The next Turing tests: Reimagining conceptions and measures of AI

New approaches to evaluating progress of AI should be built around diversity, partnership, and systemicity

By Jose Hernandez-Orallo[1,2], Katherine M. Collins[1,3,4], Lucy Cheke[1,5], Laura Weidinger[6], Umang Bhatt[1], Thomas L. Griffiths[4], Stephen Cave[1]

[1]Institute for Technology and Humanity, University of Cambridge, Cambridge, UK. [2]Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, València, Spain. [3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Princeton Laboratory for Artificial Intelligence, Princeton University, Princeton, NJ, USA. [5]Department of Psychology, University of Cambridge, Cambridge, UK. [6]Google DeepMind, London, UK.

Email: jh2135@cam.ac.uk

CITE LIKE THIS:

- Hernandez-Orallo, J.; Collins, K.M.; Cheke, L.; Weidinger, L.; Bhatt, U.; Griffiths, T.L.; Cave, S. The next Turing tests: Reimagining conceptions and measures of AI. *Science* **393**, 566-568 (2026). DOI:10.1126/science.aee3176

ABSTRACT

In 1950 Alan Turing famously proposed replacing the question “Can machines think?” with a different question: “Can a machine pass for a human?” (1). His thought experiment, in which the machine had to fool a human judge in a teletype conversation (now known as the Turing test), established a vision of human-like machine intelligence that has shaped both the popular imagination and artificial intelligence (AI) developments for over 75 years. With AI affecting ever more areas of science and society, merely contesting this vision, as has been done many times in the past, is not enough. We need to open up a new range of options for measuring progress—while at the same time democratizing who evaluates AI and for what purpose. We argue that, to achieve these goals, the next Turing tests should be built from three standpoints: diversity, partnership, and systemicity.

In 1950 Alan Turing famously proposed replacing the question “Can machines think?” with a different question: “Can a machine pass for a human?” (1). His thought experiment, in which the machine had to fool a human judge in a teletype conversation (now known as the Turing test), established a vision of human-like machine intelligence that has shaped both the popular imagination and artificial intelligence (AI) developments for over 75 years. With AI affecting ever more areas of science and society, merely contesting this vision, as has been done many times in the past, is not enough. We need to open up a new range of options for measuring progress—while at the same time democratizing who evaluates AI and for what purpose. We argue that, to achieve these goals, the next Turing tests should be built from three standpoints: diversity, partnership, and systemicity.

Turing did not mean his thought experiment to be the goal of AI, but decades of slow progress consolidated the misconception that many, if not all, limitations of AI would dissipate once an AI system passed the Turing test. Now the test has been passed, but modern AI still displays substantial limitations, from capabilities to safety. It had already been argued that this fixation on the Turing test was “harmful,” prompting calls to instead design tests for specific tasks that humans cannot or prefer not to perform (2). However, this alternative is at odds with modern, general-purpose AI, which is increasingly capable of many cognitive tasks at which humans excel. To measure this progress, AI benchmarks have proliferated, leaving the original Turing test behind. Still, the legacy of the imitation game hangs over the field (3): The targets of human-level machine intelligence (HLMI) and the related, yet more nebulous, notion of artificial general intelligence (AGI) have recently been refurbished as achieving AI that can perform all purely cognitive tasks that a human adult can. This is reflected by new waves of tests targeting

AGI (4, 5). If passing the Turing test has not led to universally beneficial AI, why should we expect otherwise now, with a similarly anthropomorphic testing paradigm?

To really avoid falling back to anthropomorphic conceptions of AI, we urgently need to consider alternative conceptions and measures that are operational with modern systems, by rethinking evaluation from three different standpoints. First, we will not revisit here the evaluation of AI systems up to or beyond human level but rather explore how to test for different cognitive profiles, composed of abstract capabilities and propensities, in diversity. Second, instead of measuring which human activities machines can replace, we will ask how we can evaluate human cognition as transformed by new kinds of thinking with AI, in partnership. Third, rather than testing how aligned AI is to human values, we advocate for the evaluation of benefits and harms of AI interventions transforming societies, in full systemicity.

Following Turing's replacement of "Can machines think?" with a different—more testable—question, we highlight contemporary questions that we think need to be replaced by new ones, each demanding new tests. The new questions are associated with thought experiments, similar to how the imitation game was a thought experiment about a kind of AI that did not exist at the time, with idealized conditions (teletype conversation) and subjects in separate categories being assimilated (woman-man with machine) (6). The original questions need to be replaced by new ones, overcoming existing blockers through new thought experiments and testing instruments, leading to a reconception of what to measure and what effective actions are to be taken to change the AI evaluation agenda (see the table). We explore this from three different standpoints below.

DIVERSITY: CAN AN AI PROFILE FIT A PARTICULAR NICHE?

A helpful metaphor is to view each new AI development as a new species. In biology, we no longer consider each species to fall on a single "scala naturae" ladder with humans (or god) at the top. Ecological diversity is useful to capture how AI systems—like other species—fill different niches, with different specialisms and weaknesses, and are subject to specific resource limitations. However, testing each and every task that any new AI system can do would be logistically overwhelming and would fail to explain the system's behavior. Instead, tests should map the landscape of capabilities and propensities: Where does each new AI system fall along the dimensions that are relevant to existing and new forms of intelligence (7)?

This diversity is often challenged by a monolithic view of progress. Some hypothesize that as soon as one player reaches AGI, this will be followed by an accelerated, convergent self-improvement, making it harder for other organizations to compete, reinforcing a winner-takes-all mindset, and contributing to AI race dynamics (8). The race is often presented as about automating AI engineers or cracking problems only intellectuals can solve (9), rather than assessing the desirable properties that various artificial systems should meet. The pattern of building larger AI models to scale up capabilities has also led to spiraling costs and complexity associated with evaluation, constraining who can conduct such evaluations, with an impact on what gets tested and why.

We propose an alternative for making the vast diversity of potential AI behaviors more widely understandable and controllable. Each AI system can be characterized by a relatively small profile of traits (capabilities and propensities) that not only capture its behavior but also the way the system was built and trained, comparable to species-specific phylogeny and ontology with animal behavior. The use of abstract traits allows AI users, developers, and policy-makers to anticipate, for instance, that an AI agent with level 5 in finance but level 2 in metacognition

may file a hallucinated tax declaration in a case that demands metacognition level 4 (perhaps because it requires evaluation of what customer information is available). For this approach to be democratized, these profiles must be interpretable (10), the scales and dimensions must be commensurate, and the data underpinning them must be open for reanalysis (11).

Future thought experiments should explore what profiles are scientifically possible, e.g., can we imagine an AI system with high levels of metacognition that will never try to modify its own code? Or an AI system with high creativity but a low propensity to take risks? From the standpoint of diversity, the tests should answer the question, What is the profile of strengths and weaknesses of each AI system that is built, and how does it align with the requirements of the tasks at hand?

To make this a reality, aggregate scores from benchmarks are not enough: We need tests that allow for triangulations, e.g., having a task that is highly sensitive to creativity yet orthogonal to narcissism and risk aversion. To this end we need to build consortia, including academia, developers, and policy-makers, that work on an agreed catalog of traits, their scales, and mechanisms to map from benchmark results. From these catalogs, stakeholders and regulators can find consensus on what profiles we would like to be deployed, beyond monolithic thresholds on compute levels or benchmark aggregations.

PARTNERSHIP: IS A HUMAN INDIVIDUAL EMPOWERED BY AI?

The second standpoint considers AI systems and people thinking together, rather than independently. Some of the AI profiles identified from the first diversity standpoint can combine with humans in different ways, also depending on each human's characteristics, from theory of mind to neuroticism. This dyadic symbiosis raises new questions and creates additional evaluative demands. We should be able to measure different configurations of AI systems

working as partners to humans (12)—whether as co-workers or companions. The main target of such tests is not the machine, or how well it replaces equivalent human work (13), but how the joint human–machine partnership produces a new experience for humans.

Evaluating how much AI transforms human cognition suffers from scalability issues and the subjectivity of human experience. Such challenges are exacerbated by the possibility for AI to adapt to what people expect. Today's AI is typically designed and tested to please humans, rather than empower them in the long term. It is crucial to test for the latter. Yet rigorous, longitudinal evaluation requires human–AI interaction research or large real-world user logs, making it difficult to perform systematic analyses of outcomes such as addiction, suicidal events, AI psychosis, and long-term cognitive effects such as over-reliance and skill atrophy. In all these cases, feedback should come not only from the users directly involved but also from the gamut of stakeholders, such as mental health experts, teachers, etc. If we want to reshape who evaluates AI and for what, we need wider access to human–AI interaction data, which is now mostly owned by the AI developers. Changing this situation could include regulations on what human–AI interaction data can be accessed by formal audits, or mechanisms for independent collectives and syndicated users to (anonymously) access the data.

Designing and evaluating AI as thought partners can accelerate profound changes in human cognition, with new kinds of behavior and culture likely emerging in the decades ahead. Similarly, AI companies can be incentivized beyond replacement, with the cognitive transformation of humans creating new needs and services.

In this transformation of human cognition, new thought experiments should challenge the limits and conceptions of human thinking, and explore variations of the human–AI partnership, e.g., could the human have reached the same outcome without the AI partner and vice versa?

To this end, we need tests that look into the interaction between humans and AI and are sensitive to new ways of thinking (such as discussions with an artificial persona of oneself). Testing diversity and partnership are complementary, as one way of determining that a particular human's way of thought has not been diluted is to check for the preservation of the person's identity and style after long AI exposure. This requires more longitudinal analyses such as those performed in education and mental health, two domains used to monitor how cognition evolves. It is in these contexts where the procedures, infrastructure, data, and ethical standards for these kinds of evaluation can be ready to be deployed.

SYSTEMICITY: IS AN AI USE NET POSITIVE FOR SOCIETY?

The third stance views this interaction between humans and AI in highly interconnected and changing societies. Some positive local effects can lead to global negative effects: An AI system can be aligned with one human or group, yet strongly misaligned with others; a highly capable AI can have low risk when used by some people, but high risk when used by malicious players—or conversely, by millions of well-intentioned people; a machine replacing humans in the workplace could boost productivity but cause redundancies and incentivize algorithmic monocultures. The ultimate goal for evaluation is not whether a given task is perfectly imitated or solved in a better way, or whether the values of the AI system align with those of humans, as if that would prevent any societal harm caused by AI, but to understand the broad and long-term benefits (14) and risks of each new use of AI in society.

Currently, the deliberation of societal benefits and risks is insufficiently participatory, as the targets of AI penetration are dominated by the culture of task replacement and productivity, instead of societal well-being, guided by democratic societal feedback, integrating expertise and participation from different disciplines, cultures, and interests. This is again exacerbated by the

difficulty of evaluating AI in ecologically valid scenarios, which may require large-scale field experiments, unfit to the pace of AI development.

Instead, we propose simulated ecosystems for evaluation that allow more accessible testing and a wider range of interventions than those that can be performed in the real world. These sandboxes should include a diversity of AI system profiles and artificial personas (realistic human models, an area where human-like AI can be useful). This is particularly relevant for safety testing, as many sources of risk of AI will originate from rare situations with humans. Anticipating such issues in these simulated scenarios is an incentive for companies that may suffer reputational and legal consequences when issues or harms manifest during deployment. Shared domain-specific regulatory or experimental sandboxes, such as the Testing and Experimentation Facilities in the European Union (EU) (15), can allow companies and policy-makers to deploy pilots, include AI systems from different providers, and conduct (virtual) human studies, allowing smaller players to have more accessible ways of testing their AI interventions.

These “sandbox” tests should play with the counterfactuals of what would have happened with and without the AI intervention, or with different AI interventions. For instance, thought experiments about AI in society could study different degrees of penetration of the technology, exploring the paradoxes of extreme situations such as full job replacement in a particular sector. In this context, more systemic evaluations should allow us to address the question, Will people be better or worse off after a particular AI intervention? Which people, and under what conditions?

It is no surprise that AI will be key to facilitating such evaluation: Simulated testing ecosystems can be generated semi-automatically by AI tools from the description of AI safety

incidents or other situations of interest, humans can be replaced by artificial personas in these simulations, and AI tools can help guide the search for extreme or unusual situations. The creation and maintenance of these digital twins mimicking real-world AI use will require extensive infrastructure, established by sectorial or cross-sectorial consortia. Yet running tests in these platforms should be open to all stakeholders, incentivized by national and international AI strategies.

Although this systemicity requires clarity about who makes the decisions about what is good for society, thus framing evaluation as a sociotechnical problem, the other two standpoints also set new targets for research and industry that must be elicited in a participatory way. Through diversity, partnership, and systemicity (see the table), we have suggested new thought experiments, test frameworks, and actions that can play the role of the Next Turing Tests to help society understand and steer AI in the direction of desirable goals—fitting, empowering, beneficial—while managing the costs we might have to pay to create, use, and evaluate AI.

Table 1. Comparing the "imitation" framing of the original Turing test paradigm with the three framings presented in the paper: diversity, partnership, and systemicity.

Framing	Imitation	Diversity	Partnership	Systemicity
Original Question	Can machines think?	Can AI be superhuman?	Can AI replace humans?	Can we align AI with human values?
Issues and Blockers	• "Thinking" eludes precise definitions. • Many objections in 1950 to the idea of machines thinking.	• Monolithic view of AI beyond "AGI." • Frontier AI is seen as disembodied, aiming at purely digital tasks.	• Heavy reliance of AI techniques on human feedback and preferences. • Limited access to longitudinal data of human–AI interaction.	• Concentration of stakeholders for defining societal benefits and harms. • Difficulty of evaluating AI in ecologically valid scenarios.
New Question	Can a machine pass for a human?	Can an AI profile fit a particular niche?	Is a human individual empowered by AI?	Is an AI use net positive for society?
Thought Experiments	• Imagine interaction with AI through teletype, based only on behavior, ignoring substrate. • Imagine the machine is pretending to be a human.	• Imagine a particular AI profile as possible. • Imagine a particular AI profile as desirable. • Imagine how profiles align with the risks and demands of the tasks at hand.	• Imagine the human reaching the same outcome without the AI system, and vice versa. • Imagine the human after long-term AI interaction. • Imagine the human after the AI partner changes or quits.	• Imagine what would have happened with and without the AI intervention. • Imagine a full replacement of some human activity by AI. • Imagine a moratorium of AI deployment in a domain.
Testing Instruments	• Teletype interview in which a human has to distinguish between a human and a machine pretending to be a human.	• Commensurate scales to infer AI profiles that are interpretable from granular-level result data.	• Questionnaires from users and other stakeholders. • Computational cognitive models of interaction, and transcript inspections.	• AI interventions for causal understanding. • Simulated ecosystems with diversity of AI systems and human models.
What Measures	• Human likeness of an AI system. • Propensity of humans to ascribe thought to AI.	• Capabilities and propensities of an AI system. • Demands and risks of tasks.	• Human empowerment with AI. • Synergy and collaboration of human–AI partnership.	• Societal benefits and risks of AI interventions. • Reversibility of AI interventions.
Actions	• Public: Change opinion. Stimulate philosophical discussions. • Researchers and funders: Improve methods to detect AI impersonation (CAPTCHAs). Aim at various tests to determine when HLMI or AGI is achieved. • Regulators and policy-makers: Prepare for cheap cognitive labor and massive workplace replacement.	• Public: Use AI profiles for more informed choice about what AI to use and where to use it. • Researchers and funders: Prioritize evaluation science and actual testing for neglected and poorly understood AI profiles. • Regulators and policy-makers: Base decisions on profiles rather than aggregate indicators, compute, or monolithic capability thresholds.	• Public: Quantify cognitive (dis)empowerment in new human–AI experiences. • Researchers and funders: Focus on mechanisms to evaluate the partnership fit between humans and AI, especially in education and health. Monitor impact of AI on human cognition (reasoning, atrophy). • Regulators and policy-makers: Issue recommendations about the suitability of AI profiles depending on the human characteristics (age, mental health).	• Public: Play with gamified tools to explore specific AI uses and futures, avoiding technodeterminism. • Researchers and funders: Improve the realism of evaluative ecosystems and ease of use for tests that would be impractical or unethical in the real world. • Regulators and policy-makers: Explore AI policies and interventions in the digital twins. Inform the effects of moratoria and possibility of reversion of AI.

REFERENCES AND NOTES

1. A. M. Turing, Mind 59, 33 (1950).
2. P. Hayes, K. M. Ford, “Turing Test Considered Harmful” in Proceedings of the Fourteenth International Joint Conferences on Artificial Intelligence (1995), vol. 1, pp. 972–977.
3. M. Mitchell, Science 385, eadq9356 (2024).
4. D. Hendrycks et al., arXiv:2510.18212 (2025).
5. R. Burnell et al., https://blog.google/innovation-and-ai/models-and-research/google-deepmind/measuring-agi-cognitive-framework/.
6. B. Gonçalves, Minds Mach. 33, 1 (2023).
7. R. T. McCoy, S. Yao, D. Friedman, M. D. Hardy, T. L. Griffiths, Proc. Natl. Acad. Sci. U.S.A. 121, e2322420121 (2024).
8. S. Cave, S. ÓhÉigeartaigh, “An AI race for strategic advantage: Rhetoric and risks” in Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (Association for Computing Machinery, 2018), pp. 36–40.
9. L. Phan et al., Nature 649, 1139 (2026).
10. L. Zhou et al., Nature 652, 58 (2026).
11. R. Burnell et al., Science 380, 136 (2023).
12. K. M. Collins et al., Nat. Hum. Behav. 8, 1851 (2024).
13. T. Kwa et al., Adv. Neural Inf. Process. Syst. 38 (2025).
14. C. Graham, K. Laffan, S. Pinto, Science 362, 287 (2018).
15. Sectorial AI Testing and Experimentation Facilities under the Digital Europe Programme; https://digital-strategy.ec.europa.eu/en/policies/testing-and-experimentation-facilities.

ACKNOWLEDGMENTS

This commentary evolved from discussions at an event to celebrate Alan Turing’s 1950 paper, held at King’s College, Cambridge (https://www.kingselab.org/the-next-turing-tests) in October 2025. J.H.-O. receives funding from the EU, the Spanish and the Valencian government. J.H.-O.’s research is also supported by OpenAI’s grant to the “AI Progress through the Lens of Predictable AI Ecosystems” program, based at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge. K.M.C. acknowledges support from the National Science Foundation Directorate for Social, Behavioral and Economic Sciences (SBE) Postdoctoral Research Fellowships (SPRF). L.W. receives a salary from, and holds stocks of, Google LLC. S.C. acknowledges the support of the Leverhulme Trust and Stiftung Mercator.