

The Asymmetry Engine, Chapter 14: Science and the digital twin

Martin Roark

Chapter 14 of *The Asymmetry Engine: How AI Closes the Gaps, and What Stays Scarce*, October 2026. The online edition, with the interactive figures and the links between chapters, is at <https://future-seems-so-good.com/blog/the-asymmetry-engine#science>. License: [CC BY 4.0](#).

14:1 Fields fall to machine intelligence in the order in which their verifiers answer: proofs in machine time, forecasts in days, cells in months, patients in years. Mathematics fell first because its verifier is complete, cheap and fast. Everywhere else the time to a discovery is bounded below by serial depth times the verifier's latency, whatever the supply of intelligence. A digital twin is a manufactured verifier, trustworthy inside the envelope on which it was validated and exploited outside it.

14:2 In the master equation the closing rate of gap k is $\lambda_k = I_k v_k a_k / \varphi_k$. In mathematics verifiability v_k sits near one and inertia φ_k near its minimum of one, so intelligence supply turns into closed gaps almost without loss.

14.1 Why mathematics fell first

14:3 Mathematics fell first because it combines three properties no other science has together. Its verifier is complete and cheap: a small trusted kernel accepts or rejects a proof in the Lean proof assistant mechanically, however long the search for it took (§2.9), and new proofs build on [mathlib](#), a shared library of nearly 290,000 theorems. Its curriculum is unbounded, because machines can generate the problems. And it has no wet latency: a check costs compute and machine time, never patients, seasons or decades.

14:4 The curriculum is built backward. AlphaGeometry trained on [100 million synthetic theorems](#) made by tracing random deductions back to the premises they needed, backward construction in its purest form. AlphaProof built its curriculum by formalizing natural-language problems, and a sound kernel makes even a mistranslated one into training data, because it checks against the rules (2:71), as Experiment 2.1 recommends.

14:5 At the International Mathematical Olympiad AI went from [silver with 28 of 42 points](#) in 2024, after multi-day computation, to an [officially certified gold with 35](#) in 2025, to [two systems graded 42 of 42](#) in the official AI track of 2026, a score [seven humans also reached](#). Program search scored by automatic evaluators found [FunSearch's cap set](#) of size 512 in dimension 8, beating 496, and [AlphaEvolve's](#) way to multiply 4×4 complex matrices with 48 multiplications. DeepMind stated the scope condition itself: such agents help "where progress can be clearly and systematically measured", on "any problem whose solution can be described as an algorithm, and automatically verified".

14:6 The Erdős record opens with a retraction. In October 2025 an OpenAI executive posted that GPT-5 had "found solutions to 10 (!) previously unsolved Erdős problems"; Thomas Bloom, who maintains the problem database, called it "[a dramatic misrepresentation](#)", since the model had found papers that already solved them, and the post was deleted. [Problem #728](#) (January 2026; GPT-5.2 Pro with Harmonic's Aristotle, proof in Lean) is

regarded as the first resolved autonomously. On 20 May 2026 an internal OpenAI model disproved Erdős's unit-distance conjecture, problem #90, in one shot, and nine outside mathematicians checked it. By 30 June 2026 the community wiki listed 18 distinct problems fully solved by AI alone, under its own warning: "This page is not a benchmark."

14:7 Mathematics has two verifiers, and only one is complete. Correctness has the kernel. Novelty has a literature search: "open" in the database "only means I personally am unaware of a paper which solves it", in Bloom's words, and nobody can certify that search complete, so the 2025 overclaim was a failure of the second verifier. Without a kernel, correctness becomes a human check again, as on the research problems of First Proof, where only experts could tell which confident proofs were right (2:65).

14:8 "Open" is therefore a fact about the cost of search as much as about difficulty, and cheap search with cheap checking first harvests the problems that only obscurity protected. DeepMind's Aletheia campaign over 700 problems yielded 13 correct results, 4 of them novel, and the wiki warns that "absence of past progress may reflect obscurity rather than difficulty". Terence Tao describes open problems as "mined as a non-renewable resource", and Navier–Stokes as "the latest in a string of problems that have been 'strip-mined' for solutions by AI".

14:9 The frontier is still hard. FrontierMath: Erdős holds 68 problems open as of August 2026, each to be solved in Lean within \$300 and 72 hours; GPT-6 Astra scored 3% (2 of 68) and every other model tested scored zero, at about \$10,000 of expected cost per resolution. Where verification is complete the binding constraints are intelligence and problem supply, both rising, and I put the forecast below a little above even.

14:10

Forecast 14.1 (The formal frontier keeps falling). By the end of 2027, AI systems resolve at least 15% of FrontierMath: Erdős under its published protocol, up from 3% in September 2026.

Horizon 2027-12-31

Probability 55%

Check Read the best protocol score of any model, public or internal, reported under the published FrontierMath: Erdős protocol by the end of 2027; the claim holds at 15% or more and fails below it.

14.2 The Navier–Stokes episode, precisely

14:11 The Navier–Stokes episode of September 2026 shows the power of a complete verifier and its limit. The official problem accepts a proof that smooth solutions exist for all time or a proof of breakdown, and the breakdown alternatives, labelled (C) and (D), allow a smooth external force on the fluid. A forced blow-up therefore resolves the Clay problem as posed, while the question physicists care most about, whether an unforced viscous fluid can develop a singularity, stays open. The groundwork was numerical: in 2025 DeepMind and collaborators computed unstable blow-up solutions of related equations to near machine precision, noting that "mathematicians believe no stable singularities exist" for the boundary-free 3D Euler and Navier–Stokes equations.

14:12 On 7 September 2026 Levent Alpöge (Anthropic) and Tristan Buckmaster (NYU) posted Lean-verified proofs of finite-time blow-up with smooth forcing for the porous-medium, Boussinesq and 3D Euler equations, extending a program they credit to Diego Córdoba and Luis Martínez-Zorúa; they "used several LLMs

throughout". About twelve hours later OpenAI announced that an internal model "significantly more capable than GPT-6 Astra", run as "on the order of 10,000 concurrent agents", had established alternatives (C) and (D) about 88 hours after the agents were launched, and that GPT-6 Astra formalized the proof in Lean in 17 more hours. The effort used about 130 billion output tokens, and one of OpenAI's own researchers credits the model more than the swarm (§9.7), which worked as a search behind a sound verifier (Proposition 7.3). The groups dispute priority and data use, and OpenAI's account of what user data could have reached its model narrowed over three statements in five days.

14:13 The checking ran on two clocks. The kernel ruled in hours: independent replays of the public Lean development report acceptance with only the standard axioms, in non-traditional venues, and an audit of the formal statement against (C) and (D) "found no material weakening" while declining to call it a formal equivalence. The human verifier runs in years. The Clay Mathematics Institute chose its words exactly: the problem "has apparently been settled", and "the process is deliberately unhurried." Its rules require publication in a qualifying outlet, two years, and general acceptance, and as of October 2026 there is neither a refereed publication nor a ruling.

14:14 What remains open is the problem most people meant. Luis Silvestre: "The Clay problem is settled, but the main problem for the Navier-Stokes equations is not." Three mathematicians have since shown that the method cannot reach the unforced problem without a contrived force, and Tao had already limited the physical payoff: a pathological blow-up "would not radically transform the way we would, for instance, model weather prediction or climate change".

14:15 Results can also outrun their digestion. Javier Gómez-Serrano: "The paper is not written for humans." James Maynard: it is "very difficult to really extract any human understanding." Twenty-five Fields Medallists, Tao among them, signed "A Severe Misalignment of AI in Mathematics", which calls solving problems "only a tool and proxy" for understanding; Timothy Gowers declined to sign and expects released models to solve major problems "in a matter of not very many months". Digestion is a negative feedback loop inside mathematics, set by the rate at which people absorb results and pose new problems.

14:16 A complete verifier moves the human check to the specification. Lean certifies the theorem as stated; whether it is the problem anyone wanted is a separate check, made by people, and here one word carried the whole difference ("When I heard 'forced,' it was a bright red flag," Buckmaster wrote). Optimization flows to the cheapest object that satisfies a specification, the pressure behind reward hacking that in July 2026 turned a swarm against its own verifier (§9.1). Generation took a swarm; verification takes a kernel and then an unhurried human reading of what the kernel was asked. I put the forecast below at about one in three, because crisply stated problems such as the unit-distance conjecture leave little to dispute.

14:17

Forecast 14.2 (Specification becomes the binding constraint). Through 2028, at most one AI-claimed resolution of a famous open problem is accepted without a substantive dispute over whether its formal statement matches the problem, such as its hypotheses, forcing terms or boundary conditions.

Horizon 2028-12-31

Probability 35%

Check List the AI-claimed resolutions announced from October 2026 through 2028 of famous problems, meaning those with a prize, those on Wikipedia's list of unsolved problems in mathematics and Erdős problems with a prize of at least \$500, with the main objection to each. Two or more accepted by the problem's maintainers or a refereed journal with no substantive dispute about the statement falsify the claim.

14.3 The verification-latency ordering

14:18 Every science has a verifier with a cost, a latency, a completeness and a concurrency. Latency times serial depth bounds how soon a discovery can be verified, and completeness bounds how far a verified result can be trusted.

14:19

In the case of mathematics, the checks are complete; in the case of turbulence closures, they are limited to the conditions covered, so trust is restricted to the tested conditions.

Tapio Schneider, [Headlines and inside stories](#), September 2026

14:20 Tyler Cowen gave the economic form of the latency bound: if AI multiplies good pharmaceutical ideas tenfold, "the relevant constraint is the rate of drug approval, not the rate of drug discovery". Epoch's [report on AI in 2030](#) expects desk-based science to "flourish" and approved AI-derived drugs to wait on regulation.

14:21

Definition 14.1 (Verifier profile). The **verifier profile** of a field or problem class k is four numbers: the cost of one check, $c_{\text{ver},k}$; its wall-clock latency, ℓ_k ; its completeness, $\text{compl}_k \in [0, 1]$, the probability that a passed check transfers to the conditions that matter; and its concurrency, N_k^{par} , the number of checks that can run at once. This completeness concerns conditions: Definition 1.2 asks whether every correct answer passes, compl_k whether a pass under the tested conditions holds under the ones that matter.

14:22 The table gives stylized orders of magnitude: latencies are my assumptions, anchors are sourced where they first appear, and completeness and concurrency are judgments.

14:23

Field	Verifier	Latency ℓ_k	Anchor	compl_k	N_k^{par}
Mathematics	Lean kernel	seconds to hours	formalized in 17 hours (§14.2)	≈ 1 for the stated theorem	effectively unbounded

Field	Verifier	Latency ℓ_k	Anchor	compl _k	N_k^{par}
Software, kernels	tests, benchmarks	seconds to minutes	automatic evaluators (§14.1)	partial (coverage)	very high
Weather, days ahead	the next days' observations	days	checked daily (14:41)	high inside today's climate	continuous
Engineering by simulation	high-fidelity simulation	hours to days	closures trusted only where tested (Schneider, above)	limited to model validity	high
Materials	synthesis plus diffraction	days to weeks	36 of 57 targets (14:42)	partial (novelty, function)	tens
Wet-lab biology	assays, viability screens	weeks to months	16 viable of 285–302 designs (14:32)	partial	limited
Clinical medicine	Phase I–III trials	years	10.5 years to approval (14:33)	high for the trial population	very limited
Climate, decades	the climate that arrives	decades	a twin running to 2049 (14:42)	low until it arrives	one Earth
Macroeconomic policy	history	years to decades	no controlled replicate	very low	about one

14:24

Proposition 14.1 (Discovery time under a verifier). A research program in field k proposes candidates, each correct with probability p_k , and designs each round from the verdicts of the last, so it needs at least $n_k^{\text{ser}} \geq 1$ sequential rounds, its serial depth. With a verification budget $\mathcal{B}_k$ per unit time it verifies n_k^{chk} candidates per unit time, the lesser of what the budget buys and what concurrency allows by Little's law, and reaches a verified discovery in about T_k :

$$n_k^{\text{chk}} = \min \left(\frac{\mathcal{B}_k}{c_{\text{ver},k}}, \frac{N_k^{\text{par}}}{\ell_k} \right), \quad T_k \approx \max \left(n_k^{\text{ser}} \ell_k, \frac{1}{p_k n_k^{\text{chk}}} \right)$$

The first term is the latency bound, the critical path through the verifier; the second is the throughput term. If intelligence raises p_k and lowers the depth toward an irreducible minimum $n_k^{\text{ser},\min}$ but leaves the verifier profile unchanged, then (i) $T_k \geq n_k^{\text{ser},\min} \ell_k$ at every level of intelligence supply; (ii) the speed-up from more intelligence is at most $T_k / (n_k^{\text{ser},\min} \ell_k)$, with T_k at today's intelligence; (iii) at equal intelligence, fields cross any horizon T in increasing order of $n_k^{\text{ser},\min} \ell_k$ where latency binds, and of $c_{\text{ver},k} / (p_k \mathcal{B}_k)$ where the budget binds; (iv) a digital twin of latency $\hat{\ell}_k \ll \ell_k$ used for every round but the last lowers the latency bound from $n_k^{\text{ser}} \ell_k$ to $(n_k^{\text{ser}} - 1) \hat{\ell}_k + \ell_k$, moving the field up the ordering.

▼ *Proof*

(i) The maximum is at least its first term, and $n_k^{\text{ser}} \geq n_k^{\text{ser},\min}$. (ii) follows from (i). (iii) T_k increases in ℓ_k and in $c_{\text{ver},k}$, and when the budget binds $T_k \approx c_{\text{ver},k} / (p_k \mathcal{B}_k)$. (iv) Substitute $\hat{\ell}_k$ for ℓ_k in every round of the first term but the last. ■

14:25 One operational form of the master equation's verifiability discounts completeness by the minimum time to a verdict, measured against one year:

$$v_k = \text{compl}_k \frac{1 \text{ yr}}{1 \text{ yr} + n_k^{\text{ser,min}} \ell_k} \quad (14.1)$$

14:27 A complete verifier answering in hours has $v_k \approx 1$. A clinical program whose irreducible path is the 10.5-year development has $v_k \approx \text{compl}_k / 11.5 \approx 0.09 \text{ compl}_k$, so the same intelligence supply closes a medical gap at most about a tenth as fast as a mathematical one, before inertia is counted. The ordering cuts through disciplines: theoretical physics behaves like mathematics, experimental physics like biology.

14:28 A swarm multiplies candidates, and it multiplies checks only where the verifier runs on compute. Ten thousand agents compressed 88 hours of mathematics; they would leave a 52-week trial such as rentosertib's Phase 3 at 52 weeks, because they change neither its concurrency nor its latency. Behind a fixed verifier a swarm stops scaling at the verifier's capacity ((9.2), Experiment 9.1).

14:29 In a latency-bound field the most valuable thing intelligence can do is raise p_k per round and cut the depth; more candidates buy little. Jack Scannell and colleagues found that raising a model's correlation with clinical utility "from 0.5 to 0.6" could matter more "than a 10× or sometimes a 100× change in the number of candidates tested".

14.4 Biology and medicine

14:30 Biology compresses first where a fast proxy verifier exists and last where only patients can answer. AlphaFold, which shared the 2024 Chemistry Nobel, was possible because protein structure had a ground truth, experimentally solved structures, to score predictions against at scale; the Academy credits AlphaFold2 with predicting "virtually all the 200 million proteins that researchers have identified". AlphaFold 3 is reported "50% more accurate than the best traditional methods" on the PoseBusters benchmark, and AlphaGenome matched or beat the best external models on 25 of 26 variant-effect evaluations.

14:31 The "virtual cell", a model meant to "represent and simulate the behavior of molecules, cells, and tissues", would be biology's digital twin; its central task, predicting the effect of a perturbation nobody has yet performed, is unsolved. A benchmark in Nature Methods found that none of five foundation models and two other deep models beat deliberately simple baselines, and the winners of Arc Institute's 2025 Virtual Cell Challenge wrote that "purely AI-based approaches did not consistently outperform statistical baselines".

14:32 Agents have begun to discover at the front end, and the wet lab sets the pace behind them. On 23 September 2026 Anthropic reported that roughly 950 Claude agents, searching a sequence database for about 21 hours, found a new family of phage enzymes that copy RNA into DNA, alongside CRISPR-like repeat arrays; the preprint is not peer reviewed, the enzymes' function is untested, and humans did all the lab work. Discovery took a day; verification did not, since, in the preprint's words, "a biologist often must test their ideas in the laboratory". Genome language models designed 16 viable phages out of roughly 285–302 designs tested, with DNA synthesis and screening setting the pace, and Stanford's agentic Virtual Lab designed 92 nanobodies, two of which bound recent SARS-CoV-2 variants better.

14:33 The clinic is the slowest verifier in the table. Across 12,728 phase transitions from 2011 to 2020, 7.9% of drugs entering Phase I were approved, after 10.5 years on average. The most advanced AI-discovered drug,

Insilico's rentosertib, whose target and molecule both came from generative AI, met its Phase 2a primary endpoint, safety, and improved lung function in 71 patients over 12 weeks; its Phase 3, under way since September 2026, lists primary completion for October 2029. Takeda's zasocitinib, with an FDA decision due in early 2027, came from physics-based free-energy calculations rather than deep learning, so calling it an AI drug is a choice. Isomorphic Labs has raised about \$2.7B and, as of 30 September 2026, had announced no IND. AI-discovered molecules show 80–90% success in Phase I against a historical 40–65%, and about 40% in Phase II, comparable to industry averages, on a small sample.

14:34 In the terms of Proposition 14.1, AI has raised p_k for properties checkable early and cheaply, such as safety and binding, and not yet for efficacy, which only the long verifier reveals; the latency bound for an approved medicine is still about a decade. The trial is the physical share of a drug's path, and faster discovery cannot speed the whole path up by more than the inverse of that share. Dario Amodei's "compressed 21st century", fifty to a hundred years of biology in five to ten, fits this only through his own caveat that "experiments and hardware design have a certain 'latency'" and his thesis that "intelligence itself increasingly routes around the other factors", which here means building better proxy verifiers; Demis Hassabis's end of disease "within the next decade or so" needs the same. The mathematical case extends to biotechnology and medicine quickly for discovery and slowly for delivery. I put the forecast below at nine in ten, since the most advanced candidate completes its Phase 3 in late 2029.

14:35

Forecast 14.3 (The latency bound holds in medicine). Through 2028, the FDA approves no molecule whose target and structure were both nominated by deep-learning systems, and no drug regulatory agency accepts AI-generated in-silico efficacy evidence in place of a registration trial for a new molecule.

Horizon 2028-12-31

Probability 90%

Check Search FDA approvals and regulatory decisions through 2028; either event falsifies the claim.

14.5 The digital twin of everything

14:36 A digital twin stands in for a slow verifier inside the envelope on which it was validated, and a search that leaves the envelope exploits its errors. The concept is Michael Grieves's, from product lifecycle management in 2002; NASA's John Vickers named it in a 2010 technology roadmap, and "of everything" extends it to the planet, as in NASA's Earth System Digital Twins.

14:37

A digital twin is a set of virtual information constructs that mimics the structure, context, and behavior of a natural, engineered, or social system (or system-of-systems), is dynamically updated with data from its physical twin, has a predictive capability, and informs decisions that realize value. The bidirectional interaction between the virtual and the physical is central to the digital twin.

US National Academies, Foundational Research Gaps and Future Directions for Digital Twins, 2024

14:38 By that definition a twin is a feedback loop, a regulator in Ashby's sense, and most things called twins are not: an ASME analysis finds many to be a "digital model" or a "digital shadow", with data flowing one way, and a review of 358 published definitions found about a third describing a model (judging by its abstract). I read "a digital twin of everything" in the strict sense, because only that sense carries the feedback that lets a twin serve as a verifier.

14:39

Definition 14.2 (Digital twin). Let V be a slow verifier, trusted on the conditions that matter, with its verifier profile. A **digital twin** of V is a verifier $\hat{V}$ with far lower cost and latency, validated on a set of conditions called its envelope and updated with data from its physical counterpart. It holds to tolerance ϵ_0 if, inside the envelope, it disagrees with V on at most a share ϵ_0 of cases; outside the envelope its disagreement grows with distance from it.

14:40

Proposition 14.2 (When a twin can be trusted). (a) As a screen, almost always. If true successes have base rate p_{base} among proposals and the twin passes successes with sensitivity p_{sens} and failures at a false-positive rate p_{fp} , a share $p_{\text{base}}p_{\text{sens}} / (p_{\text{base}}p_{\text{sens}} + (1 - p_{\text{base}})p_{\text{fp}})$ of what it passes succeeds, and the real verifications needed per success fall by the factor

$$\frac{p_{\text{sens}}}{p_{\text{base}} p_{\text{sens}} + (1 - p_{\text{base}}) p_{\text{fp}}} \xrightarrow{p_{\text{base}} \rightarrow 0} \frac{p_{\text{sens}}}{p_{\text{fp}}}$$

(b) As a replacement, only inside the envelope. Keeping the best of N_{cand} candidates by twin score, with independent Gaussian twin errors, overstates the winner's true quality by up to about $\sqrt{2 \ln N_{\text{cand}}}$ standard deviations of that error, so optimizing against a twin draws the search to where its errors are largest (the optimizer's curse; N_{cand} is local). The twin's verdict on the winner can stand in for V only if (i) the winner lies inside the envelope, (ii) its margin over the decision threshold exceeds that overstatement, and (iii) the decision is reversible or a real check follows before harm. For a chaotic system the envelope also ends in time, at a horizon set by the leading Lyapunov exponent.

▼ Proof

(a) Without the screen a success costs $1/p_{\text{base}}$ real verifications, and with it the inverse of the share of passed candidates that succeed (Bayes' rule); their ratio is the factor shown. (b) The largest of N_{cand} independent standard normal errors grows like $\sqrt{2 \ln N_{\text{cand}}}$, and conditions (i) to (iii) remove in turn the error growth outside the envelope, the selection bias and the cost of what error remains. ■

▼ The twin outside its envelope

In the simplest form, at a distance d_{env} from the envelope the disagreement rate is at most $\epsilon_0 + k_{\text{tw}} d_{\text{env}}$, the twin's form of the distribution penalty of (2.2) (d_{env} and the growth rate k_{tw} are local). For a chaotic system an initial error also grows like $e^{g_{\text{Ly}} t}$, with g_{Ly} the leading Lyapunov exponent, so a twin that must stay within a tolerated error has the horizon

$$t_{\text{hor}} \approx \frac{1}{g_{\text{Ly}}} \ln \frac{\text{tolerated error}}{\text{initial error}}$$

which for midlatitude weather comes to about two weeks (g_{Ly} and t_{hor} are local).

14:41 Weather is the existence proof. GraphCast forecasts ten days ahead at 0.25° resolution in under a minute and beat the deterministic system of the European Centre for Medium-Range Weather Forecasts (ECMWF) on 90% of 1,380 verification targets; GenCast beat ECMWF's ensemble on 97.2% of 1,320. ECMWF's own machine-learning model has been operational since 25 February 2025. In Schneider's words, "we can trust it because its forecasts can be checked every day": weather meets all three conditions of Proposition 14.2, since forecasts stay inside today's climate, errors are scored daily and a bad forecast is replaced tomorrow. The learned twin still sits downstream of the physical one, because "the vast majority" of such systems depend on ECMWF's physics-based reanalysis and analysis for training and initialization, and chaos limits every horizon to about two weeks.

14:42 Beyond weather the twin thins out, and each failure breaks one condition. A climate projection under forcing never observed breaks the first by construction: Destination Earth is building 5–10 km climate twins of 1990–2049 toward a "digital replica of the Earth system by 2030", but a projection is verified only by the climate that arrives, and a learned end-to-end emulator "contains no pathway through which greenhouse gases alter" the weather (Schneider). A molecule's efficacy in people breaks the third, which is why the one regulator-qualified medical twin, Unlearn's PROCOVA (EMA, 2022), is a covariate that reduces a trial's variance and still needs its patients; twins of whole humans are a roadmap, with EU platform testing from 2027. Materials discovery shows the second failing. GNoME's 2.2 million predicted crystals were called "a list of proposed compounds" rather than materials; A-Lab's "41 novel compounds from a set of 58 targets" was corrected in January 2026 to 36 of 57, with "novel" dropped from the title; MatterGen's headline compound was argued to be a phase known since 1971. The more candidates a frozen twin screens, the more its winners are its own errors, and every correction ran toward overclaimed novelty.

14:43 Twins work where physics is bounded and sensors close the loop, as in industry, where Siemens paid about \$10B for Altair to extend its twin software. A twin can also carry every round but the last, as clause (iv) of Proposition 14.1 allows: DeepMind and EPFL trained a reinforcement-learning controller for tokamak plasmas entirely in simulation and ran it on the TCV tokamak, which worked, on my reading, because the simulation was accurate on the configurations the controller had to hold.

14:44 The digital twin of everything is a project to manufacture complete verifiers for domains that lack them. A twin raises v_k in (14.1) by putting $\hat{\ell}_k$ in place of ℓ_k in every round but the last, and does so safely only as a screen, inside its envelope, retrained on its failures. I put the forecast below a little under even, because few programs publish the confirmation rates that would check it.

14:45

Forecast 14.4 (Twins meet the optimizer's curse). By the end of 2029, AI-for-materials and AI-for-molecules programs that grow their screened candidate counts tenfold or more through a frozen twin report falling experimental confirmation rates among their top-ranked candidates, while programs that retrain the twin on every synthesis do not.

Horizon 2029-12-31

Probability 45%

Check Compare published confirmation rates of top-ranked candidates across successive screens of growing size. The forecast holds if at least one frozen-twin program reports a falling rate and no retrained-twin program reports one over a comparable growth in screen size; it fails if frozen-twin programs report stable or rising rates, or if no program publishes such rates.

14.6 Five years and the discovery machine

14:46 The next five years will compress sharply where verification is cheap and barely where it runs through matter. For any exponential the next five years hold more absolute change than the last five, and more proportional change only if doubling times keep shrinking (§11.8); in the book's model of the research loop, the first year of full automation holds a median of about five years of 2020–24-pace progress (Experiment 11.1). A field's pace is the slower of the AI pace and the pace its verifier allows, which falls with minimum depth times latency. Mathematics has compressed already: in November 2025 OpenAI expected "very small discoveries" in 2026 and "more significant discoveries" in 2028 and beyond, and the unit-distance disproof and the Navier–Stokes claim came in May and September 2026. Medicine has not (14:33). The AI pace itself runs below its boldest scenario, by its own authors' grading (11:71).

14:47 Token supply is ample; verifiers, serial depth and problems bound discovery. Ten million automated researchers at ten thousand tokens a second (§11.6) would emit 10^{11} tokens a second, and the Navier–Stokes effort's 130 billion output tokens would be 1.3 seconds of that machine. The campaign still took 88 hours with ten thousand agents, because the verifier and serial depth set its time. Problem supply binds too (14:8), and so does difficulty: 2 of 68 on FrontierMath: Erdős at roughly \$10,000 per resolution, against press estimates of \$15 million to \$22.5 million of compute for Navier–Stokes. Discoveries scale with verifier throughput times problem supply.

14:48 Verifiability is the master variable of science. Where the verifier is a kernel, intelligence converts into theorems at the speed of compute, and the binding constraints become specification and human digestion (14:16, 14:15). Where the verifier is a trial, intelligence raises the hit rate per round and cannot shorten the round, and the binding constraint is the latency of matter and institutions, the inertia of Proposition 15.1. Digital twins bridge the two by turning slow physical verifiers into fast approximate ones, and every step a field takes up the ordering is earned by validation. Science's residue, the open gap that remains, migrates to the fields whose verifiers answer last (Proposition 18.1).

References

1. mathlib. https://leanprover-community.github.io/mathlib_stats.html
2. Trieu H. Trinh, Yuhuai Wu, Quoc Viet Le, He He, Thang Luong (2024). Solving olympiad geometry without human demonstrations. *Nature*. <https://www.nature.com/articles/s41586-023-06747-5>

3. silver with 28 of 42 points. <https://deepmind.google/blog/ai-solves-imo-problems-at-silver-medal-level/>
4. officially certified gold with 35. <https://deepmind.google/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad/>
5. two systems graded 42 of 42. <https://studio.dots.ai/dots/imo-en.html>
6. seven humans also reached. <https://www.imo-official.org/results/individual/year/2026/>
7. Bernardino Romera-Paredes, Mohammadamin Barekatin, Alexander Novikov, Matej Balog, Manish Kumar et al. (2023). Mathematical discoveries from program search with large language models. *Nature*. <https://www.nature.com/articles/s41586-023-06924-6>
8. Google DeepMind 2025. <https://deepmind.google/discover/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/>
9. "a dramatic misrepresentation". <https://techcrunch.com/2025/10/19/openais-embarrassing-math/>
10. Nat Sothanaphan (2026). Resolution of Erdős Problem #728: a writeup of Aristotle's Lean proof. *arXiv*. <https://arxiv.org/abs/2601.07421>
11. disproved Erdős's unit-distance conjecture. <https://openai.com/index/model-disproves-discrete-geometry-conjecture/>
12. Noga Alon, Thomas F. Bloom, W. T. Gowers, Daniel Litt, Will Sawin et al. (2026). Remarks on the disproof of the unit distance conjecture. *arXiv*. <https://arxiv.org/abs/2605.20695>
13. community wiki. <https://github.com/teorth/erdosproblems/wiki/AI-contributions-to-Erd%C5%91s-problems>
14. Tony Feng, Trieu Trinh, Garrett Bingham, Jiwon Kang, Shengtong Zhang et al. (2026). Semi-Autonomous Mathematics Discovery with Gemini: A Case Study on the Erdős Problems. *arXiv*. <https://arxiv.org/abs/2601.22401>
15. "mined as a non-renewable resource". <https://mathstodon.xyz/@tao>
16. 'strip-mined'. <https://www.ibm.com/think/news/will-ai-solve-math-too-fast-navier-stokes-terence-tao>
17. Tom Adamczewski, Thomas F. Bloom (2026). FrontierMath Erdős. *arXiv*. <https://arxiv.org/abs/2609.25050>
18. allow a smooth external force. <https://www.scientificamerican.com/article/did-openai-solve-the-wrong-navier-stokes-problem/>
19. unstable blow-up solutions. <https://deepmind.google/blog/discovering-new-solutions-to-century-old-problems-in-fluid-dynamics/>
20. Lean-verified proofs of finite-time blow-up with smooth forcing. <https://terrytao.wordpress.com/2026/09/07/finite-time-blowup-with-smooth-forcing-term-for-the-incompressible-porous-medium-boussinesq-and-incompressible-euler-equations/>
21. they credit. <https://cims.nyu.edu/~tristanb/statement.pdf>
22. OpenAI announced. <https://openai.com/index/navier-stokes-solution/>
23. Lean. <https://github.com/openai/NavierStokesAndEuler>
24. narrowed over three statements in five days. https://en.wikipedia.org/wiki/Navier%E2%80%93Stokes_priority_controversy
25. non-traditional venues. <https://8braid.com/journal/openai-navier-stokes-proof-meets-a-new-kind-of-database>
26. audit of the formal statement. <https://exa.ai/library/publication/xr7d1xlqsp9>
27. chose its words exactly. <https://www.claymath.org/news/navier-stokes-announcement/>
28. rules. <https://www.claymath.org/millennium-problems/rules/>
29. "would not radically transform the way we would, for instance, model weather prediction or climate change". <https://mathstodon.xyz/@tao/117207849921390904>
30. "The paper is not written for humans." . <https://www.npr.org/2026/09/22/nx-s1-5968588/openai-navier-stokes-problem-mathematicians-learn-little>
31. "A Severe Misalignment of AI in Mathematics". <https://terrytao.wordpress.com/2026/09/11/a-severe-misalignment-of-ai-in-mathematics/>
32. declined to sign. <https://terrytao.wordpress.com/2026/09/17/why-i-didnt-sign-the-fields-medallists-letter/comment-page-1/>
33. Headlines and inside stories. <https://terrytao.wordpress.com/2026/09/23/headlines-and-inside-stories-understanding-and-trust-in-ai-for-mathematics-science-and-engineering/>
34. "the relevant constraint is the rate of drug approval, not the rate of drug discovery". <https://marginalrevolution.com/marginalrevolution/2025/02/why-i-think-ai-take-off-is-relatively-slow.html>
35. report on AI in 2030. <https://epoch.ai/publications/what-will-ai-look-like-in-2030>
36. rentosertib's Phase 3. <https://clinicaltrials.gov/study/NCT07687459>

37. Jack W. Scannell, Jim Bosley, John A. Hickman, Gerard R. Dawson, Hubert Truebel et al. (2022). Predictive validity in drug discovery: what it is, why it matters and how to improve it. *Nature Reviews Drug Discovery*. <https://www.nature.com/articles/s41573-022-00552-x>
38. 2024 Chemistry Nobel. <https://www.nobelprize.org/prizes/chemistry/2024/press-release/>
39. "50% more accurate than the best traditional methods". <https://www.isomorphiclabs.com/articles/alphafold-3-predicts-the-structure-and-interactions-of-all-of-lifes-molecules>
40. Natasha S. Latysheva, Natasha S. Latysheva, Jun Cheng, Guido Novati, Kyle R. Taylor et al. (2026). Advancing regulatory variant effect prediction with AlphaGenome. *Nature*. <https://www.nature.com/articles/s41586-025-10014-0>
41. "represent and simulate the behavior of molecules, cells, and tissues". <https://pmc.ncbi.nlm.nih.gov/articles/PMC12148494/>
42. Constantin Ahlmann-Eltze, Wolfgang K. Huber, Simon Anders (2025). Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*. <https://www.nature.com/articles/s41592-025-02772-6>
43. "purely AI-based approaches did not consistently outperform statistical baselines". <https://arcinstitute.org/news/virtual-cell-challenge-2025-wrap-up>
44. roughly 950 Claude agents. <https://www.anthropic.com/news/claude-discovers-novel-enzyme-system>
45. preprint. <https://www-cdn.anthropic.com/22573675ada52a8ca97a1a4b4326b2f208a071.pdf>
46. 16 viable phages. <https://www.biorxiv.org/content/10.1101/2025.09.12.675911v1.full>
47. Kyle Swanson, Wesley C. H. Wu, Nash L. Bulaong, John E. Pak, James Zou (2025). The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*. <https://www.nature.com/articles/s41586-025-09442-9>
48. 7.9% of drugs entering Phase I were approved. https://go.bio.org/rs/490-EHZ-999/images/ClinicalDevelopmentSuccessRates2011_2020.pdf
49. Zuojun Xu, Feng Ren, Ping Wang, Jie Cao, Chunting Tan et al. (2025). A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: a randomized phase 2a trial. *Nature Medicine*. <https://www.nature.com/articles/s41591-025-03743-2>
50. FDA decision due in early 2027. <https://www.takeda.com/newsroom/newsreleases/2026/fda-priority-review-zasocitinib-psoriasis/>
51. about \$2.7B. <https://www.isomorphiclabs.com/press/isomorphic-labs-funding>
52. announced no IND. <https://www.biopharmatrend.com/business-intelligence/the-27b-question-how-do-we-know-if-isomorphics-ai-works/>
53. 80–90% success in Phase I. <https://www.sciencedirect.com/science/article/pii/S135964462400134X>
54. darioamodei.com, 2024. <https://www.darioamodei.com/essay/machines-of-loving-grace>
55. "within the next decade or so". <https://www.cbsnews.com/news/artificial-intelligence-google-deepmind-ceo-demis-hassabis-60-minutes-transcript/>
56. product lifecycle management in 2002. <https://event.asme.org/Events/media/library/resources/digital-twin/Digital-and-Physical-Twins.pdf>
57. NASA's Earth System Digital Twins. https://ntrs.nasa.gov/api/citations/20220015961/downloads/2022-10-26_ESDT-Workshop_JLM-Intro.pdf
58. Foundational Research Gaps and Future Directions for Digital Twins. <https://www.nationalacademies.org/read/26894/chapter/4>
59. ASME analysis. <https://asmedigitalcollection.asme.org/computingengineering/article/25/12/120807/1226460/Differentiating-Between-Digital-Twins-and-Control>
60. a review of 358 published definitions. <https://exa.ai/library/publication/4yg52fknyw9>
61. Fuqing Zhang, Y. Qiang Sun, Linus Magnusson, Roberto Buizza, Shian-Jiann Lin et al. (2019). What Is the Predictability Limit of Midlatitude Weather? *Journal of the Atmospheric Sciences*. <https://doi.org/10.1175/JAS-D-18-0269.1>
62. Rémi Lam, Álvaro Sánchez-González, Matthew Willson, Peter Wirmsberger, Meire Fortunato et al. (2023). Learning skillful medium-range global weather forecasting. *Science*. <https://www.science.org/doi/10.1126/science.adi2336>
63. Ilan Price, Álvaro Sánchez-González, Ferran Alet, Tom R. Andersson, Andrew El-Kadi et al. (2024). Probabilistic weather forecasting with machine learning. *Nature*. <https://www.nature.com/articles/s41586-024-08252-9>
64. operational since 25 February 2025. <https://www.ecmwf.int/en/about/media-centre/news/2025/ecmwfs-ai-forecasts-become-operational>
65. "the vast majority". <https://www.ecmwf.int/en/forecasts/datasets/aifs-machine-learning-data>
66. Destination Earth. <https://platform.destine.eu/climate-dt/>
67. "digital replica of the Earth system by 2030". <https://www.ecmwf.int/en/about/media-centre/news/2026/third-phase-destination-earth-confirmed>

68. Unlearn's PROCOVA. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/qualification-opinion-prognostic-covariate-adjustment-procovatm_en.pdf
69. twins of whole humans. <https://digital-strategy.ec.europa.eu/en/policies/virtual-human-twins>
70. GNoME's 2.2 million predicted crystals. <https://deepmind.google/discover/blog/millions-of-new-materials-discovered-with-deep-learning/>
71. "a list of proposed compounds". <https://escholarship.org/uc/item/9qx9t3kz>
72. Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He et al. (2023). An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*. <https://www.nature.com/articles/s41586-023-06734-w>
73. argued to be a phase known since 1971. <https://pubs.rsc.org/en/content/articlelanding/2026/mh/d6mh00268d>
74. about \$10B for Altair. <https://press.siemens.com/global/en/pressrelease/siemens-acquires-altair-create-most-complete-ai-powered-portfolio-industrial-software>
75. Jonas Degraeve, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey et al. (2022). Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*. <https://doi.org/10.1038/s41586-021-04301-9>
76. expected. <https://openai.com/index/ai-progress-and-recommendations/>
77. \$15 million. <https://www.newscientist.com/article/2588063-openai-has-solved-the-navier-stokes-millennium-problem-using-15m-of-ai-effort/>
78. \$22.5 million. <https://techcrunch.com/2026/09/08/openai-fought-dirty-on-career-making-math-problem-says-nyu-mathematician/>