Intelligent Computing
A SCIENCE PARTNER JOURNAL

PERSPECTIVE

What Should Replace the Turing Test?

Philip N. Johnson-Laird[1,2] and Marco Ragni[3*]

[1]Department of Psychology, Princeton University, Princeton, NJ 08540, USA. [2]Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, USA. [3]Technische Universität Chemnitz, Thüringer Weg 11, 09126 Chemnitz, Germany.

*Address correspondence to: marco.ragni@hsw.tu-chemnitz.de

Citation: Johnson-Laird PN, Ragni M. What Should Replace the Turing Test? *Intell. Comput.* 2023;2:Article 0064. https://doi.org/10.34133/icomputing.0064

Submitted 15 July 2023
Accepted 9 October 2023
Published 10 November 2023

Today, chatbots and other artificial intelligence tools pass the Turing test, which was Turing's alternative to trying to answer the question: can a machine think? Despite their success in passing the Turing test, these machines do not think. We therefore propose a test of a more focused question: does a program reason in the way that humans reason? This test treats an "intelligent" program as though it were a participant in a psychological study and has 3 steps: (a) test the program in a set of experiments examining its inferences, (b) test its understanding of its own way of reasoning, and (c) examine, if possible, the cognitive adequacy of the source code for the program.

Most computer people know of the Turing test [1]. Turing devised it as an alternative to the question: can machines think? In the standard test, you communicate from a terminal to a human and to a computer program, and your task is to identify which is which from their answers to your questions. If you cannot do any better in this task than in the analogous task of deciding which is which—a man and a woman similarly communicating with you—then the program passes the test. To cognitive scientists, the test fails to address thinking—and Turing knew that it did. Various chatbots pass the test. For example, one of us (P.N.J.-L.) was sent a text, and thought its author was a human plagiarist. However, it was the output of a program using a large language model. Given that such algorithms do not reason in the way that humans do, the Turing test and any others it has inspired are obsolete.

We propose to replace the Turing test with a more focused and fundamental one to answer the question: do programs reason in the way that humans reason? The mechanisms of human reasoning are controversial in cognitive science, but the theory of mental models, e.g., [2,3], has led to experiments showing that it is not based on any standard logic. Here are 3 tell-tale cases:

- Humans infer possibilities from compound assertions [4], e.g., inferences of the sort:

It is cold or cloudy.
Therefore, possibly it is cold.

It is sensible, but it is invalid in all standard modal logics, which deal with possibilities.

- Humans reject logically valid inferences if they refer to a possibility for which premises provide no support [4,5], e.g.,

Possibly it is hot.
Therefore, possibly it is hot or humid or both.

- Humans condense consistent possibilities into one [6], again contrary to all standard modal logics, e.g.,

Few customers ate steak or sole.
Therefore, few customers ate steak.

If "few" is replaced with "most", reasoners reject the inference. Indeed, logic cannot predict what conclusions humans will infer. That is because any premises validly imply infinitely many different conclusions (i.e., conclusions that are true in all cases in which the premises are true). The theory of mental models in sum is this: people build mental models of the possibilities to which premises refer and tend to draw conclusions that hold in at least one of these possibilities without excluding any of the other possibilities to which they refer (see https://www.modeltheory.org/). Likewise, standard logics have no procedure for withdrawing a valid conclusion. To use the jargon: logic is monotonic; everyday reasoning is non-monotonic, because each possibility referred to by, say, a disjunction holds in default of knowledge to the contrary. Therefore, each model of a possibility can, in principle, be withdrawn, but at least one possibility must hold for the disjunction to be true.

Evidence from experiments corroborates the model theory, and programs simulate all its essentials, which go much further than our sketch here [7–10]. How can we convert the theory into tests to answer our fundamental question? Many of the pertinent tests depend on the difference between sensible inferences that follow from mental models and inferences that are valid in standard logics but nonsensical; e.g., It is raining; therefore, it is raining or there is a rhinoceros in your bath, or both. The answer to our fundamental question can be obtained by a straightforward process consisting of 3 steps:

1. Test the program in a series of psychological experiments about reasoning

These experiments should be those, such as the examples cited above, for which human reasoning differs significantly from logic. If the program differs from humans, we have answered the question. It does not reason like a human. However, if its performance does not differ significantly from human reasoning, we go to the second step.

2. Test the program's understanding of its own way of reasoning

These experiments should call for the program to "introspect" on its own performance. Here is an example:

Please answer the following question:

If Ann is intelligent, does it follow that Ann is intelligent or she is rich, or both?

If the program rejects this inference (as humans do) even though it is logically valid, then the next question is:

Why do you think that the inference does not follow?

A sign of human-like reasoning is this sort of answer:

Nothing in the premise supports the possibility that Ann is rich.

If the program passes this test, the third step is analytic.

3. Examine the source code for the program

If it contains the same major components of the programs that are known to simulate human performance, that evidence is decisive. These components include an intuitive system for rapid inferences, a deliberative system for more thoughtful reasoning, including the withdrawal of conclusions that evidence refutes, and a system for modulating the interpretation of terms such as "or" as a consequence of their context and general knowledge (e.g., [8]). If it relies instead on some sort of deep learning, then the answer is equivocal—at least until another algorithm is able to explain how the program reasons. If its principles are quite different from human ones, it has failed the test.

In sum, we propose to replace the original Turing test with an examination of a program's reasoning. We treat it as a participant in a series of cognitive experiments, and, if need be, we submit its code to an analysis that is an analog of a brain-imaging study.

Acknowledgments

The authors are grateful to David Guerth, who has implemented an algorithm for reasoning about possibilities according to the theory of mental models. They also thank Ruth Byrne; Sunny Khemlani, who maintains the model theory's website; and many others for their contributions to the theory.

Funding: This research was partially funded to MR by the research initiative "Instant teaming between humans and production systems" of the Chemnitz University of Technology and co-financed by the Saxony State Ministry of Science and Art (grant SMWK3-7304/35/3-2021/48192).

Author contributions: P.N.J.-L. and M.R. discussed, formulated, and developed the presented concepts. P.N.J.-L. wrote a first version.

Competing interests: The authors declare that they have no competing interests.

References

1. Turing AM. Computing machinery and intelligence. *Mind*. 1950;LIX(236):433–460.
2. Johnson-Laird PN. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge (MA): Harvard University Press; 1983.
3. Johnson-Laird PN, Khemlani S, Byrne RMJ. Truth, verification, and reasoning. *Proc Natl Acad Sci*. 2023;120(40):1.
4. Hinterecker T, Knauff M, Johnson-Laird PN. Modality, probability, and mental models. *J Exp Psychol Learn Mem Cogn*. 2016;42(10):1606–1620.
5. Ragni M, Johnson-Laird PN. Reasoning about epistemic possibilities. *Acta Psychol*. 2020;208.
6. Johnson-Laird PN, Ragni M. Possibilities as the foundation of reasoning. *Cognition*. 2019;193.
7. Johnson-Laird PN, Quelhas AC, Rasga C. The mental model theory of free choice permissions and paradoxical disjunctive inferences. *J Cogn Psychol*. 2021;33(8):951–973.
8. Johnson-Laird PN, Bucciarelli M, Mackiewicz R, Khemlani SS. Recursion in programs, thought, and language. *Psychon Bull Rev*. 2021;29:430–454 .
9. Khemlani S, Johnson-Laird PN. Reasoning about properties: A computational theory. *Psychol Rev*. 2020;129(2):289–312 .
10. Guerth D. mModalSentential. 2019 [Accessed 11 Oct 2023] https://github.com/CognitiveComputationLab/cogmods/tree/master/modal/2019_guerth