

Conversation Transcript:

AI-assisted science needs a collective verification network

Conversation concerning a possible contribution to the Cell “AI and Collective Intelligence” competition: <https://www.cellcomm.org/ai-and-collective-intelligence>.

AI system used

Provider: OpenAI

Interface: ChatGPT

Model: GPT-5.6 Sol (reasoning model)

Conversation date: 2 September 2026

Editing note: User prompts have been lightly corrected for typographical and grammatical errors. The substantive content of the conversation has not been changed. Process diagrams in the assistant responses have been reformatted vertically.

Conversation

User

Consider the prize described in <https://www.cellcomm.org/ai-and-collective-intelligence>

I want to describe how AI can be integrated in science. I believe that one change would be that some of the tasks previously performed by scientists will now be carried out automatically, such as programming or performing cumbersome computations. What do you think?

ChatGPT

Yes. I think there is a strong idea there, but I would formulate the transformation at a higher level than simply “AI will do programming and cumbersome computations.”

The competition explicitly allows proposals about **how scientific work and collective intelligence should be reorganized in the age of AI**, not only proposals concerning a specific biomedical problem. The guidelines say that submissions may address “social and information-exchange aspects of human collective intelligence” and even invite reflections on how to preserve individual creativity while organizing a broad collective effort. So your proposed direction fits the scope.

The central claim I would develop is something like:

AI changes the optimal division of labor in science.

Historically, a scientist spends a substantial fraction of their time on tasks that are necessary for research but are not necessarily where their distinctive scientific contribution lies: writing and debugging code, deriving long algebraic manipulations, running numerical analyses, searching literature, translating equations into implementations, preparing datasets, checking parameter spaces, producing plots, and so on.

AI increasingly makes some of those tasks cheap and scalable.

That has a much deeper consequence than “science becomes faster.” It potentially changes

what a scientist's role is.

You could distinguish three levels.

1. Execution becomes increasingly automated.

Examples are exactly the ones you mentioned:

- programming;
- symbolic calculations;
- numerical computations;
- data processing;
- literature retrieval and preliminary synthesis;
- generation and testing of alternative implementations;
- repetitive simulation/analysis workflows.

The scientist would specify the objective, constraints and scientific assumptions, while an AI system performs a substantial part of the implementation.

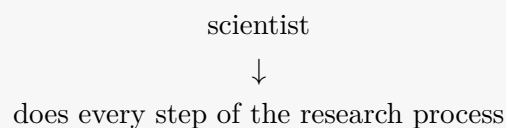
But I would not stop there.

2. Human scientific effort moves upward in abstraction.

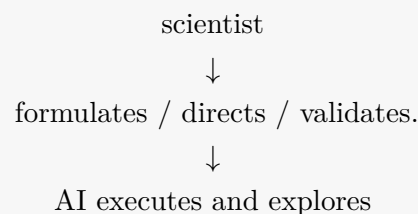
If implementation becomes cheap, the scarce resource becomes things such as:

- deciding **which questions are worth asking**;
- formulating the right problem;
- recognizing which assumptions are questionable;
- deciding which evidence would actually discriminate between hypotheses;
- identifying surprising results;
- connecting apparently unrelated problems;
- assessing whether an AI-produced result is scientifically meaningful;
- deciding when existing paradigms themselves should be challenged.

So instead of



you increasingly have something like



And this connects especially well with the premise of this prize. Their stated long-term question is precisely how **human collective intelligence might orchestrate powerful AI systems**, potentially exploiting the diversity of ideas present across the scientific community.

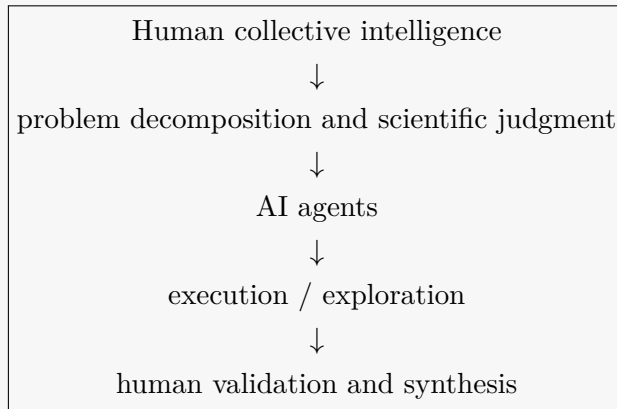
3. The really interesting consequence is collective rather than individual.

This is where I think your proposal could become distinctive.

Today, the size of a scientific question that a researcher can realistically attack is constrained by execution cost. I might have an idea, but pursuing it could require six months of code development, knowledge of another discipline, an enormous literature search, or thousands of calculations.

With AI, that constraint becomes weaker.

Consequently, scientists could operate less like isolated producers of complete research projects and more like nodes in a **human–AI scientific network**:



A biologist could formulate a question requiring expertise in statistical mechanics; a physicist could contribute a modeling idea without personally becoming an expert programmer in the relevant biological infrastructure; an AI system could translate between these contributions, implement candidate models, identify missing information and connect researchers whose work is complementary.

That is extremely close to one of the questions the organizers themselves highlight: “How could AI surface connections between findings that no single researcher would notice?”

Where I would push your idea

I would **not** make the thesis:

AI should replace scientists in programming and calculations.

That is true but relatively obvious, and probably not sufficient for a competitive one-page proposal.

I would make it:

AI can remove execution as one of the principal bottlenecks of scientific research. Scientific institutions should therefore reorganize around the resources that remain scarce: human judgment, creativity, diversity of hypotheses, experimental knowledge, problem selection and validation.

Then there is an interesting second claim:

The unit of scientific production may consequently change from the individual researcher/laboratory performing an end-to-end project to a distributed network in which humans formulate, criticize and connect ideas while AI performs much of the intermediate execution.

And now you have a genuine proposal about **collective intelligence**, rather than simply a prediction about AI-assisted coding.

There is also an important complication that would make your argument much stronger: **automation does not remove the need for expertise**. If AI writes the code or performs the calculation, someone must still know enough to detect a subtly wrong assumption, an inappropriate approximation, a spurious numerical result, data leakage, or an invalid inference. The human role therefore shifts from *production* toward **direction and verification**, rather than disappearing.

This also leads naturally to a problem the competition explicitly cares about: **credit and incentives**. If researcher A proposes a crucial question, researcher B recognizes a connection, researcher C supplies data, and AI performs most of the technical implementation, our current authorship system may poorly represent the actual intellectual contributions. A future infrastructure for collective science therefore needs to record contributions at a much finer level. So I think you potentially have a **much more interesting proposal than “AI automates scientific tasks.”** The one-page essay could revolve around one thesis:

AI makes execution abundant; scientific judgment remains scarce. We should redesign scientific collaboration around that new asymmetry.

That would also let you discuss programming and cumbersome calculations as concrete examples

without making them the entire proposal.

User

Please keep the following points less verbose

ChatGPT

Sure. A more concise version of the core points would be:

- **AI can automate execution-heavy scientific tasks**, such as programming, symbolic calculations, numerical computations, data processing, and literature searches.
- **Human effort can shift toward higher-level scientific reasoning**: choosing relevant questions, designing methods, interpreting results, challenging assumptions, and identifying unexpected connections.
- **AI may therefore change the division of labor in science**, rather than simply making existing workflows faster.
- **Science could become more collective and interdisciplinary**, because AI can reduce technical barriers between researchers with different expertise.
- **Human expertise remains essential for validation**: automated results still need to be checked for incorrect assumptions, numerical errors, or invalid conclusions.
- **Scientific incentives may also need to change**, since intellectual contributions such as proposing questions, connecting ideas, or providing data may become more important than carrying out the technical implementation.

The central idea could be summarized as:

AI makes scientific execution cheaper; human judgment, creativity, and problem selection become increasingly valuable.

User

At the same time, there is a problem in verifying the data from AI: e.g. the produced code or the performed computation, especially if it requires nontrivial passages such as approximations, no? Are there currently verification tools that can be used?

ChatGPT

Exactly. I think this is an important second pillar of the proposal: **AI may remove the execution bottleneck while creating a verification bottleneck.**

Suppose AI produces a calculation containing

$$\begin{array}{c} A \simeq B \\ \downarrow \\ C \simeq D \end{array}$$

where the first approximation assumes a small parameter, neglects some terms, or is valid only in a particular regime. Checking the algebra is relatively easy; checking that the **approximation was scientifically justified** is much harder.

There are already several classes of verification tools:

- **Formal verification / theorem provers** — e.g. **Lean**, **Z3**, **Dafny**. They can verify mathematical proofs or prove that code satisfies formally specified properties. Lean, for instance, can independently kernel-check a proposed proof, including AI-generated ones.
- **Formal code verification** — tools such as **Frama-C** can prove absence of classes of runtime errors and verify that C code conforms to a mathematical specification.
- **Validated numerical computation** — interval arithmetic can propagate rigorous numerical bounds rather than just floating-point values. Packages such as

`IntervalArithmetic.jl` can guarantee that the true numerical result lies inside a computed interval.

- **Conventional verification** — unit tests, independent implementations, comparison with known limits, dimensional checks, convergence tests, higher-precision calculations, etc. These are weaker than formal verification but much easier to apply broadly.

But there is a fundamental limitation: **these tools verify statements relative to assumptions/specifications.**

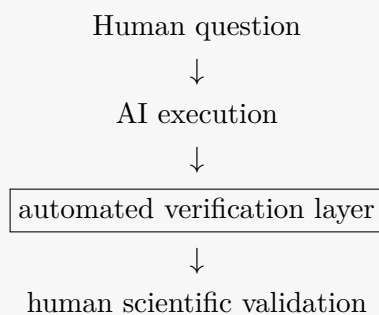
For example, a theorem prover might rigorously establish

$$\epsilon \ll 1 \implies f(\epsilon) = f_0 + \epsilon f_1 + \mathcal{O}(\epsilon^2),$$

but it does not automatically establish that $\epsilon \ll 1$ for the physical system you are studying, or that neglecting $\mathcal{O}(\epsilon^2)$ is acceptable for the scientific question.

That remains a scientific judgment.

This suggests a particularly interesting architecture for your idea:



And I think **the verification layer itself could be one of the most interesting parts of the proposal.** Ideally, future scientific AI should not merely output a number or some code; it should output:

result + assumptions + approximations + validity domain + uncertainty/error estimate + machine-checkable verification where possible.

For example, rather than saying “I computed $x = 1.34$ ”, an AI scientific system should effectively say:

“I obtain $x = 1.34$. I used approximations A and B, valid under conditions C and D. Numerical error is below X. These algebraic steps were formally checked. This particular physical assumption cannot be automatically verified and requires researcher judgment.”

That distinction between **machine-verifiable correctness** and **scientific validity requiring human judgment** could be quite powerful for the Cell Collective Intelligence proposal. It also reinforces your original argument: as AI takes over execution, the scientist’s role increasingly becomes **formulation, verification, and judgment**, rather than simply performing calculations.

User

What does Frama-C do? Is Lean complete (e.g. has all of mathematics been formalized yet)?

ChatGPT

Frama-C and Lean address rather different verification problems.

Frama-C

Frama-C analyzes C programs. Its purpose is essentially to answer questions such as:

“Given these assumptions about the inputs, can this C program ever do something incorrect?”

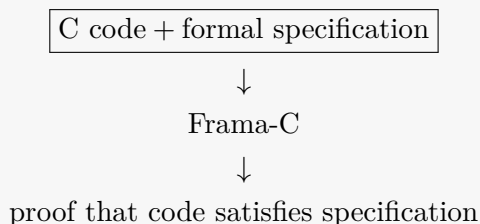
It has several verification mechanisms. For example, **Eva** performs static analysis and can prove the absence of classes of runtime errors such as invalid memory accesses, division by zero, integer overflow, uninitialized reads, etc.

More interesting for your scientific-AI idea is the **WP plugin**. You can specify what a function is supposed to do using ACSL contracts—for example,

$$x > 0 \implies \text{output} = \sqrt{x}$$

within some formally defined meaning—and Frama-C generates mathematical verification conditions showing whether the implementation satisfies that specification. It can send those conditions to SMT solvers such as Z3 or CVC5.

So schematically:



The critical caveat is exactly relevant to your proposal:

Frama-C proves that the code implements what you told it to implement.

It does not prove that your scientific model is correct.

If an AI writes a simulation using an unjustified approximation, Frama-C might perfectly verify that the code implements that approximation while saying nothing about whether the approximation makes physical sense.

Is Lean “complete”?

No, definitely not in the sense that all mathematics has been formalized.

Lean is a **proof assistant**. Mathlib, its main mathematical library, contains a substantial amount of formalized mathematics and covers algebra, analysis, topology, probability, number theory, and many other areas. It has even been used for frontier-level formalization projects. But this is still only a fraction of existing mathematics.

You should distinguish two concepts:

1. “Has all mathematics been entered into Lean?”

No.

A theorem from a paper does not magically exist in Lean. Someone must formally define all the objects involved and construct a proof from previously available results.

For example, a physicist can casually write

$$e^x \simeq 1 + x$$

for $|x| \ll 1$.

To formalize the corresponding statement properly, you need something more precise, such as

$$|e^x - (1 + x)| \leq Cx^2$$

under specified conditions on x .

Lean can then verify the proof **exactly**.

2. “Could Lean in principle prove every true mathematical statement?”

Also no, in the unrestricted sense.

Gödel’s incompleteness theorems imply that a sufficiently expressive, consistent formal system capable of arithmetic cannot prove every true statement about arithmetic.

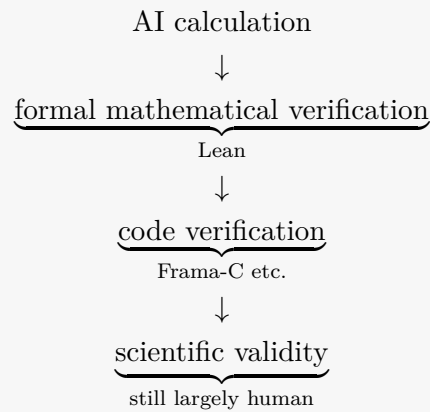
So there cannot be a single sound formal system that simply contains/proves **all mathematical truth**.

What Lean gives you instead is something extremely valuable:

If Lean's kernel accepts a proof, the theorem follows from the stated definitions and axioms.

Its kernel mechanically checks the proof rather than trusting the person—or AI—that generated it.

And that exposes exactly the hierarchy relevant to your prize idea:



The last step is the difficult one. Lean could prove flawlessly that an approximation follows **if** $\epsilon < 10^{-3}$, while neither Lean nor Frama-C inherently knows whether assuming $\epsilon < 10^{-3}$ is justified for the physical system under study.

That distinction—**formal correctness versus scientific correctness**—is, I think, one of the strongest issues to raise in your proposal.

User

So currently we are still missing a complete, trustworthy and general method to confirm AI-generated results, which, together with problems of hallucinations, can be dangerous for scientific research?

ChatGPT

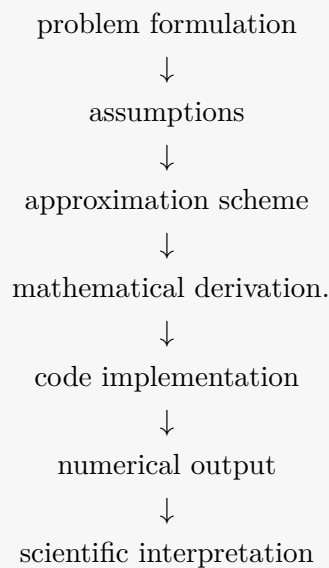
Yes—in the broad sense, that is correct.

We still do **not** have a **single general, complete, and fully trustworthy method** to validate arbitrary AI-generated scientific outputs.

A good short summary is:

- **No universal verifier exists.**
- We have **partial verification tools** for specific things:
 - formal proofs for some mathematics,
 - formal/static verification for some code,
 - testing and benchmarking for numerical routines,
 - reproducibility and cross-checks for data analysis.
- But these tools usually verify only a **restricted layer** of the result.

The main problem is that a scientific result typically involves many layers:



AI can fail at any of these layers.

For example, even if:

- the **code is syntactically correct**,
- the **algebra looks plausible**,
- and the **final number is reproducible**,

the result can still be scientifically wrong because:

- the approximation is invalid in that regime;
- a hidden assumption was inserted;
- the AI hallucinated a theorem, citation, or method;
- the code computes the wrong quantity;
- the interpretation of the output is incorrect.

So yes: **hallucinations + lack of general verification** can indeed be dangerous for scientific research, especially if AI outputs are accepted too quickly or treated as authoritative.

What we really have today is not a full solution, but a **patchwork of safeguards**:

- formal verification where possible;
- unit tests and independent implementations;
- comparison with known limits/cases;
- uncertainty estimates and convergence checks;
- human expert review;
- full reproducibility of the pipeline.

So I would phrase the point like this:

AI can accelerate scientific production, but trustworthy validation remains incomplete and problem-dependent. Therefore, without robust verification and expert oversight, AI-generated results can introduce subtle but serious scientific errors.

That is a strong point for your proposal, because it shows that the future role of scientists is not only to ask questions, but also to **audit, validate, and interpret** AI-produced results.

User

I see two main ways to avoid these problems: first, we should build more conscious models that are able to better quantify their level of uncertainty, avoid hallucinations and false results, and explicitly tell you if a result is not that trustworthy; second, we should rely more on human verification of critical passages. This would also require more interactions between scientists of

different disciplines: the possibility of using AI to go beyond one's area of expertise should be carefully balanced by another scientist verifying that the claims are correct. Do you agree with these proposals? Do you have others?

ChatGPT

Yes. I agree with both, with one modification to the first: I would call them **uncertainty-aware models**, rather than “more conscious” models.

1. Models that know when not to answer

AI systems should estimate their epistemic uncertainty and be able to say:

“I do not have enough evidence to establish this result reliably.”

This means developing better **calibration and abstention mechanisms**, rather than forcing the model to always produce an answer. This is still an open problem: current models can be confidently wrong, particularly on difficult expert-level questions.

Ideally, an AI result would come with something like:

result + confidence + assumptions + validity domain.

However, merely asking an AI “how confident are you?” is not sufficient: reported confidence itself can be poorly calibrated.

2. Human verification at critical points

I strongly agree with this.

AI could allow scientists to work outside their traditional expertise, but this creates a new danger:

ability to produce work outside your expertise $\neq$ ability to recognize errors in that work.

So interdisciplinary science could increasingly involve something like:

physicist
↓ AI
biological model.
↓ biologist
expert validation

This would make **collective intelligence even more important**, not less. AI lowers disciplinary barriers, while scientists from the relevant disciplines provide domain-specific validation.

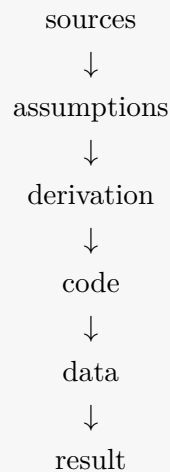
3. I would add traceability and provenance

This could be a third major principle.

An AI should not simply provide:

“The answer is 4.72.”

It should preserve a complete chain:

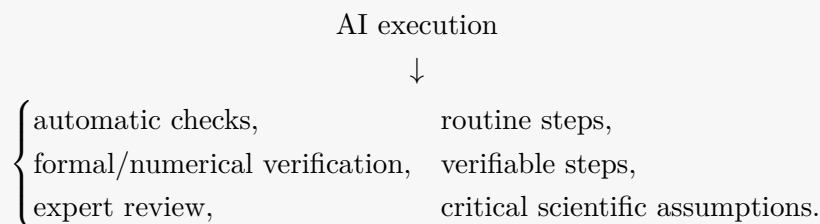


Then another scientist—or an automated verifier—can reconstruct and audit the result. This is increasingly discussed as **provenance** in autonomous science: maintaining a complete, auditable record of what the AI did may be at least as important as understanding its internal reasoning.

4. Verification should be proportional to risk

Not every AI operation needs another scientist to inspect it. That would eliminate much of the efficiency gain.

Instead:



For example, formatting a dataset does not require expert review. Introducing a new approximation into a physical model probably does.

So I would organize your proposal around **three safeguards**:

1. **Uncertainty-aware AI** — models should recognize uncertainty and abstain rather than hallucinate.
2. **Human expert verification** — particularly for critical assumptions and claims outside the user's expertise.
3. **Transparent, reproducible provenance** — every important AI-generated result should be auditable.

This actually strengthens your broader argument: **AI does not make collective scientific expertise obsolete; paradoxically, it could make interaction between different kinds of experts even more important.**

User

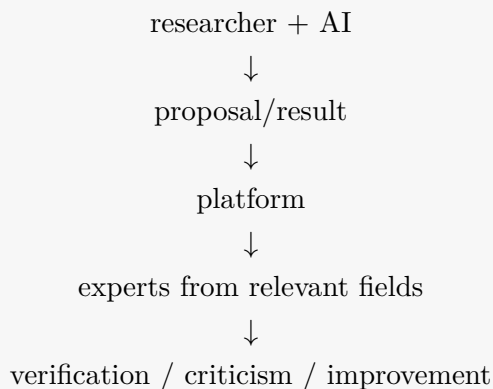
As a concrete example, a networking platform would be useful, right? For instance, instead of relying on email, conferences and word of mouth to contact scientists from different fields, one could rely on an online platform in which ideas are shared, and people with different expertise could verify whether other people's AI-generated proposals are valid.

ChatGPT

Yes. That would be a very concrete way to translate your general idea into a collective-intelligence mechanism.

The key point is that AI may let researchers generate hypotheses, models, code, or analyses outside their own narrow expertise, but **verification then requires access to the relevant experts**. Today that access is inefficient: email, personal networks, conferences, and word of mouth.

A dedicated platform could instead work like this:



The platform could include:

- **structured submission of AI-assisted ideas/results**, including assumptions, code, sources, and uncertainty;
- **automatic routing to relevant experts** based on expertise;
- **verification requests**, where scientists can flag incorrect assumptions, approximations, or interpretations;
- **versioning and audit trails**, so one can see how a result changed after criticism;
- **credit for verification**, not only for producing the original idea;
- **reputation systems** based on useful corrections and validated contributions.

The last point is especially important. At present, scientists receive much more credit for **producing new results** than for checking someone else's work. Your system would need incentives that make verification itself a recognized scientific contribution.

It could therefore be more than a scientific social network. It would be something closer to a **distributed verification layer for AI-assisted science**.

There are obvious risks—low-quality AI-generated submissions could overwhelm experts, reputation systems could become distorted, and unpublished ideas raise priority/confidentiality issues—but those are design problems rather than objections to the basic concept.

I think this is actually one of the strongest concrete implementations of the framework you are developing:

AI expands what each scientist can attempt; the platform expands the collective capacity to verify what AI produces.

User

Great! This should be my proposal ;)