CROX · WHITE PAPER

Loss of Situational *Awareness*

Why fixing the model will not make AI safe, and what aviation did instead

ADAM FIELD · CROX · NOVEMBER 2026

crox.io/loss-of-situational-awareness

Summary

AI safety is failing because it is solving the wrong problem. Almost the whole effort, from alignment research to guardrails to safety training to the government institutes that test frontier models, is aimed at one thing: making the model fail less. That is component engineering. It is necessary, it has a ceiling, and the ceiling is nowhere near safe.

Aviation spent its first half-century on the same mistake. Most accidents were filed as pilot error and the answer was better pilots: stricter selection and more training. The accident rate plateaued. It fell when the industry stopped trying to perfect the component and designed the system around its failure: redundancy, confidential reporting, just culture, investigations that asked why a decision made sense at the time. No part of a modern airliner is as reliable as the airliner. Safety is a property of the system.

This paper applies that lesson to deployed AI. Guardrails and training are the engine certificate, not the safety case. "The model hallucinated" is this decade's "pilot error", a verdict that closes the inquiry instead of opening it. And the thing AI most lacks is the thing aviation built first, in 1976: confidential near-miss reporting, held outside the regulator, because the model is a crew member that will never file a report.

The wrong problem

In June 2024 a former OpenAI researcher published a long essay arguing that superintelligence arrives before 2030 and that the US government takes over the effort by 2028. Situational Awareness was read in every lab and most ministries. Its safety chapters have two concerns: aligning a system smarter than its makers, and keeping the weights out of Chinese hands.

Pilots use that phrase too, and not as a compliment. When an accident report says a crew lost situational awareness, it means they were competent, the aircraft was serviceable, and they were watching the wrong thing while the real picture changed around them. Hold on to that.

Strip away the scale and the geopolitics and look at what kind of problem it thinks safety is. It is a problem inside the model. Make the model want the right things. Keep the model away from the wrong people. The rest of the field, the part not thinking about superintelligence, has made the same choice at smaller scale. Safety training so the model refuses. Guardrails to filter what it says. Evaluations to count how often it fails. Red teams to find the holes before someone else does. Regulators and government institutes that test the model, certify it, and stop at the point of deployment.

Every one of these tries to make the component fail less. None of them asks what happens when it fails anyway.

It fails anyway. Fine-tuning on as few as ten examples, at a cost of under a dollar, has been shown to strip out safety training. Adversarial prompts found on one model transfer to others. No frontier lab claims a jailbreak rate of zero, and every published patch has been followed by another. A filter on a probabilistic system is itself probabilistic. You can push the failure rate down. You cannot certify it to zero, and nobody who has worked near one believes you can.

Now look at the failures that have actually hurt people. In February 2024 a Canadian tribunal found Air Canada liable for negligent misrepresentation after its chatbot invented a bereavement fare rule. The airline's position, as the tribunal member put it, amounted to treating the chatbot as a separate legal entity responsible for its own actions. That tells you nobody inside owned what it said, and when the customer reported the error there was no path for that report to reach the system. In late 2021 Zillow shut its home-buying business after its pricing model drifted and the company wrote down hundreds of millions of dollars. The missing control was not a better model. It was a kill criterion written down before the drift, and a person whose job was to call it. In the Netherlands an algorithm flagged tens of thousands of families for childcare benefit fraud, with nationality among its risk factors. Officials acted on the flags for years, warnings from inside went unanswered, and in January 2021 the government resigned.

A better model would not have prevented any of these. A better system would have caught every one: someone watching the output against policy, someone with the standing to stop, a channel for the people who saw it first.

That is the argument of this paper. AI safety is failing because it is trying to perfect the component, while the harm arrives through the system around it.

I flew 737s before I worked in technology. The industry I came from had this exact problem fifty years ago and solved most of it. Not with better pilots. With a system that assumed pilots would err and was built to catch the error before it mattered.

The rest of this paper is about that system: why aviation gave up on the perfect component, what it built instead, and what in it transfers to AI.

Aviation's 1970s

It was not virtue that made aviation safe. It was a decade so bad that the industry ran out of ways to blame the dead.

For most of aviation's first half-century, the response to a crash was to find the component that failed and improve it. When the component was a person, the report said pilot error and the fix was selection and training. Boeing's own tally of the jet age's first three decades put the flight crew as primary cause in close to three quarters of hull losses with a known cause. The

fixes had been applied for a generation, and the accident rate, after falling steeply through the 1960s, flattened in the early 1970s and stayed there. Better pilots had stopped working, because the pilots were already good.

In December 1972 an Eastern Air Lines crew flew a serviceable L-1011 into the Everglades while all three of them worked on a burnt-out nose gear indicator bulb. In March 1977 two 747s collided on a fogbound runway at Tenerife and 583 people died; the captain who began his take-off roll without clearance ran KLM's flight training, and when his flight engineer asked whether the Pan Am was clear of the runway he was overruled in two words. In December 1978 a United DC-8 ran out of fuel over Portland while the captain worked a gear problem and his crew hinted, and hinted, and never said it plainly.

The aircraft and the airlines differ. The failure is identical: competent people in a cockpit where the system depended on each of them being right, with no mechanism for the moment one of them was wrong.

What came next was built in a hurry and against resistance. In 1976 NASA began running a confidential reporting system on behalf of the FAA, placed deliberately outside the regulator so a pilot could admit a mistake without the agency that licensed him ever learning his name. A 1979 NASA workshop put a name to the cockpit problem, and by 1981 United was running the first comprehensive crew resource management programme in the United States, teaching captains to invite challenge and first officers to make it. Over the next two decades came threat and error management, the Swiss cheese model, formal safety management systems, and an investigation culture that asked why a decision made sense to the crew rather than whose fault it was.

Pilots hated much of it at first. Reporting your own error felt like confession. Being taught by a psychologist how to speak to your first officer felt like an insult. It stuck because the accident rate fell and kept falling, and because the system protected the people who used it.

AI is in its 1970s, and it is making the 1970s mistake. Its reports say hallucination where ours said pilot error, and its fixes are the ones we tried first: a better component, more training, tighter rules about what it may do. The difference is that the failures are quiet. A hull loss leads the evening news. A model that misprices ten thousand homes or miscodes a thousand discharge summaries is a line in a quarterly report, if it appears anywhere. The pressure that forced aviation to abandon the perfect component is absent, which means AI gets to abandon it voluntarily or wait for its Tenerife.

No part is as safe as the aircraft

Here is the number that should reorganise how anyone thinks about AI safety. A transport aircraft is certified so that any failure that would prevent continued safe flight and landing is extremely improbable, which the certification guidance on both sides of the Atlantic defines

as the order of once in a billion flight hours. The engines on the 737 I flew, among the most reliable machines ever built, shut down in flight about once every 300,000 to 500,000 hours on the manufacturer's own figures.

Those two numbers are several thousand times apart, and the gap between them is the whole of aviation safety. Nobody closed it by building a billion-hour engine. It was closed with a second engine, an aircraft that flies on one, crews trained for the shutdown, a reporting system that catches the precursors, and a certification regime that treats the failure as expected rather than exceptional. The aircraft is safer than any part of it because safety was designed as a property of the system.

AI safety is trying to build the billion-hour engine. Every guardrail, every round of safety training, every eval is an attempt to push the component's failure rate down, and the field measures its progress in exactly those terms: refusal rates, jailbreak rates, benchmark scores. The numbers improve. They do not approach zero, and the distance between where they sit and where safe would be is the distance aviation stopped trying to close with the component fifty years ago.

This is not an argument against guardrails, any more than aviation is against reliable engines. Certify the engine. Then stop pretending the certificate is the safety case.

What aviation built

Six habits do most of the work of closing that gap. None of them makes the component better. Each one assumes it will fail. I describe each as aviation practises it, then as it would look wrapped around a deployed model.

Confidential reporting, protected by a just culture

The accident is the tip. The near miss is the data, and you only get the data if the person reporting is confident it will not be used against them. NASA's ASRS has taken millions of reports since 1976. In the UK, CHIRP has done the same since 1982, as a charity the CAA funds but does not run, for the same reason. Just culture draws the line in advance: honest error is reported and learned from, reckless disregard is still sanctioned, and the line is written down before anyone crosses it.

For AI: a channel where an engineer or a clinician can say "the model nearly did X and I caught it", read weekly, with report volume tracked as a health metric where rising is good. Almost no deploying organisation has one. The nearest thing is a Slack message that gets fixed quietly and forgotten.

Brief the threat before you fly

Every sector starts with a threat and error brief. Not the normal procedure but the abnormal: what is different today, where will it bite, what do we do when it does. Thirty seconds, aloud, every time.

For AI: a pre-deployment brief per release. Three threats, each with a named watcher and a named action. Most teams have a launch checklist, which is the normal procedure, and a feeling, which is not a brief.

Crew resource management

CRM is the discipline of getting the right information out of the most junior person in the room while the most senior person still has time to act on it. Its tools are mundane: standard challenge phrases, a required response, an expectation that the captain asks before deciding.

For AI: who in the team can stop a release, in what words, and what happens to them afterwards. If the answer is "the data scientist decides" or "the founder decides", the authority gradient is the one that killed people at Tenerife.

Layered defences

James Reason's Swiss cheese model (1990) is familiar and I make no apology for it. Each defence has holes; accidents happen when the holes line up. The 737 MAX is the modern lesson: a flight control law acting on a single sensor, no redundancy, crews not told the system existed. One slice, nothing behind it.

For AI: the layers are evals, guardrails, human review, monitoring and rollback. Most teams have two of the five and believe they have four. "Monitoring" often means a dashboard nobody opens.

Investigate for why it made sense

The lazy reading of Air France 447 is that a pilot held the stick back. The investigation found iced pitot tubes, an autopilot that dropped out at altitude, a flight regime the crew had never been trained to hand-fly, and a stall warning that went silent as the stall deepened. Every action made sense to the person taking it, with what they could see. The fixes went to the aircraft, the training and the warning logic, not to a memo saying "do not stall".

For AI: "the model hallucinated" is this decade's "pilot error". It names the component and closes the file, and like pilot error it is usually true and never sufficient. An incident review needs three questions instead. What did the people involved know at the time? What did the system make easy or hard? What changes to the system, not the people? If the output is "be more careful" or "we need a better model", the review has failed.

A safety management system

Since 2013 ICAO has required operators to run a formal safety management system: a policy with named accountable executives, a risk process, an assurance loop that checks the controls are real, and a promotion effort that keeps the culture alive when nothing has gone wrong for a while. The SMS is the wrapper that stops the other five habits decaying.

For AI: this is the governance layer most organisations believe they have because they have an AI policy. A policy is the first of four components. The other three are where the work is.

The human in the loop

The AI industry's favourite safety control is a person. A clinician signs off the generated note. A lawyer checks the drafted clause. A reviewer approves the output before it reaches the customer. Human in the loop, the slide says, and the room relaxes.

Aviation ran this experiment for forty years and wrote down what happened.

Lisanne Bainbridge's 1983 paper Ironies of Automation made the case before glass cockpits were common: the more you automate, the more the tasks left to the human are the ones automation cannot do, and the less practised the human is at doing them. Parasuraman and Riley's 1997 review gave the field its taxonomy, the use, misuse, disuse and abuse of automation, and the failure modes that literature went on to document are now standard. Complacency, where the operator stops checking a system that is usually right. Automation bias, where the operator trusts the system over their own eyes. Skill fade, where the hand-flying muscle atrophies because it is never used.

Then the accidents arrived to confirm the papers. Air France 447's crew had never been trained to hand-fly at high altitude, because in normal operation the aircraft never asks them to. Asiana 214 struck the sea wall at San Francisco in 2013 with experienced pilots aboard who believed the autothrottle was holding their speed; it had dropped into a mode in which it would not. An American Airlines training captain, Warren Vanderburgh, had been warning about "children of the magenta line", pilots who follow the flight director rather than fly the aircraft, since 1997.

Now read the NHS guidance on ambient scribes. The control is that the clinician reviews the generated note before it enters the record. A tired doctor, at the end of a clinic, reading twenty notes that are usually right, looking for the one that is not. That is the Asiana cockpit. It is precisely the monitoring task aviation learned humans are worst at, and we have written it into the design as the main defence.

I am not arguing against human review. I am arguing that a human in the loop is a layer, not the layer, and that a layer made of a bored person checking a mostly-right machine is thinner than it looks. Aviation's answer was to design for engagement: require the human to act rather

than observe, sample rather than rubber-stamp, hand-fly on purpose, and never let the automation be the only thing between the aircraft and the ground.

The AI equivalent is designable. Make the reviewer confirm specific facts rather than approve a block of text. Seed known errors into the stream and measure how many are caught. Rotate who reviews what, and watch approval time per item: when it collapses towards zero, that is a warning, not a productivity gain.

None of this is in the AI safety literature, because the AI safety literature is about the model. Aviation's literature is about the person sitting next to the machine. That is the chapter AI needs most, and it was written before most of today's engineers were born.

Where the analogy breaks

I would not trust a paper on this subject that did not spend a section arguing against itself. Five things aviation had that AI does not, one thing aviation got wrong, and one difference that cuts the other way.

Bodies. Aviation's system was paid for in hull losses. A crash is unambiguous, public and counted. A model that quietly denies the wrong people credit for a year produces no wreckage and no board of inquiry. The pressure that built ASRS does not exist for AI, and nobody should expect it to arrive on schedule.

A regulator with teeth. A 737 cannot fly without a type certificate, an airline cannot operate without an operating certificate, and a pilot cannot work without a licence the regulator can suspend. Deployed AI has nothing comparable. The EU AI Act's high-risk obligations now start in December 2027 for standalone systems and August 2028 for systems embedded in regulated products. The MHRA regulates the slice of medical AI that qualifies as a device. Most deployed models sit in none of these.

A stable machine. Every 737 of a given type behaves the same way, and when one is found unsafe the fleet is grounded. A model changes under its operator: new versions, prompt drift, data drift, a vendor update nobody announced. Aviation's reporting system assumes the thing you report on tomorrow is the thing you flew today. With AI that assumption fails weekly, and one failure mode replicates across every output at once.

An operator who notices. Aviation's entire reporting apparatus runs on a human in the seat who felt the aircraft do something wrong and chose to say so. The model is in the seat now. It does not notice and it does not report. Whatever reporting system AI builds has to be designed around the humans adjacent to the model, the reviewer, the support agent, the clinician, and around instrumentation that does the noticing for them.

A profession. Pilots carry a licence, belong to unions that fought for just culture, and stay in the industry for thirty years. AI product teams turn over in eighteen months and carry nothing between employers. Culture is hard to build in a workforce that does not stay.

And aviation is not finished. The 737 MAX showed that a mature safety system can still be captured by commercial pressure and a regulator that delegated too much to the manufacturer. Anyone holding aviation up as a model should hold MCAS up beside it. The lesson is not that aviation got it right. It is that aviation built a system that eventually told it when it had got it wrong, and 346 people still died before it did.

One difference runs in the paper's favour. Aviation could measure its components. An engine's failure rate is a number with a confidence interval, and the whole certification edifice rests on being able to produce it. A language model's failure rate is not a number of that kind. It depends on the prompt, the deployment, the data that arrived that week and the version the vendor pushed overnight, and the evals that claim to measure it are themselves contested. If you cannot bound the component, the component cannot be the safety case. Aviation chose the system when it still had the option of trusting its parts. AI does not have that option.

So the transfer is of methods, not outcomes. The methods were designed for a world with bodies, regulators, stable machines, operators who notice and a profession that stays. Each one needs rework before it fits a world with none of those. The next section is that rework.

What to build

Two levels. What a deploying team can build next quarter without anyone's permission, and what the industry needs that no single team can build.

At team level: a safety management system for one model

Start with one model in production, not the whole estate.

1. Map the layers. Draw evals, guardrails, human review, monitoring and rollback as five slices. For each one: is it real, who owns it, when did it last fire, what happened when it did. Mark the holes honestly. Most teams find two real slices.
2. Brief each release. Three named threats, each with a named watcher and a named action, written before the deploy and read aloud at it. Keep them. They are the beginning of your hazard log.
3. Open a near-miss channel. "The model nearly did X and I caught it." Reviewed weekly by someone with authority to change the system. Report volume is a health metric, and rising is good.
4. Sign a just-culture statement. One paragraph from the most senior person on the team: honest error is reported and learned from, reckless disregard is still sanctioned, and here

is where the line sits. It is tested the first time someone uses the channel, and everyone will be watching what happens to them.

5. Review incidents with three questions. What did the people involved know at the time? What did the system make easy or hard? What changes to the system, not the people?
6. Design the human layer for engagement. Confirm facts, do not approve blocks. Seed known errors and measure the catch rate. Watch approval time per item and treat its collapse as a warning.
7. Get observed. Aviation audits normal operations under non-jeopardy terms, with trained observers in the jumpseat counting threats and errors on routine flights. The equivalent for an AI team is a periodic outside look at how the model is actually run, as opposed to how the policy says it is. Leadership almost never has that data.

None of this needs a regulator or a standard. All of it needs a senior person willing to be told things they did not want to know.

At industry level: confidential near-miss reporting for AI

The gap I keep coming back to is the one aviation closed first, in 1976. There is no confidential, non-punitive, cross-organisation channel for the people closest to deployed AI to report what nearly went wrong.

What exists is crash-model reporting. The AI Incident Database collects public failures after the fact. The OECD's incidents monitor scrapes the news. The EU AI Act will require providers to report serious incidents to authorities, which is the equivalent of mandatory accident reporting and will capture a fraction of what practitioners see. Nothing collects the near miss from the engineer who saw it, protects their identity, and feeds the pattern back to everyone else.

Aviation's design answers the obvious objections. It runs outside the regulator, as NASA ran ASRS, so that a report cannot become evidence. Identifying details are stripped at intake and the original destroyed. De-identified patterns are published without company names, so a reviewer in one hospital trust learns what a reviewer in another caught. And early reporters get something back: ASRS gave a waiver of certain sanctions for an event the pilot had reported, and that one incentive is a large part of why it worked.

The UK could do this at the scale of one sector in a year. Clinical AI is the obvious first sector, because the safety-case culture already exists under DCB0129 and DCB0160, and because the people reviewing AI output there are clinicians who have filed incident reports their whole careers.

I have built one of these controls myself, at a much smaller scale, and it taught me the lesson of this paper the hard way. rogerwilco is a small open-source tool for teams building software with AI agents. Before an agent does anything it cannot undo, a merge, a deploy, a migration, the tool checks what is about to happen against what is actually on disk and in the record, and

refuses if they differ. The research behind it found that asking the agent to account for itself added nothing on capable models (0 of 140 runs improved) and corrupted the account on weaker ones (51 of 129 relayed statements altered). What caught the mistakes was reading the machine's state directly: 16 of 16 pre-send checks against the record. The flight data recorder does not ask the pilot what happened. Neither should the system around a model.

I have more confidence in the team-level list than the industry-level one. The team-level list has been tested, in a different industry, for fifty years. The industry-level one is a proposal, and the right response to a proposal is to argue with it.

Liability and discovery

Every lawyer who reads the previous section will have the same objection, and it is a good one. A near-miss log is a discoverable document. The engineer who writes "the model nearly prescribed the wrong dose and I caught it" has created evidence that the organisation knew its system could prescribe the wrong dose. The general counsel's rational instruction is to write nothing down. Just culture dies in the first disclosure request.

Aviation hit the same wall and legislated through it. In the United States, ASRS reports are de-identified by NASA before the FAA sees them, and a filed report earns the reporter a waiver of certain sanctions. In Europe, Regulation 376/2014 on occurrence reporting requires member states to protect reporters from proceedings for unpremeditated error, restricts the use of occurrence reports in disciplinary action, and writes just culture into law. The UK carried that regime over after leaving the EU. The principle in each case is the same: the safety value of the information outweighs its evidential value, except for wilful misconduct and what the regulation calls a manifest, severe and serious disregard of an obvious risk.

AI has none of this. A deploying organisation that builds a near-miss channel today is doing so against its own legal advice, and the first time that channel surfaces in litigation, every other general counsel will shut theirs.

Three partial answers are available now, none sufficient on its own. Legal professional privilege can cover a review conducted at the direction of counsel, which is how many organisations already handle post-incident investigations, though it sits uneasily with weekly operational review. Contractual confidentiality within a sector scheme, of the kind CHIRP operates, raises the bar without making the material undiscoverable. And de-identification at intake, done by a third party, changes what exists to be discovered.

The full answer is statutory. Safety reporting protection for AI, modelled on the aviation regime, limited to honest error, with wilful misconduct and serious disregard of obvious risk carved out. That is a legislative ask, and it is the one I would put to any minister who says they care about AI safety.

Until then the practical advice is uncomfortable. Build the channel anyway, keep it narrow, route it through counsel where you can, and accept that a safety culture carries legal risk in a world that has not yet decided to protect it. Aviation accepted that risk for years before the law caught up.

Close

Go back to the Everglades, Tenerife and Portland. Nobody in those cockpits was incompetent. Each one held people who could see something was wrong and had no way, in that culture and that moment, to make it matter. Nobody fixed it by making the pilots better, because the pilots were already good. The industry fixed it by building a system that assumed the pilot would err, and stopped blaming them for being human.

AI is spending its safety budget on a better pilot. The model will improve, it will keep failing, and the harm will keep arriving through the system around it, unwatched and unreported.

Your model is a new crew member. It is confident and mostly right, and it will never file a report. Somebody has to build the system around it.

Appendix

Aviation to AI deployment

AVIATION PRACTICE	WHAT IT DOES	DEPLOYED-AI EQUIVALENT
Component certification (engine reliability, CS 25.1309)	Bounds the failure rate of one part; never the safety case on its own	Guardrails, safety training, evals, red teams
Pilot error verdict	Attributes the failure to the component and closes the inquiry	"The model hallucinated"
ASRS (US), CHIRP (UK)	Confidential, non-punitive near-miss reporting held outside the regulator	Team near-miss channel now; sector scheme proposed
Just culture	Written line between honest error and reckless disregard, drawn in advance	Signed one-paragraph statement, tested at first use
Threat and error brief	Pre-sector briefing of what is abnormal today and what the crew will do about it	Pre-deployment brief per release: three threats, watchers, actions
Crew resource management	Challenge and response across the authority gradient	Who can stop a release, in what words, and what happens to them
Swiss cheese model	Layered defences; harm when holes align	Evals, guardrails, human review, monitoring, rollback
LOSA	Observation of normal operations under non-jeopardy terms	Periodic outside observation of how the model is actually run

AVIATION PRACTICE	WHAT IT DOES	DEPLOYED-AI EQUIVALENT
Safety management system (ICAO Annex 19)	Policy, risk management, assurance, promotion	Governance beyond the AI policy document
Automation complacency and bias	Operator stops checking a mostly-right system	Reviewer rubber-stamping generated output
Flight data recorder	Records what the machine did, independent of the crew's account	rogerwilco: check the agent's claimed act against disk and record
Type certificate, airworthiness directive, grounding	Fleet-wide response to a discovered defect	Model version control, rollback, vendor change notification
Regulation 376/2014 protections	Legal protection for safety reporters	Does not yet exist for AI

Claims checked

Checked on 30 October 2026 against the sources linked below, by a reviewer who had not seen the draft. Where the check corrected a claim, the text above carries the corrected wording.

CLAIM	STATUS	SOURCE OPENED
Catastrophic failure conditions must be extremely improbable, of the order of 1 in 10^9 per flight hour	Confirmed. The figure lives in the advisory material (AC 25.1309-1A/1B, AMC 25.1309), not the rule text	FAA AC 25.1309-1A
CFM56-7B in-flight shutdown rate, 0.002 to 0.003 per 1,000 engine hours	Confirmed. ETOPS ceilings are 0.05, 0.02 and 0.01 per 1,000 hours by diversion time	CFM press release; 14 CFR Part 25 Appendix K
Flight crew as primary cause in most hull losses; rate plateau from the early 1970s	Corrected. Boeing's cumulative 1959 to 1987 figure is 73%; there is no 1970s-only figure; the rate flattened in the early 1970s	Boeing data via Flight Safety Foundation; Boeing Statistical Summary 1959 to 2001
Eastern Air Lines 401, December 1972	Confirmed, with precision: the altitude-hold function disengaged while the crew worked on the nose gear indicator	NTSB AAR-73-14
Tenerife, March 1977, 583 dead, take-off without clearance	Corrected. The first officer's challenge was answered; the flight engineer's doubt was overruled	FAA Lessons Learned
United 173, December 1978; origin of CRM	Confirmed, with precision: NTSB recommendation A-79-47; United's 1981 programme was the first comprehensive one in the US	NTSB AAR-79-7; FAA CRM history
ASRS: 1976, NASA-run for the FAA, waiver of sanctions, 2.3 million reports	Confirmed	ASRS programme briefing; ASRS immunity policy
CHIRP	Corrected. Founded 1982 by the CAA and the RAF Institute of Aviation Medicine; an independent charity since 1996; CAA funded	CHIRP history; CAA on CHIRP
Reason, Human Error, 1990	Confirmed (secondary source)	Swiss cheese model

CLAIM	STATUS	SOURCE OPENED
Bainbridge, Ironies of Automation, 1983	Confirmed	Crossref record
Parasuraman and Riley, 1997	Confirmed as citation; the terms complacency and automation bias are not in its abstract, so the text now credits the wider literature	Crossref record
Air France 447	Confirmed, including the stall warning behaviour and the absence of high-altitude manual handling training	BEA final report, 2012
Asiana 214	Confirmed. Autothrottle in HOLD; three of 291 passengers killed	NTSB AAR-14/01
Vanderburgh, children of the magenta, 1997	Confirmed (secondary sources only)	AOPA
737 MAX, MCAS on a single sensor, pilots not told, 346 dead, regulatory capture	Confirmed	House Transportation Committee final report, September 2020
ICAO Annex 19	Confirmed. Adopted February 2013, applicable November 2013	ICAO
Regulation 376/2014 reporter protections	Corrected. The statutory exceptions are wilful misconduct and a manifest, severe and serious disregard of an obvious risk; UK assimilated law	Article 16
Qi et al., 2023, fine-tuning strips safety training	Corrected. Ten examples for under twenty cents	arXiv 2310.03693
Zou et al., 2023, transferable adversarial attacks	Confirmed	arXiv 2307.15043
Jailbreak rates above zero after every patch	Not verified; the text now says no lab claims a rate of zero	none
Moffatt v. Air Canada, 2024 BCCRT 149	Corrected. Damages of C$812 for negligent misrepresentation; the separate-legal-entity line is the tribunal's characterisation of Air Canada's position	Tribunal decision
Zillow Offers, November 2021	Confirmed. $304m write-down plus $240m to $265m expected; 25% of staff	Zillow Q3 2021 shareholder letter
Dutch childcare benefits	Confirmed. About 26,000 wrongly accused; dual nationality processing ruled unlawful in July 2020; resignation 15 January 2021; 2017 warnings ignored	Parliamentary inquiry report, December 2020; NL Times on the regulator's finding
NHS England ambient scribing guidance v3, 29 July 2026	Confirmed	NHS England
EU AI Act high-risk dates, 2 December 2027 and 2 August 2028	Confirmed. The omnibus is adopted law: Regulation (EU) 2026/1744, in force 27 July 2026	EUR-Lex
AI Incident Database, OECD monitor, AI Act Article 73	Confirmed: public submissions, news scraping and mandatory reporting to regulators respectively; none is confidential near-miss reporting	AIID; OECD; Article 73

CLAIM	STATUS	SOURCE OPENED
rogerwilco study figures	Confirmed by the author against the study notes: 0 of 140, 51 of 129, 16 of 16	Author's own studies, 2026

Sources read for this draft

- Situational Awareness: The Decade Ahead, Leopold Aschenbrenner, June 2024
- NHS England: guidance on AI-enabled ambient scribing products
- Bird & Bird: ambient scribing in the NHS, a compliance framework
- Taylor Wessing: the EU Digital Omnibus on AI