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

Artificial Intelligence and the Transformation of Human Autonomy: an analysis of free will and moral responsibility in data-driven decision-making

Tayyebeh Gholami[1] ✉ | **Seyed Hassan Hosseini Sarvary**[2]

1. Corresponding Author, Postdoctoral Researcher, Department of Philosophy of Science, Sharif University of Technology, Tehran, Iran. Email: t.gholami@staff.sharif.edu.
2. Professor, Department of Philosophy of Science, Sharif University of Technology, Tehran, Iran. Email: hoseinih@sharif.edu.

Article Info

Article type:
Research Article

Article history:
Received 04 May 2026
Received in revised from 30 May 2026
Accepted 03 June 2026
Published online 14 July 2026

Keywords:
Artificial Intelligence, Autonomy, Free Will, Moral Responsibility

ABSTRACT

The proliferation of Artificial Intelligence (AI) systems in daily life has raised fundamental questions regarding the boundaries of human autonomy and the meaning of free will. This paper demonstrates that intelligent systems, by imperceptibly shaping preferences and rearranging the horizon of choice, transform and redefine the form and scope of the experience of human autonomy. Employing an analytical-critical approach and utilizing the framework of AI ethics, this study examines three key axes: (1) the relationship between algorithmic prediction and the conditions for the possibility of human choice; (2) the ethical consequences and the redefinition of responsibility distribution in algorithmically guided environments; and (3) the necessity of developing responsible design principles and regulatory frameworks to preserve human decision-making independence. The theoretical contribution of the paper lies in utilizing a network approach to agency and responsibility, which facilitates a more precise analysis of the "responsibility gap" and a redefinition of the roles of humans and machines in complex decision-making processes. The results indicate that human autonomy undergoes a structural transformation in the age of AI; humans remain responsible, yet significant segments of the decision-making chain—from data collection to algorithmic recommendation—lie outside their direct control.

Cite this article: Gholami, T. & Hosseini Sarvary, H. (2026). Artificial Intelligence and the Transformation of Human Autonomy: an analysis of free will and moral responsibility in data-driven decision-making. *Journal of Philosophical Investigations*, 20 (55), 265-284. https://doi.org/10.22034/jpiut.2026.72239.4495

 Publisher: University of Tabriz.

Extended Abstract

Introduction

The proliferation of Artificial Intelligence (AI) systems in daily life has raised fundamental questions regarding the boundaries of human autonomy and the meaning of free will. Algorithms are no longer passive data-processing tools but active agents in shaping consumer preferences, medical diagnoses, and even legal judgments. This paper demonstrates that intelligent systems, by imperceptibly shaping preferences and rearranging the horizon of choice, transform and redefine the form and scope of the experience of human autonomy. The central research question is: If human decisions are increasingly influenced by machine prediction and guidance, can we still defend the possibility of autonomous moral action?

Methodology and Theoretical Framework

Employing an analytical-critical approach and utilizing the framework of AI ethics, this study examines three key axes: (1) the relationship between algorithmic prediction and the conditions for the possibility of human choice; (2) the ethical consequences and the redefinition of responsibility distribution in algorithmically guided environments; and (3) the necessity of developing responsible design principles and regulatory frameworks to preserve human decision-making independence. The theoretical contribution of the paper lies in utilizing a network approach to agency and responsibility (Actor-Network Theory), which facilitates a more precise analysis of the "responsibility gap" and a redefinition of the roles of humans and machines in complex decision-making processes.

Main Arguments and Findings

First, algorithmic prediction structurally weakens human autonomy. Intelligent systems, through personalization and filtering, present users with an "algorithmic version" of choices. The phenomenon of "automation bias" leads individuals to trust system outputs even when evidence of error exists. Consequently, the normative capacity for ethical accountability is gradually eroded. The paper argues that what is weakened is not merely empirical freedom of choice but the Kantian condition of the "autonomy of practical reason."

Second, the paper addresses the "responsibility gap." Using Constantinescu et al.'s four-condition Aristotelian test (causality, freedom, knowledge, and deliberation), the analysis shows that Artificial Moral Advisors (AMAs) lack the intrinsic intentionality and practical wisdom required for full moral agency. While AI systems have meaningful ethical impacts, they cannot be held fully responsible, yet humans cannot bear complete responsibility either. This paradox reveals that the ethical problem cannot be reduced to individual choice.

Third, the paper introduces Actor-Network Theory (ANT) as a solution. Moving beyond the dualism of human versus machine, agency and responsibility are redefined as relational constructs emerging from networks of human and non-human actors (algorithms, data, institutions, designers, users). In this framework, responsibility is distributed across multiple layers: the design level (developers who shape choice horizons), the institutional level (regulators ensuring accountability), and the user level (individuals evaluating options).

Fourth, the paper identifies three levels of autonomy—formal (existence of options), cognitive (understanding consequences), and normative (evaluating alternatives)—and

demonstrates that algorithms may preserve the first while structurally undermining the second and third, creating an "illusion of autonomy."

Theoretical Contribution and Novelty

The paper's novelty lies in redefining autonomy and moral responsibility within a "network-normative" framework. Unlike existing approaches that either emphasize individual human agency or seek to attribute independent agency to machines, this research offers a model of "distributed responsibility" in which agency is understood not as an intrinsic property of an agent but as a relational construct emerging from the interaction of humans, algorithms, and institutional structures. This framework bridges the gap between descriptive analysis and moral prescription by providing normative criteria for reallocating responsibility across design, regulation, and action levels.

Conclusion

The results indicate that human autonomy undergoes a structural transformation in the age of AI; humans remain responsible, yet significant segments of the decision-making chain—from data collection to algorithmic recommendation—lie outside their direct control. The main challenge emerges not in the technical capability of machines but in the normative reconfiguration of the relationship between humans, algorithms, and moral responsibility. Understanding this transformation can reshape future research in the philosophy of mind, technology ethics, and algorithmic governance. Any theory of moral responsibility that continues to rely on an individualistic model of control and choice will prove inadequate for explaining action in the algorithmic lifeworld.

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

هوش مصنوعی و دگرگونی خودمختاری انسان: تحلیلی از اراده آزاد و مسئولیت اخلاقی در تصمیم‌سازی داده‌محور

طیبه غلامی[۱] ✉ | سید حسن حسینی سروری[۲]

۱. نویسنده مسئول، پژوهشگر پسادکتری، گروه فلسفه علم، دانشگاه صنعتی شریف، تهران، ایران. رایانامه: mailto:t.gholami@staff.sharif.edu
۲. استاد گروه فلسفه علم، دانشگاه صنعتی شریف، تهران، ایران: رایانامه: . mailto:hoseinih@sharif.edu

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۴۰۵/۰۲/۱۴
تاریخ بازنگری: ۱۴۰۵/۰۳/۰۹
تاریخ پذیرش: ۱۴۰۵/۰۳/۱۳
تاریخ انتشار: ۱۴۰۵/۰۴/۲۴

کلیدواژه‌ها:
هوش مصنوعی، خودمختاری، اراده آزاد، مسئولیت اخلاقی

چکیده

گسترش سامانه‌های هوش مصنوعی در زندگی روزمره، پرسش‌های بنیادینی درباره حدود خودمختاری انسانی و معنای اراده آزاد مطرح کرده است. این مقاله نشان می‌دهد که سامانه‌های هوشمند با جهت‌دهی نامحسوس ترجیحات و بازآرایی افق انتخاب، شکل و دامنه تجربه خودمختاری انسانی را دگرگون و بازتعریف می‌کنند. پژوهش حاضر با رویکرد تحلیلی–انتقادی و بهره‌گیری از چارچوب اخلاق هوش مصنوعی، سه محور را بررسی می‌کند: (۱) نسبت میان پیش‌بینی الگوریتمی و شرایط امکان انتخاب انسان؛ (۲) پیامدهای اخلاقی و بازتعریف توزیع مسئولیت در محیط‌های هدایت‌شده الگوریتمی؛ و (۳) ضرورت تدوین اصول طراحی مسئولانه و چارچوب‌های نظارتی برای حفظ استقلال تصمیم‌گیری انسانی. مشارکت نظری مقاله در بهره‌گیری از رویکرد شبکه‌ای به عاملیت و مسئولیت است که امکان تحلیل دقیق‌تر «شکاف مسئولیتی» و بازتعریف نقش انسان و ماشین در تصمیم‌گیری‌های پیچیده را فراهم می‌آورد. نتایج نشان می‌دهد که خودمختاری انسانی در عصر هوش مصنوعی دچار دگرگونی ساختاری می‌شود؛ انسان همچنان مسئول است، اما بخش‌هایی از زنجیره تصمیم‌سازی—از جمع‌آوری داده تا پیشنهاد الگوریتمی—خارج از کنترل مستقیم او قرار دارند. به این ترتیب، چالش اصلی نه در توانایی فنی ماشین‌ها، بلکه در بازپیکربندی هنجاری رابطه انسان، الگوریتم و مسئولیت اخلاقی ظهور می‌کند و درک آن می‌تواند مسیر پژوهش‌های آتی در فلسفه ذهن، اخلاق فناوری و حکمرانی الگوریتمی را متحول سازد.

استناد: غلامی، طیبه و حسینی سروری، سید حسن. (۱۴۰۵). هوش مصنوعی و دگرگونی خودمختاری انسان: تحلیلی از اراده آزاد و مسئولیت اخلاقی در تصمیم‌سازی داده‌محور، *پژوهش‌های فلسفی*، ۲۰ (۵۵)، ۲۸۴-۲۶۵. https://doi.org/10.22034/jpiut.2026.72239.44950

ناشر: دانشگاه تبریز.

مقدمه

نفوذ روزافزون سامانه‌های هوش مصنوعی در فرآیندهای تصمیم‌گیری فردی و اجتماعی، پایه‌های مفهومی فلسفه اخلاق ــ به‌ویژه مفاهیم «اراده آزاد»، «خودمختاری» و «مسئولیت اخلاقی» ــ را با چالشی بی‌سابقه روبرو ساخته است. منظور از «اراده آزاد» در این نوشتار، توانایی عامل انسانی برای انتخاب میان مسیرهای بدیلِ دست‌کم به لحاظ ذهنی ممکن است؛ «خودمختاری» به ظرفیت تصمیم‌گیری مبتنی بر باورها و خواست‌های تأمل‌شدهٔ خود فرد اشاره دارد؛ و «مسئولیت اخلاقی» به قابلیت نسبت دادن ستایش یا نکوهش به عامل در قبال کنش‌هایش گفته می‌شود. هرچند که خود مقاله در ادامه نشان خواهد داد که این تعاریف سنتی در مواجهه با الگوریتم‌ها نیازمند بازتعریف هستند. امروزه الگوریتم‌ها دیگر ابزارهایی منفعل برای پردازش داده نیستند، بلکه به بازیگرانی فعال در شکل‌دهی به ترجیحات مصرفی، تشخیص‌های پزشکی و حتی قضاوت‌های حقوقی تبدیل شده‌اند. پژوهش‌های اخیر نشان می‌دهد که این سامانه‌ها با طراحی چارچوب انتخاب، برجسته‌سازی گزینه‌های خاص و حذف گزینه‌های دیگر، می‌توانند زمینه «توهم آزادی» را فراهم آورند؛ وضعیتی که در آن فرد احساس می‌کند آزادانه انتخاب می‌کند، حال آنکه افق انتخاب او از سوی منطق الگوریتمی محدود و جهت‌دهی شده است (والور، ۲۰۱۶؛ وانگ، ۲۰۲۴). این پدیده پرسشی بنیادین را مطرح می‌سازد که هسته اصلی مسئله این مقاله است: اگر تصمیمات انسانی به‌طور فزاینده‌ای تحت تأثیر پیش‌بینی و هدایت ماشین‌ها قرار گیرد، آیا می‌توان همچنان از امکان «کنش اخلاقی خودمختار» دفاع کرد؟ به عبارت دیگر، آیا اخلاق سنتی که بر مبنای انتخاب آگاهانه و مستقل بنا شده، در عصر الگوریتم‌ها همچنان اعتبار دارد؟

در سال‌های اخیر، پژوهش‌های گسترده‌ای در فضای آکادمیک ایران به بررسی ابعاد فلسفی و اخلاقی هوش مصنوعی اختصاص یافته است که نشان‌دهنده اهمیت یافتن این موضوع در حوزه فلسفه علم و اخلاق کاربردی است. در این میان، مسئله «عاملیت اخلاقی» و چگونگی انتساب مسئولیت به ماشین‌ها یکی از کانون‌های اصلی بحث بوده است. پژوهشگران در تلاش بوده‌اند با تکیه بر مبانی فلسفی سنتی و اسلامی، مشخص کنند که آیا ماشین‌های هوشمند می‌توانند فاعل اخلاقی مستقلی باشند یا خیر. برای مثال، عاشوری کیسمی و پرویزی (۱۴۰۳) در بررسی عاملیت اخلاقی هوش مصنوعی عام، به چالش‌های نظری نسبت‌دادن صفات اخلاقی به سیستم‌های هوشمند پرداخته‌اند و حسینی سورکی (۱۴۰۳) نیز با واکاوی معیارهای عاملیت اخلاقی ماشین‌ها، تلاش کرده است مرزهای فاعلیت انسانی و ماشینی را روشن سازد. در کنار این مباحث، تلاش برای همسوسازی فناوری با ارزش‌های دینی و ملی نیز محوریت داشته است؛ به‌گونه‌ای که قوامی‌پور و محمودی (۱۴۰۳) و همچنین میری بالاجورشری و همکاران (۱۴۰۳) با واکاوی سند اخلاق هوش مصنوعی یونسکو و مبانی اخلاق اسلامی، به دنبال تدوین چارچوب‌هایی بومی برای پاسخگویی به مسائل اخلاقی در عصر الگوریتم‌ها بوده‌اند. با این حال، اگرچه این پژوهش‌ها به خوبی برجسته کرده‌اند که هوش مصنوعی نمی‌تواند فاعل اخلاقی کامل به شمار رود، اما اغلب بر دوگانگی «انسان در برابر ماشین» تمرکز داشته‌اند و کمتر به ساختار «رابطه‌ای» و «شبکه‌ای» تصمیم‌سازی پرداخته‌اند.

در سطح بین‌المللی نیز ادبیات اخلاق هوش مصنوعی طی سال‌های اخیر رشد زیادی داشته است. تلاش‌های اولیه عمدتاً بر تدوین رهنمودها و چارچوب‌های کلی اخلاقی متمرکز بود؛ کاری که در آثار جامعی همچون *«دستنامه آکسفورد در اخلاق هوش مصنوعی»* (دوبر، ۲۰۲۰) و مقالات ارزیابی‌گر این دستورالعمل‌ها (هاگندورف، ۲۰۲۰) مشهود است. با گذر زمان، تمرکز پژوهش‌ها به سمت مسائل عمیق‌تر فلسفی حرکت کرد؛ از جمله بررسی ماهیت باورهای هوش مصنوعی (ما و والتون، ۲۰۲۴)، چالش‌های اعتماد و تفسیرپذیری (چویی، ۲۰۲۳) و پیوند میان اخلاق و معرفت‌شناسی ماشین‌ها (روسو). در این میان، مسئله «مسئولیت اخلاقی» و «شکاف مسئولیتی» (الستی، ۱۴۰۳) یکی از پرچالش‌ترین مباحث بوده است. پژوهشگرانی همچون تولون (۲۰۲۲) و

نیراپ (۲۰۲۳) کوشیده‌اند با بازنگری در مفاهیم مسئولیت آینده‌نگر[۱] و انسان‌محوری متواضعانه، راه‌حل‌هایی برای این شکاف بیابند. همچنین، نقش هوش مصنوعی به عنوان «توسعه‌دهنده ذهن» (ولد و هرناندز، ۲۰۲۲) یا واسطه در روابط انسانی (باتیستی، ۲۰۲۵) نشان می‌دهد که این فناوری نه تنها ابزار عملی، بلکه عاملی است که ماهیت تجربه انسانی را تغییر می‌دهد. با این وجود، بررسی انتقادی این ادبیات نشان می‌دهد که بخشی از رویکردهای موجود نیز همچنان در چارچوب‌های فردگرایانه یا دوگانه‌انگارانه باقی مانده‌اند؛ به این معنا که یا به دنبال نسبت دادن مسئولیت به ماشین به عنوان یک موجود مستقل هستند، یا سعی دارند با حفظ کامل اختیار برای انسان، نقش ماشین را نادیده بگیرند. آنچه در این ادبیات کمتر مورد توجه دقیق قرار گرفته، تحلیل دقیق «ساختار شبکه‌ای» تصمیم‌گیری است؛ ساختاری که در آن انسان و ماشین نه به عنوان دو موجود جداگانه، بلکه به عنوان عناصری درگیر در یک فرآیند ترکیبی عمل می‌کنند.

نوآوری اصلی این مقاله در بازتعریف مفهوم خودمختاری و مسئولیت اخلاقی در قالب یک چارچوب «شبکه‌ای-هنجاری» است. برخلاف رویکردهای موجود که یا بر عاملیت فردی انسان تأکید دارند یا در پی نسبت‌دادن نوعی عاملیت مستقل به ماشین‌ها هستند، این پژوهش با توسعه تحلیلی نظریه شبکه کنشگر، مدلی از «مسئولیت توزیع‌شده» ارائه می‌دهد که در آن، عاملیت نه به‌عنوان ویژگی ذاتی یک فاعل، بلکه به‌عنوان برساخته‌ای رابطه‌ای در بستر تعامل انسان، الگوریتم و ساختارهای نهادی فهم می‌شود. این چارچوب نشان می‌دهد که مسئله مسئولیت در عصر هوش مصنوعی، بیش از آنکه به لحظه تصمیم‌گیری نهایی مربوط باشد، به سطح طراحی شرایط امکان انتخاب و معماری تصمیم‌سازی بازمی‌گردد. به‌طور مشخص، تمایز این پژوهش با کاربردهای پیشین نظریه شبکه کنشگر در اخلاق فناوری در آن است که این مقاله، به‌جای توصیف توزیع عاملیت، معیاری هنجاری برای بازتخصیص مسئولیت در سطوح طراحی، تنظیم‌گری و کنش ارائه می‌دهد و بدین‌ترتیب شکاف میان تحلیل توصیفی و تجویز اخلاقی را پر می‌کند.

این مسئله زمانی پیچیده‌تر می‌شود که هوش مصنوعی از مرز پیش‌بینی رفتار عبور کرده و وارد حوزه «عاملیت» و تصمیم‌سازی شود. سامانه‌های هوشمند امروزی قادرند بر پایه تحلیل کلان‌داده‌ها، تصمیماتی اتخاذ کنند که پیش از این در حیطه انحصاری قضاوت انسان بوده و درجه‌ای از استقلال عملیاتی از خود نشان دهند (چسترمن، ۲۰۲۰). این استقلال فنی ماشین، یک پارادوکس اخلاقی ایجاد می‌کند: از یک سو، ماشین‌ها فاقد «قصد اخلاقی» و «آگاهی» هستند و نمی‌توان آن‌ها را فاعل اخلاقی مستقلی دانست؛ و از سوی دیگر، تأثیر شگرف آن‌ها بر فرآیند تصمیم‌گیری انسان—از طریق محدودسازی گزینه‌ها و ایجاد اتکای شناختی—مانع از آن می‌شود که مسئولیت کامل را بر عهده انسان بگذاریم. چسترمن (۲۰۲۰) این وضعیت را «شکاف مسئولیتی»[۲] می‌نامد؛ شکافی که در آن نه ماشین پاسخگو است و نه انسان را می‌توان کاملاً مسئول دانست. او این چالش‌ها را در سه دسته عملی، اخلاقی و مشروعیتی[۳] طبقه‌بندی می‌کند و نشان می‌دهد که بحران اصلی، فقدان چارچوبی برای توزیع مسئولیت در سیستم‌های ترکیبی انسان-ماشین است.

با این حال، رویکرد صرفاً حقوقی یا تقسیم مسئولیت به صورت دوتایی (انسان یا ماشین)، برای حل این معضل کافی نیست. آنچه در ادبیات موجود کمتر مورد توجه قرار گرفته، نحوه بازپیکربندی خودمختاری انسان در تعامل با الگوریتم‌هاست. این مقاله استدلال می‌کند که هوش مصنوعی لزوماً خودمختاری را حذف نمی‌کند، بلکه آن را به مثابه ویژگی‌ای شبکه‌ای و وابسته به محیط بازتعریف می‌کند؛ بازتعریفی که در آن افق انتخاب نه تنها توسط باورها و خواست‌های فرد، بلکه به طور

---

[۱] forward-looking responsibility
[۲] Responsibility Gap
[۳] legitimacy

فزاینده‌ای توسط معماری الگوریتمی شکل می‌گیرد. بر این اساس، مسئله اصلی این پژوهش، تبیین چگونگی حفظ استقلال انسانی در شرایطی است که بخشی از فرآیند تصمیم‌سازی به ماشین واگذار شده است.

در راستای پاسخ به این پرسش، مقاله حاضر با رویکردی تحلیلی–انتقادی، سه محور را به ترتیب منطقی دنبال می‌کند:

۱. نسبت میان پیش‌بینی الگوریتمی و شرایط امکان انتخاب انسان را بررسی می‌کند تا نشان دهد چگونه شخصی‌سازی داده‌محور، ساختار روان‌شناختی انتخاب را بازآرایی می‌کند.

۲. با تحلیل مفهوم شکاف مسئولیتی، محدودیت‌های نظریه‌های مسئولیت اخلاقی در محیط‌های هدایت‌شده الگوریتمی آشکار ساخته می‌شود.

۳. سرانجام، با بهره‌گیری از رویکرد «عاملیت ترکیبی»، چارچوبی بدیل برای توزیع مسئولیت پیشنهاد می‌گردد؛ چارچوبی که نشان می‌دهد انسان می‌تواند با طراحی مسئولانه و نظارت بر الگوریتم‌ها، استقلال هنجاری خود را حفظ کند.

نتیجه نهایی مقاله آن است که خودمختاری در عصر هوش مصنوعی دچار دگرگونی ساختاری می‌شود و پاسخگویی اخلاقی نیازمند گذار از «مسئولیت فردی» به «مسئولیت شبکه‌ای[۱]» است.

۱. پیش‌بینی الگوریتمی و بازتولید ساختاری خودمختاری

پیش‌بینی رفتار انسان، کانونی‌ترین سازوکاری است که از طریق آن سامانه‌های الگوریتمی بر فرآیند تصمیم‌گیری فردی اثر می‌گذارند. الگوریتم‌های معاصر—به‌ویژه در شبکه‌های اجتماعی، پلتفرم‌های تجارت الکترونیک و سامانه‌های حمل‌ونقل هوشمند—دیگر صرفاً نقش «ابزار پیشنهاددهنده» را ایفا نمی‌کنند؛ بلکه با تحلیل کلان‌داده‌ها و الگوهای رفتاری، به‌صورت فعال مسیرهای تصمیم‌گیری را جهت‌دهی کرده و در شکل‌گیری گزینه‌هایی که فرد با آن‌ها مواجه می‌شود، مداخله مستقیم دارند (وانگ، ۲۰۲۴، ۷–۵). این مداخله زمانی اهمیت می‌یابد که توجه کنیم ماهیت شخصی‌سازی‌شده این تصمیمات، موجب می‌شود افراد غالباً بدون آگاهی در میان گزینه‌هایی حرکت کنند که پیش‌تر توسط الگوریتم پالایش، فیلتر و اولویت‌بندی شده‌اند (لو، ۲۰۲۲، ۶–۳). در نتیجه، فرد نه با مجموعه‌ای بی‌واسطه از بدیل‌های ممکن، بلکه با «نسخه‌ای الگوریتمی‌شده» از انتخاب‌ها روبه‌رو می‌شود؛ نسخه‌ای که خود محصول مفروضات، اهداف و معیارهای پنهان طراحان سامانه است.

این محدودسازی گزینه‌ها، تنها زمانی به یک چالش اخلاقی بدل می‌شود که با گرایش‌های روان‌شناختی انسان ترکیب شود. شواهد پژوهشی نشان می‌دهند که حتی زمانی که کاربران آگاه‌اند توصیه‌ای توسط الگوریتم تولید شده، همچنان احتمال بیشتری دارد آن را بپذیرند. این پدیده که «سوگیری اتوماسیون»[۲] نامیده می‌شود، وضعیتی است که در آن انسان‌ها تمایل دارند به خروجی سیستم‌های خودکار اعتماد کنند، حتی زمانی که شواهدی از خطا وجود دارد (کامینگز، ۲۰۱۲، ۳–۲). نکته حائز اهمیت آن است که آنچه در ظاهر به‌عنوان اعتماد به یک ابزار تکنیکی جلوه می‌کند، در سطحی عمیق‌تر به مسئله‌ای فلسفی بدل می‌شود: تقابل میان خودمختاری انسانی و سازوکارهای نامرئی‌ای که شرایط امکان انتخاب را از پیش صورت‌بندی می‌کنند. در چنین موقعیت‌هایی، فرد—حتی با برخورداری از دانش و توان قضاوت مستقل—به‌تدریج داوری سامانه را بر داوری خویش مقدم می‌دارد. این اعتماد نه به‌واسطه اجبار بیرونی، بلکه در قالب نوعی گرایش شناختی شکل می‌گیرد که آزادی انتخاب را به سطحی

---

[۱] منظور از «مسئولیت شبکه‌ای» در اینجا نه نفی مسئولیت فردی، بلکه صورتبندی از مسئولیت است که در درون شبکه‌ای از کنشگران انسانی و غیرانسانی (طراحان، کاربران، نهادها، الگوریتم‌ها و داده‌ها) توزیع می‌شود، بدون آنکه انتساب نهایی مسئولیت به عامل انسانی از میان برود. این مفهوم برگرفته از خوانش انتقادی نظریه شبکه‌کنشگر (ANT) در اخلاق هوش مصنوعی است.

[۲] Automation Bias

از «توهم انتخاب» تقلیل می‌دهد. این پدیده از آن جهت اهمیت دارد که اعتماد شناختی به سامانه‌های خودکار، فرآیند ارزیابی انتقادی گزینه‌ها را تضعیف کرده و موجب می‌شود فرد بدون بازاندیشی کافی، پیشنهاد الگوریتمی را به‌عنوان گزینه بهینه بپذیرد؛ امری که مستقیماً ظرفیت خودمختاری هنجاری را محدود می‌سازد.

مسئله اصلی نه کاهش مطلق اختیار، بلکه جابه‌جایی محل اعمال آن است؛ اختیار از سطح انتخاب آگاهانه و لحظه‌ای فاعل، به سطح طراحی پیشینی و غیرشفاف منتقل می‌شود. این جابه‌جایی پیامدی مهم‌تر از «کاهش تعداد گزینه‌ها» دارد: آنچه دستخوش تغییر می‌شود، شرایط هنجاری امکان کنش خودمختار است. در تلقی‌های کلاسیک از خودمختاری — از کانت (۱۹۷۲، ۵۵) تا خوانش‌های معاصر — یکی از شروط ضروری خودمختاری آن است که فاعل بتواند دلایل کنش خود را بازاندیشی کرده و نسبت به آن‌ها موضع انتقادی بگیرد. بدیهی است که این شرط به تنهایی برای تحقق خودمختاری کافی نیست و شروط دیگری همچون استقلال از اجبار بیرونی و خودقانون‌گذاری نیز لازم است، اما استدلال این مقاله آن است که محیط‌های پیش‌بینی‌شده الگوریتمی حتی همین شرط اول را نیز تضعیف می‌کنند؛ زیرا دلایل کنش پیش از آنکه موضوع تأمل فاعل قرار گیرند، در سطحی پیشینی و نامرئی صورت‌بندی می‌شوند.

استدلال اصلی این بخش را می‌توان به‌صورت فشرده چنین بازسازی کرد:

۱) خودمختاری اخلاقی مستلزم امکان بازاندیشی انتقادی نسبت به دلایل کنش است.

۲) در محیط‌های الگوریتمی، دلایل کنش در سطحی پیشینی و از طریق سازوکارهای شخصی‌سازی و پیش‌بینی صورت‌بندی می‌شوند.

۳) این صورت‌بندی پیشینی، امکان دسترسی و فاصله‌گیری انتقادی فاعل از دلایل کنش را محدود می‌سازد.

بنابراین،

(۴) خودمختاری اخلاقی در این محیط‌ها دچار تضعیف ساختاری می‌شود، حتی اگر تجربه پدیداری اختیار حفظ شده باشد.

این وضعیت را می‌توان نوعی «خودمختاری تضعیف‌شده ساختاری»نامید؛ وضعیتی که در آن عاملیت انسانی به‌صورت ظاهری حفظ می‌شود، اما ظرفیت هنجاری آن برای پاسخ‌گویی اخلاقی به‌تدریج فرسوده می‌گردد. بنابراین، قدرت فناوری نه در حذف مستقیم آزادی، بلکه در جهت‌دهی تدریجی و نامحسوس آن نهفته است. الگوریتم‌ها چارچوبی می‌سازند که درون آن تصمیم‌گیری معنا پیدا می‌کند؛ چارچوبی که در آن فرد همچنان خود را مختار می‌پندارد، در حالی که افق انتخاب‌های او از پیش مهندسی شده است. این تحلیل نشان می‌دهد که اگر اختیار انسان در نتیجه پیش‌بینی الگوریتمی به‌صورت نامرئی بازآرایی شود، دفاع از اراده آزاد و مسئولیت اخلاقی نیازمند بازتعریف مفاهیم سنتی است؛ موضوعی که در بخش بعدی با بررسی مفهوم «شکاف مسئولیتی» پیگیری خواهد شد.

۲. انتقال عاملیت و چالش مسئولیت اخلاقی

یکی از پیچیده‌ترین پیامدهای نفوذ هوش مصنوعی، بروز چالش در تبیین مسئولیت اخلاقی تصمیماتی است که تحت هدایت الگوریتم‌ها شکل می‌گیرند. در حوزه‌هایی مانند پزشکی یا حمل‌ونقل هوشمند، سامانه‌های هوشمند پیشنهادهایی ارائه می‌دهند که مستقیماً بر تصمیم نهایی انسان تأثیر می‌گذارند؛ از توصیه یک مدل تشخیصی به پزشک تا فعال‌سازی ترمز اضطراری در خودروهای خودران. در چنین شرایطی، پرسش اساسی این است که مسئولیت پیامدهای این تصمیمات بر عهده کیست؟ آیا انسانِ تصمیم‌گیرنده مسئول است، طراح سیستم، یا خود سامانه؟ این پرسش ما را با مفهومی شناخته‌شده به نام «شکاف مسئولیتی» روبرو می‌سازد؛ وضعیتی که در آن بین توانایی پیش‌بینی پیامدهای یک تصمیم و امکان انتساب آن به یک فاعل اخلاقی مشخص،

فاصله‌ای معنادار ایجاد می‌شود (آلبارادا، ۲۰۲۵). این شکاف زمانی تشدید می‌شود که سامانه‌ها پیچیده‌تر از آن باشند که نظارت کامل انسان بر آن‌ها ممکن باشد، اما همزمان فاقد استقلال کافی برای احراز عنوان «فاعل اخلاقی مستقل» باشند. در نتیجه، مسئولیت انسانی در فرآیند تصمیم‌گیری در فضایی میان‌تهی معلق می‌ماند و ماهیت تصمیمات اتخاذ شده مبهم و چندلایه می‌گردد.

برای درک عمیق‌تر این شکاف، باید به ماهیت وجودی سامانه‌های هوش مصنوعی توجه کرد. این سامانه‌ها موجوداتی با ویژگی‌های متناقض‌نما هستند: آن‌ها فاقد نیت، خودآگاهی یا همدلی انسانی‌اند، اما در عین حال قادرند پیامدهای اخلاقی معناداری ایجاد کنند. از این رو، آن‌ها نه ابزارهای صرف‌اند و نه عاملان اخلاقی کامل، بلکه در ناحیه‌ای میانی قرار می‌گیرند که رفتارشان می‌تواند با هنجارهای اخلاقی سازگار باشد، بدون آنکه بتوان مسئولیت اخلاقی کامل را به آن‌ها نسبت داد (گوگوشین، ۲۰۲۱، ۴۶). در این چارچوب، می‌توان از نوعی «عاملیت اخلاقی محدود» سخن گفت؛ عاملیتی که صرفاً در سطح عملکردی معنا دارد، اما فاقد قصد درونی و خودفهمی اخلاقی است. برای اثبات دقیق این محدودیت و روشن‌تر شدن ابعاد شکاف مسئولیتی، می‌توان از چارچوب تحلیلی کانستانتینسکو و همکاران (۲۰۲۲) بهره جست.

کانستانتینسکو و همکاران با ارائه «آزمون مسئولیت اخلاقی» مبتنی بر چهار شرط ارسطویی—علّیت، آزادی، دانش و تعمق—ابزاری برای سنجش دقیق عاملیت اخلاقی فراهم می‌کنند (کانستانتینسکو و همکاران، ۲۰۲۲، ۲۰) منطق این آزمون آن است که تنها موجوداتی قادر به برآوردن هم‌زمان همه این شرایط هستند که می‌توانند به‌عنوان فاعل اخلاقی کامل و موضوع نسبت‌دادن مسئولیت تلقی شوند. اعمال این آزمون بر «مشاوران اخلاقی مصنوعی»[1] (AMAs) نشان می‌دهد که چرا شکاف مسئولیتی پدید می‌آید: این سامانه‌ها در هیچ‌یک از شرایط چهارگانه موفق نمی‌شوند. نخست، آن‌ها شرط علّیت را برآورده نمی‌کنند، زیرا منشأ اولیه کنش و هدف‌گذاری همواره بیرون از سامانه و در اختیار کاربران یا طراحان است. دوم، آزادی آن‌ها به‌طور ساختاری محدود است؛ مشاوران اخلاقی مصنوعی در چارچوب معماری‌های ازپیش‌تعیین‌شده عمل می‌کنند و نمی‌توانند به معنای ارسطویی «خودآغازگر» باشند. افزون بر این، هرچند این سامانه‌ها به حجم عظیمی از اطلاعات دسترسی دارند، فاقد آگاهی زمینه‌ای و تجربه زیسته‌ای هستند که شرط معرفتی مسئولیت اخلاقی را محقق می‌کند. مهم‌تر از همه، AMA ها از «حکمت عملی»[2] محروم‌اند؛ یعنی نمی‌توانند معنای اخلاقی یک کنش را در نسبت با شرایط خاص و تعارض ارزش‌ها به‌درستی بسنجند (کانستانتینسکو و همکاران، ۲۰۲۱، ۸۰۹).

نتیجه‌گیری منطقی این تحلیل آن است که اگرچه AMAها واجد نوعی عاملیت عملکردی محدود هستند، اما مسئولیت اخلاقی پیامدها به‌صورت غیرقابل‌انتقالی بر عهده انسان باقی می‌ماند. این وضعیت پارادوکس مرکزی چالش مسئولیت در عصر هوش مصنوعی را آشکار می‌سازد: سامانه‌ها هرچه بیشتر در فرایند تصمیم‌گیری مشارکت داده می‌شوند، سهم انسان در کنترل مستقیم کمتر می‌شود، در حالی که بار مسئولیت اخلاقی همچنان تماماً بر دوش او باقی می‌ماند. این تحلیل نشان می‌دهد که هوش مصنوعی صرفاً دامنه تصمیم‌گیری انسانی را محدود نمی‌کند، بلکه ساختار آن را در محیطی چندلایه و داده‌محور بازتعریف می‌کند. بنابراین، مسئله اخلاقی دیگر قابل فروکاهی به انتخاب فردی نیست؛ بلکه راه‌حل باید در چگونگی توزیع مسئولیت، تضمین شفافیت الگوریتمی و طراحی مسئولانه سامانه‌ها جستجو شود؛ موضوعی که بخش پایانی مقاله به آن اختصاص خواهد داشت. با این حال، پرسش مهم این است که آیا چارچوب‌های نظری موجود واقعاً قادرند این گذار از مسئولیت فردی به مسئولیت

---

[1] Artificial Moral Advisors (AMAs)

[2] Phronesis

توزیع‌شده را به‌طور مفهومی تبیین کنند یا خیر؟با این حال، بخش قابل‌توجهی از ادبیات موجود—از جمله تحلیل‌های چسترمن و گوگوشین —عمدتاً در چارچوبی باقی می‌مانند که مسئله مسئولیت را یا به سطح انتساب به یک فاعل مشخص تقلیل می‌دهد، یا در بهترین حالت به توزیع آن میان چند عامل انسانی بسنده می‌کند. محدودیت مشترک این رویکردها آن است که نقش «معماری تصمیم‌سازی» را به‌عنوان سطحی پیشینی که امکان‌های کنش را شکل می‌دهد، به‌طور نظام‌مند در تحلیل وارد نمی‌کنند. در نتیجه، اگرچه این دیدگاه‌ها به‌درستی بر وجود شکاف مسئولیتی تأکید دارند، اما در تبیین منشأ ساختاری آن ناکام می‌مانند.

۳. نظریه شبکه کنشگر و بازتعریف شبکه‌ای عاملیت

در چنین بن‌بستی، نیاز به چارچوبی احساس می‌شود که بتواند از سطح تحلیل فاعل منفرد فراتر رفته و ساختار رابطه‌ای تصمیم‌سازی را تبیین کند. برای فهم عمیق‌تر پدیده «شکاف مسئولیتی» و تبیین امکان توزیع مسئولیت در سیستم‌های مبتنی بر هوش مصنوعی، چارچوب‌های کلاسیک اخلاق که بر فاعل انسانی منفرد تمرکز دارند، دیگر کفایت نمی‌کنند. در اینجا، نظریه شبکه کنشگر[۱]امکان گشایش افق تحلیلی تازه‌ای را فراهم می‌سازد؛ افقی که با عبور از دوگانه سنتی «انسان در برابر ماشین»، عاملیت و مسئولیت (رک. خزاعی، ۱۳۹۸) را به‌عنوان برساخته‌هایی شبکه‌ای و رابطه‌ای بازتعریف می‌کند. این رویکرد ریشه در آثار برونو لاتور[۲](۲۰۰۵) و میشل کالون[۳](۱۹۸۶) دارد که با نقد دوگانه‌انگاری انسان/ فناوری، عاملیت را به‌عنوان پدیده‌ای برآمده از شبکه روابط بازتعریف می‌کنند. این مبنا امکان تحلیل شبکه‌ای مسئولیت در سامانه‌های هوش مصنوعی را فراهم می‌سازد (رک. محمدی و مقدم حیدری، ۱۳۹۱).

براساس نظریهٔ شبکه کنشگر (ANT)، کنش اجتماعی نه حاصل اراده یا قصد یک فاعل منفرد، بلکه نتیجه تعامل پویای شبکه‌ای از کنشگران انسانی و غیرانسانی است. در این چارچوب، فناوری، الگوریتم، داده، قانون، نهادهای سیاسی، کاربران انسانی و شرکت‌های فناورانه همگی به‌عنوان «کنشگر»[۴]در نظر گرفته می‌شوند. این کنشگران از طریق فرآیندی که در نظریهٔ شبکه کنشگر «ترجمه»[۵]نامیده می‌شود، منافع، نقش‌ها و هویت‌های یکدیگر را بازتعریف و هم‌راستا می‌کنند تا کنش جمعی امکان‌پذیر شود. ترجمه در این معنا به فرآیندی اشاره دارد که طی آن هیچ کنشگری به‌تنهایی مسلط نیست، بلکه قدرت و عاملیت در دل شبکه و از خلال روابط آن تولید می‌شود (وانگ، ۲۰۲۴، ۱۱-۵).

پیامد نظری مهم این رویکرد آن است که عاملیت دیگر ویژگی ذاتی یک فاعل انسانی مستقل تلقی نمی‌شود، بلکه پدیده‌ای است که از دل روابط شبکه‌ای میان عناصر انسانی و فناورانه سربرمی‌آورد. در چنین برداشتی، الگوریتم‌ها و سامانه‌های هوش مصنوعی نه صرفاً ابزارهایی خنثی، بلکه عناصری فعال در شکل‌دهی به امکان‌ها و محدودیت‌های کنش انسانی‌اند. البته این امر به معنای نسبت دادن قصد اخلاقی یا آگاهی هنجاری به ماشین‌ها نیست، بلکه تأکید بر این نکته است که نقش آن‌ها در تولید پیامدهای اخلاقی را نمی‌توان نادیده گرفت. از این منظر، مسئولیت اخلاقی نیز ماهیتی توزیع‌شده، رابطه‌ای و چندسطحی پیدا می‌کند. دیگر نمی‌توان مسئولیت را به‌طور کامل به یک تصمیم‌گیر انسانی یا یک سامانه الگوریتمی نسبت داد؛ بلکه مسئولیت

---

[۱] Actor-Network Theory – ANT

[۲] Bruno Latour. برخی از آثار برونو لاتور به فارسی ترجمه شده است. از جمله: «پرس و جو درباره شیوه‌های هستی» (ترجمه رویا منجم، تهران، نشر علم، ۱۳۹۴)؛ و«مذاکره با اشیاء» (ترجمه رحمان شریف‌زاده، تهران، نشر نی، ۱۴۰۳).

[۳] Michel Callon

[۴] actant

[۵] translation

در شبکه‌ای از کنشگران توزیع می‌شود که هر یک سهمی در شکل‌گیری تصمیم نهایی دارند. این تحلیل، بنیان‌های فردگرایانه نظریه‌های سنتی مسئولیت را به چالش می‌کشد و ضرورت بازاندیشی در مدل‌های حکمرانی، پاسخ‌گویی و تنظیم‌گری را برجسته می‌سازد (وانگ، ۲۰۲۴، ۱۱-۵).

بااین‌حال، باید توجه داشت که نظریه شبکه کنشگر نیز محدودیت‌های خود را دارد؛ به‌ویژه در محیط‌های واقعی شهری، تعادل قدرت میان بازیگران انسانی و غیرانسانی ممکن است نابرابر باشد و برخی سامانه‌ها یا داده‌ها می‌توانند نقش نامتناسبی در شبکه ایفا کنند. بنابراین، تحلیل شبکه‌ای برای جلوگیری از تمرکز قدرت در دست فناوری، نیازمند بررسی دقیق قدرت، دسترسی و تأثیر واقعی هر عامل در عمل است.

کاربرد این نظریه در زمینه شهرهای هوشمند و مشاوران اخلاقی مصنوعی (AMA) هاچشم‌انداز روشن‌تری ارائه می‌دهد. در این شبکه‌ها، بازیگرانی همچون دولت‌ها، شرکت‌های توسعه فناوری، ساکنان شهری، داده‌ها و سامانه‌های هوشمند در تعامل‌اند. AMAها، هرچند فاقد نیت، خودآگاهی یا حکمت عملی کامل‌اند، اما به‌عنوان عاملان غیرانسانی با عاملیت محدود، می‌توانند تصمیمات انسانی را تسهیل، گزینه‌ها را تحلیل و پیامدهای عملی را شفاف‌سازی کنند. در این سطح، AMAها «ابزارهای شبکه‌ای با ظرفیت عملکردی محدود» به شمار می‌روند که مشارکت آن‌ها در شبکه، بر توزیع مسئولیت و شکل‌گیری نتایج اخلاقی تأثیر می‌گذارد، بدون آنکه مسئولیت اخلاقی کامل به آن‌ها قابل نسبت باشد. حتی آینده‌پژوهی نظریه شبکه کنشگرنشان می‌دهد که حتی اگر روزی هوش مصنوعی به سطحی از خودآگاهی برسد، عاملیت اخلاقی کامل آن بدون تعامل با شبکه انسانی و توانایی حکمت عملی مستقل تحقق نخواهد یافت.

در ادبیات معاصر اخلاق هوش مصنوعی، این دیدگاه که مسئولیت باید به‌صورت شبکه‌ای میان کنشگران مختلف سازمان‌دهی شود، مورد تأیید قرار گرفته است (گوگوشین، ۲۰۲۵، ۱۱۷۲-۱۱۷۰). چنین رویکردی امکان آن را فراهم می‌سازد که بدون فروکاستن عاملیت انسانی، واقعیت‌های فناورانه عصر الگوریتم‌ها به رسمیت شناخته شوند. بر این اساس، نظریه شبکه کنشگر راهبردهای عملی زیر را برای کاهش ابهام مسئولیتی ارائه می‌دهد:

۱. توزیع مسئولیت: پذیرش اینکه AMAها به‌عنوان عاملان ابزارمند، مسئولیت را در شبکه تسهیل می‌کنند، اما مسئولیت نهایی و هدف‌گذاری همواره بر عهده انسان باقی می‌ماند.

۲. شفافیت الگوریتمی و توضیح‌پذیری: شفاف کردن فرآیند تصمیم‌گیری هوش مصنوعی برای سایر کنشگران شبکه تا امکان نظارت و فاصله‌گیری انتقادی فراهم شود.

۳. طراحی مسئولانه و اخلاقی: تدوین دستورالعمل‌ها و قوانین اخلاقی برای توسعه شهرهای هوشمند و سیستم‌های هوش مصنوعی، با هدف ارتقای همکاری متوازن بین عاملان انسانی و غیرانسانی.

برای روشن‌تر شدن این بازتعریف شبکه‌ای، می‌توان به یک مثال عملی در مدیریت حمل‌ونقل هوشمند اشاره کرد. تصور کنید سامانه‌ای AMA در مدیریت ترافیک، پیشنهاد مسیریابی اضطراری برای خودروهای اورژانسی ارائه می‌دهد. در این سناریو، تصمیم نهایی توسط اپراتور انسانی اعمال می‌شود، اما AMA نقش تحلیلگر و تسهیل‌کننده اطلاعات را ایفا می‌کند. در اینجا، مسئولیت نهایی توزیع‌شده باقی می‌ماند؛ AMA به دلیل تحلیل دقیق‌تر داده‌ها سهمی در موفقیت عمل دارد، اما اپراتور و طراحان سیستم به دلیل اعمال کنترل نهایی و تعیین اهداف، مسئولیت هنجاری را بر عهده دارند. این مثال نشان می‌دهد که حتی در سناریوهای پیچیده، می‌توان بدون نسبت دادن مسئولیت کامل به ماشین، نقش فعال آن را در شبکه پذیرفت. در نهایت، این تحلیل نشان می‌دهد که تعامل انسان و هوش مصنوعی نه تنها محدودیت‌های تصمیم‌گیری انسانی را مشخص می‌کند،

بلکه با بازتعریف عاملیت و مسئولیت، فرصت‌هایی برای حکمرانی اخلاقی و توسعه پایدار در محیط‌های پیچیده و داده‌محور فراهم می‌آورد.

۴. چالش‌های اخلاقی خودمختاری و توزیع مسئولیت

بر اساس تحلیل چسترمن، چالش بنیادین هوش مصنوعی نه در سلب مستقیم «قدرت تصمیم‌گیری» از انسان، بلکه در بازتعریف پنهان مبنای اخلاقی اختیار انسانی نهفته است. در شرایطی که تصمیمات انسانی حاصل تعامل با سامانه‌هایی هستند که گزینه‌ها را به‌صورت نامرئی اولویت‌بندی، محدود یا جهت‌دهی می‌کنند، پرسش از امکان کنش اخلاقی به معنای کلاسیک آن—به‌ویژه در چارچوب اخلاق کانتی—دوباره مطرح می‌شود. در اخلاق کانت، کنش اخلاقی مستلزم عمل «از سر عقل» و با «آگاهی از امکان بدیل» است؛ شرطی که تضمین می‌کند فاعل نه صرفاً واکنش‌گر، بلکه خودآیین و مسئول باشد (چسترمن، ۲۰۲۰، ۲۳۰–۲۳۳).

در این چارچوب، می‌توان استدلال کرد که آنچه در محیط‌های الگوریتمی تضعیف می‌شود، نه صرفاً آزادی انتخاب به معنای تجربی، بلکه شرط کانتی «خودآیینی عقل عملی» است؛ زیرا اگر فاعل نتواند منشأ دلایل کنش خود را به‌صورت مستقل مورد ارزیابی قرار دهد، نسبت او با قانون اخلاقی از نسبت «تقنین‌گر» به نسبت «پیرو» تنزل می‌یابد. بدین‌ترتیب، چالش هوش مصنوعی را باید نه در سطح محدودسازی گزینه‌ها، بلکه در سطح تضعیف ظرفیت تقنین عقلانی فاعل تحلیل کرد.

مطالعات تجربی نشان می‌دهند که این نگرانی صرفاً نظری نیست. الگوریتم‌ها از طریق ایجاد آنچه «توهم خودمختاری»[۱] نامیده می‌شود، بستری فراهم می‌کنند که در آن کاربران احساس انتخاب آزادانه دارند، در حالی که دامنه گزینه‌های پیش‌روی آنان از پیش در چارچوبی الگوریتمی و ارزش‌بار[۲]شکل گرفته است (لو، ۲۰۲۲، ۱۸–۸). در چنین شرایطی، شرط «آگاهی از امکان بدیل» به‌طور ساختاری تضعیف می‌شود. این وضعیت نشان می‌دهد که خودمختاری انسانی را نمی‌توان مفهومی یک‌لایه تلقی کرد. در تعامل با سامانه‌های هوش مصنوعی، دست‌کم می‌توان میان سه سطح از خودمختاری تمایز گذاشت: خودمختاری صوری (وجود گزینه‌ها)، خودمختاری شناختی (فهم دامنه و پیامدها)، و خودمختاری هنجاری (ارزیابی اخلاقی بدیل‌ها). الگوریتم‌ها ممکن است سطح نخست را حفظ کنند، اما از طریق طراحی رابط کاربری و وزن‌دهی ارزش‌بار، سطح دوم و سوم را تضعیف می‌کنند. بنابراین، «توهم خودمختاری» نه یک خطای روان‌شناختی فردی، بلکه پیامد ساختاری معماری تصمیم‌سازی الگوریتمی است که اختیار اخلاقی را تهی می‌سازد، بدون آنکه فاعل انسانی آگاهانه از آن محروم شده باشد.

این تحلیل صرفاً در سطح نظری باقی نمی‌ماند، بلکه در موارد عینی نیز پیامدهای قابل‌توجهی دارد. این تضعیف ساختاری خودمختاری، چالش‌های جدی را در نهادهای عمومی و حقوقی نیز ایجاد کرده است. برای مثال، در برنامه ربو–دِبت[۳] در استرالیا، الگوریتم محاسبه بدهی‌ها برداشت ناقصی از قوانین داشت و باعث خطاهای گسترده شد. یا در قانون مالیات انگلستان، دادگاه تأکید کرد که اطلاعیه پرداخت مالیات باید توسط مأمور انسانی صادر شود و تصمیم کامپیوتری کافی نیست (چسترمن، ۲۰۲۰، ۲۳۲). این موارد نشان می‌دهند که حتی در سطح قانونی، تصمیم‌گیری صرفاً الگوریتمی فاقد مشروعیت است و حضور انسان برای تضمین مشروعیت فرآیند ضروری به نظر می‌رسد.

---

[۱] illusion of autonomy
[۲] value-laden
[۳] Robo-debt

تحلیل نظام‌مند این وضعیت نشان می‌دهد که تلاش برای یافتن یک «فاعل منفرد» مسئول در این سیستم‌های پیچیده، ما را به بن‌بست نظری می‌کشاند. به جای آن، مسئله باید در چارچوب‌های جدیدی بازتعریف شود. نخست، می‌توان سیستم‌های هوش مصنوعی را به‌عنوان نوعی «نهاد جمعی» در نظر گرفت که تصمیمات آن‌ها حاصل کنش یکپارچه شبکه‌ای است. در این صورت، مسئله مسئولیت به این پرسش تبدیل می‌شود که چگونه می‌توان چنین نهاد جمعی‌ای را پاسخگو کرد. دوم، اگر مسئولیت همچنان در سطح افراد جستجو شود، با پدیده «پراکندگی مسئولیت» مواجه می‌شویم؛ وضعیتی که تعیین فرد مسئول را عملاً ناممکن می‌سازد.

این بن‌بست تحلیلی، دقیقاً جایی است که نظریه شبکه کنشگر (ANT) راهکار ارائه می‌دهد. همان‌طور که در بخش پیشین بحث شد، ANTبا عبور از الگوی فردمحور، توجه را به شبکه‌ای از کنشگران انسانی و غیرانسانی معطوف می‌کند. از این منظر، آنچه به‌عنوان «شکاف مسئولیتی» توصیف می‌شود، نه فقدان مسئول، بلکه ناتوانی چارچوب‌های سنتی در درک عاملیت شبکه‌ای است. نکته اساسی—چنان‌که آلباردا (۲۰۲۵) تأکید می‌کند—این است که شکاف مسئولیت پیامد اجتناب‌ناپذیر ماهیت فنی هوش مصنوعی نیست، بلکه حاصل نحوه سازمان‌دهی انسانی پیرامون این فناوری و چارچوب‌های مفهومی ما برای فهم عاملیت است. به بیان دیگر، شکاف مسئولیت نه یک سرنوشت فناورانه، بلکه نتیجه انتخاب‌های نهادی، طراحی و هنجاری است که ما می‌توانیم تغییر دهیم.

برای خروج از این بن‌بست، رویکردهای جدید اخلاق هوش مصنوعی بر سه مطالبه کلیدی تأکید می‌کنند که می‌تواند مبنای طراحی و حکمرانی قرار گیرد:

۱. حفظ خودمختاری انسانی: تضمین عقلانیت و امکان مداخله معنادار انسان در تصمیمات، به‌ویژه در لایه‌های پیشینی طراحی که افق انتخاب را شکل می‌دهند.

۲. حق توضیح: تضمین شفافیت الگوریتمی به‌عنوان شرط لازم برای پاسخ‌گویی اخلاقی، تا کاربران بتوانند دلایل پیشنهادهای ماشین را درک کنند.

۳. همسویی ارزشی: جلوگیری از تعارض میان منطق الگوریتمی و ارزش‌های انسانی از طریق نهادینه‌سازی اخلاق در فرآیند توسعه سیستم‌ها.

در نتیجه، تقلیل مسئله خودمختاری به این پرسش که «چه کسی تصمیم نهایی را گرفته است» گمراه‌کننده است. مسئله اصلی در سیستم‌های مبتنی بر هوش مصنوعی، چگونگی شکل‌گیری افق تصمیم‌پذیری است؛ افقی که پیشاپیش برخی گزینه‌ها را معنادار و برخی دیگر را نامرئی می‌سازد. تا زمانی که این لایه پیشینیِ تصمیم‌سازی مورد توجه قرار نگیرد، هرگونه انتساب مسئولیت سطحی و ناکافی خواهد بود. تحلیل اخلاقی باید از لحظه تصمیم نهایی فراتر رود و به فرآیندهای طراحی، تنظیم و ترجمه ارزش‌ها در شبکه انسان—ماشین بپردازد.

جمع‌بندی نهایی این بخش آن است که هوش مصنوعی نمایانگر شکل تازه‌ای از عاملیت است که نه با الگوی فاعل اخلاقی انسانی همسان است و نه در قالب ابزار صرف قابل فهم. این «عاملیت ترکیبی»[1]مستلزم بازتعریف چارچوب‌های اخلاقی سنتی ماست. تنها با به‌رسمیت شناختن این موقعیت میانی است که می‌توان از فروکاست مسئولیت به انسان یا واگذاری نادرست آن به ماشین پرهیز کرد و به سوی مدل‌های واقع‌بینانه‌تر پاسخ‌گویی اخلاقی حرکت نمود (برتونچینی و سرافیم، ۲۰۲۳، ۷).

[1] hybrid agency

۵. راهکارهای کلان: از حکمرانی جمعی تا توانمندسازی انسان

تحلیل‌های پیشین نشان داد که چالش‌های اخلاقی هوش مصنوعی نه صرفاً حاصل پیشرفت فنی، بلکه نتیجه بازپیکربندی رابطه میان عاملیت، مسئولیت و ساختارهای تصمیم‌سازی است. در چنین شرایطی، پاسخ‌های اخلاقی نمی‌توانند به راه‌حل‌های فردمحور یا تکنیکی محدود شوند، بلکه مستلزم رویکردهایی کلان در سطح حکمرانی، طراحی و توانمندسازی انسانی هستند. مراد از «مسئولیت شبکه‌ای» در این مقاله، نه صرفاً تقسیم مسئولیت میان افراد یا نهادها، بلکه نوعی صورت‌بندی از مسئولیت است که در درون شبکه‌ای از کنشگران انسانی و غیرانسانی شکل می‌گیرد. در این چارچوب، مسئولیت نه ویژگی یک فاعل منفرد، بلکه پیامد روابط، ترجمه‌ها و تعاملات میان عناصر مختلف شبکه است؛ به‌گونه‌ای که عاملیت و پاسخگویی به‌صورت برساخته‌ای رابطه‌ای و توزیع‌شده در کل شبکه تولید می‌شود. آنچه این پاسخ‌ها را از توصیه‌های پراکنده متمایز می‌سازد، نسبت آن‌ها با سطوح مختلف مسئله عاملیت و مسئولیت است. همان‌طور که نشان داده شد، شکاف مسئولیتی نه در یک نقطه خاص، بلکه در سه سطح به‌طور هم‌زمان شکل می‌گیرد: سطح نهادیِ حکمرانی و پاسخ‌گویی، سطح طراحی و معماری تصمیم‌سازی، و سطح کنشگر انسانی و ظرفیت خودآیین او. از این رو، هرگونه راهکار اخلاقی اگر تنها یکی از این سطوح را هدف بگیرد، ناگزیر ناقص خواهد بود. سه مسیر پیشنهادی در این بخش، پاسخی لایه‌مند و هماهنگ به این سه سطح‌اند.

نخست، ضرورت طراحی چارچوب‌های حکمرانی مشترک و چندسطحی مطرح می‌شود. پژوهش‌های اخیر در حوزه مدیریت و تنظیم‌گری هوش مصنوعی تأکید می‌کنند که مسئولیت عملکرد سامانه‌های هوشمند باید در شبکه‌ای از کنشگران انسانی و نهادی توزیع شود؛ شبکه‌ای که طراحان، توسعه‌دهندگان، کاربران، نهادهای ناظر و سیاست‌گذاران را در بر می‌گیرد (برتونچینی و سرافیم، ۲۰۲۳، ۷). در این رویکرد، هوش مصنوعی نه به‌عنوان یک ابزار خنثی و نه به‌مثابه یک فاعل اخلاقی مستقل در نظر گرفته می‌شود، بلکه به‌عنوان عنصری از یک سیستم اجتماعی—فناورانه فهم می‌شود که پیامدهای آن تنها از خلال حکمرانی جمعی قابل پاسخ‌گویی است. اهمیت حکمرانی جمعی دقیقاً در همین‌جاست که با منطق عاملیت شبکه‌ای هم‌راستا می‌شود. اگر تصمیم‌ها محصول تعامل شبکه‌ای‌اند، پاسخ‌گویی نیز نمی‌تواند بر مبنای مدل خطی علت—فاعل—پیامد سازمان یابد. حکمرانی چندسطحی تلاشی است برای هم‌شکل‌سازی ساختار پاسخ‌گویی با ساختار واقعی تولید تصمیم. برای تحقق این امر، نهادهای نظارتی باید از رویکردهای صرفاً بازدارنده به سمت رویکردهای مشارکتی حرکت کنند؛ به‌گونه‌ای که در کمیته‌های اخلاق و شوراهای نظارت بر هوش مصنوعی، نه تنها کارشناسان فنی، بلکه جامعه‌شناسان، فیلسوفان و نمایندگان شهروندان نیز حضور داشته باشند تا بتوانند ابعاد هنجاری و اجتماعی الگوریتم‌ها را ارزیابی کنند. در غیاب چنین هم‌خوانی‌ای، مسئولیت یا به‌طور ناموجه بر دوش کاربران نهایی می‌افتد یا در پیچیدگی نهادی مستهلک می‌شود؛ هر دو حالت، بازتولیدکننده شکاف مسئولیتی‌اند، نه حل‌کننده آن.

دوم، بازتعریف نقش الگوریتم‌ها از «راهنمای رفتار» به «توانمندساز تصمیم‌گیری انسانی» اهمیت ویژه‌ای دارد. در ادبیات معاصر اخلاق طراحی، این تحول به‌عنوان گذار از سیستم‌های هدایت‌گر به سیستم‌های تقویت‌کننده قضاوت انسانی توصیف شده است (برتونچینی و سرافیم، ۲۰۲۳، ۷). در این چارچوب، هدف از طراحی هوش مصنوعی نه جایگزینی تصمیم انسانی، بلکه گسترش افق شناختی اوست. سامانه‌های توانمندساز، با شفاف‌سازی معیارهای تصمیم‌گیری، نمایش فعالانه گزینه‌های بدیل و آشکارسازی محدودیت‌ها و سوگیری‌های خود، به کاربر امکان می‌دهند تا آگاهانه‌تر و مسئولانه‌تر تصمیم بگیرد. از منظر تحلیلی،

طراحی توانمندساز پاسخی مستقیم به مسئله «توهم خودمختاری» است. اگر مشکل اصلی، نه حذف اختیار بلکه شکل‌دهی نامرئی افق تصمیم‌پذیری است، آنگاه اخلاق طراحی باید معطوف به آشکارسازی همین افق باشد. برای نمونه، یک سیستم هوشمند پزشکی که به پزشک پیشنهاد تشخیصی می‌دهد، نباید صرفاً نتیجه نهایی را ارائه دهد؛ بلکه باید میزان اطمینان[۱] خود را نمایش دهد، فاکتورهای کلیدی در تصمیم‌گیری را برجسته کند و حتی گزینه‌های دیگری که در نظر گرفته اما رد شده‌اند را فهرست کند. سامانه‌ای که معیارهای خود، منطق اولویت‌بندی و بدیل‌های حذف‌شده را مرئی می‌سازد، نه‌تنها در فرآیند تصمیم مداخله می‌کند، بلکه امکان داوری اخلاقی درباره خودِ مداخله را نیز فراهم می‌آورد. در این معنا، توضیح‌پذیری[۲] صرفاً یک الزام فنی برای رفع باگ نیست، بلکه شرط امکان مسئولیت اخلاقی و اعمال استقلال انسانی است.

برای روشن‌تر شدن سازوکار «توزیع شبکه‌ای مسئولیت»، می‌توان به نمونه‌ای از کاربرد هوش مصنوعی در حوزه پزشکی اشاره کرد. در سامانه‌های تشخیص و درمان مبتنی بر هوش مصنوعی، انتساب مسئولیت در صورت بروز خطا یا آسیب، قابل فروکاستن به یک عامل منفرد نیست و باید در قالب ساختاری چندسطحی تحلیل شود.

در سطح نخست، پزشک انسانی قرار دارد که بسته به میزان استقلال سامانه، در یکی از دو نقش اصلی ایفای وظیفه می‌کند: یا در نقش ناظر و تأییدکننده نهایی خروجی‌های سامانه، یا در نقش همکار تصمیم‌گیر که به‌صورت تعاملی با پیشنهادهای الگوریتمی عمل می‌کند. در حالت نخست، مسئولیت نهایی همچنان بر عهده پزشک است، زیرا او مرجع تصمیم نهایی محسوب می‌شود. با این حال، در حالت دوم و به‌ویژه در شرایطی که سامانه از سطح بالاتری از خودمختاری عملکردی برخوردار است، نسبت دادن مسئولیت کامل به پزشک با محدودیت‌های جدی مواجه می‌شود.

در سطح دوم، نهادهای درمانی و سازمان‌های ارائه‌دهنده خدمات سلامت قرار دارند که مسئولیت انتخاب، به‌کارگیری و نظارت بر سامانه‌های هوش مصنوعی را بر عهده دارند. این نهادها مکلف‌اند از انطباق سیستم‌ها با استانداردهای اخلاقی، حقوقی و بالینی اطمینان حاصل کنند. در صورت قصور در این وظایف نظارتی، مسئولیت نهادی یا سازمانی متوجه آن‌ها خواهد بود.

در سطح سوم، تولیدکنندگان و توسعه‌دهندگان سامانه‌های هوش مصنوعی قرار دارند که مسئول طراحی الگوریتم، کیفیت داده‌های آموزشی، امنیت سیستم و پیشگیری از سوگیری‌های فنی هستند. خطاهای ناشی از نقص الگوریتمی، خطا در آموزش مدل یا آسیب‌های ناشی از ضعف امنیت سایبری، در این سطح قابل انتساب است (چن و همکاران، ۲۰۲۳).

در نهایت، خود سامانه هوش مصنوعی در وضعیت فعلی فاقد جایگاه مسئولیت اخلاقی یا حقوقی مستقل تلقی می‌شود، زیرا از ویژگی‌هایی همچون اراده آزاد، خودآگاهی و قصد اخلاقی برخوردار نیست. با این حال، در برخی دیدگاه‌های آینده‌نگر، امکان بازتعریف جایگاه اخلاقی سامانه‌های بسیار پیشرفته در صورت تحقق سطوحی از خودمختاری عملیاتی(مانند توانایی یک خودروی خودران در تصمیم‌گیری فوری برای کاهش آسیب در تصادف، یا یک ربات جراح در انجام حرکت‌های مستقل) مطرح شده است. جمع‌بندی این مدل نشان می‌دهد که مسئولیت اخلاقی در سامانه‌های هوش مصنوعی، نه در یک فاعل منفرد متمرکز می‌شود و نه به‌طور کامل از میان می‌رود، بلکه در شبکه‌ای چندلایه از کنشگران انسانی و فناورانه توزیع می‌گردد. این توزیع، مستلزم بازنگری در مفاهیم سنتی مسئولیت و پذیرش منطق «عاملیت شبکه‌ای» در تحلیل تصمیم‌گیری‌های بالینی است.

سومین مسیر، توجه به سواد دیجیتال به‌عنوان زیربنای حفظ خودمختاری انسانی است. پژوهش‌ها نشان می‌دهند که در عصر الگوریتم‌ها، توانایی کنش خودآیین بدون درک سازوکارهای تأثیرگذاری فناوری عملاً ناممکن است (یئو، ۲۰۲۵، ۱۰). سواد

---

[۱] Confidence Score

[۲] Explainability

دیجیتال در این معنا صرفاً مهارت کار با ابزارها نیست، بلکه نوعی آگاهی انتقادی نسبت به نحوه شکل‌دهی الگوریتم‌ها به ترجیحات، گزینه‌ها و تصمیمات است. چنین آگاهی‌ای شامل درک مفاهیمی مانند «سوگیری الگوریتمی»، «حباب فیلتر» و «اقتصاد توجه» است. برنامه‌های آموزشی باید شهروندان را توانمند سازند تا تشخیص دهند چه زمانی یک پیشنهاد، محصول تحلیل عینی داده‌هاست و چه زمانی نتیجه منطق تجاری یا ایدئولوژیک طراحان سیستم. چنین آگاهی‌ای می‌تواند از تأثیرپذیری پنهان جلوگیری کرده و امکان مقاومت در برابر توهم خودمختاری را فراهم سازد. سواد دیجیتال، سپر دفاعی انسان در برابر قدرت جهت‌دهنده الگوریتم‌هاست و بدون آن، حتی بهترین طراحی‌های توانمندساز نیز ممکن است نتوانند از عقلانیت انسان دفاع کنند.

می‌توان چارچوب مسئولیت در سیستم‌های مبتنی بر هوش مصنوعی را در سه سطح متمایز اما به‌هم‌پیوسته تحلیل کرد: در سطح طراحی، توسعه‌دهندگان و شرکت‌های فناور با انتخاب داده‌ها، تعیین اهداف و معماری الگوریتمی، در شکل‌دهی افق انتخاب کاربران نقش دارند و مسئولیت آن‌ها ناظر بر سوگیری‌ها و جهت‌دهی پیشینی تصمیم‌هاست. در سطح نهادی، دولت‌ها و نهادهای قانون‌گذار با تنظیم‌گری، تضمین شفافیت و اعمال نظارت، مسئول جلوگیری از تمرکز قدرت الگوریتمی و حفظ پاسخ‌گویی‌اند. در سطح کاربر، افراد به‌عنوان تصمیم‌گیرندگان نهایی، مسئول ارزیابی و انتخاب میان گزینه‌های ارائه‌شده هستند، هرچند این انتخاب‌ها در چارچوبی از پیش‌ساخته و الگوریتمی‌شده صورت می‌گیرد.

بر این اساس، مسئولیت اخلاقی در نظام‌های مبتنی بر هوش مصنوعی نه در یک فاعل منفرد متمرکز می‌شود و نه از میان می‌رود، بلکه به‌صورت شبکه‌ای و چندسطحی توزیع می‌گردد. کاهش شکاف مسئولیتی مستلزم هم‌راستاسازی ساختارهای پاسخ‌گویی با این واقعیت شبکه‌ای، شفاف‌سازی فرآیندهای طراحی، و تقویت توانایی کاربران در درک و ارزیابی تصمیم‌های الگوریتمی است.

نتیجه‌گیری

در جمع‌بندی این پژوهش می‌توان گفت مسئلهٔ «ارادهٔ آزاد در عصر هوش مصنوعی» در درجهٔ نخست پرسشی درباره توانایی ماشین‌ها برای اراده نیست؛ بلکه مسئله، بازتعریف موقعیت انسانی در شبکه‌ای از سیستم‌های پیش‌بین، پیشنهاددهنده و تصمیم‌ساز است. محیط‌های هوشمند مفهوم آزادی را ابداع نکرده‌اند، بلکه آن را در قالبی نو بازتولید و تقویت کرده‌اند. این ادعا که انسان خود را مختار می‌پندارد، حال آنکه انتخاب‌هایش حاصل مجموعه‌ای از علل و شرایط پیشینی است، امروز در معماری‌های داده‌محور و الگوریتم‌های یادگیری ماشین صورتی فناورانه یافته است. این دگرگونی را می‌توان نه‌تنها در سطح فناوری، بلکه در سطح زیست‌شناختی و شناختی انسان نیز مورد بازاندیشی قرار داد.

این نگرانی را می‌توان با استناد به یافته‌های علوم اعصاب نیز تقویت کرد. آزمایش‌های بنجامین لیبت (رک. لیبت الف، ۲۰۰۴، ۱۲۴) نشان می‌دهد که فعالیت‌های عصبی آغازگر یک عمل، پیش از آگاهی فرد از تصمیم شکل می‌گیرند؛ موضوعی که سم هریس (رک. هریس ۲۰۱۰؛ ۲۰۱۲) با اتکا به آن استدلال می‌کند ارادهٔ آزاد بیش از آن‌که یک قدرت علی واقعی باشد، تجربه‌ای ذهنی است. در چشم‌انداز پژوهش حاضر، هوش مصنوعی نوعی «لیبتِ فناورانه» ایجاد می‌کند؛ انسان همچنان خود را آغازگر تصمیم‌ها می‌داند، حال آنکه بخش مهمی از فرایند تصمیم‌سازی پیش از ورود آگاهی او، در لایه‌های پنهان الگوریتمی رخ داده است. با این حال، این تفسیرها اگرچه نشان‌دهنده تقدم برخی فرآیندهای ناخودآگاه بر آگاهی‌اند، اما لزوماً به نفی کامل خودمختاری هنجاری منجر نمی‌شوند؛ زیرا مسئولیت اخلاقی نه صرفاً به آغاز علّی کنش، بلکه به توانایی بازاندیشی، ارزیابی و اصلاح دلایل کنش وابسته است.

ادبیات معاصر در حوزهٔ اخلاق هوش مصنوعی نیز مؤید همین دگرگونی است. چن و همکاران (۲۰۲۳) از فرسایش مرزهای مسئولیت در سامانه‌های هوشمند سخن می‌گویند؛ وضعیتی که در آن «هیچ‌کس» مسئول نهایی تلقی نمی‌شود. گوگوشین (۲۰۲۱، ۲۰۲۳) این وضعیت را با مفهوم «کنشگران غیرقابل‌مجازات» توصیف می‌کند؛ جایی که الگوریتم‌ها اثرگذارند اما پاسخ‌گو نیستند و کاربران در موقعیت عاملی قرار می‌گیرند که اختیارشان تضعیف شده است. پژوهش‌هایی در حوزه شهرهای هوشمند، تصمیم‌سازی شخصی‌شده و خودمختاری دیجیتال (وانگ، ۲۰۲۴؛ لو، ۲۰۲۲؛ چسترمن، ۲۰۲۰) نشان می‌دهند که مسئله نه اجبار مستقیم، بلکه بازآرایی تدریجی «شرایط امکان انتخاب» است. انسان آزاد به نظر می‌رسد، اما افق انتخاب او در لایه‌های داده، پیش‌بینی و طراحی الگوریتمی شکل می‌گیرد.

بر این اساس، ادعای اصلی پژوهش را می‌توان به‌صورت دقیق‌تری صورت‌بندی کرد: خودمختاری انسانی در عصر هوش مصنوعی نه به‌کلی از میان می‌رود و نه دست‌نخورده باقی می‌ماند، بلکه دچار دگرگونی ساختاری می‌شود. انسان همچنان مسئول شناخته می‌شود، اما بخش‌هایی اساسی از زنجیرهٔ علی تصمیم، از جمع‌آوری داده تا پیشنهاد الگوریتمی، خارج از کنترل مستقیم او قرار دارد. این گسست میان مسئولیت و اختیار، بنیادی‌ترین چالش اخلاقی در زیست‌جهان دیجیتال معاصر است.

مشارکت نظری این مقاله در آن است که با بهره‌گیری از رویکرد شبکه‌ای به عاملیت و مسئولیت، مسئلهٔ ارادهٔ آزاد را از سطح «کنترل فردی» به سطح «طراحی شرایط انتخاب» منتقل می‌کند. در این چارچوب، اخلاق هوش مصنوعی نه صرفاً به رفتار کاربران یا قابلیت‌های فنی سامانه‌ها، بلکه به ساختارهای پنهان تصمیم‌سازی، سوگیری‌های الگوریتمی و پیش‌فرض‌های طراحی معطوف می‌شود؛ همان لایه‌هایی که بیشترین نقش را در شکل‌دهی انتخاب‌ها ایفا می‌کنند. از این رو، هر نظریه‌ای از مسئولیت اخلاقی که همچنان بر الگوی فردگرایانه کنترل و انتخاب تکیه دارد، در تبیین کنش در زیست‌جهان الگوریتمی ناکافی خواهد بود؛ و تنها با پذیرش عاملیت شبکه‌ای و انتقال کانون تحلیل به سطح طراحی شرایط انتخاب است که می‌توان از امکان پاسخ‌گویی اخلاقی معنادار در عصر هوش مصنوعی دفاع کرد.

منابع

الستی, کیوان. (۱۴۰۳). هوش مصنوعی و کنترل معنادار بشری: سازوکارهای واکنش‌پذیر به دلایلِ نسبت-به-عامل-خنثی و امکان ردگیری عامل(های) مسئول، پژوهش‌های فلسفی –کلامی، ۲۶(۴)، ۲۷-۵۴. https://www.doi.org/10.22091/jptr.2025.11378.3139

بلبلی قادیکلایی، سمیه و پارسانیا، حمید. (۱۴۰۲). مروری نظام‌مند بر دلالت‌های اخلاقی استفاده از هوش مصنوعی در فناوری‌های دیجیتال و نسبت آن با اخلاق شکوفایی. *راهبرد اجتماعی فرهنگی*، ۱۲ (۴۸)، ۷۹۸-۷۷۱. https://doi.org/10.22034/scs.2022.160772

حسینی سورکی، سید محمد. (۱۴۰۳). معنا و معیار عاملیت اخلاقی ماشین‌های هوشمند. *پژوهش‌های فلسفی کلامی*، ۱۰۲، ۱۰۸-۸۷. https://doi.org/10.22091/jptr.2025.12466.3256

خزاعی, زهرا. (۱۳۹۸). Agency and Virtues، *پژوهش‌های فلسفی-کلامی*، ۲۱(۳)، ۱۱۹-۱۴۰. https://doi.org/10.22091/jptr.2019.4673.2190

زرگر، زهرا. (۱۴۰۴). رابطه عواطف و ظرفیت اخلاقی در فناوری‌های هوش مصنوعی. *پژوهش‌های فلسفی*، ۵۰، ۴۰-۲۲. https://doi.org/10.22034/jpiut.2024.63626.3875

عاشوری کیسمی، محمدعلی و پرویزی، مریم. (۱۴۰۳). بررسی عاملیت اخلاقی هوش مصنوعی عام. *پژوهش‌های علم و دین*، ۱۵ (۱)، ۱۵۱-۱۲۵. https://www.doi.org/10.30465/srs.2024.49282.2162

قوامی‌پور، محدثه و محمودی، امیررضا. (۱۴۰۳). واکاوی ظهور هوش مصنوعی و تاثیرات آن در جامعه و دین. *پژوهش‌های علم و دین*، ۱۵ (۱)، ۲۱۳-۲۴۴. https://doi.org/10.30465/srs.2024.48601.2135

محمدی، شادی و مقدم حیدری، غلامحسین (۱۳۹۱)، «بازسازی و بررسی مقولات متافیزیکی نظریه شبکه‌کنشگر برونو لاتور»، *فصلنامه ذهن*، ۱۳(۵۲)، ۱۷۲-۱۳۹.

میری بالاجورشری، سیده مهشید و محمودی، امیررضا. (۱۴۰۳). واکاوی مسائل اخلاقی در زمینه هوش مصنوعی با نگاهی به اخلاق اسلامی. *مطالعات اخلاق کاربردی*، ۱۴(۲)، ۱۲۳-۹۷. https://doi.org/10.22081/jare.2024.68803.1901

References

Alasti, K. (2024). Artificial Intelligence and Meaningful Human Control: Agent-Neutral-Reason-Responsive Mechanisms and the Possibility of Tracing Responsible Agent(s), *Journal of Philosophical Theological Research (JPTR)*, 26(4), 27-54. (in Persian) https://www.doi.org/10.22091/jptr.2025.11378.3139

Albareda, J. L. (2025). Uncovering the gap: challenging the agential nature of AI responsibility problems. *AI and Ethics*. Springer. https://doi.org/10.1007/s43681-025-00685-w

Ashouri Kisomi, M. A., & Parvizi, M. (2024). Investigation of the ethical agency of AGI. *Science and Religion Studies*, 15(1), 125-151. (in Persian) https://www.doi.org/10.30465/srs.2024.49282.2162

Battisti, D. (2025). Second-Person Authenticity and the Mediating Role of AI: A Moral Challenge for Human-to-Human Relationships? *Philosophy & Technology*.

Bertoncini, A. L. C., & Serafim, M. C. (2023). Ethical content in artificial intelligence systems: A demand explained in three critical points. *Frontiers in Psychology, 14*, 1074787. https://doi.org/10.3389/fpsyg.2023.1074787

Bolboli Qadikolaei, S. & Parasnath, H. (2023). A systematic review of the ethical implications of using artificial intelligence in digital technologies and its relationship with the ethics of flourishing. *Socio-Cultural Strategy*, *12*(3), 771-798. (in Persian) https://doi.org/10.22034/scs.2022.160772

Callon, M. (1986). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St Brieuc Bay. In J. Law (Ed.), Power, Action and Belief: A *New Sociology of Knowledge* (196–223). London: Routledge.

Chen, A., Wang, C., & Zhang, X. (2023). Reflection on the equitable attribution of responsibility for artificial intelligence-assisted diagnosis and treatment decisions. *Intelligent Medicine*, *3*(02), 139-143.

Chesterman, S. (2020). Artificial intelligence and the problem of autonomy. *Notre Dame Journal on Emerging Technologies*, 1, 210–250.

Choi, D. Y. (2023). The trustworthiness of AI: Comments on Simion and Kelp's account. Asian Journal of Philosophy, 2(1), 1–9.

Constantinescu, M., Vică, C., Uszkai, R., & Voinea, C. (2022). Blame it on the AI? On the moral responsibility of Artificial Moral Advisors. *Philosophy & Technology*, 35 (1), 1-34. https://doi.org/10.1007/s13347-022-00529-z

Constantinescu, M., Voinea, C., Uszkai, R., & Vică, C. (2021). Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context. *Ethics and Information Technology, 23*(4), 803-814.

Cummings, M. (2012), Automation Bias in Intelligent Time Critical Decision Support Systems, AIAA 1st Intelligent Systems Technical Conference, *Session: IS-15: Human Interaction with Intelligent Systems.*

Dubber, M. D., Pasquale, F., & Das, S. (Eds.). (2020). *The Oxford handbook of ethics of AI*. Oxford University Press.

Ghavamipour Sereshkeh, M. & Mahmooudi, A. (2024). Analyzing the emergence of artificial intelligence and its effects on society and religion. *Science and Religion Studies*, *15*(1), 213-244. (in Persian) https://doi.org/10.30465/srs.2024.48601.2135

Gogoshin, D. L. (2021). Robot responsibility and moral community. *Frontiers in Robotics and AI*, *8*, https://doi.org/10.3389/frobt.2021.768092

Gogoshin, D. L. (2023). A challenge for the scaffolding view of responsibility. *Ethical Theory and Moral Practice*, *26*(1), 73-90. https://doi.org/10.1007/s10677-022-10340-6

Gogoshin, D. L. (2024). Patient preferences concerning humanoid features in healthcare robots. *Science and Engineering Ethics*, *30*(6),48-62. https://doi.org/10.1007/s11948-024-00508-x

Gogoshin, D. L. (2025). A way forward for responsibility in the age of AI. *Inquiry*, *68*(4), 1164-1197. https://doi.org/10.1080/0020174X.2025.2198690

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Harris, S. (2010). *The Moral Landscape*. New York: Free Press.

Harris, S. (2012). *Free Will*, New York: Free Press

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

Kant, I. (1972). *The Moral Law: Kant's Groundwork of the Metaphysic of Morals* (H. J. Paton, Trans.). London: Hutchinson University Library.

Khazaie, Z. (2019). Agency and Virtues, *Journal of Philosophical Theological Research (JPTR)*, 21(81), 119-140. https://doi.org/10.22091/jptr.2019.4673.2190

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

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, 8(4), 529 – 539. https://doi.org/10.1017/S0140525X00044903

Libet, B. (2004a). Intention to act: Do we have free will? In *Mind time: The temporal factor in consciousness* (pp. 123-156). Harvard University Press.

Libet, B. (2004b), *Mind time*, MA: Harvard University Press.

Lu, W. (2022). Inevitable challenges of autonomy: Ethical concerns in personalized algorithmic decision-making. *AI & Society*, 37 (3), 1203-1215. https://doi.org/10.1007/s00146-021-01330-w

Ma, W., & Valton, V. (2024). Toward an Ethics of AI Belief. *Philosophy and Technology*, 37(76), 1–28. https://doi.org/10.1007/s13347-024-00762-8

Miri Balajorshari, S. M., & Mahmoudi, A. (2024). Analysis of ethical challenges in the field of artificial intelligence with an approach to Islamic ethics. *Journal of Applied Ethics Studies*, *14*(2), 97-123. (in Persian) https://doi.org/10.22081/jare.2024.68803.1901

Mohammadi, Sh. & Moghadam Heidari, Gh. (2012), Reconstruction and Investigation of Metaphysical Categories of Bruno Latour's Actor Network Theory, *Mind Quarterly*, 13 (52), 139-172. (in Persian)

Nyrup, R. (2023). Trustworthy AI: a plea for modest anthropocentrism. Asian Journal of Philosophy, 2(2), 1–10.

Russo, F., Schliesser, E., & Wagemans, J. (forthcoming). Connecting ethics and epistemology of AI. *AI and Society*, 1–19.

Tollon, F. (2022). Is AI a Problem for Forward Looking Moral Responsibility? The Problem Followed by a Solution. In Communications in Computer and Information Science. pp. 307–318. Cham.

Vallor, S. (2016). New social media and the technomoral virtues. In *Technology and the virtues: A philosophical guide to a future worth wanting* (pp. 159-187). Oxford University Press.

Vold, K., & Hernandez-Orallo, J. (2022). AI Extenders and the Ethics of Mental Health. In M. Ienca & F. Jotterand (Eds.), Ethics of Artificial Intelligence in Brain and Mental Health.

Wang, B. (2024). Ethical reflections on the application of artificial intelligence in the construction of smart cities. *Journal of Engineering*, Vol. 2024, 8207822. https://doi.org/10.1155/2024/8207822

Yeo, l. (2025). Autonomy and the question of free will in the age of artificial intelligence. *American Research Journal of Humanities and Social Sciences*, 11 (1), 1-6. https://doi.org/10.21694/2378-7031.25001.

Zargar, Z. (2025). On the Relation of Emotion and Moral Capacity in Artificial Intelligence Technologies. *Journal of Philosophical Investigations*, *19*(50), 19-40. (in Persian) https://doi.org/10.22034/jpiut.2024.63626.3875