The Impact of Artificial Intelligence on the Epistemic Structure of Modern Science

Franc Mali
University of Ljubljana – Faculty of Social Sciences

Abstract

In this paper, we first focus on how the rapid integration of AI into scientific research processes is leading to an epistemic transformation of established scientific categories and principles that have been deeply rooted in our understanding of science since the modern scientific revolution of the mid-seventeenth century. In the second part, we address the implications of these epistemic changes for scientists' ability to explain and interpret AI outputs (e.g., predictions and explanations). For the first time in the history of human thought, the question of who counts as a "knower" in science and what counts as a legitimate "epistemic operation" has become open in ways previously unknown. With a touch of irony, one might say that we are at a turning point in human history at which scientists increasingly assume the role of students, while machines take on the role of teachers. Indeed, many solutions generated by machines contain new ideas and strategies that humans must learn from AI.

Keywords

AI-driven science, epistemology, scientific prediction, AI black box, explainability of AI outputs, closed-loop AI science cycle

1 Introduction

Artificial intelligence (AI)1 is rapidly making its way into every sphere and every level of scientific research. AI is not only changing the speed at which research is conducted—something that everyone working in science is already enthusiastic about—but is also changing our very understanding of what constitutes scientific knowledge, how such knowledge can be demonstrated and explained, and how it can serve as a basis for scientific predictions and explanations. Throughout the centuries, there have been ongoing epistemological debates among theoreticians of science concerning the questions of what scientific knowledge is, how it is formed and justified, and under what conditions it emerges. Of course, the importance attached to each of these questions has varied depending on the historical context. However, with the entry of AI into scientific research processes, we find ourselves at an entirely new turning point. This turning point invites us to rethink our position as scientists in processes of knowledge production.

In this regard, we could say that if Nicolaus Copernicus displaced human beings from the centre of the cosmos and Charles Darwin displaced them from their privileged place in nature, AI now seems to be "dethroning" human beings from their position as the only intelligent entities. Alan Turing's assumption from more than 70 years ago—namely, that human beings might not embody the only possible form of intelligent and rational agency—seems to have become increasingly relevant in recent years. To be clear, contradictory views still exist among experts as to whether such a situation will arise in the near future or whether it has, to some extent, already arisen. Nevertheless, after all these transformations, the role of scientists as knowledge workers can no longer remain the same as it was in the past.

In the first part of our discussion, we will focus on how the rapid integration of AI into scientific research processes is leading to an epistemic transformation of established scientific categories and principles that have been deeply rooted in our understanding of science since the modern scientific revolution of the mid-seventeenth century. In the second part, we will address the implications of these epistemic changes for scientists' ability to explain and interpret AI outputs (e.g., predictions and explanations). With a touch of irony, one might say that we are at a turning point in human history at which scientists must learn from machines. Many solutions generated by machines contain new ideas and strategies that humans must learn from AI.

2 Some Points of Difference Between New AI-Based and Traditional Epistemology of Science

Very generally speaking, the epistemic structures underlying contemporary scientific knowledge can still be traced back to the well-known Kantian distinction between a priori and a posteriori knowledge. The former type of knowledge has traditionally been regarded as independent of experience (e.g., mathematical truths), whereas the latter is regarded as knowledge that relies on empirical evidence (e.g., human experience and observations made using scientific instruments). This means that, after Kant, the concept of the epistemic justification of scientific knowledge evolved in the context of various epistemological theories, such as empiricism, rationalism, positivism, inductivism, deductivism, falsifiability, hermeneutics, or combinations of these approaches [1].

However, AI appears to disrupt these basic pillars of the traditional epistemological view. AI models are transforming the very architecture of modern scientific knowledge production [2] [3] [4] [5]. Unlike traditional epistemology, which derives its explanatory and predictive power from empirically or theoretically grounded laws, AI-based epistemology relies primarily on principles of probability. Before discussing this

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.cogni.123

[1] Although I use the term AI inclusively, in this discussion it refers particularly to its most recent developments over the last few years, including analytical tools such as transformers, large language models (LLMs), and multimodal generative models that rely on techniques such as gradient descent.

radical transformation brought about by AI, we need to address some of the reasons why it is occurring. Let us mention two significant reasons.

First, the question of who counts as a "knower" in science and what counts as a legitimate "epistemic operation" has suddenly become open in ways previously unknown. With AI, the very architecture of knowledge production is being transformed [6]. For example, large language models (LLMs) have begun to (self-)learn from vast corpora of textual data, allowing them to extract patterns and semantics directly from real-world human language use. Second, generative AI has brought about an extraordinary increase in the capacity to solve complex problems across all domains of science. It has also demonstrated its capacity in other spheres of human life to cope with complex, nonlinear, sometimes unstable, and often self-organizing systems in nature, technology, and society [7].

The claim that the complex nature of the phenomena under investigation inherently calls for the assistance of AI can be illustrated by numerous practical cases. For example, given the impressive recent results achieved by AI models in solving complex mathematical problems, it seems that machines are encroaching on one of the last intellectual bastions of humankind. About four years ago, the first versions of LLMs could not even count properly. They subsequently began solving all kinds of school-level mathematical problems, followed by problems in university mathematics. They went on to solve all the problems of the Mathematical Olympiad. In the last year, they have succeeded in solving mathematical puzzles that had long stumped even the world's best researchers (e.g., the latest LLMs by OpenAI disproved a conjecture formulated by the Hungarian mathematician Paul Erdős in 1946 [8]. Jacob Tsimerman, who has recently been awarded the Fields Medal, an honour comparable to the Nobel Prize, has said in an interview with *Neue Zürcher Zeitung*: "Mathematics can't go back to the pre-AI era. For centuries, mathematics had been an economy of scarcity: proofs of new theorems were rare and therefore represented the ultimate goal for mathematicians. Now, an age of abundance is beginning. AI systems are producing correct mathematical proofs on an assembly line" [9].

3 Epistemic Transformation in AI-Driven Science: Predictions Without a Theoretical Foundation, Relying Solely on Big Data and Algorithms

The radical epistemic shift in contemporary science resulting from the emergence of AI can best be observed and evaluated through the transformation of basic epistemological categories, such as scientific prediction, scientific explanation, causality, and probability. Let us consider the category of scientific prediction. Theoreticians of science often state that the predictive power of modern science is one of its defining features [10] [11]. Faith in the predictive power of science was initially based on the expectation that scientific knowledge would eliminate chance and thereby bring nature or society under full control. The most rigorous version of this view was already advocated by classical deterministic thinkers. For example, the French mathematician Pierre-Simon Laplace argued that if we possessed complete knowledge of our world, we would be able to attain a comprehensive understanding of both past and future events.

In modern theories of science, such strict determinism is no longer heuristically productive. Modern scientific predictions are not formulated in the spirit of Laplace's omniscient demon; rather, they are usually interpreted as inductively generalized conclusions (in the spirit of Hume's inductivism [12]) or as deductive inferences aimed at falsifying or confirming theoretical hypotheses (in the spirit of Popper's critical rationalism [13]).

Nevertheless, despite all these methodological variations, the epistemic structure of modern science is grounded in one common principle: the assumption that predictions are always grounded in subject-specific knowledge and, in this sense, are theory-based. As Jan C. Schmidt puts it: "Zwar mögen Prognosen mitunter auf statistischen, nicht auf deterministischen Gesetzmäßigkeiten aufbauen und somit probabilistisch sein, immer jedoch basieren sie auf einem Gegenstandswissen und sind in diesem Sinne theoriebasiert" [14]. Predictions therefore play an important role in processes of planning and decision-making. Auguste Comte, a classical sociological theorist, already expressed the following idea: to know in order to predict, to predict in order to act.

With AI, an epistemic transformation of prediction is taking place. Predictions in the context of AI are probabilistic by design. Although AI models remain strongly oriented towards forecasting, they enable predictions without necessarily providing a theoretical justification for their statements about the future [15]. Even when the newest AI models produce extremely reliable predictions within a particular scientific domain, this does not necessarily mean that they understand the broader underlying principles behind those predictions. Unlike the epistemological assumptions of traditional science—which derives its predictive power from established, empirically validated, and theoretically grounded knowledge about the phenomenon being predicted—the AI forecasting process relies primarily on probability.

In the context of AI forecasting, probabilistic predictions are based on big data and the use of correlations. These data-driven, theory-free predictions show some similarities to inductive reasoning. However, this does not amount to full-fledged inductivism, which aims to infer all future observations from a finite number of measurements and thereby derive general laws [12]. They are therefore distinct from predictions derived from classical deterministic or contemporary stochastic models, which are assumed to be theoretically grounded and constructed on the basis of an understanding of the underlying regularities governing natural or social systems.

As many theoreticians of science have observed, AI models provide algorithms that are primarily oriented towards forecasting (and, on this basis, towards decision-making processes), rather than towards extracting explanatory information, whether causal or otherwise [16], [17], [18]. The existence of causal relationships is, at best, merely assumed, without such connections actually being identified or understood. As Veda C. Storey notes, the probabilistic nature of AI models means that their outputs will vary in ways that are not—and perhaps never will be—fully explainable.

Several practical cases demonstrate that even when current AI systems produce reliable predictions within a particular scientific

domain, this does not necessarily mean that they understand the broader underlying principles behind those predictions. For example, a recent study showed that a generative AI model was able to predict the motion of a planet orbiting a star, yet it was unable to reconstruct the fundamental physical laws governing the motion of these celestial bodies [19]. This situation contrasts with cases from the beginnings of modern natural science. Centuries ago, astronomers such as Johannes Kepler discovered geometric regularities that enabled accurate predictions of the future positions of planets. Isaac Newton later built upon these discoveries to formulate fundamental physical laws that not only allowed for the prediction of planetary motion but also explained the physical principles governing the universe more broadly.

This means that, in the epistemology of classical natural science, it was assumed to be possible to move from predicting sequences to understanding the deeper mechanisms that generate them. According to Kaiser and co-authors, the emergence of accurate predictions that are not directly derived from highly intelligible theories can also be illustrated by the case of AlphaFold 2 [20]. AlphaFold 2 is an algorithm that uses deep learning to predict protein folding with superhuman accuracy but cannot directly produce theoretical contributions to the science of protein folding. The list of such cases is long and extends across different scientific fields, from physics and biotechnology to oceanography and climate science.

The changing epistemic structure of AI-based science remains a relatively unexplored theoretical topic and requires much greater attention. AI models based on self-supervised learning have great potential for solving various scientific problems because they can leverage massive amounts of unlabelled data and transfer their knowledge to low-data regimes. However, these transfer-learning schemes can be ad hoc, lack theoretical guidance, and remain vulnerable to shifts in underlying distributions.

4 AI-Driven Science Reconfigures Both the Role of Scientists in the Research Process and the Epistemic Function of Machines

Even before the emergence of generative AI (GAI), various applications of machine-learning models in science were already raising questions about the relationship between humans and machines. Even experts in the field did not have complete access to all aspects of the functioning of machine-learning models. They did not necessarily understand the pathways through which these models proceeded from input to output. For example, Jenna Burrell wrote that classifications in traditional AI models were considered opaque in the sense that, if one was a recipient of an algorithm's output—that is, its classification decisions—one rarely had any concrete sense of how or why a particular classification had been derived from the inputs [21]. With regard to the question of opacity deriving from the internal "logic" of traditional machine learning, Jenna Burrell distinguished among three forms of opacity: (1) opacity as intentional corporate or state secrecy, (2) opacity as technical illiteracy, and (3) opacity arising from the characteristics of machine-learning algorithms and the scale required to apply them usefully.

Many solutions generated by traditional machine learning contained strategies that humans had to learn from machines. Nevertheless, the increased integration of GAI into scientific practice over the past few years has led to a situation in which the question of who counts as a "knower" in the research process and what counts as a legitimate "epistemic operation" has suddenly become open in ways not seen since the modern scientific revolution. With a touch of irony, one might say that we are at a turning point in human history at which scientists increasingly assume the role of students, while machines take on the role of teachers.

The new epistemic structure of AI-driven science has significant implications for the question of how scientists can achieve explainability of AI outputs. Scientists, like other users, often find themselves in situations in which they rely on AI strategies that they do not fully understand. They trust the achieved goals and empirical payoffs, but not necessarily the methodological derivations. In other words, they are impressed by the operational efficiency of AI models but cannot fully explain how the outputs have been generated or which inputs have been critical in producing them. For scientists, AI models therefore operate as "black boxes."

In the expert literature, we encounter various suggestions for how to address the problem of bringing transparency to the existing AI "black box." For example, Richard Heersmink and co-authors differentiate among three notions of transparency intended to facilitate explanations of the inner "logic" of AI: reflective, phenomenological, and informational transparency [5]. We will not engage in a more detailed discussion of this issue here.

The lack of explainability and interpretability[2] of the AI "black box" can lead to numerous negative practical consequences (e.g., a lack of trust among users), legal consequences (e.g., difficulties for regulatory bodies in determining whether a particular system processes data fairly and securely), and, last but not least, negative epistemological and methodological implications. With regard to the negative epistemological and methodological implications, it is important to highlight problems such as the inability to properly assess biases, difficulties in correctly interpreting model results, and limited applicability in areas where model outputs must be understood prior to real-world implementation.

When it comes to real-world applications, the issue of AI transparency becomes particularly significant. The degree of transparency, for instance, directly affects the extent to which we can place trust in AI systems deployed in high-stakes contexts, such as autonomous driving and medical diagnosis. An additional factor contributing to the lack of transparency is that proprietary AI models, which are controlled by large AI corporations, largely obscure their training data and design choices. Their susceptibility to external scrutiny is therefore limited. Open-source AI models are more transparent, but they require significant computational resources, which are unevenly

[2] Daniel Giacobbe and coautors insist that, in order to fully understand the challenges and opportunities associated with explainability, a distinction should be made between interpretability and explainability [22]. A model is interpretable when each step in the computations leading to the eventual prediction, as well as the relative contribution of the different features, can be easily inferred from a "white-box" model, whereas explainability refers to attempts to explain, as far as possible, how a "black-box" model produces its output. In this sense, interpretability can be understood as full explainability.

distributed across countries and institutions. The lack of such resources can create risks of misuse and misinterpretation.

In any case, the AI black box is not merely a technical issue. It represents a major epistemic challenge for modern science. Will scientists still be motivated to deepen their explanations of the natural and human worlds once AI reaches the level of general human intelligence? As stated by Hector Zenil and co-authors, the integration of new information into the corpus of existing scientific knowledge remains one of the most intellectually demanding aspects of scientific discovery [23]. Here, we encounter not only the challenge of simply appending new data to old knowledge but also the challenge of reconciling and reinterpreting existing knowledge and even overturning entire paradigms. In the context of AI-led discovery based on algorithms and big data, such reconciliation and reinterpretation cannot be taken for granted.

As we have already argued, we are reaching a point at which GAI-driven science is beginning to evolve according to a "logic" that human scientists can exploit but no longer fully understand. Scientists are currently faced with the dilemma of how much control to relinquish in order to accelerate discovery while keeping outputs relevant to humans. But what will happen in the future if GAI models take over the entire process of scientific research? Trends towards the development of GAI-driven science in which models autonomously perform tasks across the entire research pipeline—from reviewing the literature on a given topic and formulating hypotheses to testing solutions, writing scientific papers, and ultimately even carrying out peer-review tasks—are becoming increasingly evident.

If and when we enter such a closed-loop AI science cycle, even greater uncertainty will arise regarding the explainability of AI-generated scientific outputs. Will scientists ultimately be entirely excluded from the explainability and interpretability of AI-generated scientific outcomes? It must be acknowledged that "AI Scientist" models tested in the field of machine-learning research have already demonstrated the ability to autonomously perform tasks across such closed-loop cycles. For all the reasons mentioned above, we need to prevent AI models from completely replacing human scientists in processes of scientific discovery. Instead, we need a form of development in which human interpretative depth and ethical awareness, on the one hand, and machine speed and memory, on the other, can coexist and co-evolve.

5 Conclusion

In our discussion, we have sought to point out that the recent development of AI-driven science is leading to significant changes in the epistemic structures of modern science. To understand these changes, it is necessary to examine the concrete methodological tools used in science. We have directed our attention to the changing role of scientific prediction. Considering that, from antiquity onwards, people have sought to foresee the future and predict forthcoming events, it is not surprising that the strong functional relationship between prediction and theoretical explanation is also highly relevant in modern science. However, with the emergence of generative AI, an epistemic transformation of prediction is taking place. AI predictions rely on statistical correlations without necessarily being provided with a strong explanatory basis. In this paper, we have sought to show how these changes are occurring and what their current and potential future implications may be for scientists' ability to explain and interpret AI outputs.

Literature

[1] Franc Mali, 2006. Epistemologija družbenih ved. Razlaga in razumevanje. Ljubljana: Založba FDV.

[2] Gabriella Punziano, 2025. Adaptive Epistemology: Embracing Generative AI as a Paradigm Shift in Social Science. Societies, MDPI, vol. 15(7): 1-18; DOI: https://doi.org/10.3390/soc15070205

[3] Ammar Younas and Yi Zeng, 2024. A Philosophical Inquiry into AI-Inclusive Epistemology. (June 23, 2024). Available at SSRN: DOI: https://ssrn.com/abstract=4902415; http://dx.doi.org/10.2139/ssrn.4902415

[4] Eddy Keming Chen, Mikhail Belkin, Leon Bergen, David Danks, 2026. Does AI already have human-level intelligence? The evidence is clear. Nature, Feb;650(8100):36-40. DOI: https: doi 10.1038/d41586-026-00285-6.

[5] Richard Heersmink, Barend de Rooij, María Jimena Clavel Vázquez, Matteo Colombo, 2024. A phenomenology and epistemology of large language models: transparency, trust, and trustworthiness. Ethics and Information Technology, 26, 41; DOI: https://doi.org/10.1007/s10676-024-09777-3

[6] Dan Hendrycks et al. 2025. A Definition of AGI; arXiv:2510.18212v3 [cs.AI] 3 Dec 2025

[7] Minsu Kim and James Thorne, 2024. Epistemology of Language Models. Do Language Models Have Holistic Knowledge? ACL, August 11-16: 12644–12669; DOI: https://aclanthology.org/2024.findings-acl.751.

[8] Theresa Palm, 2026. Was wird aus der Mathematik, wenn KI alles besser weiß? Sueddeutsche Zeitung, 13. August, 2026.

[9] Christoph Drösser, 2026. KI in der Mathematik: Jahrhundertkrise oder Neuanfang? Neue Zuericher Zeitung, 14. August, 2026.

[10] Peter Janch, 1997. Kleine Philosophie der Naturwissenschaften. München: Beck Verlag

[11] Bunge Mario, 1996. Finding Philosophy in Social Sciences.New Haven and London: Yale University Press.

[12] David Hume, 2007. An Enquiry Concerning Human Understanding. Oxford: Oxford University Press.

[13] Karl R. Popper, 1971. Prognose und Prophetie in Sozialwissenschaften. In: E. Topitsch (Ed.), Logik der Sozialwissenschaften, 113-126. Koeln und Berlin: Kiepenheuer & Witsch Verlag.

[14] Jan C. Schmidt, 2022. Wandel und Kontinuität von Wissenschaft durch KI. In: C. Gethman, P.Buxman, J. Dirstearth (Ed), Künstliche Intelligenz in der Forschung, 79-122. Berlin: Springer Verlag.

[15] Carl F. Gethmann, 2022. Zur Frage der Ersetzbarkeit des Menschen durch KI in der Forschung. In: C. Gethman, P.Buxman, J. Dirstearth (Ed), Künstliche Intelligenz in der Forschung, 43-79. Berlin: Springer Verlag.

[16] Sanja Srećković, Andrea Berber, Nenad Filipović, 2022. The Automated Laplacean Demon: How ML Challenges Our Views on Prediction and Explanation. Minds and Machines, 32:159–183; DOI: https://doi.org/10.1007/s11023-021-09575-6.

[17] Hanchen Wang, Tianfan Fu, Yuanqi Du et al., 2023. Scientific discovery in the age of artificial intelligence. Nature, 620(7972):47-60. DOI: https//: doi: 10.1038/s41586-023-06221-2.

[18] Mario Krenn and Heather Champion, 2026. Philosophy of Autonomous Science:Ten Questions for the Coming Age of Artificial Scientists. Daedalus, 155 (1-2): 335–349. DOI: https://doi.org/10.1162/DAED.a.991

[19] Veda C. Storey, 2026. AI tools boost scientists' impact but narrow focus. Nature, 649(8099):1111-1112. DOI: doi: 10.1038/d41586-025-04092-3.

[20] Bryan Kaiser, Tailin Wu, Maike Sonnewald, Colin Thackray, Skylar Callis, 2024. A Moonshot for AI Oracles in the Sciences. arXiv:2406.17836 [cs.AI]; DOI: https://doi.org/10.48550/arXiv.2406.17836.

[21] Jenna Burrell, 2016. How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. Big Data & Society, 3(1), 1–12. DOI: https: // doi.org/10.1177/2053951715622512.

[22] Daniele R. Giacobbea, Yudong Zhangc, José de la Fuentee, 2023. Explainable artificial intelligence and machine learning: novel approaches to face infectious diseases challenges, Annals of Medicine, 55(2):2286336; DOI: hhtps: doi: 10.1080/07853890.2023.2286336.

[23] Hector Zenil, Jesper Tegner, Felipe Abraho, 2026. The future of fundamental science led by generative closed-loop artificial intelligence. Frontiers in Artificial Intelligence, 9 : 1-16; DOI: https://doi.org/10.3389/frai.2026.1678539.