Philosophy of Science, Institute for Humanities and Cultural Studies (IHCS)
Biannual Journal, Vol. 15, No. 2, Autumn and Winter 2025-2026, 285-319
https://www.doi.org/10.30465/ps.2026.54442.1827

Responsibility Gap in AI:

Revising Social Roles as the Basic Solution

Zahra Zargar*

Saeedeh Babaii**

Abstract

About two decades ago, "The Gap of Responsibility" as a problem was first introduced by Matthias to refer to an ethical challenge of AI. Due to their ability to learn, AI technologies are able to function beyond the control and anticipation of designers and users. Matthias argues that in this case, the necessary and sufficient conditions of responsibility would not be satisfied either for designers or for users, and that is the responsibility gap. The problem of the responsibility gap has invited many researchers to explore solutions. There is significant divergence among current accounts of the responsibility gap, which mainly results from their basic theoretical assumptions. In this paper, we categorize the responses to three approaches: dissolving the problem, filling the gap by attributing responsibility to AI, and filling the gap by attributing responsibility to human agents. By analyzing the social and practical aspects of the concept of responsibility, we defend a compound, contextual, and gradual notion of responsibility. On this basis, we claim that the first and second approaches are not successful, due to being grounded on poor concepts of responsibility. Hence, the third approach is more plausible. Moreover, since in the familiar cases of responsibility gap we fill the gap through revising social roles, in the case of AI-based responsibility gap we can follow a

* Assistant Professor, Department of History & philosophy of science, Institute for Science and Technology Studies, Shahid Beheshti University, Tehran, Iran (Corresponding Author), Z_zargar@sbu.ac.ir

** PhD, Philosophy of Science and Technology, Institute for Humanities and Cultural Studies, Tehran, Iran & Administrator of Ethics Committee in Fahm Research Group, Saeedeh.babaii@gmail.com

Date received: 04/04/2026, Date of acceptance: 05/09/2026

similar solution too, as suggested in some works of the third approach. Finally, we discuss some of remained challenges for the third approach.

Keywords: Responsibility, The Responsibility Gap, AI Ethics, AI Agency, Social Roles.

Introduction

In his paper "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata" Matthias introduces the concept of the responsibility gap. He starts with a definition of responsibility and its necessary and sufficient conditions and proceeds to show how the development of AI can pose challenges for attribution of responsibility. Matthias asserts that we can fairly attribute responsibility to someone only when she has control over her behaviors and their consequences in a proper sense. Moreover, one should know certain relevant facts about her actions, and freely choose one way of acting among many. In usual cases of employing machines, if the operator uses the machine according to the manufacturer's manual, the manufacturer is responsible for the consequences, and otherwise, the operator is (Matthias, 2004: 175). But for the learning machines, this is not the case. Matthias examines four kinds of AI design, including symbolic systems, connectionism, genetic algorithms. and autonomous agents. For each case, he demonstrates that due to their learning capacities, the conditions of responsibility are not satisfied, neither for manufacturers nor for operators. In fact, he claims that these machines are so opaque and complex that no human meets the necessary conditions for being responsible for the consequences of their functions. And this is the gap of responsibility.

Materials & methods

In the last two decades, many works have been dedicated to the AI responsibility gap and plausible answers for it. There is significant divergence among these accounts, which mainly comes from their basic theoretical assumptions. In this paper, we categorize the responses to three main approaches: dissolving the problem, filling the gap by attributing responsibility to AI, and filling the gap by attributing responsibility to human agents. By analyzing the social and practical aspects of the concept of responsibility, we defend a compound, contextual, and gradual notion of responsibility. On this basis, we defend the third approach. Moreover, we argue that in the case of the AI responsibility gap, like other

responsibility gaps in the history of our social life, revising social roles is the basic pattern for overcoming the gap.

Discussion & Result

- Dissolving the Problem

Some authors don't empathize with Matthias about the essence and importance of the responsibility gap. Kiener, for example, argues that the real problem about AI and responsibility isn't the gap of responsibility. Instead, it is the abundance of responsibility. Since every person who causally participates in hurting someone and has a moral obligation not to do so is responsible, these two conditions are satisfied by many (Kiener, 2025: 368).

On the other hand, Danaher and Munch et al. accept Matthias' articulation of the responsibility gap, but assert that not only isn't the gap evil, but it is even good. The main idea is that AI can perform tasks, thereby taking the burden of responsibility off humans' shoulders. This in turn decreases the psychological pressures of responsibility and provides more joy and comfort for human agents, which is morally good (Danaher, 2022; Munch et al., 2023).

- Attributing Responsibility to AI

A main line of responses to Matthias' challenge is to suggest reasons for filling the gap by attributing responsibility to AI. Here we have two approaches: property-oriented and relational. In the first approach, scholars introduce some properties (like autonomy, agency, etc.) as the basic conditions of responsibility, and then try to show that AI can actually or possibly meet these conditions (Haselager, 2005; Rodogno, 2016; List & Pettit, 2011; Floridi, 2016). The relational approach, instead, considers the status of AI in the network of our social relations. If we perceive an AI technology as if it is a responsible agent, and thereby make sense of its action, then we have good reasons to attribute responsibility to it (Sullins, 2006).

- Attributing Responsibility to Humans

For many scholars, attribution of responsibility to AI sounds implausible. On one hand, they make arguments to reject the responsibility of AI, and on the other hand, they develop accounts for preserving human responsibility in the age of AI technologies. Sparrow suggests Human-in-the-loop as a way of preventing the responsibility gap. He argues that the final action of AI should always be done only after the confirmation of a human agent (Sparrow, 2007). Also, Human-on-the-loop

emphasizes that a human agent can only supervise or veto the function of AI (Scharre, 2018). But due to different factors like automation bias and over-trust of humans in AI, it seems that in both approaches the human control of the output wouldn't be effective and real. As a response to these problems, meaningful human control (MHC) was suggested as a solution concentrating on designing the socio-technical network in a way that meaningful human control is preserved. Two conditions of tracking (designing the system for being reason-responsive and tracking humans' moral intents) and tracing (designing the system such that there is always a human who understands the system's function and its consequences) are assumed to serve this goal (Santoni de Sio & van den Hoven, 2018; Santoni de Sio and Mecacci, 2021, 1063–1076).

- Revising Social Roles as the Basic Solution

To find the best approach, it is proper to have a closer look at the concept of responsibility. As a social construct, this concept serves to reinforce people's will for promoting moral values by putting some pressure on them. To do this task in the best way, we need a rich concept of responsibility which, despite monist accounts that equate it with one single notion (i.e., answerability) or pluralist accounts that consider it as having multiple distinct types (i.e., accountability, culpability, etc.), assumes responsibility as a compound notion which is constituted by attributability, accountability and answerability. A compound, gradual, and contextual notion of responsibility aligns with our daily use of the word and also makes it adequate for performing its role in morality. This thick notion discredits the first and second approaches toward the responsibility gap, which respectively dissolve it and attribute responsibility to AI. Both approaches are based on a thin notion.

On the other hand, the responsibility gap has familiar examples in the history of our social life. Humans have been hurt by various natural forces, where *prima facie* no human agent was responsible for that. But social roles, as dynamic parts of social structure, make it possible to provide a human with the requirements of responsibility, i.e., the needed knowledge and power, to take responsibility and prevent those harms as much as possible. Hence, in the case of the AI responsibility gap, the practical and successful solution should follow this basic pattern too: revising social roles and their corresponding norms in order to prevent harms which are currently out of our control.

Conclusion

Contemplating the role of responsibility in our morality and our ordinary use of the term leads us to a compound, gradual, and contextual account of responsibility. This rich concept limits meaningful attribution of responsibility only to human agents. Hence, to overcome the gap, we should find ways for preserving human responsibility in AI systems. A familiar way of doing so is revising social roles, which includes defining new tasks, providing new resources of knowledge and power, and making responsibility attribution reasonable.

Bibliography

Asaro, P. (2012). "On banning autonomous weapon systems: human rights, automation, and the design of ethical algorithms." International Review of the Red Cross.

Aristotle. (1984). Nicomachean ethics. In J. Barnes (Ed.), The complete works of Aristotle (Vol. 2, pp. 1729─1867). Princeton University Press.

Bales, A. (2025). "Against willing servitude: Autonomy in the ethics of advanced artificial intelligence." The Philosophical Quarterly.

Bringsjord Selmer, (2008), Robots: The Future Can Heed Us, *AI and Society*. Vol. 22 No.4, 539-550.

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

Danaher, J. (2016). Robots, law and the retribution gap. Ethics and Information Technology, 18(4), 299-309 https://doi.org/10.1007/s10676-016-9403-3

Danaher, J. (2022). Tragic choices and the virtue of techno-responsibility gap. Philosophy and Technology, 35(26), 1-26

Davidson, D. (2001). Essays on actions and events. Oxford University Press.

De Cremer, D. (2024). The AI-savvy leader: 9 ways to take back control and make AI work. Harvard Business Review Press.

Dignum, V. (2022). Responsible artificial intelligence: How to develop and use AI in a responsible way. Springer Nature.

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. Engaging Science, Technology, and Society, 5, 40–60.

European Commission. (2024). Report on liability for artificial intelligence and other emerging digital technologies.

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of responsibility in distributed moral actions. Philosophical Transactions.

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

Gunkel, D. J. (2023). Person, thing, robot: A philosophy of responsible AI. MIT Press.

Haselager, W. F. (2005). Robotics, philosophy and the problems of autonomy. Pragmatics & Cognition, 515-532

Heath, J. (2021). The failure of traditional environmental philosophy. Res Publica, 28, 1–16.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?. Ethics and information technology, 11(1), 19-29.

Jeppsson, S. (2022). Accountability, answerability, attributability: On different kinds of moral responsibility. In Oxford handbook of moral responsibility (pp. 73-90)

Johnson, D. G., & Verdicchio, M. (2019). AI, agency and responsibility: The VW fraud case and beyond. AI & Society, 34(3), 639-647.

Johnson, D. G., & Verdicchio, M. (2023). Ethical AI is not about AI. Communications of the ACM, 66(2), 32-34.

Kiener, M. (2025). AI and responsibility: No gap, but abundance. Journal of Applied Philosophy, 42(1), 357-374

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. University of Pennsylvania Law Review, 165(3), 633-705.

Latour, B. (2005). Reassembling the social: An introduction to actor-network-theory. Oxford University Press.

List C. and Pettit P., (2011), *Group Agency: The Possibility, Design, and Status of Corporate Agents*, Oxford University Press.

Matthias, A, (2004), The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata, *Ethics and Information Technology*, p 175-183.

Munch, L., Mainz, J., & Bjerring, J. C. (2023). The value of responsible gaps in algorithmic decision-making. Ethics and Information Technology, 25(21), 1-11.

Nissenbaum, H. (1996). Accountability in a computerized society. Science and Engineering Ethics.

Rodogno R., (2016), Social robots, fiction, and sentimentality, *Ethics and Information Technology,* p. 257–268.

Roff, H. M. (2016). The strategic robot problem: Networked power and democratic accountability. Journal of Military Ethics.

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. Philosophy & Technology, 34(4), 1057-1084.

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical framework. Frontiers in ICT, 5, 15.

Scharre, P. (2018). Army of none: Autonomous weapons and the future of war. W. W. Norton & Company.

Sharkey, N. (2010). Saying no! to killer robots. Journal of Military Ethics, 9(4), 369-383.

Shoemaker, D. (2011). Attributability, answerability and accountability: Toward a wider theory of moral responsibility. Ethics, 121, 602-632.

Smith, A. M. (2015). Responsibility as answerability. Inquiry, 58(2), 99-126.

Smith, A. (2012). Attributability, answerability, and accountability: In defense of a unified account. Ethics, 122(3), 575-589

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

Strawson, P. F. (1974). Freedom and resentment. In Freedom and resentment and other essays (pp. 1─25). Oxford University Press.

Sullins, J. (2006). When is a robot a moral agent? International Review of Information Ethics, 6, 24-30

Tigard, W. D. (2021). There is no techno-responsibility gap. Philosophy and Technology, 34,. 589-607

Van Inwagen, P. (1983). An essay on free will. Oxford University Press.

Verbeek, P.-P. (2011). Moralizing technology: Understanding and designing the morality of things. University of Chicago Press.

Watson, G. C. (1996). Two faces of responsibility. Philosophical Topics, 24, 227-248.

فلسفهٔ علم، پژوهشگاه علوم انسانی و مطالعات فرهنگی
دوفصل‌نامهٔ علمی (مقالهٔ پژوهشی)، سال ۱۵، شمارهٔ ۲، پاییز و زمستان ۱۴۰۴

شکاف مسئولیت در هوش مصنوعی:
دفاعی از پاسخ‌های مبتنی بر بازآرایی نقش‌های اجتماعی

زهرا زرگر*
سعیده بابایی**

چکیده

مسئله «شکاف مسئولیت» نخستین بار حدود دو دهه پیش توسط ماتیاس و برای اشاره به چالشی اخلاقی در ارتباط با فناوری‌های هوش مصنوعی معرفی شد. قابلیت یادگیری این فناوری‌ها به آنها اجازه می‌دهد خارج از کنترل و پیش‌بینی طراحان و کاربران عمل کنند. ماتیاس معتقد است در چنین شرایطی، شروط لازم و کافی مسئولیت‌پذیری نه برای کاربران و نه طراحان برآورده نمی‌شود، و بر این اساس نمی‌توان هیچ انسانی را مسئول دانست. در نتیجه، ما با شکاف مسئولیت مواجهیم. مسئله شکاف مسئولیت توجه محققان بسیاری را در حوزه اخلاق فناوری به خود جلب کرده و آثار متعددی در ارتباط با آن منتشر شده است. مجموع راهکارهای ارائه‌شده درباره شکاف مسئولیت، رویکردهای واگرا و متفاوتی دارند و زمینه‌های این تفاوت عمدتاً به مفروضات نظری و اساسی آنها بازمی‌گردد. در مقاله حاضر سه رویکرد اصلی نسبت به مسئله شکاف مسئولیت شامل انحلال مسئله، انتساب مسئولیت به هوش مصنوعی، و انتساب مسئولیت به عامل انسانی، مورد بررسی قرار می‌گیرد. با تحلیل ابعاد اجتماعی و کارکردی مفهوم «مسئولیت» از تعبیری ترکیبی، زمینه‌مند و درجه‌مند از این مفهوم دفاع شده و بر این اساس، دو رویکرد نخست از آن جهت که بر تعبیر درستی

* استادیار، گروه تاریخ و فلسفهٔ علم، پژوهشکدهٔ مطالعات بنیادین علم و فناوری، دانشگاه شهید بهشتی، تهران، ایران (نویسندهٔ مسئول)، Z_zargar@sbu.ac.ir

** دکترای فلسفهٔ علم و فناوری، پژوهشگاه علوم انسانی و مطالعات فرهنگی، تهران، ایران، و مدیر کارگروه اخلاق، گروه پژوهشی فهم، Saeedeh.babaii@gmail.com

تاریخ دریافت: ۱۴۰۵/۰۱/۱۵، تاریخ پذیرش: ۱۴۰۵/۰۲/۱۹

از مسئله استوار نیستند، نقد می شوند. علاوه‌براین، ادعا می شود چنان که در موارد مشابه پیدایش شکاف مسئولیت، «بازآرایی نقش‌های اجتماعی» راهبرد کلی و انسان-محور برای حل مسئله است، در مورد شکاف مسئولیت ناشی از فناوری‌های هوش مصنوعی نیز راه حل اساسی از این جنس است. در نهایت برخی چالش‌های موجود در رویکرد سوم مورد بحث قرار می‌گیرد.

کلیدواژه‌ها: مسئولیت‌پذیری، شکاف مسئولیت، اخلاق هوش مصنوعی، عاملیت هوش مصنوعی، نقش‌های اجتماعی.

۱. مقدمه

مسئولیت‌پذیری (**Responsibility**) مفهومی اساسی در اخلاق است که با نسبت دادن پیامدهای اخلاقی افعال به افرادی که افعال از آنها سرزده است، آنها را به سمت رعایت ارزش‌های اخلاقی سوق می‌دهد. پیشرفت روزافزون فناوری و سرعت بالای گسترش آن در جهان، انتساب مسئولیت در عصر فناوری را به چالش مهمی برای متفکران تبدیل کرده است. پیچیده‌تر شدن فناوری‌ها و شکل‌گیری شبکه‌های اجتماعی-تکنیکی باعث شده زنجیره علّی منتهی به هر رویداد، مرکّب از افراد، نهادها، و مصنوعات متعددی باشد. در نتیجه، تشخیص اینکه مسئولیت هر رویداد به عهده کیست، به چالشی دشوار بدل شده‌است. در شرایطی که نتوان مسئول یا مسئولان رویداد را تشخیص داد، هدایت فرایندهای منتهی به آن رویداد و یا پیشگیری از وقوع رویدادهای نامطلوب و غیراخلاقی نیز با مسئله مواجه خواهد بود.

این مسئله در ادبیات اخلاق فناوری به عنوان چالش «توزیع مسئولیت[۱]» (**Responsibility Distribution**) شناخته می‌شود. چالش مزبور با ظهور فناوری‌های هوش مصنوعی شکل تازه‌ای پیدا کرده و به معضل «شکاف مسئولیت» بدل شده‌است؛ معضلی که نخستین بار آندریو ماتیاس آن را معرفی و صورت‌بندی کرده است. نسل جدید هوش مصنوعی توسط کلان‌داده‌ها آموزش می‌بیند و عوامل محیطی بسیاری که فرد خاصی بر آنها کنترل ندارد بر عملکرد این فناوری‌ها تأثیر می‌گذارند؛ طوری که گویی فناوری‌ها دارای سطحی از خودمختاری هستند. علاوه‌براین، قدرت هوش مصنوعی در تحلیل داده‌ها، تصمیم‌گیری، و کنترل سامانه‌های کلان باعث شده وظایف مهم و روزافزونی مانند مدیریت سامانه‌های حمل‌ونقل، مدیریت سامانه‌های بهداشت‌ودرمان عمومی، برنامه‌نویسی، و... به هوش

مصنوعی واگذار شود. پرسش اینجاست که آیا در مواردی کـه عملکـرد هـوش مصـنوعی خارج از کنترل و پیش‌بینی انسان است، می‌توان بخشـی از مسـئولیت را بـه عهـده هـوش مصنوعی گذاشت؟

پاسخ به این پرسش از یک سو بر تحلیل مفهـوم مسـئولیت و شـرایط آن، و از سـویی دیگر بر بررسی احراز یا عدم احراز این شرایط توسط هوش مصنوعی متّکی است. بر ایـن اساس، بخشی از تلاش‌های صورت‌گرفته صرفِ تأمل بر مفهوم مسئولیت، مؤلفه‌های آن، و مفاهیم هم‌بسته با آن شده‌است. زیرا اینکه آنچه به هـوش مصـنوعی مـی‌تـوان نسـبت داد، مسئولیت است یا ویژگی دیگری، محلّ مناقشات بسیار است. مفهوم مسئولیت به مثابه یک مفهوم کلی، و هم‌چنین، به طور خاص در مواجهه با هوش مصنوعی، مفهومی بسیط نیست؛ بلکه منظومه‌ای از مفاهیم هم‌بسته حول آن وجـود دارد. در هسـته ایـن منظومـه، عاملیـت (**Agency**) به عنوان پـیش‌شـرط ضـروری مسـئولیت قـرار دارد کـه در سـنت فلسـفی بـر قصدمندی و اراده استوار است (Davidson, 2001: 47; Aristotle, 1984). با این حال، در ادبیات هوش مصنوعی، این تلازم مورد چـالش قـرار گرفتـه اسـت؛ چنـان‌کـه فلوریـدی در کـار مشترکش با سندرز ابتدا با تفکیک این دو، هوش مصنوعی را عامل اخلاقی فاقد مسئولیت دانست (Floridi & Sanders, 2004)، اما در اثر بعدی خود برای اجتناب از شکاف مسـئولیت، بر پیوند ناگسستنی عاملیت و مسئولیت تأکید کرد (Floridi, 2016: 8-9). مفهـوم عاملیـت در ساحت عمل با دو مفهوم پاسخ‌گـویی (**Accountability**) و جـواب‌گـویی (**Answerability**) پیوند می‌خورد. پاسخ‌گویی ماهیتی رابطه‌ای و نظارتی دارد و به ضرورت وجود سـاختاری برای شناسایی و ردیابی عامل‌های انسانی (طراحان و مالکـان) در قبـال پیامـدهای سیسـتم اشـاره دارد (Nissenbaum, 1996; Kroll et al., 2017: 657–675; Dignum, 2022). در مقابـل، جواب‌گویی بر قابلیـت عقلانـی عامـل بـرای ارائـه دلیـل و دفـاع از منطـق تصـمیمات و رفتارهایش استوار است (Smith, 2015: 102-104) و در مورد سیستم‌های هوش مصنوعی، بار تبیین‌گری را مستقیماً بر عهده توسعه‌دهندگان قرار می‌دهد. در نهایت، مفاهیم تقصیرپذیری (**Blameworthiness**) و مجازات‌پذیری (**Punishability**) ابعاد واکنشی مسـئولیت را تشـکیل می‌دهند. تقصیرپذیری، که به طور سنتی منوط به وجود امکان‌های جـایگزین (Principle of Alternative Possibilities) (Van Inwagen, 1983) و ویژگی‌های شخص‌وارگی (**Personhood**) (Strawson, 1962/1974) است، در مواجهه با هوش مصنوعی با چالش مواجه می‌شود؛ چون مجازات تنها زمانی معنا می‌یابد که عامل، ظرفیت درک رنج یـا شخصـیت حقـوقی داشـته

باشد (Sparrow, 2007: 71-73; Gunkel, 2023: 112-118). از این رو، در نظام‌های مصنوعی، مفهوم سنتی مجازات جای خود را به تنظیم رفتاری و سازوکارهای اصلاحی از طریق طراحی سیستم می‌دهد (Danaher, 2016).

محققین مختلف بر اساس تعابیری که از مفاهیم فوق و ارتباط آنها با مسئولیت در نظر می‌گیرند، راهکارهای متفاوتی در مورد شکاف مسئولیت ارائه می‌دهند. دیدگاه‌های ارائه‌شده را می‌توان در سه دسته کلی جای داد: رویکردهایی که به نفع انحلال این مسئله استدلال می‌کنند، یعنی آن را معضلی منفی یا واجد اهمیت نمی‌دانند؛ رویکردهایی که راه حل مسئله را انتساب مسئولیت به هوش مصنوعی می‌دانند؛ و رویکردهایی که با انتساب مسئولیت صرفاً به عامل انسانی راه حل‌هایی ارائه می‌دهند. هدف مقاله حاضر، اولاً شرح مختصر هر سه رویکرد و اشکالات وارد به آنها، و ثانیاً دفاع از رویکردهایی است که مبتنی بر بازآرایی نقش‌های اجتماعی بوده و از این طریق مسئولیت عامل انسانی را حفظ می‌کنند.

برای این منظور، در قسمت بعدی صورت‌بندی ماتیاس از معضل شکاف مسئولیت ارائه می‌شود. در قسمت سوم مقاله دیدگاه‌هایی که قائل به انحلال مسئله هستند، شرح داده می‌شوند. قسمت چهارم مقاله، به دو رویکردی می‌پردازد که راه حل مسئله را در انتساب مسئولیت به هوش مصنوعی، یا انسان معرفی می‌کنند. نهایتاً در قسمت پنجم، تحلیلی از مفهوم مسئولیت پیشنهاد می‌شود که ترکیبی، زمینه‌مند و درجه‌مند است. ادعا می‌شود اولاً این مفهوم دارای جامعیت و انطباق با کاربردهای این مفهوم در سیاق مباحث اخلاق اجتماعی است. ثانیاً بر اساس این تعبیر، رویکرد سوم نسبت به مسئله شکاف مسئولیت مبتنی بر فهم صحیح‌تری نسبت به هسته اصلی مسئله است. علاوه‌براین، از آنجا که راه حل مسئله شکاف مسئولیت در حالت کلی از طریق «بازآرایی نقش‌های اجتماعی» قابل حصول است، در مورد شکاف مسئولیت ناشی از فناوری نیز چنین راهکارهایی که در برخی مصادیق رویکرد سوم مورد توجه قرار گرفته است، ثمربخش خواهد بود. در نهایت به برخی چالش‌های موجود برای این راهکار نیز اشاره خواهد شد.

۲. صورت‌بندی ماتیاس از معضل شکاف مسئولیت

ماتیاس در مقاله «شکاف مسئولیت: انتساب مسئولیت به اعمال ماشین‌های خودکار یادگیرنده» (Matthias, 2004)، مفهوم شکاف مسئولیت را معرفی می‌کند. او در این مقاله با ارائه تعریفی از مسئولیت و شروط لازم برای انتساب آن، نشان می‌دهد چطور ظهور

فناوری‌های هوش مصنوعی یادگیرنده می‌توانند انتساب مسئولیت را متحول ساخته و با چالش همراه کنند.

ماتیاس در تعریف مفهوم مسئولیت، می‌گوید مسئول دانستن یک فرد بابت یک عمل، معادل این است که یا آن فرد باید بتواند تبیینی از مقاصد و باورهای خود در مورد انجام آن عمل ارائه دهد، یا به‌نحو درستی موضوع گرایش‌های واکنشی (**Reactive attitudes**) خاص مثل رنجش، سپاس‌گزاری و ... قرار گیرد. امّا در صورتی می‌توان به‌درستی و مطابق عدالت فردی را مسئول دانست، که آن فرد بر رفتار خود و پیامدهای ناشی از آن، به معنایی مناسب، کنترل داشته باشد. برای برآورده شدن شروط مسئولیت، لازم است که فرد به حقایق خاصی در ارتباط با عمل معرفت داشته باشد، و همچنین بتواند آزادانه یک عمل را از میان چندین گزینه ممکن انتخاب کرده و مطابق آن عمل کند (Matthias, 2004: 175).

با تکیه بر تعریف و شروط فوق برای مسئولیت، ماتیاس مسئولیت پیامدهای استفاده از ماشین را به متصدّی (**Operator**) استفاده از ماشین، یا تولیدکننده (**Manufacturer**) ماشین نسبت می‌دهد. در مواردی که متصدّی، ماشین را مطابق دستورالعملی که تولیدکننده به او ارائه داده استفاده می‌کند، به‌طور ضمنی مسئولیت استفاده پیامدهای استفاده از ماشین را بر عهده می‌گیرد. اگر ماشین مطابق دستورالعمل کار نکند، مسئولیت پیامدهای آن بر عهده تولیدکننده است (Matthias, 2004: 175). امّا در مورد ماشین‌های خودکار یادگیرنده، وضعیت متفاوت است.

ماتیاس با بررسی چهار نمونه از ماشین‌های یادگیرنده (مریخ‌نورد با قابلیت یادگیری در مسیریابی، آسانسورهای انطباق‌پذیر در برج‌های بزرگ، تشخیص خودکار سرطان ریه بر اساس تصاویر میکروسکوپی نمونه‌هایی از بافت بدن، نسخه پیشرفته ربات متحرک AIBO به عنوان حیوان خانگی) نشان می‌دهد ماشین‌هایی از این دست چون قابلیت یادگیرندگی دارند، می‌توانند بدون مداخله انسان مسیری را برای عمل انتخاب کرده و مطابق آن عمل کنند. قوانینی که این ماشین‌ها طبق آن عمل می‌کنند، توسط تولیدکننده تعیین نمی‌شود و خود ماشین‌ها آن را تغییر می‌دهند. پس نمی‌توان به‌شیوه سنتی، متصدّی، یا تولیدکننده را بابت پیامدهای عملکرد ماشین مسئول دانست. زیرا وقتی آنها کنترلی بر خروجی ماشین ندارند، مسئول دانستن آنها خلاف عدالت است. به این ترتیب، در این موارد با «شکاف مسئولیت» مواجه می‌شویم (Matthias, 2004: 176-177).

ماتیاس سپس با بررسی چهار نوع طراحی هوش مصنوعی شامل سامانه‌های نمادین (**Symbolic Systems**)، اتصال‌گرایی (**Connectionism**)، الگوریتم‌های ژنتیکی و برنامه‌نویسی ژنتیکی، و عاملان خودمختار)، بر اساس ویژگی‌های فنی این ماشین‌ها برقرار نبودن شروط انتساب مسئولیت در مورد پیامدهای استفاده از آنها را نشان می‌دهد. برای مثال شبکه‌های عصبی شفاف نیستند و نمی‌توان شیوه تولید خروجی بر اساس ورودی را به‌سادگی دریافت. همچنین در این ماشین‌ها، وقوع خطا بخشی از فرآیند آزمون و خطا و یادگیری، و در نتیجه اجتناب‌ناپذیر بوده و مسئولیت آن متوجه عاملی انسانی (طراح، کاربر، یا متصدّی) نیست. در مورد عاملان خودمختار نیز کنترل سامانه تا حدی به محیط پیرامون آن وابسته است، و تعامل با عوامل متکثر محیطی می‌تواند چنان پیچیده باشد که پیش‌بینی یا کنترل تأثیرات محیط بر ماشین ناممکن شود (Matthias, 2004: 182).

علاوه‌براین، ماتیاس متذکر می‌شود که در بسیاری از موارد، حفظ نظارت انسانی بر عملکرد ماشین می‌تواند علی‌الاصول، یا به دلایلی اقتصادی، ناممکن باشد. برای مثال وقتی ماشین‌ها نسبت به متصدّی استفاده برتری اطلاعاتی داشته باشند (مثل سامانه‌های کنترل پرواز مبتنی بر رادار)، یا در مواردی که سرعت پردازش و متغیرهای عملکرد به‌قدری زیاد باشد که یک انسان نتواند در زمان واقعی بر عملکرد سامانه نظارت کند، مسئله نظارت و کنترل انسانی بر ماشین منتفی خواهد بود. این در حالی که است نمی‌توانیم از بهره‌مندی از چنین سامانه‌هایی چشم‌پوشی کنیم؛ چون انجام وظایف پیچیده مستلزم متکی بودن بر چنین ماشین‌هایی است. از سوی دیگر، نسبت دادن مسئولیت به انسان‌هایی که نمی‌توانند بر عملکرد ماشین‌ها کنترل کافی داشته باشند، خلاف عدالت است. بر این اساس، ماتیاس بر لزوم چاره‌اندیشی برای ترمیم شکاف مسئولیت در اخلاق و قانون‌گذاری تأکید می‌کند (Matthias, 2004: 83).

۳. شکاف مسئولیت: انحلال مسئله

برخی محققین در مواجهه با طرح مسئله ماتیاس، معضل بودن شکاف مسئولیت را – به‌نحوی که ماتیاس آن را صورت‌بندی کرده است– زیر سؤال برده و به نوعی برای منحل کردن این مسئله استدلال کرده‌اند. سه موضع متفاوت در قبال این موضوع را می‌توان ذیل این رویکرد جای داد:

- مسئله مذکور در واقع «شکاف» مسئولیت نیست، بلکه چالشی دیگر مرتبط با مسئولیت است.
- شکاف مسئولیت «فناوری‌بنیان» (**Techno-Responsibility Gap**) نیست[۲].
- شکاف مسئولیت «شر» نیست.

نمونه‌ای از موضع اول را می‌توان در کار کینر دید. کینر مسئولیت‌پذیری را معادل جواب‌گویی در نظر می‌گیرد و برای مسئولیت‌پذیری شخص در قبال یک آسیب، دو شرط برمی‌شمرد: اینکه شخص در ایجاد آن آسیب نقشی علّی داشته باشد؛ و اینکه شخص برای جلوگیری از وقوع آن آسیب الزامی اخلاقی داشته باشد (Kiener, 2025: 358). او با تکیه بر این دو شرط نتیجه می‌گیرد در مورد عملکرد هوش مصنوعی، انسان‌های بسیاری دو شرط فوق را برآورده کرده و واجد مسئولیت هستند. زیرا همه افرادی که در زنجیره رویدادهای منتهی به آسیب زدن هوش مصنوعی، مشارکت علّی دارند، شرط علّیّت را برآورده می‌کنند و تمام افرادی که در توسعه و استفاده از هوش مصنوعی نقش دارند، دارای الزامی اخلاقی برای جلوگیری وقوع آسیب توسط هوش مصنوعی هستند (Kiener, 2025: 368). پس مسئله ناشی از هوش مصنوعی نه شکاف مسئولیت، بلکه در اصل «وفور» (**abundance**) مسئولیت است و مواجهه با آن و حل آن راهکار متفاوتی می‌طلبد[۳].

تیگارد نماینده‌ای از موضع دوم است. او معتقد است شکاف مسئولیت، مسئله‌ای ناشی از ظهور فناوری نبوده و مهم هم نیست. او توصیف استراسون (**Strawson**) از مسئولیت اخلاقی را مفروض می‌گیرد که مطابق آن، «مسئولیت» کارکرد اجتماعی گرایش‌های طبیعی انسانی نسبت به اراده خیر، شر، یا بی‌تفاوتی دیگران است. شکاف مسئولیت زمانی رخ می‌دهد که ما در حالی که دارای چنین گرایش‌هایی (مثل خشم، سرزنش، تحسین یا قدردانی) هستیم، نتوانیم هدف مناسبی برای گرایش خود پیدا کنیم. این حالت اغلب هنگامی رخ می‌دهد که مسبب وضعیت پیش‌آمده، غایب، ناشناخته، یا اساساً ناموجود باشد، یا گزینه‌های موجود برای مسئولیت به دلیل شرایط غیرمعمولی که عمل در آن رخ داده معذور بوده، یا به دلیل ویژگی‌های نامعمول خود از مسئولیت معاف باشند (Tigard, 2021: 591-593). این وضعیت در زندگی روزمره ما در موارد متعددی رخ می‌دهد، مانند مواقعی که فردی دچار بدشانسی اخلاقی شده و بدون داشتن نیت سوء یا امکان پیش‌گیری از آسیب، مسبب آسیب رسیدن به دیگری می‌شود و معذور است، یا به دلیل خردسالی یا بیماری روانی از مسئولیت معاف است. پس شکاف مسئولیت، پدیده‌ای است که در زندگی

روزمره و خارج از چارچوب اعمال فناورانه نیز به کرّات با آن مواجه می‌شویم و بخشی عادی از زیست اخلاقی ما است.

تیگارد تلاش می‌کند نشان دهد در مورد فناوری‌های هوش مصنوعی نیز شکاف مسئولیت معضل بغرنجی ایجاد نکرده و مفهوم مسئولیت اخلاقی را به خطر نمی اندازد. برای این کار، او تعریفی دقیق‌تر از شرایط وقوع شکاف مسئولیت ارائه می‌دهد. مطابق این تعریف، شکاف مسئولیت زمانی ایجاد می‌شود که اولاً مسئول دانستن شخص یا اشخاصی به میزان خاص D برای یک وضعیت متناسب به نظر برسد. ثانیاً گزینه‌هایی برای مسئولیت‌پذیری در قبال آن وضعیت وجود داشته باشند، امّا مطابق فهم معمول ما، میزانی که آنها مسئول این وضعیت هستند، با D هم‌خوانی نداشته باشد (Tigard, 2021: 598). تیگارد همچنین مسئولیت را شامل سه حالت پاسخ‌گویی، جواب‌گویی، و انتساب‌پذیری (**attributability**) می‌داند (Tigard, 2021: 598-599). شکاف مسئولیت در صورتی رخ می‌دهد که دو شرط شکاف مسئولیت در مورد یکی از تعابیر فوق برآورده شود.

تیگارد سپس هر یک از تعابیر فوق را درباره هوش مصنوعی خودمختار بررسی کرده تا وجود یا عدم شکاف مسئولیت را بیازماید. او می‌گوید در مورد تعبیر «پاسخ‌گویی»، در مواردی که شرط اول برقرار است، شرط دوم می‌تواند برقرار نباشد. برای مثال وقتی هوش مصنوعی آسیبی ایجاد می‌کند، می‌توان به افرادی که مخاطرات پیش‌بینی‌ناپذیرش را پذیرفته یا مشارکتی علّی در تولید نتیجه داشته‌اند، مسئولیت نسبت داد و آنها را موظف به جبران دانست (Tigard, 2021:601). در مورد تعبیر «جواب‌گویی» نیز، می‌توان از هوش مصنوعی درباره شیوه عملکردش دلایلی طلبید. در پاسخ، هوش مصنوعی می‌تواند «اطلاعاتی» را ارائه کند. یا ممکن است به جای دریافت اطلاعات از هوش مصنوعی، از متخصصان انسانی بخواهیم در مورد شیوه عملکرد هوش مصنوعی توضیح دهند[۴]. تیگارد می‌پذیرد که ارائه اطلاعات یا توضیح عاملان انسانی، شکل استانداردی از «جواب‌گویی» نیستند، اما معتقد است در مورد عاملان انسانی نیز گاهی شرط جواب‌گویی با اغماض برقرار است و این از مسئولیت‌پذیری عاملان کم نمی‌کند. بنابراین او در مورد هوش مصنوعی جواب‌گویی را به‌عنوان فرآیندی اجتماعی و سیاسی معنادار می‌بیند (Tigard, 2021:602).

در مورد تعبیر «انتساب‌پذیری» نیز اگرچه شرط دوم برقرار است، امّا شرط اول برقرار نیست. به عبارت دیگر، انتساب (یعنی نسبت دادن یک عمل به شخصیت عامل) در مورد هویات مصنوعی متناسب نبوده و ناسازگار است، زیرا این هویات به معنای واقعی دارای

شخصیت نیستند و مانند کودکان و بیماران روانی از مسئولیت اخلاقی معاف هستند (Tigard, 2021:603). اگر ما نسبت به هوش مصنوعی گرایش‌هایی مشابه گرایش‌هایمان نسبت به عاملان اخلاقی داریم، باید در گرایش‌های خود تجدید نظر کنیم. بر اساس تحلیل حالات فوق، تیگارد تهدید مسئولیت توسط فناوری‌های نوظهور را خیلی کم‌خطرتر از چیزی می‌بیند که غالب نویسندگان برآورد کرده‌اند (Tigard, 2021:605).

در حالی که تیگارد شکاف مسئولیت در مورد هوش مصنوعی را معضلی مهم نمی‌داند، برخی محققین آن را حتی دارای ارزش نیز می‌دانند. داناهر، مانچ و همکاران در این دسته قرار می‌گیرند (Danaher, 2022; Munch et al. 2023). داناهر در دفاع از مزایای شکاف مسئولیت از اصطلاح «فضیلت شکاف مسئولیت» استفاده کرده و برای نشان دادن آن به تأمل بر ماهیت تصمیم‌گیری و عمل اخلاقی می‌پردازد. به عقیده او در بسیاری از موارد ما با تنگناهای اخلاقی مواجهیم که در آنها باید میان ملاحظات اخلاقی مختلف دست به انتخاب بزنیم؛ در این وضعیت‌های تراژیک، انتخاب ما ناگزیر به از دست رفتن برخی ارزش‌های اخلاقی منتهی می‌شود. چنین انتخابی عامل اخلاقی را تحت فشار روانی قرار می‌دهد.

در مواجهه با تنگناهای اخلاقی، دو راهبرد اصلی پیش روی ماست: تفویض، یا مسئولیت‌پذیری. در راهبرد تفویض، ما بار روانی و اخلاقی تصمیم‌گیری در موارد تراژیک را از روی دوش خود برداشته و روی دوش هویتی می‌گذاریم که آن را «بهتر» تحمل می‌کند. این واگذاری می‌تواند بر اساس دلیل، یا تصادفی باشد[۵]. امّا در رویکرد مسئولیت‌پذیری، ما لزوم تصمیم‌گیری در تنگناهای اخلاقی را پذیرفته و پذیرش مسئولیت چنین تصمیم‌هایی را نوعی فضیلت اخلاقی می‌دانیم. داناهر هر یک از این دو رویکرد را دارای مزایا و معایبی می‌داند. تفویض، بار روانی و اخلاقی را کم کرده، امّا تکرار آن می‌تواند توان تفکر و عمل اخلاقی را تضعیف کند. از سوی دیگر، مسئولیت‌پذیری فضیلتی اخلاقی است، امّا می‌تواند سبب شود افراد در قبال پیامدهای منفی اجتناب‌ناپذیر تصمیمات تراژیک مسئول دانسته شده و ناعادلانه مجازات شوند. بنابراین، در مواجهه با تنگناهای اخلاقی، باید به تعادلی میان این دو رویکرد رسید (Danaher, 2022: 12-14).

فناوری‌های هوش مصنوعی می‌توانند کفه مزایای اخلاقی را به نفع تفویض سنگین کنند. این فناوری‌ها گزینه مناسبی برای تفویض مبتنی بر دلیل هستند، چون ما با علم به توانمندیشان، انجام یک وظیفه اخلاقی را بهشان محوّل کرده و تصمیم‌گیری در موقعیت‌های تراژیک را به آنها می‌سپاریم. آنچه دقیقاً در این موارد تفویض را ممکن

می‌سازد، وجود شکاف مسئولیت است؛ از یک سو نمی‌توان فناوری‌ها را مسئول دانست، و از سوی دیگر نمی‌توان انسان‌ها را مسئول دانست (Danaher, 2022: 16). به این ترتیب، واگذاری تصمیم‌گیری به فناوری‌های هوش مصنوعی بار اخلاقی و روانی را در موارد بسیاری از دوش انسان‌ها برمی‌دارد.

با الهام از کار داناهر، مانچ و همکارانش نیز از مزیت شکاف مسئولیت دفاع می‌کنند. آنها با طرح ادعایی قوی‌تر از ادعای داناهر، می‌گویند به‌طور کلّی پذیرش مسئولیت می‌تواند برای فرد بار روانی منفی داشته باشد و به همین دلیل ارزش اخلاقی منفی دارد. اینکه فرد خطاکار تنبیه شده یا در باقی زندگی خود متحمّل پیامدهای منفی پذیرش مسئولیت خطا مثل سرزنش، عذاب وجدان یا از دست دادن شغل و منزلت اجتماعی شود، فشاری روانی برای او ایجاد می‌کند (Munch et al., 2023: 5). پس بهتر آن است که در فرآیندهای خطاپذیری که می‌توانند منجر به بی‌عدالتی شوند (مانند تصمیم‌گیری برای اعطای وام به مستحق‌ترین افراد)، انسان‌ها را از تصمیم‌گیری معاف کنیم. تفاوت کار مانچ و همکارانش با داناهر این است که آنها سودمندی شکاف مسئولیت را محدود به تنگناهای اخلاقی ندانسته و آن را در طیف وسیعی از تصمیم‌گیری‌های اخلاقی، مثبت می‌دانند.

۴. شکاف مسئولیت: حل مسئله

برخلاف محققینی که قائل به انحلال مسئله مورد اشاره ماتیاس هستند، گروهی دیگر شکاف مسئولیت را به عنوان مسئله‌ای که ناشی از ظهور فناوری‌های هوش مصنوعی است پذیرفته و تلاش می‌کنند برای آن پاسخی پیدا کنند. دیدگاه‌های ارائه شده ذیل این رویکرد را می‌توان شامل پاسخ‌هایی که مسئولیت را به هوش مصنوعی نسبت می‌دهند، و پاسخ‌هایی که مسئولیت را محدود به انسان می‌دانند، دسته‌بندی کرد.

۱.۴ انتساب مسئولیت به هوش مصنوعی

استدلال‌های مدافع مسئولیت‌پذیری هوش مصنوعی عموماً بر دو گام استوارند: تبیین شروط انتساب مسئولیت و سپس انطباق این شروط بر سیستم‌های هوش مصنوعی. این تلاش‌ها را می‌توان در دو رویکرد کلی طبقه‌بندی کرد:

الف) رویکـرد متمرکـز بـر ویژگـی‌هـا (**Property-focused**): در ایـن دیـدگاه، انتسـاب مسئولیت منوط به احراز ویژگی‌های هستی‌شـناختی خاصـی در فاعـل اسـت. ایـن ویژگی‌ها طیفی از آگاهی (Himma, 2009)، خودمختاری (Haselager, 2005) و حالات قصدمندانه خاص (Rodogno, 2016) تا حکمت عملی (List & Pettit, 2011) یا ترکیبی از آنها را در بر می‌گیرند. نمونه‌ای از این رویکـرد، نظریـه مسـئولیت‌پـذیری بـدون تقصـیر فلوریـدی اسـت. وی بـا فروکاسـتن مسـئولیت بـه پاسـخ‌گـویی علّـی (**Being causally accountable**)، سه ویژگی خودمختاری، تعامل با محیط و یـادگیری را شروط لازم و کافی برای مسئولیت‌پذیری می‌داند(Floridi, 2016: 6-7) .

ب) رویکرد ربطی(**Relational Approach**): در مقابل، دیدگاه‌های ربطی بر ایـن فـرض بنیادین استوارند که شأن اخلاقی و مسئولیت یک هویت، نه ناشـی از ویژگـی‌هـای درونی عینی، بلکه محصول نحوه ادراک، مواجهه و ارتباط ما با آن در شبکه تعاملات اخلاقی است. در این چارچوب، همان‌گونه که در نظـام‌هـای اجتمـاعی بـه برخـی عاملان غیرانسانی (ماننـد حیوانـات دسـت‌آمـوز) بـر اسـاس نقـش و عملکردشـان مسئولیت نسبت داده می‌شود (Sullins, 2006: 24-25)، هوش مصنوعی نیز در صورت تصدّی نقش‌های اجتماعی که مستلزم پذیرش مسـئولیت اسـت، مـی‌توانـد مسـئول قلمداد شود

۲.۴ انتساب مسئولیت صرفاً به انسان‌ها

در مقابل گروه قبل که از امکان واگذاری مسئولیت به هوش مصنوعی دفاع می‌کنند، برخـی فیلسوفان بر این باورند که مسئولیت تنها به انسان‌ها تعلق دارد. آنها معتقدند هوش مصنوعی فاقد آگاهی، خودمختاری و قصدمندی واقعی است و نمـی‌توانـد درک مناسـبی از اخـلاق داشته باشد. همچنین، چون هوش مصنوعی دارای ظرفیت تحمـل رنـج اخلاقـی نیسـت و مجازات برایش بی‌معناست، شروط پذیرفتن مسئولیت را دارا نیست. در نتیجه، باید مسـئله شکاف مسئولیت را به توزیع مسئولیت میـان انسـان‌هـا بازگردانـد. بـه عبـارتی، طراحـان، برنامه‌نویسان و کاربران هوش مصنوعی همگی به نوعی در خطاهای احتمالی این فناوری‌ها مسئول هستند و باید پیامدهای ناشی از استفاده از این فناوری‌ها را پذیرا باشند (De Cremer, 2024 Verbeek, 2011; Bringsjord, 2008).

از جمله مدافعین این رویکرد که می‌توان آنها را در دسته شک‌گرایان هستی‌شناختی (**Ontological Skeptics**) جای داد، جانسون و وردی‌چیو هستند. آنها معتقدند سامانه‌های هوش مصنوعی صرفاً عامل‌هایی مصنوعی هستند، زیرا در قالب مدارهای الکترونیک از پیش طراحی‌شده عملکرد خود را انجام می‌دهند نه کنش‌های طبیعی انسانی. جانسون و وردی‌چیو برای توضیح این موضوع، بین سه نوع عاملیت تفکیک قائل می‌شوند، امّا تأکید دارند که فقط نوع سوم، یعنی عاملیت سه‌گانه، با مسئولیت پیوند دارد (Johnson & Verdicchio, 2019):

۱) در فهم وسیع (عاملیت علّی)، عاملیت به معنای ورود هویات به روابط علّی و ایجاد تغییر در جهان است؛ این نگاه اجازه می‌دهد تأثیرات عینی فناوری بر شکل‌گیری جهان را به‌رسمیت بشناسیم.

۲) در فهم محدود (عاملیت ارادی)، عاملیت منحصر به کنش‌های ارادی و حالات ذهنی انسان است و انتساب آن به هوش مصنوعی، صرفاً تعبیری استعاری قلمداد می‌شود.

۳) مدل عاملیت سه‌گانه: عاملیتی است که انسان با مشارکت فناوری و برای دست‌یابی به هدفی مشخص، بروز می‌دهد. عاملیت علّی و ارادی هیچ یک برای پوشش‌دهی مفهوم عاملیت بسنده نیستند، و ترکیب هر دو نوع عاملیت، به حل بهتر مسئله کمک می‌کند. در عاملیت سه‌گانه، سه مؤلفه را می‌توان از یکدیگر تفکیک کرد: یک) کاربری که می‌خواهد به هدفی مشخص برسد و وظیفه دستیابی به هدف را به طراح واگذار می‌کند. دو) طراحی که یک مصنوع را برای رسیدن به این هدف مشخص می‌سازد. سه) مصنوعی که اثر علّی لازم را برای رسیدن به هدف تولید می‌کند.

به باور مخالفان مسئولیت‌پذیری هوش مصنوعی، واگذاری مسئولیت به این سامانه‌ها نه معنا دارد و نه فایده. چرا که رفع شکاف هستی‌شناختی میان مدل‌های محاسباتی و موجودات واجد احساس، مستلزم اراده‌ای است که هوش مصنوعی فاقد آن است. هوش مصنوعی صرفاً دارای عاملیت علّی است و از آنجا که مفاهیمی چون «قصد سوء» یا «بی‌توجهی» اموری ارادی هستند، مسئولیت نهایتاً متوجه انسانی است که طراحی و به‌کارگیری سیستم را مجاز کرده است. از منظری کارکردی، مسئولیت سازوکاری برای ایجاد انگیزه در افراد است تا مسئولانه رفتار کنند؛ لذا این بار باید بر عهده کسانی باشد که توانایی تضمین و کنترل خروجی‌ها را دارند. جانسون و وردی‌چیو (۲۰۲۳) تأکید می‌کنند که برای حل چالش شکاف مسئولیت، باید به جای تمرکز بر رفتار ماشین، بر ساختارهای

انسانی و سازمانی پشت آن متمرکز شد. در این نگاه، هوش مصنوعی بخشی از یک سامانه اجتماعی–تکنیکی بزرگتر است که تحلیل اخلاقی آن تنها با در نظر گرفتن کنشگران انسانی، هنجارها و رویه‌های حاکم بر آن معنا می‌یابد.

بیلز (۲۰۲۵) با رویکردی هستی‌شناختی، واگذاری عاملیت به هوش مصنوعی و ظهور شکاف مسئولیت را نوعی بندگی داوطلبانه (**willing servitude**) می‌نامد که به معنای پایان خودمختاری اخلاقی انسان است. وی استدلال می‌کند که مسئولیت نباید به سامانه‌های غیرانسانی منتقل شود، زیرا این امر با جوهر اراده آزاد بشری در تضاد است. از نظر او، مسئولیت تنها زمانی معنا دارد که برخاسته از خودمختاری باشد و از آنجا که هوش مصنوعی فاقد این ویژگی است، هرگونه انتساب مسئولیت به ماشین یا فرار از مسئولیت انسانی، از نظر اخلاقی باطل است. بیلز تأکید می‌کند که ارزش تصمیم‌گیری نه فقط در نتیجه، بلکه در فرایند تصمیم‌گیری توسط یک موجود صاحب اراده نهفته است. لذا حتی اگر کارایی هوش مصنوعی بالاتر باشد، سپردن مسئولیت به آن باعث تهی شدن اخلاقی جامعه خواهد شد.

رویکردی دیگر با نام انسان–در–حلقه (**Human in the loop**) شناخته می‌شود. رابرت اسپارو (Sparrow, 2007) استدلال می‌کند که استفاده از سامانه‌های هوشمند مستقل در تصمیمات حساس (مانند سلاح‌های خودمختار) غیراخلاقی است. زیرا این سامانه‌ها به دلیل فقدان درک رنج، مجازات‌پذیر نیستند و همچنین منطقی نیست انسان‌ها را برای رفتارهای پیش‌بینی‌ناپذیر یک ماشین کاملاً خودمختار مجازات کنیم. برای جلوگیری از این شکاف مسئولیت، ضرورت دارد که عمل نهایی صرفاً با تأیید مستقیم انسان انجام شود تا مسئولیت اخلاقی جان انسان‌ها حفظ گردد.

رویکرد انسان–بر–حلقه (**Human on the Loop**) در عین شباهت، برآمده از نگاهی منتقدانه به رویکرد انسان–در–حلقه است. مدافعانی چون پل شار (Scharre, 2018) معتقدند در حوزه‌هایی که سرعت واکنش ماشین از درک انسان فراتر می‌رود (مانند جنگ‌های سایبری)، ماشین به‌صورت خودکار عمل می‌کند و متصدّی انسانی تنها قدرت نظارت و وتو دارد. با این حال، این رویکرد با چالش‌های جدی روبه‌روست. اعتماد بیش از حد ناظر به ماشین (سوگیری اتوماسیون) نظارت را به تأییدی بی‌چون‌وچرا تبدیل می‌کند. علاوه‌براین، ناظر انسانی که فقط به نمایشگر نگاه می‌کند و فعالیت فیزیکی در تصمیم‌گیری ندارد، به‌تدریج تمرکزش را از دست می‌دهد و از بافتار و جریان واقعی اتفاقات عقب

می‌ماند و هنگام بروز بحران نمی‌تواند به‌سرعت تصمیم درستی بگیرد. در نتیجه، نظارت انسانی به کنترل صوری تقلیل یافته و ناظر انسانی عملاً به سپر بلای اخلاقی تبدیل می‌شود. یعنی تمام مسئولیت قانونی و اخلاقی خطاها بر دوش اوست، در حالی که نقش ناچیزی در پیشگیری از آن‌ها داشته است (Roff, 2016; Asaro, 2012; Elish, 2019; Sharkey, 2010: 370-371 Scharre, 2018).

برای رفع نواقص رویکرد انسان–در–حلقه و انسان–بر–حلقه، گروه دیگری از نظریه‌پردازان می‌کوشند با معرفی مفهوم «کنترل معنادار انسانی» (**Meaningful Human Control- MHC**) حل مسئله را یک گام جلو ببرند (Roff, 2016; Asaro, 2012; Sharkey, 2010). سانتونی و فن دن هوفن مفهوم کنترل معنادار انسانی را از عمل تأیید بدون درک وقایع و حضور فیزیکی بی‌معنا متمایز می‌کنند و راهکاری ارائه می‌دهند که نشان می‌دهد چگونه می‌توان مسئولیت را حتی در سامانه‌های بسیار پیچیده و خودکار، هم‌چنان متوجه انسان‌ها دانست. آن‌ها دو شرط ضروری برای برقراری مسئولیت انسانی معرفی می‌کنند (Santoni de Sio & van den Hoven, 2018):

- شرط ردیابی(**Tracking**): این شرط فشار را از روی متصدّی برداشته و به لایه طراحی می‌برد: سامانه باید به‌گونه‌ای طراحی شود که رفتار آن، در پاسخ به دلایل و ارزش‌های انسانی باشد (**Reason-Responsive**). یعنی اگر شرایط محیطی یا ارزش‌های ما تغییر کرد، رفتار سامانه نیز تغییر کند. در واقع، ماشین باید اهداف و نیات اخلاقی انسان را «ردیابی» کند.
- شرط ردزنی (**Tracing**): باید بتوان دست‌کم یک انسان را در سلسله‌مراتب طراحی یا بهره‌برداری پیدا کرد که چگونگی کارکرد سامانه را درک کند و پیامدهای اعمالش را (به عنوان طراح یا متصدّی) بداند. این شرط در پاسخ به یکی از مشکلات رویکرد انسان–در–حلقه ارائه شده؛ این که ممکن است متصدّی در حلقه باشد امّا به دلیل سرعت بالا یا پیچیدگی سامانه، فقط دکمه تأیید را بدون فکر فشار دهد.

ادعا این است که این دو شرط با قرار دادن رفتار ماشین در امتداد اراده و کنترل انسان، مسئله شکاف مسئولیت را حل می‌کنند. برآورده شدن این شروط مستلزم حفظ سطوحی از شفافیت و تبیین‌پذیری در هوش مصنوعی است و بر شیوه‌های توسعه هوش مصنوعی قیودی وضع می‌کند. سانتونی و فن دن هوفن معتقدند مسئولیت باید در کل «سامانه فنی–اجتماعی» توزیع شود و به مرحله طراحی نیز بازگردد. سانتونی در اثر بعدی خود که با

مشارکت مکاچی نگاشته است، دلالت‌های کنترل معنادار انسانی را در ارتباط با شکاف مسئولیت شرح می‌دهد. از جمله نتایج شرط ردزنی، طراحی یک سامانه فنی-اجتماعی به‌نحوی است که عاری از شکاف‌های مسئولیت باشد. در چنین سامانه‌ای عاملان انسانی برای انجام مسئولیت خود ظرفیت و انگیزه لازم را پیدا می‌کنند. برای مثال، اگر فردی در یک سامانه فناورانه مسئول پاسخ‌گویی به جامعه است، باید آگاهی و فهم لازم از شیوه کارکرد هوش مصنوعی در این سامانه را کسب کند و برای کسب این ظرفیت باید از او حمایت شود (Santoni de Sio and Mecacci, 2021, 1063–1076).

۵. حل شکاف مسئولیت از طریق بازآرایی نقش‌های اجتماعی

همان طور که مرور دیدگاه‌های مختلف در بخش‌های پیشین نشان می‌دهد، یکی از سرچشمه‌های اختلاف نظر درباره پاسخ مسئله شکاف مسئولیت، اختلاف درباره مفهوم مسئولیت و لوازم آن است. در این قسمت، نخست تعبیری از مسئولیت معرفی شده و از آن دفاع می‌شود. دوم، بر اساس این تعبیر، ادعا می‌شود از رویکردهای سه‌گانه بررسی‌شده (انحلال مسئله شکاف، انتساب مسئولیت به هوش مصنوعی، انتساب مسئولیت صرفاً به انسان) رویکرد سوم مبتنی بر دریافت صحیح‌تری نسبت به مسئله شکاف بوده و به همین دلیل نسبت به رویکردهای دیگر مرجح است. در نهایت، در میان پاسخ‌های رویکرد سوم، از کارآمدی پاسخ‌هایی که مبتنی بر بازآرایی نقش‌های اجتماعی هستند (مثل کنترل معنادار انسانی) دفاع شده و همچنین چالش‌های پیش روی آنها شرح داده می شود.

۱.۵ مسئولیت به عنوان مفهوم ترکیبی، زمینه‌مند و درجه‌مند

مسئله شکاف مسئولیت در زمینه مباحث اخلاق فناوری مطرح شده‌است، که حوزه‌ای اجتماعی از اخلاق است. بنابراین در ارتباط با این مسئله، باید مفهوم مسئولیت به‌نحوی تحلیل شود که هم با کاربردهای رایج آن در مورد مسائل اخلاقی اجتماعی سازگار باشد، و هم برای هدف کارکردی این مفهوم در جامعه -یعنی تنظیم رفتار انسان‌ها برای تحقق ارزش‌های اخلاقی - کارآمد باشد. بر این اساس، کارکرد و معنای مسئولیت اخلاقی در بستر اجتماع، محوریت پیدا می‌کند. با در نظر گرفتن کاربردهای مفهوم مسئولیت اخلاقی در حوزه مسائل اجتماعی، می‌توان سه ویژگی را برای این مفهوم برشمرد: ترکیبی بودن، زمینه‌مند بودن، و درجه‌مند بودن.

ترکیبی بودن: همان طور که در مقدمه مقاله نیز مورد اشاره قرار گرفت، مفهوم مسئولیت با مفاهیم دیگری (همچون عاملیت، پاسخ‌گویی، جواب‌گویی، تقصیرپذیری و...) قرابت و هم‌بستگی دارد. برخی محققان در رویکردی یگانه‌انگارانه (**moral responsibility monism**)، نسبت مفهوم مسئولیت با این مفاهیم را ترادف در نظر گرفته‌اند و مثلاً مسئولیت را به‌معنای جواب‌گویی در نظر گرفته‌اند (Smith, 2012). در این حالت مسئولیت مفهومی یگانه و بسیط فرض شده‌است . برخی دیگر، مفهوم مسئولیت را مفهومی چندگانه (**moral responsibility pluralism**) و مفاهیم نزدیک به آن را «انواع مختلفی» از مسئولیت در نظر گرفته‌اند (Watson, 1996; Jeppsson, 2022). برای مثال، شومیکر، مسئولیت را شامل سه نوع پاسخ‌گویی، جواب‌گویی، و انتساب‌پذیری می‌داند (Shoemaker, 2011). بر این اساس، در یک موقعیت می‌توان نوعی از مسئولیت را به عاملی نسبت داد امّا نوعی دیگر را نسبت نداد. برای مثال، اگر فردی مبتلا به بیماری روانی به دیگران آسیب بزند، او را مسئول به‌معنای مجازات‌پذیر نمی‌دانیم چون بابت آنچه از او سر زده تقصیری ندارد. امّا به‌معنای انتساب‌پذیری او را مسئول می‌دانیم، زیرا عملِ انجام‌شده به‌درستی به او منتسب است. اینجا مسئولیت عنوانی کلی است که چند مفهومِ متمایز مصداق آن هستند.

علاوه بر دو رویکرد فوق، می‌توان مسئولیت را «ترکیبی از چند مفهوم» در نظر گرفت. در این حالت، مفهوم مسئولیت، شامل چند مفهوم و قوام‌یافته توسط آنها است. در چنین تعبیری، هرگاه مسئولیت به عاملی نسبت داده می‌شود، تمام مؤلفه‌های مسئولیت نیز در این انتساب مرتبط و معنادار هستند، هرچند همه آنها به یک‌اندازه برجسته و مهم نباشند.

در نظر گرفتن مسئولیت به عنوان مفهومی مرکّب از پاسخ‌گویی، جواب‌گویی و انتساب‌پذیری چند مزیت دارد. اول اینکه کاربرد این مفهوم طیف وسیعی از موارد را پوشش می‌دهد. اگر مسئولیت را صرفاً مترادف یکی از مفاهیم فوق در نظر بگیریم، بخشی از کاربردهای مفهوم مسئولیت، از دایره این مفهوم خارج می‌شوند. مثلاً اگر مفهوم مسئولیت را صرفاً انتساب‌پذیری در نظر بگیریم، کاربردهایی از مفهوم مسئولیت اخلاقی که در آن با ارجاع به این مفهوم به دنبال اثبات مجازات‌پذیری یک عامل هستیم، معنا پیدا نمی‌کنند. ضمناً دربرگرفتن مؤلفه‌های مختلف موجب می‌شود مفهوم مسئولیت هم به معنای گذشته‌نگر برای جبران پیامدهای نامطلوب گذشته، و هم به معنای آینده‌نگر و برای پیش‌گیری از پیامدهای نامطلوب در آینده کارکرد داشته باشد.

از سوی دیگر، می‌توان به رویکرد چندگانه که تکثر در تعابیر مسئولیت را می‌پذیرد نیز، اشکالاتی وارد کرد. مطابق این رویکرد، ما در انواع مختلف مسئولیت‌پذیری، با مفاهیم جداگانه‌ای سروکار داریم و هنگامی که به مسئولیت یک عامل حکم می‌دهیم، یک نوع از انواع مسئولیت را در نظر داشته و سایر انواع به این حکم نامرتبط هستند. در حالی که به نظر می‌رسد استفاده ما از مفهوم مسئولیت در موارد مختلف به یکدیگر مرتبط باشد. مثلاً هنگامی که یک بیمار روانی را که به دیگران آسیب زده مسئول می‌دانیم، در حالی که با این حکم عمدتاً بر انتساب عمل به او تأکید داریم، امّا در وهله اول مجازات‌پذیری و جواب‌گویی را نیز در مورد او مرتبط و معنادار می‌دانیم. هرچند، می‌توان بنا به دلایلی (که قابل بحث و مناقشه هستند)، ضریب مجازات‌پذیری یا جواب‌گویی را کاهش داد و یا در مواردی به صفر رساند. امّا نمی‌توان گفت در این مورد مسئولیت، صرفاً معادل انتساب عمل است و انواع دیگر مسئولیت‌پذیری نامرتبط هستند. در واقع ترکیبی در نظر گرفتن مفهوم مسئولیت، هم با این واقعیت که حکم به مسئول بودن یک عامل در موقعیت‌های مختلف می‌تواند دلالت‌های متفاوتی داشته باشد سازگار است، و هم این شهود را که مؤلفه‌های مختلف مسئولیت در تمامی مواردی که مسئولیت به یک عامل نسبت داده می‌شود مرتبط و مطرح هستند، توضیح می‌دهد. بر این اساس، می‌توان گفت مفهوم مسئولیت اخلاقی به‌عنوان یک برساخت اجتماعی، شامل هر سه مؤلفه فوق می‌شود[۶] و اگر مفهومی رقیق‌تر از این را در نظر بگیریم، نوعی شبه‌مسئولیت را در نظر گرفته‌ایم که تمام کارکرد اجتماعی مفروض برای مفهوم مسئولیت را نخواهد داشت.

زمینه‌مندی: زمینه‌مندی مفهوم مسئولیت به این نکته اشاره دارد که میزان ضریب هر یک از مؤلفه‌های مفهوم مسئولیت می‌تواند بسته به زمینه تغییر کند. زمینه می‌تواند نوع عمل، نوع عامل، و فراتر از آن، قراردادهای عرفی جامعه‌ای باشد که مفهوم مسئولیت در آن به کار می‌رود. علاوه‌براین موارد، می‌دانیم که برای هر یک از مؤلفه‌های مسئولیت، شرایط لازم و کافی متفاوتی وجود دارد. برای مثال، برای آن که عملی به یک عامل انتساب‌پذیر باشد، لازم است آن عامل منشاء علّی عمل باشد. امّا این شرط برای پاسخ‌گویی کافی نیست. برای پاسخ‌گویی، لازم است که عامل امکان عمل به‌نحوی دیگر (**To do otherwise**) را نیز داشته باشد. امّا اینکه چه آستانه‌ای از برآورده شدن این شروط، برای تحقق یکی از مؤلفه‌های مسئولیت‌پذیری کافی است، به زمینه بستگی دارد. برای مثال، اینکه میزان سهم علّی یک

عامل از چه آستانه‌ای باید بالاتر باشد که بتوانیم آن را منشاء علّی عمل دانسته و مؤلفه انتساب‌پذیری را درباره آن برقرار بدانیم، می‌تواند بسته به قراردادهای اجتماعی متغیر باشد.

درجه‌مندی: درجه‌مند بودن مفهوم مسئولیت نیز به این واقعیت اشاره دارد که شروط لازم و کافی برای تحقق مؤلفه‌های مسئولیت می‌توانند به درجات متفاوتی محقق شوند. بر این اساس، مسئولیت، امری صفر و یکی نیست و عاملان می‌توانند به درجات متفاوتی نسبت به یک عمل مسئول باشند. این حالت در مورد مسئولیت جمعی (**Collective Responsibility**) که در آن مسئولیت میان عاملان متعددی توزیع می‌شود و هر یک سهمی از مسئولیت را بر عهده می‌گیرد، به خوبی مشهود است.

۲.۵ بررسی پاسخ‌های ارائه‌شده به مسئلهٔ شکاف مسئولیت

با پذیرش مفهوم مسئولیت با ویژگی‌های فوق، می‌توان دیدگاه‌های مرورشده در قسمت‌های پیشین مقاله را مورد بررسی قرار داد. ذیل رویکرد انحلال شکاف مسئولیت، سه ادعا مطرح شد که به ترتیب مسئله را نه شکاف مسئولیت بلکه توزیع مسئولیت می‌دانست؛ مسئله شکاف مسئولیت را فناوری‌بنیان نمی‌دانست؛ و شکاف مسئولیت را دارای مزیت و خیر می‌دانست.

دیدگاه کینر که مسئله مورد نظر ماتیاس را توزیع مسئولیت می‌دانست، مبتنی بر رویکرد یگانه‌انگار نسبت به مفهوم مسئولیت بوده و آن را معادل جواب‌گویی در نظر می‌گرفت. بر این اساس، این مفهوم رقیق مصادیق متعددی پیدا می‌کرد و مسئله به جای شکاف، به توزیع مسئولیت تبدیل می‌شد. چنان که گفته شد، چنین تعبیری از مسئولیت، کامل نبوده و تمام کارکردهای اجتماعی مفهوم مسئولیت را ندارد. همین نقد، درباره دیدگاه تیگارد نیز وارد است. تیگارد که رویکردی چندگانه نسبت به مفهوم مسئولیت داشت نیز به‌نحوی دیگر این مفهوم را رقیق کرده است. او نیز با جداسازی مؤلفه‌های مسئولیت از یکدیگر، نشان می‌دهد که اگر مسئولیت مترادف هر یک از این مفاهیم باشد، آنگاه در مورد اعمال فناورانه هوش مصنوعی نیز می‌توان عواملی انسانی را مسئول دانست. در هر دو استدلال فوق، اگر مفهوم مسئولیت را ترکیبی از هر سه مؤلفه بدانیم، دیگر نمی‌توان ادعا کرد در مورد تمام اعمال فناورانه، عاملی انسانی وجود دارد که تمام مؤلفه‌های مسئولیت‌پذیری در مورد او مرتبط و معنادار بوده و تا حد قابل قبولی برآورده شوند، بنابراین مسئله شکاف مسئولیت در وهله اول، هم‌چنان مطرح خواهد بود.

داناهر، و مانچ و همکاران، در استدلال‌های خود شکاف مسئولیت را واجد خیر می‌دانستند. منظور از خیر، کاستن از بار روانی انسان‌ها است. داناهر، فشار روانی ناشی از مسئولیت را در مورد تصمیم‌گیری تراژیک شر دانسته، و مانچ و همکاران این فشار را فراتر از موقعیت‌های تراژیک، در هر حال شر می‌دانستند. به نظر می‌رسد این رویکرد با اساسِ مفهوم مسئولیت اخلاقی در تعارض است. چنان که گفته شد، مفهوم مسئولیت، کارکردیِ اجتماعی دارد و آن تنظیم اعمال افراد به‌نحوی است که خروجی‌های مطلوب بیشینه و خروجی‌های نامطلوب کمینه شود. این ابزار مفهومی، از طریق تأثیرات روانی بر افراد (ترس از عقوبت و امیدواری به پاداش) هدف مورد نظر را محقَق می‌سازد. بنابراین، استدلال فوق پیش از آنکه پاسخی به مسئله شکاف باشد، حمله‌ای علیه مفهوم مسئولیت اخلاقی و عقلانیت آن بوده، و از موضوع شکاف مسئولیت خارج است. هرچند نقد این رویکرد مجال مجزایی می‌طلبد، امّا می‌توان به همین مقدار بسنده کرد که حذف مفهوم مسئولیت هم از منظر فایده‌گرایانه به سبب آن که ضرر کمتر (فشار روانی) را با ضرر بیشتر (پیامدهای رنج‌آور بزرگ) جایگزین می‌کند قابل قبول نیست، و هم از منظر اخلاق فضیلت‌گرا به این دلیل که افراد را از ممارست مسئولیت‌پذیری و کسب این فضیلت محروم می‌کند، مردود است.

دومین رویکرد بررسی‌شده، با انتساب مسئولیت به هوش مصنوعی، شکاف مسئولیت را پر می‌کرد. این رویکرد را می‌توان از جهت تعهدات ضمنی و پیامدهای آن مورد نقد قرار داد. از یک سو، استدلال‌هایی که از امکان‌پذیری مسئولیت هوش مصنوعی دفاع می‌کنند، عموماً مصداقی از بسط‌گرایی اخلاقی (**Moral Expansionism**) هستند. در این نوع استدلال‌ها برای تعمیم دادن مسئولیت اخلاقی از قلمرو انسان‌ها به قلمرو امور نا–انسانی، به موارد مرزی توسل می‌شود. ادعا می‌شود اگر درباره برقراری شرایط لازم و کافی مسئولیت‌پذیری نتوانیم میان موارد مرزی انسان (مانند کودکان) و هوش مصنوعی (در حالت قوی آن) تفاوتی قائل شویم، آنگاه باید دایره مسئولیت اخلاقی را از حوزه انسانی به حوزه فرا–انسانی بسط دهیم. اشکالی که به این استدلال وارد می‌شود این است که با به کار بستن مکرر این استدلال، قلمرو امور مسئولیت‌پذیر پی‌درپی بسط می‌یابد تا جایی که تمام امور را در برمی‌گیرد؛ که نتیجه‌ای نامقبول است [۷].

از سویی دیگر، با توجه به گسترش کارکردهای هوش مصنوعی در ساحات مختلف، اطلاق مسئولیت به هوش مصنوعی می‌تواند به طفره رفتن انسان از مسئولیت‌پذیری در

حوزه‌های مختلف منتهی شود. برای مثال، اگر دستیارهای هوش مصنوعی به حدی پیشرفته شوند که در اغلب موقعیت‌های روزمره و حرفه‌ای بتوانند به جای عاملان انسانی تصمیم‌گیری کنند و مسئولیت این تصمیم‌ها را نیز بر عهده بگیرند، انسان‌هایی که به چنین دستیارهایی مجهز هستند، دلیل کمی برای دخیل کردن خود در این تصمیم‌ها و پذیرفتن پیامدهای ناشی از آن خواهند داشت. این وضعیت به پس‌رفت اخلاقی انسان و تحلیل رفتن قوای درک و عمل اخلاقی منتهی خواهد شد. بنا بر مجموع این دلایل، می‌توان گفت هرچند فناوری‌های هوش مصنوعی در چگونگی توزیع مسئولیت میان انسان‌ها مدخلیتی جدی دارند، امّا خود هویاتی نا–مسئول (**A-responsible**) بوده و مسئولیت به‌معنای رایج در عرف، درباره آنها مطرح نمی شود.

اشکالات وارد به دو رویکرد فوق، ما را به سمت رویکرد سوم، یعنی انتساب مسئولیت صرفاً به عامل انسانی، سوق می‌دهد. این پاسخ، تعبیری غلیظ از مسئولیت را فرض گرفته است؛ تعبیری که با کاربردهای مفهوم مسئولیت در عرف سازگار است و همان طور که در آغاز این قسمت بیان شد، از چند مؤلفه تشکیل شده‌است. پذیرش این مفهوم غلیظ، انتساب مسئولیت را دشوار کرده و مسئله شکاف را به عنوان مسئله‌ای درخور توجه که انتساب مسئولیت به هوش مصنوعی نیز آن را حل نمی‌کند، برجسته می‌سازد.

۳.۵ بازآرایی نقش‌های اجتماعی، پاسخی سنتی به مسئلهٔ شکاف مسئولیت

نگاهی به صورت‌بندی اولیه مسئله شکاف، گویای آن است که این مسئله معطوف به دغدغه وقوع پیامدهای نامطلوبی است که اجتناب از آنها لازم است، امّا ظاهراً هیچ عاملی شرایط پذیرش مسئولیت آن را ندارد. نکته قابل توجه این است که چنین دغدغه‌ای نوظهور نبوده و در تاریخ زندگی بشر سابقه‌ای طولانی دارد. پیامدهای ناشی از رویدادهای طبیعی غیرمترقبه مثل سیل، زلزله، سرمازدگی، خشک‌سالی و... ، نمونه بارزی از خسارت‌هایی هستند که نمی‌توان کسی را بابت آن مسئول دانست. بر این اساس، تدبیر جوامع انسانی نسبت به چنین بلایایی، می‌تواند الگو و سرنخی برای حل مسئله شکاف به دست دهد.

هرچند بلایای طبیعی ماهیتاً عاملی انسانی ندارند و در وهله اول هیچ یک از اعضای جامعه انسانی شرایط پذیرش مسئولیت عواقب ناشی از این بلایا را ندارد، امّا پس از شناسایی این بلایا و زمینه‌ها و الگوهای وقوعشان، جوامع برای مقابله با آسیب‌های این بلایا نقش‌های اجتماعی جدیدی را تعریف کرده یا در نقش‌های اجتماعی موجود اصلاحاتی

ایجاد می‌کنند؛ به‌نحوی که پیامدهای وقوع این بلایا را تا حد ممکن کم کنند. یک «نقش اجتماعی» متشکل از هنجارهای گوناگون اجتماعی و بایدها و نبایدهایی است که پذیرنده نقش بر اساس آنها قضاوت می‌شود. برای مثال نقش همسر، آموزگار، شهروند، نماینده مجلس، همگی نقش‌هایی اجتماعی هستند که در جوامع گوناگون و زمان‌های مختلف، شامل هنجارهای متفاوتی می‌شوند. از آنجا که «نقش‌های اجتماعی» به عنوان عناصری پویا در ساختار اجتماعی، قابل حذف و اضافه و تصحیح هستند، تغییر مسئولیت‌های اجتماعی از طریق تحول در نقش‌های اجتماعی صورت می‌گیرد. مثلاً با تعریف یک مسئولیت اجتماعی جدید، مانند اضافه کردن «مسئولیت مقاوم‌سازی ساختمان‌ها در مقابل زلزله» به مسئولیت‌های نقش «شهردار»، فردی که به واسطه نقش اجتماعی‌اش چنین مسئولیتی را پذیرفته است، ملزم است لوازم انجام این مسئولیت را نیز کسب کند. به عبارت دیگر، هرچند او در وهله اول واجد معرفت و قدرت لازم برای پیش‌گیری از خسارات ناشی از زلزله نباشد، امّا مطابق تعبیر آینده‌نگرانه و فعّال از مسئولیت، پس از پذیرش نقش اجتماعی مربوطه، وظیفه دارد معرفت و قدرت لازم برای ایفای نقش خود را نیز کسب کند و جامعه نیز در این مسیر با او همراهی کند.

در مورد پیامدهای نامطلوب ناشی از استفاده از هوش مصنوعی نیز تدابیری از این سنخ می‌تواند به راهکارهای نتیجه‌بخش منتهی شود. خلاء وجود اشخاص انسانی با شرایط مسئولیت‌پذیری، از طریق اصلاح آرایش نقش‌های اجتماعی قابل پر شدن است. این بازآرایی، می‌تواند هم شامل تعریف نقش‌هایی جدید شود که وظیفه‌ای اختصاصی در مورد کنترل پیامدهای اخلاقی هوش مصنوعی دارند، و هم می‌تواند شامل اصلاح نقش‌های موجود (مثل کاربر، طراح، تولیدکننده) شود، به‌نحوی که اشخاص در گستره عمل خود نسبت به پیامدهای اخلاقی استفاده از هوش مصنوعی و نقش علّی کنش‌هایشان بر آنها آگاه شده و به نحو مقتضی عمل کنند. همچنین، متصدیان این نقش‌ها برای کسب لوازم نقش خود باید پشتیبانی شوند. این لوازم شامل دسترسی به اطلاعات مورد نیاز برای تخمین مخاطرات، سواد اخلاقی برای تشخیص انواع پیامدهای نامطلوب، و همچنین قدرت عمل برای پیشگیری از وقوع پیامدهای نامطلوب خواهند بود. در این صورت، میزان مسئولیت نقش‌های مختلف نیز به درجات متفاوتی و متناسب با میزان برآورده شدن شرایط مسئولیت‌پذیری توسط نقش‌های گوناگون خواهد بود. برای مثال، اگر نقشی به طور خاص برای شناسایی ابعاد اخلاقی و مخاطرات احتمالی یک فناوری هوش مصنوعی تعریف شده

باشد و متصدّی آن در مورد طراحی و توسعه این فناوری قدرت اعمال نظر نیز داشته باشد، نسبت به پیامدهای آن فناوری درجه بالاتری از مسئولیت دارد، در حالی که کاربرانی که با تولید داده‌های ورودی به فناوری جهت می‌دهند نیز مسئولیت دارند، اما به درجاتی به‌مراتب پایین‌تر.

رویکردی که مسئولیت را صرفاً به انسان نسبت می‌دهد، به‌طور کلی واجد ظرفیت ارائه پاسخ‌هایی از این جنس است. برای مثال، کاکلبرگ با اتخاذ چنین رویکردی و نا–مسئول دانستن هوش مصنوعی (Cockelbergh, 2020: 2055)، بر ضرورت توجه به تعبیر آینده‌نگرانه از مسئولیت و شکل‌دهی به فناوری و محیط اجتماعی پیرامون در راستای مسئولانه بودن کاربرد و توسعه هوش مصنوعی تأکید می‌کند (Cockelbergh, 2020: 2053). نسخه‌های تکامل‌یافته‌تر این رویکرد، یعنی کنترل معنادار انسانی، با در نظر گرفتن جوانب چندگانه مفهوم مسئولیت، راه حلی دقیق‌تر، از جنس تعریف و تصحیح نقش‌های اجتماعی پیشنهاد می‌کنند. چنان که ذیل شرط ردزنی مطرح شد، حفظ جایگاهی انسانی که شرایط لازم برای مسئولیت‌پذیری را داشته باشد، از لوازم کنترل معنادار انسانی است. بنابراین، این رویکرد علاوه‌براین که تعبیری از مسئولیت را فرض گرفته که به تعبیر رایج و عرفی نزدیک‌تر بوده و از این نظر صورت‌بندی بهتری از مسئله ارائه داده است، راه حل بهتری را نیز پیشنهاد می‌دهد؛ راه حلی که با آنچه در طول تاریخ جوامع انسانی در شرایط مشابه در پیش گرفته و از آن نتیجه گرفته‌اند، هم‌خوانی دارد.

با این حال، این رویکرد نیز با چالش‌هایی مواجه است. یکی از این چالش‌ها، فراهم کردن لوازم مسئولیت‌پذیری است. همان طور که گفته شد، یکی از شروط مسئولیت‌پذیری آگاهی است. در مورد هوش مصنوعی، برآورده شدن این شرط مستلزم تبیین‌پذیری هوش مصنوعی است. عملکرد هوش مصنوعی باید تا حد قابل قبولی تبیین‌پذیر بوده و دامنه امکان‌های عملکردی آن برای فردی که مسئولیت می‌پذیرد، روشن باشد. به عبارت دیگر، هوش مصنوعی تا حد ممکن نباید خروجی‌هایی غیرقابل پیش‌بینی داشته باشد. تبیین‌پذیری از یک جهت قیودی فنی بر طراحی وضع کرده، و از جهتی دیگر مستلزم ملاحظاتی در سیاست‌گذاری است، زیرا سرعت یک مزیت رقابتی مهم در بازار فناوری است و همین موجب می‌شود برآورده شدن برخی شاخص‌های اخلاقی مثل تبیین‌پذیری، از اولویت خارج شوند. در مورد شرط کنترل نیز از آنجا که ملاحظات اخلاقی اغلب در نقطه مقابل ملاحظات اقتصادی قرار می‌گیرند و با وضع محدودیت‌های مختلف می‌توانند از برخی

مزایای رقابتی توسعه‌دهندگان هوش مصنوعی بکاهند، تدابیری برای صیانت از قدرت عمل نقش‌هایی که بناست ضامن اخلاقی بودن هوش مصنوعی باشند، لازم است.

علاوه‌براین، برای تکمیل و تضمین اجرایی شدن راه حل‌های مبتنی بر بازآرایی نقش‌های اجتماعی، آگاهی‌بخشی عمومی درباره مسئولیت گروه‌های مختلف ذی‌نفع از جمله طراحان و سازندگان، سیاست‌گذاران و متخصصان فلسفه و اخلاق فناوری، و همچنین کاربران، بسیار ضروری به نظر می‌رسد. ترویج گفتمان مسئولیت‌پذیری در قبال توسعه و کاربرد فناوری‌های هوش مصنوعی در سطوح دانشگاهی، سیاستی، حقوقی و قانونی و همین‌طور، حوزه عمومی از مؤلفه‌هایی است که در تأثیرگذاری واقعی‌تر رویکرد مسئولیت معنادار ضروری است. این باور باید در ذهن انسان‌هایی که در عصر هوش مصنوعی می‌زیند نهادینه شود که این فناوری‌ها صرفاً جنبه خدمت‌رسانی در زیست روزمره را ندارند، بلکه شبکه عاملان –به تعبیر برونو لاتور (Latour, 2005)– مسئول و اثرگذار را تغییر می‌دهند و روابط کنشی و واکنشی جدیدی را شکل می‌دهند که نیازمند به‌رسمیت شمردن هستند. بنابراین بسیار مهم است که گروه‌ها و ذی‌نفعان مختلف با کسب آگاهی از نقش‌هایی که در قبال تولید خروجی‌های هوش مصنوعی دارند، مسئولیت خود را بپذیرند و رفتارها و تصمیمات خود را به‌نحوی مسئولانه‌تر تنظیم کنند. در عین حال، به‌لحاظ حقوقی باید تفاوت سطح برآورده شدن شروط مسئولیت میان کنشگران متفاوت لحاظ شود، و نقش‌هایی که اختصاصاً برای حفظ کنترل معنادار تعریف شده‌اند، وزن عمده پاسخ‌گویی و جواب‌گویی به جامعه و جبران خسارات احتمالی ناشی از عملکرد هوش مصنوعی را بر عهده بگیرند.

۶. نتیجه‌گیری

شکاف مسئولیت به عنوان مسئله‌ای ناشی از ظهور فناوری‌های هوش مصنوعی با قابلیت یادگیری و خودمختاری توسط ماتیاس معرفی شد. در حدود دو دهه‌ای از که صورت‌بندی این مسئله می‌گذرد، دیدگاه‌های مختلفی در واکنش به این مسئله ارائه شده‌اند. مجموع این دیدگاه‌ها را می‌توان ذیل سه رویکرد تقسیم‌بندی کرد: انحلال مسئله شکاف، پاسخ‌گویی با انتساب مسئولیت به هوش مصنوعی، و پاسخ‌گویی با انتساب مسئولیت به انسان.

مفهوم مسئولیت در کارکرد عرفی و اجتماعی، مفهومی ترکیبی، زمینه‌مند و درجه‌مند است. بر این اساس و با فرض تعبیری قوی و کارآمد از مفهوم مسئولیت، تنها رویکرد سوم،

یعنی انتساب مسئولیت به انسان، صورت‌بندی دقیق و درستی از مسئله شکاف مسئولیت دارد.

علاوه‌براین، چنین رویکردی ظرفیت ارائه راه حلی فعالانه و آینده‌نگر نسبت به مسئله شکاف مسئولیت دارد. با اصلاح و بازآرایی نقش‌های اجتماعی، می‌توان مسئولیت‌های جدیدی تعریف کرد و با تجهیز افراد به لوازم این مسئولیت‌ها شکاف مسئولیت در مورد اعمال فناورانه هوش مصنوعی را پر کرد؛ مشابه همان تدبیری که جوامع انسانی در طول تاریخ در مواجهه با پیامدهای آثار ناشی از عوامل غیرانسانی داشته‌اند.

با این حال، رسیدن به راه حل‌هایی با موفقیت عملی در این چارچوب، مستلزم غلبه بر چالش‌های مختلفی است. یکی از این چالش‌ها، طراحی هوش مصنوعی تبیین‌پذیر است. ابهام‌زدایی از سازوکار تولید خروجی هوش مصنوعی تا حد ممکن، و حفظ آگاهی نسبت به امکان‌های عملکردی هوش مصنوعی از لوازم مهم برای حفظ مسئولیت‌پذیری است. این چالش، هم مستلزم راهکارهایی فنی و هم سیاست‌گذارانه است. علاوه‌براین، صیانت از قدرت عمل نقش‌های متولّی اخلاقی بودن هوش مصنوعی نیز چالش‌برانگیز است. زیرا تقیّد به رعایت ملاحظات اخلاقی در توسعه فناوری، بالقوّه می‌تواند مزیت رقابتی را کاهش دهد و به همین دلیل جدی گرفته نشود. برای رفع این چالش نیز، باید تدابیری اندیشید.

پی‌نوشت‌ها

۱. چالش توزیع مسئولیت با نام مسئله «دست‌های بسیار» نیز شناخته می شود.

۲. تیگارد ادعا می‌کند شکاف مسئولیت مسئله‌ای نیست که به دنبال توسعه فناوری‌های هوش مصنوعی پدید آمده باشد، بلکه در زندگی روزمره و فعالیت‌های غیرفناورانه نیز مطرح می‌شود.

۳. کینر معتقد است معضل اصلی در ارتباط با وفور مسئولیت این است که از یک سو به دلیل محدودیت‌های عملی نمی‌توان تمام افراد مسئول را مورد بازخواست قرار داد و از سوی دیگر بازخواست کردن بخشی از افراد نیز با شائبه بی‌عدالتی و دل‌بخواهی بودن همراه است (Kiener, 2025: 369).

۴. هوش مصنوعی با قابلیت جواب‌گویی به عنوان هوش مصنوعی تبیین‌پذیر (Explainable AI / XAI) شناخته می‌شود.

۵. مثلاً استفاده از سکه برای تصمیم‌گیری در مورد اینکه به کدام یک از دو بیمار اورژانسی با شرایط یکسان و در حال مرگ رسیدگی کنیم، نمونه‌ای از تفویض بر اساس تصادف است. اما اگر به جای

سکّه، تصمیم‌گیری را به فردی که از ما شایستگی بیشتری دارد واگذار کنیم، تفویض بر اساس دلیل است.

۶. اگر در سیاق بحث حاضر، مفهوم مسئولیت اخلاقی را به عنوان ابزاری برساختی برای تنظیم اعمال افراد در جامعه در نظر بگیریم، آنگاه می‌توان گفت که سه مؤلفه مذکور به‌نحو امکانی مقوّم مفهوم مسئولیت هستند و برای مثال اگر در جامعه‌ای دیگر مفهوم مسئولیت اخلاقی به‌نحو دیگری به کار بسته شود، ترکیبی از مؤلفه‌هایی دیگر خواهد بود.

۷. چنین استدلالی علیه بسط‌گرایی اخلاقی، اصالتاً در حوزه اخلاق محیط زیست و در نقد بسط دایره امور اخلاقی از انسان‌ها به تمام طبیعت ارائه شده‌است (Heath, 2021: 8) امّا با تغییراتی می‌توان آن را در مورد بسط مسئولیت‌پذیری به قلمرو مصنوعات نیز طرح کرد.

کتاب‌نامه

Asaro, P. (2012). "On banning autonomous weapon systems: human rights, automation, and the design of ethical algorithms." International Review of the Red Cross.

Aristotle. (1984). Nicomachean ethics. In J. Barnes (Ed.), The complete works of Aristotle (Vol. 2, pp. 1729–1867). Princeton University Press.

Bales, A. (2025). "Against willing servitude: Autonomy in the ethics of advanced artificial intelligence." The Philosophical Quarterly.

Bringsjord Selmer, (2008), Robots: The Future Can Heed Us, *AI and Society*. Vol. 22 No.4, 539-550.

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

Danaher, J. (2016). Robots, law and the retribution gap. Ethics and Information Technology, 18(4), 299-309 https://doi.org/10.1007/s10676-016-9403-3

Danaher, J. (2022). Tragic choices and the virtue of techno-responsibility gap. Philosophy and Technology, 35(26), 1-26

Davidson, D. (2001). Essays on actions and events. Oxford University Press.

De Cremer, D. (2024). The AI-savvy leader: 9 ways to take back control and make AI work. Harvard Business Review Press.

Dignum, V. (2022). Responsible artificial intelligence: How to develop and use AI in a responsible way. Springer Nature.

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. Engaging Science, Technology, and Society, 5, 40–60.

European Commission. (2024). Report on liability for artificial intelligence and other emerging digital technologies.

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of responsibility in distributed moral actions. Philosophical Transactions.

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

Gunkel, D. J. (2023). Person, thing, robot: A philosophy of responsible AI. MIT Press.

Haselager, W. F. (2005). Robotics, philosophy and the problems of autonomy. Pragmatics & Cognition, 515-532

Heath, J. (2021). The failure of traditional environmental philosophy. Res Publica, 28, 1–16.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?. Ethics and information technology, 11(1), 19-29.

Jeppsson, S. (2022). Accountability, answerability, attributability: On different kinds of moral responsibility. In Oxford handbook of moral responsibility (pp. 73-90)

Johnson, D. G., & Verdicchio, M. (2019). AI, agency and responsibility: The VW fraud case and beyond. AI & Society, 34(3), 639-647.

Johnson, D. G., & Verdicchio, M. (2023). Ethical AI is not about AI. Communications of the ACM, 66(2), 32-34.

Kiener, M. (2025). AI and responsibility: No gap, but abundance. Journal of Applied Philosophy, 42(1), 357-374

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. University of Pennsylvania Law Review, 165(3), 633-705.

Latour, B. (2005). Reassembling the social: An introduction to actor-network-theory. Oxford University Press.

List C. and Pettit P., (2011), *Group Agency: The Possibility, Design, and Status of Corporate Agents*, Oxford University Press.

Matthias, A, (2004),), The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata, *Ethics and Information Technology*, p 175-183.

Munch, L., Mainz, J., & Bjerring, J. C. (2023). The value of responsible gaps in algorithmic decision-making. Ethics and Information Technology, 25(21), 1-11.

Nissenbaum, H. (1996). Accountability in a computerized society. Science and Engineering Ethics.

Rodogno R., (2016), Social robots, fiction, and sentimentality, *Ethics and Information Technology,* p. 257–268.

Roff, H. M. (2016). The strategic robot problem: Networked power and democratic accountability. Journal of Military Ethics.

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. Philosophy & Technology, 34(4), 1057-1084.

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical framework. Frontiers in ICT, 5, 15.

Scharre, P. (2018). Army of none: Autonomous weapons and the future of war. W. W. Norton & Company.

Sharkey, N. (2010). Saying no! to killer robots. Journal of Military Ethics, 9(4), 369-383.

Shoemaker, D. (2011). Attributability, answerability and accountability: Toward a wider theory of moral responsibility. Ethics, 121, 602-632.

Smith, A. M. (2015). Responsibility as answerability. Inquiry, 58(2), 99-126.

Smith, A. (2012). Attributability, answerability, and accountability: In defense of a unified account. Ethics, 122(3), 575-589

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

Strawson, P. F. (1974). Freedom and resentment. In Freedom and resentment and other essays (pp. 1–25). Oxford University Press.

Sullins, J. (2006). When is a robot a moral agent? International Review of Information Ethics, 6, 24-30

Tigard, W. D. (2021). There is no techno-responsibility gap. Philosophy and Technology, 34,. 589-607

Van Inwagen, P. (1983). An essay on free will. Oxford University Press.

Verbeek, P.-P. (2011). Moralizing technology: Understanding and designing the morality of things. University of Chicago Press.

Watson, G. C. (1996). Two faces of responsibility. Philosophical Topics, 24, 227-248.