Journal of Philosophical Investigations
Print ISSN: 2251-7960 Online ISSN: 2423-4419
Homepage: https://philosophy.tabrizu.ac.ir

The Possibility of Moral Responsibility in Artificial Intelligence: a philosophical explanation of the responsibility gap

Sajedeh Firoozbakht[1] | **Seyed Mostafa Shahraeini**[2] ✉

1. PhD in Philosophy, University of Tabriz, Tabriz, Iran. Email: sajede8897869@gmail.com
2. Corresponding Author, Professor of Western Philosophy Department, Research Institute for Humanities and Cultural Studies, Tehran, Iran. Email: m-shahraeini@Ihcs.ac.ir

Article Info

Article type:
Research Article

Article history:
Received 14 December 2025
Received in revised from 11 January 2026
Accepted 15 January 2026
Published online 14 July 2026

Keywords:
Algorithmic Accountability, Participatory Agency, Moral Responsibility, Moral Mediation, Artificial Intelligence.

ABSTRACT

In recent years, the emergence and rapid use of Artificial Intelligence (AI) systems have posed a fundamental question for moral philosophy and the technology: how can moral responsibility be meaningfully articulated in a context where decision-making and action are carried out by non-human entities? In the classical philosophical tradition, moral responsibility has been intrinsically linked to human agency, intention, consciousness, and free will; however, the advent of intelligent technologies has profoundly challenged this traditional linkage. Aiming to re-conceptualize moral responsibility in relation to artificial intelligence, this article seeks to advance a philosophical account of the challenges surrounding it—commonly referred to as the "responsibility gap". The research methodology adopted in this article is analytical–interpretive and grounded in the philosophy of technology and AI ethics. The findings indicate that within human–machine interactive networks, classical models of responsibility are no longer sufficient. Instead, responsibility must be understood within a relational, intersubjective, and distributed framework encompassing both human and non-human agents. In the subsequent stage, through the analysis of two applied case studies, concrete instances of the responsibility gap are examined. Based on this analysis, the article proposes a three-level model of moral responsibility—design-level, organizational-level, and institutional-level responsibility. The study aims to demonstrate that, in the face of emerging technologies, the question of a single moral agent gives way to an inquiry into a network of relations, influences, and reciprocal forms of accountability. In conclusion, the article argues that the possibility of moral responsibility in artificial intelligence can only be defended if one moves beyond the classical individual-centered approach and reconceptualizes responsibility as an institutional and participatory mechanism embedded within the full constellation of human and non-human factors.

Cite this article: Firoozbakht, S. & Shahraeini, M. (2026). The Possibility of Moral Responsibility in Artificial Intelligence: A Philosophical Explanation of the Responsibility Gap. *Journal of Philosophical Investigations*, 20 (55), 241-264. https://doi.org/10.22034/jpiut.2026.70724.4376

 Publisher: University of Tabriz.

Extended Abstract

A philosophical analysis of the possibility of moral responsibility in the context of Artificial intelligence (AI), with particular attention to what has been described in contemporary literature as the "responsibility gap" is one of the most important topics of the ages. As AI systems increasingly participate in decision-making processes that have morally significant consequences—such as in criminal justice, healthcare, finance, and public administration—traditional frameworks of moral responsibility appear insufficient to account for how responsibility should be attributed when harm occurs. The article argues that this difficulty is not accidental or merely practical, but instead reflects a deeper conceptual tension between classical theories of moral responsibility and the socio-technical nature of AI systems. The discussion begins by outlining the dominant assumptions underlying classical accounts of moral responsibility. Traditionally, moral responsibility presupposes the existence of a human moral agent who possesses intentionality, rational deliberation, sufficient epistemic access to relevant facts, and control over actions and their outcomes. On this view, responsibility is typically attributed to an individual who freely performs an action with foreseeable consequences. The article shows that these assumptions are increasingly challenged by AI systems, especially those based on machine learning, which operate in opaque ways, adapt over time, and generate outcomes that are not fully predictable or traceable to a single human decision. Against this background, the article introduces the concept of the responsibility gap, understood as a situation in which morally significant harm occurs, yet no single agent appears to satisfy the standard conditions for moral responsibility. Importantly, the article emphasizes that this gap should not be interpreted as evidence that responsibility disappears altogether. Rather, it indicates that prevailing moral concepts may be ill-suited to technologically mediated forms of action. The responsibility gap thus functions as a philosophical problem that calls for conceptual revision rather than moral resignation.

The article then critically examines several prominent responses to the responsibility gap. One such response involves attributing a form of moral or quasi-moral agency to AI systems themselves. Proponents of artificial moral agency argue that sufficiently autonomous systems may bear responsibility in their own right. The article rejects this approach, contending that current AI systems lack the normative capacities—such as moral understanding, intentionality, and accountability—that are required for genuine moral agency. Treating AI as morally responsible risks obscuring human and institutional accountability rather than clarifying it. Another response focuses on locating responsibility exclusively in human actors, such as programmers, designers, operators, or end users. While this strategy preserves the human-centered nature of moral responsibility, the article argues that it often relies on overly simplistic causal models. In complex AI systems, outcomes emerge from interactions among multiple agents, technical components, datasets, organizational practices, and regulatory environments. Isolating responsibility at a single point in this network fails to capture how moral influence is distributed across the system as a whole. To move beyond these limitations, the article draws on contemporary philosophy of technology, particularly theories of technological mediation, actor–network theory, and the philosophy of information. These approaches emphasize that technologies are not morally neutral tools but actively shape human perception, reasoning, and action. AI systems, in this sense, mediate moral agency by influencing what agents can know,

what options are available to them, and how decisions are framed. Responsibility, therefore, must be understood as emerging within human–technology relations rather than residing solely in isolated individuals. Building on this insight, the article argues for a relational and distributed conception of moral responsibility. On this view, responsibility is not eliminated by automation but reconfigured across multiple levels, including system design, data governance, organizational deployment, and institutional oversight. The article illustrates this claim through discussion of real-world domains such as algorithmic risk assessment in criminal justice and AI-assisted medical decision-making. In these cases, morally relevant outcomes depend on a chain of decisions involving engineers, policymakers, institutions, and users, none of whom alone fully control the system's behavior.

The article further distinguishes between individual moral responsibility and forms of systemic or institutional responsibility. While individuals remain morally relevant, many ethical failures associated with AI arise from structural conditions, such as incentive systems, regulatory gaps, or organizational cultures. Consequently, the article argues that ethical analysis must shift from a narrow focus on blame toward broader questions of accountability, governance, and responsibility allocation. In its concluding sections, the article proposes a multi-level framework for moral responsibility in AI. This framework integrates ethical design principles, organizational responsibility, and public governance mechanisms. Rather than asking whether AI systems themselves can be morally responsible, the article emphasizes the importance of algorithmic accountability, transparency, and participatory oversight. Moral responsibility, on this account, is best understood as a shared and ongoing practice embedded in socio-technical systems. The article concludes that the responsibility gap does not undermine the possibility of moral responsibility in AI but reveals the need for a revised ethical framework. By reconceptualizing responsibility as relational, distributed, and institutional, the article offers a philosophically grounded response to one of the most pressing ethical challenges posed by contemporary artificial intelligence.

پژوهش‌های فلسفی
شاپا چاپی: ۷۹۶۰-۲۲۵۱ شاپا الکترونیکی: ۴۴۱۹-۲۴۲۳
Homepage: https://philosophy.tabrizu.ac.ir
Journal of Philosophical Investigations

امکان مسئولیت اخلاقی در هوش مصنوعی: تبیین فلسفی شکاف مسئولیت

ساجده فیروز بخت[۱] | مصطفی شهرآیینی[۲]✉

۱. دکتری فلسفه و پژوهشگر اخلاق هوش مصنوعی، دانشگاه تبریز، تبریز، ایران. رایانامه: sajede8897869@ gmail.com

۲. نویسنده مسئول، استاد فلسفه پژوهشکده فلسفه غرب، پژوهشگاه علوم انسانی و مطالعات فرهنگی، تهران، ایران. رایانامه: m-shahraeini@Ihcs.ac.ir

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۴۰۴/۰۹/۲۳
تاریخ بازنگری: ۱۴۰۴/۱۰/۲۱
تاریخ پذیرش: ۱۴۰۴/۱۰/۲۵
تاریخ انتشار: ۱۴۰۵/۰۴/۲۴

کلیدواژه‌ها:
پاسخ‌گویی الگوریتمی، فاعلیت مشارکتی، مسئولیت اخلاقی، میانجی‌گری اخلاقی، هوش مصنوعی.

چکیده

در سال‌های اخیر، ظهور و گسترش سامانه‌های هوش مصنوعی پرسشی بنیادی را پیش‌روی فلسفه اخلاق و فلسفه فناوری نهاده است: چگونه می‌توان از مسئولیت اخلاقی در بستری سخن گفت که در آن، تصمیم‌گیری و کنش از سوی موجوداتی غیرانسانی انجام می‌شود؟ در سنت فلسفی کلاسیک، مسئولیت اخلاقی همواره با فاعلیت انسانی، قصد، آگاهی و آزادی اراده پیوند داشته، اما ورود فناوری‌های هوشمند این پیوند سنتی را به چالش کشیده است. مقاله حاضر با هدف بازتعریف مسئولیت اخلاقی در نسبت با هوش مصنوعی، می‌کوشد در جهت تبیین فلسفی چالش‌های پیرامون آن یا به‌اصطلاح «شکاف مسئولیتی» قدم بردارد. روش پژوهش در این مقاله، تحلیلی–تفسیری و مبتنی بر فلسفه فناوری و اخلاق هوش مصنوعی است. یافته‌های مقاله نشان می‌دهد که در شبکه‌های تعاملی انسان ـ ماشین، مدل‌های کلاسیک مسئولیت، دیگر کفایت نداشته و در عوض، می‌بایست آن را در بستر رابطه میان‌ذاتی و توزیع‌شده میان فاعلان انسانی و غیرانسانی فهمید. در مرحله بعد، با تحلیل دو نمونه کاربردی، مصادیق واقعی شکاف مسئولیتی بررسی شده و در نهایت، مدل سه‌سطحی مسئولیت اخلاقی طراحی، سازمانی و نهادی پیشنهاد می‌گردد. این پژوهش در پی آن است نشان دهد که در مواجهه با فناوری‌های نوین، پرسش از فاعل منفرد جای خود را به پرسش از شبکه‌ای از روابط، اثرگذاری‌ها و پاسخ‌گویی‌های متقابل می‌دهد. در خاتمه، مقاله چنین استنتاج می‌کند که امکان مسئولیت اخلاقی در هوش مصنوعی تنها در صورتی قابل دفاع است که از رویکرد فردمحور کلاسیک فاصله گرفته و مسئولیت به‌عنوان سازوکاری نهادی، مشارکتی و در بستر تمامی عوامل انسانی و غیرانسانی بازتعریف گردد.

استناد: فیروزه‌بخت، ساجده و شهرآیینی، سید مصطفی. (۱۴۰۵). امکان مسئولیت اخلاقی در هوش مصنوعی: تبیین فلسفی شکاف مسئولیت، *پژوهش‌های فلسفی*، ۲۰ (۵۵)، ۲۴۱-۲۶۴. https://doi.org/10.22034/jpiut.2026.70724.4376

ناشر: دانشگاه تبریز.

مقدمه

پرسش از امکان و معنای مسئولیت اخلاقی در هوش مصنوعی یکی از جدی‌ترین و بنیادی‌ترین چالش‌های فلسفه فناوری معاصر[۱]است، زیرا مستقیما مفاهیمی را هدف می‌گیرد که در سنت اخلاقی کلاسیک، بدیهی و کم‌چالش تلقی می‌شدند. از زمان پیدایش سیستم‌های خودآموز و الگوریتم‌های تصمیم‌گیر، تمایز میان «فاعل انسانی دارای نیت، اراده و آگاهی» و «ابزار فاقد آن» دیگر قادر به توضیح ساختار واقعی کنش اخلاقی در جهان فناورانه امروزی نیست. در بسیاری از نظام‌های هوشمند، کنش نه محصول تصمیم مستقیم یک فاعل انسانی واحد، بلکه حاصل برهم‌کنش و تعامل پیچیده طراحان، کاربران، داده‌ها، مدل‌های یادگیرنده، زیرساخت‌های نهادی و محیط‌های اجرایی است. این پدیده، که در ادبیات فلسفی معاصر با عنوان «فاعلیت توزیع‌شده»[۲]میان همه فاعلان یا عاملان شناخته می‌شود، مرزهای سنتی اخلاق و مسئولیت را به پرسش کشیده[۳] و موجب می‌شود که نسبت‌های علّی و هنجاری میان تصمیم، کنش و پیامد اخلاقی دیگر به سادگی روشن و قابل ردیابی نبوده و به‌آسانی نتوان مسئول نهایی امور را مشخص کرد[۴] در چنین وضعیتی، تعیین این‌که «چه کسی واجد مسئولیت اخلاقی است؟ طراح، کاربر، سازمان، یا خود سامانه؟» به طرز بنیادینی مبهم می‌گردد. این امر در نظر اندیشمندان و فیلسوفان معاصر به «شکاف مسئولیتی یا شکاف در مسئولیت»[۵]معروف شده است (فلوریدی و سندرز، ۲۰۰۴، ۳۶۴)، شکافی که نه ناشی از بی‌دقتی انسانی، بلکه نتیجه ساختار فناورانه کنش در سامانه‌های هوشمند بوده و مسئله اصلی‌اش، این است که در شرایطی که کنش اخلاقی حاصل تعامل انسان و ماشین است، مسئولیت چگونه باید تعریف و تخصیص یابد؟

اهمیت فلسفی این مسئله از این‌جا ناشی می‌شود که چهارچوب‌های کلاسیک مسئولیت اخلاقی، اعم از رویکردهای فضیلت‌محور، وظیفه‌گرایانه یا پیامدگرایان، همگی بر پیش‌فرض وجود فاعلی واحد با نیت آگاهانه و کنترل معنادار بر کنش استوارند. هنگامی که این پیش‌فرض در عمل فرو می‌ریزد، نه‌تنها انتساب مسئولیت دشوار می‌شود، بلکه خود معنای مسئولیت اخلاقی نیازمند بازاندیشی خواهد شد. در واقع، اگر مسئولیت را هم‌چون سنت ارسطویی و کانتی (با وجود اختلاف در تفسیر معنایی این مفهوم که در ادامه و در قسمت مبانی فلسفی مسئولیت اخلاقی، روشن خواهد شد)، وابسته به نیت و اراده انسانی بدانیم، چگونه می‌توانیم آن را به سامانه‌هایی نسبت دهیم که آگاهی یا نیت ندارند، اما تصمیم‌های اخلاقی مهم می‌گیرند؟ به عبارت دیگر، نسبت دادن مسئولیت از منظر رویکردهای سنتی اخلاقی به سامانه‌هایی که فاقد آگاهی و قصد اخلاقی هستند، ناموجه به‌نظر می‌رسد؛ اما اگر مسئولیت را صرفا به سازوکارهای نهادی یا ساختاری فرو بکاهیم، آن‌گاه نقش فاعلیت انسانی، وجدان اخلاقی و پاسخ‌گویی فردی به‌طرز نگران‌کننده‌ای تضعیف می‌شود (یوناس، ۱۹۸۴، ۱۲۵–۱۲۲)؟ از این‌رو، مسئله مسئولیت در هوش مصنوعی نه صرفا یک مشکل فنی یا حقوقی، بلکه در بنیان خود مسئله‌ای فلسفی است که فلسفه فناوری را ناگزیر می‌سازد که مفاهیمی چون فاعلیت، کنش، و مسئولیت را در پرتو فناوری‌های خودمختار بازتعریف کند. در پاسخ به این چالش،

---

[1] Philosophy of technology

[2] distributed agency

[۳] در فلسفه کلاسیک، مسئولیت اخلاقی مبتنی بر سه مؤلفه کلیدی است: آگاهی از کنش، اراده آزاد، و نسبت علّی میان فاعل و پیامد عمل. این سه شرط، در بستر هوش مصنوعی دیگر به سادگی برقرار نیستند.

[۴] در واقع، در بسیاری از تصمیم‌های حساس، از اعطای وام بانکی گرفته تا حوزه مربوط به خدمات شهری، بهداشت، پزشکی و تشخیص بیماری یا حتی صدور احکام قضایی، دیگر فقط انسان‌ها تصمیم‌گیرنده نیستند و الگوریتم‌ها و سامانه‌های هوشمند نیز به‌صورت فعال در زنجیره تصمیم‌گیری مشارکت می‌کنند. این دگرگونی سبب شکل‌گیری وضعیتی می‌شود که درصورت وقوع آسیب، به‌آسانی نتوان مسئول اخلاقی را تعیین نمود؛ به عبارت دیگر، تشخیص مسئول اخلاقی در سامانه‌های مبتنی بر هوش مصنوعی به‌آسانی ممکن نبوده، و در صورت وقوع حادثه، مشخص نیست که مسئولیت متوجه چه کسی است؟ کاربران؟ سازمان طراح‌کننده؟ نهاد نظارتی مربوطه؟ الگوریتم‌ها و سامانه‌ها؟

[5] Responsibility Gap

این مقاله با الهام از دیدگاه نظریه‌پردازانی چون جیمز مور[۱] و لوچیانو فلوریدی[۲] پیشنهاد می‌کند که مسئولیت اخلاقی در عصر هوش مصنوعی باید در چهارچوبی شبکه‌ای، میانجی‌گرایانه و چندسطحی فهم شود. در این چهارچوب، که گاه از آن با عنوان «اخلاق توزیع‌شده»[۳] یاد می‌شود، مسئولیت نه به یک فاعل منفرد، بلکه به شبکه‌ای از عاملان انسانی و غیرانسانی نسبت داده می‌شود. رویکردهای پدیدارشناسانه[۴] و پساپدیدارشناسانه[۵] در فلسفه فناوری، به‌ویژه در آثار دون آیدی[۶] و پیتر–پل فربیک[۷] این ایده را تقویت می‌کنند که فناوری‌ها صرفا ابزارهای خنثی نیستند، بلکه به‌طور فعال در شکل‌دهی تجربه، ادراک و تصمیم‌گیری اخلاقی انسان‌ها نقش میانجی ایفا می‌کنند.[۸] در ادامه، بیش‌تر به توضیح این رویکردها خواهیم پرداخت.

مقاله حاضر می‌کوشد ابتدا از ره‌گذر تحلیل تطبیقی دیدگاه‌های فربیک، فلوریدی، لاتور و دیگر اندیشمندان، بازخوانی تازه‌ای از مفهوم مسئولیت اخلاقی در هوش مصنوعی ارائه دهد. چنین خوانشی، علاوه بر توضیح مفاهیم میانجی‌گری اخلاقی[۹] و مسئولیت توزیع‌شده،[۱۰] بر سه مفهوم بنیادی دیگر نیز، استوار است: الف): فاعلیت مشارکتی،[۱۱] به‌عنوان شکلی از فاعلیت که در آن کنش اخلاقی محصول تعامل چندسطحی میان انسان، ماشین و ساختارهای اجتماعی است؛ ب): ادغام بین‌فهمی،[۱۲] که بر امکان هم‌فهمی و هم‌تصمیمی انسان و سامانه‌های هوشمند دلالت دارد؛ ج): پاسخ‌گویی الگوریتمی،[۱۳] به‌عنوان الگوی نوینی برای فهم اخلاقی مسئولیت در سطح سیستم‌ها و شبکه‌ها به‌کار می‌رود. سپس، سامانه‌های COMPAS و IBM Watson Health به‌جهت مسأله شکاف مسئولیتی به‌صورت اجمالی معرفی شده، و در خاتمه نیز، مدل سه‌سطحی مسئولیت اخلاقی پیشنهاد گردیده و چنین استنتاج می‌شود که واکاوی مفاهیم مذکور و تحلیل نسبت آن‌ها با سنت‌های اخلاقی کلاسیک، نشان می‌دهد که مفهوم مسئولیت اخلاقی در عصر هوش مصنوعی نیازمند بازسازی معرفت‌شناسانه است؛ به‌گونه‌ای که اخلاق می‌بایست نه به‌عنوان مجموعه‌ای از قواعد بیرونی، بلکه به‌مثابه رابطه‌ای زنده و درحال تکوین میان انسان و فناوری در نظر گرفته شود.

---

[1] James Moor
[2] Luciano Floridi
[3] distributed morality
[4] Phenomenological theory
[5] Post-phenomenological theor
[6] Don Ihde
[7] Peter- Paul Verbeek

[۸] از نظر فربیک، ابزارهای فناورانه شریک در فاعلیت در ساختار کنش اخلاقی محسوب می‌شوند. این دیدگاه راه را برای فهم تازه‌ای از مسئولیت می‌گشاید: مسئولیتی که در آن انسان و فناوری در فرآیند اخلاقی درهم تنیده‌اند، و پاسخ‌گویی اخلاقی نه در سطح فردی، بلکه در سطح روابط و میانجی‌گری‌ها شکل می‌گیرد (فربیک، ۲۰۰۵ ۱۳۰–۱۱۱؛ فربیک،۲۰۱۱، ۸۰–۶۷). افزون بر این، نظریه شبکه کنش‌گر یا ((Actor-Network theory (ANT) برونو لاتور (Bruno Latour) با تاکید بر «غیرانسان‌ها» یا (non-humans) به‌عنوان کنش‌گران شبکه، بنیان‌های متافیزیکی انسان‌محور اخلاق را به چالش کشیده است. از این منظر، مسئولیت اخلاقی باید در چهارچوبی گسترده‌تر از انسان تعریف شود؛ چهارچوبی که در آن، هر عنصرِ درگیر در فرآیند تصمیم‌گیری، از نرم‌افزار و الگوریتم گرفته تا داده و ساختار نهادی، در شکل‌گیری پیامد اخلاقی سهم دارد.

[9] Moral Mediation
[10] Distribute responsibility
[11] participatory agency
[12] interunderstanding integration

[۱۳] algorithmic accountability پاسخ‌گویی الگوریتمی به مجموعه‌ای از رویه‌ها و سازوکارها اشاره دارد که هدف‌شان شفاف‌سازی، نظارت، و پاسخ‌گو ساختن الگوریتم‌ها یا مدل‌های هوش مصنوعی در برابر کاربران، نهادهای نظارتی یا جامعه است. این مفهوم عمدتاً حول محور کد، داده، طراحی و عملکرد الگوریتم‌ها متمرکز بوده و چهار عنصر کلیدی دارد: الف): قابلیت ممیزی، به این معنا که آیا می‌توان بررسی کرد که الگوریتم‌ها چگونه تصمیم می‌گیرند؟ ب): شفافیت فنی، به معنای مستندسازی ورودی‌ها، پارامترها، منطق تصمیم‌گیری؛ ج): تخصیص مسئولیت مهندسی، به این معنا که چه کسی کد را نوشته است؟ یا چه کسی مدل را آموزش داده است؟ د): سازوکار بازنگری، به این معنا که آیا امکان تصحیح خطاها یا پاسخ‌گویی حقوقی وجود دارد؟

پیشینه پژوهش

هوش مصنوعی از آن دسته موضوعاتی است که جنبه‌های گوناگونی برای بحث و گفت‌وگو دارد، و در این زمینه، در زبان فارسی با ادبیات متنوع و بسیار غنی روبه‌رو هستیم. بررسی‌ها نشان می‌دهد که پژوهش‌های قابل‌توجهی در زمینه امکان آگاهی، فاعلیت، شخص‌بودگی، پراگماتیسم، عواطف و ظرفیت اخلاقی در هوش مصنوعی و حتی در حوزه‌های تخصصی آموزش، معماری، بهداشت و سلامت، هنر، بانکداری و عدالت کیفری انجام پذیرفته است[۱] اما آن‌چه در ادبیات هوش مصنوعی در زبان فارسی مغفول مانده یا کم‌تر بدان پرداخته‌اند، موضوع مسئولیت اخلاقی در سیستم‌های مبتنی بر هوش مصنوعی است؛ موضوعی که در منابع غربی به‌ویژه در زبان انگلیسی با عنوان «حاکمیت مسئولانه هوش مصنوعی» یا «هوش مصنوعی مسئول» با هدف مدیریت مسئولانه هوش مصنوعی از طریق تعریف شیوه‌های ساختاری با هدف توسعه تحقیقات آینده فراوان به‌چشم می‌خورد. با این‌حال، نکته متمایزکننده و مهم مقاله پیش‌رو، تلاش برای توضیح و تبیین روشن و منسجم مسئولیت اخلاقی، ضمن تمرکز بر شکاف مسئولیتی همراه با نمونه‌های موردی می‌باشد.

روش پژوهش

این مقاله از روش تحلیلی–فلسفی بهره می‌گیرد و در زمره پژوهش‌های نظری در اخلاق هوش مصنوعی و فلسفه فناوری قرار می‌گیرد. تمرکز اصلی پژوهش بر تحلیل مفهومی مسئولیت اخلاقی در بستر سامانه‌های هوش مصنوعی است، نه بررسی تجربی یا فنی. روش غالب مقاله تحلیل استدلالی و مقایسه‌ای است؛ به این معنا که هر موضع نظری در پرتو معیارهای فلسفی مسئولیت اخلاقی سنجیده می‌شود. در عین حال، مقاله از رویکردی بین‌رشته‌ای محدود اما هدفمند بهره می‌برد و از مفاهیم رایج در مطالعات هوش مصنوعی و حکمرانی فناوری استفاده می‌کند، بی‌آن‌که وارد تحلیل‌های فنی یا مطالعات موردی تجربی شود. هدف نهایی روش‌شناسی مقاله، روشن‌کردن امکان یا عدم امکان نسبت‌دادن مسئولیت اخلاقی و چگونگی تفسیر آن در شرایط کنش‌های خودکار و پیچیده هوش مصنوعی است.

یافته‌های پژوهش

یافته نخست مقاله تأیید وجود شکاف مسئولیت اخلاقی در سامانه‌های هوش مصنوعی است. مقاله نشان می‌دهد که با افزایش خودمختاری و پیچیدگی تصمیم‌گیری در این سامانه‌ها، الگوهای سنتی نسبت‌دادن مسئولیت اخلاقی کارایی خود را از دست می‌دهند. یافته دوم آن است که هوش مصنوعی فاقد شرایط لازم برای برخورداری از مسئولیت اخلاقی مستقل و منفرد است. بر اساس تحلیل فلسفی، سامانه‌های هوش مصنوعی، علی‌رغم توانایی یادگیری و تصمیم‌گیری، فاقد قصد اخلاقی، آگاهی هنجاری و درک الزام اخلاقی‌اند و بنابراین نمی‌توان آن‌ها را به‌عنوان مسئول اخلاقی اصیل در نظر گرفت. یافته سوم نشان می‌دهد که نسبت‌دادن مسئولیت اخلاقی به یک فاعل انسانی منفرد مانند طراح یا کاربر نهایی نیز، در بسیاری از موارد ناکافی است، و کنش‌های مبتنی بر هوش مصنوعی حاصل شبکه‌ای پیچیده از تصمیمات انسانی، نهادی و فناورانه‌اند و تقلیل مسئولیت به یک فرد، موجب نادیده‌گرفتن ابعاد ساختاری و سیستمی مسئولیت می‌شود. چهارمین یافته مهم مقاله، ضرورت بازاندیشی در مفهوم سنتی مسئولیت اخلاقی است. مقاله نتیجه می‌گیرد که شکاف مسئولیت نه با نفی مسئولیت، بلکه با بازتعریف آن قابل پاسخ

---

[۱] این عناوین کلّی عبارتند از: «مدیریت آموزشی» (عسگری و بهرامی، ۱۴۰۴)، «بهبود فرآیندهای معاونت مدارس» (ایدلی، ۱۴۰۴)، «آموزش هوشمند» (سمرباف‌زاده، ۱۴۰۴)، «تاثیر یادگیری ماشین بر طراحی و ساخت‌وساز» (سپهری قوجه، ۱۴۰۳)، «رویکرد فراترکیب در مدیریت سالمندان» (شرفی و محمدیاری، ۱۴۰۴)، «تولید محتوای مفهومی در انیمیشن» (عمرانی و شاه‌کرمی،۱۴۰۴)، «چگونگی افزایش عملکرد نظارتی در سیستم بانکداری ایران» (نصر اصفهانی، قائمی اصل، منتظر و اسماعیلی، ۱۴۰۴)، «سیاست‌گذاری عدالت کیفری برای ضابطین دادگستری» (امیریان فارسانی،۱۴۰۴)

است. مسئولیت اخلاقی باید به‌صورت توزیع‌شده، سیستمی و چندسطحی فهم شود؛ به‌گونه‌ای که نقش انسان‌ها، نهادهای نظارتی و ساختارهای تصمیم‌گیری هم‌زمان در نظر گرفته شود.

۱. مبانی فلسفی مسئولیت اخلاقی: از فاعل انسانی تا فاعلیت فناورانه

«مسئولیت اخلاقی» در سنت فلسفی، مفهومی است که ریشه در در درک انسان از فاعلیت، آزادی و نسبت علّی میان کنش و پیامد دارد. از ارسطو تا کانت، با وجود اختلاف در مبنای مسئولیت اخلاقی – از تاکید ارسطو بر اختیار عملی و آگاهی گرفته تا تاکید کانت بر خودآیینی عقلانی و قانون اخلاقی– این مفهوم اجمالا به انسانی نسبت داده می‌شود که آگاهانه و با قصد و نیت، عملی را انجام داده و در برابر پیامد آن پاسخ‌گو است. در کتاب *اخلاق نیکوماخوس* ارسطو، «مسئولیت» تنها زمانی معنا دارد که کنش‌گر بتواند میان گزینه‌های مختلف تصمیم بگیرد و بداند چرا و برای چه هدفی عمل می‌کند (ارسطو، ۱۸۹۵، ۱۱۱۰الف–۱۱۱۱ب). کانت نیز در کتاب *بنیاد مابعدالطبیعه اخلاق*، مسئولیت اخلاقی را با مفهوم «اراده آزاد» و «قانون اخلاقی درونی» مرتبط دانسته و آن را ملازم خودآیینی عقلانی می‌داند[۱](کانت، ۱۷۸۵، ۴۱).

در عصر فناوری‌های خودمختار، این الگوی انسان‌محور به چالش کشیده شده است. با ظهور هوش مصنوعی و سامانه‌های تصمیم‌گیر خودآموز، دیگر نمی‌توان به‌سادگی نسبت میان فاعل، کنش و پاسخ‌گویی را در قالب سنتی توضیح داد (مور، ۲۰۰۶، ۱۶). مور از نخستین فیلسوفانی بود که به این مسئله اشاره کرد و مفهوم «فاعل اخلاقی مصنوعی»[۲]را مطرح نمود. از نظر او، هرچند ماشین‌ها واجد آگاهی و نیت انسانی نیستند، اما در سطحی از پیچیدگی می‌توانند در فرآیندهای اخلاقی مؤثر باشند و لذا نوعی فاعلیت تابعی یا ثانویه[۳]را دارا هستند. این رویه و دیدگاه به‌ویژه در فلسفه فناوری معاصر گسترش یافت. دون آیدی در چهارچوب پدیدارشناسی فناوری[۴]نشان می‌دهد که ابزارهای فناورانه یا تکنولوژیک در تجربه و تصمیم انسانی میانجی و واسطه هستند و جهت‌گیری ادراک و کنش ما را دگرگون می‌کنند (آیدی، ۱۹۹۰، ۳۵–۳۰). بر این اساس، فناوری صرفا ابزاری خنثی نیست، بلکه در ساختار کنش اخلاقی سهیم است.

فربیک ضمن بسط دیدگاه فوق در آثار خود به‌ویژه در *اخلاق فناوری*، با به‌کارگیری مفهوم «میانجی‌گری اخلاقی» از فناوری به‌عنوان «میانجی اخلاقی» یاد کرده است؛ یعنی عاملی که در شکل‌گیری ادراک، نیت و تصمیم اخلاقی دخیل است (فربیک، ۲۰۱۱، ۹۱). به تعبیر او، «فناوری‌ها در خود فرآیند اخلاقی حضور دارند و آن را هم تولید می‌کنند (فربیک، ۲۰۱۱، ۱۱۷). فناوری صرفا ابزاری در دست انسان نیست، بلکه در خود کنش انسانی مداخله می‌کند، آن را شکل می‌دهد، و حتی شریک در فاعلیت می‌شود[۵]

در سوی دیگر، برونو لاتور با نظریه شبکه کنش‌گر اساساً تمایز میان انسان و غیرانسان را در فاعلیت به پرسش کشید. در نگاه او، فاعل» نه موجودی مستقل، بلکه گره‌ای در شبکه‌ای از کنش‌ها و میانجی‌گری‌هاست که شامل انسان‌ها، اشیا، نهادها و فناوری‌ها می‌شود (لاتور، ۲۰۰۵، ۷۲– ۵۴). به عبارت دیگر، کنش اجتماعی نه حاصل اراده یک فاعل انسانی واحد و متکی بر

---

[۱] در چهارچوب فلسفی ارسطویی و کانتی، فاعلیت اخلاقی امری صرفاً انسانی تلقی می‌شود، زیرا تنها انسان است که توانایی درک قانون اخلاقی و انتخاب بر مبنای آن را دارد. بنابران، می‌توان گفت که اجمالا در این سنت‌ها، در عین یکسان نبودن مفهوم فاعلیت نزدشان، فاعل اخلاقی به انسانی اطلاق می‌گردد که از آگاهی، قصد و قدرت تصمیم‌گیری آزاد برخوردار است؛ و مسئولیت اخلاقی‌اش بر اساس همان سه مؤلفه فوق تعیین می‌گردد.

[2] Artificial moral agent

[3] Derived agency

[4] Phenomenology of technology

[۵] به عبارت دیگر، فناوری فقط وسیله نیست، بلکه شریک کنش است. در واقع، فناوری‌ها میان انسان و جهان قرار می‌گیرند و این رابطه را تغییر، بازسازی یا جهت‌دار می‌کنند. اشیاء فناورانه گرچه نیت ندارند، اما جهت‌دهی به تجربه و کنش انسانی دارند.

مسئولیت او، بلکه محصول تعامل شبکه‌ای از عامل‌ها یا فاعلان انسانی و غیرانسانی، مثل اشیا، فناوری‌ها، نرم‌افزارها و نهادها است (لاتور، ۲۰۰۵، ۲۵–۵). از این منظر، فاعلیت، مفهومی توزیع‌شده است، و در نتیجه، مسئولیت اخلاقی نیز باید در سطح شبکه‌ای از عوامل و در آن ساختار بازتعریف شود، نه در سطح فرد انسانی.[۱]

لوچیانو فلوریدی نیز در چهارچوب *فلسفه اطلاعات*[۲] مفهوم «فاعل اطلاعاتی»[۳]را طرح کرده است. در نظام اخلاق اطلاعاتی او، همه موجودات اطلاعاتی ـ اعم از انسان، ماشین و ساختارهای داده‌ای ـ در یک فضای اخلاقی مشترک کنش می‌کنند (فلوریدی، ۲۰۱۳، ۷۵– ۷۱). بنابراین، مسئولیت اخلاقی در این نظام نه به فرد، بلکه به شبکه‌ای از کنش‌گران تعلق دارد که در تولید و انتقال اطلاعات نقش دارند. فلوریدی در این‌باره می‌نویسد:

> در اخلاق اطلاعاتی، هر کنش‌گر اطلاعاتی نسبت به تغییرات در فضای اطلاعاتی پاسخ‌گو است، حتی اگر آن کنش‌گر یک سامانه غیرانسانی باشد (فلوریدی، ۲۰۱۳، ۸۹).

مقصود از بازنگری‌های فوق در تغییر مفهوم مسئولیت، این است تا نشان دهد که مفهوم سنتی مسئولیت، مبتنی بر فردگرایی و آگاهی انسانی، دیگر پاسخ‌گوی واقعیت‌های کنونی نیست. امروزه، کنش اخلاقی در بستری رخ می‌دهد که در آن مرز میان فاعل و ابزار در هم می‌شکند. در نتیجه، تحلیل مسئولیت اخلاقی باید از مدل انسان‌محور به مدل رابطه‌محور و میان‌ذاتی منتقل شود. همان‌گونه که فیلیپ بری[۴]اشاره می‌کند، اخلاق فناوری باید به‌جای تمرکز بر نیت فاعل انسانی، به بررسی ساختارهای سیستمی و تعاملات چندعاملی یا چندفاعلی بپردازد که پیامدهای اخلاقی را رقم می‌زنند و هیچ‌گاه به‌تنهایی عمل نمی‌کنند (بری، ۲۰۱۲، ۳)؛ فلذا، کنش اخلاق در نظر بری، همواره درون ساختارهای فناورانه و نیز، اجتماعی به‌وقوع می‌پیوندد (بری، ۲۰۱۲، ۳۱۷–۳۰۵).[۵]

بر اساس تمامی مباحث فوق، می‌توان چنین استنتاج نمود که مبانی فلسفی مسئولیت اخلاقی در عصر هوش مصنوعی بر سه چرخش بنیادین استوار است: ۱): چرخش از فاعل مفرد انسانی و شخصی به فاعلیت شبکه‌ای[۶] در پرتو نظریه لاتور؛ ۲): چرخش از قصد و نیت فردی به میانجی‌گری فناوری، در اندیشه فربیک و آیدی؛ ۳): چرخش از انسان‌محوری به اطلاعات‌محوری، در فلسفه اطلاعات فلوریدی. این سه چرخش، شالوده گذار از مسئولیت به‌مثابه‌ی «خاصیت روان‌شناختی انسان» به مسئولیت به‌مثابه‌ی «پدیده‌ای هستی‌شناختی و سیستمی» را فراهم می‌سازند بدین‌سان، اخلاق در بستر هوش مصنوعی، دیگر نه به‌عنوان قضاوت درباره نیت و قصد افراد، بلکه به‌مثابه تحلیل روابط درهم‌تنیده میان انسان، ماشین و جامعه معنا می‌یابد.

---

[۱] این دیدگاه راه را برای درک تازه‌ای از مسئولیت می‌گشاید که دیگر بر محور نیت یا قصد فردی نمی‌چرخد، بلکه بر مبنای نسبت‌های علّی، ارتباطی و سیستمی میان کنش‌گران شکل می‌گیرد. لاتور بر این نکته تاکید می‌کند که باید به فناوری‌ها نیز به‌مثابه کنش‌گران نگاه کرد، چرا که آن‌ها در تعیین نتایج عمل، مشارکت علّی دارند و بنابراین، در تحلیل مسئولیت نیز نمی‌توان آن‌ها را حذف کرد (لاتور، ۲۰۰۵، ۶۰–۲۸).

[2] *Philosophy of information*

[3] informational agen

[4] Philip Brey

[۵] به عبارت دیگر، ساختارهای فناورانه، از جمله الگوریتم‌ها، صرفاً ابزارهای خنثی نیستند، بلکه بخشی از ساختار تصمیم‌گیری و تأثیرگذاری اجتماعی محسوب می‌شوند، و بنابراین، باید در سطح نهادهای طراح، نظارتی و اجرایی نیز پاسخ‌گو باشند. این نوع تحلیل، به سطحی بالاتر از مسئولیت فردی، مانند برنامه‌نویس یا اپراتور می‌رود و بر ساختارهای نهادی و طراحی فناورانه تمرکز می‌کند (بری، ۲۰۱۲، ۱۳–۱).

[6] Networked agency

۲. فاعلیت و عاملیت در عصر هوش مصنوعی: از میانجی‌گری تا فاعلیت مشارکتی

یکی از بنیادی‌ترین پرسش‌های فلسفی درباره هوش مصنوعی آن است که آیا سامانه‌های مبتنی بر هوش مصنوعی را می‌توان فاعل اخلاقی دانست یا خیر؟ پاسخ به این پرسش، به‌ویژه پس از دگرگونی‌های معرفتی در فلسفه فناوری، دیگر نمی‌تواند صرفاً سلبی باشد.[۱] این ادعا از آن جهت مطرح می‌شود که فیلسوفانی همانند فربیک نشان می‌دهند که فناوری نه ابزاری منفعل، بلکه عنصری میانجی در شکل‌گیری کنش و داوری اخلاقی است، که رابطه انسان و جهان را بازپیکربندی می‌کند. از نظر او، در هر کنش اخلاقی دست‌کم دو گونه فاعلیت درهم‌تنیده‌اند: فاعلیت انسانی و فاعلیت فناورانه. او این درهم‌تنیدگی را میانجی‌گری اخلاقی می‌نامد؛ فرآیندی که در آن فناوری‌ها ارزش‌ها، ادراکات و تصمیم‌های اخلاقی ما را شکل می‌دهند (فربیک، ۲۰۱۱، ۹۴-۹۰).[۲]

برونو لاتور نیز، در نظریه شبکه کنش‌گر همین موضوع را از چشم‌اندازی هستی‌شناختی بسط می‌دهد. او می‌نویسد: «هر کنش از مسیر روابط متقابل عبور می‌کند و هیچ کنشی از آنِ یک موجود واحد نیست» (لاتور، ۲۰۰۵، ۷۲-۵۴)؛ فاعلیت و کنش، نه ویژگی فردی بلکه نتیجه شبکه‌ای از تعاملات فاعلان و عاملان انسانی و غیر انسانی است، و به همین خاطر، مسئولیت باید بین تمامی عوامل دخیل در وقوع رفتار توزیع شود[۳] اما برای این‌که بتوان فاعلیت مشارکتی را به سطح مسئولیت اخلاقی ارتقا داد، باید شرطی دیگر با عنوان «بین‌فهمی»[۴] نیز برقرار باشد.[۵] فلذا، می‌توان گفت که فناوری صرفاً شیء یا ابزار نیست، بلکه بخشی از فرآیند درک و داوری است (فلوریدی، ۲۰۱۳، ۸۵). فلوریدی در نظریه «اخلاق اطلاعات» مدعی است که زیست‌بوم اطلاعاتی، بستری است که در آن کنش‌گران انسانی و غیرانسانی از طریق تعامل با یک‌دیگر و تبادل اطلاعات، درک مشترکی از ارزش‌ها و پیامدها می‌سازند (فلوریدی، ۲۰۱۳، ۷۵-۷۱). او بر این باور است:

---

[۱] در واقع، برخی سامانه‌های فناورانه، نظیر الگوریتم‌ها و ربات‌ها، به میزانی از خودمختاری رفتاری و کنش‌گری هدف‌محور دست یافته‌اند که در سطح عملیاتی، فعل اخلاقی انجام می‌دهند و می‌توانند تصمیماتی بگیرند که پیامد اخلاقی داشته باشند، ولی لزوما «نیت اخلاقی» ندارند (فلوریدی و سندرز، ۲۰۰۴، ۳۵۵-۳۵۱)؛ بر این مبنا، میان فاعل اخلاقی و مسئول اخلاقی تمایز وجود دارد، یعنی برخی از فاعلان و کنش‌گران مانند فاعل مصنوعی ممکن است فاعل اخلاقی باشند و فعل اخلاقی انجام دهند، و در عین حال، مسئول اخلاقی نباشند و به‌درستی نتوان مسئولیت اخلاقی افعال را بدان‌ها نسبت داد (فلوریدی و سندرز، ۲۰۰۴، ۳۶۴).

[۲] بر این اساس، فناوری‌ها دارای بُعد اخلاقی هستند و انسان دیگر تنها فاعل اخلاقی نیست، بلکه در یک میدان هم‌کنشی با فناوری‌ها عمل می‌کند. در واقع، الگوریتم‌ها نه‌تنها واسطه کنش ما هستند، بلکه خود در تولید ارزش‌های تصمیم‌گیرانه، مانند عدالت، کارایی یا سود دخالت دارند. این وضعیت در نظر فربیک، همان فاعلیت مشارکتی تلقی شده و در آن، اخلاق نه در برابر فناوری، بلکه درون روابط انسان و فناوری رخ می‌دهد که پیش‌تر بدان پرداخته شد، که (فربیک، ۲۰۱۱، ۱۱۷). این نگاه، زمینه مفهومی لازم برای طرح فاعلیت مشارکتی را فراهم می‌کند؛ فاعلیتی که در آن، مرزهای سنتی میان فاعل و ابزار فرو می‌ریزند و کنش اخلاقی محصول تعامل و هم‌تصمیمی میان عاملان و فاعلان گوناگون شده و انسان‌ها و ماشین‌ها در تصمیم‌گیری‌های اخلاقی به‌صورت مشارکتی عمل می‌کنند (فربیک،۲۰۱۱، ۱۱۷). از همین‌روی، گفته می‌شود که هر تحلیل اخلاقی که بخواهد مسئولیت را به‌شیوه‌ای سازگار و هماهنگ با عصر جدید و مبتنی بر تعامل انسان- ماشین توصیف کند، باید فناوری را به‌عنوان کنش‌گر در ساخت موقعیت‌های اخلاقی به رسمیت شناخته و نوعی فاعلیت اخلاقی را برای آن در نظر بگیرد.

[۳] در پرتو این دیدگاه، فاعلیت به‌مثابه فرآیندیشبکه‌ای، مشارکتی و توزیع‌شده میان عوامل انسانی و غیرانسانی فهم می‌شود، و در نتیجه مسئولیت اخلاقی نیز باید درون همین ساختار مشارکتی تحلیل شود (لاتور، ۲۰۰۵ ، ۷۲-۵۴). در نظر لاتور، هر آن‌چه که بتواند تفاوت ایجاد کند و اثر بگذارد، فاعل است؛ فاعل، یعنی هر موجودی که در شبکه موجب تفاوت، تغییر یا کنش می‌شود، خواه انسان یا اشیاء باشد، خواهد فناوری و و الگوریتم باشد، و خواه سند، داده و نهاد یا شرکت سازنده آن باشد (لاتور، ۲۰۰۵ ، ۷۱-۶۳). به عبارت دیگر، فاعلیت در نظر برونو لاتور، نه خاصیت و ویژگی فردی، بلکه متعلّق به شبکه و به تمامی پیوندها و عوامل مرتبط، و به‌اصطلاح، شبکه‌ای و مشارکتی است، و هیچ فاعلی به‌طور مستقل از دیگری عمل نکرده و هرکنشی، نتیجه همکاری و هم‌پیکری مجموعه‌ای از عوال انسانی و غیرانسانی است.

[4] interunderstanding

[۵] این مفهوم، که در پیوند با فلسفه هرمنوتیکی و پدیدارشناسی متأخر قابل توضیح است، بر امکان هم‌فهمی و هم‌تعبیر میان کنش‌گران متفاوت دلالت دارد. در نظام‌های هوش مصنوعی، چنین هم‌فهمی به‌صورت فنی در قالب «ادغام شناختی» یا Cognitive integration و از منظر فلسفی در قالب «ادغام بین‌فهمی» بروز می‌کند. ادغام بین‌فهمی را می‌توان گونه‌ای «هماهنگی تفسیری» دانست که از طریق آن، انسان و سامانه هوشمند در بستر تعاملات متقابل، به درکی مشترک از وضعیت و هدف کنش دست می‌یابند. این مفهوم را می‌توان با نظریه میانجی‌گری فربیک و نیز مفهوم «زیست‌بوم اطلاعاتی» یا infosphere فلوریدی مرتبط دانست.

جهان به‌مثابه زیست‌بوم اطلاعات است، و هر موجود اطلاعاتی در جهان نقشی در حفظ یا تخریب ارزش‌های اطلاعاتی دارد، و مسئولیت باید متناسب با این سهم تعیین شود (فلوریدی، ۲۰۱۳، ۹۷-۹۲).

از این منظر، فاعلیت مشارکتی در هوش مصنوعی مستلزم سه مؤلفه کلیدی است: الف): میانجی‌گری ادراکی[1] که در آن، فناوری مسیر ادراک و قضاوت انسانی را بازتعریف می‌کند (فربیک، ۲۰۱۱، ۱۰۳). ب): تعامل تفسیری[2] که بر اساس آن، میان انسان و ماشین نوعی گفت‌وگوی تفسیرمحور شکل می‌گیرد (آیدی، ۱۹۹۰، ۴۴-۴۲). ج): پاسخ‌گویی الگوریتمی، که بر مبنای آن، مسئولیت اخلاقی از سطح فردی به سطح شبکه‌ای و سازمانی انتقال می‌یابد (بری، ۲۰۱۲، ۵). در چهارچوب چنین فاعلیتی، مسئولیت اخلاقی نه در مقام نسبت‌دادن تقصیر یا نیت، بلکه به‌مثابه پاسخ‌گویی در برابر رابطه‌ها و فرآیندهای درهم‌تنیده تعریف می‌شود. به تعبیر فلوریدی «اخلاق در عصر اطلاعات، اخلاق روابط است نه اراده‌ها» (فلوریدی، ۲۰۱۳، ۹۲). در نتیجه، مسئولیت نیز به‌جای «داشتنِ نیت» به «پاسخ‌گو بودن نسبت به پیامدهای میان‌ذاتی» بدل می‌شود. این تحول مفهومی، ما را به مفهومی نزدیک می‌کند که می‌توان آن را «اخلاق مشارکتی»[3]نامید: اخلاقی که در آن کنش‌گران انسانی و غیرانسانی، به‌طور هم‌زمان در شکل‌گیری تصمیم اخلاقی دخالت دارند و هر یک سهمی از مسئولیت را بر دوش می‌کشند. چنین اخلاقی به تعبیر فربیک، محصول «بازتوزیع فاعلیت در سطح روابط است» (فربیک، ۲۰۰۵، ۱۱۲).[4]

۳. مسئولیت اخلاقی و پاسخ‌گویی الگوریتمی: از فرد به سیستم

۳-۱. تمایز مفهومی: مسئولیت اخلاقی در برابر پاسخ‌گویی الگوریتمی

در سنت اخلاقی کلاسیک، با همه اختلافی که در معنا و تفسیر واژه مسئولیت نزد فلاسفه وجود دارد، اجمالا می‌توان آن را مفهومی در نظر گرفت که پیشینه‌اش به دوران یونان باستان باز گشته و مستلزم نسبت علیّتی روشن بین نیت فاعل و پیامد عمل، و نیز صفات درونی چون آگاهی و اختیار است؛ مسئولیت اخلاقی با آگاهی، دانایی و اختیار همراه است (کانت، ۱۷۸۵، ۴۰۵- ۳۹۷؛ ارسطو، ۱۹۸۵، ۱۱۰۹ب۱۱۱۰۳۰آ۱۵)[5] در برابر این سنتِ شخص‌محور و مبتنی بر فاعل انسانی صرف، مفهوم پاسخ‌گویی الگوریتمی پدیدار می‌شود که در دهه گذشته به کلیدواژه اخلاق هوش مصنوعی بدل گردیده، و معیارش شفافیت، توضیح‌پذیری، توصیف‌پذیری فرآیندهای تصمیم‌گیری سامانه‌ها، ردیابی[6]مسیرهای علت‌مند تصمیم، تخصیص مسئولیت در

---

[1] Perceptional mediation
[2] Hermeneutic interaction
[3] participatory ethics

[4] بنابراین، در عصر هوش مصنوعی، فاعلیت اخلاقی نه ویژگی ذاتی سوژه، بلکه رابطه‌ای پویا و در حال دگرگونی میان انسان، ماشین و ساختارهای اجتماعی است. این درک تازه از فاعلیت، امکانِ گذار از الگوی سنتی مسئولیت به الگویی میان‌ذاتی و شبکه‌ای را فراهم می‌آورد؛ الگویی که در آن، اخلاق به فرآیند بین‌فهمی، هم‌فهمی و هم‌پاسخ‌گویی بدل می‌شود.

[5] ارسطو، ارزش اخلاقی را در اعمالی می‌بیند که برخاسته از خرد عملی و فضیلت اخلاقی بوده و تنها کسانی که آگاهانه، آزادانه و از روی انتخاب خردمندانه عمل می‌کنند، مسئول اعمال خود هستند (ارسطو، ۱۹۸۵ ، ۱۱۱۳ب -۱۱۱۰آ) . در نظر کانت، انسان موجودی عقلانی، خودمختار و آزاد است که خودش می‌تواند قوانین اخلاقی را وضع کند و آن‌ها را با تکیه بر عقل خودش استنباط و انتخاب کند. از نظر او، ارزش اخلاقی یک عمل نه به نتایج آن بلکه به نیت و انگیزه درونی فرد بستگی داشته و عملی که از روی احترام به قانون اخلاقی انجام شده باشد، واجد ارزش اخلاقی است (کانت، ۱۷۸۵ ، ۴۰۵- ۳۹۷).

[6] ردیابی یا Tracing به این معناست که باید امکان ردیابی تصمیمات سیستم نسبت به انتخاب‌ها، مقاصد و مسئولیت‌های انسانی وجود داشته باشد. انسانی که سیستم را طراحی و برنامه‌ریزی یا حتی مدیریت کرده باشد، باید بتواند مسئول و پاسخ‌گو باشد؛ و اگر تصمیمی اشتباه گرفته شد، باید به‌طور شفاف بتوان مشخص کرد چه کسی، کِی و چرا آن تصمیم را اتخاذ کرده است. به‌طور مثال، اگر یک ماشین خودران تصادف کند، باید بتوان فهمید آیا اشتباه از الگوریتم بوده، یا از داده‌ی آموزشی، یا از سازنده و طراح آن که آن را نادیده گرفته است. در واقع، اگر شرط ردیابی کنار گذاشته شود، هیچ‌کسی نمیداند که چه کسی مسئول تصمیم اشتباه است (سانتونی و هوفن، ۲۰۱۸، ۹۷).

سطح سامانه و نهاد و در نهایت، قابلیت بازرسی تصمیم‌های الگوریتمی است (بینز، ۲۰۱۸، ۵۴۳). پاسخ‌گویی الگوریتمی، که مفهومی نسبتاً جدید است و تلاش می‌کند سازوکاری برای پاسخ‌گو کردن سامانه‌های تصمیم‌گیر خودکار ایجاد کند، بر این نکته تاکید می‌کند که هر تصمیمی، حتی اگر توسط ماشین یا ربات گرفته شده باشد، باید قابل پیگیری، توضیح‌پذیر و در نهایت قابل بازخواست باشد.[۱]

۳-۲. شکاف مسئولیتی: صورت‌بندی و انواع

شکاف مسئولیتی زمانی رخ می‌دهد که نتوان یک پیامد اخلاقی را به‌طور معناداری به یک فاعل مشخص نسبت داد؛ نه به‌خاطر فقدان تقصیرِ آشکار، بلکه به‌دلیل توزیع شدن عاملان یا فاعلان و عدم شفافیت در زنجیره علّی (فلوریدی و سندرز، ۲۰۰۴، ۳۶۴؛ سانتونی و مکاچی، ۲۰۲۱، ۱۲۱). نمونه‌های مشهور مانند سامانه COMPAS، که در ادامه بیش‌تر و مفصل‌تر با آن آشنا خواهیم شد، نشان می‌دهند که الگوریتم‌ها می‌توانند پیامدهای تبعیض‌آمیز تولید کنند، در حالی‌که مرجع مشخص و روشنی برای دلیل وقوع این اتفاق ندارند، چه کسی باید پاسخ‌گو باشد؟ طراح؟ داده‌پرداز؟ سازنده مدل؟ قاضی که از خروجی استفاده کرده؟[۳]

۳-۳. از چه ابزارهای نظری و عملی برای بستن شکاف می‌توان استفاده کرد؟

برای تبدیل پاسخ‌گویی الگوریتمی به نوعی مسئولیت اخلاقی معنادار، باید هم معیارهای نظری روشن شوند و هم سازوکارهای عملی برقرار گردند. به گفته برخی اندیشمندان هم‌چون سانتونی و هوفن حفظ کنترل معنادار انسانی[۴] شرط لازم مسئولیت است (سانتونی و هوفن، ۲۰۱۸، ۷). بنابراین، در سامانه‌های حساس باید مکانیسم‌های دخالت انسانی، مستندسازی تصمیم و امکان لغو خروجی الگوریتم در نظر گرفته شود. تنها از طریق این ترکیب و تعامل انسان ـ ماشین است که می‌توان به شکل متوازن میان کارایی فناوری و ارزش‌های اخلاقی حرکت کرد.[۵] در ادامه، بیش‌تر به توضیح این مطالب خواهیم پرداخت.

---

[۱] یکی از مهم‌ترین ابزارهایی که در چهارچوب پاسخ‌گویی الگوریتمی مطرح می‌شود، الزام به افشای اطلاعات است؛ یعنی: کاربران، ذی‌نفعان و نهادهای نظارتی باید بدانند که یک الگوریتم چگونه طراحی شده است، و چه داده‌هایی در آن وارد شده‌اند، و چه نوع تصمیماتی گرفته و چگونه نتایج نهایی ارائه می‌شوند. این سطح از افشا، لازمه‌ی ممیزی مستقل است و بدون آن، هیچ‌کس نمی‌تواند در برابر اشتباهات الگوریتمی پاسخ‌گو باشد. به عبارت ساده‌تر، مردم حق دارند بدانند الگوریتمی که درباره آن‌ها تصمیم می‌گیرد (مثلاً برای اعطای وام یا استخدام) بر چه اساسی این تصمیم را می‌گیرد. اگر این تصمیم ناعادلانه باشد، باید بتوانند شکایت کنند و یک نفر یا نهاد مشخص، توضیح بدهد که چرا این اتفاق افتاده است.

[۲] سامانه COMPAS یا Correctional Offender Management Profiling for Alternative Sanctions یکی از شناخته‌شده‌ترین گزارش‌های آنلاین و تحقیقی و نیز، نمونه‌های به‌کارگیری الگوریتم‌ها در نظام قضایی کیفری ایالات متحده آمریکا است، که سازمان خبری مستقل و غیرانتفاعی Propublica آن را منتشر کرده است. این سامانه به‌طور خاص برای پیش‌بینی خطر بازگشت به جرم و احتمال ارتکاب مجدد جرم توسط متهمان یا محکومان طراحی شده، و توسط یک شرکت خصوصی به‌نام Northpointe توسعه یافته و از طریق پرسش‌نامه‌ها و داده‌های کیفری، اجتماعی و شخصی افراد، نمره‌ای در سه سطح پایین، متوسط و بالا برای «ریسک بازگشت به جرم» یا « خطر ارتکاب جرم خشونت‌آمیز» در دادگاه‌ها (برای تصمیم‌گیری درباره آزادی با وثیقه)، در احکام قضایی (برای تعیین شدت مجازات یا طول دوران زندان)، و در نظام آزادی مشروط برای ارزیابی خطرات احتمالی تخصیص می‌دهد. برای مطالعه‌ی بیش‌تر رجوع کنید به Angwin, Julia, Larson, Jeff, Mattu, Surya and Kirchner, Lauren (2016) ‘Machine Bias: there’s software used acros the country to predict future criminals. And it’s biased against blacks’, Propublica

[۳] در تحلیل فلسفی، می‌توان انواع شکاف مسئولیتی را تفکیک کرد: الف): شکاف علّی یا casual gap یعنی هنگامی‌که نسبت علّی میان کنش‌گر انسانی و پیامد نامشخص یا غیرقابل اثبات است، به‌طور مثال، تصمیمات ناشی از فرآیندهای پیچیده یادگیری ماشین که قابل‌تفکیک به اقدامات و رفتارهای انسانی نیستند. ب): شکاف نظارت یا institutional gap، یعنی هنگامی‌که نهادهای نظارتی یا اپراتورها ابزارها را به‌کار می‌گیرند، اما توانایی یا اختیار اصلاح آن‌ها را ندارند. ج): شکاف نهادی، یعنی هنگامی‌که قواعد حقوقی و نهادی برای برخورد با پیامدهای سامانه‌ها ناکافی یا قدیمی‌اند.

[4] Meaningful Human Control

[۵] درک مسئولیت اخلاقی در هوش مصنوعی بدون تبیین نسبت میان انسان، طراحی و الگوریتم ممکن نیست. از دیدگاه فلسفه فناوری، هر کنش فناورانه در بستر یک «سه‌گانه میانجی» یا Triple mediation شکل می‌گیرد: فاعل انسانی، سازوکار طراحی، و سیستم خودکار پاسخ‌گویی. مسئولیت اخلاقی زمانی پایدار می‌ماند که در هر سه سطح این میانجی‌گری، کنترل و شفافیت برقرار باشد. در واقع، می‌توان از یک مثلث مسئولیت سخن گفت که اضلاع آن عبارت‌اند از: کنترل معنادار انسانی، طراحی اخلاقی، و پاسخ‌گویی الگوریتمی، که در آن، کنترل معنادار انسانی به‌مثابه پیش‌شرط مسئولیت، طراحی اخلاقی به‌مثابه بستر و سازوکار پیش‌گیرانه و پاسخ‌گویی الگوریتمی به‌مثابه ابزار بازبینی و شفافیت پسینی در نظر گرفته خواهند شد.

۳-۴. کنترل معنادار انسانی: پیش‌شرط فاعلیت اخلاقی

یکی از پیشنهادهای اصلی برای تلاش عملی در جهت تخصیص مسئولیت، بازگرداندن سطحی از کنترل انسانی است که «معنادار»[1] باشد؛ یعنی حضور انسان صرفاً صوری نباشد، بلکه شامل اطلاع، توانِ مداخله، و اختیار تغییر در خط تصمیم‌گیری باشد. شرط کنترل معنادار انسانی به‌طور خاص در مواقع حساس تصمیم‌گیری که جان، کرامت یا حقوق افراد در میان است حیاتی است.[2] در واقع، کنترل معنادار یعنی فهم، تشخیص و توان اعمال اراده مستقل انسان در فرآیند تصمیم‌سازی حفظ شود. به عبارت دیگر، سامانه یا سیستم باید ابزار تصمیم‌گیری برای افراد باشد، نه قاضی یا پزشک؛ قاضی برای صدور حکم در دادگاه‌ها به‌هنگام وقوع جرم باید بتواند اطلاعات و تحلیل‌های نهایی سامانه‌ها را رد کند یا با شواهد انسانی و بافت منطقه‌ای ترکیب کند. در غیر این صورت، سوگیری نژادی و عدم طراحی مسئولانه می‌تواند عدالت را به خطر بیاندازد. یعنی انسان باید در موقعیتی باشد که به کمک سیستم فکر کند، نه این‌که توسط سیستم و سامانه جایگزین فکر و تصمیم خود انسان شود (بری، ۲۰۱۲، ۳۱۷-۳۰۵).

۳-۵. ردگیری و ردیابی

شرط «ردگیری»[3] و ردیابی، در بستر کنترل معنادار انسانی و دخالت متداوم او تحقق می‌پذیرد. در واقع، برای این‌که کنترل انسانی واقعا معنادار باشد، صرفا حضور خودش کافی نیست؛ بلکه باید شرایطی برقرار باشد که دو مفهوم فوق را در بر بگیرند (سانتونی و هوفن، ۲۰۱۸، ۹-۷). به عبارت دیگر، برای پاسخ‌گو نگه‌داشتن افراد و نهادها در برابر رفتارهای سیستم‌های خودکار، باید شکلی از کنترل معنادار انسانی بر این سیستم‌ها حفظ شود، اما این «کنترل» صرفاً به معنای حضور انسان در حلقه تصمیم‌گیری نیست، و این حضور باید همراه با شرایط و لوازم قابلیت ردگیری و ردیابی باشد. هدف اصلی ردگیری، حصول اطمینان از همراهی و موافقت سیستم با ارزش‌های اخلاقی بوده و تضمین می‌کند که سیستم، پاسخی اخلاقی داشته باشد، نه فقط محاسبات عددی یا منطقی؛ و هدف اصلی ردیابی، تعیین مسئولیت انسانی و امکان پاسخ‌گویی اوست، و تضمین می‌کند که انسان با شناخت و مسئولیت در زنجیره اخلاقی حضور دارد. در واقع، لزوم استفاده از همین دو شرط است که این اطمینان را به‌دست می‌دهد که تمامی سیستم‌های هوشمند و فناورانه از کنترل انسان خارج نمی‌شوند و هم‌چنان در چهارچوب ارزش‌های اخلاقی باقی می‌مانند و نیز، انسان همواره پاسخ‌گوی تصمیمات گرفته‌شده خواهد بود.

---

[1] meaningful

[2] مفهوم کنترل معنادار انسانی که نخستین بار توسط سانتونی و هوفن مطرح گردید، در مقام واکنشی به خطر حذف انسان از چرخه‌ی تصمیم‌گیری و تحت عنوان شرط لازم مسئولیت اخلاقی بیان گردید؛ این اصل به معنای ایجاد سازوکارهایی است که انسان بتواند در هر لحظه از چرخه تصمیم مداخله کند، تصمیم را اصلاح کند و اثرش را بسنجد (سانتونی و هوفن،۲۰۱۸، ۷). به‌جهت بازگرداندن فاعلیت انسانی، تبدیل شدن او به ناظر فعال و نه منفعل و نهایتا، مسئول دانستن او، سه اصل باید رعایت شود: الف): آگاهی از تصمیم: در طراحی سامانه، باید مشخص شود که کجا و چگونه انسان در تصمیم‌گیری حضور دارد؛ قبل، حین یا پس از تصمیم الگوریتمی. به عبارت دیگر، کاربران باید از منطق تصمیم‌گیری الگوریتم مطلع باشند و بدانند سامانه‌ها چرا و چگونه تصمیم می‌گیرند. ب): توان مداخله: در هر مرحله از فرآیند تصمیم‌گیری سامانه‌ها، همواره باید امکان دخالت انسانی وجود داشته باشد، و کاربران باید بتوانند تصمیم الگوریتمی سامانه‌ها را متوقف یا تغییر دهند. ج): قابلیت بازبینی: تصمیم‌ها باید توسط کاربر انسانی قابل ردگیری، مستند و بازتولید باشند، و هر بار مداخله یا تأیید انسانی باید ثبت شود تا بعدا بتوان مسیر تصمیم را بازسازی کرد. بدون تحقق این سه شرط، حتی اگر فرد از نظر حقوقی مسئول شناخته شود، از نظر اخلاقی فاقد اختیار واقعی است. به‌عنوان مثال، پزشک یا قاضی نباید صرفاً مجری دستورات سیستم باشد؛ سیستم‌ها باید به‌گونه‌ای طراحی شوند که فرد بتواند در صورت لزوم، تصمیم را رد یا تغییر دهد.

[3] ردگیری یا Tracking به این معناست که سیستم باید در جهت مقاصد انسانی، اخلاقی عمل کند. یعنی خروجی سیستم باید با ارزش‌ها، اهداف و نیت‌های اخلقی انسان‌ها هم‌راستا باشد. به‌طور مثال، اگر یک سیستم تشخیص چهره برای امنیت عمومی طراحی شده باشد، فقط و فقط باید در راستای امنیت شهروندان (به‌عنوان اهداف مربوطه) و نه برای سرکوب آن‌ها یا حتی تبعیض نژادی عمل کند. به عبارت دیگر، اگر سیستم، اهداف غیراخلاقی را دنبال کند و یا حتی از مسیر کنترل انسانی منحرف شود، می‌گوییم شرط ردگیری نقض شده است. نکته‌ی مهم و قابل توجه این شرط این است که شرط ردگیری درباره هم‌راستایی اهداف با یکدیگر بوده و سیستم باید در زمان واقعی و پیوسته با ارزش‌های انسانی تنظیم شود، نه فقط در لحظه طراحی یا آموزش‌های اولیه. افزون بر این، اگر شرط ردگیری کنار گذاشته شود، سیستم ممکن است کارهایی کند یا تصمیماتی اتخاذ کند که انسان نمی‌خواهد (سانتونی و هوفن، ۲۰۱۸، ۷).

۳-۶. طراحی اخلاقی: نهادینه‌سازی ارزش‌ها در ساختار فناوری

اگر کنترل معنادار انسانی ضامن اختیار است، «طراحی اخلاقی»[1]ضامن پیش‌گیری از خطاست. در این الگو، ارزش‌های اخلاقی نه در مرحله پسینی تنظیم مقررات، بلکه در همان لحظه‌ی طراحی و توسعه وارد فرآیند طراحی و تولید الگوریتم‌ها می‌شوند. به تعبیر فربیک، فناوری‌ها حامل اخلاق هستند، فلذا، باید از ابتدا با آگاهی اخلاقی ساخته شوند (فربیک، ۲۰۱۱، ۹۷-۹۱). طراحی اخلاقی بر سه محور استوار است: الف): ترجمه ارزش‌ها به الزامات فنی: به‌طور مثال، عدالت به معنای تعادل میان دقت و شمول در داده‌ها است. ب): بازتاب اخلاق در واسط کاربر: به‌طور مثال، شفافیت از طریق نمایش نحوه تصمیم الگوریتم نشان داده می‌شود. ج): بازطراحی براساس بازخورد اخلاقی: به این معنا که سامانه‌ها باید امکان اصلاح ارزش‌ها را در طول عمر خود داشته باشند. این دیدگاه، طراحی را از فعالیت صرفاً فنی به فرآیندی اخلاقی ـ فرهنگی تبدیل می‌کند، که در آن، مهندس، فیلسوف و سیاست‌گذار همگی در نقش طراحان اخلاق سهیم خواهند بود[2](آیدی، ۱۹۹۰، ۱۱۲).

۳-۷. پاسخ‌گویی الگوریتمی: ابزار شفافیت و بازسازی مسئولیت

حتی در بهترین طراحی‌ها نیز، خطا و سوگیری محتمل است. از این‌رو، پس از مرحله طراحی، باید سازوکاری برای پاسخ‌گویی و نظارت وجود داشته باشد که شکاف میان تصمیم و نتیجه را پر کند. مفهوم پاسخ‌گویی الگوریتمی که توسط بینز و دیگر اندشمندان توسعه یافته، به این معناست که باید امکان ردیابی، توضیح و نسبت‌دهی تصمیمات هوش مصنوعی به‌نحوی فراهم شود تا در صورت بروز خطا یا آسیب، مسئولیت مشخص باشد، و سامانه‌ها باید همانند کنش‌گران اجتماعی، دلایل تصمیم‌های خود را ارائه دهند؛ که در سطح اجرایی، این امر مستلزم چند ابزار است: الف): ممیزی الگوریتمی[3] یعنی بررسی بیرونی[4]و دوره‌ای عملکرد مدل‌ها توسط نهادهای مستقل. ب): ردپای تصمیم[5] یعنی ایجاد قابلیت ردیابی برای تمام گام‌های تصمیم‌گیری الگوریتمی. ج): توضیح‌پذیری چندسطحی[6] یعنی ارائه توضیحات قابل فهم برای کاربران، متخصصان و نهادهای نظارتی به تفکیک سطح تخصص. براساس ویژگی پاسخ‌گویی الگوریتمی سیستم‌ها باید قابلیت ثبت و مستندسازی روند تصمیم‌گیری خودش را داشته باشند؛ و مشخص شود به چه دلیلی مثلا گزینه‌ی الف را پیشنهاد داده، و یا بر چه داده‌هایی تکیه کرده است. این قابلیت به ما امکان می‌دهد که در صورت بروز مشکل یا حادثه و یا هر نوع چالش و محدودیتی، بتوانیم آن را پیگیری و بررسی کنیم.

۳-۸. توزیع مسئولیت و نقش نهادها

روشن است که در سامانه‌های مبتنی بر هوش مصنوعی و تعامل انسان- ماشین، مسئولیت اخلاقی دیگر قابل انتساب صرف به یک کنش‌گر و فاعل انسانی نیست، بلکه در مجموعه‌ای از فاعلان، زیرساخت‌ها، طراحی‌ها و ساختارهای سازمانی توزیع می‌شود. تحلیل فلسفی نشان می‌دهد که توزیع مسئولیت بین افراد و نهادهای مختلف، مانند طراح، اپراتور، ناظر و سیاست‌گذار باید با

---

[1] Ethical by design

[2] بر اساس ویژگی طراحی اخلاقی، الگوریتم سامانه‌ها نباید براساس تعصبات داده‌ای (مثلاً نژاد، جنسیت، سن) تصمیم‌گیری کنند، و نیز، طراحی باید شامل فیلترهای ارزشی باشد که مانع تصمیم‌های غیراخلاقی یا ناعادلانه بشود. به‌طور مثال، در برخی سیستم‌های پزشکی دیده شده که توصیه‌ها برای افراد سیاه‌پوست به دلایل آماری متفاوت بوده، اما این تمایز غیراخلاقی و مبتنی بر تبعیض داده‌ای بوده است و در چنین مواردی، طراحی اخلاقی کمک می‌کند تا تصمیمات سامانه نه فقط از نظر کارکردی بلکه از نظر ارزش‌های انسانی نیز قابل‌قبول باشد.

[3] Algorithm audit

[4] گزارش‌ها و مداخلات پژوهشی اخیر نشان می‌دهد که فقط انجام ممیزی‌های بیرونی کافی نیست؛ باید سازوکارهایی درون‌سازمانی و نهادی ایجاد شود که شامل شفافیت طراحی (design transparency) ، استانداردهای مستندسازی (documentation standards)، بازبینی مستقل و مستمر (independent continuous review)،جبران آسیب (remediation & redress) می‌باشند. به عبارت دیگر، پاسخ‌گویی الگوریتمی باید از «بازبینی دوره‌ای» فراتر رود و به «ساختاری و فرآیندی» تبدیل شود

[5] Decision Traceability

[6] Multi-layered Explainability

قواعد نهادی همراه باشد که هم سهم هر نقش را مشخص کند و هم مسئولیت جمعی و نهادینه‌شده را تعریف نموده و آن‌ها را پاسخ‌گو سازد. این جهت‌گیری نهادی، نه‌تنها الزام‌آور است، بلکه از منظر عدالت توزیعی[1] و جلوگیری از تحمیل بی‌عدالتی بر اقشار آسیب‌پذیر نیز ضروری است. به عبارت دیگر، ممیزی‌های فنی کافی نیستند؛ و پاسخ‌گویی واقعی مستلزم نهادهایی است که بتوانند رفتار سیستم‌ها را در زمینه‌های اجتماعی و نهادی ارزیابی کنند (سلبست، فریدلر، ونکاتاسوبرامانیان و ورتسی، ۲۰۱۹، ۶۵-۶۲) به بیان دیگر، پاسخ‌گویی باید از سطح کُد به سطح حاکمیت گسترش یابد.

۴. شکاف مسئولیت اخلاقی در بسترهای واقعی: از تحلیل نظری تا مصادیق الگوریتمی

۴-۱. مسئله واقعی: بحران مسئولیت در نظام‌های تصمیم‌یار الگوریتمی

برای سنجش کارایی چهارچوب‌های نظری درباره‌ی مسئولیت اخلاقی و پاسخ‌گویی الگوریتمی، بررسی پرونده‌های واقعی ضروری است. در این بخش، از میان مثال‌های بارز شکاف مسئولیتی در حوزه عملی چون عدالت کیفری، سلامت دیجیتال، امنیت ملّی، مدیریت شهری، بهداشت و... به دو نمونه شاخص حوزه عدالت کیفری و قضایی COMPAS و پزشکی IBM Watson Health اشاره می‌کنیم. بررسی این دو نمونه نشان می‌دهد که چگونه شکاف مسئولیتی از سطح نظری و فردی به سطح کاربردی، اجتماعی و نهادی منتقل می‌شود.

۴-۲. سامانه COMPAS و مسئله‌ی عدالت توزیعی در تصمیم‌گیری قضایی

عدالت کیفری یا تصمیم‌گیری درباره سرنوشت افراد، به‌مثابه مصداق بارز حیطه عملی شکاف مسئولیت اخلاقی، یکی از مهم‌ترین کاربردهای هوش مصنوعی است که در آن، میزان ارتکاب جرم در آینده توسط افراد یا مناطق خاص پیش‌بینی می‌شود[2]؛ نمونه شناخته‌شده آن، سامانه COMPAS که توسط شرکت Northpointe تأسیس و با هدف پیش‌بینی احتمال تکرار جرم در دادگاه‌های ایالات متحده به‌کار گرفته شد. در این سامانه، الگوریتمی به کار رفته که براساس داده‌ها و اطلاعات گذشته تاریخی و بدون شفافیت کافی، احتمال ارتکاب جرم در آینده را برای هر فرد محاسبه می‌کند. الگوریتم‌های این سامانه در ابتدا، با کاهش سوگیری انسانی در قضاوت و افزایش کارایی همراه بود، اما گزارش تحقیقاتی پروپابلیکا[3] در سال ۲۰۱۶ نشان داد سامانه فوق، نه‌تنها یک ابزار فناورانه، بلکه یک ابزار فلسفی در تفسیر عدالت است که با اصل فردیت عدالت، یعنی بررسی خاص هر پرونده در تعارض است، و احتمال‌گرایی را جایگزین شواهد خاص می‌کند. هم‌چنین، این سامانه در عمل، دچار سوگیری نژادی سیستماتیک گردیده، به‌گونه تبعیض‌آمیز در شرایط مشابه، سیاه‌پوستان را در اولویت بالاتری برای ارتکاب جرم به نسبت سفیدپوستان قرار داده است[4] (انگوین، ماتو، کِریشنر، ۲۰۱۶، ۴). این وضعیت، مصداق واضح و روشن یک شکاف مسئولیتی در

---

[1] Distribute justice

[2] شکاف مسئولیت اخلاقی علاوه بر حیطه عملی و کاربردی، در موضوعات نظری نیز کارایی دارد؛ به عبارت دیگر، گستردگی و امکان وقوع چنین شکافی هم در حوزه نظری و فلسفی و هم در حوزه عملیاتی بوده و حیطه نظری و عملی، هر دو می‌توانند بستر شکاف مسئولیتی قرار بگیرند.

[3] ProPublica

[4] به عبارت دیگر، سامانه، ضمن احتمال «پرخطر» دانستن افراد سیاه‌پوست به‌نسبت سفیدپوستان، احتمال ارتکاب جرم یا رفتارهای مجرمانه توسط آن‌ها را به میزان حداقل دو برابر بیش‌تر از افراد سفیدپوست پیش‌فرض گرفته است. به‌نظر می‌رسد که مشکل اصلی نه در فقدان داده‌ها، بلکه در نبود شفافیت و نامشخص بودن مسئول و مرجع مسئولیت است. تحقیقات انجام شده بر این سامانه و سامانه‌های دیگر نشان می‌دهد که سامانه فوق، نه‌تنها بحث تبعیض و تعصب نژادی و حتی پیش‌داوری نژادی را پیش می‌کشد، بلکه نوعی عدم شفافیت و ابهام در پاسخ‌گویی و شکاف مسئولیتی را به نمایش می‌گذارد، به‌گونه‌ای که اصلا مشخص نیست چه کسی یا چه نهاد و سیستمی مقصر وضعیت پیش‌آمده است. افزون بر این، گزارش‌ها حاکی از عدم پذیرش مسئولیت اخلاقی پیامدها و نتایج به‌بارآمده الگوریتم‌های سامانه فوق توسط عوامل انسانی، اعم از برنامه‌نویسان، کارکنان دادگاه، یا حتی مقامات قضایی است، به‌نحوی که هیچ‌کدام راضی نشدند مسئولیت اخلاقی پیامدها را بپذیرند. در

زمینه علمیاتی و نهادی است، اما اهمیت فلسفی آن صرفا در دشواری انتساب مسئولیت خلاصه نمی‌شود (بینز، ۲۰۱۸، ۵۴۶)؛ بلکه نشان می‌دهد که الگوهای کلاسیک مسئولیت اخلاقی، که بر فاعل منفرد، نیت‌مند و قابل‌شناسایی مبتنی هستند، در مواجهه با سامانه‌های الگوریتمی ناکارآمد می‌شوند. در چنین نظام‌هایی، کنش اخلاقی نه محصول تصمیم یک فاعل واحد، بلکه حاصل برهم‌کنش ساختارمند میان داده‌ها، الگوریتم‌ها، سیاست‌های نهادی و کنش‌گران انسانی است. از این‌رو، مسئله صرفاً کمبود شفافیت یا ضعف پاسخ‌گویی نهادی نیست، بلکه ما با دگرگونی ماهیت فاعلیت اخلاقی مواجه هستیم؛ دگرگونی که ایجاب می‌کند مسئولیت اخلاقی نه به‌صورت فردمحور، بلکه در چهارچوبی توزیع‌شده و مشارکتی، مطابق با رویکردهای فلسفه فناوری صورت‌بندی شود؛ به عبارت دقیق‌تر، تمامی داده‌ها و کدها، سیاست‌های قضایی، و کاربران انسانی (اعم از کاربر و طراح)، همگی در فرآیند تولید تصمیم دخیل هستند، و هیچ‌کدامشان به‌تنهایی نمی‌توانند به‌مثابه فاعل اخلاقی در نظر گرفته شوند، چرا که همان‌طور که در ادامه اشاره خواهد شد، بدون شفافیت طراحی، امکان ردیابی مسئولیت از میان رفته و مسئول و مرجع نهایی مشخص نخواهد شد[۱](فربیک، ۲۰۱۱، ۱۱۹). در چنین شرایطی، بازسازی اخلاقی مسئولیت تنها از رهگذر الگوی پاسخ‌گویی الگوریتمی ممکن است، الگویی که در آن همه‌ی سطوح طراحی، داده و استفاده مشمول پاسخ‌گویی مشترک می‌شوند، نه این‌که همه عوامل دخیل باشند، اما کسی مسئول و پاسخ‌گو نباشد[۲](فلوریدی، ۲۰۱۳، ۱۰۱-۹۷).

۵. سامانه IBM Watson Health و مسئولیت اخلاقی در هوش مصنوعی پزشکی

در نظام سلامت دیجیتال نیز وضعیت از همین قرار است؛ بدین معنا که در حیطه سلامت جامعه نیز، هوش مصنوعی برای تشخیص بیماری، تجویز دارو، تصویربرداری پزشکی، تعیین اولویت‌های درمانی بیماران و پیش‌بینی مرگ‌ومیر به‌کار گرفته می‌شود. سامانه[۳] IBM Watson Health محصول شرکت IBM، با هدف پشتیبانی و کمک به تصمیم‌های درمانی پزشکان به‌ویژه در بیماران مبتلا به سرطان طراحی شد. این سیستم با استفاده از داده‌های پزشکی و مقالات علمی، پیشنهادهای درمانی ارائه می‌کرد، اما در سال ۲۰۱۸، گزارش داخلی منتشر شده از شرکت که توسط *STAT* News انجام شده بود، فاش کرد که سامانه فوق در برخی موارد پیشنهادهای «غیرایمن و بالقوه مضر» برای بیماران ارائه داده است (راس و سوئتلیتز، ۲۰۱۸، ۲). در واقع، استفاده از هوش مصنوعی در چنین وضعیت‌هایی سبب شده است که تصمیمات حیاتی به الگوریتم‌هایی سپرده شوند که باوجودی که کارکردشان برای تحقق اهدافی مانند تشخیص بیماری‌ها و... تعیین و برنامه‌ریزی شده است، اما در موقعیت‌هایی که سامانه‌ها، بیماری فرد را به‌نادرستی تشخیص دهد، یا بدون درنظر گرفتن سوابق پزشکی یا حساسیت‌های دارویی افراد، دارویی

چنین موقعیتی، سیستم و سامانه به‌عنوان یک «جعبه‌سیاه» یا black box عمل نموده و تصمیم نهایی دادگاه‌ها بر مبنای امتیازهای تولیدشده توسط الگوریتم اتخاذ گردیده است.

[۱] چنین پدیده‌ای در نظر برخی اندیشمندان مانند فلوریدی، «میانجی‌گری اخلاقی غیرقابل تقلیل» پنداشته شده، به‌گونه‌ای‌که در آن، مسئولیت میان عناصر شبکه توزیع می‌گردد.

[۲] و نتوان گفت که چه کسی باید پاسخ‌گوی این تصمیم ناعادلانه و آسیب‌های چنین تصمیمی باشد؟ قاضی‌ای که از آن الگوریتم استفاده کرده؟ برنامه‌نویس؟ داده‌های تاریخی که سوگیری در آن‌ها وجود داشته؟ شرکت سازنده Northpointe؟ هیچ‌کدام؟ یا همگی؟

[۳] سامانه IBM Watson Health یکی از پروژه‌های شاخص شرکت IBM در حوزه کاربرد هوش مصنوعی در پزشکی و بهداشت است که هدفش، ارتقاء فرآیندهای درمان، تشخیص بیماری و تصمیم‌گیری بالینی از طریق تحلیل داده‌های عظیم پزشکی با بهره‌گیری از پردازش زبان طبیعی و یادگیری ماشین است. پزشکان با کمک این سامانه می‌توانند بیماری‌ها را دقیق‌تر تشخیص دهند، و با سرعت و دقت بیش‌تری روش‌های درمانی مبتنی بر شواهد را پیشنهاد کنند و نیز، پرونده‌های بالینی را دقیق‌تر مورد تحلیل و بررسی قرار دهند. از کارکردهای این سامانه می‌توان به تحلیل پرونده‌های پزشکی الکترونیکی (EMRs)، تفسیر تصویربرداری پزشکی و یافته‌های آزمایشگاهی، بررسی ادبیات علمی پزشکی مانند مقالات، راهنماها و پروتکل‌ها و در نهایت، ارائه پیشنهادهای درمانی شخصی‌سازی‌شده برای سرطان و دیابت اشاره نمود. برای مطالعه‌ی بیش‌تر رجوع کنید به:

1- Ross, Casey (2018) ‘‘IBM Watson recommended ‘unsafe and incorret’ cancer tratments’’, *Stat News*.

2- Lohre, Steve (2022) IBM Is selling Watson Health Assets’’, *The New York Times*.

را برای بیمار پیشنهاد کند که آسیب‌زا باشد، اساسا چه کسی مسئول است؟ پزشک؟ برنامه‌نویس سامانه‌های هوش مصنوعی؟ شرکت سازنده؟ بیمارستان؟ همگی؟ یا هیچ‌کدام؟[۱](میتِلشات، آلو، واکتر و فلوریدی، ۲۰۱۶، ۱۰-۷) .

نمونه‌های فوق، نشان می‌دهند که پرسش «چه کسی مسئول است»، اگر در چهارچوب فاعل منفرد و انسانی مطرح شود، اساسا پرسشی اشتباه خواهد بود؛ زیرا کنش اخلاقی در بستر فناورانه و مبتنی بر هوش مصنوعی محصول عاملیت مشارکتی انسان– ماشین است، نه تصمیم یک سوژه مستقل. بنابراین، بازتعریف مسئولیت اخلاقی به‌مثابه پدیده‌ای شبکه‌ای و میانجی مند، نه صرفا یک راه‌حل عملی، بلکه پاسخی نظری به دگرگونی ساختار کنش اخلاقی در عصر هوش مصنوعی است. به عبارت دیگر، فرآیند تصمیم‌گیری، نتیجه فرآیندی میان‌ذاتی است که در آن، انسان و سامانه هوشمند و تمام عوامل دخیل در فرآیند تصمیم به‌صورت مشارکتی در کنش اخلاقی حضور داشته باشند. به تعبیر آیدی، در چنین موقعیت‌هایی فناوری‌ها در امتداد ادراک انسانی بوده و زاویه‌دید ما را تعیین می‌کنند؛ هم‌چنین، فناوری بخشی از «قصدیت گسترده»[۲]است، که در آن، مسئولیت نیز باید به‌صورت توزیع‌شده و مشارکتی میان عوامل انسانی و غیرانسانی تحلیل شود[۳](آیدی، ۱۹۹۰، ۵۳).

براساس تمامی مطالب گفته شده، می‌توان گفت که شکاف مسئولیتی یکی از مهم‌ترین مسائل و چالش‌های مسئولیت اخلاقی به‌شمار می‌رود. در ادامه، ضمن معرفی مدل پیشنهادی برای برطرف نمودن این شکاف، می‌کوشیم رویکردی را معرفی، تفسیر و تبیین کنیم که هم به‌لحاظ مفهومی و هم کاربردی قابلیت تحقق داشته و به‌مثابه رویکردی نوین در جهت بازسازی مفهوم مسئولیت در عصر تعامل انسان– ماشین نگریسته و عنوان شود.

۶. مدل سه‌سطحی مسئولیت اخلاقی در هوش مصنوعی

مدل سه‌سطحی مسئولیت اخلاقی با به‌کارگیری و گسترش مسئولیت در سطوح مختلف در پی آن است تا مفهوم مسئولیت اخلاقی در بستر سیستم‌های هوش مصنوعی را در حیطه فردی، کاربردی و نهادی توسعه داده و افزون بر وجه نظری مفهوم فوق، امکان اجرایی و عملیاتی آن در سطوح بالاتر را نیز فراهم و بررسی نماید. به عبارت دیگر، این مدل می‌کوشد نشان دهد که پس از بازسازی و بازتعریف مفهوم مسئولیت، و تبدیل آن از شخص محوری (به‌معنای فاعل انسانی دارای آگاهی، قصد و اراده) به توزیع‌شده در بستر رابطه‌ای و شبکه‌ای، ترجیحا لازم است مفهوم آن را در سطوح سازمان‌ها و نهادهای نظارتی نیز، مشخص و گسترش دهیم. نکته محوری این مدل آن است که سطوح سه‌گانه‌ی یادشده صرفا سازوکارهای مدیریتی یا تنظیم گرانه نیستند، بلکه هر یک، واجد سطحی متمایز از نسبت اخلاقی با کنش، پیامد و آسیب هستند. به‌عبارت دیگر، این مدل نمی

---

[۱] سامانه IBM Watson علی‌رغم تبلیغات گسترده، اساسا با چالش‌های متعدد علمی، فنی و اخلاقی هم‌چون عملکرد ناپایدار در عمل، مشکلات داده‌ای، فقدان شفافیت الگوریتمی و فروش بخش‌هایی از پروژه روبه‌رو گردید. در بیماری‌هایی مانند سرطان، پیشنهادهای درمانی سامانه فوق، یا نادرست بود و یا با پروتکل‌های پذیرفته‌شده مغایرت داشت. به‌علاوه، عدم دسترسی به داده‌های باکیفیت و سوابق کامل بیماران باعث کاهش دقت مدل‌های الگوریتمی گردید. از این گذشته، پزشکان نمی‌توانستند تصمیم‌های سامانه را به‌طور کامل درک و یا ارزیابی کنند و خود همین مسئله، اعتماد پرسنل بیمارستان به‌ویژه پزشکان را کاهش داد. افزون بر این، در سال ۲۰۲۲، شرکت IBM بخش عمده‌ای از سامانه را به شرکت سرمایه‌گذاری Francisco Partners واگذار کرد، که نشان از شکست نسبی در تحقق اهداف تجاری اولیه داشت.

[2] extended intentionality

[۳] از منظر اخلاق اطلاعاتی فلوریدی، سامانه فوق‌الذکر، نمونه‌ای است از فاعلیت اطلاعاتی غیربشری که در زیست‌بوم اطلاعاتی نقش دارد و می‌تواند منشأ تغییرات مثبت یا منفی در وضعیت اخلاقی دیگران باشد (فلوریدی، ۲۰۱۳، ۸۹). در نتیجه، اخلاق در این بستر نمی‌تواند صرفاً به نیت انسانی محدود شود، بلکه باید رابطه‌ای از پاسخ‌گویی متقابل میان انسان و ماشین برقرار کند. روشن است که در سامانه IBM زاویه‌دید الگوریتمی جایگزین بخشی از قضاوت بالینی پزشک شده و بدون این‌که سازوکار بازبینی یا کنترل معنادار انسانی به‌کار گرفته یا تضمین شود، صرفا به خروجی‌های الگوریتمی سیستم بسنده شده است

کوشد فقدان فاعل اخلاقی واحد را صرفا با توزیع وظایف جبران کند، بلکه نشان می‌دهد که خود مفهوم مسئولیت اخلاقی در بستر کنش‌های فناورانه، ماهیتی رابطه‌ای، لایه‌مند و غیرقابل‌فروکاست به یک سطح خاص دارد.

۶-۱. سطح طراحی فردی و خُرد

در این سطح، مسئولیت متوجه افراد، مهندسان و طراحان است که تصمیم‌های بنیادین درباره داده‌ها، ساختار الگوریتم و معیارهای بهینه‌سازی می‌گیرند. سه اصل در این سطح فردی ضروری است: الف): مستندسازی طراحی: هر تصمیم فنی باید در اسناد قابل ردیابی، ثبت شود تا در صورت بروز آسیب، منشأ خطا مشخص باشد (فلوریدی، ۲۰۱۳، ۱۰۱-۹۹). ب): چک‌لیست اخلاق طراحی: پیش از استقرار هر سامانه، بررسی اخلاقی داده‌ها، معیارها و پیامدها الزامی است. ج): آموزش اخلاق فناوری: مهندسان باید درک کنند که هر تصمیم فنی واجد پیامد اخلاقی است. بر اساس دیدگاه یوناس، در عصر فناوری‌های خودمختار اخلاق باید اصل هدایت‌گر طراحان باشد (یوناس، ۱۹۸۴، ۱۲۶).

۶-۲. سطح سازمانی فرآیندی و میان‌سطحی

در سطح سازمانی، مسئولیت بر عهده شرکت‌ها، دانشگاه‌ها و نهادهایی است که سامانه‌های هوشمند را توسعه یا به‌کار می‌گیرند. این سطح باید سازوکارهایی برای حسابرسی، اصلاح و جبران خطا ایجاد کند. پیشنهادهای اصلی عبارت‌اند از: الف): تشکیل شورای حسابرسی اخلاقی درون‌سازمانی که به‌طور منظم عملکرد مدل‌ها را بررسی می‌کند. ب): ثبت خودکار همه تصمیم‌ها، تغییرات داده و نسخه‌های مدل در سامانه لاگ استاندارد[۱](سلبست، فریدلر، ونکاتاسوبرامانیان و ورتسی، ۲۰۱۹، ۶۱). ج): ایجاد سازوکار اصلاح سریع[۲]برای پاسخ فوری به آسیب‌ها. به تعبیر فیلیپ بری، سازمان‌ها باید نهادهای اخلاقی یادگیرنده‌ای شوند که در برابر خطاها مسئولیت جمعی می‌پذیرند (بری، ۲۰۱۲، ۳۱۰).

۶-۳. سطح نهادی و حکمرانی کلان

در سطح سوم، مسئولیت از مرز سازمان‌ها فراتر می‌رود و به سیاست‌گذاری و قانون‌گذاری می‌رسد. این سطح شامل دولت‌ها، نهادهای نظارتی و نظام‌های حقوقی است. اصول پیشنهادی این سطح عبارتند از: الف): شفافیت اجباری: الزام شرکت‌ها به انتشار گزارش شفافیت الگوریتمی و توضیح درباره داده‌ها و معیارها. ب): تشکیل نهادهای پاسخ‌گویی هوش مصنوعی: سازمان‌های مستقل با اختیارات حقوقی برای بررسی شکایات و اعمال جریمه (گزارش AI Now، ۲۰۲۳، ۶-۴). ج): حق دسترسی به عدالت الگوریتمی: شهروندان باید بتوانند نسبت به تصمیم‌های خودکار اعتراض کنند و سازوکار ترمیم خسارت وجود داشته باشد. این سطح تحقق همان چیزی است که فلوریدی آن را حکمرانی اخلاق اطلاعات می‌نامد (فلوریدی، ۲۰۱۳، ۱۰۳).

بر این اساس، مدل سه‌سطحی مسئولیت اخلاقی در هوش مصنوعی نه جایگزینی فاعل اخلاقی واحد با مدیریت سیستمی، بلکه تبیینی فلسفی از چگونگی تداوم معنا و امکان مسئولیت اخلاقی در شرایطی است که کنش اخلاقی به‌طور ذاتی محصول تعامل انسان‌ها، مصنوعات فناورانه و ساختارهای نهادی است. در این چهارچوب، مسئولیت نه حذف می‌شود و نه صرفاً اجرایی می‌گردد، بلکه به‌مثابه مفهومی هنجاری، در سطوح مختلف کنش و تصمیم بازتوزیع و بازتفسیر می‌شود.

---

[۱] Standard logging system یا سامانه لاگ استاندارد، سیستمی است که تمام رخدادهای مهم، تصمیم‌ها، ورودی‌ها، خروجی‌ها و تغییرات سامانه‌های هوش مصنوعی یا نرم‌افزار را به‌صورت ساختاریافته ثبت می‌کند. درواقع، این سامانه یک چهارچوب ثبت، ذخیره‌سازی و مدیریت رویدادها در یک سیستم نرم‌افزاری یا سامانه مبتنی بر هوش مصنوعی است که برای ردیابی، مستندسازی، پایش، ارزیابی و پاسخ‌گویی استفاده می‌شود. این مفهوم در حوزه اعتمادپذیری هوش مصنوعی و پاسخ‌گویی الگوریتمی بسیار مهم است.

[2] Remediation

۷. پیشنهادهای سیاستی و طراحی

این بخش مکمل مدل سه‌سطحی است و هدفش ارائه پیشنهادهای عملی برای تبدیل چهارچوب نظری به سازوکارهای قابل اجرا در طراحی، سازمان و سیاست‌گذاری است. به عبارت دیگر، هدف این بخش آن است که مفاهیم نظری مسئولیت اخلاقی در هوش مصنوعی به راه‌کارهای عملی و سیاستی قابل اجرا تبدیل شوند. اگرچه فلسفه می‌تواند بنیان‌های ارزشی را روشن کند، اما بدون ترجمه نهادی و طراحی دقیق، این ارزش‌ها در سطح شعار باقی می‌مانند. گفتنی است یکی از مهم‌ترین پیشنهادهای سیاستی قابل اجرا، کنترل معنادار انسانی، طراحی اخلاقی، و پاسخ‌گویی الگوریتمی است، که پیش‌تر به‌هنگام بررسی بحث شکاف در مسئولیت بدان پرداخته شد.

۷–۱. سیاست‌های طراحی و توسعه؛ مستندسازی و شفافیت طراحی: از اخلاق نیت تا اخلاق سند

یکی از مهم‌ترین عوامل شکل‌گیری شکاف مسئولیتی، نبود ردپا یا سند در فرآیند طراحی است. وقتی سامانه‌ای تصمیمی اخلاقی نادرستی می‌گیرد، در غیاب اسناد طراحی، مشخص نیست چه کسی آن را و با چه نیتی اتخاذ کرده است (یوناس، ۱۹۸۴، ۱۲۵–۱۲۲). به تعبیر آیدی، در عصر فناوری‌های خودمختار، اخلاق مبتنی بر نیت دیگر کفایت ندارد، و باید اخلاق پیش‌بینی‌گر و مستندساز جای آن را بگیرد (آیدی، ۱۹۹۰، ۱۲۵–۱۲۲). از این منظر، الزام به مستندسازی تصمیم‌های طراحی به‌عنوان نخستین سیاست اخلاقی مطرح می‌شود. این الزام عبارتند از: الف): نگه‌داری نسخه‌های تمام الگوریتم‌ها و تغییرات آن‌ها به‌صورت زمان‌بندی‌شده؛ ب): مستندسازی معیارهای انتخاب داده و روش‌های پیش‌پردازش؛ ج): ثبت دلایل انتخاب اهداف بهینه‌سازی، مانند کاهش خطا در برابر حفظ عدالت. به عبارت دیگر، برای رفع شکاف مسئولیت، هر سامانه هوش مصنوعی باید «شناسنامه مدل»[1]و «شناسنامه داده»[2]داشته باشد (گبرو، مورگنسترن، وچیونه، وگان، والاک، دام سوم و کرافورد، ۲۰۱۸). این اسناد باید شامل منشأ داده، معیارهای انتخاب، محدودیت‌ها و شاخص‌های عملکرد باشند و در دسترس نهادهای نظارتی قرار گیرند. چنین شفافیتی نه‌تنها امکان حساب‌رسی را فراهم می‌کند، بلکه باعث می‌شود طراحان در برابر پیامدهای اخلاقی تصمیمات خود آگاهانه‌تر رفتار کنند[3](سلبست، فریدلر، ونکاتاسوبرامانیان و ورتسی، ۲۰۱۹، ۶۲–۶۰).

۷–۲. ارزیابی تأثیر اجتماعی و اخلاقی پیش از استقرار

ارزیابی تأثیر اجتماعی[4] ترجمه عملی اصل کانتی «غایت بودن انسان» است که با تکیه بر آن، می‌توان مدعی شد:

> هیچ سامانه‌ای نباید به‌گونه‌ای طراحی شود که انسان را صرفاً ابزار بهینه‌سازی سود یا سرعت قرار دهد (بینز، ۲۰۱۸، ۵۴۶).

در سطح اجرایی، این ارزیابی باید در سه مرحله انجام گیرد: الف): مرحله طراحی: یعنی شناسایی گروه‌های ذی‌نفع، مانند کاربران، اقلیت‌ها، محیط‌زیست و نهادهای عمومی. ب): مرحله توسعه: یعنی ارزیابی احتمال بروز تبعیض، حذف یا آسیب اجتماعی از طریق تحلیل داده و سنجش سناریوهای فرضی. ج): مرحله استقرار: یعنی نظارت بر آثار واقعی و بازبینی سیاست‌ها در صورت مشاهده پیامدهای ناخواسته. در واقع، اجرای ارزیابی تأثیر اجتماعی باید الزام قانونی پیدا کند و نتایج آن در دسترس عموم قرار گیرد.

---

[1] Model Card

[2] Data Sheet

[3] اجرای این سیاست، یعنی سیاست طراحی، نیازمند ابزارهای فنی مانند «طراحی اخلاقی» یا ethical by design است؛ یعنی پایگاه داده‌ای که هر تصمیم فنی را با برچسب اخلاقی ثبت کند. در صورت بروز آسیب، این سند امکان بازسازی زنجیره‌ی علّی را فراهم می‌کند و به تخصیص منصفانه مسئولیت می‌انجامد. به این ترتیب، مستندسازی، پُلی میان نظریه اخلاقی و مسئولیت حقوقی ایجاد می‌کند.

[4] Social impact assessment

تجربه اتحادیه‌ی اروپا در چهارچوب AI Act نشان داده است که چنین الزام‌هایی می‌توانند شرکت‌ها را وادار به طراحی اخلاقی‌تر کنند. به عبارت دیگر، پیش از اجرای هر سامانه با ریسک بالا، سازمان‌ها باید ملزم به انجام ارزیابی تأثیر اجتماعی[1]شوند. این ارزیابی‌ها باید ابعاد عدالت، شمول‌پذیری، تبعیض، حریم خصوصی و آثار فرهنگی را بسنجند (بینز، ۲۰۱۸، ۵۴۶). چنین ارزیابی‌هایی می‌توانند از بروز سوگیری‌ها[2]و شکاف‌های اخلاقی در مراحل اولیه جلوگیری کنند.

۷-۳. نهادهای مستقل پاسخ‌گویی و نظارت

ایجاد نهادهای مستقل برای نظارت بر هوش مصنوعی از پیشنهادهای کلیدی است. گزارش[3]AI Now تاکید می‌کند که ممیزی داخلی شرکت‌ها کافی نیست و نیاز به نهادهای بیرونی با قدرت اجرایی وجود دارد. این نهادها باید وظیفه بررسی شکایات، اجرای تحریم‌ها و تدوین استانداردهای ملّی اخلاقی را بر عهده داشته باشند (AI Now، ۲۰۲۳، ۶-۵). بر اساس گزارش این مؤسسه، به‌رغم هدف اولیه سامانه‌های مبتنی بر هوش مصنوعی فوق‌الذکر، با وجود تلاش برای ارتقاء شفافیت و بهره‌وری شهروندان، هم‌چنان ابهام در پاسخ‌گویی و نامشخص بودن مسئول و مرجع رسیدگی به وضعیت پیش‌آمده به‌چشم می‌خورد. این گزارش هشدار می‌دهد که فقدان معیارهای مسئولیت‌پذیری دقیق در چنین پلتفرم‌هایی می‌تواند میان شهروندان جامعه تبعیض ایجاد کند؛ چرا که داده‌ها بدون ملاحظه بافت اجتماعی و بدون دخالت انسانی تفسیر می‌شوند. به‌نظر می‌رسد طراحی و پیاده‌سازی الگوریتم‌های سامانه‌ها بدون لحاظ چهارچوب‌های مسئولیت در هر منطقه و بافت شهری، می‌تواند به نتایج ضد عدالت اجتماعی منتهی شود. در واقع، سیستم‌ها باید به‌گونه‌ای طراحی شوند که واسطه‌گری فناوری در اخلاق عمومی در همه‌ی مناطق شهری به رسمیت شناخته شود، اما در بسیاری از شهرها، تصمیمات مهم بدون شفافیت کافی اتخاذ می‌شوند، و شهروندان نمی‌دانند که چرا و چگونه از برخی خدمات شهروندی محروم شده‌اند[4](فربیک، ۲۰۱۱، ۱۰۲- ۸۹).

۷-۴. نهادینه‌سازی آموزش میان‌رشته‌ای اخلاق و فناوری

مسئولیت اخلاقی تنها از طریق قوانین به دست نمی‌آید؛ بلکه مستلزم فرهنگ‌سازی است. آموزش میان‌رشته‌ای برای مهندسان، فیلسوفان و مدیران باید اجباری شود تا هر گروه درک مشترکی از مخاطرات اخلاقی و اجتماعی فناوری پیدا کند (یوناس، ۱۹۸۴، ۱۲۹). اگرچه قوانین و ساختارها مهم‌اند، اما هیچ سازوکار بیرونی جایگزین قضاوت اخلاقی درونی نمی‌شود (بری، ۲۰۱۲، ۳۱۲-۳۱۰). در واقع، مهندسان باید بتوانند اثرات اجتماعی کار خود را پیش‌بینی کنند که این مهم، تنها با تربیت اخلاقی ممکن است. از این رو، لازم است برنامه‌های آموزش میان‌رشته‌ای در دانشگاه‌ها و صنایع به‌صورت رسمی نهادینه شوند، که ذیل چند پیشنهاد اجرایی شامل این موارد خواهند بود: الف): گنجاندن درس‌های «اخلاق هوش مصنوعی» و «فلسفه فناوری» در برنامه آموزشی رشته‌های مهندسی. ب): برگزاری کارگاه‌های مشترک میان گروه‌های فلسفه، حقوق و مهندسی در سازمان‌ها. ج): ایجاد

---

[1] Social impact assessment

[2] Bias

[3] AI Now یک موسسه پژوهشی در نیویورک است که به بررسی پیامدهای اجتماعی، اخلاقی و سیاسی هوش مصنوعی تمرکز کرده و سالانه آخرین و به‌روزترین تحقیقات پژوهشی‌اش را در معرض مطالعه عموم قرار می‌دهد.

[4] در همین راستا، فلوریدی پیشنهاد می‌کند که اخلاق فناوری در نهایت باید به حکم‌رانی اطلاعات منجر شده و ترجیحا می‌بایست سازوکارهایی طراحی کرد که نه‌تنها ارزش‌ها را تعریف، بلکه از آن‌ها پاسداری کنند (فلوریدی،۲۰۱۳ ، ۱۰۳-۹۷). به همین دلیل، او متذکر می‌شود که باید نهادهای مستقل پاسخ‌گویی هوش مصنوعی در سطح ملّی و بخشی تشکیل شوند. کارکرد این نهادها سه‌گانه است: الف): نظارت پیشینی: یعنی بررسی سامانه‌ها پیش از استقرار، بر اساس معیارهای شفافیت و عدالت . ب): نظارت پسینی: یعنی ارزیابی شکایات و بازرسی از عملکرد سیستم‌های مستقر. ج): اقدام اصلاحی: یعنی صدور توصیه‌نامه یا اعمال جریمه در صورت نقض اصول اخلاقی. ترکیب این نهادها باید میان‌رشته‌ای باشد، یعنی فلاسفه، حقوق‌دانان، مهندسان داده و جامعه‌شناسان می‌بایست با یک‌دیگر در مورد این توصیه‌نامه‌ها دستورالعمل‌ها مشارکت داشته باشند، هم‌چنین، استقلال مالی و ساختاری آن‌ها تضمین می‌کند که ارزیابی‌ها بی‌طرفانه انجام شود. علاوه بر آن، هر نهاد موظف است گزارش سالانه شفافیت الگوریتمی را به‌گونه‌ای منتشر کند که شامل موارد تخلف، اصلاح و نوآوری‌های اخلاقی باشد؛ چنین گزارش‌هایی در نهایت، هم ابزاری آموزشی و هم ابزار بازدارنده محسوب می‌شوند

گواهینامه‌های حرفه‌ای اخلاق فناوری[1] هدف نهایی، ایجاد نوعی اخلاق بازتابی است که درون فرآیند، طراحی نهادینه شود، نه این‌که پس از حادثه یادآوری گردد[2]

۷-۵. سازوکارهای ترمیم، جبران و یادگیری سازمانی

هیچ سازوکاری از خطا مصون نیست؛ بنابراین نظام مسئولیت اخلاقی باید نه‌تنها پیش‌گیرانه، بلکه ترمیمی و یادگیرنده باشد. پیشنهاد می‌شود که سازمان‌ها مکلف به ایجاد فرآیندهای جبران و یادگیری اخلاقی شوند. این فرآیندها شامل سه مرحله است: الف): شناسایی و اعلام خطا: یعنی هر رخداد اخلاقی یا آسیب ناشی از سیستم باید به‌صورت رسمی مستندسازی و گزارش شود. ب): جبران خسارت: یعنی در صورت تأیید آسیب، مکانیزم‌های مالی یا حقوقی جبران باید فعال شوند. ج): بازبینی طراحی: یعنی نتایج بررسی باید در طراحی نسل بعدی سیستم‌ها به‌کار گرفته شود تا خطا تکرار نشود. اجمالا چنین سازوکاری فلسفه مسئولیت را از «پاسخ به گذشته» به «یادگیری برای آینده» تغییر می‌دهد؛ چیزی که یوناس با عنوان «اخلاق پیش‌بینی» از آن یاد می‌کند (یوناس، ۱۹۸۴، ۱۳۰).

۷-۶. پیوند میان اخلاق و حکمرانی: از نظریه به نهاد

در نهایت، باید اذعان کرد که مسئولیت اخلاقی در هوش مصنوعی، تنها در صورتی معنا دارد که به بخشی از نظام حکمرانی تبدیل شود. همان‌طور که فلوری می‌گوید ما به‌جای «فناوری مسئول» به «اکوسیستم مسئولیت» احتیاج داریم» (فلوریدی، ۲۰۱۳، ۱۰۳)، یعنی شبکه‌ای از طراحان، قانون‌گذاران، ناظران، کاربران و شهروندان که هر یک وظیفه‌ای معین در حفظ ارزش‌های اخلاقی دارند؛ و باید سیاست‌هایی تدوین شود که میان اخلاق، قانون و فناوری تعامل پویایی برقرار کند. سیاست‌هایی از قبیل: الف): تدوین منشور ملّی اخلاق هوش مصنوعی با همکاری دانشگاه‌ها و نهادهای قانون‌گذار؛ ب): ایجاد پایگاه داده عمومی از پروژه‌های اخلاقی و پژوهش‌های مرتبط برای تبادل تجربه؛ ج): طراحی شاخص‌های ملّی بلوغ اخلاقی فناوری برای سنجش پیشرفت سازمان‌ها در اجرای اصول اخلاقی[3]

نتیجه‌گیری

پژوهش حاضر در ابتدا با بحث از مبانی مفهومی مسئولیت اخلاقی در سنت فلسفی آغاز نمود و نشان داد که این مفهوم در عین تفاوت در معنا و تفسیر مسئولیت اخلاقی در سنت‌های فلسفی کلاسیک بر پیش‌فرض‌هایی چون قصد، اراده آزاد و علیّت استوار بوده است. در مرحله بعد، همان‌گونه که ملاحظه گردید، ورود فناوری‌های نوینی هم‌چون هوش مصنوعی و یادگیری ماشینی به عنوان کنش‌گرانی غیرانسانی، این پیش‌فرض‌ها را با چالش مواجه ساخت و به ظهور شکاف‌هایی در انتساب مسئولیت منجر شد. پس از آن، به معرفی و تحلیل رویکردهای برجسته در پاسخ‌گویی به این وضعیت پرداخته و رویکردهایی معرفی شدند که هر یک تلاش نمودند با مفاهیمی چون مسئولیت توزیع‌شده، کنترل معنادار انسانی (شرط امکان فاعلیت)، طراحی اخلاقی (فراهم‌کننده ساختار پیش‌گیرانه) یا حساب‌رسی الگوریتمی (تضمین‌کننده شفافیت و عدالت پس از عمل) به این چالش‌ها پاسخ دهند.

با تحلیل و تفسیر رویکردهای فلسفه فناوری، چنین اذعان شد که تنها در ترکیب این سه سطح و ضلع مثلث است که می‌توان از فروپاشی مسئولیت اخلاقی جلوگیری کرد؛ به‌نحوی‌که اگر هر یک از اضلاع غایب باشد، مسئولیت یا در غیاب طراحی

---

[1] Professional Ethics Certifications

[2] این پیشنهادها و آموزش‌ها نباید صرفاً نظری باشند، بلکه باید بر حل مسئله‌ی واقعی تمرکز کنند. مهندسان باید یاد بگیرند چگونه در موقعیت‌های مبهم ، هم‌چون تعارض بین عدالت و کارایی تصمیم بگیرند.

[3] تمامی این اقدامات، اخلاق را از سطح نظری به سطح سیاست عمومی می‌کشانند و امکان ارزیابی و بهبود مستمر را فراهم می‌کنند.

اخلاقی به انسان تحمیل ناعادلانه می‌شود، یا در غیاب کنترل معنادار انسانی در شبکه فنی گم می‌شود، یا در غیاب پاسخ‌گویی الگوریتمی، در پسِ پیچیدگی فنی پنهان می‌گردد. به‌این ترتیب، مدل «مثلث مسئولیت» به‌مثابه پُلی، میان فلسفه اخلاق و حکم‌رانی فناورانه نگریسته شده، که می‌تواند بنیان مقالات را از تحلیل صرف نظری به سطحی عملیاتی و سیاست‌گذارانه ارتقا دهد.

در ادامه، با نگاهی نقادانه، کاستی‌ها، نابسندگی‌ها و محدودیت‌های این رویکردها مشخص و زمینه برای ارائه چهارچوبی نو فراهم شد. در خاتمه، چهارچوب پیشنهادی پژوهش، با بازتعریف مفاهیم بنیادین مسئولیت اخلاقی، ارائه گردید؛ چهارچوبی که بر تعامل میان انسان، ماشین و نهادهای اجتماعی و سیاسی تاکید دارد. در این میان، کاربست این چهارچوب در حوزه‌های خاص، از جمله عدالت کیفری و سلامت دیجیتال مورد بررسی قرار گرفت و با تحلیل موردی پلتفرم‌هایی چون COMPAS و IBM Watson قابلیت این چهارچوب را در مواجهه با مسائل واقعی نشان داده شد. ارزیابی دو پرونده فوق نشان داد که بحران مسئولیت اخلاقی در هوش مصنوعی نه صرفا ناشی از نقص فنی بلکه ناشی از ضعف و لزوم تغییر نگرش در درک فلسفی و تعریف سنتی فاعلیت و پاسخ‌گویی است. در هر دو نمونه، عدم شفافیت ساختاری، فقدان کنترل انسانی و نهادهای نظارتی و نیز، نهادینه‌نشدن پاسخ‌گویی، موجب فروپاشی پاسخ‌گویی و از هم‌گسیختگی پیوند میان کنش و پیامد گردیده است. به تعبیر جیمز مور، «وقتی کنش‌گران در زنجیره تصمیم‌گیری از هم گسسته‌اند، ضرورت پاسخ‌گویی از میان رفته و مسئولیت، بدون توجه به آن تعریف و تفسیر می‌گردد (مور، ۲۰۰۶، ۱۸).

منابع

امیریان فارسانی، امین. (۱۴۰۳). هوش مصنوعی و سیاست‌گذاری عدالت کیفری در ایران معاصر: فرصت‌ها و چالش‌های نوین برای کارگزاران قضایی. فصلنامه تحولات سیاسی و اجتماعی معاصر ایران، ۴(۵)، ۹۵-۸۰. https://www.noormags.ir/view/fa/articlepage/2310711

ثمر بخش زاده، سیما. (۱۴۰۳). آموزش هوشمند و نقش فناوری در کلاس درس. *ششمین کنفرانس بین‌المللی مدیریت، آموزش و پژوهش‌های آموزشی در آموزش*، ۴۹۳۳-۴۹۱۷. https://www.noormags.ir/view/fa/articlepage/2320378

سپهری قوجه، فرهاد. (۱۴۰۳). معماری و هوش مصنوعی: تأثیر یادگیری ماشین بر طراحی و ساخت. *فصلنامه معماری سبز*، ۱۰(۷)، ۲۰-۱۳. https://www.noormags.ir/view/fa/articlepage/2266583

شرفی، وحید و محمدیاری، زهره. (۱۴۰۳). کاربردهای هوش مصنوعی در مدیریت سلامت سالمندان: رویکرد فراترکیب. *مجله سالمندی ایران*، (۷۹)، ۴۲۱-۴۰۳. https://www.noormags.ir/view/fa/articlepage/2305569

عسگری، محمد و بهرامی، مهدی. (۱۴۰۳). مدیریت آموزشی در مدارس با استفاده از هوش مصنوعی. *فصلنامه پژوهش‌های راهبردی در آموزش و پرورش*، ۵۹(۳)، ۱۴۸-۱۳۱. https://www.noormags.ir/view/fa/articlepage/2328339

عیدلی، سعید. (۱۴۰۳). بررسی راهکارهای استفاده از هوش مصنوعی در بهبود فرآیند اداره مدارس ایران. *ششمین کنفرانس بین‌المللی مدیریت، آموزش و پژوهش‌های آموزشی در آموزش*، ۴۹۳۳-۴۹۱۷. https://www.noormags.ir/view/fa/articlepage/2320109

عمرانی، شکیب و امیرشاه کرمی، سید نجم‌الدین. (۱۴۰۳). فناوری هوش مصنوعی در ایده‌پردازی و تولید محتوای مفهومی در پویانمایی «کلاغ» و «مرا بغل نکن، می‌ترسم» و گنوسیس. *فصلنامه علمی-پژوهشی دانشگاه ایران*، (۴۶)، ۹۶-۷۹. https://www.noormags.ir/view/fa/articlepage/2307963

نصر اصفهانی، محمد؛ قائمی اصل، مهدی؛ منتظر، راحله و اسماعیلی، ملیکا. (۱۴۰۳). نقش قابلیت‌های توانمندساز هوش مصنوعی در عملکرد نظارتی نظام بانکی ایران. *فصلنامه مطالعات کشوری*، ۳(۴)، ۷۵۳-۷۱۳. https://www.noormags.ir/view/fa/articlepage/2313268

References

AI Now Institute. (2023). *AI Now 2023 Report: Governing AI in the Public Interest*. AI Now Institute, New York University.

Amirian Farsani, A. (2025). Artificial intelligence and criminal justice policymaking in contemporary Iran: New opportunities and challenges for judicial officers. *Quarterly Journal of Contemporary Political and Social Developments of Iran*, 4(5), 80–95. (in Persian). https://www.noormags.ir/view/fa/articlepage/2310711/

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*.

Aristotle. (1895). *The Nicomachean Ethics of Aristotle* (R. W. Browne, Trans.). George Bell and Sons.

Asgari, M., & Bahrami, M. (2025). Educational management in schools using artificial intelligence. *Quarterly Journal of Strategic Research in Education and Training*, 59(3), 131–148. (In Persian). https://www.noormags.ir/view/fa/articlepage/2328339

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency* (pp. 543–552).

Brey, P. (2012). Anticipating ethical issues in emerging IT. *Ethics and Information Technology*, 14(4), 305–317.

Eidli, S. (2025). Examining strategies for the use of artificial intelligence in improving the process of Iran's department of school. In *6th International Conference on Management, Education and Training Research in Education* (pp. 4917–4933). (in Persian). https://www.noormags.ir/view/fa/articlepage/2320109

Floridi, L. (2013). *The Ethics of Information*. Oxford University Press.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3), 349–379.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv preprint*, arXiv:1803.09010.

Ihde, D. (1990). *Technology and the Lifeworld: From Garden to Earth*. Indiana University Press.

Jonas, H. (1984). *The Imperative of Responsibility*. University of Chicago Press.

Kant, I. (1785). *Groundwork of the Metaphysics of Morals*. Cambridge University Press.

Latour, B. (2005). *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press.

Lohr, S. (2022). *IBM is selling Watson Health assets*. The New York Times.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 7–10.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4), 18–21.

Nasre Esfahani, M., Qaemi Asl, M., Montazer, R., & Esmaeeli, M. (2025). The role of artificial intelligence's enabling capabilities in the supervisory performance of Iran's banking system. *Journal of Country Studies*, 3(4), 713–753. (in Persian). https://www.noormags.ir/view/fa/articlepage/2313268

Omrani, S., & Amirshah Karami, S. N. (2024). Artificial intelligence technology in idea generation and conceptual content production in animation "Kalagh" and "Don't Hug Me, I'm Afraid", and Gnosis. *Scientific Research Quarterly of Iran University*, (46), 79–96. (in Persian). https://www.noormags.ir/view/fa/articlepage/2307963/

Ross, C. (2018a). *IBM Watson recommended 'unsafe and incorrect' cancer treatments. STAT* News.

Ross, C., & Swetlitz, I. (2018b). *IBM's Watson recommended 'unsafe and incorrect' cancer treatments*. STAT News.

Samarbafzade, S. (2025). Smart education and the role of technology in the classroom. In *6th International Conference on Management, Education and Training Research in Education* (pp. 4917–4933). (in Persian). https://www.noormags.ir/view/fa/articlepage/2320378

Santoni de Sio, F., & Mecacci, G. (2021). The responsibility gap in AI. *Philosophy & Technology*, 34(1), 105–128.

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A theoretical framework. *Frontiers in Robotics and AI*, 5, 15.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Algorithmic accountability: Moving beyond audits. *Communications of the ACM*, 62(7), 58–66.

Sepehri Ghooje, F. (2024). Architecture and artificial intelligence: The impact of machine learning on design and construction. *Green Architecture Quarterly*, 10(7), 13–20. (in Persian). https://www.noormags.ir/view/fa/articlepage/2266583

Sharafi, V., & Mohammad Yari, Z. (2025). Applications of artificial intelligence in elderly healthcare management: A meta-synthetic approach. *Iranian Journal of Aging*, (79), 403–421. (in Persian). https://www.noormags.ir/view/fa/articlepage/2305569

Verbeek, P.-P. (2005). *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. Penn State Press.

Verbeek, P.-P. (2011). *Moralizing Technology: Understanding and Designing the Morality of Things*. University of Chicago Press.