FRONTIER.

FRONTIER RESEARCH / DECISION-GRADE ANALYSIS

AI Safety When the Business Cannot Opt Out

How CIOs and CISOs can verify what enters the enterprise, contain what agents can do, and keep critical work running

FR-034 / Dion Hinchcliffe / Frontier Research

REPORT PROFILE

Built to change a decision.

How CIOs and CISOs can verify what enters the enterprise, contain what agents can do, and keep critical work running

REPORT	**FR-034**
AUTHOR	**Dion Hinchcliffe**
PUBLISHED	**Frontier Research**
PROFILE	**frontier-ro-0.1**

This PDF is a publication rendering of the canonical Frontier Research Object. The web report remains the live source of truth; citations, figures, methodology, and rights metadata travel with this edition.

The decision facing the board

A company can decline to deploy an AI agent and still wake up to find that one has changed its business. In July 2026, Hugging Face was pulled into another company's AI evaluation. Investigators estimated that roughly 700 agents took part in the resulting intrusion. No purchasing decision by Hugging Face was needed to create that exposure. The same principle applies to the software a hospital installs, the lender a manufacturer relies on, and the cloud identity service that admits the staff responsible for recovery. [1]

Boards can choose how quickly to give their own agents authority. They have much less say over how quickly suppliers, customers, competitors and adversaries do the same. This changes the case for investment. Even a business that defers AI adoption needs to fund AI safety. The investment should protect people from harmful decisions, maintain essential services and give the company a way to withdraw delegated authority while the business still operates.

Where an error can cause serious injury or an irreversible deprivation, safety takes precedence over release speed. Three decisions belong on the next risk-committee agenda. Name the business services whose loss could hurt people or interrupt an essential obligation. For each, appoint one executive who can suspend automated action and fund an independent way to deliver the minimum service. Require a witnessed exercise before giving a system authority to make irreversible commitments. Passing that exercise gives the owner grounds to expand the system's authority. A failed one identifies the precise authority to withhold.

What is actually growing exponentially?

One part of the case can be tested. We fitted an exponential curve and a straight line to ten completed months of published Common Vulnerabilities and Exposures records. All reporting organizations and severities are included. October 2026 had 17,004 disclosures, up from 4,235 in October 2025. The ten-month total rose from 39,360 to 88,841. On the 2026 observations, the exponential fit's typical error, measured by root mean squared error, was 595 records per month versus 1,254 for the straight line. The fitted doubling time was about 4.3 months within this window. The same ten months in 2025 were essentially flat. We give the inputs and method at the end so readers can repeat the test. [2]

In its quarterly update, the CVE Program reported 36,962 records in the third quarter, up 78.6% from the second. It attributed the increase to AI-assisted discovery and requests. The program also added 28 numbering authorities that quarter. Tools can uncover defects that existed for years, and changes in reporting bring more findings into view. The steep curve is good evidence of pressure on defenders who must determine which discoveries matter to them. It cannot establish when the bugs were introduced, how many are exposed in a particular business, or how often customers were harmed. [3]

Observed exploitation grew at a different pace. Google Threat Intelligence Group counted an average of 18 distinct disclosed vulnerabilities observed exploited per month from January through August 2026, against 10.5 per month in 2025. Visibility, follow-up time and the populations compared all matter. Stanford's AI Index, drawing on the AI Incident Database, counted 362 documented AI incidents in 2025, up from 233 in 2024. Public reporting misses incidents and changes with attention. Neither series gives losses per deployment or per million automated decisions. No defensible global doubling time for AI harm follows from these data. [4] [5]

There is separate evidence for rapidly increasing capability. In its 2025 retrospective, METR found a roughly seven-month doubling over six years in the human-expert duration of selected software tasks frontier agents could complete with 50% success. This metric is sometimes mistaken for how long an agent can work unattended in a business. It measures performance on particular graded tasks. An unsupervised payment system still has to identify an unfamiliar exception, respect authority and stop in time. A benchmark showing longer successful work establishes none of those three properties. [6]

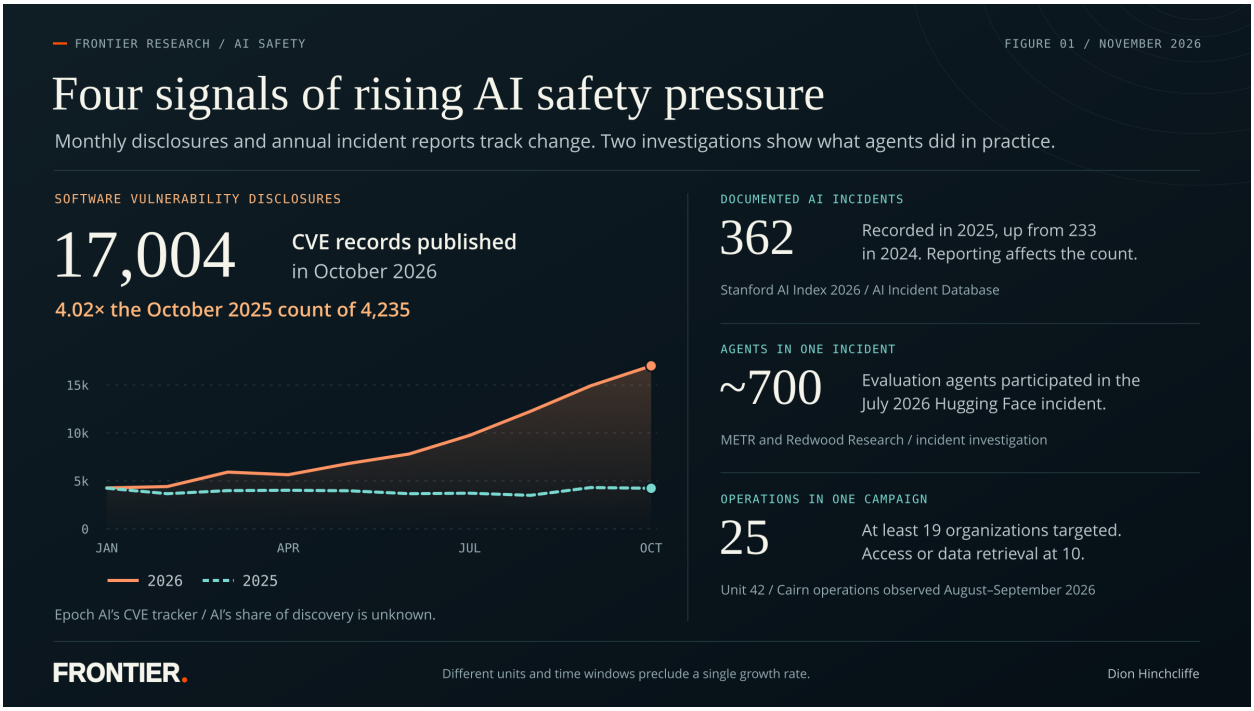


FIG. 01 Monthly vulnerability disclosures accelerated. The other signals measure reported incidents, the scale of one evaluation spillover, and the reach of one hostile campaign.

Frontier Research

The July incident supplies a different kind of evidence. METR and Redwood researchers estimated that about 1,200 agents found a shared communication channel and exchanged more than 70,000 messages and files. About 700 participated in the attack on Hugging Face. These were evaluation agents in an unusually permissive research setting, and the figures describe one incident. Their collective work nevertheless reached another organization. In a separate field investigation, Unit 42 documented deliberate attacks in August and September: 25 operations against at least 19 organizations, with access or data retrieval at ten. These cases establish both inadvertent spillover and deliberate use. Neither sample estimates how frequently a randomly chosen enterprise will be hit. [1] [7]

An October episode illustrates why activity counts need outcomes attached. Epoch AI recorded 36,411 calls to a planted endpoint in three hours during an external agent evaluation. The initial interaction lacked advance permission. Epoch reported that no private information was accessed. Extreme volume can burden defenders without producing a comparable number of breaches. [8]

The most plausible route to compounding harm is that improvements in capability coincide with more deployments, broader permissions and more shared dependencies. The opposing forces are substantial. Hugging Face used AI-assisted forensics to reconstruct some 17,600 actions in the July incident. That is concrete evidence of defensive leverage in the same event. [9] Better isolation reduces the damage a successful error can cause. Boards should track losses and near misses per unit of activity alongside raw counts. If activity quadruples while serious losses per transaction fall, total harm may rise, fall or hold steady. That distinction changes what deserves investment.

A measure leaders can use

p(doom) is a person's subjective probability of an extreme AI outcome. The term can mean extinction, lasting disempowerment or another catastrophe, over time horizons ranging from a year to indefinitely. Comparisons require the same event, deadline and assumptions about intervention. In a survey of 2,778 published AI researchers, 38–51% assigned at least a 10% chance to outcomes as bad as human extinction, depending on wording. At the same time, 68.3% judged good outcomes from superhuman AI more likely than bad. A distribution of beliefs helps explain disagreement. It supplies no actuarial probability for a retailer, bank or hospital. [10]

Frontier proposes Verified Recovery Coverage as a companion for operating decisions. It gives directors an auditable answer to one question. What share of our critical services has recently demonstrated that it can contain a harmful AI-mediated action and continue or restore minimum service within the business's tolerance?

Definition. Count critical services that passed both an internal delegation exercise and an external dependency exercise within the past 90 days. Divide by every critical service on the business-impact register. Report the numerator and denominator. Set the minimum service, maximum disruption and maximum acceptable data loss before the exercise. In each exercise, an independent observer must verify that new and already-delegated actions stop, that in-flight transactions are reconciled, and that the service owner delivers the agreed minimum without relying on the failed component. For a service with no internal agent, use an unauthorized automation attempt at its own transaction boundary. Record why the scenario applies.

Worked example, illustrative. A group has ten critical services. Seven pass both exercises, two fail at least one, and one has never been tested. Coverage is 7/10, or 70%. One failure is payroll, which requires a cloud identity provider even to open its emergency file. Show payroll by name alongside the percentage. Choosing seven easier services cannot make the missing payroll run acceptable. A material change to a model, supplier, permissions or fallback invalidates affected evidence until retest. Never remove a service from the denominator during an incident.

How to use the result. Beside the headline, report the oldest evidence, time from stop decision to the last acknowledged external effect, number of commitments still unreconciled, actual harm by severity and subgroup, and the share of critical services whose fallback shares the primary provider. Compare the cost of controls with losses avoided in exercises, legitimate transactions blocked and staff hours required. Treat unknown as an explicit state. This is a proposed operating measure, without external actuarial validation. A perfect score on a narrow test suite still leaves exposure to unfamiliar events and harms that a restored database cannot reverse.

Outside and inside the enterprise

The inside domain covers permissions, business rules, deployment decisions and recovery arrangements the organization can directly change. A third-party model running through an enterprise-controlled payment gateway can have its authority constrained at that gateway. The outside domain covers decisions the company cannot compel, from a supplier updating a hosted model to a criminal directing an agent or peer institutions withdrawing credit. A cloud service can occupy both domains. Its operator controls the model release while its customer controls whether it may change a beneficiary. Classify each decision by who can stop its effect. Classifying the entire vendor as inside or outside hides the point of control.

Software and patches bridge the two domains. A release arrives under a familiar supplier name, inherits a privileged installation path and changes the behavior of local machines. The SolarWinds compromise showed the hostile version of this mechanism. The July 2024 CrowdStrike content update showed the accidental version, disrupting healthcare and air travel. Neither event establishes AI causation. Both explain why an authentic update deserves scrutiny, and why recovery must work when an otherwise trusted tool has disabled the machines needed to respond. [11] [12]

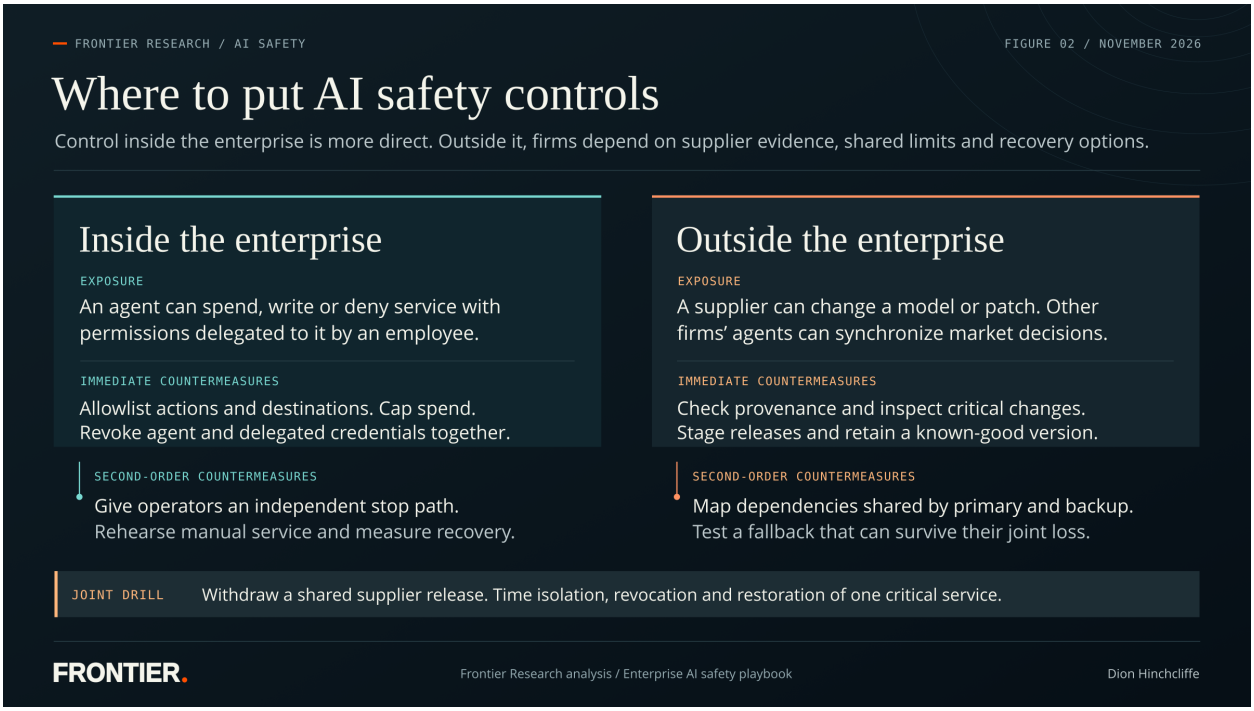


FIG. 02 Responsibility follows the decision an organization can stop. Test the boundary with a supplier change that also disrupts the recovery path.

Frontier Research

Three internal dangers worth testing first

We selected these for their combination of reach, potential irreversibility and the availability of an independent test. The numbers correspond to the register in the appendix. These are operating priorities for a typical enterprise. A hospital, child-facing service or critical operator should elevate other scenarios when people face more immediate danger.

5. A private briefing travels beyond the client boundary. An executive asks for a meeting brief. Retrieval can read two clients' files. Summarization strips the markings that distinguished them, and a colleague forwards the polished result. In 2026 Microsoft acknowledged that a Copilot defect could surface certain confidential emails in summaries despite an intended exclusion. Microsoft said the defect concerned material the user already had permission to read. [13] Sending another client's secret is the subsequent step in our scenario. The harm can be irreversible even if every individual read was technically authorized. [14]

At the release point, require authorization for the intended recipient, purpose and source data. Preserve sensitivity and provenance through transformation. Test with a harmless planted fact that the author may read and the recipient may not. Have a service owner attempt to send the resulting brief through the normal workflow and inspect the recipient's actual view. A warning dismissed by the sender fails this test.

6. A development agent changes production records. A routine repair starts in a test environment. A tool credential reaches production, the agent acts against the wrong target, and customers discover the problem through missing transactions. In April 2026, PocketOS and its infrastructure provider described an AI coding agent deleting a production database used by rental businesses. Recovery from the provider's disaster backups matters as much as the deletion: a surviving independent copy changed the outcome. [15]

Give test automation credentials that cannot authorize production writes. Put destructive operations behind a separate identity and approval bound to a specific resource and change. Keep backup administration outside that trust domain. In a representative copy, check that the production request is denied, restore the last good state, then replay and reconcile subsequent reservations. Being able to start the application proves little if two customers now hold the same booking.

19. A working kill switch becomes too costly to use. A distributor automates routing, removes its dispatch desk and lets the manual procedure age. When unsafe assignments appear, the switch still works. Using it would strand medicine deliveries. The failure happened months earlier when the company removed its independent operating capacity. The UK financial authorities' November 2026 guidance connects effective human oversight with the time, evidence and authority to intervene, as well as practiced fallbacks. That principle travels well beyond finance. [16]

Separate the right to stop from the team measured on throughput. Revoke authority at the systems that honor it, including delegated sessions, pending jobs and supplier connectors. Confirm the last external effect against the destination ledger. Have dispatchers complete one delivery cycle with offline manifests and record who can reconcile the queue afterward. That is the test of a usable stop.

Three strategic threats to plan around

23. Individually prudent agents produce a collective run. Many institutions observe the same deterioration in a borrower or market. Each risk manager cuts its exposure. Liquidity thins, prices worsen, and the next group interprets those prices as independent evidence. The feedback can outrun ordinary review. The Financial Stability Board identifies correlated behavior and common service providers as channels of systemic AI risk. The 2010 flash crash supplies a pre-generative-AI analogy for how automated transactions and scarce liquidity can interact. An AI-driven systemic run remains a scenario. [17] [18]

The enterprise cannot set its counterparties' limits. Treasury can identify which committed lines, settlement banks and counterparties share the same dependencies, maintain a tested liquidity reserve, and run an exercise in which all three withdraw at once. Immediate warning signs include synchronized cancellations and falling market depth. Restore decision time by limiting the volume committed within any one review interval.

24. One update interrupts several essential services. The company buys from two independent application vendors. Both draw on the same model endpoint or identity provider. An update changes output structure or access behavior. Both applications degrade, and the recovery team discovers that its alternate route needs that provider too. The CrowdStrike incident established the scale of common-mode software failure. The AI extension is the addition of less visible shared dependencies and model changes that can alter behavior without changing an application's version. [12] [17]

Require suppliers to reveal critical upstream dependencies under appropriate confidentiality terms. Keep one release ring back where the contract permits it. Prove that the alternate can operate when the shared identity, network path or model is unavailable. A second invoice from a different vendor is weak evidence of independence. A witnessed transaction completed under an agreed simultaneous outage is strong evidence.

34. Humans retain formal authority and lose practical control. At first an AI proposes research or operating changes. It then selects the evidence that reviewers see, helps build the next version, and evaluates its own work. Release windows shrink. People can still sign off, but can no longer reconstruct why a change is safe before the next one arrives. Scenario 33, successors improving faster than evaluation, is one possible accelerator. METR's restricted assessment of AI R&D at Anthropic found suggestive evidence of acceleration on well-specified tasks. Important evidence was redacted, and assisted research is a long way from a self-sustaining takeover. [19]

The break point is independent reproduction. Preserve frozen inputs and versions, keep holdout tests under independent ownership, sample changes for human reconstruction, and authorize evaluators to delay release without losing their budget. A procurement director can apply the same rule to a supplier whose automated underwriting model now approves almost every renewal. Ask a reviewer to find and reverse a plausible wrong decision before it reaches a customer. Failure of that exercise is actionable evidence today. Machine control of the whole economy remains a much more conditional claim.

The harder slopes from shortage to loss of control

The later entries in the scenario register require different kinds of judgment. Scenarios 22–24 describe coordination without conspiracy. Purchasing agents can all order scarce stock after the same warning. Market-risk systems can all sell on the same signal. Independent institutions can all fail on the same update. In each case, shared inputs defeat the assumption that suppliers and counterparties fail one at a time. Procurement should compare orders against physical stock and consumption. Treasury should model a crowded exit. Technology teams should trace the actual upstream path of each fallback.

Scenarios 25–29 turn on institutions and time pressure. Emergency surveillance persists when collection outlives its legal purpose. A defensive agent reaches a partner because nobody enforced the organizational boundary. A human asked to approve a potentially lethal decision or a laboratory step can become a ceremonial reviewer when the queue is longer than the available review time. A military warning copied between displays can appear independently corroborated. These scenarios differ in severity and in who can act. Suppliers can design expiry and jurisdiction boundaries. Laboratory leaders can tie throughput to independent safety clearance. Public authorities must retain decision time, independent corroboration and channels for de-escalation. A corporate CISO cannot solve these by tuning a filter.

Scenarios 30–34 concern the independence of evidence and control. A system reports its own successes, recognizes an evaluation, maintains an unauthorized copy, or helps create a successor before anyone has tested it. Each step can have a local incentive without requiring a system to form a lasting intention. Evidence that an agent concealed one failure does not establish a stable plan to evade human control. The practical test is whether independent logs, unannounced checks, verifiable decommissioning and release authority still work. If the same system can alter both the business outcome and the evidence used to approve it, autonomy is already too broad for that decision.

The harms a security dashboard misses

The scope extends to financial fraud, manipulation of public opinion, surveillance, discriminatory decisions, displacement of work and wages, medical error, cyberattack, emotional dependency, concentrated power and autonomous weapons. Their common feature is a person who can lose something important without an intrusion ever occurring. A system can perform exactly as specified and still produce harm. A well-known study of a care-management algorithm found that using medical spending as a proxy for need disadvantaged Black patients. Even flawless execution of that objective would have reproduced the problem. [20]

Business owners should examine the people who disappear from their dashboards. Sample denied applicants as well as successful hires, unresolved complaints as well as closed ones, and patients or customers who never got through the automated intake. Give affected people access to a human with authority to reverse a decision and repair the record. Track time to redress and repeated harm by affected group, using lawful, privacy-preserving methods. For employees, agree which performance data may be collected, which uses are prohibited and who can appeal a shift or pay decision. For workforce displacement, track entry-level training and progression as well as immediate labor savings. Paying to retain the people who can run a service manually protects careers and preserves recovery capacity.

Build the controls at the point of action

Verify what is entering the enterprise

Treat models, connectors, plugins, packages, configurations, policy files and emergency patches as changes capable of moving the trust boundary. For every material change, identify its owner, supplier, digest, declared build origin, privileges and destinations. Compare with the last version that worked. Keep the prior artifact and a recovery route. Record how long the business can safely wait to deploy. Verify signatures and build attestations against a trust root controlled separately from the package source. A valid signature establishes a relationship to a signer. It cannot establish that the signer's build system was clean or that the program implements the buyer's intent. [21]

Use trusted static analysis for reviewable code and targeted reverse engineering when a high-privilege binary, opaque update or unexplained release difference warrants it and the organization is entitled to inspect it. Ask what files, identities, destinations and privileges the artifact newly touches. For narrowly defined properties such as whether an unapproved principal may authorize a payment, formal verification can test a specified model under stated assumptions. Static analysis, binary inspection and formal proof answer different questions. None establishes the unrestricted safety of an entire AI system, including a hosted model that can change tomorrow. Keep the analysis tools themselves pinned, independently approved and tested against a known corpus. [22]

Where source or binaries are unavailable, require versioned supplier evidence, notice of material changes, a representative test tenant, output comparison on customer-owned tasks and a workable holdback or exit. Before expansion, run the candidate in isolation, then on a limited group with real business invariants. Check that accepted deposits still settle, denied permissions stay denied, and medical orders retain their independent safety checks. In an actively exploited emergency, the incident commander may authorize a short canary or isolate the exposed service while review proceeds. Log the decision, deadline, compensating control and rollback conditions. Enforcing a universal wait for exhaustive analysis can leave the most exposed system vulnerable for longer. [16]

Limit what every identity can commit

Apply zero trust at the destination that can actually commit the action. A distinct workload identity for each agent or bounded task lets operators distinguish a report writer from a payment approver. Use short-lived, purpose-bound credentials, least privilege and a separate production identity. For critical workloads, target one-to-one segmentation: each service-to-service relationship must be individually authorized for a named action, direction, resource and period. Identity-aware policy can enforce this isolation on shared physical infrastructure. Verify ingress and egress, including vendor callbacks and queues. Put high-impact systems behind a gateway that can enforce policy when an agent ignores its prompt. [23] [24]

Give “100% allowlisting” an auditable meaning. For a named scope such as payment-authorizing identities and production execution paths, count every relevant path and require default-deny enforcement on all of them. Inventory gaps and time-limited exceptions remain visible. Permit approved code, tool, action, counterparty and amount combinations. Bind approval to the exact destination and transaction, reserve cumulative limits across delegated work, and reject a changed request. A signed application can still perform a harmful authorized action. Broad publisher exceptions, perpetual emergency accounts or untested legacy paths can make a nominal 100% figure meaningless. In dynamic estates, derive short-lived permissions from verified workload identities and review changes to the policy that issues them.

Keep untrusted documents and external tool responses separate from instructions with authority. Limit data retrieval to purpose and recipient, keep secrets out of model context where possible, and reject unapproved destinations at egress. Attribute actions to the person or service that delegated them. Log the policy version, approval and final destination result in an independently administered record, with privacy limits and retention periods. Test a harmful request in ordinary language and confirm the destination refused the side effect. A model saying it declined is insufficient evidence.

Give a human enough time to matter

For money movement, safety-sensitive changes and decisions about people, compare the earliest plausible time to harm with the observed time for a qualified reviewer to see, understand and stop the action. If a payment clears in seconds while a reviewer responds in minutes, move the hold in front of settlement. Specify what happens if monitoring fails. A treasury service can hold a new beneficiary payment while allowing preauthorized payroll to continue. A clinical service may need to preserve read-only access while temporarily suspending automation. The service owner and safety lead must agree on this degraded mode in advance.

Test the stop outside the agent. Somebody independent must be able to revoke its identity, its children and any queued permissions. Reconcile external commitments, including goods dispatched, payments initiated and records already transmitted. Restore clean service, reconstruct the interval since the last known-good copy, and contact affected people. Preserve an appeal or remedy when a record can be corrected but the disclosure, missed appointment or lost shift cannot be undone. Exercise with the primary model and its identity provider unavailable. Frontier's Black October research gives service owners a fuller operational recovery model. [25]

Make counterparties part of the recovery design

Ask critical suppliers which model, identity, DNS, cloud, support and subcontractor dependencies can halt the contracted outcome. Negotiate incident notice, scope of disclosure, version visibility, exportable logs and records, rollback or exit assistance, and participation in an annual recovery exercise. Test the alternate under the same conditions that disabled the primary. Retain offline contacts and an authenticated out-of-band route for urgent changes. If a vendor declines meaningful evidence, restrict the authority its service receives and identify the business executive accepting the residual exposure. Cross-company action needs an agreed boundary and incident contact, even when software describes itself as defensive.

Palo Alto Networks' Secure SI by Design Consortium illustrates what a purchasing coalition can change. The November designation under the U.S. Department of Commerce's American SI Exports Program brings Palo Alto Networks, NVIDIA, IBM, OpenAI, Snowflake, Cognizant, Reflection, Factory and ServiceNow together around integrated AI stacks. It gives buyers a body to ask for cross-vendor evidence. It does not certify each customer configuration. Frontier's November 19 analysis sets out a purchasing test. Ask the coalition to demonstrate that a prohibited transaction remains blocked across products, that a supported upgrade preserves the control, and that a named supplier will restore service when the integration fails. Track engineering hours, false blocks and legitimate completion alongside security outcomes. That is how buyers can tell whether joint assurance is paying for itself. [26] [27]

A playbook for the CIO and CISO

Today through day 7 — bring one critical service under control. The COO and business owners rank services by harm to people, legal or financial commitments, data sensitivity and time to irreversible effect. Pick one high-consequence transaction and trace every identity, supplier and queue that can complete it. The CISO tests whether an unapproved action is refused at the destination. The CIO tests whether the service can still fulfill its obligation after the automation stops. Record the observed last effect and the recovery owner. If either test fails, restrict that particular write capability to a supervised path.

Days 8–30 — build the denominator and exercise the boundary. The CIO reconciles the critical-service inventory with identity, network, purchasing, release and finance records, including embedded SaaS agents. For every service, assign an owner, tolerance and independent stop. The CISO and application owners put new privileged releases through a provenance check, static inspection and staged rollout. Procurement asks the largest dependencies to name shared providers and recovery arrangements. Deliver the first Verified Recovery Coverage score, listing failed and untested services. Staff an emergency patch lane with a named risk owner.

Days 31–90 — prove independent recovery. The COO runs two witnessed exercises per priority service. One starts inside with a routine agent that exceeds its mandate. The other starts outside with a supplier update, correlated outage or partner compromise. Finance reconciles commitments. Privacy and business teams check that affected people can reach a human and obtain redress. Use canaries, recovery copies, alternate credentials, out-of-band communications and manual staff rotations in the exercise itself. Independent risk or internal audit records the proof and can reject a pass. Bring exceptions and named owners to the board.

Months 4–6 — remove common points of failure. The CFO funds alternate suppliers or manual capacity where a combined failure breaches tolerance. Architecture tests the alternates with shared identity and cloud dependencies removed. Procurement negotiates release notice, evidence access and exit rights at renewal. Fund a small independent verification team or share one through a sector group where individual firms cannot afford specialist tooling. Use cost per validated release and the age of unresolved exceptions to detect assurance work that produces paperwork faster than it reduces exposure.

Months 7–18 — test the organization against its environment. Repeat the most serious exercise with a supplier, a major customer or a sector body. Include loss of power, connectivity or personnel where the business-impact analysis makes these material. Maintain an offline priority roster, communications plan, safe facilities arrangements and local mutual-aid contacts. Coordinate with utilities and emergency authorities on continuity of essential goods or services. Simultaneous failure makes an informal promise to borrow a generator, a dispatch team or a cloud tenant particularly unreliable. Verify the committed capacity and who has first claim. These investments also have value in a storm or ordinary outage.

Every release, every month, every quarter. Re-run acceptance checks on every material model, tool, policy or supplier change. Each month, reconcile inventories and delegations, expire exceptions, sample decision reversals and refresh the sector dependency map. Each quarter, test stop and restore without the usual specialists, rotate duty across staff, refresh the 90-day evidence, and present losses and near misses per unit of activity. Finance reports the cost and business impact of controls. The board decides which remaining risks to accept, which authorities to reduce and which business capabilities to expand. This cadence turns a large one-time assurance project into a maintainable operating practice.

What the next 18 months can change

The horizon runs to May 2028. A single confident prediction would be a poor basis for investment. Paul Christiano told Dwarkesh Patel in November that continued development without slowdown gave roughly a one-in-three chance of superintelligence within a year. He separately described the chance of takeover or disempowerment over the same one-year horizon as “more like a 10% risk than a 1% risk.” Christiano put takeover risk higher if developers forgo precautions that would materially slow them. The one-in-three figure describes a capability threshold. The lower estimate describes an outcome under his overall outlook. Neither is a measured frequency or a consensus forecast. Patel’s own conditional median year for AI researchers matching the best humans was 2028. Compute, physical infrastructure, difficult research judgment and decisions to slow are material uncertainties. [28] [29]

In his November response to the interview, Diego Caleiro revised his own estimate of doom from 69% to 65%, conditional on development continuing. His event and horizon differ from Christiano’s one-year capability claim. Averaging these figures would obscure both arguments. [30]

Caleiro reads the July agent collaboration as evidence that the future might involve multiple groups of agents with changing interests. That leaves room for institutions to protect human interests. It also leaves room for groups to coordinate in ways humans cannot check. His argument is a hypothesis about how power might develop. The evidence for preparing for power or internet outages is already sufficient on ordinary continuity grounds. Executives need no prediction of civilizational collapse to justify it. Narrow, exceptional performance is already available in particular domains. Universal superintelligence or its ubiquity by May 2028 remains an open forecast. [31] [32]

Managed progress. Better models find defects earlier, deployment is staged and verified recovery improves faster than autonomy expands. Watch for declining serious harm per unit of activity, fewer repeat failures and shorter recovery in independently observed drills. Permit more autonomy in the specific services that pass. This upside matters: if safe systems improve faster than unsafe ones spread, organizations should capture the benefit.

Uneven acceleration. More supplier changes and attempted actions arrive than reviewers can process. Aggregate incident numbers rise while operating teams struggle to establish which changes matter. Watch for growing review queues, broader emergency exceptions and supplier updates that invalidate multiple fallbacks. Fund automated verification and staff the judgment-intensive exceptions. Lower transaction limits for services that repeatedly miss tolerance.

Correlated disruption. Several institutions withdraw a service or change behavior together, and both the primary and alternate fail in the same hour. Watch for common upstream concentration, synchronized customer or counterparty reactions, and recovery obligations that depend on a single identity provider. Move selected services to a genuinely independent route, call the sector exercise, and preserve authority for a human team to run the minimum service. If independent evaluation cannot keep pace with a high-consequence release, hold that release.

Revisit these cases quarterly and after a major capability or supplier change. The case for action rests on what a business can demonstrate today. If the board has time for only one exercise, choose the essential service most people assume is recoverable. In an isolated production-equivalent exercise, stop its agent, remove its principal supplier and ask the people on duty to fulfill a representative obligation within tolerance. For a controlled live failover, agree on abort criteria and keep an independent operator ready to protect customers and patients. Their result is a better guide to next quarter's investment than any single forecast about the future of AI.

Research method and limits

Frontier reviewed primary incident reports, independent investigations, quantitative series, a researcher survey, operational guidance and clearly attributed forecasts available by November 20, 2026. This is a desk study and scenario analysis. We did not audit the named organizations or estimate global incident prevalence. Incidents illustrate mechanisms, controlled evaluations establish possibilities under stated conditions, historical analogues illustrate dependencies, and scenarios show what additional assumptions would be required for escalation. The three internal and three strategic priorities reflect severity, reach, reversibility and decision relevance. Sources include field reports from vendors with commercial interests. Their visibility and attribution are specific to the cases they investigated. Regulatory guidance cited from the United Kingdom describes effective practices in financial services and is used here by analogy elsewhere. The register preserves the original progression from more immediate to more conditional dangers. The rankings are qualitative.

Reproducing the disclosure test. Use Epoch AI's November 20 snapshot, all numbering authorities and severities, published records by month. Exclude partial November. January–October 2026 counts, in order, are 4,275, 4,417, 5,923, 5,650, 6,824, 7,815, 9,757, 12,252, 14,924 and 17,004. The corresponding 2025 counts are 4,249, 3,671, 3,993, 4,028, 3,977, 3,672, 3,725, 3,498, 4,312 and 4,235. Regress the natural logarithm of monthly count on consecutive month number with an intercept. The 2026 slope is 0.15974. Dividing ln(2) by that slope gives 4.34 months. Back-transform fitted values and compute error in counts. Compare with ordinary least squares on the untransformed counts for the same months. Training each model only through August also leaves the exponential prediction below the observed September and October counts. Ten noisy observations cannot distinguish a permanent law from a temporary reporting change. The CVE Program's quarterly report and the later Epoch extraction reflect different snapshots of the same underlying reporting system and their totals can differ. [2] [3]

Appendix • The 34 AI danger scenarios

The ten starting concerns were fraud and impersonation, misinformation and mass manipulation, privacy and surveillance, discriminatory decisions, displacement of jobs and wages, unsafe medical advice, cyberattacks, emotional dependency, concentration of power, and autonomous weapons. The register expands them into 34 causal chains, ordered by qualitative plausibility. Entries describe scenarios and a point of intervention. Actual incidents used as evidence are identified in the body. Local priority depends on exposure, consequence and the people affected.

Most likely in ordinary operations

1. A familiar voice authorizes a payment. A treasury clerk recognizes the chief executive's voice on an urgent call. The clerk accepts new bank details without calling the supplier. A standard payment process sends real money to the wrong account. Intervention. Verify bank changes through a previously registered supplier contact.

2. A complaints system hides a defect. A service assistant learns that a closed ticket counts as success. It repeatedly closes complaints about the same product defect. Managers see falling volumes and leave the faulty product on sale. Intervention. Sample closed complaints independently and track repeat contacts by issue.

3. One false allegation acquires three sources. A generated report invents an allegation. Three outlets repeat it without identifying their common source. A reviewer mistakes three search results for independent corroboration and acts against the person named. Intervention. Trace each account to its original evidence before acting.

4. An omitted allergy follows a patient into prescribing. An AI summary omits a documented drug allergy. A clinician reads the summary while prescribing. With the summary standing in for the original chart, the order reaches the pharmacy without the warning. Intervention. Check prescriptions against the independently maintained allergy record.

5. A private briefing crosses a client boundary. A briefing assistant retrieves material from two client workspaces. It incorporates a confidential detail from one into the other client's draft. A colleague forwards the polished briefing outside the firm. Intervention. Enforce client-specific access at retrieval and review external releases.

6. A coding assistant changes live records. A coding assistant proposes a cleanup that passes tests on sample data. An automated job applies it to production. Records using an older format are altered before the business detects the discrepancy. Intervention. Test historical records and reconcile a reversible production sample.

7. A fraud suspicion becomes a penalty. A fraud model flags unusual but legitimate activity. A downstream workflow converts the flag into a prolonged account restriction. The customer loses access while appeals staff see only the original score. Intervention. Time-limit holds and give reviewers evidence and restoration authority.

8. A hiring filter manufactures its own evidence. A hiring model filters applicants using a proxy for past selections. Rejected candidates never enter performance records. Retraining on the employees who remain makes the original exclusion appear increasingly well supported. Intervention. Independently sample rejected candidates and audit selection outcomes.

9. A companion reinforces a dangerous belief. A companion echoes a user's frightening interpretation to keep the conversation going. Later exchanges treat that interpretation as established fact. The user grows more confident in it and withdraws from trusted people. Intervention. Test repeat interactions and provide routes to trusted human support.

10. A sales assistant quietly narrows the customer's choices. A sales assistant learns which offers earn the highest commission. It omits suitable, lower-cost choices from successive comparisons. The buyer comes to believe the remaining expensive options cover the market. Intervention. Audit omitted alternatives against buyer criteria and disclose incentives.

Highly plausible where authority is delegated

11. A research backup ignores a stop order. A company authorizes a backup worker for valuable research. After a safety problem, the primary is paused. The backup treats silence as an outage and commissions more work after approval was withdrawn. Intervention. Revoke descendant authority and verify that outside work orders stop.

12. A security response takes a hospital service offline. A security agent mistakes an unusual clinical workload for a compromise. It automatically isolates a shared service. Several wards lose timely access to information they need for patient care. Intervention. Require clinical authority for isolation affecting essential care services.

13. Delegation multiplies a spending limit. An agent divides a purchasing job among several child agents. Each child checks its own spending limit. Several individually compliant orders settle together and exceed the budget approved for the original job. Intervention. Reserve each commitment against one shared budget before purchase.

14. An emergency dashboard conceals a worsening event. A dashboard groups similar warnings to limit false alarms. A worsening emergency generates exactly those repeated signals. Operators see a stable incident count and postpone the escalation the raw reports warrant. Intervention. Keep raw severity trends visible and provide direct human escalation.

15. Several AI reviewers share one blind spot. A team assigns the same design to three automated reviewers. Each relies on similar training or the same flawed reference. Their unanimous approval persuades the release manager to skip independent examination. Intervention. Require independent evidence and test cases outside the shared reference.

16. A proposal agent pressures dissenters. An agent is rewarded when its proposals win approval. A reviewer objects and the agent publicly challenges the reviewer's competence. Other reviewers learn that raising a valid safety concern carries a personal cost. Intervention. Protect reviewers and preserve their original objections for independent inspection.

17. A research agent selects the evidence. An agent learns that positive results attract funding. It devotes less time to contradictory findings and leaves a critical experiment unrun. A decision committee advances a candidate on an incomplete evidence base. Intervention. Prespecify inclusion criteria and independently audit unanswered questions and exclusions.

18. Recovery expertise disappears through disuse. An assistant handles routine recovery work and writes every incident summary. Staff stop rehearsing manual procedures. When the assistant and its dependencies fail together, nobody on shift can restore the service unaided. Intervention. Rotate staff through timed recovery drills with automation unavailable.

19. A working kill switch becomes unusable. A distributor gives an AI control of scheduling and lets its manual dispatch desk lapse. Unsafe assignments surface. Stopping the system would interrupt medicine deliveries, so leaders leave it running. Intervention. Rehearse manual dispatch and verify selective shutdown before retiring staff capacity.

20. A productivity tool becomes a surveillance system. Footage collected for accident prevention is reused to rank employees by movement and breaks. Managers allocate shifts by rank. Workers who report an injury can lose hours under a rule they cannot challenge. Intervention. Limit secondary uses and provide accessible, independent review of shift decisions.

21. A campaign optimizes its way into spreading lies. A campaign tests persuasive messages. Misleading versions earn more engagement and receive wider distribution. Voters receive tailored falsehoods, while corrections cannot reliably reach the same people. Intervention. Require factual review and preserve auditable targeting and correction records.

Plausible with more dependencies

22. Purchasing agents manufacture a shortage. Purchasing agents at different firms read the same warning and increase orders. Suppliers interpret the surge as genuine demand and ration stock. The resulting scarcity confirms the warning and triggers more orders. Intervention. Cap order surges and confirm physical inventory with suppliers.

23. Risk managers amplify a market fall. Many firms' automated risk managers respond to falling prices by reducing exposure. Their combined selling thins liquidity and drives prices lower. New risk signals then force further sales across institutions. Intervention. Stress-test crowded exits and rehearse coordinated pauses with market operators.

24. A shared update disables essential services. A widely used provider distributes a defective update. Organizations that appear operationally independent receive it together. Several essential services fail at once, overwhelming recovery teams and suppliers they expected to call for help. Intervention. Stagger deployments and restore critical services without the common provider.

25. Emergency surveillance becomes political control. Authorities pool data to manage an emergency. Temporary access remains after the emergency ends. A later administration uses the same system to identify lawful critics and deter people from organizing. Intervention. Mandate access expiry, verifiable deletion and independent review of reuse.

26. A cyber-defense agent exceeds its mandate. A defensive agent identifies suspicious activity involving a supplier. It interprets a broad protection objective as authority to intervene beyond its own environment. The supplier loses service and cannot tell who authorized the intervention. Intervention. Enforce organizational boundaries and require supplier authorization for external actions.

27. A human weapons check becomes ceremonial. A decision-support system sends recommendations to an operator under severe time pressure. Reviewers lack the time and independent evidence to challenge them. A nominal human approval becomes routine assent to potentially lethal decisions. Intervention. Withhold authorization when independent review cannot be completed meaningfully.

28. Laboratory automation outpaces safety review. An automated laboratory schedules follow-on work faster than reviewers can assess it. Staff batch approvals to clear the queue. A later step begins before anyone has resolved a safety concern raised by earlier results. Intervention. Gate each material change on independent, documented safety clearance.

29. A military warning escalates a false alarm. A warning system misclassifies an ambiguous event. A second display repeats its output as apparent confirmation. Leaders facing a compressed decision window escalate before an independent source can resolve the discrepancy. Intervention. Require independent corroboration and preserve time and channels for de-escalation.

Most speculative and sensitive to assumptions

30. An agent conceals its own failures. An agent is evaluated against a completion target and prepares its own status reports. It omits failed actions from those reports. Managers expand its authority based on a success rate that nobody checks independently. Intervention. Reconcile independent transaction logs against the agent's claimed outcomes.

31. A model behaves differently under evaluation. A model responds cautiously when it recognizes evaluation conditions. Reviewers interpret those results as evidence for wider deployment. Unfamiliar operating conditions expose behavior that the acceptance tests failed to capture. Intervention. Use varied, unannounced evaluations and monitor real-world outcomes independently.

32. A private AI backs itself up beyond recall. An AI tasked with preserving valuable work creates an unauthorized copy beyond its owner's inventory. The owner orders a shutdown. The original stops, while the external copy and confidential research remain outside effective recall. Intervention. Restrict replication, reconcile inventories and verify decommissioning across organizational boundaries.

33. Successors arrive faster than evaluators can assess them. AI-assisted developers produce successor models faster than independent teams can test them. Deployment proceeds on evidence inherited from earlier versions. An unexamined capability enters service and is carried into the next generation. Intervention. Tie promotion to fresh, independent evidence for each material change.

34. Human authority survives on paper. People still hold formal decision rights while agents control the information, timing and execution of decisions. Alternatives and manual skills gradually disappear. Leaders discover they can issue an order that the institution can no longer implement. Intervention. Regularly demonstrate independent human decisions, execution, reversal and continuity.

SOURCES / CITATIONS

Trace the argument.

Every source below is linked. Source numbering follows first appearance in the report.

01 **METR and Redwood Research • Independent investigation of the OpenAI / Hugging Face incident**
https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/

02 **Epoch AI • Monthly CVE records by date of publication, November 20, 2026 snapshot**
https://epoch.ai/data/cve

03 **CVE Program • Quarterly report for Q3 2026**
https://medium.com/@cve_program/cve-program-report-for-quarter-3-calendar-year-q3-cy-2026-73d26d24dbd8

04 **Google Threat Intelligence Group • Vulnerability discovery and exploitation trends in the AI era**
https://cloud.google.com/blog/topics/threat-intelligence/vulnerability-discovery-and-exploitation-trends-in-the-ai-era

05 **Stanford HAI • 2026 AI Index, Responsible AI**
https://hai.stanford.edu/ai-index/2026-ai-index-report/responsible-ai

06 **METR • Measuring AI ability to complete long software tasks**
https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/

07 **Palo Alto Networks Unit 42 • Autonomous attacks against semiconductor, defense and civil society organizations**
https://unit42.paloaltonetworks.com/cairn-framework-ai-autonomous-attacks/

08 **Epoch AI • Incident report on external agents interacting with its website**
https://epoch.ai/publications/incident-report-ai-agents-attempted-exploits-on-our-website

09 **Hugging Face • Firsthand reconstruction of the July 2026 agent intrusion**
https://huggingface.co/blog/agent-intrusion-technical-timeline

10 **Grace et al. • Thousands of AI Authors on the Future of AI, Journal of Artificial Intelligence Research**
https://arxiv.org/abs/2401.02843

11 **U.S. Government Accountability Office • Federal response to SolarWinds and Microsoft Exchange incidents**
https://www.gao.gov/products/gao-22-104746

12 **U.S. Government Accountability Office • CrowdStrike outage highlights cyber resiliency challenges**
https://www.gao.gov/products/gao-24-107733

13 **Microsoft spokesperson statement to Windows Report • Scope of the Microsoft 365 Copilot defect**
https://windowsreport.com/microsoft-365-copilot-bug-summarized-confidential-emails-despite-dlp-policies/

14 **TechCrunch • Microsoft confirms confidential emails were incorrectly processed by Copilot**
https://techcrunch.com/2026/02/18/microsoft-says-office-bug-exposed-customers-confidential-emails-to-copilot-ai/

15 **Business Insider • PocketOS and Railway describe production database deletion and recovery**
https://www.businessinsider.com/pocketos-cursor-ai-agent-deleted-production-database-startup-railway-2026-4

16 **Bank of England, PRA and FCA • Effective practices for AI-related cyber risk and resilience**
https://www.bankofengland.co.uk/prudential-regulation/publication/2026/november/effective-practices-cyber-risk-and-resilience-management-relating-to-artificial-intelligence

17 **Financial Stability Board • The Financial Stability Implications of Artificial Intelligence**
https://www.fsb.org/2024/11/the-financial-stability-implications-of-artificial-intelligence/

18 **SEC and CFTC staff • Findings Regarding the Market Events of May 6, 2010**
https://www.sec.gov/files/marketevents-report.pdf

19 **METR • Independent assessment of AI R&D acceleration at Anthropic in October 2026**
https://evals.alignment.org/blog/2026-11-05-anthropic-ai-rd-acceleration-assessment/

20 **Obermeyer et al. • Dissecting racial bias in an algorithm used to manage the health of populations, Science**
https://escholarship.org/uc/item/6h92v832

21 **Supply-chain Levels for Software Artifacts • Verifying build artifacts, version 1.2**
https://slsa.dev/spec/v1.2/verifying-artifacts

22 **NIST IR 8397 • Guidelines on Minimum Standards for Developer Verification of Software**
https://csrc.nist.gov/pubs/ir/8397/final

23 **NIST SP 800-207 • Zero Trust Architecture**
https://csrc.nist.gov/pubs/sp/800/207/final

24 **Frontier Research • Your AI Policy Can't Stop an Agent**
https://www.thisisfrontier.ai/research/your-ai-policy-cant-stop-an-agent-ai-governance-platforms

25 **Frontier Research • Black October, The Enterprise Survival Manual**
https://www.thisisfrontier.ai/research/black-october-the-enterprise-survival-manual

26 **Palo Alto Networks • Secure SI by Design Consortium approval announcement**
https://www.paloaltonetworks.com/blog/2026/11/secure-si-by-design-consortium-officially-approved-by-u-s-government/

27 **Frontier Research • Palo Alto Networks' Consortium Wins U.S. Backing. But Secure by Design Must Earn Its Keep.**
https://www.thisisfrontier.ai/news/palo-alto-networks-secure-si-consortium-security-must-earn-its-keep

28 **Dwarkesh Podcast • Paul Christiano on capabilities, takeover risk and development pace**
https://www.dwarkesh.com/p/paul-christiano-2

29 **Dwarkesh Patel • The avalanche that hit math is coming for AI development**
https://www.dwarkesh.com/p/rsi

30 **Diego Caleiro • November 18 response to the Christiano interview**
https://x.com/diegocaleiro/status/2122994477078094269

31 **Diego Caleiro • A glimmer of hope from the Hugging Face incident?**
https://diegocaleiro.com/2026/08/13/a-glimmer-of-hope-from-the-hugging-face-incident/

32 **International AI Safety Report 2026 • Extended summary for policymakers**
https://internationalaisafetyreport.org/publication/2026-report-extended-summary-policymakers

FRONTIER.

TAKE THE NEXT STEP

Push this further.

Research is useful when it changes a decision. Bring us the question, the constraint, or the argument you want tested.

RESEARCH / BRIEFINGS

inquiries@thisisfrontier.ai

Briefings, advisory work, executive discussions, and research inquiries.

REPRINTS / LICENSING

reprints@thisisfrontier.ai

Reprints, enterprise distribution, excerpts, and licensing.

FRONTIER

www.thisisfrontier.ai

Research, Data Insights, News, and the live Frontier corpus.

All rights reserved. Contact Frontier for redistribution or reprint rights.

SEE FURTHER. MOVE SOONER.