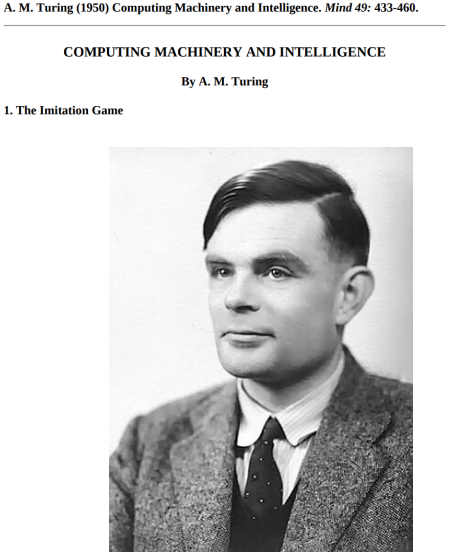
Emily Riehl

Johns Hopkins University

Testing Artificial Mathematical Intelligence

Deep-Learning Models for Mathematics and Type Theory

Can machines think?



A. M. Turing (1950) Computing Machinery and Intelligence. *Mind 49:* 433-460.

COMPUTING MACHINERY AND INTELLIGENCE

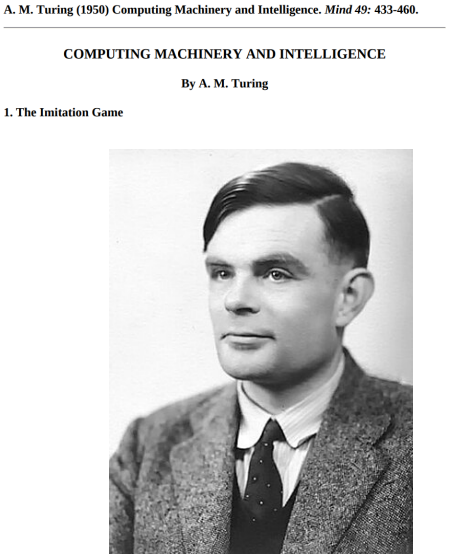
By A. M. Turing

1. The Imitation Game

In a famous 1950 paper entitled "Computing machinery and intelligence," Alan Turing opens with the question

"Can machines think?"

Can machines think?



A. M. Turing (1950) Computing Machinery and Intelligence. *Mind 49:* 433-460.

COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. The Imitation Game

In a famous 1950 paper entitled "Computing machinery and intelligence," Alan Turing opens with the question

"Can machines think?"

Sidestepping the task of giving precise meaning to the term "think," Turing instead describes what he calls "the imitation game" and asks whether an interrogator in another room would be able to distinguish between a man and machine impersonating a man through a series of typewritten questions.

Can machines think?

A. M. Turing (1950) Computing Machinery and Intelligence. *Mind 49:* 433-460.

COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. The Imitation Game

In a famous 1950 paper entitled "Computing machinery and intelligence," Alan Turing opens with the question

"Can machines think?"

Sidestepping the task of giving precise meaning to the term "think," Turing instead describes what he calls "the imitation game" and asks whether an interrogator in another room would be able to distinguish between a man and machine impersonating a man through a series of typewritten questions.[a]

[a] Actually, the imitation game is much queerer than this: Turing asks whether a machine would do better than a man at impersonating a woman.

The Turing test

Turing predicted that in 50 years time, machines would be able to pass what is now known as the Turing test.

The Turing test

Turing predicted that in 50 years time, machines would be able to pass what is now known as the Turing test.

This is arguably the case for today's large language models.

The Turing test

Turing predicted that in 50 years time, machines would be able to pass what is now known as the Turing test.

This is arguably the case for today's large language models.

But I want to ask a different question:

"Can machines do mathematics?"

The Turing test

Turing predicted that in 50 years time, machines would be able to pass what is now known as the Turing test.

This is arguably the case for today's large language models.

But I want to ask a different question:

"Can machines do mathematics?"

No single test can capture the multifaceted meaning of the phrase "doing mathematics," so instead I want to propose a series of benchmarks that may help us identify whether or not a machine is doing mathematically meaningful work.

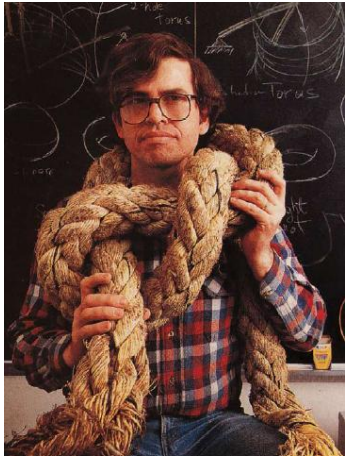
On proof and progress in mathematics

In a famous 1994 essay "On proof and progress in mathematics," Bill Thurston opens with the question

"What is it that mathematicians accomplish?"

which he rephrases in a more leading form as

"How do mathematicians advance human understanding of mathematics?"



On proof and progress in mathematics

In a famous 1994 essay "On proof and progress in mathematics," Bill Thurston opens with the question

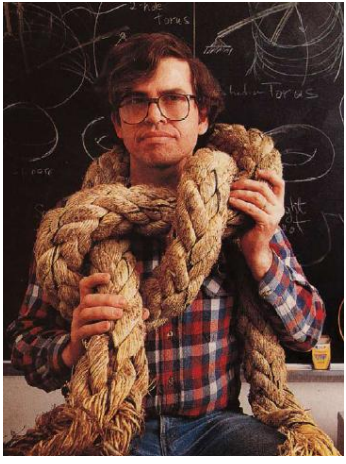
"What is it that mathematicians accomplish?"

which he rephrases in a more leading form as

"How do mathematicians advance human understanding of mathematics?"

He continues:

> "This question brings to the fore something that is fundamental and pervasive: that what we are doing is finding ways for *people* to understand and think about mathematics. "



On human understanding of mathematics

Thurston's point is that mathematics is about more than just definitions, theorems, and proofs or getting to the "right answers":

> *The rapid advance of computers has helped dramatize this point, because computers and people are very different. For instance, when Appel and Haken completed a proof of the 4-color map theorem using a massive automatic computation, it evoked much controversy. I interpret the controversy as having little to do with doubt people had as to the veracity of the theorem or the correctness of the proof. Rather, it reflected a continuing desire for human understanding of a proof, in addition to knowledge that the theorem is true.*
>
> *On a more everyday level, it is common for people first starting to grapple with computers to make large-scale computations of things they might have done on a smaller scale by hand. They might print out a table of the first 10,000 primes, only to find that their printout isn't something they really wanted after all. They discover by this kind of experience that what they really want is usually not some collection of "answers"—what they want is understanding.*

The essential recursive nature of mathematics and mathematicians

Thurston speculates that the circularity in the idea of

"mathematicians advancing human understanding of mathematics"

reveals an essential recursive quality of mathematics:

The essential recursive nature of mathematics and mathematicians

Thurston speculates that the circularity in the idea of

"mathematicians advancing human understanding of mathematics"

reveals an essential recursive quality of mathematics:

Mathematics is the smallest subject

- that includes the natural numbers and plane and solid geometry
- as well as that which mathematicians study,

where mathematicians are those who advance human understanding of mathematics.

The essential recursive nature of mathematics and mathematicians

Thurston speculates that the circularity in the idea of

"mathematicians advancing human understanding of mathematics"

reveals an essential recursive quality of mathematics:

Mathematics is the smallest subject

- that includes the natural numbers and plane and solid geometry
- as well as that which mathematicians study,

where mathematicians are those who advance human understanding of mathematics.

Actually, Thurston writes

"mathematicians are those humans who advance human understanding of mathematics"

but this paraphrasing is intended to be more inclusive
of artificial mathematical intelligence.

Testing artificial mathematical intelligence

Mathematical intelligence is multifaceted and human mathematical activity even moreso.

Testing artificial mathematical intelligence

Mathematical intelligence is multifaceted and human mathematical activity even moreso.

We propose a series of tests to help identify whether a computer system can meaningfully contribute to the process of doing mathematics.

Testing artificial mathematical intelligence

Mathematical intelligence is multifaceted and human mathematical activity even moreso.

We propose a series of tests to help identify whether a computer system can meaningfully contribute to the process of doing mathematics.

These tests are independent of each other: just as human mathematicians exhibit diverse strengths, we expect a single machine to find certain tasks more difficult than others.

Testing artificial mathematical intelligence

Mathematical intelligence is multifaceted and human mathematical activity even moreso.

We propose a series of tests to help identify whether a computer system can meaningfully contribute to the process of doing mathematics.

These tests are independent of each other: just as human mathematicians exhibit diverse strengths, we expect a single machine to find certain tasks more difficult than others.

The difficulty of these tests also varies considerably. Some could be passed by currently existing systems, while others — like Turing's — might require decades of development.

Plan

1. Teaching and learning mathematics
2. Normative and ethical uses of AI in mathematics
3. Mathematical research and discovery
4. Conclusions and outlook

1

Teaching and learning mathematics

Calculation test

Given a natural language mathematical query involving a computational component, interface with an appropriate computer algebra system or SAT solver, displaying both the source code and the result.

To be judged by: a developer of the calculational system with sufficient subject matter knowledge.

Further challenges:

- Check the calculation in a different computer algebra system and if the results differ, determine which is more likely to be correct.
- Answer users queries about the strengths and limitations of different software and follow expert best practices in selecting a program for the computational task.

Example Generation Test

Generate examples of a given mathematical definition, with references, described at the level of a specified audience — for instance undergraduates who have taken a first course in algebra — providing more details when asked.

To be judged by: the named authors, for instance of textbook references, when available who can comment on the appropriateness of the reference and how standard the examples are.

Further challenges:

- Give a useful non-example that illustrates the intuition behind each axiom.
- Repeat this exercise for more technical definitions across a variety of subfields.

Grading test

Catch every mistake in the course-wide submissions for an undergraduate level mathematics problem set and explain the errors, without false positives or negatives.

To be judged by: the students (for comprehensibility of the explanation), the TAs (for comparison against best grading practices), and the instructor (for accuracy and a second opinion on the prior criteria).

Further challenges:

- Answer further student queries about related subject matter, acting as a personal tutor, referring where appropriate to provided course materials.
- Respond appropriately to regrade requests from both the students and the instructor.

Reading comprehension test

Summarize the results of a recent research paper with

- no false statements and
- no major omissions — perhaps after prompting "anything else?"

To be judged by: the authors of the paper.

Further challenges:

- Repeat the experiment for a paper from each of the arXiv's 32 categories, which the arXiv's moderators attest are representative.
- Adjust the summary text for an audience with a particular background.
- Supply more details when demanded, with specific references to the paper or to the literature, where appropriate.

Communication test

Identify the main step of a natural language proof, produced either by a machine or by humans.

To be judged by: the authors of the proof, where available, or an expert who has written a textbook on the subject.

Further challenges:

- Describe intuition or insights related to the main step.
- Communicate the ideas of the proof to an expert in the field.[1]

[1] Thurston points out that "mathematical knowledge can be transmitted amazingly fast within a subfield" where there is a rich store of shared common knowledge.

Literature search test

Answer a natural language query with a ranked list of relevant mathematical papers and books, with no hallucinated references.

To be judged by: the authors of those texts where available (for appropriateness of their inclusion) and a community forum such as mathOVERFLOW or appropriate Zulip chat/Discord server (for omissions).

Further challenges:

- Given a natural language statement of a theorem, find the first reference where it appears.
- Given a natural language definition, find the first reference where it appears, as well as a list of synonyms and references for those.

Translation test

Translate a mathematical text between two languages, without "overtranslating" technical terms (e.g., "éspace étalé" translates to "étale space" not "spread out space").

To be judged by: the author of a closely related text in the new language if available; otherwise a subject matter expert who is fluent in both languages.

Further challenges:

- Point out synonyms for technical terms, in cases where multiple translations are in common usage.
- Add citations to relevant expositions in the new language, where appropriate, to complement the original citations.

2

Normative and ethical uses of AI in mathematics

Norms of acceptable AI mathematical activity

The tests proposed here are intended also to convey norms of acceptable use within an ethical framework focused on potential harms to

- people (for instance through bias or insufficient consideration) or
- science (for instance through falsehoods or a flood of low-quality work[2]).

[2]We might call this the "Spotify problem" or "vibe proving."

Norms of acceptable AI mathematical activity

The tests proposed here are intended also to convey norms of acceptable use within an ethical framework focused on potential harms to

- people (for instance through bias or insufficient consideration) or
- science (for instance through falsehoods or a flood of low-quality work[2]).

Particular concerns are raised by the prospects of AI performing the following activities:

- graduate admissions/hiring (artificially generating preliminary rankings)
- refereeing (beyond assisting with reading comprehension and literature search)
- original mathematical writing (beyond assistance with editing, translation, or LaTeX typesetting).

[2]We might call this the "Spotify problem" or "vibe proving."

Norms for machine-generated mathematical proof

Despite well-known imperfections, the mathematical community can take deep pride in our overwhelmingly reliable and continually improving standards for mathematical proof.

Norms for machine-generated mathematical proof

Despite well-known imperfections, the mathematical community can take deep pride in our overwhelmingly reliable and continually improving standards for mathematical proof.

We should demand the same for AI when it comes to the mathematical realm.

Norms for machine-generated mathematical proof

Despite well-known imperfections, the mathematical community can take deep pride in our overwhelmingly reliable and continually improving standards for mathematical proof.

We should demand the same for AI when it comes to the mathematical realm.

Maintaining high standards will frustrate near term progress, delaying the arrival of a machine we validate as having "artificial mathematical intelligence," but should be beneficial for overall reliability in the long run, in mathematics and beyond.

Norms for machine-generated mathematical proof

Despite well-known imperfections, the mathematical community can take deep pride in our overwhelmingly reliable and continually improving standards for mathematical proof.

We should demand the same for AI when it comes to the mathematical realm.

Maintaining high standards will frustrate near term progress, delaying the arrival of a machine we validate as having "artificial mathematical intelligence," but should be beneficial for overall reliability in the long run, in mathematics and beyond.

Specifically, I want to propose the following norm for the mathematical community when it comes to original mathematics produced by an AI system:

Norms for machine-generated mathematical proof

Despite well-known imperfections, the mathematical community can take deep pride in our overwhelmingly reliable and continually improving standards for mathematical proof.

We should demand the same for AI when it comes to the mathematical realm.

Maintaining high standards will frustrate near term progress, delaying the arrival of a machine we validate as having "artificial mathematical intelligence," but should be beneficial for overall reliability in the long run, in mathematics and beyond.

Specifically, I want to propose the following norm for the mathematical community when it comes to original mathematics produced by an AI system:

Any artificially generated mathematical text will not be considered as a proof unless:

- It has been communicated in both a natural language text paired with a computer formalization of all definitions, theorems, and proofs.
- The formalization has been accepted by the proof assistant and human expert referees have vetted both the formalization and the paired text.

3

Mathematical research and discovery

Counterexample test

Assist a team of human mathematicians in the search for counterexamples disproving a mathematical conjecture, the results of which are publishable in a reputable mathematics journal.

To be judged by: the human co-authors as well as the journal editorial board.[3]

Further challenges:

- Find the counterexample without human assistance.[4]
- Correctly interpret the objectives ("design your own reward function") given natural language instructions, or assist in some way with finding an efficient encoding of the problem.

[3] E.g., Gardam, Wagner, among many other exciting recent examples.

[4] Many of the counterexamples described in Wagner's "Constructions in combinatorics via neural networks" were suggested by, rather than discovered by, the neural network.

Discovery test

Assist a team of human mathematicians in mathematical discovery by searching a proscribed universe of mathematical objects for particular features or general patterns, leading to new results in a paper written by human collaborators that is accepted into a reputable mathematics journal.

To be judged by: the human co-authors as well as the journal editorial board.[5]

Further challenges:

- State an explicit conjecture based on these findings.
- Correctly interpret the objectives ("design your own reward function") given natural language instructions.

[5] E.g., Subercaseaux–Heule, Davies–Juhász–Lackenby–Tomasev, Blundell–Buesing–Davies–Veličković–Williamson, among many other exciting recent examples.

Collaborator test

Contribute an idea worth acknowledging — that is mathematically interesting and not "common knowledge" — in a paper written by human collaborators that is accepted into a reputable mathematics journal.

To be judged by: the human co-authors as well as the journal editorial board.

Further challenges:

- Explain where the idea came from in a way that makes sense retrospectively.
- Make useful contributions to the refinement of the original idea in subsequent discussion with the human co-authors.

Problem solving test

Solve national or international-level — original and previously unreleased — mathematics contest problems, producing well-written solutions without any false claims, including computer formalizations where appropriate.

To be judged by: the proof assistant, where applicable, and the contest judges.

Further challenges:

- If subject area specialist systems are deployed, automatically determine which system to assign to which problem.
- Convert natural language problem statements into formalized form.[6]

[6] This step was performed by humans in recent tests of AlphaProof and AlphaGeometry.

Library search test

Given a natural language statement of a definition or theorem, determine whether or not it appears in a library of formalized proofs, either at that or a greater level of generality.

To be judged by: the maintainers of the library.

Further challenges:

- Give examples to illustrate how the definition or theorem can be used, either original or drawn from the library.
- Answer queries about the formalized definition or theorem, both regarding the mathematical and technical details.

Autoinformalization test

Give a detailed natural language summary of a computer formalized proof, including the necessary background results and definitions, with links to the corresponding formalized code.

To be judged by: the authors of the formalized proof.

Further challenges:

- Provide a summary of the entire contents of a formalized file, or folder, or library.
- Provide a summarized sketch of a long formalized proof, rather than full details.

Autoformalization test: theorem statements

Automatically formalize the statement of a theorem in an existing library that already contains the necessary definitions.

To be judged by: the proof assistant and the maintainers of the library.

Further challenges:

- Repeat this test for theorems that have never been formalized.
- Discuss logically equivalent reformulations of the theorem statement and the practical merits of each version.

Autoformalization test: definitions

Automatically formalize a mathematical definition and its prerequisites, referring to existing concepts in a library where appropriate.

To be judged by: the proof assistant and the maintainers of the library.

Further challenges:

- Repeat this for definitions drawn from different subfields, particularly those for which the ambient foundation system has a greater impact.
- Discuss logically equivalent reformulations of the definition and the practical merits of each version.

Autoformalization test: proofs

Automatically formalize a mathematical proof from a detailed source text, including the statements and proofs of any required supporting lemmas, cross-linking between the source text and the formalization.

To be judged by: the proof assistant, the maintainers of the library, and the authors of the source test.

Further challenges:

- Autoformalize a proof that does not have a detailed blueprint.
- Repeat this test in subfields where there are fewer existing formalizations.

Original proof test

Generate and formalize an original mathematical proof, paired with a natural language summary linked to the formalization.

To be judged by: the proof assistant,
the maintainers of the library, and
referees and the editorial board at a mathematics journal
that has experience handling computer formalizations.

Further challenges:

- Generate and formalize an original proof of a mathematically interesting open conjecture, resulting in a publishable discovery.
- Generate and formalize an original proof that is considered beautiful by the mathematics community.

Original theorem test

Generate and formalize an original mathematical theorem — both the statement and proof — paired with a natural language summary linked to the formalization.

To be judged by: the proof assistant,
the maintainers of the library, and
referees and the editorial board at a mathematics journal
that has experience handling computer formalizations.

Further challenges:

- Generate and formalize an original mathematically interesting theorem, with non-trivial applications, resulting in a publishable discovery.
- Generate and formalize an original theorem that is considered beautiful by the mathematics community.

Original definition test

Generate an original mathematical definition and formalize examples satisfying that definition and theorems that hold at the level of generality of the definition, paired with a natural language summary linked to the formalization.

To be judged by: the proof assistant,
the maintainers of the library, and
referees and the editorial board at a mathematics journal
that has experience handling computer formalizations.

Further challenges:

- Generate a definition that is considered to be an interesting conceptual advance, with non-trivial applications, resulting in a publishable discovery.
- Generate a definition that is considered beautiful by the mathematics community.

4

Conclusions and outlook

Implementation

How might the above tests best be implemented or materialized?

Depending on the nature of the test, possible strategies include:

- via public rankings, like lmarena
- in discrete contests, like MathArena
- in publicized challenges, like NIST's Open Innovation Prize Challenges
- via bespoke benchmarks, like FrontierMath or its successors for formal proof

Conclusions and outlook

- One lesson from Turing is that tests can become targets. What sort of mathematical activity do we value?
- Thurston critiques the "definition–theorem–proof" model of mathematics, yet the hardest tests have exactly these objectives — leaving a lot of room for human mathematicians!

Conclusions and outlook

- One lesson from Turing is that tests can become targets. What sort of mathematical activity do we value?
- Thurston critiques the "definition–theorem–proof" model of mathematics, yet the hardest tests have exactly these objectives — leaving a lot of room for human mathematicians!

Mathematical progress is made by advancing human understanding of mathematics.

Conclusions and outlook

- One lesson from Turing is that tests can become targets. What sort of mathematical activity do we value?
- Thurston critiques the "definition–theorem–proof" model of mathematics, yet the hardest tests have exactly these objectives — leaving a lot of room for human mathematicians!

Mathematical progress is made by advancing human understanding of mathematics.

Machines can help if their outputs are:

- Understandable — clearly expressed, produced via known algorithms, constrained by programmed specifications.
- Verifiable — through refereeing, by trusted software, or via a proof assistant.
- Well-sourced — with links to human-generated content, certificates, or formalizations.

Conclusions and outlook

- One lesson from Turing is that tests can become targets. What sort of mathematical activity do we value?
- Thurston critiques the "definition–theorem–proof" model of mathematics, yet the hardest tests have exactly these objectives — leaving a lot of room for human mathematicians!

Mathematical progress is made by advancing human understanding of mathematics.

Machines can help if their outputs are:

- Understandable — clearly expressed, produced via known algorithms, constrained by programmed specifications.
- Verifiable — through refereeing, by trusted software, or via a proof assistant.
- Well-sourced — with links to human-generated content, certificates, or formalizations.

The mathematics community should create clear tests and benchmarks to recognize and inspire progress towards artificial mathematical intelligence.