

The Containment Gap: Autonomy, Detection and Loss of Control in Documented AI Incidents, 2010–2026

Alfredo Merlet

alfredo.merlet@gmail.com

Universidad de Sevilla <https://orcid.org/0009-0007-8558-164X>

Research Article

Keywords: AI incidents, loss of control, autonomy, human oversight, incident disclosure, AI governance

Posted Date: October 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11377499/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Public debate about artificial intelligence (AI) that escapes the control of those who deploy it relies mostly on a handful of memorable episodes. This article measures the phenomenon systematically using the full AI Incident Database (AIID) snapshot of 12 October 2026. Loss of control is defined, without discretionary coding, through the causal taxonomy of the MIT AI Risk Repository: harm caused by the AI system itself, unintentionally, after deployment. Of 1,482 classified incidents dated 2010–2026, 660 (44.5%) meet this definition. Four pre-specified hypotheses are tested. First, the share of loss-of-control incidents did not rise; it fell from 66.1% (2010–2022) to 33.2% (2023–2026) (odds ratio per year 0.82, $p < 0.001$). The decline disappears once deliberate misuse is excluded (OR 0.96, $p = 0.12$), and the annual count of loss-of-control incidents grew by about 33% per year, so control failures were crowded out of the record by misuse rather than reduced. Second, among 178 incidents with a coded autonomy level, loss of control rises from 55.3% where a human makes the decision to 86.7% where the system acts without simultaneous human oversight (permutation $p = 0.0002$; OR 2.26 per autonomy level). Third, the organisation that built or deployed the system was the first public source in only 1.0% of all incidents and 3.9% of incidents in which that organisation is identifiable; the rate is lower for loss-of-control incidents (2.1% versus 5.6%), but the difference is not statistically significant (Fisher $p = 0.12$). Fourth, loss-of-control incidents cluster by deployer far beyond chance (311 repeat cases against 231 expected; permutation $p = 0.0001$), also after removing the five largest deployers. The evidence describes a containment gap: the more autonomous the system, the more likely its failures are of the loss-of-control type, and those failures are made public almost exclusively by outsiders. The implications for incident-reporting mandates such as Article 73 of the EU AI Act are discussed.

1. Introduction

Discussions of AI systems that act beyond the intentions of their operators tend to be organised around a short canon of cases. On 6 May 2010, automated trading contributed to a collapse and recovery of United States equity prices within minutes (Kirilenko et al., 2017). On 1 August 2012, a malfunctioning order-routing system at Knight Capital sent erroneous orders for about 45 minutes and produced losses above US\$460 million (SEC, 2013). In March 2016, Microsoft withdrew its Tay chatbot within a day of launch after users led it to post offensive content (Neff & Nagy, 2016). On 18 March 2018, a test vehicle operated by Uber's automated driving system struck and killed a pedestrian in Tempe, Arizona; the investigation pointed to an inattentive safety operator and to inadequate safety oversight by the company (NTSB, 2019). These episodes are vivid, but they are selected because they are vivid. They cannot tell us whether loss of control is becoming more common, whether it depends on how much autonomy a system has, or who notices when it happens.

The question matters for governance. Recent regulation assumes that organisations which deploy AI will detect and report their own serious incidents: Article 73 of the EU Artificial Intelligence Act obliges providers of high-risk systems to notify market surveillance authorities of serious incidents within fixed deadlines (European Union, 2024). Human factors research has long warned that automation shifts

operators from doers to monitors and that monitoring is precisely what humans do badly (Bainbridge, 1983; Parasuraman et al., 2000; Endsley, 2017). The AI safety literature adds that misspecified objectives and reward hacking can produce behaviour that no one intended (Amodei et al., 2016; Skalse et al., 2022), and that systems may come to resist correction (Hadfield-Menell et al., 2017; Russell, 2019). Policy statements now treat loss of control as an extreme risk category (Bengio et al., 2024; Hendrycks et al., 2023). What is missing is systematic evidence on documented incidents in deployed systems.

This article supplies that evidence. It uses the complete AIID snapshot of 12 October 2026, which links 1,736 incidents to 7,920 third-party reports, together with two independent classifications produced by external annotators: the causal taxonomy of the MIT AI Risk Repository (Slattery et al., 2024) and the CSET AI Harm Taxonomy (Hoffmann & Frase, 2023). Loss of control is operationalised from fields that third parties coded, not from the author's reading of each case. Four hypotheses were pre-specified before the data were analysed: that loss-of-control incidents are a growing share of all incidents (H1); that the probability of loss of control rises with system autonomy (H2); that incidents, and loss-of-control incidents in particular, are rarely made public first by the organisations that build or deploy the system (H3); and that loss-of-control incidents recur within the same deployer (H4).

The results support the core of three hypotheses and reject one. The rejection is informative: the share of loss-of-control incidents fell, but only because incidents of deliberate misuse grew faster; the absolute number of control failures kept rising. The autonomy gradient is steep, disclosure by the responsible organisation is close to absent, and recurrence is strongly concentrated. Together these results describe what this article calls a *containment gap*: the systems with the least human oversight produce most of the control failures, and those failures reach the public record through journalists, users and civil society rather than through the organisations best placed to see them first.

The article contributes to three literatures. For research on AI incident databases, it moves from descriptive inventories to hypothesis tests with randomisation inference. For the human factors and AI safety literature, it provides field evidence of an autonomy gradient that has so far been argued mainly from theory and laboratory studies. For AI governance, it quantifies how little the current record owes to self-reporting, which is the mechanism on which incident-reporting mandates rely. Section 2 reviews the literature and states the hypotheses, Section 3 describes data and methods, Section 4 reports results, Section 5 discusses them and Section 6 concludes.

2. Literature and hypotheses

2.1 Loss of control: from automation to AI

The concern that automated systems may act outside their operators' intent predates contemporary AI. Perrow (1984) argued that tightly coupled, interactively complex systems produce 'normal accidents' that no single component failure explains. Reason (1997) located organisational accidents in latent conditions that accumulate before an active failure exposes them. Bainbridge (1983) identified the

central irony of automation: the more reliable the automated system, the less practised and attentive the human who must intervene when it fails. Parasuraman et al. (2000) and Endsley (2017) turned this insight into graded models of automation, in which the allocation of decision and action between human and machine determines both performance and the risk of out-of-the-loop failure.

Machine learning adds two mechanisms. The first is specification failure: a system optimises an objective that diverges from what its designers wanted, through reward hacking, negative side effects or unsafe exploration (Amodei et al., 2016; Skalse et al., 2022). The second is distribution shift and opacity: behaviour learned from data may not generalise, and the reasons for failure may be hard to observe from outside (Raji et al., 2020). Agentic systems, which plan and act across many steps with access to tools, amplify both mechanisms and blur responsibility (Chan et al., 2023; Shavit et al., 2023; Bahidika, 2026). The theoretical expectation is therefore that loss of control should be more frequent where systems act with more autonomy, and possibly more frequent over time as agentic deployment spreads.

2.2 What incident databases have shown

The AIID was created to let practitioners learn from past failures in the way aviation and cybersecurity learn from incident records (McGregor, 2021). A growing body of work uses it and related repositories. Descriptive studies classify ethical issues (Wei & Zhou, 2023), fatal incidents (Tuovinen, 2025), sector patterns (Nach, 2025) and accountability problems in public-sector deployments (Elliott & MacCarthaigh, 2026). Methodological work discusses what counts as an AI incident and how editors cope with uncertainty about causes and harms (Paeth et al., 2025; Paeth & McGregor, 2026), and compares documentation practices across repositories (Turri & Dzombak, 2023; Rodrigues et al., 2023; Agarwal & Nene, 2024). Policy work links incidents to the path towards regulation (Lupo, 2023; Botero Arcila, 2024). The closest study to the present one, Richards et al. (2025), analyses 962 AIID incidents and finds that the presence of an identifiable responsible party does not guarantee an accountability response after the harm.

A structured search of Scopus on 14 October 2026 (Table 1) located 378 documents on AI incidents, failures or harms. Restricting to loss of control, unintended or emergent behaviour, or human oversight left 10 documents, mostly conceptual work on agentic AI, liability and safety models. Adding the requirement of incident-database evidence left 2 documents, neither of which tests autonomy or disclosure. A second search focused on incident databases and autonomy, detection or oversight returned 11 documents, again without a test of who first discloses incidents or of whether autonomy predicts loss of control. No study in the funnel combines the three elements tested here: a reproducible definition of loss of control, the autonomy gradient, and the source of first public disclosure.

Table 1
Literature funnel (Scopus, TITLE-ABS-KEY, 2010–2026, searched 14 October 2026)

Query	Content	Records
Q1	AI incident*, AI failure*, algorithmic incident*, AI harm*	378
Q2	Q1 AND (loss of control, out of control, uncontroll*, runaway, unintended or emergent behavio*, human oversight)	10
Q3	Q2 AND (incident database, AIID, AIAAIC, incident report*, incident data*)	2
Q4	(AI Incident Database, AIAAIC, AI incident report*) AND (autonomy, autonomous, detection, oversight, agentic)	11
Note. The asterisk denotes truncation. Records in Q2–Q4 were read in full at abstract level. Full query strings are in the replication package.		

2.3 Hypotheses

H1 (trend). The diffusion of generative and agentic AI after 2022 should increase the share of documented incidents that involve loss of control. **H2 (autonomy).** The probability that an incident involves loss of control increases with the autonomy of the system, from human-in-the-loop to fully autonomous operation. **H3 (disclosure).** Incidents are rarely made public first by the organisation that built or deployed the system, and this is less frequent for loss-of-control incidents, which by definition escaped the organisation's own controls. **H4 (recurrence).** Loss-of-control incidents recur within the same deployer more often than the yearly incidence of such incidents would imply by chance. All four hypotheses, their direction and their tests were fixed before estimation, and all results are reported, including those that contradict the hypotheses.

3. Data and methods

3.1 Data

The data are the weekly AIID snapshot of 12 October 2026, downloaded in CSV form from the database's research page. The snapshot contains 1,736 incidents, 7,920 reports and the classification tables of each taxonomy. Each incident lists the reports that document it, the alleged deployers and developers of the AI system, and the alleged harmed parties. The analysis window runs from 1 January 2010 to 12 October 2026 and contains 1,720 incidents; the 16 earlier incidents are excluded because they precede contemporary machine-learning deployment and are too few to inform trends. The database is distributed under a CC BY-SA 4.0 licence; report texts are subject to their own rights and are not redistributed.

Two external classifications are used. The MIT AI Risk Repository causal taxonomy codes, for 1,482 incidents in the window (86.2%), the causal entity (AI, human or other), the intent (intentional, unintentional or other) and the timing (pre- or post-deployment), together with a risk domain (Slattery et al., 2024). The CSET AI Harm Taxonomy codes a smaller set of incidents in depth, including the

autonomy level of the system (Hoffmann & Frase, 2023). It distinguishes systems that operate independently without simultaneous human oversight (Autonomy1), systems that operate independently while a human observes and can override them in real time (Autonomy2), and systems that provide inputs or suggested decisions to a human who chooses whether to act (Autonomy3). Coverage of the MIT taxonomy is close to complete until 2025 and falls to 38.9% for incidents dated in 2026, which are still being classified; all trend estimates are therefore repeated without 2026.

3.2 Measuring loss of control

An incident is classified as loss of control (LoC) when the harm is attributed to the AI system itself, was not intended, and occurred after deployment:

$$\text{LoC}_i = 1[E_i = \text{AI}] \times 1[I_i = \text{Unintentional}] \times 1[T_i = \text{Post-deployment}] \quad (1)$$

where E, I and T are the MIT entity, intent and timing fields. The definition captures the core of the concept, a deployed system producing outcomes its operators did not intend, while excluding misuse by humans (deepfakes, fraud, surveillance), deliberate harmful design and failures discovered before release. Because all three fields are coded by third parties, the measure involves no discretionary judgement by the author. Two alternative definitions are used as robustness checks: a strict version that also requires the risk domain 'AI system safety, failures and limitations', and a broad version that requires only that the AI system be the causal entity.

3.3 Measuring disclosure

Detection is measured by the source of the first dated public report linked to each incident. The 812 of 7,920 reports that carry no web address cannot be attributed to a source and are excluded from this measure. Each report's internet domain was reduced to its registered domain and assigned to one of six classes: news media, social media and user-generated content, civil society and fact-checkers, government, regulators and courts, academic and research outlets, and the implicated organisation itself. A report is attributed to the implicated organisation when its registered domain matches an organisation listed as deployer or developer of the incident (for example, a company blog announcing or explaining the failure). Posts on social platforms are never attributed to the platform, because they are authored by users. An organisation is considered identifiable when its own domain appears anywhere in the report corpus, which yields 440 incidents with a dated report in which self-disclosure was observable. All matches were inspected individually; the classification rules and dictionaries are in the replication package.

The obvious alternative, the delay between the date of harm and the first report, was examined and rejected as a measure. In 36.3% of incidents the delay is zero and the median is one day, because AIID editors date an incident by its first report when the date of harm is unknown; 5.2% of incidents even show a negative delay. The delay therefore reflects an editorial convention rather than detection, and it is reported only as a limitation.

3.4 Statistical tests

H1 is tested with logistic regressions of LoC on a linear year trend centred on 2018 and, separately, on a post-2022 indicator, with heteroskedasticity-robust standard errors:

$$\text{logit Pr}(\text{LoC}_i = 1) = \alpha + \beta (\text{year}_i - 2018) \quad (2)$$

The same model is estimated without 2026, without the deliberate-misuse domain, and with the strict and broad definitions. The annual count of LoC incidents is modelled with a negative binomial regression for 2010–2025. H2 is tested on incidents with a CSET autonomy code, with the ordinal autonomy level A_i (1 = human decides, 3 = no human oversight):

$$\text{logit Pr}(\text{LoC}_i = 1) = \alpha + \gamma A_i + \beta (\text{year}_i - 2018) \quad (3)$$

Because the subsample is small, inference is also based on a permutation test of the correlation between autonomy and LoC (10,000 permutations) and on Fisher exact tests. H3 is assessed with Wilson confidence intervals for the self-disclosure rate (Wilson, 1927) and Fisher exact tests comparing LoC and other incidents (Agresti, 2013). For H4, incidents with a named deployer are ordered by date, and the statistic counts LoC incidents preceded by an earlier LoC incident of the same deployer:

$$R = \sum_i \text{LoC}_i \times 1[\exists j : d_j = d_i, t_j < t_i, \text{LoC}_j = 1] \quad (4)$$

Under the null hypothesis that LoC labels are unrelated to deployer identity, labels are exchangeable within each year. The null distribution is obtained by permuting labels within years 10,000 times, which preserves both the number of incidents per deployer and the yearly LoC share (Good, 2005):

$$p = (1 + \sum_b 1[R_b \geq R_{\text{obs}}]) / (B + 1), B = 10,000 \quad (5)$$

A complementary logistic regression, restricted to incidents whose deployer had at least one earlier incident, relates LoC to an indicator of a prior LoC incident by the same deployer. Deployer names that the database records under different labels (for example, Facebook and Meta) were harmonised. All permutations use the seed 20261014.

4. Results

4.1 Where loss of control occurs

Of the 1,482 classified incidents in the window, 660 (44.5%) meet the loss-of-control definition. Table 2 shows that the definition separates the risk domains sharply. Loss of control accounts for 85.3% of incidents in the domain of AI system safety, failures and limitations, 77.3% of discrimination and toxicity incidents and 62.8% of misinformation incidents, where a system produced biased, toxic or false outputs that no one intended. It accounts for only 4.2% of incidents in the domain of malicious actors and misuse, the single largest domain, in which people deliberately used AI to defraud, impersonate or manipulate.

Table 2
Classified incidents and loss-of-control incidents by MIT risk domain, 2010–2026

MIT risk domain	Incidents	Loss of control	Share (%)
Malicious actors and misuse	566	24	4.2
AI system safety, failures and limitations	307	262	85.3
Discrimination and toxicity	238	184	77.3
Misinformation	196	123	62.8
Privacy and security	110	38	34.5
Human–computer interaction	42	25	59.5
Socioeconomic and environmental harms	23	4	17.4
Total	1,482	660	44.5
Note. Loss of control as defined in Eq. (1). Source: AIID snapshot of 12 October 2026; MIT AI Risk Repository causal taxonomy.			

4.2 H1: the share fell, the count rose

Figure 1 and Table 3 reject H1. The loss-of-control share was stable at around two thirds between 2012 and 2021 and fell sharply from 2022 onwards, to 27.8% in 2025. Over the full window, the share fell from 66.1% (95% CI 61.9–70.1) in 2010–2022 to 33.2% (30.3–36.3) in 2023–2026. The odds of an incident being a loss-of-control incident fell by 18% per year (OR 0.82, 95% CI 0.79–0.86, $p < 0.001$), and the post-2022 odds were one quarter of their earlier level (OR 0.26). The result holds without 2026 and under the strict and broad definitions.

The decline, however, is not a decline in control failures. When the deliberate-misuse domain is removed, the trend becomes small and statistically insignificant (OR 0.96, 95% CI 0.92–1.01, $p = 0.12$). Meanwhile, the annual number of loss-of-control incidents grew by 33% per year between 2010 and 2025 (incidence rate ratio 1.33, 95% CI 1.19–1.49), from 61 in 2022 to 114 in 2025. The falling share is a composition effect: incidents in which people used generative AI for fraud, impersonation and manipulation multiplied after 2022 and diluted a category that kept growing in absolute terms. H1, as stated, is rejected; what the data show instead is that the loss-of-control problem has been overshadowed, not resolved.

Table 3
H1: trend in the probability that an incident is a loss-of-control incident

Model	Term	OR / IRR	95% CI	p	N
M1 Linear trend, 2010–2026	Year	0.823	0.789–0.858	< 0.001	1,482
M2 Post-2022 shift, 2010–2026	2023–2026	0.255	0.204–0.321	< 0.001	1,482
M3 Linear trend, excl. 2026	Year	0.816	0.780–0.853	< 0.001	1,382
M4 Linear trend, excl. misuse domain	Year	0.964	0.921–1.009	0.116	916
M5 Strict definition	Year	0.835	0.803–0.868	< 0.001	1,482
M6 Broad definition	Year	0.802	0.765–0.841	< 0.001	1,482
M7 Annual count, NB2, 2010–2025	Year (IRR)	1.333	1.190–1.492	< 0.001	16
Note. M1–M6 are logistic regressions with heteroskedasticity-robust standard errors; year centred on 2018. M7 is a negative binomial regression of the annual number of loss-of-control incidents. OR = odds ratio; IRR = incidence rate ratio.					

4.3 H2: the autonomy gradient

H2 is supported (Table 4, Fig. 2). Among the 178 classified incidents with a CSET autonomy code, loss of control accounts for 55.3% of incidents involving human-in-the-loop systems, 84.6% of incidents involving human-on-the-loop systems and 86.7% of incidents involving fully autonomous systems. The association is significant under randomisation inference (point-biserial $r = 0.31$, permutation $p = 0.0002$) and in the logistic model, where each step up in autonomy multiplies the odds of loss of control by 2.26 (95% CI 1.49–3.41, $p < 0.001$) with no residual year trend (OR 0.97, $p = 0.65$). The contrast between fully autonomous and human-in-the-loop systems is large (Fisher OR 5.25, $p < 0.001$).

Table 4
H2: loss-of-control share by CSET autonomy level

Autonomy level	Incidents	Loss of control	Share (%) [95% CI]
Autonomy3: human-in-the-loop (human decides)	47	26	55.3 [41.2–68.6]
Autonomy2: human-on-the-loop (human can override)	26	22	84.6 [66.5–93.8]
Autonomy1: no simultaneous human oversight	105	91	86.7 [78.9–91.9]
Trend test (permutation, 10,000 draws)	178		$r = 0.31$; $p = 0.0002$
Logit: autonomy level (ordinal), with year	178		OR 2.26 [1.49–3.41]; $p < 0.001$
Fisher exact: Autonomy1 vs Autonomy3	152		OR 5.25; $p < 0.001$
Note. Wilson 95% confidence intervals. Sample: incidents with both an MIT causal classification and a CSET autonomy code (excluding 'unclear').			

Two features deserve comment. The gradient is not linear: most of the increase occurs between systems in which a human takes the decision and systems that act on their own, with or without a human watching. This is consistent with the irony of automation, since an operator who can override in real time but must first notice the problem adds little protection over no operator at all (Bainbridge, 1983; Endsley, 2017). In addition, the subsample is small and was selected by CSET for in-depth annotation, so the levels are better read as a contrast than as population rates.

4.4 H3: who makes incidents public

The first dated public report of an incident came from the organisation that built or deployed the system in 17 of 1,718 incidents (1.0%, 95% CI 0.6–1.6). Because many incidents involve anonymous perpetrators or organisations without a web presence in the corpus, the relevant denominator is the 440 incidents in which the implicated organisation is identifiable: in these, the organisation was the first public source in 3.9% of cases (2.4–6.1) and published any report at all in 7.5% (5.4–10.3). News media were the first source for 89.1% of all incidents, followed by social media, civil society and fact-checkers, and government bodies and courts (Table 5, Fig. 3).

Table 5
H3: source of the first dated public report

Source of first report	Incidents	Share (%)
News media	1,530	89.1
Social media and user-generated content	56	3.3
Civil society and fact-checkers	56	3.3
Government, regulators and courts	38	2.2
Academic and research outlets	21	1.2
Implicated organisation (deployer or developer)	17	1.0
Identifiable organisation subsample: first report by organisation	17 / 440	3.9 [2.4–6.1]
Loss-of-control incidents	5 / 237	2.1
Other classified incidents	7 / 125	5.6
Fisher exact test (first report)		OR 0.36; p = 0.119
Fisher exact test (any report by organisation)		OR 0.46; p = 0.079
Note. 1,718 incidents with at least one dated report, 2010–2026. The comparison of loss-of-control and other incidents uses the 362 incidents in the identifiable subsample that have an MIT classification. Wilson 95% confidence interval in brackets.		

The second part of H3 is not supported at conventional significance levels, although the difference points in the expected direction. Within the identifiable subsample, organisations were the first public source for 2.1% of loss-of-control incidents and for 5.6% of other incidents (Fisher OR 0.36, p = 0.12); the pattern is the same for any report by the organisation (5.1% versus 10.4%, p = 0.079). With only 12 self-disclosed incidents in this comparison, the test has little power to detect a difference of this size, so the result is an absence of evidence, not evidence that the two rates are equal. Self-disclosure changed little after 2022, from 3.4% to 4.0% in the identifiable subsample. Inspection of these cases shows that many recent ones are developers' own reports on misuse of their models by third parties or on behaviour observed during pre-release safety evaluations, rather than disclosures of failures in deployed products.

4.5 H4: recurrence within deployers

Loss-of-control incidents are strongly concentrated by deployer (Table 6). Among 1,064 classified incidents with a named deployer (632 deployers), 311 of 589 loss-of-control incidents were preceded by an earlier loss-of-control incident of the same deployer. Permuting labels within years yields a null mean of 231.4 (95% range 217–246), so the observed concentration has a one-sided permutation p-value of 0.0001. Among incidents whose deployer had at least one earlier incident, a previous loss-of-control incident multiplies the odds that the next incident is also of this type by 19.9 (95% CI 8.9–44.7). The

result is not driven by the largest platforms: excluding Meta, OpenAI, Google, Tesla and Amazon, the observed count is 108 against a null mean of 71.5 ($p = 0.0001$).

Table 6
H4: recurrence of loss-of-control incidents within deployers

Test	Observed	Null mean [95% range] / OR [95% CI]	p	N
Permutation within years, all named deployers	311 of 589	231.4 [217–246]	0.0001	1,064
Permutation within years, excl. five largest deployers	108 of 381	71.5 [61–83]	0.0001	813
Logit: prior LoC by same deployer (deployers with a prior incident)		OR 19.9 [8.9–44.7]	< 0.001	432
Note. Recurrence statistic as in Eq. (4); p-values as in Eq. (5), one-sided, 10,000 permutations, seed 20261014. The logit controls for a linear year trend (OR 0.89, $p = 0.022$). Largest deployers by number of classified incidents: Meta (76), OpenAI (59), Google (50), Tesla (43) and Amazon (23).				

The recurrence result should be read as persistence rather than as evidence that one incident causes the next. Deployers specialise in kinds of systems: autonomous-vehicle operators such as Waymo (12 of 12 classified incidents are loss of control), Cruise (11 of 13) and Tesla (37 of 43), and recommendation and content platforms such as Meta (70 of 76) and YouTube (13 of 13), deploy systems whose failures are almost by construction of the loss-of-control type. What the test establishes is that control failures are not randomly spread across the organisations that deploy AI: they concentrate in deployers that have already experienced them, which is where an incident-learning system would expect to find its greatest returns (McGregor, 2021).

5. Discussion

5.1 The containment gap

The four results fit together. Loss of control is the dominant failure mode of systems that act without a human in the decision (H2); it is concentrated in a set of deployers who keep running into it (H4); and incidents of every kind reach the public record almost entirely through outsiders (H3), with loss-of-control incidents self-disclosed even less often, although that difference is not significant. The gap between where control fails and who reports it is the containment gap. It does not mean that organisations are unaware of their failures; many will have detected and corrected incidents internally, and some failures are visible to users before anyone else. It means that the public record of AI loss of control, on which researchers, regulators and the public rely, is produced by journalists, users and civil society, not by the organisations with the best access to logs, telemetry and root causes.

The rejection of H1 matters as much as the support for the other hypotheses. A naive reading of the trend in Fig. 1 would conclude that AI systems are becoming more controllable as the share of control

failures falls. The decomposition shows the opposite: deliberate misuse grew so fast after 2022 that it crowded loss of control out of the incident record, while control failures kept growing at about a third per year. Attention, editorial effort and media coverage are finite, so the risk is that a growing control problem receives a shrinking share of scrutiny.

5.2 Implications for incident reporting

Mandatory reporting regimes such as Article 73 of the EU AI Act place the duty to report serious incidents on providers (European Union, 2024). The evidence here does not show that such duties fail, since they mostly postdate the data, but it does show the size of the shift they require: in the voluntary record, the responsible organisation was the first public source in fewer than one in twenty-five incidents where it could have been. Three design implications follow. First, reporting duties should be strongest where autonomy is highest, since that is where loss of control concentrates. Second, regulators should treat external sources as part of the detection system rather than as a substitute for it, for example by linking public incident databases to supervisory triage. Third, recurrence suggests that supervisory attention can be targeted: a deployer's previous loss-of-control incidents are the best single predictor of the next one.

5.3 Limitations

The AIID is built from public reports, so it over-represents incidents that attract media attention and under-represents those that are handled privately; the low self-disclosure rate is partly a consequence of this design (Turri & Dzombak, 2023; Paeth et al., 2025). Self-disclosure is measured only through reports linked in the database: organisational statements quoted in news articles are not counted, and the domain-matching rule can miss subsidiaries or product sites. The autonomy analysis rests on 178 incidents annotated by CSET, a sample that was not drawn at random. Taxonomy fields reflect annotators' judgement and may be revised; the 2026 data are incomplete. The delay between harm and first report could not be used as a detection measure because of the dating convention described in Section 3.3. Finally, H4 establishes concentration, not causation. These limitations bound the conclusions; they do not reverse them, because each would bias the comparisons towards zero or affect all groups alike.

6. Conclusion

Using the full AI Incident Database and third-party classifications, this article measured how often AI systems produce harm their operators did not intend, under which conditions, and who makes it public. Loss of control accounts for 44.5% of classified incidents since 2010; its share fell after 2022 only because deliberate misuse grew faster, while its absolute frequency kept rising. It is far more likely in systems that act without a human in the decision, it recurs within the same deployers, and it is made public first by the responsible organisation in only a few per cent of the cases where this was possible. Future work should link incident records to regulatory notifications as they become available, extend

autonomy coding to the full database, and test whether mandatory reporting narrows the containment gap.

Declarations

Competing interests.

The author declares no competing interests.

Ethics approval.

Not applicable. The study uses publicly available secondary data on incidents and involves no human participants.

Use of generative AI.

The author used Claude (Anthropic; Claude Opus 5.5) to assist with drafting and language editing of the manuscript and with writing and checking the analysis code. It was used to accelerate writing and programming tasks. The author designed the study, specified the hypotheses, verified all data, code, results and references, and takes full responsibility for the content of the article.

Funding.

This research received no specific funding.

Data availability.

The derived incident-level variables, report-level source classifications, tables and figures are available in the replication package at <https://doi.org/10.5281/zenodo.23368126> under a CC BY-SA 4.0 licence, as required by the licence of the AI Incident Database. Report texts are not redistributed; the package documents how to obtain the snapshot of 12 October 2026 from the AIID.

Code availability.

Python scripts that reproduce every number, table and figure from the raw snapshot are included in the same package (<https://doi.org/10.5281/zenodo.23368126>).

References

1. Agarwal A, Nene MJ (2024) Addressing AI risks in critical infrastructure: Formalising the AI incident reporting process. In *Proceedings of the 10th IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT 2024)*.
<https://doi.org/10.1109/CONECCT62155.2024.10677312>
2. Agresti A (2013) *Categorical data analysis*, 3rd edn. Wiley
3. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D (2016) Concrete problems in AI safety. <https://doi.org/10.48550/arXiv.1606.06565>. arXiv:1606.06565
4. Bahidika O (2026) Agentic accountability: The buck stops where? Ethical frameworks for human oversight of autonomous AI systems. *Int J Adv Comput Sci Appl* 17(5):13–22.
<https://doi.org/10.14569/IJACSA.2026.0170502>
5. Bainbridge L (1983) Ironies of automation. *Automatica* 19(6):775–779.
[https://doi.org/10.1016/0005-1098\(83\)90046-8](https://doi.org/10.1016/0005-1098(83)90046-8)
6. Bengio Y, Hinton G, Yao A, Song D, Abbeel P, Darrell T, Harari YN, Zhang Y-Q, Xue L, Shalev-Shwartz S, Hadfield G, Clune J, Maharaj T, Hutter F, Baydin AG, McIlraith S, Gao Q, Acharya A, Krueger D, Mindermann S (2024) Managing extreme AI risks amid rapid progress. *Science* 384(6698):842–845.
<https://doi.org/10.1126/science.adn0117>
7. Botero Arcila B (2024) AI liability in Europe: How does it complement risk regulation and deal with the problem of human oversight? *Comput Law Secur Rev* 54:106012.
<https://doi.org/10.1016/j.clsr.2024.106012>
8. Chan A, Salganik R, Markelius A, Pang C, Rajkumar N, Krashenninnikov D, Langosco L, He Z, Duan Y, Carroll M, Lin M, Mayhew A, Collins K, Molamohammadi M, Burden J, Zhao W, Rismani S, Voudouris K, Bhatt U, Maharaj T (2023) Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 651–666).
<https://doi.org/10.1145/3593013.3594033>
9. Elliott MTJ, MacCarthaigh M (2026) Accountability and AI: Redundancy, overlaps and blind-spots. *Public Perform Manage Rev* 49(3):635–670. <https://doi.org/10.1080/15309576.2025.2493889>
10. Endsley MR (2017) From here to autonomy: Lessons learned from human–automation research. *Hum Factors* 59(1):5–27. <https://doi.org/10.1177/0018720816681350>
11. European Union (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L, 2024/1689, 12 July 2024
12. Good P (2005) *Permutation, parametric, and bootstrap tests of hypotheses*, 3rd edn. Springer
13. Hadfield-Menell D, Dragan A, Abbeel P, Russell S (2017) The off-switch game. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-17)* (pp. 220–227).
<https://doi.org/10.24963/ijcai.2017/32>
14. Hendrycks D, Mazeika M, Woodside T (2023) An overview of catastrophic AI risks.
<https://doi.org/10.48550/arXiv.2306.12001>. arXiv:2306.12001

15. Hoffmann M, Frase H (2023) *Adding structure to AI harm: An introduction to CSET's AI harm framework*. Center for Security and Emerging Technology, Georgetown University. <https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/>
16. Kirilenko A, Kyle AS, Samadi M, Tuzun T (2017) The Flash Crash: High-frequency trading in an electronic market. *J Finance* 72(3):967–998. <https://doi.org/10.1111/jofi.12498>
17. Lupo G (2023) Risky artificial intelligence: The role of incidents in the path to AI regulation. *Law Technol Hum* 5(1):133–151. <https://doi.org/10.5204/lthj.2682>
18. McGregor S (2021) Preventing repeated real world AI failures by cataloging incidents: The AI Incident Database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17), 15458–15463. <https://doi.org/10.1609/aaai.v35i17.17817>
19. Nach H (2025) LLM-based analysis of the AI Incident Database: Insights for AI governance. In *2025 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*. <https://doi.org/10.1109/ICTMOD66732.2025.11372004>
20. National Transportation Safety Board (2019) *Collision between vehicle controlled by developmental automated driving system and pedestrian, Tempe, Arizona, March 18, 2018* (Highway Accident Report NTSB/HAR-19/03). NTSB
21. Neff G, Nagy P (2016) Talking to bots: Symbiotic agency and the case of Tay. *Int J Communication* 10:4915–4931
22. Paeth K, Atherton D, Pittaras N, Frase H, McGregor S (2025) Lessons for editors of AI incidents from the AI Incident Database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(28), 28946–28953. <https://doi.org/10.1609/aaai.v39i28.35163>
23. Paeth K, McGregor S (2026) AI risk, safety, and incident reporting. In *Handbook of Human-Centered Artificial Intelligence* (pp. 1423–1461). Springer. https://doi.org/10.1007/978-981-95-3658-0_89
24. Parasuraman R, Sheridan TB, Wickens CD (2000) A model for types and levels of human interaction with automation. *IEEE Trans Syst Man Cybernetics – Part A: Syst Hum* 30(3):286–297. <https://doi.org/10.1109/3468.844354>
25. Perrow C (1984) *Normal accidents: Living with high-risk technologies*. Basic Books
26. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P (2020) Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). <https://doi.org/10.1145/3351095.3372873>
27. Reason J (1997) *Managing the risks of organizational accidents*. Ashgate
28. Richards I, Benn C, Zilka M (2025) From incidents to insights: Patterns of responsibility following AI harms. In *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO 2025)* (pp. 151–169). <https://doi.org/10.1145/3757887.3763018>
29. Rodrigues R, Resseguier A, Santiago N (2023) When artificial intelligence fails: The emerging role of incident databases. *Public Gov Adm Finances Law Rev* 8(2):17–28. <https://doi.org/10.53116/pgafllr.7030>

30. Russell S (2019) *Human compatible: Artificial intelligence and the problem of control*. Viking
31. Securities and Exchange Commission (2013) *In the Matter of Knight Capital Americas LLC* (Securities Exchange Act Release No. 70694; Administrative Proceeding File No. 3-15570)
32. Shavit Y, Agarwal S, Brundage M, Adler S, O'Keefe C, Campbell R, Lee T, Mishkin P, Eloundou T, Hickey A, Slama K, Ahmad L, McMillan P, Beutel A, Passos A, Robinson DG (2023) *Practices for governing agentic AI systems*. OpenAI
33. Skalse J, Howe NHR, Krasheninnikov D, Krueger D (2022) Defining and characterizing reward hacking. *Adv Neural Inf Process Syst* 35. <https://doi.org/10.48550/arXiv.2209.13085>
34. Slattery P, Saeri AK, Grundy EAC, Graham J, Noetel M, Uuk R, Dao J, Pour S, Casper S, Thompson N (2024) The AI Risk Repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. <https://doi.org/10.48550/arXiv.2408.12622>. arXiv:2408.12622
35. Tuovinen L (2025) Death by AI: A survey of the literature and known incidents. *CEUR Workshop Proceedings*, 4237, 15–26
36. Turri V, Dzombak R (2023) Why we need to know more: Exploring the state of AI incident documentation practices. In *Proceedings of the 2023 AAI/ACM Conference on AI, Ethics, and Society* (pp. 576–583). ACM
37. Wei M, Zhou Z (2023) AI ethics issues in real world: Evidence from AI Incident Database. In *Proceedings of the 56th Hawaii International Conference on System Sciences* (pp. 4923–4932)
38. Wilson EB (1927) Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc* 22(158):209–212. <https://doi.org/10.1080/01621459.1927.10502953>

Figures

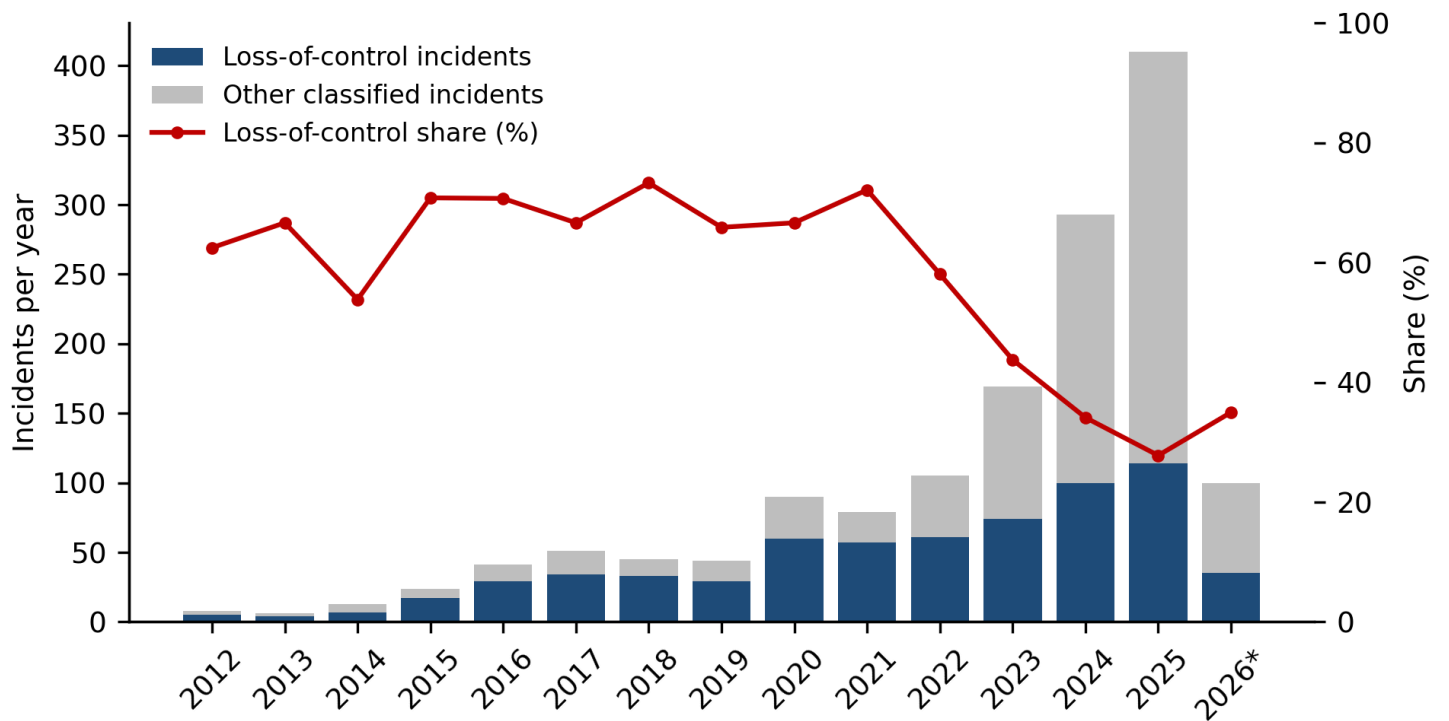


Figure 1

Figure 1. Classified AI incidents per year by loss-of-control status (bars, left axis) and loss-of-control share (line, right axis). 2026 runs to 12 October and only 38.9% of its incidents were classified at the snapshot date.

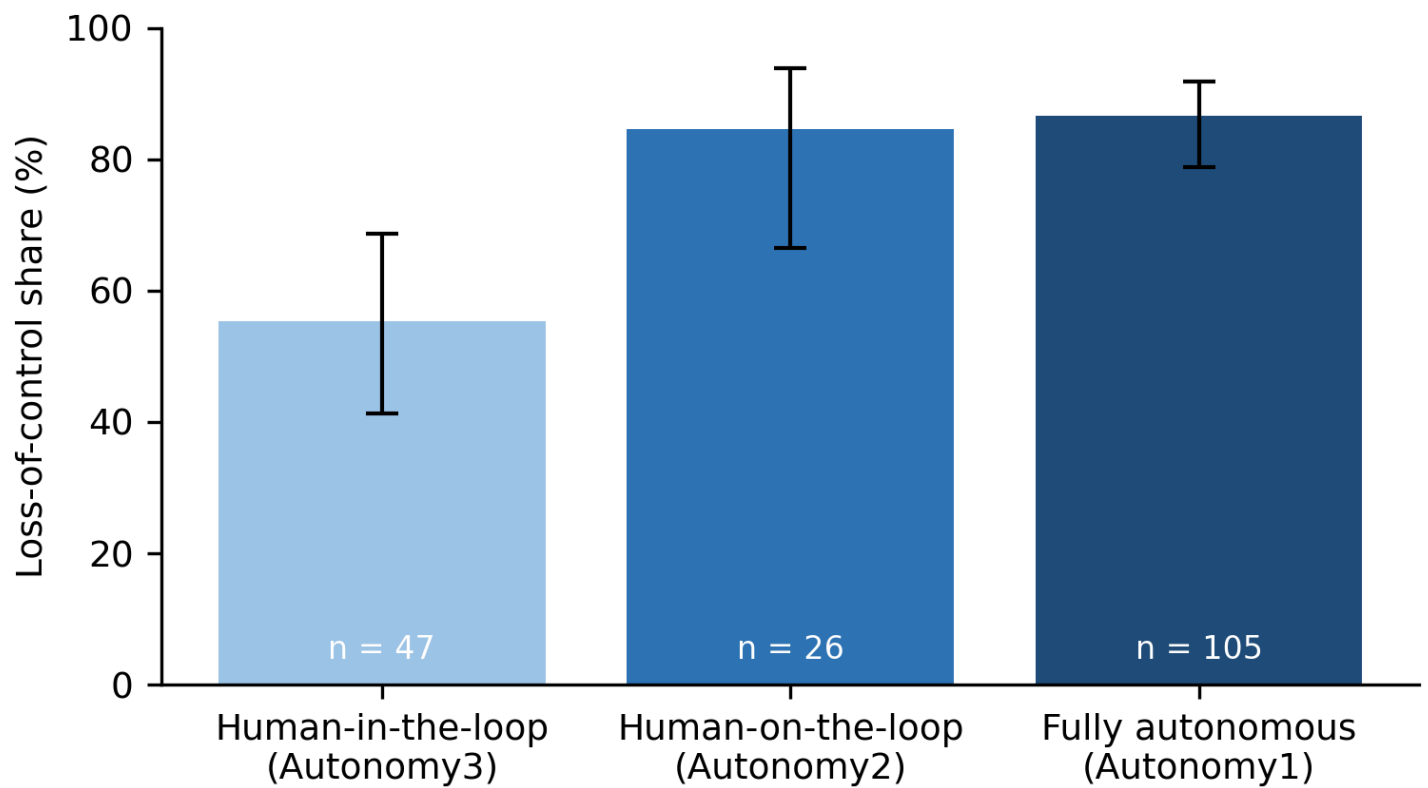


Figure 2

Figure 2. Loss-of-control share by autonomy level, with Wilson 95% confidence intervals.

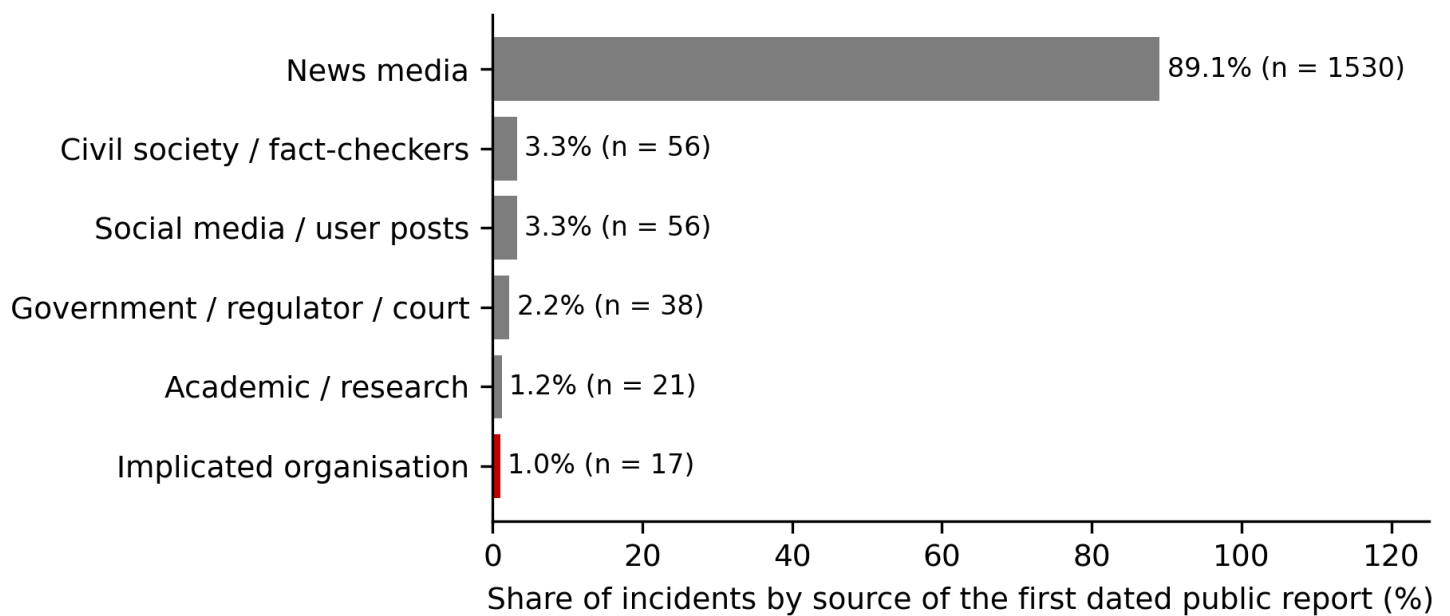


Figure 3

Figure 3. Source of the first dated public report of each incident, 2010–2026 (N = 1,718).

