Journal of Philosophical Theological Research
ISSN Print: 1735-9791 Online: 2538-2500

The Meaning and Criteria of Moral Agency in Intelligent Machines

Seyyed Mohammad Hoseini Souraki

Assistant Professor, Department of Moral Philosophy, University of Qom, Qom, Iran. m.hoseini@qom.ac.ir

Abstract

The advancements in artificial intelligence (AI) and the emergence of intelligent and superintelligent machines have significantly blurred the traditional boundaries between humans and machines. These technological developments have raised critical philosophical and ethical questions about the possibility of attributing moral agency to these systems. Can intelligent machines act morally, or should moral responsibility for their actions remain solely with their designers and operators? This paper aims to analyze these questions through a conceptual and analytical lens, focusing on the meaning, criteria, and philosophical implications of moral agency in the context of intelligent machines.

Research Article

Keywords: moral agency, AI ethics, intelligent machines, autonomy, moral responsibility.

Received: 2024/10/03 ; **Received in revised form:** 2024/11/07 ; **Accepted:** 2024/11/25 ; **Published online:** 2024/12/22

Hoseini Souraki, S.M. (2024). The Meaning and Criteria of Moral Agency in Intelligent Machines. Journal of Philosophical Theological Research. *Journal of Philosophical Theological Research (Philosophy of Ethics and Technology: challenges and prospects special Issue)*, 26(4), 83-108. https://doi.org/10.22091/jptr.2025.12466.3256

Defining Moral Agency

Moral agency refers to the capacity of an entity to make ethical decisions, act autonomously, and accept responsibility for its actions. Philosophical discussions on moral agency, rooted in the works of thinkers such as Immanuel Kant and John Searle, identify several essential components:

Consciousness and Self-Awareness: The ability to recognize oneself as an independent entity and distinguish between oneself and others.

Moral Understanding: The ability to evaluate actions based on moral principles, and distinguishing between right and wrong.

Autonomy: The capacity to make independent decisions free from external constraints.

Intentionality and Free Will: The ability to act based on deliberate choices and goals.

Moral Responsibility: The capacity to accept accountability for one's actions and their consequences.

These criteria form the foundation for evaluating whether an entity, such as an intelligent machine, can qualify as a moral agent.

Perspectives on the Moral Agency of Intelligent Machines

Denial of Moral Agency

One of the dominant perspectives in contemporary philosophy rejects the possibility of moral agency for intelligent machines. Thinkers such as John Searle and Daniel Dennett argue that machines, regardless of their computational complexity, lack the essential characteristics of moral agents. Searle's "Chinese Room" thought experiment illustrates this point by demonstrating that machines merely manipulate symbols without understanding their meaning, thus lacking true consciousness or intentionality.

From this perspective, moral agency is intrinsically tied to uniquely human attributes, such as subjective awareness, emotional sensitivity, and the capacity for moral intuition. Machines, as deterministic systems governed by algorithms, cannot possess the free will or intentionality necessary for moral responsibility. As a result, any moral actions performed by machines are ultimately attributable to their human creators.

This view also aligns with Kantian ethics, which emphasizes rationality, autonomy, and free will as prerequisites for moral agency. Kant argued that only beings capable of acting according to moral laws derived from rational deliberation could be considered moral agents.

Acceptance of Limited or Functional Moral Agency

Contrary to the denialist perspective, some scholars propose a more nuanced view that recognizes the potential for limited or functional moral agency in intelligent machines. Luciano Floridi and J. W. Sanders introduce the concept of "mind-less morality," suggesting that moral agency does not necessarily require subjective awareness or intentional states. Instead, they argue that intelligent systems can be considered moral agents within specific contexts if they exhibit the following characteristics: interactivity, autonomy, and adaptability.

This perspective posits that machines can perform morally significant actions without possessing the full range of human-like cognitive and emotional capacities. For example, autonomous vehicles can make decisions that have moral implications, such as prioritizing the safety of passengers over pedestrians in emergency situations. While these decisions are based

on pre-programmed algorithms, they can be seen as a form of functional moral agency.

Key Arguments and Counterarguments

The Role of Consciousness and Intentionality

A central debate in the discussion of moral agency for machines revolves around the role of consciousness and intentionality. Critics of machine moral agency, such as Searle, argue that without genuine consciousness and intentionality, machines cannot be considered moral agents. They contend that machines lack the ability to understand the moral significance of their actions, as they merely process information without any real comprehension.

Proponents of limited moral agency, however, argue that consciousness and intentionality are not strictly necessary for moral decision-making. They suggest that machines can be designed to follow ethical guidelines and make decisions that align with moral principles, even if they do not possess subjective experiences. This view is supported by the idea that moral agency can be understood in terms of functional roles rather than internal states.

Implications for Ethics and Society

Ethical Design and Regulation

The debate over machine moral agency has significant implications for the ethical design and regulation of AI systems. If machines are to be considered moral agents, even in a limited sense, it becomes crucial to ensure that they are designed with ethical principles in mind. This includes developing algorithms that prioritize human well-being, fairness, and transparency.

Regulatory frameworks will also need to address the challenges posed by autonomous systems, including issues of accountability, liability, and oversight. Policymakers must strike a balance between promoting innovation and ensuring that AI systems are used responsibly and ethically.

Conclusion

The question of whether intelligent machines can be considered moral agents is complex and multifaceted. While traditional philosophical perspectives emphasize the importance of consciousness, intentionality, and free will for moral agency, more contemporary views suggest that machines can exhibit a form of functional moral agency within specific contexts.

The debate has important implications for the ethical design and regulation of AI systems, as well as for broader societal attitudes towards technology. As AI continues to evolve, it will be crucial to address these ethical and philosophical questions to ensure that the development and deployment of intelligent machines align with human values and principles.

In conclusion, while machines may not possess the full range of capacities required for moral agency in the traditional sense, they can still play a significant role in ethical decision-making. The challenge lies in defining the boundaries of machine moral agency and ensuring that these systems are designed and used in ways that promote human well-being and ethical integrity.

References

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

Dennett, D. C. (2014). When HAL kills, who's to blame? Computer ethics. In F. Battaglia, N. Mukerji & J. Nida-Rümelin (Eds.), *Rethinking responsibility in science and technology*. Pisa University Press. https://doi.org/10.1400/225034.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

Graff, J. (2024). Moral sensitivity and the limits of artificial moral agents. *Ethics and Information Technology*, 26(1), 13. https://doi.org/10.1007/s10676-024-09755-9.

Jaworska, A., & Tannenbaum, J. (2023a). The grounds of moral status. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2023/entries/grounds-moral-status/

Manna, R., & Nath, R. (2021). The problem of moral agency in artificial intelligence. *2021 IEEE Conference on Norbert Wiener in the 21st Century (21CW)*, 1–4. https://doi.org/10.1109/21CW48944.2021.9532549.

Misselhorn, C. (2022b). Artificial moral agents: Conceptual issues and ethical controversy. In O. Mueller, P. Kellmeyer, S. Voeneky & W. Burgard (Eds.), *The Cambridge handbook of responsible artificial intelligence: Interdisciplinary perspectives* (pp. 31–49). Cambridge University Press. https://doi.org/10.1017/9781009207898.005.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4), 18–21. IEEE Intelligent Systems. https://doi.org/10.1109/MIS.2006.80.

Müller, V. C. (2023). Ethics of artificial intelligence and robotics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/

Schlosser, M. (2019). Agency. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/agency/

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–457. https://doi.org/10.1017/s0140525x00005756.

Sullins, J. P. (2011). When is a robot a moral agent? In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 151–161). Cambridge University Press. https://doi.org/10.1017/CBO9780511978036.013.

پژوهش‌های فلسفی کلامی
ISSN Print: 1735-9791 Online: 2538-2500

معنا و معیار عاملیت اخلاقی ماشین‌های هوشمند

سید محمد حسینی سورکی

استادیار، گروه فلسفه اخلاق، دانشگاه قم، قم، ایران. m.hoseini@qom.ac.ir

چکیده

تحولات شگفتی‌آور و شگرف در حوزه هوش مصنوعی و ظهور ماشین‌های هوشمند و سیستم‌ها و ربات‌های فراهوشمند، مرز و معیارهای انسان و ماشین را به چالش کشیده و به پرسش‌های فلسفی و اخلاقی جدیدی درباره امکان عاملیت اخلاقی ماشین‌های مصنوعی و انتساب مسئولیت اخلاقی به آنها دامن زده است. در این مقاله، پس از تبیین معنا و مؤلفه‌های عاملیت اخلاقی (خودآگاهی، خودمختاری، اراده آزاد، نیت‌مندی و مسئولیت‌پذیری)، به بررسی اهم دیدگاه‌ها درباره امکان عاملیت اخلاقی ماشین‌های پرداخته شده و اهم مبانی و مدعیات دو رویکرد اصلی در این زمینه بیان و بررسی شده است. برخی از صاحب‌نظران، ضمن نفی و انکار عاملیت اخلاقی ماشین‌ها، بر این باورند تنها انسان شروط و شرایط لازم برای عامل اخلاقی بودن را دارد و البته، برخی دیگر، ضمن بازنگری در معنای رسمی و سنتی عاملیت اخلاقی، به گونه و تقریری خاص از عاملیت اخلاقی محدود، تشکیکی و طیفی را پذیرفته و امکان اطلاق عامل اخلاقی به ماشین‌های فراهوشمند، بر این مبنا و معنا را موجه و مقبول دانسته و امکان اطلاق عامل اخلاقی را به ماشین‌ها و حتی حیوانات تعمیم و تسری می‌دهند. با استناد به محدودیت‌های کنونی ماشین‌های فراهوشمند و بی بهره بودن آنها از خودآگاهی و فقدان نیت‌مندی، شرایط لازم برای عاملیت اخلاقی کامل را نداشته و تنها به شکل مَجاز و مسامحی می‌توان امکان عاملیت اخلاقی محدود و ابزارگرایانه را برای رُبات‌ها و ماشینهای فرا/هوشمند در حوزه‌های خاص را پذیرفت.

کلیدواژه‌ها: عاملیت اخلاقی، اخلاق هوش مصنوعی، ماشین‌های هوشمند، خودمختاری، مسئولیت اخلاقی.

تاریخ دریافت: ۱۴۰۳/۰۷/۱۲ ؛ **تاریخ اصلاح:** ۱۴۰۳/۰۸/۱۷ ؛ **تاریخ پذیرش:** ۱۴۰۳/۰۹/۰۵ ؛ **تاریخ انتشار آنلاین:** ۱۴۰۳/۱۰/۰۲

▣ حسینی سورکی، سید محمد (۱۴۰۳). معنا و معیار عاملیت اخلاقی ماشین‌های هوشمند. *پژوهش‌های فلسفی- کلامی (ویژه‌نامه فلسفه اخلاق و فن‌آوری: چالش‌ها و چشم‌اندازها)*، ۲۶(۴)، ۸۳-۱۰۸. https://doi.org/10.22091/jptr.2025.12466.3256

مقدمه

تاریخ تولید و تحوّل آلات و ابزارها از ابتدایی‌ترین ابزارهای دستی تا پیچیده ترین ماشین‌های هوشمند،[1] برایند و بازتاب توانایی‌ها و ذهن خلاق بشر در پاسخ به نیازهای روزمره و غلبه بر موانع و بوده است. با ساخت اولین ابزارها، آدمی گام در مسیری نهاد که به ظهور تکنولوژی و تمدن‌های بزرگ انجامید.

چرخهٔ تطور و تکامل ابزارها و فناوری‌ها همچنان ادامه دارد و در دهه‌های اخیر شتابی شدید و شگرف داشته است. البته، باید توجه داشت که ابزارهای هوشمند امروزی، تنها نسخه‌های بروزشدهٔ ابزارهای دیروز نیستند، بلکه به واقع، کارکردها و ماهیتی فراتر از یک ابزارِ صرف یافته و چنان ارتقا یافته و در سطحی فراتر از ماشین‌های سنتی ایستاده‌اند که در نگاه بسیاری، نه‌تنها خودکار که خودبهبوددهنده و خودآیین[2] نیز به‌حساب می‌آیند. ظهور دستیارهای هوشمند[3] ربات‌های اجتماعی[4] و خودروهای خودران و خودآیین،[5] گواهی روشن بر جریان و جهشی بزرگ و بی‌وقفه در مسیر تکامل فناوری‌های نوظهور و به‌ویژه هوش مصنوعی است. این پیشرفت‌ها نه تنها مرزهای فناوری را جابجا کرده‌اند، بلکه تعامل انسان با ماشین و حتی ساختارهای اجتماعی و اقتصادی را نیز دگرگون ساخته‌اند. ماشین‌های هوشمند امروزی، افزون بر انجام محاسبات پیچیده به شکل خودکار و داشتن قابلیت‌های شناختی (ادراک و یادگیری[6] تصمیم‌گیری مستقل[7]) توان تطبیق‌پذیری و تعامل پویا با محیط[8] و بازنمایی احساسات انسانی، در بروز خلاقیت و خلق آثار هنری بدیع نیز جلوه‌هایی خیره کننده داشته‌اند (Floridi & Chiriatti, 2020; Müller, 2023).

از همین‌رو، در ادبیات علمی، این ماشین‌ها و مصنوعات بشری را با نام‌های دقیق‌تری چون سامانه‌ها و سیستم‌های هوش مصنوعی،[9] ماشین‌های فراهوشمند/ ابرهوشمند[10] ماشین‌های خودآیین[11] یاد کنند.

افزایش قابلیت‌ها و ارتقای توانمندی‌های ماشین‌های هوشمند، به مبهم و محو شدن مرز میان انسان و ماشین نیز انجامیده و رُبات‌ها و ماشین‌های هوشمند نه‌تنها می‌توانند به ما بیاموزند، بلکه می‌توانند به جای ما بیاندیشند و چه بسا برما حکمرانی کنند. از همین‌رو، برخی صاحب‌نظران، اذعان و انتظار دارند که ماشین‌های هوشمند، در نهایت، به شریک و رقیب آدمیان بدل شده و با انسان‌ها هماوردی کنند. در عصر فزونی و فسونگری مصنوعات فراهوشمند، کران‌ناپیدا بودن افق‌های توسعه و تحولات در حوزهٔ

1. Intelligent Machines
2. Automated Machines
3. Voice-based AI Assistants
4. Social Robots
5. Autonomous Vehicles
6. Learning
7. Autonomous Decision-Making
8. Dynamic Interaction
9. AI Systems
10. Superintelligent Machines
11. Autonomous Machines

هوش مصنوعی و همچنین هول و هراس مواجهه با «تکینگی تکنولوژیک»[۱] بیش از پیش، بر دهشت و دلنگرانی‌ها افزوده[۲] است. همچنین قابلیت‌ها و ابعاد در خور تأمل ماشین‌های فراهوشمند، افزون بر آنکه متفاوت بودن و مزیّت‌های آنها را در قیاس با ابزارهای ساده و صرفاً خودکارِ پیشین برجسته کرده، به بروز بحث‌های تازه‌ای دربارهٔ نقش ماشین‌ها در تولید هنجار و ارزش انجامیده و به طرح پرسش‌های فلسفی و اخلاقی از این دست دامن زده است: آیا یک ماشین هوشمند می‌تواند اخلاقی عمل کند؟ آیا ماشین‌های هوشمند شبه/خودآیین را می‌توان به‌عنوان یک عامل اخلاقیِ[۳] مسئولیت‌پذیر در نظر گرفت و مسئولیت تصمیم‌ها و رفتارها را متوجه خود ماشین‌ها کرد؟ (Artificial Moral Agents, Müller, 2023).

صرفِ نظر از چالشِ مسئولیت‌پذیری و عاملیت اخلاقی ماشین‌ها و مسئلهٔ «شکافِ مسئولیت»،[۴] برخی گام را فراتر نهاده و از امکان برخورداری ماشین‌های هوشمند از شأن اخلاقی سخن به میان آورده[۵] و باب گفت‌وگو در باب حقوق ماشین‌ها (Gunkel, 2018) را نیز گشوده‌اند؟ (Russell & Norvig, 2020; Bostrom, 2014).

در این مقاله، با نگاهی جستارگشایانه و از منظری تحلیلی، به معنا و مفهوم عاملیت اخلاقی خواهیم پرداخت و مبنا و معیار عاملیت اخلاقی و امکان عاملیت اخلاقی ماشین‌های فرا/ هوشمند را بررسی و تحلیل خواهیم کرد. بحث از امکان و انحاء عاملیت اخلاقی ماشین‌های فراهوشمند، مسئله و موضوعی بین‌رشته‌ای و محل تقاطع فلسفه ذهن، فلسفه اخلاق، علوم شناختی و هوش مصنوعی و همچنین حقوق و سیاست‌گذاری، روانشناسی است که آثار و ثمرات عملی گسترده‌ای برای طراحی، استفاده، و نظارت بر سیستم‌های هوشمند امروزی و آینده خواهد داشت.[۶]

---

۱. تکینگی تکنولوژیکال (Technological Singularity) به زمانی فرضی در آینده اشاره دارد که در آن پیشرفت فناوری، به‌ویژه هوش مصنوعی، به حدی شتاب و اوج می‌گیرد که از کنترل و درک انسان خارج می‌شود و به مرز غیرقابل کنترل و برگشت‌ناپذیر میرسد و تبعات و پیامدهای غیرقابل پیش‌بینی برای تمدن بشری خواهد داشت.

۲. برخی تمرد ماشینهای فراهوشمند و تمایل به بهبود فناوری، و همسو نبودن آنها با ارزشهای انسانی یک خطر و تهدید بالقوه دانسته‌اند (see Bostrom, 2014, pp. 7, 130).

3. Moral Agent

٤. شکاف مسئولیت (Responsibility Gap) به مشکلی اشاره دارد که در آن مشخص نیست چه کسی مسئول تصمیم‌ها و اقدامات ماشین‌های هوشمند، به‌ویژه در سیستم‌های شبه‌خودمختار، است. این مسئله با افزایش خودمختاری و پیچیدگی هوش مصنوعی، به‌ویژه در آستانه احتمال تکینگی، برجسته‌تر می‌شود.

٥. امروزه در مقام تبیین معیارهای حداقلی و لازم و کافی برای برخورداری از شأن اخلاقی، برخی افزون بر آگاهی و احساس از انگاره ای با عنوان "ontocentrism" از خودمرکزی سخن به میان آورده و گاه حاوی حامل طلاعات بودن را یکی از شروط و شرایط برخورداری از شأن اخلاقی دانسته‌اند (see Ladak, 2024, p. 217).

٦. امروزه بحثهای جدی و جدیدی در خصوص روانشناسی اخلاقی هوش مصنوعی مجال طرح یافته است (see Bonnefon et al., 2024).

ماشین هوشمند[1]/ فرا هوشمند؛ معنا و مولفه‌ها

ابزارهای ساده،[2] ابتدایی‌ترین شکل فناوری بودند که به مثابه امتداد توانایی‌های فیزیکی انسان، کاملاً وابسته به نیروی انسانی و فاقد هوشمندی و توان پردازش اطلاعات بودند. در گام بعدی، ماشین‌هایی ابداع و عرضه شدند که هرچند پیچیده‌تر از ابزارهای ساده بودند، اما عملکرد آنها[3] نیز از پیش برنامه‌ریزی شده و محدود به انجام وظیفه‌ای مشخص بود و البته، پایه‌ای برای توسعه فناوری‌های پیچیده‌تر مانند ماشین‌های هوشمند و فراهوشمند شدند. ماشین‌های هوشمند[4] و فراهوشمند[5] تأثیرات عمیقی بر انسان و جامعه نهاده و افق‌های جدیدی در توانایی‌های فناورانه گشوده‌اند.

ماشین‌های هوشمند، با استفاده از الگوریتم‌های پیشرفته و فناوری‌های پردازش اطلاعات، قادر به انجام وظایف پیچیده هستند و این ماشین‌ها را بر اساس قابلیت‌ها و ظرفیت‌هایی چون توان یادگیری، درک، تصمیم‌گیری[6] و انطباق‌پذیری، می‌توان از ابزارها و ماشین‌های سنتی و ساده متمایز ساخت.

ماشین‌های هوشمند می‌توانند از داده‌های ورودی و تجربه‌های گذشته برای بهبود عملکرد خود استفاده کنند. برای مثال، یک دستیار صوتی هوشمند مانند Siri یا Alexa می‌توانند با تحلیل درخواست‌های کاربر، پیشنهادهای بهتری ارائه دهد و با پردازش حجم بالایی از داده‌ها، الگوهای معناداری از آنها اخذ و استخراج کند.[7]

ماشین‌های هوشمند توانایی اتخاذ تصمیمات مستقل در چارچوب وظایف تعریف‌شده را دارند. این ویژگی در خودروهای خودران[8] و دیگر ماشین‌های هوشمندی که از پردازش زبان طبیعی[9] برای تعامل طبیعی با انسان‌ها و محیط استفاده می‌کنند، به وضوح دیده می‌شود. با این همه، توانایی‌های این ماشین‌ها، به رغم پیچیدگی و پیشرفته بودن، محدود به الگوریتم‌ها و داده‌ها و دستورالعمل‌هایی است که توسط انسان‌ها طراحی و بدانها اِعطا شده و برای عملکرد بهینه همچنان نیازمند نظارت و کنترل انسانی هستند (Bostrom, 2014, pp. 50–52).

بر این اساس، توان بالای پردازش داده‌ها و تحلیل اطلاعات، تعامل پویا با محیط و تصمیم‌گیری و و

---

1. Intelligent Machines

۲. همانند چکش، اهرم، قیچی، و در مرحله پیشرفته تر، ترازوهای مکانیکی، چرخ خیاطی دستی و چراغ قوه.

۳. همانند ساعت‌های دیجیتال، ماشین حساب، جاروبرقی ساده و دستگاه‌های پخش ویدئو.

4. Smart Machines

5. Superintelligent Machines

6. Decision-Making

۷. برای مثال، سیستم‌های پیشنهادگر /Recommendation Systems مانند الگوریتم‌های Netflix یا Spotify از تحلیل داده‌های کاربران برای ارائه پیشنهادهای شخصی‌سازی‌شده استفاده می‌کنند.

8. Autonomous Vehicles

9. Natural Language Processing

البته، نیاز به نظارت انسانی در این ماشین‌ها و سیستم‌ها درخور تأمل است و به رغم آنکه گاه در حوزه‌های خاصی مانند بازی شطرنج یا تشخیص چهره، عملکردی فوق‌العاده و فرابشری دارند، به اعتقاد صاحب‌نظران، هنوز به سطح هوش عمومی انسان[1] نرسیده‌اند (Bostrom, 2014, p. 50) و بعید است که بتوان با توجه به برخی محدودیتها عاملیت اخلاقی را به آنها منتسب کرد.

البته، امروزه بحث بر سر رُبات‌ها و ماشین‌های فراهوشمندی است که بنا به ادعا، از نظر عملکرد شناختی در بیشتر حوزه‌ها به شکل قابل توجهی از بهترین ذهن‌های انسانی پیشی می‌گیرند (Bostrom, 2014, p. 89).

این ماشین‌ها، در مقایسه با ماشین‌های هوشمند، سریع‌تر، کارآمدتر، برخوردار از خلاقیت و با سطح استقلال به مراتب بالاتر در تصمیم‌گیری هستند. توانایی شناختی ماشین‌های فراهوشمند[2] فراتر از همه مصنوعات بشری و قابل قیاس با انسان و دست‌کم در بسیاری از موارد فراتر از انسان است. یادگیری ماشین‌های فراهوشمندِ خودتوسعه‌دهنده[3] به آنها اجازه می‌دهد که به‌صورت مداوم عملکرد خود را ارتقا دهند. یک ماشین فراهوشمند می‌تواند در تعیین اهداف، طراحی استراتژی‌ها، و اجرای آن‌ها، مستقل و خودمختار عمل کند و ایده‌ها و راه‌حل‌هایی ارایه کند که حتی برای انسان‌ها غیرقابل تصور است (Bostrom, 2014, Ch. 3).

پردازش ابرداده‌ها، توان حل مسائل بسیار پیچیده و بویژه درک محیط و مسائل پیچیده به‌صورت چندوجهی به این ماشین‌ها این امکان را می‌دهد که در شرایط ناشناخته نیز تصمیمات هوشمندانه بگیرند و حتی قادراند برای نیل به اهداف خود، محیط خود را تغییر دهند و محدود به تعامل در چارچوب‌های از پیش تعیین‌شده نباشند.

به طور خلاصه، یادگیری چندوجهی و خودتوسعه‌دهنده، پردازش و تحلیل ابرداده‌ها در مقیاس وسیع، خلاقیت فراتر از انسان، درک عمیق از محیط و مسائل پیچیده، استقلال در تعیین اهداف و تصمیم گیری مستقل و خودآیینی، وجوه بارز و متمایز کننده ماشین‌های فراهوشمند است. بی تردید، ماشین‌های هوشمندی از این دست برخوردار از سیستم‌هایی پویا، مستقل، و شبیه به انسان خواهند بود و از همین‌رو، بسیاری بر این باورند که اگر ماشین‌هایی با چنین ابعاد و ویژگیهایی وجود داشته باشند، انتساب مسئولیت اخلاقی به آنها می‌تواند معنادار و محتمل باشد.[4]

صرف نظر از شک و تردیدها در باب سطح هوشمندی و توانمندی‌های هوش مصنوعی و ماشین‌های

1. Human-Level Machine Intelligence
2. Superintelligent Machines
3. Self-Improving

٤. وجه تمایز ابزارهای بدوی و ماشین‌های ساده، ماشین‌های هوشمند و ماشین‌های فراهوشمند، را می‌توان در سه سطح و مؤلفه کلیدی (سطح وابستگی به انسان، قابلیت تحلیل و پردازش داده‌ها و تصمیم‌گیری، و توان تعامل با محیط) خلاصه کرد.

فراهوشمند، بحث و بررسی معناداری متوجه ساختن مسئولیت اخلاقی به ماشین‌ها و امکان عاملیت اخلاقی ماشین‌های فراهوشمند، به تعریف و تلقی ما از مفهوم عاملیت و عاملیت اخلاقی و همچنین معیار و مؤلفه‌های عاملیت اخلاقی وابسته است. در ادامه، به معنا و مؤلفه‌های عاملیت اخلاقی پرداخته می‌شود.

معنا و مؤلفه‌های عامل اخلاقی/ عاملیت اخلاقی[1]

در حوزهٔ فلسفه عمل و علوم شناختی، واژهٔ «عاملیت»[2] به توانایی انجام عمل یا تصمیم‌گیری اشاره دارد، اما معنای آن بسته به زمینه متفاوت است و در این چارچوب، «عامل»[3] به کسی گفته می‌شود که عمل خاصی را با هدف و برنامه‌ای مشخص انجام می‌دهد. به بیان دیگر، عاملیت به توانایی یک عامل یا بازیگر برای عمل کردن در یک محیط خاص اشاره دارد. این مفهوم به‌طور کلی مستقل از ابعاد اخلاقی است که در آن صورت به عنوان عاملیت اخلاقی شناخته می‌شود (Schlosser, 2019).

عامل اخلاقی (کنشگر اخلاقی)[4]، در یک نگاه کلی و البته نه چندان اجماعی، به موجودی اطلاق می‌شود که توانایی اتخاذ تصمیم‌های خطیر اخلاقی، عمل بر اساس آن تصمیم‌ها، و ظرفیت پذیرش مسئولیت و پاسخگویی در قبال اعمال / ترک فعل‌های خود را داشته باشد (Wallach & Allen, 2009). این نکته را نیز نباید از نظر دور داشت که در بسیاری از موارد «عامل اخلاقی» در تقابل و تمایز با «کنش‌پذیر اخلاقی»[5] تعریف و فهم می‌شود. (Jaworska & Tannenbaum, 2023b). «کنش‌پذیر اخلاقی» موجودی است که اعمال اختیاری دیگران بر او اثر می‌گذارد، اما لزوماً توان تصمیم‌گیری یا انجام عمل اخلاقی و پذیرش مسئولیت را ندارد. به بیان دیگر، همهٔ عاملان اخلاقی باید دغدغه‌مند حقوق و رفاه اعضای جامعه اخلاقی که ارزش ذاتی دارند و به تعبیر مصطلح، برخوردار از «شأن اخلاقی[6]»هستند، باشند و در قبال آنها مسئول و پاسخگو هستند. به بیان ساده‌تر، یک کنش‌پذیر اخلاقی موضوع کنش‌های اخلاقی است، اما این توانایی و ظرفیت را ندارد که عامل تصمیم‌گیری اخلاقی باشد (Floridi & Sanders, 2004).

در واقع، همهٔ آنها که ارزش ذاتی دارند و از «شان اخلاقی»برخوردارند یا عامل اخلاقی هستند و یا کنش‌پذیر اخلاقی و هر آنچه عامل اخلاقی است، لزوماً شأن اخلاقی[7] نیز خواهد داشت و اما

1. Moral Agency
2. Agency
3. Agent
4. Moral agent
5. Moral Patient
6. Moral Status

۷. شأن اخلاقی (Moral Status) به معنای شایستگی یک موجود برای دریافت رفتار اخلاقی از سوی دیگران است. این مفهوم معمولاً به موجوداتی تعلق می‌گیرد که دارای ارزش ذاتی یا توانایی تجربه درد، لذت، یا آگاهی باشند (see Jaworska & Tannenbaum, 2023a).

آنکه شأن اخلاقی دارد، لزوماً عامل اخلاقی نیست (Müller, 2023). این که شرکت‌ها و یا حیوانات باهوش و برخوردار از هوش اجتماعی و برای مثال نخستی‌ها[1] و همچنین ربات‌های و ماشین‌های هوشمند، را می‌توان عامل اخلاقی به حساب آورد، همواره محل مناقشه و اختلاف نظر بوده است (Misselhorn, 2022a, pp. 31–32).

پیش‌فرض‌ها و مؤلفه‌های عاملیت اخلاقی

«عاملیت اخلاقی» به مجموعه‌ای از ویژگی‌های ضروری اشاره دارد که به یک موجود این امکان را می‌دهد تا به صورت آگاهانه، مستقل، و مسئولانه به‌عنوان یک عامل اخلاقی عمل کند. این ویژگی‌ها نه تنها پایه‌های نظری اخلاق را تشکیل می‌دهند، بلکه معیارهایی عملی برای تشخیص عاملیت اخلاقی در موجودات مختلف هستند. مؤلفه‌های کلیدی عاملیت اخلاقی عبارتند از:

درک اخلاقی و توانایی تشخیص درست از نادرست

این مؤلفه به توانایی تحلیل و ارزیابی موقعیت‌ها بر اساس اصول اخلاقی اشاره دارد. یک عامل اخلاقی باید بتواند اعمال خود و دیگران را بر مبنای مفاهیم اخلاقی قضاوت کند و تصمیم‌گیری‌هایی مبتنی بر ارزش‌های اخلاقی داشته باشد. این ویژگی مستلزم آشنایی با مفاهیمی مانند عدالت، انصاف، و مسئولیت‌پذیری است.

آگاهی و خودآگاهی[2]

«خودآگاهی» توانایی یک موجود برای شناخت و درک خود، از جمله افکار، احساسات، رفتارها، و ویژگی‌های شخصی، و همچنین توانایی تفکر درباره این شناخت و درک. این مفهوم شامل آگاهی از وجود خود به‌عنوان یک موجود مستقل و مجزا از دیگران در محیط پیرامون است. این ویژگی پیش‌شرط ضروری برای فهم موقعیت‌های اخلاقی و ارزیابی اعمال است. یک موجود خودآگاه می‌تواند اعمال خود را در چارچوب اخلاقی تحلیل کند، و از این طریق، به بهبود رفتار خود بپردازد. خودآگاهی همچنین به عامل این امکان را می‌دهد که نقش و وظیفه خود را در یک موقعیت اخلاقی درک کرده و بر اساس آن عمل کند.

خودمختاری/ خودآیینی

«خودمختاری» به استقلال و آزادی عمل در تصمیم‌گیری اشاره دارد، به‌گونه‌ای که عامل بتواند بدون تأثیرپذیری از محرک‌های بیرونی یا اجبار، انتخاب‌های آگاهانه و اخلاقی داشته باشد. این ویژگی یکی از

1. rimates
2. Consciousness and Self-awareness

ارکان اصلی عاملیت اخلاقی است، زیرا بدون خودمختاری، تصمیم‌گیری‌های اخلاقی فاقد ارزش ذاتی خواهند بود. خودمختاری مستلزم توانایی تفکر انتقادی و ارزیابی دلایل اخلاقی است.

ارادهٔ آزاد

«ارادهٔ آزاد» به توانایی تصمیم‌گیری و انجام اعمال آزادانه[1] بر اساس انتخاب‌های آگاهانه اشاره دارد. این مؤلفه عامل را قادر می‌سازد تا مسئولیت اعمال خود را بپذیرد و در قبال آن‌ها پاسخگو باشد. اراده آزاد مستلزم این است که عامل بتواند میان گزینه‌های مختلف انتخاب کند و تصمیم‌گیری‌های خود را بر اساس دلایل اخلاقی توجیه کند.

حساسیت عاطفی[2]

حساسیت عاطفی، به توانایی درک و پاسخ به احساسات خود و دیگران اشاره دارد. این مؤلفه به عامل این امکان را می‌دهد که در موقعیت‌های اخلاقی، نه تنها بر اساس منطق و اصول اخلاقی، بلکه بر اساس همدلی و درک احساسات نیز عمل کند. حساسیت عاطفی نقش کلیدی در تصمیم‌گیری‌های اخلاقی ایفا می‌کند؛ زیرا به عامل این امکان را می‌دهد که تأثیر اعمال خود بر دیگران را درک کند و بر اساس آن، تصمیم‌گیری‌های مسئولانه‌تری انجام دهد. این ویژگی به ویژه در موقعیت‌هایی که مستلزم همدلی و درک متقابل هستند، اهمیت پیدا می‌کند.

مسئولیت‌پذیری

مسئولیت‌پذیری به توانایی پذیرش مسئولیت اعمال و پیامدهای آن‌ها اشاره دارد. یک عامل اخلاقی باید بتواند اعمال خود را توضیح دهد، در قبال آن‌ها پاسخگو باشد، و در صورت لزوم، پیامدهای منفی اعمال خود را بشناسد و جبران کند. این ویژگی مستلزم درک ارتباط میان اعمال و نتایج آن‌ها، و همچنین تعهد به رعایت اصول اخلاقی است. البته، برخی این مولفه‌ها را در سه مولفه اصلی خلاصه کرده‌اند: ۱) خودآیینی، ۲) نیت‌مندی، و ۳) مسئولیت‌پذیری (Sullins, 2011, pp. 157–158).

بر این اساس و بطور خلاصه، عاملیت اخلاقی مفهومی چندبعدی است که مستلزم ترکیبی از درک اخلاقی، خودآگاهی، خودمختاری، ارادهٔ آزاد، و مسئولیت‌پذیری است. این مؤلفه‌ها نه تنها برای تشخیص عاملیت اخلاقی در انسان‌ها ضروری هستند، بلکه چارچوبی برای بررسی امکان عاملیت اخلاقی در ماشین‌ها و سیستم‌های هوشمند نیز فراهم می‌کنند. درک این ویژگی‌ها به ما کمک می‌کند تا مرزهای عاملیت اخلاقی را بهتر شناخته و چالش‌های مرتبط با آن را در دنیای فناوری و فلسفه اخلاق بررسی کنیم.

---

1. Decision-making and Free Will
2. Affective sensitivity

دیدگاه‌های اصلی در مورد عاملیت اخلاقی ماشین‌ها

نفی و انکار عاملیت اخلاقی ماشین‌های هوشمند

بسیاری از فیلسوفان و متفکران، با تردید و انکار به موضوع عاملیت اخلاقی ماشین‌های هوشمند نگریسته و بر این باورند که به‌رغم پیشرفت‌ها و ظرفیت‌های چشمگیر ماشین‌های به اصطلاح هوشمند نوظهور، این سیستم‌ها شرایط لازم برای پذیرش مسئولیت اخلاقی و شایستگی اطلاق عنوان «عامل اخلاقی» را دارا نیستند.

ایمانوئل کانت، عاملیت اخلاقی را تنها درخور موجوداتی می‌دانست که از عقلانیت[1] برخوردارند (Schlosser, 2019). او همچنین اراده آزاد و خودمختاری را پیش شرط مسئولیت اخلاقی می‌انگاشت و معتقد بود تنها موجوداتی که بهره‌مند از عقلانیت، خودمختاری، و اراده آزاد، هستند، می‌توانند عامل اخلاقی باشند (Talbert, 2024). لازمه این دیدگاه این است که عاملیت اخلاقی، دربارهٔ کودکان نابالغ، و همچنین انسان‌های بالغ اما بیماری که قابلیت تفکر و توان تصمیم‌گیری عقلانی و خودآیین ندارند – به رغم آنکه برخوردار از شأن اخلاقی هستند[2] – معنادار و متصور نیست. در واقع، هر عامل اخلاقی، برخوردار از شأن اخلاقی[3] و دارای ارزش ذاتی است و به همین جهت حقوق و منافع او تأمین شود و باید در تصمیم‌گیری‌های اخلاقی در نظر گرفته شود.

افزون بر کانت که در زمانی می‌زیست که حتی تصور وجود ماشین‌هایی تا این اندازه پیشرفته و هوشمند، نیز ناممکن می‌نمود، بسیاری از فیلسوفان معاصر و از جمله جان سِرل[4] (۱۹۳۲–...) – فیلسوف تحلیلی معاصر– همسو با نگاه کانت بر این اعتقادند که ماشین‌ها به جهت آنکه فاقد ویژگی‌های بنیادینی چون اراده آزاد، و نیت‌مندی[5] هستند، صلاحیت و شرایط پذیرش مسئولیت و پاسخگویی در قبال اعمال خود را ندارند و به همین دلیل، اطلاق عاملیت اخلاقی به ماشین‌های هوشمند بی‌معنا و ناموجه است. در نگاه او، ماشین‌های هوشمند، صرفاً ابزارهایی هستند که بر اساس قوانین برنامه‌ریزی‌شده عمل می‌کنند و کار آنها شبیه‌سازی و تقلید است. آن‌ها فاقد معنا یا تجربه ذهنی‌اند که برای تصمیم‌گیری اخلاقی ضروری است.

---

1. Rationality

۲. همچنین حیوانات که - بنا بر برخی مبانی– به رغم انکه قابلیت ادراک رنج و الم را دارند.

۳. رویکردها و دیدگاه‌ها درباره ملاک و معیار شأن اخلاقی، متفاوت و متنوع‌اند و برای مثال، بر مبنای انسان‌محوری در اخلاق زیستی، عضویت در گونهٔ انسان شرط لازم و کافی برای برخورداری از شأن اخلاقی است. در دهه‌های اخیر برخی همچون پیتر سینگر– فیلسوف استرالیایی معاصر– ظرفیت و قابلیت تجربه درد و لذت را شرط لازم و معیار موجه برخورداری از شأن اخلاقی دانسته‌اند (Jaworska & Tannenbaum, 2023b).

4. John R. Searle

5. Intentionality

سِرل ضمن طرح آزمایش فکری «اتاق چینی»[۱] استدلال می‌کند که ماشین‌های هوشمند، صرفاً از قواعد و نمادها پیروی می‌کنند و کار رایانه‌ها و ماشین‌های هوشمند داده‌پردازیست و آنها تنها نحو را پردازش می‌کنند، نه معنا را و صرفاً بر اساس الگوریتم‌ها و پردازش نمادها کار می‌کنند، بدون اینکه تجربه ذهنی یا درک واقعی از آنچه انجام می‌دهند داشته باشند و به همین جهت، فاقد درک واقعی و آگاهی[۲]اند.(Searle, 1980, pp. 429–430) او تأکید می‌کند که عاملیت اخلاقی همبسته و منوط به شعور و آگاهی است. تصمیم‌گیری اخلاقی به توانایی تجربه احساسات، همدلی و درک پیامدهای عمل بستگی دارد و در واقع، آنچه به عنوان اعمال به ماشین‌های هوشمند، منتسب می‌شود، نتیجه برنامه‌ریزی انسان‌هاست و نه تصمیم‌گیری خودآیین ماشین‌ها و حتی اگر یک ماشین بتواند رفتار اخلاقی را شبیه‌سازی کند، در نهایت کار آن تقلید و شبیه‌سازی است (Searle, 1980, p. 431).

برخی فیلسوفان نیز بر توان درک و احساس و حساسیت اخلاقی[۳] به عنوان مهمترین پیش‌نیازهای عاملیت اخلاقی تاکید کرده و بر محدودیت‌های این سیستم‌های هوشمند از این منظر انگشت نهاده‌اند.(Graff, 2024, pp. 9–11). حساسیت اخلاقی به توانایی تشخیص ویژگی‌های اخلاقی مرتبط در موقعیت‌های مختلف و نحوه ارتباط آن‌ها اشاره دارد و برخی غیرقابل‌کدگذاری بودن و از سنخ مهارت عملی بودن را از ویژگی‌های بارز آن دانسته‌اند؛ به این معنا که اولاً، حساسیت اخلاقی نمی‌تواند به‌صورت مجموعه‌ای از قوانین جامع و از پیش تعریف‌شده کُدگذاری شود. این به دلیل وابستگی شدید تصمیمات اخلاقی به زمینه‌های خاص و عوامل منحصربه‌فرد هر موقعیت است. به‌طور مثال، روابط انسانی دارای ویژگی‌های کیفی منحصربه فردی هستند که نمی‌توان آن‌ها را به قوانین کلی تقلیل داد.(Graff, 2024, p. 9) و ثانیاً، حساسیت اخلاقی یک مهارت عملی است که نمی‌توان آن را به سادگی آموزش داد. این مهارت شبیه به توانایی‌های مانند بداهه‌نوازی در موسیقی است و نیاز به تمرین و تجربه‌اندوزی مستمر دارد.

با توجه به آنچه دربارهٔ شروط و شرایط عاملیت اخلاقی بیان شد، می‌توان عاملیت ماشین‌های هوشمند کنونی را نیز در چارچوب این معیارها مقایسه و تحلیل کرد. این مقایسه، دلایل انکار عاملیت ماشین‌های هوشمند را آشکار می‌سازد و نشان می‌دهد که ماشین‌های هوشمند، به رغم پیشرفت‌های چشمگیر، هنوز فاصلهٔ قابل توجهی با انسان‌ها به عنوان عاملان اخلاقی دارند.

الف) اراده آزاد در مقابل جبرگرایی؛ عاملیت اخلاقی، بر مبنای تعاریف دقیق و مصطلح، بر اراده آزاد استوار است، یعنی توانایی انتخاب و تصمیم‌گیری مستقل بر اساس دلایل اخلاقی. حال آنکه، ماشین‌های

---

۱. آزمایش فکری اتاق چینی (Chinese Room) نشان داد که کامپیوترها به مثابه ماشینهای هوشمند، شاید بتوانند رفتارهای هوشمندانه از خود نشان دهند، اما این به معنای داشتن آگاهی یا فهم واقعی نیست. این آزمایش فکری چالشی جدی برای نظریه‌هایی است که ذهن را صرفاً به عنوان یک سیستم محاسباتی در نظر می‌گیرند.

2. consciousness

3. Moral Sensitivity

هوشمند کنونی متاثر و تابع جبرگرایی الگوریتمی[1] هستند و تصمیمات آن‌ها کاملاً تحت‌الشعاع برنامه‌ریزی‌های معین و از پیش تعیین‌شده قراردارد و منوط به پردازش داده‌ها است. به عبارت دیگر، ماشین‌ها نمی‌توانند خارج از چارچوب فرامین و الگوریتم‌های تعیین شده عمل کنند یا انتخاب‌هایی کاملاً خودانگیخته داشته باشند. این موضوع به فلسفه عمل و تمایز میان علل درونی و علل بیرونی بازمی‌گردد. اعمالِ عاملیت اخلاقی ناشی از علل درونی مانند اراده و نیت است، در حالی که اعمال ماشین‌ها ناشی از علل بیرونی نظیر برنامه‌ریزی و داده‌ها است.

ب) نیت‌مندی در مقابل فرایندهای مکانیکی؛ نیت‌مندی به معنای داشتن هدف و انگیزهٔ آگاهانه در انجام یک عمل است. انسان‌ها قادرند اعمال خود را بر اساس نیت‌ها و انگیزه‌های اخلاقی انجام دهند، در حالی که اعمال ماشین‌ها صرفاً نتیجهٔ پردازش داده‌ها و اجرای الگوریتم‌هاست. ماشین‌ها فاقد قصد و نیت واقعی هستند و نمی‌توانند انگیزه‌های اخلاقی را درک یا تجربه کنند. و تنها بازتاب دهنده اهداف و نیت‌های طراحان و کاربران خود هستند. این تفاوت بنیادین، عاملیت اخلاقی ماشین‌ها را زیر سؤال می‌برد.

ج) پیامد در مقابل انگیزه و نیت؛ ماشین‌ها عمدتاً بر اساس پیامدها برنامه‌ریزی می‌شوند. برای مثال، یک خودروی خودران ممکن است برای کاهش تصادفات یا بهینه‌سازی مسیر حرکت طراحی شود. این رویکرد با پیامدگرایی در اخلاق همسو است که در آن ارزش اخلاقی یک عمل بر اساس نتایج آن سنجیده می‌شود. با این حال، در اخلاق – به ویژه در نگاه وظیفه گرایانهٔ کانت – نیت و انگیزه نقش و اهمیتی ویژه و بلکه بی بدیل دارد. این تمایز نشان می‌دهد که ماشین‌ها، حتی اگر بتوانند پی‌جوی ببار آوردن بهترین پیامدها باشند، فاقد نیت اخلاقی هستند و دستکم از منظر فضیلت‌گرایی و نگاه وظیفه‌گرایانه، «عامل اخلاقی» به معنای دقیق کلمه نیستند.

د) خودآگاهی و هدف‌مندی؛ عاملیت اخلاقی مستلزم خودآگاهی است، یعنی توانایی شناخت خود و تمایز میان خود و دیگران. خودآگاهی به عامل این امکان را می‌دهد که اعمال خود را بازتاب دهد و آن‌ها را در چارچوب اخلاقی تحلیل کند؛ حال آنکه ماشین‌ها فاقد خودآگاهی و تجربهٔ ذهنی هستند. از دیدگاه توماس نیگل، خودآگاهی و تجربهٔ ذهنی ویژگی‌هایی هستند که مختص موجودات آگاهند و ماشین‌ها به دلیل ماهیت محاسباتی خود، فاقد این ویژگی‌ها هستند. به بیان دیگر، عامل اخلاقی، خودمختار و هدفمند است، در حالی که ماشین‌ها صرفاً داده‌ها را پردازش می‌کند و فاقد آگاهی پدیداری یا درک واقعی از شرایط و اعمال خود هستند و تجربه ذهنی[2] (ندارند. عاملیت ماشین‌ها تابعی از طراحی انسان است و بی‌بهره از استقلال واقعی است؛ به این معنا ماشین‌ها نمی‌توانند اهداف خود را تعیین کنند یا به‌طور مستقل عمل کنند.

1. Algorithmic Determinism
2. Subjective Experience

با توجه به این نکات و تحلیل‌ها می‌توان گفت که ماشین‌های هوشمند کنونی، به‌غم توانایی‌های چشمگیر در پردازش داده‌ها و انجام وظایف پیچیده، هنوز فاقد مؤلفه‌های اساسی عاملیت اخلاقی، همچون اراده آزاد، و خودآگاهی‌اند و به این دلایل، انتساب مسئولیت به آنها و انتظار مسئولیت‌پذیری از آنها بی‌معنا و بی‌وجه است. این محدودیت‌ها نشان می‌دهد که ماشین‌ها، در بهترین حالت، می‌توانند به عنوان ابزارهایی پیشرفته در خدمت اهداف اخلاقی انسان‌ها عمل کنند، اما نمی‌توانند به‌طور مستقل به عنوان عامل‌های اخلاقی حقیقی در نظر گرفته شوند. این موضوع چالش‌های مهمی را برای آیندهٔ اخلاق فناوری و نقش ماشین‌ها در تصمیم‌گیری‌های اخلاقی در بر دارد.

افزون بر این، ماشین‌ها نمی‌توانند احساساتی مانند همدلی، احساس گناه یا تعهد اخلاقی را تجربه کنند، حال آنکه این احساسات نقش و نفوذ مهم و اثرگذاری در تصمیم‌گیری اخلاقی دارند.

بر این اساس می‌توان، عاملیت اخلاقی (برآمده از اراده آزاد، نیت‌مندی، و مسئولیت‌پذیری) را که مختص موجودات خودآگاه و خودآیین است متفاوت با عاملیت ماشین‌ها (توانایی انجام وظایف بر اساس الگوریتم‌ها و داده‌ها و تحت کنترل برنامه‌ریزی‌های انسانی) دانست.

فیلسوفان و متفکرانی که عاملیت اخلاقی ماشین‌های هوشمند را نفی و انکار کرده‌اند، را می‌توان به دو دسته تقسیم کرد:

دسته اول، فیلسوفانی همچون سِرل[1] که قاطعانه و به صراحت، عاملیت اخلاقی را منحصر در انسان دانسته و امکان عاملیت اخلاقی ماشین‌های هوشمند را به نحو مطلق، و از اساس امکان‌ناپذیر می‌دانند. این دسته بر این نکته و ادعا اصرار می‌ورزند که بطور کلی غیر انسان‌ها و بویژه رُبات‌ها هرگز اراده مستقلی نداشته و نخواهند داشت؛ زیرا هرگز - چه اینک و چه در آینده - نمی‌توانند کاری را انجام دهند که برای انجام آن برنامه‌ریزی نشده‌اند.[2] این دسته در واقع از ایده و انگاره «منحصر به فرد و استثنایی بودن انسان[3]» در برخورداری از شأن اخلاقی و عاملیت اخلاقی دفاع می‌کنند و سِرل از پیشگامان این ایده است. به اعتقاد او، نسبت دادن عاملیت اخلاقی به ماشین‌ها باب و بهانه‌ای برای فرار از مسئولیت انسانی است و اساساً، انسان به جهت برخورداری از قوه خِرد و همچنین قابلیت‌ها و ظرفیت‌های خارق العاده و جایگاه

---

۱. راجر اسکروتن، هیلاری پاتنم، و توماس نیگل را نیز می‌توان در این دسته جای داد.

۲. البته، در تقابل تام با این نگاه و دیدگاه، برخی به این دیدگاه بسیار شاذ باور دارند که تنها ربات‌ها و ماشین‌های هوشمند، عاملان اخلاقی تمام عیار و کامل‌اند؛ زیرا کسی را می‌توان عامل اخلاقی دانست که بر مبنایی کاملاً منطقی عمل کند و تصمیم‌گیری و عمل او عاری و بری از هرگونه سوگیری و سانقه‌ها و امیال درونی و بیرونی باشد و از آنجا که آدمیان هیچ‌گاه تا این حد مُنسلخ از امیال و آزاد در مقام عمل و تصمیم‌گیری نبوده و نیستند، نمی‌توانند عاملان اخلاقی تمام‌عیار باشند و اگر رُبات‌ها دقیق طراحی شوند، به لحاظ اینکه فارغ از حُب و بغض شخصی و کاملا آزادانه و بی‌طرفانه تصمیم می‌گیرند، مصداق پارادایمی و نمونه و نوع‌نمون عامل اخلاقی‌اند (see Sullins, 2011, p. 156).

3. Human Exceptionalism

خاص در عالم، از سایر موجودات و ماشین‌ها متفاوت و متمایز است و گویی شرط انتساب مسئولیت و عاملیت اخلاقی، انسان بودن با شرایط خاص است و غیر انسان، هرگز شایسته اطلاق عامل اخلاقی نخواهد بود.

پیروان این دیدگاه استدلال می‌کنند اخلاقیات تنها در زمینه‌ای که آگاهی و قصد وجود دارد، معنا پیدا می‌کند و هوش مصنوعی هر چقدر هم که پیچیده باشد، فاقد ویژگی‌های غیرقابل جایگزین انسان است و بنابراین، نمی‌تواند برخوردار از عاملیت اخلاقی مستقل باشد (Sullins, 2011, p. 156).

در دسته دوم، فیلسوفانی همچون – دنیل دِنِت[1] (۱۹۴۲ -۲۰۲۴م) – فیلسوف آمریکای معاصر - قرار دارند که معتقدند مصنوعات هوشمند کنونی فاقد شروط و شرایط عاملیت اخلاقی و از جمله خودآیینی و نیت‌مندی واقعی هستند، اما این بدان معنا نیست که عاملیت غیر انسانها منطقا و مطلقا ناممکن باشد.

به اعتقاد دانیل دنت، رُبات‌های هوشمند کنونی چه بسا واجد حالات ذهنی و احساسی از جنس احساس شرم و گنهکاری و یا حالت‌های انگیزشی[2] و یا باورهای شناختی، و حتی حالاتی چون غفلت و سهل‌انگاری باشند، اما باز هم برای برخورداری از عاملیت اخلاقی واقعی و تمام عیار، باید واجد ویژگی قصدیت و نیت‌مندی مرتبه بالاتر[3] باشند و به تعبیری، بتوانند باورهایی درباره باورها، خواسته‌هایی درباره خواسته‌ها، و حتی ترس‌ها و امیدهای خودشان را داشته باشند، بنابر این رُبات‌ها در حال حاضر عاملان اخلاقی نیستند، اما ممکن است در آینده، توانایی امکان و شرایط کسب این عاملیت را بیابند (Dennett, 2014, pp. 350–352).

به بیان دیگر، دنت، به رغم آنکه عاملیت اخلاقی هوش مصنوعی، در حال حاضر و با فناوری‌های موجود را یک آرزو و آرمان می‌انگارد تا یک واقعیت عملی، در استثنا بودن آدمیان تردید روا داشته و با رویکردی محتاطانه، عاملیت ماشین‌های هوشمند را در بُقعه امکان نهاده و باب تحقق آن را باز گذاشته است.

این دو نگاه و دیدگاه، نشان‌دهنده اختلاف نظرهای عمیق درباره ماهیت عاملیت اخلاقی و قابلیت‌های بالقوه فناوری است.

عاملیت اخلاقی محدود

برخی نیز با عدول از تلقی رسمی و رایج درباره معنای عامل اخلاقی، در معنای «عاملیت اخلاقی» تصرف کرده و اساسا به طیفی و تشکیکی بودن عاملیت اخلاقی قائل‌اند. به این معنا که انسانها به جهت برخورداری از سطح بالایی از آگاهی و حساسیت و خودآیینی، واجد عاملیت اخلاقی کامل و تمام‌عیار

1. Daniel Dennett
2. motivational states
3. Higher Order Intentionality

هستند[1] و ماشین‌های هوشمند و فرا هوشمند، در رتبه بعدی و با فاصله ای زیاد/اندک، به عنوان ۲ سیستم‌های اخلاقی عملکردی[2] می‌توانند به سطحی از قابلیت عملکردی دست یابند که در دامنه‌ای محدود از موقعیت‌ها تصمیمات اخلاقی مناسب و قابل‌اعتماد اتخاذ کنند (Graff, 2024). بر این اساس، می‌توان این ماشین‌ها را به یک معنا برخوردار از عاملیت اخلاقی، اما محدود به حوزه‌های خاص[3] دانست.[4]

این دسته، معنایی رقیق و رنگ‌باخته از عاملیت اخلاقی را در نظر می‌گیرند و معتقدند ماشین‌های هوشمند را می‌توان به این معنا و در شرایط خاص و محدود، عامل اخلاقی به حساب آورد. آنها تأکید می‌کنند که حالات ذهنی یا نیت‌مندی[5] برای عاملیت اخلاقی ضروری نیستند. ماشین‌ها می‌توانند وارد بازی اخلاقی شوند و اخلاقی عمل کردن را شبیه‌سازی کنند، حتی اگر قاصد نباشند و «نیت» نداشته باشند (Floridi & Sanders, 2004, p. 13).

این دسته، می‌پذیرند که ماشین‌ها ممکن است مسئولیت‌پذیر نباشند و یا حتی درکی از مسئولیت اخلاقی نداشته باشند، اما این به معنای نفی عاملیت اخلاقی آنها نیست. آنها عاملیت اخلاقی را از مسئولیت اخلاقی جدا می‌کنند و استدلال می‌کنند که ماشین‌ها می‌توانند منبع اعمال اخلاقی باشند، حتی اگر مسئولیت‌پذیر نباشند. برای مثال، استدلال اصلی فلوریدی و سندرز، بر پایهٔ مفهوم «سطح انتزاع»[6] استوار است. (Floridi & Sanders, 2004, pp. 1–3) و معتقدند که اگر ماشین‌های هوشمند در یک سطح انتزاع مناسب مورد بررسی قرار گیرند، می‌توانند به‌عنوان عاملان اخلاقی در نظر گرفته شوند، حتی اگر فاقد اراده آزاد، حالات ذهنی یا مسئولیت اخلاقی باشند. آنها برای این منظور سه معیار – تعامل‌پذیری[7]، خودمختاری[8]، بهبودبخشی بر اساس تجارب پیشین (سازگاری)[9] – را برای شناسایی و تعیین عاملیت اخلاقی پیشنهاد می‌کنند (Floridi & Sanders, 2004, p. 7).

بر این مبنا، اگر یک ماشین هوشمند این سه ویژگی را در سطح انتزاعی خاص داشته باشد، می‌توان آن را یک عامل اخلاقی دانست، حتی اگر فاقد ویژگی‌هایی مانند آگاهی یا قصد باشد. این دیدگاه به انگاره و

---

1. Full Moral Agency
2. Functionally Moral Systems یا FMSs
3. Restricted Moral Domains

٤. فلوریدی، مفهوم «عاملیت اخلاقی مصنوعی محدود» (Artificial Moral Agency) را مطرح کرده است که نشان‌دهنده توانایی محدود ماشین‌ها در انجام وظایف اخلاقی در حوزه‌های خاص است، اما این توانایی به معنای برخورداری از شأن اخلاقی و یا عاملیت اخلاقی کامل نیست (see Floridi & Chiriatti, 2020).

5. Intentional States
6. Level of Abstraction/LoA
7. Interactivity
8. Adaptability
9. Adaptability

نظریهٔ «اخلاق بدون ذهن» معروف شده است[1] (Floridi & Sanders, 2004, p. 3).

«اخلاق بدون ذهن» به امکان وجود ماشین‌ها و سیستم‌های اخلاقی با تصمیم‌گیری‌های اخلاقی بدون نیاز به آگاهی اشاره دارد. این ایده در حوزه‌های فلسفه اخلاق، هوش مصنوعی و علوم شناختی محل بحث است و در پی پاسخ به این سؤال است که آیا می‌توان سیستم‌هایی طراحی کرد که قادر به تصمیم‌گیری‌های اخلاقی باشند، بی آنکه دارای آگاهی یا حتی ذهن باشند؟ «اخلاق بدون ذهن» به این معناست که یک سیستم (برای مثال، یک ربات یا نرم‌افزار) می‌تواند بر اساس قواعد، الگوریتم‌ها یا داده‌های برنامه‌ریزی‌شده، تصمیماتی بگیرد که از نظر اخلاقی قابل قبول یا حتی مطلوب باشند، بدون اینکه این سیستم دارای آگاهی، احساسات یا درک واقعی از مفاهیم اخلاقی باشد. این مفهوم با این ایده مرتبط است که اخلاق می‌تواند به عنوان یک فرآیند محاسباتی یا الگوریتمی در نظر گرفته شود؛ بدون نیاز به وجود آگاهی یا تجربه ذهنی.

برای مثال، سیستم‌های هوش مصنوعی می‌توانند بر اساس الگوریتم‌های پیچیده، تصمیماتی بگیرند که از نظر اخلاقی قابل دفاع باشند. برای مثال، یک خودروی خودران ممکن است در موقعیتی قرار بگیرد که مجبور به انتخاب بین دو گزینه باشد (مانند مسئله ترامواِ[2]). این سیستم می‌تواند بر اساس قواعد اخلاقی برنامه‌ریزی‌شده، تصمیمی بگیرد که از نظر اخلاقی درست و دقیق باشد، بدون اینکه دارای آگاهی یا درک اخلاقی باشد. همچنین برخی از سیستم‌های نرم‌افزاری برای کمک به تصمیم‌گیری‌های اخلاقی در حوزه‌هایی مانند پزشکی، حقوق یا مدیریت طراحی شده‌اند. این سیستم‌ها می‌توانند بر اساس داده‌ها و قواعد اخلاقی، توصیه‌هایی ارائه دهند، بدون اینکه دارای آگاهی یا ذهن باشند.

از این منظر، اگر یک سیستم بتواند تصمیماتی بگیرد که از نظر اخلاقی قابل قبول باشند، نیازی به داشتن آگاهی یا ذهن نیست و چه بسا، سیستم‌های هوش مصنوعی بتوانند تصمیمات اخلاقی را با دقت، و سرعت به مراتب بالاتری نسبت به انسان‌ها اتخاذ کنند؛ زیرا بنا به ادعا، تحت تأثیر احساسات، تعصبات یا محدودیت‌های شناختی قرار ندارند. به بیان دیگر، ماشین‌های هوش مصنوعی به دلیل ناتوانی در دستیابی به حساسیت اخلاقی کامل، نمی‌توانند عاملان اخلاقی کامل باشند. با این‌حال، ممکن است در حوزه‌های خاصی که حساسیت اخلاقی نقش کمتری دارد (مانند تریاژ[3] در شرایط بحرانی و یا تصمیم‌گیری‌های عمومی)، کارایی و عملکرد درست و قابل‌قبولی داشته باشند (Graff, 2024, p. 10).

بر اساس این تفکیک و بازاندیشی معنایی، می‌توان گفت که عاملیت اخلاقی در انحصار انسان‌ها

---

1. Mind-less Morality
2. trolley problem
۳. تریاژ (Triage) - مشتق از واژه فرانسوی trier به معنای «مرتب‌سازی» - به فرآیند تصمیم‌گیری و اولویت‌بندی در مدیریت منابع محدود اشاره دارد که به طور خاص، در حوزه‌های پزشکی، اورژانس و مدیریت بحران کاربرد دارد.

نیست و برخی ماشین‌ها و چه بسا رُبات‌ها نیز حَظ و حصه‌ای از عاملیت اخلاقی خواهند داشت. (Shapiro, 2006)[1] در واقع، این دسته با تصرف در معنا و مدلول عامل اخلاقی و ارائهٔ تعریفی جدید در صدد بسط معنایی این واژه بر آمده و معتقدند که گسترهٔ مفهومی «عامل اخلاقی» باید توسعه گسترش یابد تا شامل موجودات غیر انسانی و همچنین شرکت‌ها، سازمان‌ها، و ماشین‌های هوشمند و حتی حیوانات شود و اذعان دارند که چنین گسترشی می‌تواند به درک بهتر مسایل اخلاقی جدید، به‌ویژه در اخلاق کامپیوتر کمک کند (Floridi & Sanders, 2004, pp. 10–13).

بر این مبنا، عاملیت اخلاقی یک مفهوم طیفی و سیال - با پذیرش سطوح مختلف عاملیت بر اساس پیچیدگی و توانایی تصمیم‌گیری ماشین‌ها - است و ماشین‌ها می‌توانند در سطوح محدودتر این طیف قرار گیرند و به عنوان ابزارهای اخلاقی به انسان‌ها در تصمیم‌گیری اخلاقی کمک کنند.

در همین راستا، برخی با برجسته ساختن تمایز مفهومی «عاملیت عقلانی» و «عاملیت اخلاقی»، بر این باورند که ماشین‌های هوشمند می‌توانند به سطحی از عاملیت عقلانی[2] دست یابند، اما نیل به عاملیت اخلاقی همچنان مبهم و محال و دست‌کم در حد یک فرض و چالش باقی می‌ماند. همچنین در برخی تعابیر از دو گونه / دو تقریر از عاملیت اخلاقی (تقریر حداکثری/ استاندارد[3] و. تقریر حداقلی[4] یا ابزارگرایانه[5]) سخن به میان آمده است (Manna & Nath, 2021, p. 3). بر این اساس، در حالی که ماشین‌های هوشمند می‌توانند به عامل اخلاقی حداقلی یا ابزارگرایانه به حساب آیند، عاملیت اخلاقی کامل نیازمند آگاهی و حالات ذهنی است که فعلاً در ماشین‌ها وجود ندارد.

دوگانهٔ دیگری نیز همراستا با تقریر و تقسیم دوگانهٔ پیشین، مطرح شده و برخی ماشین‌های هوشمند را بر اساس سطح هوشمندی و خودمختاری، به دو دسته ماشین‌های ضمنی[6] و ماشین‌های صریح تقسیم کرده‌اند (Bonnefon et al., 2024) که البته در طبقه‌بندی و تقسیم مشهور جیمز مور - جیمز مور (۱۹۴۲-۲۰۲۴) - فیلسوف اخلاق پیشگام در حوزهٔ اخلاق کامپیوتر- نیز بازتاب یافته است.[7]

---

۱. به طور خلاصه، برخی با ذو مراتب بودن و تشکیکی انگاشتن عاملیت اخلاقی بر این باورند که هر چه یک عامل اخلاقی از ظرفیت‌های شناختی بیشتری برخوردار باشد، به مراتب مسولیت بیشتری متوجه اوست و در مقابل، به همان اندازه که ظرفیت‌های شناختی کمتری دارد، تعهدات و مسئولیت کمتری خواهد داشت. و بر این اساس، چه بسا آن دسته از حیواناتی که قادر به نوع‌دوستی متقابل هستند، مسئولیت بیشتری دارند و برای مثال دلفین‌ها که گاه با شناور نگه داشتن ملوانان در شرف غرق شدن، جان آنها را نجات می‌دهند، در خور مذمت و یا در مظان پذیرش مسولیت‌اند و بنوعی ملزم‌اند که به منافع انسان‌ها اهمیت دهند (Shapiro, 2006, p. 364).

2. Rational Agency
3. Standard Conception of Agency
4. Minimal Agency
5. Instrumentalist Conception of Agency
6. Implicit Machines
7. Explicit Machines

جیمز مور – در یک تقسیم مشهور – عوامل اخلاقی را بر اساس سطح توسعه‌یافتگی اخلاقی و توانایی سیستم‌ها برای درک و عمل بر اساس اصول اخلاقی عوامل اینگونه طبقه‌بندی و تقسیم کرده است: (Moor, 2006, pp. 19–21).

۱. عوامل اخلاقی تأثیرگذار:[1] این مصنوعات هوشمند/ نیمه‌هوشمند، تأثیرات اخلاقی یا غیر اخلاقی می‌گذارند، بدون اینکه نیت یا درک و استدلال اخلاقی داشته باشند (مانند دستگاه‌های دیجیتال ساده).

۲. عوامل اخلاقی تابع / ضمنی:[2] این ماشین‌ها فاقد درک مستقیم از اصول اخلاقی هستند و تنها به صورت مکانیکی، نتیجه‌ای را تولید می‌کنند که ممکن است با اصول اخلاقی همسو باشد. این ماشین‌های هوشمند برای جلوگیری از نتایج غیر اخلاقی برنامه‌ریزی‌شده‌اند (مانند سیستم‌های ایمنی در خودروهای خودران). و البته، به رغم آنکه سطح بالاتری از تکنولوژی را بازنمایی می‌کنند، قادر به ترجیح و تأمل و استدلال اخلاقی نیستند و تنها، به شکل غیرمستقیم از اصول اخلاقی، پیروی می‌کنند و توانایی آنها ارزیابی پیامدهای اخلاقی نیز بسیار محدود است.

۳. عوامل اخلاقی صریح:[3] ماشین‌هایی که توانایی شناسایی و پردازش اصول اخلاقی برای تصمیم‌گیری‌های اخلاقی را دارند. این ماشین‌ها به سطح بالاتری از هوشمندی و خودمختاری توسعه یافته‌اند و می‌توانند اصول اخلاقی را به طور مستقیم در تصمیم‌گیری‌های خود به کار گیرند. در این سطح، سیستم‌ها نه تنها از مجموعه‌ای از قواعد اخلاقی پیروی می‌کنند، بلکه قادر به تحلیل و تفسیر این اصول نیز هستند. توانایی درک و تحلیل اصول اخلاقی و همچنین، قابلیت ارزیابی گزینه‌های مختلف بر اساس پیامدهای اخلاقی از ویژگیهای این گونه از عوامل اخلاقی است. البته، باز هم از داده‌ها و الگوریتم‌ها برای ارزیابی عوامل اخلاقی بهره می‌گیرند و خودآیینی تام و تمام ندارند و به نوعی بالمآل غیر مستقل و تحت نظارت انسان هستند.

۴. عوامل اخلاقی کامل و تمام عیار/ قوی. این دسته که ایده‌آل‌ترین و پیشرفته‌ترین نوع عوامل اخلاقی محسوب می‌شوند، نه تنها از اصول اخلاقی آگاه هستند، بلکه دارای سطح بالایی از خودآگاهی و قابلیت قضاوت مستقل اخلاقی هستند. این ماشین‌ها می‌توانند تصمیمات خود را بر اساس زمینه، شرایط، و پیامدهای اخلاقی به صورت کاملاً مستقل اتخاذ کنند. و بالتبع، دارای ویژگی‌هایی مانند آگاهی اراده آزاد و نیت اخلاقی اخلاقی هستند که به اذعان و اقرار صاحب‌نظران، تا کنون، هوش مصنوعی به این سطح و مرحله دست نیافته و تنها انسانها را می‌توان مصداق عامل اخلاقی کامل و تمام عیار به حساب آورد (Moor, 2006, pp. 19–21).

1. Ethical Impact Agents
2. Implicit Ethical Agents
3. Explicit Ethical Agents

نتیجه‌گیری

پیشرفت‌های خیره‌کننده و خارق‌العاده در حوزهٔ هوش مصنوعی، به‌ویژه در حوزه‌هایی نظیر یادگیری عمیق، تصمیم‌گیری خودکار و تعامل انسان و ماشین، این امکان را فراهم کرده است که ماشین‌ها در موقعیت‌های مختلف تصمیم‌گیری کنند و حتی اقداماتی انجام دهند که پیامدهای اخلاقی به دنبال دارند. این تحولات، پرسش اساسی را پیش می‌کشد که آیا ماشین‌های هوشمند می‌توانند به‌عنوان عامل اخلاقی تلقی شوند؟ در این مقاله تلاش کردیم تا ضمن تحلیل دقیق مفهوم عاملیت اخلاقی و مولفه‌های آن و بررسی اقوال و دیدگاه‌ها به این پرسش مهم پاسخ دهیم.

برای آن‌که یک موجود به‌عنوان عامل اخلاقی شناخته شود، باید واجد ویژگی‌هایی نظیر آگاهی، خودآگاهی، اراده آزاد، و توانایی پذیرش مسئولیت باشد. تحلیل نظریه‌های فلسفی اخلاق، به‌ویژه در سنت کانتی، نشان می‌دهد که عاملیت اخلاقی مستلزم برخورداری از مکانیسم‌های عاطفی و شهود اخلاقی است که در انسان‌ها نقش اساسی در تصمیم‌گیری‌های اخلاقی ایفا می‌کنند. ماشین‌های هوشمند، به‌ویژه با معماری کنونی هوش مصنوعی، فاقد چنین ویژگی‌هایی هستند. آنها بر اساس الگوریتم‌ها و داده‌های از پیش تعریف‌شده عمل می‌کنند و توانایی درک مفاهیمی نظیر عدالت، همدلی، و مسئولیت اخلاقی را ندارند. به این دلیل، نمی‌توان آنها را به‌عنوان عاملان اخلاقی تمام‌عیار در نظر گرفت.

برخی فیلسوفان، به‌ویژه در سنت کانتی، عاملیت اخلاقی را به‌عنوان یک ایده‌آل انتزاعی و کامل تعریف کرده‌اند که حتی انسان‌ها نیز به‌طور کامل به آن دست نمی‌یابند. از این منظر، ممکن است ماشین‌های هوشمند با توجه به توانایی تصمیم‌گیری منطقی و بی‌طرفانه، در برخی زمینه‌ها بهتر از انسان‌ها به این استانداردهای ایده‌آل نزدیک شوند. با این حال، چنین توانایی‌هایی به معنای عاملیت اخلاقی واقعی نیست، چرا که بیشتر تصمیم‌گیری‌های اخلاقی انسان‌ها به‌طور ناخودآگاه و مبتنی بر مکانیسم‌های عاطفی انجام می‌شود. افزون بر این، عاملیت اخلاقی تنها به توانایی شناختی محدود نمی‌شود، بلکه نیازمند شناخته شدن توسط جامعه به‌عنوان یک عامل اخلاقی است. در حال حاضر، جامعه ماشین‌های هوشمند را به‌عنوان عاملان اخلاقی نمی‌شناسد و بنابراین، آنها نه از نظر درونی (فاقد مکانیسم‌های عاطفی) و نه از نظر بیرونی[1] (عدم تأیید اجتماعی) واجد شرایط عاملیت اخلاقی نیستند.

اگر عاملیت اخلاقی را به‌صورت تشکیکی و ذو مراتب در نظر بگیریم، می‌توان ماشین‌های هوشمند را در معنایی مسامحی، محدود یا استعاری به‌عنوان عامل اخلاقی تلقی کرد. در این رویکرد، انسان‌ها در صدر عاملان اخلاقی قرار دارند و به دلیل ویژگی‌های منحصربه‌فردشان، عاملان اخلاقی تمام‌عیار و بی نِد و نظیر محسوب می‌شوند. در مقابل، ماشین‌های هوشمند، به دلیل برخی شباهت‌ها در رفتارهایشان با انسان، می‌توانند تصویر رقیق و تقلیل‌یافته‌ای از عاملیت اخلاقی را به ذهن متبادر کنند. این مفاهیم در

---

۱. به عنوان یک شرط بیرونی (External Condition) (see Brożek & Janik, 2019)

برخی منابع با عنوان «عاملیت اخلاقی مصنوعی» مطرح شده است.

ماشین‌های هوشمند، به‌ویژه ماشین‌های فراهوشمند، به دلیل برخی ویژگی‌ها و از جمله، تقلید رفتار انسانی، تصمیم‌گیری‌های سریع در موقعیت‌های پیچیده، یادگیری و سازگاری و تطبیق با محیط و خودبهبودبخشی، خودآیینی نسبی و استقلال عملکردی، تعامل با انسان‌ها، ممکن است شباهت‌هایی با آدمیان داشته باشند و این ویژگی‌ها تقویت کنندهٔ احتمال و انگاره امکان عاملیت اخلاقی تمام‌عیار این ماشین‌ها باشند، با این همه باید توجه داشت که با وجود توانایی‌های بسیار، ماشین‌های هوشمند فاقد احساس و حساسیت اخلاقی و حالات و مکانیسم‌های عاطفی نظیر همدلی و شهود هستند و چنان داده‌محور و وابسته به الگوریتم‌ها هستند که نمی‌توانند به شیوه مستقل و خارج از چارچوب داده‌ها و الگوریتم‌ها اخذ تصمیم کنند و در نتیجه، توانایی ارزیابی اخلاقی مستقل ندارند. همچنین، عدم خودآگاهی و ناتوانی در پیش‌بینی و ارزیابی پیامدهای بلندمدت و پیچیده که بخشی از روند و فرایند ارزش‌گذاری اخلاقی و یکی از پیشفرض‌های عاملیت اخلاقی است، مانعی بر سر انتساب مسئولیت اخلاقی به ماشین‌های خودکار و عامل اخلاقی دانستن آنهاست.

بر این اساس، مسئولیت نهایی همچنان بر عهده انسان‌هایی است که این ماشین‌ها و سیستم‌ها را طراحی و برنامه‌ریزی کرده‌اند. این انسان‌ها هستند که ارزش‌ها و معیارهای اخلاقی را به ماشین‌ها منتقل می‌کنند و دغدغهٔ اصلی باید مسئولیت‌پذیری انسان‌ها در قبال اعمال و تصمیم‌های ماشین‌ها باشد، نه مسئولیت‌پذیری خود ماشین‌ها.

با این همه و در نهایت، اگرچه ماشین‌های هوشمند صلاحیت‌های حداقلی برای عاملیت اخلاقی تمام‌عیار را ندارند، اما می‌توان به معنایی مسامحی و محدود یا ابزارگرایانه، به‌عنوان عاملانی با پیامدهای اخلاقی در نظر گرفت. برای مثال، یک ماشین خودران که در یک تصادف احتمالی تصمیم می‌گیرد جان یک فرد را به‌جای دیگری حفظ کند، به‌ظاهر مانند یک عامل اخلاقی عمل می‌کند. اما چنین تصمیماتی صرفاً بر اساس الگوریتم‌های از پیش تعریف شده صورت می‌گیرند و نمی‌توان آن را معادل قضاوت اخلاقی یک عامل اخلاقی دانست. در حالی که ماشین‌های فراهوشمند می‌توانند در حوزه‌هایی نظیر تصمیم‌گیری بی‌طرفانه و عاری از تعصب عملکرد بهتری از انسان داشته باشند، عاملیت اخلاقی واقعی مستلزم ویژگی‌هایی است که هنوز به‌طور انحصاری در وجود انسان‌ها یافت می‌شود. بنابراین، تا زمانی که ماشین‌ها فاقد آگاهی، احساسات، و توانایی ارزیابی اخلاقی مستقل هستند، نمی‌توانند به‌عنوان عاملان اخلاقی تمام‌عیار شناخته شوند.

توجه به این نکته نیز مهم است که بحثهای ناظر به «عاملیت اخلاقی» همواره تحت الشعاع رویکرد و انگارهٔ فیزیکالیسم[1] و نفی دوگانگی روح وبدن بوده و بعید نیست که در چارچوب تفکری که تمامی

1. Physicalism

موجودیت‌ها و پدیده‌ها و از جمله ذهن، آگاهی، و اخلاق را چیزی جز فرایندهای فیزیکی در مغز نمی‌انگارد و در چارچوب مادی، توجیه و توضیح‌پذیر می‌داند، عاملیت اخلاقی و اراده آزادِ ماشین‌ها و ممکن و معنادار باشد و به فرایندهای فیزیکی و الگوریتمی تقلیل یابد، اما از زاویه نگاهی که نفس انسان را یک جوهر غیرمادی و مستقل از بدن به حساب می‌آورد، عاملیت اخلاقی مصنوعات بی‌روح و انتساب مسئولیت اخلاقی به ماشین‌ها بعید و ناممکن می‌نُماید.

تعارض منافع

نویسنده هیچ‌گونه تعارض منافعی گزارش نکرده است.

References

Bonnefon, J.-F., Rahwan, I., & Shariff, A. (2024). The Moral Psychology of Artificial Intelligence. *Annual Review of Psychology*, 75(1), 653–675. https://doi.org/10.1146/annurev-psych-030123-113559

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

Brożek, B., & Janik, B. (2019). Can artificial intelligences be moral agents? *New Ideas in Psychology*, 54, 101–106. https://doi.org/10.1016/j.newideapsych.2018.12.002

Dennett, D. C. (2014). When HAL kills, who's to blame? Computer ethics. In *Rethinking responsibility in science and technology / edited by Fiorella Battaglia, Nikil Mukerji, Julian Nida-Rümelin*. Pisa University Press. https://doi.org/10.1400/225034

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, 14(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Graff, J. (2024). Moral sensitivity and the limits of artificial moral agents. *Ethics and Information Technology*, 26(1), 13. https://doi.org/10.1007/s10676-024-09755-9

Gunkel, D. (2018). Can machines have rights? In *Living Machines: A Handbook of Research in Biomimetic and Biohybrid Systems* (pp. 596–601). https://doi.org/10.1093/oso/9780199674923.003.0063

Jaworska, A., & Tannenbaum, J. (2023a). The Grounds of Moral Status. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2023). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2023/entries/grounds-moral-status/

Jaworska, A., & Tannenbaum, J. (2023b). The Grounds of Moral Status. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2023). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2023/entries/grounds-moral-status/

Ladak, A. (2024). What would qualify an artificial intelligence for moral standing? *AI and Ethics*, 4(2), 213–228. https://doi.org/10.1007/s43681-023-00260-1

Manna, R., & Nath, R. (2021). The Problem of Moral Agency in Artificial Intelligence. *2021 IEEE Conference on Norbert Wiener in the 21st Century (21CW)*, 1–4. https://doi.org/10.1109/21CW48944.2021.9532549

Misselhorn, C. (2022a). Artificial Moral Agents: Conceptual Issues and Ethical Controversy. In S. Voeneky, P. Kellmeyer, O. Mueller, & W. Burgard (Eds.), *The Cambridge Handbook of Responsible Artificial Intelligence* (1st ed., pp. 31–49). Cambridge University Press. https://doi.org/10.1017/9781009207898.005

Moor, J. H. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, 21(4), 18–21. IEEE Intelligent Systems. https://doi.org/10.1109/MIS.2006.80

Müller, V. C. (2023). Ethics of Artificial Intelligence and Robotics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2023). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/

Schlosser, M. (2019). Agency. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/agency/

Searle, J. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3), 417–457. https://doi.org/10.1017/s0140525x00005756

Shapiro, P. (2006). Moral Agency in Other Animals. *Theoretical Medicine and Bioethics*, 27(4), 357–373. https://doi.org/10.1007/s11017-006-9010-0

Sullins, J. P. (2011). When Is a Robot a Moral Agent? In M. Anderson & S. L. Anderson (Eds.), *Machine Ethics* (pp. 151–161). Cambridge University Press. https://doi.org/10.1017/CBO9780511978036.013

Talbert, M. (2024). Moral Responsibility. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2024). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2024/entries/moral-responsibility/