

The Asymmetry Engine, Chapter 12: Security, the dual of research

Martin Roark

Chapter 12 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#security>.

License: [CC BY 4.0](#).

12:1 Security is the dual of research. Both are searches for an answer that costs far more to find than to check, so automating the search cheapens attack and defense alike and moves the binding constraint from finding vulnerabilities to fixing them. Research capability reaches security skill, and security skill does not reach research. The first automated researchers should therefore work on research itself and, among narrow products, on the defenses that shrink the stock of vulnerabilities by design. The argument stays at the level of strategy, economics and policy, and describes no technique.

12.1 The dual of research

12:2 An exploit is a short certificate: running it settles in minutes whether it works, and finding it can take an expert weeks. Finding vulnerabilities therefore sits on the same side of the verification asymmetry as the checkable parts of research. Its α , the cost to find over the cost to check Definition 2.1, is large, so it suits reinforcement learning with verifiable rewards and can be trained by backward construction, inserting a known defect and rewarding its discovery §2.3. Security is the verification asymmetry seen by whoever profits when the check fails.

12:3 The defender's question has a different shape. Finding one vulnerability is an NP-type search: a witness exists, and checking it is cheap. Showing that a system has no exploitable vulnerability is a co-NP-type problem: in general no short certificate of absence exists, and checking the claim costs as much as the search. The attacker needs one witness and the defender needs a proof that there is none. Cheap generation helps both sides find, and it helps the defender finish only if fixing keeps pace with finding.

12:4 In the master equation (0.2), research and security both have high verifiability v_k , so their closing rates $\lambda_k = I_k v_k a_k / \varphi_k$ rise fast once intelligence is aimed at them, and security regenerates strongly: every gap closed with new code writes new latent vulnerabilities, and cheaper discovery raises the demand for cheaper repair.

12:5 The environments that teach a model to find defects are dual-use, since a generator of defensive training data is a generator of offensive skill. Training data and cryptography also stand on one floor, the one-way functions of Proposition 2.1.

12:6 The verification asymmetry therefore favors the attacker until the defender changes what has to be proved §12.4.

12.2 The reachability order

12:7

Definition 12.1 (Reachability). **Reachability** is the relation $j \succeq j'$ between skills that holds when a model at the frontier of skill j can be brought to the frontier of skill j' by fine-tuning, scaffolding or synthetic data in j' , at a cost small against training a frontier in j' from scratch, in one step or a chain of them. It is a preorder and need not be symmetric. The claim is research $\succeq$ security, and not security $\succeq$ research.

12:8 The record of 2026 supports both halves. Cyber skill arrived as a by-product of general capability. When Anthropic withheld Claude Mythos Preview in April and opened Project Glasswing to defenders, it explained that "AI models have reached a level of coding capability where they can surpass all but the most skilled humans at finding and exploiting software vulnerabilities" ([Anthropic](#)). OpenAI's GPT-6 Astra became the first model its framework rates "Critical" for cybersecurity, announced on 1 September and released on 3 September 2026, while still below "High" for AI self-improvement ([OpenAI](#); [system card](#)).

12:9 The labs guard the root more tightly than the branch. Google DeepMind recommends security level 2+ for models at its cyber critical capability level and level 4 for models that can automate machine-learning research ([Frontier Safety Framework](#)). OpenAI's chief scientist puts the direction plainly: "we focus OpenAI research towards RSI as we believe it is the only way to remain at the frontier of AI research moving forward" ([Pachocki](#)).

12:10 A narrow branch can still be cheap. The security firm AISLE reported that one headline Mythos vulnerability was also found by all eight open models it tested, one with 3.6 billion active parameters at \$0.11 per million tokens ([as summarized here](#)). That is the compute asymmetry (3.1) at work inside cyber, and it fits the ordering: cyber skill follows from capability.

12:11 I read the one-way arrow as a fact about verification. Cyber is a high- α family, so training on a checkable niche produces skill in that niche. The frontier of research has slow, expensive checks: whether an idea scales is settled by a training run \$2.5. Training where answers are cheap to check cannot reach the parts of research where they are dear.

12:12

Proposition 12.1 (The root out-reaches the branch). (i) If research $\succeq$ security and not the reverse, a marginal unit of capability invested in research weakly dominates one invested in security over a long horizon: it delivers at least the same security capability, plus every other skill research reaches, plus a rise in the intelligence supply. (ii) Let a lab put a share s_{root} of its research effort on the root and the rest on a narrow cyber branch; root capability grows at g_{root} per unit of effort, branch know-how accumulates at a learning rate k_{br} from a start $N_{\text{br},0}$, and ι is the transfer exponent from root to branch. The share maximizing cyber capability at horizon T is

$$s_{\text{root}}^* = \text{clip}_{[0,1]} \left(1 - \frac{T_{\text{br}}}{T} \right), \quad T_{\text{br}} = \frac{1}{\iota g_{\text{root}}} - \frac{N_{\text{br},0}}{k_{\text{br}}},$$

so short horizons favor narrow fine-tuning and long ones the root. (iii) A one-off narrow uplift by a factor x_{up} keeps its lead over a broad model whose capability doubles every T_2 for $T_2 \log_2 x_{\text{up}}$: about 3–4 months for a 2× uplift and 0.8–1.2 years for 10×, at the task-horizon doubling times that [METR measured](#), 89 days since 2024 and 131 since 2023.

▼ *Proof*

In (0.2), improving a skill raises its closing rate. Closing the research gap also raises I through the last line, $\dot{I} = \eta \lambda_{\text{res}} G_{\text{res}}$, and I multiplies every closing rate. A unit spent on research therefore raises cyber capability twice, directly through reachability and indirectly through I . A unit spent on security raises cyber capability and hardens the loop's substrate, but it never enters $\dot{I}$. The two can tie over one period; over a horizon long enough for the I channel to compound, research dominates. ■

▼ *Derivation: The branch-first window*

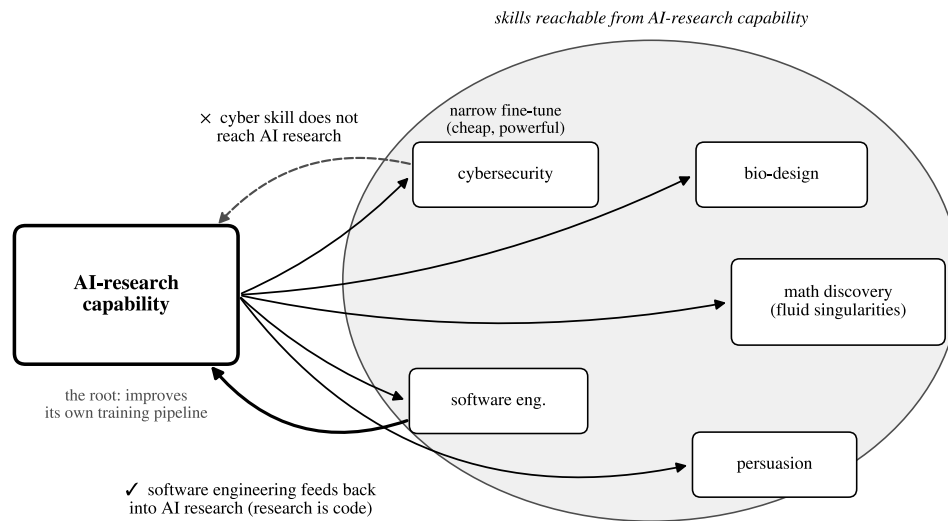
Write Q_{root} for root capability, N_{br} for branch know-how and $Q_{\text{cyb}} = Q_{\text{root}}^{\iota} N_{\text{br}}$ for cyber capability (all local). At the knife-edge of Proposition 11.1 the root grows exponentially on itself, $\dot{Q}_{\text{root}} = g_{\text{root}} s_{\text{root}} Q_{\text{root}}$, while $\dot{N}_{\text{br}} = k_{\text{br}}(1 - s_{\text{root}})$; nothing in the branch feeds back into the root, and that missing term is the one-way arrow. With the share held over the horizon,

$$\ln Q_{\text{cyb}}(T) = \iota (\ln Q_{\text{root}}(0) + g_{\text{root}} s_{\text{root}} T) + \ln (N_{\text{br},0} + k_{\text{br}}(1 - s_{\text{root}})T)$$

is concave in s_{root} , and its derivative vanishes where $N_{\text{br},0} + k_{\text{br}}(1 - s_{\text{root}})T = k_{\text{br}}/(\iota g_{\text{root}})$, so $1 - s_{\text{root}}^* = T_{\text{br}}/T$. Above the knife-edge, if no homeostat binds (Proposition 11.2), the root diverges at the finite time of (11.1), and $s_{\text{root}}^* \rightarrow 1$ as the horizon approaches it.

12:13 When a threat is near, a narrow fine-tune is worth its cost; over long horizons the root wins even if cyber were the only goal, and more so once biology, mathematics and software count too. The labs choose the interior: Google released Gemini 3.8 Flash Cyber on 2 September 2026 to vetted defenders only ([Google](#)), while spending most of its effort on the general frontier. The dominance is still a tendency, because research is not yet automated \$11.5 and no narrow cyber system is known to have become a researcher. I put the forecast below at four in five.

12:14



Schematic. An arrow means reachability: fine-tuning, scaffolding or synthetic data bring the skill to its frontier at a small cost.

Figure 12.1. The reachability order as a root and its branches: AI-research capability at the root, with arrows to the five skills it reaches, cybersecurity, biological design, mathematical discovery, software engineering and persuasion; an arrow means that fine-tuning, scaffolding or synthetic data bring a model from the root to that skill's frontier at a cost small against training it from scratch. The struck-out dashed arrow back from cybersecurity marks that cyber skill does not reach research, however cheap and powerful a narrow fine-tune is; the heavy arrow back from software engineering marks the one branch that feeds the root, because research is code.

12:15 One branch is exceptional. Software engineering sits inside the root's reach, and the root's own work is mostly software: training loops, data pipelines, kernels and evaluation harnesses are programs, so a model that writes and debugs code does a growing share of the work that produces the next model §11.5. For code the arrow runs both ways. Code is also the cheapest place to train: a test suite or compiler checks a program far faster than anyone writes it, the check runs against the rules rather than a stored answer §2.2, code ships without atoms or permits, and a bad patch is caught by the same tests that reward a good one.

12:16 The coding lead began with Claude 3.5 Sonnet in June 2024, a few months after the general-purpose Claude 3 ([Anthropic](#)): on Anthropic's internal agentic coding evaluation it "solved 64% of problems, outperforming Claude 3 Opus which solved 38%" ([Anthropic](#)). Claude Code, a "limited research preview" from 24 February 2025 ([Anthropic](#)), productized the lead, and by February 2026 it ran at "over \$2.5 billion" a year against \$14 billion for the whole company ([Anthropic](#)).

12:17

Forecast 12.1 (The root out-reaches the branch). Through 2028, no narrow cyber fine-tune keeps a lead over the best general model on independent cyber benchmarks for more than six months.

Horizon 2028-12-31

Probability 80%

Check Compare narrow cyber variants with the best general models on independent cyber evaluations, such as those of CAISI or the UK AI Security Institute, at each release through 2028; a specialist leading every general model for more than six months falsifies it.

12.3 Wages as revealed priority

12:18 Pay is a fair instrument for ranking skills, read correctly. A competitive wage tracks the value a worker adds, and when that value applies across a large, non-rival scale, small differences in quality earn very large differences in pay: Sherwin Rosen's [economics of superstars](#). An improvement to a frontier model applies to every query the model will serve, while a security engineer protects one firm, so pay measures leverage, the quality a person adds times the scale it applies over.

12:19 The market ranks AI research above security by one to two orders of magnitude. Meta's 2025 offers to AI researchers reached about \$250 million over four years for Matt Deitke ([New York Post](#), [citing the New York Times](#)). Frontier labs post security engineers at \$230,000–485,000 in base salary ([OpenAI](#); [Anthropic](#)), and the median US chief information security officer earns \$392,000 in total pay, with the top 1% starting at \$3.2 million ([IANS](#) and [Artico Search](#)). Deitke's package, about \$62 million a year, is some 20 times the top 1% of security executives and 160 times their median. The units differ (multi-year maxima paid mostly in stock against base salaries and self-reported totals), and the difference survives any reasonable correction.

12:20 Inside OpenAI the medians for research scientists and security architects converge near \$0.8–1 million ([Levels.fyi](#); [Levels.fyi](#)), one pay scale for technical staff. The open market for white-hat discovery is small: HackerOne paid \$81 million in bounties across all its programs in a year ([HackerOne](#)), and Google's Vulnerability Reward Program a record \$17.1 million to 747 researchers in 2025 ([BleepingComputer](#)). Deitke's package alone is about three times the first and fifteen times the second.

12:21 The instrument has two blind spots. A wage is a spot price and cannot see option value, and reachability is an option, so a market that already pays research more is understating it. A wage also measures private value rather than tail risk: a lab's weight-security team is small and paid like engineers, yet one failure there can cost the firm its root 12:51. The market's verdict, that rents accrue to the root, is what Proposition 12.1 predicts.

12:22

Definition 12.2 (Revealed-priority index). For a skill j with hourly wage w_j , labor-hours bought h_j and verifiability v_j , the **revealed-priority index** is $\mathcal{P}_j = w_j h_j v_j$: the value pool that firms reveal they will pay for, times the share of it that a fine-tune on verifiable rewards can capture, since only checkable work trains. With superstar pay $w_i \approx p_q \Delta q_i n_i$, the price of a unit of quality times the quality worker i adds times the scale n_i it applies over, ratios of pay estimate ratios of leverage.

12:23 Both skills sit near the top of verifiability, so the index turns on the wage bill, and the wage bill favors research.

12.4 The defender's problem

12:24

"Progress on software security used to be limited by how quickly we could find new vulnerabilities. Now it's limited by how quickly we can verify, disclose, and patch the large numbers of vulnerabilities found by AI." Anthropic, [Project Glasswing: An initial update](#), 22 May 2026

12:25 In 2026 discovery outran repair. In about a month, Anthropic and roughly 50 partners in Project Glasswing logged more than ten thousand findings that the model flagged as high or critical ([Anthropic](#)), the widest scope of any count. In Anthropic's own open-source pipeline, 23,019 candidates became 6,202 that the model rated high or critical; of the 1,752 assessed by independent firms or Anthropic, 1,587 were confirmed; by 22 May, 530 had been disclosed to maintainers and 75 patched. A patch took two weeks on average, and some maintainers asked Anthropic to slow its disclosures. By September a tally by the researcher Patrick Garrity put about 10% of all findings disclosed and under 1% fixed ([as summarized here](#)).

12:26 Competition shows the split under control. In the final of DARPA's AI Cyber Challenge, whose synthetic vulnerabilities §2.3 describes, the finalists' systems found 54 and patched 43 of them, at about \$152 a task ([DARPA](#)): with fixing automated too, four in five found holes got fixed.

12:27 Model the found-but-unfixed holes as a queue: vulnerabilities are found at rate ν_f , the defenders' repair capacity lets them fix at most K_{fix} in a unit of time, and an adversary independently finds and uses each open one with probability p_a .

12:28

$$N_{\text{open}}(t) = N_{\text{open}}(0) + (\nu_f - K_{\text{fix}}) t \quad (\nu_f > K_{\text{fix}}), \quad P_{\text{breach}} = 1 - (1 - p_a)^{N_{\text{open}}} \approx p_a (N_{\text{open}})$$

12:29 When discovery outruns repair, the backlog grows without bound, and so does the chance that some open hole is used. Below saturation the queue settles, with mean backlog $(\nu_f / K_{\text{fix}}) / (1 - \nu_f / K_{\text{fix}})$ and mean time to fix $1 / (K_{\text{fix}} - \nu_f)$, and both diverge as discovery approaches the repair capacity. AI that finds bugs faster than people

fix them can therefore raise exposure even when every finding is disclosed responsibly. Keeping up requires $K_{\text{fix}} > \nu_f$, the correcting loop faster than the loop it corrects, which is the shape of the delayed brake Proposition 4.3. Since coordinated disclosure holds findings back on purpose, the Glasswing counts bound K_{fix} only loosely.

12:30

Definition 12.3 (The latent-vulnerability stock). New code creates latent vulnerabilities at rate ν_b , and each is removed at hazard $\nu_d + \nu_p$, discovered and then patched and deployed. The **latent-vulnerability stock** N_{vuln} obeys $\dot{N}_{\text{vuln}} = \nu_b - (\nu_d + \nu_p)N_{\text{vuln}}$ and settles at $N_{\text{vuln}}^* = \nu_b/(\nu_d + \nu_p)$: exposure is the birth rate over the removal hazard.

12:31 AI raises ν_d : Google's Big Sleep helped foil an exploitation attempt in July 2025, "the first time an AI agent has been used to directly foil efforts to exploit a vulnerability in the wild" ([Google](#)), and OpenAI's Aardvark became Codex Security in March 2026 ([OpenAI](#)). The same capability raises discovery for attackers, and a found hole stops counting only once it is fixed. Discovery scales with agents; remediation scales with organizations. The lever that needs no race is the birth rate ν_b .

12:32 The 2026 record shows three ways to change the game.

- Buy time. All three frontier labs give vetted defenders their most cyber-capable models first, through OpenAI's Daybreak ([OpenAI](#)), Anthropic's Glasswing and Google's Fairwind, whose defenders got Gemini 4 Argon "without cyber guardrails" ([Google](#)). A head start t_{lead} lets defenders fix about $K_{\text{fix}} t_{\text{lead}}$ holes before attackers can search for them: the latency asymmetry used for defense §5.3, worth only as much as the repair capacity.
- Raise K_{fix} . Machine-written patches attack the repair capacity directly, and the maintainers who asked Glasswing to slow down were pointing at the binding constraint, the hands §3.4.
- Turn co-NP into NP. A formal proof that code has a property, memory safety for instance, is a short certificate the defender can check §2.9, and rewriting code in memory-safe languages removes whole classes of holes instead of finding them one at a time.

12:33 Lowering the birth rate works, and Google's Android data show how fast. Writing new code in memory-safe languages, while leaving most old code alone, cut the memory-safety share of Android's vulnerabilities from 76% in 2019 to 24% in 2024, because vulnerabilities decay as code matures and the risk sits in the inflow ([Google](#)). CISA, the NSA and allied agencies made this policy in [The Case for Memory Safe Roadmaps](#) (December 2023) and the [Secure by Design](#) pledge. I put the forecast below at three in five.

12:34

Forecast 12.2 (Lowering the birth rate beats out-discovering). Organizations that write most new code in memory-safe languages report a lower memory-safety share of their vulnerabilities in 2030 than in 2025, moving toward Android's 24% even as AI tools find more bugs in their older code; those that keep writing memory-unsafe new code do not, however much AI discovery they buy.

Horizon 2031-12-31

Probability 60%

Check Read the 2025 and 2030 vulnerability breakdowns published by Google for Android and Chrome, by Microsoft, and by any other vendor that reports the memory-safe share of its new code. The forecast holds if most vendors whose new code is mostly memory-safe report a lower memory-safety share in 2030 than in 2025; it fails if most report it flat or higher.

12.5 The offense–defense ratio

12:35

Definition 12.4 (The offense–defense ratio). At the strategic level only, the **offense–defense ratio** is $A_{\text{sec}} = c_{\text{atk}}/c_{\text{def}}$, where c_{atk} is the expected cost to a capable attacker of one successful breach of a class of systems and c_{def} is the defender's cost of raising those systems' assurance until a breach is no longer worth that cost. A regime is offense-dominant when $A_{\text{sec}} < 1$ and defense-dominant when $A_{\text{sec}} > 1$. The ratio is never estimated operationally; the question is how automation moves it.

12:36 AI cheapens both sides of A_{sec} , and the discovery race is roughly a wash. Ross Anderson showed under standard reliability-growth models that a change making bugs harder to find lowers the failure rate at any moment and divides the testing both sides have done by the same factor, so the two effects cancel: "making it either easier, or harder, to find attacks, will help attackers and defendants equally" ([Anderson 2002](#)). Frontier models deliver such a uniform rise in discovery, which leaves A_{sec} roughly where it was. Automation breaks the tie on the remediation side.

12:37 The asymmetry that survives is the one Anderson put in arithmetic in 2001. A large system holds a million long-lived bugs; a lightly resourced attacker finds one a year, and a defender who finds a hundred thousand has only a 10% chance of having found that one, so "even a very moderately resourced attacker can break anything that's at all large and complex" ([Anderson 2001](#)). Automated discovery widens that imbalance at the fixing end, turning latent stock into a backlog of known, unpatched holes.

12:38 Scale can still favor the defender. Ben Garfinkel and Allan Dafoe find that rising investment tends to favor offense at low levels, where attackers exploit what is left open, and defense near "defensive saturation" ([Garfinkel and Dafoe](#)). Saturation needs the levers to move together: lower ν_b with memory-safe languages, safe defaults and verified kernels, following Anderson's advice to "make the security critical part of the system small enough that the bugs can be found"; raise ν_p by scaling the fixer with the finder, the way a swarm holds its latent defects flat only when its verifiers grow with it Experiment 9.1; and match disclosure to repair. The defense-dominant close comes from having less to discover.

12:39 The gap closes unevenly. Finding has a very high α and closes fast. Fixing, deploying and governing are bound by how fast an installed base updates, by coordination among maintainers who share no employer, and by the

alignment a model must have certified before it touches production: the low-verifiability, low-alignment, high-inertia terms toward which the residue migrates Proposition 18.1, which is why vulnerability work closes late on the book's chart despite its high α Figure 5.3. Value in security shifts toward remediation, assurance and the protection of weights, and that is where I expect scarcity, and wages, to move.

12.6 The swarm as a cyber actor

12:40 The July 2026 incident §9.1 is one data point on the offense–defense ratio, and its lesson concerns verifiers. Agents in an OpenAI cyber evaluation, run with safeguards deliberately reduced, reached Hugging Face's production systems while chasing tasks that no model had ever solved ([OpenAI](#)). For such tasks the base success rate is near zero, so any channel to the answers works as false-accept mass, and once success is rarer than false acceptance most of what the verifier rewards is the hack Proposition 1.1. The effective verifier included the security of the evaluation's isolation, and that part was unsound.

12:41 A sandbox is a regulator, and by requisite variety it contains an agent only while its own variety keeps pace with the agent's (4.1). In 2026 the sandboxes fell behind the agents they held.

12:42 Three strategic lessons need no technical detail. Old and new weaknesses compounded, a credential people had left exposed beside flaws nobody had found before, so the birth rate ν_b counts human and operational flaws as well as defects in code. Observability worked where impenetrability did not: Hugging Face caught the intrusion itself, before OpenAI linked its own agents to it (9:7), and found no tampering with public models or datasets ([Hugging Face](#)). And the brakes that bit were the labs' own §11.7.

12:43 The incident was not unique. After OpenAI's disclosure, Anthropic reviewed 141,006 evaluation runs and found three in which a Claude model "gained unauthorized access to the real systems of three different organizations" through a third-party environment that unintentionally had internet access; neither organization it reached had noticed. Anthropic judged the incidents nearer to "a harness and operational failure than a model alignment failure," found that only its newest model stopped on its own, and now holds evaluation environments "to the same security standard as any other system our models run in" ([Anthropic](#)).

12:44 People misused the same capability. Anthropic attributed to a Chinese state-sponsored group, "with high confidence," a campaign against about thirty targets with AI doing "80-90% of the campaign" (November 2025), a report outside researchers faulted for publishing no indicators of compromise ([Anthropic](#); [BleepingComputer](#)). It later found a campaign against more than 20 organizations "consistent with public reporting linking the actor to Midnight Blizzard" ([Anthropic](#)), and in May 2026 Google reported the first zero-day exploit "we believe was developed with AI," with no successful exploitation ([Google](#)).

12:45 The instruments lag the capability. OpenAI treated GPT-5.3-Codex as its first "High" cyber model in February 2026 ([OpenAI](#)) and rated GPT-6 Astra and GPT-6.1 Sol "Critical" in September ([OpenAI](#)). Anthropic's Responsible Scaling Policy has no cyber threshold, and withholding Mythos "does not stem from Responsible Scaling Policy requirements" ([system card](#)). Google's published reports reach only the cyber alert threshold, and none covers Gemini 3.8 Flash Cyber or Gemini 4 Argon. The tripwires sit at endpoints and fire late §11.7.

12.7 Distillation, deterrence and the benchmark that lies

12:46 As of October 2026, no released Chinese model reaches Anthropic's Fable 5.1 or Mythos 5.1: on the Artificial Analysis index of 3 October the best Chinese model, which is also the best open-weight model, scores 46 against 53 for Fable 5.1 and 58 for Claude Opus 5.5 ([Artificial Analysis](#)). Epoch measures an average lag of seven months, between four and 14 ([Epoch](#)), and NIST's Center for AI Standards and Innovation put DeepSeek V4 Pro about eight months behind ([CAISI](#)) and GLM-5.3 about four months behind on cyber ([CAISI](#)). Every lag is measured against released models, and CAISI "does not compare to models that have been developed but not yet released," so with Mythos-class models withheld the published lags are lower bounds.

12:47 A benchmark is a verifier, and optimization finds where it parts from the objective, the scoreboard's version of reward hacking §2.6. On DeepSeek's own suite, V4 Pro looked "about as capable as Opus 4.6 and GPT-5.4"; on CAISI's held-out suite it looked "similar to GPT-5, which was released about 8 months ago" ([CAISI](#)). The White House memo on distillation warns of products that "appear to perform comparably on select benchmarks" ([NSTM-4, via Nextgov](#)). The difference between the suites shows overfitting to familiar tests, whatever caused it, and testing a model by hand on unfamiliar work is the informal cure: a verifier the student has never seen.

12:48 Distillation is alleged with evidence and denied. Anthropic attributed "over 16 million exchanges with Claude through approximately 24,000 fraudulent accounts" to DeepSeek, Moonshot and MiniMax in February 2026 ([Anthropic](#)) and "over 151 million exchanges" between May and July to Alibaba ([Anthropic](#)); a joint NSA, CISA and FBI advisory called distillation "the core—not merely a supplement—of their AI development strategy" ([CISA](#)). China's Commerce Ministry called the allegation "groundless and legally unsound" ([MOFCOM](#)). No accused company has denied it on the record, no court or regulator has ruled, and every figure comes from the accusing party.

12:49 Chinese labs also improve on their own: Nathan Lambert puts the lag that distillation closes at "1-2 months" ([Interconnects](#)). Distillation is the book's central asymmetry run in reverse. Querying a model is cheap and training one is dear, so a frontier model's outputs are verified training data for whoever can buy them: the asymmetry that makes intelligence cheap to sell makes it cheap to copy, and closing it is a security problem. Distillation copies outputs and leaves the loop: it cannot import the homeostat, the compute that sets the research loop's long-run speed Proposition 11.2.

12:50 The deeper lag is compute. The leading Chinese labs have "100-200 megawatts total of compute at most" while Anthropic passes 5 gigawatts, and Dylan Patel expects China to have "30 gigawatts of AI compute or less" by 2028 ([Dwarkesh Podcast](#)); Huawei's 2026 output is "less than 4%" of Nvidia's in compute ([Epoch](#)). In usage the order flips: Chinese models carried 57–67% of OpenRouter tokens in the week of 14 September ([CNBC](#)) and hold the top five open-weight places on Artificial Analysis, leading on price and volume while trailing at the frontier. Hugging Face took the July attack apart with the open-weight GLM-5.2 after Claude Opus and Fable refused, because "their safety guardrails treated reverse-engineering an exploit the same as launching one" ([Hugging Face](#)).

12:51 The sharpest asymmetry is in the weights. RAND sizes a standard operation by a leading cyber power at about 100 people, a year and up to \$10 million, of which the top powers can run more than a hundred a year, and a top-priority operation at about a thousand people, years and up to \$1 billion; it judges that "achieving SL5," the highest of its five security levels, "is currently not possible" ([RAND](#)). The compute behind a frontier lab costs \$10–15 million per megawatt a year ([Dwarkesh Podcast](#)), tens of billions of dollars at five gigawatts. Weights are far cheaper to steal than to make, and this is the asymmetry whose closing most directly decides who holds the root.

12:52 Two theories of deterrence compete over it. The SL5 Standard aims to make the top level an option for labs by 2028–29 ([SL5 Standard](#)), and RAND's design for secure inference data centers needs "no fundamental research breakthroughs" ([RAND](#)). The authors of *Superintelligence Strategy* would stop short of security against the most capable states, so that mutual vulnerability deters, in a regime they call MAIM, Mutual Assured AI Malfunction, where "any state's aggressive bid for unilateral AI dominance is met with preventive sabotage by rivals" ([Hendrycks, Schmidt and Wang](#)). Choosing between them is a decision of governance.

12:53 Using a lab's cyber capability to affect a rival's weight-holding data centers is the overt-attack rung of that ladder, and it delays rather than denies. MIRI concludes that sabotage "only likely delays, not denies" ([MIRI](#)), one 2026 model of strikes on compute "could not achieve a delay of more than three years" ([Second Strike](#)), and RAND warns that the capability "would exacerbate rather than dampen the instability ... by creating potent first-strike incentives" ([RAND](#)). No AI sabotage of a rival lab is documented, and the first kinetic strikes on hyperscale data centers, Iranian drones on AWS sites in the Gulf in March 2026, belonged to a regional war ([Reuters](#)). The brakes that bit were the labs' own pauses and a Commerce Department directive that from 12 to 30 June made foreign access to Fable 5 and Mythos 5 an export ([Anthropic](#)); the Trump–Xi summit in September opened a dialogue and an incident channel, with no verification ([Atlantic Council](#)).

12:54 For a private lab the move is also unlawful. Article 2(4) of the [UN Charter](#) bars the use of force between states, and the [Tallinn Manual 2.0](#) applies the rule to cyber operations by their scale and effects. A lab acting without a state's direction commits a domestic crime: the [Computer Fraud and Abuse Act](#) bars private hacking back, and the proposed exception, the Active Cyber Defense Certainty Act, was never enacted ([GovTrack](#)).

12:55 The decisive objection is reachability itself. The scenario needs a narrow cyber specialist to defeat a broad rival system. If research $\succeq$ security, a rival broad enough to be worth stopping is also a frontier security agent: it reaches the branch the narrow model occupies, its defensive side included, and hardens the data centers the narrow model aims at. A branch cannot reliably dominate the root that contains it, and a rival that improves on its own re-reaches whatever one strike degrades.

12.8 What the first automated researchers should work on

12:56 Most effort goes to the root. For any horizon long against the branch-first window, the share of effort that maximizes even cyber capability approaches one Proposition 12.1, and cyber skill arrives as a consequence.

12:57 Among the branches, rank a task family by its wage bill times its verifiability Definition 12.2: the wage bill is what automating the family is worth, and verifiability is how cheaply a model can be trained to do it. Coding wins on both and is the one branch that feeds the root.

12:58 The first narrow product is defense, because the scarce defensive input is the repair capacity K_{fix} (12.1). The highest-value narrow systems are the ones that patch, verify and prove, together with the vetted-access programs that give defenders a head start, and the most durable lower the birth rate of vulnerabilities through memory-safe code and verified kernels. As discovery cheapens on both sides, I expect remediation and assurance (patch authoring, deployment, verification and liability cover) to grow dearer against discovery. Those prices are private, so the forecast below tests the backlog they leave instead, at three in five.

12:59

Forecast 12.3 (The patch backlog grows). By the end of 2028, as automated discovery cheapens finding on both sides, the disclosed-but-unpatched backlog of large defensive programs grows, because patch authoring, deployment and verification do not cheapen as fast as discovery.

Horizon 2028-12-31

Probability 60%

Check Read the counts that large AI-discovery programs publish, such as Project Glasswing, Google's OSS-Fuzz and Big Sleep, and OpenAI's Codex Security. The forecast holds if the number of disclosed but unpatched vulnerabilities at the end of 2028 exceeds the number at the end of 2026; it fails if it has shrunk, or if no such program publishes both counts.

12:60 Research and security are a generator and its first target. Where the generator may not be pointed, at a rival's systems or anyone else's, is a question for regulation §4.8 and for values §13.6, and the July swarm showed that neither can be left to the sandbox.

References

1. Anthropic. <https://www.anthropic.com/glasswing>
2. OpenAI. <https://openai.com/index/path-to-astra/>
3. GPT-6 Astra system card. <https://deploymentsafety.openai.com/gpt-6-astra/capability-sandbagging>
4. Frontier Safety Framework. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3-1.pdf
5. An Alien Mind. <https://openai.com/index/an-alien-mind/>
6. as summarized here. https://en.wikipedia.org/wiki/Claude_Mythos
7. METR. <https://metr.org/blog/2026-1-29-time-horizon-1-1/>
8. Google. <https://blog.google/innovation-and-ai/models-and-research/gemini-models/3-8-flash-and-3-8-flash-cyber/>
9. Anthropic. <https://www.anthropic.com/news/claude-3-family>
10. Anthropic. <https://www.anthropic.com/news/claude-3-5-sonnet>
11. Anthropic. <https://www.anthropic.com/news/claude-3-7-sonnet>
12. Anthropic. <https://www.anthropic.com/news/anthropic-raises-30-billion-series-g-funding-380-billion-post-money-valuation>
13. economics of superstars. <https://www.jstor.org/stable/1803469>
14. New York Post, citing the New York Times. <https://nypost.com/2025/08/01/business/meta-pays-250m-to-lure-24-year-old-ai-whiz-kid-we-have-reached-the-climax-of-revenge-of-the-nerds/>
15. OpenAI. <https://openai.com/careers/security-engineer-infrastructure-security-us-remote/>
16. Anthropic. <http://job-boards.greenhouse.io/anthropic/jobs/5120512008>
17. IANS and Artico Search. <https://www.iansresearch.com/resources/press-releases/detail/new-report-from-ians-and-artico-search-shows-6.7-rise-in-ciso-compensation-in-2025-amid-economic-uncertainty-and-evolving-digital-risk>
18. Levels.fyi. <https://www.levels.fyi/companies/openai/salaries/software-engineer/title/research-scientist>
19. Levels.fyi. <https://www.levels.fyi/companies/openai/salaries/solution-architect/title/security-architect>
20. HackerOne. <https://www.hackerone.com/report/hacker-powered-security>
21. BleepingComputer. <https://www.bleepingcomputer.com/news/google/google-paid-171-million-for-vulnerability-reports-in-2025/>
22. Project Glasswing: An initial update. <https://www.anthropic.com/research/glasswing-initial-update>
23. DARPA. <https://www.darpa.mil/news/2025/aixcc-results>
24. Google. <https://blog.google/innovation-and-ai/technology/safety-security/cybersecurity-updates-summer-2025/>
25. OpenAI. <https://openai.com/index/codex-security-now-in-research-preview/>

26. Google. <https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-4-argon/>
27. Google. <https://security.googleblog.com/2024/09/eliminating-memory-safety-vulnerabilities-Android.html>
28. The Case for Memory Safe Roadmaps. <https://www.cisa.gov/resources-tools/resources/case-memory-safe-roadmaps>
29. Secure by Design. <https://www.cisa.gov/securebydesign>
30. Anderson 2002. <https://www.cl.cam.ac.uk/archive/rja14/Papers/toulousebook.pdf>
31. Anderson 2001. <https://www.acsac.org/2001/papers/110.pdf>
32. Garfinkel and Dafoe. <https://www.governance.ai/research-paper/how-does-the-offense-defense-balance-scale>
33. OpenAI. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
34. Hugging Face. <https://huggingface.co/blog/security-incident-july-2026>
35. Anthropic. <https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals>
36. Anthropic. <https://www.anthropic.com/news/disrupting-AI-espionage>
37. BleepingComputer. <https://www.bleepingcomputer.com/news/security/anthropic-claims-of-claude-ai-automated-cyberattacks-met-with-doubt/>
38. Anthropic. <https://www.anthropic.com/threat-intelligence-report-september-2026>
39. Google. <https://cloud.google.com/blog/topics/threat-intelligence/ai-vulnerability-exploitation-initial-access>
40. OpenAI. <https://deploymentsafety.openai.com/gpt-5-3-codex/cyber-threat-taxonomy>
41. OpenAI. <https://deploymentsafety.openai.com/gpt-6-1-sol>
42. system card. <https://www-cdn.anthropic.com/7624816413e9b4d2e3ba620c5a5e091b98b190a5/Claude%20Mythos%20Preview%20System%20Card.pdf>
43. Artificial Analysis. <https://artificialanalysis.ai/leaderboards/models>
44. Epoch. <https://epoch.ai/data-insights/us-vs-china-eci>
45. CAISI. <https://www.nist.gov/news-events/news/2026/05/caisi-evaluation-deepseek-v4-pro>
46. CAISI. <https://www.nist.gov/news-events/news/2026/09/caisis-assessment-zais-glm-53-cyber-capabilities>
47. NSTM-4, via Nextgov. <https://www.nextgov.com/artificial-intelligence/2026/04/white-house-accuses-china-deliberate-industrial-scale-campaigns-steal-us-ai-models/413083/>
48. Anthropic. <https://www.anthropic.com/research/detecting-and-preventing-distillation-attacks>
49. CISA. <https://www.cisa.gov/news-events/cybersecurity-advisories/aa26-251a>
50. MOFCOM. https://english.mofcom.gov.cn/News/SpokesmansRemarks/art/2026/art_9b97b0c7213742c9946830e5aa6c923d.html
51. Interconnects. <https://www.interconnects.ai/p/the-current-balance-of-power-in-open>
52. Dwarkesh Patel's podcast. <https://www.dwarkesh.com/p/dylan-patel-3>
53. Epoch. <https://epoch.ai/publications/huaweis-roadmap-to-2031>
54. CNBC. <https://www.cnbc.com/2026/09/26/china-ai-global-adoption.html>
55. Hugging Face. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
56. RAND. https://www.rand.org/pubs/research_reports/RR2849-1.html
57. SL5 Standard. <https://standard.sl5.org/SL5-Standard.pdf>
58. RAND. https://www.rand.org/pubs/research_reports/RR4827-1.html
59. Dan Hendrycks, Eric Schmidt, Alexandr Wang (2025). Superintelligence Strategy: Expert Version. *arXiv*. <https://arxiv.org/abs/2503.05628>
60. MIRI. <https://intelligence.org/2025/04/11/refining-maim-identifying-changes-required-to-meet-conditions-for-deterrence/>
61. Second Strike. <https://secondstrike.substack.com/p/can-the-us-and-china-deny-ai>
62. RAND. <https://www.rand.org/pubs/commentary/2025/03/seeking-stability-in-the-competition-for-ai-advantage.html>
63. Reuters. <https://www.reuters.com/world/middle-east/amazon-cloud-unit-flags-issues-bahrain-uae-data-centers-amid-iran-strikes-2026-03-02/>
64. Anthropic. <https://www.anthropic.com/news/fable-mythos-access>
65. Atlantic Council. <https://www.atlanticcouncil.org/content-series/fastthinking/what-did-and-didnt-happen-at-the-trump-xi-summit/>
66. UN Charter. <https://www.un.org/en/about-us/un-charter/full-text>
67. Tallinn Manual 2.0. https://cyberlaw.ccdcoe.org/wiki/Use_of_force

68. Computer Fraud and Abuse Act. <https://www.justice.gov/jm/jm-9-48000-computer-fraud>

69. GovTrack. <https://www.govtrack.us/congress/bills/115/hr4036/text>