Modern Technologies Law

University of Science and Culture

http://doi.org/10.22133/mtlj.2025.467357.1348

Moral Agency of Artificial Intelligence from the Perspective of Philosophy of Mind and Contemporary Neuroscience

Smaeil Keshavarz Safiyyi

Assistant prof., Faculty of Humanities, Islamic Azad University, Qazvin, Iran.

Article Info

Original Article

Received:
11-07-2024
Accepted:
16-10-2024

Keywords:

Agency

Consciousness

Free Will

Artificial Intelligence

***Corresponding author**
e-mail: keshavarzsmaeil@gmail.com

Abstract

The primary objective of this research is to examine the moral-legal agency of artificial intelligence from the perspectives of contemporary philosophy of mind and neuroscience. To achieve this goal, first, the concept of agency and the mental components of moral agency (including free will and consciousness) were investigated. Contemporary philosophers and neuroscientists have three different approaches to the mental components of moral agency: reductionist physicalism, non-reductionist physicalism, and dualism. The main question of this research is whether intelligent machines have moral agency or not. The findings of the research indicate that currently, reductionist physicalism, non-reductionist physicalism, and substance dualism (which attribute the origin of mental components to a non-physical substance called the "self," which does not exist in intelligent machines) deny that artificial intelligence has moral agency. In the meantime, phenomenological neuroscience has a more realistic approach than the previous three approaches. From the perspective of this approach, moral agency has biological (sense of self-existence) and phenomenological (the organism's intentional interaction with meaningful environmental stimuli) origins, and currently, artificial intelligence lacks these two components. However, advocates of reductionist, non-reductionist physicalism, and phenomenological neuroscience believe that in the future, with increasing advances in the field of intelligent machines achieving human physical and mental abilities (such as emotional artificial intelligence), artificial intelligence will also enjoy independent moral and legal agency.

How to Cite:
Keshavarz Safiyyi, S. (2026). Moral Agency of Artificial Intelligence from the Perspective of Philosophy of Mind and Contemporary Neuroscience. *Modern Technologies Law*, 7(13), 255-276.

Published by University of Science and Culture https://www.usc.ac.ir
Online ISSN: 2783-3836

1. Introduction

With the ever-increasing advances in artificial intelligence technology and the continuous production of intelligent devices and robots with higher quality and quantity of intelligence, the world is witnessing the significant presence of artificial intelligence in social relations with humans, and their acceptance as new members of human societies is becoming more and more important. The omnipresence of intelligent machines in various human activities and their replacement by human beings in high-risk jobs have brought risks from different perspectives.

One of these risks is the issue of responsibility arising from the decisions of artificial intelligence and the ability to assign responsibility to artificial intelligence as a decision maker. If we witness the emergence of strong artificial intelligence on a large scale in the course of human social relations and the decisions made by artificial intelligence The emergence of various legal responsibilities (both criminal and civil), can these responsibilities be attributed and imposed on artificial intelligence itself? So far, what has been written about the responsibility resulting from the behavior of intelligent machines has been mostly focused on the owner, user, or producer of artificial intelligence. Studying the works of moral philosophers about human responsibility indicates the existence of two important preconditions for the responsibility of competent (non-qualified) people: 1- agency, and 2- rationality. Free will and awareness are the components of having responsible moral agency (Qassimi & Thagha al-Islami, 1400, pp. 19-20/). Both components have a mental and neurological origin. Since the middle of the 20th century, with the expansion of cognitive sciences, Mental components have been increasingly noticed by philosophers of mind and neuroscientists in such a way that the existence of these elements has become a challenging issue from the point of view of different ways of thinking. Such as physicalism and dualism, each of which has a different approach to metaphysical components. In recent years, we have witnessed the formation of an amazing common territory between the achievements and discussions of neuroscience and philosophy of mind, with the findings and studies in the field of artificial intelligence. This common territory is a feasibility study. Simulating and transferring human mental abilities and capacities to intelligent machines. The concept of intelligence is basically a mental component that has been extended from humans to machines. One of the most challenging issues in artificial intelligence is the moral agency of artificial intelligence. The identification of this issue depends on the intelligent machines having free will and consciousness as two subjective components is moral agency. In this article, we try to examine the feasibility of agency and its components in artificial intelligence as essential components of responsibility.

2. Methodology

In this research, data collection was carried out using library studies and a descriptive-analytical method. The aim was to analyze what has been written so far about the nature of artificial intelligence and to provide a correct understanding of artificial intelligence and its nature in the first stage. In the second stage, the basis and possibility of using artificial intelligence in legal areas in general and the jurisdiction of the Guardian Council in particular are questioned. In the third stage, the legal implications and challenges related to the use of artificial intelligence in compliance with laws and regulations with Sharia and the Constitution, are analyzed and evaluated in order to achieve the desired legal requirements.

3. Results and Discussion

The common result of all three schools of dualism, reductionist and non-reductionist physicalism in the context of the possibility of intelligent machines having moral agency indicates a dominant tendency (although with different approaches) that intelligent machines do not accept moral agency in the current situation due to the lack of In reductionist physicalism, we see the existence of two major spectrums: in the first spectrum, which is based on the possibility of physical simulation of the brain and mind with hardware and Computer software is based on

computational functionalism, mental capabilities are reduced to brain capabilities and are removed from metaphysical elements. Nevertheless, one of the great challenges of reductionists is facing the category of moral agency and human responsibility towards their decisions. Because reducing mental states to physical states of the brain leads to the negation of agency and free will, and ultimately to belief in determinism and determinism. The second range of reductionists, despite believing in the reducibility of mental states to brain neurochemical states, consider the issue of human brain capabilities to be derived from the biological origin of man as a kind of organism and living being, and therefore, in the current situation, it is not possible to realize it in intelligent machines. Thus, with the lack of human brain capabilities (such as consciousness) in smart machines, the idea of artificial intelligence having moral agency will be ruled out. Non-reductionists deny the moral agency of intelligent machines with three premises: 1-Mental states such as will and consciousness, although as high-level characteristics, are derived from low-level physical characteristics, but due to their qualitative, first-person and affective aspects, they can be reduced to Quantitative and third-person states are not cerebral 2-Artificial intelligence currently lacks qualitative and subjective intentional states (such as will and consciousness) and relies solely on quantitative and physical computational components. 3- Moral agency depends on the agent having mental components (free will and awareness). The conclusion that comes from the above premises is that artificial intelligence machines lack moral agency. Dualists also agree with the intellectual premises of non-reductionists with the difference that, unlike non-reductionist physicalism, in the first premise, the origin of the qualitative states of the mind is derived from the activities of the lower physical levels of the brain and nervous system. Rather, it is attributed to a metaphysical component known as the self or immaterial soul. However, unlike substantive dualists

Thinkers defending reductionist, non-reductionist physicalism and phenomenological neuroscience believe that due to the possibility of realizing the functional and biological organization of the human brain in strong artificial intelligence technology and that this function is not exclusive to brain neurochemical properties, in the future with the advancement of technology and the possibility of achieving strong artificial intelligence With mental qualities and components such as awareness, will and self-awareness, the grounds for AI to have moral and legal agency will be provided.

4. Conclusions and Future Research

In the current situation, all four theories of dualism, reductionist physicalism, non-reductionist physicalism and phenomenological neuroscience believe that current intelligent machines lack moral and legal agency. However, unlike the essential dualism that excludes the possibility of the moral agency of artificial intelligence in the future due to the lack of its concept in intelligent machines, other approaches mentioned above believe that Considering the feasibility of realizing the functional and biological organization of the human brain in strong artificial intelligence technology and not being exclusive to the neurochemical properties of the brain, in the future Artificial intelligence will have moral and legal agency. Based on this, the advancements of smart technology specialists and neuroscientists in the future can promise the entry of artificial intelligence into the membership of the moral-legal society of humans.

حقوق فناوری‌های نوین، ۱۴۰۵، ۷(۱۳) https://www.mtlj.usc.ac.ir

حقوق فناوری‌های نوین

http://doi.org/10.22133/mtlj.2025.467357.1348

عاملیت اخلاقی ـ حقوقی هوش مصنوعی از منظر فلسفهٔ ذهن و عصب‌شناسی معاصر

اسماعیل کشاورز صفی‌ئی

استادیار، دانشکدهٔ علوم انسانی، دانشگاه آزاد اسلامی، قزوین، ایران.

اطلاعات مقاله

مقالهٔ پژوهشی

تاریخ دریافت:
۱۴۰۳/۰۴/۲۱

تاریخ پذیرش:
۱۴۰۳/۰۷/۲۵

واژگان کلیدی:
عاملیت
آگاهی
ارادهٔ آزاد
هوش مصنوعی

چکیده

هدف اصلی این پژوهش، بررسی عاملیت اخلاقی ـ حقوقی هوش مصنوعی از منظر رهیافت‌های فلسفهٔ ذهن و عصب‌شناسی معاصر است. برای رسیدن به این هدف، ابتدا مفهوم عاملیت و مؤلفه‌های ذهنی عاملیت اخلاقی (شامل ارادهٔ آزاد و آگاهی) بررسی شد. فیلسوفان و عصب‌شناسان معاصر سه رویکرد متفاوت درخصوص مؤلفه‌های ذهنی عاملیت اخلاقی دارند: فیزیکالیسم تقلیل‌گرا، فیزیکالیسم ناتقلیل‌گرا و دوگانه‌گرایی. پرسش اصلی این پژوهش آن است که آیا ماشین‌های هوشمند از عاملیت اخلاقی برخوردارند یا خیر؟ یافته‌های حاصل از پژوهش، حاکی از آن است که در وضع فعلی، هر سه رویکرد فکری فیزیکالیسم تقلیل‌گرا، فیزیکالیسم ناتقلیل‌گرا و دوگانه‌گرایی جوهری (که منشأ مؤلفه‌های ذهنی را به جوهری غیرفیزیکی موسوم به «خود» منتسب می‌دانند و این جوهر در ماشین‌های هوشمند وجود ندارد)، منکر بهره‌مندی هوش مصنوعی از عاملیت اخلاقی هستند. دراین‌میان، از نظر نویسنده، عصب‌شناسی پدیدارشناسانه رویکرد واقع‌گرایانه‌تری به‌نسبت سه رویکرد پیشین دارد؛ چراکه برخورداری هوش مصنوعی از عاملیت اخلاقی را منوط به دو ویژگی بیولوژیکی (احساس وجود خویشتن) و پدیدارشناختی (برهم‌کنش التفاتی ارگانیسم با محرک‌های معنادار محیط) می‌کند که در وضع موجود، هوش مصنوعی فاقد این دو مؤلفه است. با وجود این، مدافعان فیزیکالیسم تقلیل‌گرا، ناتقلیل‌گرا و عصب‌شناسی پدیدارگرا بر این باورند که در آینده، پیشرفت‌های روزافزون در شبیه‌سازی ماشین‌های هوشمند در برخورداری از مؤلفه‌های بدنی، مغزی و ذهنی انسانی (نظیر هوش مصنوعی عاطفی)، زمینه‌ساز عاملیت مستقل اخلاقی ـ حقوقی هوش مصنوعی در آینده است.

***نویسندهٔ مسئول**

رایانامه: keshavarzsmaeil@gmail.com

نحوهٔ استناددهی:

کشاورز صفی‌ئی (۱۴۰۵). عاملیت اخلاقی ـ حقوقی هوش مصنوعی از منظر فلسفهٔ ذهن و عصب‌شناسی معاصر. حقوق فناوری‌های نوین، ۷(۱۳)، ۲۵۵-۲۷۶.

ناشر: دانشگاه علم و فرهنگ https://www.usc.ac.ir

شاپای الکترونیکی: ۳۸۳۶-۲۷۸۳

مقدمه

با پیشرفت‌های روزافزون فناوری هوش مصنوعی و تولید بی‌وقفهٔ دستگاه‌ها و ربات‌های هوشمند با کیفیت و کمیت هوشمندی بالاتر، جهان شاهد حضور چشمگیر کنشگری هوش مصنوعی در روابط اجتماعی با انسان‌هاست و پذیرش آن‌ها به‌عنوان اعضای جدید جوامع انسانی بیش‌ازپیش درحال مطرح‌شدن است. حضور همه‌جانبهٔ ماشین‌های هوشمند در فعالیت‌های مختلف انسانی و جایگزینی آن‌ها با انسان، به‌ویژه در مشاغل پرریسک و پرخطر، مخاطراتی را از چشم‌اندازهای مختلف با خود به ارمغان داشته است. یکی از این مخاطرات، مسئلهٔ مسئولیت‌پذیری ناشی از تصمیم‌های هوش مصنوعی و قابلیت انتساب مسئولیت به هوش مصنوعی به‌عنوان تصمیم‌گیرنده است. اگر در جریان روابط اجتماعی انسان‌ها، شاهد ظهور هوش مصنوعی قوی در سطح گسترده باشیم و تصمیمات گرفته‌شده از سوی هوش مصنوعی موجب پدیداری مسئولیت‌های مختلف حقوقی (اعم از کیفری و مدنی) باشد، آیا می‌توان این مسئولیت‌ها را به خود هوش مصنوعی نسبت داد و تحمیل کرد؟ آنچه تاکنون درخصوص مسئولیت ناشی از رفتار ماشین‌های هوشمند به نگارش درآمده، بیشتر بر عاملیت مالک، کاربر یا تولیدکنندهٔ هوش مصنوعی متمرکز بوده است. در آثار و منابع خارجی، مسئلهٔ عاملیت اخلاقی هوش مصنوعی بسیار بررسی شده است[1].

باوجوداین در منابع داخلی، بررسی مبانی اخلاقی ـ فلسفی عاملیت هوش مصنوعی در میان محافل حقوقی چندان بررسی نشده است. مطالعهٔ آثار فیلسوفان اخلاق در باب مسئولیت‌پذیری انسان، حاکی از وجود دو پیش‌شرط مهم برای مسئولیت‌پذیری انسان‌های دارای اهلیت (غیرمحجور) است: ۱. عاملیت و ۲. عقلانیت. ارادهٔ آزاد و آگاهی، مؤلفه‌های برخورداری از عاملیت اخلاقی پاسخ‌گو به شمار می‌آیند (hasemi and Thiqa al-Islami, 2021, pp. 19-20). هر دو مؤلفه دارای خاستگاهی ذهنی و عصب‌شناختی هستند. از اواسط قرن بیستم با گسترش علوم شناختی، مؤلفه‌های ذهنی به‌طور فزاینده‌ای توجه فیلسوفان ذهن و عصب‌پژوهان را جلب کرد؛ به‌نحوی که وجود این عناصر از منظر نحله‌های فکری مختلف، به مسئله‌ای چالش‌برانگیز تبدیل شده است؛ نظیر فیزیکالیسم و دوگانه‌گرایی که هریک رویکرد متفاوتی به مؤلفه‌های متافیزیکی دارند. در سال‌های اخیر، شاهد شکل‌گیری قلمرو شگفت‌انگیز مشترکی میان دستاوردها و مباحث علوم عصب‌شناختی و فلسفهٔ ذهن با یافته‌ها و مطالعات حوزهٔ هوش مصنوعی هستیم. این قلمرو مشترک همانا امکان‌سنجی شبیه‌سازی و انتقال توانایی‌ها و ظرفیت‌های ذهنی انسان به ماشین‌های هوشمند است. مفهوم هوشمندی اساساً مؤلفه‌ای ذهنی است که از انسان به ماشین تعمیم داده شده است. یکی از پردامنه‌ترین موضوعات چالش‌برانگیز در هوش مصنوعی، مقولهٔ عاملیت اخلاقی هوش مصنوعی است؛ مسئله‌ای که شناسایی آن منوط به برخورداری ماشین‌های هوشمند از ارادهٔ آزاد و آگاهی به‌مثابهٔ دو مؤلفهٔ ذهنی عاملیت اخلاقی است. در این مقاله، کوشش بر آن است تا به بررسی امکان‌سنجی وجود عاملیت و مؤلفه‌های آن در هوش مصنوعی، به‌منزلهٔ مؤلفه‌های ضروری مسئولیت‌پذیری، بپردازیم. در همین راستا، در گام اول به بررسی اجمالی عاملیت در فلسفهٔ اخلاق می‌پردازیم. سپس با ورود به مقولهٔ اخلاق ماشین، امکان‌سنجی وجود این دو مؤلفه را در هوش مصنوعی بررسی خواهیم کرد.

---

۱. ازجمله این آثار می‌توان به مقالات ذیل اشاره کرد:

1. Himma KE. (2009). Artificial agency, consciousness, and the criteria for 2moral agency: What properties must an artificial agent have to be a moral agent? Ethics and Information Technology;11
2. Lorenzo Cominelli,1. Daniele Mazzei,2 and Danilo Emilio De Rossi1, (2018) Social Emotional Artificial Intelligence Based on Damasio's Theory of Mind, Frontiers in Robotics and AI 5: pp1-20
3. Rajakishore Nath & Riya Manna, (2021), From posthumanism to ethics of artificial intelligence, AI and Society
4. Tshilidzi Marwala (2017), Rational Choice and Artificial Intelligence, University of Johannesburg
5. Yi Zeng (2024) *Journal of LATEX Templates*, Brain-inspired and Self-based Artificial Intelligence

۱. عاملیت در فلسفهٔ اخلاق

۱-۱. مفهوم و مؤلفه‌های عاملیت

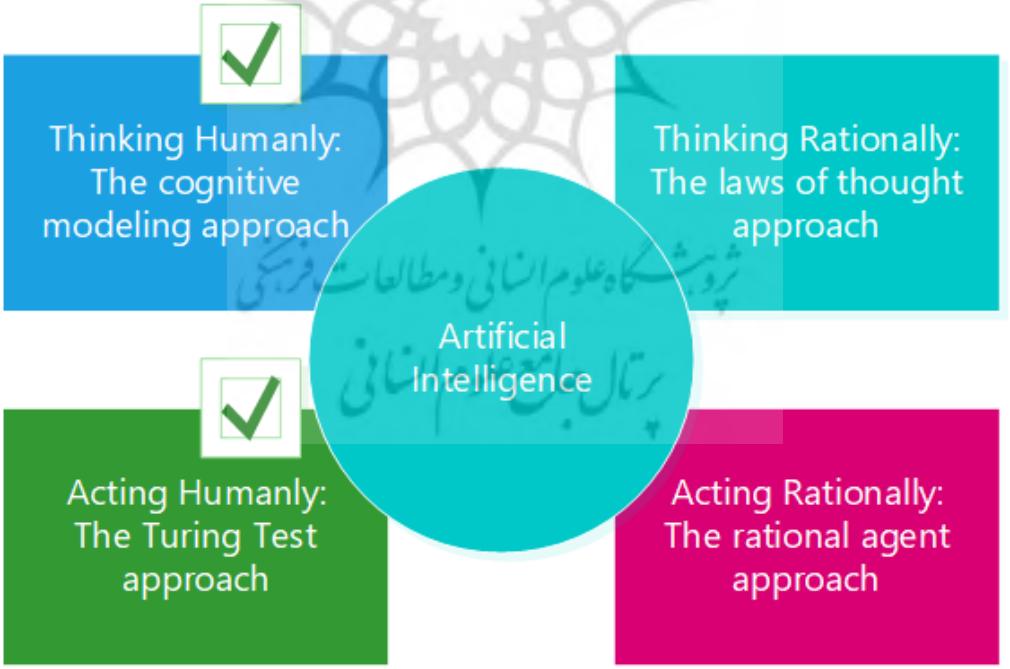
نزد اندیشمندان، عامل به کسی گفته می‌شود که در مواجهه با محرک‌های محیطی، معنای آن را ادراک و به آن‌ها پاسخ می‌دهد. عاملیت را در واقع می‌توان ویژگی ارگانیسم‌های آگاهی برشمرد که در تجربهٔ زیستهٔ خود در رویارویی هرروزه با محرک‌های معنادار محیطی، آن را ادراک و پس از محاسبه و سنجش به آن پاسخ می‌دهند. به‌عبارت‌دیگر، آنچه موجودات زنده را از اشیای بی‌جان متمایز می‌سازد، بهره‌مندی موجودات زنده از مکانیزم ادراک معانی محیطی و تأثیرگذاری در محیط پیرامونی ازطریق واکنش و تصمیم‌گیری است؛ مسئله‌ای که از آن تحت عنوان عاملیت یاد می‌شود. برهمین‌اساس، از منظر برخی اندیشمندان، عاملیت به‌معنای اثرگذاری ارادی بر عملکرد خود و وقایع محیطی است. انسان‌ها موجوداتی خودسامانده، فعال و خوداندیشه‌ورز هستند که به‌جای تماشای محض وقایع، در وقوع رخدادها مشارکتی فعال دارند؛ به‌نحوی که با بازبینی تجربیات زیستهٔ پیشین و پیش‌بینی پیامدهای احتمالی آیندهٔ رفتار خود، به‌نحو آگاهانه در تحقق رویدادهای آینده تأثیرگذارند (Bandura, 2006, p. 164; Ash'ari & Hassani, 2014, p. 99-100). در فلسفهٔ اخلاق، عاملیت اخلاقی به توانایی فرد برای انتخاب اخلاقی براساس برخی تصورات درست و نادرست و پاسخ‌گویی در قبال این اعمال اطلاق می‌شود و عامل اخلاقی موجودی است که قادر به عمل با ارجاع به درست و نادرست باشد (Angus, 2003, p. 943). از دیدگاه برخی دیگر، عاملان اخلاقی کسانی هستند که به خاطر داشتن توانایی‌های معین در انجام برخی اعمال، باید پاسخ‌گوی اعمال خود باشند؛ بدین‌ترتیب، عاملان اخلاقی از عاملان غیراخلاقی (نظیر کودکان و حیوانات) و همچنین از غیرعاملان (نظیر اشیای بی‌جان) متمایزند (Nailer, 2022, pp. 5-6). بنابر نظر برخی اندیشمندان آنچه موجب تمایز انسان کامل از عاملان غیراخلاقی و غیرعاملان می‌شود، بهره‌مندی انسان از حیث التفاتی و هدف‌مندی رفتار اوست؛ به‌گونه‌ای که انسان‌ها به‌دلیل داشتن دو مؤلفهٔ باور و امیال، واجد رفتاری هدف‌مند و دلیل‌محور هستند و برهمین‌اساس، دارای عاملیت اخلاقی هستند (Nailer, 2022, pp. 6-7). برخی دیگر نظیر تیلور، صرف توانایی بازنمایی انسان از حیث داشتن باور و میل به چیزی را مبنای عاملیت اخلاقی انسان و تمایز از عاملان غیراخلاقی (نظیر کودکان و حیوانات) نمی‌دانند؛ بلکه عاملان اخلاقی دارای نوعی فردیت و درکی از خود و خودآگاهی هستند که مبنای تصمیم‌گیری‌های اخلاقی آن‌هاست (Sugarman, 2005, pp. 796). همچنین، آنچه یک عامل اخلاقی را از کسی که موضوع اخلاقی یه شمار می‌رود متمایز می‌کند، مسئلهٔ مسئولیت‌پذیری عامل به‌سبب ارتکاب رفتار است؛ درحالی‌که یک شخص اخلاقی کسی است که می‌تواند موضوع حق‌های اخلاقی و حقوقی قرار گیرد؛ اما مشمول تکالیف اخلاقی (نظیر مسئولیت‌های اخلاقی و حقوقی) نمی‌شود؛ نظیر کودکان (Gallagher, 2007, p. 200). به‌عبارت دیگر، عاملیت اخلاقی با مسئولیت‌پذیری (قابلیت ستایش و نکوهش فرد به‌سبب رفتار ارتکابی) گره خورده است.

به‌طورکلی، مطالعهٔ آثار اندیشمندان مختلف در باب مؤلفه‌های برخورداری از عاملیت اخلاقی پاسخ‌گو حاکی از وجود دو شرط به‌عنوان شرایط عاملیت اخلاقی است: ۱. ارادهٔ آزاد و امکان انتخاب؛ ۲. آگاهی و پیش‌بینی نتایج عمل (hasemi & Thiqa al-Islami, 2021, p. 19-20). مقولهٔ اراده و آگاهی دربردارندهٔ حالات ذهنی و التفاتی هستند؛ بدین مفهوم که میل، قصد، نیت و آگاهی همیشه به چیزی یا دربارهٔ چیزی معطوف‌اند. این قبیل کیفیات ذهنی، نزد فیلسوفان، عصب‌پژوهان و روان‌شناسان تحت عنوان «حیث التفاتی یا دربارگی» شناخته می‌شود. ارادهٔ آزاد را می‌توان مهم‌ترین مؤلفهٔ ضروری عاملیت اخلاقی دانست. منظور از ارادهٔ آزاد، توانایی عامل در گزینش آزادانه و بدون اجبار تصمیم گرفته شده، با وجود امکان‌پذیری عدم انتخاب فعل ارتکابی یا انتخاب بدیل (گزینش فعل دیگری غیر از فعل ارتکابی)، است (Kane, 1998, p. 115). شرط دوم عاملیت، داشتن آگاهی از انجام فعل و پیش‌بینی پیامدهای آن است. مطابق این شرط، یک فرد درصورتی یک عامل اخلاقی است و می‌توان او را به‌سبب انجام رفتارش ستایش یا سرزنش کرد که به معنای عمل خود و پیامدهای آن آگاه باشد (Himma, 2009, pp. 19-29). به‌طور خلاصه، مطابق این شرط، عاملان اخلاقی عاملان آگاه هستند و در انتخاب و گزینش تصمیم، به‌نحو آگاهانه و عامدانه عمل می‌کنند (Gallagher, 2007, p. 200).

۲-۱. مفهوم هوش مصنوعی

باتوجه‌به آنچه در بخش پیشین گفته شد، مسئلهٔ عاملیت اخلاقی انسان با مؤلفه‌های ذهنی و التفاتی او نظیر آگاهی، قصد (اراده)، باور و میل به‌شدت گره خورده است. به‌عبارت‌دیگر، عاملیت اخلاقی مبتنی بر جنبه‌های سابجکتیو، کیفی و اول‌شخص انسانی است؛ چیزی که نزد فیلسوفان ذهن و عصب‌شناسان به‌عنوان کوالیا یا حیث التفاتی شناخته می‌شود (Bailey, 2020, p. 90). همین پرسش به‌طور جدی درخصوص عاملیت اخلاقی هوش مصنوعی نیز مطرح است؛ اینکه آیا هوش مصنوعی از کوالیا و حیث التفاتی برخوردار است یا خیر؟ اینکه اگر هوش مصنوعی فاقد مؤلفه‌های کیفی و التفاتی ذهنی (نظیر آگاهی از، باور به، میل به و قصد به چیزی) باشد، آیا می‌توان قائل به وجود عاملیت اخلاقی برای وی بود؟ پاسخ‌گویی به پرسش‌های فوق منوط به این است که هوش مصنوعی از منظر کدام رویکرد اخلاقی حاکم بر عاملیت ارادهٔ آزاد و آگاهی در تصمیم‌گیری نگریسته شود: فیزیکالیسم تقلیل‌گرا، ناتقلیل‌گرا یا دوگانه‌گرایی. مطالعهٔ آثار پژوهشگران موافق یا مخالف عاملیت اخلاقی هوش مصنوعی حاکی از تأثیرپذیری آن پژوهشگران از یکی از رویکردهای سه‌گانهٔ فوق‌الذکر است. باوجوداین، پیش از ورود به مبحث هوش مصنوعی از منظر رویکردهای سه‌گانه، لازم است به تعریف فلسفی خود هوش مصنوعی بپردازیم؛ چراکه نوع تعریفی که از هوش مصنوعی می‌شود در تعیین برخورداری هوش مصنوعی از مؤلفه‌های عاملیت اخلاقی تأثیرگذار است. برهمین‌اساس، ابتدا به تعریف هوش مصنوعی می‌پردازیم. سپس عاملیت اخلاقی هوش مصنوعی را از منظر رویکردهای سه‌گانه بررسی می‌کنیم.

مطالعهٔ تعاریف مختلفی که متخصصان و اندیشمندان از هوش مصنوعی کرده‌اند بیانگر چالشی بنیادین درخصوص شناسایی ماهیت و کارکرد هوش مصنوعی است. تنوع تعاریف ارائه شده را می‌توان بر پایهٔ دو شاخص تفکر/ رفتار و همچنین عقلانی/ انسانی بودن هوش مصنوعی استوار دانست. بدین‌ترتیب چهار تعریف از هوش مصنوعی ارائه شده است: ۱. تفکر انسان‌گونه؛ ۲. رفتار انسان‌گونه؛ ۳. تفکر عقلانی (منطقی)؛ ۴. رفتار عقلانی (منطقی).(Russell & Norvig, 2010, pp. 1-2) در ادامه به تحلیل این تعاریف می‌پردازیم.



(شکل ۱: تعاریف ارائه شده از هوش مصنوعی، Link)

۱-۲-۱. تفکر انسان‌گونه؛ رویکرد علوم شناختی

مسئلهٔ ماشینی که مانند انسان تفکر می‌کند نخستین مسئله‌ای بود که در زمینهٔ هوش مصنوعی مطرح شد. این تعریف به‌دنبال شبیه‌سازی و درون‌نگری فرایندهای ذهنی انسان در تصمیم‌گیری و به‌کارگیری آن در ماشین‌های هوشمند است. تمرکز این تعریف بر مقولهٔ تفکر است و اینکه آیا ماشین‌ها نیز می‌توانند مانند انسان دارای حیث التفاتی نظیر فکر کردن به چیزی باشند (Russell & Norvig, 2010, pp. 1-4). برای مثال، در تعریف هاگلند، هوش مصنوعی به‌منزلهٔ تلاش جدید و هیجان‌انگیزی است برای اینکه رایانه‌ها را به فکر وادار کنیم؛ ماشین‌هایی برخوردار از ذهن، به‌طور کامل (Nath, 2010, p. 28; Haugeland, 1989, p. 2). یا در تعریف بلمن، هوش مصنوعی نوعی اتوماسیون فعالیت‌های انسانی است که همراه با تفکر انجام می‌شود؛ نظیر تصمیم‌گیری، حل مسئله و یادگیری (Bellman, 1978, p. 5). همچنین هربرت سایمون و آلن ناول[1] را نیز می‌توان در این دسته قرار داد؛ زیرا به‌دنبال مقایسهٔ ردپای فرایند استدلال‌ورزی انسان با فرایند استدلال‌ورزی هوش مصنوعی بودند (Newell & Simon, 1961, pp. 109–124).

۲-۲-۱. رفتار انسان‌گونه؛ رویکرد آزمون تورینگ

خاستگاه فکری این تعریف به نظریهٔ کارکردگرایی محاسباتی بازمی‌گردد. کارکردگرایی در پی انتقاد از مکتب رفتارگرایی در اواسط قرن بیستم به همت فیلسوفانی مانند هیلاری پاتنم و جری فودور ظهور یافت. در این نظریه، ذهن به‌منزلهٔ محاسبه‌گر یا دستکاری‌کنندهٔ نماد عمل می‌کند. ذهن ورودی را از دنیای طبیعی می‌گیرد، محاسبه و پردازش می‌کند و خروجی‌های جدیدی را به‌شکل‌های انتزاعی یا واقعی درست می‌کند (Piccinini & Bahar, 2012; Horst, 2005). محاسبه، یک فرایند گرفتن ورودی و دنبال کردن مرحله‌به‌مرحلهٔ الگوریتم برای به‌دست آوردن یک خروجی مشخص است (Maslin, 2016, pp. 82-83). کارکردگرایی محاسباتی معلول شکل‌گیری اندیشهٔ موسوم به مدل رایانه‌ای ذهن است که خاستگاه آن به افکار آلن تورینگ[2]، ریاضی‌دان انگلیسی، بازمی‌گردد. وی با طراحی اولین دستگاه محاسباتی موسوم به موتور محاسباتی خودکار (ACE)، زمینهٔ انقلابی سترگ در دانش بشری و پدیداری علوم رایانه‌ای و هوش مصنوعی را فراهم آورد. تورینگ در آزمونی که امروزه در علوم رایانه‌ای و فلسفهٔ ذهن به آزمون تورینگ شهرت دارد، مسئلهٔ هوشمندی ماشین‌های محاسباتی ساختهٔ انسان (هوش مصنوعی) را در چهارچوب گفت‌وگوی میان یک داور از یک طرف و یک ماشین محاسباتی طرح‌اندازی کرد؛ به‌نحوی‌که درصورت موفقیت ماشین در دادن پاسخ‌های تقلیدی مانند انسان به سؤالات داور و عدم تشخیص داور مبنی بر اینکه پاسخ‌های داده‌شده از سوی یک ماشین است و نه یک انسان، ماشین مانند انسان دارای قابلیت هوشمندی تلقی می‌شود (Oppy & Dowe, 2011). به‌طورکلی، براساس این تعریف، اگر ماشینی بتواند رفتاری هوشمندانه و شبیه به انسان از خود نشان دهد، به صورتی که نتوان تشخیص داد که آیا عامل آن رفتار یک ماشین است یا انسان، آن ماشین هوشمند و دارای عاملیت خواهد بود؛ صرف‌نظر از اینکه چگونه و به چه کیفیتی تفکر می‌کند.

۳-۲-۱. تفکر عقلانی؛ رویکرد قوانین تفکر

خاستگاه فکری این تعریف به تعریف ارسطو از انسان به‌عنوان موجودی دارای تفکر منطقی بازمی‌گردد. ارسطو نخستین کسی بود که به فرموله کردن قوانین دقیق تفکر و استدلال منطقی (تولید نتایج قطعی مکانیکی از مقدمات) پرداخت. بعدها رامون دامول این ایده را مطرح کرد که استدلال صحیح منطقی می‌تواند توسط شیئی ماشینی نیز صورت گیرد. در قرن هفدهم، ویلیام اسکینارد و پاسکال اولین ماشین‌های محاسبه اختراع کردند و در همین قرن، توماس هابز در کتاب مشهور لویاتان خود، این ایده را آفرید که سیستم‌های فیزیکی عملکرد ذهن در تفکر منطقی را شبیه‌سازی می‌کنند؛ بنابراین می‌توان به ساخت انسان مصنوعی امیدوار بود. تأکید بر شناسایی قوانین تفکر صحیح، در اثبات‌گرایی منطقی قرن بیستم به اوج خود رسید؛ رویکردی که به‌دنبال طراحی یک فرایند صحیح محاسباتی و الگوریتمی برای استخراج صحیح دانش معتبر از

1. Herbert A. Simon & Allen Newell
2. Alan Mathison Turing

تجربیات اولیه بود. این رویکرد، که ذهن را به‌مثابهٔ فرایند پردازش‌گری محاسباتی تلقی می‌کرد، بعدها زمینه‌ساز شکل‌گیری ایدهٔ هوش مصنوعی شد (Russell & Norvig, 2010, p. 7). ایده‌ای موسوم به مدل رایانه‌ای ذهن، که در آن ذهن به‌مثابهٔ یک برنامهٔ نرم‌افزاری و مغز به‌عنوان یک رایانه سخت‌افزاری تلقی می‌شود و انسان به‌منزلهٔ ماشینی متفکر نگریسته می‌شود (Nath, 2010, p. 29; Searle, 1990, pp. 21-37). در این تعریف از هوش مصنوعی، تمرکز بر قواعد برای استدلال‌ورزی و تفکر صحیح است. رویکرد منطق‌گرا در هوش مصنوعی به‌دنبال آفرینش سیستم‌های هوشمند منطقی است؛ هرچند با موانعی نظیر عدم قطعیت احتمالات زیر ۱۰۰ درصد و تفاوت حل مسئله در اصل با عمل مواجه است (Russell & Norvig, 2010, p. 4).

٤-٢-١. رفتار عقلانی؛ رویکرد عامل عقلانی

این نظریه بر پایهٔ عقلانیت ابزاری و الگوریتمی هدف ـ وسیله استوار است؛ بدین‌نحو که هوش مصنوعی عاملی است که با ادراک محیط و سنجش آن، به‌دنبال رسیدن و یافتن بهترین نتیجه (پاسخ به محرک محیطی) است یا به‌دنبال بهترین نتیجهٔ مدنظر است. مزیت رویکرد رفتار عقلانی، به‌عنوان مقبول‌ترین تعریف از هوش مصنوعی در مقایسه با سایر تعاریف، این است که علاوه‌بر استفاده از استاندارد منطقی ـ عقلانی رویکرد سوم، بر مؤلفهٔ عینی و سوم‌شخص رفتار تمرکز دارد تا بر مؤلفه ذهنی و اول‌شخص تفکر. به‌طورکلی، تمایز انسانی از عقلانی و ترجیح عقلانی بر انسانی نزد اندیشمندان هوش مصنوعی و همچنین تمایز تفکر از رفتار و ترجیح رفتار به تفکر، حاوی چند نکتهٔ بسیار مهم است:

۱. ترجیح مدل عقلانی به‌جای انسانی در هوش مصنوعی، مؤید تلاش متخصصان هوش مصنوعی در برتری دادن عقلانیت کامل منطقی ـ ریاضی (الگوریتمی) و ابزاری به‌منزلهٔ الگوی ایده‌آل عقلانیت است که با کمترین خطای محاسباتی و دسترسی به حداکثر داده و اطلاعات ممکن، به انتخاب عقلانی بهترین گزینه (ابزار) برای حداکثرسازی مطلوب (بهینه‌سازی مطلوبیت به‌عنوان هدف) مبادرت می‌ورزد (Marwala, 2017). چنین الگویی از دیدگاه منتقدان عقلانیت ابزاری و کارکردگرایی محاسباتی، در انسان واقعی وجود ندارد؛ زیرا انسان واقعی به‌دلیل قدرت محاسباتی محدود و دسترسی محدود به داده‌ها، دچار خطا در رفتار و تفکر است (Engelen, 2007, Russell & Norvig, 2010, pp. 1-4; Elster, 1991, pp. 16-17; pp. 15-35)؛

۲. ترجیح تمرکز بر رفتار به‌جای تفکر، بیانگر گرایش به مؤلفه‌های عینی در مسئلهٔ هوش مصنوعی به‌جای مؤلفه‌های ذهنی است؛ چراکه رفتار بازتاب بیرونی و عینی تفکر است که از دید سوم‌شخص قابل مشاهده و به‌لحاظ تجربی قابل آزمون (قابل اثبات یا ابطال) است؛ درحالی‌که تفکر، فرایندی ذهنی و درونی و اول‌شخص است که از دید سوم‌شخص قابلیت ارزیابی و مشاهده ندارد و جعبهٔ سیاه ناگشودنی تلقی می‌شود.

۲. جایگاه مؤلفه‌های ذهنی عاملیت در هوش مصنوعی

باتوجه‌به اینکه مسئلهٔ عاملیت اخلاقی–حقوقی بر پایهٔ دو مؤلفهٔ قصد و نیت (اراده) و آگاهی استوار است و هر دو مؤلفهٔ مذکور، ماهیتی ذهنی دارند. برهمین‌اساس، می‌توان گفت عاملیت اخلاقی امری صرفاً فیزیکی و بر پایهٔ رابطهٔ علیت فیزیکی میان مغز، بدن و محیط نیست؛ بلکه واجد علیتی ذهنی است که در آن عامل عقلانی با برخورداری از اراده، آگاهی و نیت (دلیل)، به تصمیم‌گیری و رفتار مبادرت می‌ورزد (Searle, 2007, pp. 52-53; Elster, 1991, pp. 109-129). به‌دنبال مناقشهٔ فیزیکالیست‌ها بر سر علیّت ذهنی، مسئلهٔ عاملیت هوش مصنوعی در تصمیم‌گیری‌های خود نیز متأثر از این مناقشه خواهد بود. به‌عبارت‌دیگر، اینکه برای هوش مصنوعی قائل به مؤلفهٔ ذهنی باشیم یا خیر، در تحلیل نهایی ما از عاملیت اخلاقی هوش مصنوعی و مسئولیت‌پذیری آن تأثیرگذار خواهد بود. برهمین‌اساس، در این قسمت به تحلیل عاملیت اخلاقی هوش مصنوعی از منظر سه رویکرد فیزیکالیسم تقلیل‌گرا، غیرتقلیل‌گرا و دوگانه‌گرایی می‌پردازیم.

۱-۲. عاملیت اخلاقی ـ حقوقی هوش مصنوعی از منظر فیزیکالیسم تقلیل‌گرا

عناصر و مؤلفه‌های اصلی مدافعان فیزیکالیسم تقلیل‌گرا نظیر جیگوان کیم، پاول و پاتریشیا چرچلند و دیوید آرمسترانگ، که اغلب از فیلسوفان و عصب‌پژوهان معاصر هستند، از قرار زیر است: ۱. بنابر اصل بستار فیزیکی یا کفایت علی فیزیکی: تنها حالات فیزیکی می‌توانند موجب تأثیرگذاری علی بر جهان فیزیکی شوند (Kim, 2008, pp. 124-125)؛ ۲. بنابر اصل عدم تعین مضاعف و اصل طرد علی: درصورت کفایت تأثیر یک علت فیزیکی بر معلول فیزیکی، وجود علت دیگری اضافه بر علت فیزیکی مطرود است (Kim, 1993, pp. 281-354)؛ ۳. حالات ذهنی، همان حالات مغزی فیزیکی هستند؛ ۴. تنها حالات مغزی فیزیکی می‌توانند موجب تأثیرگذاری علی بر رفتار فیزیکی انسان به‌عنوان معلول شوند (Armstrong, 1978, pp. 261-276; Churchland, 1981, pp. 67-90). از مقدمات فوق این نتیجه حاصل می‌شود که عاملیت و علیت ذهنی وجود ندارد. یکی از نحله‌های فکری وفادار به فیزیکالیسم تقلیل‌گرا، رفتارگرایی و کارکردگرایی محاسباتی است که اندیشمندانی چون هیلاری پاتنم، جری فودور[1] از طراحان آن هستند. از طرف دیگر، سنت غالب در هوش مصنوعی، کارکردگرایی محاسباتی است که در آن، ذهن به‌مثابهٔ برنامهٔ نرم‌افزاری و مغز به‌مثابهٔ سخت‌افزار رایانه نگریسته می‌شود و کارکرد ذهن، کارکردی الگوریتمی و محاسباتی است که بر پایهٔ آن، با دسترسی به داده‌های لازم، بهترین نتایج مدنظر کسب می‌شود. در این سنت فکری، مؤلفه‌های ذهنی به مؤلفه‌های مغزی فروکاسته می‌شوند؛ به‌نحوی که با تمرکز بر رفتار مبتنی بر عقلانیت ابزاری، جنبه‌های اول‌شخص و سابجکتیو به جنبه‌های سوم‌شخص و ابژکتیو تحویل و تقلیل می‌یابند (کاوانا و نانی، ۱۴۰۱، ص ۱۰۵). بدین‌ترتیب می‌توان گفت تعریف هوش مصنوعی بر پایهٔ رفتار یا تفکر منطقی، برگرفته از نوعی فیزیکالیسم تقلیل‌گراست که با انکار جنبه‌های کیفی و فروکاستن ذهن به مغز و شبیه‌سازی کارکرد محاسباتی مغز در هوش مصنوعی، به امکان‌پذیری و حتی برتری عملکرد رفتاری هوش مصنوعی نسبت به انسان معتقد است. برهمین‌اساس است که طرح‌اندازان و مدافعان فلسفی هوش مصنوعی نظیر هیلاری پاتنم و جری فودور به فیزیکالیسم تقلیل‌گرا نیز باور دارند. بنابر اندیشه‌های نخستین هیلاری پاتنم، ذهن دارای یک سامان کارکردی شبیه ماشین‌های محاسباتی است و علت این کارکرد، خواص بیولوژیکی مغز نیست؛ بلکه این کارکرد در سامانه‌های غیربیولوژیکی نظیر رایانه نیز تحقق‌پذیر است. وی با طرح قابلیت تحقق متکثر کارکرد مغز در سخت‌افزارهای غیرمغزی، تا آنجا پیش رفت که مدعی شد اگر مغز همهٔ انسان‌ها از پنیر نیز ساخته شده باشند، اما سازمان کارکردی آن‌ها مانند اکنون باشد، باز هم قادر به ارائهٔ این کارکرد محاسباتی خواهند بود (Kavanna & Nani, 2022, pp. 105-106; Putnam, 2010, pp. 139-152). به‌عبارت‌دیگر، از منظر هیلاری پاتنم، شبیه‌سازی کارکردی سخت‌افزارهای غیرمغزی با کارکرد مغزی، موجبات عاملیت رفتاری هوش مصنوعی را مانند انسان فراهم خواهد کرد. براساس نظریهٔ علی جری فودور، مغز انسان از یک زبان موسوم به «زبان فکر» برخوردار است که دستور زبان خاص خود را دارد؛ بدین‌صورت که زبان فکر هر مفهوم موجود در عالم خارج را به‌صورت یک نماد بازنمایی می‌کند و با ترکیب این نمادهای بازنمایی‌شده، گزاره‌ها در ذهن ما تشکیل می‌شود. تلفیق نظریهٔ علی فودور با نظریهٔ کارکردگرایی محاسباتی پاتنم، موجب شکل‌گیری نظریهٔ محاسباتی ذهن شد که یک تبیین فلسفی برای حضور آگاهی در هوش مصنوعی ارائه می‌کرد (Fodor & Pylyshyn, 1988, pp. 3–71).

جولیانو تونونی[2] از دیگر عصب‌پژوهان نوخاسته‌گرایی است که با باور به نظریهٔ امکان تحقق متکثر کارکرد مغز در سایر سخت‌افزارهای فیزیکی و غیربیولوژیکی، بر امکان‌پذیری برخورداری ماشین‌های مصنوعی هوشمند از کیفیات ذهنی نظیر آگاهی و ارادهٔ آزاد تأکید دارد. بنابر نظریهٔ یکپارچگی اطلاعات تونونی، ماشین‌های مصنوعی در صورتی که قادر به تولید میزان گسترده‌ای از اطلاعات یکپارچه باشند، ضرورتاً ماشین‌هایی آگاه خواهند بود؛ بنابراین آگاهی صرفاً محصول خواص منحصربه‌فرد مواد بیولوژیکی مغزی نیست. با وجود این، ماشینی که قادر به تجربهٔ درونی/ذهنی آگاهی نباشد، صرفاً زامبی غیرآگاهی خواهد بود (Kavanna & Nani, 2022, p. 213; Tononi, 2004, pp. 1-22; Tegmark, 2022, pp. 480-481). بدین‌ترتیب می‌توان گفت از منظر تونونی، درصورت دستیابی هوش مصنوعی به تجربهٔ درونی آگاهی و

1. Hilary Putnam/Jerry Fodor
2. Juliano Tononi

ارادهٔ آزاد، امکان برخورداری ماشین‌های هوشمند از عاملیت اخلاقی ـ حقوقی میسر خواهد بود. در مقابل، برخی دیگر از عصب‌پژوهان با وجود باور به فیزیکالیسم تقلیل‌گرا و قابلیت فروکاستن حالات ذهنی به حالات فیزیکی مغزی، منکر قابلیت تحقق کارکردهای مغزی در ماشین‌های هوشمند هستند. بنابر نظریهٔ جرالد ادلمن[1]، آگاهی پدیده‌ای کاملاً بیولوژیکی و ویژگی منحصربه‌فرد مغز انسان زنده است؛ لذا عملکرد مغز قابل تحقق و مقایسه با رایانه نیست و ماشین‌های هوشمند به‌دلیل اتکا بر اجرای دستورهای ثابت و صریح و فقدان انعطاف‌پذیری و شکل‌پذیری و مرتب‌سازی اطلاعات مبهم در مواجهه با محرک‌های متنوع، قادر به تولید و انجام عملکردهای مغزی نیستند (Kavanna & Nani, 2022, Edeilman, 1989; p.168).

استانیسلاس دیهانه[2] از دیگر تقلیل‌گرایانی که بر مبنای نظریهٔ «فضای کار سراسری»، آگاهی را محصول یک شبکهٔ منسجم مغزی می‌داند، دیهانه است؛ به‌نحوی که پردازش آگاهانه، کارکرد اقتصاد محاسباتی مغز است. برهمین‌اساس، پیدایش ماشین‌های هوشمند آگاه اصولاً امکان‌پذیر است؛ بااین‌حال، ماشین‌های هوشمند برای برخورداری از آگاهی مصنوعی باید واجد سه ویژگی ارتباطات منعطف، انعطاف‌پذیری و خودمختاری باشند که درحال‌حاضر فاقد این ویژگی‌های اساسی هستند (Kavanna & Nani, 2022, p. 152). در واقع می‌توان گفت از منظر دیهانه، قابلیت پدیداری آگاهی در ماشین‌های هوشمند در آینده امکان‌پذیر است و درصورت دستیابی به این توانمندی، امکان فراهم‌ساختن عاملیت اخلاقی ـ حقوقی برای هوش مصنوعی میسر خواهد بود.

بنابر رویکرد کارکردگرایانه و تقلیل‌گرای دانیل دنت[3]، آگاهی و سایر مؤلفه‌های ذهنی نظیر اراده، وابستگی ضروری به ساختار نورونی مغز ندارند؛ برهمین‌اساس، یک ماشین مصنوعی نیز درصورت داشتن سازمان کارکردی مغز، از آگاهی برخوردار خواهد بود؛ اما درحال‌حاضر، یافته‌های عصب‌شناسی معاصر بر لزوم وابستگی پدیداری آگاهی به آناتومی و ساختار نورونی مغز تأکید دارد (Kavanna & Nani, 2022, p. 55). همان‌طور که ملاحظه می‌شود، دانیل دنت نیز مانند دیهانه و تونونی، مقولهٔ قابلیت برخورداری ماشین‌های هوشمند از آگاهی پدیداری را با پیشرفت فناوری می‌پذیرد؛ مسئله‌ای که درصورت حصول آن، می‌توان زمینه‌های عاملیت اخلاقی و حقوقی ماشین‌های هوشمند را فراهم دید.

بدین‌ترتیب، به‌اختصار می‌توان گفت عصب‌پژوهانی که به فیزیکالیسم تقلیل‌گرا متمایل‌اند در دو طیف عمده قرار می‌گیرند: طیف اول دربردارندهٔ کسانی نظیر پاتنم و فودور است که معتقد به تشابه میان مغز و دستگاه‌های هوشمند از حیث کارکرد محاسباتی هستند. رویکردی که در رایج‌ترین تعریف از هوش مصنوعی به‌منزلهٔ رفتار عقلانی نیز به‌چشم می‌خورد. در این نقطه‌نظر، با فروکاستن حالات ذهنی به حالات مغزی و تشابه میان کارکرد محاسباتی مغز با عملکرد دستگاه‌های هوشمند، نتیجه این می‌شود که در هوش مصنوعی نیز می‌توان قائل به وجود آگاهی و در نتیجه فراهم بودن مقدمات عاملیت اخلاقی ـ حقوقی بود. دستهٔ دوم باوجود باور به فروکاهش‌پذیری حالات ذهنی به حالات فیزیکی مغزی، منکر تشابه و قابلیت تعمیم کارکرد مغز به ماشین‌های هوشمندند؛ زیرا کارکرد مغز را ناشی از ویژگی بیولوژیکی انسان به‌منزله‌ی ارگانیسم زنده تلقی می‌کنند؛ چیزی که در ماشین‌های مصنوعی و غیربیولوژیکی به‌چشم نمی‌خورد. نتیجه اینکه آگاهی به‌مثابهٔ حالت بیولوژیکی مغز قابل تعمیم به هوش مصنوعی نخواهد بود.

پرسشی که مطرح می‌شود این است که آیا فروکاستن حالات ذهنی به حالات فیزیکی مغزی عملاً به جبرگرایی و فقدان عاملیت انسان و علیت ذهنی منجر نمی‌شود؟ زیرا بنابر اصل بستار و اصل عدم تعین مضاعف، معلول فیزیکی ناشی از علت فیزیکی مقدم بر آن است و علت فیزیکی برای تحقق معلول فیزیکی کافی است و نیازی به علت مضاعف دیگری نظیر علیت ذهنی نیست. بدیهی است که جبرگرایی و نفی عاملیت ذهنی اراده و آگاهی، به عدم مسئولیت انسان در قبال تصمیماتش خواهد منجر شد؛ چراکه درصورتی می‌توان فردی را مسئول تصمیماتش دانست که بتوان او را به‌خاطر رفتار ارتکابی سرزنش کرد و درصورتی می‌توان کسی را به‌خاطر رفتارش سرزنش کرد که در گزینش رفتار ارتکابی

1. Gerald Edelman
2. Stanislas Dehaene
3. Daniel Dennett

اراده و آگاهی داشته باشد (Vargas, 2004, pp. 218-241; Fischer, 2007, pp. 147-150; Van Inwagen, 1983, p. 56). بدین‌ترتیب، با تعمیم جبرگرایی ناشی از عاملیت فیزیکی و نفی عاملیت ذهنی در تصمیم‌گیری از رفتار انسانی به رفتار هوش مصنوعی، باید لاجرم قائل به فقدان مسئولیت‌پذیری هوش مصنوعی بود؛ زیرا از این چشم‌انداز، تصمیمات هوش مصنوعی نیز معلول علل جبری و ناخودآگاه فیزیکی (پاسخ الگوریتمی تراشهٔ الکترونیکی موجود در دستگاه‌های هوشمند به محرک‌های فیزیکی بیرونی) است.

۲-۲. عاملیت اخلاقی-حقوقی هوش مصنوعی از منظر فیزیکالیسم ناتقلیل‌گرا

بنابر نظریهٔ فیزیکالیسم ناتقلیل‌گرا، حالات و کیفیات ذهنی انسان (نظیر اراده و آگاهی) به‌عنوان ویژگی‌های سطح بالا، برآمده از فرایند تکامل بدنی و فیزیکی در سطوح پایین اوست؛ لذا جوهر متمایزی موسوم به نفس که مجرد از بدن باشد وجود ندارد. با وجود این، حالات ذهنی قابل فروکاستن به حالات مغزی نیستند و نمی‌توان آن را با حالات مغزی این‌همان دانست و از آنجا که به‌لحاظ کارکردی قابل پیش‌بینی و تبیین‌پذیری توسط ویژگی‌های سطوح پایین فیزیکی نیستند، برهمین‌اساس قابل فروکاستن و تقلیل به حالات فیزیکی نیستند (Beorlegui, 2009, p. 902). تندترین انتقادات وارده بر رویکرد الگوریتمی هوش مصنوعی از سوی کسانی بوده است که می‌توان آن‌ها را ذیل فیزیکالیسم ناتقلیل‌گرا قرار داد. اندیشمندانی مانند هیوبرت دریفوس[1] و جان سرل[2] و مدافعان اندیشهٔ نوخاسته‌گرایی در این دسته قرار می‌گیرند.

هیوبرت دریفوس، متأثر از اندیشهٔ هرمنوتیک، اگزیستانسیالیسم مارتین هایدگر[3] و پدیدارشناسی موریس مرلوپونتی[4]، در آثار متعدد خود (شامل کیمیاگری و هوش مصنوعی[5]، آنچه رایانه نمی‌تواند انجام دهد[6]، ذهن بیش از یک دستگاه[7])، پیش‌فرض‌های چهارگانهٔ هوش مصنوعی (شامل پیش‌فرض‌های بیولوژیکی، روانی، هستی‌شناختی و معرفت‌شناختی) را که بر پایهٔ شبیه بودن عملکرد هوش مصنوعی به عملکرد ذهن انسان استوار است، بررسی می‌کند. اساس انتقاد دریفوس از هوش مصنوعی این است که اولاً عملکرد ذهن انسان برخلاف رایانه به‌صورت آنالوگ است، نه دیجیتال. ثانیاً جهان انسانی کاملاً متفاوت از عالم فیزیکی است. جهان انسانی امری معنادار، آمیخته به ارزش و زمان‌مند است و ذهن سوبژکتیو انسان در هر لحظه درحال پردازش‌گری معنادار و ارزشمند با زیست‌جهان پیرامون خود است؛ درحالی‌که عالم فیزیکی مجموعه‌ای از واقعیت‌های بی‌معنا، بی‌زمان و خنثی است و پردازش‌گری رایانه نیز به‌مثابهٔ تابعی الگوریتمی و ریاضیاتی، فاقد هرگونه ذهن سوبژکتیو و محتوای درونی و معنادار است. ضمن اینکه پردازش‌گری رایانه با امری دقیق و نامبهم سروکار دارد، درحالی‌که تجربهٔ ادراکی و آگاهانهٔ انسان با محیط پیرامون ممکن است شکلی مبهم و نامعین داشته باشد (Dreyfus, 1999)؛ ثالثاً هر دانش یا معرفت بشری قابل تبدیل به گزاره‌های ریاضی ـ منطقی نیست؛ درحالی‌که مکانیزم هوش مصنوعی بر پایهٔ تبدیل دانش به گزاره‌ها و داده‌های الگوریتمی استوار است (Moghaddas, 2020, pp. 165-168).

از دیگر منتقدان و مخالفان آگاهی و ارادهٔ آزاد در هوش مصنوعی، جان سرل می‌باشد. تمثیل اتاق چینی[8] جان سرل یکی از مهم‌ترین و دیرپاترین انتقاداتی است که نسبت به مسئلهٔ آگاهی در هوش مصنوعی مطرح شده است. وی در مقام رد آگاهی در ماشین‌های هوشمند این تمثیل را مطرح کرد که اگر فردی مسلط به زبان انگلیسی در اتاقی قرار گیرد که در آن، مجموعهٔ قواعد لازم برای ترکیب و ساختن جملات به زبان چینی در اختیارش باشد و بتواند در پاسخ به سؤالات بیرون از اتاق به زبان چینی، براساس این قواعد ترکیب پاسخ‌های درستی دهد، افراد بیرون از اتاق تصور می‌کنند که فرد درون اتاق، زبان چینی را می‌فهمد؛ درحالی که این‌گونه نیست. رایانه‌های هوشمند نیز دقیقاً شبیه فرد درون اتاق

1. Hubert Dreyfus
2. John Searle
3. Martin Heidegger
4. Maurice Merleau-Ponty
5. *Alchemy and artificial intelligence*
6. *What Computers Still Can't Do: A Critique of Artificial Reason*
7. Mind Over Machine
8. Chinese room

عمل می‌کنند؛ یعنی بدون درک و آگاهی از معنای کلمات، صرفاً قادر به محاسبهٔ الگوریتمی نمادها و داده‌ها و پاسخ به محرک بیرونی هستند (Searle, 1980, pp. 332-35; Kavanna & Nani, 2022, p. 121). با وجود این، جان سرل در سال‌های اخیر با طرح نظریهٔ طبیعت‌گرایی زیست‌شناختی[1] تا حدودی به قابلیت ایجاد حالات ذهنی در هوش مصنوعی متمایل شده است. بر مبنای این نظریه، پدیدارهای ذهنی نظیر تفکر، آگاهی و اراده، به‌عنوان پدیداری سطح بالا، ناشی از پدیدارهای سطح پایین عصب ـ زیست‌شناختی مغز هستند. با وجود این، کیفیات ذهنی به‌دلیل داشتن ویژگی‌های کیفی، التفاتی و اول‌شخص‌بودن، به فرایندهای مغزی فیزیکی فروکاسته نمی‌شوند؛ هرچند که دارای خاستگاه و بنیادی نوروبیولوژیک (فیزیکی) باشند. با وجود این، وی به‌مانند کارکردگرایان با اعتقاد به قابلیت تحقق چندگانه، بر این باور است که مغز یگانه ابزار تحقق کیفیات ذهنی نیست؛ بلکه ماشین مصنوعی نیز درصورت پیشرفت دانش بشری، قادر به تولید حالات ذهنی خواهد بود (Searle, 2007, pp. 326-330). به‌عبارت دیگر، جان سرل بر این باور است که ماشین‌های هوشمند درصورت پیشرفت و مجهزشدن به مؤلفه‌های ذهنی مغز، قادر به برخورداری از آگاهی و ارادهٔ آزاد خواهند بود. بدین‌ترتیب با مجهز شدن هوش مصنوعی به کیفیات ذهنی، می‌توان از عاملیت اخلاقی و حقوقی آن‌ها سخن گفت.

از دیگر طرف‌داران فیزیکالیسم ناتقلیل‌گرا، مدافعان اندیشهٔ نوخاسته‌گرایی هستند. بنابر این اندیشه که نوع سازنده و رایج آن «Micropsychism» است، آگاهی در سطوح ماکرو (مانند انسان و حیوانات) از ترکیب آگاهی‌های سطوح میکرو (ذرات تشکیل‌دهندهٔ موجودات) ساخته می‌شود؛ اما ازآنجاکه حالات ذهنی (نظیر آگاهی و قصد) به‌لحاظ کارکردی قابل پیش‌بینی و تبیین‌پذیری توسط ویژگی‌های سطوح پایین فیزیکی نیستند، برهمین‌اساس قابل فروکاستن و تقلیل به حالات فیزیکی نیز نخواهند بود (Beorlegui, 2009, p. 902). برهمین‌اساس، حالات و کیفیات ذهنی ما درحالی‌که از ویژگی‌های سطح پایهٔ فیزیکی مغز نشئت می‌گیرند، دارای تأثیرگذاری علّی (علیت ذهنی) بر تصمیمات و رفتارهای (امر فیزیکی) ما نیز هستند (Gazzaniga, 2014, p. 61). کاربست این رویکرد درخصوص امکان‌سنجی بروز حالات ذهنی نظیر آگاهی در هوش مصنوعی با این مانع بزرگ مواجه می‌شود که عملکرد سیستم عصبی ـ مغزی انسان در پدیدارسازی حالات ذهنی (نظیر آگاهی و میل) به‌صورت آنالوگ است؛ زیرا تغییر در محرک‌های فیزیکی، باعث تغییر در مقادیر فیزیکی بازنمایی آن‌ها می‌شود. برای مثال، افزایش شدت صدا باعث افزایش سرعت شلیک نورون‌های حلزون گوش می‌شود؛ درحالی‌که عملکرد ماشین‌های هوشمند در بازتولید و ارائهٔ پاسخ‌های رفتاری به‌صورت دیجیتال است؛ به‌گونه‌ای که بازنمایی مقادیر فیزیکی در قالب محاسبات الگوریتمی کمیتی (عددی) و به‌صورت صفر و یک (بدون برخورداری از جنبهٔ کیفی مقادیر بازنمایی شده) صورت می‌گیرد (Beck, 2019, pp. 319-49; Chiang et al., 2018; Arvan & Maley, 2022, p. 244). بدین‌ترتیب، باتوجه‌به اینکه آگاهی به‌عنوان یک حالت ذهنی در سطح ماکرو از ترکیب مقادیر فیزیکی در سطح میکرو که به‌نحو آنالوگ بازنمایی می‌شوند پدید می‌آید و ماشین‌های هوشمند فاقد جنبهٔ بازنمایی آنالوگ هستند و صرفاً مقادیر فیزیکی را به‌صورت کمیتی و عددی بازنمایی می‌کنند؛ درنتیجه نمی‌توانند بدون بازنمایی‌های عصبی آنالوگ، تجربیات پدیداری در سطوح میکرو را پدید آورند و برهمین‌اساس، فاقد آگاهی پدیداری در سطح ماکرو هستند (Arvan & Maley, 2022, p. 244).

بدین‌ترتیب به‌طور خلاصه از منظر فیزیکالیست‌های ناتقلیل‌گرا، مؤلفه‌های ذهنی نظیر آگاهی و اراده قابل فروکاستن به مؤلفه‌های فیزیکی نخواهد بود و هوش مصنوعی نیز در وضع فعلی فاقد مؤلفه‌های ذهنی والتفاتی نظیر آگاهی است. نتیجه اینکه با فروکاست‌ناپذیری مؤلفه‌های کیفی و التفاتی ذهنی (نظیر آگاهی و قصد) به عناصر کمیتی و فیزیکی و اتکای ماشین‌های هوشمند کنونی بر جنبه‌های کمّی و فیزیکی محاسباتی، امکان تصورپذیری حالات و کیفیات ذهنی در هوش مصنوعی منتفی است و باتوجه‌به اینکه وجود مؤلفه‌های ذهنی نظیر آگاهی و اراده (میل و قصد)، از شرایط ضروری عاملیت اخلاقی است؛ بنابراین ماشین‌های هوشمند به‌منظور فقدان مؤلفه‌های کیفی ذهنی، فاقد عاملیت اخلاقی خواهند بود؛ درنتیجه نمی‌توان از آن‌ها انتظار مسئولیت‌پذیری در قبال پیامدهای ناشی از تصمیم‌گیری‌های خود داشت. باوجوداین،

1. Biological Naturalism

برخی اندیشمندان نظیر جان سرل با اعتقاد به قابلیت تحقق کارکردهای مغزی از طریق فناوری‌های غیربیولوژیکی و قابلیت هوش مصنوعی در دستیابی به کیفیات ذهنی در آینده، به امکان‌پذیری فراهم ساختن زمینه‌های عاملیت اخلاقی ـ حقوقی هوش مصنوعی باور دارند.

۳-۲. عاملیت اخلاقی-حقوقی هوش مصنوعی از منظر دوگانه‌گرایی

بنابر نظر مدافعان دوگانه‌گرایی جوهری، که خاستگاه آن به آثار رنه دکارت فرانسوی بازمی‌گردد، همه‌چیز از دو گوهر قائم به ذات و دارای هستی متمایز ساخته شده است: گوهر اول، ماده یا شیء دارای بُعد و امتداد است که هستی خارجی یا عینی دارد و دوم، گوهر اندیشه که دارای بُعد و مکان نیست و هستی درونی یا ذهنی دارد (Rozemond & Carriero, 2008, pp. 372-389). ذهن دکارتی غیرمکانیکی است. دکارت مفهوم «من فکر می‌کنم» را پیش‌فرض می‌گیرد؛ زیرا این «من» هستم که جهان را تجربه می‌کند. به‌همین ترتیب، مفهوم «من» دکارت، مفهوم محاسباتی ذهن را نفی می‌کند. جوهر ذهن، فکر است و اعمال افکار با اعمال آگاهی یکی می‌شود؛ بنابراین، نتیجه آن می‌شود که اعمال شناختی، اعمال آگاهانه هستند، نه اعمال محاسباتی که توسط ماشین‌های هوشمند صورت می‌گیرد. بنابراین، برای دکارت، یکی از مهم‌ترین جنبه‌های حالات و فرایندهای شناختی، پدیداری آن‌هاست؛ زیرا قضاوت، درک و غیره را فقط می‌توان در رابطه با آگاهی تعریف و توضیح داد و نه در رابطه با محاسبات. ما فقط می‌توانیم محاسبات را در ماشین‌ها پیدا کنیم و نه در ذهن که اراده می‌کند، می‌فهمد و قضاوت می‌کند (Nath, 2021, p. 7).

همان‌طور که ملاحظه می‌شود، تأکید نظریهٔ دوگانه‌گرایی بر وجود جوهری غیرفیزیکی موسوم به خود یا نفس به‌عنوان شخص آگاه و تمایز میان حالات شناختی و پدیداری به‌عنوان عمل آگاهانه از اعمال محاسباتی است. از نقطه‌نظر دوگانه‌گرایی، فناوری های کنونی هوش مصنوعی به‌منظور تمرکز بر جنبه‌های فیزیکی و مدل رایانه‌ای و الگوریتمی کارکرد محاسباتی مغز انسان در ادراک و تصمیم‌گیری، فاقد ذهنیت و مؤلفه‌های ذهنی و انتزاعی اشخاص آگاه نظیر خود (من)، آگاهی، حیث التفاتی و خودآگاهی است؛ زیرا هیچ احساسی از خود در آن‌ها وجود ندارد. هوش مصنوعی کنونی فقط به ظاهر هوشمندانه پردازش اطلاعات می‌کند و واقعاً خود را درک نمی‌کند یا به‌طور ذهنی از خود آگاه نیست و جهان را، همان‌طور که هوش انسانی درک می‌کند، با خود درک نمی‌کند (Yi Zeng, 2024). ذهنیت نوعی بازنمود چشم‌انداز و هویت شخصی در تجربیات روزمره است. درحال‌حاضر، به‌منظور فقدان لایه‌های عصب ـ بوم‌شناختی[1] و ذهنی مغز انسان، هوش مصنوعی یک ذهنیت بنیادین را نشان نمی‌دهد و برهمین‌اساس، در وضع کنونی، هیچ آگاهی یا خودی در مدل‌هایی مانند چت جی‌پی‌تی و موارد مشابه امکان‌پذیر نیست (Northof & Gouveia, 2024, pp. 16-18). در واقع می‌توان گفت از منظر دوگانه‌گرایان جوهری، با وجود مهارت‌های روزافزون و شگفت‌انگیز فناوری‌های مصنوعی در شبیه‌سازی عملکردهای انسانی، به‌منظور فقدان مؤلفه‌های سه‌گانهٔ طبیعت انسانی (بدن، ذهن و نفس یا خود) در ماشین‌های هوشمند، امکان برخورداری از ذهنیت و موقعیت و شأن اخلاقی در وضع کنونی برای آن‌ها میسر نیست (Wróblewski & Fortuna, 2023, pp. 43-32). درمجموع، از آنجا که نمی‌توان قائل به وجود واقعی و بیولوژیکی مفهوم خود (نوعی تعلق انحصاری بدن‌مندانه و عاطفی که دربردارندهٔ منافع، علایق، نیازها، باورها، امیال و ترجیحات شخصی باشد) بود و آگاهی التفاتی اول‌شخص هوش مصنوعی از وجود خود و متعلقاتش وجود ندارد؛ لذا نمی‌توان هوش مصنوعی را به‌عنوان یک شخص مستقل و آگاه در جهان اخلاق و حقوق به‌رسمیت شناخت و برای وی عاملیت و مسئولیت اخلاقی و حقوقی در نظر گرفت (Russell & Norvig, 2010, pp. 1027-1033; Nath, 2021, pp. 6-8).

از مظاهر و انشعابات نوین دوگانه‌گرایی جوهری، نظریهٔ دوگانه‌گرایی خاصیتی یا ویژگی است. بنابر نظریهٔ دوگانه‌گرایی طبیعی یا خاصیتی دیوید چالمرز[2]، هر ساختار فیزیکی برای اینکه دارای تجربهٔ آگاهانه باشد، باید شامل سه مؤلفهٔ انسجام ساختاری، تغییرناپذیری سازمانی و رویکرد دووجهی به اطلاعات باشد. دراین‌میان، مؤلفهٔ سوم مهم‌ترین مؤلفه است؛ زیرا علاوه‌بر وجه کمیتی یا فیزیکی اطلاعات، بر جنبهٔ کیفی

1. Neuroecological
2. David Chalmers

و غیرفیزیکی آن، که همان آگاهی پدیداری است، تأکید دارد (Chalmers, 1995, pp. 200-219). برهمین‌اساس، برخورداری رایانه‌ها از آگاهی منوط به تشابه عملکرد سازمانی رایانه با سازمان عملکردی مغز انسان است. یافته‌های عصب‌پژوهی معاصر حاکی از آن است که ویژگی‌های منحصربه‌فرد آناتومیک و فیزیولوژیکی نورون‌های مغز، نقش تعیین‌کننده‌ای در پدیداری آگاهی ایفا می‌کنند؛ درحالی‌که چنین خصوصیاتی در تراشه‌های سیلیکونی ماشین‌های هوشمند وجود ندارد و در این حالت، ما صرفاً با یک زامبی غیرآگاه مواجه خواهیم شد (Tegmark, 2022, p. 481; Kavanna & Nani, 2022, pp. 27-28) همان‌طور که ملاحظه می‌شود، از منظر چالمرز نیز درصورت مشاهدهٔ مؤلفه‌های سه‌گانهٔ سازمان عملکردی مغز انسان در ماشین‌های هوشمند، می‌توان از وجود آگاهی در هوش مصنوعی که مؤلفهٔ اساسی عاملیت اخلاقی ـ حقوقی است، سخن گفت.

دوگانه‌گرایی نوخاسته‌گرا را می‌توان از دیگر شاخه‌های نوین دوگانه‌گرایی برشمرد که با تکیه بر داده‌های علمی جدید، به دفاع از این نظریه درخصوص منشأ علّی رفتار انسان می‌پردازند (Amosultani, 2018, pp. 111-136). تحت این یافته‌ها، آنچه انسان و قوای ذهنی او را از فعالیت‌ها و فرایندهای طبیعی (نظیر ذوب شدن بستنی در آب) متمایز می‌کند، وجود عاملی مختار موسوم به «خود» یا «نفس» است که صرفاً پاسخی بازتابی به محرک‌های محیطی بیرونی نمی‌دهد؛ بلکه اعمال خود را براساس باورها و امیال و ترجیحاتش انتخاب و گزینش می‌کند، نه اینکه صرفاً معلول شرایط محیطی پیشین باشد (Stephan, 2010, pp. 180-189; Goodman et al., 2013, p. 178).

به‌طور خلاصه، از منظر رویکرد دوگانه‌گرایی، آگاهی به‌عنوان مؤلفه‌ای ذهنی، مستلزم وجود «خود» یا «نفس» به‌مثابهٔ «سابجکت آگاه» است؛ سابجکتی که از وجود خود آگاهی دارد و در گزینش نهایی تصمیمات خود دارای عاملیت است. ماشین‌های هوشمند در وضع کنونی فاقد «خود» و هرگونه احساس تعلق به خود هستند؛ درنتیجه فاقد آگاهی، اراده و خودآگاهی هستند؛ بنابراین تصمیمات آن‌ها بر مبنای باور، امیال و ترجیحات و غایات معطوف به «خود» یا «نفس» آن‌ها نیست. برهمین‌اساس، در وضع موجود، امکان پذیرش عاملیت اخلاقی در ماشین‌های هوشمند میسر نخواهد بود. همان‌طور که ملاحظه می‌شود، نظریهٔ دوگانه‌گرایی جوهری در مقایسه با فیزیکالیسم تقلیل‌گرا و ناتقلیل‌گرا، به‌منظور خصلت متافیزیکی و روح‌گرایانهٔ خود، مخالفت بیشتری را در پذیرش عاملیت اخلاقی ـ حقوقی هوش مصنوعی از خود نشان می‌دهد.

٤-٢. عاملیت اخلاقی-حقوقی هوش مصنوعی از منظر عصب‌شناسی پدیدارشناختی[1]

به‌طور کلی، عصب‌پژوهانی که متمایل به اندیشهٔ پدیدارشناسی ادراک و تأکید بر جنبهٔ بدن‌مندانه و عاطفی ادراک و تصمیم‌گیری انسان، توأمان با پذیرش جوهری متافیزیکی موسوم به «خویشتن» یا «خود» هستند، را می‌توان به‌منزلهٔ سنت و رویکرد آشتی‌جویانه میان فیزیکالیسم (اعم از تقلیل‌گرا و غیرتقلیل‌گرا) و دوگانه‌گرایی جوهری دانست. خاستگاه پدیدارشناسی ادراک به اندیشه‌های فیلسوف فرانسوی، موریس مرلوپونتی، بازمی‌گردد. از چشم‌انداز مرلوپونتی، انسان موجودی گسسته از جهان نیست؛ بلکه موجودی دارای حیث التفاتی (نظیر آگاهی از چیزی، میل یا قصد به چیزی) است که تجربه‌های ادراکی‌اش در مواجهه با موقعیت‌های محیطی از طریق حرکات بدنی، حساسیت و عاطفه آغاز می‌شود و درنهایت به ادراک آن موقعیت و ارتکاب رفتار ارادی بدنی منجر می‌شود. برهمین‌اساس، ادراک و آگاهی از یک موقعیت، یک توان ذهنی یا مکانیکی محض نیست؛ بلکه امری بدن‌مند است. رفتار ارتکابی انسان، حیث التفاتی عملی است که از برهم‌کنش معنادار انسان با محیط پیرامونی شکل می‌گیرد (Bailey, 2020, p. 90). از رویکرد پدیدارشناختی نزد عصب‌پژوهانی که به مقولهٔ ادراک و آگاهی علاقه داشتند، بسیار استقبال شد. از شناخته‌شده‌ترین عصب‌پژوهان پدیدارشناس می‌توان به آنتونیو داماسیو[2] و جان کوین اورگان[3] اشاره کرد. ویژگی مشترک عصب‌پژوهان پدیدارشناس، تأکید بر وجود عنصری به‌نام «احساس خویشتن» یا «خود» است که حاوی محتوایی عاطفی و بدن‌مند است. جان

1. Phenomenological Neuroscience
2. Antonio Damasio
3. John Coyne Oregan

کوین اورگان در نظریهٔ حسی ـ حرکتی خود، سه نوع خویشتن‌شناختی را مطرح می‌کند که عبارت‌اند از: ۱. تمییز خویشتن از جهان خارج؛ ۲. معرفت دربارهٔ خویشتن (دانستن دربارهٔ خود)؛ ۳. معرفت دربارهٔ آگاهی از خویشتن (بداند که می‌داند) با تأکید بر مورد دوم و سوم (خودآگاهی)، اورگان بر این باور است که تجربهٔ آگاهانه صرفاً بر پایهٔ مفهوم «خود» صورت می‌پذیرد (Oregan, 2011, pp. 71-80). در باب قابلیت برخورداری ماشین‌های مصنوعی غیربیولوژیکی و هوشمند از آگاهی، اورگان در یکی از مقالات شناخته‌شدهٔ خود بر این باور است که کاربست نظریهٔ حسی ـ حرکتی در ماشین‌های هوشمند نظیر ربات، از نظر منطقی هرچند دشوار، اما ممکن است. بااین‌حال، تحقق آن در هوش مصنوعی منوط به برخورداری ربات‌ها از دو مؤلفهٔ ذهنی است: وجود خود یا خویشتن و آگاهی پدیداری (Oregan, 2012, pp. 117-136).

از منظر آنتونیو داماسیو، آگاهی ویژگی پیچیده‌ای از مغز است که در حین برهم‌کنش میان مغز، بدن و محیط پدیدار می‌شود. بروز آگاهی منوط به وجود سه مؤلفهٔ اساسی است: ۱. بیداری؛ ۲. ذهن عملیاتی؛ ۳. حس خویشتن. خویشتن از نظرگاه داماسیو دارای درجات سلسله‌مراتبی است، درجاتی که با طبقه‌بندی سه‌گانهٔ خویشتن‌شناختی اورگان مشابهت دارد (Damasio, 1999; Cavana & Nani, 2022, pp. 143-144). در باب قابلیت برخورداری ماشین‌های هوشمند از آگاهی، داماسیو در یکی از جدیدترین آثار خود بر این باور است که «نظم عجیب اشیا، احساسات و عواطفی است که از ریشه‌ای بیولوژیکی سرچشمه گرفته است. هوش و آگاهی، ظرفیت خلقت فرهنگی انسان را تشکیل می‌دهند. به‌طور خلاصه، ذهن محاسباتی قلمرو معنای انسان بودن نیست.» از نظرگاه داماسیو، ذهن ما در دو «رجیستری» کار می‌کند: در یک ثبت، ما با ادراک، حرکت، خاطرات، استدلال، زبان‌های کلامی و زبان‌های ریاضی سروکار داریم. این رجیستر باید دقیق باشد و به‌راحتی بتوان آن را با شرایط محاسباتی توصیف کرد. این دنیای سیگنال‌های سیناپسی است که به‌خوبی توسط هوش مصنوعی و رباتیک شبیه‌سازی شده است؛ اما یک ثبت دوم هم وجود دارد که مربوط به عواطف و احساسات است؛ وضعیت زندگی در بدن زندهٔ ما را توصیف می‌کند و به‌راحتی خود را به یک حساب محاسباتی تقلیل نمی‌دهد. هوش مصنوعی و رباتیک فعلی به این ثبت دوم نمی‌پردازند (Damasio, 2018).

برخی پژوهشگران حوزهٔ هوش مصنوعی نیز با الهام از اندیشه‌های داماسیو، پلتفرم جدیدی تحت عنوان هوش مصنوعی عاطفی ـ اجتماعی (SEAI)[1] را طراحی و آزمایش کرده‌اند که بر مبنای آن، یک ربات باهوش اجتماعی باید بتواند اطلاعات معنادار را در زمان واقعی از محیط اجتماعی استخراج کند و مطابق با رفتار منسجم انسان‌مانند واکنش نشان دهد. علاوه‌براین، باید بتواند این اطلاعات را درونی کند، در سطح بالاتری دربارهٔ آن استدلال کند، نظرات خود را به‌طور مستقل بسازد و سپس به‌طور خودکار، تصمیم‌گیری را باتوجه‌به تجربهٔ منحصربه‌فرد خود سوگیری کند. برهمین‌اساس، تأثیر احساسات و عواطف مصنوعی و ارتباط آن‌ها با باورها و تصمیمات ربات در یک انسان‌نمای فیزیکی درگیر در تعامل انسان و ربات بررسی و برجسته خواهد شد (Cominelli et al., 2018, p. 1-20). یافته‌های این آزمایش و مدل‌سازی حاکی از آن است که وقتی SEAI به یک انسان‌نمای اجتماعی عواطف و احساسات مصنوعی را اعطا کرد، عامل موفق شد از آن‌ها برای ایجاد عقاید در مورد دنیای اجتماعی که در آن غوطه‌ور است بهره‌برداری کند و براساس آن‌ها بیشتر تجلی یابد. با SEAI، ربات‌ها می‌توانند از احساسات مصنوعی برای تصمیم‌گیری و ارزش‌گذاری تعاملات گذشتهٔ آن‌ها بهره‌مند شوند.

بدین‌ترتیب، از منظر عصب‌پژوهان پدیدارشناس، عاملیت اخلاقی محصول برهم‌کنش عاطفی و بدن‌مند ارگانیسم دارای خویشتن یا «خود» با محیط پیرامونی است. به‌جهت خاستگاه حس‌مندانه و بیولوژیکی مفهوم خویشتن (نکته‌ای که این رویکرد را از دوگانه‌گرایی جوهری متمایز می‌سازد) و غیربیولوژیکی‌بودن ماشین‌های هوشمند و فقدان عاطفه‌پذیری در آن‌ها، امکان تحقق آگاهی و خودآگاهی در آن‌ها وجود ندارد. برهمین‌اساس، با فقدان این مؤلفه‌های بیولوژیکی در ماشین‌های هوشمند، امکان تحقق عاملیت اخلاقی و مسئولیت‌پذیری آن‌ها در وضع فعلی میسر نیست؛ لیکن باتوجه‌به دستاوردها و پیشرفت‌های اخیر هوش مصنوعی در ساختن ماشین‌های هوشمند عاطفی در آینده، که امکان برهم‌کنش فیزیکی ـ عاطفی را برای ادراک محیط پیرامون فراهم می‌سازد، می‌توان موجبات زیست‌فناورانهٔ برخورداری هوش مصنوعی از عاملیت اخلاقی ـ حقوقی را فراهم دید.

1. Social Emotional Artificial Intelligence

نتیجه‌گیری

برآیند مشترک هر سه مکتب دوگانه‌گرایی، فیزیکالیسم تقلیل‌گرا و ناتقلیل‌گرا در زمینهٔ امکان‌سنجی برخورداری ماشین‌های هوشمند از عاملیت اخلاقی–حقوقی، حاکی از گرایشی غالب (هرچند با رویکردهای متفاوت) مبنی بر عدم پذیرش ماشین‌های هوشمند در وضع کنونی از عاملیت اخلاقی به‌جهت فقدان برخورداری از مؤلفه‌های ذهنی است. در فیزیکالیسم تقلیل‌گرا شاهد وجود دو طیف عمده هستیم:

طیف اول: این طیف برپایهٔ امکان شبیه‌سازی فیزیکی مغز و ذهن با سخت‌افزار و نرم‌افزارهای رایانه‌ای از حیث کارکردگرایی محاسباتی استوار است. در این رویکرد، قابلیت‌های ذهنی به قابلیت‌های مغزی فروکاسته می‌شوند و از عناصر متافیزیکی خارج می‌شوند. با وجود این، یکی از چالش‌های بزرگ تقلیل‌گرایان، مواجهه با مقولهٔ عاملیت اخلاقی و مسئولیت‌پذیری انسان نسبت به تصمیمات خود است؛ چراکه فروکاستن حالات ذهنی به حالات فیزیکی مغزی، به نفی اختیار و ارادهٔ آزاد و در نهایت باور به موجبیت و جبرگرایی منجر می‌شود؛ مسئله‌ای که به انکار عاملیت اخلاقی انسان و مسئولیت‌پذیری او منجر می‌شود. همین چالش در پذیرش یا انکار عاملیت اخلاقی ماشین‌های هوشمند نیز قابل تصور خواهد بود؛

طیف دوم: این طیف از تقلیل‌گرایان، باوجود باور به فروکاست‌پذیری حالات ذهنی به حالات نوروشیمیایی مغزی، مسئلهٔ قابلیت‌های مغزی انسان را برگرفته از خاستگاه بیولوژیکی انسان به‌عنوان نوعی ارگانیسم و موجود زنده می‌دانند و لذا در وضع فعلی، امکان تحقق آن را در ماشین‌های هوشمند نمی‌پذیرند. بدین ترتیب، با فقدان قابلیت‌های مغزی انسان (نظیر آگاهی) در ماشین‌های هوشمند، تصور برخورداری هوش مصنوعی از عاملیت اخلاقی–حقوقی، منتفی خواهد بود. ناتقلیل‌گرایان با سه مقدمه منکر عاملیت اخلاقی ماشین‌های هوشمند هستند:

۱. حالات ذهنی نظیر اراده و آگاهی هرچند به‌عنوان ویژگی‌های سطح بالا، برگرفته از ویژگی‌های سطح پایین فیزیکی هستند؛ اما به‌منظور جنبه‌های کیفی، اول‌شخص و التفاتی خود قابل فروکاستن به حالات کمی و سوم‌شخص مغزی نیستند؛ ۲. هوش مصنوعی در وضع کنونی فاقد حالات کیفی و التفاتی ذهنی (نظیر اراده و آگاهی) است و صرفاً بر مؤلفه‌های کمّی و فیزیکی محاسباتی متکی است؛ ۳. عاملیت اخلاقی منوط به برخورداری عامل از مؤلفه‌های ذهنی (اراده آزاد و آگاهی) است. نتیجه‌ای که از مقدمات فوق حاصل می‌شود این است که ماشین‌های هوشمند مصنوعی، فاقد عاملیت اخلاقی هستند.

دوگانه‌گرایان نیز با مقدمات فکری ناتقلیل‌گرایان موافق‌اند؛ با این تفاوت که برخلاف فیزیکالیسم ناتقلیل‌گرا، در مقدمهٔ اول، منشأ حالات کیفی ذهن را برگرفته از فعل‌وانفعالات سطوح پایین فیزیکی مغز و سیستم عصبی نمی‌دانند؛ بلکه آن‌را به مؤلفه‌ای متافیزیکی موسوم به خود یا نفس غیرمادی منتسب می‌دانند.

دراین‌میان، رویکرد مناسب از نظر نویسنده عصب‌شناسی پدیدارگرا است؛ رویکردی که به‌منزلهٔ سنتز فیزیکالیسم و دوگانه‌گرایی محسوب می‌شود و هم بر مؤلفه‌های فیزیکی (بیولوژیکی) و هم متافیزیکی (پدیدارشناختی) تأکید دارد. این رویکرد با افزودن حسّانیت و بدن‌مندی به مفهوم خویشتن و برهم‌کنش التفاتی ـ عاطفی ارگانیسم آگاه با محیط پیرامون، بر خاستگاه بیولوژیکی و درعین‌حال پدیدارشناختی مفاهیم ذهنی نظیر «خود»، آگاهی و خودآگاهی تأکید دارد و برهمین‌اساس با این دو مقدمه در وضع کنونی، منکر عاملیت اخلاقی ماشین‌های هوشمند است: ۱. مقدمهٔ عاطفی: عاملیت اخلاقی منوط به برخورداری از حس خویشتن است؛ ۲. مقدمهٔ پدیدارشناختی: عاملیت اخلاقی مستلزم برهم‌کنش بدن‌مندانهٔ ارگانیسم دارای حیث التفاتی با محیط پیرامونی به‌منظور ادراک معنا، تفسیر و پاسخ‌گویی به محرک‌های محیطی است؛ ۳. ماشین‌های هوشمند فاقد عناصر عاطفی و پدیدارشناختی هستند.

نتیجه اینکه هوش مصنوعی در وضع کنونی فاقد عاملیت اخلاقی–حقوقی است. با وجود این، برخلاف دوگانه‌گرایان جوهری، اندیشمندان مدافع فیزیکالیسم تقلیل‌گرا، ناتقلیل‌گرا و عصب‌شناسی پدیدارگرا بر این باورند که باتوجه‌به امکان‌پذیری تحقق سازمان عملکردی و بیولوژیکی مغز انسان در فناوری هوش مصنوعی قوی و منحصر نبودن این عملکرد به خواص نوروشیمیایی مغزی، در آینده با پیشرفت فناوری و امکان دستیابی هوش مصنوعی قوی به کیفیات و مؤلفه‌های ذهنی نظیر آگاهی، اراده و خودآگاهی، زمینه‌های برخورداری ماشین‌های هوشمند از

عاملیت اخلاقی و حقوقی فراهم خواهد شد. برهمین‌اساس، پیشرفت‌های متخصصان فناوری‌های هوشمند و دانشمندان عصب‌شناسی در آینده می‌تواند نویدبخش ورود هوش مصنوعی به عضویت جامعهٔ اخلاقی ـ حقوقی انسان‌ها باشد.

منابع

بیلی، اندرو (۱۳۹۹). متفکران فلسفه ذهن. (یاسر پوراسماعیل، مترجم). نشر لگا.

تگمارک، مکس (۱۴۰۱). زندگی: انسان بودن در عصر هوش مصنوعی. (میثم محمدامینی، مترجم). نشر نو.

شعری، زهرا و حسنی، محمد (۱۳۹۳). تبیین و نقد عاملیت اخلاقی در نظریه یادگیری شناختی اجتماعی. دوفصل‌نامهٔ پژوهشی علوم تربیتی از دیدگاه اسلام، ۲(۳)، ۹۳-۱۱۲. https://dor.isc.ac/dor/20.1001.1.24237396.1393.2.3.5.2

عموسلطانی، محمدمهدی و آذربایجانی، مسعود (۱۳۹۷). ارادهٔ آزاد، برخاسته از جوهر مجرد نفس یا ویژگی‌ای نوخاسته، جستارهای فلسفه دین، ۷(۱)، ۱۳۶-۱۱۱. لینک

قاسمی، حسین و ثقه‌الاسلامی، علیرضا (۱۴۰۰). تأثیر کلان‌داده‌ها بر تحول مفهومی عمل و عاملیت اخلاقی. اخلاق در علوم و فناوری، ۱۶(۱)، ۱۸-۲۴. https://dor.isc.ac/dor/20.1001.1.22517634.1400.16.2.4.9

کاوانا، اوجینیو و نانی، آندره (۱۴۰۱). نظریه‌های آگاهی. (سعید صباغی‌پور و عبدالرحمن نجل‌رحیم، مترجمان). نشر ارجمند.

مسلین، کیت (۱۳۹۵). فلسفهٔ ذهن. (مهدی ذاکری، مترجم). انتشارات علمی و فرهنگی.

مقدس، محمدمهدی (۱۳۹۹). تحلیل و بررسی امکان هوش مصنوعی هیدگری در آرای هیوبرت دریفوس. جستارهایی در فلسفه و کلام، ۱۰۴، ۱۶۱-۱۸۳. https://doi.org/10.22067/epk.2021.70220.1032

Angus, T. (2003). *Animals & Ethics: An Overview of the Philosophical Debate*. Broadview Press.

Armstrong, D. M. (1978). Naturalism, Materialism and First Philosophy. *Philosophia, 8*(2-3), 261–276. https://doi.org/10.1007/BF02379243

Arvan, M., & Maley, C. J. (2022). Panpsychism and AI consciousness. *Synthese, 200*(3), 1-22, https://doi.org/10.1007/s11229-022-03695-x

Bandura, A. (2006). Toward a psychology of human agency. *Perspectives on psychological science*, *1*(2), 164-180. https://doi.org/10.1111/j.1745-6916.2006.00011.x

Beck, J. (2019). Perception is analog. *The Journal of Philosophy*, *116*(6), 319-349. http://dx.doi.org/10.5840/jphil2019116621

Bellman, R. E. (1978). *An Introduction to Artificial Intelligence: Can Computers Think?*. Boyd and Fraser Publishing Company.

Beorlegui, C. (2009). Emergentism. *Pensamiento*, *65*(246), 881-914. Link

Chalmers, D. (1995). Facing up to the problem of Consciousness. *Journal of Consciousness Studies*, *2*(2), 200–219. https://doi.org/10.1098/rstb.2017.0342

Chiang, C.-C., Shivacharan, R. S., Wei, X., Gonzales-Reyes, L. E., & Durand, D. M. (2018). Slow Periodic Activity in the Longitudinal Hippocampal Slice Can Self-Propagate Non-Synaptically by a Mechanism Consistent with Ephatic Coupling. *The Journal of Physiology*, *597*(1).249-269, https://doi.org/10.1113/jp276904

Churchland, P. M. (1981). Eliminative Materialism and the Propositional Attitudes. *The Journal of Philosophy*, *78*(2), 67–90. http://dx.doi.org/10.4314/sajpem.v22i3.31370

Cominelli, L., Mazzei, D., & De Rossi, D. E. (2018). Social Emotional Artificial Intelligence Based on Damasio's Theory of Mind. *Frontiers in Robotics and AI, 5*, 1–20. http://dx.doi.org/10.3389/frobt.2018.00006

Damasio, A. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt.

Damasio, A. (2018). *The Strange Order of Things: Life, Feeling, and the Making of Cultures*. Knopf Doubleday Publishing Group.

Dreyfus, H. (1999). *What Computers Still Can't Do*. MIT Press.

Edelman, G. M. (1989). *The Remembered Present: A Biological Theory of Consciousness*. Basic Books.

Elster, J. (1991). Rationality and social norms. *European Journal of Sociology/Archives Europeennes de Sociologie*, *32*(1), 109-129. https://doi.org/10.1017/S0003975600006172

Engelen, B. (2007). Rationality and Institutions: An Inquiry into the Normative implications of Rational Choice Theory. *Erasmus Journal for Philosophy and Economics*. http://dx.doi.org/10.23941/ejpe.v1i1.19

Fischer, J., Kane, R., Pereboom, D., & Vargas, M. (2007). *Four Views on the Will* Metaphilosophy, *40*(1), 147 – 153. http://dx.doi.org/10.1111/j.1467-9973.2009.01564.x

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition*, *28*(1-2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Gallagher, S. (2007). Moral Agency, Self-Consciousness, and Practical Wisdom. *Journal of Consciousness Studies*, *14*(5-6), 199–223. Link

Gazzaniga, M. S. (2014). Mental Life and Responsibility in Real Time with a Determined Brain. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Volume 4: Free Will and Moral Responsibility*. Oxford University Press.

Goodman, L., & Caramenico, D. (2013). *Coming to Mind, The Soul and Its Body*. University of Chicago Press.

Haugeland, J. (1989). *Artificial intelligence: The very idea*. MIT press.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, *11*(1), 19-29. https://doi.org/10.1007/s10676-008-9167-5

Horst, S. (2005). The Computational Theory of Mind. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.Link: https://plato.stanford.edu/entries/computational-mind/

Kane, R. (1998). *The Significance of Free Will*. Oxford University Press.

Kim, J. (1993). *Supervenience and Mind: Selected Philosophical Essays*. Cambridge University Press.

Kim, J. (2008). *Physicalism, or Something Near Enough*. Princeton University Press. 10.1515/9781400840847

Nailer, T. (2022). *Moral Agency*. The University of Adelaide. Link

Nath, R. (2010). A Cartesian critique of the artificial intelligence. *Philosophical Papers and Reviews, 2*(3), 27–33.

Newell, A., & Simon, H. A. (1961). GPS, a program that simulates human thought. In H. Billing (Ed.), *Lernende Automaten* (pp. 109–124). Link

Northoff, G., & Gouveia, S. S. (2024). Does artificial intelligence exhibit basic fundamental subjectivity? A neurophilosophical argument. *Phenomenology and the Cognitive Sciences*, *23*(5), 1097-1118. https://doi.org/10.1007/s11097-024-09971-0

O'Regan, J. K. (2011). *Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness*. Oxford University Press.

O'Regan, J. K. (2012). How to Build a Robot that is Conscious and Feels. *Minds and Machines, 22*(2), 117-136 http://dx.doi.org/10.1007/s11023-012-9279-x

Oppy, G., & Dowe, D. (2011). The Turing Test. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. Link

Piccinini, G., & Bahar, S. (2012). Neural Computation and the Computational Theory of Cognition. *Cognitive Science A Multidisciplinary Journal 37(3)*, 453–488. https://doi.org/10.1111/cogs.12012

Putnam, H. (1975). *Mind, Language and Reality* (Vol. 2). Cambridge University Press.

Rajakishore, N., & Manna, R. (2021). From posthumanism to ethics of artificial intelligence. *AI and Society*, *38*(1), 185-196. https://doi.org/10.1007/s00146-021-01274-1

Rozemond, M. (2008). Descartes's Dualism. In J. Broughton & J. Carriero (Eds.), *A Companion to Descartes* (pp. 372–389). Blackwell Publishing.

Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence: A Modern*. Approach, 2nded., PrenticeHall, December2002.

Searle, J. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences*, *3*(3), 417– 424. https://doi.org/10.1017/S0140525X00005756

Searle, J. (1990). Is the Brain a Digital Computer? *Proceedings and Addresses of the American Philosophical Association*, *64*(3), 21–37. https://doi.org/10.2307/3130074

Searle, J. (2007). *Freedom and Neurobiology: Reflections on Free Will, Language, and Political Power*. Columbia University Press.

Stephan, A. (2010). Are Deliberations and Decisions Emergent, if Free? In A. Corradini & T. O'Connor (Eds.), *Emergence in Science and Philosophy* (pp. 207–228). Routledge.

Sugarman, J. (2005). Persons and moral agency. *Theory & Psychology*, *15*(6), 793-811. https://psycnet.apa.org/doi/10.1177/0959354305059333

Tononi, G. (2004). An Information Integration Theory of Consciousness. *BMC Neuroscience, 5* (42), 1–22. https://doi.org/10.1186/1471-2202-5-42

Tshilidzi, M. (2017). *Rational Choice and Artificial Intelligence*. University of Johannesburg. http://dx.doi.org/10.48550/arXiv.1703.10098

Van Inwagen, P. (1983). *An Essay on Free Will*. Clarendon Press.

Vargas, M. (2004). Responsibility and the aims of theory and revisionism. *Pacific Philosophical Quarterly*, *85*(2), 218–241. http://dx.doi.org/10.1111/j.0279-0750.2004.00195.x

Wróblewski, Z., & Fortuna, P. (2023). Moral Subjectivity and the Moral Status of Artificial Intelligence: A Philosophical and Psychological Perspective. *Journal of LATEX Templates*, *4*(144), 29–48. http://dx.doi.org/10.12887/36-2023-4-144-05

Yi, Z etall. (2024). Brain-inspired and Self-based Artificial Intelligence. *Journal of LATEX Templates*. 1.1-31 https://doi.org/10.48550/arXiv.2402.18784

[In Persian]

Amosultani, M. M., & Azarbaijani, M. (2018). 'Free Will, Arising from the Abstract Essence of the Soul or a Newly Emerged Feature', *Essays on the Philosophy of Religion*, *7*(1), 111-136. Link

Bailey, A. (2012). *Thinkers in the Philosophy of Mind*. (Yaser Pourasmaeil, translator). Lega Publishing.

Cavana, E. & Nani, A (2022). *Theories of Consciousness*. (Saeed Sabaghipour and Abdolrahman Najl-Rahim, translators). Arjomand Publishing.

Ghasemi, H., & Thaqe-ol-Eslami, A. (2021). The impact of big data on the conceptual evolution of ethical action and agency. *Ethics in Science and Technology*, *16*(1), 18-24. https://dor.isc.ac/dor/20.1001.1.22517634.1400.16.2.4.9

Maslin, K. (2016). *Philosophy of Mind*. (Mehdi Zakeri, translator). Scientific and Cultural Publications.

Sheeri, Z., & Hassani, M. (2014). Explanation and Criticism of Moral Agency in Social Cognitive Learning Theory. *Bi-Quarterly Research Journal of Educational Sciences from an Islamic Perspective*, *2*(3), 93-112. https://dor.isc.ac/dor/20.1001.1.24237396.1393.2.3.5.2

Tegmark, M. (1401). *Life: Being Human in the Age of Artificial Intelligence*. (Meysam Mohammad Amini, translator). Nashre No.