Interdisciplinary Studies in Ethics
JISE Journal of Interdisciplinary Studies in Ethics

What Artificial Intelligence Ethics Is Not — and What It Is?

Jalal Peykani

Department of Philosophy, Payam Noor University, Tehran, Iran. j_peykani@pnu.ac.ir

Abstract

Introduction: Verifying the claim that the general public, humanities experts, politicians, policymakers, and others have all turned their attention to the phenomenon called artificial intelligence is not particularly difficult. Consequently, a sort of consensus has formed regarding the decisive importance of artificial intelligence. However, people's intuitive understanding of AI largely stems from interactions with chatbots, which has led to the emergence of several significant misunderstandings. Examples of these misunderstandings will be elaborated later in the article. Beyond the purely technical aspects of the matter, there is an important dimension of AI commonly referred to as AI ethics. In the contemporary Iranian humanities domain, the literature on AI ethics is gradually taking shape. Nonetheless, due to the complexities of AI ethics and its interdisciplinary nature, signs of misunderstanding appear even at this early stage. Accordingly, this article first seeks to demonstrate what AI ethics is not, and subsequently endeavors to clarify what it is and its inherent nature. Finally, the main issues pertinent to it will be examined.

Findings: Understanding what AI actually is proves very challenging. Contrary to the impression that may arise at first glance, one cannot easily gain an intuitive and clear grasp of AI's essence. The greatest error is to imagine AI as something similar to human intelligence or to assume, after some interaction with popular chatbots, that AI equates to this. Equally mistaken is the notion that AI is merely a complex and fast computational machine similar to a computer. Understanding AI's nature requires technical knowledge, which most humanities specialists typically lack it. This is evident when we examine AI definitions: asking how AI operates in machines does not produce an intuitive mental picture because its mechanisms are entirely technical and technological. Hence, McCarthy defines AI as both a science and an engineering discipline. Thus, all components of the phrase *"AI ethics"* are complex: *"ethics"* is employed in a particular, nontrivial sense, and *AI* itself is a highly complex technology, far removed from intuitive comprehension. This leads humanities experts to face a serious initial obstacle. From a methodological standpoint, this necessitates interdisciplinary approaches, requiring collaboration with technical specialists

Cite this article: Peykani, J. (2025). What Artificial Intelligence Ethics Is Not — and What It Is?. *Interdisciplinary Studies in Ethics*, 1(1), p. 261-274. https://doi.org/10.48308/jiethics.2025.240435.1014

Received: 2024/11/16 ; **Received in revised form:** 2024/12/13 ; **Accepted:** 2025/01/08 ; **Published online:** 2025/04/09
Article type: Research Article jiethics.sbu.ac.ir

and engineers in this field. It is possible and indeed valid to discuss local issues regarding AI; however, attempting to root AI ethics within one's own intellectual tradition is a mistaken and flawed approach. Regardless of the correctness or incorrectness of this widespread tendency, it can be stated with certainty that an equivalent concept to AI ethics cannot be found in our own tradition. Any such attempt leads only to confusion and error. Interestingly, some even try to localize AI itself or impose their own frameworks and assumptions onto it. A prevalent misunderstanding is that, contrary to the intuitive and technical understanding, AI is generally conceived as an agent similar to a human but possessing a machine brain. Consequently, just as a human agent has an ethical system, AI machines are expected to have an ethical system as well. This is among the most fundamental misconceptions surrounding AI ethics. AI ethics primarily aims to reduce the risks posed by AI. The literature of AI ethics is saturated with warnings and concerns about AI-related dangers. These dangers mainly relate to human life, happiness, and well-being. However, sometimes these risks are exaggerated. Discussions on AI ethics can be initially and fundamentally divided into two categories. The first category involves purely theoretical issues, which are mostly philosophical and do not directly apply to industry or technology. The second category concerns practical issues that arise during the application and use of AI in industry and technology. Compared to the first category, these have less philosophical emphasis and embody ethical challenges encountered in the production of AI technologies or the construction of AI-based machines, thereby falling under practical and applied ethical questions related to AI.

Discussion: AI ethics is an approach seeking to construct practical guidelines within the AI domain to prevent outcomes that our ethical intuitions generally deem improper. However, for various reasons—including AI's reliance on machines and machine learning—these guidelines entail considerable technical complexities. Therefore, there exist significant differences between the conventional, even philosophical, understanding of ethics on one side, and the technically grounded comprehension of AI ethics on the other. Consequently, humanities specialists' engagement with AI ethics must be accompanied by caution, extensive knowledge of AI itself, and, if necessary, collaboration with AI experts.

Keywords: Ethics, Artificial Intelligence, AI Ethics, Philosophy of Technology.

References

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

Beauchamp, T. L., & Childress, J. F. (2019). *Principles of biomedical ethics* (8th ed.). Oxford University Press.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency, 149-159. https://doi.org/10.1145/3287560.3287583

Cave, S., & Dihal, K. (2020). The Whiteness of AI. *Philosophy & Technology*, 33(4), 685-703. https://doi.org/10.1007/s13347-020-00415-6

Crootof, R. (2015). The Killer Robots Are Here: Legal and Policy Implications. *Cardozo Law Review*, 36, 1837–1915.

Floridi, L., & Taddeo, M. (2016). What is data ethics?. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083), 20160360. https://doi.org/10.1098/rsta.2016.0360

Gunkel, D. J. (2018). *Robot Rights*. MIT Press.

Hurley, M., & Adebayo, J. (2017). Credit scoring in the era of big data. *Yale Journal of Law and Technology*, 18(1), 148-216. https://digitalcommons.law.yale.edu/yjolt/vol18/iss1/5/

Liao, Matthew. (2020). "A Short Introduction to the Ethics of Artificial Intelligence", in: Liao, Matthew. *Ethics of Artificial Intelligence*, Oxford University Press.

Miri Balajourshari, Seyedeh Mahshid, and Mahmoudi, Amir Reza. 2024. "Examining Ethical Issues in the Context of Artificial Intelligence with a View to Islamic Ethics," in: *Applied Ethics Research Quarterly*, Volume 14, Issue 6, pp. 97-123. [in Persian]

Molnar, P. (2019). Technological testing grounds: Migration management experiments and reflections from the ground up. *European Journal of Migration and Law*, 21(3), 329-352. https://doi.org/10.1163/15718166-12340054

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4), 18-21. https://doi.org/10.1109/MIS.2006.76

Ramazani, Majid, and Feyzi Derakhshi, Mohammad Reza. 2013. "Machine Ethics: Challenges and Approaches to Ethical Issues in Artificial Intelligence and Superintelligence," in: *Ethics in Science and Technology Quarterly*, Volume 8, Issue 4, pp. 1-9. [in Persian]

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347-1358. https://doi.org/10.1056/NEJMra1814259

van den Hoven, J. (2010). *The Handbook of Information and Computer Ethics*. Wiley

Zargar, Zahra. 2025. "The Relationship Between Emotions and Moral Capacity in Artificial Intelligence Technologies," in: *Philosophical Research*, Spring 2025, Issue 50, pp. 19-40. [in Persian]

JISE
Journal of Interdisciplinary Studies in Ethics
مطالعات میان‌رشته‌ای اخلاق
دانشگاه شهید بهشتی

اخلاق هوش مصنوعی چه نیست و چه هست؟

جلال پیکانی

گروه فلسفه، دانشگاه پیام نور، تهران، ایران. j_peykani@pnu.ac.ir

چکیده

در جهان امروز حوزه‌ای نو و میان‌رشته‌ای با عنوان «اخلاق هوش مصنوعی» شکل گرفته است که نشانه‌های ورود آن به فضای فکری ایران نیز به‌تدریج آشکار می‌شود؛ با این حال، سابقهٔ مواجههٔ ما با مباحث نو و همچنین پیچیدگی‌های این قلمرو تازه، موجب می‌شود که بروز برخی سوءتفاهم‌ها دربارهٔ چیستی اخلاق هوش مصنوعی قابل پیش‌بینی باشد. برخی آثار اندک و پراکنده‌ای که تاکنون در این زمینه منتشر شده‌اند نیز تا حدی بر این ادعا گواهی می‌دهند. این مقاله می‌کوشد عمدتاً به شیوه‌ای سلبی، چیستی، قلمرو، مسائل و اهمیت اخلاق هوش مصنوعی را روشن سازد. در این مقاله نخست نشان داده‌ایم که سوءتفاهم‌های متعددی در خصوص ماهیت این حوزه وجود دارد؛ برای نمونه، سوءفهم در معنای «اخلاق» در ترکیب «اخلاق هوش مصنوعی»، نسبت دادن مفاهیم شهودی و غیرعلمی به هوش مصنوعی و به تبع آن به اخلاق هوش مصنوعی، بی‌توجهی به جنبه‌های فنی و تکنیکی، و تلاش برای جست‌وجوی آن در سنت‌های علوم انسانیِ خود است.

با وجود تأکید بر جنبهٔ سلبی، از توصیف ایجابی این حوزه نیز به‌طور کامل چشم‌پوشی نشده است؛ بدین منظور، ضمن ارائهٔ تعریفی از اخلاق هوش مصنوعی و دسته‌بندی مسائل و پرسش‌های اصلی آن، کوشیده‌ایم ماهیت این قلمرو را شفاف‌تر سازیم. هدف اصلی مقاله، ارائهٔ تصویری واقع‌بینانه از اخلاق هوش مصنوعی و کمک به اصلاح برخی سوءتفاهم‌های رایج در این زمینه است. با توجه به نو بودن این موضوع در فضای فکری ایران و همراه شدن طرح آن با ابهام‌ها و سوءبرداشت‌ها، مقالهٔ حاضر بیش از آنکه در پی حل مسئله‌ای خاص باشد، درصدد ایضاح و تبیین ماهیت اخلاق هوش مصنوعی است.

کلیدواژه‌ها: اخلاق، هوش مصنوعی، اخلاق هوش مصنوعی، فلسفه‌ی فناوری.

استناد به این مقاله: پیکانی، جلال (۱۴۰۴). اخلاق هوش مصنوعی چه نیست و چه هست؟. *مطالعات میان‌رشته‌ای اخلاق*، ۱(۱)، ص۲۶۱-۲۷۴.
https://doi.org/10.48308/jiethics.2025.240435.1014

تاریخ دریافت: ۱۴۰۳/۰۸/۲۶ ؛ **تاریخ اصلاح:** ۱۴۰۳/۰۹/۲۳ ؛ **تاریخ پذیرش:** ۱۴۰۳/۱۰/۱۹ ؛ **تاریخ انتشار:** ۱۴۰۴/۰۱/۲۰

نوع مقاله: پژوهشی

jiethics.sbu.ac.ir

مقدمه

علوم انسانی از تحولات حوزه علم و فناوری تأثیر مستقیم می‌پذیرند و در بسیاری موارد نیز بر علم و فناوری اثرگذارند. در عرصهٔ فناوری، مسألهٔ اخلاق جایگاهی ویژه دارد و از همین رو قلمرویی با عنوان «اخلاق فناوری[1]» شکل گرفته است. در اخلاق فناوری، پرسش‌ها و مسائل اخلاقی مرتبط با فناوری مورد بررسی قرار می‌گیرند (van den Hoven, 2010: 5). اخلاق فناوری عمدتاً با رویکردی مبتنی بر نگرانی نسبت به پیامدهای اخلاقی نامطلوب فناوری پدید آمده است. بدیهی است که با توجه به گستردگی فناوری، قلمرو اخلاق فناوری نیز بسیار وسیع خواهد بود. در میان فناوری‌های نوظهور، هوش مصنوعی یکی از برجسته‌ترین پدیده‌هاست که توجه بسیاری را به خود جلب کرده است. هوش مصنوعی نه تنها از حیث فناورانه اهمیت بالایی دارد، بلکه از جمله فناوری‌هایی است که به سرعت توجه عموم جامعه را نیز برانگیخته است. شاید در تاریخ، کمتر فناوری‌ای را بتوان یافت که چنین سریع در میان تودهٔ مردم نفوذ یافته باشد. راستی‌آزمایی این ادعا دشوار نیست؛ زیرا امروزه عموم جامعه، متخصصان علوم انسانی، سیاستمداران، سیاست‌گذاران و دیگر گروه‌ها به پدیدهٔ هوش مصنوعی توجه نشان داده‌اند. کافی است نگاهی به فضای عمومی جامعه بیندازیم: گویی وارد کردن اصطلاح «هوش مصنوعی» در هر بحثی نشانهٔ به‌روز بودن و قرار داشتن در مدار پیشرفت است، در حالی که بی‌توجهی به آن معادل دورماندن از تحولات تلقی می‌شود؛ البته تصور غالب افراد ــ چه در میان متخصصان علوم انسانی و چه در میان عموم مردم ــ از هوش مصنوعی بیشتر به چت‌بات‌های در دسترس محدود می‌شود؛ با این حال، در اهمیت بنیادی این فناوری تردیدی وجود ندارد. تقریباً اجماعی شکل گرفته است که بخشی مهم از تحولات آینده جهان تحت تأثیر هوش مصنوعی خواهد بود. ویژگی دیگر این پدیده، ورود نسبتاً دفعی اصطلاح «هوش مصنوعی» به ادبیات عمومی است. درست پس از آنکه حدود سه سال پیش برخی سیاستمداران برجستهٔ جهان نسبت به این فناوری موضع‌گیری کردند و تقریباً هم‌زمان با عرضه چت‌بات‌ها، به‌ویژه (ChatGPT)، رسانه‌ها و سیاستمداران به‌طور مستمر بدان پرداختند؛ بدین ترتیب، اصطلاح «هوش مصنوعی» به سرعت وارد گفتمان عمومی شد؛ در حالی‌که باید توجه داشت که خودِ هوش مصنوعی به مثابه یک حوزه فناورانه از نیمهٔ دوم قرن بیستم شکل گرفته است، اما عمومی شدن اصطلاح آن و مواجههٔ تودهٔ مردم با چت‌بات‌ها سابقه‌ای بسیار کوتاه دارد.

بدین‌ترتیب نوعی اتفاق‌نظر دربارهٔ اهمیت تعیین‌کننده هوش مصنوعی شکل گرفته است؛ البته فهم شهودی عموم افراد از هوش مصنوعی عمدتاً بر اثر تجربهٔ استفاده از چت‌بات‌ها پدید آمده است و همین امر به ایجاد برخی سوءفهم‌های مهم منجر شده است. مصادیق این سوءفهم‌ها در ادامهٔ مقاله بررسی خواهد شد. در کنار جنبه‌های صرفاً فنی و تکنیکی، بُعد مهم دیگری از هوش مصنوعی وجود دارد که به

1. Technology Ethics

«اخلاق هوش مصنوعی[1]» شهرت یافته است. در مباحث رایج پیرامون هوش مصنوعی در ایران، به ندرت از مقولهٔ اخلاق هوش مصنوعی سخن به میان می‌آید. پیچیدگی این فناوری، کاربردی شدن آن، پیوند یافتنش با زندگی انسان‌ها، و آثار و پیامدهای آن بر وضعیت فردی و جمعی آدمیان ــ که به تدریج در حال گسترش و تحقق است ــ سبب شده تا مقولهٔ اخلاق با هوش مصنوعی پیوندی ناگسستنی یابد. نتیجهٔ این فرایند شکل‌گیری قلمرویی با عنوان «اخلاق هوش مصنوعی» است. اخلاق هوش مصنوعی بخشی فرعی از حوزهٔ گسترده‌تر «اخلاق فناوری» به شمار می‌آید؛ حوزه‌ای که خود ذیل فلسفهٔ فناوری قرار می‌گیرد. در فضای علوم انسانی ایران معاصر، ادبیات اخلاق هوش مصنوعی به تدریج در حال شکل‌گیری است. با این حال، به دلیل پیچیدگی‌های ذاتی و ماهیت میان‌رشته‌ای این حوزه، از همان آغاز نشانه‌هایی از سوءفهم نیز مشاهده می‌شود؛ بر این اساس، در این مقاله نخست می‌کوشیم نشان دهیم که اخلاق هوش مصنوعی چه چیزی نیست، سپس به تبیین چیستی و ماهیت آن خواهیم پرداخت و در نهایت مسائل اصلی این حوزه را بررسی خواهیم کرد. تا جایی که نگارنده جستجو کرده است، در زبان فارسی مقالات اندکی دربارهٔ اخلاق هوش مصنوعی منتشر شده است. نخستین مقالهٔ مرتبط در سال ۱۳۹۲ با عنوان «اخلاق ماشین: چالش‌ها و رویکردهای مسائل اخلاقی در هوش مصنوعی و ابرهوش» (رمضانی و فیضی درخشی، ۱۳۹۲) انتشار یافت که به معرفی کوتاهی از چیستی این حوزه و مسائل اصلی آن می‌پرداخت؛ اما در آن به سوءفهم‌های احتمالی پیرامون ماهیت اخلاق هوش مصنوعی اشاره‌ای نشده بود. مقالهٔ دیگری به‌تازگی منتشر شده که از نظر محتوایی غنی و با عنوان «رابطهٔ عواطف و ظرفیت اخلاقی در فناوری‌های هوش مصنوعی» ارائه شده است (زرگر، ۱۴۰۴)؛ با این حال، این مقالات نیز به طور مشخص موضوع مورد نظر مقالهٔ حاضر را پوشش نمی‌دهند؛ افزون بر اینها، چند مقالهٔ همایشی یا مروری نیز در این زمینه انتشار یافته است. همین کمبود منابع و شکل‌نگرفتن ادبیات جدی دربارهٔ اخلاق هوش مصنوعی به زبان فارسی نشان می‌دهد که هنوز در آغاز راه هستیم؛ از این‌رو توجه به سوءفهم‌های احتمالی می‌تواند اهمیت فراوانی داشته باشد.

۱. اخلاق هوش مصنوعی چه چیزی نیست؟

۱-۱. تعبیر «اخلاق» در اصطلاح «اخلاق هوش مصنوعی» و نسبت آن با اخلاق متعارف

در زبان انگلیسی دو واژهٔ نزدیک به هم وجود دارد که در فارسی غالباً هر دو به «اخلاق» ترجمه می‌شوند (Ethics): و (Morals)؛ (البته برخی متخصصان و مترجمان نکته‌سنج به این تفاوت توجه دارند)، همین امر منشأ برخی سوءفهم‌ها دربارهٔ ماهیت اخلاق در ادبیات علوم انسانی غربی شده است.

«اخلاق هوش مصنوعی» در واقع ترجمهٔ اصطلاح (Artificial Intelligence Ethics) است. نه تنها تعبیر (ethics) بر خلاف (morals) با فهم متعارف ما از اخلاق تفاوت‌هایی دارد، بلکه در ارتباط با هوش

1. AI Ethics

مصنوعی این تفاوت به‌مراتب برجسته‌تر است. در اغلب مواردی که (ethics) به صورت مضاف به کار می‌رود، مقصود معنای فراخ‌تر این تعبیر است. هرچند گاه (ethics) در پیوند با (morals) یا «خوب و بد اخلاقی» تعریف می‌شود، اما در معنای گسترده‌تر، (ethics) بر مجموعه‌ای از ضوابط و دستورالعمل‌های ناظر بر رفتار دلالت دارد؛ ضوابطی که رعایت آنها باعث می‌شود در یک قلمرو، بستر یا ساختار مشخص، اهداف و نتایج مورد نظر تحقق یابد (Beauchamp & Childress, 2019: 21). این برداشت از (ethics) به مفهوم «مقررات» نزدیک‌تر است تا اینکه به معنای (morals) باشد؛ برای نمونه، در عنوانی که تاکنون رسمی‌ترین و برجسته‌ترین مصداق عینی «اخلاق هوش مصنوعی» به شمار می‌آید، چنین فهمی از (ethics) آشکار است: اتحادیه اروپا پس از بحث‌های طولانی، اصلاحات متعدد و ارائهٔ نسخه‌های اولیه، سرانجام در سال (۲۰۲۴) نسخهٔ نهایی و شسته‌رفته‌ای از چیزی را ارائه کرد که «مجموعهٔ مقررات هوش مصنوعی اتحادیه اروپا[1]» نام گرفت[2]؛ بنابراین، دست‌کم در حوزهٔ هوش مصنوعی، تعبیر (ethics) بیشتر به «مقررات، قوانین و ضوابط» نزدیک است تا به معنایی که ما در فرهنگ خود غالباً از «اخلاق» مراد می‌کنیم، یعنی اخلاق در معنای خوب و بد اخلاقی یا همان (morals) است.

۱-۲. دشواری کسب فهمی شهودی از هوش مصنوعی

دشواری بحث تنها به معنای «اخلاق» محدود نمی‌شود؛ بلکه فهم چیستی «هوش مصنوعی» نیز بسیار دشوار است. برخلاف تصوری که در نگاه نخست ایجاد می‌شود، به‌سادگی نمی‌توان فهمی شهودی و روشن از ماهیت هوش مصنوعی به دست آورد. بزرگ‌ترین خطا آن است که هوش مصنوعی را چیزی شبیه هوش انسان تصور کنیم یا صرفاً با اندکی کار کردن با چت‌بات‌های رایج گمان بریم که «هوش مصنوعی» همان است؛ همچنین، خطاست اگر آن را صرفاً نوعی ماشین محاسبه‌گر سریع و پیچیده، مشابه رایانه، بدانیم. فهم چیستی هوش مصنوعی مستلزم آگاهی از دانش فنی و تکنیکی است؛ دانشی که بسیاری از متخصصان علوم انسانی طبیعتاً در آن تخصصی ندارند. برای درک این دشواری کافی است به تعاریف مختلف هوش مصنوعی توجه کنیم. متیو لیاو[3]، یکی از چهره‌های برجستهٔ اخلاق هوش مصنوعی، در مقدمه مجموعه‌ای از مقالات گردآوری‌شده از صاحب‌نظران این حوزه، به نقل از متخصصان هوش مصنوعی جنبه‌های متعددی را برمی‌شمرد که باید در تعریف هوش مصنوعی مدنظر قرار گیرند. جان مک‌کارتی[4] که اصطلاح «هوش مصنوعی» را وضع کرد، آن را چنین تعریف کرده است: «علم و مهندسی ساختن ماشین‌های هوشمند (Liao, 2020: 3). این تعریف چندان گویا نیست؛ از همین‌رو، لیاو تعریف

1. EU AI Act
2. See; https://artificialintelligenceact.eu/
3. Matthew Liao
4. John McCarthy

دیگری پیشنهاد می‌کند: «ساختن ماشین‌هایی که به نحوی عمل می‌کنند که اگر بنا بود انسان آن اعمال را انجام دهد، نیازمند توانایی‌های شناختی مانند تفکر، یادگیری و حل مسئله بود» (همان). هرچند این تعریف به‌ظاهر برای ما قابل فهم است، اما اگر پرسش کنیم این فرایند دقیقاً با چه سازوکاری در ماشین رخ می‌دهد، به‌سختی می‌توانیم تصویری شهودی در ذهن خود شکل دهیم. سازوکار هوش مصنوعی به‌طور کامل جنبه‌ای فناورانه و فنی دارد؛ به همین دلیل مک‌کارتی آن را «علم و مهندسی» می‌داند.

۱-۳. دشواری فهم هوش مصنوعی بدون نظر متخصصان

بنابراین همهٔ اجزای ترکیب «اخلاق هوش مصنوعی» از پیچیدگی برخوردارند: از یک‌سو، «اخلاق» در معنایی خاص و متفاوت به کار رفته است و از سوی دیگر، خود «هوش مصنوعی» فناوری‌ای بسیار پیچیده و دور از فهم شهودی ما به شمار می‌آید؛ همین امر سبب می‌شود که متخصصان علوم انسانی، دست‌کم در آغاز کار با مانعی جدی مواجه شوند؛ در نتیجه، به لحاظ روش‌شناختی این الزام وجود دارد که مطالعات در این حوزه ماهیتی میان‌رشته‌ای داشته باشند و با همکاری نزدیک میان متخصصان فنی، مهندسان و صاحب‌نظران علوم انسانی پیش بروند.

۱-٤. جستجوی اخلاق هوش مصنوعی در سنت علوم انسانی خود

طبق رویه‌ای که در دهه‌های اخیر در فضای علوم انسانی ایران رواج یافته است، بسیاری می‌کوشند در مواجهه با هر پدیدهٔ نو، نسخه‌ای بومی از آن را در سنت علوم انسانی خود بیابند. روشن است که سازگار ساختن هر علم با فرهنگ خود امری موجه است؛ اما مشکل زمانی رخ می‌دهد که تلاش شود پدیده‌ای کاملاً جدید در سنتی جستجو شود که در زمانی بسیار پیش‌تر از دوران معاصر شکل گرفته است؛ به تعبیر دیگر، سخن گفتن از «مسائل محلی» در زمینهٔ هوش مصنوعی کاملاً درست و بایسته است، اما ریشه‌یابی آن در سنت‌های گذشته رویکردی نادرست و گمراه‌کننده خواهد بود. فارغ از ارزش یا بی‌ارزشی این رویکرد پرنفوذ، دست‌کم دربارهٔ اخلاق هوش مصنوعی می‌توان با اطمینان گفت که هیچ نمونه یا همتای مشابهی در سنت‌های فکری پیشین ما وجود ندارد و هرگونه تلاش برای یافتن آن چیزی جز سردرگمی و خطا به بار نخواهد آورد. جالب آنکه برخی حتی دربارهٔ خود «هوش مصنوعی» نیز می‌کوشند ریشه‌هایی بومی بیابند یا دست‌کم چارچوب‌ها و پیش‌فرض‌های خود را بر آن تحمیل کنند[1]؛ البته این امر منحصر به

---

۱. شاید یکی از جالب‌ترین مصادیق این رویکرد را بتوان پیش نشستی عنوان کرد که در مهر ماه ۱۴۰۲ با عنوان مبانی فلسفی در توسعه هوش مصنوعی توسط مجمع عالی حکمت اسلامی با همکاری ستاد راهبردی فناوری‌های هوشمند برگزار شد:
https://aicisc.com/%d9%be%db%8c%d8%b4-%d9%86%d8%b4%d8%b3%d8%aa-%d8%b9%d9%84%d9%85%db%8c-%d9%87%d9%85%d8%a7%db%8c%d8%b4-%d8%a8%db%8c%d9%86-%d8%a7%d9%84%d9%85%d9%84%d9%84%db%8c-%d9%87%d9%88%d8%b4-%d9%85%d8%b5%d9%86%d9%88/

متفکران ایرانی نیست.[۱] برای روشن‌تر شدن موضوع می‌توان به مقاله‌ای اشاره کرد که اخیراً با عنوان «می‌توان از مسائل محلی در زمینه هوش مصنوعی سخن گفت و این کاملاً درست است، اما ریشه‌یابی آن در سنت خود شیوه‌ای است نادرست و خطا» منتشر شده است (میری بالاجورشری و محمودی، ۱۴۰۳). در این مقاله عباراتی به چشم می‌خورد از این دست: «اصول اخلاقی متعددی با رویکرد اسلامی وجود دارند که می‌توانند در طراحی هوش مصنوعی به کار گرفته شوند؛ برای مثال، اصل عدل و انصاف در اسلام یکی از اصول اخلاقی اساسی است. در طراحی هوش مصنوعی باید تلاش کرد از هرگونه تبعیض ناعادلانه در پردازش اطلاعات و تصمیم‌گیری‌های سیستم هوشمند خودداری شود. عدالت و انصاف از اصول بسیار مهم اسلامی است؛ بر اساس قوانین شرعی از مسلمانان خواسته شده است سیستم‌های هوش مصنوعی باز، مسئولانه و بی‌طرفانه ایجاد و به کار گرفته شوند». با توجه به آنچه در بالا گفتیم، در این مقاله «ethics» به‌کلی با «morals» خلط شده است؛ بنابراین، بحث «بومی‌سازی» در اخلاق هوش مصنوعی، در معنایی که گفته شد، تا حد زیادی منتفی است.

۱-۵. اخلاق هوش مصنوعی و بازآفرینی عاملی اخلاقی همچون انسان

یکی دیگر از سوفهم‌های رایج این است که بر اساس فهمی شهودی و غیرفنی، عموماً هوش مصنوعی را به مثابه فاعلی شبیه انسان، اما دارای مغزی ماشینی تلقی می‌کنیم؛ در نتیجه، درست همان‌گونه که فاعل انسانی دارای نظام اخلاقی است، چنین پنداشته می‌شود که برای ماشین نیز باید نظام اخلاقی مشابهی تعریف شود. این یکی از مهم‌ترین و در عین حال خطیرترین سوفهم‌های رایج است. علت این سوفهم غفلت از دو نکته مهم است: نخست آنکه در معنای رایج هوش مصنوعی رایج[۲]، همان هوش مصنوعی محدود[۳] است، نه هوش مصنوعی قوی[۴] یا کلی. قسم نخست محصول فناوری موجود است؛ یعنی هوش مصنوعی‌ای که خودمختاری[۵] کامل ندارد و در نهایت وابسته به انسان است و برای انجام وظایف مشخص و محدود طراحی و ساخته می‌شود (Liao, 2018: 2). اگر روزی قسم دوم هوش مصنوعی پدید آید، شاید نتوان این برداشت را به‌درستی «سوفهم» نامید.

نکتهٔ دوم آن است که اخلاق هوش مصنوعی در نسبت با انسان تعریف می‌شود؛ یعنی مجموعه‌ای از قواعد و دستورالعمل‌ها که اعمال آنها بر هوش مصنوعی در نهایت در خدمت خیر و سعادت انسانی و نه خیر و سعادت خودِ هوش مصنوعی است. هرچند گاه به‌طور پراکنده از «حقوق هوش مصنوعی» نیز

---

۱. بنگرید به: میرزا رضوان علی بیگ، ۱۴۰۲، در: https://iqna.ir/fa/news/4167996

2. Narrow
3. Strong
4. Artificial General Intelligence (AGI)
5. Autonomy

سخن گفته می‌شود، اما این مسئله فرعی است و اصل ماجرا همچنان سعادت انسان است؛ البته بر اساس محدودیت‌های تکنیکی فعلی این سخن صادق است؛ اما اگر روزی هوش مصنوعی قوی خلق شود، ممکن است این جنبه نیز دچار تغییر شود. رسانه‌ها، ادبیات و سینما ـ به‌ویژه رمان‌ها و فیلم‌های علمی-تخیلی ـ در ایجاد و تقویت این سوفهم نقش تعیین‌کننده‌ای دارند. این آثار غالباً باعث می‌شوند که از هوش مصنوعی برداشتی انسان‌واره شکل گیرد.

۱-٦. خطر یا سعادت؟

به‌طور سنتی، دست‌کم در فلسفهٔ اخلاق، مقولهٔ اخلاق به معنای (morals) عمدتاً معطوف به سعادت آدمی است؛ یعنی در پی ارائه یا یافتن مسیر نیل به سعادت انسانی است؛ به ندرت فیلسوفی در این اصل چون‌وچرا کرده است که اخلاق چیزی جز راهی برای سعادت نیست. هرچند این امر در اخلاق فضیلت برجستگی بیشتری دارد، اما در سایر رویکردهای فلسفی به اخلاق نیز حضور دارد. در مقابل، اخلاق هوش مصنوعی عمدتاً به‌دنبال کاستن از خطرات ناشی از هوش مصنوعی است. ادبیات این حوزه سرشار از هشدارها و نگرانی‌ها دربارهٔ پیامدهای منفی هوش مصنوعی است؛ پیامدهایی که بیش از هر چیز متوجه حیات، سعادت و رفاه انسان‌اند؛ البته گاه در این زمینه افراط نیز رخ می‌دهد و خطرات بزرگ‌نمایی می‌شوند. ادبیات و سینمای علمی-تخیلی نیز در این میان نقشی برجسته در پرورش چنین تصوراتی دارد. با این حال، می‌توان ادعا کرد که در ادبیات اخلاق هوش مصنوعی، نگرانی از خطرات بر خوش‌بینی به فرصت‌ها چیرگی دارد؛ به‌طور کلی، نمی‌توان انکار کرد که در میان فیلسوفان و در ادبیات فلسفهٔ تکنولوژی، سوءظن و بدبینی نسبت به فناوری بیش از حسن‌ظن و خوش‌بینی است. شاید برجسته‌ترین نمونه در این زمینه، مارتین هایدگر باشد؛ با این حال، در میان فیلسوفان تکنولوژی، اخلاق فناوری و به‌طور خاص فیلسوفان اخلاق هوش مصنوعی با خاستگاه تحلیلی (و نه قاره‌ای)، این بدبینی و نگرانی به‌مراتب کمتر به چشم می‌خورد.

۲. چیستی اخلاق هوش مصنوعی

اکنون لازم است فارغ از رویکرد سلبی، از جنبه‌ای ایجابی نیز به ماجرا بنگریم و به این پرسش پاسخ دهیم که اخلاق هوش مصنوعی چیست و کدام مسائل و مباحث را بیش از همه دربرمی‌گیرد. در بخش‌های پیشین به برخی تعاریف اخلاق هوش مصنوعی اشاره کردیم، اما همان‌طور که تصریح شد، این تعاریف چندان روشنگر و راه‌گشا نیستند. در غیاب دانش تکنیکی لازم، ساده‌ترین راه برای دستیابی به فهمی ایجابی از اخلاق هوش مصنوعی، پرداختن به مسائل اصلی آن است. بررسی این مسائل می‌تواند گامی مهم در جهت روشن‌تر شدن چیستی اخلاق هوش مصنوعی باشد. در این مسیر، نخست باید توجه داشت که مباحث اخلاق هوش مصنوعی را می‌توان در یک تقسیم‌بندی بنیادی به دو دسته

کلی تقسیم کرد:

الف). مسائل نظری و فلسفی: این دسته از مسائل، بیشتر جنبه نظری دارند و به مباحث فلسفی ناب مربوط می‌شوند. آنها به طورمستقیم در صنعت و فناوری کاربرد ندارند، بلکه بیشتر در سطح تبیین مفهومی و پرسش‌های بنیادین مطرح می‌شوند.

ب). مسائل عملی و کاربردی: این دسته در هنگام استفاده از هوش مصنوعی در صنعت و فناوری بروز می‌کنند؛ در اینجا، پرسش‌ها و چالش‌های اخلاقی عمدتاً در فرآیند طراحی، تولید و به‌کارگیری سامانه‌های مبتنی بر هوش مصنوعی مطرح می‌شوند. جنبه فلسفی این مسائل کمتر از دستهٔ نخست است، اما از نظر اهمیت و پیامدهای عملی، بسیار برجسته‌اند.

ابتدا به دستهٔ نخست مسائل می‌پردازیم. در نگاه نخست چنین به نظر می‌رسد که این دسته از مسائل، در قیاس با دستهٔ دوم، برای متخصصان علوم انسانی و به‌ویژه فیلسوفان اخلاق قابل‌فهم‌تر باشند و به دانش تکنیکی چندانی نیاز نداشته باشند؛ اما در واقعیت چنین نیست. این مسائل به‌مراتب دشوارتر از آن‌اند که در نگاه اول به نظر می‌رسند و بدون برخورداری از دانش تکنیکی عمیق در حوزهٔ هوش مصنوعی، فهم ما از آنها هرگز از سطحی ابتدایی فراتر نخواهد رفت. در ادامه به برخی از این مسائل نظری اشاره می‌کنیم. نخستین (نه لزوماً به معنای مسئله دارنده بالاترین اهمیت، بلکه صرفاً نخستین به اعتبار شمارش) این پرسش است که آیا هوش مصنوعی می‌تواند همچون انسان یک عامل اخلاقی[1] تلقی شود؟ (Cave & Dihal, 2020) پاسخ غالب متخصصان به این پرسش منفی است، به‌ویژه با توجه به این نکته که هنوز هوش مصنوعی در معنای قوی یا فراگیر، یعنی هوش مصنوعی مستقل از انسان و دارای خودمختاری، به وجود نیامده است؛ با این‌حال، برخی پژوهشگران even—در اقلیت—پاسخی مثبت داده‌اند. اگر پاسخ مثبت داده شود، بلافاصله پرسشی دیگر مطرح می‌شود: جایگاه اخلاقی این موجود فناورانهٔ شبه‌انسانی چه خواهد بود؟ شهود انسانی نشان می‌دهد که اگر موجودی غیرانسانی اما برخوردار از خودآگاهی، به‌عنوان فرآوردهٔ فناوری پدید آید، ناگزیر باید از شأن اخلاقی نیز برخوردار باشد؛ با این‌حال، در اینجا نیز اختلاف‌نظرهایی وجود دارد. برخی بر این باورند که خودآگاهی به‌تنهایی برای ایجاد شأن اخلاقی کافی نیست و وجود احساس نیز شرطی اساسی است. از این منظر، میان «خودآگاهی» به‌مثابهٔ شرط لازم و «احساس‌مندی» به‌مثابهٔ شرط لازم و کافیِ انسان‌وارگی تمایز گذاشته می‌شود؛ در مقابل، گروهی دیگر معتقدند تا زمانی که چنین ماشین هوشمندی (یعنی ماشینی دارای خودآگاهی و احساس‌مندی) ساخته نشده است، بحث و نظرورزی دربارهٔ شأن اخلاقی آن بیهوده و اتلاف وقت است. پرسش مهم دیگر این است که آیا می‌توان برای هوش مصنوعی و ربات‌های مبتنی بر آن، حقوق اخلاقی

1. Moral agent

در نظر گرفت، یا آنها را باید صرفاً ابزارهایی بی‌حقوق تلقی کرد. یکی از منابع مهم در این زمینه کتاب ارزشمند حقوق ربات‌ها[1] نوشتهٔ گانکل است (Gunkel, 2018). به نظر می‌رسد که دست‌کم در وضعیت فعلیِ رشد فناوری، این مباحث هنوز از مسائل مبتلابه فاصله دارند و اولویت چندانی ندارند؛ این پرسش، البته، به‌طور تنگاتنگ به پرسش نخست وابسته است و تا حد زیادی در ذیل آن قابل فهم است.

سومین پرسش مهم این است که آیا می‌توان بخشی از تصمیمات اخلاقی مربوط به انسان‌ها را به هوش مصنوعی واگذار کرد (Moor, 2006)؛ برخلاف دو پرسش نخست، این پرسش در وضعیت کنونیِ رشد فناوری، واقعی و قابل‌طرح است. امروزه برخی تصمیم‌گیری‌های مهم، همچون بارورسازی ابرها، سیاست‌های جمعیتی، مبارزه با جرایم و شناسایی مجرمان، پیش‌بینی افراد و گروه‌هایی که بیشتر در معرض ارتکاب جرم‌اند (Angwin et-al, 2016)، فعالیت‌های تشخیصی و درمانی در حوزهٔ سلامت (Rajkomar & Kohane: 2019) و حتی مسائل مربوط به مهاجرت (Molnar, 2019: 329)، به‌واسطهٔ هوش مصنوعی پیش می‌رود و در عمل این فناوری نقش تصمیم‌گیر ایفا می‌کند. دربارهٔ همهٔ این موارد، انتقادات بسیاری از منظر اخلاقی وجود دارد که همگی در حیطهٔ اخلاق هوش مصنوعی جای می‌گیرند؛ با این‌حال، باید توجه داشت که هرچند مصادیق یادشده ذیل مسائل کاربردی قرار می‌گیرند، اما اصل پرسش، یعنی اینکه آیا می‌توان تصمیم‌گیری‌های مهم انسانی را به هوش مصنوعی سپرد یا نه، پرسشی نظری است.

چهارمین پرسش مهم نظری به مسئلهٔ جانبداری، سوگیری و تعصب در هوش مصنوعی مربوط می‌شود؛ بدین معنا که آیا شکل‌گیری چنین تمایلاتی در هوش مصنوعی نیز، همانند انسان، از منظر اخلاقی نادرست است یا خیر؛ البته این موضوع نه‌تنها در سطح نظری، بلکه در عرصهٔ صنعت و فناوری نیز از مسائل مهم تلقی می‌شود و می‌توان آن را مسئله‌ای مشترک میان هر دو حوزه دانست. در قلمرو دستهٔ دوم، یعنی مسائل کاربردی، افزون بر مصادیقی که ذیل پرسش سوم مطرح شد، می‌توان موارد دیگری را نیز برشمرد؛ نخست، با توجه به اینکه هوش مصنوعی حجمی عظیم از داده‌ها و اطلاعات خصوصی افراد را گردآوری می‌کند، مسئلهٔ حفظ حریم خصوصی افراد و شیوهٔ مواجهه با چالش‌های مرتبط با آن اهمیت بسیار دارد (Floridi & Taddeo, 2016)؛ دومین مسئلهٔ مهم، کاربرد هوش مصنوعی در عرصه‌های نظامی و تسلیحاتی است که می‌تواند پیامدهایی به‌شدت خطرناک داشته باشد (Crootof, 2015). مداخلهٔ هوش مصنوعی در بازار و به‌ویژه تهدید جایگاه مشاغل انسانی نیز موضوع مهم دیگری است که توجه اندیشمندان را به خود جلب کرده است. برخی مسائل نیز بسیار جزئی و موردی‌اند؛ برای نمونه، دربارهٔ خودروهای خودران این پرسش مطرح است که در شرایط بحرانی، یعنی زمانی که به‌ناچار باید میان برخورد با یک کودک یا یک سالمند یکی را انتخاب کنند (Awad, 2018: 59-64)، چه تصمیمی باید اتخاذ

1. Robot rights

شود؛ به‌طور کلی، دامنه مسائل کاربردی بسیار گسترده‌تر و متنوع‌تر از مسائل نظری محض است و مواردی که در اینجا ذکر شد صرفاً به‌منزلهٔ نمونه بیان شد.

اکنون با مرور عمده مسائل ـ اعم از نظری و کاربردی ـ اخلاق هوش مصنوعی می‌توان به تصویری اولیه و نسبتاً دقیق از این حوزه دست یافت؛ با این حال، باید توجه داشت که صرف دستیابی به چنین تصویری برای کار در این زمینه کافی نیست؛ زیرا برای آنکه بتوان دربارهٔ هر مسئله به‌طور جدی تأمل و اظهار نظر کرد، نیازمند آگاهی و دانش کافی نسبت به ماهیت فنی و سازوکارهای هوش مصنوعی هستیم؛ از همین روست که بسیاری از متخصصان برجستهٔ اخلاق هوش مصنوعی، دارای پیشینه‌ای در حوزه‌های مهندسی و فناوری‌اند، هرچند حضور متخصصانی با زمینه‌های علوم انسانی نیز کم نیست. مانع مهم دیگری نیز در این میان وجود دارد که شایان توجه است: گاه ممکن است یک متخصص علوم انسانی یا فیلسوف اخلاق راه‌حلی برای مسئله‌ای پیشنهاد کند که از منظر فنی و عملی قابلیت اجرا نداشته باشد؛ بنابراین، برخورداری از دانش فنی و مهندسی در زمینهٔ هوش مصنوعی از این جنبه نیز ضرورت می‌یابد؛ البته این مانع را می‌توان تا حدودی با رویکرد کار گروهی، یعنی تشکیل تیم‌های میان‌رشته‌ای متشکل از فیلسوفان اخلاق و متخصصان فنی و مهندسی، برطرف کرد.

نتیجه

اخلاق هوش مصنوعی رویکردی است ناظر به برساختن دستورالعمل‌های کاربردی در حوزهٔ هوش مصنوعی با هدف جلوگیری از بروز پیامدهایی که عمدتاً بر اساس شهود اخلاقی، نادرست قلمداد می‌شوند؛ با این حال، به دلیل ماهیت پیچیدهٔ هوش مصنوعی ـ به‌ویژه پیوند آن با ماشین و ابتنایش بر یادگیری ماشینی ـ این دستورالعمل‌ها با چالش‌ها و ظرایف فنی بسیاری همراه‌اند؛ از این‌رو میان فهم متعارف از اخلاق و حتی برداشت‌های فلسفی از آن، با فهم تکنیک‌محور از اخلاق هوش مصنوعی تفاوتی اساسی وجود دارد؛ بنابراین، تقرّب و ورود متخصصان علوم انسانی به عرصهٔ اخلاق هوش مصنوعی مستلزم رعایت احتیاط، ملاحظات روشی و برخورداری از مطالعات کافی در قلمرو فنی هوش مصنوعی است. در بسیاری موارد نیز همکاری میان‌رشته‌ای و بهره‌گیری از دانش متخصصان این حوزه ضرورتی اجتناب‌ناپذیر به شمار می‌آید.

ملاحظات اخلاقی

این تحقیق به صورت یک پژوهش مستقل صورت گرفته، و برای انجام آن، از حمایت مالی موسسه خاصی استفاده نشده است. هم‌چنین، نویسندگان تعارض منافعی نداشته‌اند.

منابع

رمضانی، م.، و فیضی درخشی، م. ر. (۱۳۹۲). اخلاق ماشین: چالش‌ها و رویکردهای مسائل اخلاقی در هوش مصنوعی و ابرهوش. فصلنامه اخلاق در علوم و فنّاوری، ۸(۴)، ۱-۹.

زرگر، ز. (۱۴۰۴). رابطه عواطف و ظرفیت اخلاقی در فناوری‌های هوش مصنوعی. پژوهش‌های فلسفی، ۵۰، ۱۹-۴۰.

میری بالاجورشری، س. م.، و محمودی، ا. ر. (۱۴۰۳). واکاوی مسائل اخلاقی در زمینه هوش مصنوعی با نگاهی به اخلاق اسلامی. فصلنامه علمی پژوهشی مطالعات اخلاق کاربردی، ۱۴(۶)، ۹۷-۱۲۳.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

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

Beauchamp, T. L., & Childress, J. F. (2019). Principles of biomedical ethics (8th ed.). Oxford University Press.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (pp. 149–159). https://doi.org/10.1145/3287560.3287583

Cave, S., & Dihal, K. (2020). The Whiteness of AI. Philosophy & Technology, 33(4), 685–703. https://doi.org/10.1007/s13347-020-00415-6

Crootof, R. (2015). The killer robots are here: Legal and policy implications. Cardozo Law Review, 36, 1837–1915.

Floridi, L., & Taddeo, M. (2016). What is data ethics? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2083), 20160360. https://doi.org/10.1098/rsta.2016.0360

Gunkel, D. J. (2018). Robot rights. MIT Press.

Hurley, M., & Adebayo, J. (2017). Credit scoring in the era of big data. Yale Journal of Law and Technology, 18(1), 148–216. https://digitalcommons.law.yale.edu/yjolt/vol18/iss1/5/

Liao, M. (2020). A short introduction to the ethics of artificial intelligence. In M. Liao (Ed.), Ethics of Artificial Intelligence. Oxford University Press.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. IEEE Intelligent Systems, 21(4), 18–21. https://doi.org/10.1109/MIS.2006.76

Molnar, P. (2019). Technological testing grounds: Migration management experiments and reflections from the ground up. European Journal of Migration and Law, 21(3), 329–352. https://doi.org/10.1163/15718166-12340054

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

van den Hoven, J. (2010). The handbook of information and computer ethics. Wiley.