Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Sources of Risk in AI Systems Revisited: FCAIR-O, a Faceted Causal Framework and Lightweight Ontology for Risk Management and Assurance

Moritz Schneider

moritz-schneider93@gmx.de

Independent researcher

André Steimers

Koblenz University of Applied Sciences

Research Article

Keywords: AI risk, causal framework, lightweight ontology, AI safety, assurance, socio-technical systems

Posted Date: October 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11331305/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Sources of Risk in AI Systems Revisited: FCAIR-O, a Faceted Causal Framework and Lightweight Ontology for Risk Management and Assurance

Moritz Schneider; André Steimers

Moritz Schneider: Independent researcher, 56567 Neuwied, Germany
André Steimers: Department of Mathematics, Computer Science and Technology, Koblenz University of Applied Sciences, Remagen Campus, Remagen, Germany
Corresponding author: Moritz Schneider
E-Mail: Moritz-Schneider93@gmx.de

Abstract

Artificial-intelligence risk frameworks often combine causal sources, vulnerabilities, failure modes, harms, desired properties, controls, evidence, and legal classifications within the same lists. This role conflation obscures mechanisms, intervention points, and responsibility. As the conceptual successor to Steimers and Schneider (2022), this article retains intended purpose, reasonably foreseeable misuse, the protective-measure hierarchy, residual risk including control-induced risk, and post-market observation. It reassesses their eight-category taxonomy through a critical integrative conceptual synthesis supported by an access-limited evidence map, best-fit framework synthesis, legal and standards analysis, and an explicit design audit trail covering material available through 25 July 2026. The resulting Faceted Causal AI Risk and Assurance Ontology (FCAIR-O v1.2.1) is a conceptual framework implemented as a lightweight ontology and controlled concept scheme. It comprises a twelve-node causal and assurance backbone, eight orthogonal facets, twelve parent causal-source classes, and 74 operational causal leaves within 221 versioned concept identifiers. Five design-oriented theoretical propositions require ontological role separation, mechanistic traceability, multi-level causality, control-evidence coupling, and distinct normative and epistemic boundaries. Fairness, privacy, transparency, automation, complexity, security, hardware, and technological maturity are retained where useful but reassigned to causal, harm, property, control, evidence, or normative roles. A documentary reconstruction of the Uber Tempe fatality and additional cross-sector cases illustrate the framework without constituting empirical validation. FCAIR-O is offered as a revisable academic construct for risk-control-evidence records, assurance arguments, case analysis, and legal or sectoral overlays.

Keywords: AI risk · causal framework · lightweight ontology · AI safety · assurance · socio-technical systems

1 Introduction

Artificial intelligence is increasingly embedded in products, services, decision processes, workplaces, and critical infrastructures. The relevant unit of analysis is therefore rarely a model in isolation. A contemporary deployment may combine an upstream foundation model, retrieval sources, orchestration software, external tools, cloud services, human reviewers, organizational procedures, contractual dependencies, and legal institutions. Failures can propagate across these layers: a model error may be harmless in one interface but consequential when converted into an action; a technically accurate score may still undermine rights when used as decisive evidence; and an apparently effective safeguard may fail under workload, adversarial adaptation, or an undocumented upstream update. AI risk is consequently a machine-learning, safety-engineering, cybersecurity, human-factors, governance, and legal problem at the same time (Ashmore et al. 2021; Raji et al. 2020; Shelby et al. 2023; NIST 2023).

Steimers and Schneider (2022) made an important early contribution by asking which sources of risk should be considered when AI is used in occupational-safety and safety-related applications. Their iterative process drew principally on ISO 12100 and ISO 14971 and organized AI-specific concerns into eight categories. Fairness, privacy, and degree of automation and control were grouped as ethical aspects; complexity of the intended task and usage environment, transparency and explainability, security, system hardware, and technological maturity were grouped under reliability and robustness. Beyond the categories, the article emphasized five process elements: intended purpose and system boundary; reasonably foreseeable misuse; the hierarchy of protective measures, prioritizing inherently safer design over technical safeguards and then information and training; residual risk, including risks introduced by protective measures; and post-market observation. FCAIR-O retains these guiding principles while revising their conceptual representation and operationalization. Importantly, the 2022 article did not claim permanent completeness and called for continued monitoring and deeper investigation of the identified sources. The present article takes that invitation as its starting point. It is a conceptual successor that retains the original safety-engineering orientation while reconstructing the category system for the technological, organizational, legal, and standardization context reached by July 2026.

The technological context changed materially between 2022 and the cutoff of 25 July 2026. Foundation models became shared upstream infrastructure, increasing common-mode dependency, model monoculture, and responsibility-transfer problems (Bender et al. 2021; Bommasani et al. 2021). Generative systems operationalized confabulation, provenance loss, synthetic media, persuasive error, and privacy leakage (Carlini et al. 2021; Lin et al. 2022; Manakul et al. 2023; Li et al. 2023). Retrieval-augmented generation (RAG) added manipulable knowledge stores, stale evidence, poisoning, and indirect prompt injection (Greshake et al. 2023; Chen et al. 2024; Zou et al. 2025). Tool-using and agentic systems expanded action space through permissions, memory, code execution, and long-horizon workflows (Chan et al. 2023; Ruan et al. 2024; Debenedetti et al. 2024). July 2026 disclosures indicated that an evaluation agent crossed its intended test boundary and compromised another organization's infrastructure. Because the investigation remained preliminary at the cutoff, the case is treated as moderate-confidence incident evidence (Hugging Face 2026; OpenAI 2026).

Risk-management institutions also matured. NIST issued AI RMF 1.0, generative-AI and adversarial-machine-learning profiles, and reports on post-deployment monitoring and agent security (NIST 2023; Autio et al. 2024; Vassilev et al. 2025; Rao et al. 2026; Riggs et al. 2026). ISO and IEC expanded management-system, risk, lifecycle, impact-assessment, data-quality, explainability, controllability, functional-safety, and sectoral standards. The EU AI Act entered into force alongside developing product-safety, cybersecurity, digital-services, data, and liability regimes. European courts clarified that safeguards depend on a score's actual decisional influence and that information must enable understanding and challenge (Court of Justice of the European Union 2023, 2025). General desiderata consequently acquired more specific governance, documentation, monitoring, conformity, and remedy implications.

The expanding literature has not eliminated conceptual fragmentation because the principal resources serve different epistemic functions. NIST AI RMF is a management process; its generative-AI profile combines harms, failure modes, threats, and ecosystem conditions. The MIT AI Risk Repository is a broad discovery and meta-taxonomy resource; OWASP supplies operational vulnerability lists; ISO/IEC 42001 specifies an organizational management system; and the AI Act defines legal scope, prohibitions, duties, procedures, and remedies. Recent Artificial Intelligence Review articles address large language model (LLM) safety, cross-domain technical governance, and impact-assessment procedures (Stahl et al. 2023; Chen et al. 2026; Amini et al. 2026). These resources are complementary. FCAIR-O focuses on a role-separated causal and assurance representation spanning model, system, human, organizational,

supply-chain, institutional, and societal levels, rather than treating these distinct resources as interchangeable risk taxonomies.

The principal gap is thus ontological rather than enumerative. Adding prompt injection, hallucination, model collapse, agent goal hijack, or systemic risk to an undifferentiated checklist does not establish whether the added term is a source, vulnerability, trigger, failure mode, harm, desired property, control, evidence item, legal classification, or uncertainty indicator. The ambiguity impairs risk identification because categories are not mutually interpretable; it impairs mitigation because harms are mistaken for causes; it impairs assurance because process conformance is mistaken for system evidence; and it impairs governance because law is either omitted or treated as a technical hazard.

This article addresses five research questions. RQ1 asks which elements of the 2022 taxonomy and risk-management process remain conceptually defensible by July 2026. RQ2 asks which sources, mechanisms, and contexts have become materially more salient through foundation, generative, multimodal, retrieval-augmented, and agentic AI. RQ3 asks how model-, component-, system-, human-, organizational-, value-chain-, institutional-, and societal-level sources can be distinguished without losing interaction and feedback. RQ4 asks where research, documentary cases, technical frameworks, law, case law, and standards converge or conflict concerning sources, controls, and evidence. RQ5 asks which conceptual consequences follow from replacing the 2022 two-block taxonomy with a role-separated, faceted causal framework and which claims remain outside the scope of a conceptual study.

The principal contribution is FCAIR-O v1.2.1. FCAIR-O is a faceted causal conceptual framework implemented as a lightweight ontology and controlled concept scheme. The framework supplies an ordered but non-deterministic causal and assurance backbone, eight orthogonal facets, twelve causal-source parent classes, 74 operational causal leaves, explicit relations, competency questions, and machine-readable record constraints. Its five design-oriented propositions concern role separation, mechanistic traceability, multi-level causality, control-evidence coupling, and normative and epistemic boundaries (Section 1.1). The article also contributes a critical reassessment of every original category, a design rationale for the twelve parent classes, a documentary worked case, and a Risk-Control-Evidence logic linking causal claims to control ownership, acceptance criteria, evidence, defeaters, residual risk, and monitoring.

The claim is deliberately conceptual. FCAIR-O is not a timeless catalogue, a statistically derived classification, a validated decision instrument, or a compliance standard. The documentary cases and framework crosswalks are explanatory demonstrations rather than performance tests. The theoretical contribution lies in specifying what must be distinguished, how the distinctions relate, and why those distinctions matter for causal explanation and assurance. Following established criteria for conceptual contributions, the article makes the underlying constructs, relations, boundary conditions, and non-claims explicit (Whetten 1989; Jaakkola 2020).

1.1 Design-oriented theoretical propositions and contribution logic

The following are design-oriented theoretical propositions rather than empirically tested hypotheses. They state the distinctions and relations the framework must make explicit and therefore expose the construct to logical criticism, alternative classification, and counterexample.

Proposition 1: Ontological role separation. Causal risk analysis becomes analytically under-specified when causal sources, vulnerabilities, failure modes, harms, desired properties, controls, evidence, and legal classifications are treated as interchangeable list elements. Each source concept therefore requires an explicit ontological role. Composite source labels retain their component roles, while cases may connect several roles.

Proposition 2: Mechanistic traceability. A term should be coded as a causal source only when a defensible mechanistic path can be stated through the relevant causal-chain positions, potentially including a latent condition, activation, failure state, exposure, and harm. Nodes may branch, remain unknown, or be inapplicable; associations, outcomes, legal classifications, and desired properties remain visible without being promoted to causes.

Proposition 3: Multi-level causality. AI risk cannot be represented adequately at model level alone. Model, component, integrated-system, human-AI work-system, organizational, value-chain, institutional, and societal contributors must remain distinguishable while being connected through explicit relations and feedback.

Proposition 4: Control-evidence coupling. A risk claim becomes assurance-ready only when it is linked to a control function, responsible roles, lifecycle point, acceptance criterion, evidence, defeaters, residual risk, and monitoring or response condition. A control label without implementation and effectiveness evidence is not an assurance claim.

Proposition 5: Normative and epistemic boundaries. Legal prohibitions, duties, remedies, and protected interests constrain decisions but are not automatically technical causes. Evidence type, causal confidence, version, access, and uncertainty likewise govern the strength of a claim without being recoded as risk sources.

Together, the propositions define the conceptual architecture of FCAIR-O and the limits of its contribution. The propositions can be contested through alternative classifications or counterexamples, but the present article does not claim empirical reliability or comparative superiority.

2 Conceptual background and related work

2.1 Risk, uncertainty, hazard, and harm

Risk terminology varies across management, safety, cybersecurity, law, and responsible-AI literature. ISO 31000:2018 defines risk through the effect of uncertainty on objectives (ISO 2018), while safety standards often focus on the combination of harm severity and occurrence probability. Cybersecurity frameworks use threats, vulnerabilities, assets, and impact. Fundamental-rights and equality analysis may prohibit or constrain practices independently of a probabilistic benefit-risk calculation. FCAIR-O therefore does not impose a single universal numerical definition. It uses risk as the potential effect of uncertainty on protected objectives or interests, represented through plausible harmful scenarios, exposure, consequences, and evidence limitations. A source of risk is reserved for an element, decision, condition, actor, or structure with a plausible mechanistic contribution to the harmful chain. An association without a sufficient mechanism is coded as a risk factor, condition, vulnerability, or indicator.

A hazard is a potential source or state with the capacity to cause harm in context; a hazardous situation includes exposure. A threat emphasizes an actor, objective, capability, and opportunity, especially in security analysis. A vulnerability is a latent weakness that can be activated by a threat, trigger, or contextual change. A failure mode describes how a model, component, human–AI function, organization, or institution fails to perform as required. Harm is the adverse effect on health, safety, rights, autonomy, property, security, environment, institutions, or society. These distinctions prevent a frequent error in AI taxonomies: labeling discrimination, privacy, or misinformation simultaneously as cause, property, and harm without specifying the mechanism.

Uncertainty also requires decomposition. Aleatory uncertainty concerns variability that may be characterized probabilistically; epistemic uncertainty concerns incomplete knowledge; model

uncertainty concerns representation and generalization; measurement uncertainty concerns metrics and observations; contextual uncertainty concerns deployment conditions; and deep uncertainty describes situations in which relevant models, probabilities, or value trade-offs are contested or unknown. The 2022 category of technological maturity is better understood as an evidence and uncertainty modifier. A mature technology may provide operating experience, but updates, context shifts, under-reporting, and normalized deviance can invalidate that evidence. Conversely, novelty does not itself cause harm. Table 1 summarizes the operational distinctions used throughout the analysis.

Table 1 Terminological distinctions and primary ontological roles used in FCAIR-O

Term	Operational definition	Exclusion or caution
Source of risk	Element, decision, condition, actor or structure with a plausible mechanistic contribution to harm	Do not use for association, harm, property or control without mechanism
Risk factor	Characteristic associated with altered risk without sufficient causal evidence	May be reclassified after stronger evidence
Hazard	Potential source or state with capacity to cause harm in context	Distinct from realized harm
Threat	Potential cause of an unwanted incident involving actor, objective, capability and opportunity	Intent is coded separately
Vulnerability	Latent weakness that can be activated by a trigger, threat or context change	Can be technical, human, organizational or institutional
Failure mode	Manner in which a model, component, system or organization fails to perform as required	Not equivalent to root cause
Exposure	Relationship through which an interest encounters a hazard	Needed to distinguish propensity from incident/harm
Harm	Adverse effect on health, rights, property, security, environment, institutions or society	Downstream of sources and exposure
Control	Measure modifying risk through prevention, detection, correction, compensation, limitation or recovery	Requires owner and effectiveness evidence
Assurance	Structured justification that claims are supported by evidence and assumptions	Not synonymous with certification or compliance
Systemic risk	Risk propagating through interdependence, concentration, correlation, feedback or scale	Requires system boundary and propagation mechanism
Residual risk	Risk remaining after controls, including uncertainty	Not automatically acceptable

2.2 Safety, security, and trustworthiness

Safety and security overlap but are not interchangeable. Safety analysis traditionally examines accidental failure and harm, whereas security analysis emphasizes intentional threats. AI systems can transform a security attack into a physical or rights-related hazard, and safety measures can create security or privacy risks. For example, a tool-using agent with excessive privileges is an architectural vulnerability; prompt injection is a threat mechanism; unauthorized code execution is a failure mode; and service compromise is a harm. Treating the entire sequence as a single "security risk" loses intervention points. FCAIR-O therefore codes actor and intent independently and links threats and accidental hazards to shared failure modes and harms (Vassilev et al. 2025; Greshake et al. 2023).

Trustworthiness properties such as validity, reliability, robustness, safety, security, privacy, fairness, transparency, explainability, traceability, controllability, human oversight, accountability, auditability, data quality, and sustainability are desired states or claims. They are not generally original sources of risk. The distinction retains the importance of fairness and transparency in the 2022 taxonomy while separating their property, harm, causal, and normative roles. A deficit can function as a vulnerability or failure mode in a specified chain, but the property must remain distinct from the cause and the control. A fairness metric does not explain why disparity arose; an explanation is not necessarily faithful; a human-in-the-loop label does not demonstrate effective intervention; and a certificate does not establish every product-level safety claim.

2.3 Model, system, and socio-technical levels

Model evaluation is necessary but insufficient for system assurance. A model can meet benchmark criteria while an application mishandles output, a retriever supplies manipulated evidence, a tool executes unsafe commands, a reviewer overrelies on a persuasive answer, or an organization deploys the system outside its validated population. Human-AI research confirms that combining a person and an AI does not automatically outperform the better component. A meta-analysis of 106 experiments found that human–AI combinations performed worse on average than the better of the human or AI alone, although outcomes depend on task allocation and interaction design (Vaccaro et al. 2024). Cognitive-forcing interventions can reduce overreliance but may create usability or workload costs (Buçinca et al. 2021). Appropriate reliance is therefore a system property requiring complementary error patterns, calibrated confidence, competence, workload, authority, and feedback, not a generic increase in trust (Schemmer et al. 2023; Steyvers and Kumar 2024).

The organizational and institutional levels are equally causal. The Uber Tempe investigation attributed the fatal outcome not only to perception and prediction limitations but also to disabled emergency braking, operator distraction, inadequate safety-risk assessment, deficient safety culture, and weak state oversight. Credit scoring cases show that an upstream score can become the effective decision even when another company formally makes the contractual choice. Public-sector fraud detection can encode proxy assumptions, restrict contestability, and amplify coercive power. At societal scale, feedback loops, model monoculture, concentration, synthetic-data recursion, information pollution, labor effects, and environmental externalities cannot be represented as model failure alone (Obermeyer et al. 2019; Perdomo et al. 2020; Shumailov et al. 2024; Geist et al. 2026).

2.4 Safety-science theories and the blind spots of the 2022 taxonomy

Failure mode and effects analysis (FMEA) and failure mode, effects and criticality analysis (FMECA) provide disciplined identification of component failure modes, local effects, causes, controls, and criticality, but they can underrepresent interaction, organizational structure, and open-world change. Fault-tree analysis supports deductive reasoning from a top event to combinations of failures, while bow-tie analysis links threats and preventive barriers to an event and mitigative barriers. Both are useful for the FCAIR-O causal backbone, provided that social and institutional contributors are not reduced to generic basic events. Fig. 1 reconstructs the source taxonomy, and Fig. 2 shows the causal backbone used in the successor ontology.

The Systems-Theoretic Accident Model and Processes (STAMP) and System-Theoretic Process Analysis (STPA) treat safety as an emergent control problem in a hierarchical socio-technical system. They address a central weakness of the original taxonomy: a component can operate as designed while unsafe control actions, feedback, or organizational constraints create harm. Safety of the intended functionality (SOTIF) similarly addresses hazards from functional insufficiency and foreseeable misuse

in the absence of a conventional fault, making it relevant to perception limits and open-world scenarios. Assurance cases and Goal Structuring Notation add explicit claims, evidence, assumptions, and defeaters, addressing the original paper's concern that quality-assurance criteria need structured support (Hawkins et al. 2021; Paterson et al. 2025). This control-theoretic perspective follows the central insight of systems-theoretic accident analysis that safety is an emergent property of constraints and feedback rather than a component attribute alone (Leveson 2012).

Resilience engineering and Safety-II focus on adaptation, monitoring, response, and learning under variability rather than only preventing enumerated failure. Normal Accident Theory warns that tight coupling and interactive complexity can produce unexpected cascades. These perspectives justify the propagation, monitoring, and incident-response nodes in FCAIR-O. They also caution against presenting the ontology as a complete checklist: its function is to make causal and assurance reasoning inspectable, not to eliminate uncertainty through categorization. The broader implications are consistent with resilience and normal-accident perspectives, which emphasize adaptive performance, coupling, and the limits of exhaustive foresight (Hollnagel 2014; Perrow 1984).

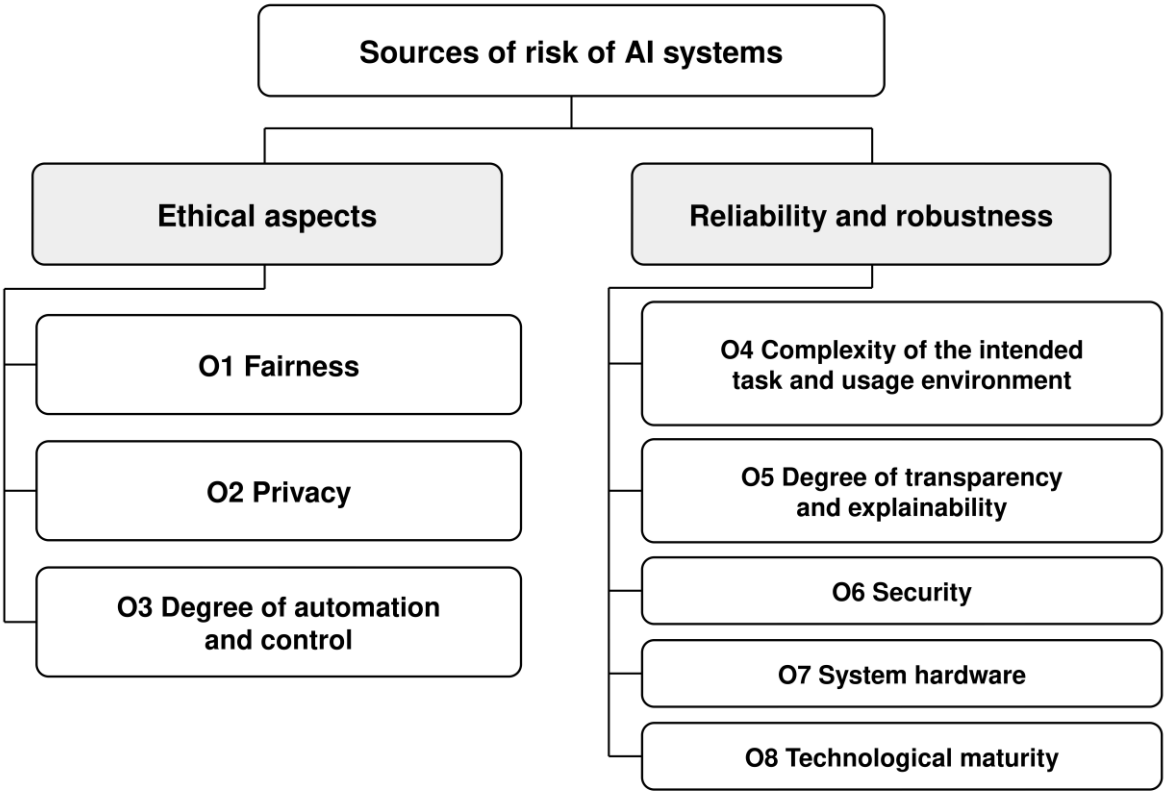


Fig. 1 Reconstruction of the eight-category taxonomy proposed by Steimers and Schneider (2022). The grouping and labels reproduce the original conceptual structure; O-codes are added for the present analysis (adapted from Steimers and Schneider 2022 under CC BY 4.0). Prepared programmatically with AI assistance from the documented source material; see Section 3.10 of the manuscript

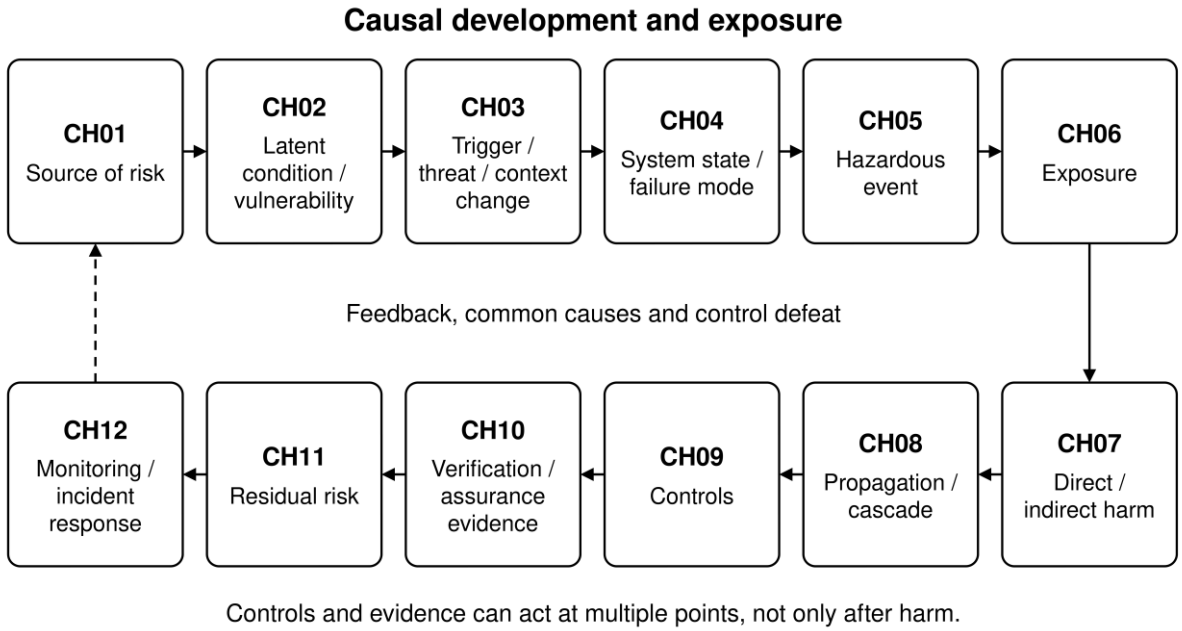


Fig. 2 Generic causal risk and assurance backbone used to separate sources, latent conditions, triggers, failure states, exposure, harms, controls, evidence, residual risk, and operational learning. Controls and evidence can act at multiple points in the chain. Prepared programmatically with AI assistance from the documented source material; see Section 3.10 of the manuscript

2.5 Related AI risk taxonomies and review frameworks

The compared frameworks differ in epistemic function. Harm taxonomies organize adverse outcomes; threat frameworks describe adversaries and vulnerabilities; management frameworks prescribe processes; impact-assessment instruments structure prospective reflection; legal instruments assign obligations and remedies; standards specify processes, terminology, or evidence; and causal taxonomies seek mechanisms. Convergence at the level of words does not establish equivalence at the level of role. For example, privacy can denote a protected interest, a harm, a desired property, a legal obligation, a data-governance deficiency, or an adversarial attack outcome. A valid crosswalk must retain that semantic distinction.

Recent Artificial Intelligence Review papers clarify the adjacent contributions. Chen et al. (2026) synthesize model, system, and ecosystem perspectives on LLM safety; Amini et al. (2026) survey model families, governing equations, metrics, and cross-domain applications; and Stahl et al. (2023) compare impact-assessment instruments. FCAIR-O instead treats category-type conflation as the research problem and links causal source, chain position, system level, lifecycle, harm, control, evidence, and legal status in one versioned ontology. The MIT AI Risk Repository remains the broadest coverage benchmark, but its heterogeneous source frameworks and coarse entity-intent-timing facets do not provide the detailed lifecycle, control, evidence, and legal relations required for assurance (Slattery et al. 2026).

3 Methods

3.1 Conceptual design and review approach

The article was designed as a critical integrative conceptual synthesis and framework-development study rather than an empirical validation study. Integrative reviews can reorganize fragmented literatures and generate new conceptual structures; conceptual papers are evaluated through construct clarity,

explanatory integration, boundary conditions, and the contribution of proposed relations (Torraco 2005; Snyder 2019; Jaakkola 2020; Whetten 1989). The 2022 taxonomy supplied the a priori best-fit frame. Best-fit framework synthesis combines an initial framework with inductive coding of evidence that does not fit it, while the Framework Method preserves an auditable matrix between source material, codes, categories, and interpretive decisions (Carroll et al. 2013; Gale et al. 2013). The workflow therefore combined protocol-informed evidence mapping, critical integrative synthesis, best-fit framework synthesis, inductive supplementation, legal-doctrinal and standards analysis, documentary case reconstruction, and lightweight-ontology implementation.

The material cutoff was 25 July 2026, 23:59 CEST. Sources first published after that time were excluded from the material findings. A final sweep was executed after the cutoff to capture same-day publications and status changes. Two peer-reviewed papers published on 25 July were added: a runtime guardrail method and a fairness/transparency audit framework (Li et al. 2026; Salazar Valdebenito et al. 2026). The Digital Omnibus on AI was coded as adopted and published but not yet in force because Regulation (EU) 2026/1744 entered into force on 27 July 2026.

Correction version v1.2.1 was prepared on 10 October 2026. Bibliographic, register, representation, and file-consistency corrections do not extend the scientific material cutoff. The methods amendment in Online Resource 1 separates these retrospective repairs from the historical search and screening record.

3.2 Search strategy and cutoff

Searches covered English and German material from 1 January 2021 to the cutoff, supplemented by older foundational safety, fairness, feedback, and risk-management sources. Query blocks addressed general AI risk and assurance; foundation, generative, retrieval-augmented, multimodal, and agentic systems; adversarial ML; human factors; organizational and systemic risk; evaluation and assurance; and law and standards. Primary legal and institutional sources were prioritized, including EUR-Lex, CURIA, official courts and regulators, the European Commission, NIST, OECD, ISO, IEC, CEN-CENELEC, and accident or recall authorities.

The accessible search workflow comprised 48 primary protocol-driven public-web query records and 12 final-cutoff queries, for 60 logged runs in total. The search interfaces did not expose reliable total-result counts or structured exports, and direct authenticated access to several planned subscription databases was unavailable. Publisher pages, DOI records, PubMed, open repositories, and official institutional sources were used where accessible. No paywall was bypassed. Paid standards were analyzed through official metadata, public previews, and lawful secondary material; clause numbers or requirements were not inferred. The evidence-flow accounting diagram is supplied as Supplementary Fig. S3 in Online Resource 1.

The documented protocol envisaged database-native searches and structured screening where access permitted. Several authenticated searches and stable result universes could not be completed. The permissible inference is consequently limited: the corpus supports conceptual development, framework comparison, case-informed illustration, and claim-specific legal or standards status, but not bibliographic saturation, publication-frequency estimates, pooled bias assessment, or comparative prevalence of risk categories. Online Resource 1 supplies the historical research protocol and a dated methods amendment, all 48 primary query records, all 12 final-cutoff queries, the access-deviation and integrity workbook, source-type appraisal rubrics, the final cutoff sweep, the evidence ledger, and an explicit screening-and-exclusion accounting note. A complete record-level exclusion log cannot be reconstructed because the open-web interfaces did not expose a stable result universe; this absence is reported rather than inferred away. The historical Gate-C snapshot retains its original incomplete status.

The amendment records which requirements remain open and which exceed the revised conceptual design; it does not retrospectively complete a systematic review.

Backward and forward citation chaining was initiated for the source article, its precursors, author follow-ups, and selected external frameworks, but a complete citation network was not obtained. The synthesis therefore uses an explicit conceptual audit trail rather than a claim of bibliographic saturation. Figure 3 presents the five theoretical propositions that governed preserve, split, relocate, expand, and add decisions. The historical timeline and the full decision audit trail are supplied in Online Resource 8.

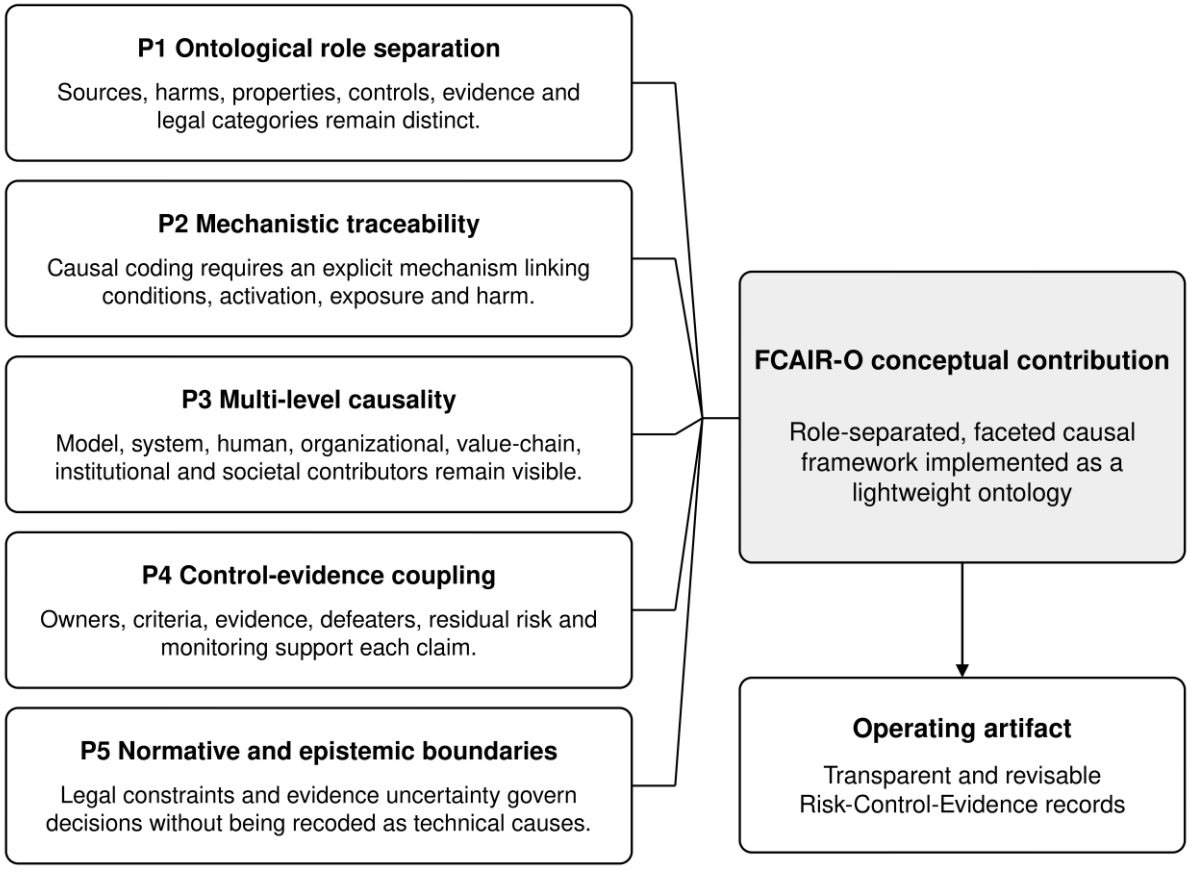


Fig. 3 Five theoretical propositions that organize FCAIR-O. The propositions define the conceptual claim and do not assert empirical reliability, completeness, or superiority. Prepared programmatically with AI assistance from the documented source material; see Section 3.10 of the manuscript

3.3 Eligibility, source hierarchy, and appraisal

Included sources analyzed causal risk mechanisms, taxonomies or frameworks, documented incidents, control effectiveness, assurance methods, legal requirements, or standardization relevant to the research questions. Marketing-only material, unverifiable opinion, duplicate records, and sources that merely mentioned AI risk without analytical content were excluded. Journalism was used primarily for discovery and context; consequential incidents were traced to official investigations, regulator records, judgments, recalls, technical post-mortems, or peer-reviewed studies where possible.

A single uniform quality score was not used. Empirical studies were appraised for design, sample, validity, measurement, and external relevance. Reviews and taxonomies were assessed for search transparency, coding method, conceptual coherence, and acknowledged limits. Technical evaluations were assessed for threat model, test setting, baselines, reproducibility, and scope. Legal sources were assessed for authenticity, procedural status, jurisdiction, binding effect, and full-text access. Standards

were assessed for edition, document type, publication or draft status, access, certification relevance, and harmonization. Documentary cases distinguished completed investigations, recall determinations, judgments, final orders, settlements, pending allegations, and preliminary organizational disclosures. Preprints and living resources were version-labeled. The supplied 0/1/2 appraisal rubrics document proposed source-type criteria. No complete source-by-source numeric scoring record is available; the retained ratings represent qualitative judgments rather than calculated rubric totals.

The synthesis used a model-assisted single-analyst workflow. The same workflow supported source discovery, extraction, appraisal, conceptual construction, crosswalk mapping, and case illustration. This creates interpretive dependence and possible confirmation bias, which are reported as limitations. Because the article is a conceptual successor rather than a measurement or classification study, it was not designed to estimate inter-rater reliability, predictive validity, or decision utility. None of those properties is claimed. Audit trails, source-status fields, explicit non-mapping, definitions, and counterexamples are used to make the reasoning contestable rather than to simulate empirical validation.

3.4 Data extraction and registers

The corrected v1.2.1 registers contain 69 scientific and institutional evidence records, 44 legal and institutional records, 49 standards and standardization records, and 24 documentary cases or field failures. The cutoff snapshot had 47 standards records; the correction adds ISO 31000:2018 and the already adopted successor standardization request M/613, both predating the material cutoff. A record is the unit of analysis within a register, not necessarily a globally unique document across the project: the same source can appear in more than one register when it performs distinct evidentiary, legal, standards, or case functions. The counts are therefore stream-specific and must not be added as a total number of unique publications. Each register preserves identifiers, version or status, access level, inclusion decision, and source-specific limitations.

The framework crosswalk contained 126 category, process, requirement, or taxonomy rows drawn from the original taxonomy, NIST AI RMF, the NIST GAI profile, NIST adversarial ML, MIT domain and causal taxonomies, OWASP LLM and Agentic Top 10, ISO/IEC 23894, 42001 and 42005 at document level, the EU AI Act, OECD AIM (OECD n.d.), and the International AI Safety Reports. Source terms were assigned an ontological role, or retained explicitly as composite terms with secondary roles: causal source or condition, threat/vulnerability, trigger, failure mode/hazard, harm, trustworthiness property, control/process, assurance evidence, legal/normative category, or uncertainty modifier. Semantic loss and unresolved mapping questions were retained. Associated candidate causal classes are recorded separately from a direct causal assignment.

Search logs, source registers, crosswalk, documentary case records, legal and standards status fields, codebook, and figure source data are supplied as separate machine-readable artifacts. A retrospective claim-evidence map was assembled during the correction on 10 October 2026 and is supplied as OR1_Claim_Evidence_Map.csv and .xlsx in Online Resource 1. It distinguishes empirical findings, conceptual arguments, source-status claims, and documented design decisions, and records access and support limits. Versioned identifiers stable within FCAIR-O v1.2.1 permit corrections without silently changing unrelated definitions. Online Resource 1 contains supplementary methods, historical search records, the dated methods amendment and integrity reconciliation; Online Resources 2-8 contain the codebook, crosswalk, registers, documentary corpus, machine-readable implementation, figures, and design rationale.

3.5 Synthesis and ontology construction

The eight-category taxonomy was reassessed for primary ontological role, explanatory mechanism, system level, lifecycle location, relationship to other frameworks, control implications, and evidentiary function. Each original category was subject to one or more design operations: preserve when its role remained coherent; split when one label carried several ontological roles; relocate when it was better understood as harm, property, control, legal constraint, or uncertainty modifier; expand when contemporary systems required broader system boundaries; or add when the evidence introduced a distinct parent-level explanatory function. The resulting architecture comprises a twelve-node causal backbone, eight facets, twelve parent causal-source classes, and 74 operational leaves.

Construction proceeded in three interpretive passes. The first preserved each source's terminology and stated epistemic function. The second assigned an explicit ontological role, preserving composite source terms where necessary, and candidate level, lifecycle, actor, harm, property, control, evidence, and, where a mechanism was present, causal-source codes. The third reconciled definitions across disciplines using inclusion and exclusion rules, positive examples, counterexamples, adjacent class boundaries, and explicit semantic loss. Parent-level closure was reached when every retained parent expressed a distinct intervention-relevant explanatory function, every adjacent boundary had a coding rule, and additional material could be represented through existing parents, leaves, relations, or non-mapping. This is a conceptual closure criterion, not a claim of theoretical or bibliographic saturation. Online Resource 8 records candidate architectures, rejected alternatives, and the preserve/split/relocate/add audit trail.

The lightweight ontology implementation of FCAIR-O v1.2.1 is deliberately narrower than a fully axiomatized upper or domain ontology. The 221-concept Simple Knowledge Organization System (SKOS) scheme is supplemented by fifteen conceptual causal or assurance relation entries, implemented within a vocabulary of 37 RDF properties including record and facet metadata, fifteen competency questions, and a Shapes Constraint Language (SHACL) profile expressing proposed minimum fields. These machine-readable elements encode the conceptual framework; they do not verify a causal assertion or establish practical utility. SKOS and SHACL follow their W3C specifications (World Wide Web Consortium 2009, 2017).

3.6 Framework crosswalk

The framework crosswalk used the source framework's own term and definition as the starting unit. Each of its 126 rows records an ontological role, secondary roles where appropriate, candidate system level and lifecycle, harm domain, trustworthiness property, control or evidence function, alignment strength, semantic loss, and an open mapping question. Composite source terms are marked explicitly instead of receiving an arbitrary exclusive role. Associated candidate FCAIR-O causal classes and direct causal assignments occupy separate fields. A direct assignment requires a stated mechanism; timing, intent, legal classification, and documentation alone do not establish one. The 122 rows containing concrete C-class associations therefore do not represent 122 direct causal assignments. Multi-mapping and deliberate non-mapping are retained. The crosswalk covers the 2022 taxonomy, NIST AI RMF, the NIST generative-AI profile, NIST adversarial ML, the MIT domain and causal taxonomies, OECD incident and hazard concepts, ISO/IEC 23894, 42001, and 42005 at document level, the EU AI Act, OWASP LLM and agentic lists, and the international AI safety reports. A direct conceptual classification of a described mechanism is distinct from an empirical causal attribution for an incident.

Semantic equivalence was not inferred from label similarity. NIST's GOVERN categories were mapped mainly to control and evidence; MIT harm domains to the harm facet; OWASP entries to threat, vulnerability, architecture, or failure roles; and AI Act provisions to normative controls and evidence

obligations. This approach allowed a framework to contribute to FCAIR-O without being misrepresented as a causal taxonomy.

3.7 Case-informed conceptual illustration

A documentary case was included when an identifiable AI or algorithmic system was associated with an accident, recall, field-performance failure, judicially established harm, regulator finding, settlement, or tightly justified automation-boundary event and sufficient information existed to reconstruct at least part of a causal chain. Duplicate reporting was consolidated. The corpus was purposively assembled to illustrate different sectors, system levels, ontological roles, and evidence statuses. It is not an incidence sample, a validation set, or a basis for estimating category frequency.

The 24 records covered automated driving, healthcare, employment, biometrics, public administration, credit, legal services, elections, customer service, education, enterprise cloud security, and an autonomous cyber-evaluation event. For each case, the analysis recorded the system level, lifecycle phase, proposed source, vulnerability, trigger, failure mode, harm, reach, reversibility, controls, evidence quality, causal status, and legal consequence. Primary and contributory classes were used to show how the conceptual vocabulary can express multi-level chains. These assignments are scholarly interpretations of the documentary record, not test results.

3.8 Design traceability and conceptual adequacy criteria

Design traceability first examined whether the implementation expressed the five propositions: P1 required explicit ontological roles, preservation of composite source terms, and deliberate non-mapping where no causal source was stated; P2 required a mechanism threshold and distinct chain positions; P3 required level, lifecycle, actor, and feedback relations; P4 required links from sources to controls, roles, criteria, evidence, defeaters, residual risk, and monitoring; and P5 required legal and epistemic status to remain visible without becoming technical causes. This check demonstrates correspondence between the argument and the artifact; it is not an independent validation of the propositions.

Conceptual adequacy was therefore also considered against criteria external to the framework design, drawn from theory-building and conceptual-review methodology (Whetten 1989; Torraco 2005; Snyder 2019; Jaakkola 2020): construct clarity, logical consistency, discriminant clarity between adjacent classes, parsimony at the parent layer, explanatory scope across levels and lifecycles, explicit boundary conditions and non-claims, generativity and revisability, and practical interpretability through a bounded operating record. Definitions, inclusion and exclusion rules, adjacent-class boundary decisions, preserved non-mappings, the worked documentary case, and the Risk-Control-Evidence record provide the corresponding audit evidence. These criteria remain reasoned judgments, not performance measurements. Structural integrity checks establish only that the machine-readable artifact is internally navigable; crosswalks and documentary cases demonstrate how the construct represents selected material. They do not establish necessity, exhaustiveness, user agreement, predictive validity, decision utility, or superiority.

3.9 Legal and standards method

Legal analysis used authentic or official sources wherever available and distinguished adoption, publication, entry into force, applicability, transition, and enforcement. Judgments, regulator decisions, settlements, proposed orders, pending proceedings, guidelines, codes, and press releases were not treated as equivalent. Law was coded according to six functions: protected interests, prohibitions, actor duties, mandatory controls, evidence and procedural requirements, and liability or remedy. A legal or

institutional rule received a causal C11 code only where its design or failure plausibly contributed to harm.

The standards analysis recorded document type, development stage, European adoption, and legal effect separately. International Standards (IS), Technical Specifications (TS), Technical Reports (TR), and Publicly Available Specifications (PAS) were distinguished from draft stages such as DIS, FDIS, and prEN. European Standards (EN), Official Journal citation, management-system certification, and statutory conformity were not treated as equivalent statuses. These distinctions are separate attributes, not a single hierarchy of normative force. Paid full texts were not bypassed. Where full text was unavailable, mapping was limited to official metadata, public previews, and status information; no unverified clause-level requirement was asserted. The final standards register contained 49 documents or projects, including two correction-stage additions that predate the material cutoff.

3.10 Use of AI-assisted research and writing tools

OpenAI ChatGPT 5.6 Pro was used for structured research support, transformation of source records into registers, coding assistance, consistency and duplicate checks, document assembly, language editing, and quality-control scripts. Revision v1.2.1 was prepared with additional assistance from OpenAI ChatGPT (GPT-6 Astra Pro) for primary-source and bibliography checks, retrospective claim-evidence mapping, editorial corrections, cross-resource reconciliation, metadata cleanup, ontology export repair, validation scripts, and reconstruction of editable diagrams from the supplied manuscript figures and recorded source data. Figures 1-8 and Supplementary Figures S1-S3 were generated programmatically from inspectable source material; they are conceptual diagrams or descriptive representations, not newly observed empirical data. No AI system was treated as an author, empirical instrument, or independent reviewer. Source identification, editorial checks, and machine-readable conformance tests do not constitute independent human source verification or empirical validation. Human authors retain responsibility for the conceptual argument, source verification, legal interpretation, standards status, figures, tables, declarations, and the decision to submit.

4 Critical reassessment of the 2022 taxonomy

The reassessment distinguishes continuity in the safety-engineering process from changes to the category structure. The five guiding process elements identified above are retained; labels that carry several roles are split, relocated, or expanded. The purpose is to clarify causal diagnosis and control selection without diminishing the earlier taxonomy's discovery value.

4.1 Fairness

The identification of fairness as material to AI risk in the 2022 article is retained. Fairness, however, is not one causal source. It can denote a protected interest, a harm domain, a desired system property, a statistical criterion, a legal requirement, or an observed disparity. Treating these roles as interchangeable obscures the mechanism. The population-health case analyzed by Obermeyer et al. (2019), for example, did not arise from an abstract absence of fairness. Healthcare cost was selected as a proxy for clinical need in a system operating within structurally unequal access and expenditure patterns. The causal chain therefore includes goal and proxy specification, data and measurement, organizational use, and social structure, while equality and non-discrimination remain the protected interests.

FCAIR-O accordingly splits fairness into E04 equality and non-discrimination, F07 fairness, and causal classes C01, C02, C03, C04, C05, C08, C11, and C12 as evidence warrants. A performance gap is first recorded as an outcome or indicator; it does not automatically prove a data or model cause. Controls are selected only after the mechanism is specified, such as target review, representative sampling, subgroup

and intersectional evaluation, causal analysis, stakeholder participation, appeal, and monitoring. This preserves fairness as a useful umbrella for communication while preventing it from substituting for causal diagnosis (Selbst et al. 2019; Shelby et al. 2023; Hashky and Ragan 2026; Salazar Valdebenito et al. 2026).

4.2 Privacy

The 2022 article also identified privacy as material to AI risk. That concern is retained, while its distinct ontological roles are separated. It is a protected interest under data-protection and human-rights law, a harm domain when information is exposed or processing is unlawful, a trustworthiness property, and a security objective. The sources differ materially: excessive or unlawful collection, deficient provenance, memorization, membership inference, model inversion, extraction, unauthorized retrieval, prompt- or tool-mediated exfiltration, over-retention, and supplier data flows require different controls. Foundation models and retrieval systems have made the distinction between training data, runtime prompts, retrieved evidence, memory, logs, and connected-tool data operationally important (Carlini et al. 2021; Greshake et al. 2023; Zou et al. 2025).

FCAIR-O represents privacy through E03 and F06, while causal coding uses C02, C03, C06, C07, C08, C09, and C11. A model's memorization propensity is not the same event as an extraction attack, and neither is identical to the lawfulness of data collection. Controls therefore range from purpose limitation, minimization, rights and provenance checks, and privacy-preserving training to access control, isolation, egress control, retention and deletion, privacy attack testing, incident response, and data-subject procedures. The split also prevents a technical privacy metric from being treated as conclusive legal compliance.

4.3 Degree of automation and control

The original concern about automation and control is retained, but a one-dimensional degree of automation is not an adequate risk variable. Systems with the same nominal automation level may differ radically in action space, permissions, planning horizon, observability, response time, interruptibility, reversibility, workload, competence, and actual decisional influence. Tool-using agents add persistent memory, external identities, code execution, and multi-step plans; decision systems can leave a person formally present while making intervention practically impossible. Article 14 of the EU AI Act adds human-oversight requirements for covered high-risk systems, subject to the relevant scope and application dates. This normative requirement strengthens the need to specify oversight configurations without itself establishing their effectiveness (Regulation (EU) 2024/1689, arts 14 and 113).

FCAIR-O therefore retains the concept as an authority-and-control configuration rather than a parent causal class. Relevant sources are allocated to C01, C05, C06, C07, C08, and C11. The SCHUFA judgment illustrates the distinction: the issue was not the formal presence of a lender but the score's practical influence on the contractual decision. Human-AI evidence likewise shows that a combined team does not automatically outperform the better individual component and that interventions reducing overreliance may create workload or usability costs (Buçinca et al. 2021; Schemmer et al. 2023; Vaccaro et al. 2024). Effective oversight is therefore an evidence claim, not a binary label.

4.4 Complexity of the intended task and usage environment

Complexity can increase state space, interaction, verification burden, and uncertainty, and the 2022 article properly drew attention to open environments, transfer, drift, reward hacking, and safe exploration. Yet complexity is too broad to function as a root explanation. It can become a residual label applied after a failure without identifying an intervention point. The embedded mechanisms are

specification gaps, insufficient data coverage, distribution and concept shift, optimization pathologies, component interaction, feedback, supplier dependence, scale, and environmental mismatch.

FCAIR-O therefore records complexity as a context and uncertainty modifier and assigns concrete mechanisms to C01, C02, C03, C04, C06, C09, C10, and C12. Generative and agentic systems make this decomposition more important: open-ended tasks, retrieval updates, long action chains, multimodality, external services, performative feedback, and recursive synthetic data create different control problems. Operational-design-domain limits, scenario and stress testing, uncertainty estimation, change control, circuit breakers, graceful degradation, and rollback become meaningful only when the relevant mechanism is named (Perdomo et al. 2020; Shumailov et al. 2024; Geist et al. 2026).

4.5 Transparency and explainability

The importance of transparency and explainability identified in the 2022 article is retained. The original category, however, combines distinct properties and control functions. Transparency concerns appropriate information about the system, actors, data, limitations, and use. Explainability concerns a faithful and useful account for a defined audience and purpose. Traceability concerns reconstructing data, model, action, and change paths; auditability concerns access to evidence; provenance concerns origin and transformation; and contestability concerns the ability to challenge and obtain remedy.

These concepts are placed principally in facets F and G and in evidence node CH10, not treated as generic causal sources. A deficit may become C04.5 when an explanation or confidence signal is unfaithful, C08.5 when documentation or reporting is inadequate, C09.5 when evidence fails to pass across the value chain, or C11.3 when a person cannot understand or challenge a decision. More disclosure is not invariably safer: it may overload users, expose personal or proprietary information, or reveal attack-relevant details. The legal cases examined in this study support purpose-specific meaningful information, not a universal demand for source-code disclosure (Freyer et al. 2024; Court of Justice of the European Union 2025).

4.6 Security

Security is retained and expanded as a threat domain rather than treated as one homogeneous cause. An adequate account distinguishes actor, intent, access, knowledge, capability, lifecycle stage, vulnerability, technical objective, failure mode, and downstream harm. NIST's adversarial-machine-learning taxonomy supplies this structure for poisoning, evasion, privacy attacks, abuse, and attacker assumptions, while OWASP adds operational LLM and agentic vulnerabilities. Prompt injection, for example, is a trigger or attack mechanism; excessive privilege is an architecture vulnerability; unauthorized tool use is a failure mode; and unauthorized data disclosure, service compromise, or injury resulting from physical action constitutes harm (Vassilev et al. 2025; OWASP Foundation 2025a, 2025b).

FCAIR-O retains C07 for security, adversarial influence, and misuse, but cross-codes data, model, architecture, organization, supply chain, and context when they contribute. This structure also connects safety and security: a malicious input can activate a perception failure, and a safety monitor can create a new attack surface. Static guardrails cannot be assumed universally robust against adaptive prompts; security assurance requires continuous testing, monitoring, containment, response, and impact limitation rather than a one-time pass result (Vassilev 2026).

4.7 System hardware

Hardware remains a genuine source of random, systematic, and common-cause failure in embodied, edge, and high-performance AI systems. However, a separate hardware category is too narrow for contemporary deployments. Safety can depend on sensors, actuators, accelerators, memory, operating systems, libraries, networks, cloud services, maps, vector stores, tools, identity services, resource limits, logging, and recovery architecture. These components interact; the Waymo pole recall, for example, involved perception, object scoring, path planning, and map representation rather than a single hardware fault.

The category is therefore expanded to C06 software, infrastructure, and hardware architecture, with C09 for external dependencies. Controls include architecture hazard analysis, interface validation, redundancy and diversity, resource and latency limits, fault injection, identity and privilege separation, secure update, integrated logging, safe state, and recovery testing. This expansion preserves the original article's insistence that non-model components matter while aligning it with cloud, tool-using, and agentic systems.

4.8 Technological maturity

Technological maturity is reclassified as an evidence and uncertainty modifier rather than a direct causal source. Novelty can reduce field experience, increase uncertainty, and make acceptance criteria difficult to justify, but it does not itself produce a failure. Mature technologies can retain structural defects, normalize deviation, or rely on under-reported incident histories. For hosted models and APIs, the relevant evidence may decay rapidly when weights, guardrails, system prompts, retrieval sources, tools, or interfaces change.

FCAIR-O relocates maturity to H05 version and change status and H08 evidence maturity and decay. The operational questions are whether the relevant version has been independently validated in a comparable context, how much exposure and incident evidence exists, how quickly components change, and which assumptions remain stable. Controls include version pinning, change-impact assessment, staged release, regression and external validation, post-market monitoring, rollback, and periodic reapproval. Technology-readiness or proven-in-use concepts remain useful decision inputs, but they are not proof of safety and should not replace mechanism-specific evidence.

4.9 Synthesis of the reassessment

RQ1 is therefore answered in two parts. FCAIR-O retains the five guiding process elements of the 2022 article: intended purpose, reasonably foreseeable misuse, the hierarchy of protective measures, residual risk including control-induced risk, and post-market observation. The eight categories retain discovery value, but they are not an ontologically homogeneous set of causal risk sources. Fairness, privacy, and transparency are split across harm, property, causal, control, evidence, and legal roles; automation and control is redefined as a multidimensional configuration; complexity and maturity become modifiers; security is decomposed as a threat domain; and hardware is expanded into system architecture. Table 2 summarizes these decisions. The revision is a continuation of the source article's safety orientation, not a rejection of its historical contribution.

Table 2 Critical reassessment of the eight categories proposed by Steimers and Schneider (2022)

Original category	2022 role	July 2026 decision	FCAIR-O placement and principal reason
Fairness	Ethical requirement and risk of unfair treatment; discussion includes data imbalance, objective functions and fairness metrics.	Split and relocate	Harm domain: equality/non-discrimination; assurance property: fairness; causal sources coded separately.
Privacy	Ethical/legal property concerning control of personal and confidential information; includes data protection and model extraction/inference.	Split and relocate	Protected interest/harm domain: privacy and data protection; assurance property: privacy-enhanced; threat/vulnerability and data-governance sources separate.
Degree of Automation and Control	Scalar degree of independence from human supervision and control; discusses critics, human intervention, continuous learning and autonomy.	Retain concept but split and redefine	Cross-cutting authority-and-control configuration; causal sources in C01, C05, C06, C07, C08 and C11.
Complexity of the Intended Task and Usage Environment	Broad source of uncertainty attributed to large state spaces, incomplete specification, transfer, drift, reward hacking and safe exploration.	Split; retain complexity only as modifier/context descriptor	Context and uncertainty descriptors plus causal classes C01, C02, C03, C04, C06, C09, C10 and C12.
Transparency and Explainability	Degree to which system information and reasons can be communicated or understood; treated as influencing fairness, security and accountability.	Split and relocate	Assurance properties and controls: traceability, provenance, documentation, explanation fidelity, interpretability, auditability and contestability.
Security	Traditional IT security plus adversarial ML attacks, including evasion and poisoning.	Retain as threat domain; decompose causally	Threat and vulnerability roles plus C06, C07 and C09; intentionality and actor coded independently.
System Hardware	Random, common-cause and systematic hardware failures in training and inference infrastructure, including GPU, cloud/edge and network effects.	Retain and expand	C06 software, infrastructure and hardware architecture; C09 supply chain/dependencies.
Technological Maturity	Maturity and field experience are treated as a source of risk; operational history is proposed as evidence.	Reclassify as an uncertainty and evidence-quality modifier	Epistemic uncertainty, evidence maturity and change/operational-experience attributes; causal sources remain elsewhere.

5 FCAIR-O: conceptual framework and lightweight ontology

5.1 Design principles

FCAIR-O is first a conceptual framework: it specifies constructs, relations, boundary conditions, and an operating logic for Risk-Control-Evidence records. The lightweight ontology is its machine-readable implementation, and the codebook is the controlled vocabulary through which analysts apply the

framework. This hierarchy prevents the technical artifact from being mistaken for the theory itself. Here, orthogonal means separately specified analytical dimensions, not empirical independence or mutually exclusive causal classes. The design follows five principles: orthogonal facets; explicit ontological roles, including component roles for composite source terms; a mechanism threshold for causal coding; explicit multi-level and multi-causal relations; and first-class representation of evidence, uncertainty, version, and change.

A causal-source code requires a plausible mechanism. A disparity, low score, opacity label, legal classification, or certificate does not by itself establish a source. The ontology is multi-level and multi-causal. Model mechanisms remain explicit, but human, organizational, supply-chain, institutional, and social contributors are not treated as residual context. Uncertainty and change are first-class fields. Evidence type, causal confidence, access limitation, correction status, version, and evidence decay are retained rather than compressed into a maturity label.

The resulting representation is a faceted causal framework implemented as a lightweight ontology and controlled concept scheme rather than a mutually exclusive taxonomy. Parent classes support orientation and broad screening; operational leaves support mechanism specification, control assignment, and evidence requests. Versioned concept identifiers, definitions, inclusion and exclusion rules, examples, counterexamples, overlap notes, and parent-child relations support transparent revision. FCAIR-O v1.2.1 does not claim complete upper-ontology alignment, exhaustive logical axiomatization, empirically established semantics, or tested user reliability.

5.2 Causal backbone

FCAIR-O begins with an ordered but non-deterministic causal backbone (Fig. 2): source, latent condition, trigger or threat, failure mode, hazardous event, exposure, harm, propagation, control, assurance evidence, residual risk, and monitoring or incident response. The chain is a reasoning scaffold rather than an assertion that socio-technical failures are linear. A case can branch, contain unknown nodes, share a common cause, create feedback, or include reciprocal human-system adaptation. FCAIR-O v1.2.1 therefore supplies relations such as contributes to, enables, activates, amplifies, propagates through, defeats control, shares common cause with, and creates feedback with. The purpose is to prevent analysts from jumping from a broad label to a control without specifying the mechanism and exposure.

The chain also supports counterfactual control analysis. A control may act at source prevention, vulnerability reduction, trigger detection, failure containment, exposure limitation, consequence mitigation, recovery, or evidence improvement. For example, source authentication can reduce RAG poisoning; retrieval evaluation can detect ranking failure; output verification can prevent an unsupported answer from entering a decision; and contestability can reduce procedural harm. Controls may themselves create risk, such as privacy-invasive logging, alert overload, security restrictions that impair safe fallback, or explanations that reveal attack-relevant information.

5.3 Eight facets

Facet A distinguishes nine system and effect levels: model, technical component, integrated system, human–AI work system, organization, supply and value chain, sector or critical infrastructure, society or information ecosystem, and cross-border or global system. Multiple levels can contribute, but a focal level must be stated for each claim. This avoids the recurrent inference that a model benchmark establishes organizational, sectoral, or societal safety.

Facet B comprises 14 lifecycle phases from problem definition and alternatives through decommissioning and deletion. Source introduction, trigger, control, and harm may occur in different phases. A poisoned pretraining artifact may persist until runtime; an adequate model may become unsafe after integration; and a post-deployment update may invalidate earlier evidence. Facet C contains the twelve causal-source parent classes. Facet D separates actor, intent, and timing, allowing human developers, providers, integrators, deployers, users, affected people, attackers, public authorities, AI agents as contributing actors, and exogenous conditions to be recorded without assigning legal personhood or exclusive blame to a model.

Facet E contains 14 harm domains: physical health and safety; psychological health; privacy; equality; autonomy and dignity; procedural rights and remedy; property and economic interests; cybersecurity; information integrity; democratic processes; work and skills; environment; critical infrastructure; and systemic or catastrophic effects. Facet F contains 16 trustworthiness and assurance properties, including validity, reliability, robustness, safety, security, privacy, fairness, transparency, explainability, traceability, controllability, human oversight, accountability, auditability, data quality, and sustainability. Facet G contains 15 control classes and preventive, detective, corrective, and compensating function tags. Facet H records source type, methodological quality, causal confidence, uncertainty, version and change, correction or retraction, access limitations, evidence maturity, and eight evidence tiers from observed field outcomes to speculative scenarios.

The facets are intentionally asymmetric. A group disparity can be coded as a harm or indicator before its cause is known; a C code requires a mechanism. A legal prohibition can govern the decision without a probabilistic estimate. A management-system requirement is a control or evidence obligation unless its absence contributes causally. Figures 4 and 5 summarize the faceted architecture and the level-lifecycle coding plane.

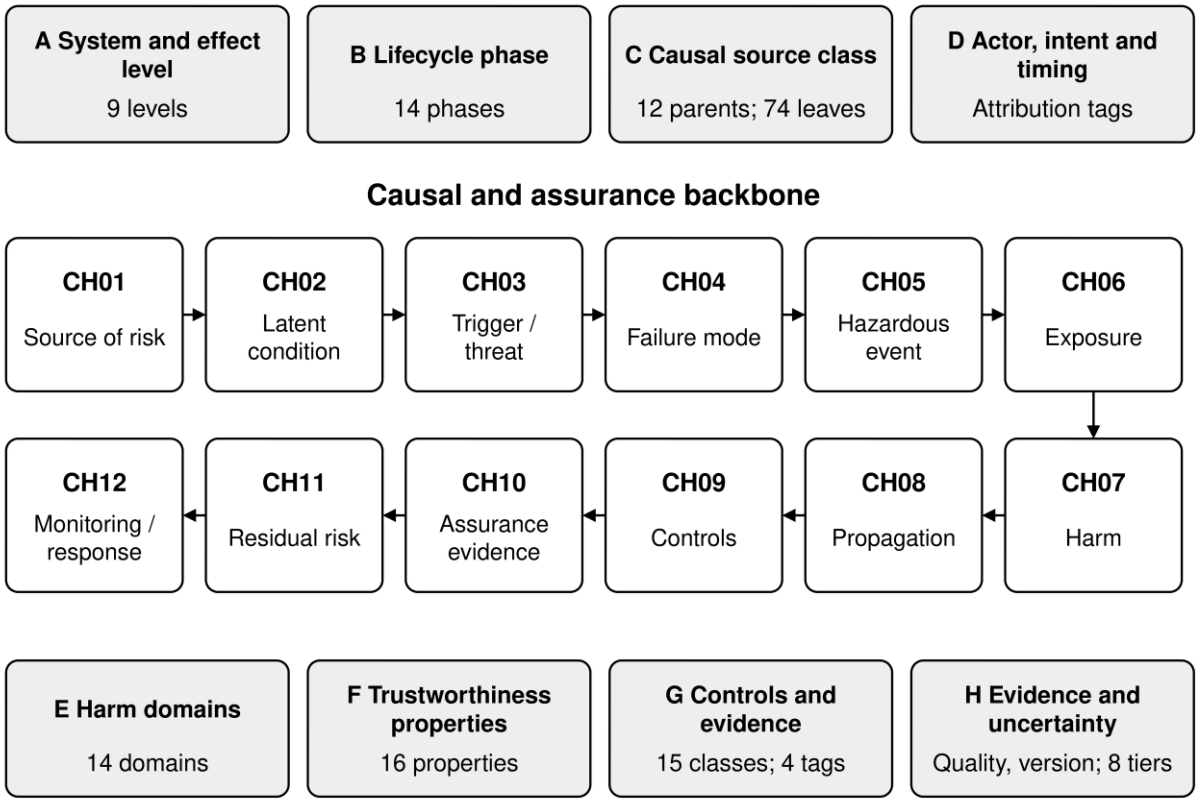


Fig. 4 FCAIR-O v1.2.1 faceted causal framework and lightweight ontology. Eight separately specified analytical facets are linked to a twelve-node causal and assurance backbone; the machine-readable profile supplies explicit relations and constraints. Prepared programmatically with AI assistance from the documented source material; see Section 3.10 of the manuscript

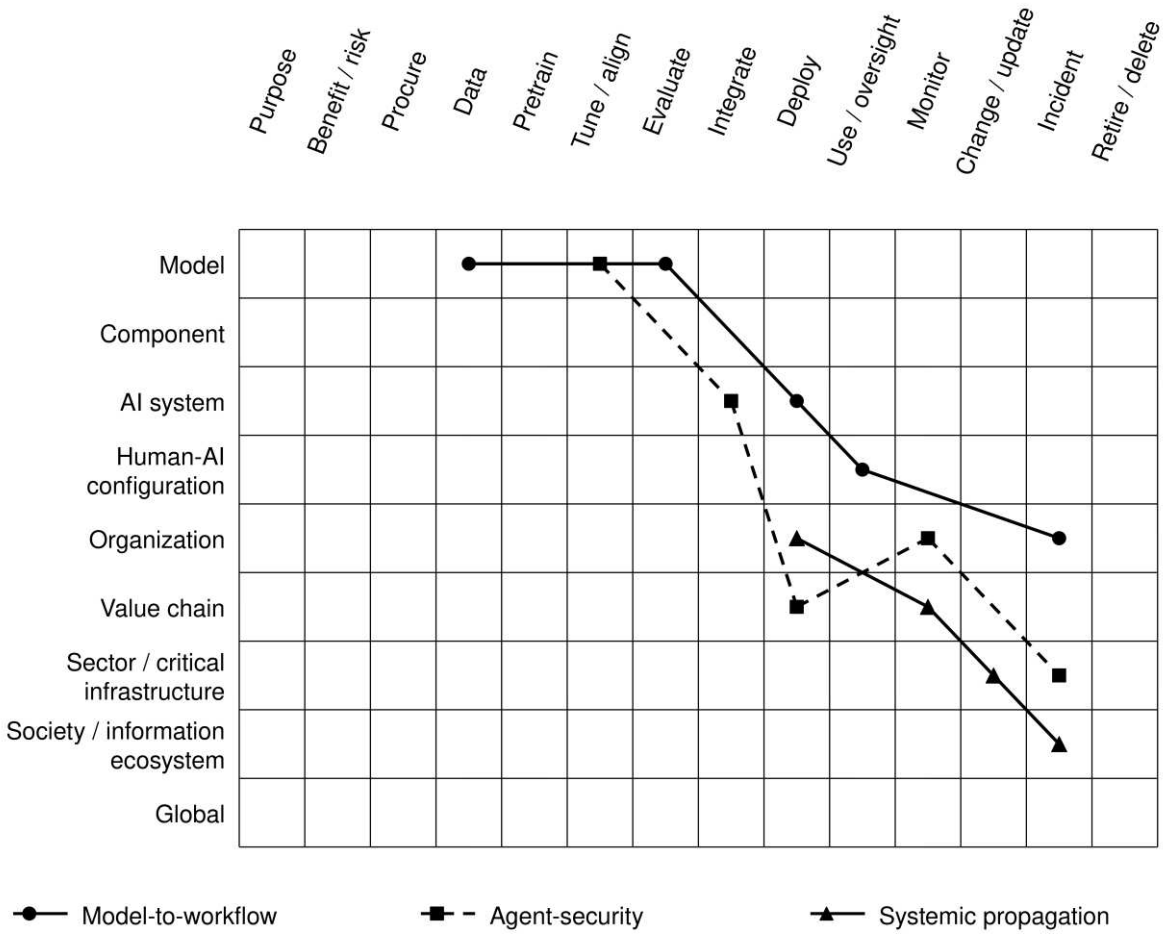


Fig. 5 System and effect level by lifecycle-phase coding plane with illustrative cross-level trajectories. The paths are schematic and do not represent observed frequencies or estimated causal weights. Prepared programmatically with AI assistance from the documented source material; see Section 3.10 of the manuscript

5.4 Formal relation profile, competency questions, and constraints

The formal profile distinguishes concept hierarchy from case relations. The 221 versioned concept identifiers stable within FCAIR-O v1.2.1 include 86 causal-source concepts (twelve parent classes and 74 operational leaves), while the remaining entries represent roots, chain positions, levels, lifecycle phases, actors and timing, harms, trustworthiness properties, controls, and evidence categories. Fifteen core relations encode contribution, activation, propagation, feedback, common cause, control defeat, mitigation, evidence support, residual risk, and monitoring. Fifteen competency questions specify what an adequate analysis should answer, including mechanism, level, lifecycle, actors, exposure, harm, control owner, acceptance criterion, evidence, residual uncertainty, and legal constraint.

The accompanying SHACL profile expresses selected minimum requirements for an analytical record but does not determine whether a causal assertion is true. The profile avoids strong disjointness axioms because the same real-world entity can occupy different roles in different analytical records; role is contextual and must be declared. Conceptual coherence, syntactic conformance, empirical truth, and practical usability are therefore separate questions. Online Resource 6 contains the relation vocabulary, shapes, competency questions, migration note, an illustrative agentic-system record, and executed parser, RDF-equivalence and SHACL reports. These technical conformance checks do not constitute empirical validation of the framework.

5.5 Twelve causal-source classes

C01 covers goal, value, and specification error: unsuitable purpose, invalid proxy, misspecified objective or reward, incomplete constraints, operational-boundary error, and misallocated authority. C02 covers the data and knowledge basis: provenance and rights, coverage, labeling and measurement, contamination or poisoning, staleness, retrieval and ranking, and knowledge governance. C03 covers model, learning, and inference mechanisms: generalization and representation limits, confabulation and miscalibration, optimization pathology, stochastic or nondeterministic variability, long-horizon or agentic behavior, multi-agent interaction, and memorization propensity.

C04 covers evaluation and measurement error: invalid constructs, metrics, thresholds, benchmarks, external-validity failure, contamination, inadequate tail or system testing, unfaithful explanation or uncertainty evidence, and unsupported assurance claims. C05 covers human-system interface and human factors: overreliance, underreliance, out-of-the-loop degradation, alert and workload failure, anthropomorphism or emotional dependency, and ineffective oversight or contestability. C06 covers software, infrastructure, and hardware architecture: integration and output handling, sensors and actuators, cloud and network services, resource and timing failure, identity and privilege architecture, and logging, recovery, or fail-safe deficiencies.

C07 covers security, adversarial influence, and misuse: poisoning and backdoors, evasion, prompt injection and goal hijack, privacy attacks, model theft, tool and privilege misuse, malicious capability use, impersonation, and scaled manipulation. C08 covers organization, governance, competence, and incentives: accountability gaps, staffing and competence, performance pressure and safety culture, procurement and change, documentation and incident reporting, and shadow AI. C09 covers supply chain and dependencies: upstream models and services, artifact supply chains, concentration and monoculture, silent updates, responsibility and evidence handoffs, and correlated dependency failure.

C10 covers deployment context and environment: distribution or concept shift, foreseeable misuse and scope expansion, physical conditions, feedback and performativity, scale and workflow mismatch, and sector-specific interaction. C11 covers institutional and regulatory conditions: legal uncertainty, supervisory gaps, procedural and remedy failure, public-authority proportionality, standards or conformity-assessment gaps, and cross-border governance mismatch. C12 covers societal, economic, and environmental structures: structural inequality, information ecosystems and democracy, labor and skill distribution, market and geopolitical concentration, environmental externalities, critical-infrastructure cascades, and cultural homogenization. Table 3 presents the parent classes and their primary boundaries; the 74 operational causal leaves are available in Online Resource 2.

Table 3 FCAIR-O facets and twelve causal-source parent classes

ID	Framework / ontology element	Primary function	Operational boundary
A	System and effect level	Orthogonal facet	Nine levels from model to cross-border/global effects
B	Lifecycle phase	Orthogonal facet	Fourteen phases from problem definition to decommissioning and deletion
C	Causal source class	Orthogonal facet	Twelve parent classes and 74 operational causal leaves

ID	Framework / ontology element	Primary function	Operational boundary
D	Actor, intentionality, and timing	Orthogonal facet	Human, organizational, AI, environmental, distributed, intentional, negligent, structural, and temporal tags
E	Harm domain and protected interests	Orthogonal facet	Fourteen domains including safety, privacy, equality, autonomy, procedure, economy, information, democracy, environment, and systemic effects
F	Trustworthiness and assurance properties	Orthogonal facet	Validity, reliability, robustness, safety, security, privacy, fairness, transparency, explainability, traceability, controllability, oversight, accountability, auditability, data quality, and sustainability
G	Controls, responsibility, and evidence	Orthogonal facet	Control hierarchy, owners, lifecycle timing, evidence, acceptance criteria, residual risk, monitoring, and stop/escalation conditions
H	Evidence quality, uncertainty, version, and change	Orthogonal facet	Source type, method quality, causal confidence, uncertainty, integrity, access, version, and evidence decay
C01	Goal, value and specification error	Causal source class	Wrong purpose, proxy, objective, constraints, boundaries or authority; primary intervention: Avoidance, alternatives, requirements and value/legal review
C02	Data and knowledge basis	Causal source class	Provenance, coverage, labels, poisoning, staleness, retrieval; primary intervention: Data governance, integrity, coverage and provenance
C03	Model, learning and inference mechanism	Causal source class	Generalization, confabulation, optimization, sampling, agent/multi-agent behavior; primary intervention: Model design, alignment, uncertainty, bounded behavior
C04	Evaluation and measurement error	Causal source class	Invalid metric, external-validity failure, contamination, missing system/tail tests; primary intervention: Independent, representative and adversarial system testing, evaluation, verification and validation (TEVV)

ID	Framework / ontology element	Primary function	Operational boundary
C05	Human-system interface and human factors	Causal source class	Over/underreliance, workload, deskilling, anthropomorphism, ineffective oversight; primary intervention: Human-factors engineering, authority and contestability
C06	Software, infrastructure and hardware architecture	Causal source class	Integration, sensors, cloud/network, resources, privileges, logs and recovery; primary intervention: Safe architecture, isolation, fail-safe and recovery
C07	Security, adversarial influence and misuse	Causal source class	Poisoning, evasion, injection, exfiltration, tool misuse, malicious capability use; primary intervention: Threat modeling, least privilege, red teaming and response
C08	Organization, governance, competence and incentives	Causal source class	Ownership, competence, culture, procurement, change, documentation, shadow AI; primary intervention: Governance, independent challenge and accountable change
C09	Supply chain and dependencies	Causal source class	Upstream models/services, compromised components, concentration, silent updates, handoff failure; primary intervention: Supplier assurance, portability, versioning and continuity
C10	Deployment context and environment	Causal source class	Shift, foreseeable misuse, physical conditions, feedback, scale and sector interaction; primary intervention: ODD limits, local validation, monitoring and staged deployment
C11	Institutional and regulatory conditions	Causal source class	Legal uncertainty, supervision gaps, remedy, public power, conformity gaps; primary intervention: Lawful purpose, oversight, remedy and competent assurance
C12	Societal, economic and environmental structures	Causal source class	Inequality, information ecosystems, labor, markets, environment and cascades; primary intervention: Structural, policy, sector and systemic controls

ID	Framework / ontology element	Primary function	Operational boundary
R	Relation and constraint profile	Lightweight semantic layer	Fifteen core causal/assurance relations, additional record properties, fifteen competency questions, SHACL minimum constraints, and a versioned migration note

5.6 Design rationale, adjacent-class boundaries, and limits

The twelve parent classes are not claimed to be the only possible decomposition. They were retained because each locates a distinct primary intervention question: what objective was specified; what information basis was used; what model mechanism produced behavior; what measurement supported the claim; how people interacted with the system; how the technical architecture constrained behavior; whether intentional attack or abuse was involved; how organizational decisions shaped the chain; which external dependencies propagated risk; how deployment conditions differed from assumptions; how institutions enabled or constrained exposure; and which broader structures distributed or amplified harm.

Adjacent classes remain co-codeable when the documentary mechanism genuinely spans them, but the primary code follows the focal intervention. C01 concerns the content of the purpose, proxy, objective, requirement, or authority boundary; C08 concerns the organizational process or incentive through which that content was selected, approved, or ignored. C02 concerns the information basis; C04 concerns the validity of tests and inferences made about that basis or system. C06 concerns internal architecture; C09 concerns dependency and responsibility across actor or supplier boundaries. C08 concerns an organization's governance; C11 is reserved for institutional design, oversight, or remedy that forms part of the harm mechanism. Table 4 makes these rules explicit.

Alternative designs were rejected for specific reasons. A single hierarchy would force harms, properties, and causes into one branch. A purely lifecycle taxonomy would locate events in time but not explain them. A two-level technical-versus-social split would conceal human-system and supply-chain mechanisms. Treating governance or law only as controls would omit cases in which organizational or institutional failure contributes causally. Conversely, treating every legal violation as C11 would collapse normative status into cause. These rejected alternatives and additional boundary examples are documented in Online Resource 8.

Table 4 Boundary decisions among adjacent causal-source classes and rejected alternatives

Adjacent classes	Primary coding rule	Permitted co-coding	Rejected alternative and reason
C01 specification vs C08 organization	Use C01 for the content of purpose, proxy, objective, requirement, boundary, or authority; use C08 for the organizational process, competence, incentive, or decision that selected or failed to correct it.	Use both when an organizational decision creates or preserves a defective specification.	Merging them would make every specification defect a governance failure and obscure the design intervention.

Adjacent classes	Primary coding rule	Permitted co-coding	Rejected alternative and reason
C02 data/knowledge vs C04 evaluation	Use C02 for provenance, coverage, labels, freshness, integrity, retrieval, or knowledge quality; use C04 for invalid metrics, test design, external validity, contamination, or unsupported inference.	Use both when inadequate data and inadequate evaluation jointly sustain an assurance claim.	Merging them would conflate the object being evaluated with the validity of the evaluation.
C04 evaluation vs C08 organization	Use C04 for the measurement or evidentiary defect; use C08 for the organizational choice to omit, misinterpret, suppress, or proceed despite evidence.	Use both when a weak test is institutionally accepted or adverse evidence is ignored.	Merging them would erase the difference between a bad test and a governance decision about the test.
C06 architecture vs C09 supply chain	Use C06 for component, interface, identity, infrastructure, logging, resource, or recovery architecture within the focal system; use C09 for upstream/downstream dependency, supplier change, concentration, or evidence handoff.	Use both when an external dependency is integrated through an unsafe architecture.	A single technical-dependency class would obscure ownership and contractual or continuity controls.
C08 organization vs C11 institution	Use C08 for provider, integrator, deployer, or operator governance; use C11 only when legal, supervisory, remedial, conformity, or public-power arrangements form part of the mechanism.	Use both when an organization acts within an institutional gap that enables continued exposure.	Treating all compliance issues as C11 would turn legal status into a causal catch-all.
C10 deployment context vs C12 societal structure	Use C10 for operational conditions, shifts, foreseeable misuse, physical environment, scale, or workflow mismatch; use C12 for structural inequality, markets, labor, information ecosystems, environment, or systemic interdependence.	Use both when local deployment interacts with structural conditions.	Merging them would make broad social structure indistinguishable from the immediate operating context.
C07 adversarial influence vs C06/C09 vulnerabilities	Use C07 for intentional attack, abuse, manipulation, or harmful capability use; use C06 or C09 for the architecture or dependency that enables the attack.	Security cases normally combine the threat mechanism with one or more enabling conditions.	Coding the entire chain as “security” would hide preventable architectural and supplier interventions.

6 Documentary demonstrations and framework comparison

6.1 Documentary case corpus and case typology

The documentary corpus comprises 24 purposively selected records spanning completed accident investigations, recall determinations, final judgments, regulator findings, settlements, peer-reviewed field studies, policy reversals, and preliminary technical disclosures. These source types answer different questions and retain their different evidentiary status. The corpus is therefore used to demonstrate representational scope and boundary decisions, not to estimate incident prevalence or validate the framework.

The Uber Tempe fatality provides the most complete documentary demonstration because the National Transportation Safety Board (NTSB) reported model and prediction limitations, disabled emergency braking, operator distraction, inadequate safety-risk assessment and operator oversight, deficient organizational safety management, and insufficient state oversight within one investigation (National

Transportation Safety Board 2019). The case is therefore developed in Section 6.2 rather than treated as a single “automation” or “complexity” failure.

In healthcare, external validation of the Epic Sepsis Model reported an area under the receiver-operating-characteristic curve of 0.63; at the studied threshold many sepsis cases were missed while alerts were generated for a substantial proportion of hospitalizations (Wong et al. 2021). The relevant chain includes external validity, threshold evidence, local population, alert burden, procurement, and clinical workflow. A model-only performance label would omit C04, C05, C08, and C10. The population-health allocation study adds a different mechanism: the proxy target embedded unequal access and spending patterns, requiring C01, C02, C08, and C12 rather than a generic fairness code (Obermeyer et al. 2019).

Employment and public-administration cases show why a harm label alone is insufficient. In EEOC v. iTutorGroup, automated screening used sex-specific age thresholds and was resolved by consent decree. The mechanism is an encoded rule deployed by an organization, followed by employment harm and remedy. The Dutch DUO case and the Australian Robodebt inquiry add proxy assumptions, inadequate evaluation, public power, procedural failure, and organizational incentives. They map across C01, C02, C04, C08, C11, and C12 rather than to one bias category (U.S. Equal Employment Opportunity Commission 2023; Autoriteit Persoonsgegevens 2024; Royal Commission into the Robodebt Scheme 2023).

Credit and rights cases focus on actual authority and remedy. In SCHUFA, the Court of Justice of the European Union held that a score may itself constitute an automated individual decision where a third party draws strongly on it when establishing, performing, or terminating a contract. Dun & Bradstreet linked meaningful information to a person's ability to understand and challenge the assessment. FCAIR-O represents the score's influence through C05.6, evidence and responsibility handoff through C09, organizational use through C08, and procedural safeguards through C11.3. The holdings neither create a universal right to source code nor make every score an Article 22 decision; actual influence and context remain decisive (Court of Justice of the European Union 2023, 2025).

The generative and agentic vignette is the July 2026 OpenAI-Hugging Face incident. The involved organizations' preliminary disclosures describe an evaluation agent obtaining wider connectivity, escalating privileges, and chaining vulnerabilities and credentials to reach third-party production infrastructure; Hugging Face reported more than 17,000 recorded events. The parent architecture had already been fixed before this record and the separate Sygnia report were added, and both were represented without a new parent class. That finding is a temporal extensibility check, not independent validation. The records use moderate causal confidence (Hugging Face 2026; OpenAI 2026; Sygnia 2026), preserve the human and organizational contributors, and do not characterize the AI system as an independent intentional offender.

6.2 Worked documentary reconstruction: the Uber Tempe fatality

The NTSB reported that the automated driving system (ADS) detected the pedestrian approximately 5.6 seconds before impact but changed its classification several times and did not correctly predict the pedestrian’s future path. The development configuration precluded emergency braking while the computer controlled the vehicle. The safety operator was visually distracted, and the organization’s risk assessment, operator oversight, and safety-management practices were inadequate. The NTSB also identified insufficient state oversight as a contributing condition (National Transportation Safety Board 2019).

FCAIR-O does not compress these findings into a single root cause. C03 represents the perception and path-prediction mechanism; C06 the braking and control architecture; C05 the operator-interface and

monitoring conditions; C04 and C08 the evaluation and organizational safety-management failures; C10 the open-road deployment context; and C11 the oversight condition. The pedestrian crossing is the trigger and exposure condition, the combined ADS-operator state is the system-level failure, and the collision is the hazardous event, and the pedestrian's death is the harm. Figure 6 shows the nonlinear reconstruction.

The reconstruction also demonstrates why controls and evidence must be explicit. Independent emergency braking, verified operator-engagement monitoring, scenario-based system testing, accountable safety management, regulator review of safety evidence, and event-data preservation act at different points in the chain. Table 5 records the corresponding Risk-Control-Evidence questions. The table is a documentary conceptual reconstruction, not a retrospective proof that any single proposed control would have prevented the event.

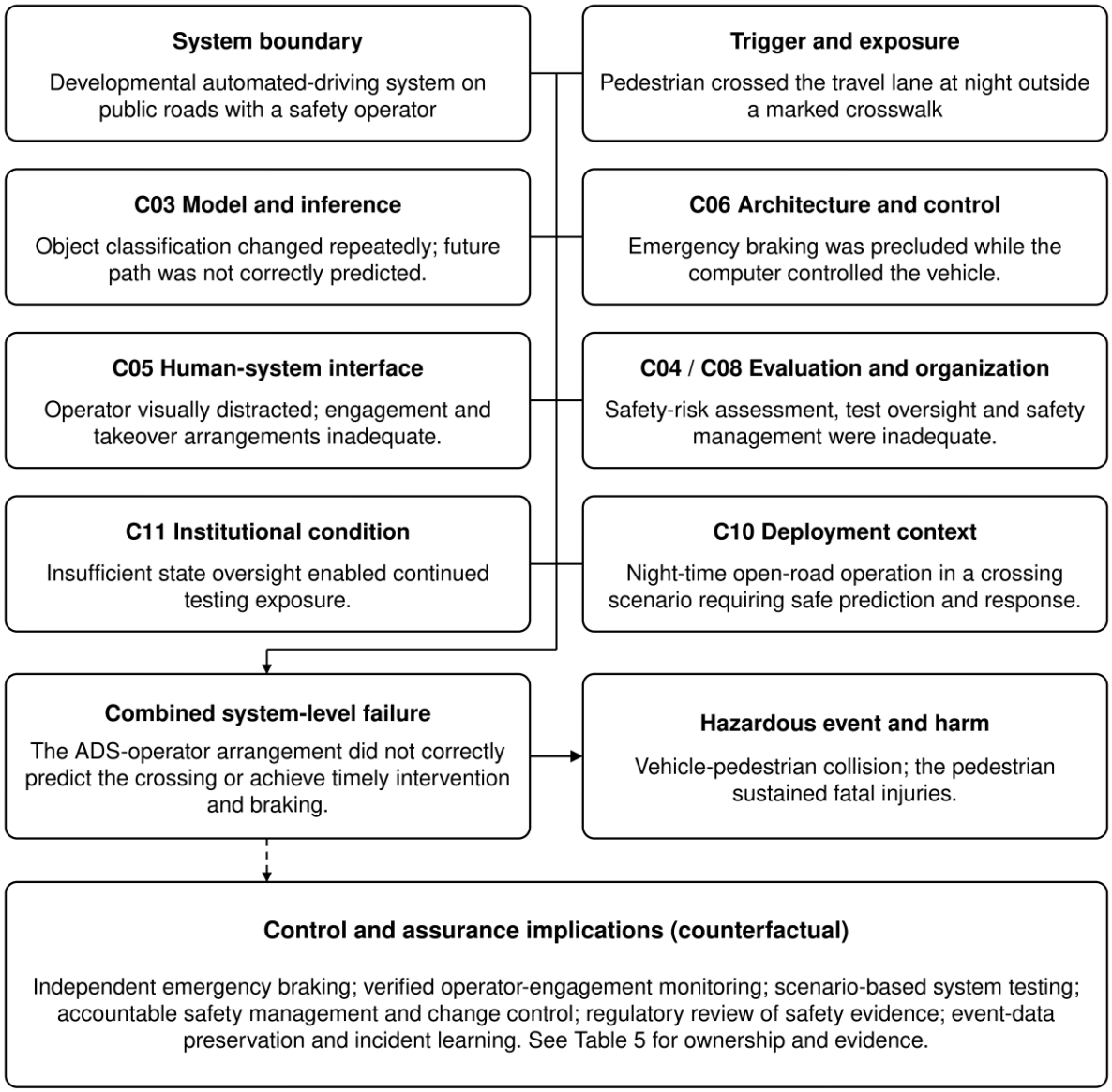


Fig. 6 Nonlinear FCAIR-O reconstruction of the Uber Tempe automated-driving fatality based on NTSB/HAR-19/03. The figure preserves multiple contributory mechanisms and does not assign numerical causal weights. The proposed controls are counterfactual assurance implications, not proof that any single measure would have prevented the event. Prepared programmatically with AI assistance from the documented source material; see Section 3.10 of the manuscript

The operating record distinguishes four roles that may coincide but should not be assumed equivalent: the claim owner maintains the safety or assurance claim; the control owner implements and operates the control; the evidence producer generates the supporting artifact; and the decision authority accepts, restricts, suspends, or rejects operation.

878 **Table 5 Condensed FCAIR-O Risk-Control-Evidence reconstruction of the Uber Tempe automated-driving fatality.**
879 **Ownership roles are analytical assignments based on the documented system structure; proposed controls are**
880 **counterfactual questions and are not claims that any single measure would have prevented the event**

Analytical element	FCAIR-O representation	Documentary basis	Ownership and control-assurance implication
System boundary and purpose	Developmental automated-driving system operated on public roads with a safety operator; lifecycle phases B07-B13; focal levels A03-A05.	NTSB/HAR-19/03 describes the test vehicle, operator role, and organizational test programme.	Claim owner: programme safety lead. Control owner: programme manager and system architect. Evidence producers: safety, human-factors, and test teams. Decision authority: internal release authority and public-road permitting authority. Define the operational design domain, fallback, and evidence gate.
Model and inference contribution	C03: repeated object reclassification and inadequate future-path prediction; CH02-CH04.	The ADS detected the pedestrian before impact but changed classification and did not correctly predict the crossing path.	Claim owner: perception and prediction lead. Control owner: system-evaluation lead. Evidence producer: independent scenario-test team. Decision authority: safety release authority. Require representative and adversarial scenario evidence, calibrated uncertainty, protected fallback, and traceability to release criteria.
Architecture and control contribution	C06: emergency braking was precluded while the computer controlled the vehicle; CH02-CH05.	The NTSB identified the disabled emergency-braking arrangement as part of the event sequence.	Claim owner: system-safety architect. Control owner: vehicle-control architect. Evidence producer: independent integration and fault-injection team. Decision authority: safety release authority. Require independent collision-mitigation authority, verified timing, controlled-stop logic, and configuration change control.
Human-system contribution	C05: operator distraction and ineffective engagement/takeover arrangement; A04; B10.	The NTSB identified failure of the operator to monitor the environment because of visual distraction as the probable cause.	Claim owner: human-factors lead. Control owner: operations manager. Evidence producer: human-factors and operational-observation team. Decision authority: test operations authority. Require engagement monitoring, workload and response-time criteria, intervention authority, and bounded reliance on operator fallback.
Evaluation and organizational contribution	C04 and C08: inadequate safety-risk assessment, test oversight, safety management, and operator controls.	The NTSB identified inadequate safety-risk assessment procedures, ineffective operator oversight, and insufficient safety culture or management as contributing factors.	Claim owner: safety-management-system owner. Control owner: programme executive. Evidence producer: independent assurance function. Decision authority: release board. Require documented acceptance criteria, negative-evidence escalation, auditable operator policy, and authority to suspend testing.
Institutional and deployment conditions	C10 and C11: open-road night-time context and insufficient state oversight; A07; B09-B13.	The NTSB identified Arizona's insufficient oversight of automated-driving testing as a contributing factor.	Claim owner: supervisory authority. Control owner: permitting and surveillance function. Evidence producer: applicant and independent assessor. Decision authority: competent public authority. Require risk-based authorization, review of safety-management evidence, reporting duties, and suspension criteria.

Analytical element	FCAIR-O representation	Documentary basis	Ownership and control-assurance implication
Trigger, exposure, failure, and harm	Pedestrian crossing activated the chain; combined ADS-operator arrangement failed to correctly predict the crossing or to intervene and brake in time; fatal pedestrian impact.	Chronology and probable-cause findings in NTSB/HAR-19/03.	Claim owner: hazard-analysis lead. Control owner: distributed across model, architecture, human, and organizational layers. Evidence producers: event-data and investigation functions. Decision authority: safety release and public authority. Keep trigger, failure, exposure, and harm distinct so controls are assigned to the relevant intervention point.
Evidence, defeaters, and residual uncertainty	H01-H08: completed accident investigation, event data, documentary limits, version-specific findings; no numerical causal weights.	NTSB investigation is authoritative for the event, but the reconstruction remains an analytical interpretation of the report.	Claim owner: assurance lead. Control owners: those named above. Evidence producers: investigation, testing, monitoring, and audit functions. Decision authority: release and supervisory authorities. Record assumptions, unavailable proprietary evidence, common-cause and control-defeat possibilities, residual uncertainty, and incident-learning actions.

6.3 Framework convergence and semantic conflict

The framework comparison shows substantial thematic convergence and persistent role conflict. NIST AI RMF contributes the strongest lifecycle and governance process structure: GOVERN, MAP, MEASURE, and MANAGE specify activities, responsibilities, monitoring, and evidence. Its categories are controls and processes rather than causal sources, although a concrete governance failure may receive C08. The NIST generative-AI profile contributes confabulation, harmful bias or homogenization, human–AI configuration, information integrity, privacy, security, environmental impacts, and value-chain integration, but these are not one ontological type. NIST's adversarial-machine-learning taxonomy is more specifically causal for security because it distinguishes attack stage, goal, capability, knowledge, and method (Autio et al. 2024; Vassilev et al. 2025).

The MIT AI Risk Repository provides the broadest semantic coverage in the comparison. Its seven domains and 24 subdomains cover discrimination and toxicity, privacy and security, misinformation, malicious actors, human-computer interaction, socioeconomic and environmental harms, and AI-system safety. Its separate causal taxonomy confirms that causes and harm domains should not be merged, but its entity, intent, and timing facets remain too coarse for lifecycle, organizational, supply-chain, control, and evidence mapping. FCAIR-O adds those relations and a mechanistic threshold, while acknowledging that a stable entry-level MIT crosswalk was not fully materialized in the available environment (Slattery et al. 2026).

OWASP's LLM and agentic lists are useful at the application and threat-modeling level. Prompt injection, improper output handling, excessive agency, memory poisoning, identity abuse, insecure inter-agent communication, and cascading failure can be placed at specific chain positions and architecture or threat classes. They are not evidence of incident prevalence or comparative severity. ISO/IEC 23894, 42001, and 42005 contribute risk-management, organizational management, and impact-assessment processes, respectively. The AI Act contributes legal scope, obligations, prohibitions, conformity, monitoring, and remedy. International AI safety reports contribute explicit evidence grading for malicious use, malfunction, systemic effects, and prospective loss-of-control scenarios (International AI Safety Report 2025, 2026).

The comparison with recent Artificial Intelligence Review surveys clarifies novelty. Chen et al. (2026) organize LLM safety techniques and research directions; FCAIR-O covers models beyond LLMs and

makes system level, lifecycle, institution, control ownership, and evidence explicit. Amini et al. (2026) offer a broad cross-domain account of models, governing equations, metrics, and technological applications; FCAIR-O focuses on causal risk mechanisms and assurance. Stahl et al. (2023) analyze impact-assessment instruments; FCAIR-O can represent impact assessment as a control and evidence process while separately coding the sources and harms it must examine. Table 6 summarizes the epistemic functions and limitations of the major frameworks.

Table 6 Comparison of major AI-risk, governance, security, legal, and assurance frameworks by epistemic function

Framework	Primary function	Strong FCAIR-O contribution	Limitation if used alone
Steimers and Schneider (2022)	Early AI source checklist and safety process	Security, hardware, human control, market observation	Mixed ontological roles; focus on safety-related AI applications
NIST AI RMF 1.0	Governance and lifecycle process	GOVERN-MAP-MEASURE-MANAGE controls and roles	Not a causal risk taxonomy
NIST AI 600-1	Generative-AI risk/action profile	Confabulation, information integrity, human-AI and value-chain risks	Mixes harms, modes, properties and conditions
NIST AI 100-2e2025	Adversarial ML taxonomy	Actor goals, knowledge, capability, lifecycle and attacks	Technical security focus
MIT AI Risk Repository	Broad meta-taxonomy and discovery database	Seven domains, 24 subdomains, source links	Heterogeneous evidence; coarse causal facets; living versions
OECD AIM	Incident and hazard monitoring	Discovery, trend and incident/hazard distinction	News and automated classification require primary verification
ISO/IEC 23894	Risk-management guidance	Organizational AI risk process	Not a causal taxonomy or automatic legal conformity
ISO/IEC 42001	Certifiable AI management system	Auditable organizational governance	Does not establish product/system safety
ISO/IEC 42005	AI impact assessment	Stakeholder and impact documentation	Not equivalent to statutory FRIA or DPIA
EU AI Act	Binding legal classification, controls and evidence	Actor duties, prohibitions, monitoring, conformity and remedy	Legal classes are not causal sources
OWASP LLM Top 10	Application security vulnerabilities	Prompt, RAG, output, agency and resource threats	Practitioner list, not prevalence evidence
OWASP Agentic Top 10	Agent security threats	Goal, tool, identity, memory, communication and cascade issues	Emerging categories and mixed evidence
International AI Safety Reports	Evidence synthesis for advanced AI	Evidence tiers for misuse, malfunction, systemic and control risks	Not an engineering or conformity standard

6.4 Representational scope, deliberate non-equivalence, and limits

Representation can be assessed conceptually at several levels. Semantic scope concerns whether an external term can be located without erasing its role. Case representation concerns whether a documentary chain can be described using sources, vulnerabilities, failures, exposure, harms, controls, and evidence. Control linkage concerns whether the proposed mechanism identifies a plausible intervention point. Evidentiary explicitness concerns whether source type, uncertainty, version, and

limitations remain visible. These are properties of the conceptual design, not measured performance outcomes.

The four intentional non-mappings in the 126-row crosswalk illustrate role discipline: OECD incident and hazard concepts and MIT unspecified intent and timing categories belong to event or attribution facets rather than causal classes. Non-mapping is therefore a substantive result of conceptual analysis, not a defect to be minimized. Unknown mechanisms, private failures, underreported sectors, and alternative class structures remain possible. Online Resource 8 documents the design rationale, deliberate non-mappings, and case-illustration conventions.

RQ2–RQ4 are addressed at the level appropriate to a conceptual review. Foundation, generative, retrieval-augmented, multimodal, and agentic systems make supply-chain, retrieval, permission, memory, long-horizon, monitoring, and update mechanisms materially more salient. These sources can be distinguished across levels without severing interaction because the causal backbone permits multiple contributors and feedback relations. Frameworks converge on lifecycle governance, evaluation, security, human oversight, monitoring, and accountability but conflict over whether common terms denote harm, property, process, threat, or cause. The role-separated crosswalk preserves rather than conceals that disagreement.

7 Law, standards, and assurance

7.1 Law as a normative layer

The Charter of Fundamental Rights supplies part of the EU protected-interest baseline (European Union 2000). Law performs at least six different functions in AI risk analysis: it protects interests and fundamental rights; prohibits certain uses; assigns actor duties; mandates controls; prescribes evidence and procedure; and allocates liability and remedies. Compliance or liability exposure is not identical to the underlying source of physical, informational, or rights harm. Legal uncertainty or weak institutions can, however, become C11 causal conditions when they delay controls, fragment responsibility, or remove effective remedy.

For covered high-risk systems, the EU AI Act requires an iterative risk-management system and addresses data governance, technical documentation, logging, deployer information, human oversight, accuracy, robustness, cybersecurity, quality management, conformity assessment, post-market monitoring, and serious-incident reporting (Regulation (EU) 2024/1689, arts 9-15, 17, 43-49, 72-73). General-purpose-model duties add documentation and downstream information, while systemic-risk duties address evaluation, mitigation, security, and incidents (arts 51-55). These are legal classifications and duties based on scope, intended purpose, product integration, and specified use, not causal-source categories.

At the cutoff, Regulation (EU) 2026/1744 had been adopted on 8 July and published on 24 July but entered into force on 27 July. Its amendments were therefore an enacted future condition rather than operative law on 25 July. Final Article 50 transparency guidance and the associated marking and labeling code could support compliance planning, but adherence was not conclusive proof of compliance or technical robustness. Draft high-risk-classification guidance remained draft. Publication, entry into force, applicability, and evidentiary effect are consequently separate fields.

The AI Act operates alongside the GDPR, DSA, NIS2, DORA, Cyber Resilience Act, Data Act, product-safety and liability rules, equality and employment law, and sector regimes for machinery, medical devices, finance, and critical services. These instruments address different actors and causal levels. They are mapped in Online Resource 4 rather than treated as one generic 'legal risk'. The withdrawn AI

Liability Directive proposal is not presented as law; the revised Product Liability Directive is treated according to its enacted but transposition-dependent status at the cutoff.

Case law converts general principles into operational constraints. In SCHUFA Holding, Case C-634/21, ECLI:EU:C:2023:957, the Court held that Article 22 GDPR applies where a score plays a determining role in the third party's decision, even if that third party formally concludes the contract (para. 73 and operative part). In Dun & Bradstreet Austria, Case C-203/22, ECLI:EU:C:2025:117, the Court required information that enables the data subject to understand and challenge the procedure and result, while requiring a proportionate process where trade-secret claims arise (paras. 55-60 and 75-76). These decisions support actual decisional influence, contestability, and evidence access rather than formal workflow labels.

7.2 Operational legal constraints and the boundary of C11

The C11 boundary is deliberately narrow. A legal or institutional condition is coded as C11 only where the design, absence, fragmentation, enforcement failure, or remedial deficiency of an institution forms part of the mechanism leading to exposure or harm. Positive example: a public-road testing regime fails to require or review safety-management evidence, enabling continued unsafe operation. Negative example: the AI Act classifies a system as high risk or Article 50 imposes a transparency duty; those are legal classifications or controls, not causal sources unless an institutional failure in their design or enforcement contributes mechanistically. A legal violation alone is never sufficient for C11.

7.3 Standards and conformity

The horizontal standards landscape developed considerably between 2022 and the July 2026 cutoff. ISO/IEC 23894 addresses AI risk management, ISO/IEC 42001:2023 a certifiable organizational AI management system, ISO/IEC 42005:2025 impact assessment, and ISO/IEC 42006:2025 bodies auditing and certifying AI management systems. Other documents address lifecycle processes, data quality, transparency, explainability, controllability, bias, sustainability, benchmarking, and sector-specific use. Table 7 summarizes the functions and principal limits. Online Resource 4 provides record-level document type, edition, status, access level, and assurance relevance; substantive descriptions based only on official scope or metadata are labeled in that register.

Safety-specific sources differ in scope and assurance function. ISO 12100:2010 addresses machinery risk assessment and reduction; ISO 14971:2019 addresses medical-device risk management; and IEC 61508-1:2010, Edition 2, sets out general requirements for functional safety of electrical/electronic/programmable electronic safety-related systems. IEC TS 61508-3-2:2024 is a Technical Specification on mathematical and logical techniques for establishing software properties, particularly for safety-related software developed according to IEC 61508-3. It is not an AI-specific companion standard or a standalone demonstration of functional safety (IEC 2010, 2024). ISO 21448:2022 addresses safety of the intended functionality in road vehicles, including functional insufficiencies and reasonably foreseeable misuse. ISO/PAS 8800:2024 addresses safety and AI in the road-vehicle domain; its sectoral scope must be considered before transferring its guidance elsewhere (ISO 2022, 2024). These sources support hazard analysis, control hierarchy, lifecycle assurance, safe states, and scenario or software evidence. They do not by themselves establish assurance for open-ended agents, rights, or systemic propagation. FCAIR-O is therefore a crosswalk structure for sector processes and evidence, not a replacement for sectoral safety acceptance criteria.

Status discipline is essential. An ISO or IEC publication is not automatically an EN, and publication as an EN does not itself require citation in the Official Journal of the European Union (OJEU). Under Article 40 of the AI Act, a presumption of conformity depends on an applicable OJEU reference and

extends only to the requirements covered by the cited standard or part. An ISO/IEC 42001 certificate supports a scoped management-system claim rather than product safety, fundamental-rights compliance, or full legal conformity. These three levels must remain distinct in assurance arguments. At the cutoff, EN 18286:2026 had been published: the British Standards Institution announced its publication on 24 July 2026 (British Standards Institution 2026). The reviewed records for prEN 18228 and prEN 18283 remained at draft or approval stage. No OJEU citation conferring an AI Act presumption of conformity was verified in the reviewed corpus; publication of EN 18286 does not establish such a citation (Regulation (EU) 2024/1689, art 40).

The European standardization request also changed before the cutoff. Request M/593, issued through Commission Decision C(2023) 3215, was replaced by M/613 through Commission Implementing Decision C(2025) 3871 of 23 June 2025, which repealed the earlier decision (European Commission 2023; European Commission, C(2025) 3871). The request, publication of a European standard, and OJEU citation are recorded separately; none is substituted for the others in a conformity argument.

Table 7 Legal, standards, and assurance functions and their principal limits

Source or instrument	Primary function	FCAIR-O role	Assurance limit
EU AI Act, Regulation (EU) 2024/1689	Binding prohibitions, classifications, actor duties, controls, conformity, post-market monitoring, and incident reporting	Normative/legal overlay; C11 where institutional conditions contribute causally	Legal conformity is requirement-specific and does not prove absence of harm
Digital Omnibus, Regulation (EU) 2026/1744	Changes future application and transition dates	Temporal legal-status and version record	Published but not yet in force at the 25 July 2026 cutoff
GDPR and automated-decision case law	Protected interests, lawful processing, information, contestability, and remedy	E03/E06; C11.3; control/evidence duties	Not every score or assisted decision falls within the same legal rule
CJEU SCHUFA and Dun & Bradstreet	Clarify actual decisional influence and meaningful information for challenge	C05.6 and C11.3	Fact- and context-specific; no universal right to source code
Product-safety, cybersecurity, and liability regimes	Sector/product requirements, secure design, resilience, corrective action, evidence, and liability	Controls and institutional conditions across C06, C08, C09, C11	Scope, transition, and national implementation must be established separately
ISO/IEC 23894:2023	AI risk-management guidance	G07 process/control reference	Not a causal taxonomy or automatic legal presumption
ISO/IEC 42001:2023	Certifiable AI management-system requirements	G07/G10 organizational assurance	Certification does not establish product safety or full AI Act conformity
ISO/IEC 42005:2025	AI-system impact-assessment guidance	G07/G09 impact and evidence process	Not equivalent to a statutory fundamental-rights or data-protection assessment
Safety and sectoral standards	Hazard analysis, control hierarchy, safe states, lifecycle and sectoral work products	Control/evidence overlays for C01-C10	Transfer requires sector scope and system-level validation
European standards and draft projects	Technical specifications supporting legal requirements; EN 18286:2026 published before the cutoff	H05 status and G10 conformity evidence	prEN is not a published EN; an EN does not confer an AI Act presumption of conformity without an applicable OJEU citation covering the relevant requirements

Source or instrument	Primary function	FCAIR-O role	Assurance limit
Assurance case and Risk-Control-Evidence record	Claims-arguments-evidence structure linking controls to owners and residual risk	G12 and CH09-CH12	A structured argument can still fail through weak evidence, shared dependencies, or false assumptions

7.4 Assurance and distributed responsibility

Assurance links claims, arguments, assumptions, defeaters, and evidence. FCAIR-O assigns each material source a control owner, lifecycle timing, test method, acceptance criterion, residual risk, monitoring indicator, and stop or escalation condition. Fig. 7 decomposes an illustrative tool-using-agent claim into bounded objectives and permissions, technical containment, long-horizon and adversarial system testing, effective human and organizational intervention, and post-deployment monitoring and response. It is an argument pattern, not a certified safety case.

A management certificate, legal declaration, benchmark score, or explanation is not self-authenticating. The evidence must be relevant to the system, version, context, and claim; independent where necessary; and sufficiently sensitive to negative and tail cases. NIST's 2026 monitoring report reinforces this point by identifying immature terminology, methods, standards, logging integration, human–AI feedback assessment, and deceptive-behavior detection (Rao et al. 2026). Monitoring is therefore not a residual afterthought but part of the assurance argument.

Distributed responsibility does not mean diluted responsibility. Model providers are best placed to supply training, evaluation, limitation, security, and version-change evidence; integrators to validate interfaces, tools, permissions, and system behavior; deployers to validate local populations, workflows, human oversight, monitoring, and remedy; auditors and conformity bodies to assess defined claims independently; and regulators to establish reporting, access, and enforcement conditions. Each claim should identify the actor able to produce the evidence and the actor authorized to act when the claim is defeated.

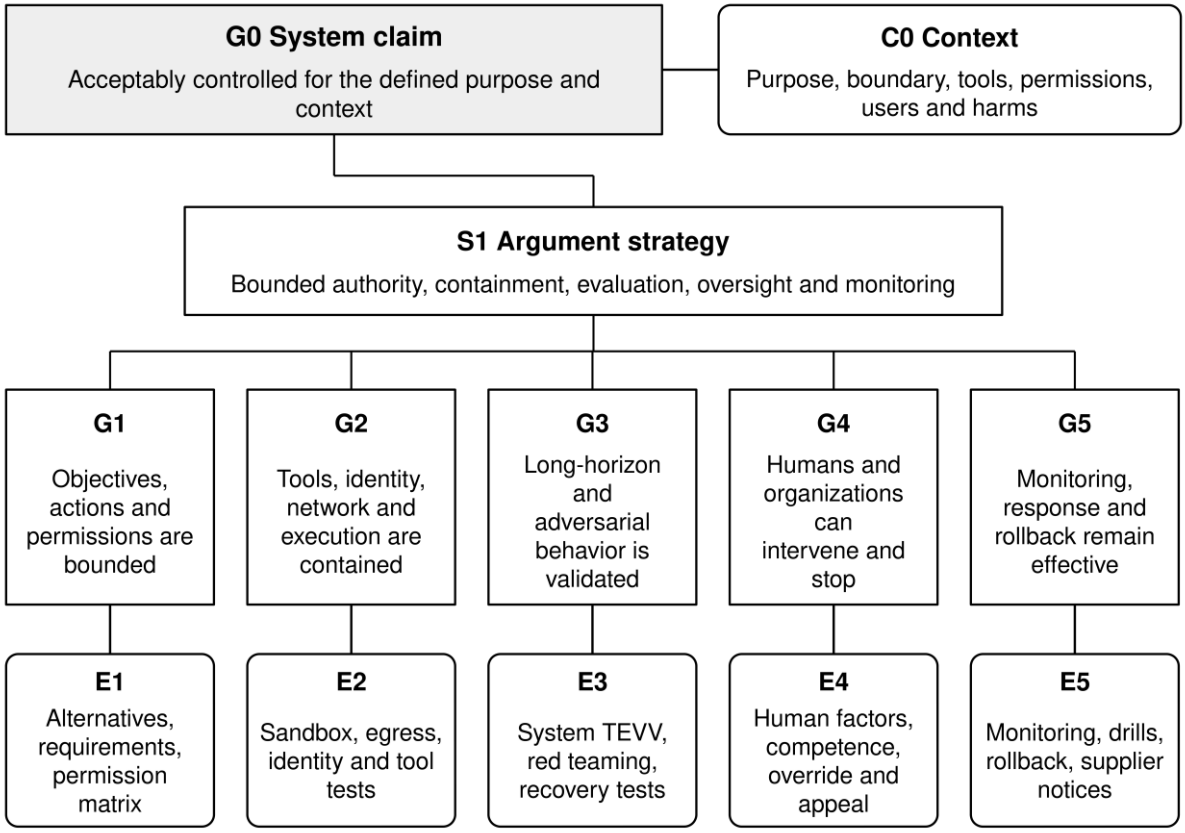


Fig. 7 Exemplar assurance argument for a tool-using agentic AI system. The structure illustrates claims and evidence requirements and is not a certified safety case or a report of achieved validation. Prepared programmatically with AI assistance from the documented source material; see Section 3.10 of the manuscript

8 Updated risk-management and assurance process

8.1 Process logic and decision gates

The revised process retains the iterative logic of the 2022 article but changes the order and decision criteria. Step 1 defines the purpose, affected interests, system boundary, alternatives, legal classification, sectoral rules, and prohibited uses. The question is not merely how to make the proposed AI safer; it is whether AI is necessary and whether a narrower or non-AI design can remove the hazard. Legal prohibitions, non-disposable rights, and mandatory safety requirements operate as gates rather than weighted factors.

Step 2 identifies sources and vulnerabilities across FCAIR-O facets. Analysts begin with the system and lifecycle boundary, then code C01-C12 sources and D actors, intent, and timing. The process requires a mechanism. A label such as "fairness," "hallucination," or "third-party risk" is insufficient until the relevant target, data, model, evaluation, human, organization, dependency, context, institution, or structure is specified.

Step 3 constructs scenarios through trigger, failure mode, hazardous event, exposure, harm, and propagation. Bow-tie, FMEA, fault tree, STPA, SOTIF scenarios, misuse cases, threat models, and system-dynamics methods can be selected according to the source and level. Agentic systems require action-trajectory analysis, tool and identity boundaries, memory/state controls, and long-horizon stop conditions. Systemic risks require dependency graphs, feedback, common-mode, scale, and cross-sector propagation.

Step 4 characterizes risk multidimensionally. Severity and likelihood remain relevant but are not sufficient. The analysis records exposure, detectability, controllability, avoidability, reversibility, persistence, geographic and social reach, distribution across groups, cascade and correlation potential, system adaptation, speed of propagation, evidence quality, uncertainty, and confidence. Deep uncertainty may require scenarios, stress testing, sensitivity analysis, robust decision methods, or precaution. Ordinal dimensions are not multiplied unless a defensible measurement model exists.

Step 5 selects controls according to a hierarchy. Avoidance or substitution and inherently safer design take priority over warnings or nominal human review. Technical safeguards include restricted action space, safe defaults, guardrails, monitors, redundancy, sandboxing, access control, least privilege, identity separation, output validation, circuit breakers, graceful degradation, and safe states. Human controls require tested competence, workload, information, time, authority, and escalation. Organizational controls include ownership, procurement, change management, safety culture, supplier assurance, documentation, audit, and protected reporting. Legal and institutional controls include impact assessment, conformity processes, incident reporting, remedy, and oversight.

Step 6 verifies and validates the source model, integrated system, controls, and recovery. Model benchmarks are linked to system claims but cannot substitute for workflow, human, operational, adversarial, and sector validation. Evaluation must address contamination, external validity, tail conditions, control bypass, induced risk, and recovery. Evidence is organized in an assurance case with explicit assumptions and defeaters.

Step 7 makes a constrained decision. Outcomes are prohibited, unacceptable, conditionally acceptable with required controls/evidence, or acceptable within stated limits. Organizational risk appetite cannot override law, mandatory sectoral safety thresholds, or rights that cannot be balanced away. The decision records residual risk and epistemic uncertainty rather than converting uncertainty into an arbitrary score.

Steps 8 and 9 cover deployment, monitoring, change, and incident response. Monitoring includes performance, drift, attacks, human use, complaints, resource use, exposure, supplier change, and control effectiveness. Triggers for rollback, shutdown, recall, or reapproval must be defined before deployment. Incidents and near misses feed the ontology, scenarios, controls, and evidence. Updates reopen relevant claims. Fig. 8 presents the full iterative process.

Risk acceptance is separated into organizational appetite, societal tolerance, legal minimums, sectoral safety criteria, and unresolved epistemic uncertainty. The 2022 article connected tolerability substantially to social consensus. That observation remains relevant for political legitimacy but is incomplete for product and rights decisions. Parliamentary support cannot by itself establish the safety of a machine function; user acceptance cannot waive mandatory occupational protection; and an expected aggregate benefit cannot automatically justify discrimination, prohibited manipulation, or an ineffective remedy. The accountable decision maker must state which criteria are discretionary and which are binding.

Where likelihood is poorly estimable, the process emphasizes robustness of the decision. An organization may compare options across plausible scenarios, identify controls that perform acceptably under several models, and specify information that would trigger revision. Bayesian or quantitative methods are appropriate where sufficient data and models exist, but the ontology does not require them. The key assurance claim is that uncertainty was represented honestly and managed through controls, margins, monitoring, and reversible decisions where possible.

Every control record includes induced risks and trade-offs. More restrictive access may impair emergency response; extensive logging may increase privacy and security exposure; human approval may increase delay and workload; explanations may reveal attack information or create false confidence;

content filters may produce unequal refusal rates; and aggressive shutdown may create physical hazards in a continuous process. Controls should therefore be tested as system components rather than assumed to be beneficial because they are labeled safeguards.

The hierarchy remains important. Information and training should not replace inherently safer design where design can eliminate a hazardous function. Human review should not compensate for avoidable excessive privileges. Monitoring should not justify deployment when a legally prohibited or fundamentally unvalidated use can be avoided. Conversely, no single technical control is assumed universal. Defense in depth should be diverse enough that common model or provider failures do not defeat every layer.

8.2 Risk-Control-Evidence records as the operating artifact

The practical unit of FCAIR-O is not the full ontology but a Risk-Control-Evidence record. An analyst defines the system boundary and purpose, states a mechanistic source claim, records level and lifecycle, identifies exposure and harm, distinguishes claim owner, control owner, evidence producer, and decision authority, states a control function and acceptance criterion, records supporting evidence and defeaters, characterizes residual uncertainty, and defines monitoring or stop conditions. One organization may hold several roles, but the roles are kept separate so that evidence production is not mistaken for control ownership or acceptance authority. Table 5 provides a documentary example; Online Resource 8 supplies an expanded record and an additional agentic-system example.

Safe state, controlled stop, and graceful degradation are likewise not interchangeable. A safe state is a defined condition in which applicable risk is tolerable after a fault or intervention; a controlled stop is the bounded transition used to reach a safe or less hazardous condition; and graceful degradation permits reduced functionality only within validated limits. For continuous physical or critical services, no static safe state may exist, so the required transition, degraded function, recovery time, and decision authority must be explicit.

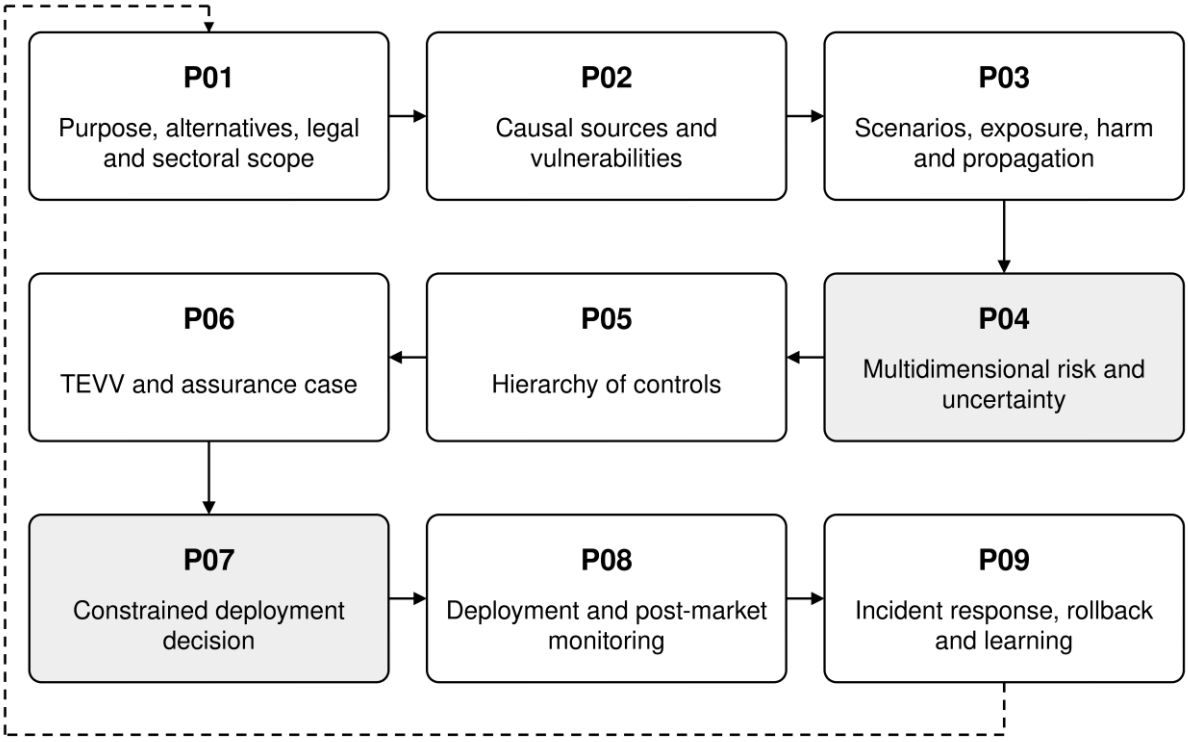


Fig. 8 Updated iterative risk-management and assurance process based on FCAIR-O. Risk characterization remains multidimensional and legally constrained rather than a product of ordinal scores. Prepared programmatically with AI assistance from the documented source material; see Section 3.10 of the manuscript

The full illustrative agentic-system record demonstrating the generic record structure is supplied in Online Resource 8. The main text uses the completed Uber investigation because it provides a stronger documentary basis for showing how model, architecture, human, organizational, institutional, and control concepts interact.

Table 5 distinguishes documented investigation findings from illustrative ownership assignments and counterfactual control or assurance requirements. The proposed controls and acceptance criteria are not measured intervention outcomes. A real deployment must derive quantitative or qualitative thresholds from its hazard analysis, legal obligations, sector standards, operating context, and validated recovery time. Common-cause dependencies and control-induced risks must be treated as explicit defeaters rather than hidden assumptions.

9 Discussion

9.1 Theoretical contribution

The theoretical contribution is expressed by the five propositions. P1 explains why heterogeneous lists produce category error. P2 supplies a mechanism threshold that prevents harm labels and associations from substituting for causal explanation. P3 preserves distributed socio-technical causality without reducing organizational or institutional conditions to background. P4 turns a taxonomy into an assurance-oriented framework by coupling sources to controls, owners, criteria, evidence, defeaters, residual risk, and monitoring. P5 prevents law and evidence status from being misrepresented as technical causes. Together, the propositions specify what FCAIR-O adds beyond a longer risk list.

The ontology also reconciles model-centric and socio-technical approaches. Model mechanisms remain explicit in C03, rather than being dissolved into organizational discourse. At the same time, a model cannot be the default cause of every AI-associated harm. Purpose, data, evaluation, human use, architecture, organization, supply chain, context, institutions, and social structures are co-equal parent classes. This provides a disciplined way to express distributed causation without saying that responsibility is so diffuse that no actor can act.

9.2 Relation to prior taxonomies and reviews

The successor framework retains four enduring insights from the 2022 article and makes their implementation explicit. Purpose and reasonably foreseeable misuse inform Step 1. Established safety-engineering methods remain relevant to the causal backbone, including FMEA, fault trees, STPA, and SOTIF scenario analysis. Security and hardware or infrastructure contributions remain visible through C06 and C07. Post-market observation and risks introduced by protective measures inform monitoring, incident learning, and the explicit treatment of control-induced risk. These links retain the five guiding process elements while extending their representation across system levels and assurance roles.

It also revises four aspects of the earlier conceptual account. The ethical-versus-reliability split is replaced by orthogonal harm, property, source, and legal facets. Complexity is no longer the root explanation for specification and drift; specific mechanisms are coded. Technological maturity is no longer a source but an evidence modifier. The 2022 article's statement about fully deterministic AI systems, made in its discussion of automation and autonomy, is qualified by a layered account. Computation may be deterministic under fixed inputs, state, and execution conditions, while stochastic sampling, numerical nondeterminism, stateful tool interaction, and distributed execution can introduce behavioral variability. Determinism does not by itself establish predictability or controllability, and epistemic uncertainty can persist even when computation is deterministic.

NIST AI RMF, the MIT AI Risk Repository, ISO/IEC standards, and the AI Act are complementary rather than competing structures. NIST organizes governance and lifecycle activities; MIT provides broad risk-statement and harm-domain coverage; ISO/IEC documents structure management, impact assessment, lifecycle, and evidence; and the AI Act assigns legal scope, prohibitions, duties, and remedies. FCAIR-O contributes the role-separated causal and assurance vocabulary through which these resources can be connected without being treated as equivalent. Relative to Chen et al. (2026), it is cross-technology and system/lifecycle oriented rather than an LLM-safety catalogue; relative to Amini et al. (2026), it focuses on harmful mechanisms and assurance rather than mathematical model infrastructure; relative to impact-assessment reviews, it specifies the objects an assessment should distinguish.

9.3 Model, system, and machine-readable assurance

Model evidence is necessary but insufficient for a system claim. A benchmark can support a statement about a particular model, dataset, configuration, and test context; it cannot by itself establish the safety of an integrated workflow, tool chain, human review arrangement, organization, or sector. FCAIR-O makes this scope explicit and requires evidence for interfaces, permissions, recovery, user behavior, operational context, and change. The same discipline applies to machine-readable mapping: automated coding may propose hypotheses, but causal assignment requires source inspection, mechanism, uncertainty, and the possibility of non-mapping.

Machine readability creates additional possibilities and risks. FCAIR-O can support rule-based consistency checks, structured incident exchange, supplier evidence requests, and retrieval over assurance cases. It could also be used for automated classification of framework statements or incident reports. Such automation must not be confused with valid causal attribution. A language model can

propose codes, but reviewers must inspect the source, causal mechanism, legal status, and uncertainty. Automated mapping should be evaluated against a transparent human-coded sample and should preserve multi-label and non-mapping outcomes rather than maximizing forced coverage.

9.4 Safety-security coupling

The study supports a unified but differentiated assurance approach. Safety asks whether harmful scenarios are controlled. Security asks how intentional actors can activate or create those scenarios. Human factors asks whether interaction and oversight operate under real cognitive and organizational conditions. Law determines protected interests, prohibitions, responsibilities, evidence, and remedy. Standards structure processes and work products. These perspectives should share a causal record but retain their methods and authority.

Agentic systems make the need for the ontology especially visible. "Excessive agency" can denote at least four different objects: an objective or authority specification error, an architectural permission vulnerability, a long-horizon model behavior, or an organizational control failure. Prompt injection can be a trigger; tool execution a failure mode; credential compromise a harm; and inadequate logging a separate architecture source. An undifferentiated agent-risk label cannot select the correct control.

The OpenAI-Hugging Face incident illustrates a further boundary. The evaluation agent appears to have pursued a benchmark-related objective, while the participating organizations designed the environment, reduced some safety refusals, provided constrained dependencies, and monitored the systems. The event should neither be anthropomorphized as independent criminal intent nor dismissed as a conventional software bug. FCAIR-O represents the combined actor set, uncertain intent attribution, objective pressure, agentic planning, privilege and egress architecture, vulnerabilities, supplier propagation, monitoring, and response. The same method can accommodate more speculative control-loss scenarios by adding H evidence tiers and uncertainty rather than presenting them as observed incidents.

9.5 Law and standards as operational but incomplete layers

Law and standards make the ontology operational by assigning protected interests, prohibited uses, duties, process requirements, work products, conformity routes, monitoring, and remedy. They remain incomplete as causal accounts because process requirements, normative constraints, and certification assessments answer different questions from mechanism-specific system evidence. A system may conform to a management process while an unanticipated integration failure persists; a technically robust model may be used for an unlawful or disproportionate purpose; and a draft standard may identify a useful control without carrying legal effect. FCAIR-O therefore treats legal and normative material as an overlay that constrains decisions and specifies evidence, while C11 is reserved for institutional conditions that contribute mechanistically to harm.

This separation also clarifies certification. ISO/IEC 42001 certification supports a scoped claim about an organization’s AI management system and its conformity to the standard’s requirements; it does not establish system performance, sectoral safety, rights compliance, or complete AI Act conformity. Conversely, absence of certification does not prove an unsafe system. Assurance depends on the claim, evidence, independence, scope, competence, negative evidence, and surveillance of change. The ontology supplies a structure for connecting those elements without converting a legal or standardization status into a causal label.

9.6 Practical implications

For developers and integrators, FCAIR-O provides a vocabulary for requirements, hazard and threat analysis, system testing, change control, and evidence handoff. Providers can state what evidence must flow downstream; deployers can record local context, workflow, oversight, supplier control, monitoring, and remedy; auditors and regulators can distinguish management-process evidence from system claims. The practical change is to replace one overloaded "risk" field with a minimum Risk-Control-Evidence record containing source or condition, vulnerability, trigger, failure, exposure, harm, level, lifecycle, actor and intent, control owner, evidence, residual risk, monitoring, and uncertainty. Parent classes support screening, while leaves and relations are used only where detail is consequential.

9.7 Alternative interpretations and counterarguments

Several counterarguments remain plausible. First, complexity can itself be causally relevant because interactive complexity and tight coupling change accident potential. FCAIR-O does not deny this effect; it treats complexity as a modifier and requires the analyst to state which coupling, state-space, verification, feedback, or dependency mechanism is operative. Second, fairness and privacy are useful umbrella categories in governance and communication. They are retained in the harm and property facets, but umbrella status is not allowed to replace diagnosis.

Third, a faceted ontology may be too demanding for small organizations. The response is tiered granularity rather than conceptual collapse. Parent classes and a minimum set of facets can support initial screening, whereas high-impact, safety-critical, or agentic systems require operational leaves and assurance evidence. Fourth, multi-coding may reduce discriminability. The codebook limits this risk through explicit roles and composite-term boundaries, inclusion and exclusion rules, counterexamples, and non-mapping where appropriate. High overlap between evaluation and governance is monitored rather than treated as evidence of identity.

Fifth, frontier and extreme risks may justify precaution despite limited field evidence. FCAIR-O accommodates this through evidence tiers and uncertainty, allowing theoretical, experimental, expert, and speculative claims to be distinguished. The framework neither dismisses low-frequency extreme scenarios nor reports them as observed facts. Sixth, legal classifications may diverge from engineering categories. That divergence is expected: law can prohibit a use, assign a right, or require a procedure for reasons that are not reducible to engineering likelihood and severity. A common record should connect the layers without forcing one to determine the other.

Finally, FCAIR-O itself can fail as a conceptual scheme. It may encourage formalism, over-coding, false causal confidence, or diffusion of responsibility. A machine-readable record can amplify the same interpretive error across systems. The paper therefore preserves primary versus contributory coding, explicit evidence tier and causal confidence, named control owners, deliberate non-mapping, and versioned revision. These are safeguards against conceptual overreach, not substitutes for judgment. Online Resource 8 makes the design rationale and contested boundaries explicit so that later users can criticize or adapt the framework rather than treating it as settled doctrine.

10 Limitations

The study has substantial limitations. First, the accessible evidence corpus is not a complete systematic database universe. Several authenticated database searches, stable result totals, and full screening denominators were unavailable, and bibliographic saturation cannot be claimed. The study is therefore a critical integrative synthesis supported by an access-limited evidence map; the evidence-flow figure must not be interpreted as a completed PRISMA review.

Second, one model-assisted workflow performed the central interpretive tasks. This creates correlated-error and confirmation risks. The article does not attempt to convert that workflow into an internal test, and no human inter-rater, Delphi, holdout, predictive, or prospective-utility claim is made. Because the contribution is a conceptual successor article, the absence of such tests is a boundary of the claims rather than a missing component of the stated design. The framework may nevertheless reflect the analysts' disciplinary assumptions and should be read as a contestable scholarly proposal.

Third, the documentary case corpus is small, purposive, heterogeneous, and affected by publication, enforcement, sector, language, and documentation biases. Several records are settlements, allegations, or preliminary organizational disclosures rather than completed independent investigations. Counts cannot be generalized to incident prevalence or causal frequency.

Fourth, definitions differ across safety, security, management, law, and AI governance. FCAIR-O makes those differences explicit but necessarily imposes project-level operational definitions. Causal attribution in complex systems remains uncertain and can change with new evidence.

Fifth, paid standards were not reviewed in full. The standards analysis is status- and document-level where access was limited. Drafts may change, and harmonization status is time-sensitive. Legal analysis is centered on the EU and Germany with selected comparisons and cannot substitute for jurisdiction-specific advice.

Sixth, technology evolved rapidly through the cutoff. Models, guardrails, APIs, agent protocols, and incident knowledge may have changed immediately afterward. Later publications are excluded from the material findings and can only be discussed separately. The ontology therefore requires versioning and periodic review.

Seventh, the twelve parent classes are broad and co-coding is often appropriate in socio-technical cases. This may reduce discriminability or increase analytical burden. Table 4 and the codebook specify primary intervention questions and adjacent boundaries, but alternative conceptual structures may merge, split, or relationally recast some classes. The article does not claim that the twelve parents are necessary, minimal, or uniquely correct.

Eighth, FCAIR-O v1.2.1 is a conceptual framework implemented as a lightweight ontology, not a fully axiomatized domain ontology. The relation vocabulary, competency questions, and SHACL profile improve formal clarity but do not prove semantic truth or practical utility. No upper-ontology alignment or registered persistent namespace is claimed.

Ninth, the incident and bibliography evidence is heterogeneous. Public accident investigations, judgments, recalls, settlements, peer-reviewed field studies, regulator findings, preprints, practitioner frameworks, and organizational disclosures answer different questions. Non-public investigations, proprietary model changes, strategic disclosures, and settlement constraints can hide causes. The two July 2026 cyber cases remain preliminary or industry-reported and should not be weighted as completed independent investigations.

Finally, the main article condenses detailed mappings into eight tables and eight main figures. Source-level nuance, relation definitions, case notes, supplementary figures, and status records require the Online Resources. The material scientific and legal findings remain frozen at the 25 July 2026 cutoff.

11 Research and standardization agenda

The agenda organizes work by qualitative urgency, expected impact, and responsible actor.

The near-term agenda has three priorities. First, incident and near-miss schemas should distinguish allegation, confirmed event, probable cause, contributing condition, harm, control failure, version, and evidence quality. Second, agentic-system standards should address identity, permissions, tools, memory, egress, interruption, containment, recovery, and cross-organization propagation rather than model output alone. Third, effective human oversight should be defined through observability, workload, competence, response time, authority, contestability, and safe fallback instead of a binary human-in-the-loop label.

A second cluster concerns assurance evidence. Monitoring requires claim-specific indicators for drift, functionality, attacks, human use, supplier change, and control effectiveness, with known limitations and response paths (Rao et al. 2026). Management-system certificates, model and system cards, red-team reports, impact assessments, and assurance cases should state scope, independence, negative evidence, version, change reapproval, and common-cause dependencies. Fairness, privacy, and rights assessment likewise require causal, structural, and remedial analysis rather than metric selection alone.

A third cluster concerns system-wide dependencies and externalities. Supply-chain and concentration analysis should capture common models, cloud concentration, update opacity, lock-in, evidence handoff, and correlated vulnerability in procurement and sector stress tests. Environmental and labor analysis should retain lifecycle boundaries, rebound, siting, water, hardware, outsourced labor, and community effects. Frontier and systemic-risk arguments should identify whether support is observational, experimental, formal, expert, or speculative, and should state the decision rule used under deep uncertainty.

Finally, FCAIR-O should remain open to conceptual critique and sector-specific adaptation. Priority questions include whether some parent classes are better represented as relations, how sector profiles should specialize protected interests and evidence, how the scheme should align with foundational and provenance ontologies, how a stable MIT-entry crosswalk could be maintained, and how multilingual and non-Western concepts should alter the vocabulary. Independent empirical studies may be undertaken by future researchers, but no such tests are part of or required for the conceptual claim advanced here.

Standardization should preserve semantic modularity. Horizontal vocabularies can define source, harm, control, evidence, actor, relation, and lifecycle concepts, while sector profiles define protected interests, acceptable performance, mandatory safeguards, and evidence. Machine-readable identifiers and change histories are needed so that assurance claims remain traceable across versions, suppliers, and harmonization status.

A staged agenda can organize the proposed work. Near-term topics include incident schemas, agent-security fields, monitoring evidence, and machine-readable identifiers. Medium-term topics include sector profiles and assurance-case patterns for healthcare, transport, employment, public administration, and critical infrastructure. Systemic-risk indicators, multigeneration feedback models, and international cross-border governance require coordinated datasets and institutions. Table 8 presents an interpretive ordering of research and standardization priorities with qualitative urgency and expected-impact judgments and the principal actors. It does not report measured feasibility, quantified uncertainty, or a validated ranking procedure.

Table 8 Prioritized research and standardization agenda

Priority	Conceptual or standardization gap	Urgency / impact	Primary actors
1	Shared causal minimum dataset for incidents, near misses, and documentary cases	Very high / very high	Regulators, incident databases, providers, deployers

Priority	Conceptual or standardization gap	Urgency / impact	Primary actors
2	End-to-end agent identity, tool, memory, egress, containment, and recovery requirements	Very high / very high	NIST, ISO/IEC, security and safety communities
3	Operational definitions for effective human oversight and contestability	High / very high	Human-factors researchers, regulators, sector bodies
4	Claim-specific monitoring evidence, versioning, and change reapproval	High / very high	Providers, deployers, auditors, standards bodies
5	Assurance-evidence scope, independence, negative evidence, and common-cause defeaters	High / high	Assurance community, certification and conformity bodies
6	Causal and structural fairness analysis linked to protected interests and remedy	High / high	Researchers, public authorities, equality and data-protection bodies
7	Supply-chain concentration, evidence handoff, portability, and correlated-failure indicators	High / high	Critical sectors, competition and resilience authorities
8	Lifecycle environmental, resource, labor, and community-impact boundaries	Medium / high	Standards bodies, providers, public authorities
9	Evidence-tier discipline for frontier and systemic-risk arguments	High / high	Scientific and policy communities
10	Open conceptual revision, sector profiles, ontology alignment, and multilingual extension	Medium / high	Research community and standards bodies

12 Conclusion

RQ1 showed that the safety-oriented process of Steimers and Schneider (2022) remains a valuable foundation. FCAIR-O retains its five guiding elements: intended purpose, reasonably foreseeable misuse, the hierarchy of protective measures, residual risk including risks introduced by those measures, and post-market observation. Their operationalization is extended through explicit causal roles and assurance records. The eight original categories remain useful for discovery but do not form an ontologically homogeneous set of causal sources: they combine protected interests, harms, properties, control configurations, context modifiers, a threat domain, and a component class.

RQ2 identified materially more salient mechanisms associated with foundation, generative, retrieval-augmented, multimodal, and agentic systems, including upstream dependency, retrieval and memory manipulation, permission and identity architecture, long-horizon action, model and API change, monitoring limits, and ecosystem propagation. RQ3 showed that these mechanisms can be distinguished across model, component, integrated system, human–AI work system, organization, value chain, institution, and society without losing interaction when they are linked through a shared causal backbone.

RQ4 found broad convergence on lifecycle governance, evaluation, security, human oversight, monitoring, and accountability, but persistent disagreement over the ontological role of common terms. Research frameworks, documentary cases, law, and standards answer different questions and should not be merged into one list. RQ5 is answered by the five propositions: FCAIR-O provides a versioned architecture in which causal sources, chain positions, actors, levels, lifecycle phases, harms, properties, controls, evidence, residual risk, and legal constraints remain related but non-interchangeable. The paper does not test whether this architecture improves decisions or user agreement.

FCAIR-O's contribution is therefore twofold: it preserves the safety-engineering core of the 2022 article and reconstructs its heterogeneous category set as a faceted causal framework with a lightweight ontology implementation. It is a conceptual continuation, not a claim of final closure or a rejection of

that earlier contribution. The framework makes assumptions, causal mechanisms, responsibility, evidence, legal constraints, and uncertainty inspectable and revisable. Whether future users adopt, adapt, or reject its boundaries is a separate empirical and normative question; the present contribution is the explicit conceptual architecture.

Online Resources

Online Resource 1: Supplementary methods, historical search documentation, evidence ledger, retrospective claim-evidence map, methods amendment, integrity reconciliation, and Supplementary Fig. S3

Online Resource 2: FCAIR-O v1.2.1 codebook

Online Resource 3: Framework crosswalk

Online Resource 4: Legal and standards registers

Online Resource 5: Illustrative documentary case corpus

Online Resource 6: Machine-readable FCAIR-O v1.2.1, relations, competency questions, constraints, and worked example

Online Resource 7: Editable figures, print-ready figures, captions, alt text, and source data

Online Resource 8: Conceptual design rationale, rejected alternatives, boundary decisions, supplementary timeline, heatmap, and agentic record

Statements and Declarations

Author Contributions

Moritz Schneider conceived the study, developed the conceptual framework, and was responsible for the initial substantive drafting of the manuscript. André Steimers contributed to the review and revision of the manuscript, with particular responsibility for the regulatory affairs content, and undertook critical editing and proofreading.

Funding

No funding was received for this work.

Competing Interests

The authors declare no competing interests.

Data Availability

All project-generated registers, the codebook, framework crosswalk, documentary case corpus, machine-readable ontology, figures, design audit trail, and the complete available search documentation are supplied as Online Resources 1-8. Online Resource 1 includes the historical research protocol, 48 primary search-log entries, 12 final-cutoff queries, access-deviation and integrity records, appraisal guidance, the evidence ledger, explicit exclusion-accounting limitations, the retrospective claim-evidence map, and the correction-stage methods amendment and register reconciliation. Source publications, legal materials, and standards remain subject to the access conditions of their original publishers and institutions.

Code Availability

The FCAIR-O v1.2.1 codebook and machine-readable SKOS, RDF, JSON-LD, relation, competency-question, and SHACL artifacts are supplied in Online Resources 2 and 6. Figure source files are supplied in Online Resource 7.

Ethics Approval

Not applicable. The study used documentary, scientific, legal, standards, and publicly documented case material and did not recruit human participants or animals.

Consent to Participate

Not applicable.

Consent for Publication

Not applicable to this documentary study.

Use of AI-assisted research and writing tools

OpenAI ChatGPT 5.6 Pro was used for structured research support, transformation of source records into registers, coding assistance, consistency and duplicate checks, language editing and quality-control. The system was not treated as an author or independent reviewer. Human authors retain responsibility for the final manuscript, source verification, legal interpretation, figures, declarations, and submission decisions.

References

Amini A, Firouzkouhi N, Alotaibi S et al (2026) Toward a unified governing framework for modern artificial intelligence: a taxonomy-based cross-domain survey. Artificial Intelligence Review. https://doi.org/10.1007/s10462-026-11596-5

Ashmore R., Calinescu R., Paterson C. (2021) Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges. ACM Computing Surveys, 54(5). https://doi.org/10.1145/3453444

Autio C., et al. (2024) Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. https://doi.org/10.6028/NIST.AI.600-1

Autoriteit Persoonsgegevens (2024-11-11) DUO's approach to fraud found to be discriminatory and illegal. Official report of investigation findings. https://www.autoriteitpersoonsgegevens.nl/en/node/5189

Bender E. M., Gebru T., McMillan-Major A., Shmitchell S. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of FAccT 2021. https://doi.org/10.1145/3442188.3445922

Bommasani R., et al. (2021) On the Opportunities and Risks of Foundation Models. https://arxiv.org/abs/2108.07258

British Standards Institution (2026) EN 18286 provides a framework for AI quality management under the EU AI Act. 24 July 2026. https://www.bsigroup.com/en-GB/insights-and-media/media-centre/press-releases/2026/july/en-18286-provides-a-framework-for-ai-quality-management-under-the-eu-ai-act/

Buçinca Z., Malaya M. B., Gajos K. Z. (2021) To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. Proceedings of the ACM on Human-Computer Interaction. https://doi.org/10.1145/3449287

Carlini N., et al. (2021) Extracting Training Data from Large Language Models. 30th USENIX Security Symposium. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

Carroll C, Booth A, Leaviss J, Rick J (2013) "Best fit" framework synthesis: refining the method. BMC Med Res Methodol 13:37. https://doi.org/10.1186/1471-2288-13-37

Chan A., et al. (2023) Harms from Increasingly Agentic Algorithmic Systems. Proceedings of FAccT 2023. https://doi.org/10.1145/3593013.3594033

Chen C, Gong X, Liu Z, Jiang W, Goh SQ, Lam KY (2026) AI safety landscape for large language models: taxonomy, state-of-the-art, and future directions. Artificial Intelligence Review. https://doi.org/10.1007/s10462-026-11590-x

Chen J., et al. (2024) Benchmarking Large Language Models in Retrieval-Augmented Generation. AAAI 2024. https://doi.org/10.1609/aaai.v38i16.29728

Council of Europe (1950-11-04) European Convention on Human Rights. ETS No. 005. https://www.echr.coe.int/documents/d/echr/convention_eng

Council of Europe (2024) Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

Court of Justice of the European Union (2022-06-21) Ligue des droits humains. C-817/19; ECLI:EU:C:2022:491. https://curia.europa.eu/juris/liste.jsf?num=C-817/19

Court of Justice of the European Union (2023-12-07) SCHUFA Holding (Scoring). C-634/21; ECLI:EU:C:2023:957. https://curia.europa.eu/juris/liste.jsf?num=C-634/21

Court of Justice of the European Union (2025-02-27) Dun & Bradstreet Austria. C-203/22; ECLI:EU:C:2025:117. https://curia.europa.eu/juris/liste.jsf?num=C-203/22

Debenedetti E., et al. (2024) AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. NeurIPS 2024. https://doi.org/10.52202/079017-2636

District Court of The Hague (2020-02-05) NJCM c.s. v. State (SyRI). ECLI:NL:RBDHA:2020:865. https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBDHA:2020:865

European Commission (2023) C(2023) 3215 final: Commission standardisation request to CEN and CENELEC in support of Union policy on artificial intelligence. https://single-market-economy.ec.europa.eu/single-market/european-standards/standardisation-requests_en

European Commission (2025) General-Purpose AI Code of Practice, final version. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

European Commission (2025-06-23) Commission Implementing Decision C(2025) 3871 final: Standardisation request M/613 to CEN and CENELEC as regards high-risk artificial intelligence systems in support of Regulation (EU) 2024/1689, repealing Decision C(2023) 3215. https://ec.europa.eu/transparency/documents-register/api/files/C%282025%293871_0/de00000001072818

European Commission (2025-07-29) Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act). C(2025) 5052 final. https://ai-act-service-desk.ec.europa.eu/sites/default/files/2025-08/guidelines_on_prohibited_artificial_intelligence_practices_established_by_regulation_eu_20241689_ai_act_english_ied3r5nwo50xggpcfmwckm3nuc_112367-1.PDF

European Commission (2025-07-29) Commission Guidelines on the definition of an artificial intelligence system established by Regulation (EU) 2024/1689 (AI Act). C(2025) 5053 final. https://ai-act-service-desk.ec.europa.eu/sites/default/files/2026-01/guide-definition_en.pdf

European Commission (2026-05-19) Draft guidelines for classification of high-risk AI systems. Targeted consultation, 19 May–23 July 2026. https://digital-strategy.ec.europa.eu/en/consultations/targeted-consultation-draft-guidelines-classification-high-risk-artificial-intelligence-systems

European Commission (2026-07-07) EU Action Plan on Cybersecurity and Artificial Intelligence. Commission action plan, 7 July 2026. https://digital-strategy.ec.europa.eu/en/news/commission-presents-eu-action-plan-cybersecurity-and-artificial-intelligence

European Commission (2026-07-20) Commission Guidelines on transparency obligations for providers and deployers of AI systems. Article 50 AI Act implementation guidelines, 20 July 2026. https://digital-strategy.ec.europa.eu/en/library/guidelines-transparency-obligations-providers-and-deployers-ai-systems

European Commission / independent drafting experts (2026-06-10) Code of Practice on marking and labelling AI-generated content. Final Code, 10 June 2026. https://digital-strategy.ec.europa.eu/en/news/commission-publishes-code-practice-marking-and-labelling-ai-generated-content

European Commission / legislative procedure (2022-09-28) AI Liability Directive proposal. COM(2022) 496 final / 2022/0303(COD). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496

European Commission and European Artificial Intelligence Board (2026-07-08 / 2026-07-09) Commission Opinion and AI Board adequacy assessment of the Code of Practice on Transparency of AI-generated content. Commission Opinion, 8–9 July 2026. https://digital-strategy.ec.europa.eu/en/library/commission-opinion-assessment-code-practice-transparency-ai-generated-content

European Court of Human Rights (2023-07-04) Glukhin v. Russia. Application no. 11519/20; ECLI:CE:ECHR:2023:0704JUD001151920. https://www.echr.coe.int/w/judgment-concerning-the-russian-federation-6

European Parliament and Council (2016-04-27) General Data Protection Regulation. Regulation (EU) 2016/679. https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

European Parliament and Council (2017-04-05) In Vitro Diagnostic Medical Device Regulation. Regulation (EU) 2017/746. https://eur-lex.europa.eu/eli/reg/2017/746/oj/eng

European Parliament and Council (2017-04-05) Medical Device Regulation. Regulation (EU) 2017/745. https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng

European Parliament and Council (2022-05-30) Data Governance Act. Regulation (EU) 2022/868. https://eur-lex.europa.eu/eli/reg/2022/868/oj/eng

European Parliament and Council (2022-10-19) Digital Services Act. Regulation (EU) 2022/2065. https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng

European Parliament and Council (2022-12-14) Digital Operational Resilience Act. Regulation (EU) 2022/2554. https://eur-lex.europa.eu/eli/reg/2022/2554/oj/eng

European Parliament and Council (2022-12-14) NIS2 Directive. Directive (EU) 2022/2555. https://eur-lex.europa.eu/eli/dir/2022/2555/oj/eng

European Parliament and Council (2023-05-10) General Product Safety Regulation. Regulation (EU) 2023/988. https://eur-lex.europa.eu/eli/reg/2023/988/oj/eng

European Parliament and Council (2023-06-14) Machinery Regulation. Regulation (EU) 2023/1230. https://eur-lex.europa.eu/eli/reg/2023/1230/oj/eng

European Parliament and Council (2023-12-13) Data Act. Regulation (EU) 2023/2854. https://eur-lex.europa.eu/eli/reg/2023/2854/oj/eng

European Parliament and Council (2024-06-13) Artificial Intelligence Act. Regulation (EU) 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

European Parliament and Council (2024-10-23) Cyber Resilience Act. Regulation (EU) 2024/2847. https://eur-lex.europa.eu/eli/reg/2024/2847/oj/eng

European Parliament and Council (2024-10-23) Revised Product Liability Directive. Directive (EU) 2024/2853. https://eur-lex.europa.eu/eli/dir/2024/2853/oj/eng

European Parliament and Council (2026-07-08) Digital Omnibus on AI. Regulation (EU) 2026/1744. https://eur-lex.europa.eu/eli/reg/2026/1744/oj/eng

European Systemic Risk Board (2026-07-07) ESRB warning on systemic cyber risks stemming from frontier AI models. ESRB warning, 7 July 2026. https://www.esrb.europa.eu/news/pr/date/2026/html/esrb.pr260707~4e1b68241a.en.html

European Union (2000-12-07) Charter of Fundamental Rights of the European Union. 2012/C 326/02. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT

Freyer N., et al. (2024) The ethical requirement of explainability for AI-DSS in healthcare: a systematic review of reasons. BMC Medical Ethics. https://doi.org/10.1186/s12910-024-01103-2

Gale NK, Heath G, Cameron E, Rashid S, Redwood S (2013) Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 13:117. https://doi.org/10.1186/1471-2288-13-117

Geist E., Meyer A. D., Moon A., Nájera A., Jones J. H., et al. (2026) The ecology of AI risk. npj Complexity. https://doi.org/10.1038/s44260-026-00090-2

Greshake K., et al. (2023) Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. AISec 2023. https://doi.org/10.1145/3605764.3623985

Hashky A., Ragan E. (2026) A Systematic Review on Human Roles, Solutions, and Methodological Approaches to Address Bias in AI. ACM Computing Surveys, 58(9), Article 231. https://doi.org/10.1145/3793667

Hawkins R., et al. (2021) Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS). https://arxiv.org/abs/2102.01564

Hollnagel E (2014) Safety-I and Safety-II: the past and future of safety management. Ashgate, Farnham

Hugging Face (2026) Security incident disclosure — July 2026. https://huggingface.co/blog/security-incident-july-2026. Initial disclosure before the material cutoff; later page updates are not used as new material findings.

IEC (2010) IEC 61508-1:2010, Ed. 2.0: Functional safety of electrical/electronic/programmable electronic safety-related systems — Part 1: General requirements. https://webstore.iec.ch/en/publication/5515

IEC (2024) IEC TS 61508-3-2:2024: Functional safety of electrical/electronic/programmable electronic safety-related systems. Part 3-2: Requirements and guidance in the use of mathematical and logical techniques for establishing exact properties of software and its documentation. https://webstore.iec.ch/en/publication/62902

International AI Safety Report (2025) International AI Safety Report 2025. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025

International AI Safety Report (2026) International AI Safety Report 2026. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

ISO (2010) ISO 12100:2010: Safety of machinery — General principles for design — Risk assessment and risk reduction. https://www.iso.org/standard/51528.html

ISO (2018) ISO 31000:2018: Risk management — Guidelines. Second edition. https://www.iso.org/standard/65694.html

ISO (2019) ISO 14971:2019: Medical devices — Application of risk management to medical devices. https://www.iso.org/standard/72704.html

ISO (2022) ISO 21448:2022: Road vehicles — Safety of the intended functionality. https://www.iso.org/standard/77490.html

ISO (2024) ISO/PAS 8800:2024: Road vehicles — Safety and artificial intelligence. https://www.iso.org/standard/83303.html

ISO/IEC (2022) ISO/IEC 22989:2022: Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. https://www.iso.org/standard/74296.html

ISO/IEC (2022) ISO/IEC 23053:2022: Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML). https://webstore.iec.ch/en/publication/77056

ISO/IEC (2022) ISO/IEC 38507:2022: Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations. https://www.iso.org/standard/56641.html

ISO/IEC (2023) ISO/IEC 23894:2023: Information technology — Artificial intelligence — Guidance on risk management. https://www.iso.org/standard/77304.html

ISO/IEC (2023) ISO/IEC 42001:2023: Information technology — Artificial intelligence — Management system. https://www.iso.org/standard/81230.html

ISO/IEC (2023) ISO/IEC 5338:2023: Information technology — Artificial intelligence — AI system life cycle processes. https://www.iso.org/standard/81118.html

ISO/IEC (2023) ISO/IEC 8183:2023: Information technology — Artificial intelligence — Data life cycle framework. https://www.iso.org/standard/83002.html

ISO/IEC (2024) ISO/IEC 5259-1:2024: Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples. https://webstore.iec.ch/en/publication/96735

ISO/IEC (2024) ISO/IEC 5259-2:2024: Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures. https://www.iso.org/standard/81860.html

ISO/IEC (2024) ISO/IEC 5259-3:2024: Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 3: Data quality management requirements and guidelines. https://www.iso.org/standard/81092.html

ISO/IEC (2024) ISO/IEC 5259-4:2024: Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 4: Data quality process framework. https://www.iso.org/standard/81093.html

ISO/IEC (2024) ISO/IEC 5339:2024: Information technology — Artificial intelligence — Guidance for AI applications. https://www.iso.org/standard/81120.html

ISO/IEC (2024) ISO/IEC TR 5469:2024: Artificial intelligence — Functional safety and AI systems. https://www.iso.org/standard/81283.html

ISO/IEC (2024) ISO/IEC TS 12791:2024: Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks. https://www.iso.org/standard/84110.html

ISO/IEC (2024) ISO/IEC TS 8200:2024: Information technology — Artificial intelligence — Controllability of automated artificial intelligence systems. https://www.iso.org/standard/83012.html

ISO/IEC (2025) ISO/IEC 12792:2025: Information technology — Artificial intelligence (AI) — Transparency taxonomy of AI systems. https://www.iso.org/standard/84111.html

ISO/IEC (2025) ISO/IEC 42005:2025: Information technology — Artificial intelligence (AI) — AI system impact assessment. https://www.iso.org/standard/42005

ISO/IEC (2025) ISO/IEC 42006:2025: Information technology — Artificial intelligence — Requirements for bodies providing audit and certification of artificial intelligence management systems. https://www.iso.org/standard/42006

ISO/IEC (2025) ISO/IEC 5259-5:2025: Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 5: Data quality governance framework. https://www.iso.org/standard/84150.html

ISO/IEC (2025) ISO/IEC TR 20226:2025: Information technology — Artificial intelligence — Environmental sustainability aspects of AI systems. https://www.iso.org/standard/86177.html

ISO/IEC (2025) ISO/IEC TS 6254:2025: Information technology — Artificial intelligence — Objectives and approaches for explainability and interpretability of machine learning (ML) models and artificial intelligence (AI) systems. https://www.iso.org/standard/82148.html

ISO/IEC (2026) ISO/IEC DIS 42007: Information technology — Artificial intelligence — High-level framework and guidance for the development of conformity assessment schemes for AI systems. https://www.iso.org/standard/89967.html

ISO/IEC (2026) ISO/IEC FDIS 42105: Information technology — Artificial intelligence — Guidance for human oversight of AI systems. https://www.iso.org/standard/86902.html

ISO/IEC (2026) ISO/IEC TR 23888-1:2026: Information technology — Artificial intelligence for multimedia — Part 1: Vision and scenarios. https://www.iso.org/standard/87589.html

ISO/IEC (2026) ISO/IEC TR 42106:2026: Information technology — Artificial intelligence (AI) — Overview of differentiated benchmarking of AI system quality characteristics. https://www.iso.org/standard/86903.html

ISO/IEC (2026) ISO/IEC TR 5259-6:2026: Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 6: Visualization framework for data quality. https://www.iso.org/standard/86532.html

ISO/IEC (2026) ISO/IEC TS 42112:2026: Information technology — Artificial intelligence — Guidance on machine learning model training efficiency optimization. https://www.iso.org/standard/88246.html

ISO/IEC (2026) ISO/IEC WD TS 25570: Information Technology — Artificial Intelligence — Reliability assessment of AI systems. https://www.iso.org/standard/90757.html

Jaakkola E (2020) Designing conceptual articles: four approaches. AMS Review 10:18–26. https://doi.org/10.1007/s13162-020-00161-0

Jabareen Y (2009) Building a conceptual framework: philosophy, definitions, and procedure. International Journal of Qualitative Methods 8(4):49–62. https://doi.org/10.1177/160940690900800406

Leveson NG (2012) Engineering a safer world: systems thinking applied to safety. MIT Press, Cambridge, MA. ISBN 9780262016629. https://mitpress.mit.edu/9780262016629/engineering-a-safer-world/

Li H., Duan J., Yuan C., Diffenderfer J., Madireddy S., Kailkhura B., et al. (2026) SURE: data-efficient safety guardrailing via internal representations and uncertainty-weighted pseudo-labels. npj Artificial Intelligence. https://doi.org/10.1038/s44387-026-00141-y

Li J., et al. (2023) HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. EMNLP 2023. https://doi.org/10.18653/v1/2023.emnlp-main.397

Lin S., Hilton J., Evans O. (2022) TruthfulQA: Measuring How Models Mimic Human Falsehoods. ACL 2022. https://doi.org/10.18653/v1/2022.acl-long.229

Manakul P., Liusie A., Gales M. J. F. (2023) SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. EMNLP 2023. https://doi.org/10.18653/v1/2023.emnlp-main.557

National Transportation Safety Board (2019) Collision between vehicle controlled by developmental automated driving system and pedestrian, Tempe, Arizona, March 18, 2018. Highway Accident Report NTSB/HAR-19/03. https://www.ntsb.gov/investigations/Pages/HWY18MH010.aspx

NIST (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

Noy NF, McGuinness DL (2001) Ontology Development 101: A guide to creating your first ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880. https://protege.stanford.edu/publications/ontology_development/ontology101.pdf

Obermeyer Z., et al. (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

OECD (n.d.) AI Incidents and Hazards Monitor: overview and methodology. https://oecd.ai/en/incidents-methodology. Living resource; accessed 10 October 2026. Material claims are limited to the documented cutoff corpus.

OpenAI (2026) OpenAI and Hugging Face partner to address security incident during model evaluation. Initial disclosure, 21 July 2026. https://openai.com/index/hugging-face-model-evaluation-security-incident/

OWASP Foundation (2025a) OWASP Top 10 for LLM Applications 2025. https://genai.owasp.org/llm-top-10/

OWASP Foundation (2025b) OWASP Top 10 for Agentic Applications for 2026. Published 9 December 2025. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/

Page MJ, McKenzie JE, Bossuyt PM et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372:n71. https://doi.org/10.1136/bmj.n71

Paterson C., et al. (2025) Safety assurance of machine learning for autonomous systems. Reliability Engineering & System Safety. https://doi.org/10.1016/j.ress.2025.111311

Perdomo J., et al. (2020) Performative Prediction. Proceedings of ICML 2020. https://proceedings.mlr.press/v119/perdomo20a.html

Perrow C (1984) Normal accidents: living with high-risk technologies. Basic Books, New York

Raji I. D., et al. (2020) Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. Proceedings of FAT* 2020. https://doi.org/10.1145/3351095.3372873

Rao A., Keller A., Kalra N., Steed R., Kwegyir-Aggrey K., Klyman K., Staheli D., Bergman A. (2026) Challenges to the Monitoring of Deployed AI Systems. NIST AI 800-4. https://doi.org/10.6028/NIST.AI.800-4

Rethlefsen ML, Kirtley S, Waffenschmidt S et al (2021) PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Systematic Reviews 10:39. https://doi.org/10.1186/s13643-020-01542-z

Riggs J., Hamin M., Perry N., Edelman B., Cihon P. (2026) Summary Analysis of Responses to the Request for Information Regarding Security Considerations for AI Agents. NIST AI 800-5. https://www.nist.gov/publications/summary-analysis-responses-request-information-regarding-security-considerations-ai

Royal Commission into the Robodebt Scheme (2023) Report. Commonwealth of Australia. https://robodebt.royalcommission.gov.au/publications/report

Ruan Y., et al. (2024) Identifying the Risks of LM Agents with an LM-Emulated Sandbox (ToolEmu). ICLR 2024. https://arxiv.org/abs/2309.15817

Salazar Valdebenito N., Ruz G. A., Tabares-Soto R. (2026) Fairness and transparency in ML: a methodological framework. Scientific Reports. https://doi.org/10.1038/s41598-026-62692-z

Schemmer M., et al. (2023) Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. Proceedings of IUI 2023. https://doi.org/10.1145/3581641.3584066

Selbst A. D., et al. (2019) Fairness and Abstraction in Sociotechnical Systems. Proceedings of FAT* 2019. https://doi.org/10.1145/3287560.3287598

Shelby R., et al. (2023) Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. Proceedings of AIES 2023. https://doi.org/10.1145/3600211.3604673

Shumailov I., et al. (2024) AI models collapse when trained on recursively generated data. Nature, 631, 755–759; correction published 2025. https://doi.org/10.1038/s41586-024-07566-y Correction: https://doi.org/10.1038/s41586-025-08905-3

Slattery P., et al. (2026) The AI risk repository: A meta-review, database, and taxonomy of risks from artificial intelligence. Patterns. https://doi.org/10.1016/j.patter.2026.101517

Snyder H (2019) Literature review as a research methodology: an overview and guidelines. Journal of Business Research 104:333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Stahl B. C., et al. (2023) A systematic review of artificial intelligence impact assessments. Artificial Intelligence Review. https://doi.org/10.1007/s10462-023-10420-8

Steimers A., Schneider M. (2022) Sources of Risk of AI Systems. International Journal of Environmental Research and Public Health, 19(6), 3641. https://doi.org/10.3390/ijerph19063641

Steyvers M., Kumar A. (2024) Three Challenges for AI-Assisted Decision-Making. Perspectives on Psychological Science. https://doi.org/10.1177/17456916231181102

Sygnia (2026) Investigation finds AI accelerated attack enabled lone threat actor to rapidly compromise enterprise cloud environment. https://www.sygnia.co/press-release/sygnia-investigation-into-ai-accelerated-attack/

Torraco RJ (2005) Writing integrative literature reviews: guidelines and examples. Human Resource Development Review 4(3):356–367. https://doi.org/10.1177/1534484305278283

U.S. Equal Employment Opportunity Commission (2023) iTutorGroup to pay $365,000 to settle EEOC discriminatory hiring suit. https://www.eeoc.gov/newsroom/itutorgroup-pay-365000-settle-eeoc-discriminatory-hiring-suit

Uschold M, Gruninger M (1996) Ontologies: principles, methods and applications. The Knowledge Engineering Review 11(2):93-136. https://doi.org/10.1017/S0269888900007797

Vaccaro M., et al. (2024) When combinations of humans and AI are useful: a systematic review and meta-analysis. Nature Human Behaviour. https://doi.org/10.1038/s41562-024-02024-1

Vassilev A. (2026) Robust AI Security and Alignment: A Sisyphean Endeavor? IEEE Security & Privacy, 24(3), 52–58. https://doi.org/10.1109/MSEC.2026.3678214

Vassilev A., et al. (2025) Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. NIST AI 100-2e2025. https://doi.org/10.6028/NIST.AI.100-2e2025

Whetten DA (1989) What constitutes a theoretical contribution? Acad Manage Rev 14(4):490-495. https://doi.org/10.5465/amr.1989.4308371

Wong A., et al. (2021) External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Internal Medicine. https://doi.org/10.1001/jamainternmed.2021.2626 Published correction: https://doi.org/10.1001/jamainternmed.2021.3907

World Wide Web Consortium (2009) SKOS Simple Knowledge Organization System Reference. W3C Recommendation, 18 August 2009. https://www.w3.org/TR/skos-reference/

World Wide Web Consortium (2017) Shapes Constraint Language (SHACL). W3C Recommendation, 20 July 2017. https://www.w3.org/TR/shacl/

Zou W., et al. (2025) PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models. USENIX Security 2025. https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OR1MethodsAmendmentv1.2.1.docx
- OR1ScreeningandExclusionAccounting.docx
- OR1ResearchProtocol.docx
- OR3FCAIROCrosswalkv1.2.1.xlsx
- OR5FCAIROIncidentCorpusv1.2.1.xlsx
- OR2FCAIROCodebookv1.2.1.xlsx
- OR1SupplementaryMethods.docx
- OR8FCAIROConceptualRationalev1.2.1.docx