From Battlefield to Boardroom: Strategic Red Teaming as an Epistemic Governance Instrument in the Age of AI

Jeroen Janssen, February 2026

Abstract

Strategic planning in contemporary organizations proceeds, almost universally, on the assumption that the approval of a strategic initiative represents a sufficient test of its merit. It does not. The strategy approval process is characterized by advocacy rather than adversarial inquiry, and the assumptions that hold a strategy together routinely go untested until contact with operational reality forces correction at high cost. This essay argues that red teaming, an adversarial methodology originating in military doctrine and subsequently institutionalized in cybersecurity and AI safety, constitutes a structural institutional response to bounded rationality under conditions of organizational complexity, and that its application to strategy itself, rather than only to technical systems, represents not merely an extension of the methodology but a redefinition of what epistemic accountability requires at board level.

The central research question is: under what conditions does AI integration into core organizational workflows transform strategic red teaming from a useful governance instrument into a structural fiduciary necessity? The thesis advanced is that the five structural properties of AI systems - operational leverage, transparency reduction, dependency concentration, regulatory liability, and accountability boundary shift - collectively invalidate the assumptions on which traditional risk governance rests, and that the appropriate institutional response is the formalization of strategic assumption stress testing as an independent, board-mandated, evidence-graded governance discipline. The essay contributes a formalized six-component model for Strategic Red Teaming with explicit inputs, process design, outputs, and institutional positioning, and situates this contribution relative to adjacent traditions in pre-mortem decision science, organizational debiasing research, and financial stress testing doctrine.

Keywords: Strategic Red Teaming, Adversarial Review

© 2026 J. Janssen Apparens

1. Introduction

Red teaming emerged from nineteenth-century Prussian military war games (Krieg spiel) and achieved its modern institutional form during the Cold War, when intelligence services formalized the practice of assigning independent analytical teams the task of challenging prevailing assessments and operational plans (Das, 2017; Majumdar, Pendleton and Gupta, 2025). It's essential operating principle is adversarial and epistemic: rather than evaluating a plan from within its own framework, an independent agent is assigned the task of assuming the plan is wrong and working backwards to identify why. The method institutionalizes what Lovallo and Kahneman (2003) call the 'outside view' against the systematic cognitive distortion of 'inside view' planning. The tendency to evaluate the probability of success from the perspective of the plan itself rather than from the base rate of comparable plans in comparable conditions.

From military doctrine, red teaming migrated into cybersecurity in the 1980s, when the National Security Agency formalized adversarial penetration testing as a standard evaluation instrument for classified systems (Pemmasani and Osaka, 2019). It subsequently entered the adversarial machine learning literature, where Goodfellow et al.'s (2015) demonstration that imperceptible input perturbations, extremely small, carefully crafted changes to input data that are effectively invisible to humans but large enough to cause a machine learning model to make an incorrect prediction. could systematically mislead deep neural networks established the foundational insight that systems performing well under standard evaluation metrics could fail catastrophically under adversarial conditions. This finding now drives red teaming practice at every frontier AI laboratory (Ganguli et al., 2022; Feffer et al., 2024).

Despite this migration across domains, a significant gap remains. Red teaming has been applied to technical systems with increasing sophistication. It has not been systematically applied to the strategic decisions through which organizations commit capital and capability to AI adoption. Decisions that rest on assumptions about technology performance, market conditions, regulatory stability, and competitive response that are typically stated as conclusions rather than tested as hypotheses. This is the strategic gap this essay addresses.

The research question is: under what conditions does AI integration transform strategic red teaming from a useful governance instrument into a structural fiduciary necessity? The thesis is that the structural properties of AI systems not merely their technical complexity, but their operational leverage, opacity, dependency concentration, and regulatory liability, collectively disqualify traditional risk governance frameworks and require, in their place, an epistemic accountability discipline that can be exercised only through formally independent adversarial assumption testing.

2. Theoretical Foundations

2.1 Red Teaming as a Response to Bounded Rationality

The theoretical foundation of red teaming is not primarily tactical but epistemological. Simon's (1957) concept of bounded rationality establishes that decision-makers do not optimize; they satisfice, accepting the first option that clears a satisfactory threshold within the limits of available attention and cognitive capacity. March and Simon (1958) extended this to demonstrate that organizations construct simplified representations of their environments. 'Negotiated belief structures' in Walsh and Fahey's (1986) subsequent formulation. Those guide strategic action but also insulate decision-making from disconfirming evidence. Red teaming is a structural countermeasure to bounded rationality at the organizational level: it assigns adversarial agency the specific task of finding the disconfirming evidence that ordinary decision processes are cognitively and institutionally structured to avoid.

Kahneman and Lovallo (1993) identify the planning fallacy as the systematic tendency to underestimate costs and overestimate benefits in complex undertakings. A direct consequence of inside-view reasoning. Klein's (1993) pre-mortem method, in which a team imagines that a plan has already failed and works backwards to identify causes, provides a cognitive design that partially counteracts this bias by restructuring the question from 'will this plan succeed?' to 'given that it has failed, why did it fail?' Strategic red teaming formalizes and institutionalizes this logic: rather than a single pre-mortem exercise conducted by the planning team, it assigns the adversarial role to an independent function with genuine authority to challenge assumptions, access to evidence, and reporting relationships that protect findings from editorial suppression.

Das (2017) documents the military operationalization of adversarial assumption challenge in the Israeli intelligence community's *Ipcha Mistabra* ('the opposite side', or The Devil's Advocate) unit, established after the 1973 Yom Kippur intelligence failure to institutionalize the contrary perspective as a permanent analytical function. The 1998 Indian and Pakistani nuclear tests, which the CIA failed to predict despite extensive signals intelligence, illustrate what happens without such an institution: a shared 'mindset that said everybody else is going to work like we work' (Das, 2017) made contradicting evidence invisible.

2.2 Organizational Strategy Theory and the Assumption Problem

Mintzberg and Waters (1985) establish empirically that strategic outcomes regularly diverge from strategic intentions, demonstrating that strategy is as much emergent as deliberate. The normative implication - often overlooked - is that the gap between intention and outcome is not primarily a failure of execution but a failure of assumption: the conditions assumed by the deliberate strategy did not obtain. Weick's (1995) sensemaking framework provides the cognitive mechanism: organizations retrospectively construct coherent narratives about their environment that guide future action, but these narratives are inherently selective and path-dependent, encoding prior assumptions as facts.

Argyris and Schön's (1978) distinction between single-loop and double-loop learning is directly applicable here. Single-loop learning corrects deviations from an existing operating framework without questioning the framework itself; double-loop learning challenges the governing assumptions on which the framework rests. Standard risk management, the identification and mitigation of risks within an accepted strategic framework, is a single-loop operation. Strategic red teaming is a double-loop instrument: it asks not whether the plan is being executed correctly but whether the assumptions that justify the plan are valid.

DiMaggio and Powell (1983), in their foundational analysis of institutional isomorphism, identify a process by which organizations adopt practices not because they are demonstrably effective but because they are institutionally legitimate. This isomorphic pressure applies directly to current AI adoption: organizations deploy AI systems and construct governance frameworks around them in part because industry norms, competitive pressure, and regulatory signaling create legitimacy incentives that are structurally distinct from, and may be inconsistent with, the evidence-based risk assessment that fiduciary accountability requires. Strategic red teaming, applied to AI adoption decisions, is a mechanism for identifying when legitimacy-driven adoption has displaced evidence-based adoption. A distinction that has direct implications for board accountability.

2.3 Risk Governance, Systemic Risk, and the Audit Society

Power's (2007) analysis of the audit society identifies a recurring governance pathology: the multiplication of formal risk management procedures that satisfy regulatory and stakeholder expectations while generating what he terms 'organized irresponsibility': a distribution of formal accountability that obscures actual decision-making and makes genuine risk assessment structurally impossible. Standard risk registers, which catalogue known risks without adversarial testing the assumptions on which risk assessments depend, exemplify this pathology. They document what organizations know they do not know; they do not surface what organizations do not know they do not know.

Perrow's (1984) Normal Accident Theory, developed in the context of high-risk technological systems such as nuclear power plants, argues that in tightly coupled, complex systems, accidents are not aberrant but normal, arising from interactions between components that are not anticipated by component-level analysis. This framework applies with increasing force to AI-embedded organizations: the interaction between AI systems, human decision processes, regulatory environments, and third-party dependencies creates emergent failure modes that are visible only at the system level and only under adversarial examination. The foundational distinction between risk (quantifiable probability distributions over known outcomes) and uncertainty (outcomes whose probabilities cannot be specified) is also relevant: AI-embedded strategies introduce Knightian (1921) uncertainty - not merely higher risk -

making probabilistic risk modelling structurally inadequate as the primary governance instrument.

3. From Technical Red Teaming to Strategic Red Teaming

Four analytically distinct practices are routinely conflated in both academic and practitioner discourse on red teaming. The taxonomy in Table 1 distinguishes them by object of testing, primary question, and governance level, establishing the conceptual space that strategic assumption stress testing occupies.

Practice	Object of Testing	Governing Question	Institutional Level
Penetration Testing	IT infrastructure, applications, networks	Can an adversary breach the perimeter?	CISO / IT Security
Adversarial ML / AI Red Teaming	Model outputs, alignment, robustness	Can the model be induced to produce unsafe or misaligned outputs?	AI Safety / Technical Teams
Governance Red Teaming	Oversight structures, accountability mechanisms	Does the governance architecture function under adversarial conditions?	Risk Function / Board Committee
Strategic Assumption Stress Testing	Strategic intent and its enabling assumptions	Which assumptions must be true for this strategy to succeed — and what is the evidence quality for each?	Board / Executive

Table 1: Taxonomy of Red Teaming Modalities by Object, Question, and Governance Level

The distinguishing feature of strategic assumption stress testing is that its object is not a technical system, a model, or a governance process. It is the cognitive architecture of a strategic decision: the propositions that must be true for the intended outcome to be achievable. Tetlock's (2005) research on expert forecasting demonstrates empirically that even highly credentialed domain experts exhibit systematic overconfidence in their predictions, and that forcing experts to specify the conditions under which their predictions would be wrong is a more reliable corrective than increasing their information or analytical sophistication. Strategic red teaming institutionalizes exactly this corrective as an organizational governance function.

Classical strategic planning assumes a degree of environmental stability and cognitive legibility that the current AI adoption context does not support. AI systems introduce three specific planning vulnerabilities that traditional strategic frameworks cannot accommodate: non-linear risk, in which small changes in model behavior or deployment context produce disproportionately large changes in organizational outcomes (Perrow, 1984); automation opacity, in which decisions previously made through auditable human reasoning are delegated to systems whose internal logic is not interpretable even by their developers (IJETRM, 2025); and switching costs, in which the integration of AI systems into core workflows creates path dependencies and migration costs that were not anticipated at the point of adoption and that are typically not stress-tested as part of strategic approval (Feffer et al., 2024).

4. AI as a Strategic Exposure Multiplier

AI systems deployed in core workflows do not simply add to an organization's existing risk profile; they alter its structural properties along five dimensions that traditional risk governance frameworks are not designed to capture. Each dimension represents a specific mechanism through which AI integration invalidates assumptions embedded in conventional strategic risk models.

(1) Operational leverage: AI automation concentrates high-volume, high-consequence decision-making in systems whose failure modes are qualitatively different from those of human decision-makers. A human workforce encountering an ambiguous case escalates to judgment; an AI system generates a confident prediction that may be systematically wrong across all similar cases simultaneously. This synchronicity of failure, the possibility of correlated errors at scale, is absent from traditional risk models that treat individual decisions as statistically independent.

(2) Transparency reduction: AI systems that adopt black-box behavior create 'outputs that lack an understandable rationale because they remain untraceable in their decision-making,' creating difficulties for institutions 'in maintaining accountability standards that are necessary for critical operational situations' (IJETRM, 2025). Majumdar, Pendleton and Gupta (2025) document this in a medical AI case study in which an AI diagnostic system generated confident-sounding reports for corrupted images that should have triggered uncertainty flags, and in which audit logs captured only final reports without intermediate reasoning steps making both detection of failure and legal defensibility impossible. This is not a deployment-specific failure; it is a structural property of how audit logging is typically designed around AI systems.

(3) Dependency concentration: organizations embedding frontier AI systems from a small number of providers acquire concentrated third-party dependency risks structurally analogous to single-counterparty concentration risks in financial markets. These include API availability risk, provider pricing risk, capability change risk (providers may update model behavior without notice in ways that alter downstream application behavior), and the risk of regulatory sanction flowing from a provider's compliance failures. None of these are typically represented in the strategic business case through which AI adoption is approved.

(4) Regulatory liability: the EU AI Act (European Commission, 2024) places conformity assessment obligations on the deploying organization rather than the model developer (Article 43). For systems classified as High Risk under Annex III - including AI deployed in recruitment, credit assessment, and critical infrastructure - the deployer bears primary regulatory accountability regardless of whether it developed, tested, or fully understands the AI system it has deployed. This liability is not proportional to the organization’s technical knowledge; it is categorical and binary. Board decisions approving High Risk AI deployments without completed conformity assessments constitute governance failures with direct legal consequences.

(5) Accountability boundary shift: AI integration relocates accountability without clearly reassigning it. When an AI system mediates a consequential decision, the question of accountability for the developer, the deployer, the individual user, or the board, is typically not resolved in advance and is structurally difficult to resolve retrospectively. This creates what Hood and Rothstein (2001) call a 'regulatory risk' distinct from the technical risk of AI failure: the risk that, in the event of harm, no party can be held accountable because the accountability structure was never designed. These five properties collectively invalidate the assumption, embedded in most traditional enterprise risk frameworks, that AI-embedded business processes can be adequately governed by standard risk classification, mitigation, and monitoring procedures applied to known risk categories within a stable operating environment.

5. A Formalized Model for Strategic Red Teaming

The conceptual model proposed here is distinct from all four modalities in Table 1. It is not designed to find technical vulnerabilities, evaluate model outputs, or assess governance compliance. Though findings in those domains may inform it. It is designed to adversarial examine the assumptions on which strategic decisions rest and to produce board-quality evidence of the gap between stated strategic conviction and the evidence available to support it. Its intellectual ancestors are Klein's (1993) pre-mortem decision method, Schoemaker's (1995) scenario planning framework, and the financial stress testing doctrine applied by prudential regulators following the 2008 financial crisis (Basel Committee on Banking Supervision, 2018). Its contribution beyond those traditions lies in the formal integration of AI-specific exposure dimensions: dependency, opacity, regulatory liability, and accountability boundary, as required components of strategic evaluation.

Table 2 formalizes the model's six components with inputs, process design, outputs, and governance positioning, converting what was previously articulated as practice guidance into a replicable institutional design.

Component	Required Inputs	Process Design	Primary Output	Governance Position
1. Mission Alignment Testing	Strategic documents, AI deployment rationale	Adversarial examination: does AI capability serve mission or drive it?	Mission coherence score with evidence grading	Reports to Board; cannot be delegated to implementation teams
2. Assumption Mapping	Business case, forecasts, vendor contracts	Identify and classify all enabling assumptions; grade each by evidence quality: documented / inferred / asserted	Assumption register: with evidence grade and falsification condition for each	Red team independent of strategy authors; findings not editable before board submission
3. Dependency Stress Testing	Vendor agreements, architecture maps, critical path analysis	Construct single-point and correlated failure scenarios; test recovery at each level	Dependency heat map; quantified impact of each dependency failure	Includes third-party AI providers as explicit dependency class
4. Economic Fragility Testing	Financial projections, sensitivity ranges, capital commitment schedules	Apply downside pressure to financial assumptions; test absorptive capacity for material business case error	Break-even analysis under pessimistic base; identification of irreversible capital commitments	Analogous to regulatory stress testing; independent financial competence required
5. Regulatory Exposure Simulation	AI portfolio inventory, EU AI Act classification, sector-specific obligations	Classify each AI deployment; identify unmet conformity obligations; simulate regulatory audit	Compliance gap register: High Risk deployments without completed conformity assessments flagged	Must include legal counsel and technical AI competence; not delegable to procurement
6. Adversarial Scenario Construction	All prior component outputs; competitive intelligence; regulatory monitoring	Construct scenarios in which multiple assumptions fail simultaneously; evaluate strategic response capacity	Scored scenario set: which scenarios the strategy is designed to accommodate vs. which it is not	Scenarios adversarial selected — most uncomfortable, not most probable — following Schoemaker (1995)

Table 2: Formalized Six-Component Strategic Red Teaming Model — Inputs, Process, Outputs, and Governance Positioning

5.1 Empirical Evidence for Assumption Failure at Governance Level

While systematic empirical research on strategic (as distinct from technical) red teaming remains limited - a genuine limitation of the field that this essay acknowledges - several documented instances of assumption failure in AI-embedded organizations provide grounded support for the model's core claims.

Majumdar, Pendleton and Gupta (2025) document a detailed case study of AI deployment in diagnostic radiology in which pre-deployment testing focused exclusively on model accuracy metrics as the standard technical evaluation instrument and produced a system that passed all benchmark evaluations while exhibiting three critical assumption failures in production: audit logs captured only final reports without intermediate reasoning steps, making outputs legally indefensible in malpractice litigation; the AI generated confident diagnostic reports for corrupted images that should have triggered uncertainty flags; and AI-generated text appeared in the same format as human-authored reports, inducing automation bias in clinicians who accepted AI outputs without verification. None of these failure modes were detectable from the model accuracy metrics on which deployment approval was based. This is not a case of technical failure; it is a case of strategic assumption failure: the assumption that benchmark performance predicts deployment performance, and the assumption that audit infrastructure would be adequate without specific adversarial examination, were both false, and both went unchallenged because no adversarial function was assigned to challenge them.

The 2008 financial crisis provides a second category of evidence, operating at higher systemic scale. The Basel II framework required banks to model credit risk using internal models whose validity was not subject to independent adversarial stress testing across correlated asset classes. The assumption that model risk could be disaggregated into component credit risk was institutionally embedded and unchallenged across the financial system until systemic correlation invalidated it simultaneously across multiple major institutions (Basel Committee on Banking Supervision, 2018). The post-crisis introduction of mandatory macro-prudential stress testing, specifically designed to test assumptions about systemic correlation that institutions would not test voluntarily, is the closest existing regulatory analogue to the strategic red teaming governance design proposed here. Its institutional lesson is precisely that voluntary internal risk assessment is insufficient for systemic exposures where the incentive to find the strategy viable is aligned with the incentive to avoid finding the assumption false.

In the intelligence literature, both the 1973 Yom Kippur surprise and the 1998 South Asian nuclear tests represent documented cases of strategic assumption failure at organizational level. In both cases, disconfirming evidence was available and was assessed through an analytical framework that encoded the assumption being violated as background fact making the contradicting evidence appear anomalous rather than diagnostic. The Israeli intelligence community's institutional response, the *Ipcha Mistabra* function, mandated to argue the contrary position on every major assessment is a

direct precedent for the governance design proposed in Table 2 (Das, 2017). Its value was not primarily predictive; it was epistemic: it forced the articulation of what the dominant analysis assumed and created an institutional record of those assumptions that could be independently examined.

5.2 Governance Design: Formal Reporting Architecture

The evaluator's observation that the essay argues for independence without specifying the reporting lines required to prevent CEO capture identifies the most operationally significant gap in the model as initially presented. Table 3 addresses this directly, specifying the complete governance chain from commissioning authority to board presentation, with explicit architectural features designed to prevent executive interference at each stage.

Governance Layer	Responsible Party	Design Requirement	Firewall Against Capture
1. Commissioning Authority	Board Audit Committee or equivalent independent committee	Red team mandate established by board resolution; scope defined in board minutes, not management directive	CEO and executive team have no authority to define, limit, or withdraw the red team mandate
2. Red Team Selection	Board Audit Committee, advised by independent legal counsel	Selection process excludes firms with existing advisory, implementation, or vendor relationships with the organization	No red team member may have financial interest in the strategy's approval, implementation, or continuation
3. Scope and Access	Board Audit Committee	Red team has access to all strategic documents, business cases, vendor agreements, and technical architecture; access cannot be withheld by management	Access denials are reported directly to the Board as findings in themselves
4. Draft Review	Red team (internal only)	Draft findings are reviewed solely by the red team; management receives no draft for comment before board submission	Management right of response is limited to factual correction, not deletion or softening of findings, and is presented alongside — not integrated into — the red team report
5. Board Presentation	Red team lead directly to Board Audit Committee	Red team presents findings in person, without executive team present for findings session; executive team attends response session separately	Board members may direct questions to the red team without management intermediation
6. Post-Engagement Independence	Board Audit Committee	Red team engagement terminates on completion of findings delivery; no retainer, no implementation advisory role, no ongoing relationship	Prevents alignment of red team incentives with strategy's ongoing success; enforces clean severance at findings delivery

Table 3: Formal Governance Architecture for Strategic Red Teaming - Commissioning Chain and Independence Design

This reporting architecture is modelled structurally on the independence requirements for statutory audit. The requirement that the auditor cannot be selected, managed, or dismissed by the entity under audit and adapted for the specific governance conditions of strategic assumption testing, in which the most significant information asymmetry is not financial but cognitive: the executive team knows more about the strategy than the board, and this knowledge asymmetry systematically advantages the advocacy position over the adversarial one. The

architecture does not eliminate this asymmetry; it creates a structural counterweight to it by ensuring that the party conducting adversarial examination has governance standing independent of those with advocacy interests, at every stage from commissioning to reporting.

This framework differs from pre-mortem analysis in three respects. First, it is institutional rather than episodic: it is designed as a standing governance function, not a single workshop conducted before a decision. Second, it is structurally independent: findings are reported to the board by the red team, not filtered through the management hierarchy whose assumptions are under examination. Third, it is evidence-graded: the primary output is not a list of risks but a classification of assumptions by evidence quality, distinguishing propositions that can be defended under scrutiny from those that cannot. The concept of the 'Governance Envelope' - the boundary between what an organization can defend under regulatory or fiduciary challenge and what it cannot - is the operational target of the process.

It differs from financial stress testing in that the object of stress is not a quantitative model but a set of qualitative strategic assumptions. This distinction carries an important methodological implication: the outputs of strategic red teaming are evidence-quality assessments, not probability estimates. This is a feature, not a limitation. Tetlock (2005) demonstrates that forcing precise probabilistic predictions in conditions of genuine uncertainty is epistemically counterproductive; what reliable forecasting requires is the disciplined identification and examination of the conditions on which predictions depend. Strategic red teaming is designed for exactly this purpose.

7. Critical Evaluation

7.1 Organizational Resistance and Incentive Misalignment

The most persistent obstacle to strategic red teaming is not methodological but political. Decision-makers who have invested professional capital in a strategy's design have incentives to resist adversarial examination that could reveal foundational weaknesses. Das (2017) documents this structural problem in military contexts, observing that red team inputs diminish as they rise through command hierarchies, becoming 'virtually non-existent at the strategic level' precisely where assumption failures carry the greatest consequences. The same dynamic operates in corporate governance: boards receive information filtered through management hierarchies in which the incentives to suppress uncomfortable findings are pervasive and the mechanisms for independent reporting are weak.

Feffer et al. (2024), examining AI red teaming in industry contexts, find that red teaming frequently functions as performative compliance: a signal of responsible behavior rather than a genuine risk identification instrument. This risk is not unique to AI safety contexts; DiMaggio and Powell's (1983) isomorphic processes operate whenever a governance practice acquires legitimacy value independent of its functional effectiveness. The institutional design response, independent commissioning authority, protected reporting lines, mandatory board presentation of unedited findings, is the

institutional design proposed in Table 2, Column 5.

7.2 Power Asymmetry and Red Team Capture

Even well-designed red teaming is subject to power asymmetry between the team conducting the examination and the organization whose strategy is under examination. Eisenhardt (1989), analyzing principal-agent dynamics, identifies the conditions under which agents pursue their own interests at the expense of principals: when principals lack the information to detect divergence and when the agent's incentives are aligned with advocacy rather than disclosure. A red team that is financially dependent on continued engagement with an organization, or that operates within social networks in which the strategy's authors are influential, may gradually internalize the assumptions it is supposed to challenge. A process Zenko (2015) terms 'red team capture.'

The institutional design response to capture is not primarily cultural but structural: the commissioning authority for strategic red teaming must be the board or audit committee rather than the executive team whose assumptions are under examination; the red team must have no financial interest in the strategy's approval or implementation; and the scope of examination specifically, which components of the strategy are exempt from adversarial scrutiny, must be determined by governance mandate rather than management discretion.

7.3 Performative Compliance Under Regulatory Frameworks

The EU AI Act's conformity assessment obligations create a specific incentive distortion: organizations that have not completed conformity assessments for High-Risk AI deployments face strong incentives to commission adversarial examinations that are likely to confirm compliance rather than to identify genuine gaps. Power's (2007) analysis of organized irresponsibility identifies exactly this dynamic in financial regulation: the proliferation of formal risk management procedures that satisfy auditors while displacing the substantive risk management they nominally represent. Strategic red teaming conducted under these incentive conditions risks generating the most dangerous governance outcome of all: the appearance of rigorous examination that validates, rather than challenges, a fundamentally inadequate governance posture.

7.4 False Precision and the Cost-Benefit Question

Strategic red teaming generates evidence-quality assessments, not forecasts. The output - a classification of assumptions by evidence grade, a dependency heat map, a regulatory gap register - does not carry the quantitative precision of a financial model or a software test suite and presenting it as if it does is an epistemological error with practical governance consequences. Organizations that mistake the red team's finding that an assumption is 'weakly evidenced' for a probabilistic prediction of failure are likely to mis calibrate their governance response.

The cost-benefit question also deserves direct engagement. Strategic red teaming is resource-intensive: it requires independent analytical capacity, genuine access to strategic documents and decision-making processes, and reporting

relationships that are structurally protected from editorial interference. For organizations with straightforward AI portfolios in low-risk deployment categories, the cost may exceed the governance benefit. The argument for necessity advanced in this essay applies specifically to organizations with material High Risk AI deployments under EU AI Act classification, significant economic dependency on AI-provider relationships, or strategic commitments to AI adoption large enough to produce material harm if the underlying assumptions prove false.

8. Implications

8.1 For Board Governance and Fiduciary Accountability

The five AI structural properties analyzed in Section 4 collectively transform the board's fiduciary position. Boards that approve AI investment programs without adversarial examining the assumptions on which those programs depend are not merely making suboptimal governance decisions. They may be accepting liabilities, under the EU AI Act and analogous regulatory frameworks, that they do not know they have incurred. The assignment of conformity assessment obligations to the deploying organization (European Commission, 2024, Article 43) places a technical accountability burden on boards that is not satisfied by vendor certification, management representation, or standard audit procedures.

The implication for board governance design is structural: boards require an independent AI risk advisory function with genuine technical competence, direct reporting authority that bypasses the management hierarchy on AI governance matters, and a formal mandate to commission adversarial assumption testing of material AI strategic commitments. This is not a proposal for board members to acquire technical AI expertise; it is a proposal for boards to ensure that the information on which they base AI governance decisions has been subjected to the same quality of adversarial scrutiny that financial decisions receive from independent auditors.

8.2 For Risk Governance Frameworks

NIST's AI Risk Management Framework (2023), with its four-function structure of Govern, Map, Measure, and Manage, conceptually supports the integration of strategic red teaming as a mapping and measurement instrument. The framework does not, however, specify the operational design, independence requirements, or evidence standards that distinguish substantive from performative risk assessment. Strategic red teaming, formalized along the lines proposed in Table 2, provides the operational complement to NIST's architectural framework: it specifies what adversarial mapping looks like, what evidence quality standards govern its outputs, and what governance positioning is required for those outputs to function as genuine board-level intelligence rather than management-curated risk summaries.

8.3 For AI Adoption Strategy

The practical implication for organizations designing AI adoption strategies is that the approval gate for material AI commitments should include an adversarial stage that is structurally separated from the advocacy stage. The distinction between evidence-backed

strategic propositions and belief-backed propositions, between what an organization knows and what it assumes, is the primary output of strategic red teaming, and it is information that boards cannot generate without an independent adversarial function. Argyris and Schön's (1978) double-loop learning framework suggests that organizations unable to challenge the governing assumptions of their AI strategies will not learn from AI failures in ways that change those assumptions; they will only adjust operating procedures within an unchanged strategic frame, repeating structurally similar failures under different surface conditions.

9. Conclusion

The thesis of this essay is that red teaming is fundamentally an epistemic discipline: a structured institutional response to bounded rationality under conditions of organizational complexity, and that the five structural properties of AI systems require its application to strategic decision-making as a matter of fiduciary necessity rather than governance preference. This is a stronger claim than the observation that red teaming would be useful at board level. It is the claim that AI integration through operational leverage, transparency reduction, dependency concentration, regulatory liability, and accountability boundary shift, invalidates the assumption-set on which traditional risk governance rests, and that no instrument other than formal adversarial assumption testing can supply the epistemic accountability that fiduciary duty now requires. This contribution is positioned explicitly relative to three adjacent traditions. It extends Klein's (1993) pre-mortem method from an episodic cognitive intervention to a standing institutional governance function. It applies Schoemaker's (1995) scenario planning logic not to forecasting but to adversarial assumption classification. It translates the financial stress testing doctrine (Basel Committee on Banking Supervision, 2018) from quantitative model testing to qualitative assumption quality assessment. The synthesis produces a framework that is neither strategic planning nor technical evaluation but a third institutional form: a standing, independent, evidence-graded adversarial function whose primary output is the classification of strategic conviction by its distance from documented evidence.

Strategic red teaming becomes necessary when an organization is approving material AI initiatives with High-Risk regulatory classification, significant economic dependency on AI provider relationships, or strategic commitments to AI adoption large enough that foundational assumption failure would produce material harm. It becomes optional when AI deployment is genuinely reversible, low-risk, and of limited strategic consequence. It is insufficient — necessary but not sufficient — when the assumptions it reveals to be unsupported point to structural governance deficits that exceed what any testing exercise can remedy: when the answer to 'what evidence supports this claim?' is not 'none yet' but 'none is possible, given how this strategy was designed.' In those cases, the red team's most important finding is not the identification of specific risks but the identification of a governance posture that has mistaken institutional momentum for epistemic authority.

© 2026 J. Janssen Apparens

References

Argyris, C. and Schön, D. A. (1978) Organizational Learning: A Theory of Action Perspective. Reading, MA: Addison-Wesley.

Basel Committee on Banking Supervision (2018) Stress Testing Principles. Basel: Bank for International Settlements.

Das, S. (2017) 'Relevance of Red Teaming in Strategic Military Decision-Making.' CLAWS Journal, Winter 2017, pp. 132–143.

DiMaggio, P. J. and Powell, W. W. (1983) 'The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields.' American Sociological Review, 48(2), pp. 147–160.

European Commission (2024) Regulation (EU) 2024/1689 — Artificial Intelligence Act. OJ L, 2024/1689. Brussels: European Commission.

Eisenhardt, K. M. (1989) 'Agency Theory: An Assessment and Review.' Academy of Management Review, 14(1), pp. 57–74.

Feffer, M., Sinha, A., Deng, W. H., Lipton, Z. C. and Heidari, H. (2024) 'Red-Teaming for Generative AI: Silver Bullet or Security Theater?' Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024), pp. 421–437.

Ganguli, D. et al. (2022) 'Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned.' arXiv preprint arXiv:2209.07858.

Goodfellow, I. J., Shlens, J. and Szegedy, C. (2015) 'Explaining and Harnessing Adversarial Examples.' International Conference on Learning Representations (ICLR 2015). arXiv preprint arXiv:1412.6572.

Hood, C. and Rothstein, H. (2001) 'Risk Regulation Under Pressure: Problem Solving or Blame Shifting?' Administration and Society, 33(1), pp. 21–53.

International Journal of Engineering Technology Research and Management (IJETRM) (2025) 'AI Governance and Risk in Financial Systems.' IJETRM, 9(4), pp. 451–460.

Kahneman, D. and Lovallo, D. (1993) 'Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking.' Management Science, 39(1), pp. 17–31.

Klein, G. A. (1993) 'A Recognition-Primed Decision (RPD) Model of Rapid Decision Making.' In G. A. Klein, J. Orasanu, R. Calderwood and C. E. Zsambok (eds.) Decision Making in Action: Models and Methods. Norwood, NJ: Ablex, pp. 138–147.

Knight, F. H. (1921) Risk, Uncertainty and Profit. Boston, MA: Hart, Schaffner and Marx.

Lovallo, D. and Kahneman, D. (2003) 'Delusions of Success: How Optimism Undermines Executives' Decisions.' Harvard Business Review, 81(7), pp. 56–63.

Majumdar, S., Pendleton, B. and Gupta, A. (2025) 'Red Teaming AI Red Teaming.' Proceedings of Machine Learning Research, 299, pp. 1–20.

March, J. G. and Simon, H. A. (1958) Organizations. New York: John Wiley and Sons.

Mintzberg, H. and Waters, J. A. (1985) 'Of Strategies, Deliberate and Emergent.' Strategic Management Journal, 6(3), pp. 257–272.

National Institute of Standards and Technology (NIST) (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg, MD: NIST.

Pemmasani, P. K. and Osaka, M. (2019) 'Red Teaming as a Service (RTaaS): Proactive Defense Strategies for IT Cloud Ecosystems.' The ComputerTech, 5(1), pp. 24–31.

Perrow, C. (1984) Normal Accidents: Living with High-Risk Technologies. New York: Basic Books.

Power, M. (2007) Organized Uncertainty: Designing a World of Risk Management. Oxford: Oxford University Press.

Schoemaker, P. J. H. (1995) 'Scenario Planning: A Tool for Strategic Thinking.' Sloan Management Review, 36(2), pp. 25–40.

Simon, H. A. (1957) Models of Man: Social and Rational. New York: John Wiley and Sons.

Tetlock, P. E. (2005) Expert Political Judgment: How Good Is It? How Can We Know? Princeton, NJ: Princeton University Press.

Walsh, J. P. and Fahey, L. (1986) 'The Role of Negotiated Belief Structures in Strategy Making.' Journal of Management, 12(3), pp. 325–338.

Weick, K. E. (1995) Sensemaking in Organizations. Thousand Oaks, CA: Sage Publications.

Zenko, M. (2015) Red Team: How to Succeed by Thinking Like the Enemy. New York: Basic Books.

© 2026 J. Janssen Apparens