

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

Allan C. Stam, University of Virginia

Abstract

Is there an intermediate risk peak before superintelligence, and should we run through it? The candidate is the cyber transition: AI that breaks code may help attackers first. In the first of four models, a peak occurs when new flaws outpace removal, then fall behind; whether AI creates or moves it is not modelled. Where adoption, running at least partly alongside capability, sets the transition's length, acceleration raises near-peak loss under a speed-set hazard unless adoption time falls with speed faster than hazard rises, and, with adoption time fixed, under a level-set hazard ending above its average; elsewhere it can help. Health benefits that arrive with the same capability, counted in DALYs with cyber harm, give a second reason to run: they can justify acceleration even where adoption sets the transition's length, but only while the benefit still waits on capability rather than on its own trials and approval, and only if the health gain from earlier arrival exceeds the added cyber harm. Second, a secret hold saves a covered operator, in exposure time, at most the shorter of the window and its deployment time, only against attackers without access, and delays search for unfound flaws by at most the window; its benchmark is safeguarded release. Third, a winner-only window raises racing only through the payoff gap it creates. Fourth, at fixed loss hazard, faster copying of the holder's current model hastens interior release; expected copying of the released model delays it for linear or power-law revenue. Where speeds sit, how hazard and adoption respond, and how much of a benefit's wait capability sets are unmeasured; no intervention ranking follows.

1. Introduction

If you have to cross a street in the rain, should you walk or run? Physics gives a clear answer: run, almost always. 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 exception is a tailwind (Bell 1976, 207–208; Zaegel et al. 2024, 1).

This paper asks whether that intuition carries over to a different storm. On the way to artificial superintelligence (ASI), by which we mean AI that outperforms the best humans at nearly all cognitive work, are there intermediate peaks in risk, periods when the danger rises and then falls? If there are, should we try to get through them faster?

The obvious candidate is the cyber transition. AI that can write and break code may help attackers first; later it might help defenders, through verified software and automated patching. Our models build the first half in as a premise, a hazard that rises with the speed of capability (Assumption A4 of Appendix A), and the evidence sections examine it. The defensive case for acceleration says to sprint through that window. A stronger version adds secrecy: race privately to the capability, hold it back, and use it with trusted partners

before the world gets it. In the rain picture, secrecy is holding the storm at the corner while your party crosses.

We take five questions in turn.

1. *Existence.* Is there an intermediate peak, or does risk plateau or climb straight on toward ASI?
2. *Overcoming.* Is the peak overcome by running faster, by lowering it, or by holding the storm back? Does clearing the near peak raise the far one?
3. *Secrecy for security.* Should a firm race to the capability in logic, coding and mathematics, hold it, and share it only with trusted defenders, given that sharing risks leakage?
4. *Secrecy and commerce.* What does the same held capability earn in commercial markets, what does each use leak, and how do those markets shape how long a firm holds and how hard firms race?
5. *The global maximum.* Is ASI the system-wide peak of risk, and do the answers to the second, third and fourth questions depend on it? We answer this question only conditionally: Proposition 4 assumes that far risk rises with speed.

Our answer, provisionally. The rain analogy fails in two places. First, the far curb is built by other people, on their own clock. Safety arrives when operators patch, retire and rewrite the installed base, and they do that on calendar time. Second, running thickens the rain. Our models take offense and defense to ride one capability clock, so faster capability arms attackers as it arms defenders.

From those two failures follows the paper's first result, which is cleanest beyond the kink, the speed past which adoption rather than capability sets the transition's length. Take adoption time as fixed, and let adoption run alongside capability, wholly ($T = \max(K/s, T_a)$) or in part. Then beyond the kink acceleration raises near-peak risk if the hazard is set by speed, and also if it is set by the capability level reached and is higher at the end of the path than on average (Section 4.3; Appendix A, Section A.7). If adoption itself speeds up with capability, acceleration can still help there, but only if adoption time falls with speed faster than the hazard rises (Appendix A, Proposition 5). Health benefits that arrive with the same capability give a second reason to run (Section 4.7). They can justify acceleration past the kink, but only while the benefit waits on capability rather than on its own trials and approval, and only if the benefit brought forward outweighs the added cyber harm. Below the kink the conditions differ: a hazard set by the level reached lets acceleration help, and a hazard set by speed must rise less than proportionally with speed, or, under mixed timing, less than the share of the transition that capability sets. If adoption waits for the capability, there is no kink, and acceleration helps under a level-set hazard, or under a speed-set hazard that rises slowly enough (Section 4.3). Where current speeds sit is unknown, because the capability distance K is unmeasured. For retirement and migration in the populations Section 6.2's series cover, critical federal legacy systems and web-protocol migration, we judge that adoption already binds; the patching series bear on a different share of adoption time and do not inform that judgment. We hold it as a provisional judgment, not an established result. On the first question the answer is narrow:

a peak can exist in the flaw-stock dynamics; whether AI capability creates or moves it is not modelled (Appendix A, Proposition 3).

Secrecy repairs the second failure, for a while, and in the model does nothing about the first. It splits the two clocks until the window closes. It protects covered operators who can deploy fixes in that time, and it delays attackers' own search for flaws nobody has yet found, for at most the window. To protect anyone outside the holder's circle from the flaws the holder has found, it must ship fixes or mitigations, and published fixes can be roadmaps to the flaws.

The commercial side changes the picture of why firms hold and release. A firm holding a frontier capability can earn from it in four places: the market for inference, defense with partners, products in other markets, and its own research. What leaks depends on the use, not on secrecy as such. And the inference market runs on two copy clocks that push release in opposite directions. At an interior optimum, which with linear revenue curves requires the released model to be copied faster than the current one by more than the loss hazard ($\kappa_{\text{post}} > \kappa_{\text{pre}} + \lambda$), and other things fixed, faster copying of the holder's current public model pushes release earlier. Faster expected copying of the model it would release lowers what release is worth, and for linear and power-law revenue curves pushes release later (Appendix A, Proposition 11). The stated copying comparative static holds the loss hazard fixed. If that hazard rises with post-release copying and release is worth more than what the holder keeps after a loss ($J(T^*) > Z$; Sections 8.2–8.3), it adds an opposing push toward earlier release. So the inference market does not, in general, force held capabilities out on a schedule set by copy speed. Appendix B extends the holding model to open- and closed-weight copies, separating the commercial lead a copy erodes from the attacker access it grants.

What we claim and how sure we are. The deductive core is four small models, each with its own objective, whose results Appendix A proves and checks numerically. Their mathematics is standard; their value is the assembly for this question. Every work in the reference list was read in full, or in full for the part cited, as the reference list and, where a version note is needed, Supplement S1 record, and Supplement S2 reports the primary-source checks claim by claim, with the data note for our calculations from CISA's catalogue. For some preprints and web pages the part cited is the abstract or a named section. Every quotation has been checked against the text read, and no claim remains pending verification. Much of the evidence on frontier cyber capability, and on restricted release, comes from one lab's reports on its own model. Results from a government competition and a government institute, independent of the lab, point the same way (DARPA 2025; AISI 2026), but neither measures harm in the field. Where we judge rather than show, we say so, and we say what would change the judgment.

Plan. Section 2 states the defensive case for acceleration, and its secret variant, at their strongest. Section 3 reviews the literature. Section 4 sets out the models and the results on speed, adoption and two peaks, and Section 4.7 adds health benefits that arrive with the same capability. Sections 5 and 6 bring the evidence on capability and on what sets the transition's length. Section 7 takes up secrecy for security, Section 8 the commercial uses of a held capability, and Section 9 why actors race anyway. Section 10 draws the policy

judgments, Section 11 the limits and the checks that would change the answer, and Section 12 concludes.

2. The defensive case for acceleration and its secret-racing variant, at their strongest

We state the case in its strongest form. The statement is ours, a reconstruction: its parts have named proponents, but no source we read states it as a whole, and none speaks of sprinting through a window. Schneier (2018, 96) argues that AI has “the potential to upend the longstanding advantage that attack has over defense on the Internet” and “the capability to tip the scales more toward defense.” Buterin (2023, “Defense-favoring worlds help healthy and democratic governance thrive”) holds that “A defense-favoring world is a better world” and ties the promotion of defense-favoring technologies to differential technology development; his cyber examples are hardened operating systems, trusted hardware and cryptography, not AI that finds vulnerabilities. Adkins, Evron and Schneier (2025, “The AI-assisted evolution of cyberdefense”) sketch a path to a world where “vulnerabilities are automatically patched even before reaching production,” while warning that AI agents threaten to tip the scales “completely in favor of cyber attackers unless new models of AI-assisted cyberdefense arise” (subtitle), and they call their phases “all speculations.”

The speed argument. Hardware already scheduled will scale models until logic, coding and mathematics are, for practical purposes, solved. That capability first favors attackers, who can find and exploit flaws faster than defenders can fix them. Later it favors defenders, through formal verification, the machine proof that code meets its specification. On this view the transition is a bounded window of danger with safety on the far side. Running through it faster shortens the time spent exposed. Slowing down, or capping compute, lengthens the window and leaves the world more exposed for longer. In its boldest form, the first actor to reach fully verified infrastructure becomes immune, so acceleration is the dominant strategy, and caps on compute make the world less safe.

The argument has force as a conditional case. The record in Section 5 shows top scores at or near the maximum on some automatically checked benchmarks, such as 100% on FrontierMath’s private Tier 4 set and 95% on ARC-AGI-2’s semi-private set, alongside unsaturated coding measures; it does not establish that logic, coding and mathematics are practically solved. The premise that the transition favors attackers is built into our models and examined as a hypothesis in Section 6; the laboratory and competition record is consistent with it. The expectation that defense will eventually dominate is shared by the lab that restricted its frontier cyber-capable model in 2026: “In the long run, we expect that defense capabilities will dominate,” though “the transitional period will be fraught” (Anthropic 2026g, “Conclusion”; see also Anthropic 2026b, “What’s next for Project Glasswing”). It remains an expectation rather than evidence. The rain intuition then does the rest: if the storm has a far side, cross it fast.

The secret-racing variant. The stronger version adds private information. A firm races to the logic, coding and mathematics capability, holds it, and shares it only with trusted defenders, who use it to find and fix flaws before attackers have the same tools. The 2026

restricted release gave this a documented case. The holder described its model as “a general-purpose, unreleased frontier model”, said it did not plan to make it generally available, and stated the aim: an “urgent attempt to put these capabilities to work for defensive purposes” (Anthropic 2026a, Introduction; “Identifying vulnerabilities and exploits with Claude Mythos Preview”). It also expected the window to be short: “Given the rate of AI progress, it will not be long before such capabilities proliferate” (Anthropic 2026a, Introduction). Secrecy, on this view, is what lets defenders run ahead of attackers on a shared clock.

The commercial variant. A third version adds commerce, and it changes the account of why firms race. A firm that holds a capability can use it privately to build products that competitors cannot yet make, in voice, video, agents, and drug and materials discovery. Meanwhile the models it has already released are being copied, so its lead in selling intelligence wears away. That erosion pressures the firm to release what it holds, to restore its commercial lead. On this reading two races run at once. The commercial race, driven by the wasting value of each released lead, pushes toward general ASI. The security race, driven by the hope of a defense-dominant position, pushes a secret race for narrow capability in logic, code and mathematics.

What the case needs. At its strongest, then, the defensive case needs four things to be true together: 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 before attackers arrive; and the commercial pressures either reinforce the security motive or leave it alone. The secret and commercial variants also need separability: that the capability in logic, code and mathematics can be raced for, held and used apart from general capability. The examples in Section 5 use general models and do not establish it (Section 8.5). The rest of the paper takes each in turn.

A differential variant. A different case asks not for faster capability overall but for defense to be accelerated faster than offense, in the spirit of differential technology development: defenders’ uptake of AI tools outpacing attackers’. Our models do not cover it. They put offense and defense on one capability clock, so faster capability arms both; separate attacker and defender speeds are listed in Appendix A, Section A.7, with no sign known. Section 11 names the observable that would bear on it.

3. Literature

Speed through a hazardous transition. Jones (2024, 576) asks “under what conditions should we continue the rapid progress of AI, and under what conditions should we stop?” Trammell and Aschenbrenner (2024, abstract) argue that acceleration “decreases the time spent at each technology level,” lowering cumulative risk when risk depends on the level of technology. When risk is instead “transitional,” posed by development itself, they show that with policy fixed the sign turns on whether the hazard’s elasticity in the speed of development is above or below one (§§4.1, 4.3); below one, “sufficiently rapid growth ... approximates an immediate, risk-free jump” (their appendix B.5.2). Proposition 1’s capability branch is, in substance, that result, with the elasticity allowed to vary; we claim nothing new there. What we add is a second clock. With their policy fixed, technology alone

sets how long the risky period lasts; their policy lever, which can mean “bans on the adoption of risky production processes” (§3.2), restricts use of risky technology, not the time operators need to deploy fixes. In our models the risky period ends only when operators have deployed fixes, on calendar time. Past the kink, the speed beyond which adoption rather than capability sets the transition’s length, acceleration only raises the near-peak loss index (Proposition 1), and any escape turns on adoption’s own response to speed (Proposition 5). As an analogy only, research on cross-country technology diffusion, which does not study the deployment of security fixes, finds that where adoption lags matter, they are “largely determined by lack of educational attainment and lack of trade openness” (Comin and Hobijn 2006, abstract). We found no estimate of whether frontier capability shortens adoption itself. Trammell and Aschenbrenner’s main channel, policy that tightens as society grows richer and yields an “existential risk Kuznets curve” (abstract), is outside our models, which hold policy fixed; our intermediate peak (Proposition 3) comes instead from a stock of flaws filling faster than it drains.

Offense and defense. Jervis (1978, 211) built his account of the security dilemma on two variables: “whether the offense or the defense has the advantage, and whether offensive postures can be distinguished from defensive ones.” Garfinkel and Dafoe (2019, 736) show that the balance can change with scale: “growth in investments will favor offense when investment levels are sufficiently low and favor defense when they are sufficiently high.” For cyber operations, Slayton (2017, 74) traces advantage to organizational processes “such as software updating, vulnerability scanning, and access management,” and calls the balance “a dyadic, not a systemic, variable,” so we claim no balance for cyberspace as a whole. Bostrom (2019, 458) speaks of vulnerabilities “opening and closing”; in his example the period ends once defense becomes “reasonably affordable.” Proposition 3 puts the contest on a calendar path, the stock of flaws against their removal and the protection of new code, and gives the condition under which the hazard rises and then falls.

Vulnerability disclosure and patching. Anderson and Moore’s (2006, authors’ manuscript, 1) survey of security economics notes that “security failure is caused at least as often by bad incentives as by bad design.” Arora, Telang and Xu (2008, 642) model a planner who sets the period a vendor has to patch, and find that “the vendor typically releases the patch less expeditiously than is socially optimal.” Brumley et al. (2008, abstract) generate exploits automatically from patches and conclude that this “should be considered practical.” On flaws not yet disclosed, Ablon and Bogart (2017, x) find that “after a year, approximately 5.7 percent have been discovered by an outside entity,” and Ozment and Schechter (2006, abstract) that foundational flaws in OpenBSD “have a median lifetime of at least 2.6 years.” Measured deployment of fixes is slow: in CERT case studies, “patches for the vulnerability were available at least a month before target sites reported intrusions” (Arbaugh, Fithen and McHugh 2000, 57), and studies of client hosts and of open-source repositories find long lags too (Nappa et al. 2015; Li and Paxson 2017; Section 6.2). Shevlane and Dafoe (2020, abstract) warn that while disclosing software flaws “often favours defence, this cannot be assumed for AI research.” This work treats flaws one at a time. Proposition 7 adds flaws in outsiders’ code that the holder never finds: a hold of the whole capability delays attackers’ search for them, for at most the window. That makes release with safeguards, not a flaw-level embargo, the comparator for any hold. Rescorla

(2004, abstract), asking whether finding security holes is worth it, could not exclude that the rate of finding them in a given program “is constant over long periods of time”; if so, many flaws stay unfound, the case Proposition 7 adds. The bound on what a window saves is the restricted mean, “the area under the survival curve” up to a horizon (Royston and Parmar 2013, 3).

AI race models and contests. Armstrong, Bostrom and Shulman (2013, abstract) model teams that “race to build the first AI,” each willing to skimp “on safety precautions if need be,” and find that “the more teams know about each others’ capabilities (and about their own), the more the danger increases.” Emery-Xu, Park and Trager (2023, 1) find that revealing rivals’ capabilities “has two opposing effects on risk.” So whether revealing a lead raises or lowers risk has no settled sign in these models. The contest machinery is standard: Tullock’s (1980, 2001 reprint, 5) lottery and the patent races of Loury (1979), Lee and Wilde (1980, 429) and Dasgupta and Stiglitz (1980). Reinganum’s (1989, 855) survey says of these models that “the addition of strategic rivals required some simplifications,” leading their authors “to adopt the exponential distribution.” Loury (1979, 406) finds that with the number of firms fixed each firm invests more than is socially optimal, and Dasgupta and Stiglitz (1980, 1) that equilibrium “may entail excessively fast research.” Propositions 9 and 10 restate such results with two additions: an expected-loss index that rises with total effort and that every racer bears, as Armstrong, Bostrom and Shulman’s disaster turns on “the safety level of the winning team” (§2), and a secret window as a prize to the winner alone. Restraint in Proposition 10 rests on permanent reversion. In Green and Porter’s (1984, 88) account, “Friedman ... outlines a strategy in which firms respond to suspected cheating ... by producing at Cournot levels thereafter.” In their own model “reversionary episodes will sometimes occur without any firm defecting, simply because of low demand” (90), the false-alarm case we exclude. Coe and Vaynman (2020, 342) show why such schemes are rare in arms: monitoring must make arming “transparent enough to assure its compliance but not so much as to threaten its security.” For Proposition 15(c) we compare the paper only with Arbatskaya and Mialon’s (2010, 23) contest won through “multiple activities”; we did not search that line further.

Private information and perishable advantages. Bargaining theory traces war to “private information about relative capabilities or resolve and incentives to misrepresent such information,” and to commitment problems (Fearon 1995, 381). Debs and Monteiro (2014, 1) show that when investments “remain undetected,” adversaries “may strike preventively even without conclusive evidence of militarization.” Axelrod and Iliev (2014, 1298) model a cyber resource whose “use today may well prevent it from being available for use later.” Arrow (1962, 614) made the general point: “The very use of the information in any productive way is bound to reveal it, at least in part.” Anton and Yao (2004, 1) show that where disclosure aids imitation, “large inventions are protected primarily through secrecy when property rights are weak.” The paper carries this logic to a held model that earns in several markets: what leaks depends on the use, and commercial use can signal that the capability exists. We found no work that compares leak rates across a capability’s uses.

Innovation races with imitation. Proposition 12 is the firm-value term of creative destruction, in which a monopoly is worth “the expected present value of the constant flow

over an interval whose length is exponentially distributed” (Aghion and Howitt 1990, §2). Aghion et al. (2001, abstract) find that “too much imitation is unambiguously growth-reducing”; Proposition 15(b) sits close to their escape-competition motive. Proposition 15’s durable share compresses the appropriation literature: Arrow’s point above, patents as “licenses for a monopoly on information for a specified period of time” (Nordhaus 1969, 70), complementary assets (Teece 1986, 285), lead time and learning curves (studies of aircraft and semiconductors reported by Levin et al. 1987, 784), first-mover advantages (Lieberman and Montgomery 1988, 41) and switching costs (Klemperer 1995, 515). Proposition 11 is optimal stopping, where “stopping corresponds to making the investment, and continuation corresponds to waiting” (Dixit and Pindyck 1994, ch. 4, §1.D). Weeds (2000, abstract) shows how fear of preemption bears on the “option value of delay,” and Moorthy and Png (1992, 345) that “sequential introduction is better than simultaneous introduction when cannibalization is a problem and customers are relatively more impatient than the seller.” We found no release problem with two copy clocks that push in opposite directions: at an interior optimum, other things fixed, erosion of the current public lead hastens release, and expected copying of the model to be released delays it for linear and power-law revenue curves.

Copying, open weights and internal deployment. Distillation trains a small model on a large one’s “soft targets” (Hinton, Vinyals and Dean 2015, 2). Query access can “duplicate the functionality” of a served model (Tramèr et al. 2016, abstract) and has recovered “the embedding projection layer (up to symmetries)” of a transformer model (Carlini et al. 2024, abstract). Yet imitation models “close little to none of the gap” on tasks thin in the imitation data (Gudibande et al. 2023, 1). Open models lagged closed ones on benchmarks “by 5 to 22 months” (Cottier et al. 2024, summary of findings). Seger et al. (2023, 6) stress that open release is “irreversible”; Kapoor et al. (2024, abstract) find research “insufficient to effectively characterize the marginal risk of open foundation models relative to pre-existing technologies.” Cheap fine-tuning can “undo the safety training of Llama 2-Chat models” (Lermen, Rogers-Smith and Ladish 2024, 1), and a RAND assessment found that “protecting models that are interacting with the internet against the most capable threat actors is currently not feasible” (Nevo et al. 2024, vi). Governance of the internal deployment of “highly advanced frontier AI systems appears absent” (Stix et al. 2025, 1). The coalition of Proposition 8 resembles a club, where at some point “congestion will set in” (Buchanan 1965, 7). Appendix B separates the commercial lead a copy erodes from the attacker access it grants (Propositions E1–E3); these are short deductions from Poisson thinning and pathwise ordering.

Where the paper sits. Most of the paper is assembly. Six results need more than a line of argument: Propositions 1, 3, 5, 7 and 11, and the bounds in Proposition 6. All use standard techniques: elasticity calculus, integrating factors, the restricted-mean identity, supporting lines of convex functions, pathwise comparison of dated exposure intervals, and deterministic optimal stopping. Propositions 9, 10 and 15(b) restate standard contest results, and Proposition 12 is a known identity. What may be new is the combination: two copy clocks with opposite effects on release, a capability-level hold that delays attackers’ search for flaws the holder never finds, and leakage that depends on use. In the searches

we made, across twenty bodies of work, each work we found treats at most one of these. That is a statement about our searches, not about the literature.

4. The model: speed, adoption and two peaks

4.1 Four models, four objectives

The questions mix a social problem (is acceleration good for the world?) with private ones (when should a firm release what it holds, and how hard should it race?). We keep them apart with four small models, each with its own objective and units (Figure 1).

Model	Results	Objective	Units
Speed and adoption	Propositions 1–5	A social loss index	Expected harmful events, or losses at a fixed value per event
Speed with health benefits, an extension of the first model	Proposition 16	A health–cyber balance: health benefit less cyber loss	Disability-adjusted life years (DALYs) averted; cyber loss at a fixed number of DALYs per harmful event
Holding the storm	Propositions 6–8	Exposure saved for operators	Exposure-time, with operator weights the analyst chooses
The contest	Propositions 9–10	Each racer’s expected payoff; welfare as declared in each result	Money-equivalent payoff
The holder’s program	Propositions 11–15	The holder’s private expected present value, at discount rate $r > 0$	Money-equivalent present value

Proposition 16 extends the first model, and only that model, with a second objective in one unit, DALYs (Section 4.7). It is a conditional balance between a health benefit and cyber harm, not a measure of total welfare, and it is used for that extension only. Results combine only where their assumptions permit, and we make no combined welfare claim across the four. Three boundaries matter. First, the held flow F counts only what the holder captures: its own security and partners’ payments, artifact and product revenue, and the private value of internal research, counted once: if the next contest’s prize is modelled in the contest model, internal research enters there, as a higher chance of winning, and its value in F is set to zero. Outsiders’ security belongs to the second model, not to F . In that model the operator weights w_j are nonnegative weights the analyst chooses, such as each operator’s criticality, the population it serves, or equal weights; the results hold for any such weights. Second, Propositions 13 and 14 assume that only involuntary loss ends the

hold; adding Proposition 11's release option requires a new derivation. Third, Propositions 6–8 take the window as given. A chosen release that makes the capability public truncates the involuntary-loss time. Proposition 6's distribution-free bound still applies, but its exponential formulas and Proposition 8's coalition objective cannot be carried over unchanged. We use those formulas only for the exogenous exponential window.

Model 1 Speed and adoption Propositions 1–5 Objective Social loss index $R_c(s) = H(s)T(s) + S$ Units Expected harmful events, or losses at a fixed value per event	Model 2 Holding the storm Propositions 6–8 Objective Exposure saved for operators Units Exposure-time, with operator weights w_j the analyst chooses	Model 3 The contest Propositions 9–10 Objective Each racer's expected payoff; welfare as declared in each result Units Money-equivalent payoff	Model 4 The holder's program Propositions 11–15 Objective Holder's private expected present value, $r > 0$ Units Money-equivalent present value
Extension of Model 1 Proposition 16 Health–cyber balance $W_L(s) = B_L(s) - V_c C_c(s)$ Units DALYs averted; V_c in DALYs per harmful event	Extension of Model 2 Appendix B, E1–E3 Open- and closed-weight copies Units Exposure units of Appendix A	No objective is summed across models; no policy ranking follows.	

Figure 1. The four models and their two extensions, each with its own objective and units. Model 1 (Propositions 1–5) measures a social loss index, $R_c(s) = H(s)T(s) + S$, in expected harmful events; Proposition 16 extends Model 1 only, with a health–cyber balance in DALYs averted; Model 2 (Propositions 6–8) measures exposure saved for operators, in exposure-time with operator weights the analyst chooses; Appendix B extends the holding model to open- and closed-weight copies (Propositions E1–E3), in the exposure units of Appendix A; Model 3 (Propositions 9–10) uses each racer's expected payoff, and Model 4 (Propositions 11–15) the holder's private expected present value at $r > 0$, both money-equivalent (Section 4.1). Results combine only where their assumptions permit; no objective is summed across models, no combined welfare claim is made, and no intervention ranking follows.

Appendix A gives the proofs; the numerical suite checks selected formulas, parameter cases and counterexamples. For each result the appendix also states the stronger version it replaces and where that version fails. The text below uses the corrected results under their stated assumptions.

4.2 The loss index

Let capability advance at speed s over a capability distance K , and let deployment need adoption time T_a of calendar time. The transition lasts $T(s) = \max(K/s, T_a)$. The kink s^* is

the speed at which the two are equal. The transition hazard is $H(s)$, rising in speed, with elasticity $\varepsilon_H = sH'/H$. We measure near-peak risk by the loss index

$$R_c(s) = H(s)T(s) + S.$$

The first term is the hazard over the transition's length. The second, S , is the step term: each capability step exposes a stock of latent flaws once. We define R_c as the model's index rather than derive it. It equals cumulative hazard when the hazard is rectangular, and is proportional to it for any hazard of fixed shape. If harmful events arrive at random, it is the expected number of events. We never treat the height of a peak alone as risk, because a spike that narrows as it rises can be higher and smaller in area.

That the step term does not depend on speed is an assumption of the model, not a result (Appendix A, Proposition 2). Writing S as a sum over steps of flaws exposed, the chance each is exploited, and the damage per exploit, it falls with any one factor when the other two are positive. Hardware-bound authentication as a way to lower exploit success, and segmentation or rate limits as ways to lower damage, are proposed intervention channels requiring separate evidence. Proposition 2 establishes only the conditional algebra. If exploit success depends on attackers' tools when a step arrives, the step term can depend on speed.

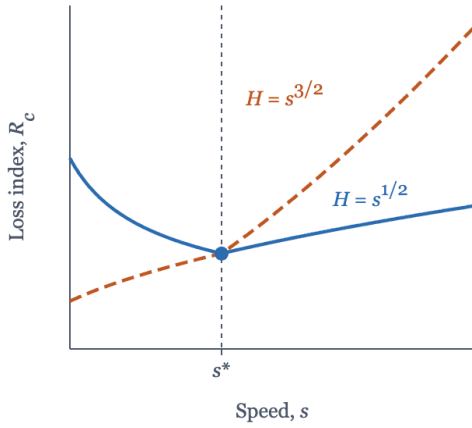
4.3 The speed threshold

Proposition 1 is the paper's base result; its capability branch is, in substance, Trammell and Aschenbrenner's (2024, §4.3) result for transitional risk with policy fixed. Take adoption time as fixed, the step term as independent of speed, the hazard as positive, differentiable and rising, and speeds as bounded below by some $s_{\min}$ under the kink. Then:

- Where capability sets the transition's length ($s < s^*$), acceleration lowers the loss index at a speed if and only if $\varepsilon_H < 1$ there. Speeding up shortens exposure and raises its intensity, and it helps only if intensity rises less than proportionally with speed.
- Where adoption sets the length ($s > s^*$), acceleration only raises it: the derivative is $H'T_a > 0$.
- A minimizing speed exists and lies at or below the kink. If $\varepsilon_H < 1$ everywhere below the kink, it is the kink itself. In general it is the lower bound, the kink, or an interior speed where $\varepsilon_H = 1$ (Appendix A, Proposition 1).

The result is elementary. Its force is in what it rules out (Figure 2). Once operators' deployment, not capability, sets how long the transition lasts, no elasticity condition rescues acceleration under fixed adoption time, with adoption running alongside capability ($T = \max(K/s, T_a)$) and a hazard set by speed. Running faster makes the rain heavier without bringing the far curb closer. Above the kink the sign is the same when the hazard is instead set by the capability level reached and is higher at the end of the path than on average, $h(K) > \bar{h}$, as it is when the hazard rises with capability; the scoped calculation is the second case below.

(a) Speed-set hazard (Proposition 1)



(b) Level-set hazard (§4.3, A.7)

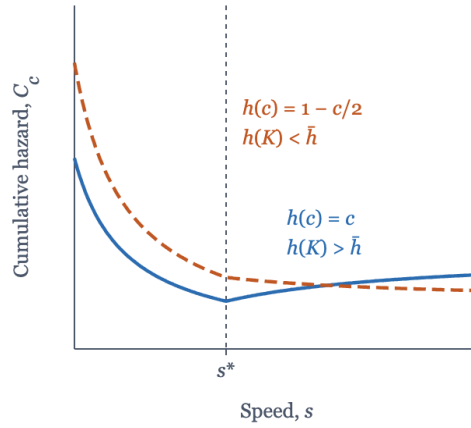


Figure 2. Near-peak loss against speed, with adoption time fixed and parallel timing. (a) The loss index $R_c(s) = H(s)T(s) + S$ for the speed-set hazards $H = s^{1/2}$ ($\varepsilon_H = 1/2$) and $H = s^{3/2}$ ($\varepsilon_H = 3/2$): below the kink s^* the index falls with speed if and only if $\varepsilon_H < 1$, so the first curve has its minimum at the kink; above the kink both rise, with derivative $H'T_a > 0$ (Proposition 1). (b) The separately scoped level-set calculation of Section 4.3 and Appendix A, Section A.7: path $c(t) = \min(st, K)$, hazard $h(c(t))$ held at $h(K)$ through the adoption tail, so that cumulative hazard is $(K/s)\bar{h} + (T - K/s)h(K) + S$; it falls with speed below the kink and, above it, rises if $h(K) > \bar{h}$ (here $h(c) = c$) and falls if $h(K) < \bar{h}$ (here $h(c) = 1 - c/2$). Panel (b) is not obtained by substituting $H = h(K)$ into panel (a). Illustrative values: $K = T_a = 1$, $S = 0$; the hazard forms are illustrative, not estimates.

Two other timings. If adoption instead starts only once the capability arrives, so that $T = K/s + T_a$, then with T_a fixed and the step term independent of speed, a marginal acceleration lowers the loss index if and only if $\varepsilon_H < (K/s)/(K/s + T_a)$. There is no kink under sequential timing (Figure 3). If the hazard is instead set by the capability level reached, a separately scoped calculation applies, not a consequence of the loss index of Section 4.2: along the path $c(t) = \min(st, K)$, with hazard $h(c(t))$ at capability level $c(t)$, fixed T_a , a speed-independent step term and the hazard held at $h(K)$ through the adoption tail, the derivative of cumulative hazard in speed above the kink has the sign of $h(K) - \bar{h}$. Here $\bar{h}$ is the average hazard over capability levels 0 to K ; Appendix A, Section A.7, lists both cases.

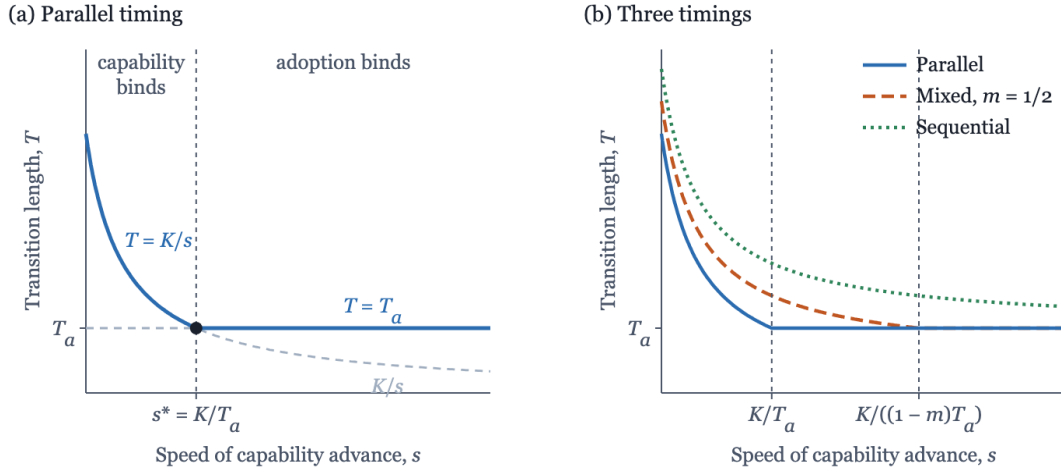


Figure 3. Transition length against speed under the three timings of adoption. (a) Parallel timing, $T(s) = \max(K/s, T_a)$: capability binds below the kink $s^* = K/T_a$ and adoption above it (Section 4.2; Proposition 1). (b) Parallel timing; mixed timing, $T = \max(K/s, (1-m)T_a) + mT_a$, with its kink at $K/((1-m)T_a)$; and sequential timing, $T = K/s + T_a$, with no kink (Section 4.3; Appendix A, Section A.7). Adoption time T_a and the share m are fixed in speed. Illustrative values: $K = T_a = 1$, $m = 1/2$.

The timing × hazard grid. The table collects the sign of the loss derivative in speed for three timings and two hazards, with adoption time, the timing share m and the step term fixed. Under mixed timing a share m of adoption waits for the capability and the rest runs alongside it, so the kink is at $K/((1-m)T_a)$ (Appendix A, Section A.7). A negative entry means that acceleration lowers the loss.

Timing	Hazard set by speed: sign below the kink / above it	Hazard set by capability level: derivative below the kink / above it
Parallel, $T = \max(K/s, T_a)$	sign of $\varepsilon_H - 1$ / positive	$-K\bar{h}/s^2$ / $K[h(K) - \bar{h}]/s^2$
Mixed, $0 < m < 1$	sign of $\varepsilon_H - (K/s)/(K/s + mT_a)$ / positive	$-K\bar{h}/s^2$ / $K[h(K) - \bar{h}]/s^2$
Sequential, $T = K/s + T_a$	sign of $\varepsilon_H - (K/s)/(K/s + T_a)$; no kink	$-K\bar{h}/s^2$; no kink

The paper’s headline answer to the speed question lies in the entries above the kink: with adoption time fixed, acceleration raises the loss under parallel or mixed timing whenever the hazard is set by speed, and whenever it is set by capability level with $h(K) > \bar{h}$. Below the kink the conditions differ. With a speed-set hazard, acceleration lowers the loss if the hazard’s elasticity is below one under parallel timing, or below the share of the transition that capability sets, $(K/s)/(K/s + mT_a)$, under mixed timing; with a level-set hazard and $\bar{h} > 0$ it always does. Under sequential timing there is no kink, and the level-set derivative is always negative. Which hazard applies is open (Figure 4). In our judgment, the speed-set form is a reduced form for a gap between attackers’ capability, which advances with speed,

and defense that runs on calendar time, a gap that widens with speed at every capability level; the level-set form drops that gap.

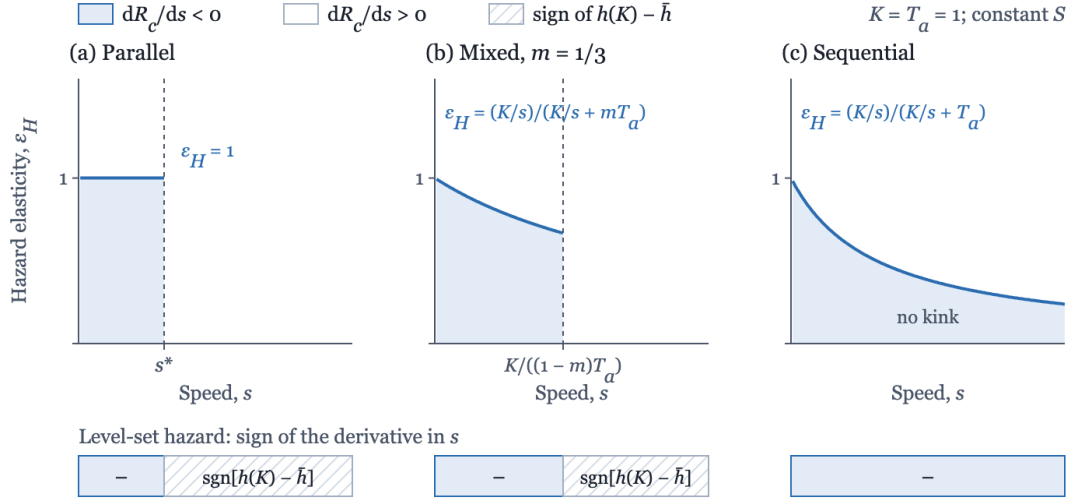


Figure 4. Where a marginal acceleration lowers the loss, by timing and hazard, with adoption time, the timing share m and the step term fixed (Section 4.3). Panels, speed-set hazard: shaded points (s, ε_H) have $dR_c/ds < 0$; the boundary is $\varepsilon_H = 1$ below the kink under parallel timing (a), $\varepsilon_H = (K/s)/(K/s + mT_a)$ below the kink $K/((1-m)T_a)$ under mixed timing (b), and $\varepsilon_H = (K/s)/(K/s + T_a)$ at every speed under sequential timing, which has no kink (c); above a kink the derivative is positive whatever ε_H . Strips, level-set hazard (the scoped calculation of Section 4.3): the derivative is $-K\bar{h}/s^2$, negative when $\bar{h} > 0$, below the kink and under sequential timing, and $K[h(K) - \bar{h}]/s^2$ above the kink; hatched strip cells take the sign of $h(K) - \bar{h}$, which is not determined, and the shading key does not apply to them. Illustrative values: $K = T_a = 1$, $m = 1/3$; ε_H is read pointwise at each speed. Which hazard applies is open.

With separate adoption populations the regime is a distribution, not one point: with weights w_j , a common hazard, constant adoption times $T_{a,j}$ and a speed away from every population's kink, acceleration lowers the weighted loss if and only if $(1 - \varepsilon_H) \sum_{\text{cap}} w_j K/s > \varepsilon_H \sum_{\text{adopt}} w_j T_{a,j}$, the sums running over the populations where capability binds and where adoption binds (Appendix A, Section A.7).

4.4 The accelerationist's escape

The defender of acceleration has an escape within the model. Perhaps faster capability also speeds adoption, through automated rollout, verified translation of legacy code, or automated rewrites. Proposition 5 gives the exact terms. Let adoption time fall with speed as $T_a = A(s/s_0)^{-\beta}$, with A the adoption time at a reference speed s_0 and $\beta \geq 0$ measuring how fast adoption time falls with speed (its elasticity in speed is $-\beta$).

- Wherever adoption binds, acceleration lowers the loss index if and only if $\beta > \varepsilon_H$: adoption time must fall with speed faster, in proportional terms, than the hazard rises. Where capability binds, β does not enter. The test is exact when the step term

does not depend on speed (Assumption A2); if it rises with speed, $S'(s) > 0$, the escape needs more: $\beta > \varepsilon_H + sS'(s)/(H \cdot T_a)$. If $\beta > \varepsilon_H$ throughout an unbounded adoption branch, loss decreases throughout that branch and no minimum is attained there; this alone does not exclude a global minimizer elsewhere.

- The escape is not that the kink moves outward. For $\beta < 1$, the kink rises with β if and only if capability binds at the reference speed. If adoption binds at current speeds and A is measured there, the kink falls as β rises, and the escape can occur while it does. For $\beta > 1$ the order of the two branches reverses (Appendix A, Proposition 5).

Proposition 1 is the case $\beta = 0$. Within the model, where adoption binds, the crux of the speed question is therefore one comparison of two elasticities. It is measurable in principle (Figure 5). As Section 6 reports, we have found no measurement of it.

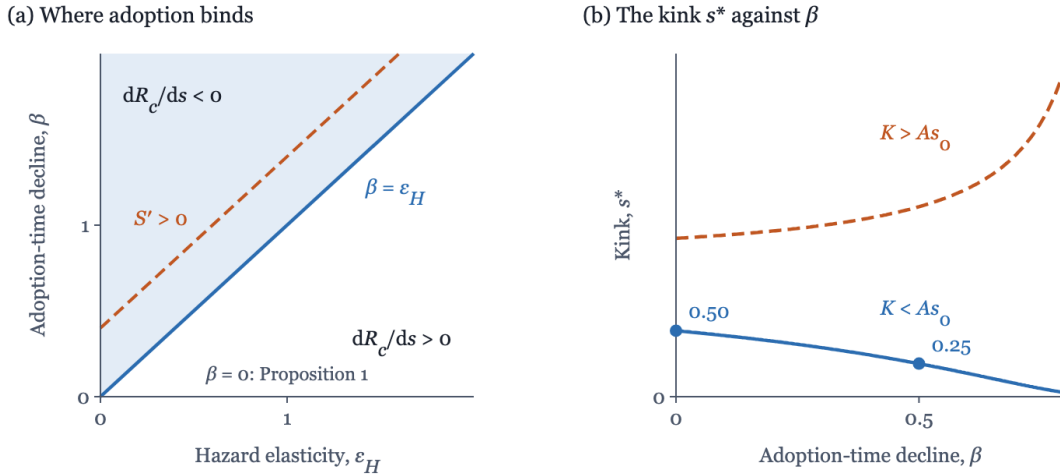


Figure 5. The accelerationist's escape (Proposition 5), with adoption time $T_a = A(s/s_0)^{-\beta}$. (a) Where adoption binds and the step term does not depend on speed, acceleration lowers the loss index if and only if $\beta > \varepsilon_H$ (shaded); if the step term rises with speed, $S'(s) > 0$, the escape needs $\beta > \varepsilon_H + sS'(s)/(HT_a)$ (dashed, drawn for one illustrative value of the added term). $\beta = 0$ is Proposition 1. (b) For $\beta < 1$ the kink $s^* = s_0(K/(As_0))^{1/(1-\beta)}$ falls as β rises when adoption binds at the reference speed, $K < As_0$ (the paper's example, $K = 1$, $A = 2$, $s_0 = 1$: raising β from 0 to 0.5 moves s^* from 0.50 to 0.25), and rises when capability binds there, $K > As_0$ (illustrative $K = 1.2$, $A = 1$, $s_0 = 1$). The escape is a fall in risk on the adoption branch, not an outward move of the kink, and it can occur while the kink falls. How hazard and adoption respond to speed is unmeasured; the paper has found no measurement of β .

4.5 Does a peak exist?

The loss index treats the transition as a block. To ask whether risk rises and then falls, we need a path. Let $E(t)$ be the stock of exploitable flaws, rising with the residual introduction of new flaws $\iota(t)$, the flaws new code introduces less than its protected share removes, and falling at the removal rate μ through patching and retirement. The hazard is φE , with φ

the exploitation intensity per flaw, on the period before ASI arrives at T_A (Appendix A, Proposition 3).

- The hazard rises and later falls before ASI if and only if, at some date, new flaws arrive faster than old ones are removed, and at a later date before ASI they arrive more slowly.
- With constant removal, introduction that never rises, and $E(0) < \iota(0)/\mu$, the path has at most one peak, and a monotone rise is possible.
- Suppose protection of new code spreads at rate γ , so that residual introduction falls from ι_0 toward a residue ι_1 , starting from an empty stock. Then a peak exists if and only if $\mu\iota_0 > \gamma\iota_1$. It is an intermediate, pre-ASI peak if it comes before T_A . After the peak, the path, continued past T_A , settles to the plateau $\phi\iota_1/\mu$.

Peak and plateau are therefore not alternatives: a peak can be followed by a plateau (Figure 6). Two policy readings follow. If protection eventually covers all new code ($\iota_1 = 0$), there is always a peak. If a residue remains, faster protection can remove the peak altogether, leaving a monotone rise to the same plateau. The plateau depends only on the residue and the removal rate, not on how fast protection spreads. So the long-run danger is set by what protection cannot reach and by how fast flaws are removed. Bounded introduction is not enough for a peak, and unbounded introduction does not rule one out (Appendix A, Proposition 3). Part (iii) starts from an empty stock. From any starting stock the plain reading is this: if today's stock already exceeds ι_0/μ , so that removal already outpaces the arrival of new flaws, no peak lies ahead in this model (Appendix A, Section A.7).

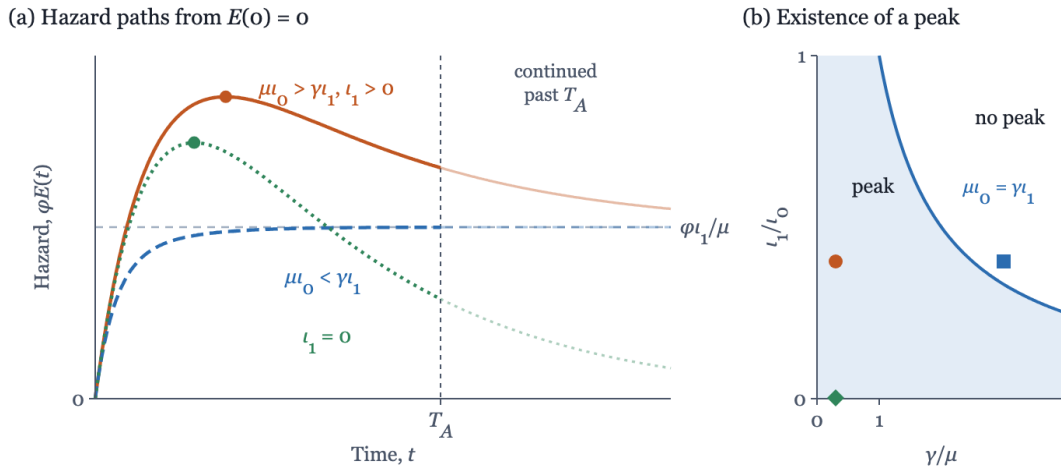


Figure 6. Flaw-stock dynamics and the existence of a peak (Proposition 3(iii)). (a) Hazard $\phi E(t)$ from an empty stock, with $dE/dt = \iota(t) - \mu E$, residual introduction $\iota(t) = \iota_1 + (\iota_0 - \iota_1)e^{-\gamma t}$ and exploitation intensity ϕ held constant. With a residue and slow protection ($\mu\iota_0 > \gamma\iota_1$) the path peaks and settles toward the plateau $\phi\iota_1/\mu$; with fast protection ($\mu\iota_0 < \gamma\iota_1$) it rises monotonically to the same plateau; with full protection ($\iota_1 = 0$) it peaks and falls toward zero. Markers are peaks at $\hat{t} < T_A$, the only peaks counted as intermediate, pre-ASI peaks; paths after T_A are the untruncated continuation, drawn faint. (b) A peak exists if and only if $\mu\iota_0 > \gamma\iota_1$; markers are the three cases of (a). Illustrative values: $\mu = \iota_0 = \phi = 1$; $\gamma = 0.3$ or 3 ;

$\iota_1 = 0.4$ or 0 ; $T_A = 6$. The model contains no speed or capability variable, so it does not say whether AI capability creates or moves the peak.

The model holds exploitation intensity per flaw constant. If attackers' capability per flaw rises faster than the stock falls, the hazard can keep rising while the stock declines. That is one channel, outside this model, by which faster capability can enter the long-run danger, and Section 6 returns to it.

So the answer to the first question is narrow. A peak can exist in the flaw-stock dynamics; whether AI capability creates or moves it is not modelled, since Proposition 3 contains no speed or capability variable.

4.6 Two peaks

Suppose the near loss and the far, ASI, loss can be put in common units, with stakes V_c and V_A and cumulative hazards $C_c(s)$ and $C_A(s)$, the far one assumed to rise in speed.

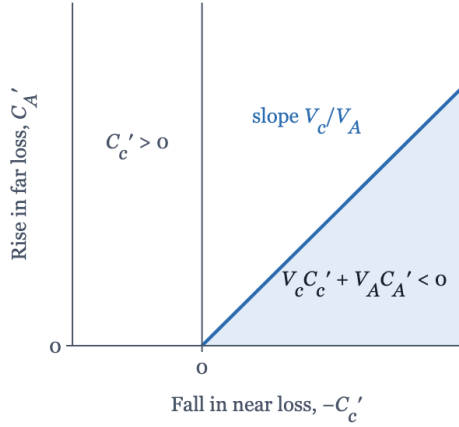
Proposition 4 gives the local test.

- A marginal acceleration lowers the total loss if and only if it lowers the near loss and $V_c/V_A > C'_A/(-C'_c)$: the ratio of stakes must exceed the ratio of the far loss's rise to the near loss's fall. Above the kink, with adoption time fixed as in Proposition 1, the near loss rises too, and no ratio of stakes justifies acceleration (for this two-loss objective; Section 4.7 adds a benefit).
- Discounting makes acceleration less attractive. With a rectangular hazard profile and step losses whose present value does not depend on speed, acceleration lowers the near loss below the kink if and only if $\varepsilon_H < x/(e^x - 1)$, with $x = rK/s$. That threshold is below one. For other fixed profiles the threshold differs but is still below one. If faster capability brings the steps earlier, the case for acceleration weakens further.
- If the far loss arrives at a date that falls with speed, acceleration also brings it nearer.
- If the two peaks are sequential catastrophes, total catastrophe probability falls with speed if and only if the fall in the near probability, weighted by survival of the far one, exceeds the rise in the far probability, weighted by survival of the near one (Appendix A, Proposition 4).

The test is local (Figure 7). It needs common units, differentiability, and no costs or benefits of speed beyond the two losses; Section 4.7 relaxes the last condition for one benefit, health gains from the same capability. A large, discrete acceleration can cut the total loss where the derivative test fails, so large moves must be judged in levels. The test also shows why the ranking of the two peaks may not matter. If adoption binds and adoption time falls with speed more slowly than the hazard rises ($\beta < \varepsilon_H$, Proposition 5), the near loss rises with speed, and both peaks point the same way. For the two-loss objective of Proposition 4, whether ASI is the system-wide peak, our fifth question, matters only where speed lowers the near peak. There Proposition 4 gives the break-even ratio of stakes, and the comparison needs a valuation of the far loss (Jones 2024, 576), which may be civilizational in scale (Bostrom 2019, 458), and which this paper does not supply. With health benefits B in the same units (Section 4.7), acceleration can improve the balance while near harm rises: with

$C'_A > 0$ it does so if and only if $B' - V_c C'_c > V_A C'_A$, so the far-peak comparison can matter there too. So we answer the fifth question only conditionally: Proposition 4 assumes that far risk rises with speed ($C'_A > 0$), and Appendix A, Section A.7, lists the reverse case.

(a) Two losses in common units



(b) Discounted near-loss threshold

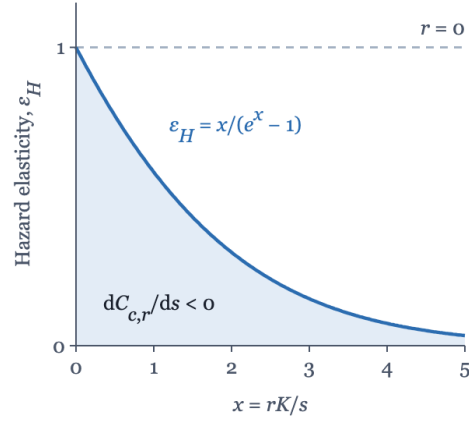


Figure 7. The local two-peak test (Proposition 4). (a) With near and far losses in common units, stakes V_c and V_A and $C'_A > 0$, a marginal acceleration lowers the total loss if and only if $C'_c < 0$ and $V_c/V_A > C'_A/(-C'_c)$ (shaded); where the near loss rises, as above the kink with adoption time fixed (Proposition 1), no ratio of stakes justifies acceleration for this two-loss objective. The ray's slope is illustrative. (b) With discounting at rate r , a rectangular hazard profile and step losses whose present value does not depend on speed, acceleration lowers the discounted near loss $C_{c,r}$ below the kink if and only if $\varepsilon_H < x/(e^x - 1)$, with $x = rK/s$; the threshold lies below the undiscounted value of one. The test is local: it needs common units and differentiability and admits no costs or benefits of speed beyond the two losses, and large moves must be judged in levels.

4.7 Life-enhancing benefits: a second reason to run

So far the only thing speed buys in the first model is a shorter stay under the rain. But the capability that arms attackers may also do good that has nothing to do with cyber defense. The example that motivates this section 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" (Jumper et al. 2021, abstract), in a field where structural coverage "is bottlenecked by the months to years of painstaking effort required to determine a single protein structure" (Jumper et al. 2021, opening paragraph). Half of the 2024 Nobel Prize in Chemistry went to Demis Hassabis and John Jumper "for protein structure prediction" (Nobel Prize Outreach 2024). Neither source says that AlphaFold shortens drug development. That capabilities of this kind bring health benefits sooner is the premise of this section, not a finding we report.

Where the time goes. A benefit of this kind reaches patients only after trials, approval, manufacture and uptake, which 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," a figure it attributes to an

earlier study (Sertkaya and Berger 2024, v). Its non-clinical stage “includes early exploratory research for target discovery,” lead optimization and animal testing; across therapeutic areas Phase 3 is the longest stage, at 38.0 months, and FDA review averages 16.2 months, a figure that includes the sponsor’s own time (Sertkaya and Berger 2024, 12). The clinical phases “may overlap” (Sertkaya and Berger 2024, 12). This does not identify how adoption overlaps capability development. The one AI-discovered drug programme whose timing we read reports about 18 months from target discovery to a preclinical candidate, and under 30 months from the start of target discovery to the end of phase 0/1 testing (Ren et al. 2024, abstract; Xu et al. 2025, introduction). That is a single programme reported by its developers, and its phase 2a trial followed on the clinical clock. We found no peer-reviewed study that estimates AI’s effect on total development time across drugs; that is a statement about our searches. Our reading of these figures, not the report’s claim, is that much of the wait for a new drug lies in stages that capability does not set.

The extension. We add a benefit to the first model and keep everything else. The coupling is full: the same speed s that advances capability over the distance K to the end of the cyber transition advances it over a distance K_B to the capability the benefit needs, and K_B may differ from K . Once that capability arrives, the benefit needs its own adoption time T_b , for trials, approval, manufacture and uptake, on the same three timings as the cyber side: parallel, $t_B = \max(K_B/s, T_b)$; sequential, $t_B = K_B/s + T_b$; or mixed, with a share m_B of the benefit’s adoption waiting for the capability, $t_B = \max(K_B/s, (1 - m_B)T_b) + m_B T_b$. The cyber and benefit timings need not match. From the date t_B on, the benefit averts y disability-adjusted life years (DALYs) per year. A DALY “combines time lost through premature death and time lived in states of less than optimal health,” and one DALY “can be thought of as one lost year of ‘healthy’ life” (WHO 2024, 6). Counting DALYs averted puts mortality and morbidity in one unit. Cyber harm enters in the same unit: V_c is now DALYs per harmful event, so $V_c C_c(s)$, with $C_c = HT + S$ the loss index of Section 4.2, is the expected health loss from the transition. Over a fixed calendar horizon L , long enough to include the whole transition, the objective is the balance

$$W_L(s) = y(L - t_B(s))^+ - V_c C_c(s).$$

This is a conditional balance, not total welfare. It assumes a constant positive flow y , a fixed V_c , a benefit that persists, and terms that add; and it counts each cyber harm once, in $V_c C_c$, never also as a cut in y . In the rain picture, the benefit is someone waiting at the far curb: a reason to arrive sooner that has nothing to do with getting less wet.

When speed helps. Write $\tau = K/s$ for the capability wait for the end of the cyber transition and $q = K_B/s$ for the capability wait for the benefit. Proposition 16 gives the change in the balance per proportional increase in speed, sW'_L , at speeds away from the kinks (Appendix A, Section A.3.5):

$$sW'_L = yq\chi_B\eta_L - V_c H(\varepsilon_H T - \chi_c \tau).$$

Here χ_c is one where capability sets the cyber transition’s length and zero where adoption does; χ_B is the same switch for the benefit; and η_L is one if the benefit arrives before the horizon. The first term is new: while capability sets the benefit’s arrival, a one percent increase in speed brings the benefit forward by one percent of q , and each year gained is

worth y . The second term is Proposition 1's cyber term, in DALYs. The table gives the four cases for both hazards, with the benefit arriving inside the horizon. A positive entry means that acceleration improves the balance, the reverse of the convention in Section 4.3's grid, which signs the loss. Under mixed cyber timing, T in the first and third rows is $\tau + mT_a$.

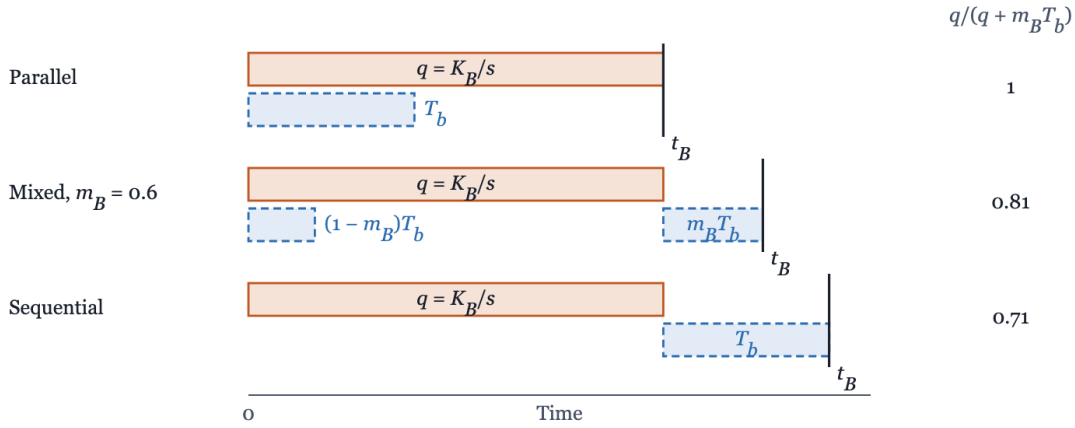
What sets each clock	Hazard set by speed: sW'_L	Hazard set by capability level: sW'_L
Capability sets both	$yq - V_c H(\varepsilon_H T - \tau)$	$yq + V_c \tau \bar{h}$
Adoption sets the cyber clock, capability the benefit	$yq - V_c H \varepsilon_H T_a$	$yq - V_c \tau [h(K) - \bar{h}]$
Capability sets the cyber clock, adoption the benefit	$V_c H(\tau - \varepsilon_H T)$	$V_c \tau \bar{h}$
Adoption sets both	$-V_c H \varepsilon_H T_a$	$-V_c \tau [h(K) - \bar{h}]$

The headline is the second row. Above the cyber kink, with a speed-set hazard, the benefit still waiting on capability and arriving inside the horizon, and no discounting, acceleration raises the balance if and only if

$$yq > \varepsilon_H V_c H T_a:$$

the benefit flow times the capability-set part of the wait for it must exceed the hazard's elasticity times the cyber loss over the adoption period. Under the level-set hazard the test is $yK_B > V_c K[h(K) - \bar{h}]$, which does not depend on speed while that row holds. Two cautions about the left side. It is yq , not y times the whole wait for the benefit: only the part of the wait that capability sets moves with speed. And m_B is not that part. On the benefit's capability branch, the capability-set share of the total wait is $q/(q + m_B T_b)$; in one illustrative case of our checks, $m_B = 0.6$ gives a share of 0.81 (Figure 8).

(a) Benefit arrival, $t_B(s)$, under three timings



(b) Reported average phase lengths, new drugs



Figure 8. The wait for a health benefit (Section 4.7; Proposition 16). (a) Arrival $t_B(s)$ under parallel, mixed and sequential benefit timing: the capability wait $q = K_B/s$ (solid) and the benefit's own adoption time T_b (dashed), of which a share m_B waits for the capability. On the benefit's capability branch the capability-set share of the total wait is $q/(q + m_B T_b)$: one under parallel timing, 0.81 for $m_B = 0.6$ in the case of the paper's checks ($q = 5$, $T_b = 2$), and $q/(q + T_b)$, here 0.71, under sequential timing; on the adoption branch it is zero, and m_B is not that share. Illustrative values. (b) Average non-clinical (31 months) and clinical (around 95 months) phase lengths for new drugs, as a report for the U.S. Department of Health and Human Services states them, attributing them to an earlier study (Sertkaya and Berger 2024, v); they do not identify how adoption overlaps capability development.

The test holds only while capability sets the benefit's arrival. Under parallel or mixed benefit timing, once speed passes the benefit's own kink, K_B/T_b , or $K_B/((1 - m_B)T_b)$ under mixed timing, the benefit waits only on trials, approval and uptake. Its term vanishes and Proposition 1's sign returns: running gets no one to the far curb sooner; whether it keeps you drier is Proposition 1's question again. Under sequential benefit timing there is no such cutoff. Below the cyber kink, benefits widen the range in which speed helps. The speed-set test becomes $\varepsilon_H < [\tau + yq\chi_B\eta_L/(V_c H)]/T$, which can hold even when $\varepsilon_H > 1$. In an illustrative case with $H = s^2$, at $s = 0.5$, $\varepsilon_H = 2$ and the cyber loss rises with speed ($C'_c = 1$), but the benefit's derivative is 16, so $W'_L = 15 > 0$.

There are now two kinks (Figure 9), the cyber kink K/T_a ($K/((1 - m)T_a)$ under mixed timing) and the benefit kink. If the cyber kink comes first, the speeds between the two are the second row, the only region above the cyber kink where benefits can justify acceleration. If the benefit kink comes first, the speeds between them are the third row: the benefit no longer responds to speed, and Proposition 1's test decides. Above both, a speed-set hazard makes acceleration lower the balance. Appendix A, Section A.3.5, covers both

orderings, coincident kinks, the date at which the benefit first arrives inside the horizon, and all nine pairs of cyber and benefit timing.

(a) Cyber kink first

Clock indicators	$\chi_c = 1, \chi_B = 1$	$\chi_c = 0, \chi_B = 1$	$\chi_c = 0, \chi_B = 0$
sW'_L , speed-set	$yq - V_c H(\varepsilon_H T - \tau)$	$yq - V_c H \varepsilon_H T_a$	$-V_c H \varepsilon_H T_a < 0$
sW'_L , level-set	$yq + V_c \tau \bar{h}$	$yq - V_c \tau [h(K) - \bar{h}]$	$-V_c \tau [h(K) - \bar{h}]$
	K/T_a	K_B/T_b	Speed, s

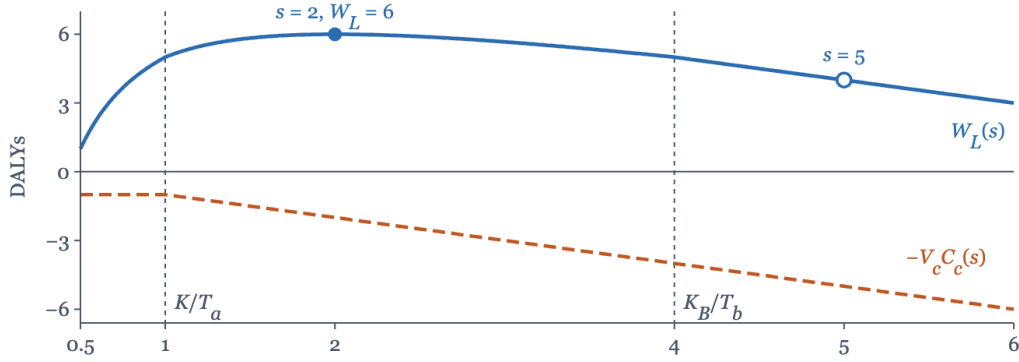
(b) Benefit kink first

Clock indicators	$\chi_c = 1, \chi_B = 1$	$\chi_c = 1, \chi_B = 0$	$\chi_c = 0, \chi_B = 0$
sW'_L , speed-set	$yq - V_c H(\varepsilon_H T - \tau)$	$V_c H(\tau - \varepsilon_H T)$	$-V_c H \varepsilon_H T_a < 0$
sW'_L , level-set	$yq + V_c \tau \bar{h}$	$V_c \tau \bar{h}$	$-V_c \tau [h(K) - \bar{h}]$
	K_B/T_b	K/T_a	Speed, s

Figure 9. Branches of the health–cyber balance as speed rises, by which clock binds (Proposition 16; Appendix A, Section A.3.5): (a) cyber kink first, (b) benefit kink first. Each cell gives sW'_L , the change in the balance per proportional increase in speed, for a benefit arriving inside the horizon and without discounting; χ_c and χ_B equal one where capability sets the cyber and the benefit clock. A positive entry means that acceleration improves the balance, the reverse of the sign convention of the loss grid in Figure 4. The outlined cell is the only region above the cyber kink where benefits can justify acceleration; under the speed-set hazard it does so if and only if $yq > \varepsilon_H V_c H T_a$. Above both kinks the speed-set entry is negative. Kinks are drawn for parallel timing; under mixed timing they are $K/((1 - m)T_a)$ and $K_B/((1 - m_B)T_b)$, and $T = \tau + mT_a$ where capability sets the cyber clock. Sequential benefit timing has no benefit kink. Entries are local derivatives away from the clock and horizon kinks, with T_a , T_b , m and m_B fixed in speed and S constant; Proposition 5's adoption that responds to speed is not assumed.

An optimum above the kink is possible, not guaranteed. Take illustrative values, not estimates: $K = T_a = V_c = y = T_b = 1$, $K_B = 4$, parallel timing on both clocks, $H = s$, $S = 0$, $L = 10$, and speeds from 0.5 to 6. The cyber kink is at $s^* = 1$ and the benefit kink at 4. The balance is $9 - 4/s$ below the cyber kink, $10 - 4/s - s$ between the kinks, and $9 - s$ above the benefit kink. Its unique maximum is at $s = 2$, twice the cyber kink, where $W_L = 6$, against 1, 5, 5 and 3 at speeds 0.5, 1, 4 and 6. At that optimum the cyber loss is rising; the benefit pays for it. The same example refutes the claim that benefits always justify speed. At $s = 5$ the benefit is large in level, $B_L = 9$, but it no longer responds to speed, so $W'_L = -1$, and raising speed from 5 to 6 lowers the balance from 4 to 3. Health benefits in levels are not a marginal argument for speed (Figure 10).

(a) Balance $W_L(s)$ and cyber term alone



(b) The two terms of sW_L'

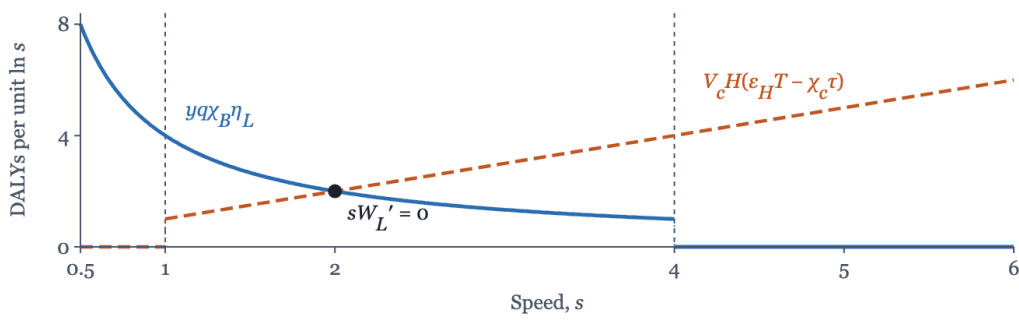


Figure 10. The illustrative example of Section 4.7: $K = T_a = V_c = y = T_b = 1$, $K_B = 4$, parallel timing on both clocks, $H = s$, $S = 0$, $L = 10$, speeds from 0.5 to 6; illustrative values, not estimates. (a) The balance $W_L(s)$ is $9 - 4/s$ below the cyber kink at 1, $10 - 4/s - s$ between the kinks, and $9 - s$ above the benefit kink at 4; its unique maximum is $W_L = 6$ at $s = 2$, against 1, 5, 5 and 3 at $s = 0.5, 1, 4$ and 6. The dashed line is the cyber term alone, $-V_c C_c(s)$, on the same axis. At $s = 5$ the benefit level is $B_L = 9$, but it no longer responds to speed, and $W_L' = -1$. (b) The two terms of $sW_L' = yq\chi_B\eta_L - V_c H(\varepsilon_H T - \chi_c \tau)$: the benefit term is $4/s$ while capability sets the benefit's arrival and zero above the benefit kink; the cyber term is zero below the cyber kink and s above it. They cross at the optimum, where $yq = \varepsilon_H V_c H T_a$ and $sW_L' = 0$. An optimum above the kink is possible, not guaranteed.

Discounting. Discounting both health gains and cyber losses from the same date at rate r , with $A_r(u) = (1 - e^{-ru})/r$, the test above the cyber kink, on the benefit's capability branch and with step losses whose present value does not depend on speed, becomes $yqe^{-rt_B} > V_c H \varepsilon_H A_r(T_a)$; steps that come earlier with speed add to the right side (Proposition 16(ii)). Without the benefit term, and with parallel timing, the threshold below the kink is Proposition 4(ii)'s $x/(e^x - 1)$. Discounting lowers the credit for a late benefit, but it changes the cyber marginal loss too, so its effect on the net sign has no general direction. WHO's current burden estimates use "No discounting for time" (WHO 2024, 7), so r here is our modelling choice.

Two peaks. With a far loss in the same units, the derivative of the balance is $B' - V_c C'_c - V_A C'_A$. With $C'_A > 0$, acceleration improves it if and only if $B' - V_c C'_c > V_A C'_A$. Proposition 4's

requirement that the near loss fall belongs to its two-loss objective; with benefits, near harm can rise while the balance improves. We add no far-loss result.

What the extension leaves out. Cyber harm can be measured in health terms.

Ransomware attacks on hospitals “decrease hospital volume by 17–24 percent during the initial attack week,” and “Among patients already admitted to the hospital when a ransomware attack begins, in-hospital mortality increases by 34–38 percent” (Neprash, McGlave and Nikpay 2026, abstract). After data breaches, 30-day acute myocardial infarction mortality “increased as much as 0.36 percentage points during the 3-year window following a breach,” which the authors associate with remediation (Choi, Johnson and Lehmann 2019, abstract). Targeted hospital cyberattacks “may be associated with disruptions of health care delivery at nontargeted hospitals within a community” (Dameff et al. 2023, abstract). None of these studies estimates DALYs per attack, so no source we read gives V_c . Their findings point to a link we do not model: attacks on health systems may cut the realized benefit flow y , which the extension holds constant. Two further limits are named, not modelled (Appendix A, Section A.7): narrow tools, such as specialized structure-prediction models, could advance the benefit on a clock of their own, breaking the coupling; and the capability that speeds drug discovery may also aid biological attack, a risk outside this balance.

The answer, conditionally. Health benefits from the same capability add a second reason to run, and it can outweigh a rising cyber loss, even past the cyber kink. But it does so only while the benefit still waits on capability rather than on its own trials, approval and uptake; past the cyber kink, with a speed-set hazard, benefits arriving inside the horizon and no discounting, only by the margin $yq > \varepsilon_H V_c HT_a$, and in other cases by Proposition 16’s master tests. Once the benefit’s own clock sets its pace, faster capability buys no earlier benefit, and Proposition 1’s test alone decides: above the cyber kink, with a speed-set hazard, it adds cyber harm (with a level-set hazard, when $h(K) > \bar{h}$). Where current speeds sit relative to either kink, and what y , V_c and q are, we have not measured.

5. Evidence I: capability and its limits

Rapid progress on selected checked tasks. By late September 2026, Epoch’s own runs on the private FrontierMath sets 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 top scores, and FrontierMath was developed with OpenAI funding; OpenAI has exclusive access to a subset of it (Epoch AI 2026a; 2026b, “About”). On ARC-AGI-2’s semi-private set the best score was 95% (ARC Prize 2026). An IMO 2026 perfect score comes from a lab’s own account of official marking (dots team 2026). Five PutnamBench entries report all 672 Lean problems solved, four with solutions and one in no-answer mode; the leaderboard records, as we saved them, do not document public verification. Two entries say their proofs were shared privately for verification, and a third that its proofs were shared privately, without stating a purpose (PutnamBench 2026). These are reported submissions, not independently established proof totals. FrontierSWE and FrontierCode remained unsaturated, at best scores of 65.5% and 54.6% (FrontierSWE 2026; Cognition 2026; Epoch AI 2026a). Epoch’s dated rows cover models released over about four months for FrontierSWE V2 and eight months for FrontierCode: in our judgment, too little for a trend fit. On agentic coding,

METR's 50% horizon is the task duration, measured by human expert completion time, at which an agent is predicted to succeed half the time. Its warning that "Measurements above 16 hrs are unreliable with our current task suite" limits extrapolation, and the page has not been updated since 8 September 2026 (METR 2026). Fitting-window and historical-version comparability remain unestablished here, so we give no doubling-rate estimate.

Two cautions keep this from proving too much. Scaling laws are fitted to loss, not skill (Kaplan et al. 2020, abstract; Hoffmann et al. 2022, abstract). Benchmark performance is only partly predictable from compute: across a tenfold increase in training compute, average absolute extrapolation errors were 6 percentage points for aggregate BIG-Bench Hard performance and 18 percentage points for individual BIG-Bench tasks (Owen 2024, abstract). DeepSeek reports improved reasoning capabilities through reinforcement learning after pre-training (DeepSeek-AI 2025, abstract), a lab's report on its own models; and on problems where a smaller model already has some success, compute spent at inference can let it outperform a model fourteen times larger (Snell et al. 2024, abstract). The record shows rapid progress on selected automatically checked tasks, not a cross-domain ranking of progress rates. Cheap checking as an explanation is a hypothesis, without causal identification here. For research mathematics and real software, our untested judgment is that specifying what to check may be a bottleneck.

Total decision procedures have limits. Rice's theorem rules out a total algorithm deciding, from an arbitrary partial recursive function enumerating a set, whether that set has a property possessed by some but not all recursively enumerable sets (Rice 1953, §6, Theorem 6, Corollary B and its interpretation). Rice explicitly leaves room for special methods for particular functions. This is a mathematical limit, not an empirical ceiling on useful verification.

The cited software proofs have particular scopes and assumptions. seL4's proofs assume that the kernel's few lines of assembly, the hardware, the in-kernel TLB and cache management, and the boot code are correct, with DMA off or trusted; the security proofs require specified system-configuration conditions; and the confidentiality proof "makes no guarantees about the absence of covert timing channels" (seL4 Project n.d., "What are the proof assumptions?" and "Does seL4 have zero bugs?"). Random testing found six wrong-code bugs in CompCert 1.6's unverified front end and a gap in the PowerPC semantics against which its back end is verified (the assembler caught the resulting out-of-range values), while finding no wrong-code errors in the early-2011 development version (Yang et al. 2011, §3.1). A study of three verified distributed systems found sixteen bugs, caused by violated assumptions mostly at the interface of verified and unverified components, and no protocol bugs in the verified parts (Fonseca et al. 2017, abstract, §4.4). These reports illustrate boundaries of the cited proofs; they do not establish that verification cannot cover specifications, interfaces, hardware or configuration in other systems. A guarantee reaches the installed base only for systems to which the proof and its assumptions apply.

General models in the reviewed examples. The security-relevant examples reviewed here use general frontier models; the holder described its restricted 2026 model as "a general-purpose, unreleased frontier model" (Anthropic 2026a, Introduction). These examples do not establish firms' available choices. Specialization and distinct offensive and

defensive development paths remain possible. Nor does the full PutnamBench result show specialization: of the five prover systems reporting it, three name general frontier models in their harnesses (GPT-5.6 Sol, DeepSeek V4, Claude Opus 5), two do not name their model, and the leaderboard records, as we saved them, do not document public verification (PutnamBench 2026). Whether narrow code systems could supply the security-relevant capability independently of general frontier models needs evidence. Section 8 examines the uses of a held general capability.

6. Evidence II: the transition and what sets its length

6.1 The attacker-favoring transition, as a hypothesis

That the transition favors attackers is the hypothesis under examination. The following are reported laboratory and competition outcomes, not a matched comparison with a human workforce; the laboratory figures are the lab's own reports, and independent replication remains unverified here:

- DARPA reported that in the final of its AI Cyber Challenge (August 2025), finalists' automated systems found 54 unique synthetic vulnerabilities in the final's 63 challenges and patched 43 of them, and found 18 real, non-synthetic vulnerabilities, for which they supplied 11 patches (DARPA 2025, "Deep Dive" and editor's note). DARPA's editor's note counts 63 synthetic vulnerabilities in the final, correcting an earlier 70 (Figure 11).
- Anthropic reported in February 2026, in the post released with Claude Opus 4.6, that it had used Claude to find more than 500 high-severity vulnerabilities in open-source software, some in codebases that fuzzers had run against for years, and had validated them (Anthropic 2026f, opening paragraphs and "Setup"). The validation was done by Anthropic's own and external human security researchers, after sanitizer checks and model de-duplication. The post gives examples, not a list, so the count, the de-duplication and the severity ratings are the lab's, and we have not verified them independently. Its April report says Mythos Preview developed working JavaScript-shell exploits 181 times, against two successes in several hundred attempts for Opus 4.6, on JavaScript-engine bugs in Firefox 147 that Opus 4.6 had found, in a test harness without the browser's process sandbox or other defense-in-depth mitigations (Anthropic 2026g, "The significance of Claude Mythos Preview for cybersecurity" and note 1). These count successful attempts, not distinct vulnerabilities. Since 181 exceeds the 112 unique reports from that work (Anthropic 2026h), the count must include repeats.
- In its June evaluation of 18 Firefox security patches, Anthropic reported that Mythos Preview built eight working exploits. Its timing plots take each vulnerability's fastest successful proof-of-concept and exploit stages from three trials per stage; the first plotted exploit came in under an hour and the eighth at roughly twelve hours (Anthropic 2026c, "N-days on Firefox"). Inputs included public patch diffs, component and severity metadata, and instrumented vulnerable and patched builds of standalone jsshell, rather than the full browser. The patches had long been public: the study kept "only bugs whose fixes had been public in Mozilla's source

repository for at least 90 days” (Anthropic 2026c, “N-days on Firefox”). These are evaluation runtimes, not elapsed field-exploitation times; independent replication and field transfer remain unverified.

- The UK AI Security Institute reported that, by its February 2026 estimate, the human-expert duration of cyber tasks that frontier models could complete with 80 percent reliability in its narrow cyber suite, under a 2.5-million-token cap per task, had doubled every 4.7 months since late 2024 (AISI 2026). It notes that the tasks “cover only some of the capabilities relevant to real-world cyberattacks,” that the exponential fit “is not a future prediction, nor a fixed law,” and that omitting any single model from the pre-Mythos fit moves the estimate between 4.1 and 5.0 months (AISI 2026). It adds that two later models, Claude Mythos Preview and GPT-5.5, “substantially exceeded both doubling rate trends,” and that “it is unclear whether this represents a new, faster trend” (AISI 2026). The estimate is not an unrestricted trend.

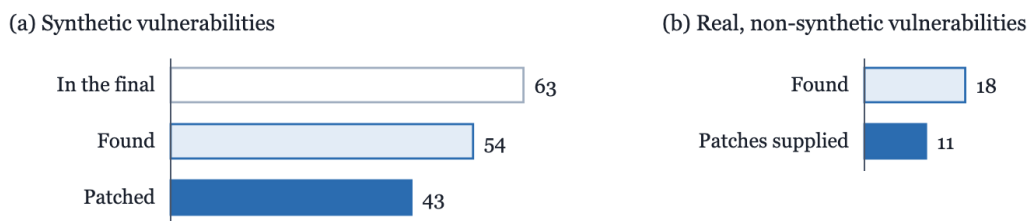


Figure 11. Final of DARPA’s AI Cyber Challenge, August 2025, as DARPA reported it (DARPA 2025, “Deep Dive” and editor’s note). (a) Synthetic vulnerabilities: the editor’s note counts 63 in the final, correcting an earlier 70; finalists’ automated systems found 54 unique synthetic vulnerabilities in the final’s 63 challenges and patched 43 of them. (b) Real, non-synthetic vulnerabilities: 18 found, for which the systems supplied 11 patches. The panels count different things. These are competition outcomes, not a matched comparison with a human workforce, a measured offense–defense balance or realized harm (Section 6.1).

For the 138 vulnerabilities disclosed in 2023 that it tracked as exploited in the wild, Mandiant reported an average time-to-exploit, measured before or after patch release, of five days after excluding 15 statistical outliers (13 zero-days and two n-days); including them, the average was 47 days (Charrier and Weiner 2024, “Time-to-Exploit”). The post does not say whether its earlier averages were filtered the same way, so comparisons with historical means require consistent filtering and samples.

These results show capability. They do not show a measured offense–defense balance, and they do not show realized harm. Anthropic reported that an AI-orchestrated campaign succeeded “in a small number of cases” (Anthropic 2025, announcement post, opening paragraphs). Counts of exploitation in the wild after 2023 exist but do not isolate AI. Google’s threat-intelligence group tracked 90 zero-days exploited in the wild in 2025, against 78 in 2024 as its 2025 review counts them (Charrier et al. 2026, executive summary). Counting n-days as well, it recorded 127 exploited vulnerabilities in 2025 and 141 from January to August 2026, a rise that since May 2026 has, in its reading, tracked the growth in disclosures (Grunewald, Mazumdar and Vanderlee 2026, “In-the-Wild Exploitation”). CISA added 186 entries to its catalogue of known exploited vulnerabilities in

2024 and 245 in 2025 (CISA 2026). These count vulnerabilities observed or listed as exploited, not incidents or harm. None of the sources we opened gives an aggregate count of AI-enabled exploitation; the same group calls exploitation of AI-discovered flaws “an early indicator rather than an established trend” (Grunewald, Mazumdar and Vanderlee 2026, “AI as the Hunter”). The balance itself cannot be measured for “cyberspace” as a whole: it “can be understood only in the context of specific adversaries with distinctive goals and levels of capability” (Slayton 2017, 106), and it can shift with the level of investment (Garfinkel and Dafoe 2019, 736). Much of the quantitative evidence above comes from one lab reporting on its own model. The DARPA and UK AI Security Institute results are independent of the lab and point the same way (DARPA 2025; AISI 2026). We therefore hold, with high confidence, that the capability exists, and with low confidence what it does to net harm in the field.

Disclosed and known-patched counts. The same lab’s disclosure dashboard reported 6,157 vulnerabilities disclosed across 591 open-source projects as of 2 October 2026, adding, “To our knowledge, 516 of these have been patched” (Anthropic 2026e, overview paragraph beginning “As of October 2, 2026”). This is a dated dashboard count, preserved in the archive snapshot of 7 October 2026. The cumulative counts do not measure cohort patch rates, patch capacity, deployment, or the share of systems still exposed. The lab itself calls the number of patches created a lagging indicator of progress (Anthropic 2026e, process explanation, “In our triage process”), and says that it discloses only a subset of what its models find, because “independent human triage and review is the rate limiting step” (Anthropic 2026e, overview paragraph beginning “As of October 2, 2026”).

6.2 What sets the transition’s length

The deployment evidence concerns different populations and endpoints, and it bears on different parts of adoption time. Under mixed timing (Section 4.3) only the share of adoption that runs alongside capability, $(1 - m)T_a$, sets the kink, so only that share can make adoption bind; the share that waits for the capability, mT_a , lengthens the transition without moving the kink. In our judgment, patching flaws that the capability finds is plausibly sequential, and retirement and migration plausibly parallel (Appendix A, Section A.7). We sort the series accordingly. All of them suggest long lags (Figure 12).

Patching, which bears on the share that waits for the capability:

- Kenna Security and the Cyentia Institute reported a median vulnerability remediation time of 90 days, with about one-third (32.5%) still open after 180 days, from a survival analysis of 190 million vulnerabilities at 12 Kenna customer organizations, with data to late 2018 (Kenna Security and Cyentia Institute 2019, 4, 18–19). Remediation medians bear on the patching part of the adoption time T_a . The cohort is 12 firms on one vendor’s platform, so censoring and denominators require checking before transferring these estimates to other populations.
- Anthropic, taking Windows Autopatch timelines as “likely on the faster side of patching management today,” states that a patch typically takes seven days before it is “shared out to 90% of enrolled devices in a fleet” (Anthropic 2026c, “N-days on Windows”). Microsoft’s documentation of Autopatch groups describes a

configurable default schedule, with deferral and deadline periods ranging from one to 20 days; the Microsoft pages we examined do not state the seven-day figure (Microsoft 2026). Update distribution bears on part of an operator's deployment time τ_j (Proposition 6). Distribution is not confirmed installation or remediation, so this is not comparable to the preceding endpoint.

- On client hosts, Nappa et al. (2015, abstract, §I), using five years of data on 8.4 million hosts and 1,593 vulnerabilities in ten client applications, found that “the median time to patch half of the vulnerable hosts is 45 days” and that “the median fraction of vulnerable hosts patched when exploits are released is at most 14%”; “only 28% of the patches in our study reach 95% of the vulnerable hosts during our observation period” (§VIII). In open-source repositories, “over 50% of CVEs were patched more than a week before public disclosure” (Li and Paxson 2017, §5.1.2), so attackers who watch the repositories “can often get a jump of weeks to months on targeting not-yet-patched systems” (abstract). In earlier CERT case studies, “Many systems remain vulnerable to security flaws months or even years after corrections become available,” though the authors judge that the intrusions such deployment would prevent “would likely only be nuisance attacks by script kiddies” (Arbaugh, Fithen and McHugh 2000, 58). These measure patch creation and deployment, not retirement or migration.

Retirement and migration, which bear on the share that runs alongside it:

- In June 2019 GAO identified ten critical federal legacy systems most in need of modernization; as of February 2025, agencies reported having completed three of the ten modernizations (GAO 2025, Highlights and table 2). This bears on the retirement and migration part of T_a for those systems, not on patch remediation.
- The share of Chrome navigations using HTTPS rose from about 30–45% in 2015 to about 95–99% around 2020, and has since largely plateaued (Thompson et al. 2025). Post-quantum key agreement secured nearly 2% of TLS 1.3 connections to Cloudflare in early March 2024 (Westerbaan 2024) and, on a different denominator, 52% of the human-generated Web traffic Cloudflare observed in early December 2025 (Belson 2025). Chrome announced the phase-out of SHA-1 certificates in September 2014 and planned to remove support in Chrome 56, around the end of January 2017, about two and a half years later (Palmer and Sleevi 2014; Whalley 2016). Native IPv6 access to Google first exceeded 50% on a single day, 28 March 2026 (50.10%), about 27 years after the RFC 2460 specification of December 1998 (we date the lag from RFC 2460, the specification we read; dating it from an earlier IPv6 specification would lengthen it), while the Internet Society reported a 43% average across its IPv6 data sources on 21 April 2026 (Google 2026; Deering and Hinden 1998; Ford 2026). These are protocol migrations, which bear on the protection spread rate γ of Proposition 3 or on the T_a of a migration-type defense. These populations, starting dates and milestones differ. They establish neither a harmonized half-adoption distribution nor a causal effect of centralization.

Time to exploit (§6.1)

Mandiant, average time to exploit
138 exploited vulnerabilities disclosed in 2023

Patching (§6.2; judged sequential)

Windows Autopatch, to 90% of fleet
update distribution, not installation

Client hosts, median to patch half
8.4 million hosts, ten client applications

Median remediation time
12 organizations, data to late 2018

Retirement and migration (§6.2; judged parallel)

SHA-1 certificates in Chrome
announced Sep 2014 → planned removal, Chrome 56

HTTPS share of Chrome navigations
≈30–45% in 2015 → ≈95–99% around 2020

GAO critical federal legacy systems
June 2019 list → Feb 2025: 3 of 10 done

IPv6 access to Google
RFC 2460, Dec 1998 → first day >50%, Mar 2026

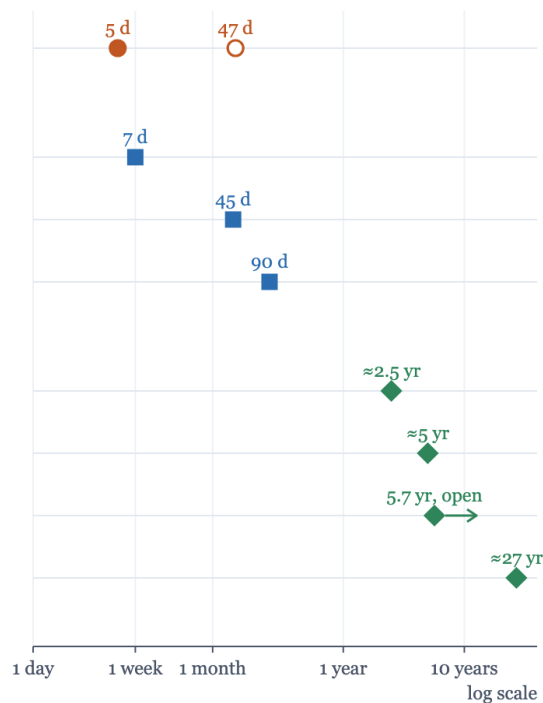


Figure 12. Reported lags on one logarithmic scale, placed together only to show their order of size (Sections 6.1–6.2). The rows come from different populations and measure different things: an average time to exploit for the 138 vulnerabilities disclosed in 2023 that Mandiant tracked as exploited in the wild, measured before or after patch release, with 15 outliers excluded (filled marker, 5 days) and included (open marker, 47 days) (Charrier and Weiner 2024); update distribution to 90% of enrolled devices under Windows Autopatch, taken as “likely on the faster side,” which is not confirmed installation (Anthropic 2026c; the Microsoft pages examined do not state the seven-day figure); the median time to patch half of vulnerable client hosts (Nappa et al. 2015); median remediation time at 12 organizations on one vendor’s platform (Kenna Security and Cyentia Institute 2019); and four retirement and migration milestones: the SHA-1 phase-out from announcement to planned removal, about two and a half years (Palmer and Sleevi 2014; Whalley 2016); the rise in the HTTPS share of Chrome navigations, 2015 to around 2020 (Thompson et al. 2025); the ten critical federal legacy systems GAO identified in June 2019, of which three were modernized by February 2025, an interval still open (GAO 2025); and IPv6 access to Google first exceeding 50% on a single day, about 27 years after RFC 2460 (Google 2026; Deering and Hinden 1998). Distribution is not installation, migration milestones are not common adoption quantiles, and none of these lags locates current speeds relative to the kink. Assigning patching to the sequential share and retirement and migration to the parallel share is the paper’s judgment (Section 6.2), not a measured timing decomposition.

One proposed reason, a judgment, is an externality: “Systems are particularly prone to failure when the person guarding them is not the person who suffers when they fail” (Anderson and Moore 2006, authors’ manuscript, 1).

By the paper's own judgment, then, the patching series bear on the sequential share and do not show that adoption binds; that question rests on the retirement and migration series alone. We read those as support for a provisional judgment: adoption already binds, in the sense of Proposition 1, for the populations they cover, critical federal legacy systems and web-protocol migration. That is a judgment, not an established result. The lags show that deployment is slow. They do not identify where current speeds sit relative to the kink, because the capability distance K is not measured. Nor do they estimate whether adoption is accelerating enough to satisfy Proposition 5's condition $\beta > \varepsilon_H$. What would change the judgment is evidence that AI deployment tools are shortening retirement and migration times sharply, across sectors rather than in the fastest channels.

6.3 Long-run exposure

The far curb may not be permanent. Proposition 3 makes the point in its own model: the plateau φ_{t_1}/μ depends on residual flaw introduction, the new flaws that protection cannot reach. In CISA's catalogue of known exploited vulnerabilities (version of 21 October 2026), among entries with a usable CWE tag, the share in authentication, authorization and access-policy classes rose from 19% (171 of 920 entries added from November 2021 through 2023) to 27% (180 of 659 added from January 2024 to 20 October 2026). The memory-corruption share fell from 33% (299 of 920) to 19% (125 of 659); 133 and 42 entries without a usable CWE are excluded. These shares are our calculation from CISA's CWE tags (CISA 2026), using judgment-based class lists given in Supplement S2's data note. An entry with several weaknesses goes to the first matching class in the order memory corruption, injection and input handling, access policy; another order would move any one class's share by less than 1.9 points in the earlier period and 1.1 in the later. Two checks leave the direction unchanged (Figure 13). Excluding the catch-all CWE-20 from the usable denominator, as NVD-CWE-noinfo is, leaves 839 and 648 usable entries; memory corruption, injection and input handling, access policy and other then number 299, 287, 171 and 82 in the earlier period and 125, 284, 180 and 59 in the later, so access policy goes from 20% to 28% and memory corruption from 36% to 19%. Keeping only entries listed in their CVE identifier's year or the next, a check on CISA's listing of older flaws, memory corruption goes from 35.2% (160 of 454) to 19.7% (105 of 532) and access policy from 19.8% (90 of 454) to 28.0% (149 of 532). The mix reflects CISA's listing practice as well as attackers.

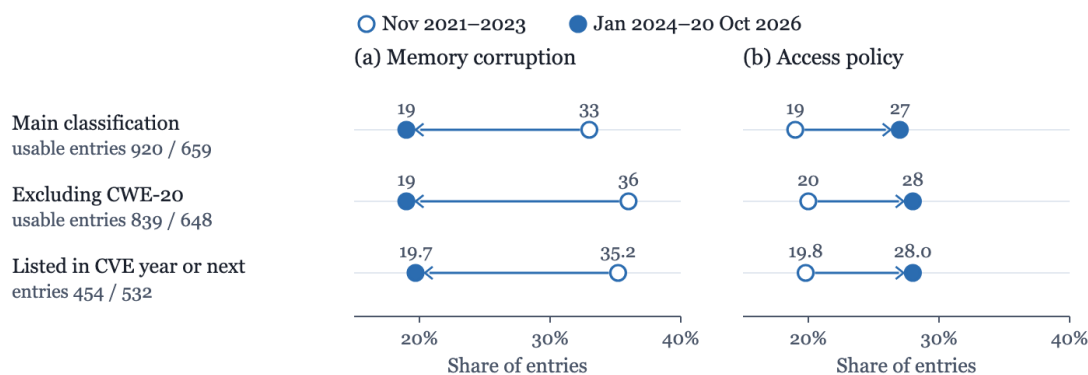


Figure 13. Shares of two weakness classes among entries in CISA’s catalogue of known exploited vulnerabilities added from November 2021 through 2023 and from January 2024 to 20 October 2026 (version of 21 October 2026; Section 6.3). Rows are the main classification, as shares of entries with a usable CWE tag, and the paper’s two checks: excluding the catch-all CWE-20 from the usable denominator, and keeping only entries listed in their CVE identifier’s year or the next. Memory corruption falls and access policy (authentication, authorization and access-policy classes) rises in all three. The shares are the paper’s calculation from CISA’s CWE tags (CISA 2026), with judgment-based class lists given in Supplement S2’s data note and an entry with several weaknesses assigned to the first matching class; the mix reflects CISA’s listing practice as well as attackers.

So the answer to the first question stays narrow: a peak can exist in the flaw-stock dynamics, under the conditions of Proposition 3, and whether AI capability creates or moves it is not modelled. Whether a peak is followed by a lower plateau, or by a second rise, depends on the residue of new code that protection cannot reach and on attackers’ capability per flaw.

6.4 Channels outside the cited proofs, and candidate defenses

The cited software proofs do not close every human or AI-agent attack channel. Scattered Spider actors call IT help desks “to reset passwords and/or transfer MFA tokens” (FBI et al. 2023, “Technical Details”). FinCEN states that criminals have reportedly used generative AI to impersonate “an executive or other trusted employee” and instruct victims “to transfer large sums or make payments” (FinCEN 2024, 5). Against adaptive attackers, Nasr et al. bypassed 12 recent defenses against jailbreaks and prompt injections, ten of them evaluated on prompt-injection benchmarks, “with attack success rate above 90% for most” (Nasr et al. 2025, abstract and table 1). CaMeL, an architectural defense, solved 77% of AgentDojo tasks “with provable security,” against 84% for an undefended system (Debenedetti et al. 2025, abstract). These rates come from different threat models and task sets and are not directly comparable. They establish no category-wide ceiling on defenses. Broader architectures and controls remain unresolved.

Lowering the peak does not 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” (Sandbrink et al. 2022, 1). Patch distribution, legacy retirement, memory-safe languages for new code, post-

quantum and hybrid migration, hardware-bound authentication, and verified kernels and libraries need no frontier capability; that is our judgment. The capability-dependent defenses, AI vulnerability discovery, AI patching and AI proof synthesis, are, in our judgment and for general frontier models (Section 5), not separable, or for AI patching only partly: the same models serve attackers.

6.5 The crux

The defensive case for acceleration is right where adoption binds only if frontier capability shortens adoption time faster, in proportion, than it raises the hazard: Proposition 5's $\beta > \varepsilon_H$, or more if the step term rises with speed (Section 4.4). Anthropic reports that “independent human triage and review is the rate limiting step” in its disclosure program (Anthropic 2026e, overview paragraph beginning “As of October 2, 2026”). On verified translation of legacy code, the evidence we read comes from benchmarks. Khatry et al. (2025, abstract) report a benchmark of 100 C repositories, with correctness assessed using test cases; OpenAI o1 solved 15 tasks in a single-shot setting. Yang et al. (2024, abstract) report that Claude-2 with VERT passed bounded model checking on 42% of 1,394 competitive-programming benchmark programs. The AI Cyber Challenge’s automated patches (Section 6.1) are patch creation on the defense side; they do not measure β , which concerns deployment. We have no measurement of β . The crux of the speed question is therefore open, and it is empirical.

7. Holding the storm: secrecy for security

7.1 The head start is worth at most its length

Secrecy splits the clocks. The holder reaches the capability and holds it. A share π of attack effort has parallel access anyway, for example states with their own frontier capability. The rest arrive when the window W closes, by open-weights catch-up, a rival’s release, a leak or theft. The hazard that ends the window adds up over these causes as intensities of mutually exclusive exits; a shared shock counts once.

Proposition 6 bounds what a window can do. Take harm as linear in exposure time, coverage as independent of the window, and attackers who resume at unchanged intensity, with no backlog, when the window closes. A covered operator that needs time τ_j to deploy fixes saves expected exposure $(1 - \pi)E[\min(W, \tau_j)]$, which can never exceed $(1 - \pi)\min(E[W], \tau_j)$. If the window ends at constant hazard λ , the saving is $(1 - \pi)(1 - e^{-\lambda\tau_j})/\lambda$. As a share of the operator’s baseline exposure to all attackers, it is asymptotically equal to $1 - \pi$ for operators who deploy well within the window, and to $(1 - \pi)/(\lambda\tau_j)$ for those who need many windows’ length to deploy (Appendix A, Proposition 6). The head start is worth at most its own length, and only against attackers who lacked access (Figure 14). The closed form is the familiar restricted-mean identity (Royston and Parmar 2013, 3).

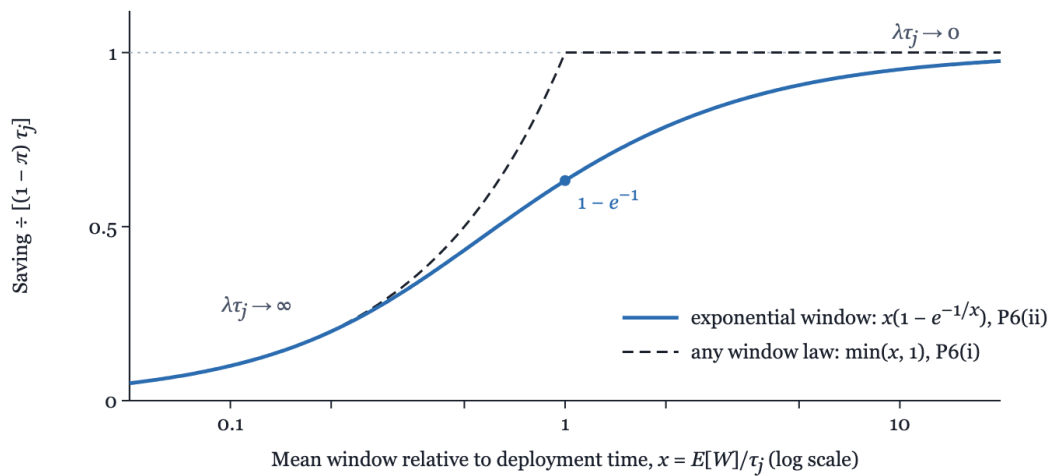


Figure 14. The head start is worth at most its length. The saving of a covered operator with deployment time τ_j , as a share of its baseline exposure τ_j and divided by $1 - \pi$, against $x = E[W]/\tau_j$: for an exponential window the normalized saving is $x(1 - e^{-1/x})$ (Proposition 6(ii)), below the distribution-free bound $\min(x, 1)$ (Proposition 6(i)), which also gives the two limits of Proposition 6(iii). Harm is linear in exposure, coverage is independent of the window, and attackers resume at unchanged intensity without backlog (A6–A9); the normalization requires $\pi < 1$, and the saving accrues only against attackers without parallel access. The curve is exact; where any operator lies on the horizontal axis is not measured, and benchmark catch-up does not locate it (Section 7.1).

Two further parts of the proposition carry the policy weight. In aggregate, if fast operators carry share α of baseline exposure, and every other operator has coverage at most $\bar{c}$, the share of total exposure a window can remove is at most $(1 - \pi)[\alpha + \bar{c}(1 - \alpha)]$, whatever the window’s distribution. And if harm is convex in exposure, “worth at most its length” fails in harm units, because the head start removes the most expensive slice of exposure; a bound in terms of the slope of harm at full exposure replaces it.

Who gains. Other things equal, fast patchers gain most in proportion. Whether they also gain most in total depends on coverage and baseline exposure, and that needs evidence. The holder’s partners are likely to be large firms that already patch fast, which is a judgment.

Benchmark catch-up, not a measured security window. The lab restricted Mythos Preview on 7 April 2026 (Anthropic 2026a, Introduction). GLM-5.3’s model release was 14 August; its weights became public two weeks later (CAISI 2026, opening paragraph). On the academic ExploitBench benchmark, Anthropic reported that GLM-5.3 developed end-to-end exploits in 50 of 410 attempts, against Mythos Preview’s 56 of 410 (Anthropic 2026d, “GLM-5.3 can develop working exploits end to end”). Separately, CAISI placed GLM-5.3 about four months behind the US frontier among released models CAISI evaluated, on an aggregate of its cyber benchmarks, with US safeguards disabled where applicable and with trusted-access US releases counted in the frontier (CAISI 2026, key findings);

Methodological Notes). Epoch’s 2024 estimate, a 90% interval of five to twenty-two months over thirteen benchmark lag values, also concerns benchmark catch-up (Cottier et al. 2024, “Open models have lagged on benchmarks by 5 to 22 months”). We draw no trend conclusion from it. These intervals identify neither literal copying, commercial rent erosion, attacker parity nor the duration or distribution of security windows. Security-window duration remains unverified. Disclosure deadlines concern different events from capability-window survival (Arora, Telang and Xu 2008, 643; Anthropic 2026b, “Our approach to discussing Mythos Preview’s findings”). Spectre’s coordinated-disclosure timeline, from Google’s report to the vendors on 1 June 2017 (Horn 2018, opening paragraphs) to a coordinated date of 9 January 2018 that early publication pre-empted (Linton and Parseghian 2018, opening paragraphs), likewise concerns a different event. Within Proposition 6, a window short relative to deployment removes at most a small share of a covered operator’s exposure, at most $(1 - \pi)/(\lambda\tau_j)$ for an exponential window; the evidence here does not locate that ratio.

7.2 Fixes are roadmaps, and holds protect unfound flaws

To protect outsiders from flaws the holder has found, it must ship fixes or mitigations. A public patch can guide attackers (Brumley et al. 2008, abstract). The under-one-hour browser result in Section 6.1 is a reported harness runtime on an existing patch, not an observed field exploit within an hour of publication; the patched browser release shipped eighteen days after that patch (“it would’ve been 18 days before the patched Firefox 148 was even released”; Anthropic 2026c, “N-days on Firefox”). Proposition 7 compares, flaw by flaw, a hold, in which attackers without access obtain the capability at W , with release, in which they obtain it at once (Figure 15).

- *Flaws the holder finds and patches for outsiders.* With the patch published at the same date under both regimes, and the patch-to-exploit lag the same with or without the capability, holding the whole capability adds nothing over a flaw-level embargo, if attackers would not have found the flaw themselves before the patch-enabled exploit. That is the independent-discovery exception: if, under release, attackers would find the flaw before the patch-enabled exploit, holding lowers exposure, weakly in general and strictly when some attackers lack access, the window is positive and attackers’ discovery comes before the operator’s deployment. If the lag shortens once attackers hold the capability, as the harness result in Section 6.1 suggests, a hold also protects flaws the holder has found and patched, for at most the window.
- *Flaws in outsiders’ code the holder never finds.* Holding delays attackers’ own search by the window. The saving is $(1 - \pi)\min(W, (\zeta_j - d_f)^+)$, at most $(1 - \pi)\min(W, \zeta_j)$, where ζ_j is when the operator’s normal cycle would remove the flaw and d_f is how long attackers with the capability take to find it: Proposition 6’s bound again.
- *Tools that only speed deployment.* Hold publication dates and the onset of exploitation fixed, and let the tool change only how fast operators deploy published patches. Then withholding it never lowers outsiders’ exposure, and it raises exposure strictly if and only if the tool would shorten deployment during a period of positive loss (Appendix A, Proposition 7).

Three readings follow. First, a capability-level hold does protect outsiders, through the flaws nobody has yet found, for at most the window. So the benchmark for any hold is not a flaw-level embargo but release with safeguards that keep the capability from attackers, a benchmark whose feasibility is open. Second, the deployment-only result needs its conditions. If the tool instead produces the patches, and withholding it delays both the onset and the end of exposure equally, undiscounted exposure is unchanged and discounted exposure falls. Third, a coordinated disclosure date need not hold: Spectre was published six days before its coordinated date “because of existing public reports and growing speculation” (Linton and Parseghian 2018, opening paragraphs). The disclosure trade-off is well studied (Arora, Telang and Xu 2008, 642; Ablon and Bogart 2017, ix). Actual safeguard durability needs evidence (Seger et al. 2023, 12; Kapoor et al. 2024, 6).

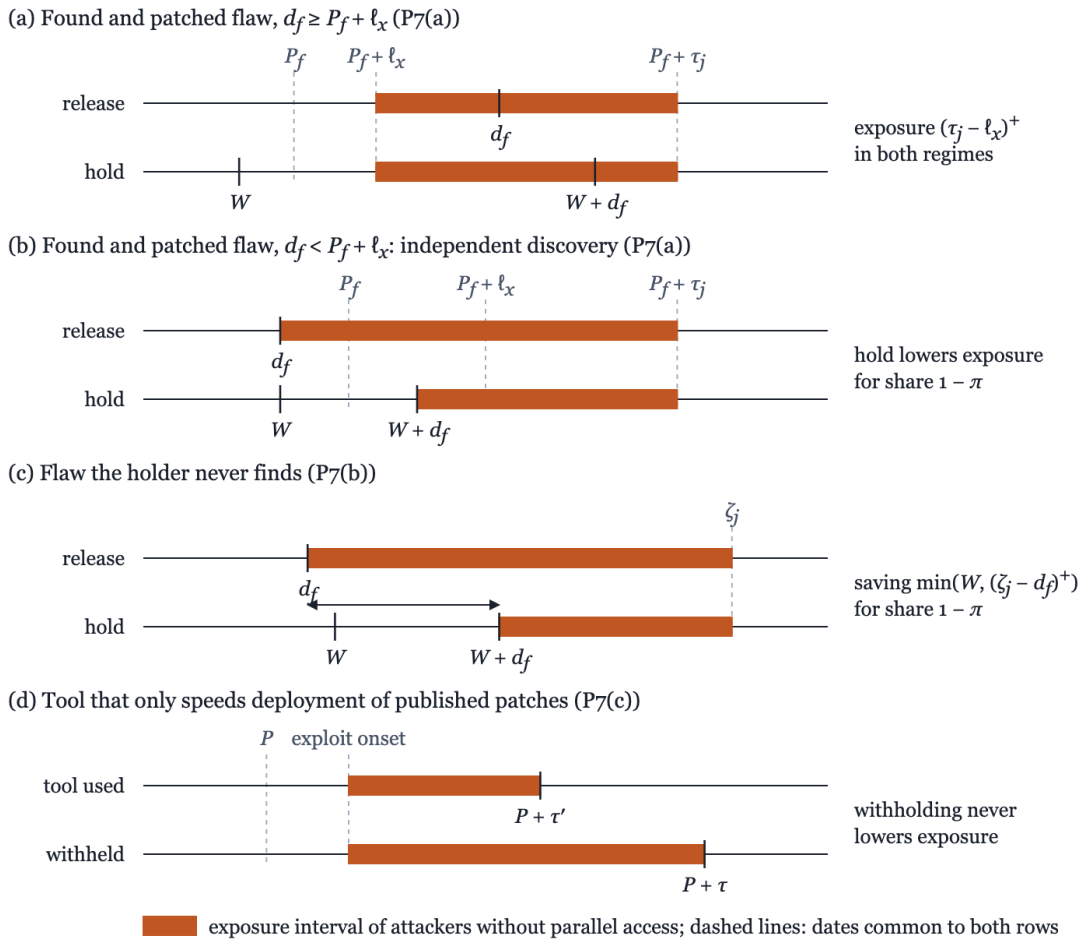


Figure 15. Fixes are roadmaps, and holds protect unfound flaws. Dated exposure intervals under release, in which attackers without access obtain the capability at once, and under a hold, in which they obtain it at W , for two flaw classes and for deployment-only tools (Proposition 7). (a) A flaw the holder finds and patches, with $d_f \geq P_f + \ell_x$: exposure is $(\tau_j - \ell_x)^+$ in both regimes, so a capability-level hold adds nothing over a flaw-level embargo. (b) The independent-discovery exception, $d_f < P_f + \ell_x$: the hold delays onset to $\min(P_f + \ell_x, W + d_f)$ and lowers exposure weakly, strictly when $\pi < 1$, $W > 0$ and $d_f < P_f + \tau_j$. (c) A

flaw the holder never finds: the hold saves $\min(W, (\zeta_j - d_f)^+)$, Proposition 6's bound again. (d) A tool that changes only how fast published patches are deployed: withholding it never lowers exposure, and raises it strictly if and only if the tool would shorten deployment during a period of positive loss. Panels (a)–(c) show attackers without parallel access, share $1 - \pi$; patch publication and the patch-to-exploit lag ℓ_x are the same under both regimes, and in (d) publication and exploit onset are fixed. If the lag shortens once attackers hold the capability, a hold also protects found and patched flaws, for at most the window (Section 7.2). Loss is linear in exposure; positions on the time axes are illustrative.

7.3 The coalition

Each partner admitted to a shared capability adds coverage and leak risk together. Proposition 8 shows that either corner can be best: sharing with no one, or with everyone eligible (Figure 16). An interior coalition is best if the first partner adds benefit and the last subtracts it. Admitting a rival raises the exposure saved if and only if it lowers the net hazard that ends the window: the rival-release hazard removed must exceed the leak hazard added. That rule holds only with coverage, deployment times, attackers' access and beneficiaries fixed, some attackers lacking access and covered exposure positive, and only if the rival can commit not to release (Appendix A, Proposition 8). If admission also covers the rival's own systems, the exposure saved can rise even while the net hazard rises (Appendix A, Section A.3.2). The structure resembles club theory, in which each member adds benefit and congestion (Buchanan 1965, 5, 7).

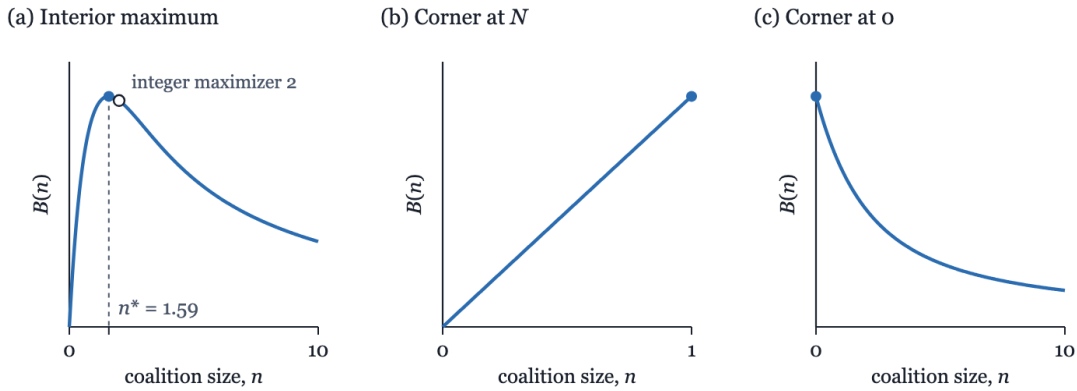


Figure 16. Either corner can be the best coalition. Exposure saved, $B(n)$, against coalition size in three examples from Section A.6: (a) coverage $1 - e^{-n}$, $\tau = 3$, $\lambda(n) = 0.5 + 0.3n$, $\pi = 0$ and $N = 10$, where $B'(0) > 0 > B'(N)$ and the maximum is interior at $n^* = 1.59$, with integer maximizer 2; (b) coverage n and $\lambda = 1 + 0.01n$ on $[0, 1]$, where B rises throughout and the largest coalition is best; (c) the holder's own systems covered at $n = 0$, partners adding coverage $0.01n$ and $\lambda = 1 + n$, where the empty coalition is best (Proposition 8(ii)–(iii)). The rule that admitting a rival raises B if and only if it lowers the net hazard ending the window holds only with coverage, deployment times, attackers' access and beneficiaries fixed, $\pi < 1$, covered exposure positive and a rival able to commit not to release (Proposition 8(iv), A13). Vertical scales differ across panels.

The 2024 RAND assessment judged protection of internet-connected models against the most capable actors infeasible at that time and recorded at least one known irreversible leak of a leading model (Nevo et al. 2024, vi, 1). Current applicability and completeness of that record remain unverified. A government retaining flaws for operations, as the US Vulnerabilities Equities Process permits (Nevo et al. 2024, 42), may give a defensive window offensive uses; that is a judgment.

7.4 Secrecy and the private race

A secret window is also a prize. Proposition 9 takes a two-player contest in which each player's chance of winning is its share of total effort and effort costs c per unit (Tullock 1980, 2001 reprint, 5), and total effort adds to an expected-loss index that every racer bears. The shared expected-loss index is our addition. Equilibrium effort is the gap between the winner's and the loser's payoffs, divided by four times the marginal cost of effort, that cost including the racer's own expected catastrophe loss. With baseline gap $v_W - v_L = D$, the positional stake, a window worth v_s to the winner alone multiplies equilibrium effort, and the expected-loss index, by $(D + v_s)/D$. A window both players receive leaves effort unchanged (Figure 17). Secrecy raises racing only by the winner–loser difference it creates (Appendix A, Proposition 9).

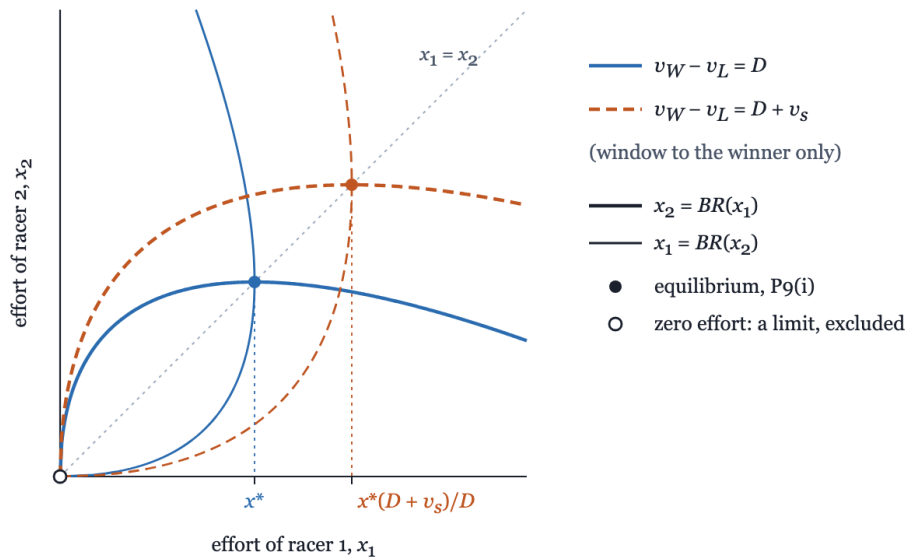


Figure 17. A winner-only window raises racing by the winner–loser gap it creates. Best responses of the two racers in the contest of Proposition 9, with payoff gap $v_W - v_L = D$ and with $D + v_s$ when a window worth v_s goes to the winner alone; equilibrium effort $x^* = (v_W - v_L)/(4(c + L\theta))$ is then multiplied by $(D + v_s)/D$ (Proposition 9(i)–(ii)). A window both players receive leaves the gap, and so effort, unchanged. Effort is in units of $D/(c + L\theta)$, and $v_s = D/2$ is illustrative. Against zero effort no best response exists (Proposition 10(iv)), so the origin is drawn open: zero is shown as a limit and excluded. Assumes A10: exponent one, linear cost and an expected-loss index linear in total effort. Whether the added racing is bad for the world depends on the welfare terms of Proposition 9(iii), which the figure does not show.

Whether that is bad for the world is a separate question, and the answer is not automatic. Welfare counts every party's real gains and losses in one unit; payments between parties are transfers and cancel. The welfare change from a winner-only window is the window's real gain to everyone except its holder, plus the holder's own real surplus, minus the cost of the extra racing net of any public value it produces. A private value larger than the world's gain does not by itself make secrecy harmful. And the claim that "even a modest escalation" by rivals outweighs the cyber gain is an empirical premise, not a theorem: it depends on the size of the far stake, the response, and the gain, in common units (Appendix A, Proposition 9).

How rivals respond to a suspected secret lead has no settled sign. One race model holds that more knowledge of capabilities makes the race more dangerous; in its public-information case the leading team takes risks when the runner-up is close, so on one reading a wide gap may calm the race (Armstrong, Bostrom and Shulman 2013, abstract, §§2.3, 3.3). Work on power shifts finds that visibility can keep the peace (Debs and Monteiro 2014, 1–2). Van den Steen's model of rational agents with differing priors shows how choosing an action can produce relative optimism about its success (Van den Steen 2004, 1141–1143). Applying that mechanism to each rival's assessment of its own approach's safety is our interpretation. A government partner that keeps flaws for operations may give rivals a preventive motive, a commitment problem of the kind bargaining theory identifies; that is a judgment (Fearon 1995, 381; Debs and Monteiro 2014, 1). One reported observation: Anthropic wrote that every similarly capable model other than GLM-5.3 had been "released with safeguards or through limited access programs" (Anthropic 2026d, "What does this mean?"). That is compatible with norm-following, independent risk management or commercial choices; one lab's statement identifies neither motive nor escalation response.

The answer to the third question remains conditional. Privately, racing to the capability and holding it may be rational; that is a judgment, since Proposition 9 shows only that a winner-only window raises equilibrium effort. For the world, a hold helps covered operators according to window length relative to deployment, and delays attackers' search for unfound flaws for at most the window. Benchmark catch-up has not established the security window's duration. A release with safeguards that keep attackers out for as long as a hold would, delays the same attackers and gives every defender the capability. Such a release, a benchmark whose feasibility is open, is a case of what Shevlane (2022, 1) calls structured capability access: instead of disseminating a system openly, the developer offers "controlled, arm's length interactions" with it, 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. Appendix A does not model it. Appendix B, Proposition E2, compares it with holding: with defensive schedules unchanged, release weakly raises outsiders' exposure, by the chance its safeguards fail times the exposure that failure adds, and it lowers exposure only if defenders' wider access improves their schedules by more than that. E2 sets aside hazards that release itself induces; an added attack-enabling hazard from serving the model would raise exposure further, weakly, at a fixed schedule (Appendix B, B-S6). Real safeguard durability and window length remain unverified.

8. One capability, four uses: the commercial edges, what leaks, and the race

8.1 Four uses

A firm holding a frontier capability can earn from it in four places.

Use	What competition is about	What it earns while held
The inference market (coding agents, chat, programming interfaces)	Better intelligence, cheaper, faster	Nothing more until released (the public model keeps earning). Once released, the lead is a wasting asset, eroded as rivals catch up
Defense, with partners	Staying ahead of attackers, who gain from the same technology	Security for the defended set, while the window lasts
Products in other markets (voice, video, agents, drug and materials discovery, design)	The product, not the model	Revenue from products competitors cannot yet make
Internal research (using the model to build the next one)	The next capability step	A faster route up the capability curve, hard to observe from outside

What leaks depends on the use, not on secrecy as such. The magnitudes in the third column are hypotheses that need evidence.

Use	What escapes	How much of the security-relevant capability escapes (judgment; needs evidence)
Release into the inference market, without effective safeguards	The capability itself, to anyone with access; copying from it can begin	Most of it, to the extent safeguards are absent or fail; how fast copies follow is unmeasured
Sharing with defense partners	Access for more people, who become insiders and theft targets; the flaws, through patches	High, and likely rising with each partner
An interactive product (an agent or a coding service)	The capability in that domain, through queries and training on outputs	High where the domain overlaps code and logic, if outputs can train a copy
An artifact product (a drug candidate, a voice model, a design)	The output, not the engine; and a signal that the engine exists	Low, especially far from code, unless the artifact can be reverse-engineered into

Use	What escapes	How much of the security-relevant capability escapes (judgment; needs evidence)
		the capability
Internal research	Little directly; possibly through faster later releases, disclosures or insiders	Small directly; its indirect effects are unmeasured

Copying from outputs is a real channel: distillation trains a student model on a teacher's outputs (Hinton, Vinyals and Dean 2015, 1–2). Copying through queries to a served model is a further channel whose size needs evidence (Tramèr et al. 2016, abstract; Carlini et al. 2024, abstract). How much of a frontier capability imitation actually copies is the magnitude this table needs (Gudibande et al. 2023, 1).

Three consequences follow, each a hypothesis to test.

1. **Artifacts may let a firm profit while holding.** Selling artifacts in markets far from code may cost little of the security edge, so firms with rich artifact markets could hold longer, and release pressure from the inference market would weaken. In the model, at an interior optimum, richer artifact sales delay release only if the loss hazard they add, theft and existence signals included, is small against what a leak destroys (Section 8.3). Sharing with partners likely has no such low-leak path, which is a judgment that needs evidence.
2. **Using an edge may reveal it, and value may attract theft.** as using a cyber resource tends to (Axelrod and Iliev 2014, 1298), **and value may attract theft.** A firm that suddenly ships a far better voice model, or finds drug candidates at an abnormal rate, may tell rivals that it holds something, and knowing that something can be done tells rivals where to look (Anton and Yao 2004, 1–2). The more a held model earns, the more it may be worth stealing. Both are hypotheses. We have found no documented case of a capability replicated soon after a public demonstration, and no measure of how theft scales with a secret's value.
3. **Artifacts may turn a temporary lead into durable assets:** patents, designs and data (Teece 1986, 285; Levin et al. 1987, 784), and locked-in customers through switching costs (Klemperer 1995, 515). This too is an empirical hypothesis. Section 8.4 shows what follows if it holds.

8.2 The holder's program

The holder's public model leads the best copy by a gap $g(t)$, which erodes at copy rate κ_{pre} before release and earns inference rent $\rho(g)$. Holding earns the flow F as well. Releasing at date T raises the lead by the step Δ , which then erodes at κ_{post} , the copy rate after release; release ends the held flow. The hold can also end involuntarily, at hazard λ , through open-weights catch-up, a rival's release, a partner's leak, theft, an existence signal from artifacts that leads rivals to the capability, or copying through products. The holder then keeps only a continuation value Z , from durable assets or nothing. All flows are discounted at $r > 0$.

8.3 Release timing on two clocks

Proposition 11 solves this as a stopping problem. An interior release date satisfies

$$F + \lambda Z + \rho(g(T)) = (r + \lambda)J(T) - J'(T),$$

where $J(T)$ is the value of releasing at T . The left side is what holding one more instant earns: the held flow, the hazard-weighted continuation value, and the rent on the current lead. The right side is what waiting costs: the forgone return on release, and the fall in release value as the current lead erodes. The date is the global optimum if the difference between the two sides crosses zero once, from above; otherwise all critical dates and both ends must be compared (Appendix A, Proposition 11). This is standard real-options reasoning (Dixit and Pindyck 1994, ch. 4, §1.D; Weeds 2000, §1), and it resembles the cannibalization problem in the timing of new products (Moorthy and Png 1992, 345).

Two clocks. The directions below are local (Figure 18): they hold at an interior optimum where that difference crosses zero simply, from above, with other primitives fixed. A change can instead move the optimum to a corner or to another crossing.

- Faster copying of the holder's current public model, κ_{pre} , wears down its inference revenue, so holding costs more every month. Release comes sooner.
- Faster expected copying of the model it would release, κ_{post} , lowers what release is worth, because the restored lead will itself erode. For linear revenue curves, and for power-law curves of every exponent, release comes later. More generally, at such a crossing and with the loss hazard fixed, release comes later whenever $x\rho'(x)$ is non-decreasing in the lead x ; linear and power-law curves both qualify. Under a smooth concave curve that saturates, such as $\rho = 1 - e^{-4x}$, it can come sooner. The exhibited saturating reversal is evaluated at $\kappa_{\text{post}} = 0$; this does not mean reversals occur only there.
- Internal research that adds no hazard delays release, if internal research's value ends at release, as assumed in Section 8.2. Richer artifact sales delay it if and only if the hazard they add, times what a leak destroys ($J - Z$), is less than one per unit of flow. A higher loss hazard brings release sooner when release is worth more than what survives a loss ($J > Z$), as the derivative in the proof of Proposition 11 shows (Appendix A).

With a single copy rate for both phases, the two channels oppose each other and the net sign is not determined. In the scans reported with the proofs, faster common copying delayed release in all 56 interior cases of a grid, and in 148 of the 161 selected draws of a wider random scan, all of them interior, against 13 in which it brought release sooner. Those are frequencies over draws selected by a stated rule, not a result.

Corners are common. With linear revenue curves, a strict interior release requires the released model to be copied faster than the current one by more than the hazard of losing the secret, $\kappa_{\text{post}} > \kappa_{\text{pre}} + \lambda$. That is necessary, not sufficient. Whether released models are in fact copied that much faster is unmeasured; the corner result below needs no such ordering. Below that rate a corner is optimal: release at once or hold until the secret is lost.

At equality, every date can be optimal. With constant flows and no public model to lose, the decision is all-or-nothing, with indifference among all dates only at the boundary; faster post-release copying pushes toward holding. Concave and convex curves can also give an interior release (Appendix A, Proposition 11).

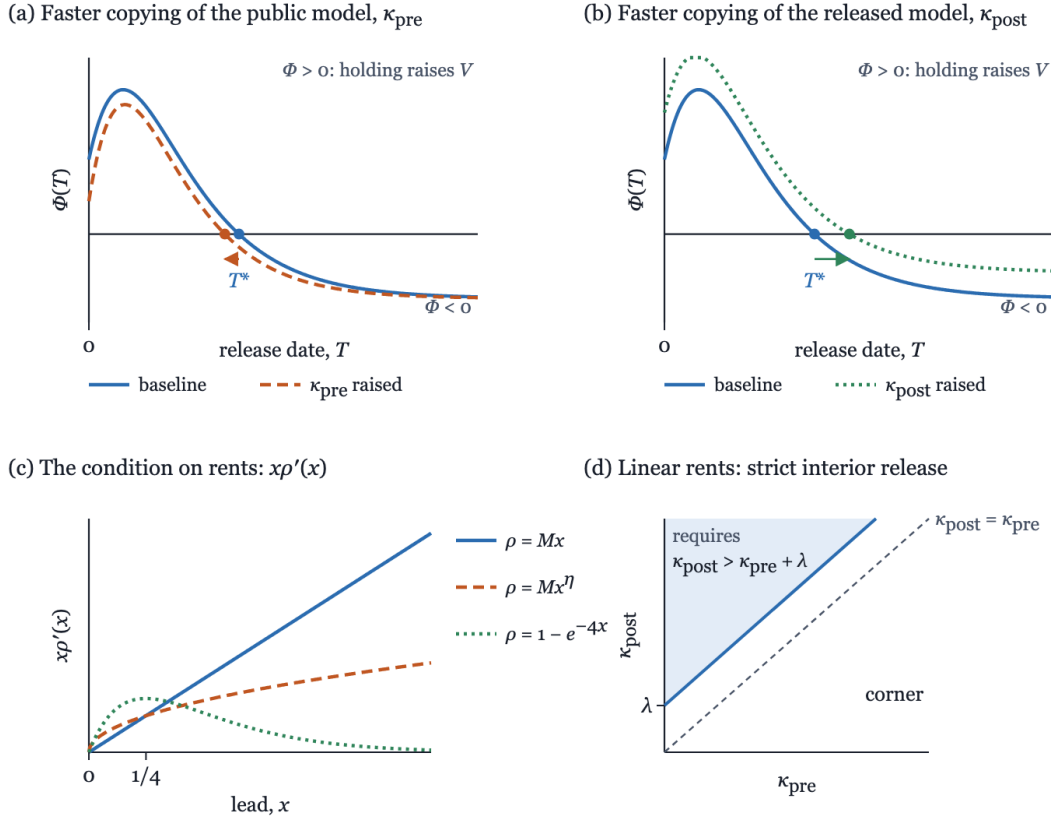


Figure 18. Release timing on two copy clocks. (a), (b) The stopping function $\Phi(T)$ of Proposition 11 for the interior example of Section A.6 (power rents $\rho = Mx^\eta$ with $\eta = 0.5$, $M = 1$, $F = 0.42$, $\Delta = 0.3$, $\lambda = 0.2$, $g_0 = 0.2$, $r = 0.05$, $Z = 0$ and $\kappa_{pre} = \kappa_{post} = 0.5$), whose downward crossing gives $T^* = 9.65$; raising κ_{pre} to 0.55 moves the crossing earlier, and raising κ_{post} to 0.52, with λ fixed, moves it later (Proposition 11(v)). (c) The condition behind the second direction: release comes later with faster post-release copying whenever $x\rho'(x)$ is non-decreasing, as for linear and power rents; for the saturating $\rho = 1 - e^{-4x}$ it falls beyond $x = 1/4$, and release can come sooner; the exhibited reversal is evaluated at $\kappa_{post} = 0$, which does not mean reversals occur only there. (d) With linear rents, a strict interior release requires $\kappa_{post} > \kappa_{pre} + \lambda$, a necessary but not sufficient condition; below that line, which includes a single copy rate for both phases, a corner is optimal (Proposition 11(iii)). With one copy rate under other rents the net sign is not determined. Directions are local: they hold at a simple downward interior crossing with other primitives fixed, and a change can move the optimum to a corner or another crossing. Whether released models are copied faster than current ones is unmeasured; the perturbations are illustrative and the vertical scale of (c) is arbitrary (Section 8.3).

What this means. The inference market does not, in general, force held capabilities out on a cadence set by copy speed. The commercial pressure to release is real, and it comes from the erosion of the current public lead. But the prospect that a new release will itself be copied quickly is, for linear and power-law revenue curves, a reason to hold. That reason holds the loss hazard fixed. At a simple downward interior crossing, if the hazard also rises with post-release copying, release comes later if and only if $A_\kappa > \lambda'(J(T^*) - Z)$, where $A_\kappa = \partial\tilde{\Phi}/\partial\kappa_{\text{post}}$ is the push toward holding at fixed λ and $\lambda' = \partial\lambda/\partial\kappa_{\text{post}}$; a parameter change can instead move the optimum to a corner or to another crossing (Appendix A, Section A.7). Which clock dominates in practice is an empirical question, and Section 11 names the test. In Proposition 11, where release makes the capability public, the realized security window is the shorter of the involuntary-loss time and the holder's release date, by definition, and Proposition 6's bound applies to it (Section 4.1). A closed release whose safeguards work does not end the attacker-access window (Appendix B, Proposition E1). A higher loss hazard shortens the first. Linking that hazard to the commercial copy rates is an assumption we do not make; the condition above states only what would follow if it held.

The commercial evidence in hand is thin. The four-month case and Epoch's range (Section 7.1) are benchmark catch-up evidence. They identify neither κ_{pre} nor κ_{post} , nor how fast commercial rent erodes. Prices per unit of capability (Cottier et al. 2025, summary), and customers' switching and retention in the market for intelligence (Demirer et al. 2025, 2–3), are the measurements the two clocks need.

Open- versus closed-weight copies. Copies reach the market in two forms: an open-weight copy publishes the weights, and a closed-weight copy serves a comparable capability behind an interface. In Proposition 11 a rival's release ends the hold whatever its form: open-weight catch-up and a closed rival's release are both loss events. Appendix B extends the holding model to draw the distinction, keeping the events that can end a hold apart from the holder's own choice. The events are a rival's comparable open-weight release, a rival's closed release whose safeguards fail, a rival's closed release whose safeguards work, theft of the weights, and a partner's leak. Each ends the holder's commercial lead; each except the working closed release also gives excluded attackers lasting capability. The holder chooses among committed holding, with no later release, immediate open release and immediate closed release. Under committed holding only, the commercial window lasts until the first of these events and the attacker-access window until the first attack-enabling one (Figure 19); under Appendix B's independent, homogeneous Poisson arrival streams, which continue after safeguarded releases, with each safeguard failure an independent draw, both are exponential. Immediate open release closes both windows at once. Immediate closed release closes the commercial window at once and leaves the attacker-access window as under holding if its safeguards work, or closes it if they fail. In every case the commercial window is no longer than the attacker-access window (Appendix B, Proposition E1). That ordering rests on a premise: the modeled thefts and leaks also supply a commercial substitute at once. Covert theft that armed attackers alone would require revising it.

Proposition E2 compares immediate closed release with committed holding from the same state, with the same future arrival streams; whether the holder's safeguards fail is

independent of those streams. With p_h the chance that the holder's safeguards fail, outsiders' expected exposure changes by $p_h S_H - \Gamma_{\text{def}}$. Here $S_H \geq 0$ is the added exposure if excluded attackers arrive at once rather than at the first attack-enabling event under the holding schedule, and Γ_{def} is the reduction from any explicitly modeled change in defensive schedules. With schedules and future access hazards unchanged, $\Gamma_{\text{def}} = 0$, and release weakly raises exposure, strictly so exactly when $p_h S_H > 0$. An added release-induced attack-enabling hazard, which E2 sets aside, would raise exposure further, weakly, at a fixed schedule (Appendix B, B-S6). Exposure here is outsiders' weighted, undiscounted exposure-time, as in Section 7, not welfare, and the comparison assumes that release changes no theft, distillation or rival-access hazard, and that defensive schedules do not depend on which events occur or on whether safeguards fail. Proposition E3 bounds what levers on covered domestic developers can do: if the attack-enabling intensity from unbound developers, theft and leaks, λ_U , is positive and unchanged, the expected attacker-access window cannot exceed $1/\lambda_U$ under any initial action studied. The extension sets aside the holder's payoff, so nothing in it carries over to Propositions 11–15. These results are not a welfare comparative static and not a policy ranking.

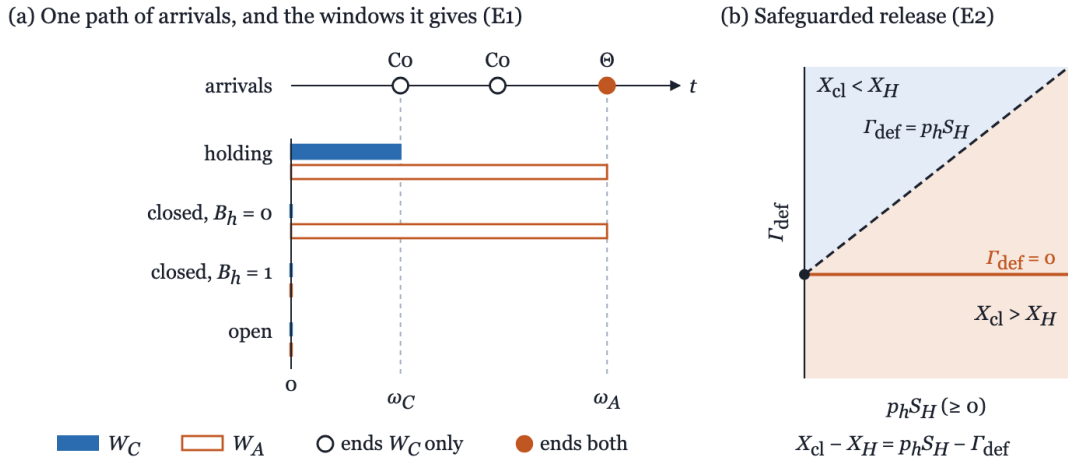


Figure 19. Two windows, and what safeguarded release does to them (Appendix B). (a) One illustrative path of the independent arrival streams: two rival closed releases whose safeguards work (C0) end the commercial lead but not attacker exclusion, and a theft (Θ) ends both, so $\omega_C \leq \omega_A$ on every path. The bars give the realized windows W_C and W_A under committed holding, immediate closed release with working ($B_h = 0$) or failed ($B_h = 1$) safeguards, and immediate open release, as in Proposition E1; zero-length windows are ticks at 0. (b) Proposition E2's comparison of immediate closed release with committed holding from the same state and arrival streams: $X_{cl} - X_H = p_h S_H - \Gamma_{\text{def}}$ with $S_H \geq 0$. With defensive schedules and future access hazards unchanged, $\Gamma_{\text{def}} = 0$, and release weakly raises outsiders' exposure, strictly when $p_h S_H > 0$; it lowers exposure only if an explicitly modeled defensive improvement exceeds $p_h S_H$. Exposure is outsiders' weighted, undiscounted exposure-time, not welfare. E2 sets aside release-induced hazards (B-S6), which would raise exposure further at a fixed schedule, and the window ordering assumes that thefts and leaks also supply a commercial substitute (B-S5). The holder's safeguard outcome B_h is independent of all primitive streams, which are the same under the actions compared; each defensive schedule is

independent of the complete primitive streams and B_h , and expected total exposure is finite. The extension sets aside the holder's payoff and ranks no interventions.

We judge, without evidence yet, that an open-weight copy erodes the commercial lead faster and further, since anyone able to host it has an outside option near the cost of compute. Whether a given open-weight release is comparable, matching the held capability in the security-relevant domain, which a benchmark lag does not show, is an empirical question. Its safeguards can be stripped (Lermen, Rogers-Smith and Ladish 2024, 1), and “abliterated versions” of the 2026 open-weight model appeared “within days” of its release (Anthropic 2026d, “GLM-5.3 lacks robust safeguards”). It cannot be recalled. A closed rival can monitor, rate-limit and recall, but these lie outside the model, which counts only lasting, usable attacker capability. Open-weight releasers may also answer to adoption, ecosystem or national strategy rather than inference margins, so their clock need not follow the holder's commercial logic; that too is a judgment. The 2026 lag was to an open-weight model; the lag to the nearest closed rival is still to be measured.

8.4 What to sell, what a secret is worth, and durable assets

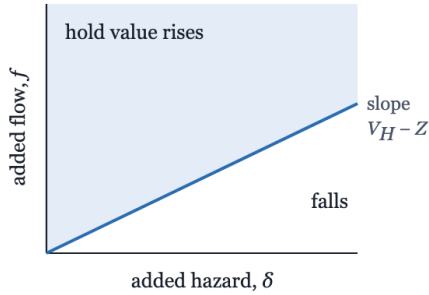
What to sell while holding. Proposition 13 evaluates a new use with constant flows, no voluntary release, and no continuation after loss, so that the hold is worth $F/(r + \lambda)$. Adding a use with flow f and added hazard δ , counting every hazard the use induces, raises that hold value if and only if $f(r + \lambda) > F\delta$: earnings per unit of added hazard must exceed what a leak destroys. With a continuation value Z , unchanged by the use, and $\delta > 0$, the threshold is $f/\delta > (F - rZ)/(r + \lambda)$: the hold value with continuation, $(F + \lambda Z)/(r + \lambda)$, minus Z . With $\delta = 0$, any $f > 0$ still passes. If theft responds to earnings, the added hazard includes it (Figure 20). But with a live release option, the hold value is the wrong opportunity cost: a use the holder commits to can pass the rule and still lower the firm's value, because the added hazard also destroys the release the holder was waiting for. An optional use could not (Appendix A, Proposition 13). The proposition evaluates given pairs of flow and hazard. It does not supply them, so nothing follows about whether interactive products or artifacts pass in practice.

What a secret is worth. Proposition 14 asks whether theft caps a secret's value. With hold value $F/(r + \lambda_0 + \lambda_\theta(F))$, where $\lambda_\theta(F)$ is the theft hazard at flow F , the value is bounded if and only if theft eventually rises at least in proportion to the flow. If theft is at least $\theta_F F$ at every flow, value stays below $1/\theta_F$, or $1/\theta_F + Z$ with durable assets. If theft rises less than in proportion, if it tracks the secret's capitalized value rather than its flow, or if security is cheap enough, the cap disappears (Appendix A, Proposition 14). The proportionality premise is ours. RAND ties the security a model needs to its capabilities and their stakes for “national security or competitiveness” (Nevo et al. 2024, vi), not to its earnings.

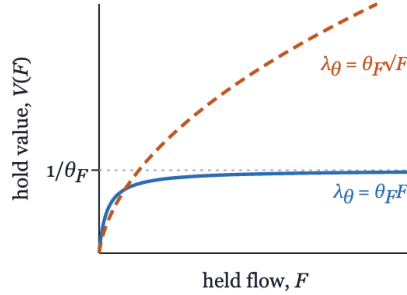
Perishable leads and durable assets. A unit flow eroding at rate κ is worth $1/(r + \kappa)$; one that does not erode is worth $1/r$ (Appendix A, Proposition 12). That is a present-value identity, the firm-value term of creative-destruction growth models (Aghion and Howitt 1990, §2). Faster copying raises a non-eroding prize's value relative to an eroding lead. It does not change the non-eroding prize's absolute value, and nothing about how hard firms pursue it follows. Whether a decisive, non-eroding prize exists is a premise.

If a share d of each step's earnings is, from the start, a perpetual component that copying cannot erode, a unit of step earnings is worth $(1 - d)/(r + \kappa) + d/r$, a valuation identity (Appendix A, Proposition 15). The durable share is assumed, not modeled. When copying erodes leads ($\kappa > 0$), it cuts two ways. It lowers the decisive prize's value relative to a step's. And it raises the value of being first at each step, which raises step racing (compare the escape-competition motive between “neck-and-neck” rivals in Aghion et al. 2001, abstract), if durable rents go to the winner alone and the step stake excludes the decisive prize's continuation value. Whether firms then pursue the decisive prize less needs substitution between the two efforts. In one specification, two contests sharing a quadratic effort cost, with racers bearing no catastrophe loss themselves, decisive effort falls as the step stake rises, and total risk rises with it if and only if $\theta_d/\theta_s < 2 + D_s/D_d$, where the θ s are external weights converting each contest's effort into an expected-loss index and the D s are the two prizes. With separate budgets, decisive effort does not move. These results are derived here; the closest literature treats a single contest won through several activities (Arbatskaya and Mialon 2010, 23). Which effect dominates in practice is empirical. The evidence to gather is frontier labs' patenting and product launches outside the inference market, timed against their capability jumps. Patents, preempted assets and buyer switching costs are channels by which a lead may become durable (Lieberman and Montgomery 1988, 41–42; Klemperer 1995, 515).

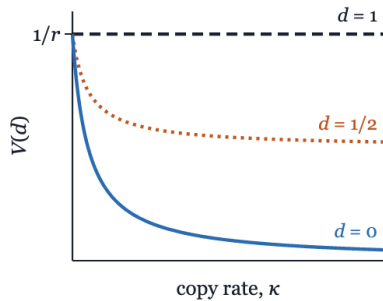
(a) Adding a use while holding (P13)



(b) Value of a secret under theft (P14)



(c) A unit of step earnings (P12, P15)



(d) Step and decisive contests (P15(c))

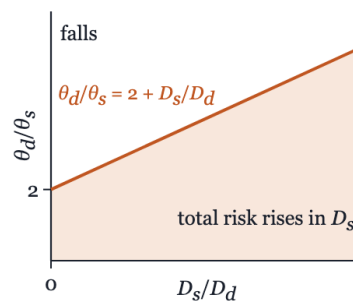


Figure 20. What to sell while holding, what a secret is worth, and durable assets (Section 8.4). (a) A use with flow f and added hazard $\delta > 0$ raises the restricted hold value if and only if $f/\delta > (F - rZ)/(r + \lambda) = V_H - Z$ (Proposition 13(i)), with constant flows, no voluntary release and a continuation value Z unchanged by the use; with a live release option, a

committed use can pass this rule and still lower the firm's value. Panel (a) illustrates $V_H > Z$, equivalently $F > rZ$; if $V_H \leq Z$, every positive added flow passes for $\delta > 0$. (b) Hold value $V(F) = F/(r + \lambda_0 + \lambda_\theta(F))$ under theft proportional to the flow, bounded by $1/\theta_F$, and under theft proportional to $\sqrt{F}$, unbounded (Proposition 14(i)–(ii)), at the Section A.6 values $r = 0.05$, $\lambda_0 = 0.1$ and $\theta_F = 0.5$; the proportionality premise is the paper's assumption. (c) Value of a unit of step earnings, $V(d) = (1 - d)/(r + \kappa) + d/r$, for durable shares $d = 0, 1/2$ and 1 at $r = 0.05$ (Propositions 12 and 15); the durable share is assumed, not modeled. (d) In the specification with two contests sharing a quadratic effort cost, racers bearing no catastrophe loss, total risk rises with the step stake if and only if $\theta_a/\theta_s < 2 + D_s/D_a$ (Proposition 15(c)); with separate budgets decisive effort does not move. Which effect dominates in practice is empirical.

8.5 Internal research, and what the commercial side does to the race

Internal use of a held model to build the next model serves both the commercial and the security motives, and it is hard to observe from outside (Stix et al. 2025, 5). That makes it a possible channel by which the general race could accelerate with little visible sign. Whether it does is a hypothesis that needs evidence; its racing effect is not modeled here. It also deepens a signaling problem: outsiders cannot easily tell holding for security from holding for commercial advantage.

Three things may push the general race: the inference market, through the erosion of each firm's current public lead; internal research use; and the strategic and supply-push channels taken up in Section 9. These are proposed channels, a judgment that needs evidence, not modeled racing effects. In the contest of Proposition 9, with costs fixed, racing rises only with a change in what winning is worth, and none of the three has been tied to that quantity.

The answer to the fourth question is therefore conditional on the capability being held. The reviewed examples do not establish that the security motive creates a separate narrow race. For a held general capability, the commercial motive does not simply erode the reason to hold: the two copy clocks pull in opposite directions, and private uses may let firms hold longer while earning. The commercial channel is one of several that may push the general race; its weight remains empirical.

9. Why actors run anyway

If acceleration does not lower the near peak where adoption binds, why do actors race? Our contest model answers in four parts (Appendix A, Propositions 9 and 10).

First, a finite prize bounds worthwhile effort when cost is linear and positive: with prize D and effort cost cx , any $x > D/c$ is strictly worse than zero if the effort changes no other benefit. Merely rising cost is insufficient (Figure 21). This bound excludes large efforts, not every possible dominant strategy under every contest rule. In the specified Tullock game, the best response varies with positive rival effort, so no effort is dominant (Appendix A, Proposition 10(iv)).

Second, a race survives in that Tullock game; with no public value from effort it wastes resources. Each racer bears only its own catastrophe loss from total racing. Starting at the symmetric equilibrium x^* and holding the other primitives fixed, a unilateral cap $0 < \bar{x} < x^*$ lowers the rival's best-response effort. The capper wins less often, but total effort falls. The zero cap is excluded: against zero, the rival has no smallest positive best response (Appendix A, Proposition 10(iv)).

Third, Proposition 10 states the result for the racers' joint payoff exactly. With a shared expected-loss index and a positional stake D , equilibrium effort is positive while the racers' joint optimum is zero. With quadratic effort costs and a public value b per unit of total effort, the equilibrium over-races if and only if $D > 8\max(b - L\theta, 0)^2/k$, where $L\theta$ is the catastrophe cost per unit of effort and k the cost parameter. It under-races if racing's uncaptured public value is large enough and the positional stake small. So racing exceeds the racers' joint optimum unless it has enough uncaptured public value (Appendix A, Proposition 10).

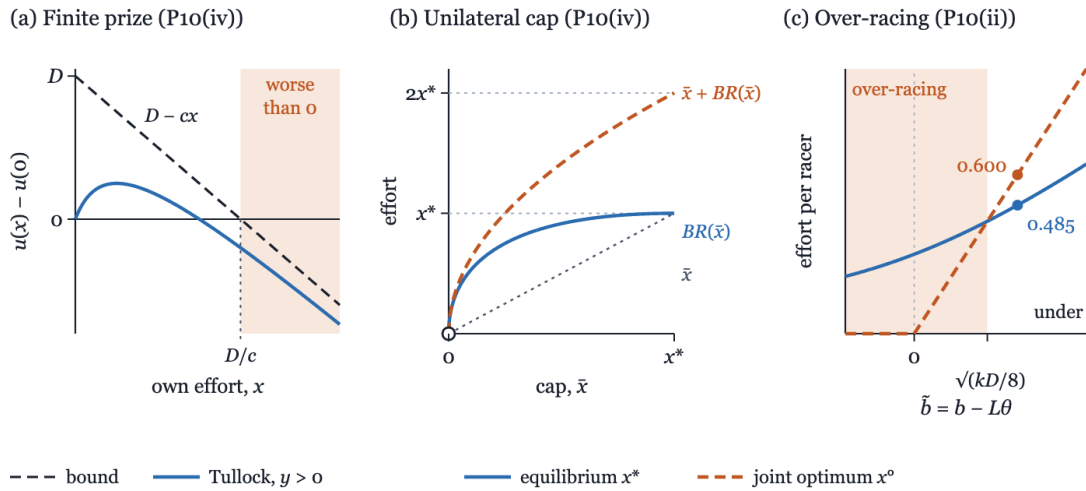


Figure 21. A finite prize still sustains a race (Section 9). (a) With prize D , linear cost cx with $c > 0$ and a loss term that does not fall with own effort, $u(x) - u(0) \leq D - cx$, so every $x > D/c$ is strictly worse than $x = 0$ (Proposition 10(iv)); the solid curve is the Tullock payoff gain against a positive rival effort y , at illustrative values. (b) In the Tullock stage game, a unilateral cap $0 < \bar{x} < x^*$, with the rival at its best response, gives $\bar{x} < BR(\bar{x}) < x^*$ and total effort $\bar{x} + BR(\bar{x}) = \sqrt{\Delta v \cdot \bar{x} / \hat{c}} < 2x^*$, rising in $\bar{x}$, with $\hat{c} = c + L\theta$; the capper wins with probability $\sqrt{\hat{c}\bar{x}/\Delta v} < 1/2$ (Proposition 10(iv)). The zero cap is excluded: against zero effort the rival has no best response, so the origin of (b) is drawn open, as a limit. Effort is in units of $\Delta v / \hat{c}$. (c) With quadratic cost $kx^2/2$ and public value b per unit of total effort, equilibrium effort x^* exceeds the racers' joint optimum x^o if and only if $D > 8\max(\tilde{b}, 0)^2/k$, with $\tilde{b} = b - L\theta$ (Proposition 10(ii)); shown for the Section A.6 example $k = 1$, $D = 0.36$, where at $\tilde{b} = 0.3$ the equilibrium effort 0.485 falls short of the optimum 0.600 . Over-racing is unconditional only without public value; the welfare criterion is the racers' joint payoff, and the results assume one functional form (A10).

Fourth, under Proposition 10's specified detection and permanent-punishment protocol, restraint requires sufficient detection (Figure 22). In the repeated race, with discount factor ϱ per period, detection of each deviation with probability v , no false alarms, and permanent reversion to racing after detection, restraint is an equilibrium if and only if v reaches the threshold $((1 - \varrho)/\varrho) \cdot 2(c + L\theta)/(c + 2L\theta)$. When that threshold exceeds one, as it does for impatient racers, no detection technology supports restraint, and under full secrecy restraint fails. In a one-shot race, restraint holds if and only if the expected sanction on detection is at least half the stake (Appendix A, Proposition 10). This is the grim-trigger logic of repeated games with imperfect detection; the false-alarm case of imperfect monitoring is excluded here (Green and Porter 1984, 88; Coe and Vaynman 2020, 342).

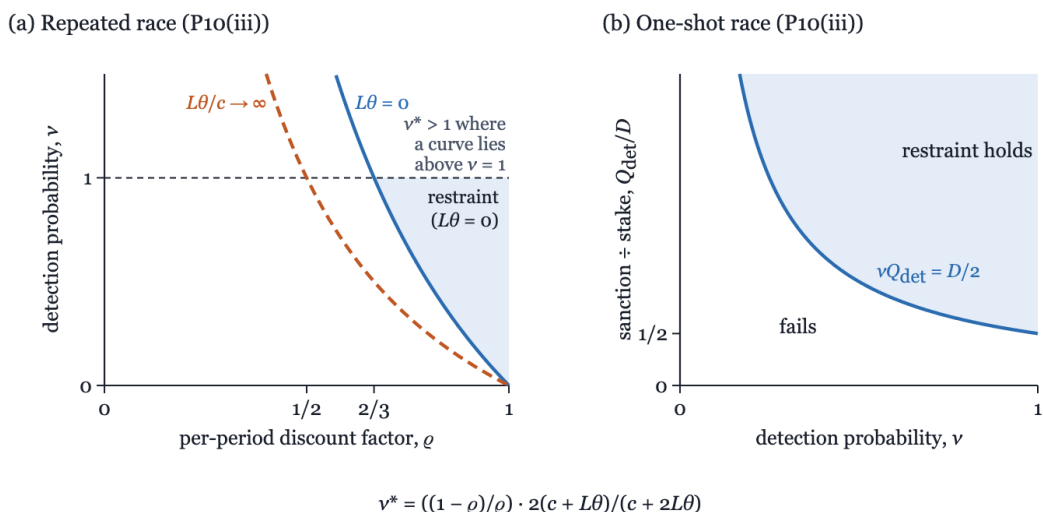


Figure 22. Restraint needs detection (Section 9). (a) In the repeated race, with per-period discount factor ϱ , detection of each deviation with probability v , no false alarms and permanent reversion to the stage equilibrium after detection, restraint is an equilibrium if and only if $v \geq v^* = ((1 - \varrho)/\varrho) \cdot 2(c + L\theta)/(c + 2L\theta)$ (Proposition 10(iii)). For $c, L\theta \geq 0$ the factor $2(c + L\theta)/(c + 2L\theta)$ lies between 1, as $L\theta/c \rightarrow \infty$, and 2, at $L\theta = 0$, so v^* lies between the two curves; it exceeds 1, and no detection technology supports restraint, for $\varrho < 1/2$ in every case and for $\varrho < 2/3$ when $L\theta = 0$. Under full secrecy, $v = 0$, restraint fails. The shaded region is the restraint region for $L\theta = 0$. (b) In a one-shot race, restraint holds if and only if the expected sanction on detection is at least half the stake, $vQ_{det} \geq D/2$. The false-alarm case of imperfect monitoring is excluded; the curve bounds are derived from the stated formula and are not measured values.

The strategic and supply-push channels sit alongside the commercial ones of Section 8. US policy frames AI as a strategic race: “The United States is in a race to achieve global dominance in artificial intelligence (AI)” (The White House 2025, 1). We judge that sunk capital and falling compute prices may push labs to ship. Neither channel has a measured share. In Proposition 9's contest, with effective effort costs fixed, a window raises racing only through the winner-loser payoff gap. METR reported that its pre-deployment evaluation of OpenAI's GPT-5 “took place over three weeks” (METR 2025, Appendix); without a comparable baseline that duration establishes neither shortening nor its cause.

10. Policy: pressures, levers, and unresolved comparisons

What the models settle. Three comparisons follow from the models without any estimate of magnitudes. First, if the aim is outsiders' exposure, tools that only speed defenders' deployment of published patches should not be withheld: with publication and exploit onset fixed, withholding them never lowers that exposure (Proposition 7(c)). Second, a window short relative to an operator's deployment time is worth little to that operator: for an exponential window its saving, as a share of the operator's exposure, approaches $(1 - \pi)/(\lambda\tau_j)$ when deployment takes many windows' length, and it comes only against attackers without access (Proposition 6). Third, a capability-level hold should be judged against release with safeguards, not against a flaw-level embargo, because a hold also delays attackers' search for flaws nobody has found, for at most the window (Proposition 7(b)). The rest of this section needs magnitudes the evidence does not yet supply, and is stated conditionally.

10.1 Pressures on holding and releasing

Most pressures on holding work through five quantities in the holder's program: what the held capability earns (F); how fast the hold ends against the holder's will (λ); the erosion of the current public lead (κ_{pre}); what release restores and how fast that erodes (Δ and κ_{post}); and the holder's effective discount rate (r). Some directions follow from the model, locally, at an interior optimum. Others rest on reasoning alone and are marked as judgments (Figure 23).

Pressure	Quantity it moves	Pushes toward
Rivals and open-weights developers copying the current public model	κ_{pre} up	Release sooner (model)
Expected fast copying of the model that would be released	κ_{post} up	Hold longer, for linear and power-law curves, with the loss hazard fixed (model). At a simple downward interior crossing, if that hazard rises with κ_{post} , release comes later if and only if $A_\kappa > \lambda'(J(T^*) - Z)$; a parameter change can move the optimum to a corner or another crossing. A saturating curve can reverse it (model)
A rival may release something comparable first	λ_r up	Release sooner, when $J > Z$ (model)
Open-weights catch-up	λ_o up	Release sooner, when $J > Z$ (model)

Pressure	Quantity it moves	Pushes toward
Theft attracted by what the secret earns	$\lambda_\theta(F)$ rising in F	A higher theft hazard at fixed F brings release sooner when $J(T^*) > Z$; a higher F with induced theft delays release if and only if $\lambda'_\theta(F)[J(T^*) - Z] < 1$, other primitives fixed (model). Caps the restricted hold value only if theft rises at least in proportion to the flow (model)
Cash burn and investor pressure for revenue	r up	In the model, sooner for linear rents; either way for power rents, because r also discounts what release is worth (model, other primitives fixed). The judgment is that cash pressure works mainly on near-term payoffs
Low customer switching costs	Steep rent in the lead	Release sooner while the public lead erodes (judgment; needs evidence)
Rich artifact markets far from code; internal research	F up, little added hazard	Hold longer (model, if the hazard added is small, and for internal research if its value ends at release, as assumed in Section 8.2)
Liability and reputational risk from releasing an offensive-capable model	Release worth less, unless with safeguards	Hold, or release with safeguards (judgment)
Defense contracts during a hold	F up	If contract flow ends at release, an increase in contract flow f_σ delays release if and only if $(d\lambda/df_\sigma)[J(T^*) - Z] < 1$, other primitives fixed (model); added partners raise the leak hazard

Pressure	Moves	sooner	later	Condition
Copying of the current public model	$\kappa_{\text{pre}} \uparrow$	←		local
Expected copying of the released model	$\kappa_{\text{post}} \uparrow$		→	linear or power ρ , λ fixed; a saturating ρ can reverse it; if λ rises, later iff $A_{\kappa} > \lambda'(J(T^*) - Z)$
A rival may release first	$\lambda_r \uparrow$	←		when $J > Z$
Open-weights catch-up	$\lambda_o \uparrow$	←		when $J > Z$
Theft hazard at fixed F	$\lambda_{\theta} \uparrow$	←		when $J(T^*) > Z$
Higher F , with induced theft	$F \uparrow$		→	iff $\lambda_{\theta}'(F)[J(T^*) - Z] < 1$
Cash burn, investor pressure	$r \uparrow$	←		linear rents
	$r \uparrow$	←	→	power rents: either sign; judgment: cash pressure acts mainly on near-term payoffs
Low customer switching costs	steep rent in lead	←-----		judgment; needs evidence
Artifact markets far from code; internal research	$F \uparrow$		→	if the added hazard is small; research: if its value ends at release (Section 8.2)
Liability, reputational risk	release worth less		-----→	judgment: hold, or release with safeguards
Defense contracts during a hold	$f_{\sigma} \uparrow$		→	iff $(d\lambda/df_{\sigma})[J(T^*) - Z] < 1$, if contract flow ends at release; partners add leak hazard

→ model direction, local - - - → judgment vertical line: the holder's release date, T^*

Figure 23. Pressures on the holder's release date (Section 10.1). Each pressure, the quantity of the holder's program it moves, and the direction it pushes the release date T^* : solid arrows are directions from the model (Proposition 11; Appendix A, Section A.7), dashed arrows judgments, and the double arrow a sign that can go either way. Model directions are local: they hold at a simple downward interior crossing, other primitives fixed, and a parameter change can move the optimum to a corner or another crossing. Conditions are those of the table in Section 10.1; $A_{\kappa} = \partial \tilde{\Phi} / \partial \kappa_{\text{post}}$ is the push toward holding at fixed λ , and $\lambda' = \partial \lambda / \partial \kappa_{\text{post}}$. For liability and reputational risk the dashed arrow shows only the hold branch of the judgment "hold, or release with safeguards". No magnitude is measured, and the pressures are not ranked.

10.2 Levers, stated conditionally

Each judgment below names its objective and missing evidence. The results do not establish an ordering among feasible interventions (Figure 24).

If the objective is lower near-peak cyber harm, defenses needing no frontier capability are candidates: patch distribution, legacy retirement, memory-safe languages for new code, post-quantum and hybrid migration, hardware-bound authentication, and verified kernels and libraries. We judge these worth evaluating; their effects and costs require evidence.

Whether acceleration also helps depends on the current regime and the unestimated hazard and adoption responses, including $\beta > \varepsilon_H$ where adoption binds (Proposition 5).

If the objective includes health benefits from the same capability, the balance of Section 4.7 is the relevant objective, and two levers differ. Faster capability brings the benefit forward only while capability sets its arrival, and, with a speed-set hazard, it raises the cyber hazard. Speeding the benefit's own adoption clock, through trial and approval capacity, manufacturing and uptake, can bring the benefit forward without, on our assumed mapping, raising the cyber hazard; that such capacity lowers T_b without changing H is an assumption, not a result. A shorter T_b raises the benefit in level weakly, and strictly only when it brings the benefit's arrival forward within the period counted; under parallel timing, while capability sets arrival, it changes nothing. It does not raise the marginal return to capability speed in the undiscounted model: while the benefit waits on capability, that return, yK_B/s^2 , does not depend on T_b . With discounting, a shorter T_b raises it somewhat when part of the adoption waits for the capability ($m_B > 0$), because the benefit then arrives earlier and is discounted less (Appendix A, Section A.3.5). Choi, Johnson and Lehmann (2019, abstract and discussion) associate worse outcomes after breaches with remediation, which they did not directly observe. Which lever should come first needs values of y , V_C , K_B/s , T_b and m_B that we do not have.

If the objective is less exposure during a window, shortening deployment, widening coverage and extending attacker exclusion are candidate levers. Proposition 6 gives conditional savings, not their priority. A short window relative to deployment removes a small share of exposure, but its bound does not show which feasible change saves most per unit cost. Mandates, central update channels and deadlines are proposed ways to shorten deployment; their effectiveness and leakage effects need evidence.

If the objective is to protect outsiders from the flaws a holder finds, flaw-level embargoes with coordinated fixes already exist. A capability-level hold should be judged against release with safeguards, not against a flaw-level embargo, since a hold also delays attackers' search for flaws nobody has found, for at most the window (Proposition 7). We judge that common safeguard standards could make that comparator more credible; their costs and effects need evidence. Evidence on how well safeguards hold against stripping and distillation would decide the comparison.

If the objective is to avoid harming outsiders by holding, tools that only speed defenders' deployment of published patches, such as, in our judgment, rollout orchestration that schedules and tracks installation, should not be held back: with publication and exploit onset fixed, withholding them never lowers exposure (Proposition 7). If a tool also produces the patches, the conclusion needs its own analysis.

If the objective is joint restraint, under Proposition 10's detection and permanent-punishment protocol, it needs verification with detection at least at the threshold of Proposition 10, and a credible penalty. Coordinated release embargoes among labs, on the multi-vendor vulnerability-embargo model, would remove domestic preemption among committed members ($\lambda_r \rightarrow 0$). A shared defense program across labs would reach more operators per window, but each added member also adds leak hazard, which shortens the

window; Proposition 8 gives the condition under which admission still raises the exposure saved (Propositions 6, 8). Shared capability evaluations among rivals may calm or inflame the race; the sign is not settled.

Government levers, if the objective is longer useful windows and wider coverage, act on the same quantities; each direction below is a judgment from the model's mapping, and program data on hold lengths and coverage would change it. 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 preemption during the window; separate waits do not. EO 14409 §3 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 (Trump 2026, §3(b)(ii)); the provision does not establish implementation. Weight-security measures, such as RAND's levels, each defined as secure against "increasingly capable categories of malicious actors" (Nevo et al. 2024, v), lower the theft hazard; on the model's mapping they make longer holds viable. Government purchase of defensive access during a hold could steer coverage toward laggard operators, countering the tilt toward the holder's fast partners. Export controls on chips or weights lower open-weights catch-up only from the jurisdictions they cover, which is a judgment. Governance of internal use of unreleased models would make a hard-to-observe use visible to a regulator (Stix et al. 2025, 3). And because held capabilities may let a few firms make products others cannot, the concentration of downstream markets deserves watching (Korinek and Vipra 2024, 22–23). Whether government-purchased access reaches laggards, and how long holds last under each lever, are measurements still to make.

(a) Comparisons that need no estimate of magnitudes

Withhold tools that only speed deployment?	never lowers outsiders' exposure	P7(c)
Window short relative to deployment	share saved $\rightarrow (1 - \pi)/(\lambda\tau_j)$	P6
Benchmark for a capability-level hold	release with safeguards, not a flaw-level embargo	P7(b)

(b) Objectives and candidate levers, unranked

Lower near-peak cyber harm	☐ defenses needing no frontier capability: patch distribution, legacy retirement, memory-safe languages, post-quantum and hybrid migration, hardware-bound authentication, verified kernels and libraries ☒ acceleration helps where adoption binds only if $\beta > \varepsilon_H$ (P5)
Health benefits from the same capability	☐ speed the benefit's own adoption clock (assumed: lowers T_b , not H) ☒ faster capability: only while capability sets arrival (P16)
Less exposure during a window	☒ shorter deployment, wider coverage, longer exclusion (P6) ☐ mandates, central update channels, deadlines: proposed
Protect outsiders from flaws a holder finds	☒ judge a hold against release with safeguards (P7) ☐ common safeguard standards
Avoid harming outsiders by holding	☒ do not hold back deployment-only tools (P7(c)) ☐ example: rollout orchestration
Joint restraint	☒ detection at least v^* and a credible penalty (P10) ☒ shared defense program: admission condition (P8) ☐ coordinated release embargoes, $\lambda_r \rightarrow 0$; shared evaluations: sign open
Longer useful windows, wider coverage: government levers	☐ liability; common pre-release evaluation windows; weight security (lower λ_θ); purchased defensive access for laggards; export controls; governance of internal use; downstream concentration
☒ conditional result of a proposition ☐ judgment, proposal or assumed mapping; needs evidence	

Figure 24. Comparisons and levers, stated conditionally (Section 10). (a) The three comparisons that follow from the models without any estimate of magnitudes (Propositions 6, 7(b) and 7(c)); the second is the exponential-window limit for operators whose deployment takes many windows' length, and it applies only against attackers without access. (b) Each objective and the levers the paper names as candidates for it, in the paper's order. Filled squares mark conditional results of a proposition; open squares mark judgments, proposals or assumed mappings whose effects, costs and leakage need evidence. Both deployment-only statements hold publication dates and exploit onset fixed; the tool changes only how fast operators deploy published patches (Proposition 7(c)). Government levers are judgments from the model's mapping. The results establish no ordering among feasible interventions.

10.3 Three limits on all of it

1. **Coverage of the clock-setters.** Cooperation and domestic mandates reach covered developers, open or closed. They leave catch-up by developers outside their coverage untouched, though controls on inputs may still slow development. That catch-up may be the main way holds end; this needs evidence. If $\lambda_U > 0$ is unchanged, levers on covered domestic developers alone cannot raise the expected attacker-access window above $1/\lambda_U$ (Proposition E3); this ranks no interventions (Figure 25). The reported four-month benchmark lag does not identify a security-window duration. An open-weight copy's safeguards can be stripped and it cannot be recalled, so under Proposition E1 a comparable rival open-weight release ends

the attacker-access window. A domestic coalition that holds while outsiders release may hand them market share, which is a judgment that needs evidence.

2. **Coordination can look like collusion.** An agreement on release timing is, in form, an agreement to restrict output; that characterization is ours. Antitrust law exists “to prevent collusion between companies that has a negative effect on consumers” (Askill, Brundage and Hadfield 2019, 9), and the tension between AI governance and competition law is “currently underexplored” (Hua and Belfield 2023, §1). Whether safety coordination would need an explicit legal safe harbor, and whether existing practice in coordinated vulnerability disclosure already provides one, are unresolved legal questions that this paper does not answer.
3. **Comparing the levers.** Proposition 6 bounds the value of a window. It supplies no intervention ranking without comparative costs, effectiveness and leakage effects, none estimated here. Long deployment lags alone identify neither the current regime nor which feasible intervention should come first.

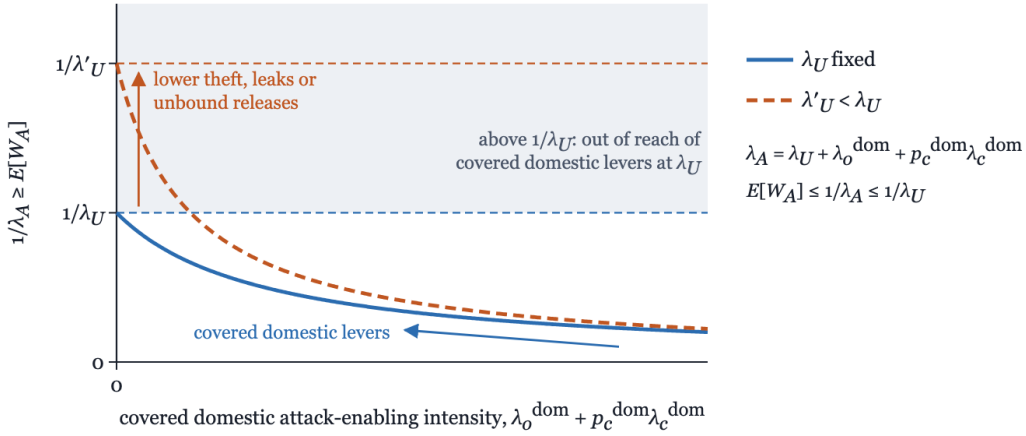


Figure 25. The reach limit of domestic levers (Appendix B, Proposition E3). The bound $1/\lambda_A$ on the expected attacker-access window against the attack-enabling intensity from covered domestic developers, $\lambda_o^{\text{dom}} + p_c^{\text{dom}} \lambda_c^{\text{dom}}$. With the attack-enabling intensity from unbound developers, theft and leaks, λ_U , positive and fixed, $E[W_A] \leq 1/\lambda_A \leq 1/\lambda_U$ for every setting of the covered domestic hazards and each initial action studied, so levers that reach only covered domestic developers cannot raise the expected window above $1/\lambda_U$. Less theft, fewer leaks or fewer unbound releases lower λ_U and raise the ceiling (dashed, $\lambda'_U = \lambda_U/2$, illustrative), without guaranteeing a longer realized window; such measures lie outside E3’s comparison. The bound ranks no interventions and does not claim that unbound catch-up is the main way holds end. Illustrative values.

A binding multilateral restraint on open-weights release of frontier cyber capability would affect a different part of the window hazard and could alter the comparison. Nothing in the sources we reviewed suggests one exists.

11. Limits, what would change the answer, and dated checks

Limits of the models. Each proposition holds under stated conditions, and the conditions do work. The loss index is a modeling choice, and the step term's independence from speed is an assumption. The peak result holds exploitation intensity per flaw constant. The window results take harm as linear in exposure, coverage as independent of the window, and attackers as resuming without a backlog. The holder's program assumes constant flows, exponential erosion of leads, constant hazard and smooth revenue curves; its comparative statics are local. The contest results assume one functional form. The four models answer to different objectives, and we have made no claim that combines them into one welfare measure. Proposition 16 puts health benefits and cyber harm in one unit, DALYs, for that extension only; it assumes a constant benefit flow, a fixed number of DALYs per harmful event and full coupling of the benefit to the capability that arms attackers, and it is a conditional balance rather than a welfare measure. Appendix A, Section A.7, and Appendix B, Section B.4, list assumptions a later note could relax, with any sign already known.

Limits of the evidence. Cited works were read in full, or in full for the part cited, and quotations from them checked. Primary-source checks are reported claim by claim in Supplement S2; no claim remains pending verification. Checking a lab's report establishes what the lab reported; independent replication of lab-reported figures remains unverified here. Much of the capability and restricted-release evidence comes from one lab's reports on its own model. Benchmark catch-up does not establish security-window duration, which remains unverified (Figure 26).

Evidence that would change the provisional judgments.

- Capability that shortens adoption itself, with adoption time falling with speed faster than hazard rises (Proposition 5).
- Windows held for years under verified multilateral restraint that binds open-weights developers.
- Commercial leads that rest mainly on things copying does not erode, such as distribution, enterprise trust and compliance, which would make release pressure weak.
- Release decisions that track the open-weights gap closely, which would suggest, without identifying it, that the earlier-release clock dominates in practice.

(a) Where each result stands

Assumed	loss index; step term independent of speed; exploitation intensity φ constant windows: harm linear in exposure, coverage independent of the window, no backlog holder: constant flows and hazard, exponential erosion, smooth rents; local statics contest: one functional form; P16: constant y , fixed V_C , full coupling
One lab's own reports	much of the capability and restricted-release evidence
Independent of the lab	DARPA AI Cyber Challenge; UK AI Security Institute (Section 6.1); not a replication of the lab's figures
Unverified	independent replication of lab-reported figures; security-window duration
Unmeasured	capability distance K ; adoption response β ; copy rates κ_{pre} , κ_{post} ; closed-weight copy lag; y , V_C , K_B/s , T_b , m_B
Provisional judgment	adoption binds, for critical federal legacy systems and web-protocol migration

(b) Evidence that would change the provisional judgments

1. adoption time falling with speed faster than the hazard rises	Proposition 5
2. windows held for years under verified multilateral restraint that binds open-weights developers	
3. commercial leads resting mainly on what copying does not erode: distribution, enterprise trust, compliance	weak release pressure
4. release decisions that track the open-weights gap closely	suggests, without identifying, that the earlier-release clock dominates

Figure 26. Where each result stands, and what would change the judgments (Section 11). (a) What the models assume; what rests on one lab's reports on its own model; what is independent of the lab, namely DARPA's AI Cyber Challenge and the UK AI Security Institute (Section 6.1), which point the same way but do not replicate the lab's figures; what remains unverified; what is unmeasured; and the provisional judgment that adoption binds for the populations the retirement and migration series cover (Section 6.2). (b) The four kinds of evidence that Section 11 names as able to change the provisional judgments, with what the paper says each would bear on. Each proposition holds under its stated conditions, and the four models are combined into no welfare measure.

Dated checks. The first three checks are predictions. Each names its window, the proposition or judgment it bears on, the direction we expect, a resolution rule fixed now (source, threshold and date), and the result that would count against the thesis (Figure 27). They are indicator forecasts: each resolves on a published count, and none by itself tests the mechanism of the proposition it bears on. The remaining checks are measurements: each names a window and the proposition it bears on, but we register no predicted value for it.

Predictions.

1. *Patches behind disclosures, to 31 December 2027 (the adoption time of Propositions 1 and 5, at its first step, patch creation, for open-source projects; Sections 6.1 and 6.5).* This is a monitoring indicator. We expect patch creation to keep lagging AI-scale disclosure. Resolution: on Anthropic's disclosure dashboard (Anthropic 2026e), the known-patched count divided by the disclosed count, on the last reading dated on or

before 31 December 2027, is below 25%. That reading must be dated on or after 1 October 2027; without one, the check does not resolve. On 2 October 2026 the ratio was 516 of 6,157, or 8.4%. Against the thesis: a ratio of 25% or more. The ratio moves with the lab’s disclosure inflow and triage pacing as well as with patching: a slower disclosure pace raises it, and a faster one lowers it, with no change in how fast projects patch. It pools all disclosures to date, so it mixes older and newer cohorts and is not a cohort patch rate (Section 6.1). It counts patches created, not deployed, so a high ratio would not show that deployment keeps pace, and it cannot identify Proposition 5’s condition $\beta > \varepsilon_H$. The dashboard counts patches only “to our knowledge”, so an undercount would bias the check toward our expectation.

2. *Exploitation still rising, calendar 2027 (Section 6.1’s attacker-favoring hypothesis; the rising phase of Proposition 3’s path).* We expect the count of exploited vulnerabilities to keep rising. Resolution: in CISA’s catalogue of known exploited vulnerabilities (CISA 2026), in the version current on 31 January 2028, entries with a dateAdded in 2027 outnumber entries with a dateAdded in 2026. Against the indicator: as many entries in 2027 as in 2026, or fewer. The count reflects CISA’s listing practice as well as attackers, and it counts first listings, not exploitation events. Two rival explanations can move it with no change in exploitation: batch listing of old flaws (in 2022, 381 of the 555 additions, 68.6%, were CVEs at least three identifier-years old; Supplement S2) and the volume of disclosures, which recent exploitation counts have tracked in one threat-intelligence group’s reading (Section 6.1).
3. *Memory-corruption share held down, calendar 2027 (Proposition 3’s residual flaw introduction).* We expect the memory-corruption share not to rise above its recent range. Resolution: applying Supplement S2’s data note unchanged to entries with a dateAdded in 2027, in the catalogue version current on 31 January 2028, the memory-corruption share of entries with a usable CWE tag is at or below 23.0%. Against the thesis: a share above 23.0%. The cut sits above each yearly share in the baseline window: 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. If the share were a fixed 19.0% (125 of 659) with independent entries, a conditional binomial calculation puts the chance that a year’s share exceeds 23.0% at 4.4–7.4% for 220 to 261 usable entries (10.1% at 2024’s count of 178); at the earlier cut of 19.0% it was about one half. That is a calculation under stated conditions, not a measured error rate. The share reflects CISA’s listing practice as well as attackers (Section 6.3), and it does not identify what lowers it or Proposition 3’s flaw-stock dynamics. If AI makes old C and C++ code cheap to harvest for flaws, the share could rise with no change in the residual introduction of new flaws (Supplement S2).

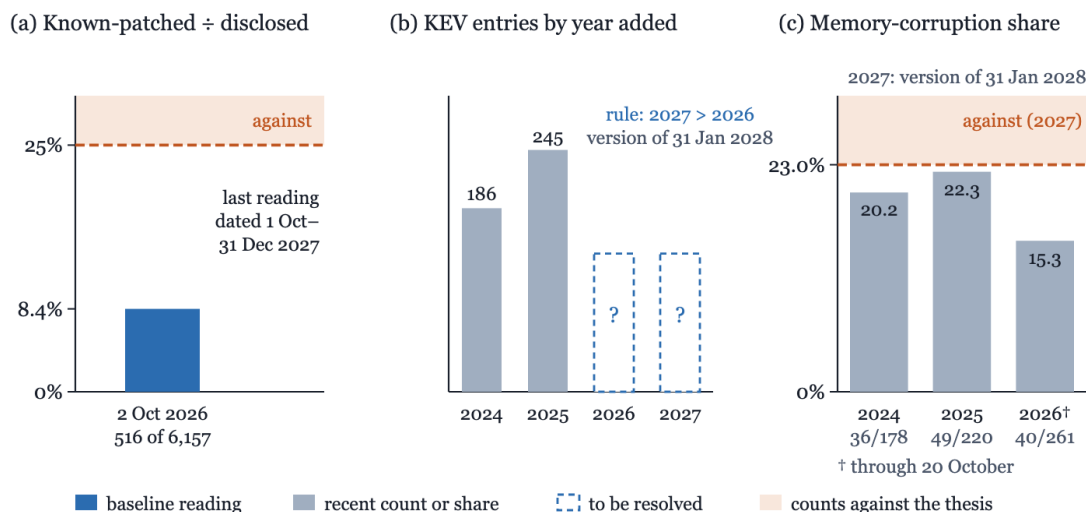


Figure 27. The three dated predictions (Section 11). (a) Known-patched count divided by disclosed count on Anthropic’s disclosure dashboard: 8.4% (516 of 6,157) on 2 October 2026; against the thesis, 25% or more on the last reading dated on or before 31 December 2027, which must be dated on or after 1 October 2027, or the check does not resolve. The count is of patches created, not deployed. (b) Entries in CISA’s catalogue of known exploited vulnerabilities by year added, with the 2024 and 2025 additions reported in Section 6.1; the prediction resolves on whether entries with a dateAdded in 2027 outnumber those in 2026 in the catalogue version current on 31 January 2028. (c) Memory-corruption share of entries with a usable CWE tag, by year added, 2024 to 2026 (2026 through 20 October); against the thesis, a 2027 share above 23.0%, with Supplement S2’s data note applied unchanged to entries with a dateAdded in 2027 in the catalogue version current on 31 January 2028. These are indicator forecasts that resolve on published counts; none tests the mechanism of the proposition it bears on, and each count also reflects listing practice, disclosure pace or triage.

Measurements.

- *Patch lags without mandates, 2027 (Proposition 5):* whether any class of operators reaches median patch lags under two weeks without a mandate.
- *The next restricted capability, through 2028 (Propositions 6 and 11):* whether the next restricted frontier capability holds more than six months before open-weights parity, and whether it ends by its holder’s own release or by parity.
- *Closed-weight copies, through 2028 (Proposition E1):* the lag from a frontier release to the nearest closed-weight copy, which has not yet been measured.
- *The two copy clocks, 2027–28 (Proposition 11):* whether the interval between frontier releases shortens as the open-weights lag shortens.
- *Uses outside the inference market, 2027–28 (Propositions 13 and 15):* whether frontier labs’ launches outside the inference market cluster in the months after capability jumps.

- *Adoption that speeds up, 2027–28 (Proposition 5)*: whether verified translation and rewriting reach production code, and whether AI patching narrows the gap between flaws found and flaws fixed.
- *Slow adoption populations, 2027–28 (Propositions 1 and 5; Appendix A, Section A.7)*: adoption times for maintainers, small operators and operators of operational technology, which none of the series in Section 6.2 covers, to locate the regime across populations.
- *Differential uptake, 2027–28 (Section 2)*: whether defenders' uptake of AI tools outpaces attackers', the observable for the differential variant that our models do not cover.
- *Where the wait for a health benefit goes, 2027–28 (Proposition 16)*: three quantities, measured separately for AI-assisted drug programmes: the capability wait K_B/s , the benefit's own adoption time T_b (trials, approval, manufacture and uptake), and the share m_B of that adoption that waits for the capability. On the benefit's capability branch, the capability-set share of the total wait, $q/(q + m_B T_b)$ with $q = K_B/s$, is derived from them and is not m_B ; on the adoption branch the share is zero. No source we read gives any of the three, or the DALYs per harmful cyber event, V_c .

12. Conclusion

Running through the rain works because the far curb is fixed and the rain does not thicken. Our models take neither to hold on the road to superintelligence: their premises are that the curb is built by operators patching, retiring and rewriting on their own clock, and that running arms attackers as it arms defenders. A peak can exist in the flaw-stock dynamics: one occurs when new flaws for a time arrive faster than old ones are removed, and later more slowly (Proposition 3). Whether AI capability creates or moves it is not modelled, and whether a lower plateau or a second rise follows depends on the code protection cannot reach and on attackers' capability per flaw. On speed, the result is cleanest where adoption, not capability, sets the transition's length. There, with adoption time fixed and adoption running at least partly alongside capability, running raises near-peak risk, whether the hazard is set by speed or by the capability level reached and is higher at the end of the path than on average; with a speed-set hazard it helps only if it speeds adoption faster than it raises the hazard. Where capability sets the length, running can help: always for a level-set hazard, and for a speed-set one if and only if the danger rises less than proportionally with speed under parallel timing, or, under mixed timing, less than the share of the transition that capability sets ($\varepsilon_H < (K/s)/(K/s + mT_a)$). For retirement and migration in the populations the series cover, the evidence suggests, provisionally, that adoption binds; the capability distance that would locate the kink is unmeasured, and we have found no measurement of the elasticity that would decide the matter. For the two-loss objective of Proposition 4, whether ASI is the larger peak matters only where speed lowers the near one, and there the comparison needs a valuation of the far loss that this paper does not supply (Proposition 4). Health benefits from the same capability give a second reason to run: they can justify speed even while near harm rises, and the far-peak comparison then matters too, but only while the benefit still waits on capability rather than on its own trials,

approval and uptake, and only if the benefit brought forward outweighs the added cyber loss (Proposition 16).

Holding the storm at the corner helps for a while under the model's conditions. It protects covered operators who can cross in time and delays attackers' search for unfound flaws, for at most the window. The duration of that security window remains unverified; benchmark catch-up does not measure it. A release with safeguards that last as long as a hold delays the same attackers and gives every defender the capability; it beats holding only if that access speeds defenders' fixes by more than the expected cost of the safeguards failing (Appendix B, Proposition E2). The commercial side does not force release on a cadence set by copy speed: at an interior optimum, expected copying after release is, for linear and power-law revenue curves, a reason to hold.

If the aim is lower near-peak harm, faster deployment, wider coverage and defenses needing no frontier capability are candidate interventions. Whether acceleration also helps depends on the unestimated hazard and adoption responses. The models give conditional results; the available evidence does not rank feasible interventions. The storm may yet have a far side. The question is who builds the curb, and how fast.

Declarations

Funding. This work received no funding.

Competing interests. The author has no competing interests to declare.

Code and data availability. The replication code and the saved data it reads will be deposited at Zenodo [Zenodo DOI to be added on deposit]. Paths are relative to the deposit. MODEL/code/check_model.py checks Appendix A's formulas, parameter cases and counterexamples and prints MODEL/code/check_model_out.txt; MODEL/code/race_model.py reproduces the earlier draft's contest model and prints MODEL/code/race_model_out.txt. MODEL/code/check_benefits.py checks Proposition 16 and Appendix C and prints MODEL/code/check_benefits_out.txt. The canonical check of Appendix B is MODEL/EXT/code/check_ext.py, which prints MODEL/EXT/code/check_ext_out.txt; the copy in MODEL/EXT/rerun-trim/code/ is byte-identical, and the other archived copies implement an earlier, broader extension and are not canonical. M8/kev_v5.py, with the class lists in M8/kev_mix.py, computes the shares, yearly counts and binomial calculations of Sections 6.3 and 11 and Supplement S2 from the saved CISA catalogue, version 2026.10.21 (M8/sources/G4-kev_live.json and its CSV), and prints M8/kev_v5_out.txt.

References

Ablon, Lillian, and Andy Bogart. 2017. *Zero Days, Thousands of Nights: The Life and Times of Zero-Day Vulnerabilities and Their Exploits*. RR-1751. RAND Corporation. <https://doi.org/10.7249/RR1751>.

Adkins, Heather, Gadi Evron, and Bruce Schneier. 2025. "Autonomous AI Hacking and the Future of Cybersecurity." *CSO*, 8 October 2025.

<https://www.csoononline.com/article/4069075/autonomous-ai-hacking-and-the-future-of-cybersecurity.html>. Reproduced at Schneier on Security, <https://www.schneier.com/essays/archives/2025/10/autonomous-ai-hacking-and-the-future-of-cybersecurity.html>.

Aghion, Philippe, and Peter Howitt. 1990. "A Model of Growth Through Creative Destruction." NBER Working Paper 3223 (January 1990). <https://doi.org/10.3386/w3223>. (Published in revised form as *Econometrica* 60 (2), 1992, <https://doi.org/10.2307/2951599>.)

Aghion, Philippe, Christopher Harris, Peter Howitt, and John Vickers. 2001. "Competition, Imitation and Growth with Step-by-Step Innovation." *Review of Economic Studies* 68 (3): 467–492. <https://doi.org/10.1111/1467-937X.00177>.

AISI (AI Security Institute, UK Department for Science, Innovation and Technology). 2026. "How Fast Is Autonomous AI Cyber Capability Advancing?" Blog post, 13 May 2026. <https://www.aisi.gov.uk/blog/how-fast-is-autonomous-ai-cyber-capability-advancing>.

Anderson, Ross, and Tyler Moore. 2006. "The Economics of Information Security." Authors' manuscript, 11 pages. University of Cambridge Computer Laboratory. <https://www.cl.cam.ac.uk/~rja14/Papers/sciecon2.pdf>. The manuscript prints no date; the text cites it as "Anderson and Moore 2006, authors' manuscript." (Published as *Science* 314 (5799): 610–613, <https://doi.org/10.1126/science.1130992>.)

Anthropic. 2025. "Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign." Announcement post, published 13 November 2025; edited 14 November 2025. <https://www.anthropic.com/news/disrupting-AI-espionage>.

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

Anthropic. 2026b. "Project Glasswing: An Initial Update." Published 22 May 2026, the date printed on the page. <https://www.anthropic.com/research/glasswing-initial-update>.

Anthropic. 2026c. "Measuring LLMs' Impact on N-Day Exploits." Research post, Frontier Red Team, by Winnie Xiao, Tim Abbott, Nicholas Carlini, Newton Cheng, David Forsythe, Keane Lucas, Milad Nasr and Shikhar Sakhuja, published 8 June 2026. <https://www.anthropic.com/research/n-days>.

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

Anthropic. 2026e. "Anthropic's Coordinated Vulnerability Disclosure Dashboard." Counts as of 2 October 2026; page last updated 2 October 2026, 19:47 UTC. <https://red.anthropic.com/2026/cvd/>.

Anthropic. 2026f. "Evaluating and Mitigating the Growing Risk of LLM-Discovered 0-Days." Frontier Red Team research post, by Nicholas Carlini, Keane Lucas, Evyatar Ben Asher,

Newton Cheng, Hasnain Lakhani, David Forsythe and Kyla Guru. Published 5 February 2026; author list edited 6 February 2026. <https://www.anthropic.com/research/zero-days>.

Anthropic. 2026g. “Assessing Claude Mythos Preview’s Cybersecurity Capabilities.” Frontier Red Team research post, by Nicholas Carlini, Newton Cheng, Keane Lucas, Michael Moore, Milad Nasr, Vinay Prabhushankar and Winnie Xiao (the first of the page’s two name lines; the second lists nineteen further names). Published 7 April 2026; author list edited 9 April 2026. <https://red.anthropic.com/2026/mythos-preview/> (redirects to <https://www.anthropic.com/research/mythos-preview>).

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

Anton, James J., and Dennis A. Yao. 2004. “Little Patents and Big Secrets: Managing Intellectual Property.” *RAND Journal of Economics* 35 (1): 1–22. <https://doi.org/10.2307/1593727>.

Arbatskaya, Maria, and Hugo M. Mialon. 2010. “Multi-Activity Contests.” *Economic Theory* 43 (1): 23–43. <https://doi.org/10.1007/s00199-008-0424-y>. Published online 2008.

Arbaugh, William A., William L. Fithen, and John McHugh. 2000. “Windows of Vulnerability: A Case Study Analysis.” *Computer* 33 (12): 52–59. <https://doi.org/10.1109/2.889093>.

ARC Prize. 2026. “ARC Prize – Leaderboard.” Web page, <https://arcprize.org/leaderboard>, with its ARC-AGI-2 data file for the semi-private set, <https://arcprize.org/media/data/leaderboard/v2.json>. Accessed 23 October 2026 (data file generated 21 October 2026).

Armstrong, Stuart, Nick Bostrom, and Carl Shulman. 2013. *Racing to the Precipice: A Model of Artificial Intelligence Development*. Technical Report #2013-1. Future of Humanity Institute, October 2013. (Published as *AI & Society* 31 (2): 201–206, 2016, published online 2015, <https://doi.org/10.1007/s00146-015-0590-y>.)

Arora, Ashish, Rahul Telang, and Hao Xu. 2008. “Optimal Policy for Software Vulnerability Disclosure.” *Management Science* 54 (4): 642–656. <https://doi.org/10.1287/mnsc.1070.0771>.

Arrow, Kenneth J. 1962. “Economic Welfare and the Allocation of Resources for Invention.” In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, by the Universities-National Bureau Committee for Economic Research and the Committee on Economic Growth of the Social Science Research Council, 609–626. Princeton University Press for the National Bureau of Economic Research. <https://www.nber.org/chapters/c2144>.

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

Axelrod, Robert, and Rumen Iliev. 2014. "Timing of Cyber Conflict." *Proceedings of the National Academy of Sciences* 111 (4): 1298–1303.
<https://doi.org/10.1073/pnas.1322638111>.

Bell, David E. 1976. "60.21 Walk or Run in the Rain?" *The Mathematical Gazette* 60 (413): 206–208. <https://doi.org/10.2307/3617483>.

Belson, David. 2025. "The 2025 Cloudflare Radar Year in Review: The Rise of AI, Post-Quantum, and Record-Breaking DDoS Attacks." Cloudflare Blog, 15 December 2025.
<https://blog.cloudflare.com/radar-2025-year-in-review/>.

Bostrom, Nick. 2019. "The Vulnerable World Hypothesis." *Global Policy* 10 (4): 455–476.
<https://doi.org/10.1111/1758-5899.12718>.

Brumley, David, Pongsin Poosankam, Dawn Song, and Jiang Zheng. 2008. "Automatic Patch-Based Exploit Generation Is Possible: Techniques and Implications." In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, 143–157. IEEE.
<https://doi.org/10.1109/SP.2008.17>.

Buchanan, James M. 1965. "An Economic Theory of Clubs." *Economica*, n.s., 32 (125): 1–14.
<https://doi.org/10.2307/2552442>.

Buterin, Vitalik. 2023. "My Techno-Optimism." Blog post, 27 November 2023.
https://vitalik.eth.limo/general/2023/11/27/techno_optimism.html.

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

Carlini, Nicholas, Daniel Paleka, Krishnamurthy Dvijotham, Thomas Steinke, Jonathan Hayase, A. Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Itay Yona, Eric Wallace, David Rolnick, and Florian Tramèr. 2024. "Stealing Part of a Production Language Model." arXiv:2403.06634v2, 9 July 2024.
<https://doi.org/10.48550/arXiv.2403.06634>. Version 2 carries the line for the *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235; the PMLR pages were not checked.

Charrier, Casey, and Robert Weiner. 2024. "How Low Can You Go? An Analysis of 2023 Time-to-Exploit Trends." Mandiant, Google Cloud Blog, Threat Intelligence, 15 October 2024. <https://cloud.google.com/blog/topics/threat-intelligence/time-to-exploit-trends-2023>.

Charrier, Casey, James Sadowski, Zander Work, Clement Lecigne, Benoît Sevens, and Fred Plan. 2026. "Look What You Made Us Patch: 2025 Zero-Days in Review." Google Threat Intelligence Group, Google Cloud Blog, 5 March 2026.
<https://cloud.google.com/blog/topics/threat-intelligence/2025-zero-day-review/>.

Choi, Sung J., M. Eric Johnson, and Christoph U. Lehmann. 2019. "Data Breach Remediation Efforts and Their Implications for Hospital Quality." *Health Services Research* 54 (5): 971–980. <https://doi.org/10.1111/1475-6773.13203>.

CISA (Cybersecurity and Infrastructure Security Agency). 2026. "CISA Catalog of Known Exploited Vulnerabilities." Catalog version 2026.10.21, released 21 October 2026. https://www.cisa.gov/sites/default/files/feeds/known_exploited_vulnerabilities.json.

Coe, Andrew J., and Jane Vaynman. 2020. "Why Arms Control Is So Rare." *American Political Science Review* 114 (2): 342–355. <https://doi.org/10.1017/S000305541900073X>. Published online December 2019.

Cognition. 2026. "FrontierCode Leaderboard," with methodology revisions and changelog. Accessed 23 October 2026. <https://cognition.com/frontiercode>.

Comin, Diego, and Bart Hobijn. 2006. "An Exploration of Technology Diffusion." NBER Working Paper 12314 (June 2006). Cambridge, MA: National Bureau of Economic Research. <https://doi.org/10.3386/w12314>. (Published in revised form as *American Economic Review* 100 (5), 2010.)

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

Cottier, Ben, Josh You, Natalia Martemianova, and David Owen. 2024. "How Far Behind Are Open Models?" Epoch AI report, 4 November 2024. <https://epoch.ai/publications/open-models-report>.

Dameff, Christian, Jeffrey Tully, Theodore C. Chan, et al. 2023. "Ransomware Attack Associated With Disruptions at Adjacent Emergency Departments in the US." *JAMA Network Open* 6 (5): e2312270. <https://doi.org/10.1001/jamanetworkopen.2023.12270>.

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

Dasgupta, Partha, and Joseph Stiglitz. 1980. "Uncertainty, Industrial Structure, and the Speed of R&D." *Bell Journal of Economics* 11 (1): 1–28. <https://doi.org/10.2307/3003398>.

Debenedetti, Edoardo, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. 2025. "Defeating Prompt Injections by Design." arXiv:2503.18813v2, 24 June 2025. <https://doi.org/10.48550/arXiv.2503.18813>.

Debs, Alexandre, and Nuno P. Monteiro. 2014. "Known Unknowns: Power Shifts, Uncertainty, and War." *International Organization* 68 (1): 1–31. <https://doi.org/10.1017/S0020818313000192>.

DeepSeek-AI, Daya Guo, Dejian Yang, et al. 2025. “DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.” arXiv:2501.12948v2, 4 January 2026. <https://doi.org/10.48550/arXiv.2501.12948>.

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

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

Dixit, Avinash K., and Robert S. Pindyck. 1994. *Investment under Uncertainty*. Princeton University Press. <https://doi.org/10.1515/9781400830176>.

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

Emery-Xu, Nicholas, Andrew Park, and Robert Trager. 2023. “Uncertainty, Information, and Risk in International Technology Races.” *Journal of Conflict Resolution*, online first, 1–29. Published online 17 November 2023. <https://doi.org/10.1177/00220027231214996>. (In print as 68 (10): 2019–2047, November 2024; the body’s pinpoint is to the online-first pagination.)

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

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

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

Fearon, James D. 1995. “Rationalist Explanations for War.” *International Organization* 49 (3): 379–414. <https://doi.org/10.1017/S0020818300033324>.

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

Fonseca, Pedro, Kaiyuan Zhang, Xi Wang, and Arvind Krishnamurthy. 2017. “An Empirical Study on the Correctness of Formally Verified Distributed Systems.” In *Proceedings of the Twelfth European Conference on Computer Systems (EuroSys ’17)*, 328–343. ACM. <https://doi.org/10.1145/3064176.3064183>.

Ford, Mat. 2026. "18 Years Later, IPv6 Reaches Majority." Internet Society Pulse, 21 April 2026. <https://pulse.internetsociety.org/en/blog/2026/04/18-years-later-ipv6-reaches-majority/>.

FrontierSWE. 2026. "FrontierSWE V2" leaderboard. Accessed 23 October 2026. <https://www.frontierswe.com/>.

GAO (U.S. Government Accountability Office). 2025. *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>.

Garfinkel, Ben, and Allan Dafoe. 2019. "How Does the Offense-Defense Balance Scale?" *Journal of Strategic Studies* 42 (6): 736–763. <https://doi.org/10.1080/01402390.2019.1631810>.

Google. 2026. "IPv6 Adoption Statistics." Data file behind <https://www.google.com/intl/en/ipv6/statistics.html>, https://www.google.com/intl/en_ALL/ipv6/statistics/data/adoption.js. Accessed 23 October 2026 (last row 21 October 2026).

Green, Edward J., and Robert H. Porter. 1984. "Noncooperative Collusion under Imperfect Price Information." *Econometrica* 52 (1): 87–100. <https://doi.org/10.2307/1911462>.

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

Gudibande, Arnav, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. "The False Promise of Imitating Proprietary LLMs." arXiv:2305.15717. <https://doi.org/10.48550/arXiv.2305.15717>.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. 2015. "Distilling the Knowledge in a Neural Network." arXiv:1503.02531. <https://doi.org/10.48550/arXiv.1503.02531>.

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, et al. 2022. "Training Compute-Optimal Large Language Models." arXiv:2203.15556. <https://doi.org/10.48550/arXiv.2203.15556>.

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

Hua, Shin-Shin, and Haydn Belfield. 2023. "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)*, 596–605. ACM. <https://doi.org/10.1145/3600211.3604694>. Pages per Crossref.

Jervis, Robert. 1978. "Cooperation under the Security Dilemma." *World Politics* 30 (2): 167–214. <https://doi.org/10.2307/2009958>.

Jones, Charles I. 2024. "The AI Dilemma: Growth versus Existential Risk." *American Economic Review: Insights* 6 (4): 575–590. <https://doi.org/10.1257/aeri.20230570>.

Jumper, John, Richard Evans, Alexander Pritzel, et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 596 (7873): 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. "Scaling Laws for Neural Language Models." arXiv:2001.08361. <https://doi.org/10.48550/arXiv.2001.08361>.

Kapoor, Sayash, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, Rumman Chowdhury, Alex Engler, Peter Henderson, Yacine Jernite, Seth Lazar, Stefano Maffulli, Alondra Nelson, Joelle Pineau, Aviya Skowron, Dawn Song, Victor Storch, Daniel Zhang, Daniel E. Ho, Percy Liang, and Arvind Narayanan. 2024. "On the Societal Impact of Open Foundation Models." arXiv:2403.07918. <https://doi.org/10.48550/arXiv.2403.07918>.

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

Khatry, Anirudh, Robert Zhang, Jia Pan, Ziteng Wang, Qiaochu Chen, Greg Durrett, and Isil Dillig. 2025. "CRUST-Bench: A Comprehensive Benchmark for C-to-safe-Rust Transpilation." arXiv:2504.15254v3, 1 October 2025. <https://doi.org/10.48550/arXiv.2504.15254>.

Klemperer, Paul. 1995. "Competition When Consumers Have Switching Costs: An Overview with Applications to Industrial Organization, Macroeconomics, and International Trade." *Review of Economic Studies* 62 (4): 515–539. <https://doi.org/10.2307/2298075>.

Korinek, Anton, and Jai Vipra. 2024. "Concentrating Intelligence: Scaling and Market Structure in Artificial Intelligence." NBER Working Paper 33139 (November 2024). <https://doi.org/10.3386/w33139>.

Lee, Tom, and Louis L. Wilde. 1980. "Market Structure and Innovation: A Reformulation." *Quarterly Journal of Economics* 94 (2): 429–436. <https://doi.org/10.2307/1884551>.

Lermen, Simon, Charlie Rogers-Smith, and Jeffrey Ladish. 2024. "LoRA Fine-Tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B." arXiv:2310.20624v2, 22 May 2024, whose first page reads "Published at ICLR 2024 Workshop on Secure and Trustworthy Large Language Models." <https://doi.org/10.48550/arXiv.2310.20624>.

Levin, Richard C., Alvin K. Klevorick, Richard R. Nelson, and Sidney G. Winter. 1987. "Appropriating the Returns from Industrial Research and Development." *Brookings Papers on Economic Activity* 1987 (3): 783–831 (pages include the comments and discussion). <https://doi.org/10.2307/2534454>.

Li, Frank, and Vern Paxson. 2017. "A Large-Scale Empirical Study of Security Patches." In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17)*, 2201–2215. ACM. <https://doi.org/10.1145/3133956.3134072>.

Lieberman, Marvin B., and David B. Montgomery. 1988. "First-Mover Advantages." *Strategic Management Journal* 9 (S1): 41–58. <https://doi.org/10.1002/smj.4250090706>.

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

Loury, Glenn C. 1979. "Market Structure and Innovation." *Quarterly Journal of Economics* 93 (3): 395–410. <https://doi.org/10.2307/1883165>.

METR. 2025. "Details about METR's Evaluation of OpenAI GPT-5." 7 August 2025. <https://metr.org/evaluations/gpt-5-report/>.

METR. 2026. "Task-Completion Time Horizons of Frontier AI Models." Web page; notice of 8 May 2026 on measurements above 16 hours; marked "no longer actively updated" on 8 September 2026. Accessed 23 October 2026. <https://metr.org/time-horizons/>.

Microsoft. 2026. "Manage Windows Autopatch Groups." Microsoft Learn, Windows deployment documentation, page dated 5 August 2025 and updated 14 May 2026. <https://learn.microsoft.com/en-us/windows/deployment/windows-autopatch/manage/windows-autopatch-manage-autopatch-groups>.

Moorthy, K. Sridhar, and I. P. L. Png. 1992. "Market Segmentation, Cannibalization, and the Timing of Product Introductions." *Management Science* 38 (3): 345–359. <https://doi.org/10.1287/mnsc.38.3.345>.

Nappa, Antonio, Richard Johnson, Leyla Bilge, Juan Caballero, and Tudor Dumitraş. 2015. "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. <https://doi.org/10.1109/SP.2015.48>.

Nasr, Milad, Nicholas Carlini, Chawin Sitawarin, Sander V. Schulhoff, Jamie Hayes, Michael Ilie, Juliette Pluto, Shuang Song, Harsh Chaudhari, Ilia Shumailov, Abhradeep Thakurta, Kai Yuanqing Xiao, Andreas Terzis, and Florian Tramèr. 2025. "The Attacker Moves Second: Stronger Adaptive Attacks Bypass Defenses Against LLM Jailbreaks and Prompt Injections." arXiv:2510.09023v1, 10 October 2025. <https://doi.org/10.48550/arXiv.2510.09023>.

Neprash, Hannah, Claire McGlave, and Sayeh Nikpay. 2026. "Hacked to Pieces? The Effects of Ransomware Attacks on Hospitals and Patients." *American Economic Journal: Economic Policy* 18 (1): 256–281. <https://doi.org/10.1257/pol.20240594>. The part cited is the published abstract.

Nevo, Sella, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott. 2024. *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*. RR-A2849-1. RAND Corporation. <https://doi.org/10.7249/RAA2849-1>.

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

Nordhaus, William D. 1969. *Invention, Growth, and Welfare: A Theoretical Treatment of Technological Change*. Cambridge, MA: MIT Press.

Owen, David. 2024. "How Predictable Is Language Model Benchmark Performance?" arXiv:2401.04757. <https://doi.org/10.48550/arXiv.2401.04757>.

Ozment, Andy, and Stuart E. Schechter. 2006. "Milk or Wine: Does Software Security Improve with Age?" In *Proceedings of the 15th USENIX Security Symposium*, Vancouver, July 2006, 93–104. USENIX Association. <https://www.usenix.org/conference/15th-usenix-security-symposium/milk-or-wine-does-software-security-improve-age>. No DOI; the version read has no page numbers.

Palmer, Chris, and Ryan Sleevi. 2014. "Gradually Sunsetting SHA-1." Google Online Security Blog, 5 September 2014. <https://security.googleblog.com/2014/09/gradually-sunsetting-sha-1.html>.

PutnamBench. 2026. "PutnamBench Leaderboard" and its data file `results.json`. Trishul Lab. <https://trishullab.github.io/PutnamBench/leaderboard.html>; <https://trishullab.github.io/PutnamBench/results.json>. Accessed 23 October 2026.

Reinganum, Jennifer F. 1989. "The Timing of Innovation: Research, Development, and Diffusion." In *Handbook of Industrial Organization*, vol. 1, edited by Richard Schmalensee and Robert Willig, 849–908. Elsevier. [https://doi.org/10.1016/S1573-448X\(89\)01017-4](https://doi.org/10.1016/S1573-448X(89)01017-4).

Ren, Feng, Alex Aliper, Jian Chen, et al. 2024. "A Small-Molecule TNIK Inhibitor Targets Fibrosis in Preclinical and Clinical Models." *Nature Biotechnology* 43 (1): 63–75. Published online 8 March 2024; issue dated 2025. <https://doi.org/10.1038/s41587-024-02143-0>.

Rescorla, Eric. 2004. "Is Finding Security Holes a Good Idea?" Paper presented at the Third Workshop on the Economics of Information Security (WEIS 2004).

Rice, H. G. 1953. "Classes of Recursively Enumerable Sets and Their Decision Problems." *Transactions of the American Mathematical Society* 74 (2): 358–366. <https://doi.org/10.1090/S0002-9947-1953-0053041-6>.

Royston, Patrick, and Mahesh K. B. Parmar. 2013. "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: 152. <https://doi.org/10.1186/1471-2288-13-152>.

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

Schneier, Bruce. 2018. "Artificial Intelligence and the Attack/Defense Balance." *IEEE Security & Privacy* 16 (2): 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.

Seeger, Elizabeth, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K. Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, Markus Anderljung, Ben Bucknall, Alan Chan, Eoghan Stafford, Leonie Koessler, Aviv Ovadya, Ben Garfinkel, Emma Bluemke, Michael Aird, Patrick Levermore, Julian Hazell, and Abhishek Gupta. 2023. "Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives." Centre for the Governance of AI. arXiv:2311.09227v1, 29 September 2023.
<https://doi.org/10.48550/arXiv.2311.09227>. The same report is posted at SSRN, <https://doi.org/10.2139/ssrn.4596436>.

seL4 Project. n.d. "Frequently Asked Questions." Web page. Accessed 23 October 2026.
<https://sel4.systems/About/FAQ.html>.

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

Shevlane, Toby. 2022. "Structured Access to AI Capabilities: An Emerging Paradigm for Safe AI Deployment." arXiv:2201.05159v1, 13 January 2022.
<https://doi.org/10.48550/arXiv.2201.05159>. Version 2 (11 April 2022) is retitled "Structured Access: An Emerging Paradigm for Safe AI Deployment" and was not read.

Shevlane, Toby, and Allan Dafoe. 2020. "The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?" arXiv:2001.00463v2, 9 January 2020. <https://doi.org/10.48550/arXiv.2001.00463>. (Published in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*, 173–179. ACM.
<https://doi.org/10.1145/3375627.3375815>.)

Slayton, Rebecca. 2017. "What Is the Cyber Offense-Defense Balance? Conceptions, Causes, and Assessment." *International Security* 41 (3): 72–109.
https://doi.org/10.1162/ISEC_a_00267. Issue dated Winter 2016/17.

Snell, Charlie, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. "Scaling LLM Test-Time Compute Optimally Can Be More Effective than Scaling Model Parameters." arXiv:2408.03314. <https://doi.org/10.48550/arXiv.2408.03314>.

Stix, Charlotte, Matteo Pistillo, Girish Sastry, Marius Hobbhahn, Alejandro Ortega, Mikita Balesni, Annika Hallensleben, Nix Goldowsky-Dill, and Lee Sharkey. 2025. "AI Behind Closed Doors: A Primer on the Governance of Internal Deployment." Apollo Research. arXiv:2504.12170. <https://doi.org/10.48550/arXiv.2504.12170>.

Teece, David J. 1986. "Profiting from Technological Innovation: Implications for Integration, Collaboration, Licensing and Public Policy." *Research Policy* 15 (6): 285–305.
[https://doi.org/10.1016/0048-7333\(86\)90027-2](https://doi.org/10.1016/0048-7333(86)90027-2).

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

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

Trammell, Philip, and Leopold Aschenbrenner. 2024. "Existential Risk and Growth." Global Priorities Institute Working Paper No. 13-2024 (December 2024; dated 7 December 2024). <https://globalprioritiesinstitute.org/existential-risk-and-growth-aschenbrenner-and-trammell>. Author order is from the title page (Trammell first); the work is also cited as Aschenbrenner and Trammell. The body pinpoints by section.

Tramèr, Florian, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. 2016. "Stealing Machine Learning Models via Prediction APIs." arXiv:1609.02943v2, 3 October 2016, the extended version. <https://doi.org/10.48550/arXiv.1609.02943>. (Published in *Proceedings of the USENIX Security Symposium 2016*, 601–618, pages per OpenAlex.)

Trump, Donald J. 2026. "Promoting Advanced Artificial Intelligence Innovation and Security." Executive Order 14409 of 2 June 2026. *Federal Register* 91 (108): 34565–34567, 5 June 2026. FR Doc. 2026-11415. <https://www.federalregister.gov/documents/2026/06/05/2026-11415/promoting-advanced-artificial-intelligence-innovation-and-security>.

Tullock, Gordon. 1980. "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. Read in the reprint in *Efficient Rent-Seeking: Chronicle of an Intellectual Quagmire*, edited by Alan A. Lockard and Gordon Tullock (Springer, 2001), 3–16, https://doi.org/10.1007/978-1-4757-5055-3_2.

Van den Steen, Eric. 2004. "Rational Overoptimism (and Other Biases)." *American Economic Review* 94 (4): 1141–1151. <https://doi.org/10.1257/0002828042002697>.

Weeds, Helen. 2000. "Strategic Delay in a Real Options Model of R&D Competition." Warwick Economic Research Paper 576. University of Warwick, 20 September 2000. (Published as *Review of Economic Studies* 69 (3): 729–747, 2002, <https://doi.org/10.1111/1467-937X.t01-1-00029>.)

Westerbaan, Bas. 2024. "The State of the Post-Quantum Internet." Cloudflare Blog, 5 March 2024. <https://blog.cloudflare.com/pq-2024/>.

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

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

Xu, Zuojun, Feng Ren, Ping Wang, et al. 2025. “A Generative AI-Discovered TNIK Inhibitor for Idiopathic Pulmonary Fibrosis: A Randomized Phase 2a Trial.” *Nature Medicine* 31 (8): 2602–2610. <https://doi.org/10.1038/s41591-025-03743-2>.

Yang, Aidan Z. H., Yoshiki Takashima, Brandon Paulsen, Josiah Dodds, and Daniel Kroening. 2024. “VERT: Verified Equivalent Rust Transpilation with Large Language Models as Few-Shot Learners.” arXiv:2404.18852v2, 25 May 2024. <https://doi.org/10.48550/arXiv.2404.18852>.

Yang, Xuejun, Yang Chen, Eric Eide, and John Regehr. 2011. “Finding and Understanding Bugs in C Compilers.” In *Proceedings of the 32nd ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI '11)*, 283–294. ACM. <https://doi.org/10.1145/1993498.1993532>.

Zaegel, Matteo, Mathis Vehils-Vinals, Hugo Guastalla, Benjamin Benabou, and Auguste Gires. 2024. “Should You Walk, Run or Sprint in the Rain to Get Less Wet?” *European Journal of Physics* 45 (2): 025802. <https://doi.org/10.1088/1361-6404/ad06bf>. Published online 21 December 2023.

Appendix A. A Model of Speed, Secrecy and Intermediate Risk Peaks

Propositions P1–P15 here are Propositions 1–15 of the main text; A1–A14 are assumption labels, not section numbers. P16 and A15–A18, added in this version, follow the same convention.

A.1 Overview

Should the world speed through a dangerous technological transition, or slow it down, hold back the capability that drives it, or lower the danger directly? A runner crossing a street in the rain gets less wet by running, because running cuts the time under falling rain while the rain swept up in front is the same at any speed. We ask when that intuition carries over to the transition in which AI that can write and break code helps attackers before it helps defenders. We use four small sub-models: speed and adoption, holding a capability back, racing for it, and the holder’s private use of what it holds. We examine fifteen claims (P1–P15), each an informal step in the defensive case of Section 2 or in an earlier, superseded draft of this project; “as first stated” below refers to that informal version. Four hold, as identities or once conditions they left unstated are added. Eleven fail in at least one part as stated, and for each we prove a corrected version or show that none is needed. Two corrections change the qualitative story. Faster copying does not in general bring a held capability’s release sooner, and holding a capability back does protect outsiders’ unfound flaws. P16 (Section A.3.5) is not one of those claims: it extends sub-model Speed with health benefits that arrive with the same capability, it has no earlier informal version, and its proofs are in Appendix C.

A.2 Objectives and set-up

A.2.1 Whose payoff, in what units

The fifteen claims mix a social question (is acceleration good for the world?) with a private one (when should a firm release what it holds?). We keep them apart. We use four sub-models, each with its own objective, and say where they connect.

Sub-model	Results	Objective	Units	Discounting
Speed: speed and adoption	P1–P5	Social loss index	Expected harmful events, or losses at fixed value per event	Undiscounted in P1–P3 and P5; P4 adds rate r
Speed, extended: health benefits	P16	Health–cyber balance: health benefit less cyber loss	DALYs averted; V_c in DALYs per harmful event	Undiscounted over a calendar horizon L ; discounted at r on both sides
Hold: holding the storm	P6–P8	Exposure saved for operators	Exposure-time, with operator weights the analyst chooses	Undiscounted, with discounted forms stated
Contest: the contest	P9–P10	Each racer’s expected payoff; welfare as declared in each result	Money-equivalent payoff	Static game, except the repeated game in P10
Program: the holder’s program	P11–P15	The holder’s private expected present value	Money-equivalent present value	Rate $r > 0$ throughout

Two accounting rules keep the sub-models from double counting. First, the flow a held capability earns, F , contains only what the holder captures: its own security value and any payments from partners (f_σ), artifact revenue (f_a), interactive-product revenue (f_p), and the private value of internal research not already counted elsewhere (f_u). Security that accrues to outsiders is social value and enters sub-model Hold, not F . Second, internal research changes future capability. If the next contest’s prize is modeled in sub-model Contest, research enters there, as a higher chance of winning, and f_u is set to zero; counting it in both places counts the next race’s prize twice. The operator weights w_j in sub-model Hold are nonnegative weights chosen by the analyst, for example by criticality, by population served, or equally; every result holds for any such weights.

Several combinations are not licensed. The hold values in P13 and P14 assume that only involuntary loss ends the hold, so they cannot be combined with P11's release option without re-deriving them, and Section A.3.4 shows that the combination changes P13's answer. The window W in P6–P8 is exogenous. If a fixed release date T truncates an involuntary-loss time L , the realized window is $W_T = \min(L, T)$. P6(i)'s distribution-free bound still applies. If L is exponential with rate $\lambda > 0$, then $E[\min(W_T, \tau)] = (1 - e^{-\lambda \min(T, \tau)})/\lambda$, which equals P6(ii)'s formula only when $T \geq \tau$. P7's pathwise comparisons retain their stated date and access conditions; its exponential expectation must use the truncated law. P8's coalition objective must be rebuilt with this restricted mean before combining coalition and release choices; changing the optimal release date with coalition size requires an additional derivative term. The private value of a window that drives racing in P9 belongs to sub-model Program, the social gain it is compared with to sub-model Hold, and they meet only through the welfare change defined in P9.

P16's balance is a fifth objective inside sub-model Speed, used for that extension only. Its unit, DALYs averted, carries both the mortality and the morbidity that health benefits and cyber harm involve. It is a conditional health–cyber balance under A15–A18, not total social welfare, and it is not combined with the objectives of sub-models Hold, Contest or Program. A cyber harm counted in $V_c C_c$ is not also subtracted from the health flow y .

A.2.2 The loss index

Near-peak risk in sub-model Speed is the loss index

$$R_c(s) = H(s) \cdot T(s) + S,$$

with the speed s , duration T , hazard H and step term S of Section A.2.4. We define R_c as the model's loss index rather than derive it. It equals cumulative hazard when the hazard is rectangular, at height H for duration T , and is proportional to it for any hazard of fixed shape $H \cdot g(t/T)$ with $\int g = 1$. If harmful events arrive as a Poisson process, R_c is the expected number of events, $1 - e^{-R_c}$ the probability of at least one, and R_c times the loss per event the expected loss. The three move together in sign, but only the count and the expected loss add across periods. A spike that narrows as it rises can be higher and smaller in area, so we never treat height alone as risk.

A.2.3 Information, hazards and commitment

The holder observes its own capability and effort. Rivals and attackers observe artifacts the holder sells, patches it publishes, and who has been granted access, but not the capability or the effort. Nothing between rivals is enforceable unless a verification technology detects a deviation (P10). A rival's release hazard falls to zero inside a coalition only if the rival can commit not to release, and we flag each use of that assumption.

The hold ends at the first of several events: open-weights catch-up, a rival's release, a leak or theft, inference from visible artifacts, and copying through interactive products. We write the window hazard as

$$\lambda(n, F, f_a, f_p) = \lambda_o + \lambda_r(n) + \lambda_\ell(n) + \lambda_\theta(F) + \delta_a + \delta_p.$$

Each term is a cause-specific intensity, the rate of that exit given the shared history and no exit so far. Intensities of mutually exclusive exits add without any independence assumption, and the window is exponential when their total is constant. Marginal hazards of causes that share a shock do not add. If one break-in both leaks the weights and enables theft, two causes with marginal rate 0.2 each give a first-arrival rate of 0.2, not 0.4.

A.2.4 Notation

s : speed of capability advance, on $[s_{\min}, \infty)$. K : capability distance to the end of the transition. T_a : adoption time (constant in P1; $A(s/s_0)^{-\beta}$ in P5, A being its value at a reference speed s_0). $T(s) = \max(K/s, T_a)$. s^* : the kink where $K/s = T_a$. $H(s)$: transition hazard; $\varepsilon_H = sH'/H$ its elasticity. S : the step term. R_A : far (ASI) loss. V_c, V_A : values at stake in the near and far losses. r : discount rate. $E(t)$: stock of exploitable flaws. $v(t)$: rate at which new code introduces flaws. $\chi(t)$: protected share of new code. $\iota = v(1 - \chi)$: residual introduction. μ : removal rate. φ : exploitation intensity per flaw. T_A : arrival of ASI. W : window during which attackers without parallel access lack the held capability. λ : hazard ending the window. π : share of attack effort with parallel access. τ_j, w_j, c_j : operator j 's deployment time, weight and coverage probability. n : coalition size. ℓ_x : exploit lag from patch to exploit. d_f : attackers' own discovery time once they hold the capability. ζ_j : date at which operator j 's normal cycle removes a flaw. x_i : racing effort. c, k : linear and quadratic effort cost. v_W, v_L : winner's and loser's payoffs. D : positional stake. v_s : private value of a window; σ_s : its real-surplus part. θ : catastrophe term per unit of total effort. L, L_A : a racer's own and society's catastrophe loss. b : uncaptured public value per unit effort. v : detection probability. ϱ : per-period discount factor. F : held private flow. $g(t)$: lead of the holder's current public model. $\kappa_{\text{pre}}, \kappa_{\text{post}}$: copy rates of the public model before release and of the released model after it (κ when equal). Δ : the capability step released. $\rho(x)$: inference rent at lead x . $J(T)$: value of releasing at T . Z : holder's continuation value after involuntary loss. f, δ : added flow and added hazard of a new use. d : durable share of a step's earnings. D_s, D_d : step and decisive prizes. y : health-benefit flow, DALYs averted per year (P16; b above is taken). K_B : capability distance to the capability the benefit needs. T_b : the benefit's adoption time. m, m_B : shares of cyber and benefit adoption that wait for the capability. $t_B(s)$: the date the benefit arrives. $\tau = K/s, q = K_B/s$: capability waits for the end of the cyber transition and for the benefit (P16 only; τ_j above is a deployment time). χ_c, χ_B : indicators that capability sets the cyber and the benefit clock (unrelated to $\chi(t)$ above). L : calendar horizon in P16 (in P9–P10, a racer's loss). η_L : indicator that the benefit arrives before L . W_L, W_r : P16's undiscounted and discounted balances (not the window W). B_L, B_r : benefit levels (not P8's $B(n)$). $A_r(u) = (1 - e^{-ru})/r$. S_r : present value of the step losses. $h(c), \bar{h}$: hazard at capability level c , and its average over $[0, K]$.

A.2.5 Assumption ledger

#	Assumption	Kind	Used in
A1	$R_c = HT + S$ is the loss index (exact for a rectangular hazard)	Modeling choice	P1–P5, P16

#	Assumption	Kind	Used in
A2	S does not depend on s	Modeling choice	P1, P2, P4, P5, P16
A3	T_a is constant (P1) or $A(s/s_0)^{-\beta}$ (P5)	Modeling choice; the value of β is an empirical premise	P1, P5, P16
A4	H is positive and differentiable, with $H' > 0$	Modeling choice	P1, P4, P5, P16
A5	Flaw stock follows $dE/dt = \iota(t) - \mu(t)E$ with $\iota \geq 0$, $\mu(t) > 0$; hazard φE with $\varphi > 0$ constant	Modeling choice	P3
A6	Harm is linear in exposure time	Modeling choice; convexity is treated in P6	P6–P8
A7	The total window intensity is constant; its terms are cause-specific intensities of mutually exclusive exits, with common shocks counted once	Modeling choice	P6, P8, P11, P13, P14
A8	Attackers with parallel access hold share π of attack effort; the rest arrive when the window closes, with unchanged intensity afterward	Modeling choice; π is an empirical premise	P6–P8
A9	Coverage is independent of the window	Modeling choice	P6, P8
A10	Two-player Tullock contest, exponent one, linear cost, equal split at zero total effort; catastrophe term θX , a linear expected-loss index in total	Modeling choice	P9, P10, P15

#	Assumption	Kind	Used in
	effort that is a probability only while $\theta X \leq 1$		
A11	The holder's flows are constant while held; inference rent is C^2 on $(0, g_0 + \Delta]$ and continuous on $[0, g_0 + \Delta]$	Modeling choice	P11–P15
A12	Leads erode exponentially when copied	Modeling choice; the rates are empirical premises	P11, P12, P15
A13	Rival commitment inside a coalition	Empirical premise	P8
A14	Adoption binds at current speeds; windows last months	Empirical premise; no result here depends on it	P5 remark; Section A.5
A15	A constant health-benefit flow $y > 0$ begins at $t_B(s)$ and persists; V_c is a fixed number of DALYs per harmful event; the health and cyber terms add	Modeling choice; the values of y and V_c are empirical premises	P16
A16	One speed s advances both distances K and K_B (full coupling)	Modeling choice	P16
A17	The benefit's adoption time T_b and share m_B , and the cyber share m , are fixed in speed; each clock is parallel, mixed or sequential, and the two may differ	Modeling choice; the values are empirical premises	P16
A18	No cyber harm is counted both in $V_c C_c$ and as a cut in y	Accounting rule	P16

A.3 Results

A.3.1 Speed, adoption and two peaks (P1–P5)

P1, the speed threshold. *In plain terms:* when capability sets the transition's length, speeding up shortens the exposure and raises its intensity, and it lowers risk only if intensity rises less than proportionally with speed. Once adoption sets the length, speed only raises intensity.

Proposition 1. Under A1–A4 with constant T_a , let $s^* = K/T_a$ and let speeds range over $[s_{\min}, \infty)$ with $0 < s_{\min} < s^*$. For $s < s^*$, $dR_c/ds = (KH/s^2)(\varepsilon_H(s) - 1)$, so acceleration lowers R_c at s if and only if $\varepsilon_H(s) < 1$. For $s > s^*$, $dR_c/ds = H'(s)T_a > 0$. At s^* the left and right derivatives are $(HT_a/s^*)(\varepsilon_H - 1)$ and $H'T_a$. A minimizer of R_c exists and lies in $[s_{\min}, s^*]$. If $\varepsilon_H < 1$ on $[s_{\min}, s^*)$, it is s^* . In general it is $s_{\min}$, s^* , or an interior point where $\varepsilon_H = 1$.

The original claim is true under conditions it left unstated: constant adoption time, a speed-independent step term, a differentiable hazard with $H' > 0$, the loss index taken as exact, and an elasticity condition that holds pointwise rather than as a global constant. The lower speed bound matters: without it a minimizer need not exist (appendix).

P2, the step term. *In plain terms:* each capability step exposes a stock of latent flaws once, whatever the speed.

As first stated, P2 said that no speed reduces S and that only fewer exploitable flaws do. The first sentence is a modeling assumption (A2), not a result. The second is false. Write $S = \sum_k N_k \psi_k d_k$, summed over steps, with N_k the flaws a step exposes, ψ_k the chance one is exploited, and d_k the damage per exploit. Halving d with N and ψ fixed halves S ; halving ψ with the other factors fixed does the same. Segmentation and rate limits as ways to lower d , and hardware-bound authentication as a way to lower ψ , are proposed intervention channels whose effects require separate evidence. And if exploit success at a step depends on the tools attackers hold when it arrives, S depends on speed after all. *Corrected statement:* S is independent of s when the number of steps and the per-step exposure, success and damage are; it falls with any of N , ψ or d when the other two are positive.

P3, the existence of a peak. *In plain terms:* risk rises while new code adds flaws faster than patching and retirement remove them, and it falls once protection has cut new flaws enough. A peak is guaranteed when protection spreads more slowly than removal works; when protection spreads faster, a peak still occurs unless enough new code stays unprotected.

The original could not be evaluated: its set-up defined only a peak height $H(s)$, not a path of risk. We specify one (A5): $dE/dt = \iota(t) - \mu(t)E$, with hazard $h = \varphi E$, on the pre-ASI domain $[0, T_A)$. Part (i) lets the removal rate vary over time; parts (ii) and (iii) take μ constant.

Proposition 3. (i) The hazard rises and later falls before T_A if and only if there are dates $t_0 < t_1 < T_A$ with $\iota(t_0) > \mu E(t_0)$ and $\iota(t_1) < \mu E(t_1)$. At any interior maximum $\iota = \mu E$, and if ι

and μ are differentiable, $\iota' \leq \mu' E$. (ii) If μ is constant, ι is differentiable and non-increasing, and $E(0) < \iota(0)/\mu$, the path first rises and has at most one peak: E is non-decreasing up to some date and non-increasing after it, and a monotone rise is allowed. (iii) Suppose $\mu, \gamma > 0$ and protection spreads at rate γ , so that $\iota(t) = \iota_1 + (\iota_0 - \iota_1)e^{-\gamma t}$ with $\iota_0 > \iota_1 \geq 0$, and $E(0) = 0$. Then a peak exists if and only if $\mu\iota_0 > \gamma\iota_1$. For $\mu \neq \gamma$ it occurs at

$$\hat{t} = \ln[(\mu\iota_0 - \gamma\iota_1)/(\gamma(\iota_0 - \iota_1))]/(\mu - \gamma),$$

and for $\mu = \gamma$ at $\hat{t} = \iota_0/(\mu(\iota_0 - \iota_1))$. It is an intermediate pre-ASI peak if $\hat{t} < T_A$. After the peak, the untruncated path, continued past T_A , settles to the plateau $\phi\iota_1/\mu$.

As first stated, P3 said that an interior pre-ASI maximum exists only if residual flaw introduction stays bounded, and that otherwise risk plateaus. Both halves are false, on an unbounded horizon and on a finite one: introduction can be unbounded while the stock peaks and falls back toward zero, or while it grows without limit (appendix). Bounded introduction is not enough for a peak either: $\iota = 1$, $\mu = 0.5$ rises monotonically to a plateau. Peak and plateau are also not alternatives: in (iii), a peak is followed by a plateau.

Part (iii) has a policy reading. If protection eventually covers all new code ($\iota_1 = 0$), there is always a peak. If a residue remains, faster protection can remove the peak, leaving a monotone rise to the same plateau $\phi\iota_1/\mu$ with no overshoot; the plateau depends only on the residue and the removal rate.

P4, two peaks. *In plain terms*: faster capability may lower the near peak and raise the far one, and the comparison needs common units, discounting, and attention to the size of the move.

Proposition 4. Let the social loss be $V_c C_c(s) + V_A C_A(s)$, with $V_c, V_A > 0$, C_c and C_A the near and far cumulative hazards, and $C'_A > 0$. (i) A marginal acceleration lowers the loss if and only if $C'_c < 0$ and $V_c/V_A > C'_A/(-C'_c)$. Above s^* , $C'_c > 0$ under P1 and no ratio of stakes justifies acceleration. (ii) With discounting at rate r and a rectangular hazard profile, the near loss becomes $H(1 - e^{-rT})/r + S_r$, where $S_r = \sum_k S_k e^{-rt_k}$ is the present value of the step losses at their dates t_k . If S_r does not depend on s , then below s^* acceleration lowers the near loss if and only if $\varepsilon_H < x/(e^x - 1)$ with $x = rK/s$. This threshold lies below one, so discounting makes acceleration less attractive. If faster capability brings the steps earlier, $S'_r(s) > 0$ adds to the derivative and works further against acceleration. (iii) If the far loss arrives at $T_A(s) = K_A/s$, its present value $V_A C_A e^{-rT_A}$ has derivative $V_A e^{-rT_A} [C'_A + rC_A K_A/s^2]$: acceleration also brings the far loss nearer. (iv) If the two peaks are sequential catastrophes, with P_c the probability of the near one and P_A that of the far one conditional on surviving the near one, total catastrophe probability is $1 - (1 - P_c)(1 - P_A)$, and acceleration lowers it if and only if $-P'_c(1 - P_A) > P'_A(1 - P_c)$.

As first stated, P4 was true as a local criterion under unstated conditions: common units, differentiability, and no costs or benefits of speed outside the two losses. As a test for discrete accelerations it can mislead, a limit of any derivative criterion rather than an error in this one: a large move can cut the total loss where the derivative test fails (appendix). Its “break-even ratio of stakes” was undefined until the stakes were separated from the technical hazards, as in (i).

P5, the accelerationist's escape. *In plain terms:* if faster capability also speeds adoption, acceleration can lower risk even where adoption binds. Whether it does depends on how fast adoption responds relative to hazard, not on where the kink moves.

Proposition 5. Under A1, A2 and A4, let $T_a = A(s/s_0)^{-\beta}$, $\beta \geq 0$, where A is the adoption time at a reference speed s_0 , held fixed as β varies. (i) For $\beta \neq 1$, the kink is $s^* = s_0(K/(As_0))^{1/(1-\beta)}$. For $\beta < 1$, capability binds below s^* and adoption above it, as in P1. For $\beta > 1$, the order reverses. For $\beta = 1$, the branches never cross unless $K = As_0$, and one binds everywhere. (ii) For $\beta < 1$, s^* rises with β if and only if $K > As_0$, that is, if and only if capability binds at the reference speed. (iii) Wherever adoption binds, $d\ln(HT_a)/d\ln s = \varepsilon_H - \beta$, so acceleration lowers R_c if and only if $\beta > \varepsilon_H$, equivalently $|\varepsilon_{T_a}| > \varepsilon_H$. Wherever capability binds, β does not enter dR_c/ds . (iv) For general $T_a(s)$, the adoption-branch elasticity is $\varepsilon_H + \varepsilon_{T_a}$. P1 is the case $\beta = 0$. If $\beta > \varepsilon_H$ throughout an unbounded adoption branch, loss decreases throughout that branch and no minimum is attained there; this alone does not exclude a global minimizer elsewhere.

As first stated, P5 said that if T_a falls with s fast enough, s^* rises. That is false in general: with $K = 1$, $A = 2$ and $s_0 = 1$, raising β from 0 to 0.5 moves s^* from 0.50 down to 0.25. If A is measured at current speed and adoption binds there, as A14 holds, then $K < As_0$ and s^* falls with β for every $\beta < 1$. The original also aimed at the wrong quantity. The escape is a fall in risk on the adoption branch, which happens if and only if $\beta > \varepsilon_H$, and it can occur while s^* falls.

A.3.2 Holding the storm (P6–P8)

P6, the head start is worth at most its length. *In plain terms:* an operator who needs time τ to deploy fixes, and whose attackers are held off for a window W , saves at most the shorter of the two, and only against attackers who lacked access.

Proposition 6. Under A6–A9, a covered operator with deployment time τ_j saves expected exposure $(1 - \pi)E[\min(W, \tau_j)]$. (i) For any non-negative W , $E[\min(W, \tau)] \leq \min(E[W], \tau)$. (ii) If W is exponential with rate λ , $E[\min(W, \tau)] = (1 - e^{-\lambda\tau})/\lambda$. With discounting at r , the saving is $(1 - \pi)(1 - e^{-(r+\lambda)\tau})/(r + \lambda)$. (iii) As a share of the operator's baseline exposure τ_j to all attackers, the saving is asymptotically equal to $1 - \pi$ when $\lambda\tau_j \rightarrow 0$ and to $(1 - \pi)/(\lambda\tau_j)$ when $\lambda\tau_j \rightarrow \infty$. (iv) Aggregate. Let G be total saving and $E_0 = \sum w_j \tau_j$ baseline exposure. Let fast operators carry share α of E_0 , and let every other operator have coverage at most $\bar{c}$ and deployment time at least τ_s . Then

$$G/E_0 \leq (1 - \pi)[\alpha + (1 - \alpha)\bar{c}\min(1, 1/(\lambda\tau_s))] \leq (1 - \pi)[\alpha + \bar{c}(1 - \alpha)],$$

for any window law with mean $1/\lambda$. (v) If harm is a convex function g of exposure, measured as exposure time weighted by attack intensity, the saving in harm is $g(\tau) - g(\tau - (1 - \pi)w) \leq (1 - \pi)wg'(\tau)$, with $w = \min(W, \tau)$ and $g'(\tau)$ any subgradient of g at τ .

As first stated, the formula in P6 was true under unstated conditions: linear harm, an exponential window independent of coverage, and attackers who resume at unchanged intensity, with no backlog, when the window closes. The stated limits were false when $\pi >$

0. Against baseline exposure, the fast-operator limit is $1 - \pi$, not one; it is one only if the denominator counts exposure to attackers without access. The corollary was true under conditions it left unstated. Read “exposure concentrated among slow operators” as a small share α held by fast operators, and “coverage among fast ones” as coverage at most $\bar{c}$ for everyone else. Then (iv) is the corollary, whatever the window. Under convex harm, “worth at most its length” fails in harm units, because the head start removes the most expensive slice of exposure; (v) restores a bound. The factorization behind $c_j E[\min(W, \tau_j)]$ also needs A9, since coverage and the window can be correlated (appendix). The closed form resembles the restricted-mean-lifetime identity of survival analysis, and we do not claim it as new.

P7, fixes are roadmaps. *In plain terms:* a published patch tells attackers where the flaw is. For flaws the holder finds and fixes, holding the whole capability adds nothing over a flaw-level embargo if attackers would not independently discover them before patch-enabled exploitation. But for flaws it never finds, holding the capability delays attackers’ own search.

We compare two regimes, flaw by flaw, with explicit dates: *hold*, in which attackers without access obtain the capability at W , and *release*, in which they obtain it at 0. Exposure to a flaw is the time between the first usable exploit and the operator’s deployed fix, and loss is linear in it.

Proposition 7. (a) *Flaws the holder finds and patches for outsiders.* The patch is published at P_f under both regimes and the operator deploys at $P_f + \tau_j$. The patch-to-exploit lag ℓ_x is the same under both regimes, with or without the capability. If attackers’ own discovery under release, d_f , comes no earlier than $P_f + \ell_x$, exposure is $(\tau_j - \ell_x)^+$ under both regimes, and holding the capability adds nothing over the flaw-level embargo. If under release attackers would have found the flaw before the patch-enabled exploit, $d_f < P_f + \ell_x$, holding lowers exposure weakly, and strictly if $\pi < 1$, $W > 0$ and $d_f < P_f + \tau_j$. (b) *Flaws in outsiders’ code the holder never finds.* Exposure runs from the attackers’ discovery to ζ_j . Holding delays discovery by attackers without access by W , saving $(1 - \pi)\min(W, (\zeta_j - d_f)^+) \leq (1 - \pi)\min(W, \zeta_j)$, with expectation $(1 - \pi)(1 - e^{-\lambda(\zeta_j - d_f)^+})/\lambda$ for an exponential window. This is P6’s bound again. (c) *Defensive-only tools.* Hold publication dates and exploit onset fixed, and let the tool change only how fast operators deploy published patches. Then withholding it never lowers outsiders’ exposure, and it raises exposure strictly if and only if the tool would shorten deployment of some flaw during an exposure interval with positive loss.

A lag that shortens with the capability. The common lag in (a) is an assumption. If the lag shortens once attackers hold the capability, as the harness result in Section 6.1 suggests, a hold also protects flaws the holder has found and patched, for at most the window. The results above keep the common lag; Section A.7 lists the relaxation.

As first stated, P7 had two sentences. The first said that if exposure begins at patch publication, a capability-level hold adds nothing for outsiders beyond a flaw-level embargo. As a conditional it is true by construction: its antecedent makes exposure start at

publication, which is case (a). As an unconditional policy claim it is false whenever the capability is a live attack channel: with 100 flaws, 90 of them never found, the hold cuts expected exposure from 435 to about 413.6 though no patch ships for the 90. In closed form, with $m_\lambda(u) = (1 - e^{-\lambda u})/\lambda$, release gives 435.000000 and the hold, with A8's mixture applied to every flaw, $435 - 0.7[10m_3(0.06) + 90m_3(4.8)] = 413.615642$ (Section A.6 gives the inputs). An alternative composition, which treats the ten patched flaws as wholly excluded from attack during the hold, gives 413.450912; the counterexample survives either way. Those 90 lie outside the antecedent, so the example refutes the policy reading, not the conditional, and P7(b) extends the claim to them. The second sentence said that holding defensive tools makes outsiders strictly worse off. That is true only weakly, under (c); a tool that does not shorten deployment changes nothing. The two sentences need not conflict: under a fixed publication schedule they are consistent, the second only weakly. They pull apart when the tool produces patches and withholding it delays exposure's onset and completion equally, leaving its duration unchanged: discounted exposure falls. The disclosure trade-off is familiar from work associated with Arora, Telang and Xu, and we claim nothing new about it.

P8, the best coalition. *In plain terms:* each partner admitted to a shared capability adds coverage and leak risk together. Whether some intermediate coalition is best depends on how the two change at the ends of the range.

Let $B(n) = (1 - \pi) \sum_j c_j(n) w_j m(\lambda(n), \tau_j)$, where $m(\lambda, \tau) = (1 - e^{-\lambda \tau})/\lambda$, $m_\lambda < 0$, and $\lambda(n)$ includes $\lambda_r(n) + \lambda_\ell(n)$.

Proposition 8. (i) $B'(n) = (1 - \pi) \sum_j w_j [c'_j(n) m_j + c_j(n) m_\lambda \lambda'(n)]$. A lemma: one differentiation from the definition. (ii) For n continuous on $[0, N]$, an interior maximum exists if $B'(0) > 0 > B'(N)$. It is unique if B is strictly concave, or, more weakly, if $B' > 0$ on $[0, n^*)$ and $B' < 0$ on $(n^*, N]$, which excludes a flat stretch at the top. (iii) For integer n , the maximizer is interior if $B(1) > B(0)$ and $B(N-1) > B(N)$. If the differences $B(n+1) - B(n)$ are strictly decreasing, there are at most two maximizers, and they are adjacent; the maximizer is unique if, in addition, no difference is zero. (iv) Hold coverage, deployment times, attack access and beneficiaries fixed, and let $\pi < 1$ and covered exposure $\sum_j c_j w_j \tau_j$ be positive. Then admitting a rival raises B if and only if it lowers λ : the rival-release hazard removed exceeds the leak hazard added. Removing the rival's release hazard requires that the rival commit not to release (A13).

As first stated, P8's interior optimum does not follow from rising coverage and a shrinking window: either corner can be best (appendix). The rival rule was true only with everything except hazards held fixed; if admission also covers the rival's own systems, B can rise while net hazard rises too. In the integer case a zero difference produces a tie: $B = (1, 2, 2, 1)$ has strictly decreasing differences and two maximizers.

A.3.3 Racing (P9–P10)

P9, secrecy and the private race. *In plain terms:* a secret window raises racing only as far as it widens the gap between what the winner and the loser get. Whether that is bad for the world depends on what the extra racing costs, what it produces, and how rivals respond.

We use a two-player contest in which player i wins with probability $x_i/(x_i + x_j)$, with an equal split when both efforts are zero (A10). The payoff is $p_i v_W + (1 - p_i)v_L - cx_i - L\theta(x_i + x_j)$, with $v_W > v_L$ and $c + L\theta > 0$. The winner's payoff v_W includes the positional stake D ; outsiders are not players.

Proposition 9. (i) The unique pure-strategy equilibrium has $x^* = (v_W - v_L)/(4(c + L\theta))$. With N symmetric players, it is $(v_W - v_L)(N - 1)/(N^2(c + L\theta))$. (ii) With baseline gap $v_W - v_L = D$, a window worth v_s to the winner only multiplies equilibrium effort, and the expected-loss index θX^* , by $(D + v_s)/D$. A window that both players receive leaves effort unchanged. A window of which the loser gets v_s^L multiplies effort by $(D + v_s - v_s^L)/D$. (iii) Let G be the window's real gain to everyone except its holder, before any payments to it (sub-model Hold), and let b be the uncaptured public value of racing. The prize v_s that drives effort is the holder's real surplus σ_s , such as its own security, plus any payments from outsiders. Welfare counts every party's real gains and losses in one unit, as the social loss of P1–P5 does; payments are transfers and drop out. The welfare change from a winner-only window is

$$\Delta W = G + \sigma_s - (c + L_A\theta - b) \cdot v_s/(2(c + L\theta)).$$

A pure transfer has $\sigma_s = 0$. If rivals also change the expected-loss index by ξ , a further $-L_A\xi$ enters, and the window is harmful if and only if $\xi > [G + \sigma_s - (c + L_A\theta - b)\Delta X]/L_A$, with ΔX the added total effort and σ_s summed over holders.

As first stated, P9's multiplier was true under unstated conditions: a winner-only window, exponent one, linear cost, and a catastrophe term linear in total effort. With a hazard convex in total effort (θX^2), the multiplier is 1.38, not 1.5. Its second sentence was false: a private window value larger than the world's gain does not make secrecy harmful. With a window both players receive, racing is unchanged and welfare rises; with a winner-only window and public value b larger than $c + L_A\theta$, it rises too, whether v_s is surplus or a transfer (appendix). The third sentence was ill-posed, because the rival response, the far stake and the cyber gain were in different units. Its conclusion, that "even a modest escalation dominates", is an empirical premise stated as a conclusion, not a theorem. With a window both players receive, so that $\Delta X = 0$, at $\xi = 0.005$, $L_A = 100$ and $G = 1$, welfare still rises, by at least $G - L_A\xi = 0.5$ whatever the window's private value; harm needs $\xi > 0.010$ with transfers and $\xi > 0.020$ if each holder's $v_s = 0.5$ is real surplus. The effort formula is the textbook Tullock comparative static, and the race between holders resembles the memoryless patent races of Loury, Lee and Wilde, and of Dasgupta and Stiglitz and Reinganum. We claim nothing new for any of it.

P10, the private race in general. *In plain terms:* rational racers race more than is good for them jointly when the race decides who bears a loss and adds a shared risk. Verification can support restraint, but only with enough detection and a credible penalty.

Proposition 10 (restated from an earlier draft's contest model; the game is stated in full in the proof). Let $D > 0$. (i) With a shared expected-loss index θX and loss L to every racer, the equilibrium effort is $D/(4(c + L\theta)) > 0$, while the racers' joint optimum is zero. The welfare criterion is the racers' joint payoff; adding outsiders' losses strengthens the result.

(ii) With quadratic cost $kx^2/2$ and public value b per unit of total effort to every racer, let $\tilde{b} = b - L\theta$. The equilibrium over-races if and only if $D > 8\max(\tilde{b}, 0)^2/k$, and under-races if $\tilde{b} > 0$ and $D < 8\tilde{b}^2/k$. (iii) In the repeated game with per-period discount factor $0 < \varrho < 1$, detection with probability ν after each deviation, no false alarms, and permanent reversion to the stage equilibrium after detection, restraint is an equilibrium if and only if $\nu \geq \nu^* = ((1 - \varrho)/\varrho) \cdot 2(c + L\theta)/(c + 2L\theta)$. When $\nu^* > 1$, as for small ϱ , no detection technology supports restraint; under full secrecy ($\nu = 0$) it fails. In a one-shot race, a sanction Q_{det} on detection supports restraint if and only if $\nu Q_{\text{det}} \geq D/2$. (iv) *Bounds and best responses (Section 9)*. For any contest rule that awards D with some probability in $[0, 1]$, with cost cx , $c > 0$, and a loss term that does not fall with own effort, every $x > D/c$ earns strictly less than $x = 0$; a cost that merely rises need not exclude any effort. In the stage game of (i), with $\Delta v = D$ and $\hat{c} = c + L\theta$, the best response to rival effort $x > 0$ is $BR(x) = \max\{0, \sqrt{\Delta v \cdot x/\hat{c}} - x\}$, which varies with x , so no effort is dominant; against $x = 0$ no best response exists. BR rises on $(0, x^*)$, where $x^* = \Delta v/(4\hat{c})$, and $\bar{x} < BR(\bar{x}) < x^*$ for $0 < \bar{x} < x^*$. So a unilateral cap $0 < \bar{x} < x^*$, with the rival at its best response, gives total effort $\sqrt{\Delta v \cdot \bar{x}/\hat{c}} < 2x^*$, rising in $\bar{x}$, and a winning chance for the capper of $\sqrt{\hat{c}\bar{x}/\Delta v} < 1/2$.

As first stated, P10 said that equilibrium speed exceeds the social optimum. That is false when racing has uncaptured public value, as (ii) shows for $D < 8\tilde{b}^2/k$. It also said that verification is the condition for restraint. Read as a necessary condition within (iii), that is true. Read as a sufficient one, it is false: with $\nu = \nu^*/2 > 0$, deviation still pays. The restraint result is the grim-trigger logic of repeated games with imperfect detection, and we claim nothing new for it.

A.3.4 The holder's program (P11–P15)

P11, release timing. *In plain terms:* holding earns a flow and risks losing everything to a leak. Releasing earns inference rents that copying then erodes. Whether faster copying hastens release depends on which copying is faster.

Set-up (sub-model Program). The holder's current public model has a lead $g(t) = g_0 e^{-\kappa_{\text{pre}} t}$ over the best copy, eroded by copying before release; the held model itself is not copied. Inference rent is $\rho(g)$, increasing, with $\rho(0) = 0$, and C^2 on $(0, g_0 + \Delta]$. While holding, the holder earns $F + \rho(g(t))$. Releasing at T raises its lead to $g(T) + \Delta$, which then decays at κ_{post} , the copying after release. The release value is $J(T) = \Pi(g(T) + \Delta)$, where $\Pi(x) = \int_0^\infty e^{-rs} \rho(xe^{-\kappa_{\text{post}} s}) ds$, which satisfies $r\Pi = \rho - \kappa_{\text{post}} x\Pi'$. Release ends the held flow. The hold can also end involuntarily, at hazard λ , when a leak, theft, open-weights parity or a rival's release makes the capability public; the holder's lead then falls to zero, and it keeps only a continuation value Z , durable assets or zero.

Proposition 11. (i) A planned release at T is worth $V(T) = \int_0^T e^{-(r+\lambda)t} [F + \lambda Z + \rho(g(t))] dt + e^{-(r+\lambda)T} J(T)$. Then $V'(T) = e^{-(r+\lambda)T} \Phi(T)$, with $\Phi(T) = F + \lambda Z + \rho(g(T)) - (r + \lambda)J(T) + J'(T)$. An interior release satisfies

$$F + \lambda Z + \rho(g(T)) = (r + \lambda)J(T) - J'(T),$$

which reduces to $F = (r + \lambda)J(T) - J'(T)$ when $Z = 0$ and there is no old public model. It is the global optimum if Φ crosses zero once, from above, along the path. If the zeros of Φ are isolated, the optimum is the best of $T = 0$, the dates where Φ crosses zero from above, and holding until involuntary loss; in general, compare all critical dates and both ends. This holds under A11; if ρ jumps, the dates at which $g(T)$ or $g(T) + \Delta$ crosses a jump point join the candidates (proof). (ii) *Constant flows*. With no old public model, Φ is constant, and the problem is bang-bang. The holder releases at once if $F + \lambda Z < (r + \lambda)\Pi(\Delta)$, never releases voluntarily if the inequality is reversed, and is indifferent among all release dates at equality. Faster post-release copying lowers $\Pi(\Delta)$ and pushes toward holding. (iii) *Linear rents*, $\rho(x) = Mx$. Here $\Phi = F + \lambda Z - (r + \lambda)M\Delta/(r + \kappa_{\text{post}}) + Mg(T)(\kappa_{\text{post}} - \kappa_{\text{pre}} - \lambda)/(r + \kappa_{\text{post}})$. A strict interior maximum requires $\kappa_{\text{post}} > \kappa_{\text{pre}} + \lambda$; at equality, with $F + \lambda Z = (r + \lambda)M\Delta/(r + \kappa_{\text{post}})$, $\Phi \equiv 0$ and every date is optimal. With one copy rate the optimum is a corner. (iv) *Concave rents*, $\rho(x) = Mx^\eta$ with $0 < \eta < 1$. Interior releases can arise, because the gain from releasing, $\rho(g + \Delta) - \rho(g)$, grows as copying erodes the old lead. (v) *Comparative statics at an interior optimum where Φ crosses zero simply and from above, other primitives fixed*. Release comes later with richer internal use that does not raise any hazard, if internal research's value ends at release, as assumed in Section 8.2. It comes earlier with faster copying of the public model before release. A higher discount rate r brings release sooner at any interior optimum with linear rents; with power rents its effect can have either sign. It comes later with faster copying of the released model, with λ held fixed, whenever $x\rho'(x)$ is non-decreasing; linear and power rents both satisfy this. The exhibited saturating reversal is evaluated at $\kappa_{\text{post}} = 0$; this does not mean reversals occur only there. Richer artifact sales delay release if and only if $\lambda'(f_a)(J(T^*) - Z) < 1$, with $\lambda'(f_a)$ the hazard added per unit of artifact flow; with $Z = 0$ this is $\lambda'(f_a)J(T^*) < 1$.

The sign of κ . With one copy rate for both phases, the two channels oppose each other and the net sign is not determined. Faster copying delayed release in all 56 interior cases of our grid, and in 148 of the 161 selected draws of a wider random scan, all of them interior, against 13 in which it brought release sooner. These are frequencies over the ranges we sampled, which Section A.6's table lists with the selection rule, not a property of the model. In the bang-bang case of (ii), faster copying can stop release altogether.

As first stated, P11 was false in three of its four parts. "Release when the inference revenue forgone exceeds F " compares a flow with a flow and leaves out capitalization and leak risk; a forgone flow above F can still make holding optimal (appendix). "Sooner the faster the copying" is false. In an interior case with one copy rate ($\eta = 0.3$, $\Delta = 0.2$, $g_0 = 0.2$, $\kappa = 0.3$, $\lambda = 0.1$), $dT^*/d\kappa = +86$: faster copying delays release. A constant-flow example, in which faster copying ends voluntary release, has no old public model and so lies outside the original's mechanism. "Later the richer the artifact uses" is true for internal use, and false for artifacts when theft responds to earnings: with $\lambda' = 0.75$, $dT^*/df_a = -45.5$ while $dT^*/df_u = +122$. The corollary, that the realized window is the shorter of the leak and release times, is true by definition. Optimal stopping of this kind is standard real-options analysis, as in Dixit and Pindyck, and we claim no novelty for the method.

P12, perishability and the decisive prize. *In plain terms*: a lead that others copy is worth less than one they cannot.

Proposition 12. For $r > 0$, a unit flow eroding at rate κ is worth $1/(r + \kappa)$, and a non-eroding one $1/r$. Their ratio $(r + \kappa)/r$ rises in κ .

As first stated, P12 was a present-value identity, not a result, plus a premise that the decisive prize does not erode. It needs $r > 0$. It concerns cash flows: under P11's power rents, a lead eroding at κ yields revenue eroding at $\eta\kappa$, worth $Mx^\eta/(r + \eta\kappa)$. As copying speeds up, the decisive prize keeps its absolute value; only its value relative to an eroding lead rises. "The pull toward a decisive prize rises with κ " is therefore a statement about relative valuation, not about pursuit: in separate contests, faster copying cuts incremental effort and leaves decisive effort unchanged. The identity is the firm-value term of Aghion and Howitt's creative-destruction model, and we do not claim it as new.

P13, what to sell while holding. *In plain terms:* a new use of a held capability is worth adding if what it earns, per unit of leak risk it adds, exceeds what a leak would destroy.

Proposition 13. With constant flows, a hold value $F/(r + \lambda)$, no voluntary release and no continuation after loss, adding a use with flow f and added hazard δ raises the hold value if and only if $f/(r + \lambda) > F\delta$. For $\delta > 0$ this is $f/\delta > F/(r + \lambda)$; for $\delta = 0$, any $f > 0$ passes. (i) With continuation Z after loss, unchanged by the use, the hold value is $V_H = (F + \lambda Z)/(r + \lambda)$, and for $\delta > 0$ the threshold is $f/\delta > (F - rZ)/(r + \lambda) = V_H - Z$; for $\delta = 0$, any $f > 0$ still passes. In words, earnings per unit of added hazard must exceed what a leak destroys. (ii) If theft responds to earnings, so that the use adds hazard $\delta + \theta_F f$, the condition is $f/(r + \lambda) > F(\delta + \theta_F f)$, and with continuation Z , unchanged by the use, $f/(r + \lambda) > (\delta + \theta_F f)(F - rZ)$. (iii) With a live release option, the hold value is the wrong opportunity cost, and the rule must be re-derived from P11.

As first stated, P13's ratio rule, $f/\delta > F/\lambda$, was stated for an undiscounted hold, and on that domain it is exact under unstated conditions: no release option, no continuation, and δ including every induced change in hazard. With $r > 0$ the threshold becomes $F/(r + \lambda)$, an extension rather than a refutation of the original algebra. Its two qualitative corollaries, that interactive products pass only when the hold is nearly worthless and that artifacts in distant domains almost always pass, were false: each depends on magnitudes the model does not supply (appendix). With a live release option, the rule can approve a use that lowers the firm's value, if the holder commits to the use for the rest of the hold. In P11's interior example, a use that the rule approves lowers the optimal value (appendix), because the added hazard also destroys the release the holder was waiting for. A purely optional use could not lower the optimal value, since the holder could decline it; what fails is the rule that approves it.

P14, the value of a secret is capped by those who want to take it. *In plain terms:* if a secret's earnings attract theft in proportion, earning more makes it a bigger target, and its value cannot exceed a ceiling. A ceiling exists if and only if theft eventually rises at least in proportion to what the secret earns.

Proposition 14. Let $r + \lambda_0 > 0$ and $\lambda_\theta(F) \geq 0$. (i) The hold value $V(F) = F/(r + \lambda_0 + \lambda_\theta(F))$ is bounded over $F \geq 0$ if and only if $\liminf_{F \rightarrow \infty} \lambda_\theta(F)/F > 0$. (ii) If $\lambda_\theta(F) \geq \theta_F F$ for fixed $\theta_F > 0$, then $V < 1/\theta_F$ at every F . With $\lambda_\theta = \theta_F F$ it rises in F toward $1/\theta_F$. The cap

survives discounting. (iii) Under (ii), with durable assets $Z \geq 0$ that survive leakage, $V = F/(r + \lambda) + [\lambda/(r + \lambda)]Z < 1/\theta_F + Z$ at every F .

As first stated, P14 was false as worded, because a hazard that merely rises with the flow does not cap value; it is true for its linear example and for any $\lambda_\theta \geq \theta_F F$. With theft proportional to $\sqrt{F}$, which rises with the flow, value is unbounded, as it is even for $\lambda_\theta = \theta_F F/(1 + \ln(1 + F))$; (i) gives the exact condition. The cap also needs θ_F fixed rather than chosen, and value that a leak extinguishes. If theft responds to capitalized value V rather than flow, $V = F/(r + \lambda_0 + \theta_F V)$ grows like $\sqrt{F/\theta_F}$. If the holder can buy security cheaply enough, value grows without bound in F . “Its value is therefore capped” needs the proportionality premise, which the RAND assessment of model-weight security informs but does not establish.

P15, durable assets. *In plain terms:* if a step’s earnings can be turned into assets that copying cannot erode, a temporary lead is worth more. The decisive prize then looms relatively smaller, and each step’s race grows. What that does to total risk depends on how effort is allocated.

Proposition 15. For $r > 0$ and a durable share $d \in [0, 1]$, a unit of step earnings is worth $V(d) = (1 - d)/(r + \kappa) + d/r = (r + d\kappa)/(r(r + \kappa))$. We model the durable share as a perpetual, non-eroding component of earnings from the start; reading it as banked temporary earnings would need a conversion technology and capital accounting, which we do not model. Changes in d are strict when $\kappa > 0$; at $\kappa = 0$, $V(d) = 1/r$. (a) $(1/r)/V(d) = (r + \kappa)/(r + d\kappa)$, which falls in d . (b) In a step contest whose winner alone gets $mV(d)$, $m > 0$, equilibrium effort $mV(d)/(4(c + L\theta))$ rises in d . (c) Suppose a firm splits effort between a step contest with prize $D_s \geq 0$ and a decisive contest with prize D_d , at joint cost $k(x_s + x_d)^2/2$, with $k, D_d > 0$. Racers bear no catastrophe loss here; $\theta_s, \theta_d > 0$ are external weights converting each contest’s effort into a linear expected-loss index. For $D_s > 0$ the first-order conditions give $x_s = D_s/(4kY)$, $x_d = D_d/(4kY)$ and $Y = \sqrt{(D_s + D_d)/(4k)}$, where $Y = x_s + x_d$ is each firm’s total effort and $X_s = 2x_s$ and $X_d = 2x_d$ are totals across the two firms; at $D_s = 0$, $x_s = 0$ and the same formulas hold. Decisive effort falls as D_s rises, and total risk $\theta_s X_s + \theta_d X_d$ rises with D_s if and only if $\theta_d/\theta_s < 2 + D_s/D_d$. With separate budgets, decisive effort $\sqrt{D_d/(4k)}$ does not depend on d .

As first stated, P15’s valuation was an identity, exact per unit of flow under exponential erosion. Part (a) is true as a statement about relative value. It is not a result about reduced pursuit of a decisive lead, which needs substitution between efforts as in (c). Part (b) was true under unstated conditions. Durable rents must widen the winner-loser gap, and the step stake must exclude continuation value from the decisive prize; otherwise step effort or the stake can fall in d (appendix). The total-risk sign stays unsigned without an effort-allocation model, and (c) gives the condition in one such model. The valuation draws on the appropriation literature associated with Arrow, Nordhaus and Teece, and the logic of locked-in customers resembles Klemperer’s switching costs. Neither (a) nor (b) is claimed as new.

A.3.5 Health benefits on a second adoption clock (P16)

P16, benefits and two adoption clocks. *In plain terms:* if the capability that raises attackers' hazard also brings a health benefit, speed earns a second return, an earlier benefit. That return exists only while the benefit waits on capability rather than on its own trials, approval and uptake, and acceleration helps only if the return outweighs the cyber loss speed adds.

Set-up (sub-model Speed, extended). For this extension only, health benefits and cyber harm are counted in DALYs averted, so that mortality and morbidity enter together. The benefit is a constant flow $y > 0$ DALYs averted per year, and $V_c > 0$ is DALYs per harmful cyber event (A15). All distances and adoption times are positive, and one speed $s > 0$ advances both distances K and K_B (A16). With $\tau = K/s$, $q = K_B/s$ and shares m , $m_B \in [0, 1]$,

$$T = \max\{\tau, (1 - m)T_a\} + mT_a, \quad t_B = \max\{q, (1 - m_B)T_b\} + m_B T_b,$$

where a share of 0 is parallel timing and a share of 1 sequential. T_a , T_b , m and m_B are fixed in speed (A3, A17); P5's adoption that responds to speed is not assumed. Away from clock kinks, $\chi_c = 1$ if the capability term binds in T and 0 otherwise, and χ_B likewise for t_B ; under sequential timing the indicator is one at every finite speed. Thus $sT' = -\chi_c\tau$ and $st'_B = -\chi_B q$. m_B is the share of the benefit's adoption that waits for capability, not the capability-set share of the total wait.

Undiscounted benefits are counted over a fixed calendar horizon $L > 0$: $B_L(s) = y(L - t_B(s))^+$ and $W_L(s) = B_L(s) - V_c C_c(s)$. To read W_L as one calendar ledger, take $L \geq \sup T(s)$ over the speed interval considered and include the step losses within L . Otherwise W_L is a complete-transition cyber ledger with benefits counted only through L ; a common shorter horizon would need the cyber integral truncated too, which adds a kink the formulas below do not cover. Write $\eta_L = 1$ if $t_B < L$ and 0 otherwise. The speed-set specification is P1's loss index, $C_c = H(s)T(s) + S$, with $H > 0$ continuously differentiable, $H' > 0$ and constant $S \geq 0$. The level-set alternative is the separately scoped calculation of Section A.7: capability path $c(t) = \min(st, K)$, a continuous nonnegative hazard $h(c)$, held at $h(K)$ through the adoption tail, so that $C_c = \tau\bar{h} + (T - \tau)h(K) + S$, with $\bar{h} = K^{-1} \int_0^K h(c)dc$. It is not obtained by substituting $H = h(K)$ into the speed-set model. Each equation is an expected-event or linear expected-health-loss calculation, not a sum of catastrophe probabilities; there are no flaw-stock dynamics, no benefit flow that responds to harm, no survival conditioning, no defensive response and no new far-risk model. A cyber harm counted in $V_c C_c$ is not also subtracted from y (A18).

Proposition 16. Under this set-up, at a speed away from clock and horizon kinks:

(i) *Undiscounted.* For the speed-set hazard, $sW'_L = yq\chi_B\eta_L - V_c H(\varepsilon_H T - \chi_c \tau)$. For the level-set hazard, $sW'_L = yq\chi_B\eta_L - V_c \tau[(1 - \chi_c)h(K) - \bar{h}]$. Acceleration improves the balance exactly when the relevant right side is strictly positive, worsens it when strictly negative, and is marginally neutral at equality. Above a finite cyber kink, while benefit arrival is capability-set and inside the horizon, the speed-set test is $yq > \varepsilon_H V_c H T_a$ and the level-set test is $yq > V_c \tau[h(K) - \bar{h}]$. Beyond a parallel or mixed benefit kink the benefit term is zero

and P1's cyber sign returns; sequential benefit timing has no finite cutoff. Below a cyber kink the speed-set test is $\varepsilon_H < [\tau + yq\chi_B\eta_L/(V_cH)]/T$, which can hold when $\varepsilon_H > 1$; without an active benefit term it is $\varepsilon_H < \tau/T$, the criterion of Section 4.3.

(ii) *Discounted.* For $r > 0$, discount health gains and cyber losses from the same date. With $A_r(u) = (1 - e^{-ru})/r$, the perpetual discounted benefit $B_r = (y/r)e^{-rt_B}$, $C_{c,r} = HA_r(T) + S_r$ and $W_r = B_r - V_cC_{c,r}$,

$$sW'_r = yq\chi_B e^{-rt_B} - V_cH[\varepsilon_H A_r(T) - \chi_c \tau e^{-rT}] - V_c sS'_r.$$

For a finite horizon, replace B_r by $B_{r,L} = (y/r)(e^{-rt_B} - e^{-rL})^+$ and multiply its derivative term by η_L ; the cyber terms are unchanged when the transition and the steps counted fall within L . With $S'_r = 0$, the speed-set threshold is $\varepsilon_H < [\chi_c \tau e^{-rT} + (yq\chi_B\eta_L/(V_cH))e^{-rt_B}]/A_r(T)$, with $\eta_L = 1$ in the perpetual case; with dated steps, subtract sS'_r/H from the numerator. Above the cyber kink and on the benefit's capability branch, the test is $yqe^{-rt_B} > V_cH\varepsilon_H A_r(T_a) + V_c sS'_r$. The discounted level-set test is (C3) of Appendix C, not this formula.

(iii) *Kinks and optimizers.* On a closed bounded interval $I = [s_{\min}, s_{\max}]$ with $s_{\min} > 0$, each objective attains a maximum and a minimum if its primitives and S_r are continuous. An optimizer is an endpoint, a clock or horizon kink, or a smooth stationary point, possibly part of a flat set. At a kink s_0 , a local maximum requires $W'_-(s_0) \geq 0 \geq W'_+(s_0)$, and strict inequalities suffice; reversed inequalities give the conditions for a minimum. At coincident kinks, use all the simultaneous changes. An optimum above the cyber kink is possible, not guaranteed.

The timing combinations. Each clock enters the derivative only through its indicator and its duration, so the four branch cells of Section 4.7's table cover all nine pairs of cyber and benefit timing and both hazards; in the mixed rows $T = \tau + mT_a$, not τ . The speed-set cell with adoption setting both clocks is strictly negative. In the level-set column, the cell with capability setting the cyber clock and adoption the benefit is positive if $\bar{h} > 0$; the cell with adoption setting both is negative if $h(K) > \bar{h}$, as for any nondecreasing, nonconstant continuous h , zero at equality, and positive if the endpoint hazard is below the average. The first table gives each clock's kink; the second, which branches occur in which order.

Timing	Cyber clock: kink; duration below it and above it	Benefit clock: kink; duration below it and above it
Parallel	K/T_a ; $T = \tau$, then T_a	K_B/T_b ; $t_B = q$, then T_b
Mixed, share strictly between 0 and 1	$K/((1 - m)T_a)$; $T = \tau + mT_a$, then T_a	$K_B/((1 - m_B)T_b)$; $t_B = q + m_B T_b$, then T_b
Sequential	No finite kink; $T = \tau + T_a$	No finite kink; $t_B = q + T_b$

Timing pairs, cyber/benefit	Branches as speed rises
Parallel or mixed on both clocks, cyber kink first	Capability sets both; then adoption sets the cyber clock and capability the benefit; then

Timing pairs, cyber/benefit	Branches as speed rises
	adoption sets both
Parallel or mixed on both clocks, benefit kink first	Capability sets both; then capability sets the cyber clock and adoption the benefit; then adoption sets both
Parallel or mixed on both clocks, coincident kinks	Capability sets both below the kink; adoption sets both above it
Parallel or mixed cyber, sequential benefit	Capability sets both below the cyber kink; adoption sets the cyber clock and capability the benefit above it
Sequential cyber, parallel or mixed benefit	Capability sets both below the benefit kink; capability sets the cyber clock and adoption the benefit above it
Sequential on both	Capability sets both at every speed

The horizon. If $L \leq T_b$, the benefit yields no positive flow within the horizon, at any speed and under any timing. If $L > T_b$, it first arrives inside the horizon as speed crosses $k_L = K_B/(L - m_B T_b)$, on its capability branch and below any finite benefit kink; insert that activation kink into either ordering. At $L = T_b$, reaching the plateau at the horizon activates nothing.

Kinks. At a kink s_0 , take all primitives at s_0 and substitute the indicators of each side. The change in sW' from the lower-speed side to the higher is: at a cyber kink alone, $-V_c H \tau$ for the speed-set hazard and $-V_c \tau h(K)$ for the level-set one; at a benefit kink alone, $-yq$ under either hazard if arrival is inside the horizon, and zero if it is not; at coincident cyber and benefit kinks, the sum of the two; and at the horizon-activation kink alone, $+yq$. For discounted objectives, multiply the cyber change by e^{-rT} , the benefit change by e^{-rt_B} and the activation change by e^{-rL} . These assume no further kink in H or S_r ; where there is one, its one-sided derivatives enter too. No sign is assigned to a kink itself.

An illustrative optimum and a counterexample. With the illustrative values of Section 4.7, not estimates ($K = T_a = V_c = y = T_b = 1$, $K_B = 4$, parallel timing on both clocks, $H = s$, $S = 0$, $L = 10$, $I = [0.5, 6]$), cyber completion and benefit arrival fall within the horizon throughout, the cyber kink is $s^* = 1$ and the benefit kink is 4. Then $W_L = 9 - 4/s$ on $[0.5, 1]$, $10 - 4/s - s$ on $[1, 4]$ and $9 - s$ on $[4, 6]$. The derivative is $4/s^2 > 0$ below 1, $4/s^2 - 1$ on $(1, 4)$ and -1 above 4. The unique global maximum is $s_0 = 2 > s^* = 1$, with $W_L = 6$, against 1, 5, 5 and 3 at $s = 0.5, 1, 4$ and 6. The one-sided derivatives are 4 and 3 at $s = 1$, and $-3/4$ and -1 at $s = 4$. Cyber loss is rising at the optimum; the benefit balances its marginal rise. The same example refutes “benefits always justify speed”: at $s = 5$, $B_L = 9$, but its derivative is zero, so $W'_L = -1$, and the balance falls from 4 to 3 as speed rises from 5 to 6. With $H = s^2$ instead, at $s = 0.5$, $\varepsilon_H = 2 > 1$, $C'_c = 1 > 0$, $B'_L = 16$ and $W'_L = 15 > 0$: benefits can outweigh a rising cyber loss below the cyber kink even when hazard rises more than proportionally with speed.

Two readings that fail. Benefits in levels do not justify speed: the example at $s = 5$ has a large benefit and a falling balance. And the left side of the above-kink test is not the benefit flow times the whole wait for the benefit. It is yq . On the benefit's capability branch the capability-set share of the total wait is $-d\ln t_B/d\ln s = q/(q + m_B T_b)$, so the test can also be read as $y t_B [q/(q + m_B T_b)] > \varepsilon_H V_c H T_a$; for parallel timing the ratio is one, and for sequential timing $m_B = 1$. In one case of our checks, $m_B = 0.6$ goes with a capability-set share of 0.806.

P4 and the objective. With a far loss in the same health units, $W'_{\text{full}} = B' - V_c C'_c - V_A C'_A$. For $C'_A > 0$, acceleration improves this extended objective exactly when $B' - V_c C'_c > V_A C'_A$. P4(i)'s requirement that the near cyber loss fall is a condition for its two-loss objective; with benefits, near cyber loss may rise while the balance improves. No new far-loss, catastrophe-composition or cross-model welfare result is supplied, and P1–P15 keep their own objectives and assumptions.

Adoption levels and the marginal return to speed. At fixed s , shortening T_b weakly advances arrival and weakly raises the benefit level, strictly only when it advances arrival during a period that is counted. With $t_B < L$, $\partial B_L / \partial T_b = -y m_B$ on the capability branch and $-y$ on the adoption branch; under parallel timing a shorter T_b has no level effect while capability binds. On the undiscounted capability branch, $B'_L = y K_B / s^2$ does not depend on T_b , so a shorter T_b does not raise the marginal return to speed while the branch stays the same. Discounted, $\partial B'_r / \partial T_b = -r m_B B'_r$ on an active capability branch: a shorter T_b raises the discounted marginal return when $m_B > 0$, through earlier dating, and leaves it unchanged under parallel timing. That trial and approval capacity lowers T_b without changing H is an assumed intervention mapping, not a result of P16. It raises benefit levels when it advances counted arrival, but it does not by itself establish a favorable speed threshold or rank interventions, and regime changes must be evaluated separately.

A.4 Summary of results

Six results need an argument of more than a line: P1, P3 (corrected), P5 (corrected), P6's aggregate and convex-harm bounds, P7 (restated) and P11 (derived). Even these use standard techniques: elasticity calculus, integrating factors for a linear stock equation, the restricted-mean identity, supporting lines of convex functions, pathwise comparison of dated exposure intervals, and deterministic optimal stopping. Their value is the assembly for this question, not new mathematics. Five are lemmas: P4, P8, P13, P14 and P15(c). P2 is an assumption. P6's closed form, P12 and P15's valuation are identities, and P11's corollary is a definition. P9, P10 and P15(b) restate standard contest results. P16, added in this version, is elementary calculus on P1's set-up with a second adoption clock; Appendix C gives its proofs.

Result	Status	What it assumes	What would break it
P1 speed threshold	Substantive (elementary)	A1–A4; constant T_a ; speeds bounded below	Adoption that responds to speed (then P5); a hazard of changing shape

Result	Status	What it assumes	What would break it
P2 step term	Assumption	A2	Exploit success or damage that depends on attackers' tools when a step arrives
P3 peak on a path	Substantive	A5; constant φ	Attacker capability $\varphi(t)$ rising faster than the stock falls
P4 two peaks	Lemma	Common units; differentiability; S_r independent of speed	Discrete moves (use levels); costs of speed outside the two losses
P5 adoption escape	Substantive	$T_a = A(s/s_0)^{-\beta}$	Measured β below ε_H removes the escape
P6 head start	Standard identity; substantive bounds	A6–A9	Attack backlog after the window; correlated coverage and leakage
P7 three flaw sets	Substantive (conceptual)	Linear loss; dated events; fixed publication in (c)	Attackers who find flaws before patches even under a hold
P8 best coalition	Lemma	B differentiable in (i)–(ii); commitment for rivals	Coverage gains from admitting the rival
P9 secrecy and racing	Standard	A10; winner-only window; σ_s as surplus	Shared window; convex hazard; public value of racing
P10 private race	Standard (restated)	The stated stage and repeated games; joint-payoff welfare	Uncaptured public value; detection below v^*
P11 release timing	Substantive (derived)	A11, A12; constant λ ; rent shape	Linear rents with one copy rate (corner); post-release copying slower than pre-release
P12 perishability	Identity	$r > 0$; non-eroding decisive prize	A decisive prize that also erodes
P13 marginal use	Lemma	No release option; δ	A live release option;

Result	Status	What it assumes	What would break it
		counts induced hazard	continuation value
P14 theft cap	Lemma	$\liminf \lambda_\theta(F)/F > 0$; fixed θ_F	Theft growing less than in proportion; theft on value; cheap security
P15 durable assets	Identity and standard; lemma in (c)	Winner-only durable rents; separable stake	Durable rents to both firms; continuation value in the step stake
P16 benefits and two clocks	Substantive (elementary); extends P1	A1–A4, A15–A18; fixed T_a, T_b, m, m_B ; arrival inside the horizon for the undiscounted tests	Narrow tools that decouple the benefit from attackers' capability; attacks that cut y ; adoption that responds to speed

A.5 What the corrections imply

Three corrections bear on the larger argument. First, P11 does not support the claim that the inference market's copy clock pushes held capabilities out sooner. At an interior optimum where Φ crosses zero simply, other primitives fixed, copying of the public model before release pushes release earlier, but, for linear and power rents, copying of the released model delays it, by lowering what release is worth; with one copy rate, faster copying delayed release in 148 of the 161 draws of our random scan and brought it sooner in 13, all interior: frequencies under the selection rule of Section A.6's table, not a result. The commercial argument has to be restated: with linear rents, a strict interior maximum needs a released model copied faster than the old one by more than the leak hazard, and concave and convex rents can also produce one. Second, P7(b) shows that a capability-level hold does protect outsiders' unfound flaws, for at most the window. The relevant comparison for policy is therefore hold against release with safeguards that keep the capability from attackers, not hold against a flaw-level embargo. Third, P14's ceiling on the value of a secret exists if and only if theft eventually rises at least in proportion to what the secret earns, and it disappears if theft tracks capitalized value or if security is cheap enough. The thesis that faster capability raises the near peak where adoption binds, under parallel or mixed timing, survives, with P5's condition $\beta > \varepsilon_H$ as the exact terms of the accelerationist's escape. P6's bound on what a held window can do for slow operators also survives, now in quantitative form.

A.6 Proofs

Our replication code, `MODEL/code/check_model.py` (paths are relative to the deposit named in the Declarations), numerically checks selected formulas, parameter cases and counterexamples below. Its printed output is `MODEL/code/check_model_out.txt`. The code uses one random-number generator with seed 20261022, so each random draw depends on the draws before it. The table gives the inputs behind the numbers quoted in the text, as the code sets them, and the outputs as the code prints them. Rents are $\rho(x) = Mx^\eta$ throughout P11 and P13, with $Z = 0$ and one copy rate $\kappa = \kappa_{\text{pre}} = \kappa_{\text{post}}$ unless stated.

Result	Inputs	Printed output
P7, the 100-flaw example	$\lambda = 3, \pi = 0.3$, attackers' discovery $d_f = 0.2$, removal date $\zeta = 5$, patch date $P_f = 0.25$, $\ell_x = 0.01$, $\tau = 0.25$; 10 flaws found and patched, 90 never found	Release 435.000000 and hold 413.615642, closed forms, with A8's mixture applied to every flaw; Monte Carlo estimate 413.576079 (standard error 0.047015) from 200,000 draws of W . Alternative composition, with the ten patched flaws wholly excluded from attack during the hold: hold 413.450912 (closed form), 413.411339 (Monte Carlo). Release is unchanged and the counterexample survives
P8, the interior coalition	Coverage $c(n) = 1 - e^{-n}$, weight 1, $\tau = 3$, $\lambda(n) = 0.5 + 0.3n$, $\pi = 0$, $N = 10$	$B'(0) = 1.554$, $B'(10) = -0.0245$, $n^* = 1.5893$; integer maximizer 2
P9, convex hazard	$D = 1$, $c = 1$, $L = 1$, $\theta = 0.4$ (so $L\theta = 0.4$), winner-only $v_s = 0.5$, hazard θX^2	Multiplier 1.3779
P11, the +86 example	$\eta = 0.3$, $\Delta = 0.2$, $g_0 = 0.2$, $\kappa = 0.3$, $\lambda = 0.1$, $r = 0.05$, $M = 1$, $F = 0.5510843151750464$ (shown to four places as 0.5511 in the code's note), set by the grid's selection rule $F = -[h(g_0) + h(0)]/2$ (next row); it is the first grid case in loop order where common κ delays release	$T^* = 18.957$, $dT^*/d\kappa = +86.229$
P11, the 56-case grid	$\eta \in \{0.3, 0.5, 0.7\}$, $\Delta \in$	56 interior cases: faster κ_{pre}

Result	Inputs	Printed output
	$\{0.2, 0.5\}, g_0 \in \{0.2, 0.5, 1.0\}, \kappa \in \{0.3, 0.6, 1.0\}, \lambda \in \{0.1, 0.3\}, r = 0.05, M = 1$: 108 combinations. Selection: with $h = \tilde{\Phi} - F$, a combination is kept only if $h(g_0) > h(0)$, and $F = -[h(g_0) + h(0)]/2$, so that $\tilde{\Phi}$ is positive at the starting lead and negative as the lead vanishes; it counts if the release date is interior before and after each perturbation of size $10^{-4}\kappa$	brings release sooner in 56, faster κ_{post} delays it in 56, and faster common κ delays it in 56 and brings it sooner in 0
P11, the random scan	400 draws: η uniform on $[0.15, 0.95]$, Δ on $[0.05, 2]$, g_0 on $[0.05, 3]$, κ on $[0.05, 3]$, λ on $[0.02, 1]$, r on $[0.01, 0.2]$, $M = 1$. Selection: a draw is kept if $h(g_0) > h(0)$, with $F = -[h(0) + u(h(g_0) - h(0))]$, u uniform on $[0.05, 0.95]$; it counts if the release date is interior before and after the perturbation	161 draws kept, all interior before and after the perturbation; common κ brings release sooner in 13 and delays it in 148. The solver compares each date with holding in a cancelled form that avoids floating-point loss, and extends the search horizon until a bound on Φ rules out any later maximum
P11 interior example; P13	$F = 0.42, \eta = 0.5, \Delta = 0.3, \kappa = 0.5, \lambda = 0.2, g_0 = 0.2, r = 0.05, M = 1$. P13's committed use: $\delta = 0.02, f/\delta = 2.147$	$T^* = 9.6470, V^* = 2.580742$. With the use: no-release value $2.57443 \rightarrow 2.57462$; optimal value $2.58074 \rightarrow 2.57767$; release date $9.647 \rightarrow 10.490$
P12, separate contests	$r = 0.05, c = 1$, incremental prize $1/(r + \kappa)$, decisive prize 5, $\kappa \in \{0.2, 0.5, 1.0\}$	Incremental effort 1.000, 0.455, 0.238; decisive effort 1.250
P14, sublinear theft	$r = 0.05, \lambda_0 = 0.1, \theta_F = 0.5$, theft hazard $\theta_F \sqrt{F}$, $F = 10^8$	$V = 19,999$, against the linear-theft cap $1/\theta_F = 2$
P14, bought security	As above with theft $\theta_F e^{-\sigma F}$, security σ on a grid 0 to 39.9 at unit cost, $F = 10^6$	$V = 6,666,635$

P1. For $s < s^*$, $R_c = KH(s)/s + S$, and $dR_c/ds = K(sH' - H)/s^2 = (KH/s^2)(\varepsilon_H - 1)$, since $K, H > 0$. For $s > s^*$, $R_c = HT_a + S$ and $dR_c/ds = H'T_a > 0$, so no minimizer lies above s^* . The one-sided derivatives follow by evaluating at s^* and using $K/s^* = T_a$. R_c is continuous on the compact interval $[s_{\min}, s^*]$, so it attains a minimum there, and that minimum is global. If $\varepsilon_H < 1$ on $[s_{\min}, s^*)$, R_c falls on that interval, so s^* minimizes it. Otherwise a minimizer is an endpoint, $s_{\min}$ or s^* , or an interior local minimum. At an interior local minimum $dR_c/ds = 0$, so $\varepsilon_H = 1$. When the points with $\varepsilon_H = 1$ are isolated, dR_c/ds keeps one sign on each side of each such point, because a derivative has the intermediate-value property. A local minimum is then a crossing of 1 from below, and a crossing from above is a local maximum. Without isolation the sign can oscillate on both sides of a strict minimum, as for $R_c = c + S + (s - s_0)^4(2 + \sin(1/(s - s_0)))$ with $c > 0$, which $H = s(R_c - S)/K$ produces with $H' > 0$ near s_0 . It can also vanish on an interval: with $K = 4$, $T_a = 1$, $s_{\min} = 1$ and $H = s \cdot q(s)$, $q(s) = 1 + (2 - s)_+^2/10 + (s - 3)_+^2/10$, $H' > 0$ and $R_c = 4q(s) + S$ below $s^* = 4$, so the minimizers are exactly $[2, 3]$, where $\varepsilon_H = 1$ throughout and no point is a strict crossing. Without the lower bound the infimum need not be attained: with $H = s^{1.4}$, $R_c = Ks^{0.4} + S$ on $(0, s^*]$, and the infimum is approached as $s \rightarrow 0$. ■

P2. $S = \sum_k N_k \psi_k d_k$ has no argument s under A2. It is linear in each of N_k , ψ_k , d_k , and increasing in each when the other two are positive. ■

P3. (i) E is continuously differentiable, with $E' = \iota - \mu E$. If $E'(t_0) > 0 > E'(t_1)$, the maximum of E on $[t_0, t_1]$ is interior, so $h = \varphi E$ rises and then falls. Conversely, rising and then falling requires such dates. At an interior maximum $E' = 0$, so $\iota = \mu E$. If ι and μ are differentiable, so is E' , and $E'' = \iota' - \mu'E - \mu E' = \iota' - \mu'E \leq 0$. (ii) With μ constant and ι differentiable, $E'' = \iota' - \mu E'$, so $(e^{\mu t} E')' = e^{\mu t} (E'' + \mu E') = e^{\mu t} \iota' \leq 0$. Hence $e^{\mu t} E'$ is non-increasing: once $E' \leq 0$ it stays ≤ 0 , and once $E' < 0$ it stays negative. Since $E'(0) > 0$, E' is positive up to some date and non-positive after it, so E rises to at most one peak, possibly flat, and never rises again. If E' stays positive, the path is monotone. (iii) With $E(0) = 0$ and constant $\mu \neq \gamma$, $E(t) = (\iota_1/\mu)(1 - e^{-\mu t}) + (\iota_0 - \iota_1)(e^{-\gamma t} - e^{-\mu t})/(\mu - \gamma)$. Differentiating gives $E' = [-\gamma(\iota_0 - \iota_1)e^{-\gamma t} + (\mu\iota_0 - \gamma\iota_1)e^{-\mu t}]/(\mu - \gamma)$, with $E'(0) = \iota_0 > 0$. $E' = 0$ if and only if $e^{(\mu-\gamma)t} = (\mu\iota_0 - \gamma\iota_1)/(\gamma(\iota_0 - \iota_1))$. If $\mu > \gamma$, the right side exceeds one, since $\mu\iota_0 - \gamma\iota_1 > \gamma(\iota_0 - \iota_1)$ is equivalent to $\mu > \gamma$, and $\hat{t} > 0$. If $\mu < \gamma$, a positive root needs the right side in $(0, 1)$, which holds if and only if $\mu\iota_0 > \gamma\iota_1$. When $\mu > \gamma$, $\mu\iota_0 > \gamma\iota_1$ holds automatically. So a peak exists if and only if $\mu\iota_0 > \gamma\iota_1$, at $\hat{t}$, and $E \rightarrow \iota_1/\mu$ as $t \rightarrow \infty$ along the continued, untruncated trajectory, a limit that lies beyond T_A when T_A is finite. For $\mu = \gamma$, $E(t) = (\iota_1/\mu)(1 - e^{-\mu t}) + (\iota_0 - \iota_1)te^{-\mu t}$, and $E' = e^{-\mu t}[\iota_0 - \mu(\iota_0 - \iota_1)t]$, which changes sign once, at $\hat{t} = \iota_0/(\mu(\iota_0 - \iota_1))$. Counterexamples to the original, on $[0, \infty)$: with $\iota = e^t$ and $\mu = e^{2t}$, introduction is unbounded, yet the stock peaks at $t = 0.73$ and falls by 90% by $t = 3$; with $\iota = 1 + t$ and $\mu = 0.1$, introduction is unbounded and the stock grows without limit, with no plateau. Both need an unbounded horizon. On a finite one, take $T_A = 1$, $\mu = (1 - t)^{-2}$ and $\iota = 1 - 2t + t/(1 - t)$: introduction is positive and unbounded on $[0, 1)$, yet $E = t(1 - t)$ peaks at $t = 1/2$ and falls back toward zero, with no plateau. To verify it, $E = t(1 - t)$ gives $E' = 1 - 2t$ and $\mu E = t/(1 - t)$, so $\iota = E' + \mu E = 1 - 2t + t/(1 - t)$. Its minimum on $[0, 1)$, at $1 - t = 1/\sqrt{2}$, is $2\sqrt{2} - 2 > 0$, and $\iota \rightarrow \infty$ as $t \rightarrow 1$. ■

P4. (i) $d/ds[V_c C_c + V_A C_A] < 0$ if and only if $V_c C'_c < -V_A C'_A$. With $V_c, V_A > 0$ and $C'_A > 0$, this requires $C'_c < 0$ and then reads $V_c/V_A > C'_A/(-C'_c)$. (ii) With $x = rK/s$ and S_r independent of s , $d/ds[H(1 - e^{-x})/r] = H'(1 - e^{-x})/r - He^{-x}K/s^2$. This is negative if and only if $\varepsilon_H(1 - e^{-x}) < xe^{-x}$, that is, $\varepsilon_H < x/(e^x - 1)$, and $x/(e^x - 1) < 1$ for $x > 0$. A speed-dependent S_r adds $S'_r(s)$ to the derivative. The form $H(1 - e^{-rT})/r$ assumes a rectangular hazard profile. For a general fixed profile, hazard $H \cdot g(t/T)$ with $\int_0^1 g(u)du = 1$, the discounted near loss is $HT \int_0^1 e^{-rTu} g(u)du + S_r = HT\psi(x) + S_r$, with $\psi(x) = \int_0^1 e^{-xu} g(u)du$. Since $d\ln[HT\psi(rT)]/d\ln s = \varepsilon_H - 1 - x\psi'(x)/\psi(x)$, acceleration lowers it below s^* if and only if $\varepsilon_H < 1 + x\psi'(x)/\psi(x)$. The threshold is profile-dependent: at $x = 1$ it is 0.58 for the rectangle, 0.39 for $g(u) = 2u$ and 0.72 for $g(u) = 2(1 - u)$. It is below one for every profile, since $\psi' < 0$, and the rectangle $g \equiv 1$ gives $x/(e^x - 1)$. (iii) Differentiate $V_A C_A(s)e^{-rK_A/s}$. (iv) With P_A conditional on surviving the near catastrophe, total catastrophe probability is $1 - (1 - P_c)(1 - P_A)$, whose derivative is $P'_c(1 - P_A) + P'_A(1 - P_c)$. Discrete counterexample: $H = s^{1.5}$ on $[1, 2]$ and $H = 2^{1.5} + 0.01(s - 2)$ beyond, $K = 20$, $T_a = 1$, $R_A = 0.01s$. The derivative test fails at $s = 1$, yet moving from $s = 1$ to $s = 10$ cuts the total loss from 20.0 to 5.9. This H is not differentiable at $s = 2$, where H' jumps from 2.12 to 0.01, so it violates A4 there; it illustrates the limit of a derivative test for large moves, and (i) applies only where A4 holds. ■

P5. $K/s = A(s/s_0)^{-\beta}$ if and only if $(s/s_0)^{1-\beta} = K/(As_0)$. The ratio $(K/s)/T_a = (K/(As_0))(s/s_0)^{\beta-1}$ falls in s for $\beta < 1$ and rises for $\beta > 1$, which orders the branches. For $\beta = 1$ it is the constant $K/(As_0)$. $d\ln s^*/d\beta = \ln(K/(As_0))/(1 - \beta)^2$ has the sign of $\ln(K/(As_0))$. On the adoption branch, $\ln(HT_a) = \ln H + \ln A - \beta \ln(s/s_0)$, and S does not depend on s (A2), so dR_c/ds has the sign of $\varepsilon_H - \beta$. ■

P6. $\min(W, \tau) \leq W$ and $\min(W, \tau) \leq \tau$ pointwise; take expectations. $E[\min(W, \tau)] = \int_0^\tau P(W > t)dt = (1 - e^{-\lambda\tau})/\lambda$ for an exponential window, and the discounted version is $\int_0^\tau e^{-rt} e^{-\lambda t} dt$. Without access, attackers arrive at W . A covered operator that has deployed by τ is then exposed for $(\tau - W)^+$ instead of τ , saving $\min(W, \tau)$ on share $1 - \pi$ of attack effort. The limits follow from $(1 - e^{-u})/u \rightarrow 1$ as $u \rightarrow 0$ and $\rightarrow 1/u$ as $u \rightarrow \infty$. (iv) Bound fast operators' saving by $w_j \tau_j$. For the rest, $c_j \leq \bar{c}$ and $m(\lambda, \tau_j) \leq \tau_j \min(1, 1/(\lambda\tau_j)) \leq \tau_j \min(1, 1/(\lambda\tau_s))$; sum and divide by E_0 . (v) Weighted by attack intensity, exposure falls from τ to $\tau - (1 - \pi)w$, since share π of attackers is unaffected. For convex g and any subgradient $g'(\tau)$, the supporting line gives $g(\tau - (1 - \pi)w) \geq g(\tau) - (1 - \pi)wg'(\tau)$. Counterexamples: with $g(x) = x^2$, $\pi = 0$, $\tau = 10$ and $W = 1$, the saving in harm is 19, not $g(1) = 1$. If the window is shorter when operators are covered (rate 2 against 0.5, coverage 0.5, $\tau = 2$), the expected saving is 0.245 rather than the factorized 0.439. ■

P7. (a) Exposure is $(P_f + \tau_j - \min(P_f + \ell_x, A_f))^+$, where A_f is attackers' own discovery date: d_f under release, $W + d_f$ under hold. If $d_f \geq P_f + \ell_x$, both minima equal $P_f + \ell_x$. If $d_f < P_f + \ell_x$, the hold raises the minimum, for share $1 - \pi$, from d_f to $\min(P_f + \ell_x, W + d_f)$; exposure falls weakly, and strictly when $\pi < 1$, $W > 0$ and $d_f < P_f + \tau_j$. With $\pi = 1$ nothing changes. (b) Exposure is $(\zeta_j - A_f)^+$. The difference between release and hold is $\min(W, u)$ with $u = (\zeta_j - d_f)^+$ for attackers without access, and zero for the share π with

access. For an exponential window, $E[\min(W, u)] = (1 - e^{-\lambda u})/\lambda$, which is zero when $\zeta_j \leq d_f$. (c) Publication P and exploit onset are fixed and do not depend on the tool. With the tool available from t , deployment completes at $\min(P + \tau, \max(P, t) + \tau')$, with $\tau' \leq \tau$, which is non-decreasing in t . Exposure therefore weakly rises with withholding. It rises strictly if and only if the completion date changes inside an interval of positive loss. If instead the tool produces the patches, and withholding it delays exposure's onset and completion by the same amount, leaving its duration unchanged, undiscounted exposure is unchanged and discounted exposure falls, from 0.879 to 0.795 in our example. The condition matters: if exploitation begins at 0 independently of publication and operators deploy when the patch appears, delaying publication from 1 to 2 at $r = 1$ raises discounted exposure from $1 - e^{-1}$ to $1 - e^{-2}$. ■

P8. (i) Differentiate under the sum, with $m_\lambda = [\lambda \tau e^{-\lambda \tau} - (1 - e^{-\lambda \tau})]/\lambda^2 < 0$, since $1 - e^{-u} > u e^{-u}$ for $u > 0$. (ii) By continuity B has a maximum on $[0, N]$. The endpoint signs exclude both ends. If $B' > 0$ on $[0, n^*)$ and $B' < 0$ on $(n^*, N]$, B is strictly increasing and then strictly decreasing, so n^* is the unique maximizer; strict concavity implies this pattern. In the interior example of our replication code, $B'(0) = 1.55 > 0 > B'(10) = -0.02$, $n^* = 1.59$, and B is not concave but B' changes sign once. (iii) If $B(1) > B(0)$, then 0 is not a maximizer, and likewise for N . With strictly decreasing differences $\Delta_n = B(n+1) - B(n)$, at most one Δ_n is zero. B rises while $\Delta_n > 0$ and falls once $\Delta_n < 0$. The maximizers are therefore the first n with $\Delta_n \leq 0$, together with $n+1$ if $\Delta_n = 0$: at most two, adjacent, and unique if no difference is zero. If every difference is positive, the unique maximizer is N ; otherwise use the characterization above. (iv) With every c_j fixed, B depends on n only through λ . Each $m(\lambda, \tau_j)$ with $\tau_j > 0$ falls strictly in λ , so with $1 - \pi > 0$ and some $c_j w_j \tau_j > 0$, B falls strictly in λ . Counterexamples: with $c = n$ and $\lambda = 1 + 0.01n$ on $[0, 1]$, B rises throughout and the largest coalition is best; with the holder's own systems covered at $n = 0$, partners adding coverage $0.01n$ and $\lambda = 1 + n$, the empty coalition is best. If admission also covers the rival's own systems, B rises from 0.51 to 0.87 while net hazard rises from 1.0 to 1.1. ■

P9. Let $\Delta v = v_W - v_L > 0$ and $\hat{c} = c + L\theta > 0$. Against a rival effort $x_j > 0$, player i 's payoff is strictly concave in $x_i \geq 0$, with second derivative $-2\Delta v x_j / X^3$. A profile with one player at zero and the other positive is not an equilibrium: against a zero rival, any positive effort wins for sure, so the positive player gains by cutting its effort, at saving $\hat{c}$ per unit. The profile $(0, 0)$ is not an equilibrium either: each player gets half of Δv , and a small effort ε wins all of it, gaining $\Delta v/2 - \hat{c}\varepsilon > 0$. So both efforts are positive and both first-order conditions hold: $\Delta v x_j / X^2 = \hat{c} = \Delta v x_i / X^2$, hence $x_i = x_j$, and $x^* = \Delta v / (4\hat{c})$. By strict concavity this profile is an equilibrium, and it is the only pure one. With N players the condition is $\Delta v (X - x_i) / X^2 = \hat{c}$, and at a symmetric profile, $X = Nx$, it gives $x^* = \Delta v (N - 1) / (N^2 \hat{c})$. Welfare: $X^* = \Delta v / (2\hat{c})$, so a winner-only prize v_s adds $\Delta X = v_s / (2\hat{c})$, and exactly one player receives it. Outsiders get G minus their payments and the holder gets σ_s plus those payments, so the payments cancel and welfare is $G + \sigma_s + (b - c - L_A \theta)X$ plus terms that do not depend on the window. Netting payments out of G while dropping the holder's receipt would leave a fictitious loss equal to the payments. Each unit of total effort costs c , adds θ to the expected-loss index at social loss L_A , and yields public value b .

Counterexamples, with $D = c = L = 1$, $\theta = 0.4$ and $L_A = 10$: a window both players receive, with $v_s = 5 > G = 1$, leaves racing unchanged and raises welfare by 11 if each v_s is real surplus, or by 1 if it is a transfer; a winner-only window with $v_s = 5$ and $b = 6 > c + L_A\theta = 5$ raises welfare by 7.8, or 2.8 as a transfer. In the escalation example, where both players receive a window with $v_s = 0.5$, welfare rises by $G + 2\sigma_s - L_A\xi$: 1.5 with real surplus and 0.5 with transfers, so harm needs $\xi > 0.020$ and $\xi > 0.010$ respectively. ■

P10. The stage game. Two racers choose $x_i \geq 0$. Racer i wins D with probability x_i/X , with an equal split at $X = 0$, pays cx_i , and bears expected loss $L\theta X$. (i) The stage equilibrium is P9 with $\Delta v = D$, so $x^* = D/(4(c + L\theta))$. Because the prize is always awarded, the joint payoff is $D - (c + 2L\theta)X$, maximized at $X = 0$. (ii) With cost $kx^2/2$ and public value b per unit of X to each racer, racer i 's payoff is $Dx_i/X - kx_i^2/2 + \tilde{b}X$. The symmetric first-order condition is $kx = \tilde{b} + D/(4x)$, so $x^* = [\tilde{b} + \sqrt{\tilde{b}^2 + kD}]/(2k)$. The joint payoff $D - kx^2 + 2\tilde{b} \cdot 2x$ at a symmetric profile is maximized at $x^\circ = \max(2\tilde{b}/k, 0)$. If $\tilde{b} \leq 0$, $x^* > 0 = x^\circ$. If $\tilde{b} > 0$, $x^* > x^\circ$ if and only if $\sqrt{\tilde{b}^2 + kD} > 3\tilde{b}$, that is, $kD > 8\tilde{b}^2$. (iii) *The repeated game.* The stage game repeats with discount factor $0 < \rho < 1$. Restraint is $x = 0$ in every period, worth $U_C = D/2$ per period. The stage equilibrium is worth $U_N = D/2 - (c + 2L\theta)x^* = U_C - D(c + 2L\theta)/(4(c + L\theta))$. A deviation is any positive effort against a restrained rival. It wins D for sure, so its gain over U_C approaches $D/2$ as the effort shrinks. After each period with a deviation, detection occurs with probability ν . There are no false alarms, and detection triggers permanent reversion to the stage equilibrium from the next period, which is credible because it is an equilibrium of every subgame. By the one-shot deviation principle, restraint is an equilibrium if and only if $D/2 \leq \rho\nu(U_C - U_N)/(1 - \rho) = [\rho\nu/(1 - \rho)] \cdot D(c + 2L\theta)/(4(c + L\theta))$, that is, $\nu \geq \nu^* = ((1 - \rho)/\rho) \cdot 2(c + L\theta)/(c + 2L\theta)$. Since $\nu \leq 1$, restraint cannot be supported when $\nu^* > 1$. In a one-shot race with sanction $Q_{\det}$ on detection, deviation gains $D/2$ and costs $\nu Q_{\det}$ in expectation. Under-racing example: $b = 0.5$, $L\theta = 0.2$, $k = 1$ and $D = 0.36$, below the threshold 0.72, give equilibrium effort 0.485 against an optimum of 0.600. (iv) Let racer i 's payoff be $p_i D - cx_i - \ell_i(x_i)$ plus terms that do not depend on x_i , with $p_i \in [0, 1]$ and ℓ_i non-decreasing. Then $u_i(x) - u_i(0) \leq D - cx - p_i(0)D + \ell_i(0) - \ell_i(x) \leq D - cx < 0$ for $x > D/c$. A cost that merely rises is not enough: under the Tullock rule, against rival effort $y > 0$, with no loss term and cost $(D/2)(1 - e^{-x})$, $u(x) - u(0) = Dx/(x + y) - (D/2)(1 - e^{-x}) \rightarrow D/2 > 0$ as $x \rightarrow \infty$, so arbitrarily large efforts beat zero. In the stage game, against $y > 0$, racer i 's payoff is strictly concave in x_i (P9's proof), with $u'(0) = \Delta v/y - \hat{c}$. The first-order condition $\Delta v \cdot y/(x + y)^2 = \hat{c}$ gives $x = \sqrt{\Delta v \cdot y/\hat{c}} - y$ when that is positive, and the corner $x = 0$ when $y \geq \Delta v/\hat{c}$. So BR is single-valued and not constant: positive for small y and zero for $y \geq \Delta v/\hat{c}$. No effort is a best response to every rival effort, so none is dominant; against $y = 0$ there is no best response (P9's proof). On the positive branch, $0 < y < \Delta v/\hat{c}$, $dBR/dy = 1/2 \sqrt{\Delta v/(\hat{c}y)} - 1$, which is positive if and only if $y < \Delta v/(4\hat{c}) = x^*$; beyond the branch $dBR/dy = 0$, and at its endpoint $y = \Delta v/\hat{c}$ BR is not differentiable. Also $BR(y) - y = \sqrt{\Delta v \cdot y/\hat{c}} - 2y > 0$ if and only if $y < x^*$, and $BR(x^*) = x^*$, so $\bar{x} < BR(\bar{x}) < x^*$ on $(0, x^*)$. Total effort is $\bar{x} + BR(\bar{x}) = \sqrt{\Delta v \cdot \bar{x}/\hat{c}}$, rising in $\bar{x}$ and below $\sqrt{\Delta v \cdot x^*/\hat{c}} = \Delta v/(2\hat{c}) = 2x^*$. The capper wins with probability $\bar{x}/\sqrt{\Delta v \cdot \bar{x}/\hat{c}} = \sqrt{\hat{c}\bar{x}/\Delta v} < 1/2$. Our replication code reproduces these closed forms by best-response search and by the value recursion of a stationary deviation, and the

earlier draft's contest code also runs with zero failures, including its best-response and cap checks (MODEL/code/race_model_out.txt). ■

P11. Holding to T earns $F + \rho(g)$ while neither release nor loss has occurred, which has probability $e^{-\lambda t}$. Loss at rate λ yields Z . Release at T yields $J(T)$. Discounting gives $V(T)$. Then $V'(T) = e^{-(r+\lambda)T} [F + \lambda Z + \rho(g(T))] - (r + \lambda)e^{-(r+\lambda)T} J(T) + e^{-(r+\lambda)T} J'(T)$. If Φ is positive before a single crossing and negative after, V rises then falls, which gives the global maximum. If the zeros of Φ are isolated, every local maximum of V on $(0, \infty)$ is a crossing from above, so comparing those with $T = 0$ and $T \rightarrow \infty$ finds the optimum. *Smoothness.* The formula for V' uses only that ρ is continuous on $(0, g_0 + \Delta]$ and that J is differentiable. When $\kappa_{\text{post}} > 0$, continuity is enough, because $\Pi(x) = \kappa_{\text{post}}^{-1} x^{-r/\kappa_{\text{post}}} \int_0^x u^{r/\kappa_{\text{post}}-1} \rho(u) du$ is then differentiable and satisfies $r\Pi = \rho - \kappa_{\text{post}} x \Pi'$. When $\kappa_{\text{post}} = 0$, $\Pi = \rho/r$, so ρ must be differentiable, and $\rho \in C^1$ makes Φ continuous. Part (v) uses a continuously differentiable $\tilde{\Phi}$ near G^* , which $\rho \in C^1$ gives when $\kappa_{\text{post}} > 0$ and $\rho \in C^2$ gives in general. If ρ jumps and $\kappa_{\text{post}} > 0$, V stays continuous but has kinks at the dates where $g(T)$ or $g(T) + \Delta$ crosses a jump point; if $\kappa_{\text{post}} = 0$, J itself jumps where $g(T) + \Delta$ does. Those dates join the candidate set. A kink can be a local maximum only if $\Phi(T-) \geq 0 \geq \Phi(T+)$, and it is a strict one if $\Phi(T-) > 0 > \Phi(T+)$. With $\rho(x) = x + 1\{x \geq 1\}$, $r = \lambda = \kappa_{\text{pre}} = \kappa_{\text{post}} = \Delta = 1$, $g_0 = 1.1$, $F = 9/4$ and $Z = 0$, $\Pi(x) = x/2 + (1 - 1/x)^+$, and Φ jumps from $+1/2$ to $-1/2$ at $T_0 = \ln 1.1$, where $g = 1$. $V(T_0) = 1.6129$ exceeds $V(0) = 1.5738$, $V(\infty) = 1.5784$ and the value at every other date, so the strict global optimum is a kink that the smooth candidate set misses. The sign of Φ establishes this. On each side of the kink $\Phi = c - g/2 + (g + 2)/(g + 1)^2$, with $c = 1/4$ for $g > 1$ and $c = -3/4$ for $g < 1$, and $d\Phi/dg = -1/2 - (g + 3)/(g + 1)^3 < 0$, so Φ rises in T on each side. Before the kink $\Phi \geq \Phi(g = 1.1) > 0.40$, so V rises up to T_0 . After it Φ rises from $-1/2$ toward $5/4$, so it has a single zero, upward, a local minimum of V , beyond which V rises toward $V(\infty)$. The global maximum is therefore the larger of $V(T_0)$ and $V(\infty)$, which is $V(T_0)$; the grid in our replication code illustrates it. (ii) With no old public model, J is constant and Φ is constant; at $\Phi = 0$, V is constant in T . (iii) $\Pi(x) = Mx/(r + \kappa_{\text{post}})$ and $J' = -\kappa_{\text{pre}} g M/(r + \kappa_{\text{post}})$; substitute. The g term rises in T if $\kappa_{\text{post}} < \kappa_{\text{pre}} + \lambda$ and is constant at equality, so Φ can fall through zero only if $\kappa_{\text{post}} > \kappa_{\text{pre}} + \lambda$; at equality Φ is the constant $F + \lambda Z - (r + \lambda)M\Delta/(r + \kappa_{\text{post}})$. Rents need not be linear or concave: with $\rho = x^2$, $r = \lambda = \kappa_{\text{pre}} = 1$, $\kappa_{\text{post}} = 3$, $\Delta = 1$, $g_0 = 3$, $F = 1/7$ and $Z = 0$, $\Pi(x) = x^2/7$ and $\Phi = (3g^2 - 6g - 1)/7$, which crosses zero once, downward, at $g^* = 1 + 2/\sqrt{3}$, a strict global interior release. (v) Write $\tilde{\Phi}(g) = F + \lambda Z + \rho(g) - (r + \lambda)\Pi(g + \Delta) - \kappa_{\text{pre}} g \Pi'(g + \Delta)$ and $T^* = \ln(g_0/G^*)/\kappa_{\text{pre}}$, with $\tilde{\Phi}(G^*) = 0$. At a simple downward crossing in T , $\tilde{\Phi}$ is strictly increasing in g , so a parameter that raises $\tilde{\Phi}$ lowers G^* and delays release. $\partial \tilde{\Phi} / \partial f_u = 1 > 0$. With $\lambda = \lambda_0 + \lambda(f_a)$, $\partial \tilde{\Phi} / \partial \lambda = Z - \Pi(G^* + \Delta)$, so $\partial \tilde{\Phi} / \partial f_a = 1 - \lambda'(f_a)(\Pi(G^* + \Delta) - Z)$. $\partial \tilde{\Phi} / \partial \kappa_{\text{pre}} = -g \Pi' < 0$, and T^* also scales with $1/\kappa_{\text{pre}}$, so release comes sooner. For the discount rate, with Z fixed: under linear rents, $\partial \tilde{\Phi} / \partial r = -M[\Delta(\kappa_{\text{post}} - \lambda) + g(\kappa_{\text{post}} - \kappa_{\text{pre}} - \lambda)]/(r + \kappa_{\text{post}})^2$, which is negative at any linear interior, since (iii) requires $\kappa_{\text{post}} > \kappa_{\text{pre}} + \lambda$; release comes sooner. Under power rents, with $x = g + \Delta$, $\partial \tilde{\Phi} / \partial r = Mx^{\eta-1}[\kappa_{\text{pre}} \eta g - (\eta \kappa_{\text{post}} - \lambda)x]/(r + \eta \kappa_{\text{post}})^2$, which has either sign. For $\rho = Mx^\eta$ with any $\eta > 0$, convex powers included, $\Pi(x) = Mx^\eta/(r + \eta \kappa_{\text{post}})$, and with $x = g +$

$\Delta, \partial \tilde{\Phi} / \partial \kappa_{\text{post}} = M\eta x^{\eta-1}[(r + \lambda)x + \kappa_{\text{pre}}\eta g]/(r + \eta\kappa_{\text{post}})^2 > 0$ at every $g \geq 0$, so release comes later. More generally, write $q(y) = y\rho'(y)$ and $y = xe^{-\kappa_{\text{post}}t}$. Then $\partial \Pi / \partial \kappa_{\text{post}} = -\int_0^\infty t e^{-rt} q(y) dt < 0$ and $\partial \Pi' / \partial \kappa_{\text{post}} = -\int_0^\infty t e^{-(r+\kappa_{\text{post}})t} q'(y) dt$, which is ≤ 0 when q is non-decreasing. Hence $\partial \tilde{\Phi} / \partial \kappa_{\text{post}} = -(r + \lambda) \partial \Pi / \partial \kappa_{\text{post}} - \kappa_{\text{pre}} g \partial \Pi' / \partial \kappa_{\text{post}} > 0$, with λ fixed, and release comes later; power rents, linear ones included, have $q' = M\eta^2 x^{\eta-1} > 0$. For other rents the sign can reverse: with $\rho = 1 - e^{-4x}$, $r = \lambda = \kappa_{\text{pre}} = \Delta = 1$, $g_0 = e$, $Z = 0$, $\kappa_{\text{post}} = 0$ and $F = 1 + e^{-4} + 2e^{-8}$, $\Pi = \rho$ and $\tilde{\Phi}(g) = F - 1 - e^{-4g}(1 - 2e^{-4} + 4e^{-4}g)$. The last term has g -derivative $-e^{-4g}(12e^{-4} - 4 - 16e^{-4}g) > 0$, so Φ falls strictly in T on $[0, \infty)$ and crosses zero once, downward, at $g = 1$. This sign argument establishes that $T^* = 1$ is the strict global optimum, against $T = 0$ and holding until loss as well; the scan over $T \in (0, 6]$ in our replication code illustrates it. The right derivative of T^* in κ_{post} at $\kappa_{\text{post}} = 0$ is $-3/(e^4 + 1) < 0$, so faster copying of the released model brings release sooner. Here $x\rho'(x) = 4xe^{-4x}$ falls for $x > 1/4$. This reversal is evaluated at $\kappa_{\text{post}} = 0$; it does not mean reversals occur only there. With one copy rate the two effects have opposite signs. These signs are local: they hold at such a crossing, other primitives fixed, and need not hold globally, where a parameter change can move the optimum to a corner, to another crossing or to another branch. Counterexamples: with constant flows, $F = 0.4$, a forgone flow $M\Delta = 1$, $r = 0.05$, $\lambda = 0.2$ and $\kappa = 1$, the optimum is to hold until involuntary loss, and raising κ from 0.5 to 1.0 moves it from immediate release to no voluntary release. With $Z \neq 0$ (F lowered by λZ so that T^* is unchanged), the artifact term without Z gets the sign wrong in two of four test cases. ■

P12. $\int_0^\infty e^{-(r+\kappa)t} dt = 1/(r + \kappa)$, and $\int_0^\infty e^{-rt} dt = 1/r$ for $r > 0$; at $r = 0$ the non-eroding value is infinite. With $\rho = Mx^\eta$ and $x(t) = x_0 e^{-\kappa t}$, revenue is $Mx_0^\eta e^{-\eta\kappa t}$, worth $Mx_0^\eta/(r + \eta\kappa)$. In separate contests for an incremental and a decisive prize, raising κ from 0.2 to 1.0 cuts incremental effort from 1.00 to 0.24 and leaves decisive effort at 1.25. ■

P13. $(F + f)/(r + \lambda + \delta) > F/(r + \lambda)$ if and only if $(F + f)(r + \lambda) > F(r + \lambda + \delta)$, that is, $f(r + \lambda) > F\delta$. With Z : $(F + f + (\lambda + \delta)Z)/(r + \lambda + \delta) > (F + \lambda Z)/(r + \lambda)$ if and only if $(f + \delta Z)(r + \lambda) > \delta(F + \lambda Z)$, that is, $f/\delta > (F - rZ)/(r + \lambda)$. With theft, replace δ by $\delta + \theta_F f$ in either form. Counterexamples to the corollaries, at $r = 0$: with $F = 50$, $\lambda = 1$ (so $F/\lambda = 50$), $f = 1000$ and $\delta = 10$, the use passes and the hold is worth 50 before it and 95.45 after; with $f = \delta = 0.001$ and $F/\lambda = 100$, the use fails. Release option: in P11's interior example, a use with $\delta = 0.02$ and $f/\delta = 2.147$, committed for the rest of the hold, raises the no-release hold value from 2.5744 to 2.5746, yet lowers the optimal value from 2.5807 to 2.5777. ■

P14. Let $a = r + \lambda_0 > 0$. (i) If $\liminf \lambda_\theta(F)/F = 2q > 0$ (or $+\infty$), there is F_1 with $\lambda_\theta(F) \geq qF$ for $F \geq F_1$, so $V \leq 1/q$ there, while $V \leq F_1/a$ on $[0, F_1]$. If the $\liminf$ is zero, take $F_n \rightarrow \infty$ with $\lambda_\theta(F_n)/F_n \rightarrow 0$. Then $V(F_n) = 1/(a/F_n + \lambda_\theta(F_n)/F_n) \rightarrow \infty$. (ii) At $F = 0$, $V(0) = 0 < 1/\theta_F$. For $F > 0$, $F/(a + \lambda_\theta(F)) \leq F/(a + \theta_F F) < F/(\theta_F F) = 1/\theta_F$, since $a > 0$. (iii) With durable $Z \geq 0$ and $\lambda = \lambda_0 + \lambda_\theta(F)$, under (ii)'s $\lambda_\theta(F) \geq \theta_F F$, the value is $(F + \lambda Z)/(r + \lambda) = F/(r + \lambda) + [\lambda/(r + \lambda)]Z$. The first term is below $1/\theta_F$ by (ii), and $\lambda/(r + \lambda) \leq 1$, so $V < 1/\theta_F + Z$ at every F . Sublinear theft: $F/(a + \theta_F \sqrt{F}) \rightarrow \infty$; at $\theta_F = 0.5$, $V = 19,999$ at

$F = 10^8$ against a cap of 2. With $\theta_F(\sigma) = \theta_{F,0}e^{-\sigma}$ bought at cost σ , value at $F = 10^6$ exceeds 6×10^6 . Theft on value: $\theta_F V^2 + aV - F = 0$, so $V = [-a + \sqrt{a^2 + 4\theta_F F}]/(2\theta_F)$. ■

P15. $V(d) = (1 - d)/(r + \kappa) + d/r = (r + d\kappa)/(r(r + \kappa))$, so $(1/r)/V(d) = (r + \kappa)/(r + d\kappa)$, and $V'(d) = \kappa/(r(r + \kappa))$, positive when $\kappa > 0$. (b) Apply P9 with $v_W - v_L = mV(d)$. (c) Each firm's payoff is $D_s p_s + D_d p_d - k(x_s + x_d)^2/2$, with Tullock win probabilities p and no private catastrophe cost; if firms bore $L\theta_s$ and $L\theta_d$ per unit of effort, those terms would enter the first-order conditions and change the allocation. For $D_s > 0$, P9's corner argument applies to each contest, with marginal cost $k(x_s + x_d) > 0$, so both efforts are positive and the first-order conditions at a symmetric profile are $D_s/(4x_s) = kY$ and $D_d/(4x_d) = kY$, with $Y = x_s + x_d$. At $D_s = 0$ the step contest offers nothing, so $x_s = 0$, and the decisive condition gives $x_d = Y = \sqrt{D_d/(4k)}$, the formulas' value at $D_s = 0$; derivatives there are one-sided. Summing gives $Y^2 = (D_s + D_d)/(4k)$. Total risk is $2(\theta_s D_s + \theta_d D_d)/(4kY)$. Its derivative in D_s has the sign of $8k\theta_s Y^2 - \theta_s D_s - \theta_d D_d = \theta_s D_s + (2\theta_s - \theta_d)D_d$. This is positive if and only if $\theta_d/\theta_s < 2 + D_s/D_d$. Counterexamples to (a) and (b): with separate budgets, decisive effort stays at 0.71 at every d while step effort rises; if each firm banks durable rents from its own products, step effort falls from 0.023 to 0; if winning a step carries a share of a decisive prize that durable assets dilute, the stake falls in d , from 51.8 to 20.0. ■

A.7 Assumptions a later note could relax

Each entry names an assumption, says in one line why relaxing it matters, and gives any sign already known.

- Sequential timing, $T = K/s + T_a$ (A3; P1, P5). Adoption may begin only once the capability arrives. With T_a fixed and S independent of s , a marginal acceleration then lowers the loss index if and only if $\varepsilon_H < (K/s)/(K/s + T_a)$, and there is no kink.
- A hazard set by the capability level rather than by speed (A1, A4). Risk may depend on the level reached, as in the case described in Section 3. In an explicitly scoped alternative calculation, not a consequence of A.2.2's index, take the path $c(t) = \min(st, K)$, hazard $h(c(t))$, fixed T_a , S independent of s , and hazard $h(K)$ through the adoption tail. Above the kink $s^* = K/T_a$, the derivative of cumulative hazard in speed has the sign of $h(K) - \bar{h}$, where $\bar{h} = K^{-1} \int_0^K h(c)dc$ is the average hazard over capability levels 0 to K .
- A joint path model, with P3's φ , γ and μ indexed to capability $s \cdot t$ (A5). It would tie the peak of P3 to the speed of P1–P5, and let speed enter the long-run danger through attackers' capability per flaw. No sign is known.
- Separate attacker and defender speeds, and adoption brought on by crises (A3, A4). One capability clock is what makes running thicken the rain, and crisis-driven adoption would make T_a respond to realized harm. No sign is known.
- Far risk falling with speed, $C'_A < 0$ (P4). P4 assumes $C'_A > 0$, so the fifth question is answered only under that assumption. P4(i)'s derivative $V_c C'_c + V_A C'_A$ still applies: above the kink, where $C'_c > 0$, acceleration would lower the total loss if and only if $V_A/V_c > C'_c/(-C'_A)$.

- A loss hazard rising with copy speed (A7; P11). P11(v)'s copying comparative statics hold λ fixed. If λ rises with κ_{post} , the derivative of $\tilde{\Phi}$ gains $(\partial\lambda/\partial\kappa_{\text{post}})(Z - J(T^*))$, an opposing push toward earlier release when $J(T^*) > Z$. At a downward crossing, $dT^*/d\kappa_{\text{post}} = -[\tilde{\Phi}_{\kappa_{\text{post}}} + \lambda'(Z - J)]/\Phi_T$, with $\lambda' = \partial\lambda/\partial\kappa_{\text{post}}$, so, at a simple downward interior crossing, release still comes later if and only if $A_\kappa > \lambda'(J(T^*) - Z)$, where $A_\kappa = \partial\tilde{\Phi}/\partial\kappa_{\text{post}}$ at fixed λ ; a parameter change can instead move the optimum to a corner or another crossing. For power rents with $Z = 0$, that is $\lambda' < \eta[(r + \lambda) + \eta\kappa_{\text{pre}}G^*/(G^* + \Delta)]/(r + \eta\kappa_{\text{post}})$.
- Internal research whose value survives release (A11; P11). P11 assumes that release ends the held flow, internal research included. If a flow F_{post} survives both release and involuntary loss, F enters the stopping condition as $F - F_{\text{post}}$; if it survives release only, as $F - (r + \lambda)F_{\text{post}}/r$. Here Z excludes the newly retained flow. The copy-rate signs are unchanged. Research that survives both becomes neutral for timing. Research that survives release only brings release sooner: a unit of it moves the stopping condition by $-\lambda/r$, a derivative of the condition, not of the release date.
- Mixed timing, $T = \max(K/s, (1 - m)T_a) + mT_a$ with $0 \leq m < 1$ (A3; P1). A share m of adoption must wait for the capability and the rest runs alongside it; $m = 0$ is P1, and $m = 1$ is the sequential case above, with no kink. With T_a , m and S fixed, the kink is at $K/((1 - m)T_a)$. Below it, acceleration lowers the loss index if and only if $\varepsilon_H < (K/s)/(K/s + mT_a)$; above it, as in P1, acceleration only raises it. In our judgment, patching flaws that the capability finds is plausibly sequential, and retirement and migration plausibly parallel. Section 4.3 tabulates the cells for both hazards.
- Removal limited by triage capacity Q , $dE/dt = \iota - \min(\mu E, Q)$ (A5; P3). Take P3(iii)'s inflow path from an empty stock, constant μ and Q , and an untruncated horizon. If $\iota_1 > Q$, there is no plateau and the stock grows without limit. If $\iota_1 < Q$, the plateau φ_{ι_1}/μ is unchanged and any peak is no lower than without the cap. Section 6.5 reports one lab's statement that human triage is the rate-limiting step in its disclosure program.
- A non-empty starting stock (A5; P3(iii)). P3(iii) starts from $E(0) = 0$. From $E(0) = E_0 \geq 0$, with $q = \mu[\iota_0 - (\mu - \gamma)E_0] - \gamma\iota_1$: for $\mu \neq \gamma$ a future peak exists if and only if $\mu E_0 < \iota_0$ and $q > 0$, at $\hat{t} = \ln[q/(\gamma(\iota_0 - \iota_1))]/(\mu - \gamma)$; for $\mu = \gamma$ it exists if and only if $\mu E_0 < \iota_0$, at $\hat{t} = (\iota_0 - \mu E_0)/(\mu(\iota_0 - \iota_1))$; it is pre-ASI if $\hat{t} < T_A$. If today's stock already exceeds ι_0/μ , no peak lies ahead.
- Foreign states choosing π and λ_U strategically (A8; E3). Both are fixed here, though a state could answer a hold by building parallel access or by stealing the weights. No sign is known.
- Maintainers, small operators and operators of operational technology as separate adoption populations (A3, A14). The deployment series of Section 6.2 cover different populations, and a single adoption time can hide the slowest. With $R = \sum_j w_j H(s) \cdot \max(K/s, T_{a,j}) + S$, fixed nonnegative weights, a common $H(s)$, constant adoption times and a speed away from every population's kink, acceleration lowers R if and only if $(1 - \varepsilon_H) \sum_{\text{cap}} w_j K/s > \varepsilon_H \sum_{\text{adopt}} w_j T_{a,j}$, the sums running over the

populations where capability binds and where adoption binds. Section 11 lists the measurement.

- An exploit lag that shortens once attackers hold the capability (P7; E2). Section 6.1's harness result suggests it. A hold then also protects flaws the holder has found and patched, for at most the window. E2's sign result is unchanged.
- Narrow tools that decouple the benefit from attackers' capability (A16; P16). If specialized models, such as structure predictors, advance K_B on a clock of their own, general capability speed no longer brings the benefit forward, and P16's benefit term does not attach to it. Not modelled; no sign is known.
- Dual-use biological risk (A15; P16). The capability that speeds drug discovery may also aid biological attack, a risk outside the health–cyber balance. Not modelled.
- Attacks on health systems that cut the realized benefit flow (A15, A18; P16). Section 4.7's hospital studies show that cyber harm reaches health directly; P16 holds y constant and has no mapping from cyber harm to realized benefit. Not modelled.

Appendix B. Extension: Open- and Closed-Weight Copies

P-numbers and assumption labels such as A7–A9 refer to Appendix A.

B.1 Scope

An open-weight copy publishes the weights; a closed-weight copy serves a comparable capability behind an interface. Both can end the holder's commercial lead. Only an open copy, or a closed copy whose safeguards fail, gives excluded attackers lasting capability. We retain three results: the two windows, their separation by safeguarded release, and the reach limit of domestic levers. Definitions and exposure units follow Appendix A.

We impose the following restrictions to keep the extension focused.

Scope	Restriction
B-S1: private values	We set aside the holder's payoff and event-specific continuations, retained flows, safeguard costs, liability, public-model rents and erosion. We leave their effects on private incentives and release choices to future work. We assert no P11–P15 carry-over or private-versus-social ranking.
B-S2: release policy	We study committed holding and immediate release in a fixed mode. Holding includes no later voluntary release, including after a safeguarded rival preemption. We leave planned future dates, post-preemption release, optimized timing across modes,

Scope	Restriction
	mode switching, corners, ties and branch changes to future work. Allowing an attack-enabling release after preemption could only shorten attacker exclusion at fixed future hazards; the optimized responses have no general direction. No continuation value hides a later release.
B-S3: defensive accounting	We study linear, undiscounted exposure at stated flaw-level schedules. We set aside coalition optimization, admission rules, aggregate coverage bounds, convex-harm and discounted refinements, and any general P6–P8 carry-over claim. Schedule changes are explicit in E2; their effect is measured there rather than presumed beneficial.
B-S4: event dynamics	We set constant, independent primitive arrival streams and independent safeguard failures. We leave endogenous hazards, common-shock dynamics beyond counting each shock once, joint-law refinements and comparative statics beyond the three results to future work. There is no general sign for their combined effects.
B-S5: usable access and public loss	We study failures that give lasting, usable attacker capability, with unchanged subsequent attack intensity and no backlog. Temporary or revocable interface abuse is outside scope. We assume theft and leaks also supply a commercial substitute immediately; secret theft that only arms attackers could end attacker exclusion before commercial leadership, so the window ordering would need revision.
B-S6: release-induced changes	For the comparison in E2, we set future access hazards equal under release and holding, and schedules independent of the full event streams and the holder's safeguard outcome. We leave adaptive schedules and release-induced changes in theft, distillation or rival access to future work. An added attack-enabling hazard would raise exposure weakly at a fixed

Scope	Restriction
	schedule; a defensive improvement can lower it as specified in E2.
B-S7: policy reach	We hold the unbound attack-enabling intensity strictly positive and fixed. We leave changes to that intensity, zero-floor cases, ceiling-attainment characterizations, alternative proposed ceilings, commercial ceilings and intervention rankings outside this extension. The reach bound is conditional on which developers and routes the levers cover.

B.2 Events and assumptions

B.2.1 One event system, two windows

Event	Intensity	Ends commercial lead	Ends attacker exclusion
O: rival open-weight comparable release	λ_o	Yes	Yes
C1: rival closed release whose safeguards fail	$p_c \lambda_c$	Yes	Yes
C0: rival closed release whose safeguards work	$(1 - p_c) \lambda_c$	Yes	No
Θ : theft of weights	λ_θ	Yes	Yes
L: partner leak	λ_ℓ	Yes	Yes

The underlying open and closed rival releases are independent homogeneous Poisson streams. Each closed arrival independently fails with probability p_c . Theft and leak streams are independent of these and each other. Streams continue after safeguarded arrivals. An event with several effects is counted once, as required by A7; independent streams here are a stronger restriction than A7 alone. Copies inferred from artifacts or products belong to the appropriate release type, without adding a second count for the same event. Event inclusion, especially theft's commercial effect, is a modeling premise.

Let ω_C be the first event in the table and ω_A the first attack-enabling event, excluding C0. Write

$$\lambda_C = \lambda_o + \lambda_c + \lambda_\theta + \lambda_\ell,$$

$$\lambda_A = \lambda_o + p_c \lambda_c + \lambda_\theta + \lambda_\ell.$$

Rates are nonnegative, $0 \leq p_c \leq 1$, and $\lambda_A > 0$, hence $\lambda_C > 0$. A zero-rate component never arrives. The two windows share events; they are dependent.

The holder's initial action is committed holding, immediate open release, or immediate closed release. A closed release fails with indicator B_h , independent of all primitive streams, with $P(B_h = 1) = p_h$, $0 \leq p_h \leq 1$. Open release always enables attackers. Working safeguards mean $B_h = 0$; "closed release" alone does not guarantee success. Future primitive streams are the same under the actions being compared.

B.2.2 Exposure and defensive schedules

The *commercial window* W_C lasts until a comparable substitute becomes commercially available, including the holder's own release. The *attacker-access window* W_A lasts until excluded attackers obtain lasting capability. Share $\pi \in [0, 1]$ of attack effort already has parallel access and is unaffected by this exclusion, as in A8. Excluded attackers arrive at unchanged intensity, with no backlog.

A defensive schedule Σ includes coverage and all publication, discovery, deployment and removal dates. For a patched flaw f at operator j , define exposure for excluded attackers by

$$A_f(\Sigma, W) = \min(P_f + \ell_x, W + d_f),$$

$$D_{fj} = P_f + \tau_{fj}, \quad e_{fj}(\Sigma, W) = (D_{fj} - A_f(\Sigma, W))^+.$$

Here P_f is patch publication, ℓ_x the patch-to-exploit lag, d_f the attackers' discovery time after access, and τ_{fj} deployment time after publication. For an unfound flaw, use $e_{fj}(\Sigma, W) = (\zeta_j - W - d_f)^+$, with removal date ζ_j . Parallel-access exposure uses onset $\min(P_f + \ell_x, d_f^\pi)$ for patched flaws and d_f^π for unfound flaws, independently of W . The extension's d_f^π records discovery by attackers already holding access; it need not equal d_f . The patch-to-exploit lag ℓ_x is common: it is the same whether or not attackers hold the capability. If the lag shortens once attackers hold the capability, as the harness result in Section 6.1 suggests, a hold also protects flaws the holder has found and patched, for at most the window. E2's sign result is unchanged, because exposure onset stays non-decreasing in W ; the equations above keep the common lag.

Let $X(\Sigma, W)$ sum these exposures with nonnegative operator/flaw weights, weighting excluded and parallel-access effort by $1 - \pi$ and π . Assume finite expected total exposure. Thus X is Appendix A's weighted exposure-time, not private revenue or an aggregate welfare index. Schedule distributions may differ across policies, but each is independent of the complete primitive event streams and B_h . This strengthens A9 to cover the shared event system; independence merely from the first attacker arrival is not our assumption.

B.3 Three results

E1. The two windows

Proposition E1. Under the stated event assumptions, ω_C and ω_A are exponential with rates λ_C and λ_A , respectively, and $\omega_C \leq \omega_A$ on every path. The realized windows are:

Holder's action	W_C	W_A
Committed holding	ω_C	ω_A
Immediate open release	0	0
Immediate closed release, working safeguards	0	ω_A
Immediate closed release, failed safeguards	0	0

In particular, $W_C \leq W_A$. Under holding, their means are $1/\lambda_C$ and $1/\lambda_A$. An open copy or a closed copy with failed safeguards ends attacker exclusion as well as commercial leadership; a working safeguarded closed copy ends commercial leadership alone.

The exponential law concerns the exogenous first arrivals. Immediate closed release with a positive failure probability gives W_A an atom at zero; it is not simply an exponential window at a substituted hazard.

E2. Safeguarded release splits the windows

Compare immediate closed release with committed holding from the same initial state, coupling them to the same future primitive streams. Write their access dates as W_{cl} and W_H , with $W_H = \omega_A$. Define

$$X_H = E[X(\Sigma_H, \omega_A)], \quad X_{cl} = E[X(\Sigma_{cl}, W_{cl})],$$

$$S_H = E[X(\Sigma_H, 0) - X(\Sigma_H, \omega_A)],$$

$$\Gamma_{def} = E[X(\Sigma_H, W_{cl}) - X(\Sigma_{cl}, W_{cl})].$$

Proposition E2. A working safeguarded release ends the commercial window immediately and leaves the attacker-access window at ω_A . Allowing failure probability p_h ,

$$W_{cl} = 0 \text{ if } B_h = 1, \text{ and } W_{cl} = \omega_A \text{ otherwise;}$$

$$X_{cl} - X_H = p_h S_H - \Gamma_{def}, \text{ with } S_H \geq 0.$$

With defensive schedules and future access hazards unchanged, $\Gamma_{def} = 0$, so release weakly raises outsiders' exposure. The increase is strict exactly when $p_h S_H > 0$. With perfectly working safeguards and the same schedule, exposure is equal. Release lowers exposure exactly when its explicitly modeled defensive improvement satisfies $\Gamma_{def} > p_h S_H$.

Defensive improvement may mean earlier deployment, broader coverage or other schedule changes, but its sign and size must come from the schedule. Uncovered, unfound flaws can contribute to S_H ; a formula counting only covered operators would require an additional assumption and is not used here. No holder-profit or intervention ranking follows from this exposure comparison.

The safeguards here are those of structured access, in which the developer must “prevent the user from circumventing these restrictions through modification or reverse engineering” of the system (Shevlane 2022, 1); p_h is the chance that they fail. Shevlane also notes that “it is often impossible to completely isolate the software-as-a-tool such that no

information about the workings of the tool leaks out during use” (Shevlane 2022, 9 n. 2). Leakage of that kind is a release-induced access hazard, which B-S6 sets aside.

E3. The reach limit of domestic levers

Split each rival release stream into independent homogeneous Poisson streams from covered domestic developers (*dom*) and unbound developers (*out*), so that $\lambda_o = \lambda_o^{\text{dom}} + \lambda_o^{\text{out}}$ and $\lambda_c = \lambda_c^{\text{dom}} + \lambda_c^{\text{out}}$. Each closed arrival is marked independently with failure probability p_c^{dom} or p_c^{out} , so $p_c \lambda_c = p_c^{\text{dom}} \lambda_c^{\text{dom}} + p_c^{\text{out}} \lambda_c^{\text{out}}$. Define the fixed unbound attack-enabling intensity

$$\lambda_U = \lambda_o^{\text{out}} + p_c^{\text{out}} \lambda_c^{\text{out}} + \lambda_\theta + \lambda_\ell > 0.$$

λ_U is the unbound attack-enabling intensity. Covered domestic levers may change λ_o^{dom} , λ_c^{dom} and p_c^{dom} . Unbound release intensities and failure probabilities, theft and leaks stay fixed. Consequently

$$\lambda_A = \lambda_U + \lambda_o^{\text{dom}} + p_c^{\text{dom}} \lambda_c^{\text{dom}}.$$

Proposition E3. For every setting of the covered domestic hazards and each initial action studied here,

$$E[W_A] \leq 1/\lambda_A \leq 1/\lambda_U.$$

Levers that reach only covered domestic developers cannot raise the expected attacker-access window above $1/\lambda_U$. This conditional reach limit ranks no interventions and makes no claim that unbound catch-up is the dominant termination route. Measures that change theft, leaks or unbound developers change λ_U and fall outside this comparison.

B.4 Assumptions a later note could relax

Starred items look most worth a note.

- ★ Theft or leak that arms attackers without a commercial substitute (B-S5). Covert theft is plausible in practice, and attackers would then gain access while the holder still leads commercially.
- ★ Release-induced access hazards (B-S6). Serving a model may raise theft or distillation hazards; this is likely in practice and bears directly on E2.
- ★ The holder’s private payoff and release timing (B-S1, B-S2). Whether holders choose safeguarded release at all depends on them; they connect the extension to P11.
- Adaptive defensive schedules (B-S3, B-S6). Defenders may speed deployment after a release; this is how a defensive improvement would arise.
- Convex or discounted harm (B-S3). Convexity makes marginal harm depend on baseline exposure; discounting makes calendar timing matter.
- Temporary or revocable interface abuse (B-S5). It is common in practice for served models.

- Endogenous or correlated hazards (B-S4). A release can provoke rival releases, and common shocks can tie streams together.
- Levers on theft, leaks or unbound developers, including a zero floor (B-S7). These are real policy levers, and E3 gives no bound when $\lambda_U = 0$.

B.5 Proofs

E1. Independent marking of each closed Poisson arrival splits it into independent streams C1 and C0 with rates $p_c\lambda_c$ and $(1 - p_c)\lambda_c$. No event in a set of independent Poisson streams has arrived by t with probability equal to the product of the component survivals. Thus $P(\omega_C > t) = e^{-\lambda_C t}$ and $P(\omega_A > t) = e^{-\lambda_A t}$. The attack-enabling event set is a subset of the commercial event set on the same paths, giving $\omega_C \leq \omega_A$. Integrating these survivals gives the stated means. The holder's initial action produces exactly the windows in the table: voluntary release supplies a commercial substitute, while only open release or a failed closed safeguard supplies attacker capability. Working safeguards leave the future access streams unchanged. The table and exogenous ordering give $W_C \leq W_A$ in every case. ■

E2. For every fixed schedule, raising W weakly delays excluded attackers' onset, whether it is $\min(P_f + \ell_x, W + d_f)$ or $W + d_f$. It therefore weakly lowers each positive-part exposure. Nonnegative weights preserve this ordering; the parallel-access share is unaffected. Hence $S_H \geq 0$. Under immediate closed release, E1 gives the stated mixture for W_{cl} . Independence of B_h from both the streams and the schedule yields

$$E[X(\Sigma_H, W_{cl})] = p_h E[X(\Sigma_H, 0)] + (1 - p_h) E[X(\Sigma_H, \omega_A)].$$

Subtract X_H and then add the schedule change at the release window. By the definition of Γ_{def} , the result is $X_{cl} - X_H = p_h S_H - \Gamma_{\text{def}}$. Identical schedule distributions, independent of the complete event system and safeguard outcome, give $\Gamma_{\text{def}} = 0$. The weak, strict and improvement conditions follow directly. This argument includes unfound flaws regardless of coverage. It relies on unchanged future access streams; a changed access law cannot be absorbed into a claimed defensive improvement. ■

E3. The attack-enabling events from unbound developers, theft and leaks form a subset of all attack-enabling events. Their first arrival ω_A^{out} has exponential rate λ_U by independence of the dom and out substreams and of their marks. Thus $\omega_A \leq \omega_A^{\text{out}}$ pathwise. Every action in E1 gives $W_A \leq \omega_A$. Taking expectations gives $E[W_A] \leq E[\omega_A] = 1/\lambda_A \leq E[\omega_A^{\text{out}}] = 1/\lambda_U$. The fixed subset is what limits covered levers; the inequality supplies no comparison of costs, defensive improvements or other intervention effects. ■

MODEL/EXT/code/check_ext.py, the canonical copy named in the Declarations, checks finite instances of E1–E3 using marked primitive arrivals, direct flaw-level exposure and numerical survival integration. For each result it also confirms that the check fails when the formula it tests is deliberately altered. Its output is MODEL/EXT/code/check_ext_out.txt. These numerical checks supplement the proofs; they do not validate the event or policy premises empirically.

Appendix C. Proofs for Proposition 16

Notation and set-up are those of Appendix A, Section A.3.5.

P16(i). On a capability branch each duration is its capability distance divided by s plus a fixed adoption tail; on its adoption branch it is constant. Differentiating gives $sT' = -\chi_c \tau$ and $st'_B = -\chi_B q$. The derivative of $y(L - t_B)^+$ is $-yt'_B$ for arrival before L and zero for arrival after it, so $sB'_L = yq\chi_B\eta_L$. For the speed-set hazard, with S constant,

$$sC'_c = sH'T + HsT' = H(\varepsilon_H T - \chi_c \tau). \quad (C1)$$

Multiplying by V_c and subtracting from sB'_L gives the speed-set formula. For the level-set hazard, the change of variables $c = st$ turns the hazard integral over the capability ramp into $(1/s) \int_0^K h(c)dc = \tau\bar{h}$, and the adoption tail contributes $(T - \tau)h(K)$. Differentiating,

$$sC'_c = -\tau\bar{h} + (\tau - \chi_c \tau)h(K), \quad (C2)$$

which gives the level-set formula. Since s, H, V_c and T are positive, rearranging gives every strict threshold and its equality case. With $\chi_c = 0, T = T_a$, which gives the two tests above the cyber kink. On the level-set row where adoption sets the cyber clock and capability the benefit, multiplying the test by s gives $yK_B > V_c K[h(K) - \bar{h}]$, which does not depend on speed while that row holds. The four cells of Section 4.7's table are substitution, keeping the mixed or sequential duration. ■

P16(ii). Integrating the constant flow over $[t_B, \infty)$, or over $[t_B, L]$ when $t_B < L$, gives B_r or $B_{r,L}$; its speed derivative times s is $yq\chi_B e^{-rt_B}$, with η_L in the finite case. Since $A'_r(T) = e^{-rT}$,

$$s[HA_r(T) + S_r]' = H\varepsilon_H A_r(T) - H\chi_c \tau e^{-rT} + sS'_r.$$

Subtracting V_c times this from the benefit derivative gives (ii)'s formula, and dividing by $V_c HA_r(T) > 0$ gives the thresholds. ■

Dated steps. A step mass that is constant undiscounted need not be constant discounted. For fixed nonnegative step weights S_j at differentiable dates $t_j(s)$, $S_r(s) = \sum_j S_j e^{-rt_j(s)}$ and $sS'_r = -r \sum_j S_j st'_j(s) e^{-rt_j(s)}$. If the dates are fixed in calendar time, $S'_r = 0$. If capability sets them, $t_j = k_j/s$, then $sS'_r = r \sum_j S_j (k_j/s) e^{-rk_j/s} \geq 0$, strictly when some positive mass has a positive distance, and earlier steps subtract from sW'_r . Step weights must be in expected-event units before multiplication by V_c , or converted once to DALYs, not both.

The discounted level-set alternative. Discount the actual path, including its constant endpoint-hazard tail:

$$C_{c,r}^{\text{lev}} = (1/s) \int_0^K e^{-rc/s} h(c)dc + h(K)[A_r(T) - A_r(\tau)] + S_r.$$

Define a derivative-weighted level average, which need not be nonnegative, $\bar{h}_{r,D}(s) = K^{-1} \int_0^K e^{-rc/s} (1 - rc/s) h(c)dc$. Then $sd/ds[(1/s) \int_0^K e^{-rc/s} h(c)dc] = -\tau\bar{h}_{r,D}$, and the discounted tail's derivative times s is $h(K)\tau[e^{-r\tau} - \chi_c e^{-rT}]$. Continuity and boundedness

on the finite capability domain justify differentiating under the integral. Adding the step derivative and subtracting from the benefit derivative,

$$s(W_r^{\text{lev}})' = yq\chi_B e^{-rt_B} - V_c \tau \{h(K)[e^{-r\tau} - \chi_c e^{-rT}] - \bar{h}_{r,D}\} - V_c sS_r', \quad (\text{C3})$$

with η_L on the benefit term in the finite-horizon version. (C3) covers parallel, mixed and sequential cyber timing, whatever the benefit timing, and its right side being positive is the discounted level-set test. A level-set hazard that does not depend on speed as a function of capability does depend on it in discounted calendar time, so the undiscounted level-set signs do not carry over without (C3). In one illustrative case of our checks, s times the undiscounted derivative is 5 and s times the discounted one is -0.099 .

Limits and the case without benefits. As $r \downarrow 0$, the discounted speed-set and level-set derivatives converge to the undiscounted ones, when the same benefit-activity indicator is used and the fixed step masses have bounded differentiable dates; this follows by dominated convergence and $A_r(T) \rightarrow T$. Finite-horizon benefit levels converge to B_L . The perpetual level B_r diverges as $r \downarrow 0$, although its speed derivative has the stated finite limit. Discounting lowers the marginal credit for a late benefit and changes the cyber marginal loss too, so there is no general ranking of the net sign as r changes. With $S_r' = 0$, $\chi_c = 1$ and no active benefit term, the speed-set threshold reduces to $\varepsilon_H < r\tau/(e^{rT} - 1)$; with parallel timing, $T = \tau$, this is P4(ii)'s $x/(e^x - 1)$. Above the cyber kink, without an active benefit term, the speed-set discounted derivative is negative whenever $S_r' \geq 0$.

The kink claims. The maximum of continuous clock functions is continuous. At a clock crossing each one-sided derivative is given by its own branch, so substituting an indicator of one on the lower-speed side and zero on the higher gives the changes listed in Section A.3.5. At a horizon crossing only η_L changes. When crossings coincide, the changes are simultaneous and add. The one-sided conditions for a maximum follow from the signs of the difference quotients; strict opposite signs make the kink's value strictly greater than all nearby values, a strict local maximum. On the benefit's capability branch, the horizon equation $t_B = L$ gives $k_L = K_B/(L - m_B T_b)$, and $L > T_b$ places it below the benefit kink. ■

P16(iii). Each objective is continuous on the compact positive interval when its primitives are continuous, so by the extreme value theorem it attains both extrema. At a differentiable interior extremum $W' = 0$. The remaining candidates are the endpoints and the kinks. Plateaus and nonisolated critical sets are possible, so all stationary points must be included, not only isolated crossings. ■

Existence needs both bounds. Compactness gives existence, not uniqueness, and not location above the cyber kink. Without an upper bound a maximizer need not exist: with sequential benefit timing, parallel cyber timing, $K = T_a = K_B = T_b = V_c = y = 1$, $S = 0$, $L = 10$ and $H(s) = 1 - 1/(2s)$ on $[1, \infty)$, which has $H > 0$ and $H' > 0$, the balance is $W_L = 8 - 1/(2s)$, strictly increasing with an unattained supremum of 8. A positive lower bound is needed too; P1's proof gives a failure without it.

Sufficient conditions for an interior maximum. These follow from the mean value theorem on the smooth pieces and continuity across the finitely many clock and horizon kinks.

1. If the right derivative at $s_{\min}$ is positive and the left derivative at $s_{\max}$ is negative, neither endpoint maximizes, so at least one global maximum is interior. It may be a kink or part of a flat set.
2. At a smooth stationary speed s_0 , $W''(s_0) < 0$ suffices for a strict local maximum. More generally, a derivative positive just below s_0 and negative just above it suffices, including at a clock kink with strict one-sided signs; equality in a one-sided sign alone does not.
3. If $W' > 0$ at all smooth speeds below s_0 and $W' < 0$ at all smooth speeds above it, with continuity across the finitely many kinks, s_0 is the unique global maximum. A smooth candidate on one branch needs its local sign pattern and then a comparison with all other candidates and the endpoints; a first-order condition alone is not a global result.
4. A concrete condition above the cyber kink. Take $H(s) = as^\alpha$, with $a, \alpha > 0$, on the branch where adoption sets the cyber clock and capability the benefit, with arrival inside L . There $W_L = y(L - m_B T_b) - yK_B/s - V_c a T_a s^\alpha - V_c S$ and $sW'_L = yK_B/s - \alpha V_c a T_a s^\alpha$, which is strictly decreasing in s and vanishes at $s_0 = (yK_B/(\alpha V_c a T_a))^{1/(\alpha+1)}$. If s_0 lies strictly inside that branch, it is the branch's unique strict maximum, with $W''(s_0) = -(\alpha + 1)yK_B/s_0^3 < 0$. For finite benefit timing this needs $\max(s_{\min}, k_c, k_L) < s_0 < \min(s_{\max}, k_B)$, with k_c the cyber kink and k_B the benefit kink; sequential benefits have no upper benefit kink, and with $L \leq T_b$ the branch does not exist. The candidate is a global maximum if it also beats the other candidates, or if condition 3 holds on the whole interval. If instead sW'_L is positive just above k_c and negative just below k_B , with arrival already active, continuity gives an interior maximum on that branch, unique there because sW'_L strictly decreases. Section 4.7's example is the case $a = \alpha = 1$.

The same compactness and one-sided conditions apply to the discounted objectives and to the level-set alternative; they do not imply that a smooth optimum above the kink exists in those cases.

Numerical checks. MODEL/code/check_benefits.py (paths relative to the deposit named in the Declarations) uses only Python's standard library, deterministic grids, direct calendar-time hazard integration and finite differences, and prints MODEL/code/check_benefits_out.txt. It checks the undiscounted and discounted derivatives for all nine timing pairs and both hazards, with a largest relative error of 2.44×10^{-9} ; both kink orderings and coincident kinks for $m \in \{0, 0.35\}$ and $m_B \in \{0, 0.6\}$, with one-sided derivatives, continuity and the changes at each kink; horizon activation and the case $L \leq T_b$; the thresholds, including equality, with and without discounting and dated steps; the limits as $r \downarrow 0$ and the reduction to P4(ii); the level and marginal effects of a shorter T_b ; the illustrative optimum, with a grid maximum at $s = 2$ where $W_L = 6$, $W'_L(2) = 0$ and $W''_L(2) \approx -1$; the counterexample at $s = 5$; the power-hazard condition for $\alpha = 0.4, 1.0$ and 2.3 ; and the unbounded case. It also confirms that six deliberately wrong formulas fail, among them a benefit that keeps a speed derivative after adoption binds, benefits discounted while cyber losses are not, and benefits counted before they arrive. All 129 lines of output, the summary included, report PASS. The parameter values are illustrative, never

estimates; the checks supplement the proofs and do not test the model's empirical premises.

Supplement: evidence snapshot and versions read

Supplement to “Running Through the Rain: Speed, Secrecy and Risk in the AI Cyber Transition” (Allan C. Stam, University of Virginia), draft v6, 25 October 2026. Section, proposition and reference citations refer to the paper. Supplement S1 records the version of each work we read where the paper’s reference list needs that stated; Supplement S2 is the claim-by-claim source audit, with the data note behind our calculations from CISA’s catalogue. Both are dated: they describe the sources as we read them in October 2026.

S1. Versions read

Entries are keyed by the paper’s in-text citation and follow the reference list’s order. References not listed here have no separate version note.

- **Adkins, Evron and Schneier 2025.** The reprint on schneier.com, with the CSO original saved for byline, date and subtitle, fetched 23 October 2026. The article is not paginated; the body pinpoints its subtitle and section headings. No DOI; Crossref and OpenAlex hold no record.
- **Aghion and Howitt 1990.** NBER Working Paper 3223 (January 1990). The published version was revised after refereeing, so the body cites a section, not an equation number. The published page range is not given, because only its first page could be confirmed (Crossref, OpenAlex).
- **Aghion et al. 2001.** Harvard DASH accepted manuscript, October 2000.
- **AISI 2026.** Saved copy fetched 23 October 2026. The 4.7-month figure is AISI’s February 2026 estimate as reported in this post; the fit is shown as a figure, not published as data.
- **Anderson and Moore 2006.** Authors’ 11-page manuscript (Cambridge Computer Laboratory). The body’s quotation and p. 1 pinpoint refer to this manuscript, not to the edited Science text.
- **Anthropic 2025.** Saved copy of the announcement post, fetched 22 October 2026. The quoted phrase “in a small number of cases” is in this post, not in the accompanying November 2025 PDF report. This is Anthropic’s account of the campaign.
- **Anthropic 2026a.** Saved copy. This is a lab’s report on its own model.
- **Anthropic 2026b.** Saved copy. This is a lab’s report on its own model.
- **Anthropic 2026c.** Saved copy, whose text matched the live page on 22 October 2026. This is a lab’s report on its own model.
- **Anthropic 2026d.** Saved copy. These are Anthropic’s reported evaluations, including its runs of the external ExploitBench benchmark.
- **Anthropic 2026e.** Overview and process explanation; archive snapshot dated 7 October 2026, 12:53:16 UTC, <https://web.archive.org/web/20261007125316/https://red.anthropic.com/2026/cvd/>. No archive capture dated 2 October is held. The live page fetched 22 October

2026, 19:55:56 UTC, was byte-identical to that archive capture. These are the lab's dated disclosure and patch counts, not a cohort patch rate or a measure of deployment.

- **Anthropic 2026f.** Saved copy fetched 23 October 2026. This is a lab's report on its own model.
- **Anthropic 2026g.** Saved copy fetched 23 October 2026. This is a lab's report on its own model.
- **Anthropic 2026h.** Saved copy fetched 23 October 2026. This is a lab's report on its own model.
- **Arbatskaya and Mialon 2010.** The published version.
- **Arbaugh, Fithen and McHugh 2000.** Author-hosted PDF of the published *IEEE Computer* layout, with printed folios 52–59. Author order follows the printed byline; Crossref and OpenAlex list McHugh, Fithen, Arbaugh.
- **ARC Prize 2026.** Data file generated 21 October 2026, downloaded 23 October 2026.
- **Armstrong, Bostrom and Shulman 2013.** Future of Humanity Institute Technical Report #2013-1 (October 2013). The body pinpoints by section because the pages differ.
- **Arrow 1962.** NBER chapter scan (nber.org/chapters/c2144). The DOI 10.1007/978-1-349-15486-9_13 is for a 1972 reprint, not the version read.
- **Askill, Brundage and Hadfield 2019.** arXiv v1.
- **Belson 2025.** Page fetched 23 October 2026. Cloudflare's own traffic data.
- **Brumley et al. 2008.** BitBlaze project PDF, which has no page numbers. Pages are per OpenAlex.
- **Buterin 2023.** Saved copy fetched 23 October 2026. The essay is not paginated; the body pinpoints its section headings. No DOI; Crossref and OpenAlex hold no record.
- **Carlini et al. 2024.** arXiv:2403.06634v2 (9 July 2024), which carries the ICML/PMLR 235 line.
- **Charrier and Weiner 2024.** Saved article text.
- **Charrier et al. 2026.** Saved page text, fetched 23 October 2026. Counts are detected and disclosed zero-days, cutoff 31 December 2025.
- **CISA 2026.** JSON and CSV feeds, fetched 23 October 2026 (1,754 entries; dateAdded 3 November 2021 to 20 October 2026). Class shares in the body are our calculation from the catalogue's CWE tags (Supplement S2, data note); CISA publishes none.
- **Cognition 2026.** Saved copy, 23 October 2026. The leaderboard was launched on 17 July 2026. Cognition's own models (SWE-1.6, SWE-2) are among those listed.
- **DARPA 2025.** Saved copy fetched 23 October 2026, carrying an undated editor's note that corrects the number of synthetic vulnerabilities in the final from 70 to 63. This is the competition organizer's report of its results; ARPA-H's page of 4 September 2025 repeats the same counts.
- **Debenedetti et al. 2025.** Abstract page, fetched 23 October 2026; the passage cited is in the abstract. Preprint.

- **DeepSeek-AI et al. 2025.** arXiv v2 (4 January 2026) abstract page, 23 October 2026; the passage cited is in the abstract. This is a lab's report on its own model.
- **Dixit and Pindyck 1994.** Chapters 1, 2, 4, 5 and 8 only. The body cites chapter 4 (pp. 93–132) by section.
- **dots team 2026.** Saved copy, 23 October 2026. This is a lab's report on its own model; no IMO record of the marking was located.
- **Epoch AI 2026a.** Export downloaded 23 October 2026, files dated 23 October 2026. The FrontierMath rows are Epoch's own evaluations on the private sets; the FrontierSWE and FrontierCode rows are Epoch's copies of the benchmark owners' leaderboards.
- **Epoch AI 2026b.** Saved copy, 23 October 2026. FrontierMath was developed with funding from OpenAI, which has exclusive access to a subset of the benchmark.
- **FBI et al. 2023.** Saved page text of the July 2025 revision, fetched 23 October 2026.
- **FinCEN 2024.** Saved PDF, fetched 23 October 2026.
- **Fonseca et al. 2017.** Author-hosted PDF, abstract and §4.4, 23 October 2026.
- **Ford 2026.** Page fetched 23 October 2026.
- **FrontierSWE 2026.** Saved copy, 23 October 2026.
- **GAO 2025.** A copy of the PDF downloaded 21 October 2026; gao.gov refused scripted retrieval on 23 October 2026. Modernization statuses are agency-reported.
- **Google 2026.** File fetched 23 October 2026, last row 21 October 2026. Share of users reaching Google over IPv6.
- **Grunewald, Mazumdar and Vanderlee 2026.** Saved page text, fetched 23 October 2026. Covers vulnerabilities disclosed 1 January 2025 to 31 August 2026.
- **Gudibande et al. 2023.** arXiv v1.
- **Hinton, Vinyals and Dean 2015.** arXiv v1.
- **Hoffmann et al. 2022.** arXiv v1 abstract page, 23 October 2026; the passage cited is in the abstract.
- **Horn 2018.** Full post, including n. 1, which excludes Meltdown from the initial report.
- **Kaplan et al. 2020.** arXiv v1 abstract page, 23 October 2026; the passage cited is in the abstract.
- **Kapoor et al. 2024.** arXiv v1 (27 February 2024). An ICML 2024 position-paper version exists and was not checked.
- **Kenna Security and Cyentia Institute 2019.** PDF fetched 23 October 2026 (© 2019 Cyentia Institute); Key Findings, p. 4, and the survival analysis, pp. 18–19. Commissioned by Kenna Security, which supplied the data. The remediation-time figures are a survival analysis of 190 million vulnerabilities at 12 Kenna customer organizations, with data to late 2018.
- **Khatry et al. 2025.** Abstract page, fetched 23 October 2026; the passage cited is in the abstract.

- **Lermen, Rogers-Smith and Ladish 2024.** arXiv:2310.20624v2 (22 May 2024), the workshop version. It was first posted in October 2023, and OpenAlex confirms the author order.
- **Li and Paxson 2017.** Author-hosted camera-ready PDF carrying the ACM copyright block and DOI. Its pages carry no proceedings folios, so the body pinpoints sections.
- **Linton and Parseghian 2018.** Opening of the post; the mitigation-status table was not in the saved text.
- **METR 2025.** Saved copy fetched 23 October 2026 (the earlier address, evaluations.metr.org/gpt-5-report/, redirects here); Appendix, “METR’s access to GPT-5”. This is an external evaluator’s account of its own access period.
- **METR 2026.** Saved copy, 23 October 2026.
- **Microsoft 2026.** Page fetched 23 October 2026. Vendor documentation of configurable defaults, not a measured rollout.
- **Nappa et al. 2015.** Authors’ 17-page copy, numbered 1–17 rather than with the proceedings folios, so the body pinpoints sections.
- **Nasr et al. 2025.** Abstract page and v1 PDF (table 1), fetched 23 October 2026. Preprint.
- **Nordhaus 1969.** Chapter 5, “The Economics of Patents” (pp. 70ff). No DOI; the edition is confirmed by 1970–71 reviews.
- **Owen 2024.** arXiv v1 abstract page, 23 October 2026; the passage cited is in the abstract.
- **Palmer and Sleevi 2014.** Page fetched 23 October 2026.
- **PutnamBench 2026.** Both downloaded 23 October 2026. Entries are submitter-reported.
- **Rescorla 2004.** The 18-page workshop paper (sos-vo.org PDF). A shorter version, not read, appeared in *IEEE Security & Privacy* in 2005, <https://doi.org/10.1109/MSP.2005.17>. Workshop name per the paper’s acknowledgements and Anderson and Moore’s reference [12] (WEIS 3, 2004).
- **Rice 1953.** Full text from the AMS archive PDF. The body pinpoints the theorem and its interpretation because the text read has no printed page numbers. The garbled proof of Theorem 6 is not quoted.
- **Sandbrink et al. 2022.** SSRN working paper, with the title shown here; the later Science and Public Policy article was not read or cited.
- **Schneier 2018.** The author’s reprint of the March/April 2018 column on schneier.com; the print text was not compared. The column is one page.
- **Seger et al. 2023.** arXiv:2311.09227v1, the same report.
- **seL4 Project n.d.** Saved copy, 23 October 2026, the answers to “Does seL4 have zero bugs?” and “What are the proof assumptions?”. This is the project’s account of its own proofs.
- **Shevlane 2022.** arXiv v1 (13 January 2022), all 26 pages; the body cites the abstract on p. 1. The current arXiv version, dated 11 April 2022, is titled “Structured Access: An Emerging Paradigm for Safe AI Deployment” and was not read.

- **Shevlane and Dafoe 2020.** arXiv:2001.00463v2.
- **Snell et al. 2024.** arXiv v1 abstract page, 23 October 2026; the passage cited is in the abstract.
- **Stix et al. 2025.** arXiv v1.
- **The White House 2025.** PDF fetched 23 October 2026; Introduction, p. 1.
- **Thompson et al. 2025.** Page fetched 23 October 2026. Google’s own Chrome telemetry.
- **Tramèr et al. 2016.** arXiv:1609.02943v2, the extended version (<https://doi.org/10.48550/arXiv.1609.02943>). The body pinpoints the abstract.
- **Trump 2026.** Federal Register full text, fetched 23 October 2026; §3.
- **Van den Steen 2004.** Published text. The body cites the choice-driven optimism mechanism; its application to lab safety is the paper’s own interpretation.
- **Weeds 2000.** Warwick Economic Research Paper 576 (20 September 2000). The body pinpoints by section.
- **Westerbaan 2024.** Page fetched 23 October 2026. Cloudflare’s own connection data.
- **Whalley 2016.** Page fetched 23 October 2026. States the planned removal date; the actual release date was not checked.
- **Yang et al. 2024.** Abstract page, fetched 23 October 2026; the passage cited is in the abstract. A later ASE 2025 version was not read.
- **Yang et al. 2011.** Author preprint PDF, §3.1 “Testing CompCert”, 23 October 2026.

S2. Source audit and KEV data note

The table lists the empirical claims of Sections 2, 4.7 and 5–11 that rest on outside records, as this draft states them. For each it gives the primary source; whether that source is independent of the lab, meaning the developer or operator whose model or system it reports on; and the claim’s status after our claim-by-claim check. VERIFIED means the record says what the text says. CORRECTED means the record supports a narrower or reworded claim, which is the one the text now states. DERIVED means our calculation from the record. JUDGMENT means our own reading, labeled as such in the text and consistent with the record. Checking a report establishes what it reports, not that its figures are right. Model results are proved in Appendices A and B and are not listed. Claims we could not support were cut and do not appear.

Claim, as stated in the text	Source	Independent of the lab?	Status
The restricting lab expects defense to dominate in the long run, after a fraught transition (Section 2)	Anthropic 2026g	No	VERIFIED

Claim, as stated in the text	Source	Independent of the lab?	Status
That expectation is not evidence (Section 2)	—	NA	JUDGMENT
AI may upend attack's advantage and tip the scales toward defense (Section 2)	Schneier 2018	NA	VERIFIED
A defense-favoring world is better; promoting defense-favoring technology as differential technology development; cyber examples that are not AI vulnerability-finding (Section 2)	Buterin 2023	NA	VERIFIED
A speculative path to automatic patching before production, with a warning that AI may tip the scales to attackers (Section 2)	Adkins, Evron and Schneier 2025	NA	VERIFIED
No source read states the case as a whole or speaks of sprinting through a window (Section 2)	—	NA	JUDGMENT
The holder called its model general-purpose and unreleased, did not plan general availability, and expected the capability to proliferate soon (Sections 2, 5, 7.1)	Anthropic 2026a	No	VERIFIED
AlphaFold as the	Jumper et al. 2021,	No; the developers'	VERIFIED

Claim, as stated in the text	Source	Independent of the lab?	Status
first computational method to predict protein structures regularly with atomic accuracy where no similar structure is known; determining one structure took months to years (Section 4.7)	abstract and opening paragraph	report on their own method	
Half of the 2024 Nobel Prize in Chemistry to Hassabis and Jumper “for protein structure prediction” (Section 4.7)	Nobel Prize Outreach 2024	Yes	VERIFIED
Clinical phase about 95 months against 31 non-clinical, a figure the report attributes to an earlier study; the non-clinical stage includes target discovery; Phase 3 the longest stage at 38.0 months; FDA review 16.2 months including the sponsor’s time; clinical phases may overlap (Section 4.7)	Sertkaya and Berger 2024, pp. v, 12	NA	VERIFIED
About 18 months from target discovery to a preclinical candidate, in one programme (Section 4.7)	Ren et al. 2024, abstract	No; the developer’s report on its own programme	VERIFIED
Under 30 months	Xu et al. 2025,	No; the developer’s	VERIFIED

Claim, as stated in the text	Source	Independent of the lab?	Status
from the start of target discovery to the end of phase 0/1 testing, followed by a phase 2a trial (Section 4.7)	introduction	report on its own programme	
Ransomware attacks cut hospital volume by 17–24 percent in the attack week and raise in-hospital mortality by 34–38 percent among patients already admitted (Section 4.7)	Neprash, McGlave and Nikpay 2026, published abstract	NA	VERIFIED
30-day acute myocardial infarction mortality up as much as 0.36 percentage points in the 3 years after a breach, which the authors associate with remediation they did not directly observe (Sections 4.7, 10.2)	Choi, Johnson and Lehmann 2019, abstract and discussion	NA	VERIFIED
Targeted hospital cyberattacks may be associated with disruptions at nontargeted hospitals (Section 4.7)	Dameff et al. 2023, abstract	NA	VERIFIED
The DALY combines premature death and time in less than optimal health, one DALY being one lost year of healthy life; WHO's current	WHO 2024, pp. 6, 7	NA	VERIFIED

Claim, as stated in the text	Source	Independent of the lab?	Status
estimates use no discounting for time (Section 4.7)			
Best FrontierMath scores of 93.7% on Tiers 1–3 and 100% on Tier 4, by OpenAI models, on a benchmark OpenAI funded and partly holds (Section 5)	Epoch AI 2026a; 2026b	Yes; OpenAI’s funding and access are stated	VERIFIED
Tier 4’s private set has 41 problems (Section 5)	Epoch AI 2026b	Yes	DERIVED
Best ARC-AGI-2 semi-private score of 95% (Section 5)	ARC Prize 2026	Yes	VERIFIED
An IMO 2026 perfect score, on a lab’s own account of official marking (Section 5)	dots team 2026	No	VERIFIED
Five PutnamBench entries report all 672 Lean problems; the records do not document public verification; three name general frontier models (Section 5)	PutnamBench 2026	No; submitter-reported	CORRECTED
FrontierSWE and FrontierCode unsaturated, at 65.5% and 54.6%, with about four and eight months of dated rows (Section 5)	FrontierSWE 2026; Cognition 2026; Epoch AI 2026a	Partly; Cognition lists its own models on FrontierCode	CORRECTED
Too little data for a trend fit (Section 5)	—	NA	JUDGMENT
METR’s 50%	METR 2026	Yes	VERIFIED

Claim, as stated in the text	Source	Independent of the lab?	Status
horizon, its 16-hour warning, and no update since 8 September 2026 (Section 5)			
Scaling laws are fitted to loss (Section 5)	Kaplan et al. 2020; Hoffmann et al. 2022	NA	VERIFIED
Extrapolation errors of 6 and 18 percentage points (Section 5)	Owen 2024	NA	VERIFIED
Improved reasoning through reinforcement learning after pre-training (Section 5)	DeepSeek-AI 2025	No	CORRECTED
Test-time compute can let a smaller model outperform a fourteen-times larger one (Section 5)	Snell et al. 2024	NA	VERIFIED
Rice’s theorem and its room for special methods (Section 5)	Rice 1953	NA	VERIFIED
seL4’s proof assumptions; no guarantee about covert timing channels (Section 5)	seL4 Project n.d.	No	CORRECTED
CompCert 1.6’s front-end bugs and PowerPC semantics gap (Section 5)	Yang et al. 2011	Yes	CORRECTED
Sixteen bugs in three verified distributed systems, mostly at interfaces (Section 5)	Fonseca et al. 2017	Yes	VERIFIED

Claim, as stated in the text	Source	Independent of the lab?	Status
AI Cyber Challenge final: 54 unique synthetic vulnerabilities found in the final's 63 challenges and 43 patched (63 synthetic vulnerabilities in all, by DARPA's corrected count); 18 real found and 11 patched (Section 6.1)	DARPA 2025	Yes	VERIFIED
More than 500 high-severity vulnerabilities found with Claude and validated by human researchers (Section 6.1)	Anthropic 2026f	No	CORRECTED
181 working exploits against two successes, in a harness without the browser's sandbox (Section 6.1)	Anthropic 2026g	No	VERIFIED
The 181 successes include repeats (Section 6.1)	Anthropic 2026g; 2026h	No	DERIVED
Eight Firefox exploits, timed by the fastest of three trials per stage; patches public for at least 90 days (Section 6.1)	Anthropic 2026c	No	CORRECTED
A 4.7-month doubling of the 80%-reliability cyber horizon, with AISI's own caveats (Section	AISI 2026	Yes	VERIFIED

Claim, as stated in the text	Source	Independent of the lab?	Status
6.1)			
Average time-to-exploit of five days, or 47 with outliers (Section 6.1)	Charrier and Weiner 2024	NA	VERIFIED
A campaign that succeeded “in a small number of cases” (Section 6.1)	Anthropic 2025	No	VERIFIED
Counts of exploitation in the wild after 2023 exist but do not isolate AI (Section 6.1)	Charrier et al. 2026; Grunewald, Mazumdar and Vanderlee 2026; CISA 2026	Yes	CORRECTED
The balance depends on adversaries and shifts with investment (Section 6.1)	Slayton 2017; Garfinkel and Dafoe 2019	NA	VERIFIED
Much of the quantitative evidence comes from one lab (Section 6.1)	—	NA	JUDGMENT
6,157 vulnerabilities disclosed and 516 known patched; patches a lagging indicator; human triage the rate-limiting step (Sections 6.1, 6.5)	Anthropic 2026e	No	VERIFIED
Median remediation of 90 days and 32.5% open after 180 days, at 12 firms (Section 6.2)	Kenna Security and Cyentia Institute 2019	NA	VERIFIED
Seven days to 90% of devices, as Anthropic’s reading of Windows Autopatch; the	Anthropic 2026c; Microsoft 2026	Anthropic: yes, about Microsoft’s channel; Microsoft: no	CORRECTED

Claim, as stated in the text	Source	Independent of the lab?	Status
Microsoft pages examined do not state it (Section 6.2)			
Client hosts: 45 days to patch half of vulnerable hosts; at most 14% patched when exploits are released; 28% of patches reach 95% of hosts (Section 6.2)	Nappa et al. 2015	Yes	VERIFIED
Over half of CVEs patched in repositories more than a week before disclosure; attackers can get a jump of weeks to months (Section 6.2)	Li and Paxson 2017	Yes	VERIFIED
Patches available at least a month before intrusions; systems vulnerable months or years after corrections; the prevented intrusions likely nuisance attacks (Section 6.2)	Arbaugh, Fithen and McHugh 2000	Yes	VERIFIED
Patching series bear on the share of adoption that waits for the capability; retirement and migration series on whether adoption binds (Section 6.2)	—	NA	JUDGMENT
Three of ten critical federal legacy systems modernized by February 2025 (Section 6.2)	GAO 2025	Yes; agency-reported status	VERIFIED

Claim, as stated in the text	Source	Independent of the lab?	Status
HTTPS in Chrome from 30–45% to 95–99% (Section 6.2)	Thompson et al. 2025	No	VERIFIED
Post-quantum key agreement on nearly 2% of TLS 1.3 connections to Cloudflare in March 2024 and, on a different denominator, 52% of the human-generated Web traffic Cloudflare observed in December 2025 (Section 6.2)	Westerbaan 2024; Belson 2025	No	CORRECTED
About two and a half years planned from SHA-1 announcement to removal (Section 6.2)	Palmer and Sleevi 2014; Whalley 2016	No	DERIVED
IPv6 access to Google past 50% on one day, about 27 years after RFC 2460 (Section 6.2)	Google 2026; Deering and Hinden 1998; Ford 2026	Google: no; Internet Society: yes	DERIVED
The externality quotation (Section 6.2)	Anderson and Moore 2006	NA	VERIFIED
Shares of access-policy and memory-corruption classes among known exploited vulnerabilities, with CWE-20 excluded and restricted by CVE identifier year as checks (Section	CISA 2026; the data note below	Yes	DERIVED

Claim, as stated in the text	Source	Independent of the lab?	Status
6.3)			
Batch listing in 2022 and the yearly memory-corruption shares behind Predictions 2 and 3, with the conditional binomial calculations (Section 11)	CISA 2026; the data note below	Yes	DERIVED
Help-desk calls to reset passwords and transfer tokens (Section 6.4)	FBI et al. 2023	Yes	VERIFIED
Generative AI reportedly used to impersonate executives (Section 6.4)	FinCEN 2024	Yes	CORRECTED
Twelve defenses against jailbreaks and prompt injections bypassed by adaptive attacks (Section 6.4)	Nasr et al. 2025	Yes	CORRECTED
CaMeL solves 77% of AgentDojo tasks with provable security, against 84% undefended (Section 6.4)	Debenedetti et al. 2025	No (developer's own evaluation)	VERIFIED
Defenses that need no frontier capability (Sections 6.4, 10.2)	Sandbrink et al. 2022 for the principle	NA	JUDGMENT
A benchmark of 100 C repositories, correctness assessed by test cases, with 15 solved single-shot by the best model (Section 6.5)	Khatry et al. 2025	Yes	CORRECTED

Claim, as stated in the text	Source	Independent of the lab?	Status
Claude-2 with VERT passed bounded model checking on 42% of 1,394 competitive-programming programs (Section 6.5)	Yang et al. 2024	No (developer's own evaluation)	CORRECTED
GLM-5.3 released 14 August 2026, weights two weeks later; about four months behind the US frontier among released models CAISI evaluated (Section 7.1)	CAISI 2026	Yes	VERIFIED
ExploitBench: 50 of 410 against 56 of 410 (Section 7.1)	Anthropic 2026d	No	VERIFIED
Open models' benchmark lag: a 90% interval of five to twenty-two months (Section 7.1)	Cottier et al. 2024	Yes	CORRECTED
Spectre's report and coordinated date (Sections 7.1, 7.2)	Horn 2018; Linton and Parseghian 2018	No	VERIFIED
RAND, in 2024: protecting internet-connected models against the most capable actors "currently not feasible", current applicability unverified; at least one known leak; the equities process (Section 7.3)	Nevo et al. 2024	Yes	VERIFIED
More information	Armstrong, Bostrom	NA	VERIFIED; the

Claim, as stated in the text	Source	Independent of the lab?	Status
raises the danger in one race model; a wide gap may calm the race on one reading (Section 7.4)	and Shulman 2013		reading is JUDGMENT
Every similarly capable model other than GLM-5.3 released with safeguards or limited access (Section 7.4)	Anthropic 2026d	No	VERIFIED
Structured access as the developer's controlled interaction (Section 7.4; Appendix B)	Shevlane 2022	NA	VERIFIED
Abliterated versions of the open-weight model within days (Section 8.3)	Anthropic 2026d	Yes; a competitor's report on another developer's model	VERIFIED
US policy frames AI as a race (Section 9)	The White House 2025	NA	VERIFIED
METR's GPT-5 evaluation took three weeks (Section 9)	METR 2025	Yes	VERIFIED
Sunk capital and falling compute prices may push labs to ship (Section 9)	—	NA	JUDGMENT
EO 14409 §3: a voluntary framework for access up to thirty days before release (Section 10.2)	Trump 2026	NA	VERIFIED

Data note: weakness classes in CISA's catalogue (Sections 6.3 and 11). DERIVED. These are our calculations, run on 23 October 2026, from the CISA catalogue of known exploited vulnerabilities, version 2026.10.21 (released 21 October 2026; 1,754 entries; dateAdded 3 November 2021 to 20 October 2026). An entry's weaknesses are its cwes strings of the

form CWE-*n*. Entries with no CWE, or only NVD-CWE-noinfo or NVD-CWE-Other, are excluded. An entry with several weaknesses takes the first matching class in the order memory corruption, injection and input handling, access policy, other. Periods are by dateAdded: through 2023, and from 2024. The class lists are our judgment:

- *Memory corruption* (43 CWEs): 119–127, 131, 134, 170, 190, 191, 401, 415, 416, 457, 466–469, 476, 562, 590, 680, 690, 761–763, 786–788, 805, 806, 822–825, 843, 908, 909, 1284.
- *Injection and input handling* (59 CWEs, including the catch-all CWE-20): 20, 22–24, 35, 36, 59, 61, 73, 74, 77–80, 88–91, 93–98, 113, 116, 117, 138, 140, 141, 150, 158, 159, 172, 176, 177, 179, 180, 184, 434, 436, 444, 470, 502, 564, 601, 611, 641, 643, 652, 776, 829, 913, 915, 917, 918, 1236, 1321, 1336.
- *Authentication, authorization and access policy* (109 CWEs, including the exposure CWEs 200, 201, 209 and 215): 200, 201, 209, 215, 250, 254–257, 259, 260, 264, 266–277, 279–294, 302–309, 312, 313, 319, 345–347, 352, 359, 384, 425, 441, 497, 521–523, 532, 538, 539, 548, 549, 551, 552, 565, 566, 598, 612, 613, 620, 639, 640, 642, 648, 668, 669, 672, 693, 703, 732, 749, 798, 799, 836, 840–842, 862, 863, 912, 922, 923, 939, 940, 1021, 1188, 1220, 1230, 1231, 1233, 1390–1392.
- *Other classified*: any other CWE.

	2021–23	2024–26
Entries	1,053	701
Excluded, no usable CWE	133	42
Usable (the denominator)	920	659
Memory corruption	299 (32.5%)	125 (19.0%)
Injection and input handling	368 (40.0%)	295 (44.8%)
Authentication, authorization and access policy	171 (18.6%)	180 (27.3%)
Other classified	82 (8.9%)	59 (9.0%)

Entries whose weaknesses fall in more than one of the first three classes, resolved by the order above, number 17 and 7. Any other order changes any one class's count by at most those numbers, so its share by at most 17/920 and 7/659, that is, by less than 1.9 and 1.1 percentage points.

CWE-20. Excluding the catch-all CWE-20 from the usable denominator, as NVD-CWE-noinfo is, the usable counts are 839 and 648 (214 and 53 excluded). The four classes, memory corruption, injection and input handling, access policy and other, then number 299, 287, 171 and 82 of 839 in 2021–23 and 125, 284, 180 and 59 of 648 in 2024–26; the memory-corruption and access-policy shares are 35.6% and 20.4% in 2021–23 and 19.3% and 27.8% in 2024–26.

Listing of older flaws. The feed has no publication date, so we restrict by CVE identifier year. Keeping entries listed in their identifier's year or the next, the usable counts are 454 and

532; memory corruption is 160 (35.2%) and 105 (19.7%), and access policy 90 (19.8%) and 149 (28.0%). Additions numbered 311 in 2021 (from 3 November), 555 in 2022, 187 in 2023, 186 in 2024, 245 in 2025 and 270 in 2026 through 20 October. In 2022, 381 of the 555 additions (68.6%) were CVEs at least three identifier-years old; across 2021–22 the share was 49.7% (430 of 866).

Yearly shares. By year of dateAdded, the memory-corruption share of usable entries, with CWE-20 included, was 32.0% (86 of 269) in 2021, 33.1% (161 of 486) in 2022, 31.5% (52 of 165) in 2023, 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.

Prediction 3's cut. With a fixed stationary share $p = 125/659$ and independent entries, the probability that one year's share exceeds a cut is a conditional binomial calculation, not a measured error rate. At the earlier cut of 19.0% it is 51.2%, 50.9% and 49.4% at $n = 178$, 220 and 261 usable entries, and pooling two years leaves it at 49.3% ($n = 440$) and 47.4% ($n = 522$). At 23.0% it is 10.1% at $n = 178$ and between 4.4% and 7.4% at every n from 220 to 261. If AI makes old C and C++ code cheap to harvest for flaws, the share could rise with no change in the residual introduction of new flaws; that is a hypothesis.

Two separate implementations give identical counts, and the catalogue's JSON and CSV feeds agree on every entry's dateAdded.