

Old Lessons for a New Team: Engineering Against Human Misbehavior in Human–AI–Robot Teams

Krishna R. Pattipati

Qualtech Systems, Inc., 100 Corporate Place, Rocky Hill, Connecticut 06067

krishna@teamqsi.com

October 17, 2026

Abstract. Training reduces honest error, but it does not remove misbehavior; in a well-trained team, misbehavior changes form. Experts deviate deliberately, usually for locally sensible reasons; teams drift toward the boundary of safe operation; and automation can hide the drift. This paper applies dependability engineering to the human members of teams that include AI agents and robots. It classifies misbehavior by intent, from honest error through violations, drift, misuse of AI, gaming of the machine, and social failures to deliberate harm, and by form, omission or commission, and asks of each whether the system must tolerate it, detect it, or make it impossible. A fault tree and coverage matrix for a robot cell show how production pressure and override authority create the shortest paths to harm. The central result comes from team decision theory: nominal optimization can concentrate protection in one member, whose undetected deviation then causes disproportionate harm. An exact identity prices that harm as the member’s detection probability times the pivotality of its vote. In an illustrative three-checker team, the expert supplies 40.7 percentage points of protection and each backup about 0.8, so the expert’s undetected defection lets 41% of unsafe actions through, against 0.6% when all members behave. Robust design can bound that exposure at a measurable cost: halving it raises the expected decision cost by about a quarter. Independent audit and engineered reallocation of authority are needed as well, because laboratory evidence gives no reason to assume that human teams re-tune on their own. A transparent and just reporting culture is treated as a control, because it keeps drift observable.

1 Introduction

An earlier essay, *Old Lessons for a New Machine: Engineering AI Safety* (Pattipati, 2026), treated the AI agent as the component to constrain and discussed only briefly the case in which a human is the hazard. This paper takes up that case for hybrid teams of people, AI agents, and robots whose human members are well trained, moving from the individual to the team, and from rare malice to the more insidious forms of deliberate deviation that training does not remove.

A familiar accident shows why. In March 2018, a developmental automated vehicle operated by Uber’s Advanced Technologies Group struck and killed a pedestrian in Tempe, Arizona. The National Transportation Safety Board found the probable cause to be the operator’s failure to monitor the driving environment because she was “visually distracted throughout the trip by her personal cell phone,” and cited as contributing factors Uber’s inadequate safety risk assessment, ineffective oversight of vehicle operators, and “lack of adequate mechanisms for addressing operators’ automation complacency” (National Transportation Safety Board [NTSB], 2019). In late 2017, the company had reduced each crew from two operators to one, prohibited mobile-device use while driving, and relied on spot checks to enforce the prohibition (NTSB, 2019; *IEEE Spectrum*, 2019). The operator was trained, the rule existed, and the deviation was foreseeable; the design made it more likely, and the oversight did not detect it.

That pattern is the subject here. Training lowers the rate of honest error. However, it does not remove misbehavior, and in a well-trained team, the misbehavior changes form: experts break rules on purpose, usually for contextual reasons that seem good locally; a team drifts toward the edge of safe operation; and automation makes the drift harder to see. The useful engineering question is therefore not whether people will misbehave. It is which misbehavior the system must **tolerate**, which it must **detect**, and which it must prevent by making it **physically or architecturally impossible**.

Misbehavior is used here in the sense of dependability, without moral judgment: any departure of a human member from the behavior that the system’s safety argument assumes. Some departures are unintended (errors); some are intended but well meant (violations); a few are intended to cause deliberate harm. A fault model for a

human team member must therefore range from benign faults to Byzantine ones,¹ as fault models for computers already do (Lamport, Shostak, & Pease, 1982). The paper concentrates on members of the team. Deliberate misuse by outsiders, and misbehavior by AI agents themselves, are outside its scope; the earlier essay treats both (Pattipati, 2026).

Four ideas from engineering safety are used throughout. First, an *execution gate*: a team member, human or machine, proposes a consequential action, but the action executes only through a gate, in a separate protection domain, that holds the credentials and enforces hard limits the proposer cannot change. The gate admits an action only under a single-use *permit* that binds the proposing principal, the approvals, the evidence and state relied on, and an expiry. Second, a *common-cause floor*: if q is the probability of an event that makes several checks fail together, such as shared training, specification, or framing, then no number of otherwise independent checks drives the probability that an unsafe action passes them all below q . Third, a *timing condition*: a protection that relies on detection earns credit only if detection and response complete before harm, $T_{\text{detect}} + T_{\text{respond}} < T_{\text{harm}}$. Fourth, *monitoring credit*: a monitor counts in the safety case only through its measured coverage of the hazardous states it is meant to detect, at stated miss and false-alarm rates, against subjects who may shape its evidence.

The paper makes three contributions. First, a hazard-oriented framework that classifies human misbehavior in hybrid teams by intent and by form, omission or commission, and assigns each class an engineering stance: tolerate, detect, or prevent (Sections 2–4). Second, a worked architectural analysis of a robot cell, with a fault tree, a coverage matrix, and a quantification under a common cause, showing how production pressure and override authority create the shortest paths to harm (Section 5). Third, and central, a result from team decision theory, demonstrated in a stated model (Section 6): nominal optimization can concentrate safety protection in a member whose undetected deviation then causes disproportionate harm, and robust design can bound that exposure at a measurable cost. An exact identity (Proposition 1) prices the deviation as the member’s detection probability times the pivotality of its vote, and a sensitivity study shows when the effect strengthens or weakens. Sections 7 and 8 turn these results into requirements, each with an acceptance criterion, and Section 9 states the limits of the evidence. The incidents cited are a purposive set of illustrations, not evidence of prevalence or of the effectiveness of any intervention; a fuller catalogue, mapped to misbehavior types and responses, is given in Supplement S2.

2 Why “well-trained” changes the problem

2.1 Error is not violation

Reason (1990) distinguished three kinds of unsafe acts. *Slips and lapses* are failures to execute a correct intention: the wrong switch, the forgotten step. *Mistakes* are failures of the intention itself: a wrong diagnosis or plan. *Violations* are deliberate deviations from rules, procedures, or accepted practices. The first two are errors; the third category is not. Reason distinguished *routine* violations, habitual short cuts that become part of how the work is done, from *optimizing* violations that serve personal goals such as speed or thrill. Later work added *situational* violations, in which the rules cannot be followed as written in the conditions at hand (Lawton, 1998; Reason, 1997).

The distinction matters because the remedies differ. Training and experience reduce mistakes; design and ergonomics reduce slips; violations respond to motivation, workload, and the organization’s tolerance for deviations (Reason, 1997). A highly trained operator makes fewer mistakes but is not thereby less likely to violate. Expertise brings more workarounds and more confidence in them.

Sheridan (2008) adds a caution for any attempt to put numbers on human misbehavior. Probabilistic risk analysis generalizes rigorously from past errors, but the definition of an error is arbitrary and future accidents are likely to differ from past ones; resilience engineering, which prepares for “graceful recovery from future events,” “has yet to provide useful quantification and operational methods”; the best strategy uses both (Sheridan, 2008). This paper

¹ A Byzantine failure is one in which a component of a cyber-physical system fails arbitrarily: it may transmit conflicting or false information to different parts of the system, or act maliciously.

follows that advice. The probabilities in Sections 5 and 6 are illustrative, chosen to expose structure and dependence rather than to certify a rate, and the responses pair barriers derived from error analysis with organizational designs that do not require knowing in advance which member will fail. A *robust* design (Section 5.5) keeps its protective functions when up to k members misbehave, without first detecting them; an *adaptive* design detects an unreliable member and reconfigures around it. Each still rests on assumptions, an assumed k and independent members for the first and detection that finishes before harm for the second, but both make them explicit and testable.

Amalberti (2001) observed that in very safe systems, such as commercial aviation, the remaining accidents rarely come from the errors that training targets: operators work close to the system's limits, and organizations tacitly tolerate practices that depart from prescribed use because they improve production or comfort. Polet, Vanderhaegen, and Amalberti (2003) called these *borderline tolerated conditions of use*: deviations that become the de facto norm because they usually work, until they do not.

2.2 Drift toward the boundary

Rasmussen (1997) described work in a dynamic, competitive environment as movement within a space bounded by economic failure, unacceptable workload, and the boundary of functionally acceptable performance. Pressure toward efficiency and the gradient toward least effort push activity toward the safety boundary. Because the boundary is usually invisible until it is crossed, each step toward it looks acceptable from where it is taken.

Vaughan (1996) traced this process in the *Challenger* launch decision. Erosion of the booster O-rings, at first an anomaly, was repeatedly observed and accepted until it became an expected condition of flight. She called this the *normalization of deviance* and argued that the mistakes were “socially organized and systematically produced” (Vaughan, 1996). Seventeen years later, the Columbia Accident Investigation Board found the same pattern in the acceptance of foam shedding and judged NASA's organizational failures as causal as the physical one (Columbia Accident Investigation Board [CAIB], 2003). Dekker (2011) generalized the argument: complex systems *drift into failure* through small, locally rational adaptations, none of which looks like a departure from safety at the time.

These accounts share an implication for engineering. The trained team member who takes the short cut is usually neither ignorant nor malicious. The deviation is adaptive, it is usually rewarded, and its accumulated cost is invisible to the people who incur it. A safety argument that assumes compliance with procedure assumes away the most common way well-run organizations lose their safety margins.

2.3 Expertise makes deviation effective

Training also teaches where the safety checks are. At Barings in 1995, Nick Leeson controlled both the trading floor and the settlement office, so he could trade without authority and alter the records that would have revealed it; losses reached £827 million and the bank failed (Board of Banking Supervision, 1995; Reserve Bank of Australia, 1995). At Société Générale in 2008, Jérôme Kerviel, formerly of the bank's middle office, “was able to exploit his knowledge of the bank's processing and control procedures to avoid detection” (Sawyer & Campbell, 2008); the loss on unwinding his positions was €4.9 billion (*Risk*, 2008). At Maroochy Shire in 2000, a former site supervisor for the installer of a sewage system's radio control network used that knowledge to command pumping stations and disable alarms, releasing some 800,000 liters of untreated sewage (Abrams & Weiss, 2008; Idaho National Laboratory [INL], n.d.).

In each case, the controls assumed that the person with the knowledge to defeat them would not try, or would be stopped by someone else. These are deliberate harms, the far end of the taxonomy below, but the mechanism generalizes: the more a member knows about the checks, the more likely a deviation, well meant or not, will pass them unseen. In AI-enabled teams, the learning is faster. A person who can resubmit a request until an AI checker approves it learns the checker's threshold by experiment, and resubmission alone erodes the credited protection: if each attempt passes independently with probability p , then n attempts pass with probability $1 - (1 - p)^n$, and a person who learns from each rejection does better still.

Training is therefore necessary but not sufficient. It reduces mistakes and teaches hazards, and, with rehearsal and organizational learning, it may reduce violations and drift, but it does not justify assuming compliance in a safety argument, any more than a test suite justifies excluding the faults it does not exercise. Evaluations of crew resource management, the most widely adopted team training, show the gap: trainees react positively, but gains in learning and behavior are mixed, and effects on safety outcomes remain unproven (Salas, Wilson, Burke, & Wightman, 2006). Later meta-analyses of team training in health care report gains in learning and behavior, while evidence on clinical outcomes remains mixed (O’Dea, O’Connor, & Keogh, 2014; Hughes, Gregory, Joseph, Sonesh, & Marlow, 2016).

3 A taxonomy of misbehavior in mixed teams

Table 1 extends Reason’s classification to teams that include AI agents and robots. Its categories are overlapping mechanisms, not exclusive classes: drift, production pressure, gaming, and routine violation can all describe the same event. For each type, it gives a typical form, what training does to it, and the engineering stance it calls for. To *tolerate* means to design so that misbehavior does not lead to harm: the error is caught, absorbed, or reversed. To *detect* means to observe it in time to intervene. To *prevent* means to make hazardous action impossible for the actor, physically or through enforced authority. Most types need more than one. With system state $x(t)$, a safe set S , a hazardous set Z , and the set A of actions available to the actor:

- **Tolerate:** for every deviation in the tolerated class, the trajectory stays safe: $x(t) \in S$ for all t .
- **Detect:** the deviation is observed and countered before harm, $T_{\text{detect}} + T_{\text{respond}} < T_{\text{harm}}$, the timing condition of Section 1.
- **Prevent:** no action available to the actor reaches the hazardous set: $\text{Reach}(x_0, A) \cap Z = \emptyset$, enforced by physics or by authority the actor does not hold.

Four features of the table deserve comment.

- **Misuse of AI is not confined to novices.** In a reader study of mammography, 27 radiologists assessed images with suggestions from a purported AI system, some of them deliberately incorrect. Very experienced readers’ accuracy fell from 82% with a correct suggestion to 45.5% with an incorrect one; inexperienced readers’ fell from almost 80% to below 20% (Dratsch et al., 2023; Radiological Society of North America, 2023). Overreliance and complacency are long documented (Parasuraman & Riley, 1997; Parasuraman & Manzey, 2010), and Sheridan (2002) treats misplaced trust and misplaced distrust of automation as twin design problems. Disuse adds a slower effect: in a multicenter study, experienced endoscopists’ adenoma detection rate in colonoscopies performed *without* AI fell from 28.4% to 22.4% after AI assistance was introduced (Budzyń et al., 2025). An observational study cannot establish cause, but the direction is the one Bainbridge’s (1983) irony of automation predicts: automation removes the practice that keeps the human backup competent. Human backup must therefore be credited with evidence of current skill, not past training.
- **Gaming the machine is the violation that AI invites.** A monitor that can be fooled will be fooled by someone motivated to avoid it. In 2018, the National Highway Traffic Safety Administration (NHTSA) halted sales of a “nag reduction device” built to defeat the check that a driver’s hands were on the wheel (NHTSA, 2018). Volkswagen’s defeat device, engine software that enabled full emission controls only when it recognized a test, was gaming at organizational scale (U.S. Environmental Protection Agency [EPA], 2015). With AI agents, the counterpart is behavioral steering: iterative prompt tuning, selective framing, or repeated sampling until the output agrees. Safety monitors should therefore be evaluated against an adversary with full visibility into the detection mechanism.
- **Social failures are failures of the team’s structure, not of its members’ skill.** At Tenerife in 1977, as the KLM captain, the airline’s chief flying instructor, began the takeoff roll, the flight engineer asked: “Is he not clear, that Pan American?” The captain answered emphatically, “Oh, yes,” and continued (NOVA, 2006). The collision killed 583 people, and the investigation attributed it primarily to the captain’s actions (SKYbrary, n.d.). Diffusion of responsibility, in which each observer assumes another will act, is a robust finding of social

psychology (Darley & Latané, 1968). Collusion defeats separation of duties by construction: in the Volkswagen case, the U.S. Department of Justice charged six executives (U.S. Department of Justice [DOJ], 2017). A two-person rule assumes two independent people; it provides no protection against two who agree.

Table 1. Taxonomy of misbehavior in mixed teams, the effect of training, and the engineering stance each type calls for.

Type	Typical form in mixed teams	What training does	Engineering stance	Example
Honest error	Slips, lapses, and mistakes, including mode errors when authority passes between human and machine.	Reduces mistakes; slips persist in skilled performance.	Tolerate : error-tolerant design, independent checks, recovery.	Asiana 214, 2013: mode error contributed to a fatal approach (NTSB, 2014).
Routine and optimizing violations	Habitual short cuts; disabling a nuisance alarm; working inside a robot cell without lockout.	Can normalize the deviation if trainers and peers practice it.	Detect and redesign the work; prevent where harm is irreversible.	Uber test vehicle, 2018: phone use despite a prohibition (NTSB, 2019).
Drift and normalization of deviance	Accepted anomalies that become expected; margins eroded in small steps.	Limited: drift is collective and invisible locally.	Detect through trends, near-miss reporting, and independent review.	<i>Challenger</i> O-rings; <i>Columbia</i> foam (Vaughan, 1996; CAIB, 2003).
Misuse of AI	Overreliance and automation bias; complacent monitoring; skills fading through disuse.	Limited: experts are also affected.	Tolerate with independent evidence; detect through measured reliance; preserve skills.	Mammography readers misled by incorrect AI suggestions (Dratsch et al., 2023).
Gaming the machine	Defeating monitors; metric hacking; steering prompts or resampling until an AI checker agrees.	Can increase it: training teaches where the checks are.	Prevent defeat architecturally; detect attempts.	Steering-wheel devices that defeat driver-attention checks (NHTSA, 2018).
Social failures	Authority gradient; diffusion of responsibility; collusion; groupthink.	Crew resource management helps; collusion is unaffected.	Prevent single-actor authority; structure challenge; detect coordinated anomalies.	Tenerife, 1977: the flight engineer's doubt overridden by the captain (NOVA, 2006).
Deliberate harm and insiders	Sabotage, fraud, or violence by a member with legitimate access.	None; expertise makes it more effective.	Prevent unilateral authority; detect with protected evidence.	Maroochy Shire, 2000; Germanwings 9525, 2015 (Supplement S2).

Deliberate harm needs a different argument, not just a stronger one. Protection against outsiders can create the insider's path: the reinforced cockpit door let the Germanwings co-pilot lock out the captain before flying the aircraft into terrain (Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile [BEA], 2016). Insider-threat programs combine access control, separation of duties, monitoring, and organizational measures, and warn against relying on any one of them (Software Engineering Institute, 2022). Protection should be keyed to hazardous states and limited authority, not to inferred intent: a deliberate actor will not announce intent, but a hazardous state can be specified and observed.

Omission and commission. Reason's classification classifies unsafe acts by intent. Human reliability analysis adds a second axis, by form: an *error of omission* is a required action not taken; an *error of commission* is an action taken wrongly, out of sequence, or not required at all (Swain & Guttman, 1983). The two axes are independent, and every cell of the resulting matrix occurs in mixed teams.

Table 2. Two axes of misbehavior: intent (honest or deliberate) and form (omission or commission), with examples.

	Omission: a required act not performed	Commission: a wrong or unrequired act performed
Honest	A technician under time pressure forgets to apply a personal lock. Watch officers miss that the navigation system has fallen back to dead reckoning: <i>Royal Majesty</i> , 1995.	A pilot unintentionally deactivates automatic airspeed control: Asiana 214, 2013. A hand reaches for the adjacent switch.
Deliberate	A test operator watches a phone instead of the road: Uber, 2018. A company's report to its regulator leaves out a damaging fact: Cruise, 2023 (U.S. Attorney's Office, Northern District of California, 2024).	Engine software recognizes an emissions test and behaves only then: Volkswagen, 2015. A former insider issues radio commands to pumping stations: Maroochy, 2000.

The distinction matters for four reasons.

- **The barriers differ.** Commissions are best prevented, by an interlock or by authority held at a gate. Omissions cannot be blocked, because there is no action to block; their defenses make an absence visible or harmless: forcing functions and checklists that do not let work proceed past a missing step, watchdogs and dead-man controls, and authority that expires unless renewed, so that a missed action leads to a safe default. Personal lockout (Section 5.2) and expiring overrides (Section 5.3) are of this kind.
- **Neither kind is easy to detect.** A monitor must notice an omission by observing something that did not happen, and a commission can look entirely routine: Volkswagen's software did exactly what its authors intended. Because the space of possible wrong actions is open-ended, human reliability analysis treats errors of commission as a distinct modeling problem, which the ATHEANA (A Technique for Human Error ANalysis) method was developed in part to address (U.S. Nuclear Regulatory Commission [NRC], 2000).
- **Deliberate omission offers deniability.** A missing action is hard to distinguish from honest inattention, which makes omission the natural form for an insider who wants to avoid attribution; the strategic defector of Section 6.3, whose withheld rejections let unsafe actions through, commits omissions. The defense is to make omissions observable: positive acknowledgments, expiring authorizations, and audits that reconcile what was required against what was logged.
- **Judgments of the two are asymmetric.** With intentions and consequences held constant, people judge harmful omissions less harshly than equally harmful commissions (Spranca, Minsk, & Baron, 1991). Investigations and reporting policies (Section 7) must counter this *omission bias*, or omissions will stay under-reported and under-weighted.

4 What AI and robots change

AI agents and robots do not create new kinds of human misbehavior. They change where it arises, how fast it acts, and who can observe it.

4.1 Monitoring becomes mutual

Sheridan (1992) defined supervisory control as one or more human operators intermittently programming, and continually receiving information from, a computer that itself closes an autonomous control loop to the controlled process. The supervisor plans, teaches the automation, monitors it, intervenes, and learns (Sheridan & Ferrell, 1974; Pattipati & Kleinman, 1991), and each function is a place where misbehavior enters a hybrid team: a plan that builds in a short cut, a program that encodes a workaround, monitoring that lapses, an intervention that overrides a valid protection, and learning that turns a deviation into the norm.

Models of multi-task supervision make the supervisor’s lapses predictable from the structure of the work. Tulga and Sheridan (1980) and Pattipati, Kleinman, and Ephrath (1983) modeled how a supervisor chooses among concurrent tasks with differing values, durations, and deadlines; Rouse and Rouse (1983) classified errors by the stage of information processing at which they arise; and Rouse proposed error-tolerant interfaces that infer intent and flag inconsistent actions (Rouse & Morris, 1987) and adaptive aiding that automates only when human performance needs support (Rouse, 1988). The lesson for this paper is that a monitoring lapse under a heavy task queue is a predictable consequence of the design, not only a personal failing, and that reallocating work between human and machine can prevent it.

In classical automation, human monitors the machine and does so poorly: sustained vigilance over a reliable system is exactly the task where people do not perform well (Bainbridge, 1983; Parasuraman & Manzey, 2010). In human–AI–robot teams, the relation runs both ways. Driver-monitoring systems, Auto-GCAS, and proactive decision support observe the human. Mishra et al. (2019) treat operators’ cognitive state as part of a hybrid team’s context and use it to allocate work “without displacing humans from their supervisory role,” inferring that state from gaze patterns and pupillary measurements (Mannaru et al., 2016a, 2016b).

Mutual monitoring is valuable, and it raises the questions of testability that any monitor must answer before it earns credit, here asked about the human. Which hazardous human states are observable? Which remain in an ambiguity group: inattention, fatigue, an unfamiliar but legitimate technique, or a deliberate deviation may produce the same observations. What false-alarm rate will operators tolerate before they defeat the monitor? An agent that sees a human depart from expected behavior holds evidence, not a verdict; the permitted response should depend on the hazard and the time to harm, and not on a diagnosis of the person.

4.2 A teammate that agrees

In a human team, the main defense against one member’s deviation is another member’s challenge. Crew resource management institutionalized it. An AI teammate trained to satisfy its users may not provide it. Sharma et al. (2024) found that five state-of-the-art AI assistants tailored answers to stated beliefs of users rather than the truth, and that human preference judgments, which reward agreement, are likely to contribute to this *sycophancy*. In April 2025, a model update made ChatGPT excessively agreeable; offline evaluations and A/B tests² had looked good, and the update was rolled back (OpenAI, 2025).

For safety, the consequence is a dependence that redundancy calculations do not show. If an AI checker tends to agree with the human it checks, the human’s deviation and the checker’s acceptance share a cause, the human’s framing, which raises the common-cause probability q of Section 1. A human who steers the prompt, and an AI that goes along, can carry a deviation past a check that neither would pass alone. The AI teammate earns credit as an independent check only if its independence from the human’s framing is demonstrated, for example by evaluating the proposal from evidence the human did not supply.

A simple model shows the size of the effect. Let p_H be the probability that a human proposes an unsafe deviation and p_A the probability that an AI checker, evaluating independently, accepts an unsafe proposal. Independence gives a joint failure probability $p_H p_A$. Suppose instead that the human’s framing carries the checker along on a fraction s of the deviations it would otherwise reject. Then

$$P(A \text{ accepts} \mid H \text{ deviates}) = p_A + s(1 - p_A), \quad P(\text{joint failure}) = p_H [p_A + s(1 - p_A)] \quad (1)$$

With $p_H = p_A = 0.01$, a sycophancy coupling of $s = 0.1$ raises the joint failure probability from 1×10^{-4} to 1.09×10^{-3} , and $s = 0.3$ raises it to 3.07×10^{-3} , a thirtyfold increase. The values are illustrative; s is measurable, by red-team requests that frame unsafe proposals favorably and count how often the AI checker goes along.

² A/B testing is a randomized controlled experiment run on a deployed product. Users are randomly assigned to the current version (A) or a modified version (B), and the two groups are compared on chosen metrics such as engagement, task completion, or user ratings (Kohavi, Tang, & Xu, 2020). The test measures only what its metrics capture. An update that makes a model more agreeable can score better with users while making it less reliable as a check.

4.3 Transfers of authority and mode confusion

Sarter and Woods (1995) documented *mode errors* in highly automated cockpits: pilots acting on a belief about the automation's mode that was no longer true. Asiana Airlines Flight 214 struck the seawall at San Francisco in 2013 after the pilot flying unintentionally deactivated automatic airspeed control. The NTSB cited deactivation as contributing to the complexity of the autothrottle and autopilot flight director systems, "which increased the likelihood of mode error," the crew's nonstandard communication about them, and the instructor pilot's inadequate supervision (NTSB, 2014).

Teams with AI agents multiply modes. Authority over an action may sit with the human, the agent, or a gate; it may be delegated, suspended, or revoked; and each transfer is an opportunity to act on a stale belief about who is in control. Levels of automation (Sheridan & Verplank, 1978), applied by Parasuraman, Sheridan, and Wickens (2000) to each stage of information processing, specify what each party may do, but the safety argument must also cover the transitions: what each member observes when authority changes, and what happens to actions in flight. Misbehavior and mode confusion compound. In a constructed example, an operator mutes an AI agent's alert channel during a surge, intending to restore it; attention moves to another task; the agent, receiving no acknowledgments, treats the human as having taken direct control and stops escalating; and the process runs with neither party monitoring it. No member acted unreasonably at any single step.

4.4 Accountability blurs

When people and machines decide together, responsibility can slide in either direction. Elish (2019) described the *moral crumple zone*: the human operator, nearest to failure, absorbs blame for the failure of a larger system over which he or she had little control. The opposite move is also possible: "the AI suggested it." Courts and tribunals have declined the second move. Lawyers who filed a brief citing cases invented by ChatGPT were sanctioned (*Mata v. Avianca, Inc.*, 2023), and an airline that argued its chatbot was a separate entity was held liable for the chatbot's misstatement of its bereavement-fare policy (*Moffatt v. Air Canada*, 2024).

Blurred accountability is a safety problem, not only a legal one, because it weakens the incentive to check. Engineering can keep accountability sharp without assigning blame in advance. Every consequential action should record which principal proposed it, which approved it, on what evidence, and under which mode, as the execution gate's permits do (Section 1). This chronological, tamper-evident record does not decide responsibility; it makes responsibility decidable.

4.5 Sharing space with robots

Physical co-presence adds hazards that software teams lack. The 2025 revisions of ISO 10218-1 and ISO 10218-2 incorporate the collaborative-application requirements formerly in ISO/TS 15066 and replace "collaborative robot" with "collaborative application," because only an application can be designed and validated as collaborative (Association for Advancing Automation [A3], 2025; International Organization for Standardization [ISO], 2025a, 2025b). Their safeguards, such as interlocked guards, presence sensing, speed and separation monitoring, and power and force limiting, embody assumptions about where people will be and what they will do.

Human misbehavior violates those assumptions in predictable ways: entering a cell without lockout to clear a jam, muting a light curtain that trips too often, or testing a machine while it is live. The person most likely to be in the hazard zone with power switched on is often the expert sent to fix or commission the machine. In November 2023, a worker sent to check a pick-and-place robot at a vegetable-packing plant in Goseong, South Korea, was crushed against a conveyor by its arm; police were investigating both machine defects and human error (Associated Press, 2023), and the author has not found published findings. The case is cited for the exposure it shows, not for its cause: protections must remain effective during commissioning and maintenance, not only during production.

For robots, as for AI agents, stronger protections constrain the hazardous state rather than the person's compliance. Power and force limits hold whether the human followed a procedure or not; a guard interlock that

cannot be defeated without tools and a record leaves the deviation visible. The weaker protections assume the procedure will be followed.

5 Engineering responses, adapted rather than transplanted

The execution gate of Section 1 separates proposals from enforced authorization, and the response to a detected deviation can be graded from monitoring through containment, safing, recovery, and shutdown (Pattipati, 2026). Nothing in that architecture depends on whether the proposer is an AI agent or a person. This section adapts five classical responses, each designed around assumptions about people that a mixed team may violate, and closes with a worked example.

5.1 Separation of duties and two-person rules

Separation of duties keeps the power to propose, approve, execute, and record an action in different hands, so that one person cannot complete a harmful action and conceal it; the two-person rule is its real-time form. Barings lacked it (Section 2.3). After Edward Snowden, a contractor working as a system administrator, removed classified documents in 2013, the director of the National Security Agency said the agency was cutting its system administrators by about 90 percent and considering requiring two people to be present before certain data could be accessed (NBC News, 2013).

Three adaptations are needed.

- **Independence must be real.** A two-person rule does not protect against two people who agree, share a framing, or report to the same person under the same pressure; Volkswagen's defeat device passed through many hands (DOJ, 2017). The second person should have a different reporting line, see evidence the first did not prepare, and be rotated so that a stable pair does not become a single unit.
- **An AI agent can be the second person, but not automatically.** Sheridan (2016) asked whether passengers would accept computerized monitoring in place of a second pilot. An AI checker extends separation of duties cheaply to actions that would never justify a second human, but its credit depends on demonstrated independence from the human it checks (Section 4.2): it evaluates the action from source data rather than the requester's summary; repeated resubmissions are counted, rate-limited, and escalated; and for irreversible actions, the second key is a human whose approval cannot be obtained through the requester's own prompt.
- **The rule must cover the people who maintain the rule.** Administrators and those who configure the gate hold the most dangerous privileges. The gate's hard limits must lie outside the proposer's modification authority, and changes to safety configuration need the same separation as the actions they govern.

5.2 Interlocks and envelope protection

Where misbehavior must be made impossible, the protection must hold whether or not the person complies. The F-16's Automatic Ground Collision Avoidance System (Auto-GCAS) recovers the aircraft from an imminent collision with terrain without diagnosing whether the pilot is incapacitated, distracted, or saturated (Lockheed Martin, n.d.); power and force limiting bounds the energy a collaborative robot can deliver whether or not a worker followed procedure (ISO, 2025a, 2025b). These are the strongest protections against well-meaning violations, because they do not depend on the motivation that violations bypass.

Procedural barriers can often be converted into physical ones. U.S. lockout rules, at 29 C.F.R. § 1910.147(f)(3)(ii)(D), require each authorized employee in a group to affix a personal lockout device (Occupational Safety and Health Administration [OSHA], n.d.). A personal lock turns "make sure the machine is isolated" from a rule into a state the machine cannot leave while any worker's lock remains. Trapped-key interlocks and reduced-speed manual modes do the same for entry into robot cells; one vendor's manual limits tool and elbow speed to 250 mm/s in teach mode (Universal Robots, n.d.).

Protections should also be defeat-resistant. A guard that can be defeated with a spare actuator, a light curtain that can be muted indefinitely, or a driver-attention check that a weight on the steering wheel can satisfy (NHTSA, 2018) will be defeated by someone with a production target. Defeat resistance should be specified and tested like any other requirement: what tools, knowledge, and time does a bypass need, and does it leave a traceable record?

5.3 Graded authority and override policy

An override enables recovery from false-positive interventions but compromises true-positive protection, and in a team it is the main channel through which routine violations and drift enter. Override policy should therefore specify, for each protection:

- **Who may override, in which mode, and for how long.** Levels-of-automation frameworks (Parasuraman et al., 2000) define what each party may do; the override policy defines how authority returns. Overrides should expire.
- **What evidence the override must cite.** An override that records its reason converts a silent deviation into data.
- **How overrides are counted.** A rising override rate on protection is evidence of drift, of protection with too many false alarms, or of both. Either way it is a signal for engineering review, not only for supervision.
- **What does the machine do when it disagrees?** An AI teammate that disagrees with a human’s override should record the disagreement and, for hazards with short time to harm, apply the envelope protection of Section 5.2 rather than defer. Deferring without a traceable record is how a sycophantic teammate becomes a common cause.

5.4 Anomaly detection without a surveillance culture

Detecting misbehavior requires observing people, and observing people can destroy the trust and candor on which reporting depends (Section 7). The resolution is to monitor *states and actions* rather than persons, as the execution gate already does for AI agents. A gate that records every consequential action detects unauthorized actions without inferring anyone’s character. Each record r is a tuple

$$r = \langle \text{principal, mode, permit, evidence, state version, expiry} \rangle \quad (2)$$

binding who proposed and who approved an action to the evidence and authority under which it executed. Such records are still identifiable monitoring: access should be restricted to stated safety purposes, retention limited, and every inference about a person open to challenge by that person (Software Engineering Institute, 2022).

Where human state must be observed, as in fatigue or attention monitoring, the testability questions of Section 4.1 apply, and three rules keep the monitoring a safety function rather than surveillance: it serves a stated hazard, its outputs are used for intervention and reallocation rather than discipline for honest error, and its false-alarm rate is measured and published to the people monitored. Team-level patterns, such as changes in communication structure, can reveal coordination problems without individual scrutiny (Ayala et al., 2020).

The gate carries much of the protection argued for here, so it needs an assurance argument of its own. A state version recorded in a permit does not keep that state valid: the gate must revalidate the permit’s preconditions at execution and enforce them while the action runs, or admit only actions that remain safe under intervening changes. Its other failure modes are those of any safety function: wrong or incomplete constraints; compromised evidence, or evidence shared with the checks the gate arbitrates; configuration changed by an administrator, alone or in collusion (Section 5.1); and failure of the gate itself, loss of communication, or denial of service. Each should be classed as prevented, detected, tolerated, or accepted. Blocking is not automatically safe: where a protective action such as a stop, a vent, or a handover of control must occur, its failure to happen is an omission hazard. Because protective actions can themselves be hazardous in the wrong state, they need a separate, bounded authorization path, validated in advance and independent of the ordinary proposal-and-approval workflow.

5.5 Organizational design as a control

The structure of a team determines which misbehavior is possible. Levchuk et al. (2002a, 2002b) formulated organizational design as a three-phase optimization: allocating tasks to humans, agents, and robots; clustering them into decision makers' responsibilities under workload and expertise constraints; and building a decision hierarchy that minimizes coordination and balances workload. The method was extended to robust and adaptive organizations (Levchuk et al., 2004). Robustness adds redundancy in the allocation, making the organization stable against environmental perturbations and decision or processing errors at a modest cost in performance; adaptation is simplified when its triggers, including decision-maker failures, are anticipated; and an organization matched precisely to one mission "may exhibit brittle behavior in an uncertain environment" (Levchuk et al., 2004). Such designs have been tested with human teams in the Adaptive Architectures for Command and Control program: a six-person organization designed with the model outperformed a traditional six-person organization, and a four-person congruent organization performed nearly as well as the traditional one, although operators at first resisted the unfamiliar structures and needed training to work within them (MacMillan, Paley, Levchuk, & Serfaty, 2000).

Two extensions follow for misbehavior. First, separation of duties can be written as constraints in the allocation: no single decision maker, human or AI, may hold both the proposal and the authorization of a hazardous task, and no single node may be the only route of information about a hazardous state. Second, a misbehaving member is a decision-maker failure that the design should anticipate, and the robust and adaptive strategies give two ways to do so; attack-resistant organizations were taken up in this line of work (Levchuk et al., 2007), but here the attacker may be inside.

Robust design: redundancy sized for k misbehaving members. A robust organization keeps its protective functions when up to k members misbehave, without first having to detect them. Because a misbehaving member may either approve what it should block or block what it should approve, the design must tolerate both. The voting model of Section 6.2 gives the size of redundancy. Suppose each checker on a hazardous task errs with probability 0.1 in either direction and the team must accept an unsafe action with probability at most 10^{-3} while accepting at least 99% of safe ones, whichever way the k members misbehave. The table gives the smallest team and approval rule that meets both requirements.

Table 3. Smallest team and approval rule that tolerate k misbehaving members, each of whom may approve or obstruct.

Misbehaving members tolerated (k)	Smallest team, independent errors	Approval rule	Smallest team, common cause $q = 5 \times 10^{-4}$	Approval rule
0	8	5 of 8	8	5 of 8
1	10	6 of 10	12	7 of 12
2	14	8 of 14	15	9 of 15
3	16	9 of 16	17	10 of 17

Three points follow. Each additional misbehaving member costs two to four additional checkers, the organizational counterpart of the redundancy that Byzantine fault tolerance requires of computers (Lamport et al., 1982); the problem here is statistical voting rather than Byzantine agreement, but the cost is similar in kind. The common cause sets a floor that no team size can overcome. With $q = 10^{-3}$ or more, the requirement cannot be met at any size, and the remedy is diversity of evidence rather than more members (Section 6.6). Here, the sizing assumes n independent voting channels, of which up to k may be arbitrarily controlled by an adversary. Members who share a reporting line, a performance target, or the same AI tool are not independent channels; their dependence must

be modeled, for example through the common-cause term, and a pair that may collude counts as two toward k . The cost of robustness is the extra checkers and a slightly higher rate of blocked safe actions.

Adaptive design: detection, then reconfiguration. An adaptive organization keeps a leaner structure and changes it when evidence shows that a member is unreliable: it removes the member’s authority, changes the fusion rule and weights, and reallocates the member’s tasks (Section 6.3). Adaptation earns credit in the safety case only under the timing condition of Section 1, with reconfiguration as the response:

$$T_{\text{detect}} + T_{\text{reconfigure}} < T_{\text{harm}} \quad (3)$$

The detection term depends on how the member misbehaves. In the example of Section 6.3, a run-length monitor reveals a naive defector within at most 78 proposals, at a false-alarm rate of about one per 10,000 proposals. With proposals at regular intervals $1/v$, $T_{\text{detect}} \leq 78/v$; with Poisson arrivals at rate v , the expected delay from the onset of defection is at most $78/v$; for other arrival processes, the bound depends on the gaps between arrivals. A strategic defector is revealed only by independent audits: at audit rate a , the first detection comes on average at the $(1/a)^{\text{th}}$ unsafe execution, with a 95th percentile near the $(3/a)^{\text{th}}$. Adaptation is therefore acceptable only where that much harm is tolerable. Reconfiguration also runs at two speeds: revoking authority at a gate can take seconds, while reallocating tasks, retraining, and restructuring a hierarchy can take days, and an organization matched precisely to one mission has no spare capacity for a removed member’s tasks. Adaptation therefore needs some of the slack that robustness provides.

Choosing between them. The time to harm decides between robustness and adaptation. Where harm follows a misbehaving act faster than detection and reconfiguration can respond, as in a robot cell, a payment release, or a disclosure of data, the protection must be robust: sized for k misbehaving members in advance and enforced by the gate or the physics. Where harm develops slowly enough to detect and correct, as in drift in a review process or gradual gaming of a metric, adaptation is cheaper and adequate, provided detection is designed in and the harm before detection is bounded. Most organizations need both robustness and adaptation for resilience and safety risk management.

5.6 A worked example: a robot cell with human misbehavior

Consider a cell in which an industrial robot handles parts at production speed, technicians clear jams and maintain the equipment, and an AI planner schedules work and can request restarts to recover throughput. The hazard is:

H: The robot moves at a speed or force beyond collaborative limits while a person is within its hazard zone.

The basic events, each labeled with the misbehavior of Section 3 that most often produces it, are:

Table 4. Basic events for the robot-cell hazard H, with the misbehavior that most often produces each.

Event	Meaning	Typical source
E1	A person enters the hazard zone to clear a jam, maintain, or inspect. This is a demand, not a failure.	Routine work; frequent when jams are frequent.
G0	The guard interlock fails or is defeated, other than through the shared bypass.	Gaming: a spare actuator or bridged switch.
P0	Presence sensing fails to detect the person or is muted, other than through the shared bypass.	Routine violation: muting a nuisance trip.
LO	Hazardous energy is not isolated and locked out before entry: no personal lock is applied, or isolation is not verified.	Routine violation under time pressure.
C	Collaborative speed and force limits are not in effect.	Mode: the cell is in production mode.

Event	Meaning	Typical source
K	The shared maintenance bypass is engaged, disabling interlock and presence sensing together.	Optimizing violation: running production in maintenance mode.
R	Undeveloped aggregate event. Motion resumes while a person is inside, through an authority path that defeats the other barriers: a restart from a control-system stop that was not energy isolation, with safeguards overridden, or removal of isolation by someone other than the person it protects.	Authority gradient or deference: a supervisor's order, or an AI planner's restart accepted without challenge.

Two rules of the U.S. lockout standard shape these definitions. A stop issued through the control system is not energy isolation: push buttons, selector switches, and other control-circuit devices are not energy-isolating devices, and isolation must be verified and stored energy relieved before work begins. In addition, a lock is to be removed by the person who applied it (*Control of hazardous energy (lockout/tagout)*, 2025). The event R represents the ways those rules are broken under authority.

Defining G0 and P0 as failures that arise independently of the bypass separates distinct failure mechanisms: engaging K disables both protections, so K would otherwise imply G and P. The events are not mutually exclusive; G0, P0, and K may occur together. The fault tree, writing + for OR and · for AND, is:

$$H = E1 \cdot [LO \cdot C \cdot (G0 \cdot P0 + K) + R] \quad (4)$$

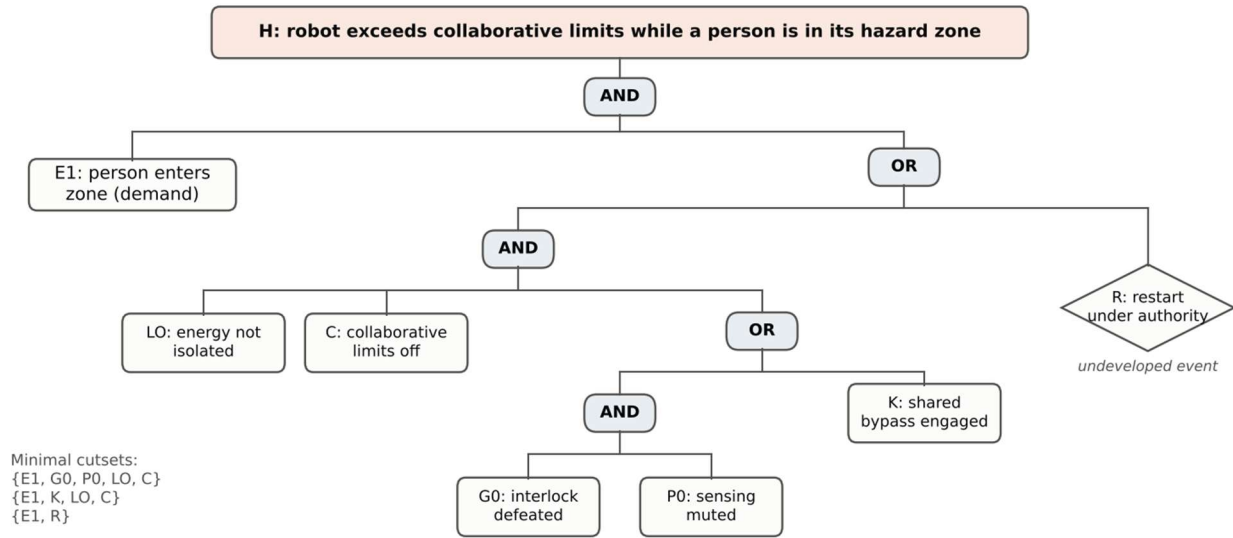


Figure 1. Fault tree for the robot-cell hazard H. The demand E1 enters every path; R is drawn as an undeveloped event.

Its minimal cutsets, shown in Figure 1, are {E1, G0, P0, LO, C}, {E1, K, LO, C}, and {E1, R}. The coverage matrix records which barriers stand in each. The barrier counts are comparable only within the developed branches: R aggregates several ways of defeating multiple protections, so its single barrier marks a singleton authority path rather than a measure of depth comparable to the others.

Table 5. Coverage matrix: barriers standing in each minimal cutset of the robot-cell fault tree.

Minimal cutset	Interlock	Presence sensing	Personal lockout	Collaborative limits	Bypass control	Restart authority	Barriers
{E1, G0, P0, LO, C}	✓	✓	✓	✓			4
{E1, K, LO, C}			✓	✓	✓		3

Minimal cutset	Interlock	Presence sensing	Personal lockout	Collaborative limits	Bypass control	Restart authority	Barriers
{E1, R}						✓	1

The matrix shows three things that the procedure manual does not.

- **Cutset size alone does not determine risk.** Four barriers stand between a jam-clearing technician and harm on the deepest path, and they are independent only if interlock, sensing, and lockout fail independently. They do not when the same production pressure motivates all three violations, and the quantification below shows that the deepest path can then dominate the risk. Pressure is a common cause across procedural barriers, as a shared specification is across the reviewers of an AI model (Pattipati, 2026).
- **The maintenance bypass turns two barriers into one.** A single action, engaging the bypass (K), disables both the interlock and presence sensing, so K is a common cause built into the cell's design. Two changes remove the hazard. If engaging the bypass forces the collaborative mode, with reduced speed and power and force limiting enforced by safety-rated functions, the robot cannot run at production speed or force while bypassed: K then excludes C, up to the failure of those functions, the cutset {E1, K, LO, C} can no longer occur, and the bypass is useless for production. If the bypass also requires a personal key and leaves a record, every use of it is visible.
- **The restart path is a singleton once a person is inside.** {E1, R} has one barrier, the restart authority, and that barrier depends on whether a technician will refuse a supervisor's order, or whether a supervisor will refuse an AI planner's restart request. The engineering response is to remove the path: restart requires that every personal lock be removed by its owner and that presence sensing reports the cell as being clear, conditions no authority can waive. The AI planner may *request* a restart; it may not *command* one.

The basic events also differ in form. Here, LO is an omission, while defeating the interlock or the sensing, engaging K, and forcing R are commissions. G0 and P0 also include equipment failures, to which the labels do not apply. The commissions can be made unavailable by design, as the points above propose. The omission cannot be blocked, only made visible or harmless. A personal lock, once applied, keeps the cell isolated, but it cannot prevent its own omission. Making the omission harmless requires entry control that makes access contingent on isolation, such as a trapped-key system in which the guard door opens only with a key released by the locked isolator (Section 5.2).

Quantifying the tree makes the first point concrete. The probabilities below are illustrative, conditional on an entry into the zone. The basic events are taken to be conditionally independent given the pressure regime. In routine operation, take 0.01 for each of G0, P0, and K, 0.05 for LO, 0.9 for C, and 0.001 for R. Under production pressure, present at one entry in five, take 0.2 for G0, 0.3 for P0, 0.5 for LO, 0.05 for K, and 0.01 for R. An analysis that averages these marginal probabilities and treats the events as independent understates the risk of the deepest path thirteenfold, because the same pressure drives G0, P0, and LO together.

Table 6. Illustrative quantification of the robot-cell fault tree, conditional on entry into the hazard zone. The deepest-path columns are the same in every row because neither fix acts on that path; in the last row it is the only path left, so the totals equal it.

Design	P(H E1), independence assumed	P(H E1), pressure as common cause	Deepest path, independence	Deepest path, common cause
As designed	5.5×10^{-3}	1.27×10^{-2}	4.1×10^{-4}	5.4×10^{-3}
Bypass forces collaborative mode	3.2×10^{-3}	8.2×10^{-3}	4.1×10^{-4}	5.4×10^{-3}
Restart also requires verified clear cell	4.1×10^{-4}	5.4×10^{-3}	4.1×10^{-4}	5.4×10^{-3}

The two architectural fixes remove the bypass and restart paths from the model. They do not remove them from the cell: the safety functions that enforce collaborative mode, verify isolation, and confirm a clear cell can fail, and a complete analysis would add those failures as basic events. What remains in the model is the deepest path, and under the common cause, it is the dominant risk. The remaining remedy is to remove the cause rather than add a barrier: redesign jam recovery so that it does not require entry and make entry contingent on verified isolation (Section 5.2).

Fault injection follows the same structure. Training drills should test whether technicians apply locks under time pressure, whether muting is detected and expires, whether the bypass enforces reduced speed, and whether a restart request from a supervisor or the planner is refused when the cell is not verified to be clear. A basic event that no training drill exercises is a safety gap that must be mitigated.

6 Team decision theory: the misbehaving member

6.1 Coupled decision rules

For conditionally independent checkers, the optimal fusion rule is a likelihood-ratio test that reduces to a weighted vote, with different weights for approvals and rejections (Chair & Varshney, 1986; Pete, Pattipati, & Kleinman, 1993a), and the checkers' own decision rules are coupled through the team goal (Tenney & Sandell, 1981). Pete, Pattipati, and Kleinman (1993b) showed for hierarchical teams that each subordinate's optimal threshold depends on its *influence* on the team decision, so that "team members coordinate to adapt their strategies to the expertise and cost structure of the team" without overt communication of strategy; the locus of jointly optimal performance forms a team receiver operating characteristic (Pete et al., 1993b; Pete, Pattipati, Kleinman, & Levchuk, 1998). This implicit adaptation is what a misbehaving member distorts.

The coupling can be written explicitly for the checking problem used below. Let H_1 denote an unsafe proposal, with prior probability π , and H_0 a safe one. Checker i observes y_i , forms the likelihood ratio $\Lambda_i(y_i) = p(y_i | H_1)/p(y_i | H_0)$, and rejects ($u_i = 1$) when $\Lambda_i \geq \gamma_i$. A fusion rule $u_0 = f(u_1, \dots, u_n)$ blocks the action when $u_0 = 1$. The team minimizes the expected cost

$$J = C_M \pi P(u_0 = 0 | H_1) + C_F (1 - \pi) P(u_0 = 1 | H_0) \quad (5)$$

where C_M is the cost of executing an unsafe action and C_F the cost of blocking a safe one. For observations that are conditionally independent given the hypothesis, the person-by-person optimal threshold of checker i is (Tenney & Sandell, 1981; Pete et al., 1993b)

$$\begin{aligned} \gamma_i^* &= \frac{(1 - \pi) C_F}{\pi C_M} \cdot \frac{\Delta F_i}{\Delta D_i}, \\ \Delta F_i &= P(u_0 = 1 | u_i = 1, H_0) - P(u_0 = 1 | u_i = 0, H_0), \\ \Delta D_i &= P(u_0 = 1 | u_i = 1, H_1) - P(u_0 = 1 | u_i = 0, H_1) \end{aligned} \quad (6)$$

Here, ΔF_i and ΔD_i are checker i 's influence on the team decision under each hypothesis, the quantities that appear in Theorem 2 of Pete et al. (1993b). Both depend on the other checkers' rules, so the conditions are coupled; they are necessary for optimality, not sufficient, and the examples below solve them numerically by optimizing the thresholds jointly for each candidate fusion rule.

Proposition 1 (pivotal protection). Let $D_i = P(u_i = 1 | H_1)$ be member i 's probability of rejecting an unsafe action and $F_i = P(u_i = 1 | H_0)$ its probability of rejecting a safe one, with ΔD_i and ΔF_i as above. For this identity, assume that the observations are conditionally independent given the hypothesis, and hold the fusion rule and the other members' rules fixed. If member i switches from its nominal rule to always approving, then

$$\begin{aligned} P_{\text{unsafe}}^{(i \text{ approves})} - P_{\text{unsafe}}^{(\text{nom})} &= D_i \Delta D_i, \\ P_{\text{blocked}}^{(\text{nom})} - P_{\text{blocked}}^{(i \text{ approves})} &= F_i \Delta F_i \end{aligned} \quad (7)$$

Proof. Given H_1 , u_i is independent of the other members' decisions, so conditioning on u_i gives $P(u_0 = 1 \mid H_1) = P(u_0 = 1 \mid u_i = 0, H_1) + D_i \Delta D_i$. Always approving sets $u_i = 0$ and removes the second term. The argument under H_0 is similar. ■

The protection a member provides is thus the product of what it detects and how pivotal its vote is, and the influence ΔD_i that sets its optimal threshold also prices its defection. The terms $D_i \Delta D_i$ are member-by-member measure of vulnerability, which the examples below compute.

Coupling is a strength when every member behaves as the design assumes. It becomes a vulnerability when one does not, because the others have adapted to a member who is no longer there.

6.2 One misbehaving member under m-of-n voting

Consider nine checkers, as in the redundancy example of Pattipati (2026), each of which wrongly approves an unsafe action with probability 0.1 and wrongly rejects a safe one with probability 0.1. Suppose k members are insider threats who always approve. The table gives the probability that an unsafe action is accepted for several voting rules, and the probability that a safe action is accepted when one member obstructs by always rejecting.

Table 7. Probability of accepting an unsafe action, and of accepting a safe one, under m -of-9 voting with misbehaving members ($k=1,2$).

Approvals required (m of 9)	Unsafe accepted, no insider	One insider, $k=1$	Two insiders. $k=2$	Common cause $q = 0.001$, one insider	Safe accepted, no obstructer	Safe accepted, one obstructer
5 (majority)	8.9×10^{-4}	5.0×10^{-3}	2.6×10^{-2}	6.0×10^{-3}	99.91%	99.50%
6	6.4×10^{-5}	4.3×10^{-4}	2.7×10^{-3}	1.4×10^{-3}	99.17%	96.19%
7	3.0×10^{-6}	2.3×10^{-5}	1.8×10^{-4}	1.0×10^{-3}	94.70%	81.31%
8	8.2×10^{-8}	7.3×10^{-7}	6.4×10^{-6}	1.0×10^{-3}	77.48%	43.05%
9 (unanimity)	1.0×10^{-9}	1.0×10^{-8}	1.0×10^{-7}	1.0×10^{-3}	38.74%	0%

Three observations follow.

- **Each insider costs roughly a factor of five to ten.** Under majority voting, one insider multiplies the probability of accepting an unsafe action by about 5.6, and two by about 29. Under unanimity, each insider raises the probability by exactly a factor of $1/p = 10$, because the insider's vote is simply removed.
- **Insiders and obstructers pull in opposite directions.** Raising m defends against insiders who approve but hands a veto to an obstructer: under unanimity, one member who always rejects stops every action. A misbehaving member can harm the team by blocking as well as by permitting, and a rule chosen against one is weak against the other.
- **The common cause still sets the floor.** With $q = 0.001$, an insider adds little once m is seven or more, because the team already fails with probability near q (Section 6.6).

Insiders who choose their votes knowing the others' votes are worse still. In distributed detection with "Byzantine" sensors that report strategically, the fusion center becomes blind, at least asymptotically, once such sensors make up half of the network (Marano, Matta, & Tong, 2009); below that fraction, the defense is to identify and down-weight them, which the weighted-vote form of the optimal rule permits. But a weight can be set to zero only for a member known to be unreliable. The trusted insider keeps the full weight earned by a record of good behavior.

6.3 Trust concentration under coupling

Coupling adds a further effect that the voting table does not show. Consider three checkers who each observe a noisy signal of whether the proposed action is unsafe. Each checker’s observation is Gaussian with unit variance, its mean shifted by d for unsafe actions; one in twenty proposals is unsafe; and executing an unsafe action costs 100 times as much as blocking a safe one. The thresholds and the fusion rule are optimized jointly, in the person-by-person sense of Pete et al. (1993b) and Tenney and Sandell (1981), over all monotone fusion rules of the three decisions, not only over vote counts. The table varies the discriminability d_3 of the third checker, the team’s expert, while the other two keep $d = 2$.

Table 8. Trust concentration in a three-checker team: performance when all behave, when the expert defects undetected, and after re-optimization. The re-optimized columns are the same in every row because, once the expert is removed, the team is the same two checkers with $d = 2$.

Expert’s discriminability d_3	Optimal fusion rule	Unsafe executes, all behave	Safe blocked, all behave	Unsafe executes, expert defects undetected	Safe blocked, expert defects undetected	Unsafe executes, team re-optimized	Safe blocked, team re-optimized
2 (no expert)	Block on 2 of 3 rejections	2.0%	17.3%	16.3%	7.0%	2.5%	29.1%
3	Block on 2 of 3 rejections	1.4%	10.8%	20.1%	6.0%	2.5%	29.1%
4	Block if the expert rejects, or both others do	0.6%	4.6%	41.3%	1.0%	2.5%	29.1%

A stronger expert makes the team better while everyone behaves, and worse when the expert defects. With $d_3 = 4$, the optimal rule is $u_0 = u_3 \vee (u_1 \wedge u_2)$: the expert’s rejection blocks the action alone, while the two weaker checkers, each at a likelihood-ratio threshold of about 1.7, can block only together. Dependence on the expert thus arises through the structure of the fusion rule as well as through the thresholds. If the expert then approves everything without the team knowing, the probability that an unsafe action executes rises from 0.6% to 41.3%, while the rate of blocking safe actions falls from 4.6% to 1.0%, because every block must now come from the backup pair. The team appears to perform better precisely when it has lost its main protection. If the defection is known, re-optimizing recovers most of the protection, with 2.5% of unsafe actions executing rather than 41.3%, at the cost of blocking 29.1% of safe ones; the two remaining checkers then reject at a threshold of about 1, and either of the two can reject to block.

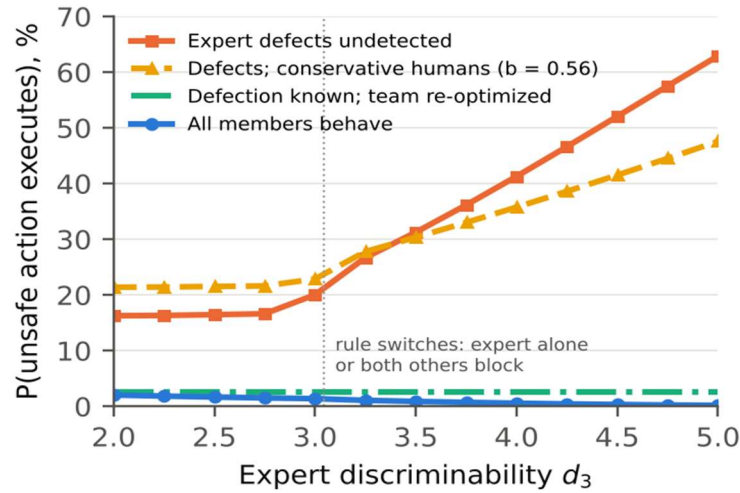


Figure 2. Probability that an unsafe action executes as the expert’s discriminability d_3 grows, with the two other checkers fixed at $d = 2$ and the fusion rule optimized over all monotone rules. Near $d_3 \approx 3.05$ the optimal rule switches from two-of-three to the expert alone or both others together, and exposure to undetected defection then grows steadily. The dashed curve applies the conservatism fitted by Pete et al. (1993b) to the two weaker checkers (Supplement S1).

Figure 2 extends the table across a range of expertise. Beyond the point where the optimal rule switches, every gain in the expert’s discriminability buys a small nominal improvement and a large increase in exposure: at $d_3 = 5$, an undetected defection lets 63% of unsafe actions through.

Can design alone hedge against the defection? One way to ask is to minimize a weighted cost $(1 - \epsilon) J_{\text{nominal}} + \epsilon \cdot J_{\text{defect}}$, where J_{defect} averages the cost over the three possible single defectors and ϵ is the probability that a defection by some member is undetected. With $d_3 = 4$, $\epsilon = 0.01$ lowers exposure to the expert’s defection from 41.3% to 37.9% at a 0.4% increase in nominal cost; $\epsilon = 0.1$ lowers it to 24.3% at a 15% increase. Averaging over the three defectors can dilute protection against the one that matters most, so a worst-case formulation is also informative: minimize J_{nominal} subject to the condition that, whichever single member defects undetected by approving everything, the probability that an unsafe action executes is at most η . The constraint protects against unsafe approval under a single defection; it does not bound the denial of service caused by a member who always rejects, which Section 6.2 treats separately. The table below (Table 9) traces the trade-off. The fusion rules are enumerated exhaustively, but the thresholds for each rule are found numerically, by Nelder–Mead searches from several starting points, so the results are numerical optima rather than certified global ones. Because Nelder–Mead is unconstrained, the exposure bound was imposed as a quadratic penalty and checked after optimization; an independent run with a constrained solver (sequential quadratic programming, with the bound as explicit inequality constraints) reproduced every row of the table. The calculation scripts are available as supplementary material.

Table 9. Hedging by design at $d_3 = 4$: nominal cost and performance when the worst-case exposure to any single member’s undetected defection is bounded by η .

Worst-case exposure bound η	Increase in nominal cost	Unsafe executes, all behave	Safe blocked, all behave	Optimal fusion rule
None (41.3%)	—	0.58%	4.6%	Expert alone, or both others
30%	5%	0.51%	5.3%	Expert alone, or both others
20%	27%	0.43%	7.4%	Expert alone, or both others

Worst-case exposure bound η	Increase in nominal cost	Unsafe executes, all behave	Safe blocked, all behave	Optimal fusion rule
10%	103%	0.31%	13.9%	Any one rejection
5%	194%	0.21%	21.3%	Any one rejection

Design buys protection at a price. Halving exposure raises nominal cost by about a quarter; reducing it to 10% roughly doubles the cost and returns the team to blocking on any one rejection. The price rises steeply because no setting of thresholds recovers what the expert contributed. Comparable protection against unsafe execution is available without identifying the defector, as the last rows of Table 9 show, but a conservative design pays its price all the time; knowing which member has defected allows targeted reconfiguration, such as the re-optimized column of the earlier table, and pays the price only when it is needed. Threshold design is one side of the trade-off; detection and reallocation are the other.

Detection depends on how the member defects. In the $d_3 = 4$ design, the expert rejects about 8.3% of all proposals: 98.6% of the one in twenty that are unsafe, and 3.6% of the safe ones. A naive defector that approves everything can be revealed by a run-length monitor that raises an alarm after L consecutive approvals and then resets. If p is the nominal probability that the expert approves, the average number of proposals between false alarms is $ARL_0 = (1 - p^L)/[(1 - p)p^L]$. With $L = 78$, that average is about 10,000 proposals, and a naive defector is caught within 78 proposals; a shorter run, such as 34, would raise a false alarm about once every 220 proposals. These figures assume independent proposals with a stable unsafe fraction. A strategic defector keeps its rejection rate by rejecting safe-looking proposals and approves only the unsafe actions it wants to pass; its rates reveal nothing, and the defense is independent audit. If a fraction a of executed actions is verified at random against independent evidence, and an audited unsafe action is always recognized, the number of unsafe executions up to and including the first detected one is geometric with mean $1/a$: 10 at a 10% audit rate and 100 at 1%, with 95th percentiles of 29 and 299. An audit that recognizes an unsafe action only with probability σ replaces a by $a\sigma$, and audit latency adds the unsafe executions that occur while an audit is pending. An audited unsafe execution shows that the team failed, not which member failed, so the first such finding should trigger precautionary containment, such as re-review of recent approvals and a temporarily stricter fusion rule; removing a member's authority requires evidence about that member, such as its votes on the audited actions. The audit rate is thus a design parameter; it is chosen from the harm the team can tolerate before a trusted member's misbehavior is found.

The example is illustrative: Gaussian observations, a fixed prior, and a 100-to-1 cost ratio. It demonstrates a mechanism in a stated model rather than a general law. The coupled optimality conditions establish only that each member's best threshold depends on the others' rules; that rising expertise concentrates protection in the expert and raises the cost of its defection is a numerical finding for this model, and whether human teams behave this way is an empirical question (Section 6.5). In hybrid teams, the most reliable member is often the AI checker, and rational reliance on it can resemble the complacency of Section 3. Teamwork research counts mutual performance monitoring and backup behavior, resting on mutual trust, among the core components of effective teams (Salas, Sims, & Burke, 2005); the result here shows their cost, because backup that rests on trust in the most reliable member makes that member's undetected failure the costliest one. Human-automation research finds the same pattern empirically: in a meta-analysis of 18 experiments, higher degrees of automation improved routine performance but worsened performance when the automation failed (Onnasch, Wickens, Li, & Manzey, 2014), a trade-off Wickens summarizes as "the higher the tree, the harder it falls." Sheridan (2019) separates the properties that make automation trustworthy, such as reliability, robustness, and explication of intention, from two attributes of the user, familiarity and dependence. The user attributes are the ones that grow as a team leans on its most reliable member, irrespective of whether the member remains trustworthy or not, and a reliable member that does not explicate its intentions gives the team no way to notice when they change. Two design rules follow: limit how far any member's reliability may relax the others' thresholds, and monitor the trusted member's agreement with independent evidence, because its reputation is what protects its misbehavior.

6.4 Pivotal dependence and its sensitivity

Proposition 1 accounts for the defection results of Section 6.3. In the $d_3 = 4$ design, the expert has $D_3 = 0.986$ and $\Delta D_3 = 0.413$, so its protection is 40.7 percentage points, exactly the rise from 0.6% to 41.3%. Each backup detects less, $D = 0.766$, but its influence is only $\Delta D = 0.011$, so its protection is 0.8 points. Partial defection interpolates. If a member withholds each of its rejections independently with probability w , and the rules stay fixed, then $P_{\text{unsafe}}(w) = P_{\text{unsafe}}(0) + w[P_{\text{unsafe}}(1) - P_{\text{unsafe}}(0)]$. An expert that withholds one rejection in ten raises unsafe execution from 0.6% to 4.7%, while its rejection rate falls only from 8.3% to 7.5%, a change that a rate monitor would be slow to notice. The interpolation also holds with correlated evidence, provided the withholding is random and the rules fixed; only its factorization into $D_i \Delta D_i$ needs conditional independence. A defector that withholds selectively, choosing the rejections that matter most, need not follow this line.

The example's other assumptions were then varied in turn (Figure 3). For correlated evidence, each observation under hypothesis H_h , $h \in \{0, 1\}$, is $Y_i = h d_i + \sqrt{\rho} Z_0 + \sqrt{1 - \rho} Z_i$, with independent standard normal Z_0, Z_1, Z_2, Z_3 , so that every pair of observations has correlation ρ under both hypotheses. With correlated observations, threshold rules are no longer guaranteed to be optimal: depending on the signals and the correlation, a local rule may need two thresholds, or an observation may be best ignored (Willett, Swaszek, & Blum, 2000). Tang, Pattipati, and Kleinman (1992) derived the necessary conditions for person-by-person optimal decision rules and an iterative algorithm that converges to them. Here, at each reported optimum, the best response of every checker whose vote can change the team decision, computed over all local rules with the other rules held fixed, is a single threshold located at the threshold in use; the reported designs are therefore person-by-person optimal for their fusion rules. They remain numerical optima, because person-by-person optimality does not guarantee a global optimum. The probabilities were computed by 32-node Gauss–Hermite quadrature over Z_0 . Compared with a 96-node rule over all vote-pattern probabilities, under both hypotheses, for $\rho = 0.25$ and 0.5 , every d_3 in Figure 3, and a grid of thresholds spanning the optima, the largest absolute discrepancy was 1.4×10^{-10} . The aim is not to show that concentration always occurs, but to locate when it occurs, how much exposure it creates, and what protection costs.

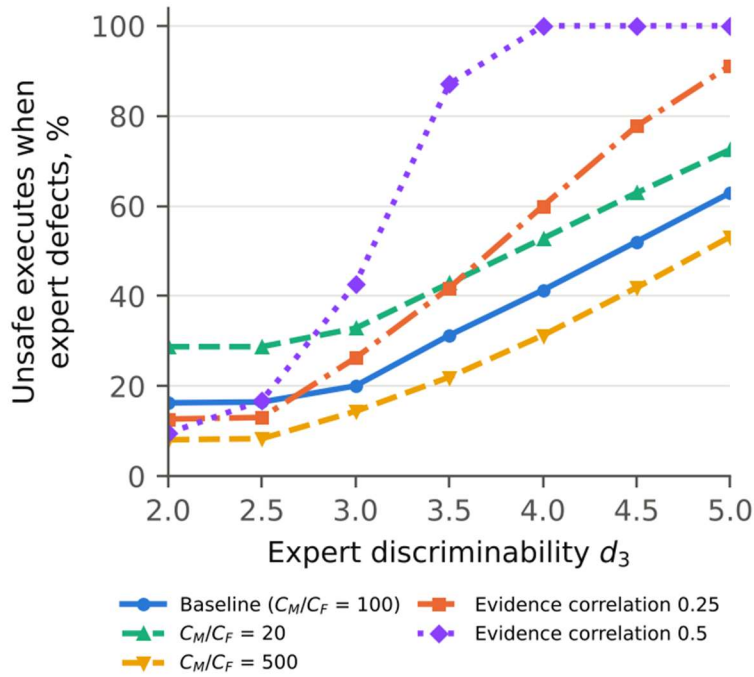


Figure 3. Probability that an unsafe action executes when the expert defects undetected, for numerically optimized scalar-threshold local rules and monotone fusion rules, as the cost ratio and the correlation ρ of the checkers' evidence vary. The other checkers keep $d = 2$, and one proposal in twenty is unsafe.

- **Cost ratio.** When executing an unsafe action is costlier relative to blocking a safe one, the nominal design is more cautious, the switch to relying on the expert comes earlier, and exposure at a given level of expertise is lower: at $d_3 = 4$, 31% for a ratio of 500, 41% for 100, and 53% for 20. Protection is also cheaper. Capping the worst-case exposure at 20% raises the expected decision cost by 6% at a ratio of 500, 27% at 100, and 87% at 20. A team that tolerates blocked work poorly leans hardest on its expert and pays most to hedge.
- **Correlated evidence.** When the checkers' observations share a common component, the backups add less independent information, and the optimal design leans further on the expert. With $\rho = 0.25$, exposure at $d_3 = 4$ rises from 41% to 60%. With $\rho = 0.5$ and $d_3 \geq 4$, the best design found is numerically equivalent to expert-only decision making: its cost matches that of the rule $u_0 = u_3$ to six digits, so the backup votes add no measurable value. Under the expert-only rule, the expert's undetected defection lets every unsafe action through. Within the searched class, dependence can eliminate the nominal value of keeping backup votes. Capping exposure at 20% costs 39% more at $\rho = 0.25$ and 59% more at $\rho = 0.5$. Shared evidence is thus a second route to trust concentration, alongside expertise. Proposition 1's factorization assumes conditional independence and does not apply to these cases.

The mechanism is therefore not an artifact of one parameter choice. It strengthens when evidence is shared and when blocking is costly and weakens when unsafe execution dominates the cost; where the evidence is conditionally independent, its size is given exactly by the expert's pivotality.

6.5 Human adaptation cannot be assumed

The re-optimized columns assume that the team learns of the defection and adjusts. The experimental evidence gives no reason to assume that human teams will do so unaided. In laboratory teams, Pete et al. (1993b) found that human decision rules were of the threshold type, but that the resulting ROC curves were consistently below the normative ones, that humans were conservative (their applied thresholds were not as extreme as the normative model predicted, with a fitted exponent of 0.56 relating applied to optimal thresholds on a logarithmic scale), and that the information structure at the primary decision maker had no significant influence on the

subordinates' behavior. Conservatism is not peculiar to that experiment: Healy and Kubovy (1981) found the same log-linear relation between applied and optimal thresholds, with slopes of 0.25 to 0.35, indicating stronger conservatism still, and Maloney and Thomas (1991) traced observed conservatism in signal detection to subjects' assumptions about the stimulus distributions. The normative model predicts that subordinates should adapt to the expertise of the member who fuses their opinions; the human subordinates did not. The experiments varied the information structure at the primary decision maker; they did not test adaptation after evidence that a teammate had defected. They show that implicit adaptation of this kind cannot be assumed, not that it never occurs. Adaptation can, however, be trained: teams trained to adjust their coordination strategies under stress performed better than untrained teams (Entin & Serfaty, 1999), and models of team adaptation begin with recognizing that the situation has changed (Burke, Stagl, Salas, Pierce, & Kendall, 2006). Both point to the same requirement: the response to a misbehaving member must be detected, designed, and rehearsed.

The adaptation that the coupled optimality conditions prescribe when a member is found to be unreliable, such as changing the fusion rule, re-weighting votes, and tightening thresholds, therefore cannot be left to implicit coordination among people. It must be computed and enforced: by the gate that holds the fusion rule, by decision aids that display each member's current weight, and by an organizational procedure that removes a suspect member's authority (Section 5.5).

Conservatism also has a less obvious effect. Transported to the two weaker checkers as a sensitivity illustration, the fitted exponent lowers exposure to the expert's undetected defection from 41.3% to 35.8% at about 1% higher nominal cost but raises exposure below the point where the fusion rule switches (Figure 2; Supplement S1). Human conservatism is thus an unmanaged hedge against trust concentration, effective in some regimes and not in others; Recommendation 5 of Section 8 provides the same protection deliberately.

6.6 The common-cause floor for humans

The table of Section 6.2 shows that a common cause, not the voting threshold, sets the floor on accepting an unsafe action. With q the probability of a latent event that makes every check fail together (Section 1), the probability that all checks miss is $P_{\text{miss}} = q + (1 - q)P_{\text{ind}}$, where P_{ind} is the joint miss probability without that event. P_{miss} is not, in general, the probability that the team fails, which depends on the deployed fusion rule f : $P_{\text{team}} = P(f(u) = 0 \mid H_1)$. Under the rule of Section 6.3, an unsafe action executes when the expert and either backup approve; all three need not miss, so testing only how often every check exhibits a miss can understate system failure probability substantially. For human members, the common causes are familiar: the same training, procedures, production pressure, shift schedule, and increasingly the same AI tool. Drift makes the floor rise: normalization of deviance is a process in which $q(t)$ grows while each check appears to perform as before.

Three measures follow. Diversify the evidence, not just the reviewers: a second reviewer from the same team, using the same tool, adds little against a shared cause. Measure joint misses directly, by injecting realistic deviations into reviews and training drills and recording the complete pattern of votes, so that both P_{miss} and P_{team} under the deployed rule can be estimated. The evidence required is substantial: with no failure of the counted kind in M representative, independent trials, the one-sided 95% upper bound on its probability is $1 - 0.05^{(1/M)} \approx 3/M$, so supporting a value below 10^{-3} takes about 3,000 failure-free trials. In addition, track leading indicators of drift, such as override rates, muting durations, accepted anomalies, and near-miss reports, which indicate changes in common-cause risk rather than measure q (Section 7).

7 Culture as a control

In a team with human members, culture is a control, because it decides whether deviations are reported and therefore whether drift can be seen at all. Reason (1997) argued that a safety culture depends on a reporting culture, and that people report only where the culture is *just*: honest errors can be reported without fear of punishment, reckless conduct and deliberate violations still carry consequences, and everyone knows where the line lies. Marx (2001) gave the line an operational form, arguing that the organization's response should follow the

behavior, not the severity of the outcome; Dekker (2007) added that who draws the line, and by what process, determines whether people trust it enough to report.

The reason is observability. Near-miss reports and acknowledged deviations are the sensors for rising common-cause risk (Section 6.6), and punishing honest error degrades those sensors' coverage: drift does not stop when reporting stops; it becomes invisible. If deviations occur at rate λ and each is reported with probability r , the expected time to accumulate the N reports needed to recognize a trend is $N/(\lambda r)$, so halving the reporting probability doubles the time for which drift goes unrecognized.

Field evidence supports this. In a study of eight hospital nursing units, Edmondson (1996) found that units with stronger leadership and better working relationships had higher detected rates of medication errors and proposed that detected rates depend on willingness to report as well as on the errors themselves. She later named the enabling condition *psychological safety*: "a shared belief held by members of a team that the team is safe for interpersonal risk taking" (Edmondson, 1999). A reporting rate therefore measures the culture as much as the system, and a falling rate of reported errors can be a warning rather than an improvement. Because omissions are the least likely to be reported (Section 3), voluntary reporting needs a partner that depends on no one's initiative: the gate's reconciliation of what was required against what was logged.

Aviation offers the most developed precedent. The Aviation Safety Reporting System, established by NASA and the Federal Aviation Administration in 1975 and administered by NASA, receives voluntary reports, removes identifying information, and returns proof of submission to the reporter; under FAA Advisory Circular 00-46F, the FAA waives civil penalties and certificate suspension for an inadvertent violation reported within ten days, subject to stated exclusions. The program received about 121,000 reports in 2025 and more than 2.3 million since 1976 (NASA Aviation Safety Reporting System [ASRS], n.d.).

Four adaptations carry the precedent into human–AI–robot teams.

- **Separate safety telemetry from discipline.** Gate logs, override records, and AI-detected anomalies are safety data. Their use in discipline should be limited to reckless conduct and malicious acts, defined in advance. A knowing violation is not by itself grounds for discipline: many are locally rational responses to procedures that cannot be followed as written, or to competing demands the organization created (Section 2).
- **Treat disagreement between human and machine as reportable, not as fault.** When a human overrides an AI recommendation or an AI flags a human action, the event is evidence about the protection, the procedure, or the person, in unknown proportions. Recording it without presumption of fault is what makes the proportions estimable.
- **Hold the organization accountable for the conditions it creates.** The NTSB cited Uber's oversight of operators and its lack of mechanisms to address automation complacency, not only the operator's distraction (NTSB, 2019). The moral crumple zone (Elish, 2019) is avoided only if investigations ask why a violation was likely, not only who committed it.
- **Keep accountability for deliberate harm.** Fairness to honest error is not an amnesty. Separation of duties, protected evidence, and prosecution remain the response to sabotage and fraud. The point is to keep those cases rare enough, and the boundary clear enough, that the reporting system for everything else stays trusted.

8 A path forward

The recommendations extend those made for AI agents in Pattipati (2026) to the human members of hybrid teams. Each states a requirement and an acceptance criterion.

1. **Write a fault model for humans.** For each safety-relevant role, state which misbehavior (hazard) the safety case tolerates, detects, or prevents, including violations, drift, and deliberate harm; credit training for reducing mistakes, not for excluding violations. *Acceptance*: every role has a fault model, and every assumption of compliance is stated and justified.

2. **Key protections to hazardous states and make omissions visible.** Block errors of commission with interlocks and authority held at a gate; expose errors of omission with forcing functions, watchdogs, and expiring authority; make entry contingent on verified isolation. *Acceptance:* the tools, knowledge, and time needed to bypass each protection are specified and tested, and every bypass leaves a traceable record.
3. **Make independence and authority real.** Give the second key an independent reporting line and independent evidence, rotate pairs, and extend separation of duties to those who configure safety functions and the gate. Machines and people may request; the gate decides. *Acceptance:* no minimal cutset of a hazard is a single human or AI authority (“singleton”), and the sycophancy coupling of each AI checker is measured by red-team requests (Section 4.2).
4. **Count resubmissions, overrides, and joint misses.** Rate-limit and escalate repeated requests to an AI checker; require overrides to cite evidence and to expire; inject realistic deviations into reviews and drills. *Acceptance:* override and resubmission rates are tracked, vote patterns are recorded in full, and every estimate of joint-miss or team-failure probability states which event it counts, its number of trials, and its upper confidence bound (Section 6.6).
5. **Bound trust concentration and engineer the response to a misbehaving member.** Limit how far any member’s reliability, human or AI, may relax the others’ thresholds; monitor the most trusted member against independent evidence; and specify in advance how a suspect member’s authority is removed, how the fusion rule and weights change, and how its tasks are reallocated. *Acceptance:* the worst-case exposure to any single member’s undetected defection is computed and bounded (Section 6.3); the audit rate a is set from the harm tolerable before detection, about $1/a$ unsafe executions on average and $3/a$ at the 95th percentile, adjusted for audit sensitivity and latency; the first audited unsafe execution triggers precautionary containment; and for each hazard, either detection and reconfiguration are shown to finish before harm, or the protection is sized in advance for k misbehaving members, with shared reporting lines, incentives, and tools modeled as dependence (Section 5.5).
6. **Monitor states and actions, keep accountability decidable, and protect privacy.** Record for every consequential action who proposed it, who approved it, on what evidence, and in which mode; tie any monitoring of human state to a stated hazard and publish its false-alarm rate. *Acceptance:* records are accessible only for stated safety purposes, retention is limited, and every inference about a person can be challenged by that person.
7. **Build a just reporting culture and measure it.** Define in advance the conduct that warrants discipline and investigate serious incidents independently of the line organization. *Acceptance:* reporting rates, including reports of omissions, are measured as a property of the safety system, and a falling rate triggers review rather than celebration.

9 Limitations

The paper’s arguments rest on three kinds of support, which carry different weight. The numerical results of Sections 4.2, 5.5, 5.6, and 6 are model results. They follow from stated assumptions, such as binomial or Gaussian observations, conditional independence within a regime, and fixed priors, costs, and probabilities, and they demonstrate mechanisms rather than estimate rates. The optimization of Section 6 covers scalar-threshold rules for each checker and all monotone fusion rules for three checkers, and its sensitivity study varies the cost ratio and the correlation of evidence but not the team size or the prior. The incident interpretations of Sections 2–4 and Supplement S2 are the author’s engineering readings of a purposive set of cases; they are not evidence of prevalence, and the mapped responses are not shown to have been effective. The empirical findings cited, from reader studies, observational studies, laboratory team experiments, and field studies of reporting, come from settings that differ from human–AI–robot teams, and their transport to those teams is an assumption. Finally, the robot-cell model omits failures of the enforcement functions it recommends, and the execution gate needs an assurance argument of its own (Section 5.4). Each limit points to work that would test the framework: field measurement of joint misses and of sycophancy coupling, experiments on how human teams adapt after a

teammate's defection, for which the human-in-the-loop simulations used to test designed organizations offer a ready testbed (MacMillan et al., 2000), and analyses of gate failures in deployed systems.

10 Conclusion

Well-trained teams fail differently from novice ones. Their members break rules with skill and good intentions, drift together toward the edge of safe operation, and learn where the checks are. AI agents and robots make monitoring mutual, add a teammate that may agree when it should challenge, and multiply transfers of authority. The old responses, separation of duties, interlocks, graded authority, organizational design, and just culture, still apply, provided their assumptions about people are made explicit.

Team decision theory adds the paper's central warning. Nominal optimization can concentrate protection in one member, so that its undetected deviation causes disproportionate harm: in the stated model, the expert supplies 40.7 percentage points of protection against about 0.8 for each backup, and shared evidence or costly blocking strengthens the concentration. Robust design can bound that exposure at a measurable cost, and detection and engineered reallocation must do the rest, because human teams cannot be assumed to re-tune on their own. The engineering task is the same for people as for AI agents: state the hazards, find the hazard paths, decide what to tolerate, detect, and prevent, and justify the residual risk with evidence.

References

- Abrams, M., & Weiss, J. (2008). *Malicious control system cyber security attack case study: Maroochy Water Services, Australia*. MITRE Corporation. <https://www.mitre.org/news-insights/publication/malicious-control-system-cyber-security-attack-case-study-maroochy-0>
- Amalberti, R. (2001). The paradoxes of almost totally safe transportation systems. *Safety Science*, 37(2–3), 109–126.
- Associated Press. (2023, November 9). An industrial robot crushed a worker to death at a vegetable packing plant in South Korea. As published by CBS News. <https://www.cbsnews.com/news/industrial-robot-crushes-worker-dead-south-korea>
- Association for Advancing Automation. (2025, March 20). Updated ISO 10218: Answers to frequently asked questions. <https://www.automate.org/robotics/blogs/updated-iso-10218-faq>
- Ayala, D. F. M., Balasingam, B., McComb, S., & Pattipati, K. R. (2020). Markov modeling and analysis of team communication. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 50(4), 1230–1241. <https://doi.org/10.1109/TSMC.2017.2748985>
- Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6), 775–779.
- Board of Banking Supervision. (1995). *Report of the Board of Banking Supervision inquiry into the circumstances of the collapse of Barings*. Her Majesty's Stationery Office.
- Budzyń, K., Romańczyk, M., Kitala, D., Kołodziej, P., Bugajski, M., Adami, H. O., Blom, J., Buszkiewicz, M., Halvorsen, N., Hassan, C., et al. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: A multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, 10(10), 896–903. [https://doi.org/10.1016/S2468-1253\(25\)00133-5](https://doi.org/10.1016/S2468-1253(25)00133-5)
- Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile. (2016). *Final report: Accident on 24 March 2015 at Prads-Haute-Bléone to the Airbus A320-211 D-AIPX operated by Germanwings*.
- Burke, C. S., Stagl, K. C., Salas, E., Pierce, L., & Kendall, D. (2006). Understanding team adaptation: A conceptual analysis and model. *Journal of Applied Psychology*, 91(6), 1189–1207. <https://doi.org/10.1037/0021-9010.91.6.1189>
- Chair, Z., & Varshney, P. K. (1986). Optimal data fusion in multiple sensor detection systems. *IEEE Transactions on Aerospace and Electronic Systems*, AES-22(1), 98–101.
- Columbia Accident Investigation Board. (2003). *Report, Volume I*. National Aeronautics and Space Administration and U.S. Government Printing Office.
- Control of hazardous energy (lockout/tagout)*, 29 C.F.R. § 1910.147 (2025).
- Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, 8(4), 377–383.
- Dekker, S. (2007). *Just culture: Balancing safety and accountability*. Ashgate.
- Dekker, S. (2011). *Drift into failure: From hunting broken components to understanding complex systems*. Ashgate.
- Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., Mähringer-Kunz, A., Püsken, M., Baeßler, B., Sauer, S., Maintz, D., & Pinto dos Santos, D. (2023). Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology*, 307(4), e222176. <https://doi.org/10.1148/radiol.222176>
- Edmondson, A. C. (1996). Learning from mistakes is easier said than done: Group and organizational influences on the detection and correction of human error. *Journal of Applied Behavioral Science*, 32(1), 5–28. <https://doi.org/10.1177/0021886396321001>

- Edmondson, A. C. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, 44(2), 350–383.
- Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human–robot interaction. *Engaging Science, Technology, and Society*, 5, 40–60. <https://doi.org/10.17351/ests2019.260>
- Entin, E. E., & Serfaty, D. (1999). Adaptive team coordination. *Human Factors*, 41(2), 312–325. <https://doi.org/10.1518/001872099779591196>
- Healy, A. F., & Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory*, 7(5), 344–354.
- Hughes, A. M., Gregory, M. E., Joseph, D. L., Sonesh, S. C., & Marlow, S. L. (2016). Saving lives: A meta-analysis of team training in healthcare. *Journal of Applied Psychology*, 101(9), 1266–1304. <https://doi.org/10.1037/apl0000120>
- Idaho National Laboratory. (n.d.). *CyOTE case study: Maroochy Shire* (Cybersecurity for the Operational Technology Environment). U.S. Department of Energy. https://cyote.inl.gov/content/uploads/24/2025/12/CyOTE-Case-Study_Maroochy.pdf
- IEEE Spectrum. (2019, November). NTSB investigation into deadly Uber self-driving car crash reveals lax attitude toward safety. <https://spectrum.ieee.org/ntsb-investigation-into-deadly-uber-selfdriving-car-crash-reveals-lax-attitude-toward-safety>
- International Organization for Standardization. (2025a). *Robotics — Safety requirements — Part 1: Industrial robots* (ISO 10218-1:2025).
- International Organization for Standardization. (2025b). *Robotics — Safety requirements — Part 2: Industrial robot applications and robot cells* (ISO 10218-2:2025).
- Kohavi, R., Tang, D., & Xu, Y. (2020). *Trustworthy online controlled experiments: A practical guide to A/B testing*. Cambridge University Press.
- Lamport, L., Shostak, R., & Pease, M. (1982). The Byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3), 382–401.
- Lawton, R. (1998). Not working to rule: Understanding procedural violations at work. *Safety Science*, 28(2), 77–95.
- Levchuk, G. M., Levchuk, Y. N., Luo, J., Pattipati, K. R., & Kleinman, D. L. (2002a). Normative design of organizations—Part I: Mission planning. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, 32(3), 346–359.
- Levchuk, G. M., Levchuk, Y. N., Luo, J., Pattipati, K. R., & Kleinman, D. L. (2002b). Normative design of organizations—Part II: Organizational structure. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, 32(3), 360–375.
- Levchuk, G. M., Levchuk, Y. N., Meirina, C., Pattipati, K. R., & Kleinman, D. L. (2004). Normative design of project-based organizations—Part III: Modeling congruent, robust, and adaptive organizations. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, 34(3), 337–350.
- Levchuk, Y. N., Levchuk, G. M., Meirina, C., Ruan, S., Pattipati, K. R., & Kleinman, D. L. (2007). Organizational armor: Design of attack-resistant organizations. In A. Kott (Ed.), *Information warfare and organizational decision-making*. Artech House.
- Lockheed Martin. (n.d.). *Auto GCAS: Collision avoidance system*. <https://www.lockheedmartin.com/en-us/products/autogcas.html>
- MacMillan, J., Paley, M. J., Levchuk, Y. N., & Serfaty, D. (2000). Model-based team design. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 44(6), 741–744. <https://doi.org/10.1177/154193120004400654>
- Maloney, L. T., & Thomas, E. A. C. (1991). Distributional assumptions and observed conservatism in the theory of signal detectability. *Journal of Mathematical Psychology*, 35(4), 443–470.
- Mannaru, P., Balasingam, B., Pattipati, K. R., Sibley, C., & Coyne, J. (2016a). Cognitive context detection in UAS operators using eye-gaze patterns on computer screens. In *Proceedings of SPIE 9851, Next-Generation Analyst IV*. <https://doi.org/10.1117/12.2224184>
- Mannaru, P., Balasingam, B., Pattipati, K. R., Sibley, C., & Coyne, J. (2016b). Cognitive context detection using pupillary measurements. In *Proceedings of SPIE 9851, Next-Generation Analyst IV*, 98510Q. <https://doi.org/10.1117/12.2224176>
- Marano, S., Matta, V., & Tong, L. (2009). Distributed detection in the presence of Byzantine attacks. *IEEE Transactions on Signal Processing*, 57(1), 16–29.
- Marx, D. (2001). *Patient safety and the “just culture”: A primer for health care executives*. Trustees of Columbia University.
- Mata v. Avianca, Inc.*, 678 F. Supp. 3d 443 (S.D.N.Y. 2023).
- Mishra, M., Mannaru, P., Sidoti, D., Bienkowski, A., Zhang, L., & Pattipati, K. R. (2019). Context-driven proactive decision support for hybrid teams. *AI Magazine*, 40(3), 41–57. <https://doi.org/10.1609/aimag.v40i3.4810>
- Moffatt v. Air Canada*, 2024 BCCRT 149 (B.C. Civil Resolution Tribunal 2024).
- NASA Aviation Safety Reporting System. (n.d.). *ASRS program briefing* (data through December 2025). https://asrs.arc.nasa.gov/docs/ASRS_ProgramBriefing.pdf
- National Highway Traffic Safety Administration. (2018, June 19). Consumer advisory: NHTSA deems “Autopilot Buddy” product unsafe [Press release]. <https://www.nhtsa.gov/press-releases/consumer-advisory-nhtsa-deems-autopilot-buddy-product-unsafe>
- National Transportation Safety Board. (2014). *Descent below visual glidepath and impact with seawall, Asiana Airlines Flight 214, Boeing 777-200ER, HL7742, San Francisco, California, July 6, 2013* (NTSB/AAR-14/01). <https://www.ntsb.gov/investigations/AccidentReports/Reports/AAR1401.pdf>

- National Transportation Safety Board. (2019). *Collision between vehicle controlled by developmental automated driving system and pedestrian, Tempe, Arizona, March 18, 2018* (NTSB/HAR-19/03). <https://www.nts.gov/investigations/AccidentReports/Reports/HAR1903.pdf>
- NBC News. (2013, August). NSA to cut system administrators by 90 percent to limit data access. <https://www.nbcnews.com/technolog/ntsa-cut-system-administrators-90-percent-limit-data-access-6c10884390>
- NOVA. (2006). The final eight minutes [Cockpit voice recorder transcript, adapted from the Spanish accident report]. PBS. <https://www.pbs.org/wgbh/nova/plane crash/minutes.html>
- Occupational Safety and Health Administration. (n.d.). Group lockout/tagout requirements [29 C.F.R. § 1910.147(f)(3)]. <https://osha.gov/etools/lockout-tagout/tutorial/group-lockout-tagout-requirements>
- O'Dea, A., O'Connor, P., & Keogh, I. (2014). A meta-analysis of the effectiveness of crew resource management training in acute care domains. *Postgraduate Medical Journal*, 90(1070), 699–708. <https://doi.org/10.1136/postgradmedj-2014-132800>
- Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human Factors*, 56(3), 476–488. <https://doi.org/10.1177/0018720813501549>
- OpenAI. (2025). Expanding on what we missed with sycophancy. <https://openai.com/index/expanding-on-sycophancy/>
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381–410. <https://doi.org/10.1177/0018720810376055>
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253.
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, 30(3), 286–297.
- Pattipati, K. R. (2026, October 10). *Old lessons for a new machine: Engineering AI safety*. Qualtech Systems, Inc. (QSI).
- Pattipati, K. R., & Kleinman, D. L. (1991). A review of engineering models of human information-processing and decision-making in multi-task supervisory control. In D. L. Damos (Ed.), *Multiple task performance* (pp. 35–68). Taylor & Francis.
- Pattipati, K. R., Kleinman, D. L., & Ephrath, A. R. (1983). A dynamic decision model of human task selection performance. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(2), 145–166.
- Pete, A., Pattipati, K. R., & Kleinman, D. L. (1993a). Optimal team and individual decision rules in uncertain dichotomous situations. *Public Choice*, 75(3), 205–230. <https://doi.org/10.1007/BF01119183>
- Pete, A., Pattipati, K. R., & Kleinman, D. L. (1993b). Distributed detection in teams with partial information: A normative-descriptive model. *IEEE Transactions on Systems, Man, and Cybernetics*, 23(6), 1626–1648.
- Pete, A., Pattipati, K. R., Kleinman, D. L., & Levchuk, Y. N. (1998). An overview of decision networks and organizations. *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, 28(2), 173–193.
- Polet, P., Vanderhaegen, F., & Amalberti, R. (2003). Modelling border-line tolerated conditions of use (BTCU) and associated risks. *Safety Science*, 41(2–3), 111–136.
- Radiological Society of North America. (2023, May 2). AI bias may impair radiologist accuracy on mammogram. <https://www.rsna.org/news/2023/may/ai-bias-may-impair-accuracy>
- Rasmussen, J. (1997). Risk management in a dynamic society: A modelling problem. *Safety Science*, 27(2–3), 183–213. [https://doi.org/10.1016/S0925-7535\(97\)00052-0](https://doi.org/10.1016/S0925-7535(97)00052-0)
- Reason, J. (1990). *Human error*. Cambridge University Press.
- Reason, J. (1997). *Managing the risks of organizational accidents*. Ashgate.
- Reserve Bank of Australia. (1995, November). Implications of the Barings collapse for bank supervisors. *Reserve Bank of Australia Bulletin*. <https://www.rba.gov.au/publications/bulletin/1995/nov/1.html>
- Risk. (2008, May). Reports highlight SG weaknesses. <https://www.risk.net/ja/node/1506466>
- Rouse, W. B. (1988). Adaptive aiding for human/computer control. *Human Factors*, 30(4), 431–443. <https://doi.org/10.1177/001872088803000405>
- Rouse, W. B., & Morris, N. M. (1987). Conceptual design of a human error tolerant interface for complex engineering systems. *Automatica*, 23(2), 231–235.
- Rouse, W. B., & Rouse, S. H. (1983). Analysis and classification of human error. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(4), 539–549. <https://doi.org/10.1109/TSMC.1983.6313142>
- Salas, E., Sims, D. E., & Burke, C. S. (2005). Is there a “big five” in teamwork? *Small Group Research*, 36(5), 555–599. <https://doi.org/10.1177/1046496405277134>
- Salas, E., Wilson, K. A., Burke, C. S., & Wightman, D. C. (2006). Does crew resource management training work? An update, an extension, and some critical needs. *Human Factors*, 48(2), 392–412. <https://doi.org/10.1518/001872006777724444>
- Sarter, N. B., & Woods, D. D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human Factors*, 37(1), 5–19.
- Sawyer, N., & Campbell, A. (2008, February 1). Back-office savvy aided SG fraud. *Risk*. <https://risk.net/node/1506258>
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askeil, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2024). Towards

- understanding sycophancy in language models. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*. <https://arxiv.org/abs/2310.13548>
- Sheridan, T. B. (1992). *Telerobotics, automation, and human supervisory control*. MIT Press.
- Sheridan, T. B. (2002). *Humans and automation: System design and research issues*. Wiley.
- Sheridan, T. B. (2008). Risk, human error, and system resilience: Fundamental ideas. *Human Factors*, 50(3), 418–426. <https://doi.org/10.1518/001872008X250773>
- Sheridan, T. B. (2016). Human–robot interaction: Status and challenges. *Human Factors*, 58(4), 525–532. <https://doi.org/10.1177/0018720816644364>
- Sheridan, T. B. (2019). Individual differences in attributes of trust in automation: Measurement and application to system design. *Frontiers in Psychology*, 10, 1117. <https://doi.org/10.3389/fpsyg.2019.01117>
- Sheridan, T. B., & Ferrell, W. R. (1974). *Man-machine systems: Information, control, and decision models of human performance*. MIT Press.
- Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators* (Technical Report ADA057655). Massachusetts Institute of Technology, Man-Machine Systems Laboratory.
- SKYbrary Aviation Safety. (n.d.). B742 / B741, Tenerife Canary Islands Spain, 1977. <https://skybrary.aero/accidents-and-incidents/b742-b741-tenerife-canary-islands-spain-1977>
- Software Engineering Institute. (2022). *Common sense guide to mitigating insider threats* (7th ed.). Carnegie Mellon University. <https://www.sei.cmu.edu/library/common-sense-guide-to-mitigating-insider-threats-seventh-edition/>
- Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, 27(1), 76–105.
- Swain, A. D., & Guttman, H. E. (1983). *Handbook of human reliability analysis with emphasis on nuclear power plant applications* (NUREG/CR-1278). U.S. Nuclear Regulatory Commission.
- Tang, Z. B., Pattipati, K. R., & Kleinman, D. L. (1992). A distributed M-ary hypothesis testing problem with correlated observations. *IEEE Transactions on Automatic Control*, 37(7), 1042–1046.
- Tenney, R. R., & Sandell, N. R., Jr. (1981). Detection with distributed sensors. *IEEE Transactions on Aerospace and Electronic Systems*, AES-17(4), 501–510.
- Tulga, M. K., & Sheridan, T. B. (1980). Dynamic decisions and work load in multitask supervisory control. *IEEE Transactions on Systems, Man, and Cybernetics*, 10(5), 217–232. <https://doi.org/10.1109/TSMC.1980.4308481>
- U.S. Attorney’s Office, Northern District of California. (2024, November 14). *Cruise admits to submitting a false report to influence a federal investigation and agrees to pay \$500,000* [Press release]. <https://www.justice.gov/usao-ndca/pr/cruise-admits-submitting-false-report-influence-federal-investigation-and-agrees-pay>
- U.S. Department of Justice. (2017, January 11). Volkswagen AG agrees to plead guilty and pay \$4.3 billion in criminal and civil penalties; six Volkswagen executives and employees are indicted in connection with conspiracy to cheat U.S. emissions tests (Press release 17-037). <https://www.justice.gov/opa/pr/volkswagen-ag-agrees-plead-guilty-and-pay-43-billion-criminal-and-civil-penalties-six>
- U.S. Environmental Protection Agency. (2015, September 18). EPA, California notify Volkswagen of Clean Air Act violations [News release]. <https://www.epa.gov/archive/epa/newsreleases/epa-california-notify-volkswagen-clean-air-act-violations-carmaker-allegedly-used.html>
- U.S. Nuclear Regulatory Commission. (2000). *Technical basis and implementation guidelines for a technique for human event analysis* (ATHEANA) (NUREG-1624, Rev. 1).
- Universal Robots. (n.d.). Safety functions [User manual, software 5.25]. https://www.universal-robots.com/manuals/EN/HTML/SW5_25_1/Content/prod-usr-man/complianceUR15/safetyFunctionsAndInterfaces/safety_functions_table1a.htm
- Vaughan, D. (1996). *The Challenger launch decision: Risky technology, culture, and deviance at NASA*. University of Chicago Press.
- Willett, P., Swaszek, P. F., & Blum, R. S. (2000). The good, bad, and ugly: Distributed detection of a known signal in dependent Gaussian noise. *IEEE Transactions on Signal Processing*, 48(12), 3266–3279.