The Challenge of AI Scientists to Scientific Epistemology: Postprint

Authors: Duan Weiwen

Date: 2024-04-03T00:00:00+00:00

Abstract

[Purpose/Significance] This study aims to investigate the challenges that AI scientists may pose to scientific epistemology. [Method/Process] Scientific discovery has long been a central theme in artificial intelligence research. The next frontier in AI-driven scientific discovery is the development of AI scientists—autonomous systems capable of automated scientific discovery whose research quality is indistinguishable from that of the best human scientists. Following a review of relevant applications of artificial intelligence in scientific research, this paper explicates the most salient characteristics of AI scientists and the core of their research agenda. On this basis, it proposes that AI scientists engender two fundamental epistemological transformations: a leap in AI capabilities and an AI-driven paradigm shift in scientific research. [Results/Conclusion] Discussions concerning relevant issues in scientific epistemology must transcend general philosophical argumentation. In anticipation of the impending AI scientific revolution, this paper proposes four epistemological strategies: suspending negative criticism, attending to transitional challenges, dynamically tracking potential breakthroughs, and seeking superior analogies.

Full Text

Artificial Intelligence Scientists' Challenge to Scientific Epistemology

Duan Weiwen1,2,3 1. School of Philosophy, University of Chinese Academy of Social Sciences, Beijing 102445 2. Institute of Philosophy, Chinese Academy of Social Sciences, Beijing 100732 3. Shanghai Laboratory for Artificial Intelligence, Shanghai 200232

Abstract

[**Purpose/Significance**] This study aims to explore the potential challenges that artificial intelligence (AI) scientists may pose to scientific epistemology.

[**Method/Process**] Scientific discovery has long been a central theme in AI research. The next frontier is the development of AI scientists—systems capable of autonomous and automated scientific discovery whose research quality is indistinguishable from that of the best human scientists. After reviewing relevant applications of AI in scientific research, this paper elaborates on the most important characteristics of AI scientists and the core of their research agenda. On this basis, it proposes that AI scientists will bring about two fundamental epistemological changes: a leap in AI capabilities and a paradigm shift in AI-driven scientific research.

[**Results/Conclusions**] Discussions on relevant epistemological issues must move beyond general philosophical debates. In anticipation of the impending AI scientific revolution, this paper proposes four epistemological strategies: suspending negative criticism, focusing on transitional challenges, dynamically tracking potential breakthroughs, and seeking better analogies.

Keywords: AI for Science; AI scientists; scientific epistemology; automated science

1. The Epistemological Significance of AI Scientists and the "Science of Science"

The emergence of non-human AI models like ChatGPT marks a milestone in scientific development. This naturally evokes memories of RoboCup, the robot soccer World Cup initiated by Japanese systems biologist Hiroaki Kitano. Inspired by RoboCup, Kitano proposed a new grand challenge in 2016: developing AI systems capable of making major scientific discoveries, even Nobel Prize-winning achievements [1]. In 2017, the USC Information Sciences Institute hosted a workshop on "Thoughtful Artificial Intelligence," co-chaired by Yolanda Gil and Ross King [2], who advocate for "automating science" [3]. This call launched a global initiative for the Nobel Turing Challenge. Kitano further elaborated in *Nature* that the Nobel Turing Challenge should develop highly autonomous AI systems capable of top-tier scientific research, systems biology being a prime candidate. These systems are expected to make major discoveries and potentially win Nobel Prizes by 2050 [4].

Scientific discovery has always been a crucial theme in AI research. As Kitano notes, this journey spans nearly 80 years, from early systems like DENDRAL and META-DENDRAL to later developments such as EURISKO and MYCIN. The DENDRAL project, initiated in the 1960s by Stanford AI researchers Edward Feigenbaum and Nobel laureate Joshua Lederberg, aimed to help organic chemists identify unknown molecules through mass spectrometry and existing chemical knowledge. One of its main sub-projects, Meta-DENDRAL, designed for the Viking Mars mission, sought to build an automated scientific research system for remote environments inaccessible to human scientists. This interdisciplinary effort included machine learning pioneer Bruce Buchanan and renowned chemist Carl Djerassi, representing a landmark in cross-disciplinary collaboration.

Another notable system is BEACON, which Simon promoted to formalize inductive reasoning for discovering empirical laws, typically employing data-driven heuristic methods. Despite controversies—since such systems fit scientific equations to cleaned data, differing significantly from real discovery processes—BEACON remains a milestone in AI-driven scientific discovery. More recently, researchers have developed automated experimental systems in biology and other fields that can generate hypotheses, plan experiments, and execute closed loops. For instance, systems have discovered Hamiltonians, Lagrangians, and conservation laws in geometry and momentum without any prior knowledge of physics or kinematics [5]. While limited to single datasets or specific tasks with constrained resources, these systems clearly indicate that the next step for AI in scientific discovery is developing fully automated scientific discovery systems that will revolutionize research methods. Accelerating this historical process requires bold, provocative goals through clear mission statements—passing the Turing test for AI scientists, where their research quality becomes indistinguishable from top human scientists.

The vision of AI scientists carries profound epistemological significance. In the philosophy of science and science and technology studies, what Kitano calls the "Science of Science"—the science of scientific processes, or metascience—has traditionally focused on human scientific activities. However, building autonomous AI scientists creates an entirely new scientific form: automated scientific research. As Feynman said, "What I cannot create, I do not understand." By creating alternative forms of scientific discovery, we accelerate science and reveal its more general nature, benefiting civilization. This represents a transformation from AI as cognitive automation tools to autonomous cognitive agents. Scientific activities and epistemology will no longer necessarily be human-centered, potentially shifting toward AI-dominated processes based on agent experience, perception, and reasoning.

2. The Vision of Scientific Automation: The Feigenbaum Test and AI Capability Challenges

Whether called machine discovery, discovery science, or scientific AI, the common vision is continuously advancing scientific automation. From the AI perspective, robot scientists or AI scientists are experimental subjects within AI and robotics research. From the scientific discovery perspective, they become the subjects conducting research and experiments [10]. This dual role frames the AI capability challenge: if we still view AI scientists as merely imitating human scientists, the challenge can be expressed as a scientific version of the Turing test. However, the Nobel Turing Challenge aims higher—creating AI systems with autonomous research capabilities that can make Nobel-level discoveries independently or collaboratively.

Given current AI development, passing even a standard Turing test remains difficult, let alone a Nobel Turing test. In 2003, Feigenbaum—"father of expert systems" and 1994 Turing Award winner—proposed the Feigenbaum Test as an alternative. Building on Jim Gray's vision of computational intelligence, Feigenbaum grounded his test in two principles: both human and artificial intelligence can be reduced to computational intelligence, and the test should focus on an AI system's reasoning quality within a specific discipline. The test involves an elite scientist from the National Academy of Sciences evaluating an AI system alongside a human colleague, asking specialized scientific questions (e.g., explaining a theory or astrophysical phenomenon). If the judge cannot reliably distinguish the AI from the human with probabilistic accuracy, the challenge is met. However, Feigenbaum acknowledged this was a grand challenge; a superintelligent computer would paradoxically be easily distinguished by its superior inductive and problem-solving abilities, meaning the test's success would also signal its obsolescence.

The real question is how to achieve the leap to superintelligent systems. Feigenbaum identified the bottleneck as knowledge bases. Twenty years earlier, AI's "science deficit" led to machine learning, but it shifted from symbolic concepts to statistical patterns, leaving large knowledge bases—critical for computational intelligence—largely untouched. To address this, he proposed two challenges: building large knowledge bases through text reading, and reducing knowledge engineering costs by orders of magnitude, potentially extracting knowledge from the World Wide Web. Though he didn't foresee connectionist deep learning and generative AI, he presciently saw the Web as the world's largest database, including cultural-historical texts and current events—"the apple on the knowledge tree" for knowledge engineers [11].

3. The Paradigm Shift in AI-Driven Scientific Research and Its Epistemological Challenges

The paradigm shift in AI-driven research, whether called "AI for Science" or the "fifth paradigm," brings revolutionary transformation. Advocates like King and Kitano describe a future where AI dominates through closed-loop automated systems running the full research cycle—from hypothesis generation to experimental validation and reinterpretation—working faster than humans while reducing bias and opening new discovery spaces. Achieving this requires overcoming enormous technical challenges: strategic goal selection, generating boundary-pushing hypotheses, designing novel experiments beyond prototypes, and communicating discoveries understandably to humans [13].

Robot scientists represent the physical embodiment of this vision. King's first-generation Robot Scientist Adam autonomously generated functional genomics hypotheses about yeast and tested them experimentally. The second-generation Eve focused on drug discovery for neglected tropical diseases, using econometric models to make screening economically viable [15]. The third-generation Genesis aims to automate thousands of parallel closed-loop cycles to learn computational models of eukaryotic cells. Another system, Coscientist, developed by Daniil A. Boiko at Carnegie Mellon, can plan and execute chemical research cycles from simple prompts, designing experiments for reactions, coding pipetting robots, and learning from results [16].

These systems reveal the evolution from AI tools to AI scientists. However, the best current systems cannot define their own hypothesis spaces or experimental designs, making them at best primitive forms of AI scientists. The path forward will be shaped by frontier AI developments like generative AI and large models. The key epistemological challenge is that scientific cognition may no longer be exclusively human. As computational philosopher Paul Humphreys sharply notes, "The measure of all things" is changing. As we extend our perceptual abilities, scientific epistemology will no longer be human epistemology [17]. This creates a paradox: human cognitive enhancement through AI may push humans from the center of scientific cognition and decision-making, raising new dilemmas about science's meaning and value.

Two aspects stand out: First, AI combined with other instruments constitutes cognitive augmentation for human scientists, extending human reasoning beyond natural limits. But will science become independent of human cognition entirely? Second, AI agents and human scientists are forming a network of intelligent agents for research and discovery. Future advanced AI will coordinate these networks cleverly and effectively, but this symbiosis may render our civilization more powerful yet potentially more vulnerable through over-reliance.

4. Concluding Remarks: Epistemological Strategies for the AI Scientific Revolution

The impact of this human-independent science will be revolutionary, yet recent explorations increasingly support its possibility. Discussions must move beyond abstract philosophical debates to strategies for the coming revolution:

First, suspend negative criticism. Rather than abstractly arguing whether AI can discover genuine scientific problems or innovate, focus on concrete problems and solutions in AI scientist development. Issues like AI's epistemic opacity and robustness in scientific research require attention. The path forward may involve acknowledging parallel mechanisms between machine and human intelligence, using comparisons between AI systems to reveal opacity, and exploring adversarial research on data/process interference to establish safety and stability as evaluation metrics.

Second, focus on transitional challenges. Current automated systems are narrow and task-specific, while humans offer complementary strengths through creative perspectives and non-traditional insights. As AI research systems develop, we should set progressive goals: AI tools, AI assistants, AI collaborators. Key transitional issues include how AI systems help overcome human cognitive limitations, avoid biases, address reproducibility concerns, and prevent new paradigm retreats.

Third, dynamically track potential breakthroughs. Instead of fixating on conceptual debates about absolute autonomy or whether AI can drive scientific revolutions, track emerging combinations and breakthroughs. For instance, USC computer scientist Yolanda Gil notes that large language models excel at scientific brainstorming, while others observe that models generating plausible but inaccurate information actually produce hypotheses [18]. This dynamic tracking reveals practical pathways.

Fourth, seek better analogies. To understand AI scientists' relationship to human scientists, we need new analogies beyond human/subhuman/superhuman categories. These analogies help reveal possibilities and address issues like hallucination in generative AI research. The goal is to think through various developmental possibilities rather than remaining conceptually constrained.

The AI scientific revolution requires epistemological optimization and integration. By acknowledging machine and human intelligence as parallel mechanisms, using complementary strategies, and addressing the epistemic challenges of automation, we can harness AI to make our civilization more powerful through revolutionary scientific discovery.

References

[1] KITANO H. Artificial intelligence to win the Nobel prize and beyond: Creating the engine for scientific discovery[J]. AI magazine, 2016, 37(1): 39-49.

[2] GIL Y. Thoughtful artificial intelligence: Forging a new partnership for data science and scientific discovery[J]. Data science, 2017, 1(1/2): 119-131.

[3] KING R D, ROWLAND J, OLIVER S G, et al. The automation of science[J]. Science, 2009, 324(5923): 85-89.

[4] KITANO H. Nobel Turing Challenge: Creating the engine for scientific discovery[J]. NPJ systems biology and applications, 2021, 7: 29.

[5] SCHMIDT M, LIPSON H. Distilling free-form natural laws from experimental data[J]. Science, 2009, 324(5923): 81-85.

[6] CHO A. AI in Action: AI's early proving ground: The hunt for new particles[J]. Science, 2017, 357(6346): 20.

[7] COLEY C W, EYKE N S, JENSEN K F. Autonomous discovery in the chemical sciences part I: Progress[J]. Angew. Chem., Int. Ed., 2020, 59: 22858-22893.

[8] COLEY C W, EYKE N S, JENSEN K F. Autonomous discovery in the chemical sciences part II: Outlook[J]. Angew. Chem., Int. Ed., 2020, 59: 23414-23436.

[9] AI scientist grand challenge: Summary of discussion during workshop held in February 2020[EB/OL]. [2023-08-05]. https://www.turing.ac.uk/sites/default/files/2021-02/summary_{{{of}}*{{discussion}}*}{{{workshop}}*{{2020}}*}{{{ai}}*{{scientist}}*}{{{grand}}*{{challenge}}*

[10] SOLDATOVA L N, CLARE A, SPARKES A, et al. An ontology for a robot scientist[J]. Bioinformatics (Oxford, England), 2006, 22(14): e464-e471.

[11] FEIGENBAUM E A. Some challenges and grand challenges for computational intelligence[J]. Journal of the ACM, 2003, 50(1): 32-40.

[12] GRAY J. What next? A dozen information-technology research goals[J]. Journal of the ACM, 50(1): 41-57.

[13] OECD. Artificial intelligence in science: Challenges, opportunities and the future of research[M]. Paris: OECD Publishing, 2023.

[14] KING R, ZENIL H. A framework for evaluating the AI-driven automation of science, in Artificial intelligence in science: Challenges, opportunities and the future of research[M]. Paris: OECD Publishing, 2023.

[15] KING R D, SCHULER COSTA V, MELLINGWOOD C, et al. Automating sciences: Philosophical and social dimensions[J]. IEEE technology and society magazine, 2018, 37(1): 40-46.

[16] BOIKO D A, MACKNIGHT R, KLINE B, et al. Autonomous chemical research with large language models[J]. Nature, 2023, 624(7992): 570-578.

[17] HUMPHREYS P W. Extending ourselves: Computational science, empiricism, and scientific method[M]. Beijing: People's Publishing House, 2017.

[18] MATTHEW H. Hypotheses devised by AI could find "blind spots" in research[EB/OL]. [2023-11-17]. https://www.nature.com/articles/d41586-023-03596-0.

DUAN Weiwen (1968-), Professor and doctoral supervisor at the School of Philosophy, University of Chinese Academy of Social Sciences; Researcher at the Institute of Philosophy, Chinese Academy of Social Sciences. E-mail: duanweiwencass@163.com

[**Purpose/Significance**] This study aims to explore the challenges that artificially intelligent (AI) scientists may bring to scientific epistemology.

[**Method/Process**] Scientific discovery has long been a central theme in AI research. The next frontier is the development of AI scientists—systems capable of autonomous and automated scientific discovery whose research quality is indistinguishable from that of the best human scientists. After reviewing relevant applications of AI in scientific research, this paper elaborates on the most important characteristics of AI scientists and the core of their research agenda. On this basis, it proposes that AI scientists will bring about two fundamental epistemological changes: a leap in AI capabilities and a paradigm shift in AI-driven scientific research.

[**Results/Conclusions**] Discussions on relevant epistemological issues must move beyond general philosophical debates. In anticipation of the impending AI scientific revolution, this paper proposes four epistemological strategies: suspending negative criticism, focusing on transitional challenges, dynamically tracking potential breakthroughs, and seeking better analogies.

Keywords: AI for Science; AI scientists; scientific epistemology; automated science

Note: Figure translations are in progress. See original paper for figures.