Journal of Philosophical Theological Research
ISSN Print: 1735-9791 Online: 2538-2500

Artificial Intelligence and Meaningful Human Control: Agent-Neutral-Reason-Responsive Mechanisms and the Possibility of Tracing Responsible Agent(s)

Keyvan Alasti

Assistant professor, National Research Institute for Science Policy (NRISP), Tehran, Iran.
keyvan.alasti@gmail.com

Abstract

Some ethical issues related to artificial intelligence relate to the nature of autonomy and the problem of the responsibility gap; autonomous AI systems have behaviors that would be called "decisions" if humans made them, and humans would be held responsible if those "decisions" caused harm. However, when an "action" is performed by autonomous technology, although there is still a need for responsibility, nothing may be held responsible for it. Therefore, it is suggested that autonomous technologies remain under human responsibility if they meet two (design) requirements: 1) the ability to "track" human reasons and 2) by following the chain of human reasons, a human agent can be "traced." These conditions will create many ambiguities for autonomous AI systems; for example, the moral reasons that need to be tracked are often agent-neutral, but tracing a human agent is only possible through agent-dependent reasons. In this article, by analyzing the interpretations of the two conditions, an interpretation that has not been proposed before will be defended as the least problematic interpretation. To this end, it is argued that the responsibility gap is solved when the history of events produced by autonomous systems involves a process by which the agent takes responsibility for an agent-neutral normative reason-responsive mechanism.

Research Article

Keywords: meaningful human control (MHC), ethics of artificial intelligence, ethics of technology, responsibility gap, autonomous artifacts, reason-responsiveness.

Received: 2024/10/20 ; **Received in revised form:** 2024/11/16 ; **Accepted:** 2024/12/09 ; **Published online:** 2024/12/22

Alasti, K. (2024). Artificial Intelligence and Meaningful Human Control: Agent-Neutral-Reason-Responsive Mechanisms and the Possibility of Tracing Responsible Agent(s). *Journal of Philosophical Theological Research (Philosophy of Ethics and Technology: challenges and prospects special Issue)*, 26(4), 27-54. https://doi.org/10.22091/jptr.2025.11378.3139

Introduction

According to Santoni de Sio and van den Hoven (referring to Fischer and Ravizza's compatibilist approach to human responsibility), autonomous artifacts remain meaningfully under human control if they meet two specific conditions: 1) The artifact can "track" human reasons, and 2) a human agent can be "traced" by following a chain of human reasons. Fulfilling these two conditions still faces problems and ambiguities that continue to be the subject of philosophical and empirical debate.

Two Interpretations of Tracing

In describing the second condition (originally for human responsibility), Fischer and Ravizza point out that responsibility is essentially historical. Every act has a certain causal history, and the attribution of responsibility for actions to an agent depends on the nature of the causal chain that shaped that history. The historical nature of responsibility can be interpreted (at least) in two different ways.

In one interpretation, responsibility is attributed to the person who first intentionally (i.e., with reason) performed an effective act. Therefore, an event (such as killing people on the road) includes a purposeful act (such as drinking or forcefully feeding alcohol) in the causal history of those events, and therefore responsibility for the event is attributed to the agent of that historical purposeful act. A drunk driver, although driving unintentionally, is responsible for accidents and deaths on the road because he/she drank even though he knew he wanted to drive. If he/she were forcibly fed alcohol, the agent who fed the alcohol to the driver would be responsible for the events. So, responsibility for a bad event caused by an autonomous car historically goes back to a human who made an informed decision.

In another interpretation, taking responsibility is not simply an act. Responsibility is attributed to an agent when the history of events includes a process of taking responsibility by an agent, in which a set of beliefs is formed for users or stakeholders; for example, in the process of taking responsibility, the belief is formed that an autonomous AI system belongs to a human agent. The set of beliefs is part of (not the causal chain, but) the mental or conceptual space of human beings who may not play any direct role in the causal chain.

The two interpretations will be effective in the argument presented about tracing conditions.

Taking responsibility for an agent-Neutral-Reason-responsive AI

Reasons may be motivational or normative. Motivational reasons are the reasons why an agent acted, and normative reasons are the reasons that justify an action morally or rationally. We are interested in having an AI system track normative reasons. However, distinguishing normative from motivational reasons requires a normative theory and is not directly possible for the system.

Another issue is that Fisher and Raviza's conditions of responsibility were originally defined for humans (not AI systems). Since they are defined for humans, they have assumptions that are not necessarily present in the use of artificial intelligence (AI). For example, it is assumed that when humans are responsible for their actions, not only the decision mechanism but also the reasons to which the decision mechanism responds are the agent's own.

This is crucial because, as I have argued when we attribute responsibility to AI systems (following the approach of Santoni De Sio and van den Hoven), the assumption that reasons belong to certain human agents is not necessarily true.

Such an assumption has also been challenged by others in recent years, including in the argument presented by Veluwenkamp (2022). He argues that being reason-responsive is incompatible with moral theories that consider agent-neutral normative reasons. To this end, he

distinguishes between agent-dependent reasons and agent-neutral reasons. Then he argues that tracking is important in establishing who is responsible, and if all reasons are agent-neutral, tracking cannot play its role. He concludes that if normative reasons are agent-neutral, then not all reasons traced will be normative.

Veluwenkamp's argument, by applying agent-neutral and agent-dependent division in Santoni de Sio and Van Den Hoven's MHC conditions, implicitly emphasizes that it is possible for a mechanism to track reasons that do not belong to the agent(s) that are supposed to be responsible. This conflicts with the assumption that reasons always belong to the agent.

Although it has been claimed that the tracing condition requires the reasons being tracked to be partially agent-dependent, the claim does not guarantee the opposite. Tracking reasons does not necessarily require tracing an agent. Therefore, the claim allows the mechanism to track only agent-neutral reasons.

So according to the proposed approach, tracking can be realized as a normative reason-responsive mechanism, and it can be completely agent-neutral. In other words, a normative reason-responsive mechanism can be realized without any connection with an agent, and the agent-neutral reason-responsive mechanism, although it does not guarantee to trace an agent, makes the possibility of being only responsive to normative agent-neutral reasons.

I have argued that the possibility of two different interpretations of the historical nature of responsibility (and indeed the tracing condition) raises the hope that, at least in one interpretation, there is the possibility that an agent-neutral reason-responsive mechanism also has the tracing condition.

It is argued that the relationship between an agent and reasons, (or an agent and a reason-responsive mechanism) can be differently interpreted. In other words, the relationship between a reason-responsive mechanism and an agent can be constructed during the process of "taking responsibility." So, there should be a social process during which beliefs are constructed in a way that an agent-neutral reason-responsive mechanism becomes an agent's own.

Conclusion

Although in Fischer and Ravizza's approach, both the decision mechanism (based on second condition) and the reasons (as a presupposition) are an agent's own, in the case of autonomous weapons, it can be intuitively assumed that only one of these two (that is, either the decision mechanism, or normative reasons that it tracks) during a process will be an agent's own, and based on that, the responsibility for the results of using an autonomous AI weapon can be attributed to the agent.

Regarding this, we can expect that the two conditions of Santoni de Sio and van den Haven will be fulfilled, without the reasons used going beyond the normative reasons (which are the reasons justifying actions based on a moral theory).

References

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.

Fischer, J. M., & Ravizza, M. (2000). Précis of responsibility and control: A theory of moral responsibility. *Philosophy and Phenomenological Research,* 61(2), 441-445. https://doi.org/10.2307/2653660

Frankfurt, H. (1993). What we are morally responsible for. *Perspectives on Moral Responsibility*, 286-294.

Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy*, 66(23),

829-839. https://doi.org/10.2307/2023833

Hindriks, F., & Veluwenkamp, H. (2023). The risks of autonomous machines: From responsibility gaps to control gaps. *Synthese*, 201(1), 21. https://link.springer.com/article/10.1007/s11229-022-04001-5

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, 8, 195-204. https://link.springer.com/article/10.1007/s10676-006-9111-5

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

Mecacci, G., & Santoni de Sio, F. (2020). Meaningful human control as reason-responsiveness: The case of dual-mode vehicles. *Ethics and Information Technology*, 22(2), 103-115. https://link.springer.com/article/10.1007/s10676-019-09519-w

Müller, Vincent C. (2023). Ethics of artificial intelligence and robotics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 15. https://link.springer.com/article/10.1007/s11023-022-09608-8

Veluwenkamp, H. (2022). Reasons for meaningful human control. *Ethics and Information Technology*, 24(4), 51. https://link.springer.com/article/10.1007/s10676-022-09673-8

پژوهش‌های فلسفی کلامی
Print: 1735-9791
Online: 2538-2500

هوش مصنوعی و کنترل معنادار بشری: سازوکارهای واکنش‌پذیر به دلایلِ نسبت–به–عامل–خنثی و امکانِ ردگیری عامل(های) مسئول

کیوان الستی

استادیار، مؤسسه تحقیقات سیاست علمی کشور، تهران، ایران. keyvan.alasti@gmail.com

علمی – پژوهشی

چکیده

بخشی از مسائل اخلاقی مرتبط با هوش مصنوعی به خصلت خودگردانی و مسئله شکاف مسئولیت مرتبط است. سیستم‌های خودگردانِ مجهز به هوش مصنوعی قادر به ایجاد رویدادهایی هستند که اگر انسان آنها را ایجاد کرده بود به آنها «تصمیم» گفته می‌شد؛ و اگر این «تصمیم‌ها» خسارتی به بار می‌آورد، انسانی مسئول آن می‌شد. اما وقتی تغییرات را ابزارهای تکنولوژیک خودگردان ایجاد می‌کنند، هرچند همچنان نیاز به پذیرشِ مسئولیت وجود دارد، اما ممکن است چیزی مسئول آن قلمداد نشود. لذا پیشنهاد شده که تکنولوژی‌های خودگردان در صورتی تحتِ مسئولیت انسان باقی بمانند که دو الزام را برآورده کنند: (۱) این قابلیت که دلایل انسانی را «ردیابی» کنند؛ و (۲) با دنبال کردن زنجیرهٔ دلایل انسانی یک عاملِ انسانی «ردگیری» شود. این شروط برای مصنوعات تکنولوژیک و هوش مصنوعی خودگردان ابهامات متعددی را ایجاد خواهد کرد. از آن جمله این است که دلایلِ اخلاقی‌ای که باید ردیابی شوند غالباً نسبت به عامل(ها) خنثی هستند، اما ردگیری یک عاملِ انسانی تنها از طریق دلایل وابسته–به–عامل امکان‌پذیر است. در این مقاله با تحلیل تعبیرهایی که از دو شرط مطرح‌شده ارائه می‌شود از یک تعبیر که پیش از این پیشنهاد نشده به عنوان کم‌مسئله‌ترین تعبیر از جهت حل مسئلهٔ ذکرشده دفاع خواهد شد. برای این کار استدلال می‌شود که شکاف مسئولیت زمانی پر می‌شود که تاریخچه رویدادهایی که توسط سامانه‌های خودگردان ایجاد می‌شود شامل فرایندی باشد که طی آن عامل مجموعه‌ای از دلایل هنجاری را به عنوان دلایل قابل‌قبول خود بپذیرید. اگر دلایلی که مطابق آنها سیستم به صورت خودگردان رفتار می‌کند متعلق به یک عامل بشری باشد، آنگاه مسئولیت رویدادهایی که ایجاد می‌شود نیز متعلق به آن عامل خواهد بود.

کلیدواژه‌ها: کنترل معنادار بشری، اخلاق هوش مصنوعی، اخلاق تکنولوژی، مسئولیت‌پذیری، شکاف مسئولیت، مصنوعات خودگردان، واکنش‌پذیری به دلیل.

تاریخ دریافت: ۱۴۰۳/۰۷/۲۹ ؛ **تاریخ اصلاح:** ۱۴۰۳/۰۸/۲۶ ؛ **تاریخ پذیرش:** ۱۴۰۳/۰۹/۱۹ ؛ **تاریخ انتشار آنلاین:** ۱۴۰۳/۱۰/۰۲

▣ الستی، کیوان (۱۴۰۳). هوش مصنوعی و کنترل معنادار بشری: سازوکارهای واکنش‌پذیر به دلایلِ نسبت–به–عامل–خنثی و امکانِ ردگیری عامل(های) مسئول. *پژوهش‌های فلسفی–کلامی (ویژه‌نامه فلسفه اخلاق و فن‌آوری: چالش‌ها و چشم‌اندازها)*، ۲۶(۴)، ۲۷–۵۴. https://doi.org/10.22091/jptr.2025.11378.3139

مقدمه

سامانه‌های مجهز به هوش مصنوعی، بیش از آن که «هوش» به معنای سنتی آن باشند، یک مصنوعِ تکنولوژیک (از نسل جدید) هستند، و همانند هر تکنولوژی جدیدی مسائل اخلاقی و اجتماعی متعددی در اطراف توسعهٔ آن‌ها مطرح می‌شود. بخشی از مسائل اخلاقی مرتبط با سامانه‌های مجهز به هوش مصنوعی به خصلت خاص خودگردانی[1] و مسئله شکاف مسئولیت[2] مربوط می‌شود. مسئله این است که سیستم‌های خودگردان مجهز به هوش مصنوعی، از جمله اسلحه‌های خودگردان یا اتومبیل‌های خودران، قادر به ایجاد رویدادهایی هستند که اگر انسان آن‌ها را ایجاد کرده بود نام «تصمیم» بر آن‌ها گذاشته می‌شد، به نحوی که اگر این «تصمیم‌ها» خسارتی به بار می‌آورد، انسانی مسئول، پاسخگو یا مدیون می‌شد. اما وقتی این تغییرات را ابزارهای تکنولوژیکی ایجاد می‌کنند که به لحاظ فنی (و نه الزاماً اجتماعی) خودگردان هستند، هرچند همچنان نیاز به پذیرشِ مسئولیت وجود دارد، ممکن است کسی (یا چیزی) مسئول آن‌ها قلمداد نشود. این مسئله‌ای است که تحت نام «شکاف مسئولیت» شناخته می‌شود و از آنجا که بخشی از کنترل اجتماعی از طریق برساخت‌هایی همانند «مسئولیت» صورت می‌گیرد، شیوع آن می‌تواند مسئله‌ساز باشد.

امیدی که ممکن است در حل این مسئله وجود داشته باشد تکیه بر این حقیقت است که اموری که توسط هوش مصنوعی خودگردان انجام می‌شود در نهایت فارغ از دخالت (هرچند غیرمستقیم) انسان‌ها نیست. و حضور افراد انسانی در حلقهٔ تصمیم[3] این امید را ایجاد می‌کند که شاید بتوان به آن فرد (یا افراد) انسانی مسئولیت نسبت داد. با این حال، این امید تنها در صورتی محقق خواهد شد که فرد (یا افراد) انسانی حاضر در حلقه تصمیم توانایی *کنترل* موضوعی را که نسبت به آن مسئول هستند داشته باشد (Wu *et al.*, 2022; Dautenhahn, 1998).

سانتونی د سیو و ون دن هاون (Santoni de Sio & Van den Hoven, 2018)، با رویکردی که وام‌دار دیدگاه فیشر و رویتزا (Fischer & Ravizza, 1998) در خصوص مسئولیت است، پیشنهاد داده‌اند که تکنولوژی‌های خودگردان در صورتی می‌توانند تحت کنترل انسان باقی بمانند (و یا به تعبیری تنها زمانی می‌توان به سیستمِ خودگردان مجهز به هوش مصنوعی مسئولیت نسبت داد) که سازوکار آن به نحوی طراحی شده باشد که دو الزام طراحی[4] را محقق کرده باشند: (۱) این قابلیت که دلایل انسانی را «ردیابی» کنند و (۲) با دنبال کردن زنجیرهٔ دلایل انسانی یک عامل انسانی «ردگیری» شود.

هرچند این دو شرط مطابق و معادل شرایط مسئولیت‌پذیری در برخی نظریات مرتبط با انسان هستند،

1. Autonomy
2. Responsibility Gap
3. human in the loop
4. design requirement

به کار بردن آنها برای مصنوعات تکنولوژیک و هوش مصنوعی خودگردان (و نه انسان) ابهامات متعددی را ایجاد خواهد کرد. تحلیل این ابهامات نشان می‌دهد که می‌توان از آنچه مطرح می‌شود تعبیرهای متفاوت و در نتیجه عملکرد و نتایج متفاوت استخراج کرد.

در این مقاله با تحلیل و استخراج تعبیرهایی که از دو شرط مطرح‌شده می‌توان ارائه کرد، از یک تعبیر که پیش از این پیشنهاد نشده (به عنوان کم‌مسئله‌ترین تعبیر) دفاع خواهد شد. در این مقاله استدلال خواهد شد که شکاف مسئولیت زمانی پر خواهد شد که تاریخچهٔ رویدادهایی که سامانه‌های خودگردان ایجاد کرده‌اند شامل فرایندی باشد که طی آن عاملْ مجموعه دلایل خاصی از دلایل هنجاری را به عنوان دلایل قابل‌قبول خود بپذیرید. اگر دلایلی که مطابق آن‌ها سیستم خودگردان رفتار می‌کند متعلق به یک عامل بشری باشد، آنگاه مسئولیت رویدادهایی که ایجاد می‌شود نیز متعلق به عامل خواهد بود.

برای روشن ساختن این استدلال ابتدا شرح مختصری از مسئله فناوری‌های خودگردان مجهز به هوش مصنوعی و مسئله شکاف مسئولیت ارائه خواهد شد. سپس با مطرح کردن دیدگاه فیشر و رویتزا در مورد مسئولیت‌پذیری انسان و به کار بستن آن در مورد مسئولیت‌پذیری سیستم‌های مجهز به هوش مصنوعی توسط سانتونی د سیو و ون دن هاون، دو شرط مطرح‌شده برای مسئولیت‌پذیر بودن این سیستم‌ها شرح داده می‌شود. در بخش بعدی، تعابیر متفاوت از دو شرط به ترتیب ذکر خواهد شد و استدلال خواهد شد که این تعابیر مانعةالجمع هستند. در نهایت، با شرح دیدگاه ولوونکامپ دربارهٔ ارتباط میان دو شرط ردگیری و ردیابی –این پیشنهاد که فرایند ردگیری در فرایند ردیابی نقش عمده‌ای ایفا می‌کند– از پیشنهاد جدید به عنوان تعبیر قابل‌قبول دفاع خواهد شد.

آندریاس ماتیاس (Matthias, 2004) اولین بار صورت‌بندی مهمی از موضوع شکاف مسئولیت در هوش مصنوعی ارائه می‌کند. او استدلال کرده است که ماشین‌های یادگیری[1]، که مستقل از انسان و بر اساس شبکه‌های عصبی عمل می‌کنند، موقعیتی را ایجاد کرده‌اند که در آن طراح یا اپراتور –که پیش از این مسئولیت اَعمال ماشین به او نسبت داده می‌شد– دیگر قادر به پیش‌بینی رفتار ماشین در آینده نیست، و لذا دیگر نمی‌تواند از نظر اخلاقی مسئول قلمداد شود. بر این مبنا یا نباید این نوع از ماشین‌ها را توسعه داد یا باید وجود این شکاف در فضای مسئولیت‌پذیری را پذیرفت.

از سوی دیگر، این امکان که خود مصنوع به نحوی طراحی شود که بتوان به آن مسئولیت نسبت داد نیز از مواردی است که کمتر امیدبخش قلمداد شده است. در این مورد افرادی همانند سولینز، فلوریدی و ساندرز، آندرسون، و جانسون (Sullins, 2011; Floridi & Sanders, 2004; Anderson, 2011; Johnson, 2006) مسئلهٔ نسبت دادن عاملیت اخلاقی، و به تبع آن مسئولیت اخلاقی، را به مصنوعات خاصی مثل هوش مصنوعی بررسی کرده‌اند. جانسون (Johnson, 2006) معتقد است که هرچند مصنوعات هوش مصنوعی

1. Machin Learning

هویت‌های اخلاقی هستند، به این معنا که موضوع احکام اخلاقی هستند، اما نمی‌توان به آنها عاملیت اخلاقی و در نتیجه مسئولیت اخلاقی نسبت داد. فلوریدی و ساندرز (Floridi & Sanders, 2004) نیز هرچند نسبت دادن عاملیت اخلاقی به هوش مصنوعی را امکان‌پذیر می‌دانند، قابلیت مسئولیت‌پذیری را مجزا و غیرقابل‌اطلاق به هوش مصنوعی قلمداد می‌کنند. از این جهت فرض نسبت دادن مسئولیت به افراد انسانی (و نه مصنوع) هنوز اقبال بیشتری دارد. از این جهت مسئله شکاف مسئولیت هوش مصنوعی همچنان نیازمند نوعی از کنترل بشری تصور می‌شود.

با این حال کاربر و طراح، یا هر فرد دیگری که در حلقه تصمیم قرار دارد، برای این که قادر باشد مسئولیت اعمال مصنوع را بپذیرد، باید از چرایی عمل هوش مصنوعی توضیحی (غالباً اخلاقی) در دست داشته باشد. مسئلهٔ توضیح‌پذیری هوش مصنوعی نیز از مواردی است که در مورد آن صحبت شده و نقدهای بسیاری به آن مطرح شده (برای مثال، نک. Robbins, 2019) و اهمیت آن نیز مورد مناقشه قرار گرفته است (Barman *et al.*, 2024).

شروط ون دن هاون و سانتونی د سیو، چنان که شرح داده می‌شود، رویکرد متفاوتی برای پرداختن به موضوع شکاف مسئولیت در خصوص فناوری‌های خودگردان است.

هوش مصنوعی خودگردان، مفهوم خودگردانی و شکاف مسئولیت

شرح مسئله

تکنولوژی‌های خودگردان تکنولوژی‌هایی هستند که رفتارهای برآمده از آن‌ها مستقل از دخالت مستقیم یک کاربر انسانی و صرفاً در نتیجهٔ پردازش‌های درونی آن‌ها تعیین می‌شود. عامل‌های بیرونی، از جمله کاربرها و اپراتورها، تنها به صورت غیرمستقیم در تعیین این که تکنولوژی خودگردان چه رفتاری نشان دهد نقش دارند. به عبارت دیگر، اگر مصنوع (یا سیستم) به نحوی باشد که بتوان مرزی میان محدوده «درون» و «بیرون» آن قائل شد، یا بتوان سازوکارهای «درونی» را از عامل‌های «بیرونی» مجزا کرد، آنگاه مصنوع (یا سیستم) خودگردان آن‌هایی هستند که به نحوی طراحی (و ایدئال) شده‌اند که رفتار آن‌ها غالباً مبتنی بر سازوکارها و پردازش‌های درونی است.

رفتارهای این تکنولوژی‌ها، به واسطهٔ خودگردان بودنشان، ممکن است پیش‌بینی‌پذیر نباشد. پیش‌بینی‌ناپذیری رفتار آن‌ها غالباً به واسطهٔ ویژگی‌های سازوکارها و نوع پردازش‌های درونی آن‌ها ایجاد می‌شود، که از میان این ویژگی‌ها می‌توان به (۱) پیچیدگی ساختار درونی، (۲) سرعت پردازش بالا، و (۳) کدر بودن سازوکار درونی و موارد مشابه دیگر اشاره کرد. لذا اگر یک سامانه مجهز به هوش مصنوعی، بعد از انجام عملیات و پردازش درونی، بین دو گزینهٔ الف و ب مثلاً الف را «انتخاب» کند، آنگاه هرچند هیچ ارادهٔ آزادی در این انتخاب دخیل نبوده است، اما به واسطهٔ آنچه نتیجهٔ پردازش درونی این سیستم است می‌توان سناریوی انتخاب جایگزین (یعنی ب) را نیز برای آن تصور کرد، فارغ از این که

آن سازوکار درونی قابل‌مشاهده یا قابل‌فهم توسط افراد بیرونی باشد یا خیر.

به طور خلاصه، می‌توان گفت که نوعی عدم‌قطعیت و لذا پیش‌بینی‌ناپذیری در رفتارهای این مصنوعات وجود دارد که به واسطهٔ ویژگی‌های سازوکارهای درونی آنها ایجاد شده است. پردازش‌های درونی در سیستم‌های خودگردان مجهز به هوش مصنوعی غالباً قرار است به نحوی مستقل از عامل‌های بیرونی و به دور از دخالت مستقیم بشر انجام شوند و رفتاری که در نهایت از این پردازش حاصل می‌شود رفتاری قابل‌اعتماد و به یک معنا «عقلانی» باشد.

(الف) شکاف مسئولیت: هر جا عدم‌قطعیت در رفتار با اثرات بالا وجود داشته باشد، لازم است مسئولیت نتایج را بتوان به کسی نسبت داد. در ادبیات فلسفی غالباً خصلت خودگردانی را شرط لازم برای امکان نسبت دادن مسئولیت قلمداد کرده‌اند. به عبارت دیگر، هر آنچه مسئولیت‌پذیر است، الزاماً شرط خودگردانی را دارا خواهد بود. همچنین کنترل اجتماعی موجودات خودگردان با سازوکاری از جمله نسبت دادن مسئولیت انجام می‌شود که نه تنها نیازمند خودگردانی، بلکه حافظ آن نیز هست. به عبارت دیگر، نسبت دادن مسئولیت ابزاری اجتماعی است که از طریق آن موجوداتی که درصدی از خودگردانی را دارا هستند (از انسان‌ها گرفته تا نهادهای اجتماعی و...) بدون این که خودگردانی‌شان به خطر بیفتد، تحت «کنترل» اجتماعی قرار می‌گیرند.

با این حال خودگردانی یک موجود غالباً شرط کافی را برای این که بتوان مسئولیت به آن نسبت داد ایجاد نخواهد کرد (Muller, 2023). به عبارت دیگر، برای مسئولیت‌پذیری شرایطی بیشتر از خودگردانی لازم است. از این جهت، رفتاری بسیار مؤثر ممکن است از این سازوکارهای خودگردان سر بزند بدون این که این قابلیت وجود داشته باشد که به آن‌ها مسئولیتی نسبت داده شود. از این جهت امکان کنترل اجتماعی اَعمالی که به واسطهٔ هوش مصنوعی خودگردان انجام می‌شود از طریق واژه‌هایی همانند «مسئولیت» دیگر امکان‌پذیر نیست.

فارغ از این که تکنولوژی‌های دارای هوش مصنوعی درآینده در شرایطی باشند که بتوان به آن‌ها مسئولیت -به معنایی که به انسان نسبت داده می‌شود- نسبت داد یا نه،[1] این تکنولوژی‌ها (به خصوص مواردی همانند اسلحه‌های خودگردان که تأثیر عملکرد آن‌ها وسعت زیادتری دارد) در حال حاضر موجودند، بدون این که بتوان به آن‌ها مسئولیت نسبت داد. این موضوعی است که اصطلاحاً با عنوان «شکاف مسئولیت» شناخته می‌شود و نیازمند پاسخ است. یکی از راه‌حل‌های ارائه‌شده ناظر به این موضوع با نام «کنترل معنادار بشری»[2] شناخته شده است.

(ب) خودگردان بودن و شکاف مسئولیت: این که رفتار یک سیستم این قابلیت را دارد که به صورت

---

۱. در این مورد برخی فلاسفه امید بیشتری دارند و برخی دیگر به متحقق شدن آن مشکوک هستند.

2. meaningful human control

درونی گزینش‌هایی داشته باشد، هرچند ممکن است در تعریف خودگردان بودن بگنجد، اما الزاماً مسئله اخلاقی‌ای مانند شکاف مسئولیت را ایجاد نخواهد کرد. برای مثال، چیزی شبیه به دماسنج یا ترموستات، که سازوکار درونی آنها پیچیدگی کمی دارد، نیز در حال گزینش میان دو گزینه (مثل الف یا ب) است. با این حال، شرایط آن‌ها از این جهت متفاوت است که پردازش درونی آن‌ها ساده و پیش‌بینی‌پذیر است.

همچنین این طور نیست که در هر شرایطی که سازوکارْ گزینشی و پیش‌بینی‌ناپذیر[۱] (یا به سختی پیش‌بینی‌پذیر) است، مسئله شکاف مسئولیت نیز حاضر باشد. گاهی مسئله‌ای که با آن مواجه هستیم مرتبط با ریسک است، نه عدم‌قطعیت. ریسک را از این جهت که محاسبه‌شدنی است متفاوت با عدم‌قطعیت می‌دانند. برای مثال، ما به تاس (که می‌تواند به عنوان سازوکاری برای گزینش از میان ۱ تا ۶ در نظر گرفته شود) خودگردانی نسبت نمی‌دهیم. لذا اگر در یک سناریوی فرضی نتیجه پرتاب یک تاس اثرات وسیعی در یک جامعه می‌گذارد، باز وضعیت ایجادشده را دارای شکاف مسئولیت قلمداد نمی‌کنیم. زیرا در مورد نتایج حاصل از تاس انداختن امکان محاسبه و اندازه‌گیری ریسک وجود دارد.

اما رفتار (یا انتخاب) هوش مصنوعی اولاً این تصور را ایجاد می‌کند که از سازوکاری منطقی و عقلانی حاصل شده و در نتیجه می‌توان با استفاده از همان منطق توضیحی منطقی برایش ارائه کرد، در حالی که در بسیاری موارد الگوریتم آن قابل‌دریافت توسط فاهمه عادی انسان نیست. و ثانیاً از آنجا که نتیجه دارای عدم‌قطعیت است، امکان واسپاری موضوع به فرایندهای فنی‌تر «محاسبه ریسک» وجود ندارد.[۲]

ناسازش‌گرایی: تقابل اراده آزاد و موجبیت

تا پیش از این مفهوم مسئولیت به مقوله‌ای خاص‌تر از آنچه امروز خودگردانی می‌نامند یعنی به اراده آزاد اشاره داشت، که مشخصاً مرتبط با انسان بود. از این جهت انسان تنها موجودی بود که می‌شد به آن مسئولیت نسبت داد. اراده آزاد، به معنای امکان انتخاب آزاد میان دو گزینه، با مفاهیم دیگری مثل موجبیت، که ناشی از پیشرفت‌های فیزیک در قرن هفدهم است، ناسازگاری دارد. بنا بر اصل موجبیت، شرایط حال حاضر جهان را شرایط پیش از آن تعیین کرده است. دست‌کم از منظر کسانی که «ناسازش‌گرا[۳]» خوانده می‌شوند، این اصل با اراده آزاد و این که عامل‌ها بتوانند بر مبنای تصمیم شخصی مسیر از پیش تعیین‌شده را تغییر دهند در تضاد است. از ناسازش‌گرایانی که موجبیت را دارای اهمیت می‌دانند نسبت دادن مسئولیت به افراد ناممکن است.

---

۱. در تعداد وقوع محدود.

۲. هندریکس و ولوونکامپ (Hindriks & Veluwenkamp, 2023) همچنان استدلال می‌کنند که موضوع هوش مصنوعی خودگردان بیشتر از این که به شکاف مسئولیت مرتبط باشد به ریسک مربوط می‌شود.

3. incompatibilism

سازش‌گرایی، همگرایی موجبیت و مسئولیت‌پذیری

سازش‌گرایی: در مقابل ناسازش‌گرایی، رویکرد دیگری وجود دارد که مسئولیت‌پذیری حاصل از ارادهٔ آزاد و موجبیت علّی را قابل جمع می‌داند. مطابق این رویکرد، حتی با وجود پذیرش موجبیت علّی در جهان، همچنان این امکان وجود دارد که به انسان‌ها مسئولیت نسبت داده شود. این رویکرد را سازش‌گرایی[1] می‌نامند. سازش‌گرایی برقرار بودن موجبیت علّی را فرض می‌گیرد،[2] و همزمان بر این اصرار دارد که می‌توان به برخی موجودات، به طور مشخص انسان‌ها، مسئولیت نسبت داد. در پاسخ به این سؤال که چه چیزی، یا چه خصلتی، مرز میان موجوداتی که شهوداً به آن‌ها مسئولیت نسبت می‌دهیم و موجودات دیگری که تمایلی شهودی به نسبت دادن مسئولیت به آن‌ها نداریم را تعیین می‌کند، دیدگاه‌های متفاوتی از ابتدای دوره مدرن (مثل دیوید هیوم و تامس هابز) تا دورهٔ معاصر (مثل هری فرانکفورت) مطرح شده است.

در روایت کلاسیک سازش‌گرایی، متعلق به فلاسفهٔ متقدم‌تر مانند هیوم و هابز، آنچه شهود ما را برای نسبت دادن مسئولیت به افراد، در شرایطی که موجبیت پذیرفته شده است، تحریک می‌کند، نه وجود اراده آزاد، بلکه شرطی خفیف‌تر یعنی خودگردان بودن (یا نبودن) است (Santoni de Sio and Van den Hoven, 2018, p. 5). شهودی که در پس این ایده وجود دارد این است که کسانی که به آنها مسئولیت نسبت می‌دهیم الزاماً به معنای فلسفی دارای ارادهٔ آزاد نیستند، بلکه کسانی‌اند که می‌توان خصلت خودگردانی را به آنها نسبت داد. به عبارت دیگر، رفتار آن‌ها محصول فرایندی است که در درون مغز انجام می‌شود. سازوکار درون-مغزیِ فردی ممکن است این انگیزه را برای او ایجاد کند که مثلاً وجهی از پول را به کسی که فکر می‌کند به آن نیاز دارد تحویل دهد. اما شرایط زمانی متفاوت خواهد بود که این فرد با اسلحه تهدید شده باشد تا همان مقدار پول را تحویل دهد. در موقعیت اول، او مسئول عملش است، اما به فرد در موقعیت دوم چنین مسئولیتی نسبت داده نمی‌شود. نوع عملی که فرد اول انجام می‌دهد، از جمله خواست او برای تحویل پول، به نحوی است که می‌توان او را مسئول عملش قلمداد کرد. این در حالی است که کسی که با اسلحه تهدید شده چنین مسئولیتی ندارد. هرچند فرض ابتدایی موجبیت در رفتارهای انسانی همچنان وجود دارد، اما شرایط مسئولیت‌پذیری فرد در موقعیت اول، چون عمل او صرفاً مبتنی بر سازوکارهای علّی درونی بوده، با فرد در موقعیت دوم که تهدید شده تا عملی را انجام دهد متفاوت قلمداد می‌شود. اگر این روایت، که وجه تمایز را خودگردانی می‌داند، جدی بگیریم، دست‌کم در خصوص تکنولوژی‌های خودگردان مسئله‌ای به نام شکاف مسئولیت وجود نخواهد داشت.

در رویکردهای کلاسیک، شرطِ خودگردان بودن مهم‌ترین شرط برای مسئولیت‌پذیری قلمداد شده است. اما دیدگاه‌های جدیدتر نشان می‌دهند که شروط دیگری نیز با همان درجه از اهمیت وجود دارد.

---

1. compatibilism

۲. قطعاً یافته‌های فیزیک کوانتوم در شرایط این مسئله تغییری ایجاد نمی‌کند.

روایت کلاسیک شهود ناقصی از موضوع ایجاد می‌کند. زیرا مثال‌هایی می‌توان زد که این تصور را که خودگردانی برای داشتن مسئولیت کافی است تضعیف کنند.[۱] افرادی که دارای بیماری‌های حاد روانی هستند و به همین سبب نه عقلانیت بلکه هیجاناتی بسیار قدرتمند رفتارهای آن‌ها را هدایت می‌کند، یا افرادی که به واسطهٔ یک ترس بیمارگون شدید از نجات دادن یک کودک از خطر تصادف اجتناب می‌کنند، به معنایی که گفته شد، دارای خودگردانی هستند. به عبارت دیگر، هرچند رفتار آن‌ها نیز از سازوکار درونی محصور در بدن آن‌ها حاصل شده، اما اطلاع ما از وجود یک بیماری روانی خاص باعث می‌شود آنها را مسئول اعمالشان قلمداد نکنیم. و این انتظار را نیز داشته باشیم که در صورتی که دعوایی حقوقی علیه رفتار این افراد مطرح شد، اطلاع دستگاه قضایی از بیماری مورد نظر حتی شاهدی برای تبرئه آنان محسوب شود. از این جهت نسبت دادن مسئولیت به شرایطی بیش از صرفِ مبتنی بودن رفتار بر یک سازوکار علّی درونی نیازمند است. چیزی که وجه تمایز ایجاد می‌کند این است که رفتار مسئولانه بر مبنای دلیل‌های عقلانی انجام می‌شود. لذا ردیابی نوعی از انتخاب عقلانی (مبتنی بر دلیل) به شروط مسئولیت‌پذیری اضافه خواهد شد. لذا دلایل عقلانی، فراتر از صرف علت‌ها، نقشی در مسئولیت‌پذیر بودن ایفا خواهند کرد. به طور خلاصه، به موجوداتی مسئولیت نسبت داده می‌شود که رفتار آن‌ها به دلایل حساس، یا واکنش‌پذیر، باشد. وابسته بودن مسئولیت‌پذیری به واکنش‌پذیری به دلیل[۲] به این معناست که عاملْ زمانی مسئولیت خواهد داشت که اعمالی که انجام می‌دهد حاصل حضور و واکنش به دلایل انسانی و در نتیجه در سازگاری و هماهنگی با این دلایل باشد. لذا به قرائتی از سازش‌گرایی نیاز خواهد بود که این ملاحظه اخیر در آن منظور شده باشد.

خوانش هری فرانکفورت و فیشر و رویتزا

قرائت قابل‌قبول‌تر از سازش‌گرایی، که در رویکرد کنترلِ معنادارِ بشری نیز استفاده شده، متعلق به فیشر و رویتزا (Fischer & Ravizza, 1998) است. این دو برای شرح دیدگاه خود از مثال‌ها و رویکرد فیلسوف سازش‌گرای پیش از خود به نام هری فرانکفورت استفاده زیادی کرده‌اند. هری فرانکفورت (Frankfurt, 1969) تلاش کرد از طریق ارائهٔ مثال‌هایی از سازش‌گرایی دفاع کند. ادعای فیشر و رویتزا این است که مسئولیت‌پذیری وابسته به حدی از توانایی کنترل نزد عامل نسبت به موضوع مسئولیت است. کنترل در معنای قوی آن وابسته به شرایطی است که در آن عامل قادر باشد یک عمل بدیل را به جز آنچه به صورت بالفعل انجام شده (یا می‌شود) انجام دهد. با این حال، مثال‌های فرانکفورت نشان می‌دهد که شرط لازم برای مسئولیت‌پذیر بودن داشتن **کنترل** به معنای قوی آن نیست، بلکه کنترل معنای ضعیف‌تری

---

۱. با مطرح کردن اهمیت سازوکارهای اجتماعی در مسئولیت‌پذیری افراد، فرض اختیار نیز به تنهایی برای مسئولیت‌پذیری کفایت نخواهد کرد. کسی که اختیار دارد شرط لازم را دارد، نه شرط کافی؛ شرط کافی را جامعه مشخص می‌کند.

2. reasen responsiveness

نیز می‌تواند داشته باشد (که با موجبیت سازگاری دارد) و فیشر و رویتزا آن را کنترل هدایتگر[1] (در مقابل کنترل نظارتگر[2]) نامیده‌اند (Fischer & Ravizza, 2000, p. 441). این نوع کنترل می‌تواند شرط لازم برای مسئولیت‌پذیری را محقق کند.

برای فهمِ مثالِ هری فرانکفورت، فردی (بگویید **الف**) را تصور کنید که مغز او توسط یک عصب‌شناس (بگویید **ب**) به نحوی دست‌کاری شده تا در نهایت از میان گزینه‌های متفاوت عملی را که مورد نظر عصب‌شناس است انجام دهد. برای مثال، عصب‌شناس تمایل دارد که الف امروز از خانه خارج شود و به محل کارش برود، اما عصب‌شناس تا زمانی که الف مطابق دلایل شخصی خود به محل کارش برود مداخله‌ای در تصمیم‌گیری او نمی‌کند، اما وقتی دلیلی کافی برای الف پیدا شود که در خانه ماندن امروز او را توجیه کند دخالت خواهد کرد و مثلاً با تراشه‌ای که پیش از این در مغز الف کار گذاشته است مانع از انتخاب دوم او خواهد شد. روشن است که در این مثال، الف کنترل نظارت‌گر بر عملش نخواهد داشت، اما همچنان تا زمانی که سر کار رفتن را با دلایل خاص خود هدایت می‌کند مسئول عملش نیز قلمداد می‌شود (Frankfurt, 1969, p. 834).

لذا در شرح فیشر و رویتزا از رویکردِ فرانکفورت، کنترل نظارتگر مستلزم وجود دو قوه (یا توانایی) متفاوت است: قوهٔ اول این است که عامل می‌تواند عملی را که در حال انجام آن است هدایت کند و قوهٔ دوم این امکان را برای او ایجاد می‌کند که در صورتی که تصمیم گرفته باشد آن عمل خاص را انجام ندهد یا به نحو دیگری انجام دهد. سازش‌گرایانی همانند فیشر و رویتزا معتقدند که مسئولیت‌پذیر بودن عامل تنها مستلزم وجود امکان هدایت کردن عمل و در نتیجه نیازمند کنترل هدایتگر است (Fischer & Ravizza, 1998, p. 31).

مفهوم واکنش‌پذیری نسبت به دلیل در چنین شرایطی مطرح می‌شود. اگر عامل را واکنش‌پذیر نسبت به دلیل بدانیم به این معناست که او نسبت به دلیل‌های کافی واکنش خواهد داشت و لذا از این جهت کنترل او کنترل نظارتگر (یعنی دارای هر دو قوه) است و به عبارت دیگر دارای اراده آزاد در نظر گرفته می‌شود، که روشن است با موجبیت همخوانی نخواهد داشت. در مثال‌های سازش‌گرایانهٔ فرانکفورت، عامل همچنان نسبت به دلیل واکنش‌پذیر نیست، چرا که وجود یا عدم وجود دلیل در عملی که در نهایت الف به عنوان یک عامل انجام خواهد داد تأثیری ندارد.

پیشنهاد فیشر و رویتزا این است که می‌توان واکنش‌پذیر بودن به دلیل را نه به عامل، بلکه به سازوکاری علّی که به یک تصمیم منتهی می‌شود نسبت داد (Fischer & Ravizza, 1998, p. 39). منظور از سازوکار سلسله علت‌هایی است که هم سازوکارهای علّی موجود در دستگاهِ شناختی عامل را شامل می‌شود و هم

1. guidance control
2. regulative control

در تعیین تصمیم نهایی نقشی علّی ایفا می‌کنند. سازوکار تصمیم‌گیری ممکن است مجموعهٔ رویدادهایی به لحاظ علّی پیوسته به هم را شامل شود، از گذشته تا زمان حال، که به دستگاه شناختی عامل می‌رسد. حال اگر این سازوکار علّی به دلایلِ انسانی واکنش‌پذیر باشد، به این معنا که مسیر آن در شرایطی که دلیل‌های کافی وجود دارد مبتنی بر آن دلایل تنظیم شود، آنگاه می‌توان گفت که عامل که بر مبنای این سازوکار عملی را انجام می‌دهد، نسبت به آن عمل مسئولیت‌پذیر است. تمایز میان عامل و سازوکارِ تصمیم‌گیری می‌تواند برای درک بهتر مسئولیت‌پذیری در سیستم‌های مجهز به هوش مصنوعی، به عنوان سازوکار تصمیم، نیز بسط پیدا کند.

این شرط که سازوکار تصمیم باید نسبت به دلایل واکنش‌پذیر باشد مواردی مشابه افراد دارای ترس بیمارگون یا بیماری روانی حاد را از دایره مسئولیت‌پذیران حذف می‌کند. این موارد از این جهت حذف می‌شوند که سازوکار درون-مغزی آن‌ها که به تصمیم منجر می‌شود به دلایل عقلانی مرتبط نیست و تصمیم مبتنی بر عوامل کاملاً علّی (شاید با غلبه هیجانات) گرفته شده است.

این شرط در کنار شرط دیگری قرار می‌گیرد که می‌گوید چرا شهود ما مسئولیت را به افرادی که دارای خودگردانی نیستند و اعمالشان از بیرون تحمیل شده است نسبت نمی‌دهد. مطابق این شرط، سازوکار تصمیم‌گیری باید متعلق به خودِ عامل باشد. به عبارت دیگر، زمانی عامل در مورد تصمیمی که توسط سازوکار گرفته شده مسئولیت‌پذیر خواهد بود که آن سازوکار (یا بخش قابل‌ملاحظه‌ای از آن) را *متعلق به* خود قلمداد کند، و به عبارت دیگر سازوکاری درونی باشد.

چون فرض این است که موجبیتی علّی برقرار است پس همه عوامل، از جمله دلایل و قصدهای بشری، با موجبیت سازگارند و بنابراین با رویدادهای پیش از خود متعین شده‌اند. در یک زنجیرهٔ علّی، هر آنچه علّت یک پدیده یا رویداد محسوب می‌شود، با علّت‌های قبل از خود متعین شده است. در این مورد قصدها، یا به اعتباری دلایل بشری، هم ممکن است با شرایط قبل از خود متعین شوند و در تعیین شرایط پس از خود نقش ایفا کنند. با این حال، همیشه می‌توان جهان‌های ممکنی داشت که یکی از این علت‌ها، از جمله دلایل و قصدهای افراد، به واسطه شرایط پیچیده آن جهان محقق نشود.

برای مثال زنجیرهٔ رویدادهای الف، ب و ج را تصور کنید. برای سادگی الف را رویداد تهدید شدن توسط فردی از راه دور، ب را قصد عامل به از بین بردن تهدید، و ج شلیک اسلحه به سمت تهدیدکننده در نظر بگیرید. حال می‌توانید امکان دیگری را نیز تصور کنید که رویداد الف مستقیماً، یا با جایگزین شدن ب با شرایط دیگری به جز قصد، رویداد ج را ایجاد کرده باشد. و ب را می‌توانید افتادن آفتاب ساحلی بر چشم، یا بی‌اختیاری حاصل از مصرف الکل قلمداد کنید. در این شرایط در این سلسله علل، قصد یا دلیلی در زنجیره علّی وجود ندارد، بلکه رویداد ج ناخودآگاه رخ داده است. این فرض می‌تواند این تصوّر را ایجاد کند که در جهان‌های ممکنی که دلیل‌ها حضور ندارند ممکن بود زنجیره علّی به نحوی دیگر شکل بگیرد. از این جهت زنجیره علّی بدون این که اراده آزادی دخالت داشته باشد به دلیل واکنش‌پذیر خواهد بود.

فیشر و رویتزا شرح می‌دهند که مسئولیت خصلتی تاریخی دارد. راننده‌ای که هوشیاری خود را بـر اثـر مصرف الکل از دست داده، اگر تصادف مرگباری داشته باشد مسئول آن رخداد خواهد بود، چرا که با توجـه به این که می‌دانسته قرار است رانندگی کند باید از نوشیدن زیاد صرف نظر می‌کرده است. حال اگر به کسی، مانند قهرمان فیلم *شمال از شمال‌غربی*[1]، الکل خورانده باشند و او را در اتومبیلِ در حال حرکتی رهـا کـرده باشند، مسئولیت تصادف‌های احتمالی بـر عهـده او نخواهـد بـود (Fischer & Ravizza, 1998, p.195) و روشن است که اگر می‌توانست رخدادهای گذشته را به قاضی بقبولاند تبرئه می‌شد. از این جهـت، مطـابق این دیدگاه، تاریخ یک عمل باید شامل پذیرفتن مسئولیت توسط عامل باشد.

دو قوه متفاوت واکنش‌پذیری به دلیل: واکنش‌پذیر بودن به دلیـل از دو قـوه متفـاوت و نامتقـارن تشـکیل شده است. قوهٔ اول پذیرندگی داشتن[2] و قـوه دوم بازکنشـی بـودن[3] اسـت[4]. پذیرنـدگی بـه معنـای ظرفیـت بازتشخیص دلایلی است که وجود دارد و بازکنشی بودن ظرفیت ترجمه کردن دلایل به یک انتخـاب اسـت (Fischer & Ravizza, 1998, p. 69). این دو قوه نامتقارن هستند به این معنـا کـه شـرایط ممکنـی می‌تـوان داشت که پذیرندگی وجود داشته باشد، اما بازکنشـی نسـبت بـه آن انجـام نگیـرد. قـوه پذیرنـدگی بایـد در جهان‌های ممکن بیشتری به نسبت بازکنشی بودن برقرار باشد، چرا که تنهـا موجـودی می‌توانـد مسـئولیت داشته باشد که بتوان برای او موقعیت‌هایی را تصور کرد که قـدرت پـذیرش دلیل‌هـای کـافی را دارد، حتـی اگر این پذیرش به هیچ تغییری منجر نشود. افرادِ دارای ضعف اراده در ایـن مـورد قابل‌توجـه هسـتند. مـثلاً افرادی که دلایل قدرتمندی را در مورد مضرات سیگار کشیدن پذیرفته‌اند، اما همچنان به کشـیدن آن ادامـه می‌دهند.

مصنوعات خودگردان و مسئولیت‌پذیری

سانتونی دِ سیو و ون دن هاون (Santoni de Sio and Van den Hoven, 2018) از این مـدل بـرای اسـتدلال به نفع این که می‌توان به سیستم‌های فنی-اجتماعی مجهز به هوش مصـنوعی مسـئولیت نسـبت داد، و بـه عبارت دیگر مسئله شکاف مسئولیت مرتبط با خودگردانی ایـن سیسـتم‌ها را حـل کـرد، اسـتفاده کرده‌انـد. سیستم‌های هوش مصنوعی در نهایت سازوکاری علّـی هسـتند و مطـابق آنچـه شـرح داده شـد، اگـر ایـن سازوکار علّی یعنی این سیستمِ مجهز به هوش مصنوعی، واکنش‌پذیر به دلیل باشد (شرط اول) و همچنـین به نحوی به عامل انسانی متعلق باشد (شرط دوم)، آنگاه می‌توان به این سیستم، که شامل انسان و سـازوکار

1. North by Northwest
2. Receptivity
3. Reactivity
٤. در این مقاله، واژه «پاسخ» برای Account، واژه «واکنش» برای Response و واژه «بازکنش» بـرای Reaction بـه کـار بـرده شده است.

مکانیکی است، مطابق تعریفی که ارائه شد مسئولیت نسبت داد.

(الف) شرط اول: اگر سازوکار، که در اینجا سلسله علت‌های غالباً درونی سیستم خودگردان مجهز به هوش مصنوعی است، به دلایل واکنش‌پذیر باشد، به یک معنا می‌توان به این سیستم مسئولیت نسبت داد، یعنی به همان معنا که نه به یک فرد دارای اختلال روانی بلکه به فردی که اعمالش مبتنی بر دلیل است مسئولیت نسبت می‌دهیم. لذا این شرط تمایزی را میان افراد مسئول و افرادی که اعمال آن‌ها هرچند مرتبط با علت‌های درونی است اما با اختلال‌هایی خارج از کنترلشان هدایت می‌شود قائل می‌شود. دست‌کم تا زمانی که در مورد این شرط صحبت می‌شود، به نظر می‌رسد که الزامی در این مورد که سازوکار به فرد خاصی متعلق باشد وجود ندارد. سانتونی د سیو و ون دن هاون به جای واژهٔ واکنش‌پذیری به دلیل از واژهٔ قابلیت ردیابی[1] استفاده می‌کنند، که آن را از نظریهٔ صدق نوزیک گرفته‌اند. این واژه از قضا برای مهندسان نیز شناخته شده است. ردیابی دلایل در مهندسی چیزی شبیه به کاری است که دماسنج با دما انجام می‌دهد. تغییرات دماسنج وابسته به تغییراتی است که در دما رخ می‌دهد. از این جهت می‌توان به سیستمی که مصنوعات خودگردان بخشی از آن است نیز، نه به عنوان عامل و نه به عنوان موجودی که به آن آزادی اراده نسبت داده شده باشد، بلکه به عنوان سازوکاری مکانیکی-الکترونیکی که به تصمیم منجر می‌شود، مسئولیت نسبت داد، مشروط بر این که آن مصنوعات خودگردان مجهز به هوش مصنوعی، به نحوی، نسبت به دلایل واکنش‌پذیر باشند، یا چنان که به صورت فنی گفته می‌شود *ردیابی‌کننده* دلایل باشند. سانتونی د سیو و ون دن هاون از معنای «شناخت» نزد نوزیک برای ارائهٔ الگویی در مورد ردیابی استفاده می‌کنند. در رویکرد نوزیک، و در پاسخ به مشکلات معرفت‌شناختی مشهوری که گتیه در خصوص تعریف شناخت مطرح کرده، برای این که بتوان به شناخت رسید فرد باید بتواند دائماً وضع امور جهان را ردیابی کند، و صدق جملات را بر مبنای آن اصلاح کند. با این حال، ردیابی کردن وضع امورِ جهان برای صدق عبارت‌ها الزاماً به معنای ردیابی کردن دلیل‌ها نیست.

(ب) شرط دوم: همان طور که گفته شد، کارکرد شرط دوم این است که موجودات مسئول را از موجوداتی که تصمیم آن‌ها تحت تأثیر کنترل‌های بیرونی است مجزا کند. از این جهت، سازوکار تصمیم، هر چه که باشد، باید بخشی از سازوکار درونی عامل و جزء یکپارچه‌ای از او قلمداد شود. به عبارت دقیق‌تر، «متعلق به عامل» باشد. از این جهت زمانی که تحت تأثیر یک تهدید، یا فشارها و اصرارهای دیگران عملی انجام می‌شود، سازوکار تصمیم متعلق به عامل نیست، یا دست‌کم عامل آن را متعلق به خود قلمداد نمی‌کند. عامل به عنوان بخشی از سیستم در مورد سازوکارهایی که او به وسیله آن‌ها تصمیم می‌گیرد پذیرش مسئولیت خواهد داشت. به عبارت دیگر، سیستم به واسطهٔ حضور عامل و دیگر شروطی که برقرار خواهند بود، مسئولیت اجتماعی خواهد داشت. عامل مسئولیت سازوکارهایی را که از طریق

1. Tracking

آن‌ها تصمیم می‌گیرد بر عهده خواهد داشت.

سانتونی د سیو و ون دن هاون در پاسخ به این چالش که چطور می‌توان شرط دوّم فیشر و رویتزا برای مسئولیت‌پذیری را در خصوص سازوکارهای تصمیم‌گیری مبتنی بر هوش مصنوعی اعمال کرد به نحوی که سیستم خودگردان تصمیم‌گیر بخشی از سازوکار درونی و متعلق به یک عامل محسوب شود، به مفهوم شناخت بسط‌یافته[1] (Clark & Chalmers, 1998) اشاره می‌کنند. مدافعان شناخت بسط‌یافته دلایلی ارائه می‌کنند که مطابق آن شناخت یک عامل صرفاً بر خصلت‌های درونی مغز استوار نیست، بلکه ابزارهایی که برای کمک به توانایی‌هایی شناختی انسانی استفاده می‌شود، حتی یک دفتر یادداشت برای فرد مبتلا آلزایمر، نیز بخشی از قوه شناخت او را تشکیل می‌دهد. از این جهت، مطابق این رویکرد سازوکارهای خودگردان مجهز به هوش مصنوعی نیز می‌توانند بخشی از قوه شناخت یک سرباز در میدان جنگ یا راننده در هدایت مسیر اتومبیل باشد.

بحث و تحلیل مباحث

بازتعبیر دیدگاه

به عنوان نوآوری مقاله از هر دو شرح تعبیرهای متفاوتی ارائه خواهد شد و در نهایت استدلال خواهد شد که شرط ردگیری را می‌توان به نحوی متفاوت از رویکردهای گذشته فهمید، به نحوی که برخی از مشکلات موجود در تعابیر قبلی حل شود. برای این منظور اجازه دهید شرایط ردیابی را از حیث منطقی بیشتر بررسی کنیم.

وقتی موجبیت را برقرار می‌دانیم به معنای تأکید بر این عقیده است که جهان دارای نوعی از بستگی علّی است، یعنی هر چیزی یا هر رویدادی با مجموعه‌ای از رویدادهای پیش از خود، که علت آن رویداد محسوب می‌شوند، و در درون همین جهان هستند، متعین می‌شود. به عبارت دیگر، می‌توان گفت علیّت و زنجیره‌های علّی-فیزیکی تنها روابطی هستند که رویدادهای جهان را رقم می‌زنند. این حکم در خصوص رویدادهایی که به عنوان «رویدادهای آزادانه بشری» می‌شناسیم نیز صادق خواهد بود. به عبارت دیگر، هیچ افسونی خارج از دایره روابط موجبیتی مبتنی بر قوانین علّی-فیزیکی تعیین‌کننده شرایط و رویدادهای عالم که شامل رویدادهای مرتبط با تکنولوژی‌ها نیز می‌شود نخواهد بود.

سازوکار تصمیم نیز یک سازوکار علّی است. در خصوص موضوع این مقاله سازوکار تصمیم را می‌توان همان تکنولوژی خودگردان مجهز به هوش مصنوعی قلمداد کرد. در این مورد خاص، منظور از تکنولوژی نه یک سیستم فنی-اجتماعی، بلکه مصنوعی خودگردان مجهز به هوش مصنوعی است، که برای مثال می‌تواند یک اسلحه خودگردان یا یک ماشین خودران باشد. لذا آنچه سازوکارِ تصمیم نامیده

1. Extended Cognition

می‌شود، در اینجا اسلحه خودگردان مجهز به هوش مصنوعی، بخش شاید مجزا و ایدئال‌شده‌ای از این زنجیره بزرگ علّی موجود است؛ ایدئال‌شده به این معنا که حتی‌الامکان از تأثیرات علّی بیرونی مثلاً محیط اطراف منفک شده است.

در این میان، فرض واکنش‌پذیر بودن سازوکارِ تصمیم به دلیل قابل تفسیر به این خواهد بود که (۱) دست‌کم یک جهانِ ممکنِ دیگر وجود داشته باشد که در آن اگر دلایل کافی حاضر باشد، عامل آن را بپذیرد، و (۲) از میان جهان‌هایی که این شرایط را دارند، دست‌کم یک جهان ممکن وجود داشته باشد که نسبت به آن دلایل بازکنشی داشته باشد، یعنی مسیر تغییر می‌کند (Fischer & Ravizza, 1998). همان طور که گفته شد، تعریف واکنش‌پذیری به دلیل و یکی از شروط خاص آن، یعنی بازکنشی بودن، در حالی ارائه می‌شود که پذیرش موجبیت به این معناست که هیچ امکانی به جز آنچه به صورت بالفعل رخ می‌دهد مجاز نیست. به بیان دیگر، جهان ممکنی به جز جهان بالفعل وجود ندارد. از این جهت نیاز به شرحی هست که در شرایطی که موجبیت فرض گرفته شده چطور می‌توان از امکان‌ها و از جهان‌های ممکن صحبت کرد.

زنجیره‌های علّی و ارتباط با موجبیت را به ناچار به نحو خاصی تحلیل خواهیم کرد. در مرتبه اول، صحبت از موجبیت حکم مشهور لاپلاس را به خاطر می‌آورد که اگر فیزیک‌دانی با ظرفیت ذهنی بسیار بالا می‌توانست در یک لحظه از زمان شرایط کل ماده را بداند، می‌توانست آینده را پیش‌بینی کند (Hoefer, 2024). به عبارت دیگر، در هر لحظه از جهان شرایط ماده، به طور کلّی، به نحوی است که به همراه قوانینی که در این جهان حاکم است، می‌تواند لحظه و لحظات بعدی و شرایط جدید جهان را به صورت کامل معین کند. با همین روال، شرایط ماده در لحظه جدید نیز همهٔ آنچه را در لحظهٔ بعدی وجود خواهد داشت معین خواهد کرد، و الی آخر. این تصور می‌تواند این ادعا را توجیه کند که شرایطِ مثلاً دهمین لحظه نیز از ابتدا، و فارغ از چگونگی رخ دادن لحظات ۱ الی ۹، با شرایط لحظه اول متعین شده است. لذا اگر هیولای لاپلاس به همهٔ جزئیات ماده در یک لحظه خاص دانش کافی داشت، می‌توانست همهٔ لحظات بعدی و مثلاً رویدادهای لحظه دهم را نیز پیش‌بینی کند. به عبارت دیگر، رخدادها وابسته به مسیر نیستند. روشن است که در این مورد قصدهای کنشگران، که می‌تواند همان دلایلی فرض شود که ما زنجیره علّی را به آن‌ها واکنش‌پذیر قلمداد می‌کنیم، نیز بخشی از شرایطی هستند که از پیش معین شده‌اند.

با این حال، شناخت و تعبیر ما از یک زنجیره علّی غالباً کافی نیست. لذا، هرچند به لحاظ هستی‌شناختی باور به موجبیت همچنان وجود دارد، اما یک تحلیلگر عادی تنها قادر است یک زنجیرهٔ علّی را زنجیره‌ای از رویدادهایی پی‌درپی ببیند که یکی پس از دیگری و هرکدام با تحریک رویداد قبلی رخ می‌دهند. برای نمونه، شرایطی را که با ضربه زدن به توپ بیلیارد سفید ایجاد می‌شود و رویدادهای حرکت توپ‌های بعدی بعد از برخورد توپ سفید را در نظر بگیرید. در این حالت، بعد از ضربه زدن به

توپ سفید شماره ۱، حتماً رویدادهایی مرتبط با برخورد این توپ با توپ‌های شماره ۲ تا ۷ باید وجود داشته باشد تا رویداد هشتم که افتادن توپ قرمز در چال است نیز رخ دهد. به عبارت دیگر، رویدادهایی میانی حتماً باید رخ داده باشند که رویداد هشتم نیز رخ دهد. یک رویداد به معنای تغییر در یک خصلت از میان بی‌نهایت خصلت در یک لحظه و مکان است که باعث تغییر در خصلت‌های دیگر در لحظات و مکان‌های بعدی خواهد شد. روشن است که هیچ هیولای لاپلاسی قادر نیست تنها با دانستن شرایطِ رویداد اول، و بدون دانستن شرایط کل مواد در آن لحظه، آنچه را برای رویداد هشتم رخ می‌دهد پیش‌بینی کند. چون رویداد یک، بدون وجود رویدادهای میانی و شرایطی که آنها را معین کرده، رویداد هشتم را معین نخواهد کرد. از این جهت در ادامه صحبت از امکان صحبت از جهان‌های ممکن (معرفتی) می‌شود، چون در خصوص چگونگی رخ دادن رویدادهای میانی اطلاعات کافی وجود ندارد. زنجیرهٔ علّی به واسطهٔ بی‌نهایت خصلت ادراک‌نشده‌ای که اطراف آن وجود دارد و ما به آن شناخت کافی نداریم، ممکن بود به انحای دیگری مسیرش را انتخاب کند.

هرچند طراحان مصنوعات خودگردان، برای این که شرط خودگردانی را محقق کنند، تلاش می‌کنند که سازوکارهای مورد نظر را تا حد امکان نسبت به عناصر بیرونی دخیل در شرایط «ایدئال» کنند، به عبارت دیگر سازوکار را به نحوی طراحی کنند که تأثیر علّی عناصر بیرونی در آن کاهش یابد، ولی سازوکار اسلحهٔ خودگردان به واسطه ورودی‌های متفاوت و پیچیدگی‌های سازوکار درونی، مسیرها و در نتیجه امکان‌های معرفتی متفاوتی را تا رسیدن به آنچه «تصمیم» قلمداد می‌شود طی می‌کند.[۱] این فرض امکانیتی مستقل از فرض وجود عامل‌ها را در این مسیر ایجاد می‌کند.[۲]

واکنش‌پذیری به دلیل: با استفاده از همین معنا از امکانیت، می‌توان واکنش‌پذیر بودن به دلیل را نیز بازتعریف کرد. اولین تصور مرتبطی که ایجاد خواهد شد این است که دلایل در این سازوکار باید در زنجیره علت‌ها نقش علّی داشته باشند و تعیین‌کننده نتیجهٔ تصمیم باشند. اما دلایل انسانی چطور در این سازوکارهای تصمیم دخیل می‌شوند؟ زنجیرهٔ علّی که سازوکار اسلحه‌های خودگردان نیز بخشی و البته نه همه آن را تشکیل می‌دهد، همانند همه تکنولوژی‌ها، در یک سیستم فنی-اجتماعی که همچنین متشکل از افراد انسانی است عمل می‌کند. در این حالت، دخالت داشتن یا در حلقه تصمیم قرار داشتن انسان را می‌توان به دو صورت متفاوت تصور کرد:

ـ از یک سو، می‌توان انسان‌ها را موجوداتی صرفاً روان‌شناختی و تحت سیطره کامل هیجانات و حالات درونی تصور کرد که در میانهٔ این زنجیره قرار گرفته‌اند.

---

۱. روشن است که این الزاماً به معنای وجود اراده آزاد (چه در مورد انسان و چه در مورد اسلحه خودگردان مجهز به هوش‌مصنوعی) نیست.

۲. چرا که اگر موجبیت فرض گرفته شده باشد، عامل‌ها نیز امکان کنترل نظارتگر و لذا تغییر دادن مسیر را نخواهند داشت.

ـ از سوی دیگر، می‌توان آن‌ها را افرادی تصوّر کرد که به واسطهٔ دلایل، که ممکن است قصدهای بشری باشند، در این زنجیره حضور دارند.

هر دو موردِ تصورشده در بالا با موجبیت سازگارند، چون رخداد شکل گرفتن قصدهای انسانی نیز می‌تواند با رویدادهای قبل از خود معین شود. لذا در یک تعبیر، آنچه فیشر و رویتزا، با ارجاع به فرانکفورت، کنترل هدایتگر می‌نامند، به این صورت قابل فهم است که قصدهای بشری نه به معنای دخالت اراده‌ای آزاد، بلکه به معنای حلقه‌ای که به طور علّی با حلقه‌های قبلی متعین شده است، در مسیرِ زنجیرهٔ علّی حضور دارند. این یعنی به همان صورتی که حرکت توپ ۲ و حرکت توپ ۳ در میانه مسیر توپ‌های اول و چهارم قرار گرفته بود، قصدهای بشری نیز به عنوان رویدادهایی مؤثر در مسیر و تاریخچه یک عمل قرار خواهند گرفت، هرچند خودشان از شرایط فیزیکی و علّی پیش از خود نتیجه شده‌اند.

علت این که دلیل‌ها، در این تعبیر خاص، قصدهای انسانی قلمداد شده‌اند توجه به این موضوع است که تنها در صورتی می‌توان شرط بازکنشی بودن را لحاظ کرد که دلایل در زنجیرهٔ علّی اثر داشته باشند، یا به عبارت دیگر دارای قوه علّی باشند. این تعبیر مشابه شرحی است که سانتونی د سیو و مکاچی (Mecacci & Santoni de Sio, 2019) ارائه کرده‌اند.[۱] قصدها، اگر آن‌ها را نزدیک به عمل قلمداد کنیم، به واسطهٔ انسان چنین شرایطی را دارا هستند. از این جهت، زنجیره‌ای علّی را که قصدهای بشری برخی حلقه‌های آن را تشکیل می‌دهند، در مقایسه با زنجیره‌های عاری از قصد، می‌توان به مثابه واکنش‌پذیری این زنجیره به دلیل قلمداد کرد. این فرض در خصوص واکنش‌پذیر بودن را فعلاً در این مرحله رها می‌کنیم تا به شرط دوم یعنی ردگیری برسیم. در نهایت از آنچه در خصوص هر دو ایده گفته شد، نتیجه در مورد هوش مصنوعی گرفته خواهد شد.

مسئله دو تعبیر از شرط ردگیری

شرط ردگیری محقَق شدن آن شهودی را تضمین می‌کند که مطابق آن افراد تا جایی *مسئول* قلمداد می‌شوند که اعمالشان نه تحت اجبار شرایط «بیرون»، بلکه از سازوکارهای «درونی» خودشان منتج شده باشد. مطابق آنچه گفته شد، فیشر و رویتزا مسئولیت‌پذیری را دارای ماهیتی ذاتاً تاریخی قلمداد می‌کنند. ذاتاً تاریخی بودن با مفهومِ دیگری به نام *پذیرش مسئولیت* مرتبط است. پذیرش مسئولیت شرط لازمی برای مسئولیت‌پذیر کردن افراد است و بر مبنای آن‌ها یک عامل نتایج حاصل‌شده از یک سازوکار را به عنوان یک «تصمیم» درونی می‌پذیرد. در ادامه استدلال خواهد شد که از «ذاتاً تاریخی بودنِ» مسئولیت‌پذیری، با توجه به مثال‌ها و ایده‌هایی که فیشر و رویتزا مطرح کرده‌اند، می‌توان دو معنای متفاوت

---

۱. این تنها راه فهم دیدگاه سانتونی د سیو و ون دن هاوون نیست. ولو این که این تعبیر مشکلات دیگری ایجاد خواهد کرد که بحث و ارائه راه‌حل برای آن موضوع مقاله دیگری خواهد بود.

و بعضاً ناسازگار با یکدیگر را برداشت کرد.

(الف) تعبیر اوّل از ردگیری (پذیرش مسئولیت به عنوان عمل): تعبیر اول را می‌توان از برخی مثال‌های روشنِ فیشر و رویتزا استخراج کرد (Fischer & Ravizza, 1998, p.195). مثال راننده‌ای که الکل نوشیده یا به او الکل خورانده‌اند را به خاطر آورید. در این مثال، فرض این است که رویدادهایی که رخ خواهد داد، از جمله تصادفات و کشته شدن افراد در جاده، تاریخچهٔ علّیِ ظاهراً معینی دارند که نسبت دادن مسئولیت آن‌ها به یک عامل وابسته به چگونگی زنجیره علّت‌هایی است که آن تاریخ را شکل داده است. در این حالت، مسئولیت به عهده کسی است که اولین بار عمل مؤثری را از روی قصد، یا به عبارتی با یک دلیل، انجام داده است. در سناریوی اول از این مثال، این عمل نوشیدن الکل توسط راننده است. اما در سناریوی دوم، عاملی که مسئولیت دارد نه خود راننده بلکه کسی است که الکل را به او خورانده است. لذا وقایع تاریخی که به حوادث بد، مثل کشتار افراد در جاده، منجر می‌شود شامل یک عمل قصدمند (مثل خوردن یا خوراندن الکل) در تاریخچهٔ علّی آن حوادث است، و لذا مسئولیت آن حوادث به عامل آن عمل قصدمند تاریخی نسبت داده می‌شود. مطابق این تعبیر، مسئولیت حوادثِ بدِ ناشی از ماشین خودگردان مجهز به هوش مصنوعی نیز به انسانی برمی‌گردد که تصمیمی را از روی دلیل، در زنجیرهٔ علّی‌ای که به حادثه ختم شده، گرفته است.

(ب) تعبیر دوم از ردگیری (پذیرش مسئولیت به عنوان رابطه تعلق): مثال‌هایی از قسم رانندهٔ الکلی صرفاً شهود را برای تأیید این ادعا که مسئولیت ماهیتی تاریخی دارد تحریک می‌کنند، و الزاماً و به طور خاص برای موضوع پذیرش مسئولیت مطرح نشده‌اند. با این حال، در بحث پذیرش مسئولیت، فیشر و رویتزا، بر خلاف تعبیر قبل، می‌گویند پذیرشِ مسئولیت صرفاً یک *عمل* نیست، بلکه مسئولیت زمانی به یک عامل نسبت داده می‌شود که تاریخچهٔ رویدادهایی که مسئولیتش پذیرفته می‌شود *شامل فرایند پذیرش مسئولیت توسط یک عامل* هم باشد. در فرایند پذیرش مسئولیت، نه یک عمل، بلکه مجموعه‌ای از باورها شکل می‌گیرند (Fischer & Ravizza, 1998, p. 208). برای مثال، این باور شکل می‌گیرد که سازوکارِ خاصِ تصمیم متعلق به یک عامل است. در فرایند پذیرش مسئولیت، این باور شکل می‌گیرد که مثلاً یک سیستم خودگردان مجهز به هوش مصنوعی متعلق به یک فرد انسانی (اپراتور یا طراح یا ...) است. روشن است که مجموعه باورها نه بخشی از زنجیرهٔ علّی، بلکه بخشی از فضای ذهنی و مفهومی افراد بشر است، که شاید هیچ نقش مستقیمی در زنجیره علّی ایفا نکند. قرار است باورها به نحوی شکل بگیرند که رابطهٔ «متعلق بودن»، که همان شرط دوم فیشر و رویتزا است، متعین شود. این یعنی در یک فرایند و با گذشت زمان، در ذهن بازیگران موجود این باور شکل گرفته باشد که سازوکار تصمیم[1] به یک عامل خاص تعلق

---

۱. سازوکار اسلحه خودگردان تنها بخشی از چیزی است که به آن زنجیرهٔ علّی گفته می‌شود. برای مثال، اسلحهٔ خودگردان تنها بخش خاصی از سازوکاری خواهد بود که به یک عمل، مثل کشتن یک کودک، منجر می‌شود.

دارد و از این جهت او مسئول «تصمیم‌هایی» خواهد بود که از این سازوکار خاص بروز می‌کند.

این دو تعبیر به نحو اصولی با یکدیگر متفاوت هستند. و آن طور که شرح داده خواهد شد، تأثیر آن وقتی در سیستم‌های هوش مصنوعی به کار بسته شود بیشتر مشهود خواهد بود.

در تعبیر اوّل، از این جهت مسئولیت‌پذیری دارای خصلتی تاریخی قلمداد می‌شود که یک نقطه تعیین‌کننده، نه در ویژگی‌های بالفعل، بلکه در میان یکی از حلقه‌های زنجیره علّی که به حادثه مورد نظر منجر شده وجود دارد. در این تعبیر، در شرایطی می‌توان ادعا کرد شکاف مسئولیت پر شده که با دنبال کردن زنجیرهٔ علّی بتوان به اولین دلیل (قصد) مؤثر در اقدامات رسید. در مثال رانندهٔ الکلی ردیابی عمل قصدمند نوشیدن راننده در گذشته شرایط مسئولیت‌پذیری متفاوتی را به نسبت زمانی ایجاد خواهد کرد که ردگیری ما را به عمل قصدمند خورانده شدن الکل توسط دیگران برساند. و هر دو مورد با سناریوی دیگری که هیچ عمل قصدمندی در این میان مطرح نباشد متفاوت خواهد بود (مثلاً الف تصادفاً و بدون این که آگاه شود الکل نوشیده باشد). بخشی از حوادثی که ممکن است در مورد مصنوع خودگردان مجهز به هوش مصنوعی رخ دهد با سازوکار مکانیکی-الکترونیکی مصنوع خودگردان به حالت ایدئال درآمده‌اند. اگر این مصنوع، به دلیل واکنش‌پذیر باشد، مطابق آنچه گفته شد، بخشی از این زنجیره حاوی قصدهای انسانی خواهد بود که این انتظار را ایجاد می‌کند که یکی از دلایل (قصدها) تعیین‌کنندهٔ دقیق عامل (یا عامل‌های) مسئول باشد. از سوی دیگر، اگر این مصنوع شرط اول فیشر و رویتزا، یعنی واکنش‌پذیری به دلیل، را نداشته باشد، آنگاه شرایطش مشابه راننده مستی است که هوشیاری تصمیم را از دست داده است. از این جهت مسئولیت اعمال این سیستم به بخشی از تاریخچه حوادث برمی‌گردد، مثلاً به زمانی که مصنوع توسط بشر طراحی شده، که پیش از به کار گرفتن مصنوع خودگردان هستند.

اما در تعبیر دوّم، ممکن است در درون زنجیرهٔ علّی‌ای که به یک حادثهٔ ناگوار، مثل کشته شدن کودکان در جنگ، ختم می‌شود هیچ رویدادی وجود نداشته باشد که تعیین کند آیا عاملی مسئول آن حوادث هست یا خیر. به عبارت دیگر، ممکن است کلِ زنجیرهٔ علّی-تاریخی در هر دو حالت یکسان باقی بماند، یعنی در حالتی که مسئولیت پذیرفته شده است و در حالت دوم که مسئولیت پذیرفته نشده است). تنها موردی که باعث می‌شود مسئولیت توسط عاملی پذیرفته شود شکل‌گیری مجموعه‌ای از باورها نزد عامل و عامل‌های دیگر در بین افراد و در اجتماع است که روی هم تعیین کرده‌اند که سازوکار مورد نظر متعلق به عامل است و از این جهت عامل به نحوی پاسخگوی نتایج این سازوکار است. از این جهت، نه زنجیره علی-فیزیکی، بلکه فضای ذهنی و اجتماعی است که مسئولیتی را به یک عامل نسبت می‌دهد.

به طور خلاصه می‌توان گفت که در تعبیر اول، یکی از حلقه‌های همان زنجیره علّی-فیزیکی است که تعیین می‌کند که آیا مسئولیتی وجود دارد یا خیر. و اگر کسی می‌توانست این زنجیره را ردگیری کند عامل مسئول را کشف می‌کرد. در تعبیر دوم، مسئله شکل گرفتن و تعیین حدود هویت‌ها است که کمتر به امور

واقع و زنجیره علّی-فیزیکی و بیشتر به شبکه باورهای فردی و اجتماعی مرتبط است.

(ج) مشکلات تعابیر مطرح‌شده: این دو تعبیر مانعةالجمع هستند. زیرا در یکی قطعاً لازم است که خصلتی در زنجیرهٔ علّی وجود داشته باشد و دیگری به طور کامل مجزا از زنجیره علّی و در فضای ذهنی و فرهنگی شکل می‌گیرد.

از سوی دیگر، هر دو تعبیر مشکلاتی به همراه دارند. تعبیر اول نیازمند وجود یک و تنها یک رویداد قصدمند مشخص در تاریخچه علّی است که تعیین‌کننده دقیق عامل باشد. اما هیچ تضمینی نیست که چنین خصلتی در زنجیرهٔ علّی شکل گرفته باشد. پذیرش تعبیر دوم ما را با مسئله‌ای دیگر در توضیح مواجه خواهد کرد: چه شرایط اجتماعی خاصی باید حاکم باشد که یک عامل مسئولیت یک سازوکار را به صورتی که پرکننده شکاف باشد بپذیرد. به نظر می‌رسد که چنین امری نیازمند تعاملات اجتماعی از نوع خاصی است که نتیجه آن باید تخصیص یافتن سازوکارهای تصمیم به عامل‌هایی باشد که در زنجیره تصمیم برای اعمال آینده حضور دارند.

این به نحوی بخشی از تنظیمات اجتماعی برای توسعه تکنولوژی خواهد بود که به واسطه نبود اطلاعات همواره از فرایند اصلی توسعه تکنولوژی عقب خواهند ماند. و لذا نوع دیگری از شکاف مسئولیت را، این بار در قلمرو توسعه، ایجاد خواهد کرد (Owen *et al.*, 2013)، که به نظر می‌رسد به مسئله توسعه مسئولانه به معنای کلی‌تر مربوط می‌شود و ارتباط آن با مصنوعات خودگردان باید در جای دیگری مورد بحث قرار گیرد (Alasti, 2024).

پذیرش مسئولیت سازوکارهای واکنش‌پذیر به دلایل نسبت-به-عامل-خنثی

ولوونکامپ (Veluwenkamp, 2022) استدلال دقیقی مطرح می‌کند که بر مبنای آن مسئله ردگیری به نحو خاصی تعبیر می‌شود. او می‌گوید شرط ردگیری با شرط دیگر یعنی ردیابی نسبت تنگاتنگی دارد. در واقع مطابق استدلال او، برای این که ردگیری (یعنی شرط دوم) به درستی انجام شود نیاز است ردیابی دلیل‌ها (یعنی شرط اول) انجام شده باشد. دلیل‌هایی که ردیابی می‌شوند نباید نسبت به عامل خنثی باشند. در ادبیات فلسفهٔ عمل، میان دلیل‌های نسبت-به-عامل-خنثی[1] و دلیل‌های وابسته-به-عامل[2] تمایز قائل می‌شوند. برای مثال، اگر دلیل من برای این که دوستی را ملاقات کنم این باشد که او دوست من است، این دلیلی وابسته-به-عامل خواهد بود، چرا که این دلیل همین عمل را برای فردی دیگر موجه نمی‌کند. اما در مثالی دیگر، اگر رفتن نزد دوستم برای این باشد که رفتن پیش دوستان شادی‌آور است، آنگاه این خصلت شادی‌آور بودن دیدار دوستان دلیلی است که می‌تواند هر فرد دیگری را نیز برای رفتن پیش یک دوست موجه کند. از این جهت این دلیل نسبت به عامل خنثی است (Veluwenkamp, 2022, p.7). بر

1. agent neutral
2. agent dependent

مبنای همین تقسیم ولوونکامپ استدلال می‌کند که همه دلیل‌هایی که ردیابی می‌شوند نمی‌توانند دلیل‌های نسبت-به-عامل-خنثی باشند. چرا که در غیر این صورت نمی‌توانند ما را به سمت یک عامل خاص هدایت کنند. اگر از جزئیات استدلال او بگذریم، آنچه برای بحث حاضر اهمیت دارد این است که دلیل‌های وابسته-به-عامل همیشه *متعلق به یک عامل* هستند و در نتیجه ما را به سمت آن عامل خاص هدایت خواهند کرد. از این جهت، اگر اسلحه خودگردان قادر باشد دلیل‌ها را، که بخشی از آنها وابسته-به-عامل هستند، ردیابی کند، پس می‌تواند به یک عامل خاص نیز مرتبط کند. از این جهت، یک سازوکار، همانند اسلحه خودگردان، زمانی به دلیل واکنش‌پذیر خواهد بود که این ارتباط با عامل را نیز ایجاد کند، یا به عبارتی، زمانی که دلیل‌ها را ردیابی می‌کند عمل ردگیری را نیز انجام بدهد. به طور خلاصه، نتیجه ولوونکامپ این است که میان این دو شرط (یعنی شرط ردگیری و ردیابی) ارتباط مستقیمی وجود دارد.

می‌توان ابهامی را که در دو تعبیر از ردگیری مطرح شد این بار در خوانش ولوونکامپ نیز بازیافت. به عبارت دیگر، می‌توان این خوانش را نیز به نحوی با هر دو تعبیر مطرح‌شده در این مقاله سازگار قلمداد کرد.

شرط دوم فیشر و رویتزا می‌گوید سازوکار باید متعلق به عامل باشد. اما شرح فیشر و رویتزا مربوط به *انسان* و مسئولیت‌پذیری انسان است، و نه سیستم‌های مجهز به هوش مصنوعی. طبیعتاً برخی موارد که در این شرط در مورد انسان پیش‌فرض قرار گرفته، در مورد مصنوعات تکنولوژیک وجود ندارد، و وقتی از دستگاه شناختی انسان به سیستم‌های خودگردان مجهز به هوش مصنوعی می‌رسیم، به تدریج این پیش‌فرض‌ها دارای اهمیت خواهند شد. برای مثال، این پیش‌فرض اهمیت پیدا می‌کند که در شرایط کنترل هدایتگر نه تنها سازوکار تصمیم‌گیری، بلکه دلایل نیز متعلق به عامل هستند. این مورد در شرح فیشر و رویتزا پیش‌فرض قرار گرفته و تنها بر متعلق بودن سازوکارِ تصمیم به عامل تاکید شده است.

اما در استدلال ولوونکامپ به نظر می‌رسد، بدون این که ذکر شود، شهودهای از پیش فرض‌شده به بحث کشیده می‌شود و به دنبال آن این که دلایل می‌توانند متعلق به عامل بشری خاصی باشند یا خیر نیز محل بحث است.

با این حال، همان طور که گفته شد، می‌توان مورد ولوونکامپ را نیز به هر دو تعبیر اشاره‌شده تقلیل داد: آنچه مورد نظر ولوونکامپ است این است که دلیل‌ها به نحوی ما را به صورت متعین به سمت یک عامل خاص سوق دهند، درست مثل تعبیر اول که در آن در حال کشف این هستیم که آیا در زنجیرهٔ علّی رویداد قصدمندی مرتبط با یک عامل خاص وجود دارد یا خیر. اما مورد ولوونکامپ را می‌توان به نحو دیگری نیز فهمید. هرچند در رویکرد فیشر و رویتزا هم سازوکار (مطابق شرط دوم) و هم دلیل (مطابق پیش‌فرض) متعلق به عامل است، اما در مورد اسلحه‌های خودگردان شهوداً می‌توان تصور کرد که تنها یکی از این دو، یعنی یا سازوکار تصمیم یا دلیل، متعلق به عامل باشد و همچنان مسئولیت نتایج اقدامات

هوش مصنوعی به عامل نسبت داده شود. برای مثال، می‌توان تصور کرد که در شرح ولوونکامپ صرفاً دلیل متعلق به عامل باشد، یعنی دلایل وابسته-به-عامل باشند. و همین برای برقرار بودن شرط دوم کافی باشد. به عبارت دیگر، یک عامل شهوداً نسبت به هر سازوکاری در تصمیم، حتی مجزا و نامتعلق به عامل، که به دلایل آن عامل واکنش[1] نشان دهد و نسبت به آن بازکنش[2] داشته باشد، مسئول خواهد بود.

از سوی دیگر این امکان نیز وجود دارد که رابطهٔ مشخص میان تصمیمِ گرفته‌شده و دلیل‌های مرتبط با عامل (همانند تعبیر دوم) نه در زنجیرهٔ علّیِ سازوکار تصمیم، بلکه در فضای ذهنی-اجتماعیِ عامل‌های بشری متعین شود. یعنی در نهایت یک عامل بشری در یک فرایند اجتماعی بپذیرد که همهٔ دلایل (حتی اگر مستقل-از-عامل باشند) متعلق به اوست.

دلایل مورد اشاره این بار دلایل هنجاری[3] و اخلاقی هستند، نه دلایل انگیزشی[4]. دلیل هنجاری و دلیل انگیزشی تمایز روشنی دارند؛ برای مثال، اگر از رانندهٔ اتومبیلی بپرسیم که «برای چه به چپ پیچیدی؟» از دلیل انگیزشی او سؤال کردیم. اما اگر از او بپرسیم «آیا برای پیچیدن به چپ دلیلی داشتی؟» از دلیل هنجاری او پرسیده‌ایم (Veluwenkamp, 2022, p. 2). غالباً به ازای هر عملی یک دلیل انگیزشی وجود دارد، اما از میان آنها تنها دلیل‌های هنجاری قادر هستند که عمل را به لحاظ اخلاقی توجیه کنند.

با توجه به تقسیم مورد اشاره، باید انتظار داشته باشیم که سیستم‌های خودگردان نسبت به دلایل هنجاری واکنش داشته باشند. ویژگی دلیل‌های هنجاری این است که نسبت-به-عامل-خنثی هستند و به نظر می‌رسد که نمی‌توانند در تعبیر اول از پیشنهاد ولوونکامپ مشارکت داشته باشند. اما مطابق پیشنهاد ارائه‌شده در این مقاله، یعنی استفاده از تعبیر دوم برای مرتبط کردن دلایل به یک عامل، و نه سازوکار، می‌توان این تصور را داشت که دلایلْ هنجاری و لذا نسبت-به-عامل-خنثی باشند و در یک فرایند اجتماعی به یک عامل خاص مرتبط شوند. در چنین حالتی، سیستمی خواهیم داشت که می‌تواند نسبت به دلایل هنجاری واکنش‌پذیر باشد و نتیجهٔ ردگیری دلایل نیز به عامل مشخصی خواهد رسید.

نتیجه‌گیری

در این مقاله با شرح دیدگاه سانتونی د سیو و ون دن هاون از کنترل معنادار بشری در مورد اسلحه‌های خودگردان مجهز به هوش مصنوعی، و از دو شرطی که برای پر کردن شکاف مسئولیت ارائه شده (یعنی ردیابی و ردگیری)، تعبیرهای متفاوتی ارائه و بررسی شد. بیان شد که شرح ردگیری را به دو صورت متفاوت می‌توان فهمید. یک تعبیر شرایطی را شرح می‌دهد که در آن زنجیرهٔ علّی تاریخی که به یک عمل

1. Responsive
2. Reaction
3. Normative
4. Motivational

منجر می‌شود دارای خصلت خاصی است که در صورت وجود آن خصلت، یک عامل دارای مسئولیت خواهد بود. در تعبیر دوم، زنجیره علّی هرگز دخیل نخواهد بود و تغییر در فضای ذهنی و باورهای اجتماعی ایجاد می‌شود، به نحوی که این باورها کمک می‌کنند که یک سازوکار، متعلق به یک عامل قلمداد شود و در نتیجه مسئولیت تبعات آن به عامل نسبت داده شود. همچنین پیشنهاد شد که رابطهٔ تعلق می‌تواند نه رابطه‌ای میان عامل و سازوکار تصمیم، بلکه رابطه‌ای میان عامل و دلایل هنجاری قلمداد شود. از این پیشنهاد دو تعبیر متفاوت ارائه شد که در یکی زنجیره علّی دخیل است و در دیگری فضای باورها درگیر می‌شود. در نهایت پیشنهاد شد تعبیری که در آن دلایل هنجاری و اخلاقی به واسطه شکل‌گیری فضای ذهنی و باورهای اجتماعی به عامل نسبت داده می‌شود می‌تواند مسائل مطرح‌شدهٔ پیشین را دست‌کم روی کاغذ و در سطح نظری حل کند.

تعارض منافع

نویسنده هیچ‌گونه تعارض منافعی گزارش نکرده است.

References

Alasti, K. (2024). Moral dilemmas, amoral obligations, and responsible innovation; Two-dimensional "human control" over "autonomous" socio-technical systems. *Ethics, Policy & Environment*, 1-23. https://doi.org/10.1080/21550085.2024.2430151

Anderson, S. L. (2011). Philosophical concerns with machine ethics. In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 162-167). Cambridge University Press. https://doi.org/10.1017/CBO9780511978036.014

Barman, K. G., Wood, N., & Pawlowski, P. (2024). Beyond transparency and explainability: On the need for adequate and contextualized user guidelines for LLM use. *Ethics and Information Technology*, 26(3), 47. https://link.springer.com/article/10.1007/s10676-024-09778-2

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19. https://www.jstor.org/stable/3328150

Dautenhahn, K. (1998). The art of designing socially intelligent agents: Science, fiction, and the human in the loop. *Applied Artificial Intelligence*, 12(7-8), 573-617. https://doi.org/10.1080/088395198117550

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.

Fischer, J. M., & Ravizza, M. (2000). Précis of responsibility and control: A theory of moral responsibility. *Philosophy and Phenomenological Research*, *LXI*(2), 441-445. https://doi.org/10.2307/2653660

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14, 349-379. https://link.springer.com/article/10.1023/B:MIND.0000035461.63578.9d

Frankfurt, H. (1993). What we are morally responsible for. In J. M. Fischer & M. Ravizza (Eds.), *Perspectives on moral responsibility* (pp. 286-294). Cornell University Press.

Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy*, 66(23), 829-839 https://doi.org/10.2307/2023833

Hindriks, F., & Veluwenkamp, H. (2023). The risks of autonomous machines: from responsibility gaps to control gaps. *Synthese*, 201(1), 21. https://link.springer.com/article/10.1007/s11229-022-04001-5

Hoefer, C. (2024). Causal determinism. In N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Summer 2024 Edition). URL = https://plato.stanford.edu/archives/sum2024/entries/determinism-causal/

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and information technology*, 8, 195-204. https://link.springer.com/article/10.1007/s10676-006-9111-5

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, 6, 175-183. https://link.springer.com/article/10.1007/s10676-004-3422-1

Mecacci, G., & Santoni de Sio, F. (2020). Meaningful human control as reason-responsiveness: The case of dual-mode vehicles. *Ethics and Information Technology*, 22(2), 103-115. https://link.springer.com/article/10.1007/s10676-019-09519-w

Müller, V. C. (2023). Ethics of artificial intelligence and robotics. In N. ZAlta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2023 Edition). URL = https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/

Owen, R., Stilgoe, J., Macnaghten, P., Gorman, M., Fisher, E., & Guston, D. (2013). A framework for responsible innovation. In R. Owen, J. R. Bessant & M. Heintz (Eds), *Responsible innovation: Managing the responsible emergence of science and innovation in society* (pp. 27-50). Wiley.

Robbins, S. (2019). A misdirected principle with a catch: Explicability for AI. *Minds and Machines*, 29(4), 495-514. https://link.springer.com/article/10.1007/s11023-019-09509-3.

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, *5*. https://doi.org/ 10.3389/frobt.2018.00015.

Santoni de Sio, F., Mecacci, G., Calvert, S., Heikoop, D., Hagenzieker, M., & van Arem, B. (2022). Realising meaningful human control over automated driving systems: A multidisciplinary approach. *Minds and Machines*, 33(4), 587-611. https://link.springer.com/article/10.1007/s11023-022-09608-8

Sullins, J. P. (2011). When is a robot a moral agent. *Machine Ethics*, 6, 151-161.

Veluwenkamp, H. (2022). Reasons for meaningful human control. *Ethics and Information Technology*, 24(4), 51. https://link.springer.com/article/10.1007/s10676-022-09673-8

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135, 364-381. https://doi.org/10.1016/j.future.2022.05.014