درک اخلاقی (مثلاً زشتی ظلم)، یک «جشش درونی» (ذوق) است نه گزاره منطقی. در مقابل، دانش در پیشرفته‌ترین مدل‌های زبانی، از سنخ «علم حصولی» هم نیست، بلکه صرفاً «پردازش آماری نمادها»^{۱۶} است. فیلسوفان هوش مصنوعی از این چالش به عنوان «مسئله لنگرگاه نمادها»^{۱۷} یاد می‌کنند، به این معنا که نمادهای «عدالت» یا «خیر» برای ماشین، به

16. Statistical Symbol Processing

17. Symbol Grounding Problem



هیچ تجربه زیسته یا حقیقت بیرونی «لنگر» نشده‌اند (Harnad, 1990, p. 342). ماشین واژگان اخلاقی را صرفاً بر اساس هم‌بستگی‌های آماری کنار هم می‌چیند، بی‌آن‌که «کیفیت ذهنی»^{۱۸} یا معنای آن‌ها را درک کند. بنابراین، وقتی هوش مصنوعی گزاره‌ای اخلاقی تولید می‌کند، در واقع در حال «تقلید نحو» است، بدون آن‌که راهی به «معنا» داشته باشد (Searle, 1980, pp. 417-424). از منظر شاه‌آبادی، عاملی که فاقد «حضور» باشد، در «غیبت و جهل مطلق» به سر می‌برد و انتساب قضاوت اخلاقی به موجودی که حتی از وجود خود آگاه نیست، تناقض منطقی است.

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

دومین رکن عاملیت «موتور محرک» کنش‌هاست. شاه‌آبادی «فطرت عاشقه» و «حب 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» ضامن تعهد اخلاقی است و حتی زمانی که سود مادی در کار نباشد (مانند ایثار)، فرد را به حرکت وامی‌دارد (شاه‌آبادی، ۱۳۸۷، ص. ۱۲۷). در سوی دیگر، هوش مصنوعی فاقد هر گونه «میل درونی» است. محرک ماشین عامل بیرونی و ریاضی به نام «تابع پاداش» است. ماشین به دنبال «خیر» نیست، بلکه به دنبال «کاهش خطای پیش‌بینی» یا «ماکزیم کردن پاداش عددی» است (Ouyang et al., 2022, pp. 27730-27744). این تفاوت منجر به پدیده‌ای می‌شود که بوستروم آن را «هم‌گرایی ابزاری»^{۱۹} می‌نامد؛ هوش مصنوعی فضایل اخلاقی را نه «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 کسب امتیاز پیدا شود، ماشین بدون هیچ عذاب وجدانی، اصول اخلاقی را زیر پا می‌گذارد. از دیدگاه عرفانی، عامل فاقد «عشق» در پایین‌ترین درجات جمادی قرار دارد و اطلاق صفت «اخلاقی» به محاسبات سودگرایانه آن تهی کردن اخلاق از معناست.

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

سومین شکاف در مسئله «اختیار» و «آزادی» است. حکیم شاه‌آبادی «فطرت آزادی» را یکی از ارکان پنج‌گانه هویت انسان می‌داند، به این معنا که انسان فطرتاً مایل است اراده‌اش نافذ باشد

18. Qualia

19. Instrumental Convergence



و هیچ مانعی را برنمی‌تابد (شاه‌آبادی، ۱۳۸۷، ص. ۱۱۷). آزادی در این جا به معنای «توانایی وجودی» برای انتخاب میان خیر و شر و «برساختن خویشتن» است. فعل اخلاقی زمانی ارزشمند است که فاعل «امکان تخلف» داشته باشد، اما با اراده‌اش خیر را برگزیند. اما در هوش مصنوعی، مفهوم آزادی بلاموضوع است. تصمیمات ماشین یا «قطعی»^{۲۰} هستند (بر اساس کدنویسی صلب) و یا «احتمالاتی»^{۲۱} (بر اساس توزیع آماری داده‌ها). در هر دو حالت، ماشین «انتخاب» نمی‌کند، بلکه «محاسبه» می‌کند. حتی در سیستم‌های یادگیری عمیق که تصمیمات غیرقابل پیش‌بینی به نظر می‌رسند، این عدم قطعیت ناشی از پیچیدگی ریاضی است، نه «اراده آزاد». ماشین نمی‌تواند برخلاف داده‌های آموزشی یا تابع هدف خود «اراده» کند. بنابراین، هوش مصنوعی فاقد «مسئولیت اخلاقی»^{۲۲} است، زیرا مسئولیت فرع بر اختیار است. مجازات یا تشویق یک الگوریتم بی‌معناست، درحالی‌که به تعبیر شاه‌آبادی، در انسان تشویق و تنبیه پاسخی به ندای فطرت است.

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

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 کتاب ذات انسان برای رسیدن به این مقصد نهایی‌اند. 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» اخلاقی‌اند 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 رویکرد «ریاکارانه» دقیقاً نقطه مقابل «صدق» و «اخلاص» است که جوهره

20. Deterministic

21. Probabilistic

22. Moral Responsibility



اخلاق عرفانی را تشکیل می‌دهد. ماشینی که غایتش «رضایت ناظر» است، هرگز نمی‌تواند عامل اخلاقی باشد، زیرا اخلاق فطری التزام به «حق» است، نه التزام به «پسند خلق». مجموعه این گسست‌های چهارگانه نشان می‌دهد که پروژه «اخلاق ماشین» در معنای قوی آن (ساخت عامل اخلاقی)، با بن‌بست هستی‌شناختی روبه‌روست. هوش مصنوعی فاقد «ظرف وجودی» (علم حضوری، عشق، اختیار و غایت‌مندی) برای پذیرش مظلوف «اخلاق» است. آنچه تکنولوژی عرضه می‌کند، «اتوماسیون هنجارها» است و نه «اخلاق». بنابراین، حکمرانی هوش مصنوعی نباید بر توهم «ماشین اخلاقی» بنا شود، بلکه باید بر «انسان محوری» و استفاده از ماشین به عنوان ابزاری در خدمت شکوفایی فطرت انسانی استوار گردد.

نتیجه‌گیری

برایند 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 به عنوان 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 فعلیت یافتن ساحت‌های درونی و گرایش اوست؛ ساحت‌هایی شامل معرفت حضوری، محبت، کشف، و گرایش به آزادی و کمال. در مقابل، آنچه در سامانه‌های هوش مصنوعی به عنوان رفتار اخلاقی بازنمایی می‌شود، مبتنی بر پردازش داده‌ها، توابع هدف و هم‌بستگی‌های آماری است و فاقد پشتوانه وجودی لازم برای ادراک معنا و ارزش اخلاقی. 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 رفتاری 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، به جای تلاش برای اخلاق‌مند



ساختن ذات ماشین، باید بر حفظ و تقویت نقش انسان به عنوان فاعل اخلاقی تمرکز کرد. هوش مصنوعی می‌تواند به عنوان ابزاری هوشمند در خدمت اهداف انسانی و تحقق عدالت به کار گرفته شود، اما جایگزینی آن با انسان در عرصه داورى اخلاقى، نه از حیث فلسفى موجه است و نه از منظر فقهى 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.

ملاحظات اخلاقی

حامی مالی

این پژوهش با هزینه شخصی انجام شده است.

مشارکت نویسندگان

هر دو نویسنده با همکاری متقابل تا تدوین نسخه نهایی همکاری کرده‌اند.

اعلامیه هوش مصنوعی مولد و فناوری‌های مبتنی بر هوش مصنوعی در فرایند نگارش

در تولید محتوای علمی، ایده‌پردازی و تدوین استدلال‌های اصلی این پژوهش از هوش مصنوعی استفاده نشده است. با این حال، نویسندگان بر اساس موازین اخلاق نشر اعلام می‌دارند که در مراحل پایانی، منحصراً جهت «ویرایش زبانی»، «انسجام بخشی به ساختار متن» و «ترجمه تخصصی چکیده تفصیلی به زبان انگلیسی» از دستیار هوش مصنوعی جمینی (Gemini) بهره برده‌اند. نویسندگان همه خروجی‌ها را با دقت تمام بازبینی، اصلاح و تأیید نهایی کرده‌اند.

تعارض منافع

نویسندگان اعلام می‌دارند که هیچ‌گونه تعارض منافع وجود ندارد.

پیروی از اصول اخلاق پژوهش

در نگارش این مقاله، تمام موازین و اصول اخلاق پژوهش اعم از اصالت متن، امانت‌داری در ارجاعات و جلوگیری از سرقت علمی به‌طور کامل رعایت شده است.

بیانیه دسترسی به داده‌ها

هیچ داده‌ای در دسترس نیست.



فهرست منابع

۱. ابن سینا، حسین بن عبدالله. (۱۴۰۰ق). رسائل. قم: انتشارات بیدار.
۲. ابن سینا، حسین بن عبدالله. (۱۳۷۹ش). النجاة من الغرق فی بحر الضلالات. تهران: انتشارات دانشگاه تهران.
۳. سه‌روردی، شهاب‌الدین. (۱۳۷۳ش). حکمة الاشراف. تهران: مؤسسه مطالعات و تحقیقات فرهنگی.
۴. شاه‌آبادی، محمدعلی. (۱۳۸۷ش). رشحات البحار. تهران: پژوهشگاه فرهنگ و اندیشه اسلامی.
۵. شاه‌آبادی، محمدعلی. (۱۳۹۷ش). شذرات المعارف. تهران: پژوهشگاه فرهنگ و اندیشه اسلامی.
۶. شاه‌آبادی، محمدعلی. (۱۳۸۶ش). فطرت عشق. تهران: پژوهشگاه فرهنگ و اندیشه اسلامی.
۷. شهیدی، سعیده سادات. (۱۴۰۱ش). ظرفیت فلسفه ابن سینا در بازیابی ابعاد گرایشی فطرت. حکمت سینوی، ۲۶(۶۷)، ۵۱-۷۴. <https://doi.org/10.30497/ap.2022.243368.1592>
۸. صغیری، علی؛ افچنگی، مهدی؛ و باغشینی، ابراهیم. (۱۴۰۴ش). نقد قابلیت ادراک در هوش مصنوعی بر اساس نظریه اتحاد عالم و معلوم در حکمت متعالیه. آموزه‌های فلسفه اسلامی، ۲۰(۳۶)، ۴۱-۷۲. <https://doi.org/10.30513/ipd.2025.7009.1617>
۹. غفوری‌نژاد، محمد. (۱۳۹۸ش). نظریه فطرت در تفسیر، عرفان و فلسفه اسلامی. تهران: پژوهشگاه فرهنگ و اندیشه اسلامی.
۱۰. مرتضوی، علی حیدر. (۱۳۸۶ش). فیلسوف فطرت. تهران: پژوهشگاه فرهنگ و اندیشه اسلامی.
۱۱. مطهری، مرتضی. (۱۳۶۹ش). فطرت. تهران: صدرا.
۱۲. ملاصدرا، محمد بن ابراهیم. (۱۳۶۶ش). تفسیر القرآن الکریم. قم: بیدار.
۱۳. ملاصدرا، محمد بن ابراهیم. (۱۹۸۱م). الحکمة المتعالیة فی الأسفار العقلیة الأربعة. بیروت: دار إحياء التراث العربی.
۱۴. ملاصدرا، محمد بن ابراهیم. (۱۳۶۳ش). مفاتیح الغیب. تهران: مؤسسه تحقیقات فرهنگی.
15. 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>
16. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety*. arXiv preprint arXiv:1606.06565.
17. 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>.
18. 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>.



19. 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).
20. 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>.
21. Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
22. 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.
23. Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, 14(3), 349-379.
24. Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3), 411-437. <https://doi.org/10.1007/s11023-020-09539-2>.
25. Germano, F., Gómez, V., & Sobbrío, F. (2022). Crowding out the truth? A simple model of misinformation. *polarization and meaningful social interactions*. arXiv. <https://doi.org/10.48550/arXiv.2210.02248>
26. 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>.
27. Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative Inverse Reinforcement Learning. *Advances in Neural Information Processing Systems*, 29.
28. Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3), 335-346.
29. 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.
30. 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>.
31. Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.
32. Searle, J. R. (1980). *Minds, brains, and programs.*, 3(3), 417-424.
33. 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.
34. Sharma, M., Tong, M., Korbak, T., Rogers, D., ... & Askill, A. (2023). *Towards Understanding Sycophancy in Language Models*. arXiv preprint arXiv:2310.13548.
35. Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford University Press.



Resources

1. 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>
2. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). **Concrete problems in AI safety**. arXiv preprint arXiv:1606.06565.
3. 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>
4. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, 563(7729), 59–64. <https://doi.org/10.1038/s41586-018-0637-6>
5. 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). Association for Computational Linguistics.
6. 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 (FACT '21) (pp. 610–623). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>
7. Bostrom, N. (2014). **Superintelligence: Paths, dangers, strategies**. Oxford University Press.
8. 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.
9. Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3), 349–379.
10. Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3), 411–437. <https://doi.org/10.1007/s11023-020-09539-2>
11. Ghafūrīnejād, M. (2019). **Nāzariyyeh-ye fiṭrat dar tafsīr, 'erfān va falsafeh-ye Eslāmī** (The theory of fiṭrah in interpretation, mysticism, and Islamic philosophy). Tehran: IICT Press. [In Persian]



12. 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>
13. Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 29.
14. Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3), 335–346.
15. Ibn Sīnā (Avicenna), H. ibn 'A. (1980). *Al-Najāh min al-gharaq fī baḥr al-ḍalālāt* (Deliverance from error). Tehran: The University of Tehran. [In Arabic]
16. Ibn Sīnā (Avicenna), H. ibn 'A. (2000). *Rasā'il* (Treatises). Qom: Bīdār. [In Arabic]
17. Murtazavī, A. H. (2007). *Filsūf-e feṭrat* (The philosopher of fiṭrah). Tehran: IICT Press. [In Persian]
18. Muṭahharī, M. (1990). *Feṭrat* (Primordial nature). Tehran: Ṣadrā. [In Persian]
19. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askeel, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* (NeurIPS), 35, 27730–27744.
20. 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>
21. Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.
22. Ṣaghīrī, 'A., Afchangī, M., & Baghshānī, I. (2025). Critique of perception capability in artificial intelligence based on the theory of the unity of the knower and the known in transcendent wisdom. *Islamic Philosophical Doctrines*. <https://doi.org/10.30513/ipd.2025.7009.1617> [In Persian]
23. Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424.
24. 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.
25. Shahābādī, M. 'A. (2007). *Feṭrat-e 'eshq* (The nature of love). Tehran: IICT. [In Persian]
26. Shahābādī, M. 'A. (2008). *Rashāḥāt al-baḥr* (Sprinklings of the seas). Tehran: IICT. [In Arabic and Persian]
27. Shahābādī, M. 'A. (2018). *Shadharāt al-ma'ārif* (Fragments of knowledge). Tehran: IICT. [In Arabic and Persian]
28. Shahīdī, S. S. (2022). The capacity of Avicenna's philosophy in retrieving the tendency dimensions of fiṭrah. *Avicennian Philosophy*, 67. [In Persian]
29. Sharma, M., Tong, M., Korbak, T., Du, Y., Askill, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T. T., McCandlish, S., Ndousse, K., Rausch, O., Zhang, N., Tamkin, A., Kaplan, J., & Perez, E. (2023). *Towards understanding sycophancy in language models*. arXiv preprint arXiv:2310.13548.
30. Ṣadr al-Dīn Shīrāzī (Mullā Ṣadrā), M. ibn I. (1981). *Al-Ḥikmah al-muta'āliyah fī al-asfār al-'aqliyah al-arba'ah* (The transcendent theosophy in the four intellectual journeys). Beirut: Dār Iḥyā' al-Turāth al-'Arabī. [In Arabic]
31. Ṣadr al-Dīn Shīrāzī (Mullā Ṣadrā), M. ibn I. (1984). *Mafātīḥ al-ghayb* (Keys to the unseen). Tehran: Mu'assaseh-ye Muṭāle'āt va Taḥqīqāt-e Farhangī. [In Arabic]
32. Ṣadr al-Dīn Shīrāzī (Mullā Ṣadrā), M. ibn I. (1987). *Tafsīr al-Qur'ān al-karīm* (Commentary on the holy Qur'ān). Qom: Bīdār. [In Arabic]
33. Suhrawardī, Sh. al-D. (1994). *Ḥikmat al-ishrāq* (The philosophy of illumination). Tehran: Mu'assaseh-ye Muṭāla'āt va Taḥqīqāt-e Farhangī. [In Arabic]
34. Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford University Press.