شناخت

Shenakht

Shenakht is a persian word which means knowledge

Print ISSN: 2008-7322 Online ISSN: 2588-7203

Causality in the Age of Machine Learning: A Philosophical Analysis of Causal Discovery by Artificial Intelligence

Sayyed Mahdi Biabanaki

Assistant Professor, Department of Philosophy of Science, University of Isfahan, Isfahan, Iran.
Email: mehdibiabanaki@gmail.com

Article Info

Article type:
Research Article

Article history:
Received: 31 January 2026
Accepted: 09 June 2026
Published Online: 27 June 2026

Keywords:
Causality, Causal Discovery, Intervention, Artificial intelligence, Causal AI

ABSTRACT

Recent advances in machine learning, particularly in data-driven and predictive modeling, have led to remarkable successes across a wide range of scientific and practical domains. At the same time, these successes have raised fundamental questions in the philosophy of science concerning the role of explanation and causality in a scientific practice increasingly centered on statistical prediction. In response to the limitations of purely correlation-based models a growing body of work under the heading of Causal Artificial Intelligence has sought to reintroduce causal structure into machine learning.

This paper offers a philosophical analysis of the notion of causality employed in contemporary machine learning, with particular attention to causal discovery and intervention-based frameworks. It argues that the causal relations represented or inferred by machine learning models should be understood neither as direct representations of the world's fundamental causal mechanisms nor as merely instrumental devices for improving predictive performance. Instead, the paper proposes an alternative interpretation according to which causality in machine learning constitutes an intervention-dependent, mid-level explanatory structure.

On this account, the epistemic legitimacy of causal models derives not from strong metaphysical commitments, but from their ability to remain invariant under interventions, support counterfactual reasoning, and enable out-of-distribution generalization. By situating causal modeling between predictive success and metaphysical realism, the paper shows how causal concepts can play an indispensable explanatory and decision-theoretic role in data-driven science without exceeding the epistemic limits of machine learning. This interpretation provides a coherent philosophical framework for understanding the renewed significance of causality in contemporary artificial intelligence.

Cite this article: Biabanaki, S. M. (2026). Causality in the Age of Machine Learning: A Philosophical Analysis of Causal Discovery by Artificial Intelligence. *Shenakht*, 19(91/2), 151-174.
http://doi.org/10.48308/kj.2026.243371.1404

Extended Abstract

The rapid development of machine learning and artificial intelligence has transformed contemporary scientific practice by enabling the extraction of highly complex statistical patterns from large-scale datasets. In many domains, including medicine, economics, social sciences, and natural language processing, machine learning systems have demonstrated remarkable predictive performance without relying on explicit theoretical assumptions about the underlying mechanisms of the phenomena they model. This success has intensified a broader methodological tendency within contemporary science that some scholars describe as a "predictive turn," in which predictive accuracy increasingly becomes a dominant criterion for evaluating scientific models. Within such a framework, explanation and causal understanding risk being treated as secondary concerns or, in some cases, as dispensable elements of scientific inquiry. However, the practical success of purely predictive systems simultaneously raises fundamental philosophical questions regarding the relationship between prediction, explanation, and causation, particularly when intelligent systems are expected not merely to classify or predict, but also to support intervention, decision-making, and reasoning under changing conditions. The present study addresses this problem by examining the philosophical status of causality in the age of machine learning and by investigating whether contemporary artificial intelligence systems genuinely discover causal structures or merely identify highly efficient statistical regularities.

The article begins from the observation that classical machine learning systems have largely been designed around statistical association and predictive optimization. Traditional machine learning models generally operate by identifying correlations within observational data and constructing predictive mappings without incorporating explicit assumptions concerning the causal mechanisms generating those data. Although such approaches may achieve impressive performance within training environments, they frequently encounter difficulties when deployed under conditions involving distributional shifts, environmental changes, or interventions. Problems such as out-of-distribution generalization, fragility under changing contexts, and limitations in counterfactual reasoning have revealed the inadequacy of purely correlation-based systems for many scientific and practical purposes.

In response to these limitations, recent developments in causal artificial intelligence have attempted to integrate concepts such as intervention, invariance, structural dependence, and counterfactual reasoning into machine learning systems. Building upon the work of Pearl, Woodward, Schölkopf, and others, these approaches seek to move beyond predictive associations and toward models capable of supporting explanatory understanding and stable generalization. Concepts such as structural causal

models, intervention calculus, and causal discovery have become central tools within this emerging research program. However, despite rapid technical advances, many discussions within causal AI implicitly assume philosophical commitments that often remain unexamined. Technical studies frequently presuppose that causal models represent genuine structures of reality, while alternative perspectives treat causal structures merely as practical tools for improving performance. Consequently, important philosophical questions remain unresolved: What kind of causality is being discovered by AI systems? What sort of representational status should be attributed to causal models in machine learning? And under what conditions can causal structures extracted from data be regarded as epistemically meaningful?

To address these questions, this study analyzes contemporary causal AI through the lens of philosophy of science and examines several influential philosophical approaches concerning the nature of scientific representation and causal explanation. The discussion first considers a realist interpretation according to which successful causal models provide increasingly accurate representations of real causal structures underlying observable phenomena. Within this perspective, the success of causal representation learning and causal discovery methods may appear to support a form of scientific realism, since such systems seem capable of identifying stable structures that persist beyond particular datasets or environments. Nevertheless, this position faces significant challenges. Causal discovery algorithms rely on strong assumptions such as causal sufficiency, faithfulness conditions, and structural stability, many of which are difficult to verify independently. Therefore, moving from predictive success to ontological commitment risks introducing forms of optimistic realism that may exceed what available evidence can justify.

The article then examines interventionist and pragmatic perspectives, especially those associated with Woodward's account of causation. Interventionist theories attempt to avoid heavy metaphysical commitments by defining causality in terms of manipulability and control rather than in terms of hidden mechanisms or necessary connections. This perspective provides important practical advantages for machine learning because it explains how causal reasoning can guide decision-making and intervention without requiring complete knowledge of underlying reality. However, if causality is defined solely through intervention and operational utility, an important conceptual problem emerges. The distinction between genuinely causal systems and highly sophisticated predictive systems may become blurred if successful intervention alone becomes the principal criterion for causal legitimacy.

In response to these limitations, the article proposes a distinct philosophical framework described as a critical-structural perspective. This approach attempts to preserve insights from realism, interventionism, and pragmatic accounts while avoiding their respective weaknesses. The proposed view rejects both naïve realism and extreme

instrumentalism. It argues that causal models in machine learning should not be understood as complete representations of the essential nature of reality, nor merely as computational devices for prediction. Instead, causal models are interpreted as representations of stable structural relations whose epistemic significance derives from their capacity to support intervention, explanation, and robust generalization across changing environments.

A crucial component of this argument involves distinguishing between causality at the level of models and causality at the level of target systems. A model may contain causal organization internally without necessarily claiming complete correspondence with external reality. The existence of causal structure within a model therefore does not automatically imply metaphysical truth regarding the world itself. Rather, the epistemic value of causal structures depends on whether they capture stable relationships capable of remaining invariant under meaningful interventions and environmental changes. Within this framework, causality becomes less a matter of uncovering ultimate reality and more a matter of identifying explanatory structures that remain robust under transformation.

The paper further argues that concepts such as intervention, invariance, and out-of-distribution generalization are central for understanding why causal models differ fundamentally from static predictive models. Purely statistical systems often fail because they learn relationships tied to particular distributions rather than relationships reflecting deeper structural dependencies. Causal models, by contrast, attempt to identify patterns whose stability persists under changing conditions. Consequently, causality is interpreted not primarily as a metaphysical label attached to hidden entities but rather as an epistemic criterion for robustness, explanatory power, and rational action.

Ultimately, the study concludes that causality in machine learning should not be interpreted either as the direct discovery of ultimate truths about reality or as a merely instrumental computational technique. Instead, causality should be understood as the representation of intervention-dependent stable structures that enable explanation, generalization, and rational decision-making in data-driven environments. Such a position offers a balanced philosophical framework capable of integrating practical developments in artificial intelligence with enduring concerns in philosophy of science regarding scientific representation, explanation, and the nature of causal understanding.

شناخت

Shenakht

Shenakht is a persian word which means knowledge

Print ISSN: 2008-7322 Online ISSN: 2588-7203

علّیت در عصر یادگیری ماشین: تحلیلی فلسفی از کشف علّیت توسط هوش مصنوعی

سیدمهدی بیابانکی

استادیار، گروه فلسفه علم، دانشگاه اصفهان، اصفهان، ایران. رایانامه: mehdibiabanaki@gmail.com

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۴۰۴/۱۱/۱۱

تاریخ پذیرش: ۱۴۰۵/۰۳/۱۹

تاریخ انتشار: ۱۴۰۵/۰۴/۰۶

کلیدواژه‌ها:

علیت، کشف علیت، مداخله، هوش مصنوعی، هوش مصنوعی علّی

چکیده

پیشرفت‌های اخیر در یادگیری ماشین، به‌ویژه در چارچوب مدل‌های داده‌محور و پیش‌بینانه، موفقیت‌های چشمگیری در حوزه‌های مختلف علمی و کاربردی به همراه داشته است. بااین‌حال، این موفقیت‌ها، هم‌زمان، پرسش‌های بنیادینی را در سطح فلسفه علم برانگیخته‌اند، به‌ویژه دربارهٔ جایگاه مفاهیمی چون تبیین و علّیت در علمی که به‌طورِ فزاینده بر پیش‌بینی آماری تکیه دارد. در واکنش به محدودیت‌های مدل‌های صرفاً همبستگی‌محور (ازجمله شکنندگی دربرابرِ تغییر محیط و ناتوانی در پاسخ‌گویی به پرسش‌های مداخله‌ای) ادبیات نوظهوری تحت عنوان «هوش مصنوعی علّی» شکل گرفته است که می‌کوشد ساختارهای علّی را به‌طورِ صریح وارد یادگیری ماشین کند.

این مقاله، با رویکردی تحلیلی در تقاطع فلسفه علم و پژوهش‌های معاصر هوش مصنوعی، به بررسی ماهیت علّیتی می‌پردازد که در مدل‌های یادگیری ماشین، به‌ویژه در حوزه کشف علّیت، بازنمایی یا استنتاج می‌شود. استدلال اصلی مقاله این است که علّیت، در این زمینه، نه باید بازنمایی مستقیم سازوکارهای بنیادین جهان فهم شود و نه صرفاً به‌مثابهٔ ابزاری مهندسی برای بهبود پیش‌بینی. درمقابل، مقاله پیشنهاد می‌کند که علّیت در یادگیری ماشین را می‌توان به‌طورِ موجه ساختاری تبیینی، میان‌سطحی و وابسته به مداخله تفسیر کرد که اعتبار معرفت‌شناختی خود را از ناوردایی تحت مداخله و توانایی تعمیم فراتوزیعی به دست می‌آورد. در پایان، مقاله نشان می‌دهد که این تفسیر بدیل می‌تواند هم موفقیت عملی رویکردهای علّی در یادگیری ماشین را توضیح دهد و هم از تعهدات متافیزیکی سنگین پرهیز کند و، بدین‌ترتیب، چارچوبی منسجم برای فهم علّیت در علم داده‌محور معاصر فراهم آورد.

استناد: بیابانکی، سیدمهدی (۱۴۰۵). علّیت در عصر یادگیری ماشین: تحلیلی فلسفی از کشف علّیت توسط هوش مصنوعی. شناخت، ۹۱(۲)۱۹، ۱۵۱-۱۷۴

DOI: http://doi.org/10.48308/kj.2026.243371.1404

ناشر: دانشگاه شهید بهشتی

مقدمه

در دهه‌های اخیر، یادگیری ماشین و به‌ویژه روش‌های مبتنی بر داده‌های کلان، به یکی از مؤثرترین ابزارهای معرفت علمی و تصمیم‌گیری عملی بدل شده‌اند. این روش‌ها در حوزه‌هایی متنوع (از بینایی ماشین و پردازش زبان طبیعی تا پزشکی، اقتصاد و سیاست‌گذاری) موفقیت‌های چشمگیری در پیش‌بینی و طبقه‌بندی از خود نشان داده‌اند. بااین‌حال، همین موفقیت عملی، پرسش‌های بنیادینی را در سطح فلسفه علم برانگیخته است؛ به‌ویژه دربارهٔ نسبت میان پیش‌بینی، تبیین و علّیت در علوم داده‌محور.

به‌طورِ سنتی، علّیت از مفاهیم محوری در فلسفه علم بوده است. از نقد هیوم بر پیوند ضروری میان علت و معلول گرفته تا تلاش‌های قرن بیستم برای صوری‌سازی علّیت در قالب‌های آماری و منطقی، همواره این پرسش مطرح بوده است که علم چگونه و تا چه حد می‌تواند از همبستگی‌های مشاهده‌ای به روابط علّی دست یابد. در این سنت، علّیت نه صرفاً ابزاری محاسباتی بلکه مفهومی مرتبط با تبیین علمی، کنترل پدیده‌ها و فهم سازوکارهای مولد جهان تلقی شده است.

یادگیری ماشین کلاسیک، با تمرکز بر بهینه‌سازی پیش‌بینی و استخراج الگوهای آماری، در نگاه نخست فاصله‌ای آشکار با این دغدغه‌های فلسفی دارد. بسیاری از مدل‌های موفق یادگیری ماشین، بدون تعهد صریح به ساختارهای علّی، قادر به تولید پیش‌بینی‌های دقیق‌اند. این وضعیت باعث شده است برخی پژوهشگران از نوعی «چرخش پیش‌بینانه» در علم معاصر سخن بگویند و حتی تبیین و علّیت را اموری ثانوی یا زائد تلقی کنند (Weiskopf, 2022; Shmueli, 2010; Douglas, 2009). بااین‌حال، محدودیت‌های این رویکرد (به‌ویژه ناتوانی در تعمیم فراتوزیعی، شکنندگی دربرابرِ تغییر محیط و دشواری در تصمیم‌گیری مداخله‌ای) به‌تدریج آشکار شده است.

در واکنش به این محدودیت‌ها، از اواخر دهه ۲۰۱۰، موجی از پژوهش‌ها در حوزه «هوش مصنوعی علّی»[1] شکل گرفته است. این پژوهش‌ها، با الهام از چارچوب‌های علّی در آمار و فلسفه علم، می‌کوشند مفاهیمی چون مداخله، ناوردایی و ساختار علّی را وارد مدل‌های یادگیری ماشین کنند. هدف این رویکردها صرفاً افزایش دقت پیش‌بینی نیست بلکه دستیابی به مدل‌هایی است که بتوانند به پرسش‌های مداخله‌ای پاسخ دهند، در محیط‌های جدید پایدار بمانند و نوعی تبیین ارائه کنند.

درواقع، «هوش مصنوعی علّی» به رویکردی در پژوهش‌های معاصر هوش مصنوعی اطلاق می‌شود که می‌کوشد مفاهیم علّی، نظیر ساختار علّی، مداخله و استدلال خلاف‌واقع را به‌طورِ صریح در مدل‌های یادگیری ماشین وارد کند. برخلافِ یادگیری ماشین کلاسیک که عمدتاً بر بهینه‌سازی پیش‌بینی مبتنی بر همبستگی‌های آماری تمرکز دارد، هوش مصنوعی علّی هدف خود را توسعه مدل‌هایی قرار می‌دهد که قادر به تحلیل اثر مداخلات، حفظ ناوردایی تحت تغییر محیط و تعمیم فراتوزیعی باشند (Pearl, 2000; Peters & Schölkopf, 2017). این رویکرد در واکنش به

[1] Causal AI

محدودیت‌های مدل‌های صرفاً داده‌محور پدید آمده و امروزه یکی از مسیرهای کلیدی برای پیوند میان پیش‌بینی، تبیین و کنترل عقلانی در سیستم‌های هوشمند شناخته می‌شود (Schölkopf et al., 2021).

باوجودِ رشد سریع این ادبیات، پرسش‌های فلسفی بنیادین دربارهٔ ماهیت علّیتی که در یادگیری ماشین مدل‌سازی یا «کشف» می‌شود اغلب به‌صورتِ ضمنی و مفروض باقی مانده‌اند. بسیاری از آثار فنی، بدون تحلیل فلسفی صریح، فرض می‌کنند که مدل‌های علّی بازنمایی‌هایی از ساختار واقعی جهان‌اند یا، برعکس، علّیت را صرفاً ابزاری مهندسی برای بهبود عملکرد سیستم‌ها می‌دانند. این دو تلقی (واقع‌گرایانه و ابزارگرایانه) هریک پیامدهای معرفت‌شناختی و روش‌شناختی متفاوتی دارند که به‌ندرت به‌طورِ نظام‌مند بررسی شده‌اند.

مسئله محوری این مقاله دقیقاً در همین نقطه شکل می‌گیرد: علّیتی که در مدل‌های یادگیری ماشین بازنمایی یا استنتاج می‌شود چه جایگاهی در فلسفه علم دارد؟ آیا باید آن را بازنمایی واقع‌گرایانه‌ای از سازوکارهای علّی جهان فهم کرد یا صرفاً ابزاری پیش‌بینی‌گر و تصمیم‌ساز در چارچوب علم داده‌محور؟ و اگر هیچ‌یک از این دو تلقی به‌تنهایی رضایت‌بخش نیست، چه تفسیر بدیلی می‌تواند هم موفقیت عملی مدل‌های علّی را توضیح دهد و هم از تعهدات متافیزیکی سنگین پرهیز کند؟

این مقاله، با اتخاذ رویکردی تحلیلی در تقاطع فلسفه علم و پژوهش‌های معاصر هوش مصنوعی، استدلال می‌کند که علّیت در یادگیری ماشین نه بازنمایی مستقیم سازوکارهای بنیادین جهان است و نه صرفاً یک ابزار محاسباتی بی‌تعهد بلکه نوعی ساختار تبیینی میان‌سطحی و وابسته به مداخله است. این تفسیر می‌کوشد جایگاه علّیت را به‌گونه‌ای صورت‌بندی کند که هم با محدودیت‌های معرفتی مدل‌های داده‌محور سازگار باشد و هم نقش غیرقابل‌جایگزین علّیت در تبیین، تعمیم و تصمیم‌گیری عقلانی را حفظ کند. در ادامه، ابتدا، پیشینه پژوهش‌های فلسفی و فنی مرتبط با علّیت در یادگیری ماشین مرور می‌شود و سپس دیدگاه‌های شاخص در این حوزه تحلیل و مقایسه می‌گردند و، درنهایت، چارچوب مفهومی پیشنهادی مقاله به‌عنوانِ تفسیری بدیل از علّیت در عصر یادگیری ماشین ارائه و ارزیابی می‌شود.

پیشینه پژوهش

بحث علّیت یکی از مفاهیم بنیادی در فلسفه علم است که از آثار کلاسیک دیوید هیوم آغاز شده و تا نظریه‌های معاصر ادامه یافته است. هیوم، با نقد مفهوم پیوند ضروری میان علت و معلول، علّیت را به عادت ذهنی و استقرا فروکاست و، بدین‌ترتیب، مسئله توجیه استنتاج علّی را به یکی از مسائل محوری فلسفه علم تبدیل کرد. این مسئله در قرن بیستم با تلاش‌هایی برای صوری‌سازی و بازسازی مفهوم علّیت در چارچوب‌های منطقی و آماری پی گرفته شد، به‌گونه‌ای که علّیت به‌تدریج از یک مفهوم متافیزیکی به ابزاری روش‌شناختی در علم تبدیل گردید.

نقطه عطف معاصر در این مسیر آثار جودیا پرل است. پرل در کتاب خود[1] چارچوب «مدل‌های ساختاری علّی»[2] را معرفی می‌کند و تمایزی دقیق میان همبستگی آماری و رابطه علّی برقرار می‌سازد (Pearl, 2000, pp. 9–15). او نشان می‌دهد که داده‌های مشاهده‌ای به‌تنهایی برای استنتاج علّی کافی نیستند و برای پاسخ به پرسش‌های علّی نیازمند مداخله[3] هستیم. معرفی حساب مداخله‌ای[4] ابزار صوری‌ای فراهم می‌کند که به واسطه آن می‌توان اثر مداخلات فرضی را از روی داده‌ها و ساختارهای گرافی استنتاج کرد (Pearl, 2000, pp. 68–75). توضیح اینکه، حساب مداخله‌ای مجموعه‌ای از قواعد و ابزارهای ریاضی و گرافی است که امکان تحلیل و پیش‌بینی اثر مداخلات در یک سیستم علّی را فراهم می‌کند. در این حساب، به‌جای توجه صرف به روابط مشاهده‌ای یا همبستگی‌ها، پژوهشگر می‌تواند تأثیر تغییر مستقیم یک متغیر (مداخله) بر سایر متغیرها را براساسِ گراف علّی محاسبه کند. این چارچوب بعدها به یکی از پایه‌های اصلی پژوهش‌های علّیت در علوم تجربی و علوم داده تبدیل شد.

درمقابل، یادگیری ماشین کلاسیک عمدتاً بر پیش‌بینی آماری و استخراج الگوهای همبستگی تمرکز داشته است. بسیاری از الگوریتم‌های یادگیری ماشین، به‌ویژه در شاخه یادگیری عمیق، بدون تعهد صریح به ساختارهای علّی طراحی شده‌اند و موفقیت آن‌ها اغلب براساسِ دقت پیش‌بینی سنجیده می‌شود نه قدرت تبیین. این وضعیت باعث شده است که برخی پژوهشگران از «غلبه پیش‌بینی بر تبیین» در علم داده‌محور سخن بگویند (Shmueli, 2010). بااین‌حال، از اواخر دهه ۲۰۱۰، نقدهای فلسفی و روش‌شناختی فزاینده‌ای متوجه این رویکرد شده است.

در این زمینه، مقاله تأثیرگذار شُلکوپف[5] نقطه اتصال صریح میان فلسفه علّیت و یادگیری ماشین را برقرار می‌کند. شُلکوپف استدلال می‌کند که بسیاری از مشکلات بنیادین یادگیری ماشین، ازجمله ناتوانی در «تعمیم خارج از توزیع[6]»، ریشه‌ای علّی دارند و بدون مدل‌سازی علّیت حل نخواهند شد (Schölkopf, 2019, pp. 2–4). از نظر او، یادگیری ماشین صرفاً همبستگی‌ها را می‌آموزد، درحالی‌که تعمیم پایدار نیازمند فهم ساختار علّی مولد داده‌هاست. این دیدگاه، یادگیری ماشین را از یک ابزار صرفاً آماری به پروژه‌ای نزدیک‌تر به فلسفه علم سوق می‌دهد.

این خط فکری در آثار بعدی شُلکوپف و همکارانش بسط یافته است. او و همکارانش نشان می‌دهند که یکی از چالش‌های اساسی هوش مصنوعی معاصر یادگیری بازنمایی‌هایی است که نه‌تنها فشرده و مفید برای پیش‌بینی‌اند بلکه با متغیرهای علّی زیربنایی جهان نیز هم‌تراز باشند (Schölkopf et al., 2021, pp. 3–6). این دسته پژوهش‌ها به‌طورِ ضمنی بر پیش‌فرضی فلسفی تکیه دارند: اینکه جهان دارای ساختار علّی واقعی است و بازنمایی‌های موفق علمی باید به نوعی این ساختار را منعکس کنند، پیش‌فرضی که با واقع‌گرایی علمی همخوانی دارد.

---

[1] *Causality: Models, Reasoning, and Inference*
[2] Structural Causal Models
[3] intervention
[4] do-calculus
[5] *Causality for Machine Learning*
[6] out-of-distribution generalization

در کنار این رویکرد، پژوهش‌های متعددی به بررسی امکان «کشف علّیت»[1] از داده‌ها پرداخته‌اند. مرورهای اخیر در حوزه کشف علیت نشان می‌دهند که اگرچه الگوریتم‌های متعددی برای استخراج ساختارهای علّی پیشنهاد شده‌اند اما این روش‌ها همواره متکی بر فروض قوی (مانند عدم وجود متغیرهای مداخله‌گر پنهان یا ثبات ساختار علّی) هستند که ماهیتی فلسفی و متافیزیکی دارند (Lamsaf et al., 2025, pp. 2–5). این نکته بار دیگر نشان می‌دهد که علّیت در یادگیری ماشین صرفاً یک مسئله فنی نیست بلکه به پیش‌فرض‌های فلسفه علم وابسته است.

از منظر فلسفه علم، برخی نویسندگان استدلال کرده‌اند که موفقیت یادگیری ماشینِ غیرعلّی می‌تواند به نفع ابزارگرایی علمی تفسیر شود، دیدگاهی که علم را نه به‌عنوانِ کشف حقیقت بلکه به‌مثابهٔ ابزاری برای پیش‌بینی موفق می‌داند. بااین‌حال، پژوهش‌های جدید در حوزه هوش مصنوعی علّی نشان می‌دهند که، بدون درنظرگرفتنِ علّیت، سیستم‌های هوش مصنوعی در حوزه‌هایی مانند پزشکی، سیاست‌گذاری و علوم اجتماعی با محدودیت‌های جدی مواجه می‌شوند، به‌ویژه هنگامی که پای تصمیم‌گیری مداخله‌ای و نه صرفاً پیش‌بینی در میان است (Kesh & Whitworth, 2025, pp. 7–10).

درمجموع، پیشینه پژوهش نشان می‌دهد که بحث علّیت در عصر یادگیری ماشین در تقاطع سه سنت فکری شکل گرفته است: (۱) فلسفه کلاسیک و معاصر علم دربارهٔ علّیت و استقرا، (۲) چارچوب‌های صوری علّی در آمار و علوم داده، و (۳) نقدهای روش‌شناختی به یادگیری ماشین مبتنی بر همبستگی. این ادبیات حاکی از آن است که پرسش از توانایی هوش مصنوعی در کشف یا بازنمایی علّیت، نه‌تنها یک مسئله فنی بلکه مسئله‌ای عمیقاً فلسفی است که مستقیماً به ماهیت تبیین علمی، تعمیم و واقع‌گرایی علمی مربوط می‌شود. پژوهش‌های موجود دربارهٔ علّیت در یادگیری ماشین عمدتاً در دو مسیر پیش رفته‌اند. یک دسته از آثار مانند پرل و شلکوپف و پژوهش‌های «کشف علیت» بر توسعه چارچوب‌های صوری، الگوریتم‌ها و مدل‌های گرافی برای استنتاج یا بازنمایی روابط علّی تمرکز دارند. مسئله اصلی در این آثار، چگونگی پیاده‌سازی علّیت در سیستم‌های یادگیری ماشین است. دسته دیگر به مزایای عملی علّیت برای بهبود تعمیم‌پذیری، تبیین‌پذیری و تصمیم‌گیری در سیستم‌های هوش مصنوعی می‌پردازد و علّیت را عمدتاً به‌عنوانِ راه‌حلی برای محدودیت‌های آماری یادگیری ماشین معرفی می‌کند.

باوجودِ غنای این دو مسیر، یک خلأ نظری مهم باقی مانده است و آن این است که در ادبیات موجود به‌ندرت به این پرسش پرداخته شده است که علّیتی که در یادگیری ماشین مدل‌سازی یا «کشف» می‌شود از چه نوعی است و چه جایگاهی در فلسفه علم دارد. به بیان دقیق‌تر، بیشتر پژوهش‌ها، به‌طورِ ضمنی، فرض می‌کنند که علّیت یک ساختار عینی در جهان است که می‌توان آن را از داده‌ها استخراج کرد یا اینکه مدل‌های علّی در یادگیری ماشین بازنمایی‌های معتبری از سازوکارهای واقعی‌اند. اما این فروض به‌ندرت به‌صورتِ صریح از منظر فلسفه علم تحلیل شده‌اند. مشخص نیست که آیا «علّیت» در یادگیری ماشین به معنای علّیت واقع‌گرایانه علمی است یا صرفاً یک ابزار مدل‌سازی مفید و آیا موفقیت

[1] causal discovery

عملی مدل‌های علّی در هوش مصنوعی به نفع واقع‌گرایی علمی است یا می‌تواند همچنان با ابزارگرایی سازگار باشد. براساسِ این شکاف، پرسش اصلی پژوهش حاضر این است که آیا علّیتی که در مدل‌های یادگیری ماشین بازنمایی یا استنتاج می‌شود باید بازنمایی واقع‌گرایانه‌ای از سازوکارهای علّی جهان تفسیر شود یا صرفاً ابزاری پیش‌بینانه و مداخله‌محور در چارچوب علم داده‌محور است؟

کشف علیت

اصطلاح «کشف علیت» به حوزه‌ای از پژوهش در تقاطع آمار، یادگیری ماشین و فلسفه علم اشاره دارد که هدف آن استنتاج یا بازسازی ساختارهای علّی از داده‌ها است، به‌ویژه در شرایطی که آزمایش‌های مداخله‌ای مستقیم در دسترس نیست یا پرهزینه و ناممکن است. مسئله محوری در این حوزه آن است که چگونه می‌توان از داده‌های مشاهده‌ای (که دراصل فقط اطلاعات هم‌وقوعی در اختیار می‌گذارند) به روابط علّی جهت‌دار میان متغیرها دست یافت. در این معنا، کشف علیت نه به پیش‌بینی بلکه به کشف ساختار می‌پردازد، ساختاری که معمولاً به‌صورتِ یک گراف علّی نمایش داده می‌شود و بیانگر جهت و وابستگی علّی میان متغیرهاست (Glymour et al., 2019).

بنابراین، در این مقاله که پژوهشی در حوزه یادگیری ماشین و فلسفه علم است، کشف علّیت به حوزه‌ای اطلاق می‌شود که هدف آن استنتاج ساختارهای علّی میان متغیرها از داده‌های مشاهده‌ای است. برخلافِ روش‌های پیش‌بینی صرف که تنها روابط همبستگی را مدل می‌کنند، کشف علّیت می‌کوشد جهت و ساختار روابط علّی را شناسایی کند و فهمی از نظام علّیت موجود در داده‌ها ارائه دهد. یکی از چالش‌های اصلی کشف علّیت این است که اغلب داده‌ها مشاهده‌ای هستند، یعنی پژوهشگر تنها مشاهده می‌کند و تغییری در سیستم ایجاد نمی‌کند. در این شرایط، روابط همبستگی به‌تنهایی نمی‌توانند دلالت علّی داشته باشند و، بنابراین، روش‌ها ناگزیر به تکیه بر فروض ساختاری مانند شرط مارکوف علّی، وفاداری و عدم وجود متغیرهای پنهان هستند (Glymour et al., 2019). درحالی‌که یادگیری ماشین کلاسیک بر پیش‌بینی تمرکز دارد، کشف علیت به دنبال شناسایی جهت و ساختار روابط است، حتی اگر این کار به قیمت کاهش دقت پیش‌بینی انجام شود.

نکته ای که تذکر آن در اینجا لازم است این است که بحث از علّیت در مدل‌های یادگیری ماشین مستلزم تفکیک میان دو سطح متفاوت است: نخست، علّیت در سطح درونی مدل و، دوم، علّیت در سطح جهان هدف یا واقعیت خارجی. در سطح نخست، یک مدل می‌تواند دارای ساختاری علّی باشد، یعنی متغیرها و روابط آن به‌گونه‌ای سازمان یابند که امکان تحلیل مداخله، وابستگی جهت‌دار و استدلال خلاف‌واقع را فراهم کنند. بااین‌حال، وجود چنین ساختاری لزوماً به این معنا نیست که مدل مدعی بازنمایی کامل یا مستقیم سازوکارهای واقعی جهان است.

این تمایز در فلسفه مدل‌ها اهمیتی اساسی دارد. بسیاری از مدل‌ها واجد ساختارهای علّی درونی‌اند، بی‌آنکه تعهدی واقع‌گرایانه نسبت به جهان هدف داشته باشند. یک روایت داستانی، شبیه‌سازی رایانه‌ای یا مدل اقتصادی ممکن

است روابطی علّی را در سطح ساختار خود بازنمایی کند، درحالی‌که نسبت آن با واقعیت خارجی صرفاً ابزاری، اکتشافی یا تقریبی باشد. بنابراین، وجود سازمان علّی در یک مدل، به‌خودیِ‌خود، برای اثبات صدق متافیزیکی آن دربارهٔ جهان کافی نیست.

مقاله حاضر نیز از چنین تمایزی تبعیت می‌کند. هنگامی که از «کشف علّیت» در مدل‌های یادگیری ماشین سخن گفته می‌شود، مقصود این نیست که مدل‌ها ضرورتاً به ساختار نهایی و کامل علّیت در جهان دست می‌یابند. ادعای مقاله محدودتر و معرفت‌شناختی‌تر است: مدل‌های یادگیری ماشین می‌توانند ساختارهایی را استخراج کنند که از نظر مداخله‌ای پایدار، از نظر تبیینی کارآمد و از نظر تعمیم در محیط‌های جدید نسبتاً ناوردا باشند. از این منظر، اهمیت فلسفی این مدل‌ها نه در بازنمایی ذات نهایی جهان بلکه در توانایی آن‌ها برای حفظ روابط ساختاری پایدار تحت تغییر و مداخله است.

دیدگاه‌های شاخص دربارهٔ علّیت در یادگیری ماشین

مسئله علّیت در یادگیری ماشین را می‌توان در امتداد مناقشات کلاسیک فلسفه علم دربارهٔ واقع‌گرایی، ابزارگرایی و ماهیت تبیین علمی فهم کرد. ورود چارچوب‌های علّی به یادگیری ماشین نه‌تنها یک تحول فنی بلکه احیای پرسش‌های بنیادینی است که پیش‌تر در فلسفه علم مطرح شده بودند، اینکه آیا نظریه‌های علمی ساختار واقعی جهان را بازنمایی می‌کنند یا صرفاً ابزارهایی برای پیش‌بینی موفق‌اند و اینکه تبیین علمی چه نسبتی با پیش‌بینی دارد. دیدگاه‌های موجود دربابِ علّیت در یادگیری ماشین را می‌توان در چند رویکرد شاخص دسته‌بندی کرد.

۱. دیدگاه واقع‌گرایانه علّی: علّیت به‌مثابهٔ ساختار عینی جهان

دیدگاه واقع‌گرایانه علّی بر این فرض اساسی استوار است که روابط علّی ساختارهایی واقعی و مستقل از مدل‌ها و بازنمایی‌های ما هستند و هدف علم (و به‌تبعِ آن یادگیری ماشین) کشف یا نزدیک شدن به این ساختارهاست. در این چارچوب، علّیت نه صرفاً ابزاری برای پیش‌بینی بلکه بیانگر سازوکارهای مولدی است که پدیده‌ها را در جهان به وجود می‌آورند.

پرل از مهم‌ترین نمایندگان این موضع است. او تأکید می‌کند که تمایز میان همبستگی آماری و رابطه علّی تمایزی هستی‌شناختی است نه صرفاً مفهومی یا روش‌شناختی (Pearl, 2000, pp. 20–23). از نظر پرل، مدل‌های ساختاری علّی بازنمایی‌هایی صوری از سازوکارهای واقعی جهان‌اند و ابزارهایی مانند حساب مداخله‌ای امکان استنتاج اثر مداخلات واقعی یا فرضی را فراهم می‌کنند (Pearl, 2000, pp. 68–75). در امتداد این دیدگاه، سنت کشف علیت که توسط اسپیرتس، گلایمور و شاینز توسعه یافته است مدعی است که، با اتکا به فروضی مشخص (مانند کفایت علّی و وفاداری آماری)، می‌توان ساختار علّی واقعی سیستم‌ها را از داده‌های مشاهده‌ای استخراج کرد (Spirtes et al., 2000,

pp. 3–10). از منظر فلسفه علم، این موضع به‌روشنی در اردوگاه واقع‌گرایی علمی قرار می‌گیرد، اینکه موفقیت تجربی مدل‌های علّی در پیش‌بینی نشانه‌ای از صدق تقریبی آن‌ها تلقی می‌شود.

بااین‌حال، این دیدگاه با چالشی جدی مواجه است: فروض لازم برای کشف علّیت غالباً قوی، غیرقابل آزمون مستقل و وابسته به پیش‌فرض‌های متافیزیکی‌اند. ازاین‌رو، برخی منتقدان استدلال می‌کنند که واقع‌گرایی علّی در یادگیری ماشین بیش‌ازحد خوش‌بینانه است.

۲. دیدگاه مداخله‌ای-عمل‌گرایانه: علّیت به‌مثابهٔ قابلیت کنترل

دیدگاه مداخله‌ای، که بیش از همه با آثار جیمز وودوارد شناخته می‌شود، تلاش می‌کند مفهوم علّیت را از تعهدات متافیزیکی سنگین رها کند و آن را براساسِ نقش عملی‌اش در علم بازتعریف نماید. در این چارچوب، یک رابطه زمانی علّی است که بتوان، ازطریقِ مداخله نظام‌مند بر یک متغیر، تغییرات پایداری در متغیر دیگر ایجاد کرد (Woodward, 2003, pp. 55–60). از منظر این دیدگاه، پرسش اصلی دربارهٔ مدل‌های علّی یادگیری ماشین این نیست که آیا آن‌ها «ذات واقعی» جهان را کشف می‌کنند یا نه بلکه این است که آیا قادرند به پرسش‌های مداخله‌ای از نوع «اگر X را تغییر دهیم، Y چه خواهد شد؟» پاسخ دهند. درنتیجه، یک مدل یادگیری ماشینی (به‌اختصار ML) می‌تواند علّی تلقی شود، حتی اگر صرفاً یک بازنمایی تقریبی و سطحی از سیستم باشد.

شایان توضیح است که «مداخله» در اینجا به‌معنای ایجاد تغییری فعال و کنترل‌شده در یک متغیر به‌گونه‌ای است که اثر آن تغییر بر سایر متغیرها قابل‌بررسی باشد. این مفهوم را باید در تقابل با مشاهدهٔ صرف فهمید: در مشاهده، پژوهشگر تنها الگوهای هم‌وقوعی را ثبت می‌کند، درحالی‌که در مداخله، ساختار سیستم به‌طورِ هدفمند دستخوش تغییر می‌شود. مهم است تأکید شود که مداخله در این چارچوب به کنش آگاهانه یک عامل انسانی فروکاسته نمی‌شود بلکه به هر نوع تغییر ساختاریِ قانون‌مند اطلاق می‌شود که بتواند روابط میان متغیرها را به‌صورتِ نظام‌مند دگرگون کند، خواه این تغییر به‌وسیلهٔ آزمایش، سیاست‌گذاری، شبیه‌سازی یا حتی تغییر محیط عملیاتی یک سامانه هوش مصنوعی صورت گیرد. از این منظر، اهمیت مداخله در آن است که امکان طرح پرسش‌های خلاف‌واقع را فراهم می‌کند، پرسش‌هایی ازاین‌دست که «اگر متغیر X به‌طورِ فعال تغییر داده شود، چه اثری بر Y خواهد داشت؟». توانایی پاسخ‌گویی معتبر به چنین پرسش‌هایی از شاخص‌های اصلی تمایز مدل‌های علّی از مدل‌های صرفاً آماری است.

این رویکرد در فلسفه علم نوعی عمل‌گرایی روش‌شناختی را نمایندگی می‌کند. علّیت در اینجا مفهومی است که ارزش آن با توانایی‌اش در هدایت مداخله، آزمایش و سیاست‌گذاری سنجیده می‌شود نه با صدق متافیزیکی‌اش. چنین برداشتی با کاربردهای عملی هوش مصنوعی -به‌ویژه در پزشکی، اقتصاد و علوم اجتماعی- سازگاری بالایی دارد.

بااین‌حال، از این دیدگاه به‌دلیلِ ابهام هستی‌شناختی مورد انتقاد شده است: مشخص نیست که آیا علّیت در این چارچوب صرفاً قراردادی عملی است یا همچنان به نوعی واقعیت مستقل اشاره دارد.

۳. دیدگاه ابزارگرایانه تقویت‌شده: علّیت به‌مثابهٔ ابزار پیش‌بینی و تصمیم‌گیری

در دیدگاه ابزارگرایانه تقویت‌شده، علّیت در یادگیری ماشین نه بازنمایی ساختار واقعی جهان بلکه ابزاری مفید برای بهبود عملکرد سیستم‌ها تلقی می‌شود. این موضع را می‌توان امتدادی از «تجربه‌گرایی برساختی» ون‌فراسن دانست. ون‌فراسن، برخلافِ ابزارگرایی کلاسیک که نظریه‌های علمی را صرفاً ابزارهایی محاسباتی برای پیش‌بینی تلقی می‌کرد، معتقد است پذیرش یک نظریه علمی مستلزم باور به صدق کامل آن نیست بلکه تنها مستلزم پذیرش «کفایت تجربی» آن است (van Fraassen, 1980, pp. 12–16). براین‌اساس، یک مدل علمی تا آنجا موجه است که بتواند پدیده‌های مشاهده‌پذیر را به‌نحوِ موفقی توصیف و پیش‌بینی کند، بی‌آنکه لازم باشد ساختارهای نظری آن بازنمایی کاملی از واقعیت نامشهود جهان باشند. این تلقی برای تحلیل مدل‌های علّی در یادگیری ماشین اهمیت ویژه‌ای دارد. بسیاری از مدل‌های معاصر یادگیری عمیق، حتی درصورتِ موفقیت چشمگیر در پیش‌بینی، لزوماً به‌مثابهٔ بازنمایی‌های کامل سازوکارهای واقعی جهان فهم نمی‌شوند. از این منظر، می‌توان موفقیت معرفتی مدل‌های علّی را نه براساسِ صدق متافیزیکی کامل آن‌ها بلکه براساسِ توانایی آن‌ها در حفظ روابط پایدار، پیش‌بینی موفق و پشتیبانی از مداخلات ارزیابی کرد.

از این منظر، افزودن مؤلفه‌های علّی به مدل‌های یادگیری ماشین صرفاً باعث افزایش توان پیش‌بینی، تعمیم‌پذیری و تبیین‌پذیری آن‌ها می‌شود. اینکه این مدل‌ها واقعاً به سازوکارهای علّی جهان اشاره کنند یا نه، پرسشی زائد یا حتی بی‌معنا تلقی می‌شود. آنچه اهمیت دارد، موفقیت عملی در تصمیم‌گیری و مداخله است. بسیاری از رویکردهای معاصر در حوزه هوش مصنوعی علّی، به‌ویژه در متون سیاست‌گذاری و صنعت، تلویحاً چنین برداشتی را مفروض می‌گیرند. در این متون، علّیت ابزاری برای کاهش ریسک، بهبود تصمیم‌سازی و افزایش اعتمادپذیری سیستم‌های هوش مصنوعی است نه موضوعی برای تعهدات هستی‌شناختی (Kesh & Whitworth, 2025, pp. 7–11).

نقد اصلی وارد بر این دیدگاه آن است که توضیح نمی‌دهد که چرا مدل‌های علّی در برخی حوزه‌ها به‌طورِ پایدار موفق‌اند و آیا این موفقیت صرفاً تصادفی یا وابسته به شرایط خاص نیست. براین‌اساس، نقش علّیت در یادگیری ماشین صرفاً به کفایت پیش‌بینانه فروکاستنی نیست. ساختارهای علّی تا آنجا اهمیت دارند که بتوانند تعمیم فراتوزیعی، استدلال خلاف‌واقع و مداخله را ممکن سازند، اموری که فراتر از پیش‌بینی صرف قرار می‌گیرند.

۴. دیدگاه انتقادی–ساختاری: علّیت به‌مثابهٔ بازنمایی ساختاری محدود

دیدگاه انتقادی–ساختاری، که می‌توان آن را در خوانش فلسفی آثار شُلکوپف و نیز در سنت واقع‌گرایی ساختاری یافت، تلاش می‌کند راه سومی میان واقع‌گرایی سخت و ابزارگرایی محض بگشاید. براساسِ این رویکرد، یادگیری ماشین بدون توجه به علّیت از نظر معرفتی شکننده است اما، درعینِ‌حال، نباید مدل‌های علّی را با سازوکارهای کامل و نهایی جهان یکی دانست. شُلکوپف استدلال می‌کند که بسیاری از ناکامی‌های یادگیری ماشین –به‌ویژه در تعمیم خارج از توزیع– ناشی از نادیده‌گرفتنِ ساختارهای علّی داده‌هاست اما هم‌زمان تأکید می‌کند که این ساختارها وابسته به سطح توصیف و

مدل‌سازی‌اند (Schölkopf, 2019, pp. 6–8). درنتیجه، مدل‌های علّی ML بازنمایی‌هایی ساختاری و حداقلی از جهان ارائه می‌دهند نه توصیف‌هایی کامل از ذات اشیا.

این موضع با واقع‌گرایی ساختاری همخوانی دارد، دیدگاهی که مدعی است آنچه علم به‌درستی بازنمایی می‌کند نه ماهیت اشیا بلکه ساختار روابط میان آن‌هاست (Ladyman et al., 2007, pp. 130–135). در این چارچوب، علّیت در یادگیری ماشین واقعی است اما واقعیتی ساختاری، وابسته به مدل و محدود به سطح خاصی از توصیف. توضیح آنکه، یکی از رویکردهای مهم در فلسفه علم معاصر، که می‌تواند برای تحلیل مدل‌های علّی در یادگیری ماشین الهام‌بخش باشد، «واقع‌گرایی ساختاری» است. این دیدگاه، که در واکنش به چالش‌های واردشده بر واقع‌گرایی علمی کلاسیک شکل گرفت، استدلال می‌کند که آنچه نظریه‌های علمی با موفقیت حفظ و بازنمایی می‌کنند نه لزوماً ذات یا ماهیت درونی اشیا بلکه ساختار روابط میان آن‌هاست. براین‌اساس، موفقیت نظریه‌های علمی را نباید به‌معنای دسترسی کامل به ماهیت نهایی جهان تلقی کرد بلکه می‌توان آن را ناشی از توانایی نظریه‌ها در ضبط ساختارهای پایدار و روابط منظم میان پدیده‌ها دانست. درنتیجه، تعهد هستی‌شناختی این رویکرد نه متوجه ذات اشیا بلکه معطوف به روابط و ساختارهایی است که در نظریه‌ها حفظ می‌شوند.

این تلقی برای تحلیل مدل‌های علّی در یادگیری ماشین اهمیت ویژه‌ای دارد. هنگامی که مدل‌های یادگیری ماشین موفق می‌شوند الگوهای مداخله‌پذیر، وابستگی‌های پایدار یا روابط ناوردا را استخراج کنند، لزوماً نباید نتیجه گرفت که آن‌ها به ذات نهایی سازوکارهای جهان دست یافته‌اند. درعوض، می‌توان گفت این مدل‌ها نوعی بازنمایی ساختاری از روابط علّی ارائه می‌کنند، بازنمایی‌ای که اعتبار آن به توانایی‌اش در حفظ ساختارهای پایدار تحت تغییر و مداخله وابسته است.

نقد و مقایسه دیدگاه‌ها دربارهٔ علّیت در یادگیری ماشین

دیدگاه‌های مطرح‌شده دربارهٔ علّیت در یادگیری ماشین را می‌توان پاسخ‌هایی متفاوت به یک پرسش بنیادین دانست، اینکه مدل‌های علّی یادگیری ماشین چه نوع دانشی دربارهٔ جهان فراهم می‌کنند. نقد و مقایسه این دیدگاه‌ها نشان می‌دهد که اختلاف آن‌ها نه صرفاً در سطح فنی بلکه در سطح تعهدات فلسفه علم –به‌ویژه دربابِ واقع‌گرایی، تبیین و نقش مدل‌های علمی– ریشه دارد.

نخست، دیدگاه واقع‌گرایانه علّی ازحیثِ انسجام نظری و پیوند با سنت کلاسیک فلسفه علم نقطه قوت قابل‌توجهی دارد. این دیدگاه می‌تواند به‌خوبی توضیح دهد که چرا مدل‌های علّی در برخی حوزه‌ها –مانند پزشکی یا اپیدمیولوژی– نسبت به مدل‌های صرفاً آماری برتری دارند، زیرا آن‌ها مدعی بازنمایی سازوکارهای واقعی جهان‌اند (Pearl, 2000, pp. 20–23). بااین‌حال، نقد اصلی وارد بر این موضع آن است که موفقیت عملی مدل‌های علّی در یادگیری ماشین لزوماً به معنای صدق هستی‌شناختی آن‌ها نیست. فروضی مانند کفایت علّی یا وفاداری آماری، که برای کشف علّیت ضروری‌اند،

به‌ندرت به‌طورِ مستقل قابل آزمون‌اند و همین امر باعث می‌شود که گذار از «موفقیت پیش‌بینانه» به «صدق علّی» از منظر فلسفه علم محل تردید باشد. درنتیجه، این دیدگاه در معرض اتهام نوعی واقع‌گرایی خوش‌بینانه قرار می‌گیرد.

درمقابل، دیدگاه مداخله‌ای–عمل‌گرایانه تلاش می‌کند از این مشکل بگریزد و مفهوم علّیت را براساسِ نقش آن در مداخله و کنترل تعریف کند. مزیت اصلی این رویکرد آن است که بدون تعهد به سازوکارهای متافیزیکی پنهان، می‌تواند استفاده علّی از مدل‌های یادگیری ماشین را توجیه کند (Woodward, 2003, pp. 55–60). بااین‌حال، این دیدگاه با چالشی مفهومی مواجه است و آن اینکه اگر علّیت صرفاً با قابلیت مداخله تعریف شود، آنگاه تمایز میان مدل‌های علّی و مدل‌های بسیار پیشرفته پیش‌بینی‌گر تاحدی تضعیف می‌شود. افزون بر این، منتقدان استدلال می‌کنند که عمل‌گرایی مداخله‌ای به‌طورِ کامل توضیح نمی‌دهد که چرا برخی مداخلات موفق‌اند و برخی دیگر نه، مگر آنکه به نوعی ساختار علّی نسبتاً پایدار در جهان متعهد شویم.

دیدگاه ابزارگرایانه تقویت‌شده، با تأکید بر کارآمدی عملی مدل‌های علّی، از تعهدات هستی‌شناختی سنگین فاصله می‌گیرد و ازاین‌حیث موضعی محافظه‌کارانه‌تر به نظر می‌رسد. این دیدگاه می‌تواند توضیح دهد که چرا، در بسیاری از کاربردهای صنعتی و سیاست‌گذاری، پرسش از «واقعی بودن» علّیت اهمیتی ندارد و آنچه اهمیت دارد بهبود تصمیم‌گیری است (van Fraassen, 1980, pp. 12–16; Kesh & Whitworth, 2025, pp. 7–11). بااین‌حال، ابزارگرایی تقویت‌شده با یک مشکل کلاسیک روبه‌روست: اگر مدل‌های علّی صرفاً ابزارند، توضیح اینکه چرا این ابزارها به‌طورِ پایدار موفق‌اند دشوار می‌شود. به بیان دیگر، این دیدگاه در تبیین چرایی موفقیت Causal AI دچار کم‌عمقی تبیینی است.

در این میان، دیدگاه انتقادی–ساختاری کوششی آگاهانه برای حفظ نقاط قوت سه دیدگاه پیشین و پرهیز از ضعف‌های آن‌هاست. این رویکرد می‌پذیرد که یادگیری ماشین، بدون علّیت، از نظر معرفتی ناکافی است اما هم‌زمان تأکید می‌کند که مدل‌های علّی بازنمایی‌های کامل یا نهایی از جهان نیستند (Schölkopf, 2019, pp. 6–8). برتری این دیدگاه در آن است که می‌تواند موفقیت مدل‌های علّی را بدون تعهد به واقع‌گرایی متافیزیکی یا ابزارگرایی افراطی توضیح دهد. براین‌اساس، مدل‌های علّی ساختارهای پایدار روابط را در سطحی معین بازنمایی می‌کنند، بی‌آنکه مدعی کشف ذات اشیا باشند (Ladyman et al., 2007, pp. 130–135).

باوجودِاین، دیدگاه ساختاری نیز بی‌چالش نیست. ابهام در تعیین مرز میان «ساختار واقعی» و «ساختار وابسته به مدل» می‌تواند این رویکرد را در معرض نقد نسبی‌گرایی قرار دهد. اگر ساختارها همواره وابسته به سطح توصیف باشند، پرسش از اینکه کدام ساختارها معرفتاً ممتازند همچنان بی‌پاسخ می‌ماند.

تحلیل و بررسی

دیدگاه پیشنهادی این مقاله بر این ایده استوار است که علّیت در یادگیری ماشین نه بازنمایی مستقیم سازوکارهای بنیادین جهان است و نه صرفاً ابزاری پیش‌بینانه بلکه نوعی ساختار تبیینی است که اعتبار خود را از امکان مداخله، تعمیم و انسجام میان سطوح توصیف به دست می‌آورد.

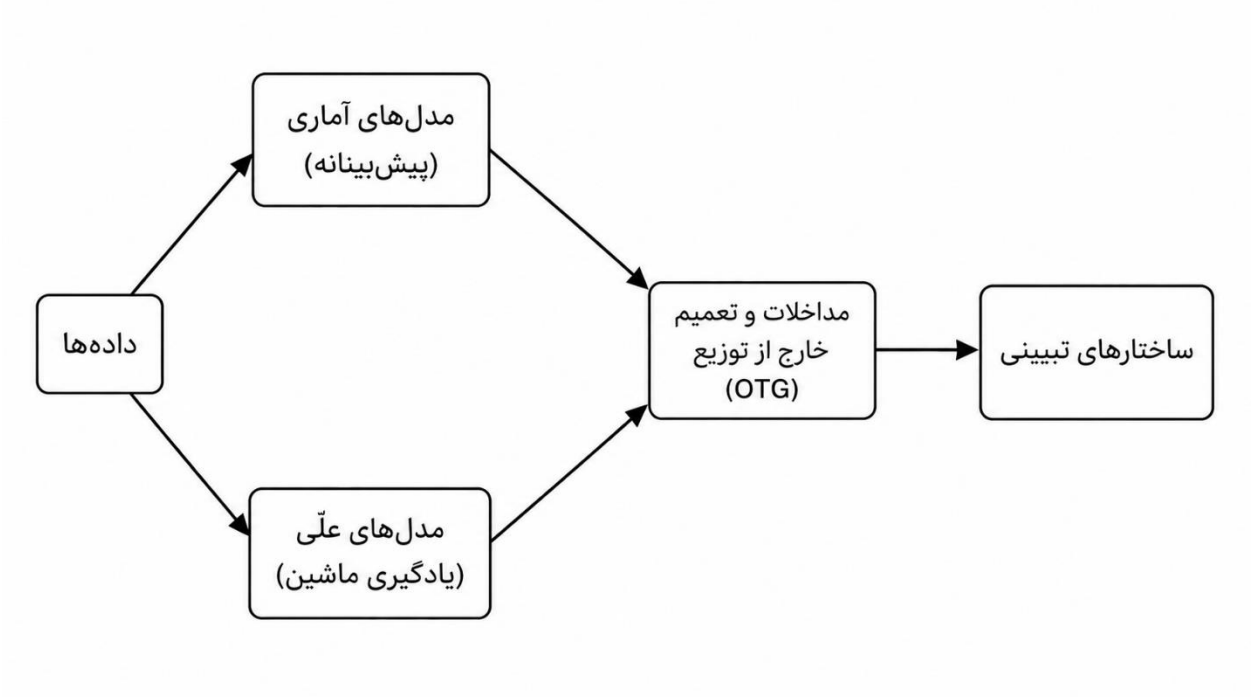
برخلافِ واقع‌گرایی علّی قوی، مدل‌های علّی یادگیری ماشین نباید به‌عنوانِ کشف سازوکارهای متافیزیکی مستقل از چارچوب مدل‌سازی فهمیده شوند. محدودیت‌های داده، فروض پیشینی، مانند کفایت علّی، و وابستگی شدید نتایج به انتخاب متغیرها نشان می‌دهد که نسبت‌دادنِ صدق هستی‌شناختی قوی به این مدل‌ها از نظر فلسفه علم چندان موجه نیست. درعینِ‌حال، دیدگاه مقاله از ابزارگرایی کلاسیک نیز فاصله می‌گیرد. موفقیت مدل‌های علّی صرفاً در پیش‌بینی خلاصه نمی‌شود بلکه در پایداری تحت مداخله، قابلیت انتقال بین محیط‌ها و تولید تبیین‌هایی است که امکان تصمیم‌گیری عقلانی را فراهم می‌کنند. این ویژگی‌ها نشان می‌دهد که علّیت در یادگیری ماشین چیزی «بیش از یک ترفند محاسباتی» است.

نقطه محوری دیدگاه مقاله این است که آنچه مدل‌های علّی بازنمایی می‌کنند «ذات» اشیا یا سازوکارهای نهایی جهان نیست بلکه ساختارهای پایدار روابط میان متغیرها در یک سطح توصیف مشخص است. این ایده مستقیماً با واقع‌گرایی ساختاری در فلسفه علم همخوان است اما با یک قید مهم و آن این است که ساختارها ازطریقِ مداخله تعریف و اعتبارسنجی می‌شوند نه صرفاً ازطریقِ برازش آماری. به بیان دیگر، اگر یک مدل علّی، تحت مداخله، رفتار پیش‌بینی‌پذیر داشته باشد، در محیط‌های جدید فرو نپاشد، و امکان پاسخ به پرسش‌های «چه می‌شود اگر...؟» را فراهم کند، آنگاه می‌توان گفت این مدل به یک ساختار علّی معتبر در آن سطح تحلیل دست یافته است –حتی اگر این ساختار نهایی یا بنیادین نباشد. تبیین در این چارچوب نه به‌معنای آشکارکردنِ علل نهایی بلکه به‌معنای قراردادنِ پدیده در شبکه‌ای از وابستگی‌های علّی فهمیده می‌شود که امکان مداخله و کنترل عقلانی را فراهم می‌کند. از این منظر، یک مدل علّی یادگیری ماشین زمانی تبیین‌گر است که نشان دهد چرا تغییر در یک متغیر تغییری معین در متغیر دیگر ایجاد می‌کند، حتی اگر این «چرا» به سازوکارهای میکروسکوپیک یا قوانین بنیادی فروکاسته نشود.

براین‌اساس، علّیت در یادگیری ماشین یک مفهوم میان‌سطحی معرفی می‌شود، نه در سطح قوانین بنیادی فیزیک و نه در سطح صرفاً همبستگی‌های آماری بلکه در سطحی که برای کنش عقلانی، تصمیم‌گیری و سیاست‌گذاری معنا دارد. این جایگاه میان‌سطحی توضیح می‌دهد که چرا علّیت هم از نظر فلسفی مناقشه‌برانگیز است و هم از نظر عملی گریزناپذیر. بنابراین، علّیت در یادگیری ماشین نه کشف حقیقت نهایی جهان است و نه صرفاً یک ابزار محاسباتی بلکه بازنمایی ساختارهای پایدارِ وابسته به مداخله است که امکان تبیین، تعمیم و کنش عقلانی را در بستر علم داده‌محور فراهم می‌کند. این موضع می‌تواند هم به بحث‌های فلسفه علم دربارهٔ واقع‌گرایی و تبیین جهت بدهد، و هم برای

پژوهشگران هوش مصنوعی چارچوب مفهومی روشن‌تری برای فهم آنچه واقعاً با هوش مصنوعی علّی انجام می‌دهند فراهم کند.

جهت روش‌تر شدن ادعای مقاله، شکل زیر را در نظر بگیرید.



شکل۱. چارچوب مفهومی فهم علّیت در یادگیری ماشین[۱]

فرآیندی که در شکل فوق نمایش داده شده است تلاشی است برای صورت‌بندی مفهومی جایگاه علّیت در یادگیری ماشین، به‌نحوی‌که هم با ملاحظات فلسفه علم سازگار باشد و هم با واقعیت‌های پژوهش‌های معاصر هوش مصنوعی. این فرآیند از داده آغاز می‌شود و به ساختارهای تبیینی وابسته به مداخله ختم می‌گردد. اما مسیر رسیدن به این نقطه، بسته به نوع مدل‌سازی، به‌طورِ معناداری متفاوت است.

در ابتدایی‌ترین سطح، همه رویکردهای یادگیری ماشین –اعم از آماری یا علّی– از داده‌های تجربی آغاز می‌کنند. این داده‌ها معمولاً حاصل مشاهدهٔ هم‌زمان مجموعه‌ای از متغیرها در شرایط خاص‌اند و، به‌خودیِ‌خود، فاقد هرگونه تفسیر علّی‌اند. داده، در این معنا، صرفاً ثبت الگوهای هم‌وقوعی است نه بیان رابطه علّی. تأکید نمودار بر «داده» به‌عنوانِ نقطه آغاز مشترک نشان می‌دهد که اختلاف میان رویکردهای مختلف نه در سطح داده بلکه در سطح صورت‌بندی مدل و نوع پرسش‌هایی است که از داده پرسیده می‌شود.

[۱] ترسیم شکل با کمک هوش مصنوعی انجام گرفته است.

یکی از مسیرهای اصلی که از داده منشعب می‌شود مسیر مدل‌های آماری یا استاتیک است. این مدل‌ها -که بخش عمده‌ای از یادگیری ماشین کلاسیک را تشکیل می‌دهند- با هدف کشف الگوهای همبستگی و بهینه‌سازی پیش‌بینی توسعه می‌یابند. این مدل‌ها را می‌توان «استاتیک» نامید، زیرا روابط میان متغیرها را به‌صورتِ ثابت و وابسته به شرایط مشاهده‌شده مدل می‌کنند. درنتیجه، هرچند ممکن است که در پیش‌بینی بسیار موفق باشند اما به‌طورِ ساختاری قادر نیستند که به پرسش‌هایی از جنس «اگر X را تغییر دهیم، چه بر سر Y می‌آید؟» پاسخ دهند.

مسیر دوم نمودار به مدل‌های علّی یادگیری ماشین اختصاص دارد. این مدل‌ها، افزون بر داده، مفروضاتی را وارد مدل می‌کنند که امکان استدلال دربارهٔ جهت‌مندی روابط، مداخله و ناوردایی را فراهم می‌سازد. گراف‌های علّی، مدل‌های ساختاری و اصل ناوردایی نمونه‌هایی از این تلاش‌اند. تفاوت بنیادین این مدل‌ها با مدل‌های آماری در این است که آن‌ها صرفاً به همبستگی بسنده نمی‌کنند بلکه می‌کوشند روابطی را صورت‌بندی کنند که دربرابرِ تغییرات فعالانه (مداخله) پایدار باقی بمانند. بااین‌حال، مقاله تأکید می‌کند که این امر نباید به‌طورِ شتاب‌زده به‌عنوانِ کشف سازوکارهای متافیزیکیِ نهایی تفسیر شود.

شایستهٔ توضیح است که «ناوردایی»[1] در اینجا به پایداری یک رابطه یا ساختار دربرابرِ تغییرات مشخص اشاره دارد. در زمینه یادگیری ماشین و علّیت، ناوردایی به این معناست که برخی روابط میان متغیرها -حتی درصورتِ تغییر شرایط، محیط یا توزیع داده‌ها- همچنان حفظ می‌شوند. در این مقاله، ناوردایی به‌عنوانِ معیاری معرفت‌شناختی برای تشخیص روابط علّی از همبستگی‌های صرف به کار می‌رود. همبستگی‌های آماری اغلب وابسته به شرایط خاص مشاهده‌اند و با تغییر محیط یا داده‌ها فرو می‌ریزند. درمقابل، به‌طورِ نظری انتظار می‌رود که روابط علّی، تحت دامنه‌ای از مداخلات و تغییرات، پایدار باقی بمانند. بنابراین، ناوردایی نه یک ویژگی مطلق بلکه مفهومی نسبی و وابسته به دامنه تغییرات است. یک رابطه زمانی ناوردا تلقی می‌شود که دربرابرِ مجموعه‌ای معین از مداخلات یا تغییرات معنادار، رفتار خود را حفظ کند. این برداشت از ناوردایی به مدل اجازه می‌دهد که، بدون ادعای کشف قوانین بنیادین طبیعت، روابطی پایدار و تبیین‌پذیر را شناسایی کند.

نقطه مرکزی نمودار جایی است که تمایز میان مدل‌ها از سطح فنی به سطح معرفتی منتقل می‌شود. در این مرحله، معیار اصلی ارزیابی مدل‌ها رفتار آن‌ها تحت مداخله و توانایی‌شان در تعمیم به محیط‌های جدید است. مدل‌های آماری معمولاً در این نقطه دچار شکست می‌شوند، زیرا ساختار آن‌ها وابسته به توزیع داده آموزش است. درمقابل، مدل‌های علّی -درصورتی‌که به‌درستی صورت‌بندی شده باشند- می‌توانند الگوهایی را حفظ کنند که در شرایط جدید نیز معتبر باقی می‌مانند. از منظر مقاله، دقیقاً در این نقطه است که مفهوم علّیت معنا پیدا می‌کند، نه به‌عنوانِ برچسبی متافیزیکی بلکه به‌عنوانِ معیاری برای پایداری و قابلیت مداخله.

[1] Invariance

درنهایت، فرآیند نمودار به «ساختارهای تبیینی» ختم می‌شود. این ساختارها نه قوانین بنیادین جهان‌اند و نه صرفاً ابزارهای محاسباتی بلکه بازنمایی‌هایی میان‌سطحی از روابط پایدارند که امکان تبیین پدیده‌ها را فراهم می‌کنند، برای کنش عقلانی و تصمیم‌گیری قابلِ‌اتکا هستند، و به سطح توصیف و دامنه کاربرد وابسته‌اند. تمایز رویکرد مقاله دقیقاً در همین‌جا آشکار می‌شود. برخلافِ رویکردهای واقع‌گرایانه قوی، مقاله ادعا نمی‌کند که این ساختارها حقیقت نهایی جهان را بازمی‌نمایانند و، برخلافِ ابزارگرایی صرف، آن‌ها را به پیش‌بینی تقلیل نمی‌دهد. درعوض، علّیت در یادگیری ماشین مفهومی تبیینی، میان‌سطحی و وابسته به مداخله فهم می‌شود.

به‌طورِ خلاصه، رویکرد مقاله از مدل‌های آماری صرف، به‌دلیلِ ناتوانی در مداخله و تعمیم، فاصله می‌گیرد؛ با واقع‌گرایی علّی سخت، به‌دلیلِ تعهدات متافیزیکی فراتر از شواهد، همراه نمی‌شود؛ و علّیت را ساختاری تبیینی تعریف می‌کند که اعتبار خود را از موفقیت در مداخله و پایداری ساختاری به دست می‌آورد. این چارچوب، همان‌گونه که در نمودار نشان داده شده است، جایگاه علّیت در یادگیری ماشین را به‌نحوی روشن می‌سازد که هم از نظر فلسفی موجه است و هم برای پژوهش‌های معاصر هوش مصنوعی راهگشاست.

ارزیابی انتقادی مدل پیشنهادی

مدل پیشنهادی این مقاله، که علّیت در یادگیری ماشین را به‌مثابهٔ «ساختار تبیینیِ وابسته به مداخله» صورت‌بندی می‌کند، در عین برخورداری از مزایای مفهومی، با مجموعه‌ای از نقدهای بالقوه مواجه است که بررسی آن‌ها برای تقویت موضع نظری ضروری است. در این بخش، مهم‌ترین این نقدها مطرح و پاسخ داده می‌شوند.

یکی از نخستین نقدها آن است که این مدل، با فاصله‌گیری هم‌زمان از واقع‌گرایی علّیِ قوی و ابزارگرایی کلاسیک، ممکن است دچار نوعی ابهام مفهومی شود. اگر علّیت نه بازنمایی واقعیت بنیادین جهان باشد و نه صرفاً ابزاری پیش‌بینانه، این پرسش مطرح می‌شود که دقیقاً چه نوع مقوله‌ای است. پاسخ مقاله به این نقد آن است که مدل پیشنهادی تعهدی هستی‌شناختی ارائه نمی‌دهد بلکه ادعایی معرفت‌شناختی دربارهٔ جایگاه علّیت در علم‌های داده‌محور مطرح می‌کند. علّیت در این چارچوب به‌منزلهٔ شیوه‌ای از سازمان‌دهی معرفت تجربی در سطحی معین تعریف می‌شود که امکان تبیین و مداخله عقلانی را فراهم می‌سازد. بدین‌ترتیب، مدل نه به واقع‌گرایی متافیزیکی متعهد می‌شود و نه به ابزارگرایی تقلیل‌گرایانه.

نقد دوم متوجه نقش محوری مفهوم «مداخله» در این مدل است. ممکن است چنین استدلال شود که وابسته‌کردن علّیت به مداخله آن را به توانایی‌ها یا کنش‌های عامل انسانی فرو می‌کاهد و، بدین‌ترتیب، علّیت را به مفهومی انسان‌محور و عمل‌گرایانه و افراطی بدل می‌کند. در پاسخ، مقاله تصریح می‌کند که مداخله در این چارچوب نه به‌معنای کنش فیزیکی یا قصد آگاهانه عامل انسانی بلکه به‌منزلهٔ تغییر ساختاریِ قانون‌مند در یک سیستم تعریف می‌شود. مداخله مفهومی معرفت‌شناختی است که به امکان ارزیابی پایداری روابط تحت تغییرات کنترل‌شده اشاره دارد نه مفهومی روان‌شناختی

یا انسان‌شناختی. این تفسیر، مفهوم علّیت را از انسان‌محوری رها می‌سازد و آن را به ویژگی‌های ساختاری سیستم‌ها پیوند می‌زند.

نقد مهم دیگر ناظر به وابستگی ساختارهای علّی به سطح توصیف است. ممکن است استدلال شود که اگر علّیت همواره به سطح تحلیل وابسته باشد، آنگاه تمایزی میان مدل‌های معتبر و نامعتبر باقی نمی‌ماند و مدل به نسبی‌گرایی معرفتی می‌انجامد. پاسخ مقاله در اینجا بر معرفی یک معیار غیرنسبی استوار است. هرچند علّیت به سطح توصیف وابسته است اما همه سطوح از اعتبار یکسان برخوردار نیستند. سطحی معرفتاً موجه است که بیشترین ناوردایی تحت مداخله و بالاترین توان تعمیم فراتوزیعی را نشان دهد. بدین‌ترتیب، وابستگی به سطح به‌معنای دل‌بخواهی یا نسبی‌بودن نیست بلکه به انتخاب سطحی اشاره دارد که از نظر ساختاری پایدارتر و از نظر تبیینی پربارتر است.

نقد دیگر آن است که درصورتِ موفقیت برخی مدل‌های آماری پیچیده در تعمیم فراتوزیعی، تمایز میان آن‌ها و مدل‌های علّی تضعیف می‌شود. اگر معیار اصلی موفقیت تعمیم باشد، آیا هر مدل تعمیم‌پذیر را می‌توان علّی دانست؟ مدل پیشنهادی این همسان‌انگاری را رد می‌کند. تعمیم شرط لازم علّیت است اما شرط کافی آن نیست. مدل علّی، افزون بر تعمیم، باید قادر به تثبیت جهت‌مندی روابط و پیش‌بینی پیامدهای مداخلاتی باشد که با الگوهای مشاهده‌شده در داده آموزشی سازگار نیستند. این توانایی پاسخ‌گویی به پرسش‌های خلاف‌واقع و مداخله‌ای خط تمایز روشنی میان مدل‌های علّی و پیش‌بینی‌گرهای صرف -حتی بسیار موفق- ایجاد می‌کند.

شایان توضیح است که «تعمیم فراتوزیعی»[1] در اینجا اشاره دارد به توانایی یک مدل برای حفظ عملکرد معتبر در شرایطی که از نظر آماری با داده‌های آموزشی آن متفاوت‌اند. برخلافِ تعمیم کلاسیک -که معمولاً به عملکرد مدل روی داده‌های جدید اما هم‌توزیع با داده آموزش اشاره دارد- تعمیم فراتوزیعی مستلزم رویارویی با محیط‌ها، زمینه‌ها یا شرایطی است که توزیع داده در آن‌ها تغییر کرده است. اهمیت این مفهوم در مقاله از آنجا ناشی می‌شود که بسیاری از شکست‌های عملی مدل‌های یادگیری ماشین دقیقاً در چنین شرایطی رخ می‌دهند. مدل‌هایی که به همبستگی‌های سطحی وابسته‌اند، اغلب در مواجهه با تغییر توزیع داده ناکارآمد می‌شوند. درمقابل، مدل‌هایی که بر ساختارهای علّی ناوردا تکیه دارند، شانس بیشتری برای حفظ عملکرد خود در محیط‌های جدید دارند. از منظر مقاله، تعمیم فراتوزیعی نه صرفاً یک معیار مهندسی بلکه نشانه‌ای معرفت‌شناختی از موفقیت مدل در بازنمایی ساختارهای پایدار روابط است. به همین دلیل، توانایی تعمیم فراتوزیعی از ملاک‌های کلیدی اعتبار علّی مدل‌ها در نظر گرفته می‌شود.

درنهایت، ممکن است این پرسش مطرح شود که آیا این مدل صرفاً توجیهی فلسفی برای رویه‌های موجود در پژوهش‌های هوش مصنوعی است یا اینکه سهم مستقلی در فلسفه علم دارد. پاسخ مقاله این است که این چارچوب بازتعریفی از جایگاه علّیت در علم‌های معاصر ارائه می‌دهد. علّیت نه به‌عنوانِ ویژگی بنیادین طبیعت و نه به‌عنوانِ ابزار

[1] Out-of-Distribution Generalization

محاسباتی بلکه به‌مثابهٔ دستاوردی معرفتی در سطحی میان‌سطحی فهم می‌شود که برای تبیین، مداخله و کنش عقلانی در علوم داده‌محور ضروری است.

بررسی انتقادی نشان می‌دهد که مدل پیشنهادی، با تصریح دقیق تعهدات معرفت‌شناختی خود، می‌تواند از خطر ابهام، انسان‌محوری و نسبی‌گرایی اجتناب کند. این مدل علّیت را به‌عنوانِ ساختاری تبیینی، سطح‌مند و مداخله‌محور معرفی می‌کند که، بدون اتکا به تعهدات متافیزیکیِ سنگین، نقشی اساسی در فهم و توسعه یادگیری ماشین ایفا می‌کند.

در اینجا لازم است که به چند پرسش اصلی درخصوصِ دیدگاه مطرح‌شده در مقاله بپردازیم. نخستین اشکالی که به ذهن می‌رسد این است که ادعای مقاله دربارهٔ «مدل‌های علّی در یادگیری ماشین» را می‌توان دربارهٔ مدل‌های علمی کلاسیک نیز مطرح کرد. درنتیجه، مقاله هنوز نشان نداده است که دقیقاً چه چیزی دربارهٔ یادگیری ماشین یا هوش مصنوعی «ویژه» است. در پاسخ باید گفت که مسئله علّیت در یادگیری ماشین و به‌ویژه در معماری‌های جدید یادگیری عمیق صرفاً بازتکرار مسئله کلاسیک علّیت در مدل‌های علمی نیست. در مدل‌سازی علمی کلاسیک، ساختار مدل معمولاً براساسِ نظریه‌های پیشینی، متغیرهای ازپیش‌تعریف‌شده و سازوکارهای نسبتاً شفاف توسط پژوهشگر طراحی می‌شود. در چنین مدل‌هایی، نسبت میان ساختار نظری، متغیرهای مدل و سازوکارهای علّی غالباً به‌طورِ مستقیم تحت کنترل معرفتی پژوهشگر قرار دارد.

درمقابل، در بسیاری از مدل‌های یادگیری ماشین، ساختارهای بازنمایی و روابط میان متغیرها به‌صورتِ داده‌محور و ازطریقِ فرایندهای بهینه‌سازی آماری شکل می‌گیرند. در این مدل‌ها، بازنمایی‌های میانی نه لزوماً براساسِ مفاهیم نظری ازپیش‌تعریف‌شده بلکه در قالب ساختارهای توزیعی و لایه‌ای تولید می‌شوند که بخش مهمی از آن‌ها مستقیماً برای انسان شفاف یا تفسیرپذیر نیستند. این ویژگی در معماری‌های جدید یادگیری عمیق تشدید می‌شود، جایی که وابستگی‌های معنایی و ساختاری رمزگذاری می‌شوند و مدل می‌تواند الگوهایی را استخراج کند که مستقیماً توسط پژوهشگر طراحی نشده‌اند. درنتیجه، مسئله فلسفی علّیت در این زمینه دیگر صرفاً این نیست که آیا مدل‌های علمی می‌توانند روابط علّی جهان را بازنمایی کنند بلکه این پرسش نیز مطرح می‌شود که چگونه ساختارهای علّی می‌توانند در دل سیستم‌هایی پدیدار شوند که فرایند تولید بازنمایی در آن‌ها تا حد زیادی خودکار، داده‌محور و غیرشفاف است. به همین دلیل، مسئله علّیت در عصر یادگیری ماشین واجد ابعادی تازه دربابِ بازنمایی، تبیین و کفایت معرفت‌شناختی مدل‌هاست.

پرسش دوم آن است که اساساً یادگیری عمیق چه چیز ویژه‌ای به مدل‌سازی اضافه می‌کند یا اساساً مکانیسم‌های هوش مصنوعی، مثل مبدل‌ها، چگونه ماهیت مدل‌های مبتنی بر یادگیری ماشین/یادگیری عمیق را تغییر می‌دهند. پاسخ آن است که ویژگی متمایز مدل‌های یادگیری عمیق صرفاً در افزایش مقیاس محاسبات یا حجم داده‌ها خلاصه نمی‌شود بلکه به نحوه تولید و سازمان‌دهی بازنمایی‌ها در این مدل‌ها مربوط است. برخلافِ بسیاری از مدل‌های کلاسیک یادگیری ماشین که بر ویژگی‌های مهندسی‌شده و متغیرهای ازپیش‌تعریف‌شده تکیه دارند، شبکه‌های عمیق ساختارهای بازنمایی

را به‌صورتِ سلسله‌مراتبی و درونی از دل داده‌ها استخراج می‌کنند. در این مدل‌ها، هر لایه می‌تواند الگوهایی انتزاعی‌تر و پایدارتر از سطح پیشین ایجاد کند و، بدین‌ترتیب، نوعی سازمان بازنمایانه چندسطحی شکل گیرد.

این وضعیت در معماری‌های مبتنی بر مبدل اهمیتی مضاعف پیدا می‌کند. در این معماری‌ها، وابستگی‌های میان عناصر داده نه ازطریقِ قواعد نمادین صریح بلکه ازطریقِ سازوکار بازنمایی‌های توزیعی رمزگذاری می‌شوند. درنتیجه، مدل قادر است ساختارهایی را استخراج کند که مستقیماً توسط پژوهشگر تعیین نشده‌اند و حتی در بسیاری از موارد به‌سادگی قابلِ‌تفسیر انسانی نیستند. همین ویژگی سبب شده است که مسئله علّیت در یادگیری عمیق صورتی متمایز از مدل‌سازی علمی کلاسیک پیدا کند. از این منظر، اهمیت فلسفی یادگیری عمیق صرفاً در موفقیت مهندسی آن نیست بلکه در این است که مرز میان «کشف ساختار» و «طراحی نظری» را دگرگون می‌کند. بسیاری از ساختارهایی که در این مدل‌ها ظاهر می‌شوند نه حاصل صورت‌بندی مستقیم انسانی بلکه برآمده از تعامل میان داده، معماری شبکه و فرایند بهینه‌سازی‌اند. همین امر مسئله بازنمایی، تبیین و علّیت را به یکی از مسائل مرکزی فلسفه هوش مصنوعی معاصر تبدیل کرده است.

پرسش سوم این است که آیا مدل‌های مبتنی بر یادگیری عمیق واجد محتوای معناشناختی و بازنمایانه هستند یا خیر؟ اگر آری، براساسِ چه نظریه‌ای دربارهٔ معنا/بازنمایی؟ در پاسخ باید بگوییم که سخن‌گفتن از «کشف» یا «بازنمایی» ساختارهای علّی در مدل‌های یادگیری ماشین مستلزم روشن‌کردنِ این پیش‌فرض فلسفی است که آیا مدل‌های یادگیری عمیق اساساً واجد محتوای بازنمایانه هستند یا خیر؟ این پرسش به‌ویژه درموردِ مدل‌های زبانی بزرگ (LLMs) محل مناقشه‌ای جدی در فلسفه هوش مصنوعی و فلسفه ذهن معاصر است. ادعای مطرح شده در این مقاله، بر یک معناشناسی قوی از بازنمایی استوار نیست. یعنی مدعی نیست که مدل‌های یادگیری عمیق واجد قصدیت یا فهم معنا به معنای کلاسیک آن هستند. بلکه مقصود از «بازنمایی» در این مقاله معنایی حداقلی‌تر است. براین‌اساس، یک مدل تا آنجا بازنمایانه تلقی می‌شود که بتواند ساختارهایی پایدار، مداخله‌پذیر و تبیین‌گر را حفظ کند و از آن‌ها برای پیش‌بینی، تعمیم و استدلال خلاف‌واقع بهره گیرد. این تلقی به رویکردهای کارکردی و ساختاری در نظریه بازنمایی نزدیک است، رویکردهایی که نقش بازنمایی را نه در شباهت ذات‌گرایانه با جهان بلکه در حفظ روابط ساختاری و ایفای کارکردهای معرفتی می‌دانند. از این منظر، بازنمایی‌های درونی مدل‌های یادگیری عمیق را می‌توان به‌مثابهٔ ساختارهایی دانست که تاحدی الگوهای پایدار و ناوردای موجود در داده‌ها را ضبط می‌کنند، حتی اگر این بازنمایی‌ها واجد معناشناسی انسانی یا شفافیت مفهومی کامل نباشند. براین‌اساس، وقتی مقاله از «کشف علّیت» در هوش مصنوعی سخن می‌گوید، مراد آن نیست که مدل‌ها لزوماً به ذات نهایی سازوکارهای جهان دست می‌یابند بلکه مقصود این است که آن‌ها می‌توانند ساختارهایی را استخراج کنند که، تحت مداخله، پایدار می‌مانند و در تبیین و تعمیم نقش معرفتی ایفا می‌کنند. بنابراین، اعتبار علّی این مدل‌ها نه بر قصدیت کامل بلکه بر کفایت ساختاری و مداخله‌ای آن‌ها استوار است.

نتیجه‌گیری

این مقاله، با تمرکز بر تحولات اخیر در یادگیری ماشین و ادبیات نوظهور هوش مصنوعی علّی، کوشید جایگاه مفهومی علّیت را در علوم داده‌محور از منظر فلسفه علم بازاندیشی کند. نقطه عزیمت تحلیل آن بود که موفقیت چشمگیر مدل‌های پیش‌بینانه آماری، هرچند در بسیاری از کاربردها کارآمد بوده است، اما به‌تنهایی برای پاسخ‌گویی به نیازهای تبیینی، تصمیم‌گیری مداخله‌ای و تعمیم به شرایط جدید کفایت نمی‌کند. همین محدودیت‌ها زمینه‌ساز بازگشت توجه به مفاهیم علّی در پژوهش‌های معاصر هوش مصنوعی شده‌اند.

دربرابرِ دو تفسیر رایج –یعنی واقع‌گرایی علّیِ قوی که ساختارهای علّی مدل‌شده را بازنمایی مستقیم سازوکارهای بنیادین جهان می‌داند و ابزارگرایی‌ای که علّیت را صرفاً وسیله‌ای محاسباتی برای بهبود عملکرد تلقی می‌کند– این مقاله رویکرد سومی را پیشنهاد کرد. براساسِ این رویکرد، علّیت در یادگیری ماشین به‌مثابهٔ ساختاری تبیینی، میان‌سطحی و وابسته به مداخله فهم می‌شود، ساختاری که اعتبار معرفت‌شناختی خود را نه از صدق متافیزیکی بلکه از ناوردایی تحت مداخله، جهت‌مندی روابط و توانایی تعمیم فراتوزیعی به دست می‌آورد.

تحلیل حوزه‌هایی چون کشف علّیت و حساب مداخله‌ای نشان داد که ساختارهای علّی در یادگیری ماشین، حتی زمانی که از داده‌های مشاهده‌ای استخراج می‌شوند، می‌توانند نقشی اساسی در تبیین، کنترل و استدلال خلاف‌واقع ایفا کنند، بی‌آنکه مستلزم تعهدات متافیزیکی سنگین باشند. از این منظر، علّیت نه امری زائد در علم داده‌محور بلکه عنصری غیرقابل‌جایگزین برای فهم پایداری، تعمیم و عقلانیت تصمیم‌گیری در سیستم‌های هوشمند است.

درنهایت، این مقاله پیشنهاد می‌کند که توجه هم‌زمان به محدودیت‌های معرفتی مدل‌های یادگیری ماشین و توان تبیینی ساختارهای علّی می‌تواند چارچوبی متوازن برای تحلیل علّیت در علم معاصر فراهم آورد، چارچوبی که هم با الزامات عملی هوش مصنوعی سازگار است و هم با دغدغه‌های فلسفه علم دربارهٔ تبیین و فهم علّی هم‌نوا باقی می‌ماند.

References

Douglas, H. E. (2009). Reintroducing prediction to explanation. *Philosophy of Science, 76*(4), 444–463. https://doi.org/10.1086/648111

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics, 10*, Article 524. https://doi.org/10.3389/fgene.2019.00524

Kesh, S., & Whitworth, M. (2025). *Causal AI: How cause and effect will change artificial intelligence*. S&P Global.

Ladyman, J., Ross, D., Spurrett, D., & Collier, J. (2007). *Every thing must go: Metaphysics naturalized*. Oxford University Press.

Lamsaf, A., Carrilho, R., Neves, J. C., & Proença, H. (2025). Causality, machine learning, and feature selection: A survey. *Sensors, 25*, Article 2373. https://doi.org/10.3390/s25082373

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. MIT Press.

Schölkopf, B. (2019). *Causality for machine learning*. *arXiv*. https://arxiv.org/abs/1911.10500

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Towards causal representation learning. *Proceedings of the IEEE, 109*(5), 612–634. https://doi.org/10.1109/JPROC.2021.3058954

Shmueli, G. (2010). *To explain or to predict? arXiv*. https://arxiv.org/abs/1101.0891

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). MIT Press.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

Weiskopf, D. A. (2022). The predictive turn in neuroscience. *Philosophy of Science, 89*(5), 1213–1222. https://doi.org/10.1017/psa.2022.39

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.