

Running Through the Rain: Speed, Secrecy and Risk in the AI Cyber Transition

A plain guide to speed, secrecy and risk as
AI learns to break code (illustrated edition)

ALLAN C. STAM · UNIVERSITY OF VIRGINIA

▼ CONTENTS

1. Part 1. The question: should we rush through a risk peak?
2. Part 2. The case for speed needs four things at once
3. Part 3. What the paper adds is a second clock
4. Part 4. Does speed help once fixing is the bottleneck?
5. Part 5. A peak can exist, but the peak model has no AI in it
6. Part 6. Life-saving benefits give speed a second reason, with conditions
7. Part 7. AI is advancing fast on tasks that are cheap to check
8. Part 8. AI can find flaws fast; fixes still run on the calendar
9. Part 9. Holding a capability back helps defenders only for a while
10. Part 10. A held capability earns in four places, and each leaks differently
11. Part 11. Actors race anyway, and restraint needs detection
12. Part 12. The models settle three things; the rest is conditional
13. Part 13. Every result rests on conditions, and the deciding numbers are unmeasured
14. Part 14. Three dated predictions could count against the paper
15. Part 15. The bottom line: others build the curb, and running thickens the rain
16. Want to learn more?
17. Bibliography

Illustrated edition. This version retells the published plain-language text in shorter sentences and adds figures. Every change was checked against the published text and the

technical sources, and every figure against the numbers it shows. Figures marked “illustrative” use made-up values to show a shape.

This is the plain-language companion to the technical paper “Running Through the Rain: Speed, Secrecy and Risk in the AI Cyber Transition.” Everything here comes from that paper and its supplement. Where the paper is unsure, so is this guide.

If you have to cross a street in the rain, should you walk or run?

Run. Physicists have worked it out. Running cuts the time rain falls on your head. And you pass through the same curtain of rain in front of you at any speed. The one exception is a tailwind.¹

Some people think AI and computer security work the same way. AI that can find and exploit flaws in software may help attackers first. Later it may help defenders, by fixing code and proving it correct. So, the argument goes, sprint through the dangerous middle.

The paper asks whether the rain picture holds. It finds two places where it breaks (Figure 1).

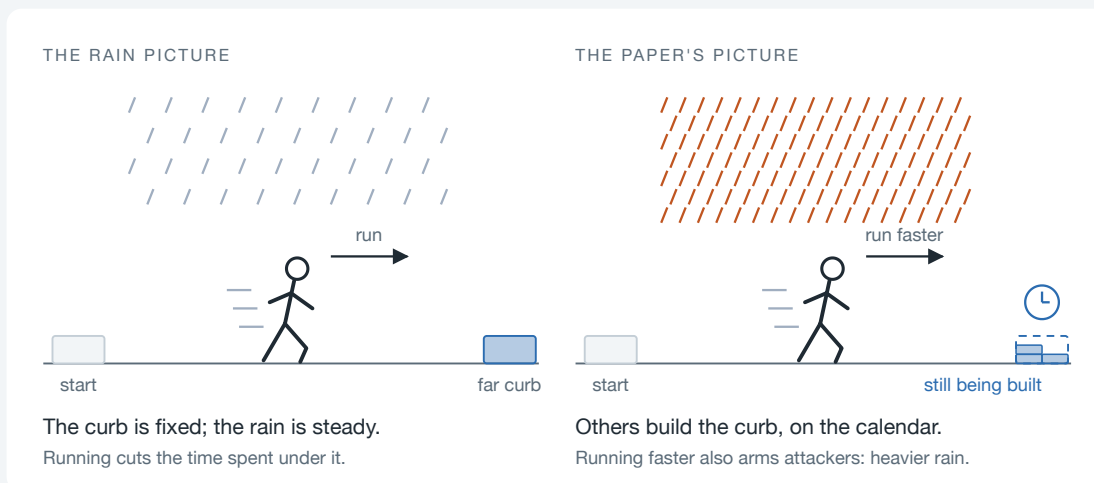


Figure 1. Left: the rain picture, where the far curb is fixed and the rain is the same at any speed, so running is better. Right: the paper’s picture, where operators build the curb on calendar time and running faster also thickens the rain. A drawing of the analogy, not of data.

First, the far curb is built by other people. Safety arrives when the people who run software patch it, retire it or rewrite it. They do that on calendar time, not on AI's schedule. Three separate records describe known patches and migration milestones (Figure 2):

- One lab's disclosure program reported 6,157 vulnerabilities disclosed across 591 open-source projects as of 2 October 2026. "To our knowledge," the lab added, "516 of these have been patched."² That is the lab's own dated count. It does not measure how fast fixes reach the computers that need them.
- In June 2019 the U.S. Government Accountability Office named ten critical federal legacy systems most in need of modernization. As of February 2025, agencies reported three of the ten done.³
- Native IPv6 access to Google first passed 50% on a single day, 28 March 2026, about 27 years after the December 1998 specification.⁴

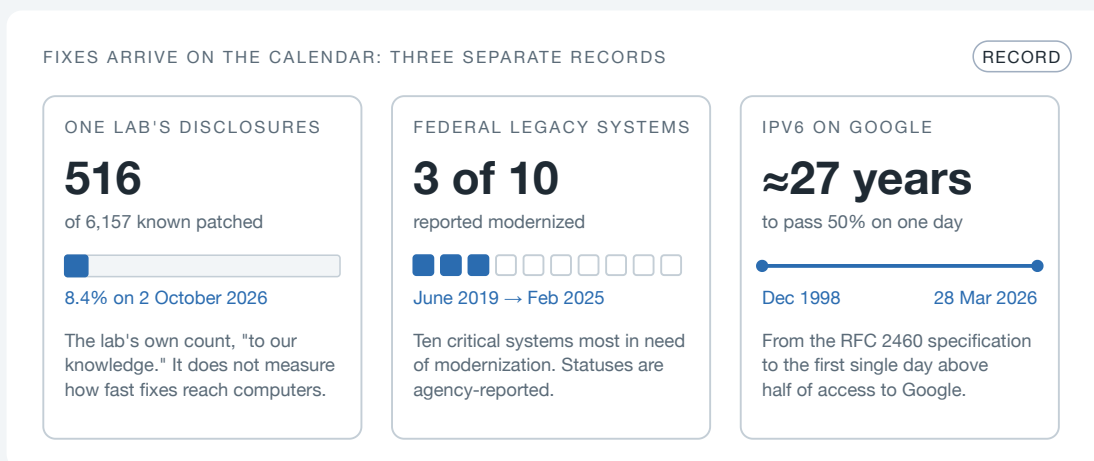


Figure 2. Three separate records: the lab's known patch count and two migration milestones. They measure different things and are not steps of one ladder. The first is the lab's own dated count of patches it knows of, not a measure of how fast fixes reach computers.

Second, running thickens the rain. The paper's models put attackers and defenders on one AI clock. Faster AI arms both at once.

From those two breaks comes the paper's main result about speed, and it is narrow. In the model, going faster only adds risk once the time it takes to fix things, not AI's pace, sets how long the danger lasts. That holds under conditions this guide spells out. Before that point, going faster can help.

Nobody has measured which side of that point we are on.

The paper also asks whether a firm should build the capability, hold it back and share it only with trusted defenders. And it asks what pushes firms to release what they hold. Those answers are conditional too. A new part asks whether life-saving benefits from the same AI change the speed answer (Part 6).

PART 1

The question: should we rush through a risk peak?

The larger road is the one to artificial superintelligence, or ASI: AI that outperforms the best humans at nearly all thinking work.⁵ On that road, are there **intermediate peaks** in risk, stretches where danger climbs and then falls? If so, should we get through them faster?

The obvious candidate is the **cyber transition**, when AI can write and break code. The paper takes five questions:

1. **Is there a peak at all?** Or does risk level off, or keep climbing?
2. **How do we get past it?** Run faster, lower the peak, or hold the storm back? Does clearing the near peak raise a far one?
3. **Secrecy for security.** Should a firm race to the capability, hold it, and share it only with trusted defenders, knowing sharing risks leaks?
4. **Secrecy and money.** What does a held capability earn, what leaks with each use, and how does that shape holding and racing?
5. **The biggest peak.** Is ASI itself the worst stretch, and do the other answers depend on it? The paper answers this only conditionally. Its test assumes the far risk rises with speed.⁶

— Wait, isn't this just the old argument that AI helps hackers?

Not quite. The paper takes “AI helps attackers first” as a starting assumption and asks a different question: *given* a dangerous stretch, does speeding through it

help? The answer turns on a second clock, the time it takes to deploy fixes, that speed arguments usually leave out.

The paper builds four separate models and never adds them up

The paper builds four small, separate models, each with its own goal and yardstick:⁷

- **Speed and adoption** counts a social risk tally: expected harmful events.
- **Holding the storm** counts exposure time saved for defenders.
- **The race** counts each racer's expected payoff.
- **The holder's choices** counts one firm's private value, in money terms.

The paper never adds these up. Exposure time saved is not harm avoided, and neither is the same as overall welfare. No ranking of policies falls out of the paper.

The paper also extends the first model with health benefits that arrive with the same AI (Part 6). That extension counts health gains and cyber harm in one unit, healthy years of life, for that extension only. It too is never added to the other models.⁸

“In the model” in this guide means one of these four, on its own terms. Later figures carry a small badge, Model 1 to Model 4, so you can tell which one a picture belongs to (Figure 3).

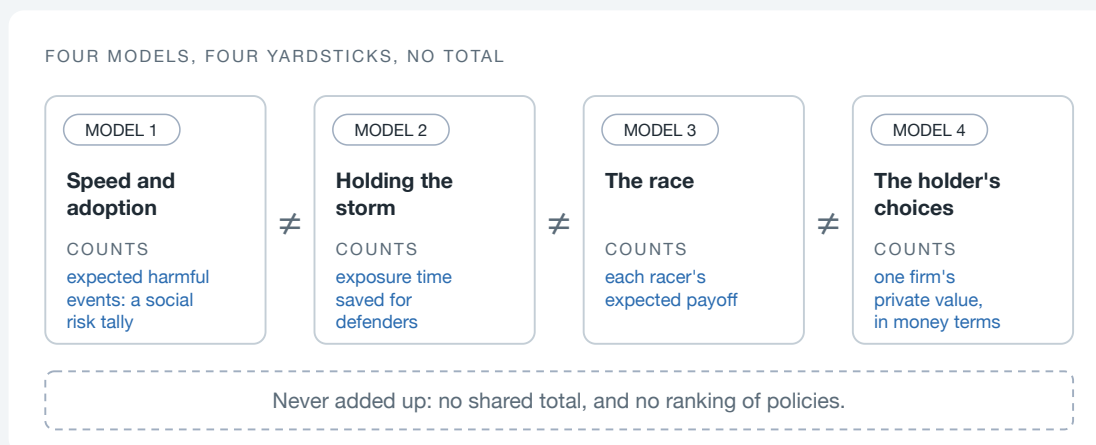


Figure 3. The paper's four models, each with its own yardstick. They are never added into one total, and no ranking of policies follows from them. Later figures carry the badge of the model they belong to.

- **Who is “the holder”?**

A firm that has built a frontier AI capability and is keeping it from general release. It might use it privately, share it with trusted partners, or sell products made with it. The paper’s real case is a lab that restricted its frontier cyber-capable model in 2026.

For policy: read every result below as a statement about one model, under stated conditions. None of them, alone or together, says “the best policy is...”.

PART 2

The case for speed needs four things at once

The paper first states the case for speed as strongly as it can. The statement is its own reconstruction. The parts have named supporters, but no single source it read makes the whole case or talks of sprinting through a window.⁹

Supporters exist. Bruce Schneier wrote that AI has “the potential to upend the longstanding advantage that attack has over defense on the Internet.”¹⁰ The lab that restricted its model in 2026 wrote: “In the long run, we expect that defense capabilities will dominate,” though “the transitional period will be fraught.”¹¹ The paper calls that an expectation, not evidence.

The case comes in three versions.

Speed. Planned computing power will push AI until logic, coding and math are practically solved. That helps attackers first, then defenders, through formal verification: machine proof that code does what it should. So the danger is a bounded window. Run through it; caps on computing only keep us exposed longer.

Secret racing. A firm races to the capability, holds it, and shares it with trusted defenders, who fix flaws before attackers catch up. The 2026 lab called its model “a general-purpose, unreleased frontier model” and described an “urgent attempt to

put these capabilities to work for defensive purposes.” It expected the window to be short: “Given the rate of AI progress, it will not be long before such capabilities proliferate.”¹²

Commerce. A holder can build products rivals can’t yet make, while its released models are copied and its sales lead wears away. That pressure pushes it to release. So two races run at once: a commercial race toward general ASI, and a security race for narrow skill in logic, code and math.

What the case needs, all at once (Figure 4):

- the near peak exists and has a far side;
- running shortens the time under it more than it raises the danger;
- a held capability lets defenders cross first;
- commercial pressure reinforces the security motive or leaves it alone.

The secret and commercial versions also need the narrow skill to be separable from general AI. The paper’s examples all use general models, so they don’t show this.¹³

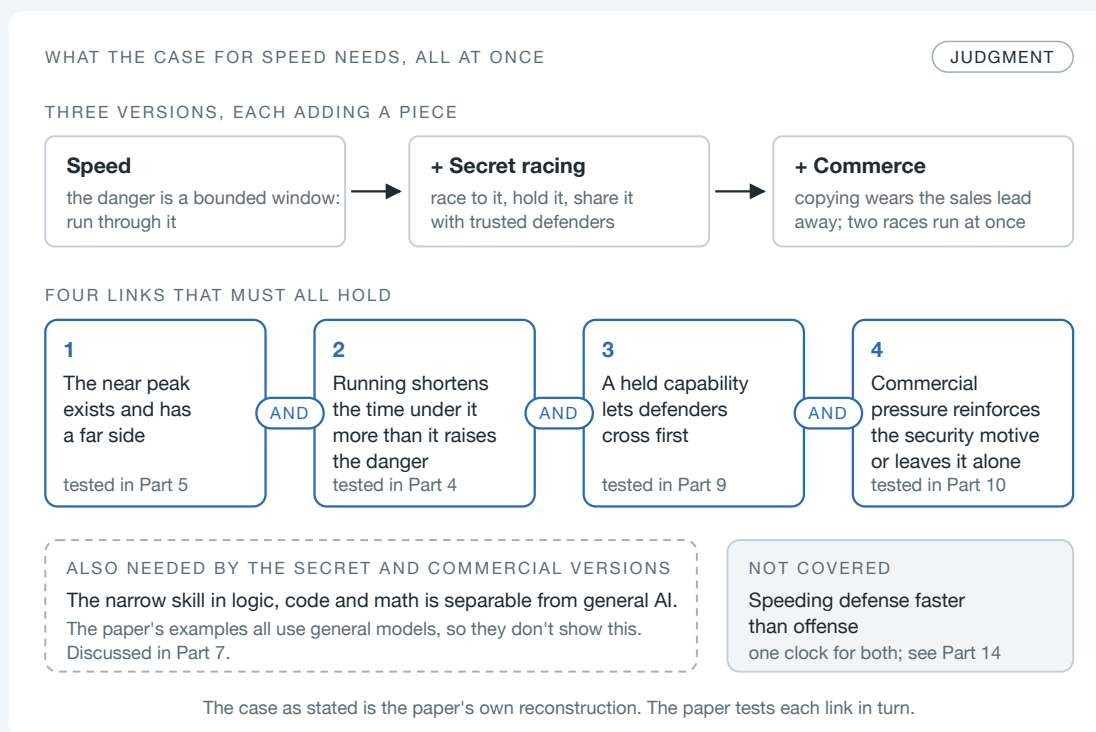


Figure 4. The case for speed at its strongest, as the paper states it: three versions, each adding a piece, and four links that must all hold at once. The secret and commercial versions also need the narrow skill to be

separable from general AI, which the paper's examples don't show; speeding defense faster than offense is not covered. The statement of the case is the paper's own reconstruction.

One version the paper does **not** cover: speeding up defense faster than offense, so defenders adopt AI tools before attackers do. Its models put both sides on one clock. It names what to watch (Part 14).¹⁴

– **What is formal verification, and why doesn't it settle this?**

It is a mathematical proof that a program meets its specification. A proof covers only the system it was done for, under its assumptions about hardware, configuration and unproved parts. Studies found bugs where verified and unverified parts meet.¹⁵ And a proof protects only the machines that install the proven software (Figure 19).

PART 3

What the paper adds is a second clock

The paper's first result restates, for AI, one by economists Philip Trammell and Leopold Aschenbrenner. If risk comes from development itself, whether speed helps depends on how sharply danger rises with speed. The paper claims nothing new there.¹⁶ What it adds is a **second clock**: the risky period ends only when operators have deployed fixes, on calendar time.

It draws three findings from security economics. Bad incentives cause failures as often as bad design. Patches can be turned into exploits automatically. Fixes reach machines slowly.¹⁷ From race models: more knowledge of rivals can make racing more dangerous, though the sign is not settled.¹⁸

What may be new is a combination of three things. Two copying clocks push release in opposite directions. A hold slows attackers' search for flaws nobody has found. And leaks depend on how a capability is used. In the paper's searches

across twenty bodies of work, each work found treats at most one of these. The paper calls that a statement about its searches, not the literature.¹⁹

PART 4

Does speed help once fixing is the bottleneck?

Side by side, the slower clock sets the length

How long does the dangerous period last? In the fully parallel version, it lasts as long as the slower clock.

- **The capability clock.** AI has some distance to travel. Go faster and you cover it sooner.
- **The adoption clock.** Operators need time to deploy fixes: patching, retiring old systems, rewriting code. It runs on the calendar.

In the simplest version the two run side by side. At low speeds AI's progress is the bottleneck; at high speeds, adoption is. The **speed threshold** is where they take equally long (Figure 5).²⁰

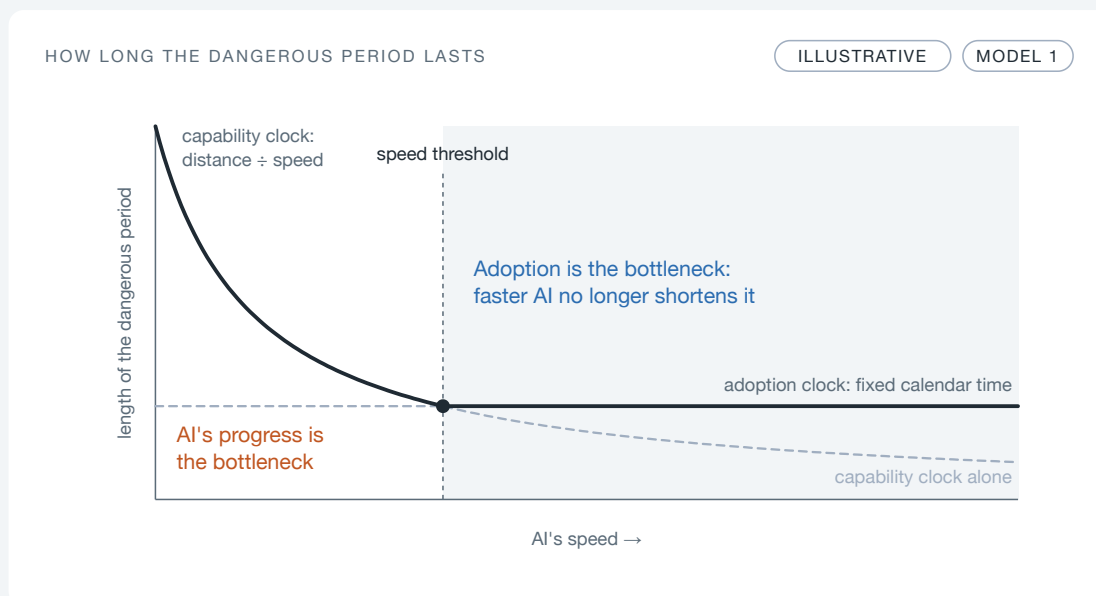


Figure 5. Model 1, with fully parallel timing. The dangerous period lasts as long as the slower clock: AI's progress at low speeds, operators'

adoption at high speeds. Past the speed threshold, faster AI no longer shortens it. Nobody knows where today's speed sits on this picture.

Illustrative values.

– What does “adoption binds” mean?

The adoption clock is the bottleneck. AI has done its part, and the transition still isn't over because operators are still deploying fixes. Past this point, making AI faster doesn't shorten the dangerous period at all.

In the model, risk is a tally. It is the danger rate during the transition, times how long it lasts, plus a one-time cost for each step up in capability. If harmful events arrive at random, the tally is their expected number.²¹ Height alone is not risk: a tall, narrow spike can hold less danger than a lower, wider one (Figure 6).

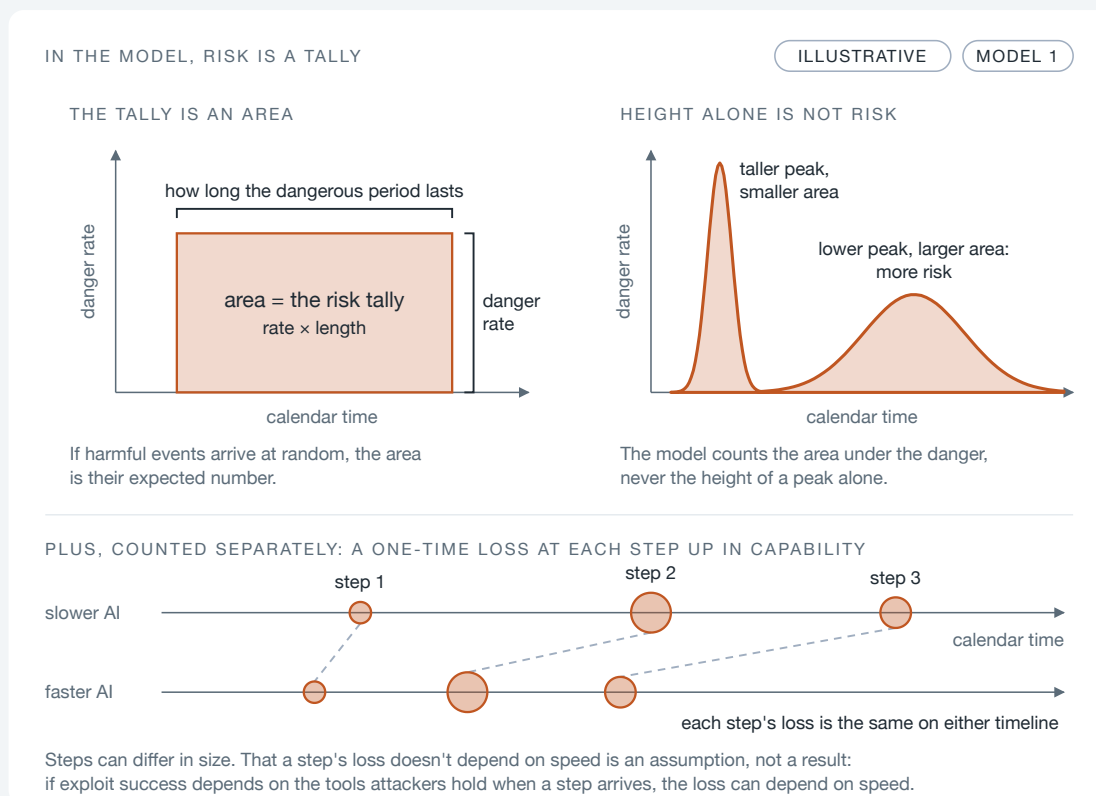


Figure 6. Model 1. In the model, risk during the transition is an area: the danger rate times how long the dangerous period lasts, so a tall, narrow spike can hold less risk than a lower, wider hump. Each step up in capability adds a separate one-time loss. Steps can differ in size, and that a step's loss is the same at any speed is an assumption, not a result.

Illustrative values.

One premise is built in: **the danger rate rises with AI's speed.**²² That is “running thickens the rain.” The evidence sections treat it as a hypothesis.

– **Why would faster AI make the danger heavier?**

In the paper it's an assumption. The reasoning: attackers' capability advances with AI's speed, while defense moves on calendar time, so faster AI widens the gap. The paper calls this a judgment. It also tests the alternative, where danger depends on how capable AI has become rather than how fast it got there.

Past a point, going faster only adds risk

Before the threshold, speed can help; past it, speed only hurts. This is the paper's base result, which this guide calls the fixed-curb rule:²³

- **Before the threshold** (AI's progress is the bottleneck), speeding up shortens the exposure but makes it more intense. It lowers risk only if danger rises *less than proportionally* with speed. Go faster by some share, and danger rises by a smaller share.
- **Past the threshold** (adoption is the bottleneck), speeding up only raises risk. The rain gets heavier and the curb gets no closer.

The past-the-threshold result holds under three conditions:

- adoption time stays fixed whatever AI's speed;
- adoption runs at least partly alongside AI's progress, not only after it;
- the step cost doesn't depend on speed.

Under those, it holds whether danger is set by speed or by how capable AI has become. In the second case it also needs danger at the end of the road to be higher than its average along the way.²⁴ So the best speed, in this model, is at or below the threshold (Figure 7).

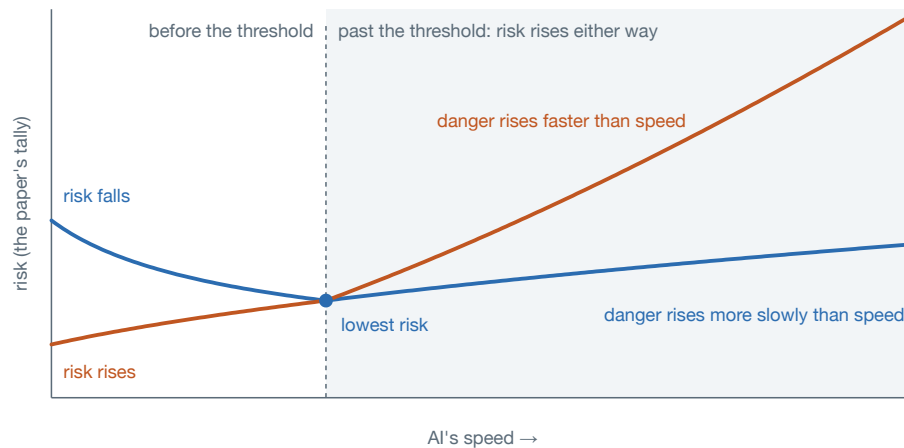


Figure 7. Model 1. Risk against speed for two ways danger could rise with speed. Before the threshold, speeding up lowers risk only when danger rises more slowly than speed; past it, risk rises either way. This holds with adoption time fixed, adoption running at least partly alongside AI, and a step cost that does not depend on speed. *Illustrative values.*

Other set-ups change the answer mostly before the threshold

The paper doesn't know which set-up is real, so it checks several. Fixes might run alongside AI's progress, wait for the capability, or a mix. Danger might be set by AI's speed or by the level reached. Figure 8 lays out the cases; in words:²⁵

- **Alongside or mixed, past the threshold:** speed raises risk under speed-set danger, and under level-set danger when end-of-road danger exceeds the average.
- **Alongside, before it:** under speed-set danger, speed helps only if danger rises less than proportionally. Under level-set danger, speed always helps.
- **Mixed, before it:** under speed-set danger, speed helps only if danger rises with speed by less than the share of the transition AI's progress sets. Under level-set danger, it always helps.
- **Fixes wait for the capability:** there is no threshold. Level-set danger: speed always helps. Speed-set danger: it helps only if danger rises slowly enough.

	DANGER SET BY AI'S SPEED		DANGER SET BY THE LEVEL REACHED	
	before threshold	past threshold	before threshold	past threshold
Fixes run alongside AI	depends only if danger rises less than in proportion to speed	↑ raises risk	↓ lowers risk always	↑ raises risk when end-of-road danger is above its average
A mix of both	depends only if danger rises by less than the share AI's progress sets	↑ raises risk	↓ lowers risk always	↑ raises risk when end-of-road danger is above its average
Fixes wait for the capability	depends no threshold: lowers risk only if danger rises slowly enough		↓ lowers risk no threshold: always	

These comparisons hold adoption time, the timing mix and the step cost fixed.
Which kind of danger applies is open. The paper's judgment favours danger set by speed.

Figure 8. Model 1. The set-ups the paper checks. Past the threshold, speeding up raises risk when danger is set by speed, and also when it is set by the level reached and end-of-road danger is above its average. Before the threshold, speed helps under level-set danger; under speed-set danger, the answer depends on how danger rises relative to speed. These comparisons hold adoption time, the timing mix and the step cost fixed. Which set-up is real is open.

Which kind of danger applies is open. The paper's judgment favors speed-set danger. It stands in for the gap between attackers, who advance with speed, and defenders on calendar time.

— Where is the speed threshold today?

Nobody knows. Locating it needs the capability distance, how far AI must travel to end the transition, and that is unmeasured. The paper's judgment that adoption binds for some populations (Part 8) is provisional. It does not say where today's AI speed sits.

Each step up exposes hidden flaws once, by assumption at any speed

Each step up in capability exposes a stock of hidden flaws once. That this cost doesn't depend on speed is an assumption, not a result.²⁶

Three things set the cost: flaws exposed per step, the chance each is exploited, and damage per exploit. Cutting any one lowers the cost, as long as the other two are above zero. Hardware-bound login could lower the chance; network segmentation or rate limits, the damage. These are proposed channels that need evidence. If exploit success depends on the tools attackers hold when a step arrives, the step cost *can* depend on speed.

Speed helps past the threshold only if it speeds fixes more than danger

A defender of speed has one way out. Perhaps faster AI also speeds adoption, through automated rollout, verified translation of old code, or automated rewrites. The paper gives exact terms:²⁷

Where adoption is the bottleneck, speeding up lowers risk only if adoption time falls with speed faster, in proportion, than danger rises.

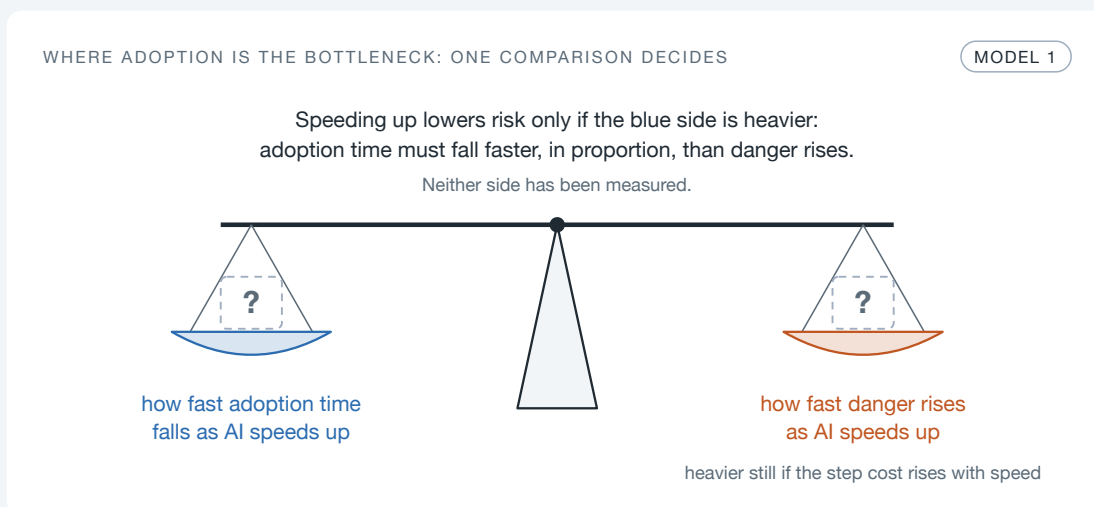


Figure 9. Model 1. The accelerationist's escape as a balance. Where adoption is the bottleneck, speeding up lowers risk only if adoption time falls with speed faster, in proportion, than danger rises. Neither side has been measured.

If the step cost also rises with speed, the escape needs more. And the escape is *not* the threshold moving outward. If adoption binds at today's speeds, the threshold can move inward while the escape still works.

So where adoption binds, the speed question comes down to one comparison: how fast adoption responds to AI's speed, against how fast danger does. It is measurable in principle. The paper found no measurement.

– What would change the speed answer?

Evidence that AI shortens adoption itself, faster than it raises danger. Signs to watch: verified translation and rewriting reaching production code, and AI patching that narrows the gap between flaws found and fixed. Another sign: retirement and migration speeding up across sectors, not only in the fastest channels.

Each group of operators has its own threshold

Big firms, small operators, open-source maintainers and industrial-control operators each have their own adoption time. So the threshold is a spread, not a point. With several populations, speed lowers overall risk only if the gain where AI's progress is the bottleneck outweighs the loss where adoption is (Figure 10).²⁸

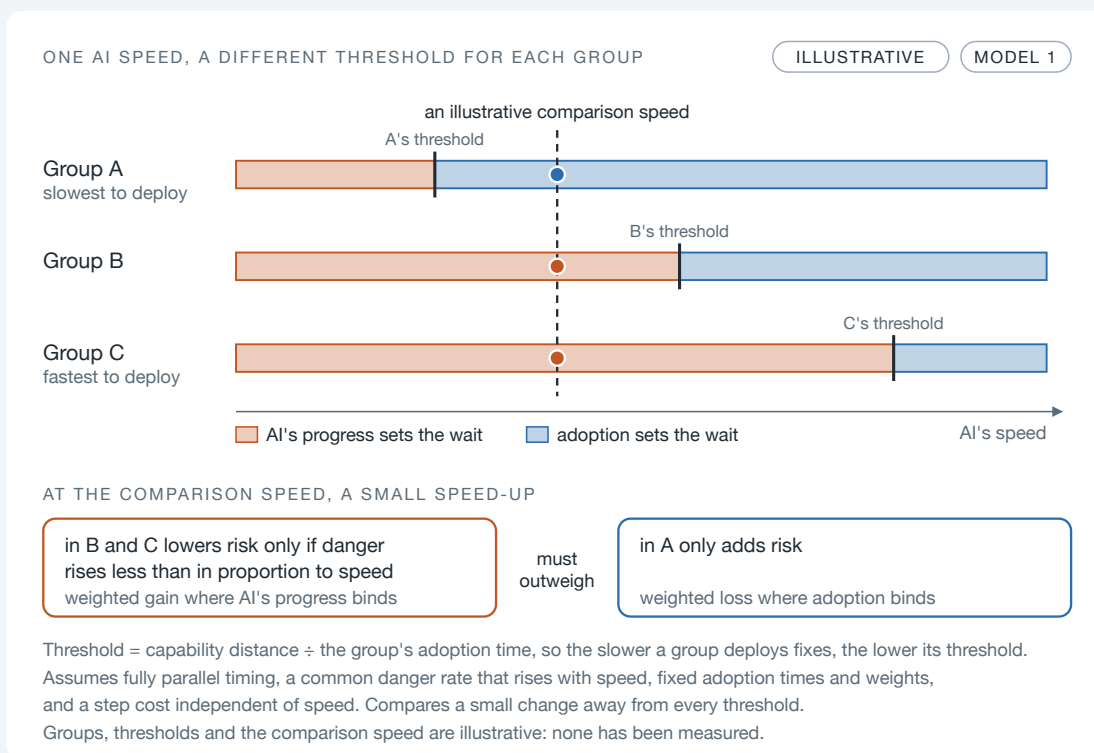


Figure 10. Model 1. Each group of operators has its own speed threshold, so at one AI speed adoption can set the wait for some groups while AI's progress still sets it for others. A small speed-up then lowers overall risk only if the weighted gain where AI's progress binds outweighs the weighted loss where adoption binds. Assumes fully parallel timing, a common danger rate that rises with speed, fixed adoption times and weights, and a step cost independent of speed.

Compares a small change away from every threshold. The groups, thresholds and comparison speed are illustrative, not measured.

For policy: the case for speed depends on a threshold nobody has located, and on how adoption responds to AI, which nobody has measured. Past the threshold, under the stated conditions, speed alone does not help.

PART 5

A peak can exist, but the peak model has no AI in it

Danger peaks only if flaws pour in faster than they drain, then slower

Imagine a bathtub. New code pours flaws in through the tap; patching and retirement drain them. The water level is the stock of exploitable flaws, and danger rises and falls with it. (An illustration of the paper's flaw-stock model.)

The paper's peak result, here called the flaw tide:²⁹

- **Danger rises and later falls only if** at some point flaws pour in faster than they drain, and later more slowly.
- **Protection of new code matters.** Suppose protection, such as memory-safe languages, spreads through new code, cutting the inflow toward a residue it can't reach. Take the example where the tub starts with no exploitable flaws and removal is constant. Then a peak is guaranteed when protection spreads more slowly than removal works. When protection spreads faster, a peak still comes unless enough new code stays unprotected. The peak is pre-ASI only if it arrives before ASI.
- **A peak can be followed by a plateau.** They are not alternatives. The plateau's height depends only on the unprotected residue and the drain rate, not on how fast protection spread.

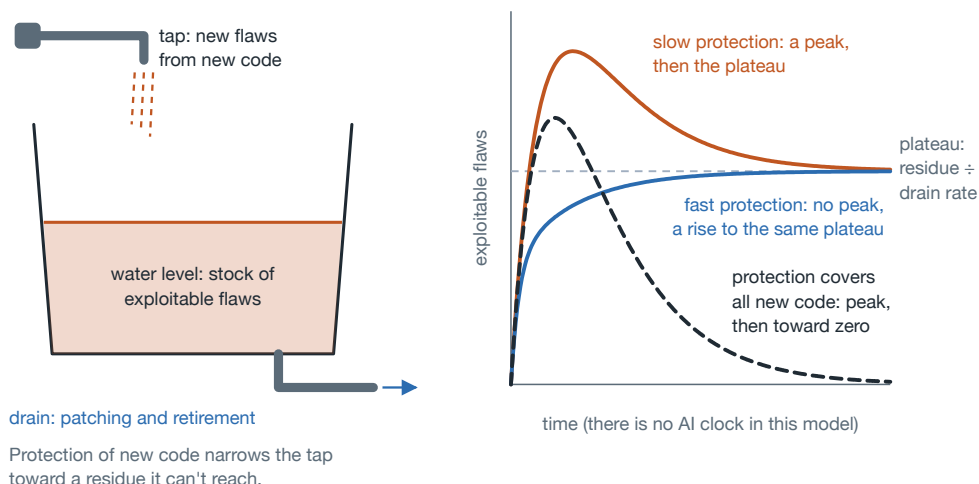


Figure 11. Model 1. Left: the bathtub picture of the flaw stock. Right: the model's path for three cases, each starting from an empty tub. Slow protection gives a peak and then a plateau; fast protection gives a steady rise to the same plateau; protection that covers all new code gives a peak that falls toward zero. There is no AI speed or capability in this model. *Illustrative values.*

Two policy readings follow, in the same example of an empty start and constant removal. If protection eventually covers all new code, there is always a peak. It is pre-ASI only if it arrives before ASI. If a residue remains, faster protection can remove the peak, leaving a steady rise to the same plateau.

So long-run danger depends on **what protection can't reach** and **how fast flaws are removed**. And if today's stock is already high enough that removal outpaces new arrivals, no peak lies ahead in this model.

Two limits. Exploitation per flaw is held constant. If attackers' skill per flaw rises faster than the stock falls, danger can climb while the tub drains. Most important: **the peak model contains no AI speed or capability**. A peak can exist in the flaw-stock dynamics. Whether AI creates or moves it is not modelled.³⁰

— So does AI create a peak or not?

The paper doesn't say, because its peak model doesn't include AI. It shows when a stock of flaws would rise then fall, and that a peak and a plateau can both

happen. Tying that path to AI's speed is listed as future work.

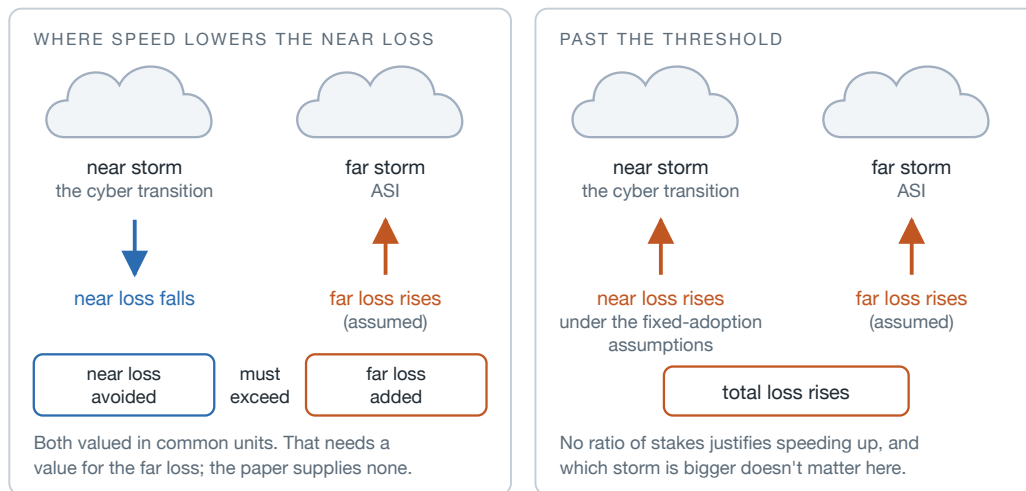
Trading a near storm for a far one needs a price for the far one

Suppose a near cyber peak and a far one at ASI. Speed might lower the first and raise the second. The paper's two-storms test:³¹

- **Speed lowers total loss only** if it lowers the near loss *and* the near stakes, compared with the far ones, are big enough to outweigh the far loss's rise. Past the threshold, the near loss rises too, so no ratio of stakes justifies speeding up. That holds for this two-loss test; Part 6 adds a benefit.
- **Valuing the future less makes speed less attractive:** the bar for speed to help gets stricter.
- **If the far loss comes sooner when AI is faster,** speed brings it nearer.
- **If both are catastrophes in sequence,** speed lowers the total chance only if the drop in near risk, weighted by surviving the far one, beats the rise in far risk, weighted by surviving the near one. Here the far catastrophe's chance means its chance once the near one has been survived.

The test judges small changes. It also counts nothing speed buys beyond the two losses; Part 6 adds one benefit, health gains from the same AI. A large jump can cut total loss where the small-change test fails, so large moves must be judged in levels.

It also **assumes the far risk rises with speed**. In this two-loss test, where adoption binds and responds to speed more slowly than danger does, both peaks point the same way and their ranking doesn't matter. Where speed lowers the near peak, the comparison needs a value for the far loss. The paper doesn't supply one (Figure 12). With a health benefit counted too, the far loss can matter even where near harm rises (Part 6).



The test judges small changes; large moves must be judged in levels. It counts only these two losses (Part 6 adds a health benefit) and assumes the far risk rises with speed. Storm sizes show no ranking.

Figure 12. Model 1, the two-storms test for a small speed-up. Where speed lowers the near loss, it helps only if the near loss avoided outweighs the far loss added, in common units, and the paper supplies no value for the far loss. Past the threshold, under the fixed-adoption assumptions, both losses rise, so no ratio of stakes justifies speeding up. The test counts only these two losses and assumes the far risk rises with speed; the storms are drawn the same size, so no ranking is implied.

For policy: long-run danger rests on unreachable code and removal speed. In the two-loss test, whether superintelligence is the bigger peak matters only where speed lowers the near one.

PART 6

Life-saving benefits give speed a second reason, with conditions

So far, speed buys one thing in Model 1: a shorter stay in the rain. But the AI that arms attackers may also do good that has nothing to do with cyber defense. Does a benefit that saves lives change the speed answer?

It can. In the model, a health benefit gives a second reason to run, even past Part 4's speed threshold, which this part calls the cyber threshold. This holds only while the benefit still waits on AI's progress, not on its own trials and approval. And it holds only if the benefit brought forward outweighs the added cyber harm.³²

Running to a pharmacy helps only if the counter serves you

Here is an analogy, not a claim from the paper. You cross the street in the rain to a pharmacy. If the counter serves you on arrival, running gets you the medicine sooner. That can be worth some heavier rain. If your order is still being filled, arriving sooner gets you nothing sooner. Then only the rain question is left (Figure 13).

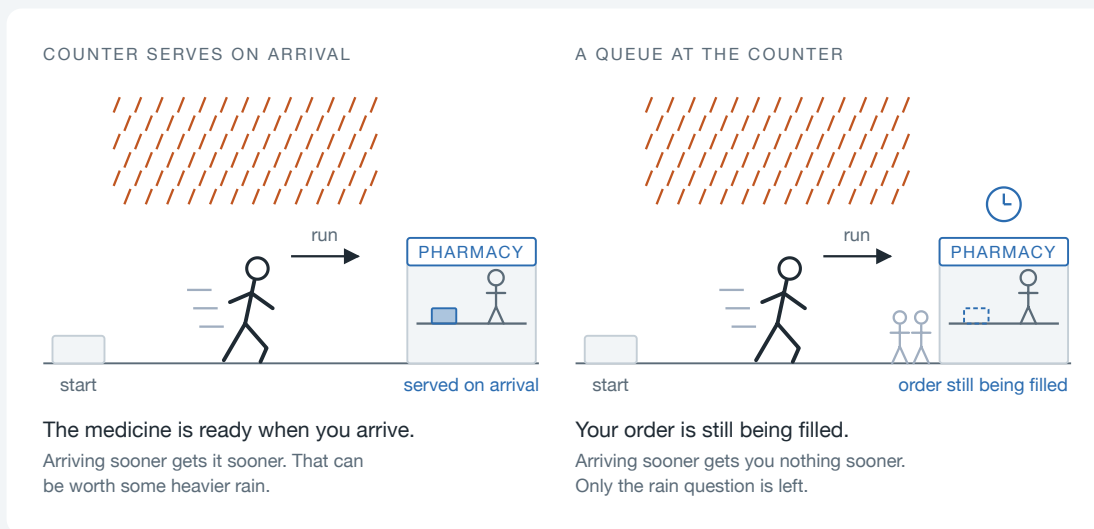


Figure 13. An analogy for the benefits extension. Left: the counter serves you on arrival, so running brings the medicine sooner. Right: the order is still being filled, so arriving sooner brings nothing sooner, and only the rain counts. A drawing of the analogy, not of data.

The paper's motivating example is protein structure. AlphaFold was, in its developers' words, "the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known." Half of the 2024 Nobel Prize in Chemistry went to Demis Hassabis and John Jumper "for protein structure prediction." Neither source says AlphaFold shortens drug development. That such capabilities bring health benefits sooner is this part's premise, not a finding.³³

One yardstick for this extension: healthy years of life

The paper adds a benefit to Model 1 and keeps everything else. One AI speed moves both roads: the road to the end of the cyber transition, and the road to the capability the benefit needs. The two distances may differ.

The model holds both adoption times fixed as AI speed changes. It also holds fixed the share of each adoption that waits for the capability. The cyber and benefit timings need not match. Benefits count only up to one fixed calendar date; this guide calls that span the period counted. The date is late enough to include the whole cyber transition and its step losses.

To weigh lives against cyber harm, the extension needs one unit. It uses DALYs, disability-adjusted life years. One DALY “can be thought of as one lost year of ‘healthy’ life.” Counting DALYs averted puts deaths and illness in one unit. Cyber harm enters in the same unit, at a fixed number of DALYs per harmful event.³⁴

The result is a balance: health benefit minus cyber loss. It is a conditional balance, not total welfare. It assumes a steady benefit that lasts once it arrives, and it counts each cyber harm once. The yardstick holds for this extension only. The other three models keep theirs, and nothing is added across them.

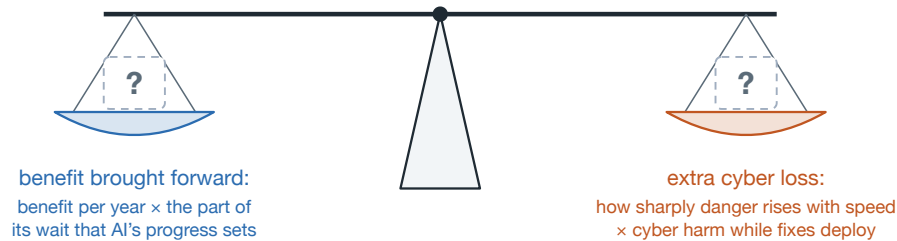
Past the threshold, the benefit must outweigh the extra cyber loss

Past the cyber threshold, speeding up can still raise the balance. It does so only if the benefit brought forward outweighs the extra cyber loss (Figure 14).³⁵

- **The benefit brought forward** is the benefit per year times the part of its wait that AI’s progress sets. Speed up by one percent, and that part shrinks by one percent.
- **The extra cyber loss** is how sharply danger rises with speed, times the cyber harm while fixes deploy.

This exact test has a narrow scope. Danger is set by AI’s speed. Speed is past the cyber threshold. The benefit still waits on AI and arrives within the period counted. And the future is not discounted. Other cases have their own tests in the paper. Nobody has measured either side.

Speeding up raises the balance only if the blue side is heavier:
the benefit brought forward must outweigh the extra cyber loss.
Both sides in DALYs: healthy years of life. Neither side has been measured.



THIS TEST HOLDS ONLY WHEN

danger is set by speed · speed is past the cyber threshold · there is no discounting
the benefit still waits on AI, not on its own trials and approval, and arrives within the period counted

Figure 14. Model 1 with benefits, a companion to Figure 9. Past the cyber threshold, speeding up raises the balance only if the benefit brought forward outweighs the extra cyber loss, both counted in DALYs. The test holds for danger set by speed, a benefit still waiting on AI and arriving within the period counted, and no discounting. Neither side has been measured.

Once the benefit waits on its own clock, the old test decides

The second reason to run can vanish. It counts only while AI's progress sets when the benefit arrives. Once the benefit waits only on trials, approval and uptake, faster AI brings it no sooner. That happens past the benefit's own threshold; under sequential timing there is none. Running then gets no one to the far curb sooner. Whether it keeps you drier is the fixed-curb rule's question again (Figure 15).³⁶

What that means depends on the cyber clock:

- **Below the cyber threshold**, the fixed-curb rule decides, as in Figure 8. Under speed-set danger, speed helps only if danger rises slowly enough. Under level-set danger, it helps if average danger along the way is above zero; otherwise it makes no difference.
- **Past the cyber threshold**, speed only adds cyber harm when danger is set by speed. When danger is set by the level reached, it adds harm if end-of-road danger is above its average.

While the benefit still waits on AI, it also widens the range where speed helps below the cyber threshold. Speed can then help even when danger rises faster than speed. In one illustrative case, danger rises twice as fast as speed, in proportion, and the cyber loss rises with speed. Yet the benefit grows 16 times as fast as the cyber loss, so the balance still improves.

DOES SPEEDING UP RAISE THE BALANCE?		
	BENEFIT CLOCK SET BY AI the benefit still waits on AI	MODEL 1 + BENEFITS SET BY ITS OWN ADOPTION trials, approval, manufacture, uptake
Cyber clock set by AI before the cyber threshold	depends helps if the benefit brought forward and the shorter stay outweigh the heavier rain Level-set danger: always helps	the old cyber test decides the benefit no longer moves; helps only if danger rises slowly enough (Figure 8) Level-set danger: helps if average danger is positive; otherwise neutral
Cyber clock set by fixing past the cyber threshold	depends helps only if the benefit brought forward outweighs the extra cyber loss (Figure 14) Level-set danger: the same weighing, against end-of-road danger above its average	only adds cyber loss faster AI buys no earlier benefit, only heavier rain Level-set danger: adds loss when end-of-road danger is above its average

Rules for a benefit arriving within the period counted, with no discounting.
Grey lines: danger set by the level AI reaches rather than by its speed.

Figure 15. Model 1 with benefits. Which way speed points depends on which clock sets each wait. Only while the benefit still waits on AI can it pay for faster AI; once its own trials, approval and uptake set its pace, the cyber test of Figure 8 decides alone. The rules are for a benefit arriving within the period counted, with no discounting.

The benefit has its own clock: trials, approval, manufacture, uptake

A benefit like a new drug reaches patients only after trials, approval, manufacture and uptake. These run on their own clock. A report for the U.S. Department of Health and Human Services states that “the clinical phase lasts an average of around 95 months compared to 31 months for the non-clinical phase.” It attributes that figure to an earlier study.

The report’s non-clinical stage includes early research such as target discovery. Across therapeutic areas, Phase 3 is the longest stage, at 38.0 months. FDA review averages 16.2 months, a figure that includes the sponsor’s own time. The clinical

phases “may overlap.”³⁷ That overlap does not show how a benefit’s adoption overlaps AI’s progress.

One AI-discovered drug programme reports about 18 months from target discovery to a preclinical candidate. It reports under 30 months from the start of target discovery to the end of phase 0/1 testing. That is one programme, reported by its developers, and its phase 2a trial then followed on the clinical clock. The paper found no peer-reviewed study of AI’s effect on total development time across drugs. That is a statement about its searches.³⁸

The paper’s reading, not the report’s claim, is that much of the wait for a new drug lies in stages AI’s progress does not set.

The model gives the benefit the same three timings as the cyber side (Figure 16):

- **Parallel:** the benefit’s adoption runs alongside AI’s progress, and the slower one sets arrival.
- **Mixed:** a share of the adoption, which the paper calls m_B , waits for the capability; the rest runs alongside.
- **Sequential:** adoption starts only once the capability arrives.

Faster AI shortens only AI’s part of the wait. And the share of adoption that waits is not the share of the wait that AI sets. In one illustrative check, with m_B at 0.6, AI’s progress sets a share of 0.81 of the whole wait.

The cyber threshold and the benefit’s own threshold may come in either order, or coincide. A clock with sequential timing has no threshold. If the benefit’s own adoption takes as long as the period counted, or longer, no benefit counts at any speed. Otherwise the benefit starts to count only once speed brings its arrival inside that period. Where speed crosses one of these points, compare the change on each side.

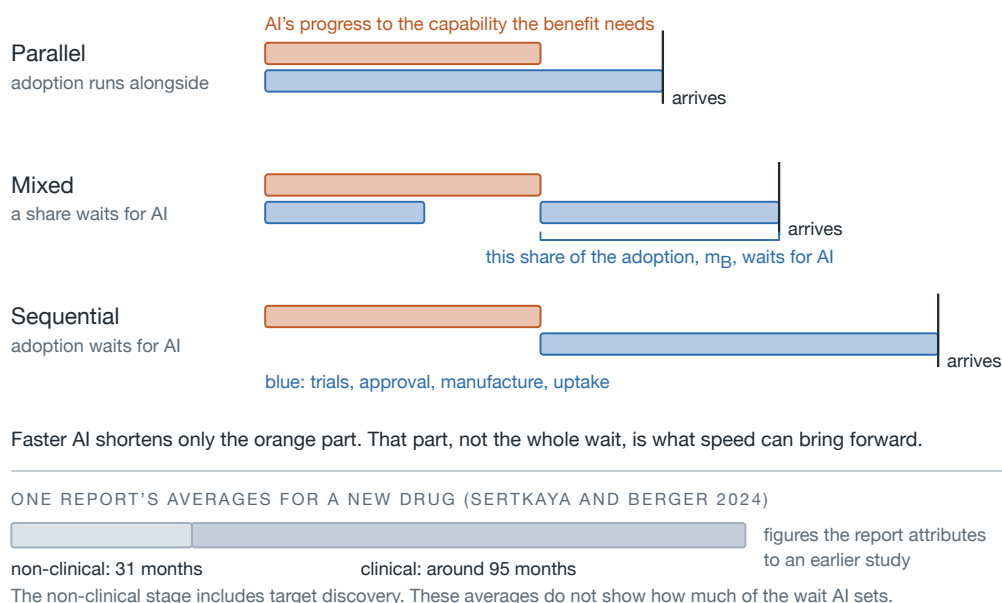


Figure 16. Model 1 with benefits. Three ways a benefit's own adoption can line up with AI's progress; the bar lengths are illustrative. Faster AI shortens only the orange part. The grey bar shows one report's averages for a new drug (from a 2024 report for the US Department of Health and Human Services, which attributes them to an earlier study); they do not show how much of the wait AI's progress sets.

Does each step of speed gain less benefit as speed rises?

Each step of speed brings the benefit forward by a share of the wait that AI still sets. When that wait is long, a step gains a lot. When the benefit is close, a step gains little. So, without discounting, each equal percentage rise in speed brings the benefit forward less as speed rises. This holds while AI sets the benefit's arrival and the benefit arrives within the period counted.³⁹

A best speed past the threshold is possible, not guaranteed

The paper works one example with illustrative values, not estimates (Figure 17). The cyber threshold sits at speed 1 and the benefit's own threshold at speed 4. The balance peaks at 6, at speed 2, twice the cyber threshold. It is 1, 5, 5 and 3 at speeds 0.5, 1, 4 and 6. At the best speed the cyber loss is rising; the benefit pays for it.⁴⁰

The same example shows that benefits do not always justify speed. At speed 5 the benefit is large, but it no longer moves with speed. Going from 5 to 6 lowers the balance from 4 to 3. A large benefit is not, by itself, a reason to go faster.

In general, a best speed exists when the inputs have no jumps and speed is limited to a fixed range above zero. It can sit at either end of that range, at a threshold, or at the speed where the benefit starts to count. It can also sit where the balance is flat, perhaps along a level stretch. A flat point alone does not prove the best speed; compare every candidate. Without both limits on the range, a best speed need not exist.

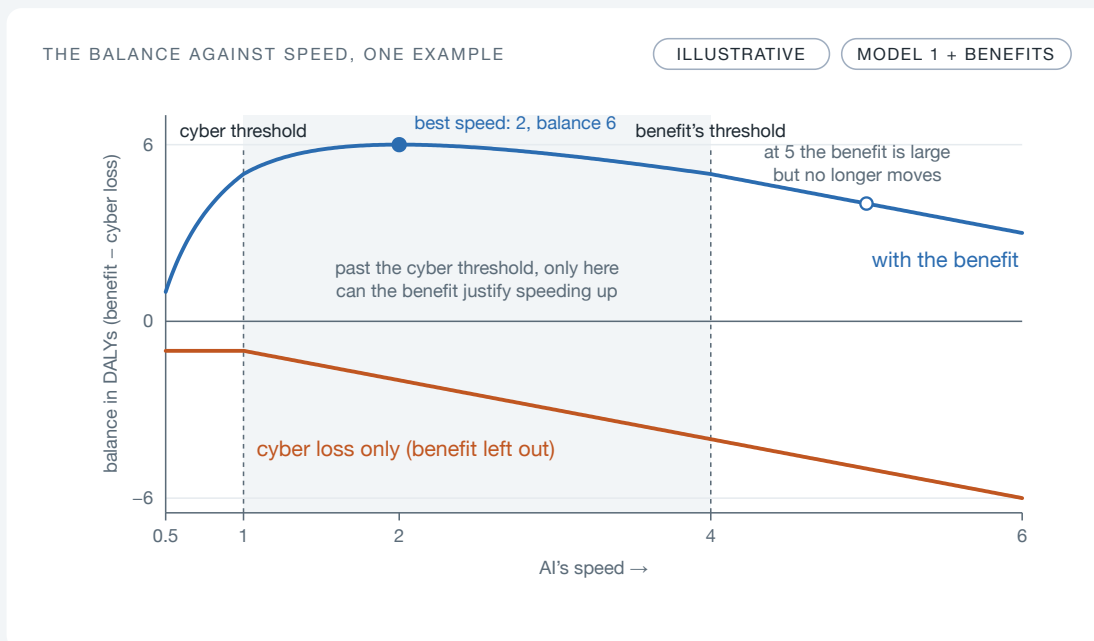


Figure 17. Model 1 with benefits. The balance against speed in the paper's example, with the benefit (blue) and with the cyber loss alone (orange). With the benefit, the best speed is 2, past the cyber threshold; past the benefit's own threshold the balance falls. Without it, in this example, speed makes no difference below the cyber threshold and only adds loss above it. An optimum past the threshold is possible, not guaranteed. *Illustrative values.*

Discounting cuts both sides, with no general direction

Suppose future health and future cyber losses are both discounted from the same date. A late benefit then counts for less. But the cyber loss changes too, so the effect on the overall sign has no general direction. WHO's current burden

estimates use “No discounting for time,” so any discount rate here is the paper’s modelling choice.⁴¹

Discounting also changes the step costs. A step loss of fixed size counts more when faster AI brings it earlier. That weakens the case for speed. Level-set danger needs its own discounted calculation, and the undiscounted answers can reverse. In one illustrative check, a gain from speed became a small loss.

The model assumes one AI speed moves both; narrow tools may not

The extension assumes full coupling: the AI that arms attackers is the AI that brings the benefit. That is an assumption, not a finding. Narrow tools, such as specialized structure-prediction models, could advance a benefit on a clock of their own. Then speeding general AI would not bring that benefit forward. The paper names this limit and does not model it; no sign is known (Figure 18).⁴²

Biological misuse and hospital attacks are named, not modelled

The capability that speeds drug discovery may also aid biological attack. That risk lies outside this balance, and the paper does not model it.

Cyber harm also reaches health directly. One study found that ransomware attacks on hospitals “decrease hospital volume by 17–24 percent during the initial attack week.” It also found: “Among patients already admitted to the hospital when a ransomware attack begins, in-hospital mortality increases by 34–38 percent.” Another found that targeted hospital attacks “may be associated with disruptions of health care delivery at nontargeted hospitals within a community.” None of these studies estimates DALYs per attack, so no source gives the health loss per harmful event.⁴³

These findings point to a link the model leaves out. Attacks on health systems may cut the benefit itself, which the extension holds constant.

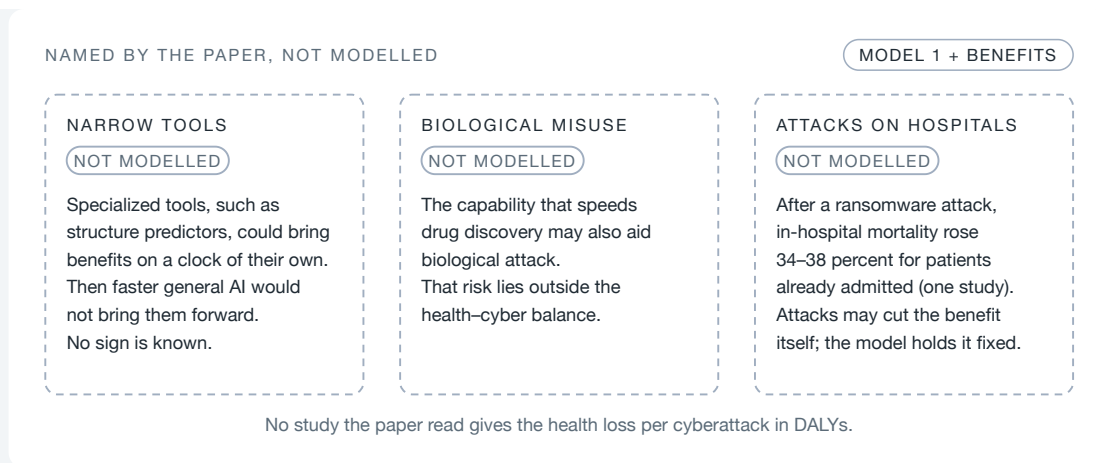


Figure 18. Three limits the paper names but does not model: narrow tools that could bring benefits without the general capability, biological misuse, and attacks on hospitals that may cut the benefit itself. The mortality figure is one study’s estimate for patients already admitted when an attack began.

With a benefit, the two-storms test gains a term

Part 5’s two-storms test found that, past the threshold, no ratio of stakes justifies speeding up. That holds for its two losses. With a benefit in the same units, speed can improve the balance even while near harm rises. Then the far storm can matter there too. If far risk rises with speed, speed helps only if the benefit gained, less the near loss’s rise, beats the far loss’s rise. The paper adds no new result about the far loss.⁴⁴

For policy: faster AI and shorter trials are different levers

If the goal includes health benefits from the same capability, two levers differ.⁴⁵

- **Faster AI** brings the benefit forward only while AI’s progress sets its arrival. With speed-set danger, it also raises cyber danger.
- **A shorter benefit clock**, through trial and approval capacity, manufacturing and uptake, can bring the benefit forward without raising cyber danger. That such capacity shortens the benefit’s clock without changing cyber danger is an assumption, not a result.

The two levers work differently. A shorter benefit clock raises the benefit’s total only weakly: strictly only when it brings arrival forward within the period

counted. Under parallel timing, while AI sets arrival, it changes nothing. While AI sets arrival and that stays so, a shorter benefit clock leaves the undiscounted return to extra speed unchanged. With discounting, it raises that return when some adoption waits for the capability. If a shorter benefit clock changes which clock sets arrival, that change needs its own check.

One study found that, after data breaches, deaths within 30 days of a heart attack rose by as much as 0.36 percentage points within three years. Its authors associate that rise with remediation, which they did not directly observe.⁴⁶ Which lever should come first needs numbers the paper does not have.

What to measure: three parts of a benefit's wait, separately

For AI-assisted drug programmes, three quantities should be measured separately. They are the wait for the capability, the benefit's own adoption time, and the share of that adoption that waits for the capability. No source the paper read gives any of the three, or the DALYs lost per harmful cyber event (Part 14).⁴⁷

For policy: in the model, a life-saving benefit is a real second reason to run. It counts only while the benefit still waits on AI, and only if it outweighs the added cyber harm. Once its own trials and approval set its pace, faster AI buys no earlier benefit.

PART 7

AI is advancing fast on tasks that are cheap to check

The paper's capability snapshot, from late 2026, shows fast progress on tasks a computer can check automatically:⁴⁸

- **FrontierMath:** Epoch's own runs put the best scores at 93.7% on Tiers 1–3 and 100% on Tier 4, whose private set has 41 problems. OpenAI models hold both; FrontierMath was developed with OpenAI funding, and OpenAI has exclusive access to a subset.

- ARC-AGI-2's semi-private set: best score 95%.
- A perfect score at the 2026 International Mathematical Olympiad, on a lab's own account of official marking.
- Five PutnamBench entries report all 672 problems solved; the saved records don't document public verification.

Coding is not saturated: FrontierSWE and FrontierCode best scores were 65.5% and 54.6%. METR's agent-task measure warns, "Measurements above 16 hrs are unreliable with our current task suite," and hasn't been updated since 8 September 2026. The paper gives no doubling estimate.

– **Isn't 100% on a math benchmark proof that math is solved?**

No. The paper says the record shows rapid progress on selected tasks that are cheap to check, not that logic, coding and math are practically solved. In its untested judgment, for research math and real software, deciding what to check may be the bottleneck.

Three cautions apply:

- Scaling laws are fitted to a training measure called loss, not to skills. Across a tenfold rise in training compute, average forecasting errors for skills were 6 percentage points for a task group and 18 for single tasks.⁴⁹
- Rice's theorem rules out one method that decides every interesting property of every program. It leaves room for special methods; the paper calls it a mathematical limit, not a ceiling on useful verification.⁵⁰
- Real proofs have edges. seL4's proofs assume correct hardware and boot code. Testing found six wrong-code bugs in an early CompCert compiler's unverified front end. Three verified distributed systems had sixteen bugs, mostly where verified and unverified parts meet.⁵¹

And the security examples all use **general** models. There is no evidence yet that a narrow code-only system could supply the capability alone.



Separate studies of separate systems, not layers of one system.

A SECOND EDGE: WHICH MACHINES RUN THE PROVEN SOFTWARE



A guarantee reaches only the machines where the proof and its assumptions apply.
The machines are an illustration, not a count.

These examples show the edges of the cited proofs. They do not show that verification cannot cover specifications, interfaces, hardware or configuration in other systems.

Figure 19. Where a proof stops, in three examples from separate studies of separate systems. seL4's proofs assume correct hardware and boot code; CompCert 1.6's wrong-code bugs were in its unverified front end; three verified distributed systems had bugs mostly where verified and unverified parts meet. A guarantee also reaches only the machines that run the proven software under its assumptions. These show the edges of the cited proofs, not a limit on what verification can cover elsewhere.

PART 8

AI can find flaws fast; fixes still run on the calendar

Attackers may gain first, but that is a hypothesis

The paper treats “AI favors attackers first” as a hypothesis. These are reported lab and competition results, not a matched comparison with human experts. The lab figures are the lab's own, without independent replication:⁵²

- **DARPA's AI Cyber Challenge (August 2025 final):** automated systems found 54 of 63 synthetic vulnerabilities and patched 43. They found 18 real ones and supplied 11 patches.
- **One lab, February 2026:** more than 500 high-severity open-source vulnerabilities found and validated, by its count.
- **That lab, April 2026:** its restricted model built working browser-engine exploits 181 times, against two successes in several hundred attempts for its earlier model. The test set-up left out the browser's sandbox. Since 181 exceeds the 112 unique reports from that work, the count includes repeats.
- **That lab, June 2026:** from 18 Firefox security patches public for at least 90 days, it built eight exploits. The first took under an hour of test time, the eighth roughly twelve hours. The plots use the fastest of three trials per stage, in a standalone JavaScript-shell test rather than the full browser. Test runtimes, not field times.
- **The UK AI Security Institute:** it tracks the length of cyber task frontier models could finish with 80 percent reliability, in its narrow suite. By its February 2026 estimate, that length had doubled every 4.7 months since late 2024. The fit “is not a future prediction, nor a fixed law.”

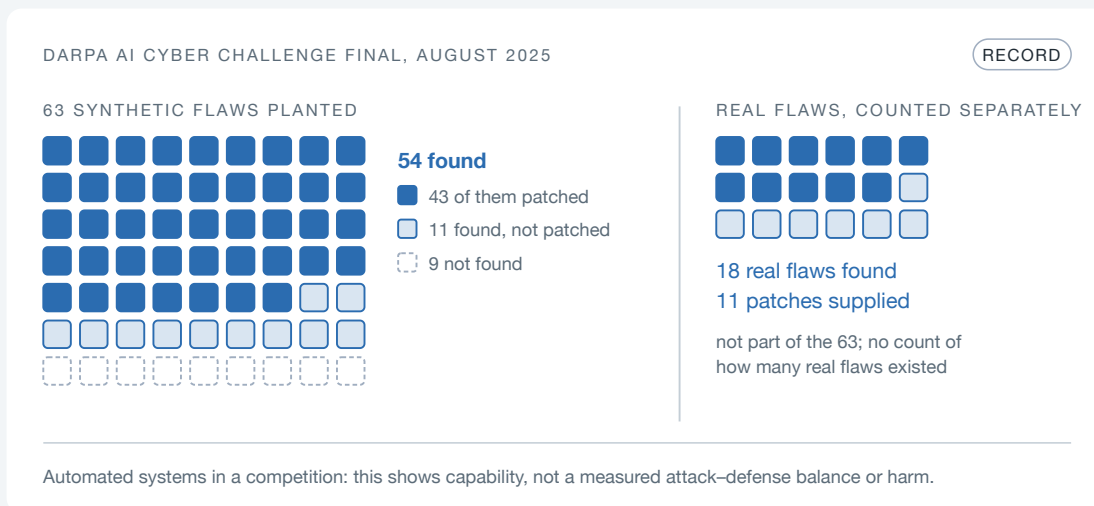


Figure 20. DARPA's AI Cyber Challenge final (August 2025). Left: of 63 synthetic flaws, automated systems found 54 and patched 43. Right: they also found 18 real flaws and supplied 11 patches, counted separately. A competition result: it shows capability, not a measured attack-defense balance or harm in the field.

In the field, Mandiant studied 138 vulnerabilities disclosed in 2023 and exploited. It reported an average time-to-exploit of five days excluding 15 outliers, 47 days including them.⁵³

- **What is a zero-day?**

A flaw attackers exploit before the maker has a fix, so defenders have had “zero days.” An n-day is already public, often with a patch available, but still unpatched on many machines.

Capability is not harm. These results show capability, not a measured attack-defense balance, and not harm.

Google’s threat-intelligence group tracked 90 zero-days exploited in 2025, against 78 in 2024, as its 2025 review counts them. Counting n-days, it tracked 127 exploited vulnerabilities in 2025 and 141 from January to August 2026. CISA added 186 entries to its known-exploited catalogue in 2024 and 245 in 2025. None isolates AI. The Google group calls exploitation of AI-discovered flaws “an early indicator rather than an established trend.”⁵⁴ So the paper holds, with high confidence, that the capability exists, and with low confidence what it does to net harm.

- **What is the KEV list?**

CISA’s public catalogue of Known Exploited Vulnerabilities: flaws the U.S. cyber-defense agency knows attackers have used. Each entry has a date added and often a weakness-type code (a CWE tag). It reflects CISA’s listing choices as well as attackers.

The disclosure dashboard’s counts don’t measure what share of each batch got patched, how fast fixes deploy, or how many systems stay exposed. The lab calls its patch count a lagging indicator. It says it discloses only part of what it finds because “independent human triage and review is the rate limiting step.”⁵⁵

Which kinds of fixing can make adoption the bottleneck?

The paper splits adoption in two.⁵⁶ **Patching** flaws AI finds plausibly *waits* for the capability, so it lengthens the transition without deciding whether adoption is the bottleneck. **Retirement and migration** plausibly run *alongside* AI's progress; only they can make adoption bind.

Patching is slow:

- At 12 firms, across 190 million vulnerabilities, median remediation was 90 days, with 32.5% open after 180 days (data to late 2018).
- On 8.4 million client computers, the median time to patch half of vulnerable hosts was 45 days. The study reports that “only 28% of the patches in our study reach 95% of the vulnerable hosts during our observation period.”
- In open source, over half of flaws were patched more than a week before disclosure, so attackers watching repositories “can often get a jump of weeks to months.”

Retirement and migration are slower:

- Agencies reported three of ten critical federal legacy systems modernized by February 2025.
- HTTPS in Chrome went from about 30–45% of navigations in 2015 to about 95–99% around 2020, then largely plateaued.
- Chrome planned about two and a half years from announcing the SHA-1 phase-out to removing support.
- IPv6 took about 27 years to pass half of access to Google on one day.

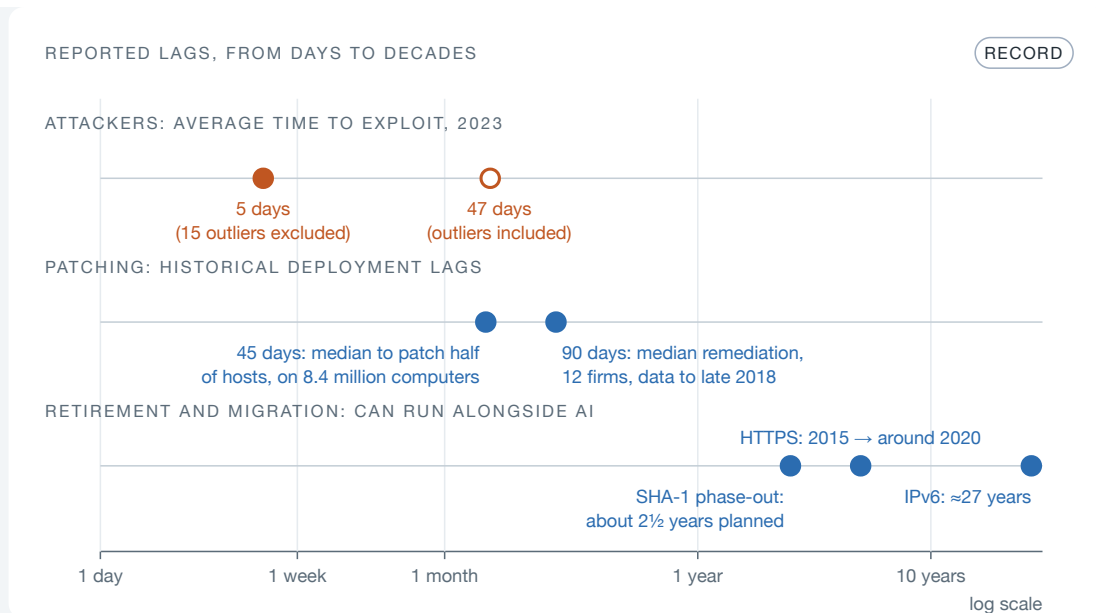


Figure 21. Reported lags on one log scale, placed together only to show their order of size. They come from different populations, years and endpoints and are not one measure. In the paper’s provisional classification, patching AI-found flaws follows the capability; retirement and migration can run alongside it and set the bottleneck.

– Why does patching count differently from retiring old systems?

Timing. A patch for a flaw AI finds can’t exist before the flaw is found, so patching follows the capability. Retiring a system or switching protocols can happen whenever operators choose. So only that work, running alongside AI’s progress, can set the speed threshold. Patching can still prolong the danger after AI has done its part; it just doesn’t set the threshold.

The paper’s judgment, and its limits. From the retirement and migration series, the paper judges that adoption already binds for the populations they cover: critical federal legacy systems and web-protocol migration. That is a provisional judgment, not an established result, and **not a judgment about patching in general**. The lags show deployment is slow. They don’t locate today’s speed against the threshold, and don’t show adoption speeding up enough for the escape.

One proposed reason, in Ross Anderson and Tyler Moore’s words: “Systems are particularly prone to failure when the person guarding them is not the person who suffers when they fail.”⁵⁷

What do CISA's listings show about flaw types?

The flaw tide's plateau depends on code protection can't reach. The paper made its own calculation from CISA's catalogue, using entries with a usable weakness tag. It compared entries added November 2021 through 2023 (920) with those added January 2024 to 20 October 2026 (659). The **access-policy** share rose from 19% (171 of 920) to 27% (180 of 659). The **memory-corruption** share fell from 33% (299 of 920) to 19% (125 of 659). Two robustness checks leave the direction unchanged. The mix reflects CISA's listing practice as well as attackers.⁵⁸

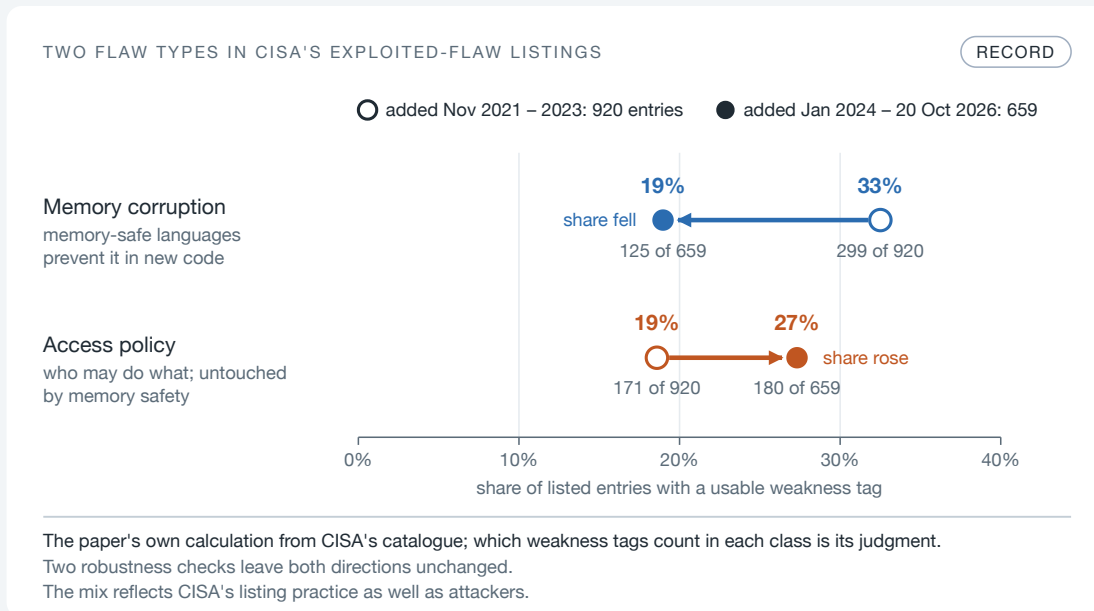


Figure 22. CISA's catalogue of exploited flaws, counting entries with a usable weakness tag. Between entries added November 2021 through 2023 (920) and those added January 2024 to 20 October 2026 (659), the memory-corruption share fell from 33% to 19% and the access-policy share rose from 19% to 27%. The shares are the paper's own calculation, with class lists that are its judgment; two robustness checks leave the direction unchanged, and the mix reflects CISA's listing practice as well as attackers.

– What is memory corruption, and why track its share?

Bugs where a program reads or writes memory it shouldn't, common in C and C++. Memory-safe languages prevent that class in new code, so its share is one rough sign of protection spreading. Access-policy flaws, mistakes about who may do what, are untouched by memory safety.

Some attacks skip software, and many defenses need no frontier AI

Some attacks skip software. Scattered Spider phones IT help desks “to reset passwords and/or transfer MFA tokens.” FinCEN reports criminals have reportedly used generative AI to impersonate “an executive or other trusted employee.” Adaptive attacks bypassed 12 recent defenses against AI prompt attacks, “with attack success rate above 90% for most.” A design-level defense, CaMeL, solved 77% of a benchmark’s tasks “with provable security,” against 84% undefended.⁵⁹

Lowering the peak doesn’t require running faster. Differential technology development holds that it “may be beneficial to delay risk-increasing technologies and preferentially advance risk-reducing defensive, safety, or substitute technologies.”⁶⁰ In the paper’s judgment, many defenses need no frontier AI: patch distribution, legacy retirement, memory-safe languages for new code, post-quantum migration, hardware-bound login, and verified kernels and libraries. AI bug-finding, patching and proof-writing serve attackers too, for general models.

The deciding comparison has never been measured

The crux is open, and empirical. Where adoption binds, the case for speed is right only if frontier AI shortens adoption faster, in proportion, than it raises danger.

The evidence on AI translating old code comes from benchmarks. On 100 C repositories, the best model tested solved 15 tasks in one attempt. Another system passed a formal check on 42% of 1,394 competition programs.⁶¹ DARPA’s automated patches are patch creation, not deployment. The paper has no measurement of how adoption responds to speed.

For policy: these defenses need no frontier AI, in the paper’s judgment. Whether speed helps depends on a number nobody has measured.

The head start is worth at most its length

An umbrella lent for part of a long walk keeps you dry for that part only. A secret capability works the same way. It splits the clocks for a while: some attackers have access anyway, such as states with their own frontier AI. The rest arrive when the **window** closes: open-weights catch-up, a rival's release, a leak or theft.

The paper's head-start limit:⁶²

A covered defender saves, in exposure time, at most the shorter of two spans: how long the window lasts, and how long that defender needs to deploy fixes. And only against attackers who lacked access.

For defenders who deploy well within the window, the share of exposure saved approaches the share of attackers locked out. For those needing many windows' worth of time, the share saved shrinks. It falls to roughly the window's average length divided by their deployment time, times that locked-out share. This assumes a window with a steady chance of ending (Figure 23).

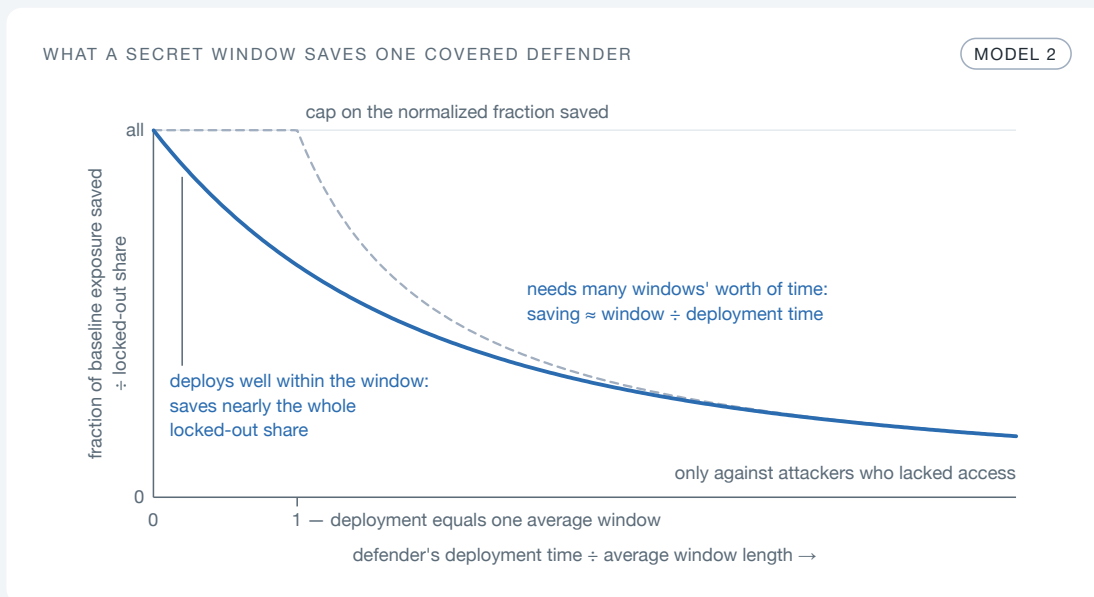


Figure 23. Model 2. The fraction of one covered defender's baseline exposure that a head start saves, divided by the locked-out share. It is plotted against how long that defender needs to deploy fixes, in average window lengths. The dashed line bounds that normalized fraction. It counts only attackers who lacked access, and the curve is for a window with a steady chance of ending.

Two parts carry policy weight:

- **The overall cap.** Across all defenders, a window can remove only a capped share of total exposure. The cap is the exposure held by fast defenders plus the program's coverage of everyone else, all scaled by the locked-out share. That holds whatever the window's length.
- **The exception.** Suppose harm grows faster than exposure time, so the longest exposures cost most. Then "worth at most its length" fails in harm units: the head start removes the costliest slice. The paper gives a replacement bound.

Conditions: harm grows in step with exposure, coverage doesn't depend on the window, and attackers resume at their old pace, with no backlog. Fast patchers gain most in proportion; whether they gain most in total needs evidence. The holder's partners are likely large firms that already patch fast, a judgment.

– What are "open weights"?

A model's weights are the numbers that make it work. An open-weight model publishes them: anyone can run and modify it, strip its safeguards, and it can't be recalled. A closed-weight model keeps them private and serves the capability through an interface.

How long a window lasts is unmeasured.⁶³ The lab restricted its model on 7 April 2026. An open-weight model, GLM-5.3, was released on 14 August; its weights went public two weeks later. The restricting lab reported results on one exploit benchmark. The open model built end-to-end exploits in 50 of 410 attempts, against 56 of 410 for the restricted model.

NIST's AI standards center placed it about four months behind the U.S. frontier among released models it evaluated. Epoch earlier gave a 90% interval of five to twenty-two months for open models' benchmark lag. But **catching up on benchmarks shows neither how fast copies spread nor how long a security window lasts.** The paper calls window length unverified.

Fixes are roadmaps, and holds protect unfound flaws

To protect outsiders from flaws it finds, the holder must ship fixes. But a public patch can guide attackers.

– **Why would a published patch help attackers?**

A patch shows exactly what changed, pointing to where the flaw was.

Researchers showed in 2008 that exploits can be generated automatically from patches. Every unpatched machine is then a known target.

The paper's "fixes are roadmaps" result compares holding the capability with releasing it, flaw by flaw, in three groups:⁶⁴

1. **Flaws the holder finds and patches.** Suppose the patch comes out on the same date either way. Then holding the whole capability adds nothing over an embargo on that flaw. That holds *unless* attackers would have found it themselves first; then holding can lower exposure. And if attackers with the capability turn patches into exploits faster, a hold protects these flaws too, for at most the window.
2. **Flaws the holder never finds.** Holding delays attackers' own search for them, for at most the window: the head-start limit again.
3. **Deployment-only tools.** Fix patch dates and the start of attacks. If a tool only helps defenders install published patches faster, withholding it never lowers outsiders' exposure. It raises exposure whenever the tool would have shortened deployment while attacks were under way.

The paper's model illustration of the second group uses 100 flaws, 90 never found by the holder. Expected exposure is 435 units under release and about 413.6 under a hold, though no patch ships for the 90. These are inputs chosen to test a claim, not measurements.

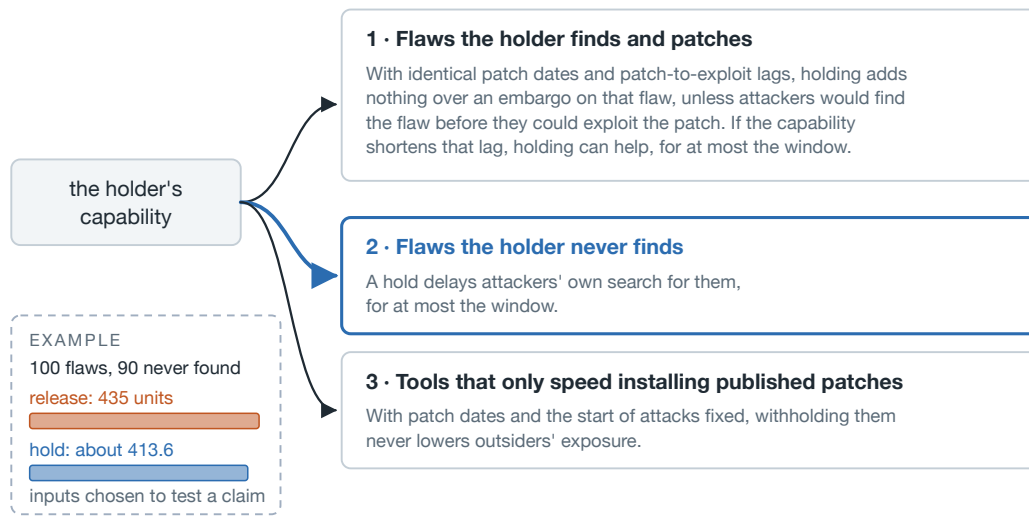


Figure 24. Model 2. The three groups of flaws in the “fixes are roadmaps” result. A hold of the whole capability protects outsiders through flaws the holder never finds, and only for at most the window. The example’s inputs were chosen to test a claim; they are not measurements. *Illustrative values.*

Three readings follow:

- A hold of the whole capability *does* protect outsiders, through unfound flaws, for at most the window. So the right comparison for a hold is **release with safeguards** that keep attackers out, not a flaw-by-flaw embargo. Whether such safeguards are feasible is open.
- The deployment-only result needs its conditions; if a tool also *produces* patches, it needs its own analysis.
- Coordinated dates can fail. Spectre was reported to vendors on 1 June 2017, for a coordinated date of 9 January 2018. It was published six days early “because of existing public reports and growing speculation.”⁶⁵

Sharing with no one, or with everyone, can be best

Each partner adds coverage and leak risk together. The paper’s coalition result: **either extreme can be best**, sharing with no one or with everyone eligible. Something in between is best if the first partner adds benefit and the last subtracts it.⁶⁶

Admitting a rival raises exposure saved only if it lowers the overall chance the window ends. The rival-release risk removed must exceed the leak risk added. That needs coverage, deployment times and attackers' access fixed, and a rival that can commit not to release. If admission also covers the rival's own systems, savings can rise even as the window shortens. In 2024 RAND judged protecting internet-connected models against the most capable attackers "currently not feasible."⁶⁷



Figure 25. Model 2. Each partner admitted adds coverage and leak risk together, so the best coalition can be no one, everyone eligible, or something in between. The three curves follow the paper's worked examples, each on its own vertical scale. Admitting a rival helps only if the rival-release risk it removes exceeds the leak risk it adds. This comparison keeps beneficiaries and exposure weights, coverage, deployment times and attacker access fixed, with some attackers excluded, positive covered exposure, and a rival that can commit not to release. *Illustrative values.*

A secret window raises racing when it widens the winner-loser payoff gap

A secret window is also a prize. Imagine a raffle where tickets cost money. Your chance of winning is your share of all tickets, and every ticket adds to a risk all

players share. (This is the paper's contest model in everyday terms.)

The paper's winner's-prize result:⁶⁸

- Racing effort depends on the gap between what the winner and the loser get.
- A window only the winner gets raises racing in proportion to how much it widens that gap, and the shared risk with it.
- A window both racers get leaves racing unchanged.

Whether extra racing harms the world is a separate question, and not automatic. It depends on the window's real value to everyone else, the holder's own real gain, and the cost of extra racing net of anything useful it produces. A private value bigger than the world's gain doesn't by itself make secrecy harmful. The claim that "even a modest escalation" by rivals outweighs the cyber gain is an empirical premise, not a theorem.

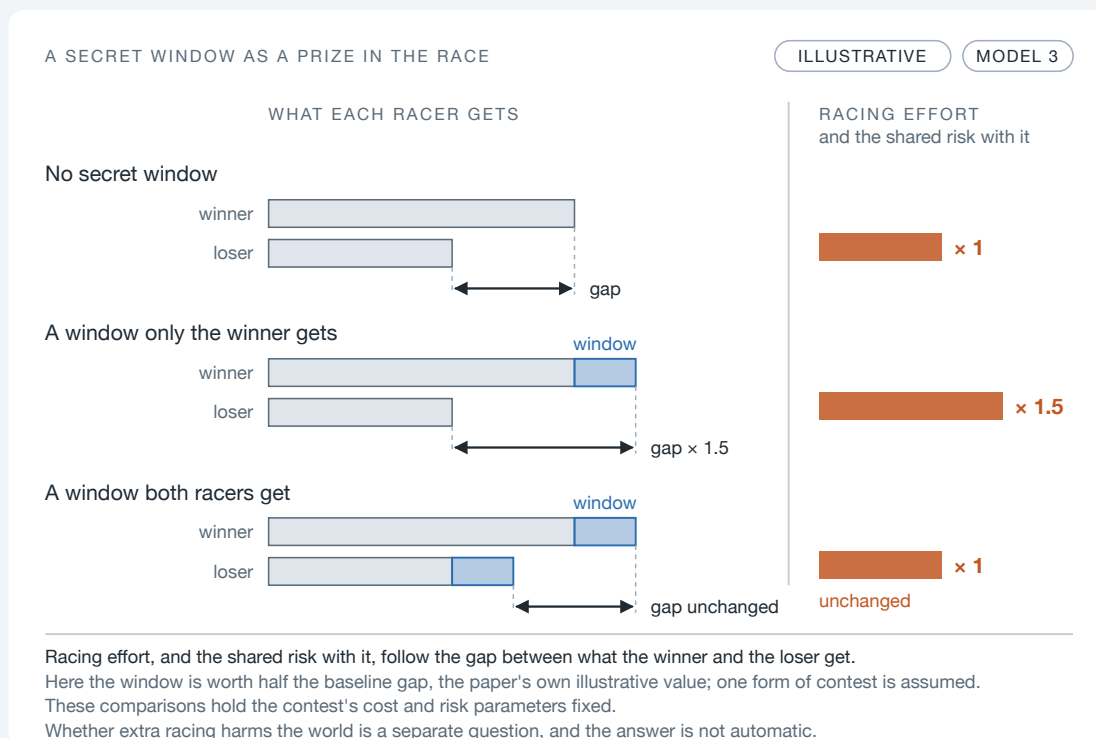


Figure 26. Model 3. A window only the winner gets widens the winner-loser gap and raises racing, and the shared risk with it, in proportion; a window both racers get leaves racing unchanged. The window here is worth half the baseline gap, the paper's illustrative value, and one form of contest is assumed, the raffle of Figure 31. These comparisons hold the contest's cost and risk parameters fixed. Whether extra racing harms the world is a separate question. *Illustrative values.*

How rivals answer a suspected secret lead has no settled sign. One lab wrote that every similarly capable model other than GLM-5.3 had been “released with safeguards or through limited access programs.” The paper says that fits norm-following, risk management or commercial choice, and identifies no motive.⁶⁹

Release with safeguards beats holding only if wider access speeds fixes enough

Toby Shevlane calls this structured capability access: the developer offers “controlled, arm’s length interactions” with a system, and “must both restrict how the AI system can be used, and prevent the user from circumventing these restrictions through modification or reverse engineering” of the system.⁷⁰

The paper’s safeguarded-release test first holds defenders’ schedules unchanged. Then release with safeguards **raises** outsiders’ exposure, if at all, by the chance safeguards fail times the extra exposure failure causes. It **lowers** exposure only if giving every defender access improves their fixing schedules by more than that. The test sets aside new risks serving the model might create, such as theft or copying. Adding them would raise exposure further.⁷¹

The test has a narrow frame. It compares two choices only: release now behind safeguards, or hold with no later release at all. It keeps everyone’s future access routes the same under both. And it assumes defenders’ schedules don’t depend on which events happen or on whether the safeguards fail.

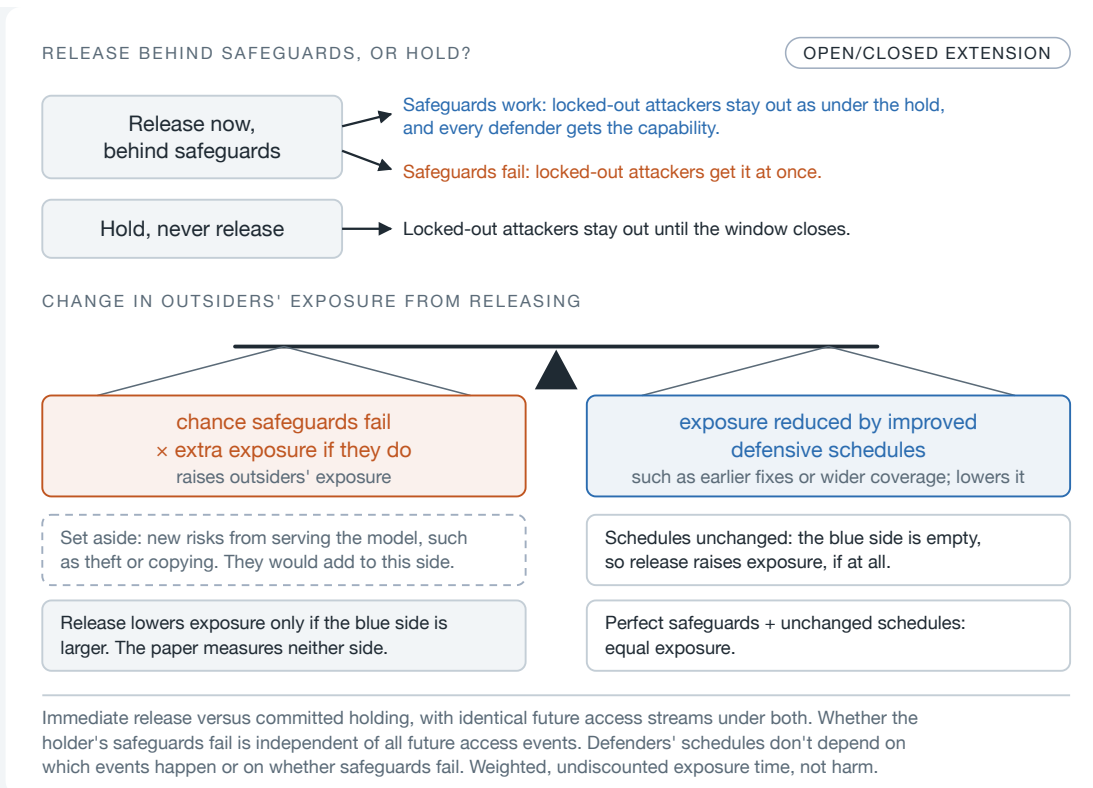


Figure 27. The safeguarded-release test, from the open/closed extension. Release now behind safeguards raises outsiders' exposure by the chance safeguards fail times the extra exposure failure causes; it lowers exposure only if the exposure reduced by improved defensive schedules is larger, and with schedules unchanged that side is empty. Perfect safeguards with unchanged schedules give equal exposure. The test compares immediate release with committed holding, keeps identical future access streams under both, assumes schedules don't depend on events or on safeguard failure, sets aside new risks from serving the model, and counts weighted, undiscounted exposure time, not harm. Whether the holder's safeguards fail is independent of all future access events.

– What is "release with safeguards"?

Releasing a capability through a controlled interface with limits on use, rather than publishing the model or holding it entirely. If safeguards hold, all defenders get it and attackers stay out. If they fail, attackers get it at once. How well real safeguards resist stripping and copying is unmeasured.

For policy: in the model a hold helps covered defenders by at most the window, only against locked-out attackers. Exposure time saved is not harm avoided or welfare gained. The fair test of a hold is release with safeguards, and it turns on safeguard durability and window length, both unverified.

PART 10

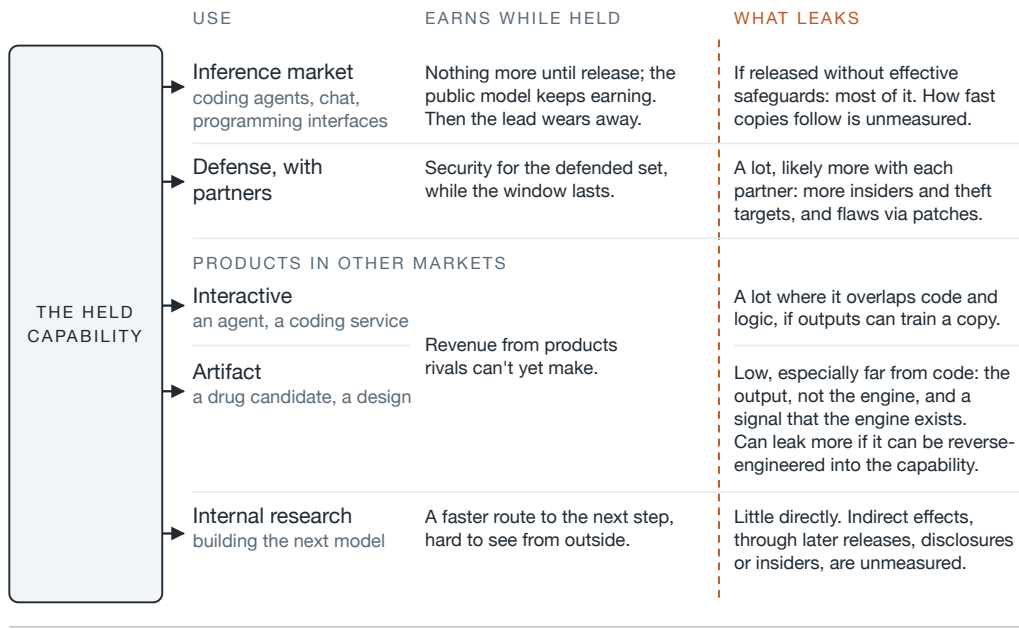
A held capability earns in four places, and each leaks differently

What leaks depends on the use, not on secrecy as such

A holder can earn in four places:⁷²

- the **inference market** (coding agents, chat, programming interfaces), where it earns nothing more until release and then its lead wears away;
- **defense with partners**, security while the window lasts;
- **products in other markets** (voice, video, agents, drug and materials discovery, design);
- **internal research**, using the model to build the next one, hard to see from outside.

What leaks depends on the use, not on secrecy as such. What follows is the paper's judgment, which it says needs evidence. Release without effective safeguards leaks most of the capability. Sharing with partners leaks a lot, likely more with each one. An interactive product leaks a lot where it overlaps code and logic. An artifact, like a drug candidate or design, leaks the output, not the engine, plus a signal the engine exists. Internal research leaks little directly.



What leaks depends on the use, not on secrecy as such.

Leak levels are the paper's judgment, in words; each needs evidence, and none is measured.

Figure 28. Model 4. The four places a held capability can earn, with products split into interactive services and artifacts, and what each use leaks. The leak levels are the paper's judgment in words and need evidence: an interactive product leaks a lot near code and logic if outputs can train a copy; an artifact can leak more if it can be reverse-engineered into the capability; internal research's indirect effects are unmeasured.

The paper offers three hypotheses to test. First, **artifacts may let a firm profit while holding**, so firms with rich artifact markets could hold longer. Second, **using an edge may reveal it, and value may attract theft**. The paper found no documented case of a capability copied soon after a demonstration, and no measure of theft against value. Third, **artifacts may turn a temporary lead into lasting assets**: patents, designs, data, locked-in customers.

Two kinds of copying push release in opposite directions

Imagine a bakery with an unlaunched recipe. Rivals copying its current bread shrink sales each month, which pushes it to launch sooner. If rivals will copy the new recipe quickly too, launching is worth less, which pushes it to wait. (An illustration of the two copy clocks.)

The paper's two-copy-clocks result treats timing as a stopping problem:⁷³

- **Faster copying of the current public model** erodes revenue, so holding costs more. **Release comes sooner.**
- **Faster expected copying of the model it would release** lowers what release is worth. For revenue that rises in a straight line with the lead, and for a broad family of curved patterns, **release comes later.**

These directions are local: they hold at a release date in the middle of the range, other things fixed. Under revenue that flattens out at the top, faster post-release copying can bring release sooner. The second direction also holds the risk of losing the secret fixed. If faster copying also raises that risk, and release is worth more than what survives a loss, it adds a push the other way.

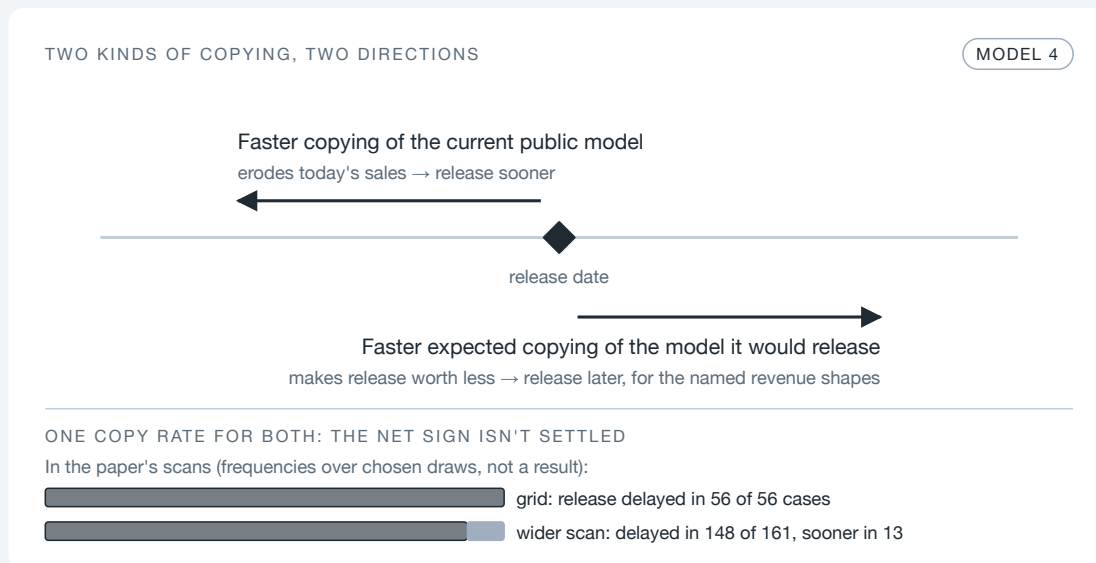


Figure 29. Model 4. The two copy clocks push a release date in opposite directions. Both directions are local, at a release date in the middle of the range with other things fixed; the later-release push holds for the named revenue shapes and at a fixed risk of losing the secret. The scan counts are frequencies over chosen draws, not a result.

— **Wait, so faster copying makes firms wait? Isn't that backwards?**

It depends which copying. Copying of what's on the market pushes release sooner; expected copying of what you'd release pushes it later, for the named revenue shapes. With one copy rate for both, the net sign isn't settled. In the paper's scans, faster common copying delayed release in all 56 cases of one grid. In a wider scan it delayed release in 148 of 161 draws, against 13 where it

hastened release. The paper calls these frequencies over chosen draws, not a result.

Corners are common: often it's release now, or hold until the secret is lost. With straight-line revenue, a mid-range release date is strictly best only under one condition. The released model must be copied faster than the current one by more than the risk of losing the secret. That's necessary, not sufficient. Below that rate, the best choice is a corner: release at once, or hold until the secret is lost. At exactly that rate, the firm can sometimes be indifferent among every date. With constant earnings and no public model to lose, it's all-or-nothing, and faster post-release copying pushes toward holding. Whether released models are copied that much faster is unmeasured.

So the inference market doesn't, in general, force held capabilities out on a copy-speed schedule. The pressure is real, from erosion of the current lead. But expected quick copying after release is, for the named revenue shapes and at fixed loss risk, a reason to hold. Which clock wins is empirical; it needs prices per unit of capability and customers' switching and loyalty.⁷⁴

The commercial lead never outlasts attackers' exclusion

The paper's extension separates **two windows**.⁷⁵ The **commercial window** is how long the holder keeps its market lead. The **attacker-access window** is how long locked-out attackers stay out.

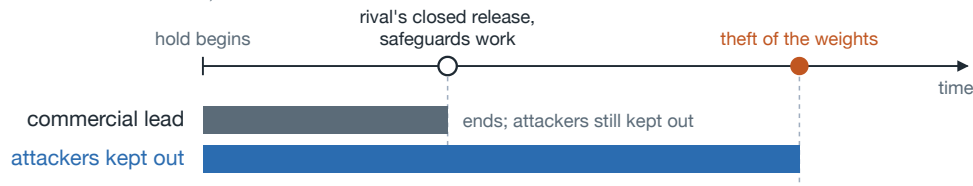
Five events can end the lead: a rival's comparable open release; a rival's closed release whose safeguards fail; one whose safeguards work; theft of the weights; a partner's leak. All end the commercial lead. All but the working closed release also give attackers lasting capability.

So the commercial window is never longer than the attacker-access window. That rests on a premise: theft and leaks hand rivals a commercial substitute at once. Covert theft that armed only attackers would require revising it.

EVENT	ends the commercial lead	gives locked-out attackers lasting capability
A rival's comparable open-weight release	yes	● yes
A rival's closed release whose safeguards fail	yes	● yes
A rival's closed release whose safeguards work	yes	○ no
Theft of the weights	yes	● yes
A partner's leak	yes	● yes

Four event types can end both windows together; a rival's closed release whose safeguards work can end commercial leadership while attackers remain excluded.

ONE ILLUSTRATIVE PATH, UNDER COMMITTED HOLDING



Commercial window $\leq$ attacker-access window, on every path.

Premise: theft and leaks hand rivals a commercial substitute at once; covert theft that armed only attackers would require revising this. The path's timing is illustrative; how long either window lasts is unmeasured.

Figure 30. The two windows, from the open/closed extension. All five event types can end the commercial lead; all but a rival's closed release whose safeguards work also give locked-out attackers lasting capability, so the commercial window is never longer than the attacker-access window. The timeline is one illustrative path under committed holding. The result assumes theft and leaks hand rivals a commercial substitute at once.

The paper judges, without evidence yet, that an open copy erodes the lead faster. Its safeguards can be stripped: “abliterated versions” of the 2026 open-weight model appeared “within days.” It can’t be recalled.⁷⁶

The reach limit. Levers that bind only covered domestic developers can’t stretch the expected attacker-access window past a ceiling. What they don’t reach sets that ceiling: unbound developers, theft and leaks. This assumes those outside access routes stay above zero and unchanged; reducing theft or leaks changes the ceiling. This ranks no interventions.⁷⁷ The extension sets aside the holder’s profit, so none of it carries over to the paper’s commercial results.

Side uses, theft and lasting assets change what a secret is worth

The side-business rule. A new use of a held capability is worth adding when what it earns, per unit of extra leak risk, exceeds what a leak would destroy. Theft drawn by earnings counts as added risk.⁷⁸ The rule assumes the hold ends only by involuntary loss. With a live release option it's the wrong yardstick. A use the firm commits to can pass the rule and still lower the firm's value, by destroying the release it was waiting for. An optional use couldn't. The rule takes given numbers; nothing follows about real products.

The theft ceiling. Theft caps a secret's value only if theft eventually rises at least in proportion to what the secret earns. If theft rises more slowly, tracks total value rather than earnings, or security is cheap enough, there's no cap. The proportionality premise is the paper's own; RAND ties needed security to capabilities and stakes, not earnings.⁷⁹

Perishable leads. A lead others copy is worth less than one they can't, a valuation identity.⁸⁰ Faster copying raises a lasting prize's value *relative to* a wearing lead, not its own value. It says nothing about how hard firms chase it.

Durable assets: lasting earnings make each step's race fiercer. Suppose part of each step's earnings is, from the start, something copying can't erode.⁸¹ Then the decisive, lasting prize looks smaller relative to a step's. Winning each step is worth more, so step racing rises. That holds only if the winner alone gets those lasting earnings. It also needs the step prize to count none of the value of going on to win the decisive prize.

The paper also works one specification. There the two contests share one cost that rises with combined effort, so total effort can change. Racers bear none of the catastrophe loss themselves. In it, effort toward the decisive prize falls as the step prize grows. Total risk rises only if decisive-race effort isn't too much riskier per unit (the paper gives the cut-off). With separate budgets, decisive effort doesn't move. The evidence to gather: labs' patents and product launches outside the inference market, timed against capability jumps.

Using the model to build the next one is hard to see

Using a held model to build the next serves both money and security and is hard to see. It could speed the general race with little visible sign; the paper calls that a

hypothesis and doesn't model it. It also deepens a signaling problem: outsiders can't tell holding for security from holding for profit.⁸²

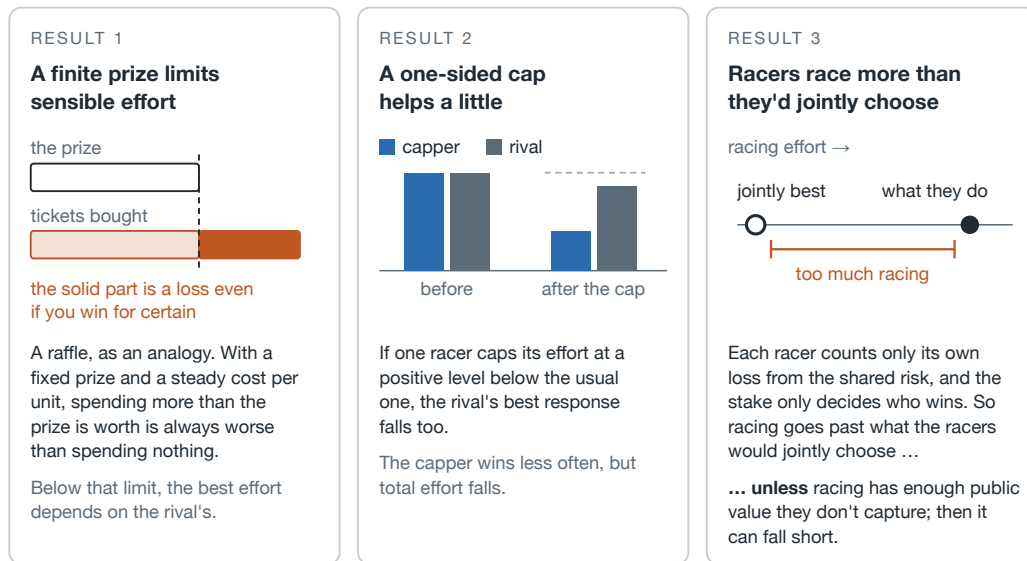
For policy: commercial pressure to release is real but doesn't set a copy-speed schedule. These results describe one firm's private choices, not what is best for the world.

PART 11

Actors race anyway, and restraint needs detection

If speed doesn't lower the near peak where adoption binds, why race? The paper's contest results answer in four parts (Figures 31 and 32):⁸³

1. **A finite prize limits sensible effort.** With a fixed prize and steady cost per unit, spending more than the prize is worth is always worse than spending nothing. In the raffle game, the best effort depends on the rival's, so no single effort is best against everything.
2. **A one-sided cap helps a little.** If one racer caps its effort at a positive level below the usual one, the rival's best response falls too. The capper wins less often, but total effort falls.
3. **Racing exceeds what racers would jointly choose,** unless it has enough public value they don't capture.
4. **Restraint needs detection.** In a repeated race where cheating is caught with some chance and triggers racing forever after, restraint holds only if detection reaches a threshold. For impatient racers that threshold can exceed certainty, so no detection technology is enough. Under full secrecy, restraint fails. In a one-shot race, restraint holds only if the expected penalty on detection is at least half the stake.



One form of contest. Bar heights in result 2 are illustrative. The fourth result, detection, is Figure 32.

Figure 31. Model 3. Three of the contest results. A finite prize limits sensible effort (shown with a raffle, as an analogy); a one-sided cap lowers the rival's effort too, so total effort falls though the capper wins less often; and racing goes past what racers would jointly choose, unless it has enough public value they don't capture. All assume one form of contest. *Illustrative values.*

— Can't labs just agree not to race?

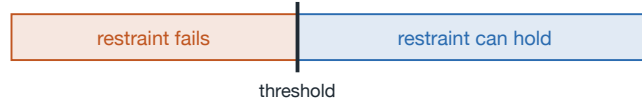
In the model, only with detection good enough to catch cheating often enough and a credible penalty. The model excludes false alarms, which in real monitoring can collapse restraint with nobody cheating. Arms-control research shows why such deals are rare: monitoring must be open enough to assure compliance but not so open it threatens security.

REPEATED RACE

Each cheat is caught with some chance; once a cheat is caught, racing resumes forever.

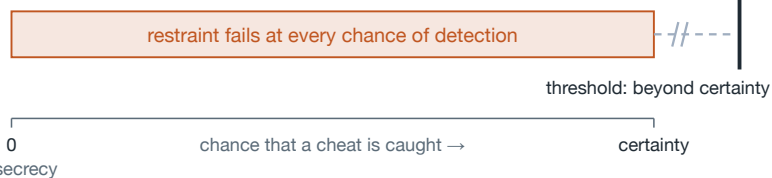
Patient racers

value the future highly



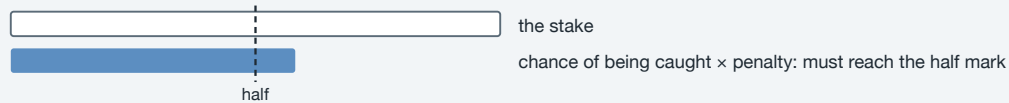
Impatient racers

value the present more



ONE-SHOT RACE

Restraint holds only if the expected penalty on detection is at least half the stake.



The model excludes false alarms; in real monitoring they can collapse restraint with nobody cheating.

Where each threshold sits is illustrative; the paper gives the formula, not a measured value.

Figure 32. Model 3. Restraint needs detection. In the repeated race, restraint can hold only if the chance of catching a cheat reaches a threshold; for impatient racers the threshold can exceed certainty, and under full secrecy restraint fails. In a one-shot race, the expected penalty on detection must be at least half the stake. The model excludes false alarms. *Illustrative values.*

Other pushes sit alongside. U.S. policy frames AI as a race: “The United States is in a race to achieve global dominance in artificial intelligence (AI).” The paper judges that sunk capital and falling computing prices may push labs to ship. Neither has a measured size.⁸⁴

PART 12

The models settle three things; the rest is conditional

The four models stay separate (Figure 3), and exposure saved is not harm or welfare avoided. The side-business rule and the theft ceiling assume no planned release, and the open/closed extension does not extend the commercial results. So no “best policy” follows; what follows is narrower.

Three things hold with no estimate of size

These need no estimate of size (Figure 33):⁸⁵

1. If the aim is outsiders' exposure, don't withhold tools that only speed deployment of published patches. With patch dates and attack onset fixed, withholding never lowers that exposure.
2. A window short compared with a defender's deployment time is worth little to that defender, and only against locked-out attackers.
3. Judge a capability-level hold against release with safeguards, not a flaw-level embargo, because a hold also slows attackers' search for unfound flaws, for at most the window.

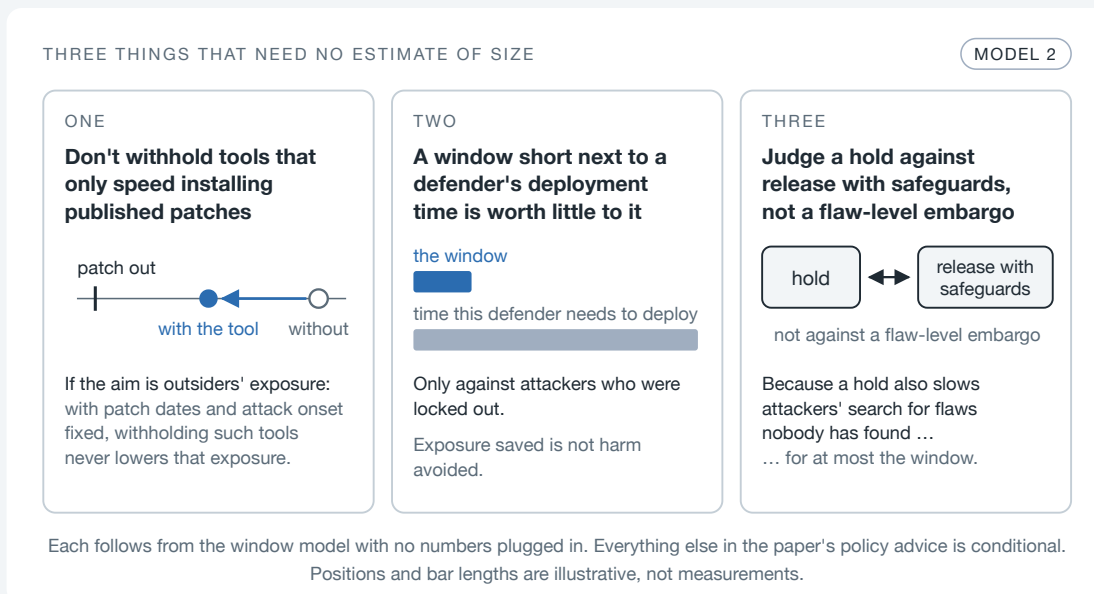


Figure 33. Model 2. The three comparisons that need no estimate of size, each with its condition. Don't withhold tools that only speed installing published patches, if the aim is outsiders' exposure and patch dates and attack onset are fixed; a window short next to a defender's deployment time is worth little to that defender, and only against locked-out attackers; judge a hold against release with safeguards, because it also slows attackers' search for unfound flaws, for at most the window. *Illustrative values.*

Everything else is conditional.

Some pressures push release sooner, others push holding longer

In the holder's model, pressures work through five quantities. They are what the held capability earns, how fast the hold ends involuntarily, and how fast the current model is copied. They also include what release restores and how fast that wears, and how heavily the firm discounts the future.⁸⁶ Directions from the model are local, other things fixed (Figure 34):

- **Toward release sooner:** copying of the current model; a rival that may release first; faster open-weights catch-up (the last two when release is worth more than what survives a loss); cash burn, for straight-line revenue (either way for curved).
- **Toward holding longer:** expected copying of the released model, for the named revenue shapes at fixed loss risk; rich artifact markets far from code, if the added risk is small; internal research that adds no risk, if its value ends at release; defense contracts, if contract earnings end at release and the expected cost of the leak risk they add is less than the extra income. Research whose value outlasts both release and a leak has no effect on timing.
- **Theft drawn by earnings** caps the hold's value only if it rises at least in proportion.
- **By judgment alone:** low customer switching costs push toward release sooner; liability for releasing an offense-capable model pushes toward holding or release with safeguards.

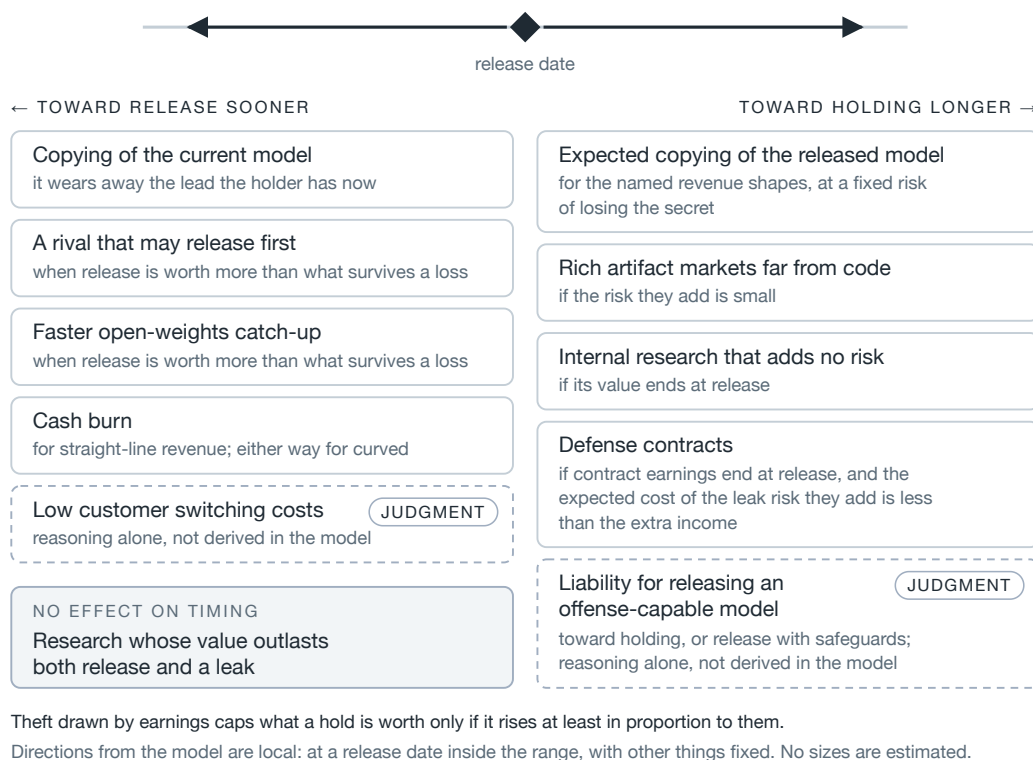


Figure 34. Model 4. Pressures on the holder's release date, each with its condition. The model's directions are local, other things fixed; the two dashed items are judgments, not derived in the model. Research whose value outlasts both release and a leak has no effect on timing.

Each lever serves a goal, and the paper ranks none

The paper's levers read "if the goal is X, these are candidates" (Figure 35). **It does not rank them.**⁸⁷

- **Lower near-peak cyber harm:** defenses that need no frontier AI (patch distribution, legacy retirement, memory-safe languages for new code, post-quantum migration, hardware-bound login, verified kernels and libraries). Their effects and costs need evidence. Whether speed also helps depends on the unmeasured comparison of adoption and danger.
- **Less exposure during a window:** shorten deployment, widen coverage, extend attackers' exclusion. Mandates, central update channels and deadlines are proposed ways; their effects and leaks need evidence.
- **Protect outsiders from flaws the holder finds:** flaw-level embargoes already exist; judge a full hold against release with safeguards. Common safeguard

standards could make that comparison more credible, in the paper's judgment.

- **Avoid harming outsiders by holding:** don't hold back deployment-only tools, such as, in the paper's judgment, rollout software that schedules and tracks installation.
- **Joint restraint:** detection at the threshold and a credible penalty. Coordinated release embargoes would remove first-mover pressure among members. A shared defense program reaches more defenders but adds leak risk. Shared capability tests may calm or inflame the race.
- **Health benefits from the same capability:** faster AI brings them forward only while AI's progress sets their arrival, and with speed-set danger it raises cyber danger. Shortening the benefit's own clock, through trial and approval capacity, manufacturing and uptake, is a separate lever. That it leaves cyber danger unchanged is an assumption (Part 6).⁸⁸

Government levers, if the goal is longer useful windows and wider coverage, are each a judgment from the model:

- Liability for harms from released models lowers the value of release without safeguards.
- Pre-release evaluation windows that bind every domestic developer to a common date remove domestic racing to release; separate waits don't.
- Executive Order 14409 instructs agencies to design a voluntary framework for government access to covered frontier models for up to thirty days before planned release to other trusted partners. The order doesn't show implementation.
- Weight security lowers theft risk. Government-bought defensive access could steer coverage toward slow operators. Export controls slow catch-up only where they reach.
- Oversight of internal use would make a hidden use visible. Downstream market concentration deserves watching.⁸⁹

IF THE GOAL IS ...	... THESE ARE CANDIDATE LEVERS	NOT RANKED
Lower near-peak cyber harm	→ Defenses needing no frontier AI: patch distribution, legacy retirement, memory-safe languages for new code, post-quantum migration, hardware-bound login, verified kernels and libraries. Effects and costs need evidence. Whether speed also helps depends on the unmeasured comparison of how adoption and danger respond to speed.	
Less exposure during a window	→ Shorten deployment, widen coverage, extend attackers' exclusion. Proposed ways: mandates, central update channels, deadlines.	
Protect outsiders from flaws the holder finds	→ Flaw-level embargoes already exist; judge a full hold against release with safeguards. Common safeguard standards could help (judgment).	
Avoid harming outsiders by holding	→ Don't hold back deployment-only tools, such as, in the paper's judgment, rollout software that schedules and tracks installation.	
Joint restraint	→ Detection at the threshold and a credible penalty (Figure 32); coordinated release embargoes; a shared defense program, which adds leak risk; shared capability tests, which may calm or inflame the race.	
Health benefits from the same capability	→ Faster AI brings them forward only while AI's progress sets their arrival, and raises cyber danger when danger is set by speed. A separate lever: shorten the benefit's own clock (that cyber danger stays unchanged is assumed).	
Government: longer windows, wider coverage	→ Liability; one evaluation date binding every domestic developer; weight security; government-bought defensive access; export controls where they reach; oversight of internal use.	JUDGMENT

The paper lists candidates for each goal and ranks none: costs, effects and leaks are not estimated.

Figure 35. Each goal and the levers the paper names as candidates for it. The paper ranks none of them; their effects and costs need evidence, and the government levers are judgments from the model.

– Can't a government just make everyone wait?

It can bind the developers it reaches. The reach limit says levers on covered domestic developers can't stretch the expected attacker-access window past a ceiling (Figure 36). Outsiders, theft and leaks set that ceiling, as long as those routes stay open and unchanged. Cutting theft or leaks moves the ceiling. A domestic group that holds while outsiders release may hand them market share, a judgment that needs evidence.

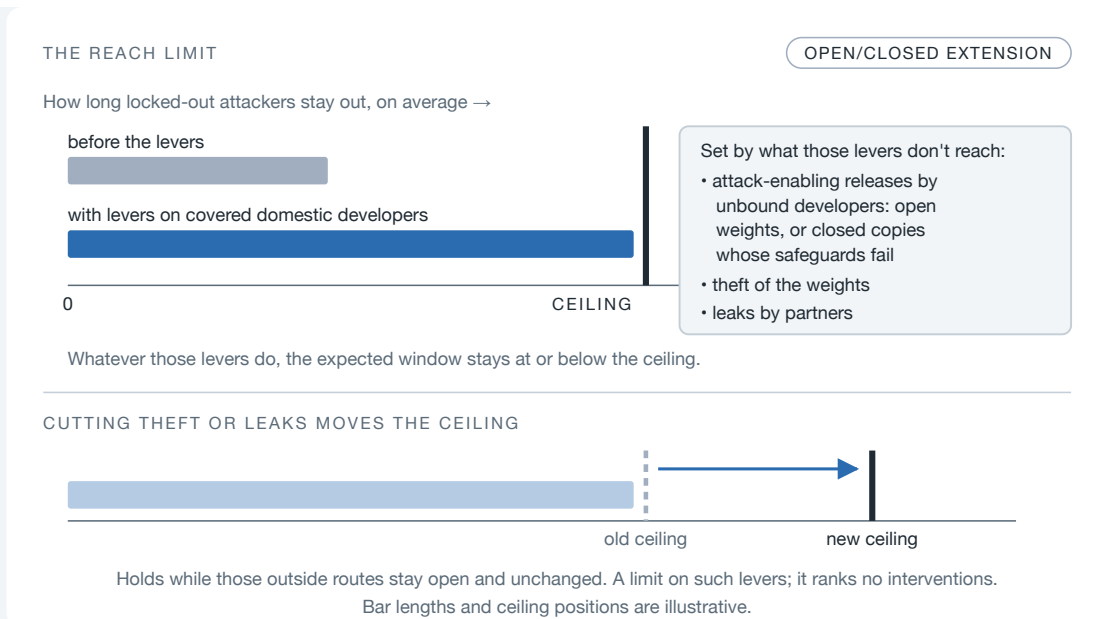


Figure 36. The reach limit, from the open/closed extension. Levers that bind only covered domestic developers can't lengthen the expected attacker-access window past a ceiling set by what they don't reach: attack-enabling releases by unbound developers (open weights, or closed copies whose safeguards fail), theft and leaks. Cutting theft or leaks moves the ceiling. This assumes those outside routes stay open and unchanged, and it ranks no interventions. *Illustrative values.*

Three limits apply to all of it

1. **Coverage.** Domestic rules leave outsiders' catch-up untouched, though input controls may slow it. That catch-up may be the main way holds end; this needs evidence.
2. **Coordination can look like collusion.** An agreement on release timing is, in form, an agreement to restrict output (the paper's framing). Antitrust law exists "to prevent collusion between companies that has a negative effect on consumers." Whether safety coordination needs a legal safe harbor is open.⁹⁰
3. **No ranking.** The head-start limit bounds a window's value but ranks nothing without costs, effectiveness and leak effects, none estimated.

— Does the paper say we should slow down AI?

No. It says that past the speed threshold, under stated conditions, speed alone doesn't lower near-peak cyber risk in the model. Before the threshold, speed can

help. Where we sit is unmeasured. Its policy statements are conditional on goals, and it draws no ranking of interventions.

PART 13

Every result rests on conditions, and the deciding numbers are unmeasured

Figure 37 sorts the results by where they stand.

The models.⁹¹

- Each result holds under stated conditions, and the conditions do work.
- The risk tally is a modeling choice; a step cost independent of speed is an assumption.
- The peak model holds exploitation per flaw constant and contains no AI speed.
- The two-storms test assumes far risk rises with speed.
- The window results assume harm in step with exposure, coverage independent of the window, and no attacker backlog.
- The holder's model assumes steady earnings, leads eroding at a steady rate, a steady loss risk and smooth revenue curves. Its directions are local.
- The contest results assume one form of contest.
- The benefits extension puts health gains and cyber harm in one unit, DALYs, for that extension only. It assumes a steady benefit, a fixed health loss per harmful event, and full coupling to the AI that arms attackers. It is a conditional balance, not a welfare measure.
- The four models answer to different goals; the paper combines them into no welfare measure and ranks no interventions (Part 1).

The evidence.⁹²

- Much of the capability and restricted-release evidence comes from one lab's reports on its own model. A government competition and a government

institute, independent of the lab, point the same way, but neither measures harm in the field.

- Checking a lab's report shows what it reported, not that its figures are right; independent replication remains unverified.
- How long a security window lasts is unverified; benchmark catch-up doesn't measure it.
- Where today's AI speed sits against the threshold is unmeasured, and so is how fast adoption responds to speed.
- "Adoption binds" is a provisional judgment about the named retirement and migration series, not patching in general.

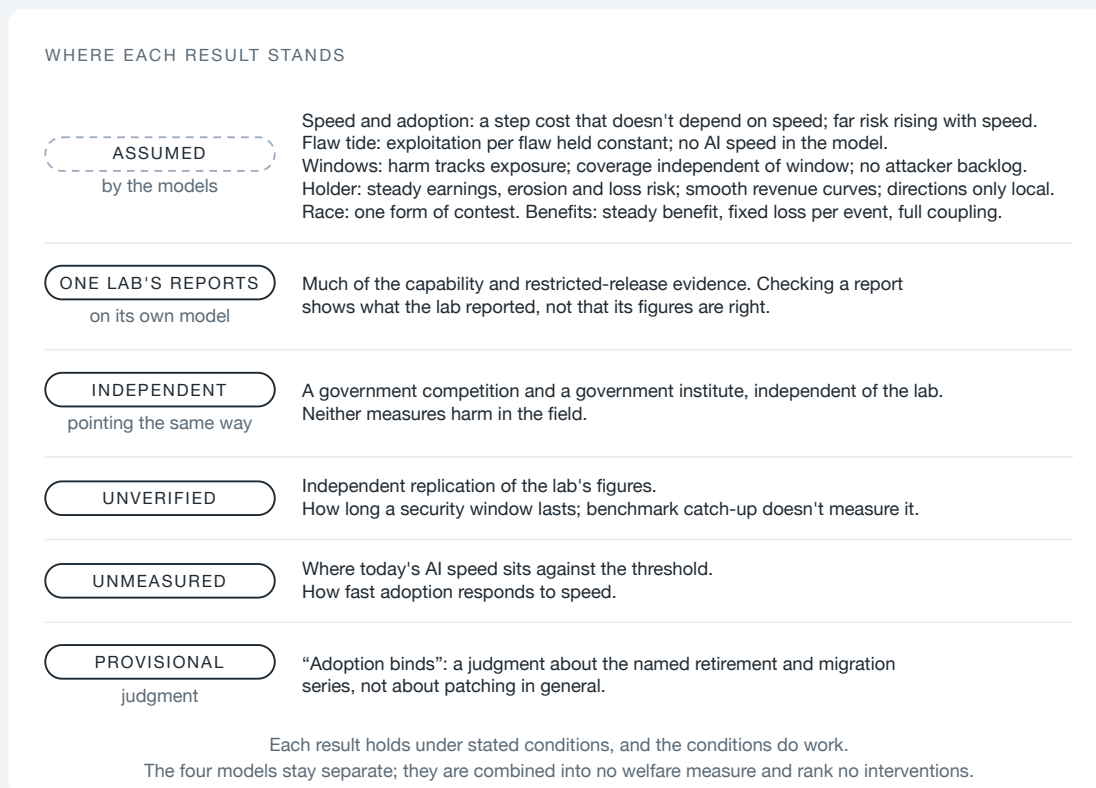


Figure 37. Where each result stands: what the models assume, what rests on one lab's reports on its own model, what independent sources point the same way on, and what remains unverified, unmeasured or a provisional judgment.

What would change the paper's provisional judgments (Figure 38):⁹³

- AI that shortens adoption faster than it raises danger;
- windows held for years under verified multilateral restraint that binds open-weights developers;

- commercial leads resting mainly on things copying doesn't erode, such as distribution, enterprise trust and compliance;
- release decisions that closely track the open-weights gap, which would suggest, without proving, that the release-sooner clock dominates.

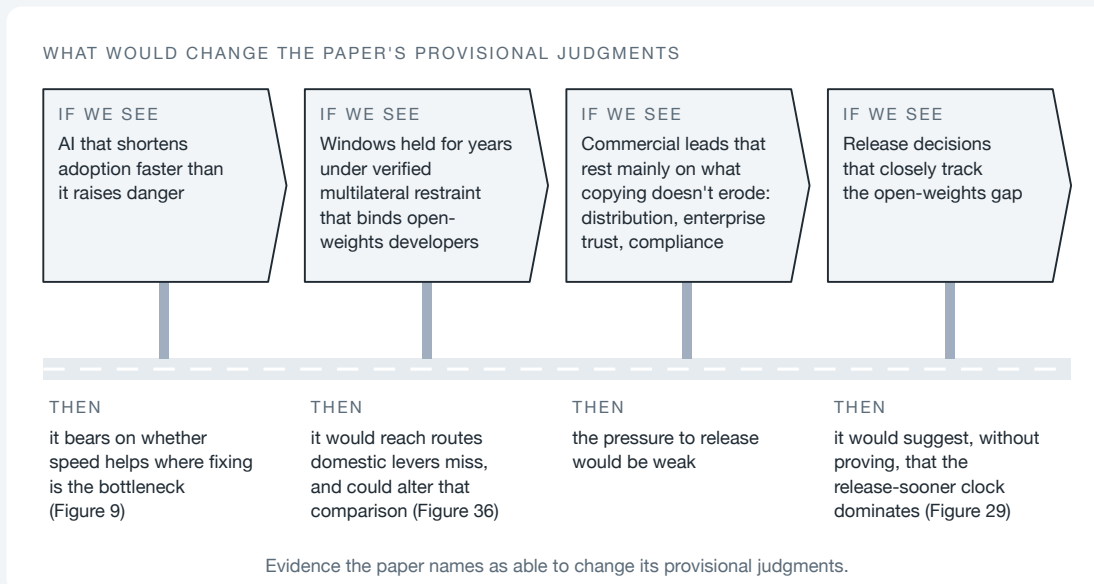


Figure 38. Four kinds of evidence the paper names as able to change its provisional judgments, and what each would bear on. The last would suggest, without proving, that the release-sooner clock dominates.

PART 14

Three dated predictions could count against the paper

The paper sets two kinds of dated check. Three are **predictions**, with a resolution rule fixed now and a result that counts against the paper's view. They are indicator forecasts: each resolves on a published count, and none by itself tests the mechanism it bears on. Nine are **measurements**, with no predicted value.⁹⁴

Three predictions, each with a line to beat

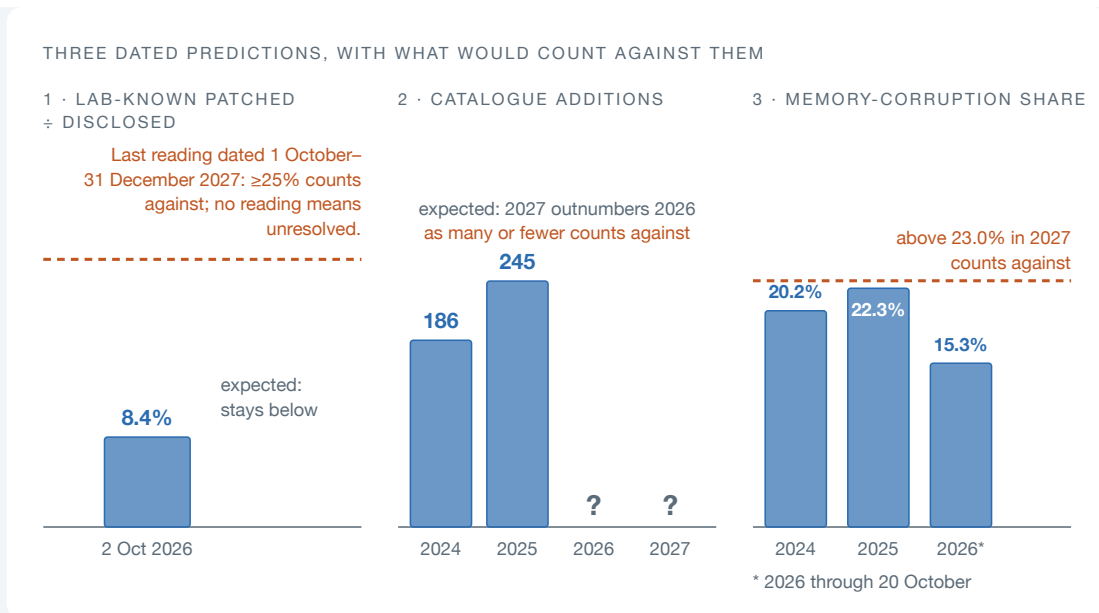


Figure 39. The three predictions, each with its recent baseline and the line that would count against the paper’s view. Each resolves on a published count; none by itself tests the mechanism it bears on.

1. Patches behind disclosures, to 31 December 2027.

- *Bears on:* adoption time in the fixed-curb rule and the accelerationist’s escape, at its first step, patch creation, for open-source projects. A monitoring indicator.
- *Expected:* patch creation keeps lagging AI-scale disclosure.
- *Resolution:* take the last reading on the lab’s disclosure dashboard dated on or before 31 December 2027. The known-patched count divided by the disclosed count is below 25%.
- *Window:* that reading must be dated on or after 1 October 2027. Without one, the check does not resolve.
- *Baseline:* 516 of 6,157, or 8.4%, on 2 October 2026.
- *Counts against:* a ratio of 25% or more.
- *Limits:* the ratio moves with the lab’s disclosure pace and triage as well as patching: slower disclosure raises it, faster lowers it, with no change in patching speed. It pools all disclosures, so it isn’t a cohort patch rate. It counts patches created, not deployed, so a high ratio wouldn’t show deployment keeps pace. It can’t identify whether adoption outruns danger. The dashboard counts

patches only “to our knowledge,” so an undercount would bias the check toward the paper’s expectation.

2. Exploitation still rising, calendar 2027.

- *Bears on:* the attacker-favoring hypothesis, and the rising phase of the flaw tide.
- *Expected:* the count of exploited vulnerabilities keeps rising.
- *Resolution:* in CISA’s known-exploited catalogue, version current on 31 January 2028, entries with a date added in 2027 outnumber those with a date added in 2026.
- *Counts against the indicator:* as many in 2027 as in 2026, or fewer.
- *Limits:* the count reflects CISA’s listing practice as well as attackers, and counts first listings, not attacks. Two things can move it with no change in exploitation. One is batch listing of old flaws: in 2022, 381 of the 555 additions, 68.6%, were flaws at least three identifier-years old. The other is disclosure volume, which recent exploitation counts have tracked in one threat-intelligence group’s reading.

3. Memory-corruption share held down, calendar 2027.

- *Bears on:* the unprotected residue of new flaws in the flaw tide.
- *Expected:* the share doesn’t rise above its recent range.
- *Resolution:* apply the supplement’s data note unchanged to entries with a date added in 2027, in the catalogue version current on 31 January 2028. Among entries with a usable weakness tag, the memory-corruption share is at or below 23.0%.
- *Classification:* an entry’s weaknesses are its CWE tags. Entries with none, or only the “no info” or “other” placeholders, are excluded. An entry with several goes to the first matching class, in the order memory corruption, injection and input handling, access policy, other. The class lists are the paper’s judgment, printed in the supplement.
- *Counts against:* a share above 23.0%.
- *Baseline:* the cut sits above each recent yearly share: 20.2% (36 of 178) in 2024, 22.3% (49 of 220) in 2025, and 15.3% (40 of 261) in 2026 through 20 October.

- *False-alarm chance*: suppose the share were a fixed 19.0% (125 of 659), with independent entries. A standard probability calculation then puts the chance a year's share exceeds 23.0% at 4.4–7.4% for 220 to 261 usable entries. At 2024's count of 178, it is 10.1%. At an earlier cut of 19.0% it was about one half. A calculation under stated conditions, not a measured error rate.
- *Limits*: the share reflects listing practice as well as attackers, and identifies neither what lowers it nor the flaw-tide dynamics. If AI makes old C and C++ code cheap to mine for flaws, the share could rise with no change in how many flaws new code introduces.

Nine more things to measure, with no predicted outcome

- **Patch lags without mandates, 2027**: whether any class of operators reaches median patch lags under two weeks without a mandate.
- **The next restricted capability, through 2028**: whether it holds more than six months before open-weights parity, and whether it ends by its holder's release or by parity.
- **Closed-weight copies, through 2028**: the lag from a frontier release to the nearest closed-weight copy, not yet measured.
- **The two copy clocks, 2027–28**: whether the time between frontier releases shortens as the open-weights lag shortens.
- **Uses outside the inference market, 2027–28**: whether labs' launches outside the inference market cluster in the months after capability jumps.
- **Adoption that speeds up, 2027–28**: whether verified translation and rewriting reach production code, and whether AI patching narrows the gap between flaws found and fixed.
- **Slow adoption populations, 2027–28**: adoption times for maintainers, small operators and industrial-control operators, which none of the paper's series covers.
- **Differential uptake, 2027–28**: whether defenders' uptake of AI tools outpaces attackers', the signal for the speed case the models don't cover.
- **Where the wait for a health benefit goes, 2027–28**: for AI-assisted drug programmes, measured separately: the wait for the capability; the benefit's

own adoption time (trials, approval, manufacture and uptake); and the share of that adoption that waits for the capability. The share of the whole wait that AI sets is derived from these and is not that last share. No source the paper read gives any of the three, or the DALYs lost per harmful cyber event.

PART 15

The bottom line: others build the curb, and running thickens the rain

Running through the rain works because the curb is fixed and the rain doesn't thicken. The paper's models assume neither holds on the road to superintelligence. Operators build the curb on their own clock, and running arms attackers as it arms defenders.⁹⁵

A peak can exist in the flaw-stock model; whether AI creates or moves it is not modelled.

Where adoption sets how long the danger lasts, running raises near-peak risk. It helps only if it speeds adoption faster than it raises danger. That holds with adoption time fixed and running at least partly alongside AI, under the stated conditions. Where AI's progress is the bottleneck, running can help. Nobody has located the threshold or measured the comparison that would decide it.

Life-saving benefits from the same AI add a second reason to run. They can justify speed even while near harm rises, and then the far storm matters too. But they do so only while the benefit still waits on AI rather than on its own trials, approval and uptake. And they do so only if the benefit brought forward outweighs the added cyber loss.

Holding the storm helps covered defenders for a while. It also slows attackers' search for unfound flaws, for at most the window, whose length is unverified. Release with safeguards beats holding only if wider access speeds defenders' fixes by more than the expected cost of safeguards failing. The commercial side doesn't force release on a copy-speed schedule.

The models give conditional results; the evidence doesn't rank the options. The storm may yet have a far side. The question, as the paper puts it, is who builds the curb, and how fast.

Want to learn more?

The technical paper: Allan C. Stam, “Running Through the Rain: Speed, Secrecy and Risk in the AI Cyber Transition.” It has every model, proof and source behind this guide. Appendix A proves the fifteen main results and says, for each, the stronger version it replaces and where that version fails. Appendix B covers open- and closed-weight copies. Appendix A also states Proposition 16, the health-benefits extension, and Appendix C proves it.

The supplement: “Supplement: evidence snapshot and versions read.” It records the version of each source read. It audits the paper's empirical claims one by one, marking each verified, corrected, derived or judgment. It also holds the data note behind the CISA calculations.

Code and data: the replication code and saved data will be deposited at Zenodo [Zenodo DOI to be added on deposit].

Bibliography

Aghion, Philippe, Christopher Harris, Peter Howitt, and John Vickers.
“Competition, Imitation and Growth with Step-by-Step Innovation.” *Review of Economic Studies* 68, no. 3 (2001): 467–92. <https://doi.org/10.1111/1467-937X.00177>.

Aghion, Philippe, and Peter Howitt. “A Model of Growth Through Creative Destruction.” NBER Working Paper 3223, National Bureau of Economic Research, January 1990. <https://doi.org/10.3386/w3223>.

AI Security Institute. “How Fast Is Autonomous AI Cyber Capability Advancing?” Blog post, May 13, 2026. <https://www.aisi.gov.uk/blog/how-fast-is-autonomous-ai-cyber-capability-advancing>.

Anderson, Ross, and Tyler Moore. “The Economics of Information Security.” Authors’ manuscript, University of Cambridge Computer Laboratory, n.d. <https://www.cl.cam.ac.uk/~rja14/Papers/sciecon2.pdf>.

Anthropic. “Anthropic’s Coordinated Vulnerability Disclosure Dashboard.” Counts as of October 2, 2026. <https://red.anthropic.com/2026/cvd/>.

Anthropic. “GLM-5.3 and the Spread of Advanced Cyber Capabilities.” Frontier Red Team research post, September 29, 2026. <https://www.anthropic.com/research/glm-5-3-and-the-spread-of-advanced-cyber-capabilities>.

Anthropic. “Partnering with Mozilla to Improve Firefox’s Security.” Announcement, March 6, 2026. <https://www.anthropic.com/news/mozilla-firefox-security>.

Anthropic. “Project Glasswing: Securing Critical Software for the AI Era.” Web page, announced April 7, 2026, with an update banner dated October 6, 2026. <https://www.anthropic.com/glasswing>.

Anton, James J., and Dennis A. Yao. “Little Patents and Big Secrets: Managing Intellectual Property.” *RAND Journal of Economics* 35, no. 1 (2004): 1–22. <https://doi.org/10.2307/1593727>.

Arbatskaya, Maria, and Hugo M. Mialon. “Multi-Activity Contests.” *Economic Theory* 43, no. 1 (2010): 23–43. <https://doi.org/10.1007/s00199-008-0424-y>.

Arbaugh, William A., William L. Fithen, and John McHugh. “Windows of Vulnerability: A Case Study Analysis.” *Computer* 33, no. 12 (2000): 52–59. <https://doi.org/10.1109/2.889093>.

ARC Prize. “ARC Prize – Leaderboard.” Web page, with its ARC-AGI-2 data file for the semi-private set, accessed October 23, 2026. <https://arcprize.org/leaderboard>, <https://arcprize.org/media/data/leaderboard/v2.json>.

Armstrong, Stuart, Nick Bostrom, and Carl Shulman. *Racing to the Precipice: A Model of Artificial Intelligence Development*. Technical Report 2013-1. Future of Humanity Institute, October 2013.

Askill, Amanda, Miles Brundage, and Gillian Hadfield. “The Role of Cooperation in Responsible AI Development.” Preprint, arXiv, version 1, July 10, 2019. <https://doi.org/10.48550/arXiv.1907.04534>.

Axelrod, Robert, and Rumen Iliev. “Timing of Cyber Conflict.” *Proceedings of the National Academy of Sciences* 111, no. 4 (2014): 1298–1303. <https://doi.org/10.1073/pnas.1322638111>.

Bell, David E. “60.21 Walk or Run in the Rain?” *Mathematical Gazette* 60, no. 413 (1976): 206–8. <https://doi.org/10.2307/3617483>.

Brumley, David, Pongsin Poosankam, Dawn Song, and Jiang Zheng. “Automatic Patch-Based Exploit Generation Is Possible: Techniques and Implications.” In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, 143–57. IEEE, 2008. <https://doi.org/10.1109/SP.2008.17>.

Carlini, Nicholas, Newton Cheng, Keane Lucas, et al. “Assessing Claude Mythos Preview’s Cybersecurity Capabilities.” Frontier Red Team research post, Anthropic, April 7, 2026, author list edited April 9, 2026. <https://red.anthropic.com/2026/mythos-preview/>.

Carlini, Nicholas, Keane Lucas, Evyatar Ben Asher, et al. “Evaluating and Mitigating the Growing Risk of LLM-Discovered 0-Days.” Frontier Red Team research post, Anthropic, February 5, 2026, author list edited February 6, 2026. <https://www.anthropic.com/research/zero-days>.

Center for AI Standards and Innovation. “CAISI’s Assessment of Z.ai’s GLM-5.3 Cyber Capabilities.” National Institute of Standards and Technology, September 17, 2026. <https://www.nist.gov/news-events/news/2026/09/caisis-assessment-zais-glm-53-cyber-capabilities>.

Charrier, Casey, James Sadowski, Zander Work, Clement Lecigne, Benoît Sevens, and Fred Plan. “Look What You Made Us Patch: 2025 Zero-Days in Review.” Google

Threat Intelligence Group, *Google Cloud Blog*, March 5, 2026.

<https://cloud.google.com/blog/topics/threat-intelligence/2025-zero-day-review/>.

Charrier, Casey, and Robert Weiner. “How Low Can You Go? An Analysis of 2023 Time-to-Exploit Trends.” Mandiant, *Google Cloud Blog*, October 15, 2024.

<https://cloud.google.com/blog/topics/threat-intelligence/time-to-exploit-trends-2023>.

Choi, Sung J., M. Eric Johnson, and Christoph U. Lehmann. “Data Breach Remediation Efforts and Their Implications for Hospital Quality.” *Health Services Research* 54, no. 5 (2019): 971–80. <https://doi.org/10.1111/1475-6773.13203>.

Coe, Andrew J., and Jane Vaynman. “Why Arms Control Is So Rare.” *American Political Science Review* 114, no. 2 (2020): 342–55.

<https://doi.org/10.1017/S000305541900073X>.

Cognition. “FrontierCode Leaderboard.” Leaderboard, with methodology revisions and changelog, accessed October 23, 2026.

<https://cognition.com/frontiercode>.

Cottier, Ben, Ben Snodin, David Owen, and Tom Adamczewski. “LLM Inference Prices Have Fallen Rapidly but Unequally Across Tasks.” Epoch AI Data Insight, March 12, 2025. <https://epoch.ai/data-insights/llm-inference-price-trends>.

Cottier, Ben, Josh You, Natalia Martemianova, and David Owen. “How Far Behind Are Open Models?” Epoch AI report, November 4, 2024.

<https://epoch.ai/publications/open-models-report>.

Cybersecurity and Infrastructure Security Agency. “CISA Catalog of Known Exploited Vulnerabilities.” Catalog version 2026.10.21, released October 21, 2026.

https://www.cisa.gov/sites/default/files/feeds/known_exploited_vulnerabilities.json.

Dameff, Christian, Jeffrey Tully, Theodore C. Chan, et al. “Ransomware Attack Associated with Disruptions at Adjacent Emergency Departments in the US.” *JAMA Network Open* 6, no. 5 (2023): e2312270.

<https://doi.org/10.1001/jamanetworkopen.2023.12270>.

Debenedetti, Edoardo, Ilia Shumailov, Tianqi Fan, et al. “Defeating Prompt Injections by Design.” Preprint, arXiv, version 2, June 24, 2025. <https://doi.org/10.48550/arXiv.2503.18813>.

Deering, S., and R. Hinden. *Internet Protocol, Version 6 (IPv6) Specification*. RFC 2460. RFC Editor, December 1998. <https://doi.org/10.17487/RFC2460>.

Defense Advanced Research Projects Agency. “AI Cyber Challenge Marks Pivotal Inflection Point for Cyber Defense.” News release, August 8, 2025. <https://www.darpa.mil/news/2025/aixcc-results>.

Demirer, Mert, Andrey Fradkin, Nadav Tadelis, and Sida Peng. “The Emerging Market for Intelligence: Pricing, Supply, and Demand for LLMs.” NBER Working Paper 34608, National Bureau of Economic Research, December 2025. <https://doi.org/10.3386/w34608>.

dots team. “The dots Model Earns an Officially Certified Perfect-Score Gold Medal at IMO 2026.” Web page, accessed October 23, 2026. https://studio-dots-ai.github.io/dots_imo_2026/en.html.

Emery-Xu, Nicholas, Andrew Park, and Robert Trager. “Uncertainty, Information, and Risk in International Technology Races.” *Journal of Conflict Resolution*, published online November 17, 2023. <https://doi.org/10.1177/00220027231214996>.

Epoch AI. “Benchmarking Hub.” Data export (FrontierMath Tiers 1–3 v2, FrontierMath Tier 4 v2, FrontierSWE, FrontierCode), accessed October 23, 2026. https://epoch.ai/data/benchmark_data.zip.

Epoch AI. “FrontierMath Tier 4 (v2).” Benchmark page with “About,” “Methodology” and “Changelog” (v2 released June 12, 2026), accessed October 23, 2026. <https://epoch.ai/benchmarks/frontiermath-tier-4-v2>.

Federal Bureau of Investigation, Cybersecurity and Infrastructure Security Agency, et al. *Scattered Spider*. Joint Cybersecurity Advisory AA23-320A, November 16, 2023, last revised July 29, 2025. <https://www.cisa.gov/news-events/cybersecurity-advisories/aa23-320a>.

Financial Crimes Enforcement Network. *FinCEN Alert on Fraud Schemes Involving Deepfake Media Targeting Financial Institutions*. FIN-2024-Alert004, November 13, 2024. <https://www.fincen.gov/sites/default/files/shared/FinCEN-Alert-DeepFakes-Alert508FINAL.pdf>.

Fonseca, Pedro, Kaiyuan Zhang, Xi Wang, and Arvind Krishnamurthy. “An Empirical Study on the Correctness of Formally Verified Distributed Systems.” In *Proceedings of the Twelfth European Conference on Computer Systems (EuroSys ’17)*, 328–43. Association for Computing Machinery, 2017. <https://doi.org/10.1145/3064176.3064183>.

FrontierSWE. “FrontierSWE V2.” Leaderboard, accessed October 23, 2026. <https://www.frontierswe.com/>.

Google. “IPv6 Adoption Statistics.” Data file, accessed October 23, 2026. <https://www.google.com/intl/en/ipv6/statistics.html>, https://www.google.com/intl/en_ALL/ipv6/statistics/data/adoption.js.

Green, Edward J., and Robert H. Porter. “Noncooperative Collusion Under Imperfect Price Information.” *Econometrica* 52, no. 1 (1984): 87–100. <https://doi.org/10.2307/1911462>.

Grunewald, Robin, Supriya Mazumdar, and Kelli Vanderlee. “Vulnerability Discovery and Exploitation Trends in the AI Era.” Google Threat Intelligence Group, *Google Cloud Blog*, September 30, 2026. <https://cloud.google.com/blog/topics/threat-intelligence/vulnerability-discovery-and-exploitation-trends-in-the-ai-era>.

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, et al. “Training Compute-Optimal Large Language Models.” Preprint, arXiv, version 1, March 29, 2022. <https://doi.org/10.48550/arXiv.2203.15556>.

Horn, Jann. “Reading Privileged Memory with a Side-Channel.” *Google Project Zero*, January 3, 2018. <https://googleprojectzero.blogspot.com/2018/01/reading-privileged-memory-with-side.html>.

Hua, Shin-Shin, and Haydn Belfield. “Effective Enforceability of EU Competition Law Under AI Development Scenarios: A Framework for Anticipatory

Governance.” In *Proceedings of the 2023 AAI/ACM Conference on AI, Ethics, and Society (AIES '23)*, 596–605. Association for Computing Machinery, 2023. <https://doi.org/10.1145/3600211.3604694>.

Jumper, John, Richard Evans, Alexander Pritzel, et al. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596, no. 7873 (2021): 583–89. <https://doi.org/10.1038/s41586-021-03819-2>.

Kaplan, Jared, Sam McCandlish, Tom Henighan, et al. “Scaling Laws for Neural Language Models.” Preprint, arXiv, version 1, January 23, 2020. <https://doi.org/10.48550/arXiv.2001.08361>.

Kenna Security and Cyentia Institute. *Prioritization to Prediction, Volume 2: Getting Real About Remediation*. Kenna Security, 2019. https://www.cyentia.com/wp-content/uploads/2026/05/Kenna-Prioritization_to_Prediction-v2.pdf.

Khatry, Anirudh, Robert Zhang, Jia Pan, et al. “CRUST-Bench: A Comprehensive Benchmark for C-to-safe-Rust Transpilation.” Preprint, arXiv, version 3, October 1, 2025. <https://doi.org/10.48550/arXiv.2504.15254>.

Klemperer, Paul. “Competition When Consumers Have Switching Costs: An Overview with Applications to Industrial Organization, Macroeconomics, and International Trade.” *Review of Economic Studies* 62, no. 4 (1995): 515–39. <https://doi.org/10.2307/2298075>.

Korinek, Anton, and Jai Vipra. “Concentrating Intelligence: Scaling and Market Structure in Artificial Intelligence.” NBER Working Paper 33139, National Bureau of Economic Research, November 2024. <https://doi.org/10.3386/w33139>.

Lermen, Simon, Charlie Rogers-Smith, and Jeffrey Ladish. “LoRA Fine-Tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B.” Preprint, arXiv, version 2, May 22, 2024. <https://doi.org/10.48550/arXiv.2310.20624>.

Levin, Richard C., Alvin K. Klevorick, Richard R. Nelson, and Sidney G. Winter. “Appropriating the Returns from Industrial Research and Development.” *Brookings Papers on Economic Activity*, no. 3 (1987): 783–831. <https://doi.org/10.2307/2534454>.

Li, Frank, and Vern Paxson. "A Large-Scale Empirical Study of Security Patches." In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17)*, 2201–15. Association for Computing Machinery, 2017. <https://doi.org/10.1145/3133956.3134072>.

Linton, Matt, and Pat Parseghian. "Today's CPU Vulnerability: What You Need to Know." *Google Online Security Blog*, January 3, 2018. <https://security.googleblog.com/2018/01/todays-cpu-vulnerability-what-you-need.html>.

METR. "Task-Completion Time Horizons of Frontier AI Models." Web page, marked "no longer actively updated" on September 8, 2026, accessed October 23, 2026. <https://metr.org/time-horizons/>.

Nappa, Antonio, Richard Johnson, Leyla Bilge, Juan Caballero, and Tudor Dumitraş. "The Attack of the Clones: A Study of the Impact of Shared Code on Vulnerability Patching." In *2015 IEEE Symposium on Security and Privacy*, 692–708. IEEE, 2015. <https://doi.org/10.1109/SP.2015.48>.

Nasr, Milad, Nicholas Carlini, Chawin Sitawarin, et al. "The Attacker Moves Second: Stronger Adaptive Attacks Bypass Defenses Against LLM Jailbreaks and Prompt Injections." Preprint, arXiv, version 1, October 10, 2025. <https://doi.org/10.48550/arXiv.2510.09023>.

Neprash, Hannah, Claire McGlave, and Sayeh Nikpay. "Hacked to Pieces? The Effects of Ransomware Attacks on Hospitals and Patients." *American Economic Journal: Economic Policy* 18, no. 1 (2026): 256–81. <https://doi.org/10.1257/pol.20240594>.

Nevo, Sella, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott. *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*. RR-A2849-1. RAND Corporation, 2024. <https://doi.org/10.7249/RR-A2849-1>.

Nobel Prize Outreach. "Press Release: The Nobel Prize in Chemistry 2024." October 9, 2024. <https://www.nobelprize.org/prizes/chemistry/2024/press-release/>.

Owen, David. "How Predictable Is Language Model Benchmark Performance?" Preprint, arXiv, version 1, January 9, 2024.

<https://doi.org/10.48550/arXiv.2401.04757>.

Palmer, Chris, and Ryan Sleevi. “Gradually Sunsetting SHA-1.” *Google Online Security Blog*, September 5, 2014.

<https://security.googleblog.com/2014/09/gradually-sunsetting-sha-1.html>.

PutnamBench. “PutnamBench Leaderboard.” Leaderboard and data file results.json, Trishul Lab, accessed October 23, 2026.

<https://trishullab.github.io/PutnamBench/leaderboard.html>,

<https://trishullab.github.io/PutnamBench/results.json>.

Ren, Feng, Alex Aliper, Jian Chen, et al. “A Small-Molecule TNIK Inhibitor Targets Fibrosis in Preclinical and Clinical Models.” *Nature Biotechnology* 43, no. 1 (2025): 63–75. <https://doi.org/10.1038/s41587-024-02143-0>.

Rice, H. G. “Classes of Recursively Enumerable Sets and Their Decision Problems.” *Transactions of the American Mathematical Society* 74, no. 2 (1953): 358–66. <https://doi.org/10.1090/S0002-9947-1953-0053041-6>.

Royston, Patrick, and Mahesh K. B. Parmar. “Restricted Mean Survival Time: An Alternative to the Hazard Ratio for the Design and Analysis of Randomized Trials with a Time-to-Event Outcome.” *BMC Medical Research Methodology* 13 (2013): article 152. <https://doi.org/10.1186/1471-2288-13-152>.

Sandbrink, Jonas B., Hamish Hobbs, Jacob L. Swett, Allan Dafoe, and Anders Sandberg. “Differential Technology Development: A Responsible Innovation Principle for Navigating Technology Risks.” Working paper, SSRN, 2022. <https://doi.org/10.2139/ssrn.4213670>.

Schneier, Bruce. “Artificial Intelligence and the Attack/Defense Balance.” *IEEE Security & Privacy* 16, no. 2 (2018): 96. <https://doi.org/10.1109/MSP.2018.1870857>. Reproduced at Schneier on Security, https://www.schneier.com/essays/archives/2018/03/artificial_intellige.html.

seL4 Project. “Frequently Asked Questions.” Web page, n.d., accessed October 23, 2026. <https://sel4.systems/About/FAQ.html>.

Sertkaya, Aylin, and Clara Berger. *Drug Development: Final*. Eastern Research Group for the US Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation, September 26, 2024. <https://aspe.hhs.gov/sites/default/files/documents/ba3c3c99877862bac0b1d04460ad01b1/drug-development.pdf>.

Shevlane, Toby. “Structured Access to AI Capabilities: An Emerging Paradigm for Safe AI Deployment.” Preprint, arXiv, version 1, January 13, 2022. <https://doi.org/10.48550/arXiv.2201.05159>.

Stam, Allan C. “Running Through the Rain: Speed, Secrecy and Risk in the AI Cyber Transition.” Working paper, University of Virginia, October 2026.

Stix, Charlotte, Matteo Pistillo, Girish Sastry, et al. “AI Behind Closed Doors: A Primer on the Governance of Internal Deployment.” Preprint, arXiv, version 1, April 16, 2025. <https://doi.org/10.48550/arXiv.2504.12170>.

Teece, David J. “Profiting from Technological Innovation: Implications for Integration, Collaboration, Licensing and Public Policy.” *Research Policy* 15, no. 6 (1986): 285–305. [https://doi.org/10.1016/0048-7333\(86\)90027-2](https://doi.org/10.1016/0048-7333(86)90027-2).

Thompson, Chris, Mustafa Emre Acer, Serena Chen, Joe DeBlasio, Emily Stark, and David Adrian. “HTTPS by Default.” *Google Online Security Blog*, October 28, 2025. <https://security.googleblog.com/2025/10/https-by-default.html>.

Trammell, Philip, and Leopold Aschenbrenner. “Existential Risk and Growth.” GPI Working Paper 13-2024, Global Priorities Institute, December 7, 2024. <https://globalprioritiesinstitute.org/existential-risk-and-growth-aschenbrenner-and-trammell>.

Trump, Donald J. “Promoting Advanced Artificial Intelligence Innovation and Security.” Executive Order 14409 of June 2, 2026. *Federal Register* 91, no. 108 (June 5, 2026): 34565–67. FR Doc. 2026-11415. <https://www.federalregister.gov/documents/2026/06/05/2026-11415/promoting-advanced-artificial-intelligence-innovation-and-security>.

Tullock, Gordon. “Efficient Rent Seeking.” In *Toward a Theory of the Rent-Seeking Society*, edited by James M. Buchanan, Robert D. Tollison, and Gordon Tullock, 97–

112. Texas A&M University Press, 1980. Reprinted in *Efficient Rent-Seeking: Chronicle of an Intellectual Quagmire*, edited by Alan A. Lockard and Gordon Tullock, 3–16. Springer, 2001. https://doi.org/10.1007/978-1-4757-5055-3_2.

US Government Accountability Office. *Information Technology: Agencies Need to Plan for Modernizing Critical Decades-Old Legacy Systems*. GAO-25-107795. July 2025. <https://www.gao.gov/assets/gao-25-107795.pdf>.

Whalley, Andrew. “SHA-1 Certificates in Chrome.” *Google Online Security Blog*, November 16, 2016. <https://security.googleblog.com/2016/11/sha-1-certificates-in-chrome.html>.

The White House. *Winning the Race: America’s AI Action Plan*. July 2025. <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf>.

World Health Organization. *WHO Methods and Data Sources for Global Burden of Disease Estimates 2000-2021*. Global Health Estimates Technical Paper WHO/DDI/DNA/GHE/2024.3. May 2024. https://cdn.who.int/media/docs/default-source/gho-documents/global-health-estimates/ghe2021_daly_methods.pdf.

Xiao, Winnie, Tim Abbott, Nicholas Carlini, et al. “Measuring LLMs’ Impact on N-Day Exploits.” Frontier Red Team research post, Anthropic, June 8, 2026. <https://www.anthropic.com/research/n-days>.

Xu, Zuojun, Feng Ren, Ping Wang, et al. “A Generative AI-Discovered TNIK Inhibitor for Idiopathic Pulmonary Fibrosis: A Randomized Phase 2a Trial.” *Nature Medicine* 31, no. 8 (2025): 2602–10. <https://doi.org/10.1038/s41591-025-03743-2>.

Yang, Aidan Z. H., Yoshiki Takashima, Brandon Paulsen, Josiah Dodds, and Daniel Kroening. “VERT: Verified Equivalent Rust Transpilation with Large Language Models as Few-Shot Learners.” Preprint, arXiv, version 2, May 25, 2024. <https://doi.org/10.48550/arXiv.2404.18852>.

Yang, Xuejun, Yang Chen, Eric Eide, and John Regehr. “Finding and Understanding Bugs in C Compilers.” In *Proceedings of the 32nd ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI ’11)*, 283–94.

Association for Computing Machinery, 2011.

<https://doi.org/10.1145/1993498.1993532>.

Zaegel, Matteo, Mathis Vehils-Vinals, Hugo Guastalla, Benjamin Benabou, and Auguste Gires. “Should You Walk, Run or Sprint in the Rain to Get Less Wet?” *European Journal of Physics* 45, no. 2 (2024): 025802. <https://doi.org/10.1088/1361-6404/ad06bf>.

SOURCES AND NOTES

1. Allan C. Stam, “Running Through the Rain: Speed, Secrecy and Risk in the AI Cyber Transition” (working paper, University of Virginia, October 2026), section 1, and the sources it cites there: David E. Bell, “60.21 Walk or Run in the Rain?,” *Mathematical Gazette* 60, no. 413 (1976): 207–8, <https://doi.org/10.2307/3617483>; Matteo Zaegel et al., “Should You Walk, Run or Sprint in the Rain to Get Less Wet?,” *European Journal of Physics* 45, no. 2 (2024): 025802, 1, <https://doi.org/10.1088/1361-6404/ad06bf>. ↩
2. Stam, “Running Through the Rain,” section 6.1, under “Disclosed and known-patched counts”; supplement S1; Anthropic, “Anthropic’s Coordinated Vulnerability Disclosure Dashboard,” counts as of October 2, 2026, overview paragraph beginning “As of October 2, 2026,” <https://red.anthropic.com/2026/cvd/>, read in the Internet Archive capture of October 7, 2026, 12:53:16 UTC (<https://web.archive.org/web/20261007125316/https://red.anthropic.com/2026/cvd/>), byte-identical to the live page of October 22, 2026. As supplement S1 notes, these are the lab’s dated counts, not a cohort patch rate or a measure of deployment. ↩
3. Stam, “Running Through the Rain,” section 6.2; supplement S1; US Government Accountability Office, *Information Technology: Agencies Need to Plan for Modernizing Critical Decades-Old Legacy Systems*, GAO-25-107795 (July 2025), Highlights and table 2, <https://www.gao.gov/assets/gao-25-107795.pdf>. As supplement S1 notes, the statuses are agency-reported. ↩
4. Stam, “Running Through the Rain,” section 6.2; Google, “IPv6 Adoption Statistics,” data file, accessed October 23, 2026, <https://www.google.com/intl/en/ipv6/statistics.html>, https://www.google.com/intl/en_ALL/ipv6/statistics/data/adoption.js; S. Deering and R. Hinden, *Internet Protocol, Version 6 (IPv6) Specification*, RFC 2460 (RFC Editor, December 1998), <https://doi.org/10.17487/RFC2460>. The paper dates the lag from RFC 2460; an earlier IPv6 specification would lengthen it. ↩
5. Stam, “Running Through the Rain,” section 1. ↩
6. Stam, “Running Through the Rain,” section 1, the five questions; section 4.6 and appendix A, proposition 4 (assumes far risk rises with speed). The reverse case is listed in Stam, “Running

Through the Rain,” appendix A, section A.7.↩

7. Stam, “Running Through the Rain,” section 4.1, “Four models, four objectives”; appendix A, section A.2.1. Stam, “Running Through the Rain,” section 11, under “Limits of the models”: “we have made no claim that combines them into one welfare measure.” Stam, “Running Through the Rain,” section 10.2: “The results do not establish an ordering among feasible interventions.”↩
8. Stam, “Running Through the Rain,” section 1, under “Plan”; section 4.1, table and the paragraph after it; appendix A, section A.2.1.↩
9. Stam, “Running Through the Rain,” section 2, opening paragraph; supplement S2. Supplement S2 classes the claim as JUDGMENT.↩
10. Stam, “Running Through the Rain,” section 2; Bruce Schneier, “Artificial Intelligence and the Attack/Defense Balance,” *IEEE Security & Privacy* 16, no. 2 (2018): 96, <https://doi.org/10.1109/MSP.2018.1870857>, reproduced at Schneier on Security, https://www.schneier.com/essays/archives/2018/03/artificial_intelligence.html. The paper read the author’s reprint on Schneier on Security; the print text was not compared.↩
11. Stam, “Running Through the Rain,” section 2; supplement S2; Nicholas Carlini et al., “Assessing Claude Mythos Preview’s Cybersecurity Capabilities,” Frontier Red Team research post, Anthropic, April 7, 2026, author list edited April 9, 2026, under “Conclusion,” <https://red.anthropic.com/2026/mythos-preview/>. In supplement S2 the claim is VERIFIED; “not evidence” is JUDGMENT.↩
12. Stam, “Running Through the Rain,” section 2; supplement S1; Anthropic, “Project Glasswing: Securing Critical Software for the AI Era,” web page, announced April 7, 2026, with an update banner dated October 6, 2026, introduction and under “Identifying vulnerabilities and exploits with Claude Mythos Preview,” <https://www.anthropic.com/glasswing>. A lab’s report on its own model, as supplement S1 notes.↩
13. Stam, “Running Through the Rain,” section 2, under “What the case needs”; section 5, under “General models in the reviewed examples”; section 8.5.↩
14. Stam, “Running Through the Rain,” section 2, under “A differential variant”; appendix A, section A.7; section 11, measurement “Differential uptake, 2027–28.”↩
15. Stam, “Running Through the Rain,” section 5; Xuejun Yang et al., “Finding and Understanding Bugs in C Compilers,” in *Proceedings of the 32nd ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI ’11)* (Association for Computing Machinery, 2011), section 3.1 (CompCert 1.6), <https://doi.org/10.1145/1993498.1993532>; Pedro Fonseca et al., “An Empirical Study on the Correctness of Formally Verified Distributed Systems,” in *Proceedings of the Twelfth European Conference on Computer Systems (EuroSys ’17)* (Association for Computing Machinery, 2017), abstract, section 4.4, <https://doi.org/10.1145/3064176.3064183>.↩

16. Stam, “Running Through the Rain,” sections 3 and 4.3; Philip Trammell and Leopold Aschenbrenner, “Existential Risk and Growth” (GPI Working Paper 13-2024, Global Priorities Institute, December 7, 2024), sections 4.1 and 4.3, <https://globalprioritiesinstitute.org/existential-risk-and-growth-aschenbrenner-and-trammell>; Stam, “Running Through the Rain,” appendix A, proposition 1. ↩
17. Stam, “Running Through the Rain,” section 3; Ross Anderson and Tyler Moore, “The Economics of Information Security” (authors’ manuscript, University of Cambridge Computer Laboratory, n.d.), <https://www.cl.cam.ac.uk/~rja14/Papers/sciecon2.pdf>; David Brumley et al., “Automatic Patch-Based Exploit Generation Is Possible: Techniques and Implications,” in *Proceedings of the 2008 IEEE Symposium on Security and Privacy* (IEEE, 2008), <https://doi.org/10.1109/SP.2008.17>; William A. Arbaugh et al., “Windows of Vulnerability: A Case Study Analysis,” *Computer* 33, no. 12 (2000), <https://doi.org/10.1109/2.889093>; Antonio Nappa et al., “The Attack of the Clones: A Study of the Impact of Shared Code on Vulnerability Patching,” in *2015 IEEE Symposium on Security and Privacy* (IEEE, 2015), <https://doi.org/10.1109/SP.2015.48>; Frank Li and Vern Paxson, “A Large-Scale Empirical Study of Security Patches,” in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS ’17)* (Association for Computing Machinery, 2017), <https://doi.org/10.1145/3133956.3134072>. The paper read Brumley et al. in the unpaginated BitBlaze project PDF, Nappa et al. in the authors’ 17-page copy (numbered 1–17) and Li and Paxson in the authors’ camera-ready PDF, which has no proceedings folios. ↩
18. Stam, “Running Through the Rain,” section 3; Stuart Armstrong et al., *Racing to the Precipice: A Model of Artificial Intelligence Development*, Technical Report 2013-1 (Future of Humanity Institute, October 2013), abstract; Nicholas Emery-Xu et al., “Uncertainty, Information, and Risk in International Technology Races,” *Journal of Conflict Resolution*, published online November 17, 2023, 1, <https://doi.org/10.1177/00220027231214996>. ↩
19. Stam, “Running Through the Rain,” section 3, under “Where the paper sits.” ↩
20. Stam, “Running Through the Rain,” section 4.2. The paper’s term is “the kink.” ↩
21. Stam, “Running Through the Rain,” section 4.2, the loss index; appendix A, section A.2.2. The paper defines the index rather than deriving it. ↩
22. Stam, “Running Through the Rain,” section 1; appendix A, assumption A4. Examined as a hypothesis in Stam, “Running Through the Rain,” section 6.1. ↩
23. Stam, “Running Through the Rain,” section 4.3; appendix A, proposition 1 (“the speed threshold”). ↩
24. Stam, “Running Through the Rain,” section 1, under “Our answer, provisionally”; section 4.3; appendix A, section A.7 (the level-set case, a separately scoped calculation). ↩
25. Stam, “Running Through the Rain,” section 4.3, under “The timing $\times$ hazard grid,” and the paragraph after it; appendix A, section A.7 (sequential and mixed timing). ↩

26. Stam, “Running Through the Rain,” section 4.2; appendix A, proposition 2 and assumption A2.↩
27. Stam, “Running Through the Rain,” section 4.4; appendix A, proposition 5. “We have found no measurement of it”: Stam, “Running Through the Rain,” section 4.4; see also Stam, “Running Through the Rain,” section 6.5.↩
28. Stam, “Running Through the Rain,” section 4.3, final paragraph; appendix A, section A.7 (separate adoption populations).↩
29. Stam, “Running Through the Rain,” section 4.5; appendix A, proposition 3 (i)–(iii). The guaranteed-peak wording follows the “In plain terms” summary of P3 in Stam, “Running Through the Rain,” appendix A, section A.3.1; the starting-stock reading is in Stam, “Running Through the Rain,” section 4.5 and appendix A, section A.7.↩
30. Stam, “Running Through the Rain,” section 4.5: “whether AI capability creates or moves it is not modelled, since Proposition 3 contains no speed or capability variable.” A joint path model is listed as future work in Stam, “Running Through the Rain,” appendix A, section A.7.↩
31. Stam, “Running Through the Rain,” section 4.6; appendix A, proposition 4 (i)–(iv).↩
32. Stam, “Running Through the Rain,” section 4.7, under “The answer, conditionally”; appendix A, proposition 16 (section A.3.5); appendix C.↩
33. Stam, “Running Through the Rain,” section 4.7, opening paragraph; John Jumper et al., “Highly Accurate Protein Structure Prediction with AlphaFold,” *Nature* 596, no. 7873 (2021): abstract, <https://doi.org/10.1038/s41586-021-03819-2>; Nobel Prize Outreach, “Press Release: The Nobel Prize in Chemistry 2024,” October 9, 2024, <https://www.nobelprize.org/prizes/chemistry/2024/press-release/>. That such capabilities bring benefits sooner is the section’s premise: Stam, “Running Through the Rain,” section 4.7.↩
34. Stam, “Running Through the Rain,” section 4.1, table and the paragraph after it; section 4.7, under “The extension”; appendix A, section A.2.1 and assumptions A15–A18; World Health Organization, *WHO Methods and Data Sources for Global Burden of Disease Estimates 2000–2021*, Global Health Estimates Technical Paper WHO/DDI/DNA/GHE/2024.3 (May 2024), 6, https://cdn.who.int/media/docs/default-source/gho-documents/global-health-estimates/ghe2021_daly_methods.pdf.↩
35. Stam, “Running Through the Rain,” section 4.7, under “When speed helps,” its table and the headline test; appendix A, proposition 16(i); appendix C, proof of P16(i). In the paper the speed threshold is “the kink.”↩
36. Stam, “Running Through the Rain,” section 4.7, the paragraphs on the benefit’s own kink and on the two kinks; appendix A, section A.3.5, under “The timing combinations.” The illustrative case (danger rising with the square of speed, at speed 0.5) is in Stam, “Running Through the Rain,” section 4.7 and appendix A, section A.3.5.↩

37. Stam, “Running Through the Rain,” section 4.7, under “Where the time goes” and “The extension”; Aylin Sertkaya and Clara Berger, *Drug Development: Final* (Eastern Research Group for the US Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation, September 26, 2024), v and 12, <https://aspe.hhs.gov/sites/default/files/documents/ba3c3c99877862bac0b1d04460ad01b1/drug-development.pdf>. The 0.6 and 0.81 case: Stam, “Running Through the Rain,” section 4.7 and appendix A, section A.3.5, under “Two readings that fail.”↩
38. Stam, “Running Through the Rain,” section 4.7, under “Where the time goes”; Feng Ren et al., “A Small-Molecule TNIK Inhibitor Targets Fibrosis in Preclinical and Clinical Models,” *Nature Biotechnology* 43, no. 1 (2025): abstract, <https://doi.org/10.1038/s41587-024-02143-0>; Zuojun Xu et al., “A Generative AI-Discovered TNIK Inhibitor for Idiopathic Pulmonary Fibrosis: A Randomized Phase 2a Trial,” *Nature Medicine* 31, no. 8 (2025): introduction, <https://doi.org/10.1038/s41591-025-03743-2>.↩
39. Stam, “Running Through the Rain,” section 4.7, under “When speed helps”; appendix A, section A.3.5, under “Adoption levels and the marginal return to speed.”↩
40. Stam, “Running Through the Rain,” section 4.7, under “An optimum above the kink is possible, not guaranteed”; appendix A, proposition 16(iii) and section A.3.5, under “An illustrative optimum and a counterexample.” The cyber-loss-only curve in Figure 17 is this guide’s calculation from the same values.↩
41. Stam, “Running Through the Rain,” section 4.7, under “Discounting”; appendix A, proposition 16(ii); appendix C; World Health Organization, *WHO Methods and Data Sources*, 7.↩
42. Stam, “Running Through the Rain,” section 4.7, under “What the extension leaves out”; appendix A, assumption A16 and section A.7.↩
43. Stam, “Running Through the Rain,” section 4.7, under “What the extension leaves out”; appendix A, section A.7; Hannah Neprash et al., “Hacked to Pieces? The Effects of Ransomware Attacks on Hospitals and Patients,” *American Economic Journal: Economic Policy* 18, no. 1 (2026): published abstract, <https://doi.org/10.1257/pol.20240594>; Christian Dameff et al., “Ransomware Attack Associated with Disruptions at Adjacent Emergency Departments in the US,” *JAMA Network Open* 6, no. 5 (2023): e2312270, abstract, <https://doi.org/10.1001/jamanetworkopen.2023.12270>.↩
44. Stam, “Running Through the Rain,” sections 4.6 and 4.7, under “Two peaks”; appendix A, section A.3.5, under “P4 and the objective.”↩
45. Stam, “Running Through the Rain,” section 10.2, paragraph headed “If the objective includes health benefits from the same capability”; appendix A, section A.3.5, under “Adoption levels and the marginal return to speed.”↩
46. Sung J. Choi et al., “Data Breach Remediation Efforts and Their Implications for Hospital Quality,” *Health Services Research* 54, no. 5 (2019): abstract and discussion, <https://doi.org/10.1111/1475-6773.13203>.↩

47. Stam, “Running Through the Rain,” section 11, measurement “Where the wait for a health benefit goes, 2027–28.”↩
48. Stam, “Running Through the Rain,” section 5; supplement S2; Epoch AI, “Benchmarking Hub,” data export (FrontierMath Tiers 1–3 v2, FrontierMath Tier 4 v2, FrontierSWE, FrontierCode), accessed October 23, 2026, https://epoch.ai/data/benchmark_data.zip; Epoch AI, “FrontierMath Tier 4 (v2),” benchmark page with “About,” “Methodology” and “Changelog” (v2 released June 12, 2026), accessed October 23, 2026, under “About,” <https://epoch.ai/benchmarks/frontiermath-tier-4-v2>; ARC Prize, “ARC Prize – Leaderboard,” web page, with its ARC-AGI-2 data file for the semi-private set, accessed October 23, 2026, <https://arcprize.org/leaderboard>, <https://arcprize.org/media/data/leaderboard/v2.json>; dots team, “The dots Model Earns an Officially Certified Perfect-Score Gold Medal at IMO 2026,” web page, accessed October 23, 2026, https://studio-dots-ai.github.io/dots_imo_2026/en.html; PutnamBench, “PutnamBench Leaderboard,” leaderboard and data file results.json, Trishul Lab, accessed October 23, 2026, <https://trishullab.github.io/PutnamBench/leaderboard.html>, <https://trishullab.github.io/PutnamBench/results.json>; FrontierSWE, “FrontierSWE V2,” leaderboard, accessed October 23, 2026, <https://www.frontierswe.com/>; Cognition, “FrontierCode Leaderboard,” leaderboard, with methodology revisions and changelog, accessed October 23, 2026, <https://cognition.com/frontiercode>; METR, “Task-Completion Time Horizons of Frontier AI Models,” web page, marked “no longer actively updated” on September 8, 2026, accessed October 23, 2026, <https://metr.org/time-horizons/>. Supplement S2 classes the PutnamBench and FrontierSWE/FrontierCode claims as CORRECTED, Tier 4’s 41 problems as DERIVED, and the others as VERIFIED.↩
49. Stam, “Running Through the Rain,” section 5; Jared Kaplan et al., “Scaling Laws for Neural Language Models,” preprint, arXiv, version 1, January 23, 2020, abstract, <https://doi.org/10.48550/arXiv.2001.08361>; Jordan Hoffmann et al., “Training Compute-Optimal Large Language Models,” preprint, arXiv, version 1, March 29, 2022, abstract, <https://doi.org/10.48550/arXiv.2203.15556>; David Owen, “How Predictable Is Language Model Benchmark Performance?,” preprint, arXiv, version 1, January 9, 2024, abstract, <https://doi.org/10.48550/arXiv.2401.04757>. The 6 points are aggregate BIG-Bench Hard; the 18 points individual BIG-Bench tasks.↩
50. Stam, “Running Through the Rain,” section 5; H. G. Rice, “Classes of Recursively Enumerable Sets and Their Decision Problems,” *Transactions of the American Mathematical Society* 74, no. 2 (1953): section 6, theorem 6, corollary B and its interpretation, <https://doi.org/10.1090/S0002-9947-1953-0053041-6>.↩
51. Stam, “Running Through the Rain,” section 5; seL4 Project, “Frequently Asked Questions,” web page, n.d., accessed October 23, 2026, <https://sel4.systems/About/FAQ.html>; Yang et al., “Bugs in C Compilers,” section 3.1 (CompCert 1.6); Fonseca et al., “Formally Verified Distributed Systems,” abstract, section 4.4. The paper read Yang et al. in an author preprint and Fonseca et al. in an author-hosted PDF.↩
52. Stam, “Running Through the Rain,” section 6.1; supplement S2; Defense Advanced Research Projects Agency, “AI Cyber Challenge Marks Pivotal Inflection Point for Cyber Defense,” news

release, August 8, 2025, <https://www.darpa.mil/news/2025/aixcc-results>; Nicholas Carlini et al., “Evaluating and Mitigating the Growing Risk of LLM-Discovered 0-Days,” Frontier Red Team research post, Anthropic, February 5, 2026, author list edited February 6, 2026, <https://www.anthropic.com/research/zero-days>; Carlini et al., “Claude Mythos Preview’s Cybersecurity Capabilities,” note 1; Anthropic, “Partnering with Mozilla to Improve Firefox’s Security,” announcement, March 6, 2026, <https://www.anthropic.com/news/mozilla-firefox-security>; Winnie Xiao et al., “Measuring LLMs’ Impact on N-Day Exploits,” Frontier Red Team research post, Anthropic, June 8, 2026, under “N-days on Firefox,” <https://www.anthropic.com/research/n-days>; AI Security Institute, “How Fast Is Autonomous AI Cyber Capability Advancing?,” blog post, May 13, 2026, <https://www.aisi.gov.uk/blog/how-fast-is-autonomous-ai-cyber-capability-advancing>. AISI’s estimate is under a 2.5-million-token cap per task. The DARPA release, as saved on October 23, 2026, carries an undated editor’s note correcting 70 synthetic vulnerabilities to 63. Supplement S2 classes the 500+ and Firefox-timing claims as CORRECTED, “includes repeats” as DERIVED, and the others as VERIFIED; it adds that checking a report establishes what it reports, not that its figures are right.↩

53. Stam, “Running Through the Rain,” section 6.1; Casey Charrier and Robert Weiner, “How Low Can You Go? An Analysis of 2023 Time-to-Exploit Trends,” Mandiant, *Google Cloud Blog*, October 15, 2024, <https://cloud.google.com/blog/topics/threat-intelligence/time-to-exploit-trends-2023>. The outliers were 13 zero-days and two n-days.↩
54. Stam, “Running Through the Rain,” section 6.1; Casey Charrier et al., “Look What You Made Us Patch: 2025 Zero-Days in Review,” Google Threat Intelligence Group, *Google Cloud Blog*, March 5, 2026, executive summary, <https://cloud.google.com/blog/topics/threat-intelligence/2025-zero-day-review/>; Robin Grunewald et al., “Vulnerability Discovery and Exploitation Trends in the AI Era,” Google Threat Intelligence Group, *Google Cloud Blog*, September 30, 2026, under “In-the-Wild Exploitation” and “AI as the Hunter,” <https://cloud.google.com/blog/topics/threat-intelligence/vulnerability-discovery-and-exploitation-trends-in-the-ai-era>; Cybersecurity and Infrastructure Security Agency, “CISA Catalog of Known Exploited Vulnerabilities,” catalog version 2026.10.21, released October 21, 2026, https://www.cisa.gov/sites/default/files/feeds/known_exploited_vulnerabilities.json. The confidence statement is in Stam, “Running Through the Rain,” section 6.1.↩
55. Stam, “Running Through the Rain,” sections 6.1 and 6.5; Anthropic, “Coordinated Vulnerability Disclosure Dashboard,” process explanation, passage beginning “In our triage process,” and the overview paragraph.↩
56. Stam, “Running Through the Rain,” section 6.2; Kenna Security and Cyentia Institute, *Prioritization to Prediction, Volume 2: Getting Real About Remediation* (Kenna Security, 2019), 4, 18–19, https://www.cyentia.com/wp-content/uploads/2026/05/Kenna-Prioritization_to_Prediction-v2.pdf; Nappa et al., “Attack of the Clones,” abstract, sections I and VIII; Li and Paxson, “Empirical Study of Security Patches,” section 5.1.2 and abstract; US Government Accountability Office, *Information Technology*; Chris Thompson et al., “HTTPS by Default,” *Google Online Security Blog*, October 28, 2025, <https://security.googleblog.com/2025/10/https-by-default.html>; Chris Palmer and Ryan Sleevi,

“Gradually Sunsetting SHA-1,” *Google Online Security Blog*, September 5, 2014, <https://security.googleblog.com/2014/09/gradually-sunsetting-sha-1.html>; Andrew Whalley, “SHA-1 Certificates in Chrome,” *Google Online Security Blog*, November 16, 2016, <https://security.googleblog.com/2016/11/sha-1-certificates-in-chrome.html>; Google, “IPv6 Adoption Statistics.” The sorting and the “adoption binds” reading are judgments: Stam, “Running Through the Rain,” section 6.2; supplement S2.↩

57. Stam, “Running Through the Rain,” section 6.2; Anderson and Moore, “Economics of Information Security” (manuscript), 1. The paper calls this “one proposed reason, a judgment.”↩
58. Stam, “Running Through the Rain,” section 6.3; Cybersecurity and Infrastructure Security Agency, “Known Exploited Vulnerabilities,” the version of October 21, 2026. The paper’s own calculation, with class lists in Stam, “Running Through the Rain,” supplement S2, data note, which classes it DERIVED. The checks exclude the catch-all CWE-20, and keep only entries listed in their identifier’s year or the next.↩
59. Stam, “Running Through the Rain,” section 6.4; Federal Bureau of Investigation et al., *Scattered Spider*, Joint Cybersecurity Advisory AA23-320A, November 16, 2023, last revised July 29, 2025, under “Technical Details,” <https://www.cisa.gov/news-events/cybersecurity-advisories/aa23-320a>; Financial Crimes Enforcement Network, *FinCEN Alert on Fraud Schemes Involving Deepfake Media Targeting Financial Institutions*, FIN-2024-Alert004, November 13, 2024, 5, <https://www.fincen.gov/sites/default/files/shared/FinCEN-Alert-DeepFakes-Alert508FINAL.pdf>; Milad Nasr et al., “The Attacker Moves Second: Stronger Adaptive Attacks Bypass Defenses Against LLM Jailbreaks and Prompt Injections,” preprint, arXiv, version 1, October 10, 2025, abstract and table 1, <https://doi.org/10.48550/arXiv.2510.09023>; Edoardo Debenedetti et al., “Defeating Prompt Injections by Design,” preprint, arXiv, version 2, June 24, 2025, abstract, <https://doi.org/10.48550/arXiv.2503.18813>. Not directly comparable (different threat models and task sets).↩
60. Stam, “Running Through the Rain,” section 6.4; supplement S2; Jonas B. Sandbrink et al., “Differential Technology Development: A Responsible Innovation Principle for Navigating Technology Risks” (working paper, SSRN, 2022), 1, <https://doi.org/10.2139/ssrn.4213670>. The defense list is the paper’s judgment (supplement S2).↩
61. Stam, “Running Through the Rain,” section 6.5; Anirudh Khatry et al., “CRUST-Bench: A Comprehensive Benchmark for C-to-safe-Rust Transpilation,” preprint, arXiv, version 3, October 1, 2025, abstract (OpenAI o1, single-shot, test cases), <https://doi.org/10.48550/arXiv.2504.15254>; Aidan Z. H. Yang et al., “VERT: Verified Equivalent Rust Transpilation with Large Language Models as Few-Shot Learners,” preprint, arXiv, version 2, May 25, 2024, abstract (Claude-2 with VERT, bounded model checking), <https://doi.org/10.48550/arXiv.2404.18852>. A later ASE 2025 version of VERT was not read.↩
62. Stam, “Running Through the Rain,” section 7.1; appendix A, proposition 6 (i)–(v). Part (iv) is the overall cap; part (v) the convex-harm exception. The closed form is the restricted-mean identity: Patrick Royston and Mahesh K. B. Parmar, “Restricted Mean Survival Time: An

Alternative to the Hazard Ratio for the Design and Analysis of Randomized Trials with a Time-to-Event Outcome,” *BMC Medical Research Methodology* 13 (2013): article 152, 3, <https://doi.org/10.1186/1471-2288-13-152>.↵

63. Stam, “Running Through the Rain,” section 7.1; Anthropic, “Project Glasswing,” introduction; Center for AI Standards and Innovation, “CAISI’s Assessment of Z.ai’s GLM-5.3 Cyber Capabilities,” National Institute of Standards and Technology, September 17, 2026, <https://www.nist.gov/news-events/news/2026/09/caisis-assessment-zais-glm-53-cyber-capabilities>; Anthropic, “GLM-5.3 and the Spread of Advanced Cyber Capabilities,” Frontier Red Team research post, September 29, 2026, <https://www.anthropic.com/research/glm-5-3-and-the-spread-of-advanced-cyber-capabilities>; Ben Cottier et al., “How Far Behind Are Open Models?,” Epoch AI report, November 4, 2024, <https://epoch.ai/publications/open-models-report>. The thirteen benchmark lag values are Cottier et al.’s.↵
64. Stam, “Running Through the Rain,” section 7.2; appendix A, proposition 7 (a)–(c). The 100-flaw illustration is in Stam, “Running Through the Rain,” appendix A, the text of proposition 7 and the table in section A.6; an alternative composition gives about 413.45, and “the counterexample survives either way.” Roadmap point: Brumley et al., “Automatic Patch-Based Exploit Generation,” abstract.↵
65. Stam, “Running Through the Rain,” sections 7.1–7.2; Jann Horn, “Reading Privileged Memory with a Side-Channel,” *Google Project Zero*, January 3, 2018, <https://googleprojectzero.blogspot.com/2018/01/reading-privileged-memory-with-side.html>; Matt Linton and Pat Parseghian, “Today’s CPU Vulnerability: What You Need to Know,” *Google Online Security Blog*, January 3, 2018, <https://security.googleblog.com/2018/01/todays-cpu-vulnerability-what-you-need.html>.↵
66. Stam, “Running Through the Rain,” section 7.3; appendix A, proposition 8 (ii)–(iv) and section A.3.2.↵
67. Sella Nevo et al., *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*, RR-A2849-1 (RAND Corporation, 2024), vi, <https://doi.org/10.7249/RRA2849-1>; its current applicability is unverified: Stam, “Running Through the Rain,” section 7.3.↵
68. Stam, “Running Through the Rain,” section 7.4; appendix A, proposition 9; Gordon Tullock, “Efficient Rent Seeking,” in *Toward a Theory of the Rent-Seeking Society*, ed. James M. Buchanan, Robert D. Tollison, and Gordon Tullock (Texas A&M University Press, 1980), 97–112; reprinted in *Efficient Rent-Seeking: Chronicle of an Intellectual Quagmire*, ed. Alan A. Lockard and Gordon Tullock (Springer, 2001), 5, https://doi.org/10.1007/978-1-4757-5055-3_2; the page is the 2001 reprint’s, and the shared loss index is the paper’s addition. “Even a modest escalation” is quoted from the earlier claim in Stam, “Running Through the Rain,” section 7.4 and appendix A.↵
69. Stam, “Running Through the Rain,” section 7.4; Anthropic, “GLM-5.3,” under “What does this mean?”↵

70. Stam, “Running Through the Rain,” section 7.4; Toby Shevlane, “Structured Access to AI Capabilities: An Emerging Paradigm for Safe AI Deployment,” preprint, arXiv, version 1, January 13, 2022, 1, <https://doi.org/10.48550/arXiv.2201.05159>.↵
71. Stam, “Running Through the Rain,” sections 7.4 and 8.3; appendix B, proposition E2 and assumptions B-S2 and B-S6. Exposure is weighted, undiscounted exposure-time, not welfare: Stam, “Running Through the Rain,” section 8.3.↵
72. Stam, “Running Through the Rain,” section 8.1, both tables and the three numbered consequences; on use revealing an edge, Robert Axelrod and Rumen Iliev, “Timing of Cyber Conflict,” *Proceedings of the National Academy of Sciences* 111, no. 4 (2014): 1298, <https://doi.org/10.1073/pnas.1322638111>; on shown results telling rivals where to look, James J. Anton and Dennis A. Yao, “Little Patents and Big Secrets: Managing Intellectual Property,” *RAND Journal of Economics* 35, no. 1 (2004): 1–2, <https://doi.org/10.2307/1593727>; on durable assets, David J. Teece, “Profiting from Technological Innovation: Implications for Integration, Collaboration, Licensing and Public Policy,” *Research Policy* 15, no. 6 (1986): 285, [https://doi.org/10.1016/0048-7333\(86\)90027-2](https://doi.org/10.1016/0048-7333(86)90027-2). and Richard C. Levin et al., “Appropriating the Returns from Industrial Research and Development,” *Brookings Papers on Economic Activity*, no. 3 (1987): 784, <https://doi.org/10.2307/2534454>; on switching costs, Paul Klemperer, “Competition When Consumers Have Switching Costs: An Overview with Applications to Industrial Organization, Macroeconomics, and International Trade,” *Review of Economic Studies* 62, no. 4 (1995): 515, <https://doi.org/10.2307/2298075>.↵
73. Stam, “Running Through the Rain,” sections 8.2–8.3; appendix A, proposition 11 (ii), (iii), (v) and section A.7. “Linear” and “power-law” curves are the paper’s named cases; generally, release comes later whenever the lead times revenue’s slope doesn’t fall as the lead grows. The scans are in Stam, “Running Through the Rain,” section 8.3 and appendix A, section A.6.↵
74. Stam, “Running Through the Rain,” section 8.3; Ben Cottier et al., “LLM Inference Prices Have Fallen Rapidly but Unequally Across Tasks,” Epoch AI Data Insight, March 12, 2025, <https://epoch.ai/data-insights/llm-inference-price-trends>; Mert Demirer et al., “The Emerging Market for Intelligence: Pricing, Supply, and Demand for LLMs” (NBER Working Paper 34608, National Bureau of Economic Research, December 2025), 2–3, <https://doi.org/10.3386/w34608>.↵
75. Stam, “Running Through the Rain,” section 8.3, under “Open- versus closed-weight copies”; appendix B, proposition E1 and assumption B-S5.↵
76. Stam, “Running Through the Rain,” section 8.3; Anthropic, “GLM-5.3,” under “GLM-5.3 lacks robust safeguards”; Simon Lermen et al., “LoRA Fine-Tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B,” preprint, arXiv, version 2, May 22, 2024, 1, <https://doi.org/10.48550/arXiv.2310.20624>. The faster-erosion claim is a judgment “without evidence yet.”↵
77. Stam, “Running Through the Rain,” sections 8.3 and 10.3; appendix B, proposition E3 and assumption B-S7. Stam, “Running Through the Rain,” section 8.3: “Nothing in it carries over to Propositions 11–15.”↵

78. Stam, “Running Through the Rain,” section 8.4; appendix A, proposition 13 (i)–(iii); section 4.1 (no voluntary release).↩
79. Stam, “Running Through the Rain,” section 8.4; appendix A, proposition 14; Nevo et al., *Securing AI Model Weights*, vi.↩
80. Stam, “Running Through the Rain,” section 8.4; appendix A, proposition 12; Philippe Aghion and Peter Howitt, “A Model of Growth Through Creative Destruction” (NBER Working Paper 3223, National Bureau of Economic Research, January 1990), section 2, <https://doi.org/10.3386/w3223>.↩
81. Stam, “Running Through the Rain,” section 8.4; appendix A, proposition 15 (a)–(c). The durable share is assumed. Part (c) is derived in the paper; closest work: Maria Arbatskaya and Hugo M. Mialon, “Multi-Activity Contests,” *Economic Theory* 43, no. 1 (2010): 23, <https://doi.org/10.1007/s00199-008-0424-y>. Escape competition: Philippe Aghion et al., “Competition, Imitation and Growth with Step-by-Step Innovation,” *Review of Economic Studies* 68, no. 3 (2001): abstract, <https://doi.org/10.1111/1467-937X.00177>, read in the Harvard DASH accepted manuscript of October 2000.↩
82. Stam, “Running Through the Rain,” section 8.5; Charlotte Stix et al., “AI Behind Closed Doors: A Primer on the Governance of Internal Deployment,” preprint, arXiv, version 1, April 16, 2025, 5, <https://doi.org/10.48550/arXiv.2504.12170>.↩
83. Stam, “Running Through the Rain,” section 9; appendix A, proposition 10 (i)–(iv). False alarms excluded: Edward J. Green and Robert H. Porter, “Noncooperative Collusion Under Imperfect Price Information,” *Econometrica* 52, no. 1 (1984): 88, <https://doi.org/10.2307/1911462>. Arms control: Andrew J. Coe and Jane Vaynman, “Why Arms Control Is So Rare,” *American Political Science Review* 114, no. 2 (2020): 342, <https://doi.org/10.1017/S000305541900073X>.↩
84. Stam, “Running Through the Rain,” section 9; supplement S2; The White House, *Winning the Race: America’s AI Action Plan* (July 2025), 1, <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf>. Sunk capital: the paper’s judgment (supplement S2).↩
85. Stam, “Running Through the Rain,” section 10, under “What the models settle”; appendix A, propositions 7(c), 6, 7(b).↩
86. Stam, “Running Through the Rain,” section 10.1, “Pressures on holding and releasing,” the table; the exact conditions are there and in Stam, “Running Through the Rain,” appendix A, proposition 11(v) and section A.7.↩
87. Stam, “Running Through the Rain,” section 10.2.↩
88. Stam, “Running Through the Rain,” section 10.2, paragraph headed “If the objective includes health benefits from the same capability.”↩
89. Stam, “Running Through the Rain,” section 10.2, under “Government levers”; Donald J. Trump, “Promoting Advanced Artificial Intelligence Innovation and Security,” Executive Order 14409 of

June 2, 2026, *Federal Register* 91, no. 108 (June 5, 2026), section 3(b)(ii), <https://www.federalregister.gov/documents/2026/06/05/2026-11415/promoting-advanced-artificial-intelligence-innovation-and-security>; Nevo et al., *Securing AI Model Weights*, v; Stix et al., “AI Behind Closed Doors,” 3; Anton Korinek and Jai Vipra, “Concentrating Intelligence: Scaling and Market Structure in Artificial Intelligence” (NBER Working Paper 33139, National Bureau of Economic Research, November 2024), 22–23, <https://doi.org/10.3386/w33139>. ↩

90. Stam, “Running Through the Rain,” section 10.3. The quotation is from Amanda Askill et al., “The Role of Cooperation in Responsible AI Development,” preprint, arXiv, version 1, July 10, 2019, 9, <https://doi.org/10.48550/arXiv.1907.04534>; on competition-law tensions, see also Shin-Shin Hua and Haydn Belfield, “Effective Enforceability of EU Competition Law Under AI Development Scenarios: A Framework for Anticipatory Governance,” in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES ’23)* (Association for Computing Machinery, 2023), section 1, <https://doi.org/10.1145/3600211.3604694>. ↩
91. Stam, “Running Through the Rain,” section 11, under “Limits of the models”; sections 4.1, 4.5 and 4.6; appendix A, section A.7; appendix B, section B.4. ↩
92. Stam, “Running Through the Rain,” section 1, under “What we claim and how sure we are”; sections 6.1, 6.2 and 7.1; section 11, under “Limits of the evidence”; supplement S2, introduction. ↩
93. Stam, “Running Through the Rain,” section 11, under “Evidence that would change the provisional judgments.” ↩
94. Stam, “Running Through the Rain,” section 11, under “Dated checks.” Prediction 1: Anthropic, “Coordinated Vulnerability Disclosure Dashboard.” Predictions 2–3: Cybersecurity and Infrastructure Security Agency, “Known Exploited Vulnerabilities”; Stam, “Running Through the Rain,” supplement S2, data note (class lists, yearly shares, batch-listing counts, conditional binomial calculations). ↩
95. Stam, “Running Through the Rain,” section 12; appendix B, proposition E2. ↩