Ethical Reflections

Vol. 7, Issue 2, No. 26, Summer 2026, pp. 27-47.

Online ISSN: 2717-1159 / Print ISSN: 2676-4810

https://jer.znu.ac.ir

Ethical Challenges of Artificial Intelligence and Social Choice Solutions for Responsible Design

Mir Hamid Almasi,[1] Mohammad Bagher Kazemi,[2] Mohammad Reza Ghaemi,[3] Fatemeh Raei [4]

Article info

DOI: 10.30470/er.2026.2065955.1450

Review Article

Submission History

Received: 15/11/2025

Revised: 22/04/2026

Accepted: 17/05/2026

Published: 23/07/2026

Authors' Contribution

Mir Hamid Almasi: Collecting data, translation, drafting of the manuscript, critical revision, reading and confirming the last version of manuscript.

Mohammad Bagher Kazemi: Collecting data, Study concept and design, drafting of the manuscript, critical revision, reading and confirming the last version of manuscript.

Mohammad Reza Ghaemi: Collecting data, translation, drafting of the manuscript, critical revision, reading and confirming the last version of manuscript.

Fatemeh Raei: Collecting data, translation, drafting of the manuscript, critical revision, reading and confirming the last version of manuscript.

Conflict of Interests

The authors declare no conflict of interest.

Extraction

This article is not extracted from thesis.

AI Usage

In this article, artificial intelligence tools were used to a limited extent and solely to assist with rephrasing sentences and making editorial corrections, as well as translating a few specialized terms for which no suitable Persian equivalents were available or for which existing translations were ambiguous.

Abstract

The rapid advancement of emerging technologies, particularly artificial intelligence (AI), has significantly challenged patterns of social interaction, human decision-making, and the ethical structures of societies. While AI demonstrates remarkable capabilities in solving complex problems, the absence of clear ethical guidelines may lead to social conflicts, ethical dilemmas, and new forms of inequality. This study adopts a descriptive–analytical approach and draws on William Ogburn's theory of cultural lag to examine the gap between technological development and the slower evolution of cultural and ethical responses. It emphasizes the necessity of addressing ethical considerations during the design and implementation of AI systems in order to ensure their responsible and socially aligned use. In this context, the paper examines two primary approaches to integrating ethics into AI. The first is the rule-based or top-down approach, which relies on predefined ethical principles and formal rules to guide AI decision-making. The second is the data-driven or bottom-up approach, which seeks to derive ethical norms and values through the analysis of human data, behaviors, and preferences. While both approaches provide valuable insights, each has limitations, particularly regarding adaptability, inclusiveness, and the risk of reflecting narrow or biased perspectives. Focusing on the concept of social choice ethics, this paper explores how collective ethical perspectives can be incorporated into the design of AI systems. It argues that participatory mechanisms, such as social choice frameworks, can help reduce the imposition of limited designer viewpoints and improve ethical representation. The study highlights three key components of this approach: the inclusion of stakeholders' positions, the measurement of ethical preferences, and the aggregation of diverse viewpoints into collective ethical outcomes. Ultimately, the paper suggests that integrating social choice mechanisms into AI design can enhance fairness, transparency, and ethical legitimacy, contributing to the development of more responsible and socially aligned artificial intelligence systems.

Keywords: Artificial Intelligence, Ethical Principles, Cultural Lag, Social Choice Ethics

1. Ph.D. student, Department of Mathematics, University of Zanjan, Zanjan, Iran, h.almasi@znu.ac.ir

2. **Corresponding author**, Associate professor, Department of Mathematics and Computer Science, University of Zanjan, Zanjan, Iran, mbkazemi@znu.ac.ir

3. Assistant professor, Department of Mathematics, University of Zanjan, Zanjan, Iran, ghaemi@znu.ac.ir

4. Assistant professor, Department of Mathematics Education, Farhangian University, Tehran, Iran, f.raei@cfu.ac.ir

Introduction

Recent advances in science and technology, particularly in artificial intelligence (AI), have profoundly transformed social interactions, human decision-making, and ethical structures (Grace et al., 2018, p. 730; Collins & Varmus, 2015, p. 795). While AI offers unprecedented capabilities for problem-solving and efficiency, the absence of comprehensive ethical guidelines can lead to social conflicts, inequality, and moral dilemmas (Floridi, 2018, p. 319; Müller, 2020, p. 4). Ethical considerations should be integrated from the early stages of AI design rather than addressed after deployment. Failure to do so may result in a "cultural lag," in which technological innovation outpaces the evolution of social, legal, and ethical norms (Ogburn, 1966, p. 17; Robinson, 1981, p. 19; Marshall, 1999, p. 82).

AI applications across domains such as smart homes (Harper, 2006, p. 5), smart agriculture (Walter et al., 2017, p. 6149), healthcare (Hossain et al., 2020, p. 127), and autonomous vehicles (Bonnefon et al., 2016, p. 1573) highlight both transformative potential and ethical risks, including accountability, privacy, and algorithmic autonomy (Allen et al., 2005, p. 568; Wallach et al., 2008, p. 570).

This study adopts a theoretical–analytical approach, grounded in Ogburn's cultural lag thesis, to examine ethical challenges in AI. It analyzes two primary approaches: the rule-based approach, based on predefined ethical principles, and the data-driven approach, based on learning from human behavior. Furthermore, the study introduces social choice ethics as a participatory framework for aligning AI systems with human values through standing, measurement, and aggregation (Baum, 2020, p. 2).

Concept of Ethics and Ethical Approaches

Ethics serves as a normative framework guiding human behavior and decision-making within social and technological contexts. In the field of artificial intelligence, ethical considerations extend beyond individual actions to include the design, deployment, and societal impact of intelligent systems (Allen et al., 2005, p. 568; Wallach et al., 2008, p. 570). Ethical frameworks provide essential guidance to ensure AI technologies operate in ways consistent with social values and human welfare. Traditionally, two dominant approaches have been applied to embed ethics in AI: the rule-based approach, which relies on predefined principles and classical ethical theories, and the data-driven approach, which derives ethical understanding from observation of human behavior and learning from societal norms (Floridi, 2018, p. 319; Müller, 2020, p. 4).

Concept of Technology and Technology Ethics

Technological systems are not neutral tools; they inherently reflect the values and assumptions embedded during their development (Grace et al., 2018, p. 730). AI, in particular, has significant societal influence, affecting decision-making in domains such as smart homes (Harper, 2006, p. 5), smart agriculture (Walter et al., 2017, p. 6149), healthcare systems (Hossain et al., 2020, p. 127), and

autonomous vehicles (Bonnefon et al., 2016, p. 1573). Consequently, ethical frameworks must consider both the technical functioning of AI systems and the social and moral consequences of their deployment (Allen et al., 2005, p. 568; Wallach et al., 2008, p. 570).

William Ogburn's Cultural Lag Thesis

Ogburn's concept of cultural lag describes the temporal gap between rapid technological development (material culture) and the slower evolution of social, legal, and ethical norms (Ogburn, 1966, p. 17). In AI, this lag creates situations where technological capabilities advance faster than societal mechanisms for accountability, fairness, and ethical governance (Robinson, 1981, p. 19; Marshall, 1999, p. 82). Addressing this gap requires integrating ethical considerations from the earliest stages of AI design to prevent potential social harm and ethical conflicts (Floridi, 2018, p. 319).

Ethical Challenges in Artificial Intelligence

The deployment of AI systems has raised numerous ethical challenges, including accountability, privacy, autonomy, and vulnerability to errors or bias (Allen et al., 2005, p. 568; Wallach et al., 2008, p. 570). Rule-based approaches attempt to encode ethical principles directly into AI algorithms, ensuring predictable and consistent behavior (Wallach et al., 2008, p. 570). However, they are limited in handling complex, context-dependent moral dilemmas that arise in real-world applications (Allen et al., 2005, p. 568). Data-driven approaches, which learn ethical norms from human behavior, provide adaptability but risk inheriting social biases or reinforcing existing inequalities (Batavia et al., 1996, p. 5).

Ethical Approaches in AI Design

The rule-based or top-down approach relies on formalized ethical principles embedded within AI systems. It is particularly useful for applications requiring strict adherence to predefined rules, such as regulatory compliance or high-stakes decision-making contexts (Wallach et al., 2008, p. 570). Nevertheless, its rigidity limits its applicability in dynamic environments where moral reasoning must adapt to contextual variations (Allen et al., 2005, p. 568).

The data-driven or bottom-up approach derives ethical behavior through machine learning models trained on human decision-making data (Batavia et al., 1996, p. 5). This approach allows AI systems to adapt to complex and evolving social norms. However, reliance on historical human data introduces the risk of bias, ethical inconsistency, and potential reinforcement of discriminatory patterns (Floridi, 2018, p. 319).

While rule-based systems offer clarity and reliability, and data-driven systems provide adaptability, neither approach alone can fully address the ethical challenges of AI. A combined or complementary framework is necessary to integrate human values effectively and to mitigate risks associated with ethical lapses or social bias (Müller, 2020, p. 4).

Social Choice Ethics as an Alternative Framework

Social choice ethics provides a participatory and collective approach to embedding ethics in AI. It emphasizes aggregating diverse ethical perspectives into a unified decision-making process, allowing AI systems to reflect societal values rather than the preferences of individual designers (Baum, 2020, p. 2).

Standing, Measurement, and Aggregation

The framework of social choice ethics is operationalized through three key mechanisms: standing, which determines whose ethical perspectives are considered; measurement, which identifies and quantifies ethical preferences; and aggregation, which combines diverse viewpoints into collective ethical guidance for AI behavior (Baum, 2020, p. 2). Incorporating these mechanisms can reduce bias, enhance transparency, and ensure broader societal alignment.

Challenges and Opportunities

Although social choice ethics offers a promising framework, practical implementation requires careful stakeholder selection, reliable measurement of ethical preferences, and robust aggregation techniques. Its successful integration depends on interdisciplinary collaboration among technologists, ethicists, policymakers, and social scientists, ensuring that AI systems are both technically effective and ethically responsible (Boualheri et al., 2022, p. 35; Jafari & Rahmati, 2025, p. 6).

Conclusion

Artificial intelligence presents unprecedented opportunities for human advancement but simultaneously raises significant ethical challenges, including bias, accountability, and privacy concerns. Traditional rule-based and data-driven approaches, while valuable, are limited in addressing complex, context-dependent moral dilemmas. This study demonstrates that social choice ethics offers a promising framework for responsible AI design by incorporating standing, measurement, and aggregation of diverse ethical perspectives. By integrating participatory and socially grounded mechanisms, AI systems can better align with societal values, enhance transparency, and reduce designer bias. Ultimately, embedding social choice principles strengthens both ethical legitimacy and societal trust in AI technologies.

References

Allen, C., Smit, I., & Wallach, W. (2005). "Artificial morality: Top-down, bottom-up, and hybrid approaches". Ethics and Information Technology, 7(3): 149–155.

Baum, D. (2020). "Social Choice Ethics in Artificial Intelligence". AI & Society, 35(1): 165–176. https://doi.org/10.1007/s00146-017-0760-1.

Batavia, H., Parag, H., Dean, A., Pomerleau, E., Charles, E., & Thorpe, E. (1996). "Applying Advanced Learning Algorithms To ALVINN". Carnegie Mellon University, The Robotics Institute.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). "The Social Dilemma Of Autonomous Vehicles". Science, 352(6): 1573–1576.

Bostrom, N. (2014). "Superintelligence: Paths, Dangers, Strategies". Oxford, UK: Oxford University Press.

Boualheri, A., Radfar, R., Alborzi, M., PourEbrahimi, A., & Dehghani, M. (1401 SH). "Explaining Ethical Decision-Making in the Context of Information Technology Management". Ta'amolat-e Akhlaqi, 3(2): 29–46.

Catch, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). "Artificial Intelligence and the 'Good Society': The US, EU, And UK Approach". Science And Engineering Ethics, 24(2): 505–528.

Collins, F. S., & Varmus, H. (2015). "A New Initiative on Precision Medicine". New England Journal Of Medicine, 372(9): 793–795.

Cordeiro, W. P. (1997). "Suggested Management Responses to Ethical Issues Raised By Technological Change". Journal Of Business Ethics, 16(12–13): 1393–1400.

Dignum, V. (2018). "Ethics In Artificial Intelligence: Introduction to The Special Issue". Ethics And Information Technology, 20(1): 1–3. https://doi.org/10.1007/s10676-018-9450-z.

Floridi, L. (2018). "Artificial Intelligence, Deepfakes and A Future of Ectypes". Philosophy & Technology, 31(3): 317–321.

Gewirth, A. (1978). "Reason And Morality". Chicago, IL: University Of Chicago Press.

Ginges, J., Atran, S., Medin, D., & Shikaki, K. (2007). "Sacred Bounds on Rational Resolution of Violent Political Conflict". Proceedings Of the National Academy of Sciences, 104(18): 7357–7360.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). "When Will AI Exceed Human Performance? Evidence From AI Experts". Journal Of Artificial Intelligence Research, 62: 729–754.

Harper, R. (2006). "Inside The Smart Home". Springer Science & Business Media.

Hossain, M. S., Muhammad, G., & Guizani, N. (2020). "Explainable AI And Mass Surveillance System-Based Healthcare Framework to Combat COVID-19 Like Pandemics". IEEE Network, 34(4): 126–132.

Hoedemaekers, R. (2001). "Commercialization, Patents and Moral Assessment Of Biotechnology Products". Medicine, Health Care and Philosophy, 2(3): 273–284.

Hubbard, F. P. (2011). "Do Androids Dream? Personhood And Intelligent Artifacts". Temple Law Review, 83: 405–441.

Izanzloo, H., Masoudi, J., & Peyk Herfeh, S. (1402 SH). "The Level of Ethical Demand with Emphasis on the Priority Component in Islamic Thought". Ta'amolat-e Akhlaqi, 4(4): 109–134.

Jafari, R., & Rahmati, H. (1404 SH). "Application of Artificial Intelligence in Ethical Counseling: From Transformative Capacities to Ethical Challenges". Faslnameh-ye Akhlaq-Pazhuhi, 8(2): 49–72.

Lin, P. (2016). "Why Ethics Matters For Autonomous Cars". In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), Autonomous Driving: Technical, Legal And Social Aspects (pp. 69–85). Berlin: Springer.

Martin, D. (2017). "Who Should Decide How Machines Make Morally Laden Decisions?". Science And Engineering Ethics, 23(4): 951–967.

Marshall, P. (1999). "Has Technology Introduced New Ethical Problems?". Journal Of Business Ethics, 19: 81–90.

MirzaMohammadi, A., & Taghavi, M. (1404 SH). Ethics of Technology: "A Critical Evaluation of the Debate between David Morrow and Joseph C. Pitt on the Value-Ladenness of Technology and Ethical Responsibility". Ta'amolat-e Akhlaqi, 6(2): 93–114.

Muehlhauser, L., & Helm, L. (2012). "Intelligence Explosion and Machine Ethics". In A. Eden, J. Søraker, J. H. Moor, & E. Steinhart (Eds.), Singularity Hypotheses: A Scientific And Philosophical Assessment (pp. 101–126). Springer.

Müller, V. C. (2020). "Ethics Of Artificial Intelligence and Robotics". Stanford University.

Ogburn, W. F. (1966). "Social Change". New York.

Robinson, J. P. (1981). "Will The New Electronic Media Revolutionize Our Daily Lives?" In R. W. Haigh, G. Gerbner, & R. B. Byrne (Eds.), Communications In The Twenty-First Century. New York, NY: John Wiley & Sons.

Verbeek, P.-P. (2006). "Materializing Morality: Design Ethics and Technological Mediation". Philosophy & Technology, 19(3): 361–380.

Walter, R., Finger, R., Huber, S., & Buchmann, N. (2017). "Smart Farming Is Key To Developing Sustainable Agriculture". Proceedings of the National Academy of Sciences, 114(24): 6148–6150.

Wallach, W., & Allen, C. (2008). "Moral Machines: Teaching Robots Right From Wrong". Oxford University Press.

Wallach, W., Allen, C., & Smit, I. (2008). "Machine Morality: Bottom-Up and Top-Down Approaches for Modelling Human Moral Faculties". AI & Society, 22(4): 565–582.

Yampolskiy, R. V. (2013). "Artificial Intelligence Safety Engineering: Why Machine Ethics Is A Wrong Approach". In V. C. Müller (Ed.), Philosophy and Theory of Artificial Intelligence (pp. 389–396). Springer.

Yudkowsky, E. (2004). "Coherent Extrapolated Volition". The Singularity Institute, San Francisco.

تأملات اخلاقی

دورهٔ هفتم، شمارهٔ دوم (پیاپی۲۶)، تابستان ۱۴۰۵، صفحات ۲۷-۴۷

شاپا الکترونیکی: ۱۱۵۹-۲۷۱۷/ شاپا چاپی: ۴۸۱۰-۲۶۷۶

https://jer.znu.ac.ir

چالش‌های اخلاقی هوش مصنوعی و راهکارهای انتخاب اجتماعی برای طراحی مسئولانه

میرحمید الماسی،[۱] محمدباقر کاظمی،[۲] محمدرضا قائمی،[۳] فاطمه راعی[۴]

اطلاعات مقاله

DOI:
10.30470/er.2026.2065955.1450

مقاله مروری

تاریخ دریافت: ۱۴۰۴/۰۸/۲۴

تاریخ اصلاح: ۱۴۰۵/۰۲/۰۲

تاریخ پذیرش: ۱۴۰۵/۰۲/۲۷

تاریخ انتشار: ۱۴۰۵/۰۵/۰۱

نقش نویسندگان

میرحمید الماسی: جمع آوری داده ها، ترجمه، نگارش اولیه و نهایی مقاله، تحلیل و نقادی.

محمدباقر کاظمی: ایده پردازی، نگارش اولیه و نهایی، تحلیل و نقادی، انجام اصلاحات مقاله.

محمدرضا قائمی: جمع آوری داده‌ها، ترجمه، نگارش اولیه و نهایی، تحلیل و نقادی، انجام اصلاحات مقاله.

فاطمه راعی: جمع آوری داده‌ها، ترجمه، نگارش اولیه و نهایی، تحلیل و نقادی، انجام اصلاحات مقاله.

حمایت مالی و تعارض منافع

مقاله حامی مالی و تعارض منافع ندارد.

استخراج

این مقاله متخذ از رساله / پایان نامه نیست.

استفاده از هوش مصنوعی

در این مقاله، از ابزارهای هوش مصنوعی به صورت محدود و تنها برای کمک در بازنویسی جملات یا اصلاحات نگارشی استفاده شده است و نیز در ترجمه چند کلمه تخصصی که معادل فارسی متناسب نداشته و یا ابهامی در ترجمه های موجود بود.

چکیده

پیشرفت سریع فناوری‌های نوین، به‌ویژه هوش مصنوعی، الگوهای تعامل اجتماعی، تصمیم‌گیری‌های انسانی و ساختارهای اخلاقی جوامع را به چالش کشیده است. در حالی که قابلیت‌های هوش مصنوعی در حل مسائل پیچیده، چشمگیر است، نبود دستورالعمل‌های اخلاقی روشن می‌تواند منجر به بروز تعارض‌های اجتماعی و نابرابری‌های جدید شود. مقاله‌ی حاضربا روش تحلیلی-توصیفی، با بهره‌گیری از «نظریه تأخر فرهنگی» ویلیام اوگبرن، به بررسی شکاف بین توسعه فناوری‌های نو و واکنش‌های فرهنگی-اخلاقی می‌پردازد و بر لزوم پرداختن به ابعاد اخلاقی در طراحی و استفاده از هوش مصنوعی تأکید می‌کند. در این راستا، دو رویکرد اصلی در پیاده‌سازی اخلاق در هوش مصنوعی بررسی می‌شود: رویکردهای قاعده‌محور که بر اصول و قواعد اخلاقی از پیش‌تعریف‌شده تکیه دارند و رویکردهای داده‌محور که بر استخراج و یادگیری تدریجی ارزش‌ها و هنجارهای اخلاقی از داده‌ها و ترجیحات انسانی استوارند. مقاله با تمرکز بر مفهوم «اخلاق انتخاب اجتماعی» به بررسی چگونگی درگیرکردن دیدگاه‌های اخلاقی جامعه در طراحی هوش مصنوعی می‌پردازد. در نهایت، این مقاله پیشنهاد می‌کند که برای کاهش تحمیل دیدگاه‌های محدود طراحان، از مکانیزم‌های مشارکتی همچون انتخاب اجتماعی استفاده شده و سه مؤلفه کلیدی آن یعنی جایگاه، اندازه‌گیری و تجمیع دیدگاه‌ها در فرآیند طراحی، لحاظ گردد.

واژه‌های کلیدی: هوش مصنوعی، اصول اخلاقی، تأخر فرهنگی، انتخاب اجتماعی.

۱. دانشجوی دکتری، گروه ریاضی وعلوم کامپیوتر، دانشکده علوم، دانشگاه زنجان، زنجان، ایران. h.almasi@znu.ac.ir

۲. **نویسنده مسئول،** دانشیار گروه ریاضی وعلوم کامپیوتر، دانشکده علوم، دانشگاه زنجان، زنجان، ایران. mbkazemi@znu.ac.ir

۳. استادیار گروه ریاضی وعلوم کامپیوتر، دانشکده علوم، دانشگاه زنجان، زنجان، ایران. ghaemi@znu.ac.ir

٤. استادیار گروه آموزش ریاضی، دانشگاه فرهنگیان، تهران، ایران. f.raei@cfu.ac.ir

مقدمه

در دهه‌های اخیر، پیشرفت‌های پرشتاب فناوری و علوم، به‌ویژه در زمینه‌ی هوش مصنوعی، زندگی بشر را از ابعاد مختلف دستخوش تحول کرده و بنیان‌های تعاملات اجتماعی، تصمیم‌گیری‌های انسانی و ساختارهای اخلاقی را به چالش کشیده است (Grace et-al., 2018, p. 730; Collins & Varmus, 2015, p.795). در حالی‌که این فناوری‌ها امکانات بی‌سابقه‌ای برای بهبود کیفیت زندگی فراهم آورده‌اند و نبود دستورالعمل‌های اخلاقی منسجم می‌تواند به نابرابری، تضاد اجتماعی و حتی بحران‌های هویتی منجر شود (Floridi, 2018, p. 319; Müller, 2020, p. 4).

مباحث اخلاقی در حوزه فناوری نباید تنها پس از گسترش کاربردهای آن مطرح شوند، بلکه لازم است از مراحل اولیه طراحی در دستور کار قرار گیرند (میری بالاجورشری و دیگران، ۱۴۰۳، ص۳). عدم توجه به این موضوع، منجر به «تأخر فرهنگی[1]» می‌شود؛ مفهومی که اوگبرن (Ogburn, 1966, p. 17) برای توصیف شکاف بین توسعه فرهنگ مادی (تکنولوژی) و فرهنگ غیرمادی (اخلاق، ارزش‌ها و قانون) به‌کار برد. در این وضعیت، فناوری به سرعت رشد می‌کند، اما ساختارهای اجتماعی و اخلاقی برای پاسخگویی به آن با تأخر واکنش نشان می‌دهند (Robinson, 1981, p. 19; Marshall, 1999, p. 82).

هوش مصنوعی با کاربردهایی در حوزه‌هایی نظیر خانه‌های هوشمند (Harper, 2006, p.5)، کشاورزی هوشمند (Walter et-al., 2017, p. 6149)، مراقبت‌های سلامت (Hossain et-al., 2020, p. 127) و وسایل نقلیه خودران (Bonnefon et-al., 2016, p. 1573) توجه عمومی و دانشگاهی را به خود جلب کرده است؛ اما همان‌قدر که قابلیت‌های آن رو به رشد است، دغدغه‌های اخلاقی نیز افزایش یافته‌اند، از جمله مسئولیت‌پذیری، حریم خصوصی، خودمختاری الگوریتمی و آسیب‌پذیری اجتماعی در برابر خطا یا سوگیری سیستم‌ها (Allen et-al., 2005, p. 568; Wallach et-al., 2008, p. 570).

در این مقاله، با استفاده از روش تحقیق نظری-تحلیلی و بر پایه مطالعه کتابخانه‌ای و اسنادی، تلاش شده است ضمن تبیین مفاهیم بنیادین اخلاق در فناوری و هوش مصنوعی، چالش‌های ناشی از تأخر فرهنگی در واکنش به پیشرفت‌های فناورانه مورد بررسی قرار گیرد. چارچوب نظری مقاله بر پایه «تز تأخر فرهنگی» ویلیام اوگبرن و دو رویکرد اصلی در اخلاق هوش مصنوعی رویکرد قاعده محور که مبتنی بر اصول از پیش تعریف‌شده و نظریه‌های کلاسیک اخلاقی است، و رویکرد داده محور که از طریق یادگیری تدریجی و مشاهده رفتار انسانی به فهم اخلاقی می‌رسد، بنا شده است. سپس، با تمرکز بر مفهوم نسبتاً نوظهور «اخلاق انتخاب اجتماعی[2]» (Baum, 2020, p.2)، سازوکارهای مشارکتی همچون جایگاه، اندازه‌گیری و تجمیع دیدگاه‌های اخلاقی تحلیل و به‌عنوان الگویی جایگزین برای تقویت تعامل میان ارزش‌های انسانی و الگوریتم‌ها در طراحی هوش مصنوعی پیشنهاد شده است.

1. Cultural lag
2. Social choice ethics

۱. چارچوب نظری و مبانی مفهومی

۱-۱. مفهوم اخلاق و رویکردهای آن

اخلاق به تصمیم گیری عملی و رفتار انسانی در وسیع‌ترین زمینه مربوط می‌شود و برترین علم اجتماعی است که جامعه‌شناسی، اقتصاد، روان‌شناسی، انسان‌شناسی و تاریخ را دربرمی‌گیرد. اخلاق مستلزم آن است که آن عملی را انجام دهید که می‌تواند معقول باشد و انتظار می‌رود به طور کلی به بهترین پیامدها منجر شود (ایزانلو و دیگران، ۱۴۰۲، ص ۱۱۲).

در فلسفه، اخلاق به توصیف آنچه برای فرد و جامعه خوب است و همچنین ماهیت وظایفی که افراد بر عهده خود و یکدیگر دارند، می‌پردازد؛ در حالی که معضل اخلاقی به معضلات اخلاقی خاصی اشاره دارد که می‌تواند بسیار پیچیده باشد. چالش‌هایی که آن‌ها به همراه دارند را نمی‌توان به راحتی حل کرد. فناوری‌های در حال بهبود مزایای متعددی را برای جامعه بشری به همراه دارند، اما ممکن است خطرات و چالش‌های منفی از جمله معضل اخلاقی پیچیده‌تر را نیز ایجاد کنند که در مورد فناوری‌های هوش مصنوعی صادق است.

دیدگاه‌های متفاوتی نسبت به اخلاق وجود دارد. طبق نظر علمای اخلاق که می‌توان آن‌ها را پدیدارشناس نامید، آنچه که خوب است، در موقعیت به دست می‌آید که برگرفته از منطق و زبان موقعیت است یا از گفت‌وگو و بحث درباره «خوبی» فی‌نفسه به دست می‌آید. این را رویکرد «قاعده محو»ر یا «اخلاق از بالا به پایین»[1] گویند. از سوی دیگر، پوزیتیویست‌ها معتقدند که ما باید دنیای واقعی را مشاهده کنیم و اصول اخلاقی را به صورت استقرایی استخراج کنیم که رویکرد «داده محور» یا «اخلاق از پایین به بالا»[2] است. در ادامه بیشتر به این امر پرداخته می‌شود.

۱-۲. مفهوم فناوری و اخلاق فناوری

منظور از «فناوری» به‌کارگیری نظام‌مند و هدفمند دانش، نظریه‌ها و مدل‌های شناختی انسان برای مداخله در جهان طبیعی و دستیابی به اهداف عملی است (Marshall, 1999, p. 82). در فلسفه معاصر، «فناوری» صرفاً به‌مثابه مجموعه‌ای از ابزارها یا مصنوعات خنثی تلقی نمی‌شود، بلکه به‌عنوان عاملی میانجی در نظر گرفته می‌شود که نحوه ادراک، کنش و تصمیم‌گیری انسان را در مواجهه با جهان شکل می‌دهد. از این منظر، «فناوری» در تعامل میان انسان و واقعیت دخالت می‌کند و از طریق طراحی و کاربرد خود، به‌طور ناگزیر حامل ارزش‌ها و ملاحظات هنجاری است. بنابراین، فناوری نه‌تنها محصول دانش فنی، بلکه بخشی از ساختارهای اجتماعی و اخلاقی است که شیوه زیست انسانی را جهت‌دهی می‌کند (Verbeek, 2006, p. 363).

اخلاق فناوری به مجموعه اصول، ارزش‌ها و قوانینی اشاره دارد که در طراحی، توسعه و استفاده از فناوری‌های مختلف برای اطمینان از استفاده اخلاقی و مسئولانه از آن‌ها به کار می‌رود. این اصول و ارزش‌ها برای حفظ حریم خصوصی، امنیت و عدالت در دنیای دیجیتال اهمیت دارند و برای جلوگیری از سوءاستفاده و مشکلات اخلاقی احتمالی در استفاده از فناوری‌ها مورد استفاده قرار می‌گیرند. زمینه‌های

1. Top-down ethics
2. Bottom-up ethics

نگرانی اخلاقی زیادی در مورد فناوری‌های جدید وجود دارد که مسائل مربوط به سازگاری فرهنگی را که توسط فناوری‌های پیشرفته امروزی مطرح می‌شود، نشان می‌دهد (احمدی و دیگران، ۱۴۰۵،ص ۶). فناوری‌های جدید، جهان‌بینی ما، الگوهای تعامل اجتماعی و روابط ما را با یکدیگر تغییر می‌دهند. چه چالش از کامپیوتر و الکترونیک، ارتباطات و هوش مصنوعی باشد، چه از مهندسی بیوژنتیک یا از برخی زمینه‌های دیگر، پیشرفت‌های تکنولوژیکی توانایی‌های ما را به‌عنوان انسان افزایش می‌دهد؛ اما اگر دستورالعمل‌هایی برای استفاده مناسب مورد توافق قرار نگیرد، پیشرفت‌های تکنولوژیکی خطر درگیری اجتماعی را افزایش می‌دهد. سرعت فزاینده تغییرات تکنولوژیک امروزی، شکاف روزافزون بین فناوری‌های جدید و دستورالعمل‌های اخلاقی پذیرفته‌شده برای استفاده از آن‌ها را افزایش می‌دهد. یک اجماع اجتماعی باید ایجاد شود تا دیدگاه اخلاقی، تأثیر گسترده و عملی در جهان اجتماعی داشته باشد. مکانیسم‌های امروزی برای توسعه گفت‌وگوی منطقی که منجر به اجماع اجتماعی شود، آهسته است و اغلب تحت‌الشعاع رسانه‌های عمومی قرار می‌گیرد. کاوش پیامدهای اخلاقی فناوری‌های جدید بخشی ضروری از فرآیندهای تغییر اجتماعی و سازگاری است. نحوه موفقیت‌آمیز بودن دستورالعمل‌های اخلاقی و ایجاد اجماع اجتماعی بر انسجام و همبستگی اجتماعی تأثیر می‌گذارد (Gerwith, 1978, p. 103). زمانی که نوآوری تکنولوژیک مستلزم یک تغییر فرهنگی بزرگ باشد، به دلیل گسست کیفی از تفکرات و شیوه‌های زندگی گذشته یا به دلیل انتشار سریع در طیف وسیعی از فعالیت‌های انسانی، تأمل زیادی باید صورت گیرد (Robinson, 1981, p. 19). اصول قدیمی انصاف و صداقت، احترام به حقوق دیگران، احترام به محیط‌زیست، و احترام به زندگی و کیفیت آن در مورد فناوری‌های جدید صدق می‌کند، اما برای روشن شدن مسائل پیچیده ناشی از پیشرفت‌های فناوری، به تجربه و بحث فکری صادقانه نیاز است (Cordeiro, 1997, p. 34). در مناقشه فلسفی درباره ماهیت ارزشی فناوری، دیدگاهی که فناوری را ذاتاً خنثی می‌داند و منشأ ارزش‌های اخلاقی را در کاربرد انسانی جستجو می‌کند، از پشتوانه نظری قوی‌تری برخوردار است (میرزامحمدی و دیگران، ۱۴۰۴). بر اساس این دیدگاه، مسئولیت اخلاقی نه در ذات ابزارهای فناورانه، بلکه در فرآیندهای اجتماعی انتخاب، نظارت و استفاده از آن‌ها معنا می‌یابد. این نگاه، اهمیت سازوکارهای مشارکتی مانند انتخاب اجتماعی را در طراحی اخلاقی هوش مصنوعی برجسته می‌سازد.

۱-۳. نظریه تأخر فرهنگی ویلیام اوگبرن

اوگبرن در کتاب *تغییر اجتماعی*[1] (1966) اصطلاح «تأخر فرهنگی» را برای به تصویر کشیدن پدیده‌ای که مشاهده کرد در نیمه اول قرن بیستم رخ می‌دهد و اکنون به نظر می‌رسد که به شدت سرعت گرفته است، ابداع کرد (Ogburn, 1966, p. 17). منظور اوگبرن از تأخر فرهنگی این بود که فرهنگ مادی با سرعت بیشتری نسبت به فرهنگ غیرمادی پیشرفت می‌کند. تأخر فرهنگی دوره‌ای است بین اختراع یک فناوری و زمانی که جامعه قضاوت اخلاقی و اجتماعی خود را نسبت به آن فناوری شکل داده است. تجهیزات فیزیکی و مراحل تولید و استفاده از آن بخشی از فرهنگ مادی است. دین، اخلاق، فلسفه، نظام‌های اعتقادی، ارزش‌ها و قانون نمونه‌هایی از فرهنگ غیرمادی هستند. تأخر فرهنگی دوره زمانی بین اختراع یک فناوری و گسترش و تثبیت تأثیر آن بر ابزارها، زیرساخت‌ها و شیوه‌های مادی جامعه است. در این مدت، ممکن است شکافی بین تکنولوژی و اخلاق وجود داشته باشد. از منظر ساختاری

1. Social Change

اجتماعی، دلایل متعددی وجود دارد که سیستم‌های اخلاقی از توسعه فناوری عقب هستند. اولین مورد این است که هزینه و پیچیدگی توسعه فناوری امروزه توسعه دهندگان فناوری را به ساختارهای شرکتی بسیار متمرکز و کنترل شده می‌کشاند. تمرکز فکری حاصل از تجهیزات، منابع و اطلاعات و کنترل اجتماعی بر توسعه‌دهندگان و فرآیند توسعه، کارایی تحقیق و توسعه یک‌نگر را فراهم می‌کند. دوم، فناوری‌های جدید امکان بازده اقتصادی بزرگ را نوید می‌دهند، اما در یک محیط بسیار رقابتی. توسعه دهندگان فناوری ابتدا برای ثبت اختراع و ارائه محصولات به بازار رقابت می‌کنند. اولین بودن در بازار تضمینی برای تسلط بر بازار نیست، اما می‌تواند یک مزیت رقابتی باشد. سوم، توسعه فناوری فرهنگ مادی به دنبال کشف و به کارگیری قوانین طبیعی جهان فیزیکی است (Marshall, 1999, p. 84).

برخلاف فرآیند توسعه فرهنگ مادی، توسعه سیستم‌های اخلاقی برای مدیریت کاربردهای فرهنگ مادی به دلایل متعدد کندتر است. اول، توسعه دستورالعمل‌های اخلاقی در یک محیط متمرکز و کنترل شده صورت نمی‌گیرد. دوم، ساختار بازار رقابتی وجود ندارد که از نظر مالی به معرفی دیدگاه اخلاقی غالب پاداش دهد. سوم، نیروهای اجتماعی که یک نظام اخلاقی می‌خواهد بر آنها تأثیر بگذارد، به اندازه جنبه‌های فیزیکی جهان قابل کنترل و مدیریت نیستند. علاوه بر این، روند توسعه اجماع اجتماعی گسترده پیرامون دستورالعمل‌های اخلاقی به طور کلی باید در انتظار معرفی و انتشار یک فناوری جدید باشد. در حالی که اخلاق‌گرایان آینده‌نگر ممکن است روندهای فناوری جدید را طرح کنند و سؤالاتی را در مورد کاربردهای اخلاقی فناوری‌های پیش‌بینی شده در آینده مطرح کنند، اما تا زمانی که یک فناوری به حدی در جامعه گسترش نیابد که به موضوعی برای بحث و مناقشه عمومی تبدیل شود، معمولاً ابعاد اخلاقی آن مورد توجه گسترده جامعه قرار نمی‌گیرد (Marshall, 1999, p. 84).

سؤالات اخلاقی مطرح شده توسط فناوری مدرن بیش از همه در زمینه‌های هوش مصنوعی، پزشکی و بیوتکنولوژی مورد بحث قرار گرفته است. ارزیابی فناوری سلامت در درجه اول به جای ارزیابی روش‌های جایگزین درمانی از نظر کارایی، عوارض جانبی و هزینه‌ها، بلکه در مورد پیامدهای اخلاقی مربوط می‌شود. لذا باید روشی ایجاد شود که این تأخر را به حداقل برساند.

۱-۴. چالش ها و مسائل اخلاقی در هوش مصنوعی

با گسترش روزافزون هوش مصنوعی و نفوذ آن در زیرساخت‌های حیاتی، ضرورت توجه به پیامدهای اخلاقی آن بیش از پیش احساس می‌شود. سیستم‌های هوشمند نه‌تنها تصمیم‌گیری‌های انسانی را تسهیل می‌کنند، بلکه در بسیاری از موارد جایگزین آن می‌شوند؛ بنابراین پرسش از «اخلاق‌پذیری» این فناوری‌ها حیاتی است.

یکی از چالش‌های برجسته، موقعیت‌هایی است که سامانه‌های هوش مصنوعی باید در آنها تصمیم‌های اخلاقی پیچیده اتخاذ کنند؛ برای مثال، خودروهای خودران ممکن است در موقعیت‌های تصادف اجتناب‌ناپذیر قرار گیرند، و این پرسش مطرح می‌شود که اولویت نجات با کدام فرد یا گروه است. پاسخ به این سؤال‌ها مستلزم تصمیم‌های برنامه‌نویسان درباره چارچوب‌های اخلاقی است که در الگوریتم‌ها گنجانده می‌شود (Lin, 2016, p. 2).

افزون بر این، ابعاد گسترده‌تری از اخلاق در حوزه‌هایی مانند: حریم خصوصی، تبعیض الگوریتمی، شفافیت تصمیم‌ها و

پاسخ‌گویی حقوقی مطرح است. توانمندی سیستم‌های هوش مصنوعی در تحلیل کلان‌داده‌ها، به نگرانی‌هایی درباره نظارت فراگیر، سوءاستفاده از اطلاعات شخصی و تثبیت سوگیری‌های نژادی یا جنسیتی منجر شده است.

در واکنش به این مسائل، نهادهای بین‌المللی و شرکت‌های بزرگ فناوری در سال‌های اخیر تلاش‌هایی برای تدوین اصول راهنمای اخلاقی انجام داده‌اند. از جمله، یونسکو در سال ۲۰۲۰ پیش‌نویس نخست «توصیه‌نامه جهانی اخلاق در هوش مصنوعی» را منتشر کرد. همچنین شرکت‌هایی چون گوگل، مایکروسافت و آمازون چارچوب‌هایی اخلاق‌محور برای طراحی محصولات هوشمند پیشنهاد داده‌اند (Catch et-al., 2018, p. 506).

برخی اندیشمندان از سه سطح کلیدی در تبیین اخلاق و هوش مصنوعی یاد کرده‌اند: الف) اخلاق با طراحی: که منظور تعبیه اصول اخلاقی در سطح الگوریتم‌ها و رفتار سیستم‌های خودکار است؛ ب) اخلاق در طراحی: که شامل رویکردهای نظارتی و تحلیلی برای سنجش تأثیرات اجتماعی سامانه‌های هوش مصنوعی است؛ج) طراحی برای اخلاق: تمامی فرآیندهای مهندسی و استانداردسازی که رفتار توسعه‌دهندگان و کاربران را سامان‌دهی می‌کند را شامل می‌شود.

در مجموع، مواجهه با مسائل اخلاقی در هوش مصنوعی نیازمند رویکردی چندسطحی و بین‌رشته‌ای است. تصمیم‌گیری صرفاً بر مبنای کارایی یا پیشرفت فنی، بدون درنظر گرفتن پیامدهای انسانی و اجتماعی، ممکن است منجر به بحران‌های اخلاقی گسترده‌ای شود. بنابراین، طراحی و پیاده‌سازی این فناوری‌ها باید از ابتدا با مشارکت فلسفه، حقوق، جامعه‌شناسی و سیاست‌گذاری اخلاقی همراه باشد.

۲. رویکردهای اخلاقی در طراحی هوش مصنوعی

در حالی که ارزیابی فناوری‌های موفق نیازمند منابع گسترده‌ای هستند، با روش ارزیابی اخلاقی فناوری می‌توان به عنوان یک جایگزین ارزیابی کم هزینه، عمل کرد. هدف این روش این است که در طول کل فرآیند توسعه، باید با توسعه دهندگان فناوری در تماس بود و در مورد رویکردهای مختلف برای مشکلاتی که به وجود می‌آیند، بحث کرد. یک ارزیابی منفرد در حل نهایی مشکل موفق نخواهد بود و برای طرح ادعاهای اخلاقی فقط در مورد محصولات نهایی کافی نیست (Hoedemaekers, 2001, p. 275). از این رو، گفت‌وگوی مستمر و ارزیابی‌های مکرر به یک ارزیابی در مقیاس بزرگ ترجیح داده می‌شود. از آنجایی که پیامدهای اخلاقی ممکن است در تمام مراحل توسعه فناوری پدید آید، یک ارزیابی اخلاقی باید کل چرخه حیات آن را بررسی کند. به منظور پوشش کل فرآیند توسعه، ارزیابی باید با تحقیق و توسعه اولیه شروع شود و علاوه بر آن شامل ثبت اختراع، بازاریابی، آزمایش و تأثیر آزمون ها بر فرد و جامعه باشد. ارزیابی فناوری اخلاقی باید از همان ابتدا در توسعه فناوری‌های جدید مشارکت داشته باشد، به طوری که بینش‌های به‌دست‌آمده می‌تواند بر طراحی فن‌آوری تأثیر بگذارد. به این ترتیب شانس شکل‌گیری اجتماعی فناوری جدید از طریق تعامل بین ارزش‌های اجتماعی و پتانسیل فناوری افزایش می‌یابد. تأخر فرهنگی را می‌توان تا حدی کوتاه کرد و فناوری‌های جدید را بهتر با محیط اجتماعی خود تطبیق داد و ادغام کرد. ارزیابی اخلاقی فناوری به معنای رویکردی واقع بینانه و متعادل به فناوری و خدماتی برای توسعه دهندگان فناوری و تصمیم گیرندگان سیاسی است. هدف ارزیابی اخلاقی فناوری، تشخیص مشکلات اخلاقی بالقوه از طریق

گفت‌وگوی مستمر به جای ارزیابی نقطه‌ای یک فناوری خاص است. نباید به هیچ نظریه اخلاقی خاصی متعهد باشد، بلکه باید به روی دیدگاه‌ها، علایق و راه‌حل‌های مختلف و مشارکت مستقیم در فرآیند توسعه باز باشد.

۲-۱. رویکرد قاعده محور

یک رویکرد اصلی به اخلاق هوش مصنوعی (AI) استفاده از انتخاب اجتماعی است که در آن هوش مصنوعی به گونه‌ای طراحی شده است که بر اساس دیدگاه‌های کلی جامعه عمل کند. این امر در اخلاق هوش مصنوعی به دو صورت: ۱) رویکرد «قاعده‌محور» یا «کل به جزء» یا «اخلاق از بالا به پایین» و ۲) رویکرد «داده محور» یا «جزء به کل» یا «اخلاق از پایین به بالا» است. رویکرد قاعده‌محور یا اراده برون‌یابی منسجم[1] که توسط محققانی مانند یودکوفسکی (۲۰۰۴)، موئل‌هاوزر و هلم (۲۰۱۲) و بوستروم (۲۰۱۴) مطرح شده است (see Yudkowsky, 2004; Muehlhauser & Helm, 2012; Bostrom, 2014)، از انتخاب دیدگاه اخلاقی برای برنامه‌ریزی اولیه خودداری می‌کند و در عوض به دنبال این است که هوش مصنوعی ارزش‌های خود را از قواعد و ارزش‌های سایر عوامل اخلاقی استخراج کند. در رویکرد قاعده محور برای خودروهای با هوش مصنوعی، اصول اخلاقی در سیستم راهنمایی خودرو برنامه‌ریزی شده است که می‌توانند یک قانون یا حکم دینی یا یک فلسفه اخلاقی کلی باشند. نکته اصلی این است که به جای اینکه یک برنامه‌نویس به خودرو دستور دهد که تحت شرایط خاص به اخلاقی‌ترین راه پیش برود، خودرو قادر خواهد بود چنین انتخاب‌های اخلاقی را بر اساس فلسفه اخلاقی که در برنامه هوش مصنوعی آن تعبیه شده است، انجام دهد (Wallach et-al., 2008, p. 570).

برنامه‌ریزی ماشینی که قادر به اتخاذ تصمیمات اخلاقی به تنهایی باشد، چه با استفاده از یک یا ترکیبی از این فلسفه‌های اخلاقی، بسیار دشوار است. اما می‌توان پرسید: «اگر انسان‌ها می‌توانند این کار را انجام دهند، چرا ماشین‌های هوشمند نمی‌توانند؟» در پاسخ، ابتدا به این نکته اشاره می‌شود که انسان‌ها می‌توانند با تصمیم‌های مبهم کنار بیایند، اما برنامه‌نویسان کامپیوتر چنین تصمیم‌هایی را به‌ویژه دشوار می‌دانند. علاوه بر این، اگرچه می‌توان استدلال کرد که افراد بر اساس یک فلسفه مشخص انتخاب‌های اخلاقی انجام می‌دهند، انسان‌های واقعی ابتدا ارزش‌های اخلاقی خود را از طریق والدین، مربیان و سایر افراد نزدیک دریافت می‌کنند. سپس این ارزش‌ها تحت تأثیر تجربیات اجتماعی، تعامل با گروه‌های مختلف و مواجهه با فرهنگ‌ها و خرده‌فرهنگ‌های جدید اصلاح می‌شوند و به تدریج ترکیب اخلاقی شخصی خود را شکل می‌دهند. علاوه بر این، این ارزش ها تحت تأثیر اصول اجتماعی خاصی هستند که محدود به هیچ یک از فلسفه‌های اخلاقی نیستند. به طور خلاصه، رویکرد از بالا به پایین بسیار غیرممکن است.

۲-۲. رویکرد داده محور

مفهوم رویکرد داده محور که خط تفکر محققانی مانند: آلن و همکاران (۲۰۰۵) و والاچ و همکاران (۲۰۰۸) است، (see Allen et-al., 2005; Wallach et-al., 2008) بر این اصل تأکید دارد که تصمیمات اخلاقی باید بر اساس شواهد، داده‌ها و تحلیل‌های تجربی شکل بگیرند و نه صرفاً بر اساس قواعد از پیش تعیین‌شده. هوش مصنوعی با رویکرد داده محور برای یادگیری اخلاق طراحی شده است،

1. Coherent extrapolated volition

شبیه به نحوه یادگیری اخلاقیات توسط کودکان انسان هنگام رشد، زیرا با محیط خود و سایر عوامل اخلاقی در تعامل است.

در رویکرد داده‌محور برای خودروهای هوشمند انتظار می‌رود ماشین‌ها یاد بگیرند که چگونه از طریق مشاهده رفتار انسان در موقعیت‌های واقعی، بدون آموزش هیچ گونه قواعد رسمی‌یا مجهز شدن به فلسفه اخلاقی خاصی، تصمیمات اخلاقی را اتخاذ کنند. این رویکرد برای جنبه‌های غیر از اخلاق یادگیری خودروهای بدون راننده اجرا شده است. به عنوان مثال، یک وسیله نقلیه خودران اولیه که توسط محققان دانشگاه کارنگی ملون ساخته شد، پس از ۲ تا ۳ دقیقه آموزش توسط یک راننده انسانی، توانست در بزرگراه حرکت کند. ظرفیت تعمیم آن به آن اجازه می‌داد در جاده‌های چهار خطه رانندگی کند، حتی اگر فقط در جاده‌های یک یا دو لاین آموزش دیده بود (Batavia et-al., 1996, p. 5).

چون جنبه‌های اخلاقی بسیار متفاوت است، از این‌رو پیشنهاد شده است که اتومبیل‌های بدون راننده می‌توانند از تصمیمات اخلاقی میلیون ها راننده انسانی، از طریق نوعی سیستم تجمیع، به عنوان نوعی تفکر گروهی یا استفاده از خرد جمعی، درس بگیرند. با این حال، باید توجه داشت که این امر ممکن است باعث شود خودروها ترجیحات نسبتاً غیراخلاقی را به دست آورند، زیرا کاملاً واضح نیست که اکثر رانندگان استانداردی را تعیین کنند که شایسته تقلید توسط خودروهای خودران جدید باشد.

رویکرد قاعده‌محور در تقابل با رویکرد داده‌محور است، چون هوش مصنوعی در آن از ابتدا به گونه‌ای برنامه‌ریزی شده است که دیدگاه اخلاقی خاصی داشته باشد و بنابراین به دنبال شناسایی دیدگاه‌های جامعه یا هیچ یک از اعضای آن نیست. در جایی که رویکرد داده‌محور مبتنی بر یادگیری از سایر کارگزاران اخلاقی است، دیدگاه‌های اخلاقی که در نهایت به آن ختم می‌شود، مجموعه‌ای از عواملی است که از آنها می‌آموزد.

۲-۳. مقایسه تطبیقی

به طور خلاصه، هر دو رویکرد قاعده محور و داده محور با مشکلات بسیار جدی روبرو هستند. این مشکلات فنی نیستند، بلکه به ساختارهای درونی فلسفه‌های اخلاقی مورد استفاده انسان‌ها مربوط می‌شوند. برای درک دقیق‌تر تفاوت میان این دو رویکرد در طراحی اخلاقی سیستم‌های هوش مصنوعی، جدول زیر نقاط تمایز کلیدی آن‌ها را مقایسه می‌کند:

جدول (۱): جدول مقایسه‌ای رویکرد قاعده‌محور و داده‌محور در طراحی هوش مصنوعی

مؤلفه	رویکرد داده محور	رویکرد قاعده محور
منطق بنیادین	یادگیری از تجربه، مشاهده رفتار انسان‌ها و استخراج اصول اخلاقی از آن	**مبتنی بر نظریه‌های اخلاقی کلاسیک (مثل فایده‌گرایی، کانتی) و تعبیه مستقیم در الگوریتم‌ها**
ویژگی اصلی	یادگیری تدریجی و انطباق با محیط و داده‌های واقعی	**تصمیم‌گیری براساس اصول ثابت و از پیش تعریف‌شده**
مزایا	انعطاف‌پذیری بالا، نزدیک به فرآیند یادگیری انسانی	**انسجام نظری، قابلیت پیش‌بینی رفتار**
چالش‌ها	خطر یادگیری رفتار غیراخلاقی، نیاز به حجم بالای داده و نظارت	**سختی در انتخاب یک نظریه اخلاقی فراگیر، عدم انطباق با تنوع فرهنگی**
نمونه کاربرد	ربات‌هایی که با تعامل مستمر با انسان رفتار خود را اصلاح می‌کنند	**خودروهای بدون راننده با قوانین اخلاقی صریح برنامه‌نویسی‌شده**
مبنای تصمیم‌گیری	داده‌های مشاهده‌شده و نتایج رفتاری قابل تحلیل	**ارزش‌های مطلق یا قوانین جهانی**

همان‌طور که جدول نشان می‌دهد، هر یک از این دو رویکرد دارای مزایا و محدودیت‌های خاص خود هستند. این محدودیت‌ها سبب شده‌اند که محققان به دنبال گزینه‌های جایگزین یا ترکیبی باشند که از نقاط قوت هر دو رویکرد بهره ببرند .در ادامه، به بررسی مفهوم «اخلاق انتخاب اجتماعی» به عنوان چنین رویکردی خواهیم پرداخت.

۳. اخلاق انتخاب اجتماعی به‌عنوان رویکرد جایگزین در طراحی اخلاقی هوش مصنوعی

۳-۱. مبانی نظری اخلاق انتخاب اجتماعی

باوم (۲۰۲۰) در مقاله «اخلاق انتخاب اجتماعی در هوش مصنوعی»[1] تعریف می‌کند: اخلاق انتخاب اجتماعی شاخه‌ای از اخلاق است که بر اساس تجمیع دیدگاه‌های اخلاقی فردی جامعه تصمیم‌گیری می‌کند و چارچوب اخلاقی صحیح را چارچوبی می‌داند که با دیدگاه‌های جمعی جامعه مطابقت دارد. این مفهوم به نحوه ترکیب دیدگاه‌های اخلاقی افراد برای رسیدن به تصمیمات جمعی و عادلانه مرتبط است و زیربنای بسیاری از نهادهای اجتماعی مانند دموکراسی و بازارهای اقتصادی را تشکیل می‌دهد. او همچنین نشان می‌دهد که مبنای هنجاری اخلاق انتخاب اجتماعی هوش مصنوعی به دلیل این واقعیت که یک دیدگاه اخلاقی کلی از جامعه وجود ندارد، ضعیف است. در عوض، طراحی هوش مصنوعی انتخاب اجتماعی با سه مجموعه تصمیم مواجه است: ۱) جایگاه: در مورد اینکه دیدگاه‌های اخلاقی شامل چه کسانی می‌شود. ۲) اندازه گیری: در مورد چگونگی شناسایی دیدگاه‌های آنها و ۳) تجمیع: در مورد چگونگی ترکیب دیدگاه‌های فردی به یک نمای واحد که رفتار هوش مصنوعی را هدایت می‌کند (Baum, 2020, p.4). این تصمیم‌ها باید در طراحی اولیه هوش مصنوعی اتخاذ شوند. هر مجموعه‌ای از تصمیم‌گیری‌ها معضلات اخلاقی دشواری را با پیامدهای عمده‌ای برای رفتار هوش مصنوعی ایجاد می‌کند، به طوری که برخی از گزینه‌های تصمیم گیری نتایج آسیب شناختی یا حتی فاجعه بار را به همراه دارند. شناخت عواملی که بر تصمیم‌گیری اخلاقی افراد در مواجهه با فناوری تأثیر می‌گذارند، می‌تواند به غنای این رویکرد اجتماعی بیفزاید. برای نمونه، پژوهش‌هایی مانند بوالهری و همکاران (۱۴۰۱) نشان می‌دهند که متغیرهایی نظیر ادراک از احتمال فاش‌شدن و کانون کنترل می‌توانند در شکل‌دهی به نیت اخلاقی افراد در محیط‌های داده‌محور نقش داشته باشند. این یافته‌ها مؤید آن است که در کنار تجمیع دیدگاه‌ها، توجه به سازوکارهای روان‌شناختی مؤثر بر رفتار اخلاقی فردی نیز می‌تواند در طراحی نظام‌های هوشمند سودمند باشد.

پرسش‌های مربوط به جایگاه، اندازه‌گیری و تجمیع باید توسط طراحان هوش مصنوعی در ابتدای فرآیند انتخاب اجتماعی پاسخ داده شود. فرآیندهای انتخاب اجتماعی می‌توانند مسائل مربوط به جایگاه، اندازه گیری و تجمیع را در نظر بگیرند. با این حال، نحوه در نظر گرفتن این مسائل با نحوه تنظیم اولیه فرآیند تعیین می‌شود. در دموکراسی‌ها این موضوع مربوط به این است که چه کسی در مورد اینکه چه کسی رأی می‌دهد و چگونه آن رأی برگزار می‌شود. به عنوان مثال، در ایالات متحده و سایر کشورها، زنان برای اولین

1. Social choice ethics in artificial intelligence

بار زمانی از حق رأی برخوردار شدند که مردان به آنها رأی دادند. به همین دلیل، طراحان هوش مصنوعی نمی‌توانند به سادگی «به هوش مصنوعی اجازه دهند آن را بفهمد». طراحان هوش مصنوعی که از اخلاق انتخاب اجتماعی استفاده می‌کنند، باید در مورد جایگاه، اندازه گیری و تجمیع انتخاب کنند. در ادامه به توضیح در مورد جایگاه، اندازه گیری و تجمیع می‌پردازیم.

۳-۲. جایگاه (Standing)

داشتن جایگاه به معنای گنجاندن اخلاق فرد در فرآیند انتخاب اجتماعی است که برای تعیین اخلاق یک هوش مصنوعی استفاده می‌شود. در تعیین جایگاه برای تصمیم‌گیری‌های اخلاقی، همواره این پرسش مطرح است که چه کسانی باید در فرایند لحاظ شوند. این موضوع زمانی حساس‌تر می‌شود که برخی از گروه‌های اجتماعی مانند نسل‌های آینده، کودکان یا اقلیت‌ها مستقیماً در فرآیند تصمیم‌گیری مشارکت ندارند، ولی تحت تأثیر عمیق نتایج آن قرار می‌گیرند. بنابراین، طراحی اخلاق در هوش مصنوعی باید به شکلی باشد که نه‌تنها دیدگاه فعلی کاربران، بلکه پیامدهای بین‌نسلی و منافع گروه‌های نادیده‌گرفته‌شده را نیز در نظر بگیرد.

دیدگاه‌ها درباره اعطای جایگاه اخلاقی و حقوقی به هوش مصنوعی بسیار متفاوت‌اند. برخی همچون هابارد معتقدند اگر یک سیستم هوشمند دارای ویژگی‌هایی مشابه موجودات دارای حقوق باشد، باید از همان جایگاه برخوردار شود (Hubbard, 2011, p. 410). مارتین نیز بر این باور است که هوش مصنوعی پیشرفته می‌تواند به سطحی از استدلال اخلاقی برسد که با انسان‌ها برابری کرده یا حتی فراتر رود (Martin, 2017, p. 953). در مقابل، یامپولسکی (Yampolskiy, 2013, p. 393) تأکید می‌کند که هوش مصنوعی نباید به مرحله‌ای برسد که بتوان برای آن حقوق یا استقلال اخلاقی قائل شد، بلکه باید به‌عنوان ابزاری مطیع و مصرف‌پذیر در کنترل انسان باقی بماند. از این منظر، اعطای جایگاه به هوش مصنوعی ممکن است چالش‌هایی در فرآیندهای تصمیم‌گیری جمعی و انتخاب اجتماعی ایجاد کند.

۳-۳. اندازه‌گیری(Measurement)

اندازه‌گیری در این زمینه به فرآیندی اشاره دارد که از طریق آن دیدگاه‌های اخلاقی یک فرد برای گنجاندن در فرآیند انتخاب اجتماعی شناسایی می‌شود. اندازه‌گیری دیدگاه‌های اخلاقی افراد می‌تواند با ابزارهایی مانند نظرسنجی، تحلیل رفتار یا داده‌کاوی انجام شود. اما به دلیل تنوع ارزش‌ها و تضاد نگرش‌ها، هیچ روش واحدی پاسخ‌گو نیست و طراحان باید به دقت انتخاب کنند که کدام روش با اهداف سامانه سازگاری بیشتری دارد.

در پاسخ به ناهماهنگی ظاهری اخلاق انسانی، برخی از اخلاق‌دانان هوش مصنوعی خواستار اندازه‌گیری نسخه ایده‌آلی از اخلاق انسانی به جای آنچه در انسان‌های واقعی مشاهده می‌شود، هستند. در واقع باید انتظار داشت که اخلاق ایده آل یک انسان با اخلاق اصلی او متفاوت باشد (Baum, 2020, p.8).

۳-۴. تجمیع (Aggregation)

در یک فرآیند انتخاب اجتماعی مهم نیست که چه کسی جایگاه دارد و دیدگاه‌های اخلاقی آنها چگونه سنجیده می‌شود، پس از انجام

اندازه‌گیری‌ها، گام نهایی این است که این دیدگاه‌های اخلاق فردی را در یک دیدگاه اخلاق اجتماعی واحد تجمیع کنیم. برای هوش مصنوعی، یک دیدگاه واحد برای تعیین چگونگی تصمیم گیری هوش مصنوعی مورد نیاز است. به عبارت دیگر، دیدگاه اخلاقی کل چیزی است که هوش مصنوعی قرار است از آن برای تصمیم گیری در مورد نحوه رفتارش استفاده کند.

مقایسه بین فردی قدرت دیدگاه اخلاقی به دلیل این واقعیت پیچیده است که مردم اغلب دارای ارزش‌های مقدس هستند که به عقیده آنها نباید با ارزش‌های دیگر معاوضه شود، دقیقاً مانند قواعد دئونتولوژیک یا اخلاق وظیفه گرا (Ginges et-al., 2007, p. 7359). برای تشکیل یک دیدگاه اخلاقی واحد، یک فرآیند تجمیع باید تضاد بین ارزش‌های محافظت شده افراد مختلف را حل کند.

۳-۵. چالش‌ها و فرصت‌ها

با وجود جذابیت نظری مفهوم اخلاق انتخاب اجتماعی، اجرای آن در عمل با چالش‌های متعددی مواجه است. نخستین چالش، تعیین دقیق دامنه مشارکت‌کنندگان در فرآیند تصمیم‌گیری اخلاقی است. در بسیاری از موارد، گروه‌هایی مانند اقلیت‌ها، نسل‌های آینده، یا جوامع حاشیه‌نشین فاقد ابزار و بسترهای لازم برای ابراز دیدگاه‌های خود هستند، و در نتیجه ممکن است اخلاق شکل‌گرفته در سامانه‌های هوش مصنوعی، بازتاب‌دهنده دیدگاه گروهی محدود باشد. چالش دوم، ابهام در روش‌های اندازه‌گیری ارزش‌های اخلاقی است. دیدگاه‌های اخلاقی افراد نه‌تنها پیچیده، بلکه متناقض، سیال و زمینه‌محور هستند. ابزارهایی مانند نظرسنجی، داده‌کاوی یا تحلیل رفتاری ممکن است نتوانند عمق، ظرافت یا نیت‌مندی پشت برخی ارزش‌ها را بازتاب دهند. این موضوع خطر ساده‌سازی و سطحی‌سازی اخلاق در سامانه‌های هوشمند را افزایش می‌دهد.در مرحله تجمیع نیز، چالش ناسازگاری ارزش‌های محافظت‌شده (مانند ارزش‌های مقدس دینی یا هویتی) مانعی جدی برای رسیدن به یک دیدگاه جمعی اخلاقی است. برخی ارزش‌ها قابل مصالحه یا میانگین‌گیری نیستند، و تلاش برای ترکیب آن‌ها ممکن است منجر به تصمیم‌های گنگ یا حتی فاجعه‌بار شود (Baum, 2020, pp.13-14).

در کنار این چالش‌ها، این رویکرد فرصت‌هایی مهم نیز فراهم می‌کند. از جمله‌ی آن‌ها، دموکراتیزه‌کردن طراحی فناوری، بازتاب تنوع فرهنگی و تسهیل گفت‌وگوی میان‌رشته‌ای درباره اخلاق در عصر دیجیتال است. استفاده از روش‌هایی مانند مشارکت شهروندی، گفت‌وگوی میان‌فرهنگی و مدل‌های تعاملی می‌تواند کارآمدی این رویکرد را افزایش دهد (جعفری و دیگران، ۱۴۰۴، ص ۶۸).

در نهایت، برای استفاده موفق از اخلاق انتخاب اجتماعی در طراحی هوش مصنوعی، باید سازوکارهای مشارکت واقعی، شفافیت در انتخاب معیارها و نظارت بینارشته‌ای به دقت طراحی و اجرا شوند. این امر نیازمند همکاری نزدیک میان طراحان فناوری، فیلسوفان اخلاق، متخصصان علوم اجتماعی و نهادهای تصمیم‌ساز است.

۴. بحث و تحلیل نهایی

در رویکرد «اخلاق انتخاب اجتماعی» فرض بر آن است که تصمیمات اخلاقی نباید صرفاً محصول ذهن طراحان یا برنامه‌نویسان هوش مصنوعی باشد، بلکه باید حاصل مشارکت گروه‌های گسترده‌تری از جامعه باشد که به‌طور مستقیم یا غیرمستقیم تحت تأثیر

این فناوری قرار می‌گیرند (Baum, 2020, p.14). این رویکرد تلاش دارد تعادل میان خودمختاری طراحان و عدالت اجتماعی را از طریق سه مؤلفه کلیدی جایگاه، اندازه‌گیری و تجمیع برقرار کند.

با این حال، این ایده‌آل مشارکتی با چالش‌های متعددی همراه است. نخست آن که تعیین جایگاه اخلاقی به معنای آن است که مشخص شود چه کسانی باید در فرآیند تصمیم‌سازی لحاظ شوند. این موضوع زمانی پیچیده می‌شود که برخی گروه‌ها مانند نسل‌های آینده یا اقلیت‌های کم‌صدا، امکان مشارکت مستقیم ندارند، ولی نتایج تصمیم‌ها به‌شدت بر آن‌ها اثر می‌گذارد (Yampolskiy, 2013, p. 393; Hubbard, 2011, p. 410).

دومین مسئله، چگونگی اندازه‌گیری دیدگاه‌های اخلاقی افراد است. روش‌هایی نظیر نظرسنجی، تحلیل رفتاری و داده‌کاوی ابزارهایی ممکن برای شناسایی دیدگاه‌ها هستند، اما هیچ‌یک به‌تنهایی توانایی درک عمق ارزش‌های اخلاقی را ندارند. برخی پژوهشگران پیشنهاد کرده‌اند که به جای سنجش اخلاق واقعی افراد، نسخه ایده‌آلی از اخلاق انسانی استخراج شود (Ginges et al., 2007, p. 7359). با این حال، فاصله بین اخلاق واقعی و ایده‌آل خود می‌تواند منشأ خطای تصمیم‌گیری باشد.

در نهایت، مهم‌ترین چالش در این رویکرد، تجمیع دیدگاه‌ها است؛ یعنی اینکه چگونه باید مجموعه‌ای از ارزش‌های متضاد، به یک تصمیم واحد و اجرایی برای سیستم‌های هوشمند تبدیل شود. دیدگاه‌های فردی اغلب مبتنی بر ارزش‌های محافظت‌شده یا اخلاق وظیفه‌گرا هستند که به‌سادگی قابل مصالحه نیستند .هرگونه مدل تجمیعی که نتواند این تضادها را به‌درستی مدیریت کند، ممکن است به تصمیماتی منجر شود که از نظر اخلاقی ناکارآمد یا حتی فاجعه‌بار باشند.

در مجموع، اخلاق انتخاب اجتماعی مسیر مهمی برای دموکراتیزه‌کردن فرآیند طراحی هوش مصنوعی ارائه می‌دهد، اما موفقیت آن منوط به پاسخ‌گویی شفاف و مسئولانه به پرسش‌هایی است که پیرامون جایگاه، اندازه‌گیری و تجمیع وجود دارد. طراحی این سازوکار باید به گونه‌ای باشد که نه‌تنها بازتاب‌دهنده دیدگاه‌های فعلی، بلکه حساس به آینده، تنوع فرهنگی و گروه‌های نادیده‌گرفته‌شده نیز باشد.

نتیجه‌گیری

توسعه شتاب‌زده فناوری‌های نوین، به‌ویژه هوش مصنوعی، بار دیگر ما را با مسئله‌ای قدیمی ولی بنیادین یعنی «تأخر فرهنگی» مواجه ساخته است. فناوری‌هایی که در حل مسائل پیچیده و تصمیم‌گیری‌های حساس، توانمندی بی‌سابقه‌ای دارند، بدون هدایت و نظارت اخلاقی، ممکن است خود منشأ بی‌عدالتی‌های تازه شوند. این مقاله تلاش کرد نشان دهد که برای پر کردن این شکاف اخلاقی، نمی‌توان صرفاً به ساختارهای سنتی تصمیم‌گیری اکتفا کرد.

رویکرد اخلاقی قاعده محور به دلیل ناتوانی در بازتاب تنوع فرهنگی، و رویکرد داده محوربه دلیل پیچیدگی در تفسیر رفتار انسانی، هر دو با چالش‌هایی مواجه‌اند. در این میان، اخلاق انتخاب اجتماعی گزینه‌ای میان‌محور است که تلاش دارد از طریق دخالت دیدگاه‌های متنوع، به تصمیم‌های اخلاقی نزدیک‌تر به اجماع اجتماعی دست یابد.

با این حال، همان‌گونه که نشان داده شد، انتخاب اجتماعی نیز پرسش‌های دشواری در مورد جایگاه (چه کسی در نظر گرفته

می‌شود)، اندازه‌گیری (چگونه ارزش‌های اخلاقی شناسایی می‌شوند)، و تجمیع (چگونه دیدگاه‌ها به تصمیم واحد منتهی می‌شوند) مطرح می‌سازد. طراحی اخلاقی در هوش مصنوعی، تنها در صورتی موفق خواهد بود که این پرسش‌ها با شفافیت و مسئولیت‌پذیری پاسخ داده شوند.

در نهایت، پیشنهاد این مقاله آن است که فرآیند طراحی و پیاده‌سازی هوش مصنوعی باید نه‌تنها فنی، بلکه عمیقاً اجتماعی و اخلاقی باشد. مشارکت عمومی، توجه به تنوع فرهنگی و بین‌نسلی، و بهره‌گیری از مدل‌های انتخاب اجتماعی، گام‌هایی مؤثر در جهت اخلاق‌محور کردن هوش مصنوعی است. بدون این تمهیدات، خطر تثبیت نابرابری‌های فناورانه و بی‌عدالتی‌های نوین بیش از هر زمان دیگری جدی خواهد بود.

منابع

احمدی، سیده معصومه؛ خلخالی، علی؛ کاظم پور، اسماعیل. (۱۴۰۲). «حاکمیت اخلاقی بر هوش مصنوعی آموزشی: مرور نظام‌مند چالش‌ها و طراحی چارچوب مدیریتی». اخلاق در علوم و فناوری. ۲۱ (۱) :۳۴-۲۵.

ایزانلو، هادی؛ مسعودی، جهانگیر؛ پیک حرفه، شیرزاد . (۱۴۰۲).«میزان مطالبه اخلاق با تاکید بر مؤلفه اولویت در اندیشه اسلامی». *تأملات اخلاقی. ۴(۴): ۱۰۹-۱۳۴.*

بوالهری، علیرضا؛ رادفر، رضا؛ البرزی، محمود؛ پورابراهیمی، علیرضا؛ دهقانی، محمود . (۱۴۰۱).«تبیین تصمیم‌گیری اخلاقی در بستر مدیریت فناوری اطلاعات». *تأملات اخلاقی*. ۳(۲): ۲۹-۴۶.

جعفری، رضا؛ رحمتی، حسینعلی. (۱۴۰۴). «کاربرد هوش مصنوعی در مشاورهٔ اخلاقی ؛ از ظرفیت‌های تحول‌آفرین تا چالش‌های اخلاقی». *فصلنامه اخلاق‌پژوهی*. ۸(۲): ۴۹-۷۲.

میری بالاجورشری، سیده مهشید؛ محمودی، امیررضا . (۱۴۰۳). «واکاوی مسائل اخلاقی در زمینه هوش مصنوعی با نگاهی به اخلاق اسلامی». *فصلنامه علمی - پژوهشی مطالعات اخلاق کاربردی*. ۲۰(۱): ۹۷-۱۲۳.

میرزامحمدی، عبدالرئوف؛ تقوی، مصطفی . (۱۴۰۴). «اخلاق فناوری: ارزیابی انتقادی مناقشه میان دیوید مورو و ژوزف. سی. پیت درباره‌ی ارزش‌باری فناوری و مسئولیت اخلاقی». *تأملات اخلاقی. ۶(۲): ۹۳-۱۱۴.*

Allen, C., Smit, I., & Wallach, W. (2005). "Artificial morality: Top-down, Bottom-up, And Hybrid Approaches". *Ethics and Information Technology*, 7(3): 149–155.

Baum, D. (2020). "Social Choice Ethics in Artificial Intelligence". *AI & Society*, 35(1): 165–176. https://doi.org/10.1007/s00146-017-0760-1.

Batavia, H., Parag, H., Dean, A., Pomerleau, E., Charles, E., & Thorpe, E. (1996). "Applying Advanced Learning Algorithms To ALVINN". *Carnegie Mellon University, The Robotics Institute*.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). "The Social Dilemma Of Autonomous Vehicles". *Science*, 352(6): 1573–1576.

Bostrom, N. (2014). "Superintelligence: Paths, Dangers, Strategies". Oxford, UK: *Oxford University Press*.

Catch, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). "Artificial Intelligence and the 'Good Society': The US, EU, And UK Approach". *Science And Engineering Ethics*, 24(2): 505–528.

Collins, F. S., & Varmus, H. (2015). "A New Initiative on Precision Medicine". *New England Journal of Medicine*, 372(9): 793–795.

Cordeiro, W. P. (1997). "Suggested Management Responses to Ethical Issues Raised by Technological Change". *Journal Of Business Ethics*, 16(12–13): 1393–1400.

Dignum, V. (2018). "Ethics In Artificial Intelligence: Introduction to The Special Issue". *Ethics And Information Technology*, 20(1): 1–3. https://doi.org/10.1007/s10676-018-9450-z.

Floridi, L. (2018). "Artificial Intelligence, Deepfakes and A Future of Ectypes". *Philosophy & Technology*, 31(3): 317–321.

Gewirth, A. (1978). "Reason And Morality". *Chicago, IL: University Of Chicago Press*.

Ginges, J., Atran, S., Medin, D., & Shikaki, K. (2007). "Sacred Bounds on Rational Resolution Of Violent Political Conflict". *Proceedings Of the National Academy of Sciences*, 104(18): 7357–7360.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). "When Will AI Exceed Human Performance? Evidence From AI Experts". *Journal Of Artificial Intelligence Research*, 62: 729–754.

Harper, R. (2006). "Inside The Smart Home". *Springer Science & Business Media*.

Hossain, M. S., Muhammad, G., & Guizani, N. (2020). "Explainable AI And Mass Surveillance System-Based Healthcare Framework to Combat COVID-19 Like Pandemics". *IEEE Network*, 34(4): 126–132.

Hoedemaekers, R. (2001). "Commercialization, Patents and Moral Assessment of Biotechnology Products". *Medicine, Health Care and Philosophy*, 2(3): 273–284.

Hubbard, F. P. (2011). "Do Androids Dream? Personhood and Intelligent Artifacts". *Temple Law Review*, 83: 405–441.

Lin, P. (2016). "Why Ethics Matters For Autonomous Cars". In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), Autonomous Driving: Technical, Legal And Social Aspects (pp. 69–85). *Berlin: Springer*.

Martin, D. (2017). "Who Should Decide How Machines Make Morally Laden Decisions?". *Science And Engineering Ethics*, 23(4): 951–967.

Marshall, P. (1999). "Has Technology Introduced New Ethical Problems?". *Journal Of Business Ethics*, 19: 81–90.

Muehlhauser, L., & Helm, L. (2012). "Intelligence Explosion and Machine Ethics". In A. Eden, J. Søraker, J. H. Moor, & E. Steinhart (Eds.), Singularity Hypotheses: A Scientific And Philosophical Assessment (pp. 101–126). *Springer*.

Müller, V. C. (2020). "Ethics Of Artificial Intelligence and Robotics". *Stanford University*.

Ogburn, W. F. (1966). "Social Change with Respect to Cultural and Original Nature". *New York*.

Robinson, J. P. (1981). "Will The New Electronic Media Revolutionize Our Daily Lives?" In R. W. Haigh, G. Gerbner, & R. B. Byrne (Eds.), Communications in The Twenty-First Century. *New York, NY: John Wiley & Sons*.

Verbeek, P.-P. (2006). "Materializing Morality: Design Ethics and Technological Mediation". *Philosophy & Technology*, 19(3): 361–380.

Walter, R., Finger, R., Huber, S., & Buchmann, N. (2017). "Smart Farming Is Key To Developing Sustainable Agriculture". *Proceedings Of The National Academy of Sciences*, 114(24): 6148–6150.

Wallach, W., & Allen, C. (2008). "Moral Machines: Teaching Robots Right from Wrong". *Oxford University Press*.

Wallach, W., Allen, C., & Smit, I. (2008). "Machine Morality: Bottom-Up and Top-Down Approaches for Modelling Human Moral Faculties". *AI & Society*, 22(4): 565–582.

Yampolskiy, R. V. (2013). "Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach". *In V. C. Müller (Ed.), Philosophy and Theory of Artificial Intelligence* (pp. 389–396). Springer.

Yudkowsky, E. (2004). "Coherent Extrapolated Volition". *The Singularity Institute, San Francisco*.