An Analysis of the Consequences of Delegating Cognitive Tasks to Artificial Intelligence and Its Impact on Moral Agency

Hamid Ahmadi Hedayat ◇

Article Type: Research Paper
Vol. 35 | **No.** 3 | **Serial** 99 | Oct. 2025
Received: 2025.12.24
Revised: 2026.03.14
Accepted: 2026.04.21
Published Online: 2026.07.26
Pages: 33-42
P-ISSN: 1027-2690
E-ISSN: 2783-4514

Abstract

Today, generative artificial intelligence (AI) is becoming integrated into nearly every aspect of daily life, representing far more than a simple technological shift. Rather, it is fundamentally reshaping human cognitive processes. From routine activities such as navigating with GPS to retrieving information from cloud-based systems, these developments reflect an increasing reliance on what cognitive science describes as cognitive offloading. This concept is based on the premise that human cognition is not confined to the brain but can extend into the environment through external artifacts and technological systems. The present study argues that when such delegation extends beyond mathematical computation and formal reasoning into the normative and ethical domain, it poses a fundamental challenge to human moral agency. In this context, AI has evolved from a cognitive aid into a potential moral advisor, consulted not only for practical decisions but also for deeply personal and value-laden ethical questions.

To understand the depth of this challenge, it is first necessary to examine the structure of moral agency. Adopting a tripartite theoretical framework, this study conceptualizes moral agency as the integration of Aristotelian practical wisdom (phronesis), Kantian autonomy, and moral accountability. The

Keywords

Artificial Intelligence, Moral Agency, Practical Wisdom (Phronesis), Autonomy, Philosophy of Technology.

◇ Assistant Professor Educational Sciences, Department of Educational Sciences, Farhangian University, Tehran, Iran (Corresponding Author)
h.hedayat@cfu.ac.ir
ORCID: 0000-0003-0814-5696

Cite This Paper: Ahmadi Hedayat, H. (2025). An Analysis of the Consequences of Delegating Cognitive Tasks to Artificial Intelligence and Its Impact on Moral Agency. *Rahyaft, 35 (3)*, 33-42. (Persian).

DOI: 10.22034/RAHYAFT.2026.12218.1641

delegation of moral deliberation to AI systematically threatens each of these three dimensions.

The first dimension, phronesis, is a virtue cultivated through lived experience and emotional engagement with particular situations. Aristotle argued that practical wisdom is not a form of theoretical knowledge that can be reduced to codified rules. Rather, it develops through engagement with the ambiguous and context-dependent circumstances of everyday life, where universal principles alone are insufficient. By providing rapid, categorized, and optimized responses, AI effectively reduces what may be described as moral friction—the hesitation, uncertainty, and emotional burden that accompany difficult moral decisions. As this friction diminishes, opportunities for moral development are likewise diminished. Just as excessive reliance on GPS has weakened individuals' spatial navigation abilities, dependence on AI for resolving ethical dilemmas may gradually erode the capacity for moral judgment in unique and complex situations. Consequently, society risks producing individuals who are technologically competent yet deficient in practical moral wisdom. The second major challenge concerns the principle of autonomy, which constitutes the cornerstone of Immanuel Kant's deontological ethics. From a Kantian perspective, an action possesses moral worth only when the moral agent legislates the moral law for themselves through rational self-governance. However, AI systems-particularly those based on complex deep-learning architectures-operate largely as black boxes, whose internal reasoning processes remain opaque even to their developers. Consequently, when individuals act upon AI-generated recommendations without understanding the underlying rationale, they shift from autonomous moral judgment toward what may be described as algorithmic heteronomy. Such dependence fosters a passive form of moral deference that conflicts with the Kantian ideal of autonomous agency. In this context, the locus of moral authority gradually shifts from the individual's rational will to an external and non-transparent system that possesses neither conscience nor moral consciousness.

The third challenge concerns responsibility and accountability. Moral agency presupposes a meaningful causal and intentional connection between the moral agent and the consequences of an action. By obscuring the attribution of responsibility, AI contributes to what has been termed the responsibility gap. This gap arises when harm results from the operation of an autonomous system, yet responsibility cannot be clearly assigned to the user, the developer, or the system itself. From a psychological perspective, AI may also function as a moral buffer. Research on automation bias suggests that individuals tend to place greater trust in algorithmic recommendations than in their own judgment and, when errors occur, are more likely to attribute responsibility to the technology rather than to themselves. As a result, the human agent risks becoming little more than an executor of algorithmic recommendations, carrying out decisions without fully experiencing their moral significance.

Furthermore, delegating moral deliberation to AI reinforces forms of technological instrumentalism and technological solutionism. This perspective assumes that complex ethical and social problems can be addressed primarily as technical challenges amenable to algorithmic solutions. However, moral judgment differs fundamentally from computational information processing. A crucial distinction must therefore be drawn between decision-making, which is primarily a computational process guided by predefined criteria, and choosing, which is an intrinsically human activity grounded in values, responsibility, and moral commitment. AI operates through processes of mathematical optimization, whereas human moral choice frequently involves tragic dilemmas, conflicting obligations, and irreducible ambiguity that cannot be adequately represented through quantitative variables. Consequently, when qualitative ethical dilemmas are delegated to algorithmic systems, they are transformed into computational problems, thereby diminishing the moral richness and normative complexity that define ethical reasoning.

In conclusion, the findings of this study suggest that the uncritical integration of AI into the normative dimensions of human life represents not merely a technological advancement but also a significant challenge to the foundations of moral agency. Moral agency is not a fixed attribute acquired once and for all; rather, it is a dynamic capacity cultivated

through continuous moral engagement and reflective practice. Excessive reliance on AI-mediated ethical reasoning may gradually weaken practical wisdom, encourage forms of algorithmic heteronomy, and widen the responsibility gap.

To address these challenges, this study proposes a twofold strategy. First, educational systems should promote critical AI literacy, enabling individuals to distinguish computational problems from normative questions and to recognize the epistemic and cognitive limitations of AI systems. Second, AI interfaces should be redesigned to incorporate reflective friction rather than maximizing efficiency alone. Instead of offering definitive prescriptive recommendations, AI systems should encourage users to pause, deliberate, and critically evaluate the ethical implications of their choices-for example, through Socratic questioning concerning possible consequences and competing values. Such an approach helps preserve AI as a cognitive and informational aid while ensuring that human beings remain the primary moral decision-makers and bear ultimate responsibility for their actions. Without such safeguards, societies risk becoming technologically sophisticated while experiencing a gradual erosion of moral judgment and ethical responsibility.

رهیافت

تحلیلی بر پیامدهای واگذاری فعالیت‌های شناختی به هوش مصنوعی و تأثیر آن در عاملیت اخلاقی

حمید احمدی هدایت ◇

نوع مقاله: پژوهشی
دوره ۳۵ | شماره ۳ | پیاپی ۹۹ | مهر ۱۴۰۴
تاریخ دریافت: ۱۴۰۴/۱۰/۰۳
تاریخ بازنگری: ۱۴۰۴/۱۲/۲۳
تاریخ پذیرش: ۱۴۰۵/۰۲/۰۱
تاریخ انتشار برخط: ۱۴۰۵/۰۵/۰۴
صفحات: ۳۳-۴۲
شاپای چاپی: ۱۰۲۷-۲۶۹۰
شاپای الکترونیکی: ۲۷۸۳-۴۵۱۴

چکیده

استفاده از هوش مصنوعی در فرایندهای تصمیم‌گیری روزمره به پدیده‌ای فراگیر در میان مردم تبدیل شده است. فرایندی که طی آن افراد در فعالیت‌های روزانه خود به‌نوعی پردازش‌های ذهنی خود را به سیستم‌های فنّاورانه بیرونی واگذار می‌کنند. اگرچه این امر در وظایف محاسباتی موجب افزایش بهره‌وری و تسریع می‌شود، اما گسترش آن به قلمرو هنجاری اخلاق باعث نگرانی و بروز پرسش‌های ژرف فلسفی خواهد شد. پژوهش حاضر با هدف تحلیل پیامدهای واسپاری پردازش‌های ذهنی عادت‌گونه به هوش مصنوعی و تأثیر ساختاری آن در عاملیت اخلاقی انسان می‌پردازد. برای دستیابی به این هدف با اتخاذ رویکردی فلسفی- تحلیلی و بهره‌گیری از چارچوب نظری سه‌وجهی مبتنی بر حکمت عملی ارسطو، خودمختاری کانت، و مفهوم مسئولیت به تحلیل رابطهٔ میان انسان و هوش مصنوعی پرداخته شد. یافته‌ها نشان از آن دارد که واگذاری تأملات اخلاقی به الگوریتم‌های هوش مصنوعی، ساختار بنیادین عامل اخلاقی را از سه طریق تضعیف می‌کند: (۱) با حذف اصطکاک اخلاقی لازم در تصمیم‌گیری‌ها، به زوال حکمت عملی منجر می‌شود (مهارت‌زدایی اخلاقی)؛ (۲) با جایگزین کردن خروجی‌های غیرشفاف ماشین به‌جای خودقانون‌گذاری مستقل، باعث می‌شود وضعیتی از ناخودمختاری الگوریتمی پدید آید؛ و در نهایت (۳) با ابهام‌آلود کردن پاسخگویی، شکاف مسئولیت روان‌شناختی ایجاد می‌کند. نتیجه اینکه با برون‌سپاری تأمل اخلاقی به هوش مصنوعی، خطر افتادن در دام نگاه ابزارانگارانهٔ افراطی به فنّاوری را به جان می‌خریم. اگرچه هوش مصنوعی می‌تواند همچون یاور و پشتیبان اطلاعاتی عمل کند، اما برخورد با آن به‌مثابهٔ نایب و تصمیم‌گیرندهٔ اخلاقی می‌تواند خطر تهی‌سازی فاعل اخلاقی را

کلیدواژه‌ها
هوش مصنوعی، عاملیت اخلاقی، حکمت عملی، خودمختاری، فلسفه فنّاوری.

◇ استادیار علوم تربیتی، گروه آموزش علوم تربیتی دانشگاه فرهنگیان، تهران، ایران (پدیدآور رابط)
h.hedayat@cfu.ac.ir
ORCID: 0000-0003-0814-5696

استناد به این مقاله: احمدی هدایت، ح. (۱۴۰۴). تحلیلی بر پیامدهای واگذاری فعالیت‌های شناختی به هوش مصنوعی و تأثیر آن در عاملیت اخلاقی. *رهیافت، ۳۵ (۳)*، صص. ۴۲-۳۳.
DOI: 10.22034/RAHYAFT.2026.12218.1641

ناشر: مؤسسه تحقیقات سیاست علمی کشور

در پی داشته باشد. توسعهٔ سواد انتقادی هوش مصنوعی کاربران و توجه به اصطکاک تأاملی در طراحی ابزارهای هوش مصنوعی برای برون‌رفت از این مسئله را می‌توان توصیه کرد.

مقدمه

بهره‌گیری از ابزارهای هوش مصنوعی مولد[1] در تاروپود زندگی روزمره، تغییری شگرف در عملکردهای شناختی انسان ایجاد کرده است. از مسیریابی در خیابان‌های شهر با جی‌پی‌اس[2] گرفته تا بازیابی اطلاعات از سرورهای ابری[3]، افراد به‌طور چشمگیری برای تقویت یا جایگزینی تلاش‌های شناختی خود به سیستم‌های فنّاورانه بیرونی متکی شده‌اند. این پدیده باربرداری شناختی[4] نامیده می‌شود که شامل استفاده از ابزارهای کمکی بیرونی برای کاهش تقاضای شناختی درونی است (Risko & Gilbert, 2016). باربرداری شناختی که در ابتدا به‌عنوان امتداد ذهن انسان در محیط مفهوم‌سازی می‌شد (Clark & Chalmers, 1998)، از تکیه بر ابزارهای ایستا به شراکتی پویا با سیستم‌های هوش مصنوعی فعال و یادگیرنده تکامل یافته است. مدل‌های هوش مصنوعی مولد همچون چت جی‌پی‌تی[5] و سیستم‌های پیچیدهٔ پشتیبان تصمیم‌گیری، اکنون نه‌تنها در مسائل واقعی یا آمایشی راهنمایی ارائه می‌دهند، بلکه به‌طور فزاینده‌ای برای مشاوره در باب مسائل پیچیده، هنجاری و عمیقاً شخصی مورد مشورت قرار می‌گیرند.

این گسترش نقش هوش مصنوعی از پروتز شناختی[6] به مشاور اخلاقی بالقوه[7]، نشان‌دهندهٔ نقطهٔ عطف مهمی است. درحالی‌که واگذاری برخی وظایف عمدتاً بی‌خطر است، واگذاری رنج تأمل اخلاقی ماهیتاً متفاوت است. فرایند استدلال اخلاقی صرفاً رسیدن به یک خروجی صحیح از مجموعه‌ای از ورودی‌ها نیست، بلکه تمرینی تکوینی است که شخصیت را شکل می‌دهد، فضیلت می‌سازد و جوهرهٔ اصلی آنچه معنای عامل اخلاقی بودن است، تشکیل می‌دهد (Vallor, 2015). با مهارت یافتن هرچه بیشتر سیستم‌های هوش مصنوعی در شبیه‌سازی استدلال اخلاقی و ارائهٔ توصیه‌های اخلاقی به‌ظاهر موجه این پرسشی اساسی مطرح می‌شود: هنگامی که کار شناختی و عاطفی تصمیم‌گیری اخلاقی به یک ماشین برون‌سپاری می‌شود در اصل چه بر سر عامل اخلاقی انسانی می‌آید؟

فرسایش بالقوهٔ عاملیت اخلاقی[8] انسان در نتیجهٔ وابستگی شناختی فزاینده به هوش مصنوعی برای دریافت هدایت اخلاقی از جمله مسائل مهمی است که می‌تواند محل کنکاش فلسفی قرار گیرد. عاملیت اخلاقی یک ظرفیت تک‌ساحتی نیست، بلکه تعاملی پیچیده از مؤلفه‌های فلسفی متمایز، شامل حکمت عملی[9]، خودمختاری[10] و مسئولیت[11] است. باربرداری تأملات اخلاقی به هوش مصنوعی در اصل هریک از این ارکان را به‌طور نظام‌مند تهدید می‌کند.

نخست، این امر توسعهٔ حکمت عملی یا (فرونسیس)[12] را به خطر می‌اندازد؛ فضیلتی ارسطویی که به‌معنای تشخیص مسیر درست عمل در موقعیت‌های خاص و جزئی است (Aristotle, 2009). فرونسیس دانشی نظری نیست که بتوان آن را محاسبه کرد، بلکه مهارتی است که از طریق تجربهٔ زیسته، درگیری عاطفی و تمرین مکرر در پیمایش ابهامات اخلاقی صیقل می‌یابد. سیستم‌های هوش مصنوعی با ارائهٔ راه‌حل‌های سریع و طبقه‌بندی‌شده، به افراد امکان می‌دهند تا این کشمکش ضروری و شخصیت‌ساز را دور بزنند و در نتیجه وضعیتی از مهارت‌زدایی اخلاقی[13] را تقویت می‌کنند (Vallor, 2016).

دوم، اینکه پدیدهٔ اصل خودمختاری اخلاقی[14] را به چالش می‌کشد؛ اصلی که سنگ بنای اخلاق کانتی است و فرض را بر این می‌گذارد که عامل اخلاقی کسی است که قانون اخلاقی را خود بر خویشتن وضع می‌کند (Kant, 1785/2012). تکیه بر یک مرجع الگوریتمی بیرونی برای جهت‌گیری اخلاقی، نوعی ناخودمختاری[15] است که در آن کانون اقتدار اخلاقی از ارادهٔ عقلانی فرد به یک سیستم بیرونی و غیرشفاف منتقل می‌شود. این امر وابستگی منفعلانه‌ای را ایجاد می‌کند که با ایده‌آل عامل اخلاقی خودقانون‌گذار[16] در تضاد است.

در نهایت، این وابستگی یک شکاف مسئولیت[17] قابل توجه ایجاد می‌کند (Sparrow, 2007; Matthias, 2004). هنگامی که تصمیمی براساس توصیهٔ هوش مصنوعی به نتیجه‌ای منفی منجر می‌شود، پاسخگویی به شکلی خطرناک مخدوش می‌گردد. کاربر می‌تواند الگوریتم را مقصر بداند، درحالی‌که الگوریتم، به‌عنوان موجودیتی فاقد آگاهی است و نمی‌تواند مسئول شناخته شود. این ابهام، پیوند حیاتی میان عمل، پیامد و پاسخگویی را که زیربنای هر چارچوب اخلاقی کارآمدی است زیر سؤال می‌برد.

بنابراین، پذیرش غیرانتقادی هوش مصنوعی به‌عنوان یک داور اخلاقی، تهدیدی مستقیم برای پرورش عاملان اخلاقی بافضیلت،

1. Generative AI
2. GPS
3. cloud-based servers
4. cognitive offloading
5. ChatGPT
6. cognitive prosthesis
7. a potential moral advisor
8. moral agency
9. practical wisdom
10. autonomy
11. responsibility
12. Phronesis
13. moral deskilling
14. Moral Autonomy
15. Heteronomy
16. self-legislator
17. Responsibility Gap

خودمختار و مسئول است. پژوهش حاضر مسیرهای مشخصی را تحلیل می‌کند که از طریق آن، این وابستگی شناختی بر عاملیت اخلاقی تأثیر می‌گذارد و استدلال می‌کند که آنچه در حال باربرداری است، صرفاً تلاش شناختی نیست، بلکه خود عمل زیستن به‌مثابهٔ یک انسان اخلاقی است.

به‌رغم گسترش روزافزون ادبیات در حوزهٔ اخلاق هوش مصنوعی، بخش عمده‌ای از پژوهش‌های پیشین بر اخلاق ماشین[1] یعنی چگونگی طراحی سیستم‌های ایمن و منصفانه یا تحلیل سوگیری داده‌ها متمرکز بوده‌اند. در مواردی هم که تأثیرات انسانی بررسی شده، غالباً به پیامدهای روان‌شناختی مجزا تقلیل یافته است. خلأ مشخصی که در ادبیات فلسفی (به‌ویژه در پژوهش‌های زبان فارسی) احساس می‌شود، نبود تحلیلی ساختاریافته از تأثیر این فنّاوری‌ها بر خود عامل اخلاقی است. نوآوری اصلی پژوهش حاضر در سطح ترکیب نظری و ارائهٔ استدلال مستقل قرار دارد؛ به این معنا که با صورت‌بندی یک چارچوب جامع سه‌وجهی (متشکل از حکمت عملی، خودمختاری و مسئولیت)، نشان می‌دهد که چگونه ایدهٔ ابزارانگاری فنّاورانه نه‌تنها در حل مسائل اخلاقی ناکام می‌ماند، بلکه به‌صورت نظام‌مند به فرسایش ساختاری عاملیت انسانی منجر می‌شود. پژوهش حاضر فراتر از یک بازخوانی مفهومی است و در اصل استدلال می‌کند که واگذاری تأملات هنجاری به ماشین با ماهیت تکوینی اخلاق از اساس در تضاد است.

روش‌شناسی پژوهش

پژوهش حاضر از روش تحلیل فلسفی و مفهومی[2] برای بررسی پیامدهای باربرداری شناختی به هوش مصنوعی و تأثیر آن در عاملیت اخلاقی انسان بهره می‌گیرد و در پی تحلیل تغییرات ساختاری زیربنایی است که این رویه ممکن است در اصل ساختار عاملیت اخلاقی ایجاد کند. ازاین‌رو، روش تحقیق کیفی و استدلالی است و بر ایضاح مفاهیم، برساختن استدلال‌های منطقی و ارزیابی انتقادی دلالت‌های هنجاری تمرکز دارد (Daly, 2010).

فرایند پژوهش در سه مرحلهٔ متمایز ساماندهی و پی ریزی شده است: مرحلهٔ نخست شامل شفاف‌سازی دقیق مفاهیم اصلی مطالعه با بهره‌گیری از ادبیات بنیادین این حوزه است. مفهوم باربرداری شناختی براساس آثار منتشرشده در فلسفهٔ ذهن و علوم شناختی، به‌ویژه ایدهٔ ذهن بسط‌یافتهٔ[3] کلارک و چالمرز (Clark and Chalmers, 1998) و بسط‌های بعدی آن توسط پژوهشگرانی چون ریسکو و گیلبرت (Risko and Gilbert, 2016)، تعریف و زمینه‌مند می‌شود. هم‌زمان، مفهوم چندوجهی عاملیت اخلاقی به مؤلفه‌های فلسفی سازندهٔ آن واکاوی می‌شود. این واکاوی مبتنی بر مروری هدفمند بر نظریه‌های اخلاقی معیار، عمدتاً اخلاق فضیلت ارسطو با تأکید بر فرونسیس یا حکمت عملی، و اخلاق وظیفه‌گرای ایمانوئل کانت با محوریت اصل خودمختاری است (Kant, 1785/2012; Aristotle, 2009). مرحلهٔ دوم، مفاهیم ایضاح‌شده برای برساختن استدلال مرکزی مقاله با یکدیگر تلفیق می‌شود. این امر شامل تطبیق مکانیسم‌های خاص باربرداری شناختی با مؤلفه‌های شناسایی‌شده عاملیت اخلاقی است. تحلیل از توصیف صرف فراتر می‌رود تا پیوندهای منطقی میان عمل برون‌سپاری تأمل به هوش مصنوعی و اثرات بالقوهٔ آن بر پرورش حکمت عملی، اعمال خودمختاری و انتساب مسئولیت را ایجاد کند. این فرایند ترکیبی نشان می‌دهد که چگونه عمل به‌ظاهر خنثای مشورت با هوش مصنوعی در اصل می‌تواند به‌طور نظام‌مند در فرایندهای تکوینی لازم برای یک شخصیت اخلاقی قوام‌یافته اختلال ایجاد کند. مرحلهٔ نهایی، موضعی انتقادی اتخاذ می‌شود تا دلالت‌های اخلاقی و اجتماعی گسترده‌تر یافته‌ها ارزیابی شود. این امر مستلزم درگیری با ادبیات معاصر در فلسفهٔ فنّاوری و اخلاق هوش مصنوعی است. مفاهیم انتقادی کلیدی مانند مهارت‌زدایی اخلاقی (Vallor, 2016) و شکاف مسئولیت (Matthias, 2004; Sparrow, 2007) به‌عنوان چشم‌انداز تحلیلی برای قاب‌بندی پیامدهای هنجاری باربرداری شناختی گسترده در قلمرو اخلاق به کار گرفته می‌شوند. درگام نهایی پژوهش از تحلیل به نقد حرکت می‌کند و به ارزیابی هنجاری چالش‌های هوش مصنوعی برای آیندهٔ رشد اخلاقی انسان می‌پردازد.

برای بررسی پیامدهای باربرداری شناختی به کمک هوش مصنوعی، نقطهٔ تمرکز روش‌شناختی این مقاله بر منطق انتخاب چارچوب سه‌وجهی استوار است. در این راستا می‌توان گفت که یک عمل اخلاقی قوام‌یافته، به سه مؤلفهٔ بنیادین نیاز دارد: (۱) بصیرت در تشخیص و فرایند (که با حکمت عملی ارسطو نمایندگی می‌شود)، (۲) استقلال در وضع قانون و مبدأ عمل (که با خودمختاری کانت تبیین می‌شود)، و (۳) عاملیت در پذیرش تبعات (که در مفهوم شکاف مسئولیت برجسته می‌شود). پژوهش در سه مرحله پیش می‌رود: ابتدا حدود مفهوم ذهن بسط‌یافته در مواجهه با سیستم‌های فعال هوش مصنوعی تدقیق می‌شود؛ سپس مکانیسم باربرداری شناختی بر چارچوب سه‌وجهی یادشده تطبیق داده می‌شود؛ و در نهایت، با استفاده از یک سناریوی تحلیلی عینی، استدلال‌های مفهومی در بستر عمل آزموده و نقد هنجاری می‌شوند.

باربرداری شناختی و ذهن بسط‌یافتهٔ فعال

مفهوم بنیادین برای درک رابطهٔ انسان و فنّاوری در این بافتار، ایدهٔ ذهن بسط‌یافته است که کلارک و چالمرز (Clark and Chalmers,

1. Machine Ethics
2. a philosophical and conceptual analysis
3. extended mind

1998) آن را صورت‌بندی کرده‌اند. آنها استدلال می‌کنند که فرایندهای شناختی همیشه محبوس در جمجمه نیستند، بلکه می‌توانند به محیط بیرون گسترش یابند. ذهن و حالت مختلف آن، ازجمله باورها، فقط محصول سازوکارها و فرایندهای درونی مغز نیستند، بلکه به جهان بیرونی و به‌خصوص مصنوعات تکنیکی‌ای که به کار می‌گیریم عمیقاً وابسته‌اند (Mohammadamini, 2022, pp. 89-90).

هنگامی که یک شیء خارجی مانند یک دفترچه یادداشت با کاربر جفت می‌شود، می‌توان آن را بخشی از سیستم شناختی در نظر گرفت (نظریهٔ کلاسیک ذهن بسط‌یافته). بااین‌حال، تعمیم این نظریه به هوش مصنوعی مولد مستلزم احتیاط فلسفی است. برخلاف دفترچه یادداشت که صرفاً یک انبار ذخیرهٔ منفعل است، سیستم‌های هوش مصنوعی شرکای شناختی فعالی هستند که خروجی‌های بدیع تولید می‌کنند (Heersmink, 2017). ازسوی‌دیگر، ممکن است استدلال شود که مشورت با هوش مصنوعی شبیه به مشورت با انسان خردمند است و نباید آن را لزوماً جایگزین قوای تأملی دانست. اما تفاوت بنیادین در اینجاست: مشورت با یک انسان دیگر، تعامل دیالکتیکی و بیناسوبژکتیوی[1] است که هر دو طرف در افق اخلاقی مشترکی زیست می‌کنند و مسئولیت‌پذیرند. اما واگذاری این فرایند به هوش مصنوعی، تقلیل گفت‌وگوی اخلاقی به پردازش محاسباتی تک‌سویه و غیرشفاف است. در اینجا، هوش مصنوعی نه به‌مثابهٔ مشاور، بلکه همچون میان‌بُر شناختی عمل می‌کند که نیاز عامل اخلاقی به درگیری فعال درونی را دور می‌زند.

چارچوب سه‌وجهی عاملیت اخلاقی

عاملیت اخلاقی سازه‌ای پیچیده است. برای تحلیل چگونگی تأثیر هوش مصنوعی بر آن، باید آن را به مؤلفه‌های اصلی و دارای تمایز کارکردی‌اش تشریح کنیم. پژوهش حاضر چارچوبی سه‌وجهی را اتخاذ می‌کند و عاملیت اخلاقی را کارکرد یکپارچهٔ حکمت عملی، خودمختاری و مسئولیت در نظر می‌گیرد. انتخاب این چارچوب پاسخی به ساختار سه‌گانه (فاعل_ فعل_ نتیجه) در فلسفهٔ اخلاق است. حکمت عملی بر قابلیت‌های درونی فاعل و رشد شخصیت او متمرکز است؛ خودمختاری بر فرایند تصمیم‌گیری و صیانت از ارادهٔ آزاد تاکید دارد و مسئولیت بر پیوند میان عمل و پیامد در عرصهٔ عمومی نظارت می‌کند. ترکیب این سه، مدلی جامع‌نگر برای ارزیابی عاملیت فراهم می‌کند که فراتر از رویکردهای تک‌بعدی است.

۱. حکمت عملی (فرونسیس): فضیلت تشخیص اخلاقی

با بهره‌گیری از اخلاق فضیلت ارسطویی، نخستین رکن عاملیت اخلاقی، فرونسیس است که غالباً به حکمت عملی یا تدبیر ترجمه می‌شود (ارسطو، اخلاق نیکوماخوس، کتاب ششم). فرونسیس هم از دانش نظری (اپیستمه)[2] و هم از مهارت فنی (تخنه)[3] متمایز است. از نظر ارسطو فرونسیس معرفتی برای رسیدن به غایت از طریق فضایل اخلاقی و عقلانی است. فرونسیس معرفتی است که توانایی تعیین خوب و بد را چه در حوزهٔ فردی و چه در حوزهٔ اجتماعی به انسان می‌بخشد. او همچنین فرونسیس را فضیلتی لحاظ می‌کند که انسان را قادر می‌سازد تا از شیوه‌های درست و مناسبی برای رسیدن به غایت مطلوب استفاده کند (Samadieh & Mollayousefi, 2017, p. 10). این فضیلت، صرفاً پیروی مکانیکی از مجموعه‌ای از قواعد عمومی ازپیش‌تعیین‌شده نیست، بلکه توانایی درک شهودی و غیرالگوریتمی است که به بافتار و جزئیات خاص یک زمینه[4] به‌شدت وابسته است. به عبارت دیگر، حکمت عملی به عامل اخلاقی امکان می‌دهد تا در میان انبوهی از اطلاعات یک موقعیت پیچیده و بی‌نظیر، فوراً تشخیص دهد که کدام ویژگی‌ها از نظر اخلاقی برجسته و تعیین‌کننده‌اند (مثلاً، تشخیص اینکه در یک شرایط خاص انسانی، نشان دادن همدردی مهم‌تر از اجرای خشک قانون است). این درک موقعیتی، امکان یک تأمل مؤثر را به‌سوی تصمیمی فضیلت‌مندانه فراهم می‌کند.

طبق نظر ارسطو، این فضیلت را نمی‌توان از روی کتاب آموخت، بلکه فقط از طریق تجربهٔ زیسته، خوگیری عاطفی و تمرین پرورش می‌یابد. این همان قوه‌ای است که به عامل امکان می‌دهد تا در نواحی خاکستری زندگی اخلاقی، جایی که قوانین ساده کفایت نمی‌کنند، مسیر خود را بیابد. همان‌طور که شنون والور[5] (۲۰۱۶) در اثر خود پیرامون فضایل تکنو- اخلاقی استدلال می‌کند، توسعهٔ فرونسیس فرایندی بدنمند[6] و موقعیت‌مند[7] است که ذاتاً به کشمکش و اصطکاک حل مسائل اخلاقی در جهان واقعی گره خورده است.

۲. خودمختاری: اصل خودقانون‌گذاری

رکن دوم خودمختاری است؛ مفهومی که محور اخلاق وظیفه‌گرای ایمانوئل کانت (۱۷۸۵/۲۰۱۲) محسوب می‌شود. برای کانت، ارزش اخلاقی یک عمل ناشی از ظرفیت عامل برای عمل کردن براساس قانونی است که خود او به‌صورت عقلانی بر خویشتن وضع می‌کند. یک عامل خودمختار، خودقانون‌گذار است و از سر وظیفه‌ای عمل می‌کند که عقل تعیین کرده است، نه متأثر از تأثیرات بیرونی همچون میل، اجبار یا اقتدار بیرونی.

1. Intersubjective

2. episteme
3. techne
4. context
5. Shannon Vallor
6. embodied
7. situated

بر اساس خودمختاری، ارادهٔ انسان آزادانه خودش را از درون مجبور می‌کند و این خودمختاری، همان قوه‌ای است که قوانین را به‌صورت وظیفه درمی‌آورد. از دید کانت ما زمانی فردی را خودمختار و آزاد می‌انگاریم که صرفاً با ارادهٔ خود نه با ارادهٔ دیگری ملزم شده باشد و افعال وی بیانگر ارادهٔ او باشد، نه اینکه ارادهٔ او توسط شخص، ماشین یا هر مرجع بیرونی دیگری تعیین و دیکته شود. بنابراین، زمانی که فردی براساس اصل خودمختاری تصمیم می‌گیرد، این اطمینان حاصل می‌شود که منشأ و خاستگاه قوانینی که رفتار او را هدایت و حتی محدود می‌کنند، در درون ارادهٔ عقلانی خود او نهفته است. او در واقع از قانونی تبعیت می‌کند که خود به‌عنوان یک فاعل خردمند برای خویشتن وضع کرده است. در نتیجه وقتی خودمختاری اراده برای فردی اعمال می‌شود، تضمین می‌کند که بیان و منشأ اصولی که او را محدود می‌کند در ارادهٔ وی نهفته است. در نتیجه مشروعیت اخلاقی مبتنی بر وجود موجودی است که بر ارادهٔ عقلانی هر فرد مبتنی است (Khazaei & Heidari, 2017, p. 58). نقطه مقابل خودمختاری، ناخودمختاری است، جایی که اراده توسط یک منبع بیرونی تعیین می‌شود. در درون این چارچوب، یک تصمیم اخلاقی اصیل صرفاً یافتن پاسخ صحیح نیست، بلکه تأیید عقلانی اصلی است که در پس آن پاسخ قرار دارد. تکیه بر یک موجودیت بیرونی مانند هوش مصنوعی برای دریافت دستورالعمل‌های اخلاقی، یک مورد پارادایمی از ناخودمختاری است که کانون اقتدار اخلاقی را از ارادهٔ درونی عامل به منبع بیرونی و غیرعقلانی منتقل می‌کند. ممکن است استدلال شود که اگر هوش مصنوعی مبانی و استدلال‌های توصیهٔ اخلاقی خود را نیز به کاربر ارائه دهد، خطر اطاعت کورکورانه رفع می‌شود. باوجوداین، مطالعات روان‌شناختی نشان می‌دهد که پدیدهٔ سوگیری اتوماسیون[1] باعث می‌شود کاربران حتی زمانی که استدلال ماشین در دسترس است، به‌دلیل تمایل شناختی به کاهش بار ذهنی، به خروجی الگوریتم به‌عنوان یک مرجع نهایی و بدون نقد واقعی تکیه کنند (Skitka et al., 2000). در این حالت، کاربر صرفاً استدلال ماشین را تأیید می‌کند، نه اینکه آن را از درون اراده خویش وضع نماید.

۳. مسئولیت: شرط پاسخگویی اخلاقی

رکن نهایی، مسئولیت است. یک موجود تنها زمانی واجد شرایط عامل اخلاقی بودن است که بتوان او را در قبال اعمالش پاسخگو دانست. پاسخگویی مستلزم وجود پیوند روشن علّی و قصدمندانه[2] میان عامل و نتیجهٔ اعمال اوست. اخلاق هوش مصنوعی معاصر بررسی کرده است که چگونه سیستم‌های خودمختار پیشرفته این پیوند را به چالش می‌کشند و به پدیده‌ای منجر می‌شوند که شکاف مسئولیت نامیده شده است (Matthias, 2004). این شکاف زمانی پدیدار می‌شود که یک سیستم خودمختار موجب آسیب می‌شود، اما انتساب منصفانهٔ تقصیر به کاربر (که ممکن است سازوکارهای درونی سیستم را درک نکند)، برنامه‌نویس (که نمی‌توانسته هر رخداد احتمالی را پیش‌بینی کند)، یا خود ماشین (که فاقد آگاهی و قصدمندی است) ناممکن باشد. عمل باربرداری شناختی به کمک هوش مصنوعی در یک بافتار اخلاقی، مستقیماً این مسئله را نشانه می‌رود و با مبهم ساختن خطوط پاسخگویی، یکی از شروط بنیادین عاملیت را تهدید می‌کند.

چارچوب سه‌وجهی پیش‌گفته به‌نوعی ابزارهای تحلیلی لازم را برای فراتر رفتن از ارزیابی ساده‌انگارانهٔ خوب در برابر بد در مورد هوش مصنوعی فراهم می‌کند. این چارچوب امکان تحلیلی دقیق و ساختارمند را میسر می‌سازد تا دریابیم که چگونه رویهٔ خاص باربرداری شناختی می‌تواند به‌طور نظام‌مند بر هریک از قوای متمایز و درعین‌حال به هم‌پیوسته‌ای تأثیر بگذارد که در کنار هم، یک عامل اخلاقی شکوفای انسانی را تشکیل می‌دهند.

یافته‌های پژوهش

بهره‌گیری از هوش مصنوعی در حلقهٔ تصمیم‌گیری اخلاقی، نه ارتقای کارکردی خنثی، بلکه فرایندی دگرگون‌ساز است که سه رکن عاملیت اخلاقی را به‌طور ساختاری دچار فرسایش می‌کند. این رویه، خطر افتادن در دام ابزارانگاری فنّاورانه را همراه دارد؛ رویکردی که مسائل پیچیدهٔ اخلاقی و هنجاری را صرفاً معضلاتی فنی می‌پندارد که با الگوریتم‌ها قابل حل‌اند (Morozov, 2013). برای درک این فرسایش ساختاری، سناریوی تحلیلی زیر را در نظر بگیرید: پزشکی متخصص در بخش مراقبت‌های ویژه (ICU) با بیمار بدحالی مواجه است که شانس بهبود پایینی دارد. بیمارستان به سیستم هوش مصنوعی پیشرفته‌ای مجهز است که با تحلیل میلیون‌ها پروندهٔ پزشکی، به‌طور قطع توصیه می‌کند که دستگاه‌های حمایت حیاتی بیمار قطع شوند تا منابع به بیماران دیگر اختصاص یابد. پزشک به‌جای درگیری عاطفی با خانوادهٔ بیمار، سنجش ارزش‌های فرهنگی و تحمل رنج تصمیم‌گیری، صرفاً به خروجی الگوریتم اعتماد می‌کند و دکمهٔ تأیید را می‌فشارد. این سناریو صرفاً یک داستان علمی— تخیلی نیست، بلکه نمایی از آیندهٔ نزدیک تصمیم‌گیری‌های واسپاری‌شده است. یافته‌ها در قالب سه پیامد اصلی دسته‌بندی می‌شوند: زوال حکمت عملی (مهارت‌زدایی اخلاقی)، تسلیم خودمختاری (ناخودمختاری الگوریتمی)، و ابهام در پاسخگویی (شکاف مسئولیت).

پیامد اول: زوال حکمت عملی (مهارت‌زدایی اخلاقی)

باربرداری شناختی به کمک هوش مصنوعی، چرخهٔ بازخوردی ضروری برای پرورش فرونسیس (حکمت عملی) را مختل می‌سازد.

1. Automation Bias
2. Intetional

همان‌گونه که در اخلاق ارسطویی تبیین شده است، فضیلت اخلاقی مشابه یک مهارت (تخنه) است که مستلزم تمرین مکرر، شکست و کشمکش درونی تأمل است (ارسطو، اخلاق نیکوماخوس). بااین‌حال، کارکرد سیستم‌های هوش مصنوعی بر پایهٔ حذف اصطکاک استوار است؛ هوش مصنوعی پاسخ‌ها را بدون ملزم کردن کاربر به پیمودن مسیر شناختی دشوار فرایند ارائه می‌دهند.

با برون‌سپاری فرایند تأمل به الگوریتم هوش مصنوعی، عامل اخلاقی عملاً دور زده می‌شود. عامل، خروجی (قضاوت اخلاقی) را دریافت می‌کند، بدون آنکه در فعالیت‌های بنیادین وزن‌دهی به ارزش‌ها، تفسیر بافتار و احساس کردن بار عاطفی تصمیم درگیر شود. والور (Vallor, 2016) در مقاله‌ای با عنوان «مهارت‌زدایی و مهارت‌افزایی اخلاقی»، این پدیده را مهارت‌زدایی اخلاقی[1] می‌نامد. همان‌طور که استفادهٔ مداوم از مسیریاب‌ها می‌تواند به‌تدریج مهارت‌های جهت‌یابی فضایی انسان را تضعیف کند، اتکای عادت‌گونه به هوش مصنوعی نیز عامل انسانی را از تمرین مداوم برای پرورش حکمت عملی باز می‌دارد. در نتیجه، اگرچه کاربر متکی به ماشین ممکن است تصمیمات درستی بگیرد، اما مسیر ارتقای او به‌سمت خبرگی اخلاقی که دریفوس و دریفوس (Dreyfus and Dreyfus, 1986) آن را نیازمند درگیری عمیق و شهودی می‌دانند با کندی و اختلال مواجه می‌شود.

پیامد دوم: تسلیم خودمختاری (ناخودمختاری الگوریتمی)

در چارچوب کانتی، ارزش اخلاقی یک عمل وابسته به تأیید عقلانی ضابطهٔ[2] نهفته در پس آن عمل توسط عامل است (Kant, 1785/2012). بااین‌حال، سیستم‌های هوش مصنوعی، به‌ویژه آنهایی که مبتنی بر یادگیری عمیق (شبکه‌های عصبی) هستند، غالباً به‌مثابهٔ جعبه‌های سیاه[3] عمل می‌کنند؛ جایی که فرایند استدلال حتی برای توسعه‌دهندگان نیز غیرشفاف است (Pasquale, 2015). هنگامی که یک عامل براساس توصیهٔ هوش مصنوعی عمل می‌کند، بی‌آنکه منطق زیربنایی آن را درک کند، نمی‌تواند ضابطهٔ عمل خود را به‌طور عقلانی تأیید نماید. او به‌جای قانون‌گذاری برای خویشتن، در حال اطاعت از یک اقتدار مرموز و غیرقابل فهم است. این امر رابطه‌ای از وابستگی ایجاد می‌کند که شبیه به پدرسالاری الگوریتمی[4] (Danaher, 2016) است؛ جایی که فنّاوری صرفاً به کاربر کمک نمی‌کند، بلکه انتخاب‌ها را براساس توابع مطلوبیت ازپیش‌برنامه‌ریزی‌شده‌ای که ممکن است با ارزش‌های اصیل کاربر همسو نباشند، به شکل نامحسوس هدایت یا دیکته می‌کند. این وابستگی، شرط خودحکمرانی لازم برای عاملیت اخلاقی اصیل را به‌طور بنیادین مخدوش می‌سازد و این تغییر حیاتی از خودمختاری به وضعیتی از ناخودمختاری الگوریتمی است.

پیامد سوم: ابهام در پاسخگویی (شکاف مسئولیت روان‌شناختی)

درحالی‌که شکاف مسئولیت غالباً در بافتارهای حقوقی بررسی می‌شود (Matthias, 2004)، این تحلیل نشان می‌دهد که باربرداری شناختی یک شکاف مسئولیت روان‌شناختی ایجاد می‌کند.

زمانی که فرد تصمیمی را به هوش مصنوعی واگذار می‌کند، از نظر روان‌شناختی خود را از نتیجه فاصله می‌دهد. مطالعات تجربی در حوزهٔ تعامل انسان و رایانه نشان می‌دهند که کاربران مستعد سوگیری اتوماسیون هستند؛ به این معنا که تمایل دارند به خروجی‌های الگوریتمی بیش از قضاوت خود اعتماد کنند و نکتهٔ حیاتی اینکه، در صورت بروز مشکل، تقصیر را به سیستم نسبت دهند (Skitka et al., 2000). هوش مصنوعی همچون یک حائل یا ضربه‌گیر اخلاقی عمل می‌کند. اگر تصمیمی به پیامدی زیان‌بار منجر شود، عامل می‌تواند ادعا کند که الگوریتم آن را توصیه کرد. این قابلیت منحرف کردن کانون علیت، شرط پاسخگویی لازم برای عاملیت اخلاقی را تضعیف می‌کند. عامل به مجری صرف ارادهٔ ماشین تبدیل می‌شود و پیوند حیاتی میان کنشگر و بار اخلاقی اعمالش قطع می‌گردد (Coeckelbergh, 2020). به‌طور خلاصه، آسایش حاصل از باربرداری شناختی با پرسش‌های فلسفی مهمی همراه است. با برون‌سپاری رنج اخلاق، ما ریسک پرورش نسلی از عاملان اخلاقی را می‌پذیریم که از نظر فنی کارآمد اما از نظر عملی فاقد حکمت و منفک از مسئولیت انتخاب‌های خویش‌اند.

ملاحظه‌ای انتقادی: آیا هوش مصنوعی ارتقادهندهٔ اخلاقی نیست؟

در برابر استدلال‌های انتقادی این پژوهش، مدافعان توسعهٔ هوش مصنوعی ممکن است استدلال کنند که این سیستم‌ها در واقع به‌مثابهٔ ابزارهای ارتقای اخلاقی[5] عمل می‌کنند. برخی اندیشمندان (Giubilini & Savulescu, 2018) استدلال می‌کنند که انسان‌ها مستعد خستگی شناختی، سوگیری‌های احساسی، تعصبات نژادی و خطاهای محاسباتی هستند. از این منظر، واگذاری قضاوت اخلاقی به یک ماشین بی‌طرف، می‌تواند به خروجی‌های اخلاقی‌تر (به‌ویژه از منظر فایده‌گرایی) منجر شود.

استدلال این پژوهش در پاسخ به این دیدگاه آن است که مدافعان هوش مصنوعی، خطای بنیادین خلط میان نتیجهٔ عمل و فاعل عمل را مرتکب می‌شوند. حتی اگر بپذیریم که یک الگوریتم می‌تواند

1. Moral Deskilling
2. Maxim
3. black boxes
4. algorithmic paternalism
5. Moral Enhancement

خروجی بهینه‌تری تولید کند، اخلاق (در سنت ارسطویی و کانتی) صرفاً بهینه‌سازی پیامدها نیست. اخلاق یک تمرین تکوینی است که در آن، رنج تصمیم‌گیری، تجربهٔ تردید و درگیری عاطفی، شخصیت فاعل را می‌سازد. ماشینی که فرایند تأمل را دور می‌زند، ممکن است عمل درست را پیشنهاد دهد اما انسان را از تبدیل شدن به یک انسان خوب و بافضیلت باز می‌دارد. بهبودی آماری نتایج، بهای سنگین تهی‌سازی سوژهٔ اخلاقی را توجیه نمی‌کند. منتقدان ممکن است استدلال کنند که هوش مصنوعی با حذف خطاهای انسانی به‌نوعی موجب ارتقای اخلاقی می‌شود. اما این دیدگاه دچار یک خلط مفهومی است: ارتقای نتیجهٔ عمل با ارتقای خود عامل اخلاقی یکسان نیست. حتی اگر ماشین تصمیمی بهینه بگیرد، این تصمیم فاقد ارزش اخلاقی است، زیرا اخلاق مستلزم آگاهی، نیت و فهم ارزش‌هاست. هوش مصنوعی صرفاً یک محاسبه‌گر است، نه یک عامل؛ بنابراین، بهبود نتایج توسط ماشین، به‌معنای رشد اخلاقی انسان نیست، بلکه صرفاً جایگزینی قضاوت انسانی با محاسبات آماری است (Danaher, 2016).

بحث و نتیجه‌گیری

پژوهش حاضر نشان داد که تلفیق غیرانتقادی سیستم‌های فعال هوش مصنوعی در زیست هنجاری انسان، صرفاً امتداد بی‌خطر ذهن نیست، بلکه پدیده‌ای است که ساختار بنیادین عاملیت اخلاقی را با بحران مواجه می‌سازد. واسپاری تأملات اخلاقی به الگوریتم‌ها، از طریق حذف اصطکاک شناختی در اصل چرخهٔ پرورش حکمت عملی را مختل می‌کند و به مهارت‌زدایی اخلاقی می‌انجامد. همچنین، با جایگزینی ارادهٔ عقلانی فرد با خروجی‌های غیرشفاف ماشین، کانون اقتدار را به بیرون منتقل می‌کند و وضعیتی از ناخودمختاری و بیگانگی الگوریتمی رقم می‌زند. در نهایت، با مبهم ساختن زنجیرهٔ علیت میان عمل و پیامد، یک شکاف مسئولیت روان‌شناختی ایجاد می‌کند که در آن انسان‌ها نقش مجری منفعل ماشین را می‌پذیرند.

پدیدهٔ باربرداری شناختی، اگرچه از نظر کارکردی برای پردازش داده‌ها و محاسبات منطقی کارآمد است، اما هنگامی که در قلمرو هنجاری اخلاق به کار گرفته می‌شود، یک خطر وجودی عمیق به همراه دارد. تنش محوری شناسایی‌شده، میان کارآمدی و شخصیت است. طراحی فنّاورانه معاصر غالباً بر کاهش اصطکاک شناختی، آسان‌تر و سریع‌تر کردن وظایف اولویت می‌دهد (Morozov, 2013). بااین‌حال، باید گفت که اصطکاک اخلاقی یعنی دشواری، تردید و بار عاطفی تصمیم‌گیری نه یک ایراد که باید حذف شود، بلکه ویژگی ضروری برای قوام‌یافتن خود اخلاقی است.

با برون‌سپاری تأمل اخلاقی به هوش مصنوعی، ما خطر افتادن در دام ابزارگرایی فنّاورانه را به جان می‌خریم؛ باوری که معتقد است مشکلات پیچیده اخلاقی و اجتماعی را می‌توان از طریق محاسبات الگوریتمی حل کرد (Morozov, 2013). این دیدگاه به اشتباه قضاوت اخلاقی را با پردازش اطلاعات یکسان می‌انگارد. همان‌طور که وایزنبام (Weizenbaum, 1976) در روزهای آغازین محاسبات رایانه‌ای به‌درستی هشدار داد، تمایزی بنیادین میان تصمیم‌گیری (یک عمل محاسباتی مبتنی بر معیارها) و انتخاب‌گری (یک عمل انسانی مبتنی بر ارزش‌ها و مسئولیت) وجود دارد. باربرداری شناختی هنجاری به هوش مصنوعی، تمایل دارد که این تمایز بنیادین را فرو بریزد. سیستم‌های هوش مصنوعی ذاتاً براساس الگوریتم‌های ریاضی کار می‌کنند که هدفشان بهینه‌سازی (بیشینه‌سازی یک متغیر مطلوب یا کمینه‌سازی خطا) براساس داده‌های کمّی است. هنگامی که انسان یک دوراهی پیچیده و کیفی اخلاقی را به چنین سیستمی واگذار می‌کند، در واقع در حال تقلیل دادن ارزش‌های متعارض انسانی، درگیری‌های عاطفی و ابهامات موقعیتی، به چند متغیر قابل محاسبه است. در این فرایند تقلیل‌گرایانه، با انتخاب‌های تراژیک و پیچیدهٔ اخلاقی همچون مسائل صرف مهندسی و بهینه‌سازی برخورد می‌شود که گویی تنها یک پاسخ قطعی دارند و صرفاً باید توسط یک پردازشگر برتر محاسبه و حل شوند.

علاوه بر این، حرکت به‌سوی ناخودمختاری الگوریتمی، ایده‌آل روشنگری در باب سوژهٔ عقلانی را به چالش می‌کشد. اگر ذهن بسط‌یافته شامل مؤلفه‌ای از هوش مصنوعی باشد که به‌صورت غیرشفاف عمل می‌کند (مسئلهٔ جعبهٔ سیاه)، عامل گسترده دیگر نمی‌تواند به‌معنای کانتی کلمه، مدعی مؤلف بودن زندگی خویش باشد. این امر نشان می‌دهد که ادغام غیرانتقادی هوش مصنوعی در زیست اخلاقی ما، صرفاً مصداقی از بسط و تقویت ظرفیت‌های شناختی (آن‌گونه که در نظریه‌های کلاسیک ذهن بسط‌یافته دربارهٔ ابزارهای منفعل مطرح می‌شد) نیست، بلکه ظرفیت آن را دارد که با غصب جایگاه ارادهٔ انسانی، درون فاعل اخلاقی را تهی سازد و او را به دریافت‌کنندهٔ منفعل هنجارهای تولیدشده توسط الگوریتم بدل کند.

پژوهش حاضر از طریق تحلیل فلسفی حکمت عملی، خودمختاری و مسئولیت نتیجه می‌گیرد که برون‌سپاری عادت‌گونه تصمیم‌گیری اخلاقی به سیستم‌های هوش مصنوعی، تهدیدی ساختاری برای یکپارچگی عامل اخلاقی ایجاد می‌کند. عاملیت اخلاقی یک ویژگی ایستا نیست، بلکه تمرینی پویا و وابسته به درگیری فعال عامل است. اخلاق میانجی‌گری‌شده توسط هوش مصنوعی، با حذف ضرورت تأمل درونی، به زوال فرونسیس (حکمت عملی) منجر می‌شود، حالتی از وابستگی (دگرآیینی) را پرورش می‌دهد و خطوط پاسخگویی را مبهم می‌سازد (شکاف مسئولیت). اگرچه هوش مصنوعی پتانسیل عظیمی برای آگاه‌سازی قضاوت انسان از طریق فراهم آوردن داده‌ها و پیش‌بینی سناریوها دارد، اما اجازه دادن به آن برای جایگزینی قضاوت انسان، به‌مثابهٔ دست شستن از بنیادین‌ترین

ظرفیت انسانی ماست: توانایی تعریف امر خیر و انتخاب پیگیری آن. خطر این نیست که هوش مصنوعی شرور شود، بلکه خطر آنجاست که انسان‌ها به‌دلیل استفاده نکردن از قوای اخلاقی خود به‌نوعی از نظر اخلاقی تهی شوند.

بر اساس یافته‌های این مطالعه، پیشنهادهای زیر ضمن اذعان به اجتناب‌ناپذیر بودن ادغام هوش مصنوعی، برای کاهش خطرات ارائه می‌شود:

۱. نظام‌های آموزشی نباید صرفاً به آموزش نحوهٔ کار با ابزارهای دیجیتال و هوش مصنوعی بسنده کنند، بلکه توسعهٔ سواد انتقادی هوش مصنوعی را در دستور کار قرار دهند. منظور از سواد انتقادی، گنجاندن پودمان‌های اخلاقی در کتب درسی است؛ بخش‌هایی آموزشی که به افراد آموزش می‌دهند چگونه محدودیت‌های شناختی ماشین را بشناسند و میان یک مسئلهٔ محاسباتی (مانند حل معادله) و یک مسئلهٔ اخلاقی (مانند تصمیم‌گیری دربارهٔ تخصیص منابع درمانی) تمایز قائل شوند.

۲. طراحان به طراحی مبتنی بر اصطکاک تأملی توجه کنند. رویکردی در طراحی رابط کاربری است که به‌جای ارائهٔ سریع‌ترین و ساده‌ترین پاسخ، مکث و درگیری ذهنی کاربر را تحریک می‌کند. به‌طور مثال، اگر کاربری از هوش مصنوعی در یک مسئلهٔ حساس انسانی مشورت خواست، سیستم به‌جای صدور دستورالعمل قطعی، خروجی خود را در قالب پرسش‌های سقراطی ارائه دهد (مثلاً اگر این تصمیم را بگیری، چه تأثیری بر حقوق اقلیت‌ها خواهد داشت؟). این طراحی مانع از اتکای منفعلانه می‌شود و عامل انسانی را در جایگاه هدایت‌گر تأمل حفظ می‌کند.

References

Aristotle. (2009). *The Nicomachean ethics* (D. Ross, English Trans.). Oxford: Oxford University Press. (Original work published in Ca. 350 B.C.E.)

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7-19. https://doi.org/10.1093/analys/58.1.7

Coeckelbergh, M. (2020). *AI Ethics*. Cambridge: The MIT Press.

Daly, C. (2010). *An introduction to philosophical methods*. Peterborough: Broadview Press.

Danaher, J. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology*, *29*(3), 245-268. https://doi.org/10.1007/s13347-015-0211-1

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer*. New York City: Free Press.

Giubilini, A., & Savulescu, J. (2018). The artificial moral advisor. *Philosophy & Technology*, 31(2), 169-188. https://doi.org/10.1007/s13347-017-0285-z

Heersmink, R. (2017). Distributed cognition and distributed morality: Agency, artifacts and systems. *Science and Engineering Ethics*, *23*(2), 431-448. https://doi.org/10.1007/s11948-016-9802-1

Kant, I. (2012). *Groundwork of the metaphysics of morals* (C. M. Mary Gregor & J. Timmermann, English Trans.). Cambridge: Cambridge University Press. (Original work published in 1785). https://doi.org/10.1017/CBO9780511973741

Khazaei, Z., & Heidari, T. (2017). Reassessing Kant's autonomy in relation to individual, moral, and political autonomy. *Journal of Philosophical Theological Research*, *19*(2), 47-67. (Persian) https://doi.org/10.22091/pfk.2017.2305.1701

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175-183. https://doi.org/10.1007/s10676-004-3422-1

Mohammadamini,M. (2022). Studying the experience of visual media use by extended Mind Theory. *Journal of Culture-Communication Studies*, *23*(58), 81-104. (Persian) https://doi.org/ 10.22083/jccs.2021.250319.3183

Morozov, E. (2013). *To save everything, click here: The folly of technological solutionism*. New York City: PublicAffairs.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Cambridge: Harvard University Press.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, *20*(9), 676-688. https://doi.orgorg/10.1016/j.tics.2016.07.002

Samadieh, M., & Mollayousefi, M. (2017). Aristotle and Phronesis (practical wisdom). Ayeneh Ma'refat: The *Journal of Islamic Philosophy and Theology*, *17*(4), 1-24. (Persian)

Skitka, L. J., Mosier, K. L., & Burdick, M. (2000). Does automation bias decision-making? *International Journal of Human-Computer Studies*, *51*(5), 991-1006. https://doi.org/10.1006/ijhc.1999.0252

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, *24*(1), 62-77.

https://doi.org/10.1111/j.1468-5930.2007.00346.x

Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy & Technology*, *28*(1), 107-124. https://doi.org/10.1007/s13347-014-0156-9

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. San Francisco: W. H. Freeman & Co.

حمید احمدی هدایت

دانش‌آموختهٔ دکتری فلسفه تعلیم و تربیت از دانشگاه شاهد تهران و هم‌اکنون استادیار گروه علوم تربیتی دانشگاه فرهنگیان است. فلسفه فنّاوری آموزشی، به‌ویژه موضوعات مرتبط با هوش مصنوعی و آموزش از علایق پژوهی ایشان است و آثار، کارگاه‌ها و سخنرانی‌های متعددی در این زمینه ارائه کرده است.


رهنت