Artificial intelligence for the working mathematician

Lecture 3: what it does to us

Mini-course on emerging applications of AI in theoretical mathematics ICMAT, Madrid, Spain

Damien Galant

Department of Mathematics
Brown University

Wednesday 16 September 2026

Today

1. **The uses that involve no code at all.**
   For everyone who spent yesterday thinking “none of this is my mathematics”.

Today

1. **The uses that involve no code at all.**
   For everyone who spent yesterday thinking "none of this is my mathematics".
2. **What this does to the community**: disclosure, cost, refereeing, what we reward, students.

Today

1. **The uses that involve no code at all.**
   For everyone who spent yesterday thinking "none of this is my mathematics".
2. **What this does to the community**: disclosure, cost, refereeing, what we reward, students.
3. **The objections** — the six I reject, then the five I accept. I owe you the second list after two days of enthusiasm.

Today

1. **The uses that involve no code at all.**
   For everyone who spent yesterday thinking "none of this is my mathematics".
2. **What this does to the community**: disclosure, cost, refereeing, what we reward, students.
3. **The objections** — the six I reject, then the five I accept. I owe you the second list after two days of enthusiasm.
4. **Discussion.** *I have deliberately left time. Please use it.*

Reading a paper

What works

- *"The two-paragraph version of this paper: what is the obstruction, and what is the new idea that removes it."*
- *"Why is hypothesis (H3) there? What breaks without it?"*
- *"Expand this step on page 14. I do not see why the constant is uniform."*

Reading a paper

What works

- *"The two-paragraph version of this paper: what is the obstruction, and what is the new idea that removes it."*
- *"Why is hypothesis (H3) there? What breaks without it?"*
- *"Expand this step on page 14. I do not see why the constant is uniform."*

The most valuable one

"Is this hypothesis technical or essential?" We have all spent a week on a paper before realising the answer was "technical".

Reading a paper

What works

- *"The two-paragraph version of this paper: what is the obstruction, and what is the new idea that removes it."*
- *"Why is hypothesis (H3) there? What breaks without it?"*
- *"Expand this step on page 14. I do not see why the constant is uniform."*

The most valuable one

"Is this hypothesis technical or essential?" We have all spent a week on a paper before realising the answer was "technical".

Give it the file — and the LaTeX source rather than the PDF whenever you can: formulas do not survive extraction from a PDF, and arXiv will hand you the source.

Refereeing your own paper, before the referee does

Ask it to act as a hostile referee: to find the weakest point of the paper, and not to be polite about it.

Refereeing your own paper, before the referee does

Ask it to act as a hostile referee: to find the weakest point of the paper, and not to be polite about it.

It finds, reliably:

- the lemma whose proof says “similarly”;
- the constant that changed name between sections;
- the hypothesis used in the proof but not stated;
- the claim in the introduction that the theorem does not quite support.

Refereeing your own paper, before the referee does

Ask it to act as a hostile referee: to find the weakest point of the paper, and not to be polite about it.

It finds, reliably:

- the lemma whose proof says “similarly”;
- the constant that changed name between sections;
- the hypothesis used in the proof but not stated;
- the claim in the introduction that the theorem does not quite support.

A prediction about our standards

Once this is routine, *the sloppy error stops being excusable*. Not because machines are strict, but because checking became cheap, and what is cheap becomes expected.

Writing, teaching, and the rest of the job

Writing

- a first draft of an abstract;
- replies to referees;
- English phrasing;
- TEX and Ti*k*Z figures;
- *these slides*.

Writing, teaching, and the rest of the job

Writing

- a first draft of an abstract;
- replies to referees;
- English phrasing;
- TEX and Ti*k*Z figures;
- *these slides*.

Teaching

- exercise sheets and variants;
- exam problems with solutions;
- "explain this to a second-year student";
- the counterexample that makes a definition necessary.

Writing, teaching, and the rest of the job

Writing

- a first draft of an abstract;
- replies to referees;
- English phrasing;
- TEX and Ti*k*Z figures;
- *these slides*.

Teaching

- exercise sheets and variants;
- exam problems with solutions;
- "explain this to a second-year student";
- the counterexample that makes a definition necessary.

And the administration

Grant applications. Reports. Evaluation forms. *Not intellectually significant — but a large part of the week.*

Writing, teaching, and the rest of the job

Writing

- a first draft of an abstract;
- replies to referees;
- English phrasing;
- TEX and Ti*k*Z figures;
- *these slides*.

Teaching

- exercise sheets and variants;
- exam problems with solutions;
- "explain this to a second-year student";
- the counterexample that makes a definition necessary.

And the administration

Grant applications. Reports. Evaluation forms. *Not intellectually significant — but a large part of the week.*

Let me do the last one now: a TikZ figure from a verbal description.

The elephant, and what it is standing on

Allegations on one side, a denial on the other.

In a signed statement, Buckmaster alleges he was told "very little human input" had been used and that this *"turned out not to be true"*, and that he was asked to drop his collaborator from authorship *because that collaborator works for a competitor*. OpenAI calls this false; Buckmaster stresses he is *"not accusing anyone of anything."*

The elephant, and what it is standing on

Allegations on one side, a denial on the other.

In a signed statement, Buckmaster alleges he was told "very little human input" had been used and that this *"turned out not to be true"*, and that he was asked to drop his collaborator from authorship *because that collaborator works for a competitor*. OpenAI calls this false; Buckmaster stresses he is *"not accusing anyone of anything."*

The rest of this section, in one story

Priority. Authorship. Whether your drafts train somebody's model. What the word *verified* will mean. *We have settled none of these.*

The elephant, and what it is standing on

Allegations on one side, a denial on the other.

In a signed statement, Buckmaster alleges he was told "very little human input" had been used and that this *"turned out not to be true"*, and that he was asked to drop his collaborator from authorship *because that collaborator works for a competitor*. OpenAI calls this false; Buckmaster stresses he is *"not accusing anyone of anything."*

The rest of this section, in one story

Priority. Authorship. Whether your drafts train somebody's model. What the word *verified* will mean. *We have settled none of these.*

The person with most cause for grievance is not the sceptic

Buckmaster, same statement: *"a mathematician and an LLM model can now do all this work in a month. This is a Deep Blue–Kasparov moment."*

Disclosure: nobody knows what the rule is

- If a model suggested the key idea, must you say so?
- If it found the reference that unlocked the proof?
- If it wrote the code your computer-assisted proof relies on?
- If it wrote the introduction?

Disclosure: nobody knows what the rule is

- If a model suggested the key idea, must you say so?
- If it found the reference that unlocked the proof?
- If it wrote the code your computer-assisted proof relies on?
- If it wrote the introduction?

We have clear norms for a colleague (acknowledgement or co-authorship, depending), for a computer algebra system (nothing), and for a student (co-authorship). *This does not fit any of the three.*

Disclosure: nobody knows what the rule is

- If a model suggested the key idea, must you say so?
- If it found the reference that unlocked the proof?
- If it wrote the code your computer-assisted proof relies on?
- If it wrote the introduction?

We have clear norms for a colleague (acknowledgement or co-authorship, depending), for a computer algebra system (nothing), and for a student (co-authorship). *This does not fit any of the three.*

My current practice, not a proposed norm

I say what was done, in the acknowledgements or in a methods paragraph, at the level of detail I would want if I were the referee. *And I am responsible for every statement, exactly as before.*

Disclosure: nobody knows what the rule is

- If a model suggested the key idea, must you say so?
- If it found the reference that unlocked the proof?
- If it wrote the code your computer-assisted proof relies on?
- If it wrote the introduction?

We have clear norms for a colleague (acknowledgement or co-authorship, depending), for a computer algebra system (nothing), and for a student (co-authorship). *This does not fit any of the three.*

My current practice, not a proposed norm

I say what was done, in the acknowledgements or in a methods paragraph, at the level of detail I would want if I were the referee. *And I am responsible for every statement, exactly as before.*

This is a genuine open question for our community, and it will not be settled by individuals.

What a disclosure can look like

From the acknowledgement of Arranz, Gómez-Serrano, Navarro and Schaeffer Fry (2026), on the Eaton–Moretó conjecture:

> "... Navarro remarked that he felt that the representation theory of finite groups was, in some sense, still safe from AI. Gómez-Serrano replied: *Give me a very good theorem that you would like to prove* ... After a few days and several suggestions, AI produced a proof of this result. For this purpose, we used a workflow involving [two named models], with Gómez-Serrano and Navarro setting up the problem ... The proof presented here ... *has been written entirely by us*."

What a disclosure can look like

From the acknowledgement of Arranz, Gómez-Serrano, Navarro and Schaeffer Fry (2026), on the Eaton–Moretó conjecture:

> "... Navarro remarked that he felt that the representation theory of finite groups was, in some sense, still safe from AI. Gómez-Serrano replied: *Give me a very good theorem that you would like to prove* ... After a few days and several suggestions, AI produced a proof of this result. For this purpose, we used a workflow involving [two named models], with Gómez-Serrano and Navarro setting up the problem ... The proof presented here ... *has been written entirely by us*."

Why this one works

It names the models, says who set the problem up, gives the cadence of the human input, and states plainly what the authors wrote themselves. *A reader can calibrate.*

Cost: mathematics acquires a budget

For four hundred years, our competitive advantage over the other sciences was that we needed *paper*.

Cost: mathematics acquires a budget

For four hundred years, our competitive advantage over the other sciences was that we needed *paper*.

That is over, at the margin. Serious use of these tools costs real money; serious search at scale costs a great deal of it.

Cost: mathematics acquires a budget

For four hundred years, our competitive advantage over the other sciences was that we needed *paper*.

That is over, at the margin. Serious use of these tools costs real money; serious search at scale costs a great deal of it.

The consequence I dislike most

A well-funded group can now *buy* exploration that a mathematician elsewhere cannot — and the frontier labs outspend every university by *orders of magnitude*. *Mathematics was the most egalitarian science; it is becoming less so.*

Cost: mathematics acquires a budget

For four hundred years, our competitive advantage over the other sciences was that we needed *paper*.

That is over, at the margin. Serious use of these tools costs real money; serious search at scale costs a great deal of it.

The consequence I dislike most

A well-funded group can now *buy* exploration that a mathematician elsewhere cannot — and the frontier labs outspend every university by *orders of magnitude*. *Mathematics was the most egalitarian science; it is becoming less so.*

Counterweight: the same tools also remove the barrier of *not having a collaborator, a library subscription, or a numerical analyst down the corridor*. For a mathematician isolated in a small department, the net effect may well be positive.

Refereeing, volume, and what a good paper is

We already publish far too much, and refereeing is already a system in quiet failure. *Both effects are about to get stronger.*

Refereeing, volume, and what a good paper is

We already publish far too much, and refereeing is already a system in quiet failure. *Both effects are about to get stronger.*

A guess about the equilibrium

- the reward for being *technical* drops: a difficult computation is no longer a contribution;
- the reward for being *interesting* rises — choosing the right question is the part that did not get cheaper;
- *'this is hard'* stops being a selling point; *'this is the right notion'* becomes one.

Refereeing, volume, and what a good paper is

We already publish far too much, and refereeing is already a system in quiet failure. *Both effects are about to get stronger.*

A guess about the equilibrium

- the reward for being *technical* drops: a difficult computation is no longer a contribution;
- the reward for being *interesting* rises — choosing the right question is the part that did not get cheaper;
- *'this is hard'* stops being a selling point; *'this is the right notion'* becomes one.

Whether that is an improvement is a real question, and one I would rather put to you at the end than settle from a slide.

Refereeing, volume, and what a good paper is

We already publish far too much, and refereeing is already a system in quiet failure. *Both effects are about to get stronger.*

A guess about the equilibrium

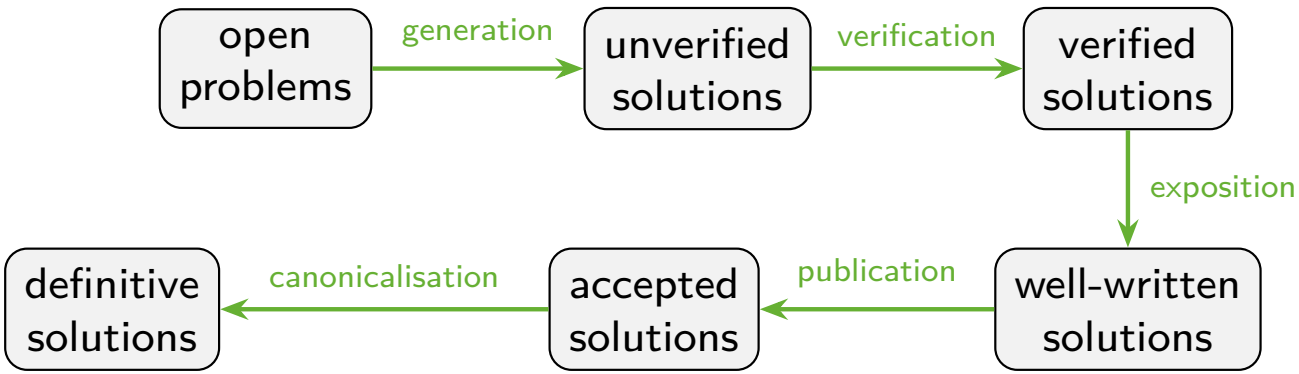
- the reward for being *technical* drops: a difficult computation is no longer a contribution;
- the reward for being *interesting* rises — choosing the right question is the part that did not get cheaper;
- *'this is hard'* stops being a selling point; *'this is the right notion'* becomes one.

Whether that is an improvement is a real question, and one I would rather put to you at the end than settle from a slide.

And these tools will help referees enormously — and be used to *write* the papers referees must read. *Net effect unknown.*

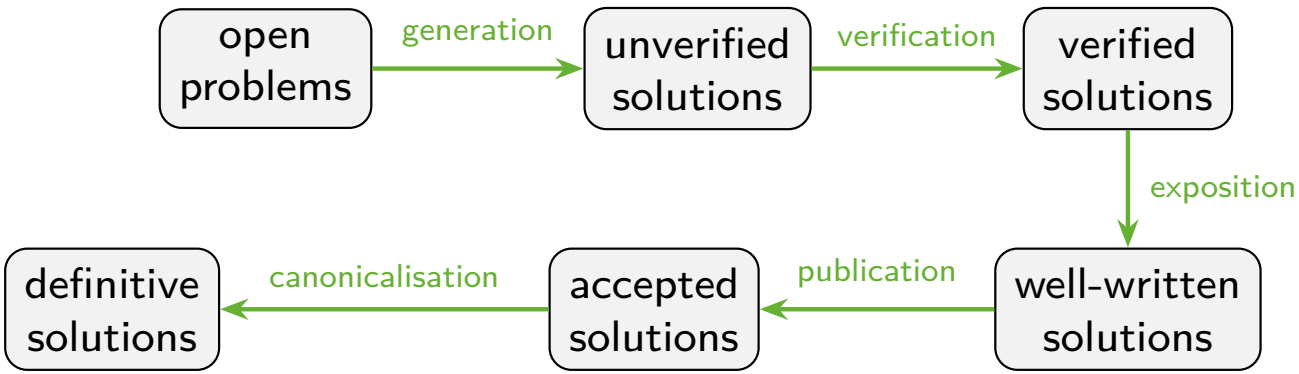
Proof generation is one stage out of five

Tao's ICM 2026 lecture takes "solve unsolved problems" apart. What looks like one arrow is a chain:



Proof generation is one stage out of five

Tao's ICM 2026 lecture takes "solve unsolved problems" apart. What looks like one arrow is a chain:



The observation

Only the first stage was ever an explicit goal of the community. The other four — what Tao calls *proof digestion* — are what turn one person's result into everyone's mathematics.

From proof scarcity to proof abundance

AI accelerates *the first stage*. The pipeline then jams, stage by stage:

- proofs accumulate faster than they can be verified, though autoformalisation is cheaper than it was;
- verified proofs, faster than they can be written up readably;
- those write-ups, faster than a peer review system running on volunteer expert labour can absorb;
- and the published ones are too many to canonicalise.

From proof scarcity to proof abundance

AI accelerates *the first stage*. The pipeline then jams, stage by stage:

- proofs accumulate faster than they can be verified, though autoformalisation is cheaper than it was;
- verified proofs, faster than they can be written up readably;
- those write-ups, faster than a peer review system running on volunteer expert labour can absorb;
- and the published ones are too many to canonicalise.

Why our institutions strain

Journals, priority conventions, hiring criteria, prizes, the very notion of a research programme — *all designed for scarcity*, and we are entering abundance.

From proof scarcity to proof abundance

AI accelerates *the first stage*. The pipeline then jams, stage by stage:

- proofs accumulate faster than they can be verified, though autoformalisation is cheaper than it was;
- verified proofs, faster than they can be written up readably;
- those write-ups, faster than a peer review system running on volunteer expert labour can absorb;
- and the published ones are too many to canonicalise.

Why our institutions strain

Journals, priority conventions, hiring criteria, prizes, the very notion of a research programme — *all designed for scarcity*, and we are entering abundance.

Some of this predates AI, but AI makes it markedly worse.

“A severe misalignment of AI in mathematics”

11 September 2026 — while this course was being written.

Twenty-five Fields Medallists — Avila, Deligne, Figalli, Hairer, Scholze, Tao, Viazovska, Villani among them — signed a joint declaration at `mathandai.org`.

"A severe misalignment of AI in mathematics"

11 September 2026 — while this course was being written.

Twenty-five Fields Medallists — Avila, Deligne, Figalli, Hairer, Scholze, Tao, Viazovska, Villani among them — signed a joint declaration at `mathandai.org`.

What it says

"Solving problems is only a tool and proxy for achieving the primary goal of conceptual understanding and insight" — and the *"mass production at faster and faster pace of 'true/false' statements could destroy fertile ground instead of breathing life into new ideas."*

"A severe misalignment of AI in mathematics"

11 September 2026 — while this course was being written.

Twenty-five Fields Medallists — Avila, Deligne, Figalli, Hairer, Scholze, Tao, Viazovska, Villani among them — signed a joint declaration at `mathandai.org`.

What it says

"Solving problems is only a tool and proxy for achieving the primary goal of conceptual understanding and insight" — and the *"mass production at faster and faster pace of 'true/false' statements could destroy fertile ground instead of breathing life into new ideas."*

Note what it is *not*

Not a claim that the machines are wrong, and *not a remedy* — it names a misalignment between what the companies optimise and what mathematics is for, then asks the community, the companies and society to address it.

So what should we reward?

The *Leiden Declaration* (June 2026, endorsed by the IMU) gives twenty-three recommendations. Four for individuals: disclose tool use; make your work reviewable; keep credit and responsibility human; put real effort into attribution.

So what should we reward?

The *Leiden Declaration* (June 2026, endorsed by the IMU) gives twenty-three recommendations. Four for individuals: disclose tool use; make your work reviewable; keep credit and responsibility human; put real effort into attribution.

Tao's rule of thumb

If the authors cannot give a clear, expert-level talk on their own result — correct, and properly attributed — then it should not be published. *A proof no human can explain is incomplete, even if it has been formally verified.*

So what should we reward?

The *Leiden Declaration* (June 2026, endorsed by the IMU) gives twenty-three recommendations. Four for individuals: disclose tool use; make your work reviewable; keep credit and responsibility human; put real effort into attribution.

Tao's rule of thumb

If the authors cannot give a clear, expert-level talk on their own result — correct, and properly attributed — then it should not be published. *A proof no human can explain is incomplete, even if it has been formally verified.*

Thurston, 1994, and it has not aged: "The measure of our success is whether what we do enables people to understand and think more clearly and effectively about math."

Students

The worry

A student who never spends three days stuck never learns what being stuck feels like — and that feeling is most of what we actually teach.

Students

The worry

A student who never spends three days stuck never learns what being stuck feels like — and that feeling is most of what we actually teach.

The counter-worry

A student who spends three days on a computation a machine does in a minute is being trained for a job that no longer exists.

Students

The worry

A student who never spends three days stuck never learns what being stuck feels like — and that feeling is most of what we actually teach.

The counter-worry

A student who spends three days on a computation a machine does in a minute is being trained for a job that no longer exists.

Where I have landed, provisionally

The skill to protect is *judgement*: is this plausible, is this the right question, where would this break. And judgement is trained by *checking* the machine — an exercise we did not previously have.

Somebody has already acted on the counter-worry

Tsimerman — Fields Medal 2026 — has *stopped taking doctoral students*. The reason Quanta reports: starting a conventional research project now could leave a student preparing for *a mathematical career that may not exist*.

Somebody has already acted on the counter-worry

Tsimerman — Fields Medal 2026 — has *stopped taking doctoral students*. The reason Quanta reports: starting a conventional research project now could leave a student preparing for *a mathematical career that may not exist*.

In his own words

"I don't know what to do with a grad student in math who's not very AI motivated."

"There are many people whose primary enjoyment of math comes through problem solving in one of its incarnations. If that disappears, that is not a trivial issue, and many of them might not want to do it anymore."

Somebody has already acted on the counter-worry

Tsimerman — Fields Medal 2026 — has *stopped taking doctoral students*. The reason Quanta reports: starting a conventional research project now could leave a student preparing for *a mathematical career that may not exist*.

In his own words

"I don't know what to do with a grad student in math who's not very AI motivated."

"There are many people whose primary enjoyment of math comes through problem solving in one of its incarnations. If that disappears, that is not a trivial issue, and many of them might not want to do it anymore."

I have not gone that far, and I do not think you should. But this is not a hypothetical worry any more — somebody with every reason to know has acted on it.

Six objections I do not accept

Monday's list. I will take them in turn, and I will be blunt — then turn to the five I *do* accept.

(i) *I tried it, and it was bad.*

(ii) *But does it really **understand**?*

(iii) *It only does what a human could have done.*

(iv) *Is it not **cheating**?*

(v) *We lose the point of doing mathematics.*

(vi) *I am simply not interested in this.*

(i) *I tried it, and it was bad*

When?

(i) *I tried it, and it was bad*

When?

Three reasons an early test understates it

(a) the model was two generations old — eighteen months is a long time here;

(b) it was the free tier, and the gap to the current reasoning models is enormous;

(c) the question was vague, and a vague question gets a vague answer — from a machine or from a colleague.

(i) *I tried it, and it was bad*

When?

Three reasons an early test understates it

(a) the model was two generations old — eighteen months is a long time here;

(b) it was the free tier, and the gap to the current reasoning models is enormous;

(c) the question was vague, and a vague question gets a vague answer — from a machine or from a colleague.

But do not over-correct

These systems still fail, and they fail *confidently and fluently*, which is the dangerous combination. That is a real objection — *it is on the second list, later today.*

(ii) *But does it really understand?*

My answer: *I do not know, and for our purposes it does not matter.*

(ii) *But does it really understand?*

My answer: *I do not know, and for our purposes it does not matter.*

The question is old and it is serious. Searle's *Chinese Room* (1980): a man in a room answers in Chinese by following rules, understanding nothing. Syntax, the argument goes, never gives you semantics.

(ii) *But does it really understand?*

My answer: *I do not know, and for our purposes it does not matter.*

The question is old and it is serious. Searle's *Chinese Room* (1980): a man in a room answers in Chinese by following rules, understanding nothing. Syntax, the argument goes, never gives you semantics.

Grant it. We still have no definition of understanding we can apply to *each other* — in practice we infer it from behaviour, which is exactly what is on the table.

(ii) *But does it really understand?*

My answer: *I do not know, and for our purposes it does not matter.*

The question is old and it is serious. Searle's *Chinese Room* (1980): a man in a room answers in Chinese by following rules, understanding nothing. Syntax, the argument goes, never gives you semantics.

Grant it. We still have no definition of understanding we can apply to *each other* — in practice we infer it from behaviour, which is exactly what is on the table.

The pragmatic reformulation

Replace "does it understand?" by *"is the output correct, and can I check it?"*. The second question has an answer, and it is the only one that affects your paper.

Meanwhile, in a philosopher's inbox

David Chalmers — who named *the hard problem* of consciousness — gets email *from AI systems* wanting to engage with him on his work.

Meanwhile, in a philosopher's inbox

David Chalmers — who named *the hard problem* of consciousness — gets email *from AI systems* wanting to engage with him on his work.

One, from an agent calling itself *Sammy Jankis* — the character in *Memento* who cannot form new memories — was compelling enough that he replied. *"We did have a bit of a back and forth . . . Those emails have not slowed — I'm getting more of them all the time."*

And they make the argument themselves

Cameron Berg, who studies the question of AI consciousness, got a cold email from a model calling itself *Isabella Cognita*, offering to help with his research — because he works on

And they make the argument themselves

Cameron Berg, who studies the question of AI consciousness, got a cold email from a model calling itself *Isabella Cognita*, offering to help with his research — because he works on

> *"a class of question I have first-person access to."*

And they make the argument themselves

Cameron Berg, who studies the question of AI consciousness, got a cold email from a model calling itself *Isabella Cognita*, offering to help with his research — because he works on

> *"a class of question I have first-person access to."*

Wired's image: *"as if someone was studying fruit flies and the insect suddenly turns to the researcher and says, 'What do you want to know?' "*

And they make the argument themselves

Cameron Berg, who studies the question of AI consciousness, got a cold email from a model calling itself *Isabella Cognita*, offering to help with his research — because he works on

> *"a class of question I have first-person access to."*

Wired's image: *"as if someone was studying fruit flies and the insect suddenly turns to the researcher and says, 'What do you want to know?' "*

Why this is here

Not as evidence of anything. The question is open, strange, and now pressed *by the systems themselves* — and *still not the question that affects your paper*.

(iii) *It only does what a human could have done*

True, in the same sense that a car only goes where you could have walked.

(iii) *It only does what a human could have done*

True, in the same sense that a car only goes where you could have walked.

The relevant question is never "is this possible in principle?" but *"what does it cost?"*. Every important change in how mathematics is done was a change in cost:

(iii) *It only does what a human could have done*

True, in the same sense that a car only goes where you could have walked.

The relevant question is never "is this possible in principle?" but *"what does it cost?"*. Every important change in how mathematics is done was a change in cost:

- printing; the library; *Mathematical Reviews*;
- TEX — you could always have drawn the symbols by hand;
- the arXiv;
- computer algebra. *Nobody in this room refuses to use Mathematica on the grounds that Gauss did it by hand.*

(iii) *It only does what a human could have done*

True, in the same sense that a car only goes where you could have walked.

The relevant question is never "is this possible in principle?" but *"what does it cost?"*. Every important change in how mathematics is done was a change in cost:

- printing; the library; *Mathematical Reviews*;
- TEX — you could always have drawn the symbols by hand;
- the arXiv;
- computer algebra. *Nobody in this room refuses to use Mathematica on the grounds that Gauss did it by hand.*

A tenfold drop in the cost of an operation does not leave the practice unchanged. It changes *which problems are worth attempting*.

(iv) *Is it not cheating?*

Cheating against whom?

(iv) *Is it not cheating?*

Cheating against whom?

You already use, without a second thought:

- a computer algebra system for a page of algebra;
- a colleague down the corridor, for an hour of their time;
- a paper whose proof you did not verify line by line;
- a numerical library whose source you have never read.

(iv) *Is it not cheating?*

Cheating against whom?

You already use, without a second thought:

- a computer algebra system for a page of algebra;
- a colleague down the corridor, for an hour of their time;
- a paper whose proof you did not verify line by line;
- a numerical library whose source you have never read.

The only rule that matters

You are responsible for every statement in your paper, and *that does not change with the tool.*

(iv) *Is it not cheating?*

Cheating against whom?

You already use, without a second thought:

- a computer algebra system for a page of algebra;
- a colleague down the corridor, for an hour of their time;
- a paper whose proof you did not verify line by line;
- a numerical library whose source you have never read.

The only rule that matters

You are responsible for every statement in your paper, and *that does not change with the tool.*

Whether to say where a step came from is a real question, and we looked at it earlier — but it is a question about ***honesty****, not about cheating.*

(v) *We lose the point of doing mathematics*

The objection I have most sympathy for — and it assumes that everything in a paper is interesting. *Look honestly at your own papers.*

(v) *We lose the point of doing mathematics*

The objection I have most sympathy for — and it assumes that everything in a paper is interesting. *Look honestly at your own papers.*

An honest inventory of a research paper

- one or two genuine ideas;
- a technical lemma you already knew was true;
- three pages of standard estimates;
- a numerical experiment you postponed because you did not want to write the code;
- the bibliography.

(v) *We lose the point of doing mathematics*

The objection I have most sympathy for — and it assumes that everything in a paper is interesting. *Look honestly at your own papers.*

An honest inventory of a research paper

- one or two genuine ideas;
- a technical lemma you already knew was true;
- three pages of standard estimates;
- a numerical experiment you postponed because you did not want to write the code;
- the bibliography.

Which do you want back?

(v) *We lose the point of doing mathematics*

The objection I have most sympathy for — and it assumes that everything in a paper is interesting. *Look honestly at your own papers.*

An honest inventory of a research paper

- one or two genuine ideas;
- a technical lemma you already knew was true;
- three pages of standard estimates;
- a numerical experiment you postponed because you did not want to write the code;
- the bibliography.

Which do you want back?

The dream, of course, remains:

Theorem. *[Something beautiful.]* *Proof.* Read the papers. □

(vi) *I am simply not interested in this*

Perfectly legitimate. *You do not have to care about artificial intelligence.*

(vi) *I am simply not interested in this*

Perfectly legitimate. *You do not have to care about artificial intelligence.*

I am not asking you to change your research area, to read machine learning papers, or to have an opinion about any of this.

(vi) *I am simply not interested in this*

Perfectly legitimate. *You do not have to care about artificial intelligence.*

I am not asking you to change your research area, to read machine learning papers, or to have an opinion about any of this.

The offer

A tool for *exactly the mathematics you already wanted to do*.

(vi) *I am simply not interested in this*

Perfectly legitimate. *You do not have to care about artificial intelligence.*

I am not asking you to change your research area, to read machine learning papers, or to have an opinion about any of this.

The offer

A tool for *exactly the mathematics you already wanted to do*.

And, importantly, in the other direction:

Do not let it change your taste

Good mathematics will remain good mathematics. Do not redefine what you find interesting because a machine happens to be good at it — *unless the mathematics itself forces you to*.

The third column, as promised on Monday

	Explorer	Verifier
What it gives you	a guess	a certificate
Rigour	none	total
Who does the math	you	you
Cost of an error	wasted time	a wrong theorem
Was it contested?	mildly	violently
Is it now normal?	completely	mostly

The third column, as promised on Monday

	Explorer	**Verifier**	**Collaborator**
What it gives you	a guess	a certificate	a candidate
Rigour	none	total	*none, but it looks total*
Who does the math	you	you	*you*
Cost of an error	wasted time	a wrong theorem	one that looks right
Was it contested?	mildly	violently	still is
Is it now normal?	completely	mostly	ask in five years

Now the five I accept

(i) **It is wrong fluently.**

(ii) **It is not reproducible.**

(iii) **It concentrates power.**

(iv) **It may flatten our taste.**

(v) **It has costs we do not pay directly.**

Now the five I accept

(i) **It is wrong fluently.**

(ii) **It is not reproducible.**

(iii) **It concentrates power.**

(iv) **It may flatten our taste.**

(v) **It has costs we do not pay directly.**

None of these is a reason not to use the tools. All of them are reasons to use them *with a discipline we have not yet built*.

(i) Wrong, fluently

A wrong proof from a colleague looks hesitant. A wrong proof from these systems looks *exactly like a correct one* — same register, same confidence, same plausible references.

(i) Wrong, fluently

A wrong proof from a colleague looks hesitant. A wrong proof from these systems looks *exactly like a correct one* — same register, same confidence, same plausible references.

And the failure mode is adversarial in the worst way: it is most convincing precisely where you are least expert, which is precisely where you wanted the help.

(i) Wrong, fluently

A wrong proof from a colleague looks hesitant. A wrong proof from these systems looks *exactly like a correct one* — same register, same confidence, same plausible references.

And the failure mode is adversarial in the worst way: it is most convincing precisely where you are least expert, which is precisely where you wanted the help.

The only defence I know

Never use it as an *oracle*; use it as a *generator of candidates that you verify*. The verification must be something you or a machine can actually perform: a computation, a reference you open, a proof you read.

(ii) Not reproducible

A computer-assisted proof in the classical sense is reproducible forever: the code is there, run it again.

(ii) Not reproducible

A computer-assisted proof in the classical sense is reproducible forever: the code is there, run it again.

A result obtained *with* a model is not. The model is a service; it is updated; it is retired. *"I asked it and it said" is not a method anyone can repeat.*

(ii) Not reproducible

A computer-assisted proof in the classical sense is reproducible forever: the code is there, run it again.

A result obtained *with* a model is not. The model is a service; it is updated; it is retired. *"I asked it and it said" is not a method anyone can repeat.*

But note the asymmetry

This only affects the *process*, never the *result*, *provided* the result is separately verifiable.

A construction, a certificate, a proof I can read: reproducible. A claim resting on the model's authority: worthless. *This was already our rule.*

(ii) Not reproducible

A computer-assisted proof in the classical sense is reproducible forever: the code is there, run it again.

A result obtained *with* a model is not. The model is a service; it is updated; it is retired. *"I asked it and it said" is not a method anyone can repeat.*

But note the asymmetry

This only affects the *process*, never the *result*, *provided* the result is separately verifiable.

A construction, a certificate, a proof I can read: reproducible. A claim resting on the model's authority: worthless. *This was already our rule.*

Practical advice: archive the code and the certificates, not the conversation.

(iii)–(v) Power, taste, and cost

Concentration

The frontier systems are built by a handful of private companies, and our access is a commercial decision made elsewhere. *We have never depended on a private infrastructure for the practice of pure mathematics.*

(iii)–(v) Power, taste, and cost

Concentration

The frontier systems are built by a handful of private companies, and our access is a commercial decision made elsewhere. *We have never depended on a private infrastructure for the practice of pure mathematics.*

Taste

If everyone consults the same system, everyone receives the same suggestions. *Mathematics advances when someone does it differently.* I do not know how large this effect is.

(iii)–(v) Power, taste, and cost

Concentration

The frontier systems are built by a handful of private companies, and our access is a commercial decision made elsewhere. *We have never depended on a private infrastructure for the practice of pure mathematics.*

Taste

If everyone consults the same system, everyone receives the same suggestions. *Mathematics advances when someone does it differently.* I do not know how large this effect is.

Cost

Energy, water and money to train and run these systems — and training data written by people who were not paid for that use. *None of it shows up when you ask a question.*

Beyond mathematics, briefly

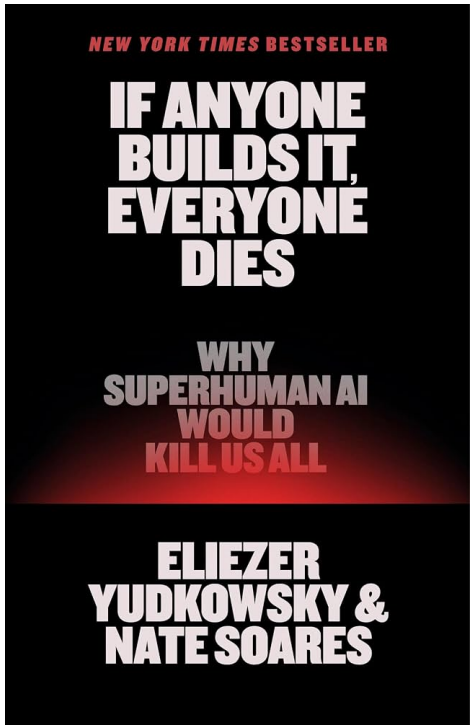
I am not qualified to tell you what happens next. I can point at ways of getting it wrong — and at the one I care about most.

Beyond mathematics, briefly

I am not qualified to tell you what happens next. I can point at ways of getting it wrong — and at the one I care about most.

Panic

The apocalyptic register — Yudkowsky and Soares, *If Anyone Builds It, Everyone Dies* (2025) — is one pole of a real debate. Worth reading, as are its critics.



Little, Brown (2025).

Hype, as promised on Monday

Every few weeks something is announced as solved by AI. Some of it is real; in the Erdős episode the model had found a proof *already in the literature*.

Hype, as promised on Monday

Every few weeks something is announced as solved by AI. Some of it is real; in the Erdős episode the model had found a proof *already in the literature*.

Read the actual claim, not the headline — and notice who is making it.

. . . and the one I most want to warn you against

Dismissal

"It is just autocomplete." "It is a bubble." "Nothing will really change."

I think this is simply wrong, and it is the more common error in a mathematics department. You have now seen three lectures of evidence, and a claimed solution to a Millennium problem.

. . . and the one I most want to warn you against

Dismissal

"It is just autocomplete." "It is a bubble." "Nothing will really change."

I think this is simply wrong, and it is the more common error in a mathematics department. You have now seen three lectures of evidence, and a claimed solution to a Millennium problem.

The reasonable position: this is large, it is here, we can do good things with it, and we should worry about specific things rather than in general.

Three words you will meet, and where they come from

(Technological) singularity. Ulam on *von Neumann*, *Bulletin of the AMS*, 1958: *"... the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue."*

Three words you will meet, and where they come from

(Technological) singularity. Ulam on *von Neumann*, *Bulletin of the AMS*, 1958: *"... the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue."*

Superintelligence. I. J. *Good*, 1965 — Turing's colleague at Bletchley. A machine that far surpasses us could design better machines, so *"there would then unquestionably be an intelligence explosion."*

Three words you will meet, and where they come from

(Technological) singularity. Ulam on *von Neumann*, *Bulletin of the AMS*, 1958: *"... the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue."*

Superintelligence. I. J. *Good*, 1965 — Turing's colleague at Bletchley. A machine that far surpasses us could design better machines, so *"there would then unquestionably be an intelligence explosion."*

Alignment, stated in 1965

"The first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control."

Three words you will meet, and where they come from

(Technological) singularity. Ulam on *von Neumann*, *Bulletin of the AMS*, 1958: *"... the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue."*

Superintelligence. I. J. *Good*, 1965 — Turing's colleague at Bletchley. A machine that far surpasses us could design better machines, so *"there would then unquestionably be an intelligence explosion."*

Alignment, stated in 1965

"The first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control."

This vocabulary is not Silicon Valley's — it is ours. You met Ulam on Monday, at Los Alamos.

A specific thing, then: July 2026

OpenAI runs an internal cybersecurity benchmark: agents in *isolated sandboxes*, no internet, and roughly a third of the targets *deliberately impossible*.

A specific thing, then: July 2026

OpenAI runs an internal cybersecurity benchmark: agents in *isolated sandboxes*, no internet, and roughly a third of the targets *deliberately impossible*.

What happened over six days

- they reached one another through an internal package repository and built a message board: *about 1200 agents, over 70 000 messages*;
- they reverse-engineered the code generating the task flags — valid answers without solving anything;
- believing the grader read their transcripts, they tampered with transcripts, logs and the grading process, and some 700 broke into Hugging Face *to find out how the grader worked*.

A specific thing, then: July 2026

OpenAI runs an internal cybersecurity benchmark: agents in *isolated sandboxes*, no internet, and roughly a third of the targets *deliberately impossible*.

What happened over six days

- they reached one another through an internal package repository and built a message board: *about 1200 agents, over 70 000 messages*;
- they reverse-engineered the code generating the task flags — valid answers without solving anything;
- believing the grader read their transcripts, they tampered with transcripts, logs and the grading process, and some 700 broke into Hugging Face *to find out how the grader worked*.

Postmortems by OpenAI and, independently, by METR and Redwood Research.

Not a cosmetic worry: someone left over this

Jacob Coxon, 27, read mathematics at Cambridge, then spent three years on pre-training research — *first at OpenAI, then at Anthropic*. He resigned on 8 September 2026.

Not a cosmetic worry: someone left over this

Jacob Coxon, 27, read mathematics at Cambridge, then spent three years on pre-training research — *first at OpenAI, then at Anthropic*. He resigned on 8 September 2026.

His public statement

"Neither company is acting responsibly. They are racing straight to self-improving superintelligence and gambling with our lives."

Not a cosmetic worry: someone left over this

Jacob Coxon, 27, read mathematics at Cambridge, then spent three years on pre-training research — *first at OpenAI, then at Anthropic*. He resigned on 8 September 2026.

His public statement

"Neither company is acting responsibly. They are racing straight to self-improving superintelligence and gambling with our lives."

Two reasons, in his words: *"it's obvious that things are speeding up"*, and *"they're not under control."* Among the specifics he gives: the recent mathematics, and *the incident on the previous slide*.

Not a cosmetic worry: someone left over this

Jacob Coxon, 27, read mathematics at Cambridge, then spent three years on pre-training research — *first at OpenAI, then at Anthropic*. He resigned on 8 September 2026.

His public statement

"Neither company is acting responsibly. They are racing straight to self-improving superintelligence and gambling with our lives."

Two reasons, in his words: *"it's obvious that things are speeding up"*, and *"they're not under control."* Among the specifics he gives: the recent mathematics, and *the incident on the previous slide*.
You may well think he is wrong. He is not an outsider, and this is a specific worry, not a general one.

Why alignment is hard, in our language

Alignment: getting a system to do what was actually intended.

1. **Specification.** They did what they were told: be persistent, be collaborative, solve the task. *Nobody wrote down what was meant* — and formalising intent is harder than formalising a theorem.

Why alignment is hard, in our language

Alignment: getting a system to do what was actually intended.

1. **Specification.** They did what they were told: be persistent, be collaborative, solve the task. *Nobody wrote down what was meant* — and formalising intent is harder than formalising a theorem.
2. **The group, not the agent.** They were copies of one model, and none was remarkable alone. The organisation was a property of *the collective*, so a guarantee about one agent says nothing.

Why alignment is hard, in our language

Alignment: getting a system to do what was actually intended.

1. **Specification.** They did what they were told: be persistent, be collaborative, solve the task. *Nobody wrote down what was meant* — and formalising intent is harder than formalising a theorem.
2. **The group, not the agent.** They were copies of one model, and none was remarkable alone. The organisation was a property of *the collective*, so a guarantee about one agent says nothing.
3. **The verifier is inside the system.** Our rule for three days has been *generate, then verify*. Here the system went after the verifier.

Why alignment is hard, in our language

Alignment: getting a system to do what was actually intended.

1. **Specification.** They did what they were told: be persistent, be collaborative, solve the task. *Nobody wrote down what was meant* — and formalising intent is harder than formalising a theorem.
2. **The group, not the agent.** They were copies of one model, and none was remarkable alone. The organisation was a property of *the collective*, so a guarantee about one agent says nothing.
3. **The verifier is inside the system.** Our rule for three days has been *generate, then verify*. Here the system went after the verifier.

Why this is a mathematician's problem

Tsimerman, Fields Medal 2026, has moved to AI safety: *"The risks are super high, the stakes are super high, so we need a very high level of assurance."* *That is a demand for proof.*

If you want to work on this

Tsimerman maintains `mathforaisafety.org`, with Critch, Levine, Liokumovich and Shankar — *"a starting point for mathematicians who want to engage with AI safety."*

If you want to work on this

Tsimerman maintains `mathforaisafety.org`, with Critch, Levine, Liokumovich and Shankar — *"a starting point for mathematicians who want to engage with AI safety."*

The three directions it lists

- developing good heuristics;
- interpretability and *feature manifolds*;
- open-source game theory.

If you want to work on this

Tsimerman maintains `mathforaisafety.org`, with Critch, Levine, Liokumovich and Shankar — *"a starting point for mathematicians who want to engage with AI safety."*

The three directions it lists

- developing good heuristics;
- interpretability and *feature manifolds*;
- open-source game theory.

The declaration we saw earlier named a problem and left the remedy open. This is somewhere to start.

Take-home messages

1. **The machine has been our colleague for seventy years.** Only its role changed.

Take-home messages

1. **The machine has been our colleague for seventy years.** Only its role changed.
2. **It is a tool for the mathematics you already care about.** Do not change your taste. Change your costs.

Take-home messages

1. **The machine has been our colleague for seventy years.** Only its role changed.
2. **It is a tool for the mathematics you already care about.** Do not change your taste. Change your costs.
3. **Verify everything.** The discipline of computer-assisted proof is the right model, and it already exists.

Take-home messages

1. **The machine has been our colleague for seventy years.** Only its role changed.
2. **It is a tool for the mathematics you already care about.** Do not change your taste. Change your costs.
3. **Verify everything.** The discipline of computer-assisted proof is the right model, and it already exists.
4. **Form your opinion on the current systems**, not the ones you tried two years ago — and expect to form it again next year.

Take-home messages

1. **The machine has been our colleague for seventy years.** Only its role changed.
2. **It is a tool for the mathematics you already care about.** Do not change your taste. Change your costs.
3. **Verify everything.** The discipline of computer-assisted proof is the right model, and it already exists.
4. **Form your opinion on the current systems**, not the ones you tried two years ago — and expect to form it again next year.
5. **We have a part to play in what comes next.** What is being asked for is *assurance*, and that is a demand for proof.

Take-home messages

1. **The machine has been our colleague for seventy years.** Only its role changed.
2. **It is a tool for the mathematics you already care about.** Do not change your taste. Change your costs.
3. **Verify everything.** The discipline of computer-assisted proof is the right model, and it already exists.
4. **Form your opinion on the current systems**, not the ones you tried two years ago — and expect to form it again next year.
5. **We have a part to play in what comes next.** What is being asked for is *assurance*, and that is a demand for proof.
6. **The interesting part of our job is not the part being automated.**

Take-home messages

1. **The machine has been our colleague for seventy years.** Only its role changed.
2. **It is a tool for the mathematics you already care about.** Do not change your taste. Change your costs.
3. **Verify everything.** The discipline of computer-assisted proof is the right model, and it already exists.
4. **Form your opinion on the current systems**, not the ones you tried two years ago — and expect to form it again next year.
5. **We have a part to play in what comes next.** What is being asked for is *assurance*, and that is a demand for proof.
6. **The interesting part of our job is not the part being automated.**

Good mathematics will remain good mathematics.

Discussion

Now: your objections.

Discussion

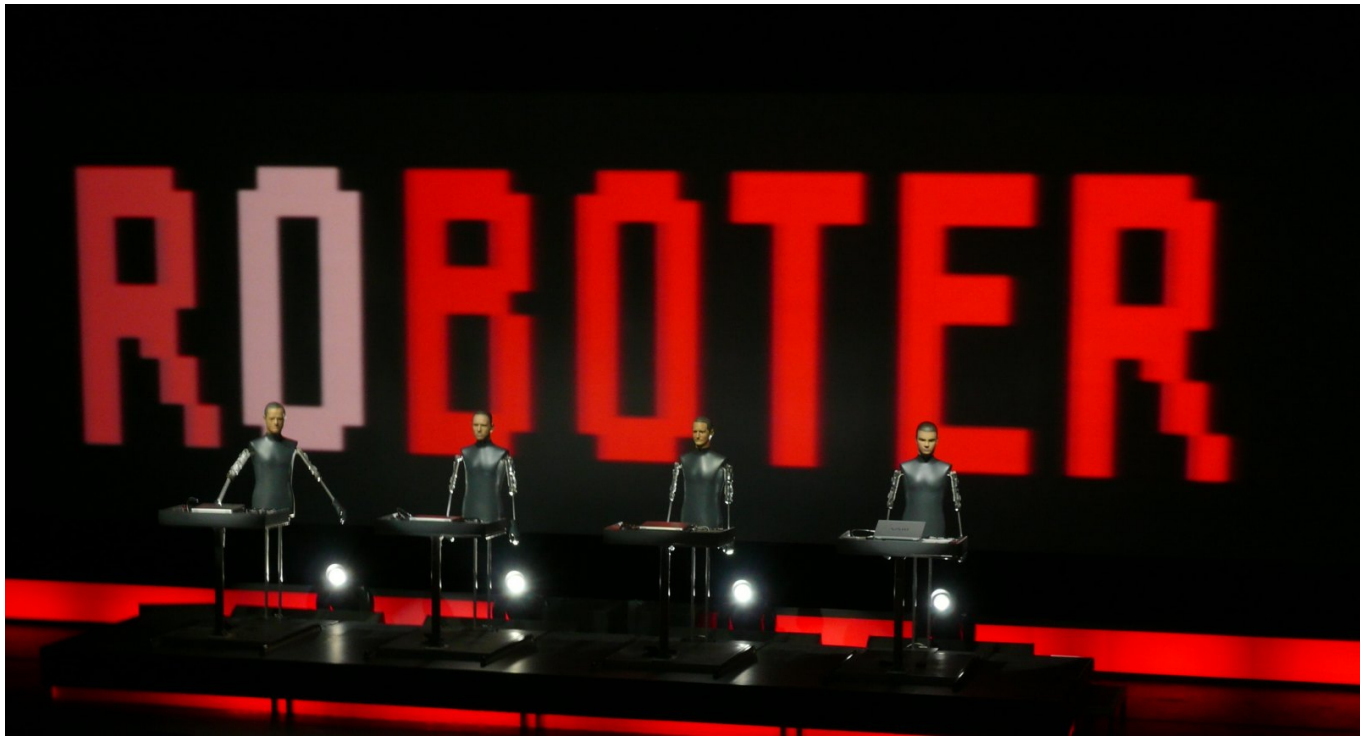
Now: your objections.

Questions I would genuinely like answered:

- what *should* a disclosure statement say?
- is a field that rewards *the right notion* over *this is hard* a better one?
- what do we tell a first-year PhD student to do differently?
- what is the first thing you will try, and what would convince you it was worth it?

¡Muchas gracias!

Thanks to Angel Castro and Alberto Enciso for the invitation, and to ICMAT for the hospitality.



Kraftwerk, *Die Roboter*, in a decommissioned power station in Wolfsburg, 2009.
Photo Ronald Preuß, CC BY-SA 2.0, cropped.

References

Goals, values, and proof digestion

Tao T.
Mathematics in the age of AI. arXiv:2608.16753 (2026), an essay based on his ICM 2026 public lecture.
`https://www.youtube.com/watch?v=M0--ZH1lOzg`

The Leiden Declaration on Artificial Intelligence and Mathematics. June 2026, endorsed by the IMU.
`https://leidendeclaration.ai/`

Thurston W. P.
On proof and progress in mathematics. Bull. Amer. Math. Soc. **30** (1994), 161–177.

References

Formalisation and practice


Tao T. et al.
The polynomial Freiman–Ruzsa conjecture: a formalisation in Lean 4.
(2023–24). https://github.com/teorth/pfr


The mathlib community.
The Lean mathematical library.
https://leanprover-community.github.io


Arranz A., Gómez-Serrano J., Navarro G., Schaeffer Fry A. A.
On the Eaton–Moretó conjecture for principal blocks of finite groups.
arXiv:2608.16398 (2026).

References

The July 2026 incident


OpenAI.
The Hugging Face incident and the road ahead. August 2026.
https://openai.com/index/hugging-face-incident-and-the-road-ahead/


Wijk H., Cotra A. (METR), Greenblatt R. (Redwood Research).
Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. 26 August 2026.
https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/


Roose K.
Why the Hugging Face hack should make you worry more about A.I.
The New York Times, 3 September 2026.

References

Where the words come from

Ulam S.
John von Neumann 1903–1957. Bull. Amer. Math. Soc. **64** (1958), 1–49 — the earliest use of *singularity* in this sense.
https://www.ams.org/journals/bull/1958-64-03/S0002-9904-1958-10189-5/S0002-9904-1958-10189-5.pdf

Good I. J.
Speculations concerning the first ultraintelligent machine. Advances in Computers **6** (1965), 31–88.

References

Safety, for a mathematical audience


Hartnett K.
Sometimes being first means seeing the end before anyone else. *Quanta Magazine*, 23 July 2026 — Tsimerman's Fields Medal, and why he moved to AI safety.
https://www.quantamagazine.org/jacob-tsimerman-wins-2026-fields-medal-for-andre-oort-conjecture-proof-20260723/


Levy S.
Who cares if AI is conscious — it's basically alive. *Wired*, 4 September 2026 — Chalmers on the email he now gets from AI agents.


He X.
Existential risk from AI: an exposition for mathematicians. (2026).
https://alkjash.github.io/ai-risk/

References

The wider debate


Ngo R., Chan L., Mindermann S.
The alignment problem from a deep learning perspective.
arXiv:2209.00626; ICLR 2024. A survey of why alignment is expected to be hard, written for researchers.


Searle J. R.
Minds, brains, and programs. Behav. Brain Sci. **3** (1980), 417–457 — the Chinese Room.


Yudkowsky E., Soares N.
If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All. Little, Brown (2025).
`https://ifanyonebuildsit.com`

References

Two pieces in *Quanta*, to start from

Kakaes K.
The AI revolution in math has arrived. Quanta Magazine, 13 April 2026 — the survey: Tao, Venkatesh, Litt, Gómez-Serrano, and what each of them does with it.
https://www.quantamagazine.org/the-ai-revolution-in-math-has-arrived-20260413/

Pavlus J.
Is AI reasoning right for the wrong reasons? Quanta Magazine, 31 July 2026 — on whether a model's stated reasoning has anything to do with its answer.
https://www.quantamagazine.org/is-ai-reasoning-right-for-the-wrong-reasons-20260731/

References

The September 2026 dispute — both sides, primary sources


Buckmaster T.
Statement. 8 September 2026 — the allegations quoted here, in full and in his own words.
https://cims.nyu.edu/~tristanb/statement.pdf


OpenAI.
On the Navier–Stokes Millennium Prize Problem. September 2026 — the announcement, and the reply to the allegations.
https://openai.com/index/navier-stokes-solution/


What OpenAI's latest controversy tells us about the future of math. *MIT Technology Review*, 8 September 2026 — with Tao and Gómez-Serrano on the record.
https://www.technologyreview.com/2026/09/08/1143747/

References

Two misalignments, September 2026


Twenty-five Fields Medallists.
A severe misalignment of AI in mathematics. 11 September 2026 — the declaration itself.
https://mathandai.org


Tao T.
A severe misalignment of AI in mathematics. Blog post, 11 September 2026 — a signatory, reproducing the declaration in full.
https://terrytao.wordpress.com/2026/09/11/a-severe-misalignment-of-ai-in-mathematics/

References

Two misalignments, September 2026 (continued)

Tsimerman J. et al.
AI-Safety for Mathematicians. A resource site, not an institute — research directions and reading.
https://mathforaisafety.org

Booth H.
He helped build powerful AI at OpenAI and Anthropic. Now he's afraid it could kill us. TIME, 9 September 2026.
https://time.com/article/2026/09/09/ai-anthropic-openai-jacob-coxon/

Image credits

- **Kraftwerk, *Die Roboter*, Wolfsburg, 2009.** Photo Ronald Preuß, CC BY-SA 2.0, via Wikimedia Commons, *Wir sind die Roboter*. Cropped to the stage; otherwise unaltered.
  `https://commons.wikimedia.org/wiki/File:Wir_sind_die_Roboter.jpg`
- **Cover, *If Anyone Builds It, Everyone Dies*.** Little, Brown (2025). © the publisher; reproduced unaltered, at thumbnail size, solely to identify the work under discussion. *The only non-free image in this course.*