Science and Religion Studies, Institute for Humanities and Cultural Studies (IHCS)
Biannual Journal, Vol. 15, No. 1, Spring and Summer 2024, 125-151
https://www.doi.org/10.30465/srs.2024.49282.2162

Investigation of the Ethical Agency of AGI

Mohammad Ali Ashouri Kisomi*
Maryam Parvizi**

Abstract

The aim of this paper is to investigate the ethical agency of artificial general intelligence (AGI). Moor classifies the ethical agency of AI into four distinct categories: 1: ethical-impact agents; 2: implicit ethical agents; 3: explicit ethical agents; and 4: full ethical agents. In this paper, we used the analytical-critical method to explore the fourth category, i.e., full ethical agents, concerning AGI. At this level, AGI acquires various abilities, which raises many ethical concerns. The results demonstrate that the idea of AGI as a full ethical agent needs to be modified. We argued that the concept of the full ethical agent needs to be modified to cover AGI, ensuring it does not fall into anthropomorphism.

Keywords: Artificial General Intelligence, Ethics of AI, Ethical Agency, Orthogonality Thesis, Neuromorphic Artificial Intelligence, Artificial Intelligence.

Introduction

Artificial intelligence (AI) is not possible without computers. When we shifted from mechanical computers to the first digital computers, we did not have a well-established image of the human brain. At that time, it was believed that the human brain was like a digital computer. However, we now know that our brain, unlike computers, does not have a central processor (CPU), operating system, software, etc. This understanding is largely due to progress in neuroscience. Today, we have a better understanding that what happens in our brain is neuronal activity trying to find different patterns (Kaku,

* Ph.D. in Philosophy, Allameh Tabataba'i University (Corresponding Author), m_ashori@atu.ac.ir
** Ph.D. in Comparative Philosophy, Allameh Tabataba'i University, maryam_parvizi@atu.ac.ir
Date received: 20/06/2024, Date of acceptance: 13/10/2024

2011). It isn't surprising that for decades, computer technologies developed with the same image of the brain as a computer.

There have been various discussions about the feasibility or impossibility of AGI (Fortnow, 2021). Some scholars believe that neuromorphic AI, due to the similarity of its structure with the human brain, can finally help us to reach the level of AGI (Pei et al., 2019; Pontes-Filho & Nichele, 2019; Pontes-Filho et al., 2022). However, if AGI becomes possible, one of the most important philosophical and ethical issues will be the matter of agency. In this paper, we will investigate ethical agency of AGI.

Materials & Methods

Moor (2006) classifies the ethical agency of AI under four categories: 1- ethical-impact agents, 2- implicit ethical agents, 3- explicit ethical agents, and 4- full ethical agents. We examine Moor's classification of ethical agency with respect to AGI.

To fulfill our objectives, this paper is divided into four sections. In the first section, AGI and a neuromorphic system are introduced. In the second section, Moor's classification is explored. Then, in the next section, the dangers of AGI narration will be examined. In the final section, our discussion and analysis are presented.

Discussion & Result

Muller and Cannon (2022) demonstrated that the two theses of "singularity" and "orthogonality" cannot lead to the argument of existential risk of AI for humans; however, the issue of agency is still noteworthy from two reasons:

The risks of the tool in the hands of human agents.

The agency of AGI. To clarify this issue, an example will be helpful:

Suppose an AGI agent (X) knows that a human agent (Y) has embedded a button to turn off in X. For X, turning off is equivalent to dying and losing agency. Therefore, probably the first thing that X tries to learn is how to disable that button or find a way to eliminate Y's ability to press the button. In this scenario, although Y has embedded a button in X's body to protect himself, this has caused X to also feel threatened by Y and need to confront or control Y or eliminate Y's control. Considering the issue that was raised about increasing the capabilities of AGI, X will gain more abilities over time than Y. As a result, it will be almost impossible for Y to control X.

A prima facie conclusion will probably be that if we want to design an AGI system, we will not have control over it. Based on the orthogonality thesis, there is no connection between intelligence and goals; that is to say, we do not know what the

goals of AGI will be. In such a scenario, the ethical codes and guidelines for developing AI will not be effective.. For instence, transparency, human supervision, etc. (Muller, 2021) are only effective if they are in line with the goals (unknown) of AGI. Because, in fact, there is no reason for the AGI agent to follow ethical codes that are not in line with its goals (Yampolskiy, 2022).

The results demonstrate that we should place the ethical agency of AGI under Moor's fourth category of ethical agency, but here we will face some problematics:
Based on the orthogonality thesis:

1: AGI can have different goals.

2: The goals of AGI may be different from or against human goals.

But based on Moor's classification, we have:

1: AI can learn ethics from humans.

2: AI will have ethical agency like a mature human.

Our finding suggests that Moor doesn't account for contradictory ethical goals and, to some extent, his classification is anthromorphic. To reformulate Moor's fourth classification and avoid anthropomorphism, two suggestions were made:

1: AGI can learn human ethical principles.

2: The ethical agent of AGI can shape its own ethical goals and principles, and these principles may be different from or contradictory with human ethical principles/codes.

Conclusion

The results of the present paper establish that Moor's classification of ethical agency is applicable to ethical impact agents, implicit agents, and explicit agents. However, his account of the full ethical agent is flawed. Based on the orthogonality thesis, it was determined that AGI may have agency and may understand human ethical principles. However, AGI's ethical principles are not necessarily aligned with human ethics. Moreover, AGI may pursue goals that are different from or contradictory to human ethical goals. Another result is that, in this context, automorphism fails under analysis and needs to be avoided.

Bibliography

Alvarado, R. (2023). AI as an Epistemic Technology. *Science and Engineering Ethics*, *29*(5), 32.

Asimov, I. (1984). The Bicentennial Man. Philosophy and Science Fiction (Philips, M., ed). New York: Prometheus Books, Buffalo.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. Minds and Machines, 22(2 - special issue 'Philosophy of AI' ed. Vincent C. Müller), 71–85.

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

Carabantes, M. (2020). Black-box artificial intelligence: an epistemological and critical analysis. *AI & society*, *35*(2), 309-317.

Chomsky, N., Roberts, I., & Watumull, J. (2023). Noam Chomsky: The False Promise of ChatGPT. The New York Times, 8.

Cervantes, J. A., López, S., Rodríguez, L. F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial moral agents: A survey of the current status. *Science and engineering ethics*, *26*, 501-532.

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. Science and engineering ethics, 26(4), 2051-2068.

Coeckelbergh, M. (2023). Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI and Ethics*, *3*(4), 1341-1350.

Dennett, D. C. (2019). Clever evolution. Metascience, 28(3), 355–358.

Descartes, Rene. (1955). The philosophical works of Descartes, Voume 1 (E.S, Haldane & G.R.T, Ross Trans.). New York: Dover Publications.

Fortnow, L. (2021). Fifty years of P vs. NP and the possibility of the impossible. *Communications of the ACM*, *65*(1), 76-85.

Goertzel, B. (2014). Artificial General Intelligence: Concept, State of the Art, and Future Prospects, Journal of Artificial General Intelligence 5(1) 1-46, DOI: 10.2478/jagi-2014-0001 Accepted 2014-3-15.

Goertzel, B. Pennachin, C. (2007). Artificial General Intelligence. Springer.

Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. In F. L. Alt & M. Ruminoff (Eds.), Advances in computers (Vol. 6, pp. 31–88). Academic Press.

Gubrud, M. A. (1997). Nanotechnology and international security. In Fifth Foresight Conference on Molecular Nanotechnology, 1.

Hagendorff, T. (2022). Blind spots in AI ethics. AI and Ethics, 2(4), 851-867.

Huang, T. J. (2017). Imitating the brain with neurocomputer a "new" way towards artificial general intelligence. *International Journal of Automation and Computing*, *14*(5), 520-531.

Ivanov, D. Chezhegov, A. Kiselev, M. Grunin, A. Larionov, D. (2022). "Neuromorphic artificial intelligence systems". *Frontiers in Neuroscience*, *16*, 1513. Doi: 10.3389/fnins.2022.959626

Kaku, M. (2011). *Physics of the future: How science will shape human destiny and our daily lives by the year 2100*. Anchor.

Kurzweil, R. (2005). The singularity is near: When humans transcend biology, New York: Penguin.

Landgrebe, J., & Smith, B. (2021). An argument for the impossibility of machine intelligence. *arXiv preprint arXiv:2111.07765*.

Longino, H. E. (2022). What's Social About Social Epistemology?. *The Journal of Philosophy*, *119*(4), 169-195.

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

Muller, Vincent. (2021). Ethics of Artificial Intelligence and Robotics. The Stanford Encyclopedia of Philosophy. Retrieved December 1, 2023, from https://plato.stanford.edu/entries/ethics-ai/

Muller, Vincent. Bostrom, Nick. (2016). Future progress in artificial intelligence: A survey of expert opinion. Fundamental issues of artificial intelligence, 555-572. https://doi.org/10.1007/978-3-319-26485-1_33

Muller. V. Cannon M. (2022). "Existential risk from AI and orthogonality: Can we have it both ways?"*WILEY*, DOI: 10.1111/rati.12320, P: 25-36.

Munn, L. (2023). The uselessness of AI ethics. AI and Ethics, 3(3), 869-877.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2023). A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435.

Pei, J., Deng, L., Song, S., Zhao, M., Zhang, Y., Wu, S., ... & Shi, L. (2019). Towards artificial general intelligence with hybrid Tianjic chip architecture. *Nature*, *572*(7767), 106-111.

Pontes-Filho, S., Nichele, S. (2019). Towards a framework for the evolution of artificial general intelligence. *arXiv preprint arXiv:1903.10410*.

Pontes-Filho, S. Olsen, K., Yazidi, A. Riegler, M. A. Halvorsen, P. Nichele, S. (2022). Towards the Neuroevolution of Low-level artificial general intelligence. *Frontiers in Robotics and AI*, *9*, 1007547.

Puaschunder, J. M. (2019). On Artificial Intelligence's razor's edge: On the future of democracy and society in the artificial age. In *Proceedings of the 12th International RAIS Conference on Social Sciences and Humanities* (pp. 37-51). Scientia Moralitas Research Institute.

Reid, D. K. (1979). Equilibration and learning. *Journal of education*, *161*(1), 51-71.

Rosenberg, L. (2023). The Manipulation Problem: Conversational AI as a Threat to Epistemic Agency. *arXiv preprint arXiv:2306.11748*.

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

Searle, John. (1998). The mystery of consciousness. London: Granta Books.

Sullins, J. P. (2011). When is a robot a moral agent. *Machine ethics*, *6*(2001), 151-161.

Thomas, Wolfgang. (2015). Algorithms: From Al-Khwarizmi to Turing and Beyond. In G. Sommaruga & T. Strahm (Eds.), Turing's revolution: The impact of his ideas about computability, 29-42.

Varlamov, O. O. Chuvikov, D. A. Adamova, L. E. Petrov, M. A. Zabolotskaya, I. K. Zhilina, T. N. (2019). Logical, philosophical and ethical aspects of AI in medicine. International Journal of Machine Learning and Computing, 9(6), 868-873.

Wallach, W. Allen, C. (2008). Moral machines: Teaching robots right from wrong. Oxford University Press.

Yampolskiy, R. V. (2022). Agi control theory. In Artificial General Intelligence: 14th International Conference, AGI 2021, Palo Alto, CA, USA, October 15–18, 2021, Proceedings 14 (pp. 316-326). Springer International Publishing.

Zucca, P. (2022). Four cognitive-ecological biases that reduce integration between medical and cyber intelligence and represent a threat to cybersecurity. *Forensic Science International: Animals and Environments*, *2*, 100046.

پژوهش‌های علم و دین، پژوهشگاه علوم انسانی و مطالعات فرهنگی
دوفصل‌نامهٔ علمی (مقالهٔ پژوهشی)، سال ۱۵، شمارهٔ ۱، بهار و تابستان ۱۴۰۳

بررسی عاملیت اخلاقی هوش مصنوعی عام

محمد علی عاشوری کیسمی*

مریم پرویزی**

چکیده

هدف از مقاله حاضر بررسی عاملیت اخلاقی هوش مصنوعی عام است. در بسیاری از پژوهش‌ها، عاملیت اخلاقی هوش مصنوعی در چهار دسته ۱- عامل تأثیر اخلاقی، ۲- عامل تأثیر اخلاقی ضمنی، ۳- عامل تأثیر اخلاقی آشکار و ۴- عامل اخلاقی کامل، تقسیم‌بندی می‌شود. در این مقاله با استفاده از روش تحلیلی-انتقادی سطح چهارم این دسته‌بندی یعنی عامل اخلاقی کامل در هوش مصنوعی عام موردبررسی قرار خواهد گرفت. در صورت امکان‌پذیر بودن هوش مصنوعی عام، چنین هوشی توانایی‌های فراوانی به دست می‌آورد، و لذا نگرانی‌های اخلاقی بسیاری وجود خواهد داشت. یکی از مهم‌ترین دلایل اهمیت مقاله حاضر این است که این دسته‌بندی عاملیت اخلاقی به تکرار مورداستفاده پژوهشگران قرارگرفته و نیاز به یک بررسی مجدد احساس می‌شود. نتایج بررسی حاضر نشان می‌دهد که اگر هوش مصنوعی عام امکان‌پذیر باشد باید آن را در دسته عامل اخلاقی کامل قرار داد. البته چنین عامل اخلاقی کاملی دو ویژگی دارد: اول، عامل اخلاقی هوش مصنوعی عام می‌تواند اصول اخلاقی انسان را یاد بگیرد و یا اصول اخلاقی دیگری از آن استنتاج کند؛ اما نباید تصور کرد که اصول اخلاقی انسان را اصول اخلاقی خود بداند؛ و دوم، عامل اخلاقی هوش مصنوعی عام می‌تواند اهداف و اصول اخلاقی خود را شکل دهد و این اصول ممکن است متفاوت/بر ضد اصول اخلاقی انسان‌ها باشد.

کلیدواژه‌ها: هوش مصنوعی عام، اخلاق هوش مصنوعی، عاملیت اخلاقی، قضیه متعامد، هوش مصنوعی نورومورفیک، هوش مصنوعی.

* دکتری، فلسفه، دانشگاه علامه طباطبائی (نویسندهٔ مسئول)، m_ashori@atu.ac.ir

** دکتری، فلسفه تطبیقی، دانشگاه علامه طباطبائی، maryam_parvizi@atu.ac.ir

تاریخ دریافت: ۱۴۰۳/۰۳/۳۱، تاریخ پذیرش: ۱۴۰۳/۰۷/۲۲

۱. مقدمه

قرن‌ها پیش از دست‌یابی به اولین سیستم هوش مصنوعی (Artificial Intelligence)، فیلسوفانی بودند که در تفکرات و پرسش‌های فلسفی آن‌ها می‌توان ریشه‌هایی از اندیشیدن به این سیستم‌ها را مشاهده کرد. زمانی که دکارت این پرسش را مطرح می‌کند که اگر یک ماشین بتواند عملکرد و ظاهری مشابه با انسان داشته باشد، چطور می‌توان این دو را از هم تمیز داد؟ (Descartes, 1955: 116) می‌توان آن را پرسشی فلسفی در خصوص هوش مصنوعی دانست که حتی پس از گذشت سال‌ها هنوز هم برای فیلسوفان موضوعیت دارد[۱].

دستیابی به هوش مصنوعی بدون وجود کامپیوترها میسر نبود. زمانی که از مسیر تولید کامپیوترهای مکانیکی (Mechanical Computer) به سمت ساخت اولین کامپیوترهای دیجیتال (Digital Computer) حرکت کردیم، تصویر درستی از مغز انسان نداشتیم. در آن زمان تصور بر این بود که مغز انسان به‌مانند یک کامپیوتر دیجیتال عمل می‌کند. اکنون و پس از گذشت چند دهه به این موضوع دست یافتیم که مغز به‌مانند یک کامپیوتر دارای یک پردازنده مرکزی (CPU)، سیستم‌عامل (Operating System)، برنامه نرم‌افزاری (Software) و غیره نیست. این موضوع تا حدود زیادی مرهون پیشرفت ما در علم اعصاب‌شناسی است؛ و امروز می‌دانیم که آنچه در مغز رخ می‌دهد، فعالیت‌های نورونی، در تلاش برای یافتن الگوهای مختلف است (Kaku, 2011). در ابتدا تکنولوژی‌های کامپیوتری با همین تصویر مغز به‌مانند کامپیوتر پیشرفت کرد و در کنفرانس دارپا (DARPA)، پایه‌های اولیه عملی هوش مصنوعی گذاشته شد، اما تا سالیان طولانی مشکل اساسی این بود که در هوش مصنوعی به دنبال سیستم‌هایی بودیم که عملکردی هوشمند، مشابه با عملکرد انسان باشند و سخت‌افزارهایی در دست داشتیم که شباهتی با مغز انسان نداشتند. برای توسعه هوش مصنوعی که بتواند عملکردی مشابه شبکه‌های عصبی مغز انسان داشته باشند، نیازمند سیستم‌های سخت‌افزار و نرم‌افزاری هستیم که بتواند به‌مانند مغز عمل کنند. به سخنی دیگر، به نظر هوشمندانه‌تر این می‌رسد که کامپیوترها را مشابه با مغز انسان بسازیم تا انتظارمان از هوشمندیِ هوش مصنوعی به واقعیت نزدیک‌تر باشد.

زمانی که از هوشمندی سخن می‌گوییم، یک دشواری اصلی این است که مقصود از هوشمندی چیست؟ برخی این سخن را منتسب به ژان پیاژه می‌دانند که می‌گوید: «هوشمندی، دانستن این است که وقتی نمی‌دانیم چه‌کار کنیم، از چه استفاده کنیم» (Zucca, 2022; Dennett, 2019). این سخن در نگاه اول ممکن است متناقض به نظر برسد، اما در حقیقت در این جمله

تناقضی وجود ندارد. انسان هوشمند، زمانی که در مواجهه خود با شرایط مختلف نمی‌داند چه‌کار کند، از مغز خود استفاده می‌کند و می‌داند باید از هوشمندی خود بهره ببرد.

اگرچه هیچ شاهدی برای یافتن این جمله در آثار پیاژه در دسترس نیست؛ اما می‌توان آن را به‌نوعی ترجمه‌ای مدرن از سخنان او دانست. پیاژه معتقد بود که هوش یک ویژگی ثابت نیست، بلکه فرآیندی از سازگاری است که به ما امکان می‌دهد دنیای اطراف خود را درک کنیم. همان‌طور که با چالش‌های جدید مواجه می‌شویم، روش‌های جدیدی برای تفکر و درک ایجاد می‌کنیم. این فرآیند همان چیزی است که پیاژه آن را «انطباق» (Accommodation) نامید. وقتی با موقعیتی مواجه می‌شویم که آن را درک نمی‌کنیم، از هوش برای یافتن راه‌حل استفاده می‌کنیم. اینجاست که مفهوم «تعادل» (Equilibration) پیاژه مطرح می‌شود. تعادل فرآیندی است که دانش موجود ما را با اطلاعات جدید متعادل می‌کند. از طریق این فرآیند است که آموختن و رشد در انسان صورت می‌گیرد (Reid, 1979). اگر این مفهوم از هوشمندی را برای حوزه هوش مصنوعی به کار ببریم، به نظر می‌رسد چنین اتفاقی در حال رخ دادن است و یا حداقل برای رسیدن به آن در تلاش هستیم.

بیش از پنجاه سال است که حول امکان‌پذیر بودن یا امکان‌ناپذیر بودن هوش مصنوعی عام[۲] (Artificial General Intelligence) مباحث گوناگونی ارائه شده است (Fortnow, 2021). دسته‌ای از پژوهشگران معتقدند که هوش مصنوعی نورومورفیک[۳] (Neuromorphic artificial intelligence) با توجه به شباهت ساختار آن با مغز انسان می‌تواند سرانجام امکان رسیدن به هوش مصنوعی عام را پدید آورد (Huang, 2017; Pei et al,. 2019; Pontes-Filho & Nichele, 2019; Pontes-Filho et al,. 2022). البته در طرف مقابل از دیرزمان برخی معتقدند که امکان دستیابی به هوش مصنوعی عام وجود ندارد (Searle, 1998; Landgrebe & Smith, 2021). اگر هوش مصنوعی در معنای عام را بتوانیم با سیستم‌های نورومورفیک به‌دست آوریم، یکی از مهم‌ترین مباحثی که با آن روبرو می‌شویم «عاملیت» (Agency) است. در خصوص عاملیت در هوش مصنوعی پژوهش‌های زیادی صورت گرفته است. دسته‌ای از پژوهشگران عاملیت را در مباحث اخلاق هوش مصنوعی (Sullins, 2011; Cervantes et al,. 2020; Moor, 2006)، برخی آن را در حوزه معرفت‌شناسی هوش مصنوعی (Alvarado, 2023; Rosenberg, 2023; Carabantes, 2020) و دسته‌ای دیگر آن را در حوزه سیاست هوش مصنوعی (Coeckelbergh, 2023; Puaschunder, 2019) و غیره موردنظر قرار داده‌اند.

در این مقاله موضوع عاملیت را از منظر اخلاقی در هوش مصنوعی عام مدنظر قرار خواهیم داد. البته قابل‌توجه است که ما قصد نداریم ممکن بودن یا نبودن هوش مصنوعی عام را اثبات کنیم، بلکه می‌خواهیم بدانیم اگر هوش مصنوعی عام زمانی امکان‌پذیر شود؛ با چه نوع عاملیت اخلاقی در این سطح روبرو خواهیم بود. مور (۲۰۰۶)، در یک تقسیم‌بندی در خصوصِ عاملیتِ اخلاقیِ هوش مصنوعی، چهار دسته‌بندی از عاملیت را متصور می‌شود: ۱- عوامل تأثیرِ اخلاقی (Ethical-impact agents)، ۲- عوامل تأثیر اخلاقی ضمنی (Implicit ethical agents)، ۳- عوامل تأثیر اخلاقی آشکار (Explicit ethical agents) و ۴- عوامل اخلاقی کامل (Full ethical agents). این دسته‌بندی بارها توسط پژوهش‌گران مورداستفاده قرار گرفته است (Muller, 2021; Wallach & Allen, 2008; Coeckelbergh, 2022;). بر اساس این دسته‌بندی در بالاترین سطح از عاملیت، هوش مصنوعی عاملیت اخلاقی کامل (بر اساس تصویری انسان‌انگارانه) را داراست. پس همان‌طور که مطرح شد، هدف از این مقاله بررسی این تقسیم‌بندی از عاملیت با توجه به هوش مصنوعی عام است. در این مسیر ما به‌دور از تفسیر انسان‌انگارانه و با نقد به آن به بررسی عاملیت اخلاقی این سیستم‌ها می‌پردازیم.

با توجه به هدف یادشده، مقاله حاضر به چهار بخش تقسیم‌بندی شده است. در بخش ابتدایی هوش مصنوعی عام و سیستم‌های نورومورفیک معرفی خواهد شد. در بخش دوم دسته‌بندی عوامل اخلاقی هوش مصنوعی موردنظر قرار خواهد گرفت. در ادامه و در بخش سوم خطرات هوش مصنوعی عام بررسی می‌شود. در بخش انتهایی نیز بحث و بررسی پیرامون دسته‌بندی عامل اخلاقی هوش مصنوعی صورت می‌گیرد.

۲. هوش مصنوعی عام

پیش از هر چیز نیاز است بدانیم مقصود از هوش مصنوعی عام چیست؟ اصطلاح «هوش مصنوعی ضعیف» برای اشاره به سیستم‌هایی استفاده می‌شود که رفتارهای «هوشمند» مشخصی را در زمینه‌ای مشخص انجام می‌دهند (Kurzweil, 2005). مفهوم «هوش مصنوعی عام[۴]» در برابر هوش مصنوعی ضعیف، به سیستم‌هایی با قابلیت تعمیم هوشمندی اشاره دارد (Goertzel, 2014: 2). هوش مصنوعی عام در تلاش برای ایجاد و مطالعه هوش مصنوعی با دامنه گسترده‌ای از هوشمندی (در سطح انسانی) و قابلیت تعمیم بالا است (Goertzel, 2014: 3). چنین سطحی از هوش مصنوعی می‌تواند قابلیت‌های فراوانی داشته باشد و از محدودیت‌های هوش مصنوعی

ضعیف رها شود. تلاش‌های مختلفی برای طراحی سیستم‌های هوش مصنوعی صورت گرفته است و یکی از جدیدترین این سیستم‌ها، هوش مصنوعی نورومورفیک است.

هوش مصنوعی نورومورفیک با تکیه و الهام از ساختار و عملکرد مغز انسان ساخته شده است. این سیستم‌ها به‌گونه‌ای طراحی شده‌اند که عملکردی مشابه با مغز انسان در یادگیری داشته باشند. با تکیه‌بر سخت‌افزاری با ساختاری مشابه مغز انسان و عملکردی مشابه شبکه‌های عصبی، این سیستم‌ها عملکردی قدرتمندتر و کارآمدتر از سایر سیستم‌های هوش مصنوعی از خود نشان می‌دهند[۵]. به‌عبارتی‌دیگر، برخلاف شیوه‌های قدیمی‌تر در هوش مصنوعی که بر پایه شبکه‌های عصبی عمیق[۶] (Deep Neural Network) در سیستم‌های کامپیوتری دیجیتال ساخته می‌شدند و از این مشکل رنج می‌بردند که سخت‌افزار آن‌ها به‌مانند شبکه‌های عصبی نبود؛ در سیستم‌های نورومورفیک، این ادعا وجود دارد که سخت‌افزار نیز بر پایه ساختار مغز انسان و ارتباط شبکه‌های عصبی ساخته می‌شود (Ivanov et al,. 2022). این موضوع سبب شده، برای اولین بار صحبت از سیستم هوش مصنوعی رود که از نظر سخت‌افزاری و نرم‌افزاری مشابه با مغز انسان است. همین امر موجب شده که بسیاری از پژوهشگران رسیدن به هوش مصنوعی عام را با سیستم‌های نورومورفیک امکان‌پذیر بدانند (Huang, 2017; Pei et al,. 2019; Pontes-Filho & Nichele, 2019; Pontes-Filho et al,. 2022). در چنین شرایطی که ادعا می‌شود هوش مصنوعی عام در حال شکل‌گیری است، موضوع عاملیت اخلاقی بسیار مهم خواهد بود. ممکن است این‌گونه استدلال شود که شناخت ما از مغز انسان بسیار محدود است و نمی‌توان ادعا کرد که سیستم‌های نورومورفیک، شباهتی به مغز انسان دارند یا خیر. اما صرف‌نظر از اینکه سیستم‌های نورومورفیک بتوانند به هوش مصنوعی عام بدل شوند یا خیر، می‌توان موضوع عاملیت اخلاقی هوش مصنوعی عام را موردنظر قرار داد. لذا ما نیز در ادامه، بر روی هوش مصنوعی عام تمرکز خواهیم کرد و بررسی‌ها را وابسته به سیستم نورومورفیک نمی‌کنیم.

۳. چهار دسته‌بندی عاملیت اخلاقی هوش مصنوعی

همان‌طور که پیش‌تر ذکر شد، جیمز مور در یک طبقه‌بندی که امروزه موردتوجه بسیاری از پژوهشگران حوزه اخلاق هوش مصنوعی است، چهار شکل عمده از عاملیت را از یکدیگر تمییز می‌دهد: ۱- عامل تأثیر اخلاقی، ۲- عامل تأثیر اخلاقی ضمنی، ۳- عامل تأثیر اخلاقی آشکار و ۴- عامل اخلاقی کامل.

در این تقسیم‌بندی، عامل تأثیرگذار اخلاقی، ماشینی است که تأثیرگذاری اخلاقی دارد و این تأثیرگذاری ممکن است با التفات (Intention) یا بدون التفات سازنده سیستم هوش مصنوعی در طراحی ماشین رخ دهد. به سخنی دیگر، در این سیستم‌ها این کاربرد ماشین توسط عامل انسانی است که تأثیرگذاری اخلاقی را مشخص می‌کند. چنین ماشینی به‌مانند یک ابزار در دست عامل انسانی است که می‌تواند از آن برای امری اخلاقی یا غیراخلاقی استفاده کند (Moor, 2006: 19). در حقیقت مور اینجا ماشین را ازاین‌جهت تحت عنوان یک عامل معرفی می‌کند که ابزار عمل اخلاقی برای انسان است.

حال زمانی که می‌خواهیم اصول اخلاقی را در کارکرد ماشین بگنجانیم و آن را از انجام اعمال غیراخلاقی باز بداریم، یک شیوه این است که در برنامه‌نویسی، کدهای اخلاقی را به‌صورت ضمنی قرار دهیم. به‌عنوان‌مثال، یک برنامه نرم‌افزاری که برای کشف تقلب طراحی می‌شود، ممکن است به‌گونه‌ای عمل کند که مجموعه‌ای از اصول اخلاقی مانند انصاف و عدم تبعیض در کدهای آن قرار بگیرد. چنین ماشینی در تعبیر مور، عامل اخلاقی ضمنی نامیده می‌شود که شناختی از عمل اخلاقی یا غیراخلاقی ندارد؛ در حقیقت بر اساس طراحی ماشین نمی‌تواند عمل غیراخلاقی از پیش مشخص شده را مرتکب شود (Moor, 2006: 19) یا احتمال چنین عملی کاهش می‌یابد[۷].

در سومین دسته عوامل اخلاقی آشکار، به ماشین‌هایی اطلاق می‌شود که از اصول اخلاقی مشخصی پیروی می‌کنند و برای یافتن این اصول برنامه‌ریزی شده‌اند (Moor, 2006: 10-20). به‌عنوان‌مثال، سیستمی را فرض کنید که در یک بیمارستان به‌کار گرفته می‌شود. در چنین سیستمی، نیاز است بر اساس وضعیت هر بیمار، سیستم تشخیص دهد هر بخش از اطلاعاتِ بیمار در اختیار فرد یا افراد مشخصی قرار بگیرد. اگر می‌خواستیم این ماشین را به شکل عامل اخلاقی ضمنی طراحی کنیم، باید ابتدا تمامی حالات ممکن را پیش‌بینی می‌کردیم و در برنامه‌نویسی می‌گنجاندیم. ازآنجایی‌که حالت‌های ممکن بر اساس شرایط مختلف بیماران مختلف، شامل احتمالات بسیار زیادی می‌شود، برنامه‌نویسی برای تمامی احتمال‌ها، امری بیهوده و بسیار دشوار به نظر می‌رسد. در چنین شرایطی، یک راه‌کار این است که برای ارزش‌های اخلاقی، مقادیر مشخص کمی تعیین کنیم و ماشین بر اساس محاسبه، بهترین تصمیم را بر اساس مقدار کمی اتخاذ کند. به عبارتی ماشین محاسبه کند که کدام عمل از نظر اخلاقی عملکردِ کمیِ بهتری دارد.

اما زمانی که از اخلاق سخن می‌گوییم تکیه‌بر مقادیر کمی، رضایت‌بخش نیست. عامل اخلاقی کامل (دسته چهارم عاملیت اخلاقی) به ماشین‌هایی گفته می‌شود که قادر به تصمیم‌گیری اخلاقی هستند. به‌عبارتی‌دیگر ماشین‌هایی که بتوانند تصمیمات اخلاقی مختلفی گرفته و این تصمیمات را به‌مانند یک انسان توجیه کنند (Moor, 2006: 20). به‌صورت خلاصه تقسیم‌بندی مور نشان می‌دهد در دو دسته اول، عاملیت انسانی نقش اساسی را ایفا می‌کند. در دسته سوم نقش ماشین در نتایج اخلاقی پررنگ‌تر خواهد بود و در دسته چهارم هوش مصنوعی نقش اصلی را ایفا می‌کند.

۱.۳ بررسی دسته‌بندی عاملیت اخلاقی

مطابق با دسته‌بندی مور، در سه دسته اول، ماشین در سطح ابزار است. به‌عبارتی‌دیگر دسته اول، چگونگی استفاده عامل انسانی از ابزار تعیین کننده است. همچنین در دسته دوم، اگرچه برخی از کدهای اخلاقی به صورت ضمنی در اختیار ماشین قرار گرفته است، اما بااین‌حال ماشین در انتخاب و به‌کارگیری این کدهای اخلاقی انتخاب ندارد، بلکه بر اساس ساختار خود عمل می‌کند. به‌عنوان‌مثال طراحی بد و یا آموزش با داده‌های سوگیرانه (Biased Data) می‌تواند منجر به اعمال و خروجی‌های غیراخلاقی ماشین شود.

ممکن است این‌طور تصور کنیم که در سیستم‌های عامل اخلاقی آشکار، پررنگ‌تر شدن نقش ماشین، نقش و مسئولیت عامل انسان را کاهش می‌دهد. در این خصوص باید توجه داشت که حتی در این سیستم‌ها نیز این عوامل انسان هستند که ماشین را طراحی، برنامه‌نویسی یا از آن استفاده می‌کنند و ضعف‌های طراحی، ارزش‌گذاری، الگوریتم، انتخاب داده‌های آموزشی و شیوه استفاده، عملکرد ماشین را مشخص می‌کنند. به‌عبارتی‌دیگر، خروجی و عمل ماشین از محاسبات و ارزش‌گذاری کمی عامل انسانی در طراحی تعیین می‌شود[۸].

در دسته چهارم، مور عاملیت هوش مصنوعی را در سطح یک عامل انسانی بالغ قرار می‌دهد. برای این منظور او یادگیری بازی شطرنج توسط کامپیوتر را مثال می‌زند. مور استدلال می‌کند که کامپیوترها، بازی شطرنج را از روی بازی شطرنج‌باز انسانی یاد گرفته‌اند و لذا می‌توان با یک طراحی مشابه آن‌ها را قادر ساخت تا اخلاق را روی عملکرد اخلاقی عوامل انسانی بیاموزند (Moor, 2006: 20-21).

یک سیستم هوش مصنوعی که عامل اخلاقی کامل باشد، می‌تواند نگرانی‌های فراوانی به وجود آورد. چنین سطحی از هوشمندی در ماشین که آن را به سطح یک عامل اخلاقی برساند،

ازنظر برخی از پژوهشگران چندان دور از دسترس نیست (Muller and Bostrom, 2016). در این شرایط، بهتر است بدانیم اگر هوش مصنوعی عام پدید آید، با چه نوع استدلال‌هایی در خصوص خطرات آن برای انسان روبرو می‌شویم.

۴. خطرات هوش مصنوعی عام

اگر هوش مصنوعی، بتواند به سطح هوشمندی انسان برسد (هوش مصنوعی عام)، از این سطح فراتر رفته و هوش مصنوعی فوق هوشمند (Superintelligence Artificial Intelligence) امکان‌پذیر می‌شود. گود این موضوع را این‌گونه بیان کرده است: «بگذارید ماشین فوق هوشمند به‌عنوان ماشینی تعریف شود که می‌تواند از تمام فعالیت‌های فکری هر انسانی هرچند باهوش پیشی بگیرد. ازآنجایی‌که طراحی ماشین‌ها یکی از این فعالیت‌های فکری است، یک ماشین فوق هوشمند می‌تواند ماشین‌های حتی بهتر از ماشین‌های ما طراحی کند. در این صورت بدون شک یک انفجار اطلاعاتی رخ خواهد داد و هوش انسان بسیار عقب‌تر از هوش مصنوعی می‌ماند؛ بنابراین اولین ماشین فوق هوشمند، آخرین اختراعی است که بشر انجام می‌دهد» (Good، ۱۹۶۵: ۳۳). در چنین شرایطی، احتمالاً باید یک نگرانی ما این باشد که با رسیدن به چنین مرحله‌ای تصمیم‌گیری در خصوص هوش مصنوعی و موضوعات اخلاقی آن دیگر به‌صورت کامل از اختیار انسان خارج می‌شود.

در این راستا ادعاهای دیگری در مورد رسیدن هوش مصنوعی به سطح ابر بشری بیان شده است. مولر و کنون ازجمله کسانی بودند که در مقاله خود استدلال‌های مطرح شده در خصوص خطر وجودی (Existential Risk) هوش مصنوعی فوق هوشمند (رک: Kurzweil, 2005; Bostrom, 2014; Russell, 2019) را موردبررسی قرار داده‌اند. این دو معتقدند این استدلال‌ها بر پایه دو مقدمه قرار دارند: قضیه متعامد (Orthogonality) و قضیه تکینگی (Singularity) (Muller & Cannon, 2022).

بر اساس قضیه تکینگی، هوش مصنوعی فوق هوشمند یک چشم‌انداز واقع‌بینانه و خارج از کنترل انسان است (Muller & Cannon, 2022: 26). برای رسیدن به نتیجه‌گیری خطر وجودی، ادعای تکینگی نیاز به یک فرض دیگر دارد. بوستروم به‌عنوان یکی از طرفداران نظریه خطر وجودی هوش مصنوعی فوق هوشمند برای انسان، قضیه متعامد را برای این منظور مطرح می‌کند. بر اساس این قضیه می‌توان گفت هر سطح از هوش می‌تواند برای هر هدفی به کار گرفته شود (Bostrom, 2017: 107). بوستروم معتقد است قضیه متعامد نشان می‌دهد که هوش

مصنوعی می‌توانند اهداف کاملاً غیرانسانی داشته باشد و درنتیجه خطرات وجودی انسان را تهدید خواهد کرد (Bostrom, 2012). مولر و کنون در برابر این استدلال‌های خطر وجود هوش مصنوعی عام برای انسان را این‌گونه صورت‌بندی می‌کنند:

مقدمه اول (قضیه تکینگی): هوش مصنوعی می‌تواند به مرحله فوق هوشمند برسد و کنترل از دست انسان خارج شود.

مقدمه دوم (قضیه متعامد): هر سطح از هوش می‌تواند برای هر هدفی استفاده شود.

نتیجه: هوش مصنوعی فوق هوشمند، خطر وجودی برای بشریت به همراه دارد.

مولر و کنون یک خلط مفهومی را در این استدلال ترسیم می‌کنند که طبق آن خطر وجودی هوش مصنوعی فوق هوشمند رد خواهد شد. ازنظر آن‌ها در این استدلال، مقدمات به دو مفهوم از هوشمندی نیاز دارد، درحالی‌که اعتبار مستلزم یک مفهوم است. به عقیده این دو متفکر با توجه به مقدمه اول (قضیه تکینگی)، هوش مصنوعی فوق هوشمند آخرین ابزاری است که انسان‌ها می‌سازند و پس‌ازآن ساخت هوش مصنوعی بر عهده هوش مصنوعی خواهد بود. به‌عبارتی‌دیگر، هوش مصنوعی در حد هوشمندی انسان یا بالاتر از انسان قرار دارد و برای ساخت هوش مصنوعی دیگر نیازی به انسان نیست. در اینجا مفهوم «هوش» به معنای «عام» مطرح است، یعنی برابر یا بیشتر از هوش انسان. بر اساس قضیه متعامد، هوش مصنوعی به‌منزلهٔ «ابزار» است که می‌توان از آن برای رسیدن به هر هدفی استفاده کرد؛ بنابراین هوش مصنوعی در این جایگاه تحت کنترل انسان است و در سطح یک ابزار باقی می‌ماند و نمی‌توان آن را به‌عنوان هوش مصنوعی عام تعبیر کرد. مولر و کنون معتقدند که این دو مقدمه را نمی‌توان با هم در جهت رسیدن به نتیجه در نظر گرفت زیرا مفاهیم متفاوتی از هوشمندی در آن‌ها موردنظر قرار گرفته است و لذا استدلال خطر وجودی هوش مصنوعی فوق هوشمند رد می‌شود (Muller & Cannon, 2022: 30-34). البته در بررسی استدلال مولر[۹] و کنون باید در نظر داشت که آن‌ها قضیه تکینگی یا متعامد را رد نمی‌کنند، بلکه نتیجه‌گیری خطر وجودی برای انسان را بر اساس استدلالی که این دو قضیه را به‌عنوان مقدمات در خود دارد را رد می‌کنند؛ اما آیا با رد این استدلال، خطر هوش مصنوعی عام برای انسان از میان خواهد رفت؟ برای این منظور توجه به قضیه متعامد می‌تواند روشنگر باشد و پس‌ازآن به نظری مجدد به استدلال مولر و کنون در بخش بحث و بررسی می‌اندازیم.

۱.۴ قضیهٔ متعامد

بوستروم قضیه متعامد را به این صورت بیان می‌کند که هوش و اهداف نهایی، محورهای متعامدی هستند که در امتداد آن‌ها عواملِ ممکن، می‌توانند آزادانه تغییر کنند. به‌عبارت‌دیگر، در اصل، هر سطحی از هوش می‌تواند با هر هدف نهایی ترکیب شود. او برای تأیید بر این قضیه سه مسیر را پی می‌گیرد (Bostrom, 2012).

مسیر اول: دیوید هیوم معتقد بود که باور (Belief) به‌تنهایی نمی‌تواند انگیزه (Motivation) برای عمل را برانگیزد و برای عمل میل (Desire) هم لازم است. یک انتقاد به قضیه متعامد می‌تواند این باشد که ممکن است هوش به دستیابی به باور منجر شود و باور ضرورتاً به ایجاد و برانگیخته شدن انگیزه منتج می‌شود، اما نظریه هیوم این انتقادات احتمالی به قضیه متعامد را تضعیف کرده و نشان می‌دهد باور و انگیزه از هم جدا هستند.

مسیر دوم: نباید این‌گونه استنباط شود که قضیه متعامد، نظریه انگیزشِ هیوم را پیش‌فرض خود قرار می‌دهد. نیازی نیست معتقد باشیم که باورها به‌تنهایی هرگز نمی‌توانند انگیزه برای عمل ایجاد کنند. کافی است فرض کنیم یک عامل، اگر میل و قدرت کافی در انجام عمل داشته باشد، می‌تواند برای دنبال کردن هر عملی دارای انگیزه باشد. حتی با فرض نادرستی نظریه هیوم، هوش بالا مستلزم کسب باورهایی نیست که به‌خودی‌خود محرک عمل باشند.

مسیر سوم: امکان ساختن یک سیستم شناختی با هوش بالا با ساختاری بی‌شباهت با عملکرد «باور» و «میل» انسان وجود دارد. اگر قادر باشیم سیستمی بسازیم که بتواند برای دنبال کردن هر هدفی انگیزه داشته باشد این امر امکان‌پذیر خواهد بود (Bostrom, 2012: 73).

لذا بر اساس این قضیه، هر سطحی از هوش می‌تواند با هر هدف نهایی ترکیب شود. از نظر بوستروم، بر اساس این سه راه، هیچ ارتباط ضروری میان سطح هوش یک عامل هوش مصنوعی و اهداف نهایی آن وجود ندارد. یک هوش مصنوعی می‌تواند بسیار هوشمند باشد اما اهداف آن کاملاً متفاوت با اهداف انسان یا حتی اهدافی مضر برای انسان باشد. سخنان بوستروم در حقیقت بر این ایده استوار است که هوش مربوط به توانایی استدلال و تصمیم‌گیری است، درحالی‌که اهداف نهایی مربوط به آنچه عامل می‌خواهد به دست آورد است. یک عامل، صرف‌نظر از سطحِ هوشِ خود، می‌تواند از توانایی‌های خود در استدلال برای پیگیری طیف گسترده‌ای از اهداف گوناگون استفاده کند (Bostrom, 2012). در برابر استدلال بوستروم و با توجه به نقد مولر و کنون، مشاهده می‌کنیم قضیه متعامد همچنان پابرجا بوده و رد نمی‌شود.

۵. بحث و بررسی

اکنون به نظر می‌رسد همه‌چیز را برای بررسی هدف اصلی این مقاله در دست داریم. درکی از هوش مصنوعی عام، دسته‌بندی عاملیت اخلاقی هوش مصنوعی و خطرات هوش مصنوعی عام همگی مشخص شده‌اند. لذا زمان بررسی دسته‌بندی مور فرا رسیده است، اما پیش از آن توجه به عاملیت با توجه به خطرات مطرح شده می‌تواند مفید واقع شود. مولر و کنون نشان دادند که از دو قضیه تکینگی و متعامد نمی‌توان به استدلال خطر وجودی برای انسان رسید؛ بااین‌حال هنوز موضوع عاملیت از دو منظر قابل‌توجه است:

۱. **خطرات ابزار در دست عامل انسانی:** مولر و کنون خطرات هوش مصنوعی زمانی که عاملیت در اختیار عامل انسانی است را رد نمی‌کنند. به بیانی دیگر، عامل انسانی می‌تواند از «ابزاری» که توانایی‌های فراوانی دارد برای مقاصد گوناگون استفاده کند. استفاده نادرست از چنین ابزاری می‌تواند برای انسان‌ها خطرات گوناگونی به وجود آورد. در این خصوص کدهای اخلاقی و سیاست‌گذاری‌های فراوانی وجود دارد، اما در چنین شرایطی یک نگرانی، ایجاد تصویر ناصحیح از عاملیت هوش مصنوعی در سطح ابزار است. به‌عبارتی‌دیگر اگر عامل انسانی سازنده یا استفاده کننده از هوش مصنوعی این تصویر را برای سایر انسان‌ها ایجاد کند که عاملیت با هوش مصنوعی است، می‌تواند از مسئولیت‌های اخلاقی خود فرار کند[۱۰].

۲. **عاملیت هوش مصنوعی عام:** اگرچه نمی‌توان جلوی پیشرفت علم را گرفت و این موضوع مطلوب هم نیست[۱۱]، اما توجه به اخلاق را نمی‌توان از نظر دور داشت. زمانی که از عاملیت هوش مصنوعی سخن می‌گوییم یک دیدگاه این خواهد بود که چنین هوشی برای ما خطرآفرین است، چراکه در کنترل ما نیست. چندان دور از انتظار نخواهد بود که هر عاملی از کنترل شدن توسط عوامل دیگر احساس خطر کند و از خود در برابر چنین شرایطی محافظت کند. لذا صرف‌نظر از استدلال بوستروم، ما باید در خصوص ساخت هوش مصنوعی دارای عاملیت بسیار محتاط باشیم. برای روشن شدن این موضوع توجه به یک مثال مفید خواهد بود:

فرض کنید یک عامل هوش مصنوعی عام (X) می‌داند عامل انسانی (Y) دکمه‌ای برای خاموش کردن در بدن X تعبیه کرده است. برای X خاموش شدن می‌تواند معادل با مردن و از دست دادن عاملیت باشد. لذا احتمالاً اولین چیزی که X تلاش می‌کند یاد بگیرد این است که چطور آن دکمه را غیرفعال کند و یا راهی بیابد که امکان Y برای فشردن دکمه را از میان ببرد. در چنین شرایطی اگرچه عامل Y برای حفاظت از خود دکمه‌ای در بدن X تعبیه کرده است، اما همین امر منجر شده است که عامل X نیز از سمت عامل Y احساس خطر کرده و برای حفاظت

از خود نیازمند مقابله یا کنترل Y و یا از میان بردن کنترل Y خواهد بود. با توجه به مبحثی که در خصوص افزایش توانایی‌هایی هوش مصنوعی عام طرح شد، لذا X توانایی‌های بیشتری در طول زمان به نسبت Y به دست خواهد آورد. درنتیجه برای Y تقریباً غیرممکن خواهد بود بتواند X را کنترل کند. البته قابل‌توجه است که اگر در ابتدا هم Y هیچ دکمه‌ای برای خاموش کردن یا کنترل X تعبیه نکرده باشد باز Y هیچ کنترلی بر روی X ندارد.

با توجه به این مثال احتمالاً یک نتیجه اولیه این خواهد بود که اگر بخواهیم یک هوش مصنوعی عام تولید کنیم، روی آن کنترلی نخواهیم داشت. اگر به قضیه متعامد رجوع کنیم، می‌دانیم که مطابق با آن میان هوشمندی و پیگیری اهداف انسانی ارتباطی وجود ندارد و به‌عبارتی‌دیگر، ما نمی‌دانیم اهداف هوش مصنوعی عام چه خواهند بود. در این شرایط ممکن است عامل فرضی X هر هدف دیگری نیز به‌جز مواردی که مطرح شد را برای خود انتخاب کند. در اینجا به نظر می‌رسد ساخت هوش مصنوعی عام نمی‌تواند چندان به سود انسان باشد. چراکه اگر این عامل بخواهد به‌مانند یک عامل انسانی از خود محافظت کند، احتمالاً کمر به نابودی انسان خواهد بست.

حال اگر به نقد کنون و مولر بازگردیم، این دو استدلال خطر وجودی هوش مصنوعی عام را به این دلیل رد کردند که هوش مصنوعی به‌عنوان یک ابزار و هوش مصنوعی عام دو مفهوم مختلف را مراد می‌کنند. در اینجا می‌توان یک ایراد به نقد کنون و مولر وارد کرد. آن‌ها این جنبه را در نظر نگرفته‌اند که اگر هوش مصنوعی عام امکان‌پذیر باشد، این هوش عام می‌تواند برای اهداف خود ابزار بسازد و از آن ابزار استفاده کند. در حقیقت اگر هوش مصنوعی عام یک عامل دارای هدف باشد، توانایی استفاده/تولید ابزار برای هدف/اهداف خود را دارد. به عبارتی می‌توان گفت چنین عاملی هم می‌تواند از ابزار استفاده کند و هم اینکه چیستی اهداف او برای عامل انسانی ناشناخته و متفاوت است. لذا به نظر می‌رسد مجدداً استدلال طرفداران خطر وجودی هوش مصنوعی فوق هوشمند معتبر می‌شود. در چنین شرایطی کدهای اخلاقی که برای سطوح هوشمندی کمتر استفاده می‌شود، چندان مفید نخواهند بود. به‌عنوان‌مثال شفافیت، نظارت انسانی و غیره (Muller, 2021) تنها در صورتی مفید واقع می‌شوند که در راستای اهداف (ناشناخته) هوش مصنوعی عام باشد. چراکه در حقیقت دلیلی ندارد هوش مصنوعی عام از کدهای اخلاقی‌ای تبعیت کند که در راستای اهدافش نیست (Yampolskiy, 2022).

اکنون برای بررسی نهایی اگر به دسته‌بندی مور رجوع کنیم، مشخص است که هوش مصنوعی عام را نمی‌توان در سه دسته ابتدایی قرار داد. این موضوع را از این منظر هم می‌توان

مطرح کرد که ساختار طراحی و سخت‌افزار سیستم‌هایی که بتوانند ما را در مسیر هوش مصنوعی عام قرار دهند با سه دسته اولی که مور موردنظر قرار می‌دهد، شباهتی ندارند. لذا احتمالاً به نظر می‌رسد هوش مصنوعی عام را باید در دسته‌بندی چهارم قرار دهیم؛ اما در اینجا با توجه به مباحث پیشین و قضیه متعامد با مشکلاتی روبرو خواهیم شد.

بر اساس قضیه متعامد:

۱. هوش مصنوعی عام می‌تواند اهداف مختلفی داشته باشد.

۲. اهداف هوش مصنوعی عام ممکن است با متفاوت/بر ضد اهداف انسان باشد.

اما بر اساس دسته‌بندی مور:

۱. هوش مصنوعی می‌تواند اخلاق را از انسان یاد بگیرید.

۲. هوش مصنوعی به‌مانند یک انسان بالغ عاملیت اخلاقی خواهد داشت.

در اینجا مشخص است که در دسته‌بندی مور یادگیری و عاملیت اخلاقی به معنای دنبال کردن اهداف اخلاقی انسانی است. درصورتی‌که مطابق با قضیه متعامد، اهداف هوش مصنوعی عام می‌تواند با اهداف انسان متفاوت باشد/برضد اهداف انسان باشد. البته می‌توان یک ایراد دیگر هم به مورد گرفت. در حال حاضر سیستم‌های یادگیری ماشین (Machine Learning) فراوانی وجود دارند که عملی مشابه با این دسته‌بندی مور انجام می‌دهند اما در سطح هوش مصنوعی عام قرار نمی‌گیرند[۱۲]. در این سیستم‌ها، اصول اخلاقی و تصمیمات اخلاقی انسان منابع یادگیری ماشین هستند و ماشین می‌تواند در موضوعات گوناگونی تصمیمات اخلاقی بگیرد. باید دقت داشت که اگرچه این ماشین‌ها می‌توانند با استفاده از الگوهای دریافت کرده، تصمیمات اخلاقی بگیرند اما چندین نکته قابل‌تأمل وجود دارد[۱۳]: ۱- این ماشین‌ها در بخش محدودی قادر به تصمیم‌گیری هستند و نمی‌توان برای همه موضوعات و چالش‌های اخلاقی از آن‌ها استفاده کرد؛ ۲- نظارت عامل انسانی در تصمیمات ماشین در استفاده از آن‌ها توصیه می‌شود چراکه تصمیمات سیستم دارای اشتباه است (Varlamov et al,. 2019: 871)؛ ۳- اتفاق‌نظر در خصوص اینکه ارزش‌های اخلاقی جهان‌شمول هستند وجود ندارد و با توجه به اینکه در این ماشین‌ها همواره از ارزش‌های اخلاقی مشخصی استفاده می‌شود، ممکن است برای برخی، نژادها، جوامع و فرهنگ‌ها تصمیمات ماشین غیراخلاقی باشد (Munn, 2023). با توجه به این نکات و محدودیت‌ها این نوع ماشین‌ها را نمی‌توان یک عامل اخلاقی کامل دانست و این دسته از ابزارها نیز به‌نوعی در دسته‌بندی سوم مور (عامل تأثیر اخلاقی آشکار) می‌گنجند.

حال که مشخص شد چهارمین دسته عاملیت هوش مصنوعی مور را نمی‌توان با هوش مصنوعی عام تطبیق داد، پیشنهاد چه خواهد بود؟ با توجه به مباحث مطرح شده، مشخص شد که این توصیف مور از دسته‌بندی چهارم بود که دارای اشکال است. به عبارتی نمی‌توان از یک هوش مصنوعی عام انتظار داشت اهداف اخلاقی انسان را دنبال کند[۱۴]، پس چنین سطحی از هوش مصنوعی را در چه دسته‌بندی قرار دهیم؟ بر اساس آنچه از قضیه متعامد به دست آمد، عاملیت هوش مصنوعی رد نمی‌شود، بلکه کنترل از دست عامل انسانی خارج می‌شود و اهداف اصلی عامل مصنوعی ممکن است بسیار متفاوت و یا حتی بر ضد انسان باشد. به‌عبارتی‌دیگر اگر بخواهیم شکل چهارم عاملیت اخلاقی هوش مصنوعی را مجدداً صورت‌بندی کنیم، باید بگوییم:

۱. عامل اخلاقی هوش مصنوعی عام می‌تواند اصول اخلاقی انسان را یاد بگیرد و یا اصول اخلاقی دیگری استنتاج کند و نباید تصور کرد که اصول اخلاقی انسان را اصول اخلاقی خود بداند.

۲. عامل اخلاقی هوش مصنوعی عام می‌تواند اهداف و اصول اخلاقی خود را شکل دهد و این اصول ممکن است با اصول اخلاقی انسان‌ها متفاوت باشد/بر ضد اصول اخلاقی انسان‌ها باشد.

۶. نتیجه‌گیری

نتایج مقاله حاضر نشان می‌دهد دسته‌بندی عاملیت اخلاقی مور برای عامل تأثیر اخلاقی، عامل تأثیر اخلاقی ضمنی، عامل تأثیر اخلاقی آشکار پاسخ‌گو است؛ اما توصیف او از عامل اخلاقی کامل دارای ایراد است. بر اساس قضیه متعامد مشخص شد که عامل اخلاقی کامل اگرچه عاملیت دارد و ممکن است مفهومی از اخلاق را درک کند، اما لزوماً این اخلاق در راستای اخلاق انسانی نبوده و ممکن است اهدافی متفاوت با اهداف اخلاقی انسان را دنبال کند. نتایج بررسی نشان می‌دهد که دسته‌بندی مور در شکل چهارم با این ایراد روبرو است که هوش مصنوعی عام را مشابه با عامل انسانی می‌داند و انتظارات عمل انسانی را از هوش مصنوعی دارد. به سخنی دیگر اگرچه عاملیت را برای هوش مصنوعی قائل می‌شود، اما عاملیت را با انسان‌انگاری هوش مصنوعی به یک مفهوم در نظر می‌گیرد. بررسی‌ها نشان می‌دهد اگر هوش مصنوعی عام را امکان‌پذیر بدانیم، علاوه بر سه دسته اول، نیازمند دسته چهارمی هستیم که عاملیت اخلاقی هوش مصنوعی عام را در آن دسته قرار دهیم. البته دو نکته قابل‌توجه است:

۱- مباحث مطرح شده امکان‌پذیر بودن هوش مصنوعی عام را تأیید نمی‌کند و فرض بر این بوده که اگر امکان‌پذیر باشد با چه نوع عامل اخلاقی روبرو هستیم و ۲- اگر زمانی هوش مصنوعی عام، امکان‌پذیر باشد، عاملیت اخلاقی آن عاملیت اخلاقی انسان متفاوت بوده و باید خطر ضدیت آن با اخلاق انسانی و یا خطرات ناشی از آن را جدی بگیریم.

با در نظر گرفتن این موارد، شکل جدید صورت‌بندی دسته چهارم عاملیت اخلاقی که متناسب با مفهوم هوش مصنوعی عام باشد به این‌صورت خواهد بود: اول آنکه این عامل اخلاقی می‌تواند اصول اخلاقی انسان را یاد بگیرد و یا اصول اخلاقی دیگری استنتاج کند؛ و نباید تصور کرد که اصول اخلاقی انسان را اصول اخلاقی خود بداند؛ دوم، این عامل می‌تواند اهداف و اصول اخلاقی خود را شکل داده و این اصول می‌تواند متفاوت/بر ضد اصول اخلاقی انسان‌ها باشد.

شایان توجه است که نتایج مقاله حاضر را باید با توجه به محدودیت‌های آن در نظر داشت. با توجه به اینکه هوش مصنوعی عام، در زمان نگارش این مقاله تحقق نیافته، سنجشِ این نتایج با مطالعه تجربی امکان‌پذیر نیست. به علاوه پرداختن به مباحث اخلاقی هوش مصنوعی عام نیازمند رویکردی میان رشته‌ای است که حوزه‌های مختلفی مانند فلسفه، اخلاق، علوم کامپیوتر، حقوق، روانشناسی و جامعه‌شناسی را در بر می‌گیرد. توجه و اهمیت به این حوزه‌ها در پژوهش، با توجه به توسعه سریع تکنولوژی از دیگر محدودیت‌های مقاله حاضر است. با در نظر گرفتن این محدودیت‌ها، پرسش‌های فراوانی پیش روی ما قرار دارد. به‌عنوان مثال، آیا توسعه هوش مصنوعی عام با عاملیت اخلاقی کامل، از نظر اخلاقی مجاز است؟ آیا هوش مصنوعی عام می‌تواند در برابر تصمیم‌های خود مسئول شناخته شود؟ و یا چه مسئولیت‌های اخلاقی بر عهده توسعه‌دهندگان این سیستم‌ها است؟ چطور می‌توان اطمینان حاصل کرد که توسعه هوش مصنوعی عام می‌تواند به بقا و رفاه انسان‌ها کمک کند؟ با توجه به اهمیت این پرسش‌ها، شایسته است پژوهش‌های متعددی در این زمینه از سوی پژوهش‌گران صورت بگیرد.

پی‌نوشت‌ها

۱. البته مقصود این نیست که آن‌چه امروز تحت عنوان هوش مصنوعی، در دست داریم، تنها ناشی از تلاش‌ها و اندیشه‌های فیلسوفان است؛ بلکه مقصود تاکید براین امر است که هوش مصنوعی از منظر فلسفی، موضوعی بسیار جدی و با سابقه‌ای طولانی است. برای دستیابی به هوش مصنوعی، اگرچه

اندیشه فیلسوفان در نوع خود قابل توجه است، اما پیشرفت و زمینه‌سازی علوم مختلف را نمی‌توان و نباید نادیده انگاشت. از پیشرفت در علوم اعصاب‌شناختی، فیزیک، الکترونیک، پایه گذاری دانش الگوریتم توسط خوارزمی تا ریاضیات هیلبرت، تلاش‌های لایب‌نیتس و اثرپذیری از او توسط جرج بول هر کدام نقشی بسیار پررنگ در دستیابی به هوش مصنوعی داشته‌اند (Thomas, 2015: 30-35).

۲. به اختصار AGI: به زبان ساده، هوش مصنوعی عام به سیستمی اشاره دارد که هوشمندی آن هم‌سطح و یا نزدیک به هوشمندی یک انسان بالغ باشد. در صورتی که این هوشمندی به سطح هوشمندی انسان در تمامی حوزه‌های عملکرد و شناخت می‌رسد و از آن فراتر رود، به چنین سیستمی هوش مصنوعی فوق هوشمند (Superintelligent Artificial Intelligence) گفته می‌شود (Bostrom, 2014: 22). شایان ذکر است که اصطلاح هوش مصنوعی عام گاهی در برخی پژوهش‌ها تحت عنوان هوش مصنوعی قوی (Strong Artificial Intelligence) نیز مراد می‌شود (Searle, 1998).

۳. به اختصار ANI.

۴. گورتزل در مورد تاریخچه واژه هوش مصنوعی عام می‌گوید در سال ۲۰۰۲ وی و کاسیو پناخین در حال ویرایش کتابی در مورد رویکردهای هوش مصنوعی قوی، با قابلیت‌های گسترده در سطح انسانی و فراتر از آن مشغول انتخاب یک عنوان بوده‌اند و با پیشنهاد شین لگ با «هوش مصنوعی عام» موافقت می‌کنند (Goertzel & Pennachin, 2007). البته وی تصریح می‌کند که بعدها متوجه می‌شود محققی به نام مارک گوبرود از این اصطلاح در سال ۱۹۹۷ در مقاله‌ای در مورد آینده فناوری و خطرات مرتبط استفاده کرده است (Goertzel, 2014: 2).

۵. این سیستم‌ها ویژگی‌های منحصربه‌فردی دارند که آن‌ها را قادر می‌سازد عملکردشان مشابه با عملکرد مغز انسان باشد. اگر بخواهیم دقیق‌تر این موضوع را بیان کنیم، این نوع هوش مصنوعی دارای ۹ ویژگی مشابه با مغز انسان است: ۱- ارتباط‌گرایی (Connectionism): این سیستم‌ها از نورون‌های به هم پیوسته مانند مغز انسان تشکیل شده‌اند. این موضوع به آنها اجازه می‌دهد تا اطلاعات را به صورت موازی و توزیع شده پردازش کنند. ۲- موازی‌سازی (Parallelism): می‌توانند اطلاعات بسیاری از پردازنده‌های مختلف را همزمان پردازش کنند. ۳- ناهمزمانی (Asynchrony): برای کار کردن نیازی به همگام‌سازی تمامی پردازنده‌ها ندارند. این موضوع به آنها اجازه می‌دهد تا عملکردی کارآمدتر و پاسخگوتر به تغییرات محیط خود باشند. ۴- ماهیت ضربه‌ای انتقال اطلاعات (Impulse nature of information transfer): مانند مغز انسان از پالس‌های الکتریسیته برای برقراری ارتباط با یکدیگر استفاده می‌کنند. این ویژگی باعث می‌شود کمتر مستعد خطا باشند. ۵- یادگیری روی دستگاه (On-device learning): می‌توانند مستقیما از محیط و بدون نیاز به دخالت انسان یادگیری را انجام دهند. این ویژگی باعث می‌شود یادگیری سرعت بیشتری داشته باشد. ۶- یادگیری محلی (Local learning): بدون نیاز به دانش عمومی، می‌توانند از محیط محلی خود بیاموزند. این ویژگی باعث می‌شود مقیاس‌پذیرتر باشند. ۷- پراکندگی (Sparsity): از میان اطلاعات پراکنده تنها مرتبط‌ترین اطلاعات را ذخیره می‌کنند. ۸- آنالوگ (Analog): از سیگنال‌های آنالوگ استفاده می‌کنند که نسبت به سیگنال‌های دیجیتال کارآمدتر و کمتر در معرض خطا هستند. محاسبات درون

حافظه (In-memory computing): می‌توانند اطلاعات را در محل حافظه علاوه بر ذخیره‌سازی، پردازش هم بکنند (Ivanov et al,. 2022).

۶. به اختصار DNN.

۷. به عبارتی ممکن است برخی از روش‌های تقلب را از پیش نشناخته باشیم و در کدهای ماشین نباشند.

۸. در برابر این سه دسته اول که شامل بیشتر سیستم‌های هوش مصنوعی می‌شود یک خطر بزرگ این است که برای ماشین عاملیتی با مانند عاملیت اخلاقی انسانی قائل باشیم. این موضوع می‌تواند منجر به سلب مسئولیت از عامل انسانی شود. این امر به صورت ضمنی منجر می‌شود که در برابر چنین سیستم‌هایی آسیب‌پذیر باشیم. برای توضیح بیشتر شاید نمونه روبات‌های چت (chatbot) بسیار روشن کننده باشد. از نمونه‌های شناخته شده و پرکاربرد این روبات‌ها می‌توان به ChatGPT، Google Bard، Perplexity و Bing اشاره کرد. در روبات‌های چت یا چت‌بات‌ها از مدل‌های زبانی بزرگ (Large Language Models) استفاده می‌شود. مدل‌های زبانی بزرگ سیستم‌های محاسباتی پیچیده‌ای هستند که می‌توانند زبان انسان را پردازش و متن‌هایی مشابه با آن تولید کنند. این متن‌ها معمولا به صورت شگفت‌انگیزی شبیه به مکالمات انسان است. این مدل‌ها بر روی مقادیر عظیمی از داده‌های متنی آموزش می‌بینند و از این داده‌ها برای یادگیری الگوها و ارتباطات بین کلمات و عبارات استفاده می‌کنند. به صورت ساده، سه مرحله اصلی برای آموزش یک مدل زبانی بزرگ وجود دارد: ۱- پیش‌آموزش (Pre-training): سیستم روی مجموعه عظیمی از داده‌های متنی آموزش می‌بیند. این داده‌ها می‌تواند شامل کتاب‌ها، مقالات، کد و سایر اشکال متن باشد. مدل در این مرحله فقط می‌توان تا الگوها و ارتباطات موجود در داده‌ها را یاد بگیرد. ۲- تنظیم دقیق (Fine-tuning): در این مرحله سیستم بر روی یک مجموعه کوچکتر از داده‌های متنی برای یک عملکرد خاص تنظیم دقیق می‌شود. به عنوان مثال یک عملکرد خاص می‌تواند شامل وظایفی مانند ترجمه، پاسخ به سوالات یا خلاصه‌نویسی باشد. نتیجه‌گیری (Inference): هنگامی که مدل تنظیم دقیق شد، می‌توان از آن برای تولید متن، ترجمه زبان‌ها، نوشتن انواع مختلف محتوا و پاسخ به سوالات استفاده کرد (Naveed et al,. 2023). با بررسی این سیستم‌ها مشخص می‌شود که آن‌چه در چت‌بات‌ها رخ می‌دهد شبیه‌سازی جملات انسانی است و نه آنکه ماشین به مانند یک انسان جواب سوالات را درک کرده و با التفات برای تصمیم‌گیری اخلاقی به انسان پاسخ دهد (Chomsky et al,. 2023). اگر عوامل انسانی در برابر پاسخ‌هایی که از این نوع هوش مصنوعی به دست می‌آورد فراموش کنند که ماشین تنها عامل اخلاقی آشکار است، آنگاه در معرض خطرات مختلفی قرار خواهد گرفت. به‌عنوان مثال احتمال دستکاری در رفتارهای انسانی با استفاده از پاسخ‌های سوگیرانه افزایش پیدا خواهد کرد.

۹. در خصوص مولر می‌توان گفت که او حتی علاوه بر اینکه ادعای تکینگی را رد نمی‌کند، بلکه در پژوهشی مشترک با بوستروم رسیدن به چنین سطحی از هوشمندی را بسیار نزدیک می‌داند (Muller and Bostrom, 2016).

۱۰. با توجه به اینکه این موضوع، هدف مقاله حاضر نیست از ادامه آن خودداری شده است و صرفا جهت باز شدن مبحث برای پژوهش‌های آتی مطرح می‌شود.

۱۱. هلن لانجینو در بررسی معرفت‌شناسی اجتماعی دانش علمی، این دانش را نیازمند تعامل افراد در جوامع علمی می‌داند. او نشان می‌دهد در روند تولید دانش علمی، روندهای تعاملی نقش تعیین کننده را دارند و جوامع علمی، عوامل کانونی معرفتی هستند. افراد در روند تولید علم به شبکه‌های پیچیده‌ای از ارتباطات وارد می‌شوند و هر کدام برای مدت زمانی در این جوامع حضور دارند و ممکن است افراد دیگری جایگزین آن‌ها شوند. در حقیقت لانجینو نشان می‌دهد شبکه‌های ارتباطی در روند تولید دانش علمی در جوامع علمی بسیار پویا بوده و همواره با یک ساختار ثابت از افراد روبرو نیستیم و رشد علم در این تعامل امری است که به تک افراد و یک شبکه منحصر به فرد متکی نیست (Longino, 2022). می‌توان به صورت ضمنی از سخنان لانجینو اینگونه برداشت کرد که پیشرفت‌های علمی اکنون بسیار پیچیده‌تر از آن هستند که به سادگی بتوان در آن‌ها خللی به وجود آورد. در حال حاضر گروه‌های مختلفی از دانشمندان در رشته‌های مختلف در حال پژوهش هستند که به صورت مستقیم یا غیرمستقیم در پیشرفت دانش هوش مصنوعی تاثیرگذار خواهند بود.

۱۲. در حال حاضر سیستم‌های هوش مصنوعی همچون Delphi موسسه Allen، Moral Machine لابراتوار رسانه دانشگاه MIT، AI Jesus دانشگاه تگزاس و شرکت گوگل، AI Fariness 360 شرکت IBM و یا Gemeni گوگل که در زمان نگارش مقاله در ابزار Bard گوگل استفاده می‌شود از نمونه‌های این نوع هوش مصنوعی هستند.

۱۳. البته محدودیت‌های این ابزارها بیش از موارد ذکر شده است. برای اطلاعات بیشتر ر.ک به (Hagendorff, 2022).

۱۴. ایزاک اسیموف در اثر داستانی خود سه قانون برای ساخت سیستم‌های روباتی که به سطح هوشمندی انسان می‌رسند طرح کرد که تحت عنوان سه قانون روباتیک (Three Laws of Robotics) در پژوهش‌ها از آن یاد می‌شود. بر اساس این قوانین: ۱- روبات اجازه ندارد به انسان آسیب برساند و یا عدم اقدام او اجازه آسیب رساندن به انسان را بدهد. ۲- روبات باید از دستورات انسان اطلاعات کند، مگر در مواردی که دستورات با قانون اول در تضاد باشد. ۳- یک روبات باید از خود محافظت کند تا زمانی که چنین حفاظتی با قانون اول و دوم مغایر باشد (Asimov, 1984). با توجه به مباحث مطرح شده، به نظر می‌رسد سه قانون آسیموف نیز برای هوش مصنوعی عام قابل استفاده نخواهد بود.

کتاب‌نامه

Alvarado, R. (2023). AI as an Epistemic Technology. *Science and Engineering Ethics*, *29*(5), 32.

Asimov, I. (1984). The Bicentennial Man. Philosophy and Science Fiction (Philips, M., ed). New York: Prometheus Books, Buffalo.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. Minds and Machines, 22(2 - special issue ‘Philosophy of AI’ ed. Vincent C. Müller), 71–85.

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

Carabantes, M. (2020). Black-box artificial intelligence: an epistemological and critical analysis. *AI & society*, *35*(2), 309-317.

Chomsky, N., Roberts, I., & Watumull, J. (2023). Noam Chomsky: The False Promise of ChatGPT. The New York Times, 8.

Cervantes, J. A., López, S., Rodríguez, L. F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial moral agents: A survey of the current status. *Science and engineering ethics*, *26*, 501-532.

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. Science and engineering ethics, 26(4), 2051-2068.

Coeckelbergh, M. (2023). Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI and Ethics*, *3*(4), 1341-1350.

Dennett, D. C. (2019). Clever evolution. Metascience, 28(3), 355–358.

Descartes, Rene. (1955). The philosophical works of Descartes, Voume 1 (E.S, Haldane & G.R.T, Ross Trans.). New York: Dover Publications.

Fortnow, L. (2021). Fifty years of P vs. NP and the possibility of the impossible. *Communications of the ACM*, *65*(1), 76-85.

Goertzel, B. (2014). Artificial General Intelligence: Concept, State of the Art, and Future Prospects, Journal of Artificial General Intelligence 5(1) 1-46, DOI: 10.2478/jagi-2014-0001 Accepted 2014-3-15.

Goertzel, B. Pennachin, C. (2007). Artificial General Intelligence. Springer.

Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. In F. L. Alt & M. Ruminoff (Eds.), Advances in computers (Vol. 6, pp. 31–88). Academic Press.

Gubrud, M. A. (1997). Nanotechnology and international security. In Fifth Foresight Conference on Molecular Nanotechnology, 1.

Hagendorff, T. (2022). Blind spots in AI ethics. AI and Ethics, 2(4), 851-867.

Huang, T. J. (2017). Imitating the brain with neurocomputer a “new” way towards artificial general intelligence. *International Journal of Automation and Computing*, *14*(5), 520-531.

Ivanov, D. Chezhegov, A. Kiselev, M. Grunin, A. Larionov, D. (2022). “Neuromorphic artificial intelligence systems”. *Frontiers in Neuroscience*, *16*, 1513. Doi: 10.3389/fnins.2022.959626

Kaku, M. (2011). *Physics of the future: How science will shape human destiny and our daily lives by the year 2100*. Anchor.

Kurzweil, R. (2005). The singularity is near: When humans transcend biology, New York: Penguin.

Landgrebe, J., & Smith, B. (2021). An argument for the impossibility of machine intelligence. *arXiv preprint arXiv:2111.07765*.

Longino, H. E. (2022). What's Social About Social Epistemology?. *The Journal of Philosophy*, *119*(4), 169-195.

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

Muller, Vincent. (2021). Ethics of Artificial Intelligence and Robotics. The Stanford Encyclopedia of Philosophy. Retrieved December 1, 2023, from https://plato.stanford.edu/entries/ethics-ai/

Muller, Vincent. Bostrom, Nick. (2016). Future progress in artificial intelligence: A survey of expert opinion. Fundamental issues of artificial intelligence, 555-572. https://doi.org/10.1007/978-3-319-26485-1_33

Muller. V. Cannon M. (2022). "Existential risk from AI and orthogonality: Can we have it both ways?"*WILEY*. DOI: 10.1111/rati.12320, P: 25-36.

Munn, L. (2023). The uselessness of AI ethics. AI and Ethics, 3(3), 869-877.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2023). A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435.

Pei, J., Deng, L., Song, S., Zhao, M., Zhang, Y., Wu, S., ... & Shi, L. (2019). Towards artificial general intelligence with hybrid Tianjic chip architecture. *Nature*, *572*(7767), 106-111.

Pontes-Filho, S., Nichele, S. (2019). Towards a framework for the evolution of artificial general intelligence. *arXiv preprint arXiv:1903.10410*.

Pontes-Filho, S. Olsen, K., Yazidi, A. Riegler, M. A. Halvorsen, P. Nichele, S. (2022). Towards the Neuroevolution of Low-level artificial general intelligence. *Frontiers in Robotics and AI*, *9*, 1007547.

Puaschunder, J. M. (2019). On Artificial Intelligence's razor's edge: On the future of democracy and society in the artificial age. In *Proceedings of the 12th International RAIS Conference on Social Sciences and Humanities* (pp. 37-51). Scientia Moralitas Research Institute.

Reid, D. K. (1979). Equilibration and learning. *Journal of education*, *161*(1), 51-71.

Rosenberg, L. (2023). The Manipulation Problem: Conversational AI as a Threat to Epistemic Agency. *arXiv preprint arXiv:2306.11748*.

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

Searle, John. (1998). The mystery of consciousness. London: Granta Books.

Sullins, J. P. (2011). When is a robot a moral agent. *Machine ethics*, *6*(2001), 151-161.

Thomas, Wolfgang. (2015). Algorithms: From Al-Khwarizmi to Turing and Beyond. In G. Sommaruga & T. Strahm (Eds.), Turing's revolution: The impact of his ideas about computability, 29-42.

Varlamov, O. O. Chuvikov, D. A. Adamova, L. E. Petrov, M. A. Zabolotskaya, I. K. Zhilina, T. N. (2019). Logical, philosophical and ethical aspects of AI in medicine. International Journal of Machine Learning and Computing, 9(6), 868-873.

Wallach, W. Allen, C. (2008). Moral machines: Teaching robots right from wrong. Oxford University Press.

Yampolskiy, R. V. (2022). Agi control theory. In Artificial General Intelligence: 14th International Conference, AGI 2021, Palo Alto, CA, USA, October 15–18, 2021, Proceedings 14 (pp. 316-326). Springer International Publishing.

Zucca, P. (2022). Four cognitive-ecological biases that reduce integration between medical and cyber intelligence and represent a threat to cybersecurity. *Forensic Science International: Animals and Environments*, *2*, 100046.