DOI:10.1145/3771100

David Benrimoh, David Harel, Nace Mikus, Peter Stone, and Ariel Rosenfeld

Opinion

The Einstein Test: A Test of AI's Ability to Generate Transformative Science

Assessing whether an AI system can independently "rediscover" known scientific breakthroughs.

Recent advances in artificial intelligence (AI) have prompted speculation about whether machines can match, or even surpass, top human creativity and insight. As we see it, a more focused question is this: Can AI generate transformative scientific breakthroughs—of the kind that require creative leaps and redefine our understanding of nature and the universe, such as relativity, evolution, or quantum mechanics? To address this question, we propose the *Einstein Test*, which assesses whether a given AI system can independently "rediscover" known scientific breakthroughs.

The Einstein Test

The Einstein Test first proposes a *retrospective* approach, where a candidate AI would be presented with a curated dataset of knowledge preceding a transformative discovery—such as the body of information available before 1905, when Einstein developed special relativity—and tasked with solving the fundamental problems that led to the breakthrough (essentially simulating the historical conditions). Success in that instance would be defined as the AI system generating a solution formally equivalent to, or superseding, the historical discovery.

The test would proceed as follows: First, the architecture, specification, and training procedures for a candidate AI system will be submitted to the testing team. Second, an expert committee will choose an unseen historical scientific breakthrough for the candidate system to try to achieve. The system would then be provided with a curated dataset, containing all available knowledge up to the development of that breakthrough, carefully excluding any post-discovery knowledge. This dataset, the preparation of which is obviously nontrivial, can be used for training and may be revisited during the test. Importantly, the candidate system will have to be carefully assessed to make sure that it does not have access during training to any external knowledge other than that provided by the expert committee (which would include the curated dataset, and any simulations of the physical world relevant to embodied or virtually embodied AIs).

IMAGE FROM SHUTTERSTOCK.COM

Third, the candidate system would be provided with initial guidance by the testing team—in whatever format its architecture requires—to address the unsolved issues faced by the scientists of that time. For example, in the case of special relativity, the AI system would be challenged to explain time dilation or the behavior of light and electromagnetic waves. This guidance would need to be carefully constructed to avoid introducing an inadvertent "guiding hand"—for example, by highlighting key pieces of information that in hindsight were crucial to the discovery, but which may not have been evidently related to the solution at the time.

Fourth, a designated expert team would act as a "research assistant" for the candidate system. If the system requests data resulting from some experiment, the team would then provide the requested data if it was available during or would have been possible in the context of the time period of the selected breakthrough; otherwise, they would respond that such an experi-

ment is not feasible using the available experimental techniques or apparatus, in which case the candidate system would be free to design and suggest its own approach. In such a case, the expert team would have to see to it that the experiment is actually carried out, and then provide the resulting data.

The candidate AI system would be allowed to run until: it states that it has accomplished the task, provides an incorrect answer that it asserts is correct, or declares it was unable to provide an answer. The candidate's answer will be formally verified and compared to the actual historical breakthrough.

Unlike the Turing Test, which assesses a machine's ability to imitate a human's normative responses successfully, or its biological modeling variant,[3] which similarly assesses the ability to faithfully model a biological system's behavior, the Einstein Test directly measures a machine's ability to conduct groundbreaking, paradigm-shifting scientific inquiry—of the kind that only a few celebrated humans have achieved. It establishes a clear, falsifiable criterion: either the candidate AI system reproduces the breakthrough or does not. As such, it mitigates much of the ambiguity associated with existing evaluation criteria of machine intelligence, such as the ARC-AGI[2] or "Humanity's Last Exam,"[6] which assesses an AI system's ability to solve highly complex human-generated challenges. We note a very recent example of an AI system proposing a microbiology hypothesis, after being given only introductory information.[4] In that instance, the AI independently arrived at a hypothesis that aligned with unpublished findings from human researchers, thereby mirroring a key aspect of the Einstein Test—the ability to rediscover transformative insights using only pre-existing knowledge. We applaud this effort, and consider it a valuable proof of concept that underscores the timely need for clear and verifiable criteria to evaluate AI's ability to generate transformative science. Other efforts, such as literature-based discovery, have also made inroads toward creating novel scientific concepts using AI.[1] In addition, in the time in which we had been thinking about the Einstein Test, leading thinkers had suggested similar ideas, suggesting some convergence of scientific thinking on this issue (see Perrigo[5] and Wolf[7]).

The Einstein Test is mainly a framework for assessing AI's ability to generate transformative science retrospectively.

Practical Considerations

Several practical considerations should be addressed, before the Einstein Test can be operationalized. We deliberately do not lay out any exact technical specifications for the test, as we believe these should be determined and agreed upon by science historians, AI experts, and leading figures in the disciplines from which the considered breakthroughs will be submitted as challenges to a candidate AI system. Key practical considerations, some of which may be significantly challenging to address, include the introduction of realistic time and resource limits, procedures for selecting the experts and the breakthroughs themselves, the means of curating the corresponding pre-discovery knowledge, the details of the procedures for determining whether the candidate system has successfully reproduced a historical discovery, and, significantly, how many challenges and of what nature should a system be required to successfully deal with in order to be labeled as having passed the Test in its full generality. We acknowledge that, in particular, ensuring that the dataset to be used in a given challenge is not contaminated by information from after the breakthrough would be a significant challenge, requiring careful scrutiny by the organizers of the challenge.

Implications and Future Directions

The Einstein Test is mainly a framework for assessing AI's ability to generate transformative science retrospectively. However, it can also serve as a gateway to understanding AI's prospective abilities. Specifically, a prospective Einstein Test would provide an AI system (ideally, one that has been successful in passing the retrospective Test) with all currently available knowledge and challenge it to generate new breakthroughs in some specified area of research. Clearly, this prospective evaluation would require substantially more resources and time, compared to the retrospective one. Following the maxim, "the best predictor of the future is the past," success in the retrospective Einstein Test could signal the potential for prospective success.

By making it possible to assess an AI system's ability to generate historical scientific milestones, we could move beyond speculation and toward a tangible, structured, empirical test of AI's ability to bring about new scientific breakthroughs. Adhering to the constraints of this test may, in addition, spur the development of novel AI architectures capable of generating significant and creative new scientific ideas. C

References

1. Akujuobi, U. et al. Link prediction for hypothesis generation: An active curriculum learning infused temporal graph-based approach. *Artif. Intell. Rev. 57*, 244 (2024); 10.1007/s10462-024-10885-1
2. Chollet, F., Knoop, M., Kamradt, G., and Landers, B. Arc prize 2024. Technical report. (2024). *arXiv preprint arXiv:2412.04604*
3. Harel, D. A Turing-Like Test for Biological Modeling. *Nature Biotechnology 23* (2005).
4. Penades, J.R. et al. AI mirrors experimental science to uncover a novel mechanism of gene transfer crucial to bacterial evolution. *bioRxiv 2025-02* (2025).
5. Perrigo, B. Demis Hassabis is preparing for AI's endgame. *Time* (2025); https://time.com/7277608/demis-hassabis-interview-time100-2025/
6. Phan, L. et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249* (2025).
7. Wolf, T. The Einstein AI model. (2025); https://thomwolf.io/blog/scientific-ai.html

David Benrimoh (david.benrimoh@mail.mcgill.ca) is an assistant professor at McGill University in Montreal, Quebec, Canada.

David Harel (david.harel@weizmann.ac.il) is president of the Israel Academy of Sciences and Humanities in Jerusalem, Israel, and an Institute Professor at The Weizmann Institute of Science, Rehovot, Israel.

Nace Mikus (nace.mikus@univie.ac.at) is a postdoctoral Fellow at Aarhus University in Aarhus, Denmark.

Peter Stone (pstone@cs.utexas.edu) is a professor of Computer Science at The University of Texas at Austin and the Chief Scientist of Sony AI in Austin, TX, USA.

Ariel Rosenfeld (ariel.rosenfeld@biu.ac.il) is an associate professor of Information Science and Applied Artificial Intelligence at Bar-Ilan University, Ramat-Gan, Israel.

© 2026 Copyright held by the owner/author(s).