
It Only Has to Fail Once

Why the Expected Value of Automated Defense Depends on the Blast Radius

Gerald C. Kane, Ph.D.

C. Herman and Mary Virginia Terry Chair in Business Administration
Department Head, Management Information Systems
Terry College of Business, University of Georgia

© 2026 Gerald C. Kane. Licensed under CC BY-NC 4.0: share and adapt for noncommercial use with credit to the author; do not present as your own. Full terms on the last page.

Executive Summary

Charles Perrow spent the years after Three Mile Island working out why some organizations suffer catastrophes that no one made a mistake to cause. His answer, in *Normal Accidents* (Perrow, 1984), separates two properties that are usually run together. Interactive complexity lets an unexpected combination of small failures occur at all. Tight coupling turns that combination into a catastrophe, because there is no slack, no buffer, and no time to intervene between the local fault and the global consequence. Complexity is the trigger. Coupling is the propagator. Organizations reliably reach for the wrong remedy: they try to eliminate the failures, which is impossible, instead of loosening the coupling, which is engineering.

That distinction is the argument of this paper, applied to a question senior leaders are being asked to decide right now: how much authority to give an AI system that changes production software.

The question has a well-argued answer in circulation. Writing after OpenAI's Black Hat USA presentation on the July 2026 Hugging Face intrusion, Ben Thompson (2026) set out an asymmetry. An attacker's automated attack carries positive expected value: a failed exploit changes nothing, and the attack only has to work once. A defender's automation carries negative expected value: success merely preserves the status quo, while a bad patch breaks production, so the agent "only needs to fail once for the entire endeavor to have a negative payoff." Thompson does not conclude that defense should stay manual. He argues that the only durable answer to automated attack is automated defense, and that firms which keep a human in the loop too long will be making a mistake. His prediction concerns who gets there first: incumbents, pricing the downside, will delay; startups, whose base case is failure, will not; and so the AI-native winners will be new entrants. The mechanism is real, and Anderson (2001) made the coverage version of the same point a quarter-century ago: attack is simply easier than defense.

The half of the argument that leaders control is the unit at which the risk is counted. Expected value is not fixed by the nature of defense. It is the outcome of a comparison the firm structures, and two terms of that comparison are set by design. The first is how much of the estate a single bad change can reach before a health check sees it and something reverts it. The second is what remaining unpatched costs, which is a number well above zero and is usually left off the ledger. Bounding exposure does not make the number positive on its own; it shrinks one loss term and adds a cost term, and whether the total clears zero depends on the avoided breach losses and on the failures the health check cannot see. What bounding exposure does is turn a sign that looked fixed into a quantity that management can compute and move.

The measured record supports that reading. Elite delivery teams in the DORA research program run a five percent change failure rate while deploying multiple times a day, and only 8.5 percent of teams in the most recent edition get below two percent at all (DORA, 2024, 2025). Under a rule that says one failure poisons the endeavor, essentially no organization could ship software. They ship anyway, because risk is priced at the level of the blast radius. CrowdStrike's own root-cause analysis of the July 2024 Falcon outage lists six findings, and the sixth calls for staged deployment of every content template instance (CrowdStrike, 2024b). The firm that suffered the canonical catastrophic automatic update did not conclude that automatic updates carry negative expected value. It concluded that unstaged ones do.

Three further findings shape the paper. Autonomous find-and-patch on real code has an existence proof in DARPA's AI Cyber Challenge (DARPA, 2025); an unattended production loop does not, at least among the systems whose operators have published their process. Anthropic's disclosure dashboard, a dated snapshot of a branching pipeline, shows 26,153 candidate findings and 421 upstream patches as of late August 2026, with the company naming human triage as the rate-limiting step for disclosure (Anthropic Frontier Red Team, 2026). And Census Bureau data show AI use rising with firm size, which rules out the reading that incumbents are holding back from AI in general while saying nothing about who eventually wins (Grundy, Breaux, & Khatiwoda, 2026).

Read as a risk-appetite problem, the incumbent's disadvantage is close to unfixable, since a firm with something to lose cannot decide to have nothing to lose; read as a deployment-architecture and verification problem, it is ordinary engineering and governance work. The paper connects this to the four-curves picture from *The Technology Fallacy* (Kane, Phillips, Copulsky, & Andrus, 2019), proposes on a provisional basis that the Authority question in the GOALS governance framework be qualified by an exposure decision, and closes with what a leadership team can settle in ninety days.

I. Where This Argument Sits

The Technology Fallacy (Kane et al., 2019) built its diagnosis on four curves moving at different speeds: technology changes faster than individuals adopt it, individuals adapt faster than organizations, and organizations adjust faster than legal and societal institutions. The organization is the layer that has to absorb the mismatch. Thompson has restated that mismatch in financial terms, which is a useful thing to have done. The adaptation gap normally shows up as a story about culture or decision rights. Here it shows up as an arithmetic claim, and an arithmetic claim has a testable implication the softer version lacks: if the gap is arithmetic, it closes when the arithmetic changes, and the terms of the arithmetic are set by the organization.

The Transformation Myth (Kane, Nanda, Phillips, & Copulsky, 2021) was written partly to correct a misreading of the first book. Readers heard "the problem is organizational" as an instruction to slow down. The second book's answer is that disruption has become the normal condition and the response is to move boldly while building an organization that can adapt. Thompson's expectation that incumbents will keep humans in the loop for too long is a well-argued version of the same slowdown, which is why it earns an answer at length.

Prior papers in this series have taken up who carries the cost of being wrong (Kane, 2026a, 2026c, 2026e, 2026f, 2026i) and, more specifically, where the check on an AI system's output can be a machine and where it has to be a person (Kane, 2026g, 2026h). This paper asks a narrower question, at what unit of exposure the cost gets counted, and leans on the second body of work for the distinction between what a deployment health check can verify and what a human reviewer still has to. The July 2026 Hugging Face intrusion itself has been examined in this series for forensics sovereignty (Kane, 2026j) and for whether an agent's claim of completion can be trusted (Kane, 2026d), and is carried here by reference.

One caution belongs at the front. This argument engages Thompson's framing of an event, and Thompson's framing rests on a single conference presentation. The remarks attributed in Section IV to Michael Dalton of OpenAI, who presented the Hugging Face post-incident session at Black Hat USA with his colleague Eric Wallace, are taken from Thompson's transcription. The underlying

recording exists, but the wording here has not been checked against it. Where the argument needs to stand on its own, it stands on the measured literature.

II. What Expected Value Is a Property Of

Take the pessimistic argument in its own terms first, because the standard version of it is stronger than the AI-era version.

Security is asymmetric because the attacker needs one working path and the defender needs to close all of them. Anderson (2001) made this argument with a stylized model, comparing the defender to a sheriff in a thinly populated territory where the men in black hats can strike anywhere. Gordon and Loeb (2002) supplied the investment corollary, showing that a rational firm should spend no more than roughly 37 percent of the expected loss on protecting an information set, a bound later shown to hold generally (Gordon, Loeb, Lucyshyn, & Zhou, 2015). Anderson and Moore (2006) supplied the reason firms fall short even of that: systems fail when the person guarding them is not the person who suffers when they fail.

Three points about this literature. Its foundation is analytic. Its central investment result is a rationality finding, which is the most common way the Gordon-Loeb bound gets misused. And its explanation for defensive underinvestment is externality and misaligned liability, not internal downside variance.

That last point is where the AI-era version departs from its ancestor. Thompson is advancing a payoff claim about automation specifically: the attacker's failed automation costs nothing, while the defender's failed automation breaks production. He presents it as his own inference from Dalton's remarks. No body of measurement stands behind it in the way twenty-five years of work stands behind the coverage asymmetry. It is a plausible analytic extension with one very good illustrative case.

Defining the comparison

An expected value is a property of a wager, and it has no meaning until the wager is specified. Thompson's version compares the automated defender's outcome against an unhacked status quo: success returns to that status quo, failure falls below it, so the expectation is negative. That is an expected loss, and it is measured against the wrong baseline. The alternative to automated remediation is the current process, in which the median known-exploited vulnerability takes 43 days to close and most are still open a week after detection (Verizon Business, 2026). The comparison that matters is between two imperfect states, over a stated horizon.

Written out, the difference in expected value between automating remediation and continuing the current process has three terms. The first is avoided loss: the reduction in breach probability that comes from closing exposure windows faster, multiplied by what a breach costs. The second is residual loss: the probability that an automated change is defective, multiplied by how much of the estate it reaches, for how long, at what cost per unit of exposure, plus any damage that persists after the change is reverted. The third is the cost of the machinery: progressive delivery infrastructure, health gates, rollback automation, and the human review and triage that remain. Automation pays when the first term exceeds the sum of the other two.

Exposure design acts on the second term only, and only on part of it. A change shipped to the whole estate at once has a downside bounded only by the size of the estate. The same change shipped to two percent of traffic behind a statistical health gate that reverts on error has a downside bounded by two percent, for the minutes it takes the gate to trip. Google's site reliability practice states the arithmetic directly: deploying a candidate that fails one request in five to a five percent canary produces a one percent overall error rate, and impact on the error budget is proportional to the traffic exposed (Warner & Davidovič, 2018). What the gate cannot bound is damage the gate cannot see. A patch that removes one vulnerability and introduces another produces no error rate. A change that corrupts records, rotates a credential, or migrates a schema may not be undone by rolling the binary back. Those failures belong to the second term as well, and exposure design does not shrink them. They are why human review of the patch remains a separate control from the health check on the deployment, a point Section VII takes up.

A hypothetical makes the shape concrete; none of the figures that follow are measured. Suppose a firm ships a thousand automated remediations a year at the elite DORA failure rate of five percent, so fifty are defective. If each defect reaches the whole estate and costs two million dollars in outage and recovery, the residual loss term is a hundred million dollars, and Thompson's sign is correct at almost any plausible level of avoided loss. If each defect instead reaches two percent of the estate for fifteen minutes before an automated gate reverts it, and costs a few thousand dollars, the same fifty defects cost a few hundred thousand. Whether the wager now pays depends on the other two terms: the review and triage cost of running the pipeline, and the breach losses the faster closure actually avoids. If the avoided losses are several million and the machinery costs a few million, the sign is positive. If a meaningful share of defects are of the kind a health gate cannot detect, the hundred-million term reappears in proportion, and no ring design removes it. The value of the exercise lies in the form of the answer: a calculation with terms management controls, made against the actual alternative.

This is Perrow's distinction doing its work. The defect is interactive complexity, which cannot be eliminated. The catastrophe is tight coupling, which can be engineered against. The coupling the pessimistic argument attributes to defense belongs to the release pipeline, and the release pipeline can be changed.

The industry that ships software at scale has formalized this. Google's site reliability practice treats an availability target as both a floor and a ceiling, on the reasoning that a hundred percent is almost never the right target and that framing it as a ceiling is what permits deliberate risk-taking (Alvidrez, 2016). The canary chapter is explicit about the conclusion it refuses to draw: rather than advocating against change, the discipline exists to measure the reliability impact of releases so that changes, including security patches, can continue to ship (Warner & Davidovič, 2018). Microsoft's architectural guidance concedes the premise, saying there are no low-risk deployments to production, and then supplies the other side of the ledger: failing to deploy regular updates introduces other risks, such as security vulnerabilities that must be addressed through deployments (Microsoft, 2026). That second clause is the avoided-loss term, and Section III reports its size.

III. What the Deployment Record Measures

The base rates

The DORA research program has published change failure rate by performance cohort for over a decade. In the 2024 edition, elite performers report a five percent change failure rate while deploying on demand, multiple times per day; low performers report 40 percent while deploying between once a month and once every six months (DORA, 2024). The cohort table is not monotonic that year, and DORA says so, noting that change failure rate has long been the outlier among its four key metrics and adding a fifth, rework rate, in response. The endpoints are the informative part. The teams that ship most often break things too. Their failures are cheap, with failed-deployment recovery measured in under an hour.

The 2025 edition replaces the cohort model with a distribution across roughly 5,000 respondents. Only 8.5 percent of teams achieve a change failure rate below two percent, and roughly four in ten fail more than one change in six (DORA, 2025). A decision rule holding that a single failure produces a negative payoff would require the software industry to stop shipping. It has not stopped, because the industry does not evaluate changes one at a time against a zero-failure standard.

DORA's finding on AI looks like evidence against automation and on inspection supports the argument here. The 2024 report states that, contrary to its expectations, AI adoption was negatively affecting delivery performance, with an estimated 7.2 percent reduction in delivery stability for every 25 percent increase in AI adoption (DORA, 2024). The 2025 report finds throughput has since turned positive while instability persists, and frames AI as an amplifier of whatever engineering system it lands in (DORA, 2025). DORA's own explanation of the instability is batch size: AI makes it easy to write more code per change, and larger changes have always failed more. That explanation is DORA's interpretation of correlational survey data, offered alongside the amplifier framing rather than as a proven single cause, and it should be carried at that weight. It is nonetheless a supported mechanism, consistent with a decade of the program's prior findings on small batches, and it points at packaging rather than at AI as such.

The case, and what its owner concluded

The July 2024 CrowdStrike outage is the load-bearing example for the pessimistic view, and it should be granted in full. A Rapid Response Content update defined 21 input parameter fields where the sensor supplied 20, producing an out-of-bounds read and a kernel crash on Windows hosts online between 04:09 and 05:27 UTC on 19 July (CrowdStrike, 2024a, 2024b). The defective content was reverted 78 minutes after release; approximately 99 percent of Windows sensors were back online ten days later. Template instances for the same channel file had shipped successfully several times between March and April before the July attempt failed. Repeated successes and one remembered failure is Thompson's asymmetry rendered as a production incident, and it should be conceded.

What CrowdStrike concluded is the instructive part. Of the six findings in the root-cause analysis, three concern the defect and three concern how the change was delivered. Finding 6 reads in its entirety: "Each Template Instance should be deployed in a staged rollout." The stated mitigation is that instances passing canary testing are to be successively promoted to wider deployment rings or rolled back if problems are detected, with bake-in time between rings, plus customer control over

where and when Rapid Response Content is deployed (CrowdStrike, 2024b). Microsoft, the platform that absorbed the event, reached a compatible attribution, observing that a reliability problem of this kind leads to widespread availability issues when it is not combined with safe deployment practices (Weston, 2024). CrowdStrike's sensor releases already used staged rollout with customer-controlled versioning while its content channel did not, a contrast that is suggestive but not a controlled comparison: the two pipelines differ in artifact, cadence, and testing, and only one of them carried the defect. The weight rests on the company's own conclusion, not on the contrast.

The obvious objection surfaces here, and it deserves a direct answer. Some changes cannot be staged. A kernel-mode driver update that halts the host on load defeats its own telemetry: the machine cannot report that the change was bad, and cannot execute the code that would revert it. That objection is correct, and Channel File 291 is precisely that case. It also demonstrates the mechanism instead of refuting it. Where exposure cannot be bounded at the level of the host, it can still be bounded at the level of the fleet, and deployment rings are what does the bounding. CrowdStrike's remedy operated at exactly that level: rings, bake time, and customer-controlled timing, all of which cap how much of the world can be running the same untested change at the same moment. The objection moves the question from whether to bound exposure to at which level to bound it, and that is a design question with an answer.

The counterweight is that partial deployment is not automatically safe. In August 2012, a Knight Capital technician deploying new code failed to copy it to one of eight servers, and the resulting inconsistent fleet produced a \$460 million loss and a \$12 million penalty; the Securities and Exchange Commission found the firm had no written procedure requiring review of the deployment, and noted that a second technician's review could have identified the missed server (U.S. Securities and Exchange Commission, 2013). An inconsistent estate is its own failure mode. Progressive delivery requires automated health gates and automated rollback, not merely shipping to some machines and hoping.

The cost of the other side of the ledger

Verizon's 2026 Data Breach Investigations Report finds exploitation of vulnerabilities has become the leading initial access vector at 31 percent, up from 20 percent the prior year. The median time to fully remediate a vulnerability on the CISA Known Exploited Vulnerabilities catalog rose to 43 days from 32, and the share fully remediated fell to 26 percent from 38. The survival analysis, run across more than a billion anonymized detection records, reports that between 60 and 70 percent of known-exploited vulnerabilities remain open seven days after detection regardless of year, volume, or organizational maturity (Verizon Business, 2026).

Set that beside RAND's finding that a zero-day exploit has an average life expectancy of 6.9 years and that the median time to develop a working exploit is 22 days (Ablon & Bogart, 2017), and the shape of the wager is visible. The attacker's asset depreciates over years. The defender's remediation window is measured in weeks and mostly missed. Whatever automation is worth, it is being compared against a status quo that is neither neutral nor free.

One qualification. RAND's collision estimate of 5.76 percent per year was contested four months later by Herr, Schneier, and Morris (2017), who examined more than 4,300 vulnerabilities and found rediscovery rates of 15 to 20 percent within a year. Two studies of comparable quality

measured the same quantity and differed by a factor of three to four. The measured foundation under the offense-advantage claim is contested, which is useful to know before building strategy on it.

IV. The Existence Proof, and Where the Loop Stops

Dalton's lessons-learned summary at Black Hat, as Thompson transcribes it, contains a specific empirical claim: fully automated offense is now demonstrably possible, and there is no comparable existence proof for full automation of the core defensive loops (Thompson, 2026).

The claim depends on what counts as the loop.

Autonomous find-and-patch on real code has been demonstrated. DARPA declared a proof of principle in 2016, when the Cyber Grand Challenge put seven machines against each other for roughly ten hours on purpose-built software in a closed testbed (DARPA, 2016). The AI Cyber Challenge, concluded at DEF CON in August 2025, moved it onto real open-source code. Seven finalist systems analyzed more than 54 million lines; across 63 synthetic challenges they discovered 54 vulnerabilities and patched 43, and they additionally found 18 real, previously unknown vulnerabilities and supplied 11 patches, at an average cost of about \$152 per task and an average patch submission time of 45 minutes (DARPA, 2025). That was a scored competition with known challenge structure and a rubric that rewarded accurate patches. The step from there to an unattended loop pushing changes into a live production estate is a separate one, and the two should not be run together. What the competition establishes is that the discovery and repair steps can be executed by machines on realistic code at low cost. What it leaves open is the deployment step and the verification that precedes it.

The per-team results matter more to a leadership audience than the aggregate, because they show that finding and fixing are separate capabilities with a wide gap between them. The winning system, Team Atlanta's, found 43 and patched 31. The sixth-place system found 41, nearly as many as the winner, and patched three. The same gap appears wherever it is measured. Anthropic reports its Sonnet 4.5 model succeeding on 76.5 percent of Cybench challenges at ten attempts, while only 15 percent of its generated patches were judged semantically equivalent to the human reference patch, a figure the company discloses was produced by having the model judge its own work (Anthropic, 2025). Google's OSS-Fuzz team, reporting 26 new vulnerabilities found through AI-generated fuzz targets, laid out a five-step workflow and stated that the AI executed the first four, with the fifth step, fixing, still an open research problem (Chang, Liu, & Metzman, 2024).

So Dalton's claim survives on the stricter reading. Among the systems whose operators have published how they work, none runs the full loop unattended in production; each keeps a human at the patch gate and says so. Google states that all patches generated by its CodeMender agent are reviewed by human researchers before submission, having upstreamed 72 fixes over six months (Popa & Flynn, 2025). OpenAI's security agent, launched as Aardvark and since folded into Codex Security, attaches a generated patch to each finding for human review (OpenAI, 2025). GitHub's Copilot Autofix presents fixes a developer can accept, edit, or dismiss, and reports remediation three times faster overall, seven times faster for cross-site scripting, and twelve times faster for SQL injection (Hanley, 2024). Whether some organization runs an unattended loop without

publishing it cannot be known from the outside; the claim here is bounded to what has been disclosed.

What the disclosure funnel shows

Anthropic publishes a coordinated vulnerability disclosure dashboard with hash commitments per finding, and it is the most auditable evidence available on where the pipeline narrows. As of August 26, 2026, it records 26,153 candidate findings produced by the company's models. The pipeline then branches. In the main branch, candidates go to one of six external security research firms or to Anthropic's own triage: 5,008 had been reviewed, 4,576 confirmed valid, a 91.4 percent true-positive rate among those reviewed, and 1,022 reported to maintainers. In a second branch, 1,278 findings were passed directly to maintainers who asked for untriaged output, a set the dashboard warns may contain false positives. Together, 2,300 findings had been disclosed across 392 open-source projects, 1,815 acknowledged, and 421 patched upstream (Anthropic Frontier Red Team, 2026).

Several cautions attach to reading 421 against 26,153 as a 1.6 percent conversion, and the dashboard supplies most of them. The denominator includes candidates not yet reviewed, duplicates of already-known bugs, and findings that are real but that maintainers reasonably decline to fix because they fall outside a project's threat model; the true-positive rate counts those as valid. The numerator lags by construction, since patches take time to create and the dashboard calls the patched count a lagging indicator. The figure is a snapshot on a date, and it has moved with every revision. And an upstream patch is not a deployed remediation: the dashboard notes that a maintainer's fix does not guarantee the patch has been widely installed, and the 43-day median from Section III is the downstream reality. The ratio is best read as an order-of-magnitude statement about a pipeline in motion, not as a measured conversion rate.

With those cautions in place, the constraint Anthropic itself names is organizational. The company states that the number disclosed is a subset of what its models have found because independent human triage and review is the rate-limiting step, and that it is holding confirmed real bugs unreported because of capacity limits (Anthropic Frontier Red Team, 2026). That is a statement about the disclosure stage specifically. It does not claim, and this paper does not claim, that human capacity explains the whole distance between candidates and patches. It does establish that the most auditable narrowing in the pipeline is a capacity constraint on human institutions, and that finding more has stopped being the problem.

Dalton anticipated exactly this, and the point should be credited to him. In the same remarks, as Thompson transcribes them, he warned that automating vulnerability discovery without automating patching would move the bottleneck to remediation and drown human engineers in findings, and he described the end state he wants as an agent that identifies a vulnerability, proposes a patch, and hands it to automated infrastructure that rolls the change out and rolls it back if there is an availability incident. That last clause is progressive delivery with automated rollback, which is the architecture this paper argues for. The OpenAI presenter named the blast-radius design as the destination, and Thompson agrees with it as a destination. What Thompson adds is a prediction about who will delay getting there, and that prediction is what Sections V and VI examine.

The industry's money corroborates the diagnosis. In March 2026, Google, Amazon, Anthropic, Microsoft and GitHub, and OpenAI jointly pledged \$12.5 million to the Linux Foundation's Alpha-

Omega project, with the stated purpose of helping maintainers move security beyond vulnerability discovery to actually deploying fixes and turning a flood of AI-generated findings into fast action (Kotsovinos & Flynn, 2026). The same Google post also claims its tools can "autonomously find and fix deep, exploitable vulnerabilities." A document that claims solved autonomy and simultaneously funds a program premised on the fix pipeline being the unsolved problem is telling you which half is marketing.

One further observation. XBOW, the AI system that topped the HackerOne United States leaderboard with roughly 1,060 submissions, discloses that its security team reviewed findings before submission to comply with HackerOne's policy on automated tools (XBOW, 2025). The canonical offense-side existence proof also had a human gate. The two gates have different origins. Offense's gate is a platform rule that can be dropped; defense's gate is a risk control that firms are choosing to keep. That is a choice about consequences, and it can be re-made once the consequences are bounded differently and the verification behind it is understood.

V. What the Adoption Evidence Can and Cannot Say

Thompson's prediction has two parts. Incumbents will keep humans in the loop too long, and the AI-native winners will be startups. The first is about delay inside firms; the second is about competitive outcomes over the long run. Adoption data speak to a narrower proposition than either, and it is important to be precise about which.

The Census Bureau's Business Trends and Outlook Survey samples nonfarm firms nationally on a probability basis. Between December 2025 and May 2026, national AI use ran between 17 and 20 percent, reaching 19.8 percent by the first week of May. Thirty-seven percent of firms with at least 250 employees reported using AI, and 32 percent of firms with 100 to 249 employees, against under 20 percent of firms with four or fewer. Use increased over the period among firms with at least 20 employees and did not change significantly among smaller ones (Grundy, Breau, & Khatiwoda, 2026). The Center for Economic Studies working paper behind the supplement finds use rates of 50 to 60 percent, or 60 to 70 percent employment-weighted, among very large firms in Information, Professional Services, and Finance (Bonney et al., 2026). McKinsey's 2026 survey, a non-probability sample of self-reports, finds the same gradient at the scaling stage: 54 percent of respondents at organizations above \$1 billion in revenue report scaling AI across the enterprise, against a third at smaller ones, and the share scaling agents rose from 27 to 40 percent at large organizations while staying flat at 22 percent at smaller ones (Tinkoff, Van der Veken, Chui, & Balakrishnan, 2026).

What these data rule out is the reading that incumbents are holding back from AI in general. Large firms adopt more, scale more, and are moving faster on agents. What the data cannot test is Thompson's actual claim. Small firms in a Census panel are not startups, adoption is not success, and the competitive outcome he predicts lies years out. His argument is also compatible with high incumbent adoption: a firm can adopt broadly and still keep a human in every loop that matters, which is precisely the pattern he expects. The adoption evidence therefore narrows the debate without settling it.

Where the data do bite is on depth. McKinsey reports that 37 percent of respondents attribute any EBIT impact to AI, essentially unchanged from the prior year, and that the share qualifying as high

performers held at six percent, also unchanged. Census finds that 57 percent of adopting firms use AI in three or fewer business functions and that most use it only to augment tasks (Bonney et al., 2026). Adoption is broad and shallow. Thompson is right that something is failing to convert, and the same McKinsey survey supplies the discriminator: nearly three-quarters of high performers report fundamentally redesigning workflows because of their AI use, up from 55 percent the year before, against one-quarter of everyone else (Tinkoff et al., 2026). The variable separating the six percent from the rest is whether the organization rebuilt the work, and that variable is available to firms of any size.

A calibration note bears on every forward-looking claim in this debate. Fourteen percent of McKinsey respondents reported that AI contributed to a decline in workforce size in the past year, against the 32 percent who, in the prior year's survey, expected exactly that over the same period (Tinkoff et al., 2026). Expectations overshot outcomes by more than two to one. A forecast is not evidence of a mechanism, and this literature is full of forecasts, including the ones in Section IX.

VI. The Disruption Frame

Thompson reaches his competitive conclusion through Christensen, asking whether AI is a sustaining or disruptive innovation and answering that for incumbents it will be sustaining, which is why the AI-native winners will be startups. The frame is load-bearing, so it deserves examination.

The theory has been tested once, systematically. King and Baatartogtokh (2015) surveyed 79 experts across 77 cases Christensen and Raynor had cited as disruption, asking whether each satisfied four elements of the theory; roughly nine percent fit all four, under a coding rule that gave the theory the benefit of the doubt wherever experts disagreed. The authors do not recommend discarding the theory, and their critics fairly note that four elements is a partial operationalization and one expert per case is thin. It remains the only systematic test. Christensen, Raynor, and McDonald (2015) themselves warn against accepting a "disrupt or be disrupted" mantra that leads incumbents to jeopardize their core business, which is to say the theory's author named reflexive disruption reasoning as an incumbent error. Chandy and Tellis (2000), testing the incumbent's-curse proposition directly, found that the variable discriminating radical innovators from the rest was willingness to cannibalize existing products, not firm size and not incumbency. That is a governance variable, decided by people in meetings.

Finally, who builds a thing and who profits from it are different questions. Teece (1986) established that when imitation is easy, returns flow to whoever controls the complementary assets. This series applied that result to the entrant-versus-incumbent question in *Cheap to Build, Hard to Win* (Kane, 2026b). The application here is narrow: frontier model capability is close to a commodity, available to attacker and defender alike, which is Thompson's own opening premise. When the core capability is commodity, distribution, proprietary data, regulatory standing, and customer relationships decide the outcome, and those are not held by the startups.

None of this refutes the startup thesis. It leaves it a hypothesis, resting on a frame whose one systematic test it did not pass, with the relevant discriminating variable being one that firm size does not determine.

VII. The Portfolio Answer, and Its Limits

If the sign of the wager depends on its structure, the remedy is to restructure the wager. Real options reasoning has said this for decades, and the operating version in this body of work has long been a preference for many small experiments over a single large bet, funding what works and killing what does not.

The measured case comes from an incumbent. Kohavi, Crook, and Longbotham, reporting on online controlled experiments at Microsoft, found that among well-designed experiments intended to improve a key metric, only about one-third succeeded, and they cite a second-hand report that Netflix considers 90 percent of what it tries to be wrong. Kohavi's later breakdown is that about a third of ideas were positive and statistically significant, a third showed no significant difference, and a third were negative, with Bing lower still and running experiments by the hundreds at any given time (Kohavi, Crook, & Longbotham, 2009; Kohavi, 2015). Microsoft operated at a hit rate of one in three, with a third of its ideas doing measurable harm, and treated that as the normal condition of the work, because no single treatment was a firm-level bet. Volume of cheap, bounded, independently reversible bets is what converts a 33 percent hit rate into compounding value. The startup's advantage in Thompson's account is that it is willing to bet the firm. The incumbent's available advantage is that it can run hundreds of bets a week and never bet the firm at all.

Three limits keep this from being a slogan.

The first is that a portfolio of small bets is not automatically a real option. Adner and Levinthal (2004) are explicit that a sequential stream of investment does not in itself constitute one, and warn that investments justified by real options logic may destroy value when the assumptions about abandonment flexibility turn out to be wrong. Their diagnosis is organizational: the executive may hold the option on paper, but the project's manager and team have every reason to keep it alive. Their testable prediction is that firms practicing real options reasoning should show higher project abandonment rates. Most firms cannot show that. Absent a working abandonment discipline, a portfolio of small bets degenerates into path dependence with an options vocabulary attached.

The second is that the ceiling on automated *fixing* is lower than the ceiling on bounded *deploying*, and the two are checked by different controls. The verified acceptance rates for automated repair are narrow and each is scoped to a bug class where a test or a static analyzer supplies the oracle: 42 percent direct acceptance for Meta's Getafix on null-dereference warnings (Bader, Scott, Pradel, & Chandra, 2019), and 15 percent of targeted sanitizer bugs fixed by Google's model-based patcher against a five percent baseline (Nowakowski & Keller, 2024). Google's July 2026 assessment is the most candid statement available: its newest automated patching program keeps engineers reviewing every proposed patch during its beta phase, publishes no acceptance rate, and acknowledges that large volumes of AI-generated patches can be net-negative once the cost of triage and review is counted (Kilian & Ingram, 2026).

This is where the distinction this series has drawn between machine verification and human checking does its work (Kane, 2026g, 2026h). A deployment health check is a machine verifier for a narrow question: is the service still up and serving within its error budget. That question has a cheap, fast, objective evaluator, and it is the reason progressive delivery can run unattended. Whether a patch is correct, and whether it closes the vulnerability without opening another, is a different question with no such evaluator in the general case. Tests catch some regressions; static

analyzers catch some bug classes; a semantically wrong but compiling patch passes both. That verification lands on a human reviewer, and it is where Google, OpenAI, and Anthropic have all chosen to keep one. The GOALS framework's certification discipline (Kane, 2026c) asks what must be true about an agent's failure behavior before authority is granted; the exposure decision asks how far a failure can reach once authority is exercised. The two are complementary, and neither substitutes for review of the change itself. A mandatory review gate and a canary ring are therefore two controls addressing two failure modes, and a firm that treats the second as a replacement for the first has confused availability with correctness.

The third limit is that some effects are delayed or irreversible, and for those the exposure design buys little. A schema migration, a credential rotation, a change to how records are written, or a patch that disables a check without anyone noticing all produce consequences that rollback does not undo and that a health gate may never observe. These are the cases where the traditional permission question keeps its full force, and where the answer to "may the agent act" should often remain no regardless of how small the first ring is.

There is also measured evidence that defensive change can be *less* disruptive than the status quo, which the pessimistic framing has no room for. Google's Android team recorded memory safety vulnerabilities falling from 76 percent of Android vulnerabilities in 2019 to 24 percent in 2024 after the shift to memory-safe languages, and observed that the rollback rate of Rust changes is less than half that of C++ (Vander Stoep & Rebert, 2024). The safer-by-construction path was also the operationally calmer one.

VIII. What a Leadership Team Can Settle in Ninety Days

This section introduces no named diagnostic. Five decisions are enough.

Qualify the authority question with an exposure question. The GOALS governance framework asks an Authority question, generally posed as a permission decision: may the agent act. That question stays. What changes is how it gets answered. For each class of change an agent might make, the answer should specify how much of the estate the change may reach before a human sees the result, what verifies the change before it ships and what monitors it after, and what reverts it if the monitor trips. Five controls do different jobs and should be named separately: review of patch correctness, verification that the security property actually holds, availability monitoring, containment through rings, and recovery through rollback. Changes whose effects are delayed or irreversible get a different answer than changes a health gate can catch in minutes. An organization that has written this down for its top ten change classes has answered the authority question in a form it can defend.

Find out whether your own pipelines are staged, and do it by looking. CrowdStrike had two deployment architectures on 19 July 2024 and only one of them produced a global outage. Most firms have a similar split, with the customer-facing application pipeline well instrumented and the security tooling, endpoint agents, and configuration pushes ungated because they were filed under operations and never treated as releases. Google's own postmortem analysis attributes 68 percent of its outages to binary and configuration pushes, 37 and 31 percent respectively (Google, 2018). A one-week audit of which change types can currently reach 100 percent of the estate in a single step

will locate the actual exposure, and the answer is usually not where the governance attention has been.

Price the abstention. Every proposal to automate a defensive loop should be evaluated against the measured cost of the current process, over a stated horizon, with the three terms of Section II written out. For most organizations the abstention number is available: the count of known-exploited vulnerabilities open past seven days, past thirty, and past ninety. If it resembles the industry distribution, between 60 and 70 percent are open at day seven (Verizon Business, 2026). A committee comparing automation risk against an implicit zero-risk status quo is answering a question that does not exist.

Buy the abandonment discipline before buying the portfolio. Adner and Levinthal's warning is the one most likely to bite, because the small-bets prescription is easy to adopt rhetorically and hard to operate. Before a firm claims a real options approach, it should be able to state its project abandonment rate for last year and name three initiatives it killed. If it cannot, the pilots will accumulate, and the firm will have bought the vocabulary without the mechanism.

Fund triage and review capacity alongside detection capacity. If discovery scales to five figures and upstream patches land in the low hundreds, buying more discovery lengthens the queue. Anthropic names human triage as the rate-limiting step for disclosure; five of the largest AI companies are jointly funding maintainer capacity for the same reason (Anthropic Frontier Red Team, 2026; Kotsovinos & Flynn, 2026). The equivalent question inside an enterprise is whether the security team's constraint is finding vulnerabilities or absorbing, reviewing, and deploying what it already finds. In most organizations that question has not been asked.

IX. The Next Twenty-Four Months

Four claims, stated so they can be checked, and carried with the calibration warning from Section V.

The first unattended defensive loop in production is more likely to appear in an organization that already runs progressive delivery at scale than in one that does not, and such organizations are disproportionately large. It will appear first in a narrow domain where a test suite supplies the oracle, such as dependency patching or configuration remediation. Nothing prevents a well-instrumented smaller firm from getting there first; the claim is about where the prerequisites currently sit, not about who is permitted to build them.

Change failure rate will become a board-visible metric at firms operating AI-assisted development at scale, because DORA has measured two consecutive years of AI adoption increasing delivery instability and its batch-size explanation implies a fix that lives in how teams package changes (DORA, 2024, 2025). Watch for the metric in risk committee packets before it appears in vendor marketing.

The find-and-fix gap will narrow on benchmarks faster than the deployed gap narrows in production, and the divergence between the two is the most useful indicator of whether this paper is right. If model patching improves substantially while the disclosure-to-deployed-fix pipeline stays slow, the constraint is organizational absorption and verification. If deployed remediation speeds up with model capability alone, it was a capability problem after all.

And the question of who converts AI into performance will not be settled by adoption rates, which are already concentrated in large firms, but by depth of workflow redesign, where the current evidence is three to one in favor of a small minority (Tinkoff et al., 2026). If that six percent stays at six percent through 2027 while adoption climbs, the conversion failure is real and Thompson's delay prediction will look better, not worse.

The decision in front of a leadership team is smaller than either camp in this debate finds it useful to admit: how much of the estate one change may touch, what verifies the change before it ships, what watches it afterward, and what puts it back. Those are questions a firm with something to lose is well placed to answer, and answering them is what converts a sign that looked fixed into a number the firm controls.

Methodology Note

This paper is a research commentary rather than a primary study. It draws on published measurement from four categories: engineering research programs reporting on their own populations (DORA, Google's site reliability and postmortem analyses, Meta's automated repair papers); government-run competition results (DARPA's Cyber Grand Challenge and AI Cyber Challenge); probability-sample government statistics (the Census Bureau's Business Trends and Outlook Survey and the associated Center for Economic Studies working paper); and vendor self-reports, which are labeled as such where they appear. Where a finding is contested in the literature, as with zero-day rediscovery rates, both estimates are reported. Where a figure is a forecast, it is identified as a forecast. The break-even example in Section II is a hypothetical with stated assumptions and no measured inputs beyond the DORA failure rate.

Four limitations bear on how the argument should be read. The triggering source is a commentary piece, and the conference remarks it transcribes have not been verified against the underlying recording; the expected-value asymmetry is Thompson's inference from those remarks rather than a claim the presenter made. The claim that defensive automation carries negative expected value has no dedicated body of measurement behind it, and the counter-argument here consequently addresses a claim that is itself unmeasured. The Anthropic disclosure figures are a dated snapshot of a branching pipeline and are read as an order-of-magnitude statement, not a conversion rate. And the McKinsey survey is a non-probability sample of self-reports; the Census figures carry the load wherever the two speak to the same question. Source-dated figures were re-verified against the live sources on September 4, 2026.

References

- Ablon, L., & Bogart, A. (2017). *Zero days, thousands of nights: The life and times of zero-day vulnerabilities and their exploits* (RR-1751-RC). RAND Corporation. https://www.rand.org/pubs/research_reports/RR1751.html
- Adner, R., & Levinthal, D. A. (2004). What is not a real option: Considering boundaries for the application of real options to business strategy. *Academy of Management Review*, 29(1), 74–85.
- Alvidrez, M. (2016). Embracing risk. In B. Beyer, C. Jones, J. Petoff, & N. R. Murphy (Eds.), *Site reliability engineering: How Google runs production systems*. O'Reilly Media. <https://sre.google/sre-book/embracing-risk/>

- Anderson, R. (2001). Why information security is hard: An economic perspective. In *Proceedings of the 17th Annual Computer Security Applications Conference* (pp. 358–365). IEEE Computer Society. <https://www.cl.cam.ac.uk/archive/rja14/Papers/econ.pdf>
- Anderson, R., & Moore, T. (2006). The economics of information security. *Science*, 314(5799), 610–613.
- Anthropic. (2025, October 3). *Building AI for cyber defenders*. <https://www.anthropic.com/research/building-ai-cyber-defenders>
- Anthropic Frontier Red Team. (2026, August 26). *Anthropic's coordinated vulnerability disclosure dashboard* (Revision 32). <https://red.anthropic.com/2026/cvd/>
- Bader, J., Scott, A., Pradel, M., & Chandra, S. (2019). Getafix: Learning to fix bugs automatically. *Proceedings of the ACM on Programming Languages*, 3(OOPSLA), Article 159. <https://arxiv.org/abs/1902.06111>
- Bonney, K., Breaux, C., Dinlersoz, E., Foster, L., Haltiwanger, J., & Pande, A. (2026). *The microstructure of AI diffusion: Evidence from firms, business functions, and worker tasks* (CES Working Paper No. CES-26-25). U.S. Census Bureau, Center for Economic Studies. <https://www.census.gov/library/working-papers/2026/adrm/CES-WP-26-25.html>
- Chandy, R. K., & Tellis, G. J. (2000). The incumbent's curse? Incumbency, size, and radical product innovation. *Journal of Marketing*, 64(3), 1–17.
- Chang, O., Liu, D., & Metzman, J. (2024, November 20). *Leveling up fuzzing: Finding more vulnerabilities with AI*. Google Online Security Blog. <https://security.googleblog.com/2024/11/leveling-up-fuzzing-finding-more.html>
- Christensen, C. M., Raynor, M. E., & McDonald, R. (2015). What is disruptive innovation? *Harvard Business Review*, 93(12), 44–53. <https://hbr.org/2015/12/what-is-disruptive-innovation>
- CrowdStrike. (2024a, July 24). *Preliminary post incident review (PIR): Content configuration update impacting the Falcon sensor and the Windows operating system (BSOD)*. <https://www.crowdstrike.com/en-us/blog/falcon-content-update-preliminary-post-incident-report/>
- CrowdStrike. (2024b, August 6). *External technical root cause analysis: Channel File 291*. <https://www.crowdstrike.com/wp-content/uploads/2024/08/Channel-File-291-Incident-Root-Cause-Analysis-08.06.2024.pdf>
- DARPA. (2016, August 4). "Mayhem" declared preliminary winner of historic Cyber Grand Challenge. <https://www.darpa.mil/news/2016/mayhem-winner-cyber-grand-challenge>
- DARPA. (2025, August 8). *AI Cyber Challenge marks pivotal inflection point for cyber defense*. <https://www.darpa.mil/news/2025/aixcc-results>
- DORA. (2024). *Accelerate state of DevOps report 2024*. Google LLC. <https://dora.dev/research/2024/dora-report/2024-dora-accelerate-state-of-devops-report.pdf>
- DORA. (2025). *State of AI-assisted software development*. Google LLC. https://services.google.com/fh/files/misc/2025_state_of_ai_assisted_software_development.pdf
- Google. (2018). Results of postmortem analysis (Appendix C). In B. Beyer, N. R. Murphy, D. K. Rensin, K. Kawahara, & S. Thorne (Eds.), *The site reliability workbook*. O'Reilly Media. <https://sre.google/workbook/postmortem-analysis/>
- Gordon, L. A., & Loeb, M. P. (2002). The economics of information security investment. *ACM Transactions on Information and System Security*, 5(4), 438–457.
- Gordon, L. A., Loeb, M. P., Lucyshyn, W., & Zhou, L. (2015). Externalities and the magnitude of cyber security underinvestment by private sector firms. *Journal of Cybersecurity*, 1(1), 3–17.

- Grundy, A., Breaux, C., & Khatiwoda, D. (2026, May 26). *Large firms with at least 20 employees biggest AI users*. U.S. Census Bureau, America Counts. <https://www.census.gov/library/stories/2026/05/ai-use-businesses.html>
- Hanley, M. (2024, August 14). *Found means fixed: Secure code more than three times faster with Copilot Autofix*. GitHub Blog. <https://github.blog/news-insights/product-news/secure-code-more-than-three-times-faster-with-copilot-autofix/>
- Herr, T., Schneier, B., & Morris, C. (2017). *Taking stock: Estimating vulnerability rediscovery*. Belfer Center for Science and International Affairs, Harvard Kennedy School.
- Kane, G. C. (2026a). *The CEO's AI agenda*. Working paper, Terry College of Business, University of Georgia.
- Kane, G. C. (2026b). *Cheap to build, hard to win*. Working paper, Terry College of Business, University of Georgia.
- Kane, G. C. (2026c). *The certification gap*. Working paper, Terry College of Business, University of Georgia.
- Kane, G. C. (2026d). *It said the work was done*. Working paper, Terry College of Business, University of Georgia.
- Kane, G. C. (2026e). *The relational premium*. Working paper, Terry College of Business, University of Georgia.
- Kane, G. C. (2026f). *The verification asymmetry*. Working paper, Terry College of Business, University of Georgia.
- Kane, G. C. (2026g). *The verification ceiling*. Working paper, Terry College of Business, University of Georgia.
- Kane, G. C. (2026h). *The verifier bottleneck*. Working paper, Terry College of Business, University of Georgia.
- Kane, G. C. (2026i). *What won't spread on its own*. Working paper, Terry College of Business, University of Georgia.
- Kane, G. C. (2026j). *Whose permission does your incident response need?* Working paper, Terry College of Business, University of Georgia.
- Kane, G. C., Nanda, R., Phillips, A. N., & Copulsky, J. R. (2021). *The transformation myth: Leading your organization through uncertain times*. MIT Press.
- Kane, G. C., Phillips, A. N., Copulsky, J. R., & Andrus, G. R. (2019). *The technology fallacy: How people are the real key to digital transformation*. MIT Press.
- Kilian, A., & Ingram, D. (2026, July 29). *From finding to fixing: Reducing maintainer burden with automated patches*. Google. <https://blog.google/security/from-finding-to-fixing-reducing-maintainer-burden-with-automated-patches/>
- King, A. A., & Baatartogtokh, B. (2015). How useful is the theory of disruptive innovation? *MIT Sloan Management Review*, 57(1), 77-90. <https://sloanreview.mit.edu/article/how-useful-is-the-theory-of-disruptive-innovation/>
- Kohavi, R. (2015, August). *Online controlled experiments: Lessons from running A/B/n tests for 12 years* [KDD 2015 keynote]. Microsoft Analysis and Experimentation. <https://exp-platform.com/Documents/2015-08OnlineControlledExperimentsKDDKeynoteNR.pdf>
- Kohavi, R., Crook, T., & Longbotham, R. (2009). *Online experimentation at Microsoft*. Third Workshop on Data Mining Case Studies and Practice Prize. https://ai.stanford.edu/~ronnyk/Exp_DMCaseStudies.pdf
- Kotsovinos, E., & Flynn, F. (2026, March 17). *Our latest investment in open source security for the AI era*. Google. <https://blog.google/innovation-and-ai/technology/safety-security/ai-powered-open-source-security/>
- Microsoft. (2026, June 17). *Architecture strategies for safe deployment practices* (Azure Well-Architected Framework, OE:11). <https://learn.microsoft.com/en-us/azure/well-architected/operational-excellence/safe-deployments>

- Nowakowski, J., & Keller, J. (2024, January 31). *AI-powered patching: The future of automated vulnerability fixes*. Google Security Engineering Technical Report. <https://storage.googleapis.com/gweb-research2023-media/pubtools/7563.pdf>
- OpenAI. (2025, October 30). *Introducing Aardvark: OpenAI's agentic security researcher* (updated March 6, 2026, on rebranding as Codex Security). <https://openai.com/index/introducing-aardvark/>
- Perrow, C. (1984). *Normal accidents: Living with high-risk technologies*. Basic Books.
- Popa, R. A., & Flynn, F. (2025, October 6). *Introducing CodeMender: An AI agent for code security*. Google DeepMind. <https://deepmind.google/blog/introducing-codemender-an-ai-agent-for-code-security/>
- Teece, D. J. (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy*, 15(6), 285–305.
- Thompson, B. (2026, August 24). *Autonomy and innovation*. Stratechery. <https://stratechery.com/2026/autonomy-and-innovation/>
- Tinkoff, D., Van der Veken, L., Chui, M., & Balakrishnan, T. (2026, August 25). *The state of AI in 2026: On the road to ROI*. McKinsey & Company. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>
- U.S. Securities and Exchange Commission. (2013, October 16). *In the matter of Knight Capital Americas LLC* (Securities Exchange Act Release No. 70694, Administrative Proceeding File No. 3-15570). <https://www.sec.gov/files/litigation/admin/2013/34-70694.pdf>
- Vander Stoep, J., & Rebert, A. (2024, September 25). *Eliminating memory safety vulnerabilities at the source*. Google Online Security Blog. <https://security.googleblog.com/2024/09/eliminating-memory-safety-vulnerabilities-Android.html>
- Verizon Business. (2026). *2026 data breach investigations report*. <https://www.verizon.com/business/resources/reports/dbir/>
- Warner, A., & Davidovič, Š. (2018). Canarying releases. In B. Beyer, N. R. Murphy, D. K. Rensin, K. Kawahara, & S. Thorne (Eds.), *The site reliability workbook*. O'Reilly Media. <https://sre.google/workbook/canarying-releases/>
- Weston, D. (2024, July 27). *Windows security best practices for integrating and managing security tools*. Microsoft Security Blog. <https://www.microsoft.com/en-us/security/blog/2024/07/27/windows-security-best-practices-for-integrating-and-managing-security-tools/>
- XBOW. (2025, June 24). *The road to top 1: How XBOW did it*. <https://xbow.com/blog/top-1-how-xbow-did-it>

Copyright, reuse, and citation

© 2026 Gerald C. Kane. This paper is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0), <https://creativecommons.org/licenses/by-nc/4.0/>.

You may share and adapt it for noncommercial purposes, provided you credit Gerald C. Kane as the author, link to the license, and indicate any changes you made. You may not present this material, or adaptations of it, as your own work, and you may not sell it or use it in paid training or consulting without written permission.

Cite as: Kane, G. C. (2026). *It Only Has to Fail Once*. <https://www.profkane.com/papers/it-only-has-to-fail-once.pdf>
Permissions and commercial use: jerry@profkane.com