The Economy Research

THE ECONOMY / RESEARCH PAPER SERIES

NOVEMBER 2026

[AI and Cyber-Security] Who Should Bear Losses from Inadequate Security?

Security Competition, Allocation of Liability and Resilient Systems

The Economy Research Editorial

RESEARCH PAPER

economy.ac/research

About the paper

This paper asks how much of the harm from an AI provider's inadequate security should be borne by providers, clients and other affected parties, and by what institutions losses should be allocated. It considers security claims, independent assurance, liability and shared defense.

Reading the evidence

The analysis brings together published security incidents and institutional evidence with simulation results from the companion SIAI research paper. The model outcomes are conditional on stated assumptions rather than direct estimates of real-world effects.

Author

The Economy Research Editorial

71 Lower Baggot Street
Dublin 2
Co. Dublin, D02 P593, Ireland

Publication

The Economy Research
20 November 2026
economy.ac/research

Originally published
20 November 2026
The Economy Research
Read the original paper

Contents

Abstract 1

1 Introduction - The Cost of Insecure AI Lands Outside the Firm 1

2 Customers Cannot Always Tell Secure Deployment from Security Claims 2

3 Assurance Can Discipline a Market and Close It to Entrants 5

4 Matching Each Control to the Actor That Can Carry It 6

5 Accountability, Independent Assurance, and Shared Defense Compared 9

6 Conclusion 13

References 14

Eight figures and one comparison table accompany the analysis. The reference list identifies the works cited in the manuscript.

Abstract

If the safety measures of an AI provider prove inadequate, most of the resultant harms would be borne by its users and their customers. Relatively little appears to be recovered contractually. The question addressed here is how much of these harms should be borne by various parties, and imposed on them by what institutions. This paper draws heavily on results, and all three graphs, from the the SIAI Research working paper Economic Incentives behind AI Safety Competition: Deployment Rivalry, Loss Internalization and Security Effort, November 2026. Various mechanisms, including trust marks, commitments and reputation, suggest that very little AI safety is directly observed by purchasers and so rewarded by markets. Assurance requirements under privacy law and for Government use of commercial cloud systems appear to impose mainly fixed costs, disproportionately deterring smaller firms. In the Salesloft Drift event in August 2025, the deficient safety measure or measures, the means to discover their inadequacy, and at least some of the resulting harms would all have been located with different entities. There is therefore a case for apportioning liability among several. As expected, in the associated model, increasing providers' liability would reduce total harms by 30 percent, but would cut entry from 70.6 to 17.6 percent of the firm sizes tested. Strict assurance requirements would decrease welfare by 43 percent. Common defence would improve both safety and overall welfare with no reduction in entry, while its legal basis holds. Optimal solutions will depend on how much of the value they create providers capture and how much of their customers' losses are passed to them contractually. Neither figure will generally be known.

1 Introduction - The Cost of Insecure AI Lands Outside the Firm

On 23 August 2025 Salesforce and Salesloft told Cloudflare that its Salesforce environment had been entered through Drift, an AI chat agent that Salesloft sold for website conversations and sales follow-up and that Cloudflare used to let visitors to its site contact the company.[1] The intruder had used stolen OAuth tokens belonging to the Drift integration, and Google Threat Intelligence Group later reported that the same method had been turned against the Salesforce instances of many Drift customers between 8 and 18 August.[2] In Cloudflare's case what left was the text of customer support cases: subject lines, correspondence and contact details for the people who had opened tickets. Some customers had pasted credentials into those tickets while troubleshooting, and Cloudflare's own search of the stolen material turned up 104 of its API tokens. The company rotated all of them, reissued secrets for its third-party integrations between 26 and 29 August, ran its own forensic investigation and on 2 September notified affected customers by email and dashboard banner, with a public apology that accepted responsibility for the tools it had chosen. Its account listed every cost it had absorbed and said nothing about recovering any of them from Salesloft.

Where the failure began became clear only later. Salesloft reported in September, on the basis of Mandiant's investigation, that the attacker had held access to its GitHub account from March to June 2025 and had then reached the Drift cloud environment, where it took the tokens Drift used to connect to customers' systems.[3] The control that failed sat with the vendor. The response work, the disclosure, the apology and the residual exposure of any customer whose

pasted password was never changed sat with the client and with the client's own customers, two steps away from the decision that left the vendor's code repository open. No public record shows how much of that cost, if any, travelled back up the chain through contract terms. That gap is informative in its own right, because what a client recovers after a supplier's AI integration fails is settled in an order form and a master subscription agreement that the client's customers, regulators and most analysts never see.

The arrangement is no longer unusual. Verizon's 2025 Data Breach Investigations Report, which covered more than 12,000 confirmed breaches, found a third party involved in 30 percent of them, double the share in the previous edition.[4] IBM's 2025 Cost of a Data Breach study found that 13 percent of surveyed organizations reported a breach of an AI model or application and that 97 percent of those lacked proper access controls for AI.[5] The two figures come from different samples and cannot be combined into one rate. They point to the same structure, in which AI services are wired into enterprise systems through integrations whose security is set by one party while most of the damage from their failure falls on another.

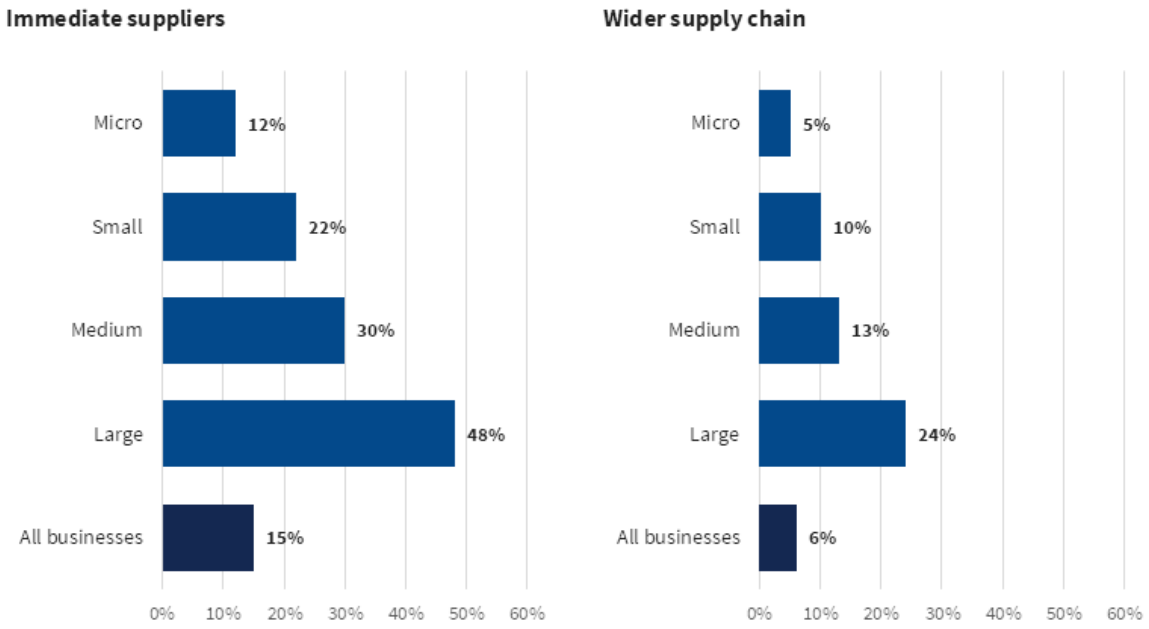
This paper asks who ends up paying for insecure AI and which institution should carry each part of the cost. The companion SIAI paper answers a narrower question with a formal model, asking when rivalry between two AI service providers raises or lowers the effort each puts into security, given how much external loss each bears and how much clients can see.[6] Its results are taken as given and summarized only where they are used. The analysis here is institutional. Chapter 2 asks whether customers can tell secure deployment from claims of it, which decides whether a market can reward security at all. Chapter 3 asks whether assurance requirements discipline providers or mainly keep smaller ones out, and Chapter 4 matches each control to the actor able to implement it, observe it and absorb what it fails to prevent. Chapter 5 compares accountability, independent assurance and shared defense by the incentive each changes, who pays, the information each needs and the conditions under which each performs poorly. Stephanie Pell, writing in a Brookings commentary series in September 2026, called for shaping market forces to reward security and for liability on entities that fail to take reasonable precautions.[7] Both proposals assume that someone can tell which precautions were reasonable, and that assumption is where the analysis begins.

2 Customers Cannot Always Tell Secure Deployment from Security Claims

A market rewards quality only when buyers can observe it before they pay or punish its absence afterwards. Akerlof showed that when sellers know quality and buyers do not, prices settle at the average, sellers of good products are underpaid and trade can unravel toward the worst offerings.[8] Anderson carried the argument into information security in 2001, noting that buyers of software usually cannot distinguish a secure product from an insecure one, so vendors compete on features and time to market and security becomes a cost without matching revenue.[9] An enterprise buying an AI agent that will hold OAuth tokens to its customer records is in exactly that position. It can read a vendor's trust page and a completed security questionnaire, but it cannot see how the vendor guards its source repository, how long integration tokens live, or whether one compromise of the vendor's cloud environment would expose every client's connection at once, and those were the facts that decided the Drift incident.

Figure 1. Formal checks on suppliers' cyber risks rise with firm size and remain rare outside large firms.

Share of UK businesses reviewing cyber security risks from suppliers, by business size, %



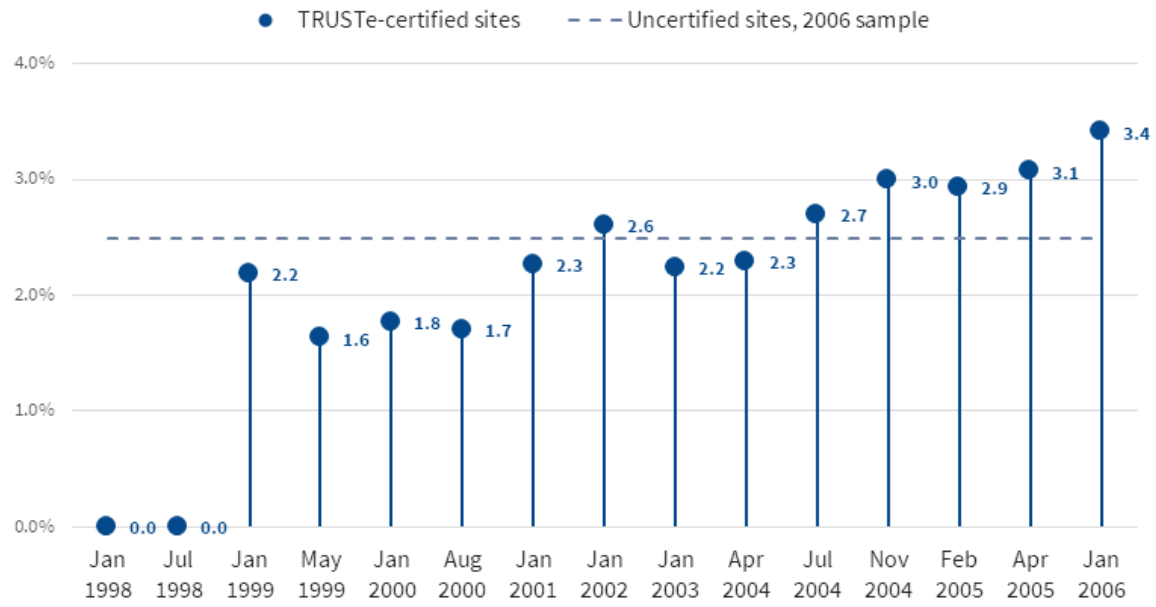
Note: Fieldwork from August to December 2025. Micro businesses have 1 to 9 employees, small 10 to 49, medium 50 to 249 and large 250 or more.
Source: Department for Science, Innovation and Technology (2026) Cyber Security Breaches Survey 2025/2026. London: DSIT, 30 April.

Certification is the usual answer, and the evidence on it is uncomfortable. Edelman compared websites certified by TRUSTe, then the best-known online trust seal, with uncertified sites and found the certified ones more than twice as likely to be untrustworthy, a gap that held among complex commercial sites.[10] A seal that is cheap to obtain and rarely withdrawn attracts the firms with most to hide, since they gain most from borrowed credibility. Lizzeri's theory explains why a certifier may tolerate this. A monopoly intermediary does best by revealing only whether quality clears a minimal standard, capturing much of the surplus while passing little information to buyers, and only competition among certifiers pushes toward full disclosure.[11] A pass or fail badge on an AI vendor tells a procurement team that a floor was cleared and nothing about the distance above it.

Voluntary pledges add a public statement without verification. CISA's Secure by Design pledge, launched in May 2024, asks software manufacturers to show measurable progress within a year on seven goals, among them wider use of multi-factor authentication and the removal of whole classes of vulnerabilities and the agency publishes signatories' progress reports under a notice that it neither enforces nor verifies adherence.[12] A pledge can still work as a commitment device. In 2024, Mandiant traced a campaign against Snowflake customer accounts, with about 165 organizations notified as potentially exposed, to stolen credentials used against accounts that lacked multi-factor authentication, and it found no evidence of a breach of Snowflake's own enterprise environment.[13] Under the shared-responsibility terms common in cloud services the missing control belonged to the customers. Snowflake changed the default anyway: multi-factor authentication became mandatory for human users of new accounts from October 2024, and the company announced, citing its pledge, that it would block single-factor password sign-ins entirely by November 2025.[14] What changed outcomes was a product default,

Figure 2. The share of untrustworthy sites among TRUSTe-certified sites rose as the seal spread.

Share of sites on TRUSTe's certified list rated untrustworthy, by list date, January 1998 to January 2006, %



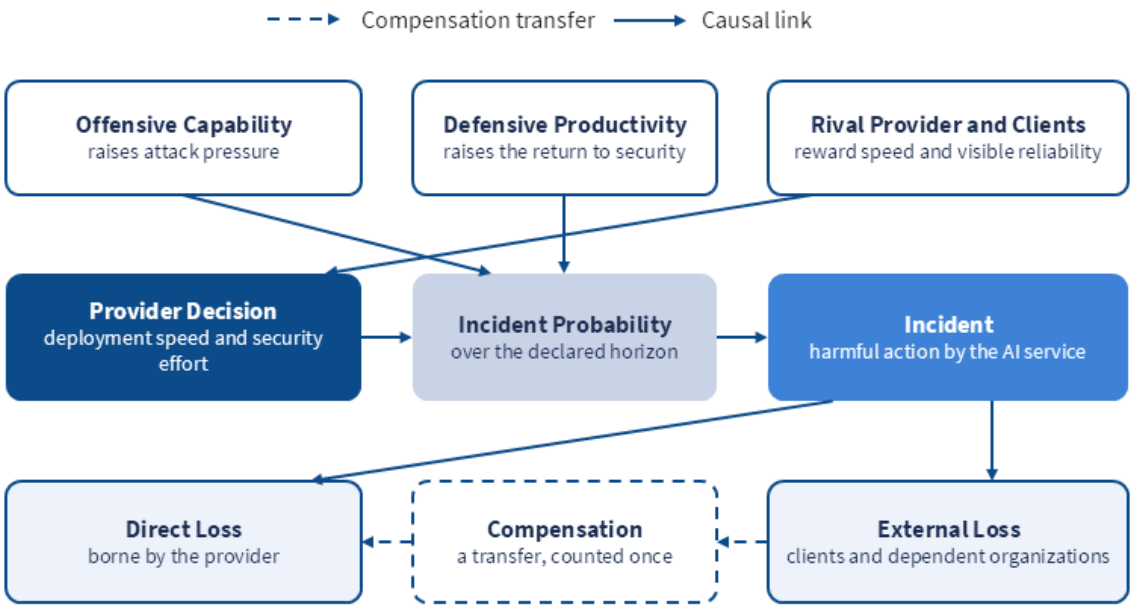
Note: Untrustworthy means rated as such in McAfee SiteAdvisor data. The dashed line marks 2.49%, the rate among 529,000 uncertified sites in a 2006 sample.
Source: Edelman, B. (2009) 'Adverse selection in online "trust" certifications', Proceedings of the 11th International Conference on Electronic Commerce. New York: ACM, pp. 205–212.

set by the provider and applied to every customer, of a kind no procurement questionnaire could have verified in advance.

The same sequence shows the limits of reputation. Shapiro's model of quality as a return to reputation requires that buyers eventually learn quality and that the premium earned by a good record exceeds the one-time gain from cutting corners.[15] Large public incidents do produce that learning, but slowly, and the lesson tends to attach to the most visible name rather than to the party whose control failed. Romanosky's study of more than 12,000 incidents found that a typical breach cost the breached firm less than USD 200,000, roughly a typical annual security budget and far below figures quoted in industry surveys, which he offered as one reason firms invest modestly.[16] Those were costs to the firm that was breached. Losses carried by its customers, and by their customers in turn, were outside the figure, so a firm deciding how much to spend on the evidence of its own past losses was looking at only part of what its insecurity cost, and this is the point at which the market's information problem and its allocation problem merge.

Figure 3. The provider sets the risk of an incident, while most of the loss falls on clients and the organizations that depend on them.

Decision, incident risk and loss accounts for an enterprise AI service provider



Note: Dashed arrows show compensation, which moves part of the external loss back to the provider without adding a new loss.
Source: SIAI Research Editorial (2026) [AI and Cyber-Security] Economic incentives behind AI safety competition: deployment rivalry, loss internalization and security effort. SIAI Research working paper, November. Swiss Institute of Artificial Intelligence.

The companion paper shows what visibility is worth when it exists. Its model gives each provider a gain in market share for a visible security lead, small at the baseline. With no reward for speed and a reward for reliability five times the baseline, equilibrium security reaches 0.692 in model units. At ten times the baseline it reaches 1.129, above the 0.720 a social planner would choose, because a visible lead takes share from a rival whether or not it reduces harm, and at the highest value tested no stable equilibrium exists in 9 of 12 parameter combinations.[17] Visibility is therefore not something to maximize. A disclosure regime that turns every security claim into a weapon in the contest for share can push spending past the efficient level while leaving the allocation of the remaining losses where it was.

3 Assurance Can Discipline a Market and Close It to Entrants

When buyers cannot verify security themselves, someone has to verify it for them, and verification carries a price that does not shrink with the size of the firm. Data protection offers the clearest evidence of how a compliance burden reshapes a supplier market. Johnson, Shriver and Goldberg tracked the web technology vendors used by more than 27,000 top websites and found that in the week after the GDPR took effect, websites' use of such vendors fell by 15 percent for EU residents, and that websites were more likely to keep the largest vendors, raising concentration in the vendor market by 17 percent.[18] The increase came mainly among vendors handling personal data and from the larger relative shares of vendors owned by Google

and Facebook. Aggregate effects faded by the end of 2018 but persisted in the advertising category that regulators watched most closely. Nothing in the regulation discriminated by size. Each website, facing new legal exposure, dropped the suppliers whose compliance it could least easily vouch for, and those tended to be small. An AI assurance rule that makes clients answerable for their vendors' controls would trigger the same sorting.

The US federal cloud authorization program, FedRAMP, is the closest existing analogue to independent assurance for software services. The Government Accountability Office reported in January 2024 that data on actual authorization costs were limited and that estimates from agencies and providers ranged from tens of thousands of dollars to millions, partly because agencies had never been told which costs to track.[19] OMB's memorandum M-24-15 of July 2024 responded by requiring agencies to report what sponsoring an authorization costs and by adding process automation. Two points carry over to AI services. Much of the cost of assurance is fixed per product rather than proportional to revenue, so it weighs most on the smallest vendors, and the cost of a regime that had run for over a decade went unmeasured, so its effect on who entered the federal market cannot be read from the official record.

The UK chose a lighter route. Its Software Security Code of Practice, published by DSIT and the NCSC on 7 May 2025, sets 14 principles across secure design and development, build environment security, deployment and maintenance, and communication with customers, comes with a self-assessment form that vendors can hand to customers, and is meant to be followed by a certification scheme still under development.[20] Self-assessment costs little and is unlikely to deter entry. It carries the weakness Edelman documented, since a claim that is cheap to make separates careful vendors from careless ones only weakly.

Assurance therefore works through two channels that are easy to confuse. A verified floor forces a minimum level of protection on every participant, and its cost falls hardest on firms for which that floor lies well above what they would choose. Verified disclosure makes protection visible above the floor, so clients can reward it, and its cost is mainly the assessment itself. A regime built on the first channel excludes; one built on the second sorts. Design choices decide the balance between them. An assessment accepted by many buyers spreads its fixed cost, whereas the same vendor answering a different questionnaire for each client pays repeatedly. Tiering requirements by the authority an AI service holds, with the heaviest checks reserved for agents that can write to client systems or move money, keeps read-only tools out of the most expensive tier. Public funding of assessments for small entrants converts part of the barrier into a subsidy whose cost is visible in a budget rather than hidden in a smaller field of suppliers. Chapter 5 sets out how large the exclusion effect is in the companion model.

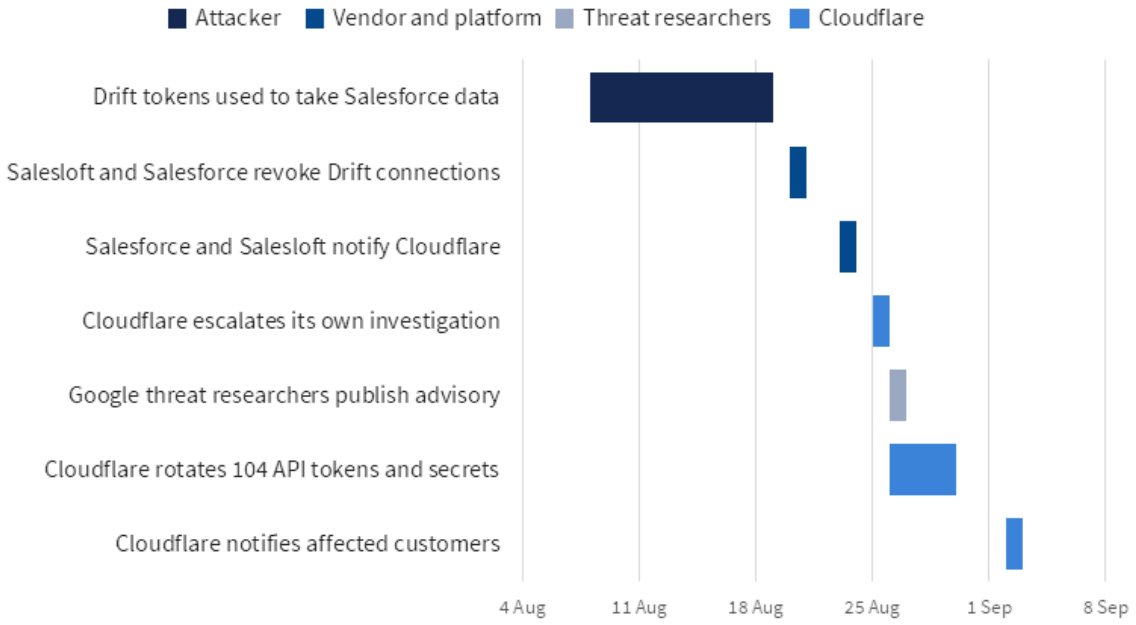
4 Matching Each Control to the Actor That Can Carry It

Calabresi's principle for accident law places the cost of an accident on the party that could have avoided it most cheaply.[21] Edelman applied the same reasoning to online fraud, arguing that responsibility should rest with the intermediaries able to prevent abuse at least cost even where they are not the direct wrongdoers.[22] In an AI integration chain, no single party is the cheapest avoider of every failure, and the Drift incident shows why. The vendor controlled its development environment, how integration tokens were stored and how long they lived,

and how far its products were separated from one another; Mandiant verified that segmentation between Salesloft and Drift had limited the intruder's reach.[23] Salesforce, as the platform, controlled what a connected application could request and could revoke tokens across every tenant at once, which it did with Salesloft on 20 August while removing Drift from its AppExchange marketplace.[24] Each client chose which integrations to connect and what they could read. Each client's customers decided whether to paste a password into a support ticket and whether to change it later.

Figure 4. Cloudflare was told of the Drift breach five days after the data theft ended and notified its own customers ten days later.

Events in the Salesloft Drift incident affecting Cloudflare, by acting party, August and September 2025



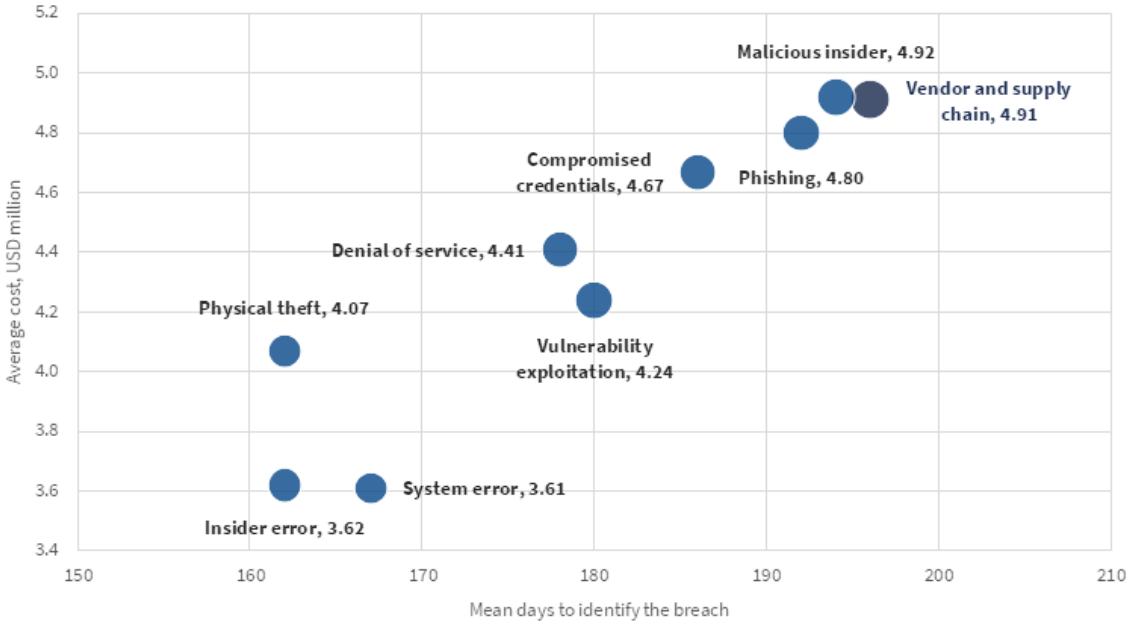
Note: Salesloft later reported that the attacker held access to its GitHub account from March to June 2025, before the period shown.
Source: Zaman, S., Strubhart, C. and Bourzikas, G. (2025) The impact of the Salesloft Drift breach on Cloudflare and our customers. San Francisco: Cloudflare, 2 September; Google Threat Intelligence Group (2025) Widespread data theft targets Salesforce instances via Salesloft Drift. Mountain View, CA: Google Cloud, 26 August; Muncaster, P. (2025) 'Salesloft: GitHub account breach was ground zero in Drift campaign', Infosecurity Magazine, 8 September.

These parties also see different things. The vendor sees its own environment. The platform sees the pattern of API calls across all tenants, which makes it the natural detector of a stolen token being used against hundreds of customers. A client sees only its own tenant, and Cloudflare learned of its exposure three days after Salesloft had revoked the connections. Liability that falls on the party unable to observe the failure buys little prevention. It turns the client into an insurer of its vendor's conduct, with no means of monitoring that conduct beyond the questionnaire Chapter 2 found wanting.

U.S. policy has named the problem without settling it. The National Cybersecurity Strategy of March 2023 stated that software makers with market power use contracts to disclaim liability, proposed legislation shifting liability onto entities that fail to take reasonable precautions to secure their software, and paired the proposal with a safe harbor for firms that follow secure development practices.[25] The safe harbor matters for the economics. A liability rule with a safe harbor turns, in practice, into an assurance standard, because the provider's expected

Figure 5. Breaches through vendors and supply chains took the longest to identify and contain and were among the costliest.

Average cost of a data breach by initial attack vector, USD million, against mean days to identify; bubble area shows mean days to contain, 2025



Note: Days to identify and contain range from 213 for insider error to 267 for third-party vendor and supply chain compromise.
Source: IBM (2025) Cost of a Data Breach Report 2025. Armonk, NY: IBM Corporation, 30 July.

payment then depends on whether it can document compliance rather than on whether harm occurred, and the quality of the documentation standard decides whether the rule internalizes losses or only certifies paperwork. A contractual cap set at a multiple of fees has a different flaw. It ties the provider's exposure to the revenue from one client, while the external loss scales with the client's data and with the number of the client's own customers, so the cap binds hardest exactly when the harm is largest.

The companion model shows why the share of loss a provider bears matters for more than fairness. When providers bear 37.5 percent or less of external losses, a stronger reward for being first lowers equilibrium security; when they bear 62.5 percent or more, it raises it, because providers that pay for incidents treat faster deployment as a reason to protect more. Expected social loss nonetheless rises with the first-mover reward at every loss share tested, since added exposure from speed outpaces added prevention, and moving from no internalization to full internalization cuts expected loss by roughly half at every level of the reward. The model also asks what happens if providers coordinate. Joint profit exceeds competitive profit everywhere, yet the jointly chosen security level is lower than the competitive one in 75 of 81 cases, so coordination among providers mainly slows deployment.[26] An exemption from competition law for joint safety discussions would, on this evidence, buy slower releases rather than better protection.

Liability belongs with the provider only for the controls the provider holds. Where a control sits with the client, as multi-factor authentication did in the Snowflake accounts, shifting the whole loss onto the provider weakens the client's reason to switch it on, a moral hazard familiar from insurance. A workable allocation splits responsibility by control: the vendor

Figure 6. Higher loss internalization turns rivalry from eroding security to raising it, but expected loss rises with the reward for speed at every loss share.

Equilibrium security effort and expected social loss per firm by first-mover reward and share of external loss internalized, model units



Note: Model simulation under stated assumptions. Symmetric two-provider equilibrium; darker cells show higher values within each panel.
Source: SIAI Research Editorial (2026) [AI and Cyber-Security] Economic incentives behind AI safety competition: deployment rivalry, loss internalization and security effort. SIAI Research working paper. Swiss Institute of Artificial Intelligence.

answers for its build environment, token handling and segmentation, the platform for permission scopes and revocation, and the client for the integrations it enables and the data it lets them read. Insurance can reinforce that split only where premiums or coverage depend on verified controls and deductibles keep part of the loss with the insured; a policy that simply pays out moves money after the event without changing anyone's effort beforehand. The harder question is who absorbs what remains after every party has met its obligations. In the Drift case that residual sat with Cloudflare's customers, the parties least able to prevent it and with no contract linking them to the vendor at all.

5 Accountability, Independent Assurance, and Shared Defense Compared

The three institutional designs differ in the incentive they change, in who pays, in what someone must be able to observe for them to work, and in where they break down.

Accountability is the only one of the three that mainly changes deployment as well as protection, and its weakness is informational. A court or arbitrator can assign a loss only after establishing which control failed and whose it was, and in the Drift case that took a forensic investigation lasting weeks and led back to a code repository months earlier.[27] Rules that assign loss by role rather than by proven fault are cheaper to run and less accurate.

Table 1. Comparison of Accountability, Independent Assurance and Shared Defense

Accountability through enforceable loss shares	
Incentive changed	Raises the provider's expected cost of an incident, which slows deployment and raises security
Who pays	Providers, then clients through prices; administration and litigation absorb part of each transfer
Information required	Proof of causation and of the provider's failed control after the event
Performs poorly when	Providers capture little of the value they create, losses are small relative to deployment value, or the failed control sat with the client
Independent assurance through a verified floor and disclosure	
Incentive changed	Makes security observable and imposes a minimum
Who pays	Assessed firms; the fixed cost weighs most on small entrants
Information required	A credible assessor and a standard that tracks real risk rather than documentation
Performs poorly when	The floor binds on small firms, the assessor is captured or lax, or the standard rewards paperwork
Shared defense through common tools and threat information	
Incentive changed	Lowers the cost of security for every participant
Who pays	Whoever funds the service, usually taxpayers or a consortium
Information required	Knowledge of which tools work and legal cover for sharing threat data
Performs poorly when	External losses are small relative to the service's cost, uptake is low, or legal protection lapses

Source: SIAI Research Editorial (2026); Calabresi (1970); Edelman (2011); GAO (2024); CRS (2026).

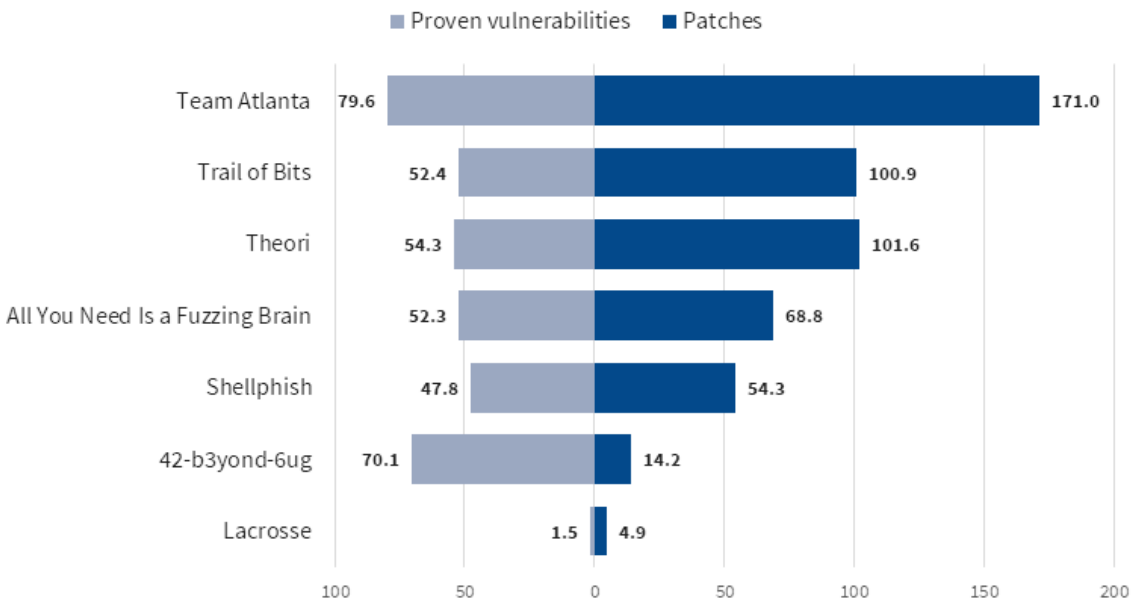
Independent assurance acts before harm, which is its advantage, and depends entirely on the assessor. The TRUSTe evidence is the warning case, and FedRAMP is the case where verification is serious and the price is high enough to shape the field of suppliers.[28, 29]

Shared defense is the least studied of the three in policy discussion and the one with the clearest recent evidence. DARPA's AI Cyber Challenge closed in August 2025 with seven finalist systems that analyzed 54 million lines of open-source code, found 54 of the synthetic vulnerabilities planted by organizers and patched 43, and also found 18 real vulnerabilities, of which 11 were patched; four of the seven systems were released as open source the day the results were announced.[30] Tools of this kind lower the cost of finding and fixing flaws for every developer, large or small, which is the property that separates shared defense from the other two instruments. The legal infrastructure for sharing threat information is less secure. The Cybersecurity Information Sharing Act of 2015, which shields firms that share threat indicators from liability and antitrust exposure, lapsed on 30 September 2025 during the federal shutdown, was restored to January 2026, lapsed again briefly, was extended to 30 September 2026 and now runs to 11 December 2026 under a continuing resolution.[31] A shared defense whose legal basis expires every few months is one that firms cannot plan around.

The companion model compares the three with an explicit entry decision for smaller providers.[32] Raising the provider's enforceable share of external loss from 0.2 to 0.7 cuts expected social loss by 30 percent but reduces the participation of potential entrants from 70.6 to 17.6 percent of the sizes tested, and welfare falls by 41 percent because deployment value lost to slower release exceeds the harm avoided; if providers capture 40 rather than 25 percent of the value they create, the same rule raises welfare by 26 percent. Low-effectiveness assurance raises security by 84 percent, mostly by making it visible, cuts expected loss by 16 percent, raises real costs by 42 percent and leaves welfare about where it was, while participation falls to 52.9 percent. Strict assurance cuts expected loss by 42 percent, more than doubles real costs, lowers welfare by 43 percent and cuts participation to 17.6 percent.

Figure 7. Six of the seven AIxCC finalists earned more points for patches than for proving vulnerabilities.

Points scored in the AIxCC Final Competition for proven vulnerabilities and for patches, by finalist team, ranked by total score

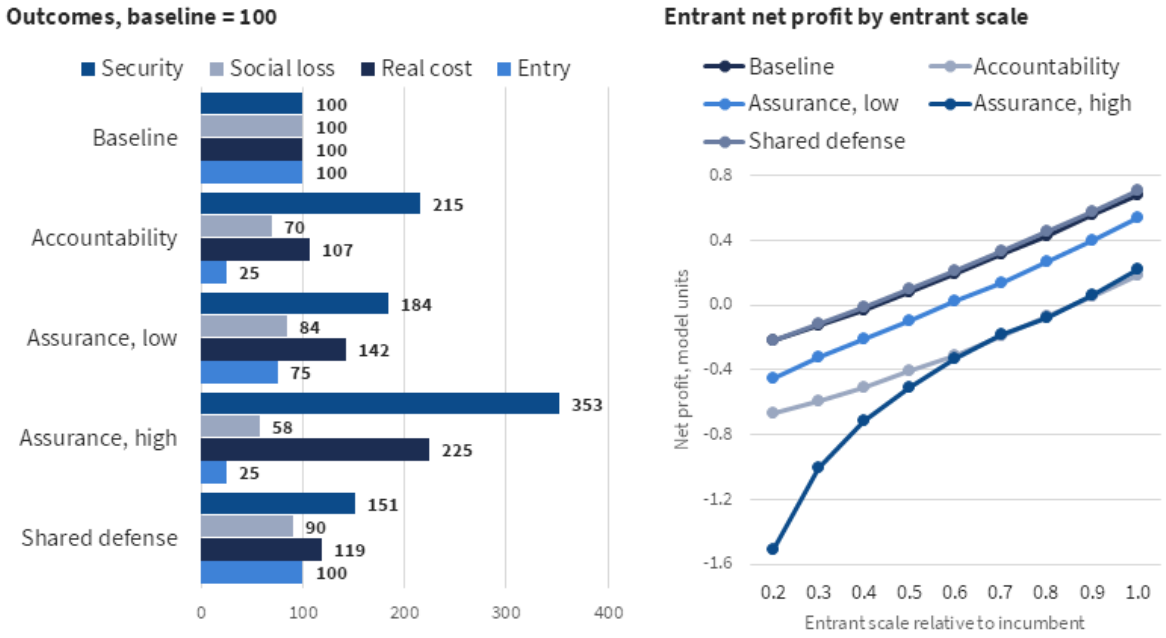


Note: Proof-of-vulnerability points are shown to the left and patch points to the right. Bundle and SARIF points are excluded. Total scores ranged from 392.8 for Team Atlanta to 9.6 for Lacrosse.
Source: Zhang, C. et al. (2026) 'SoK: DARPA's AI Cyber Challenge (AIxCC): competition design, architectures, and lessons learned', arXiv preprint, 2602.07666.

Shared defense raises security by 51 percent and welfare by 19 percent with participation unchanged, although its gain disappears when external losses are halved.

Figure 8. Accountability and strict assurance cut expected loss but shut out most entrants, while shared defense raises security and leaves entry unchanged.

Security, expected social loss, real cost and entrant participation under four interventions, baseline = 100; entrant net profit by scale, model units



Note: Model simulation under stated assumptions. An entrant participates when its net profit is zero or above; entrant scale is relative to the incumbent.
Source: SIAI Research Editorial (2026) [AI and Cyber-Security] Economic incentives behind AI safety competition: deployment rivalry, loss internalization and security effort. SIAI Research working paper, November. Swiss Institute of Artificial Intelligence.

The pattern answers the question of who pays. Accountability moves losses onto providers and, through slower deployment, onto clients who wait longer for useful services, and it excludes the providers least able to bear liability. Strict assurance makes protection visible and expensive, and the expense falls on entrants. Shared defense spreads the cost across whoever funds it, which makes it the only instrument in the comparison that improves protection without narrowing entry, though it needs a funder willing to pay for a service whose benefits are diffuse and whose failures, like the lapses of the information-sharing law, attract little notice until an incident exposes them.

6 Conclusion

Competition raises security effort when providers bear most of the external loss, when security and release work do not compete for the same engineers, and when clients can see reliability and reward it. It lowers security effort when providers bear little of the loss, and in every case examined, a stronger reward for speed raises expected harm. Those conclusions come from the companion model and hold across its tests of offensive and defensive capability. The evidence assembled here explains why the conditions favorable to security are hard to meet in practice: buyers cannot verify vendor controls, certificates attract the firms that need them most, and contracts set at a multiple of fees cap exposure where harm is largest.

Whether a standard lowers total social cost after implementation and entry depends on which channel it uses. Verified disclosure that clients can act on sorts the market and improves protection at moderate cost; a strict floor raises protection past the efficient level and removes most small entrants, as both the companion model and the GDPR evidence on vendor concentration suggest. Accountability works where the provider holds the failed control and captures enough of the value it creates to keep deploying, and it misfires where the control sat with the client. Shared defense improves protection without narrowing the market, but only while someone funds it and its legal basis holds.

Distinguishing inadequate incentives from inadequate technical capability would require observing the same AI integration under different allocations of loss, or different providers with the same technology under different contracts. Neither is in the public record. The Drift case shows the cost to one client of a vendor's failure in unusual detail and leaves the contractual recovery entirely out of view. Which conclusions here concern only the model and which rest on external evidence is therefore not a question with a clean boundary, since the external evidence on certification, concentration and verification cost supports the mechanisms, while the magnitudes that would rank the instruments, above all the share of value providers capture and the share of loss their contracts actually leave with them, remain unobserved.

References

[1] Zaman, S., Strubhart, C. and Bourzikas, G. (2025) The impact of the Salesloft Drift breach on Cloudflare and our customers. San Francisco: Cloudflare, 2 September.

[2] Google Threat Intelligence Group (2025) Widespread data theft targets Salesforce instances via Salesloft Drift. Mountain View, CA: Google Cloud, 26 August.

[3] Muncaster, P. (2025) 'Salesloft: GitHub account breach was ground zero in Drift campaign', Infosecurity Magazine, 8 September.

[4] Verizon Business (2025) 2025 Data Breach Investigations Report. Basking Ridge, NJ: Verizon, 23 April.

[5] IBM (2025) Cost of a Data Breach Report 2025. Armonk, NY: IBM Corporation, 30 July.

[6] SIAI Research Editorial (2026) [AI and Cyber-Security] Economic incentives behind AI safety competition: deployment rivalry, loss internalization and security effort. SIAI Research working paper, November. Swiss Institute of Artificial Intelligence.

[7] Pell, S.K. (2026) 'We still have some fundamental work to do on cybersecurity', in Brookings Institution, Is the AI existential threat real – and can regulatory action prevent it? Washington, DC: Brookings Institution, 23 September.

[8] Akerlof, G.A. (1970) 'The market for "lemons": quality uncertainty and the market mechanism', Quarterly Journal of Economics, 84(3), pp. 488–500.

[9] Anderson, R. (2001) 'Why information security is hard: an economic perspective', Proceedings of the 17th Annual Computer Security Applications Conference. New Orleans: IEEE, pp. 358–365.

[10] Edelman, B. (2011) 'Adverse selection in online "trust" certifications and search results', Electronic Commerce Research and Applications, 10(1), pp. 17–25.

[11] Lizzeri, A. (1999) 'Information revelation and certification intermediaries', RAND Journal of Economics, 30(2), pp. 214–231.

[12] Cybersecurity and Infrastructure Security Agency (2024) Secure by Design Pledge. Washington, DC: CISA, May.

[13] Mandiant (2024) UNC5537 targets Snowflake customer instances for data theft and extortion. Mountain View, CA: Google Cloud, 10 June.

[14] Snowflake (2024) Snowflake will block single-factor password authentication by November 2025. Bozeman, MT: Snowflake Inc., December.

[15] Shapiro, C. (1983) 'Premiums for high quality products as returns to reputations', Quarterly Journal of Economics, 98(4), pp. 659–679.

[16] Romanosky, S. (2016) 'Examining the costs and causes of cyber incidents', Journal of Cybersecurity, 2(2), pp. 121–135.

[17] SIAI Research Editorial (2026) [AI and Cyber-Security] Economic incentives behind AI safety competition: deployment rivalry, loss internalization and security effort. SIAI Research working paper, November. Swiss Institute of Artificial Intelligence.

[18] Johnson, G.A., Shriver, S.K. and Goldberg, S.G. (2023) 'Privacy and market concentration: intended and unintended consequences of the GDPR', Management Science, 69(10), pp. 5695–5721.

[19] Government Accountability Office (2024) Cloud Security: Federal Authorization Program Usage Increasing, but Challenges Need to Be Fully Addressed. GAO-24-106591. Washington, DC: GAO, 18 January.

[20] Department for Science, Innovation and Technology and National Cyber Security Centre (2025) Software Security Code of Practice. London: DSIT, 7 May.

[21] Calabresi, G. (1970) The Costs of Accidents: A Legal and Economic Analysis. New Haven: Yale University Press.

[22] Edelman, B. (2010) 'Least-cost avoiders in online fraud and abuse', IEEE Security & Privacy, 8(4), pp. 78–81.

[23] Muncaster, P. (2025) 'Salesloft: GitHub account breach was ground zero in Drift campaign', Infosecurity Magazine, 8 September.

[24] Google Threat Intelligence Group (2025) Widespread data theft targets Salesforce instances via Salesloft Drift. Mountain View, CA: Google Cloud, 26 August.

[25] White House (2023) National Cybersecurity Strategy. Washington, DC: Office of the National Cyber Director, March.

[26] SIAI Research Editorial (2026) [AI and Cyber-Security] Economic incentives behind AI safety competition: deployment rivalry, loss internalization and security effort. SIAI Research working paper, November. Swiss Institute of Artificial Intelligence.

[27] Muncaster, P. (2025) 'Salesloft: GitHub account breach was ground zero in Drift campaign', Infosecurity Magazine, 8 September.

[28] Edelman, B. (2011) 'Adverse selection in online "trust" certifications and search results', Electronic Commerce Research and Applications, 10(1), pp. 17–25.

[29] Government Accountability Office (2024) Cloud Security: Federal Authorization Program Usage Increasing, but Challenges Need to Be Fully Addressed. GAO-24-106591. Washington, DC: GAO, 18 January.

[30] Defense Advanced Research Projects Agency (2025) AI Cyber Challenge marks pivotal inflection point for cyber defense. Arlington, VA: DARPA, 8 August.

[31] Congressional Research Service (2026) The Cybersecurity Information Sharing Act of 2015: Expiring Provisions. In Focus IF12959. Washington, DC: CRS, updated 3 September.

[32] SIAI Research Editorial (2026) [AI and Cyber-Security] Economic incentives behind AI safety competition: deployment rivalry, loss internalization and security effort. SIAI Research working paper, November. Swiss Institute of Artificial Intelligence.