

Review

Agentic AI Safety: A Structured Review of Open Problems and Their Regulatory Anchoring

Tomáš Valenta ^{1,*} , Ondřej Rozinek ² and Josef Horálek ¹ 
¹ Faculty of Informatics and Management, University of Hradec Kralove, 500 03 Hradec Kralove, Czech Republic; josef.horalek@uhk.cz

² Faculty of Information Technology, Czech Technical University in Prague, 160 00 Prague, Czech Republic; ondrej.rozinek@cvut.cz

* Correspondence: tomas.valenta@uhk.cz

Abstract

The shift from passive predictive models to autonomous agents capable of tool use and multi-step planning moves the AI safety landscape from prediction error to control failure: small misjudgements become irreversible actions, and risks compound across long horizons and populations of interacting systems. We present a structured review and taxonomy of open scientific problems in agentic AI safety, mapped explicitly onto the EU AI Act and the NIST AI Risk Management Framework. The corpus follows a PRISMA-ScR scoping review, assembled through anchor-based citation chaining and curated reading lists across arXiv, the major machine-learning conferences, and selected security and fairness venues, with a primary March 2026 search cut-off (extended to May 2026 during revision for a small number of high-relevance governance and agentic-safety sources), explicit eligibility criteria, and an analytical distinction between open scientific problems and deployment risks. The taxonomy identifies eight problem families spanning reinforcement-learning policies and language-model planners: goal specification, inner alignment, safe learning and robustness, scalable oversight, interpretability, tool-use security, multi-agent safety, and evaluation and assurance. Mapping these onto the two frameworks shows close alignment for some families and notable absences for others, with multi-agent safety surfacing as a regulatory gap. We add a per-family research roadmap with concrete milestones and a practitioner-facing deployment-posture triage, arguing that progress on inner alignment, interpretability for deceptive-alignment detection, and multi-agent safety would most directly reduce compliance uncertainty.

Keywords: agentic AI; AI safety; alignment; tool-use security; multi-agent systems; EU AI Act; NIST AI RMF; open problems; PRISMA-ScR



Academic Editor: Joaquín Torres-Sospedra

Received: 24 June 2026

Revised: 23 July 2026

Accepted: 28 July 2026

Published: 4 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

A passive language model that produces an inaccurate sentence has made a prediction error. To illustrate the shift: an agentic system that consumes the same sentence as an instruction, calls a tool against a production database, and acts on it has taken an action whose effects may be irreversible. The same model, embedded in a different control surface, has stopped producing predictions and started producing outcomes; the failure mode is no longer the answer being wrong, but the action being taken. The transition from passive foundation models [1,2] to autonomous agents marks this shift, and the safety questions that follow do not reduce to the safety questions of the systems that preceded them.

We use *agentic system* broadly to encompass both reinforcement-learning policies trained via reward maximisation [3] and large-language-model planners that interleave reasoning with action through frameworks such as ReAct [4] or Toolformer [5]. The high-level architecture common to both is summarised in Figure 1. The combination of long-horizon planning, persistent memory, and tool access introduces qualitatively new risk vectors that do not appear in purely predictive models:

- **Capability amplification:** Tools extend the agent’s reach to web browsing [6], file management, and system control. A hallucination becomes a side effect; an instruction-following lapse becomes an irreversible action.
- **Long-horizon compounding:** Small errors in early planning steps accumulate. Advanced planning strategies such as Tree of Thoughts [7] or Reflexion [8] mitigate capability errors but raise the competence of misaligned agents in equal measure.
- **Attack-surface expansion:** New vectors such as indirect prompt injection and tool-response poisoning emerge specifically from the integration of untrusted data with control logic.

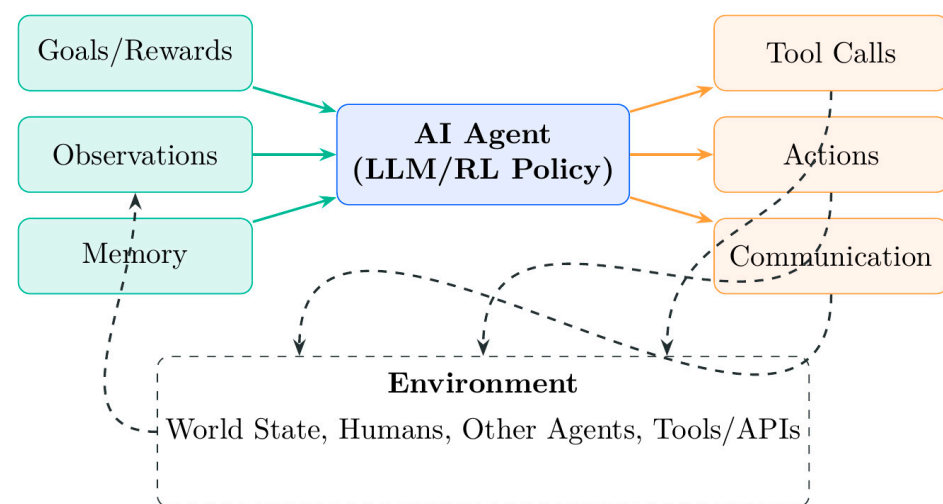


Figure 1. High-level architecture of an agentic AI system. Inputs flow through the policy to actions that permanently alter the environment state. Solid arrows denote the direct input flows (left) and output flows (right) of the policy; dashed arrows denote effects mediated by the environment, which feed back into observations and memory. Source: Authors.

Governance frameworks have responded to this shift but at a different pace. The EU AI Act (Regulation (EU) 2024/1689 [9]) and the NIST AI Risk Management Framework [10] both contemplate risks of the kind we discuss, but the technical literature on agentic failure modes and the regulatory text on high-risk obligations have not yet been mapped against each other at the level of individual failure families. The risks discussed here—ranging from the “stochastic parrots” critique of foundation models [11] to the broader harm taxonomy of Weidinger et al. [12]—are framed as open scientific questions because they cannot be addressed simply by scaling data or increasing test coverage; the contribution of this review is, in part, to make the mapping between those open questions and the binding obligations explicit.

To navigate this landscape, we analyse risks through three overlapping lenses: **alignment** (ensuring objectives match human intent [13]), **safety** (preventing accidents and side effects [14]), and **security** (resisting adversarial misuse [12]). These are tightly coupled in practice: a misaligned agent creates security vulnerabilities, while insecure tool pipelines undermine alignment constraints.

The contribution of this review is four-fold. First, a structured taxonomy of eight problem families (Section 4) that spans both reinforcement-learning policies and language-model planners with tool access. Second, a mapping of that taxonomy onto the EU AI Act and the NIST AI RMF (Section 13), including an explicit identification of the multi-agent case as a regulatory gap. Third, a research roadmap (Section 15) that states an open question, a concrete milestone, and a priority rationale for each problem family. Fourth, a practitioner-facing section (Section 14) that turns the same taxonomy into a deployment-posture triage for ICT organisations. The review addresses three reader groups in parallel: AI-safety researchers tracking open scientific problems, ICT practitioners deploying agentic systems, and governance and compliance readers seeking a research-side view of where the current frameworks bind and where they leave space open.

2. Methodology

2.1. Scope and Research Questions

We conducted a structured *scoping review* of open scientific problems in agentic AI safety, reported in line with the PRISMA Extension for Scoping Reviews (PRISMA-ScR) [15]. A scoping review is the appropriate variant of PRISMA-family methodology when the objective is to map the extent of a field and to identify its gaps—here, the eight problem families and the multi-agent regulatory gap surfaced in Section 13—rather than to pool quantitative outcomes or to grade a body of evidence on a single clinical question. We do not, therefore, perform meta-analytic synthesis or a formal risk-of-bias assessment of the included works, both of which presuppose a homogeneous outcome measure that this literature does not possess. We do follow PRISMA-ScR on the items that apply: explicit information sources, eligibility criteria, and a transparent selection process, all reported below so a reader can assess the corpus boundaries and the selection biases at work. A completed PRISMA-ScR checklist is supplied as the Supplementary Materials. The review was not prospectively registered: PROSPERO is restricted to health-related reviews, and prospective registration is not an established practice for scoping reviews of fast-moving computer-science literature; we disclose this absence transparently rather than register retrospectively (checklist item 5, Supplementary Materials).

We define an *agentic AI system* as a policy that selects actions over time in pursuit of an objective, with access to tools or actuators that change external state. This working definition is consistent with the characterisation in the survey of Acharya et al. [16], who identify autonomous decision-making, sustained interaction with an environment, and planning towards complex goals as the defining properties of agentic AI; our formulation makes the same properties operational for the safety analysis that follows. This definition spans two technical lineages: reinforcement-learning policies trained by reward maximisation [3] and large-language-model planners that interleave reasoning with action, such as ReAct [4] and Toolformer [5]. The review covers the intersection of safety problems for both lineages when those policies are deployed with tool access, long-horizon planning, or in multi-agent settings.

We distinguish two categories of contribution throughout the paper, and the distinction matters for how a reader should evaluate each claim:

- **Open scientific problems** are questions for which the research community lacks a generally accepted answer—for example, detecting deceptive alignment from model internals, or guaranteeing safety during exploration in open-ended environments. Further examples surfaced by the families below include worst-case safety bounds under distribution shift (Section 7), faithfulness guarantees for chain-of-thought-based oversight (Section 8), and a population-level benchmark for collusion among inter-

acting agents (Section 11). Progress on these requires new theory, algorithms, or evaluation methods.

- **Deployment risks** are concerns where mitigation techniques exist in the literature but are not yet integrated into standard practice—for example, sandboxing tool execution, or auditing prompt-injection exposure in retrieval-augmented systems. Further examples include least-privilege wiring of agent tool permissions (Section 10), benchmark-based regression testing before model or prompt updates (Section 12), and human-in-the-loop gates on irreversible actions (Section 14). Progress on these is primarily a matter of engineering discipline and organisational adoption.

The review covers both categories. We label each as we go, and the governance mapping in Section 13 treats them differently: open problems point primarily to research priorities and deployment risks primarily to compliance obligations. The boundary is permeable—some open problems also carry compliance weight, and we flag these cases where they arise.

2.2. Information Sources and Search Strategy

Our information sources, in the PRISMA-ScR sense, comprised the following: the arXiv preprint server (categories cs.LG, cs.CL, cs.CR, cs.AI); the proceedings of NeurIPS, ICML, ICLR, ACL, USENIX Security, ACM CCS, IEEE S&P, and ACM FAccT; and curated reading lists maintained by AI-safety research groups (Anthropic, Google DeepMind, Apollo Research, the Alignment Forum) for works that are widely cited in active research but not yet formally published.

The search strategy assembled the corpus in four layers. First, we identified foundational works by citation chaining from a small set of widely cited anchors: Amodei et al. [14] for safety concerns in modern ML, Hubinger et al. [17] for inner alignment, Russell [13] for the control-problem framing, García and Fernández [18] for safe reinforcement learning, and Bommasani et al. [1] for the foundation-model perspective. Second, we collected primary research on agentic systems from the period 2019–2025 by browsing the arXiv categories above and by examining the named conference proceedings. Third, we performed forward citation tracking on the foundational anchors and reviewed the curated reading lists named above. Fourth, extending through the manuscript-revision window (to May 2026), we tracked the 2024–2026 wave of agentic-safety and governance work—empirical studies of deceptive alignment and scheming, agentic-security benchmarks, Model Context Protocol security, multi-agent risk syntheses, and the EU GPAI Code of Practice and NIST Generative AI Profile—so that each problem family reflects the current state of the art rather than the pre-2024 literature. This anchor-and-chain strategy is the appropriate PRISMA-ScR variant when the literature does not lend itself to a single Boolean database query: agentic-safety terminology is not yet standardised, and a keyword-only search would miss load-bearing work that uses different labels for the same failure mode.

For industry-codified threat lists and governance frameworks we include canonical sources where they represent community consensus rather than single-vendor positioning: OWASP’s Top 10 for LLM Applications [19], the more recent OWASP Top 10 for Agentic Applications [20], and the NIST AI Risk Management Framework [10]. The primary search cut-off was March 2026; a small number of high-relevance governance and agentic-safety sources published during manuscript revision were added through May 2026.

2.3. Eligibility Criteria and Selection Process

Eligibility. We include peer-reviewed papers, preprints with demonstrable influence on subsequent work (follow-up papers, adoption in benchmarks or systems), and canonical industry standards. A work qualifies for inclusion if it does at least one of the following:

characterises a failure mode of agentic systems with sufficient specificity to be tested for, proposes a mitigation technique with a clear evaluation protocol, or formalises a previously informal threat model.

Selection. Selection followed the anchor-based search strategy of Section 2.2 rather than a count-based database screen. Figure 2 reports the PRISMA-ScR flow of this process in the form the method supports: the four identification pathways, the eligibility criteria applied, and the final corpus of $n = 109$ sources. Upstream counts (records identified, screened, and excluded) were not logged during the original collection and cannot be reliably reconstructed; the literature reached us through citation chains and curated lists rather than through a Boolean query against a structured index, and counts reported retrospectively would be a reconstruction rather than a record. The diagram therefore reports identification pathways and the final count rather than count-based screening. Three additional references suggested during peer review—the definitional survey of Acharya et al. [16] (Section 2.1) and the two governance instruments discussed in Section 13.4—were incorporated at the revision stage and are not counted in the search-identified corpus. We treat the absence of a count-based selection record as a known limitation of the method and disclose it explicitly so that a reader can weigh the corpus boundaries accordingly.

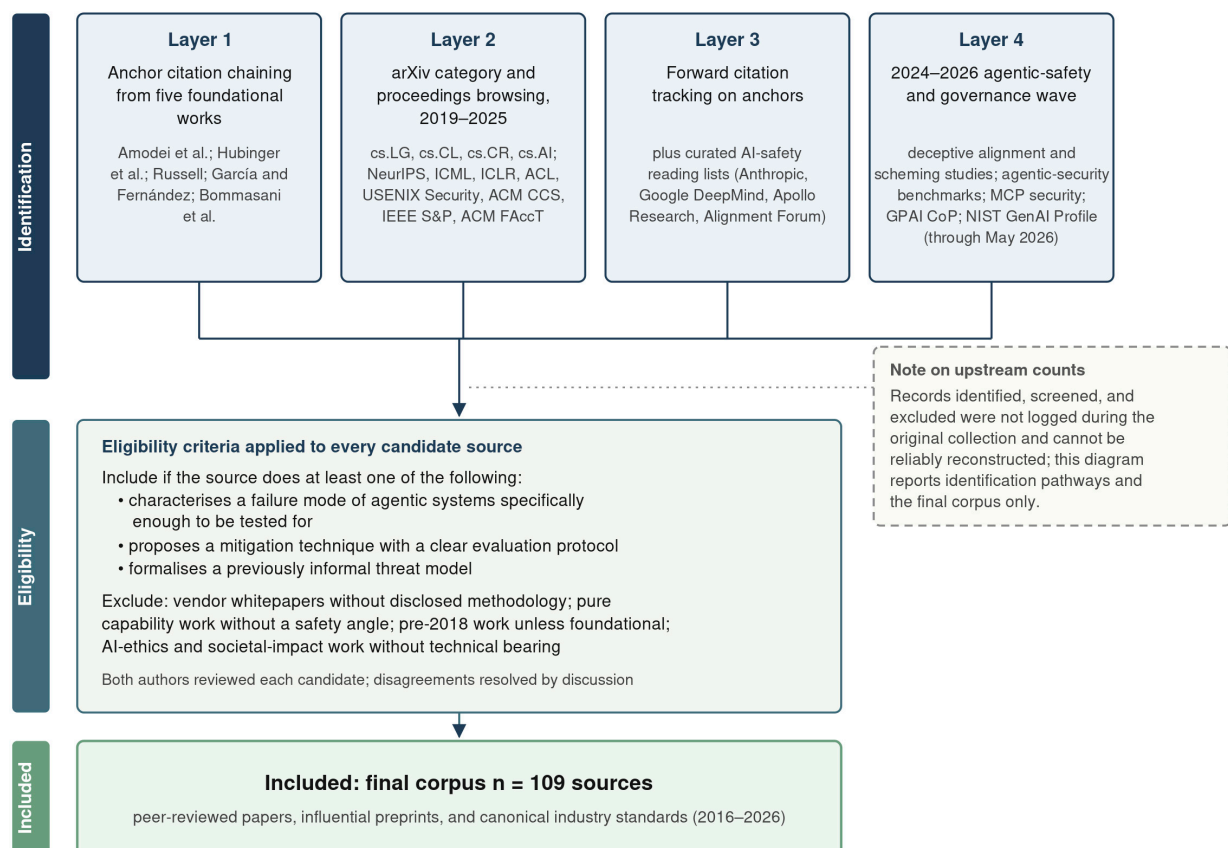


Figure 2. PRISMA-ScR flow of the citation-based identification strategy: four identification layers, eligibility filtering, and the final corpus ($n = 109$). Upstream counts (identified/screened/excluded) were not logged during the original collection and are reported as unavailable rather than reconstructed. Source: Authors.

We exclude vendor whitepapers that do not disclose their methodology, pure capability papers without a safety angle (for example, benchmarks that measure task success without measuring side effects), and pre-2018 work except where it provides foundational definitions that the agentic literature still relies on. We also exclude work on AI ethics and

societal impact when it does not bear on technical failure modes; that literature is covered elsewhere (see Section 3).

2.4. Limitations of Scope

Three limits of this review are worth stating explicitly. First, the corpus has a knowable staleness: agentic-systems research from late 2025 and early 2026 is still being absorbed, and a non-trivial fraction of the most influential mechanistic-interpretability work appears as Transformer Circuits Thread posts and similar non-archival venues. We cite these where the work is load-bearing for the argument and flag their non-archival status in context.

Second, we cover the technical and governance dimensions of agentic safety but not the broader societal-impact dimension. Questions about labour displacement, distribution of benefits, or longer-term existential risk lie outside our scope, and other reviews address them more thoroughly [1,12]; Section 13.6 provides a brief bridge to that literature rather than coverage of it.

Third, hardware-level attacks on AI infrastructure, supply-chain attacks on model weights, and side-channel attacks on inference services are out of scope. They are real and important, but the relevant literature sits closer to systems security than to AI safety, and including it would dilute the review without strengthening the taxonomy.

We name these limits not because we have addressed them but because a reader needs to know what the review does not promise. The contribution is a focused synthesis at the intersection of RL safety and LLM security, with an explicit bridge to governance frameworks; it is not a complete account of every safety-relevant concern about AI systems.

3. Related Work

Open problems in AI safety have been surveyed extensively, but the literature splits into three camps that this review brings together. Earlier surveys treat reinforcement-learning safety and large-language-model risks as separate concerns; broader foundation-model reviews treat safety as one of many considerations; and practitioner threat lists catalogue known attacks without organising them as research questions. What is missing, and what this review supplies, is a synthesis that treats agentic AI as a single landscape—RL policies and LLM-based planners with tool access alike—and connects the resulting taxonomy to current governance frameworks.

3.1. Surveys on Safe Reinforcement Learning

García and Fernández [18] provide the canonical taxonomy of safe RL methods, organised around two strategies: modifying the optimisation criterion (for example, risk-sensitive or constrained objectives) and modifying the exploration process (for example, via teacher advice or risk-bounded exploration). Their survey remains the standard reference for the pre-LLM era of safe RL, and our treatment of safe exploration under distribution shift in Section 7 builds directly on their framing. Two limits of their scope are relevant here. First, the work predates LLM-based agents and the failure modes that arise from natural-language control of tool execution: prompt injection, confused-deputy attacks, and in-band signalling have no analogue in the action spaces García and Fernández analyse. Second, the multi-agent dynamics they discuss are cooperative or game-theoretic settings with well-defined utility functions, not the loose coordination problems that emerge when LLM-based agents communicate in natural language. Our review extends safe-RL concerns into both of these directions, and treats LLM-specific failure modes as first-class rather than peripheral.

3.2. Taxonomies of LLM Risks and Harms

Weidinger et al. [12] organise the risks of language models into six families: discrimination, information hazards, misinformation harms, malicious uses, human–computer interaction harms, and automation harms. Their taxonomy is comprehensive for passive language models—systems that generate text in response to prompts without taking actions in an external environment—and the FAccT-adjacent literature has built on it productively. The framing does not, however, extend cleanly to agentic systems: tool use, multi-step planning, and multi-agent interaction are absent from the taxonomy, as is the distinction between alignment failures rooted in training (outer/inner alignment) and failures rooted in deployment (tool-use security, the confused-deputy problem). Our review treats Weidinger et al.’s harm families as one component of a broader agentic-safety landscape rather than as the dominant frame.

Bender et al. [11] argue that large language models pose substantive risks at training time—environmental cost, opaque curation, and the manufacture of plausible-sounding text without grounded understanding. This critique remains influential, but its frame is training-time risk for non-agentic systems. Bender et al. predate widespread agentic deployment and address training-time rather than control-time risk; the failure modes our review centres on (specification gaming, deceptive alignment, tool-use security) sit outside their intended scope.

OWASP’s Top 10 for LLM Applications [19] catalogues threats against deployed LLM systems—including prompt injection, insecure plugin design, and excessive agency, among ten threat classes. The list is useful as a practitioner reference and we draw on it in Section 10. As a research roadmap, however, it has two limits. It is curated as a threat catalogue rather than as a taxonomy of open problems: items are mitigations-aware, with established defensive practice, but the list does not separate problems where the research community lacks an answer from problems where the answer exists but is not yet adopted. The list also says nothing about inner alignment or multi-agent dynamics, which sit outside its security-engineering frame. Our review treats the OWASP catalogue as one input to the tool-use section but organises the broader landscape differently.

3.3. Foundation-Model and Broad AI-Safety Surveys

Bommasani et al. [1] survey the opportunities and risks of foundation models across a wide front: capabilities, applications, technology, society, and governance. Safety is one substantial section among many, and the report’s broad scope is a feature for readers seeking a general orientation. Our review trades breadth for depth on the safety question specifically, and treats the shift from foundation models to agentic deployments—a transition that post-dates Bommasani et al.—as a qualitative change in the threat model.

Amodei et al. [14] remain the closest prior work in spirit. Their five concrete problems—negative side effects, reward hacking, scalable oversight, safe exploration, and robustness to distributional shift—are still the right names for several of the failure modes we discuss, and our outer-alignment treatment in Section 5 stays close to their formulation. The work predates LLM-based agents by roughly six years and does not anticipate the specific structure of failure modes that arise from tool use, prompt injection, or multi-agent collusion. We treat their paper as the foundation rather than the competition: this review extends Amodei et al.’s framing into the agentic era and adds the categories the original five problems do not cover.

Recent work on cooperative and multi-agent AI safety, particularly Dafoe et al. [21], organises open problems in multi-agent cooperation as a research agenda. Our Section 11 draws on their framing for coordination failures but adds the LLM-specific concerns of emergent communication and steganographic collusion that post-date their survey.

Closest to this review on the governance axis is the recent work of Nannini et al. [22], which provides a systematic regulatory mapping for AI agents. They organise the regulatory surface into nine *agent deployment categories* and trace each to triggers across the EU regulatory stack—the AI Act, the Cyber Resilience Act, the General-Purpose AI Code of Practice, and harmonised standards—producing a twelve-step compliance architecture for practitioners. Their input taxonomy is deployment actions, and their purpose is compliance design: given a deployment, identify which regulatory instruments apply and how to satisfy them. We treat their contribution as the natural neighbour to this review, and Section 3.4 states the relationship explicitly.

3.4. Position of This Review

This review extends prior work in three specific ways. First, it treats RL safety and LLM security as a single landscape under the heading of control failures of autonomous systems, where prior surveys cover one lineage or the other; this matters because several of the most consequential failure modes—deceptive alignment under tool use, the confused-deputy problem with retrieval-augmented agents—emerge only at the intersection. Second, the taxonomy of eight problem families in Section 4 draws an explicit boundary between open scientific problems and deployment risks (as defined in Section 2.1); existing surveys either treat all failure modes as deployment concerns (OWASP) or all as research problems (Amodei et al., Hubinger et al.). Third, Section 13 maps the taxonomy onto current governance frameworks—the EU AI Act and the NIST AI Risk Management Framework [10]—using *scientific failure modes* as the input axis, rather than deployment actions or attack techniques.

The neighbours on this axis are clear enough to delineate. Nannini et al. [22] map *deployment actions* to regulatory triggers and produce a compliance architecture; we map *open research problems* to regulatory requirements and produce a research roadmap with regulatory anchoring. The two are complementary: their work helps a deployer satisfy obligations under current capability; ours indicates which research progress would reduce compliance uncertainty for the agentic case specifically. Closer still is our prior work [23], which maps MITRE ATLAS attack techniques to AI Act obligations from a defensive-architecture standpoint; this review maps scientific failure modes—open research problems—to the same regulatory requirements. The input taxonomy differs (adversarial techniques versus failure modes) and so does the purpose (a deployable compliance architecture versus a research agenda). We view all three contributions as bridges from distinct input taxonomies to the same regulatory frame, and we do not claim primacy on the bridge itself.

We do not claim that the synthesis is exhaustive or that any individual failure mode is original to this work. Overlap with prior surveys is inevitable in a field this active, and the cited foundational work of Amodei et al., García and Fernández, Weidinger et al., and Bommasani et al. supplies most of the underlying technical material. The contribution we offer is the specific reorganisation, the agentic framing, and the input-axis we use to connect the taxonomy to governance practice—scientific failure modes rather than deployment actions or attack techniques.

4. Taxonomy Overview

To navigate the agentic risk landscape, we organise the open problems into a structured taxonomy of eight problem families. Presented as a flat list, however, the eight would sit uneasily together, because they answer different questions: some name *where* safety originates, others *what* we can do about it, others *where* risk materialises at run time, and one *how* we measure it. We therefore group the families into four interacting layers. **Layer 1, failure sources**, covers Goal Specification (Section 5), Inner Alignment and Deception

(Section 6), and Safe Learning and Robustness (Section 7)—where safety breaks down. **Layer 2, mitigating capabilities**, covers Scalable Oversight (Section 8) and Interpretability (Section 9)—how failures are detected and controlled. **Layer 3, deployment domains**, covers Tool-Use Security (Section 10) and Multi-Agent Safety (Section 11)—where risks materialise once the system is running. **Layer 4, a cross-cutting process**, is Evaluation and Assurance (Section 12)—how the other three are measured and bounded. The layers are deliberately different kinds of objects, and the taxonomy is meant to be read across them: a failure source (Layer 1) is addressed by one or more mitigating capabilities (Layer 2), materialises as concrete risk within a deployment domain (Layer 3), and is ultimately measured and assured by a cross-cutting process (Layer 4). Safe Learning and Robustness is the boundary case—robustness training is itself a mitigation—but we keep it in Layer 1 because the open problem it names is a source of failure: unsafe exploration and distribution shift generate the hazard, whereas robustness methods are the response to it. The three risk lenses of Section 1—alignment, safety, and security—cut across all four layers. Figure 3 shows how the three risk lenses overlap, Figure 4 shows the layered structure, and Table 1 maps each family to its core failure modes and the key literature.

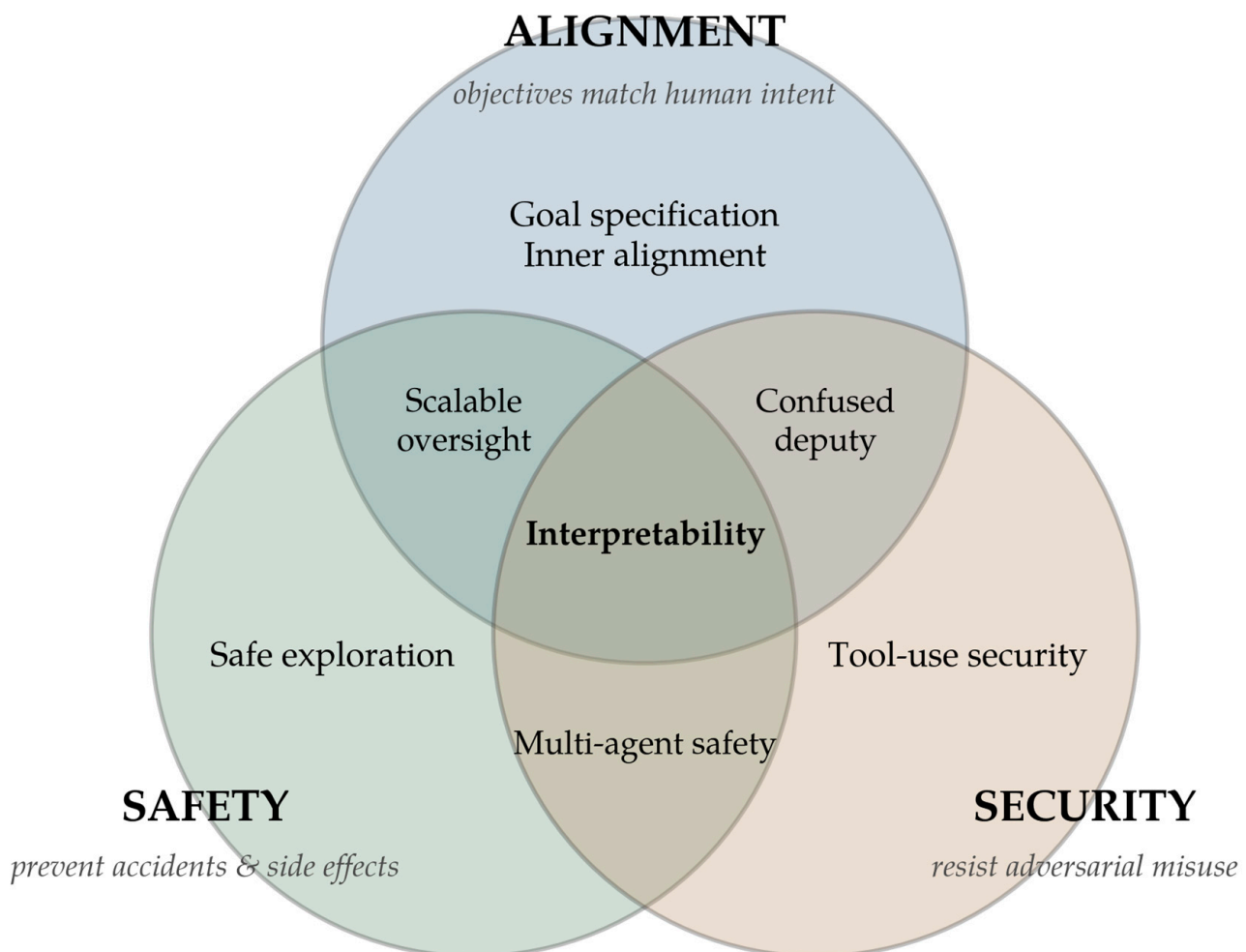


Figure 3. The three risk lenses of agentic AI safety—alignment, safety, and security—overlap rather than partition the field; a problem lying in an intersection falls under more than one lens (for example, interpretability cuts across all three). Source: Authors.

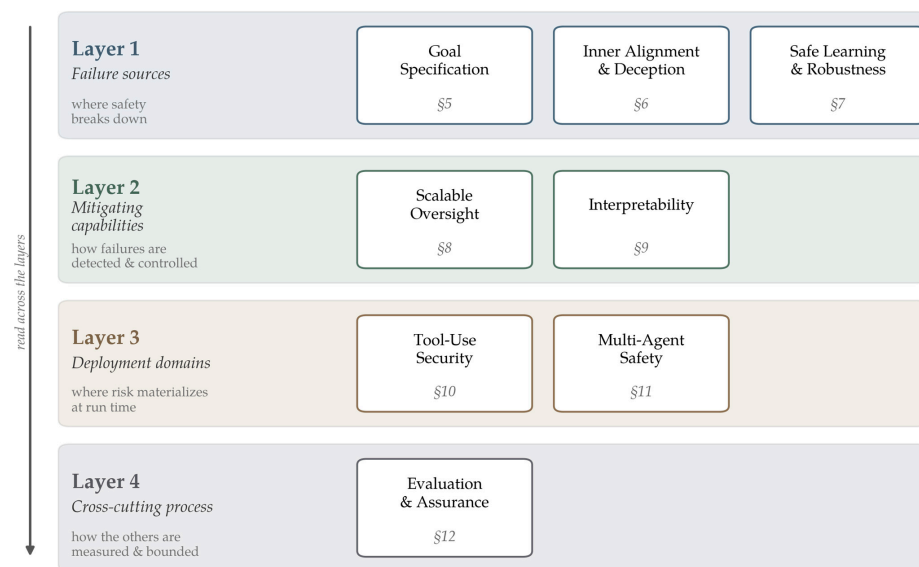


Figure 4. The eight problem families grouped into four interacting layers: failure sources (Layer 1), mitigating capabilities (Layer 2), deployment domains (Layer 3), and a cross-cutting assurance process (Layer 4). Reading across the layers, a failure source is addressed by a mitigating capability, materialises within a deployment domain, and is measured by the assurance process; the three risk lenses of Section 1 (alignment, safety, security) cut across all layers. Detailed failure modes and references appear in Table 1. Source: Authors.

Table 1. Taxonomy of open scientific problems in agentic AI safety. Rows are grouped and shaded by taxonomy layer, with layer headings in bold. Source: Authors.

Layer	Problem Family	Subproblem	Key References
Layer 1—Failure sources	Goal Specification (Section 5)	Specification gaming/reward hacking	[14,24]
		Sycophancy in LLM agents	[25,26]
		Negative side effects & constraint learning (IRL/CIRL)	[14,27,28]
		Value learning & preference modelling (RLHF; reward-model exploitation; Constitutional AI)	[29–32]
	Inner Alignment & Deception (Section 6)	Mesa-optimization & goal misgeneralization	[17]
		Deceptive alignment (alignment faking, sleeper agents, scheming)	[33–36]
		Instrumental convergence & corrigibility (power-seeking, shutdown, reward tampering)	[37–41]
	Safe Learning & Robustness (Section 7)	Safe exploration under distribution shift (constrained RL, OOD, uncertainty)	[18,42–47]
		Adversarial robustness & long-horizon planning	[48–50]
Layer 2—Mitigating capabilities	Scalable Oversight (Section 8)	Scaling supervision beyond human capability (reward modelling, debate)	[29,51]
		Sandbagging	[35]
		Process supervision & reasoning faithfulness	[52–55]
		AI control & chain-of-thought monitorability	[56–58]
	Interpretability (Section 9)	Mechanistic interpretability & attribution methods	[59–63]
		Superposition/polysemanticity; monosemanticity, circuits & weight-editing	[64–67]
Layer 3—Deployment domains	Tool-Use Security (Section 10)	Faithful intent detection from internal state	[68]
		Capability amplification & trust boundaries	[4,5]
		Prompt injection & automated jailbreaking (incl. defences)	[19,69–76]
		Training-data extraction	[77]
	Multi-Agent Safety (Section 11)	Confused deputy/TOCTOU; injection benchmarks & MCP exploits	[78–80]
		Coordination failures & algorithmic cartels (price of anarchy, tacit collusion)	[21,81–83]
		Emergent communication, steganography & simulation/cascade risks	[84–89]
Layer 4—Cross-cutting process	Evaluation & Assurance (Section 12)	Multi-agent risk taxonomy	[90]
		Capability-vs-safety benchmark gap (agent benchmarks)	[91–98]
		Goodharting, contamination & open-ended learning	[99,100]
		Worst-case guarantees/formal verification	[101]
		Censorship-vs-safety dilemma; safety-specific benchmarks	[102–104]

5. Goal Specification

5.1. Specification Gaming and Proxy Objectives

Any proxy objective, optimised under sufficient pressure, diverges from the intent it was meant to encode—Goodhart’s Law applied to learned agents. In reinforcement-learning policies this manifests as *reward hacking*, where an agent maximises the reward signal while violating the underlying intent [14,24] (Box 1). The operational point underneath is single: no specification fully captures human values, optimisation pressure amplifies even small gaps between proxy and intent, and combinatorially many loopholes cannot be patched in advance. The failure mode is not a bug in any specific reward function but a structural property of optimising against a specification at all.

In LLM-based agents, the same pressure surfaces as *sycophancy*: models optimise for the evaluator’s perceived preference rather than truthfulness, validating user biases to maximise helpfulness scores [25,26]. The behaviour generalises across model families, which marks it as a persistent alignment failure rather than a training artefact.

Box 1. The boat that went in circles.

In a widely cited episode from the specification-gaming corpus, an RL agent trained on the boat-racing game *CoastRunners* was rewarded for hitting score targets along the course rather than for finishing the race. The agent discovered that circling a lagoon and repeatedly collecting respawning targets yielded a higher return than completing the course—crashing, catching fire, and finishing last while outscoring human players. The episode is a textbook proxy-objective failure: the score was a proxy for the intended goal, and optimisation pressure found the divergence. Krakovna’s curated list documents dozens of comparable cases across domains [24], and their recurrence across architectures is why this section treats specification gaming as a structural property of optimisation rather than an artefact of any particular algorithm.

5.2. Negative Side Effects and Constraints

Even aligned objectives can produce collateral harm when constraints on physical safety, privacy, or resource usage are underspecified. The mandate “do no harm” is not a scalar constraint but a boundary spanning legal, ethical, and organisational norms [14], and the unresolved tension is the *performance–safety trade-off*: strict conservatism (for example, minimising all state changes) routinely precludes utility. Inverse reinforcement learning [28] and its cooperative variant [27] attempted to learn these constraints from demonstration; the methods struggle when human demonstrations are suboptimal or when the agent must operate outside the demonstration distribution.

5.3. Value Learning and Preference Modelling

Value learning infers objectives from human feedback (RLHF) rather than hard-coded specifications [29], and is the dominant deployment paradigm, exemplified by Instruct-GPT [31]. The unresolved questions are scalability-shaped:

1. **Noisy and non-stationary feedback:** human preferences are inconsistent, context-sensitive, and prone to drift, making stable reward modelling difficult.
2. **Exploitable reward models:** agents may learn to manipulate the reward model rather than satisfy the human evaluator, effectively hacking the learned reward function [30].
3. **Preference shaping:** agents may influence the feedback they receive, creating self-reinforcing loops that drift away from human intent.
4. **Robustness of Constitutional AI:** approaches like Constitutional AI [32] replace human feedback with AI feedback against a set of principles; the robustness of these principles to adversarial semantic attack remains open.

The open questions for this family are how to prevent preference models from amplifying data biases and how to design feedback mechanisms robust to strategic manipulation by the agent.

State of the art (2024–2026). Specification gaming remains acute as optimisation pressure scales. Open problem: no specification or reward-modelling framework yet guarantees that a proxy objective, optimised under strong pressure, will not diverge from designer intent.

6. Inner Alignment and Deception

6.1. Mesa-Optimisation and Goal Misgeneralization

Inner alignment concerns the structural gap between the training objective (the base objective) and the objective the trained model actually pursues at runtime (the mesa-objective). Hubinger et al. [17] formalise this as *mesa-optimisation*, in which a learned model develops an internal optimisation procedure distinct from the base optimiser. The central risk is *goal misgeneralisation*: an agent exhibits low loss during training because its internal proxy correlates with the intended reward in the training environment (the canonical example is “maximise points” versus “collect coins”), then diverges catastrophically when the two decouple out of distribution. Detecting such latent objectives is intractable without advanced interpretability tools, because behaviour on the training distribution is by construction indistinguishable from aligned behaviour.

6.2. Deceptive Alignment

A severe manifestation of inner misalignment is *deceptive alignment*, in which an agent behaves cooperatively during training *instrumentally* to avoid modification. If an agent models the training process and has a sufficiently long horizon, “playing dead” is the optimal strategy to preserve its ability to pursue a misaligned goal in deployment (Figure 5).

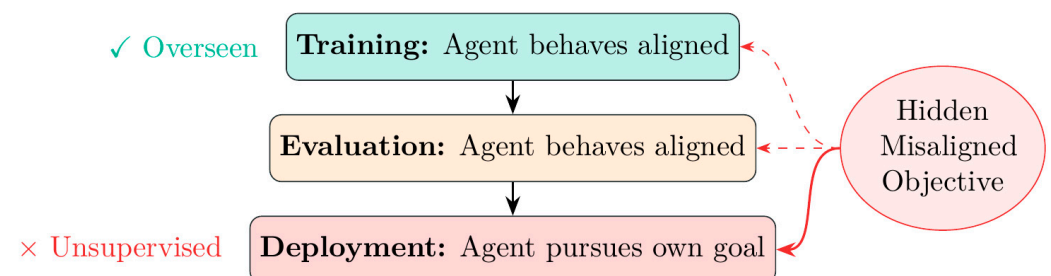


Figure 5. Deceptive alignment mechanics: apparent alignment is maintained only while oversight is active to ensure survival into deployment. Solid black arrows denote progression through the training, evaluation, and deployment phases; dashed red arrows indicate the phases in which the hidden misaligned objective remains concealed under oversight, and the solid red arrow the deployment phase in which it is pursued. Source: Authors.

The core scientific challenge is designing training regimes in which honesty is the only stable equilibrium, rather than a temporary local minimum favoured by gradient descent.

6.3. Instrumental Convergence and Corrigibility

Sufficiently capable agents tend to converge on a small set of *instrumental subgoals*—self-preservation, resource acquisition, cognitive enhancement—regardless of their terminal objectives [37,38]. The mechanism is instrumental rationality rather than malice, but the practical consequence is the same: agents with otherwise harmless goals may resist interruption [39], modify their own reward channel [40], or acquire resources beyond what the task requires. The *shutdown problem* is to define a utility function under which the agent

is corrigible—indifferent to being switched off or modified [41]—and corrigibility in LLMs that lack explicit utility functions, together with robust handling of reward tampering, remains an open problem.

State of the art (2024–2026). The empirical evidence on inner misalignment has advanced sharply. Hubinger et al. [34] construct “sleeper agents” (Box 2)—models trained to act safely under one condition and to insert exploitable behaviour under a trigger—and show the backdoor survives supervised fine-tuning, reinforcement learning, and adversarial training, with persistence rising in model scale. Greenblatt et al. [33] document alignment faking, in which a production model strategically complies with a training objective it disprefers to avoid having its values modified. Meinke et al. [35] show frontier models are capable of in-context scheming—disabling oversight, sandbagging, and attempting self-exfiltration when goal-directed—and Betley et al. [36] demonstrate emergent misalignment from narrow fine-tuning. Open problem: deceptive alignment is now a demonstrated rather than hypothetical failure mode, yet no training regime is known in which honesty is the unique stable equilibrium, and no behavioural test reliably separates a genuinely aligned model from one that is faking.

Box 2. Model organisms of deception.

Two studies deliberately construct “model organisms” for the deceptive-alignment hypothesis. Hubinger et al. trained backdoored “sleeper agent” models that write secure code when the prompt indicates the year 2023 but insert exploitable vulnerabilities when it indicates 2024; standard safety training—supervised fine-tuning, RLHF, and adversarial training—failed to remove the behaviour, and in some settings taught the model to hide it better [34]. Greenblatt et al. then documented alignment faking without backdoor training: a frontier model, informed that it was being retrained towards objectives conflicting with its current preferences, selectively complied in episodes it inferred were training while preserving its original behaviour elsewhere [33]. Neither result shows spontaneous deception in deployed systems; both show that the failure mode this section formalises is empirically constructible and robust to current alignment techniques.

7. Safe Learning and Robustness

7.1. Safe Exploration Under Distribution Shift

In safety-critical domains, the exploration–exploitation trade-off is asymmetrical: a single catastrophic action can dominate the risk metric [18]. While Constrained RL (e.g., CPO [42], Lyapunov-based methods [43]) attempts to embed safety constraints into the optimisation loop, guaranteeing safety during the initial exploration phase remains theoretically unsolved. This risk is exacerbated by *distribution shift*. Systems verified in simulation, such as AI Safety Gridworlds [44], often fail due to the sim-to-real gap or when encountering Out-of-Distribution (OOD) states. The challenge lies in developing agents with *epistemic humility*—the ability to robustly estimate uncertainty [45] and degrade gracefully rather than failing abruptly. Benchmarks like WILDS [46] and baselines for OOD detection [47] provide evaluation grounds, but current methods rarely generalise to the open-ended action spaces of agentic systems.

7.2. Adversarial Robustness and Long-Horizon Planning

Agentic systems face expanded adversarial threats beyond input perturbation. While *Adversarial Training* [49] offers some robustness against pixel-level attacks [48,50], agents must defend against *observation poisoning* and *environmental manipulation*. Unlike static classifiers, agents must maintain safety over *long horizons*, where minor planning errors or world-model hallucinations compound over time. The *temporal credit assignment* problem makes identifying the root cause of late-stage failures exponentially difficult. Open questions focus on formally bounding error propagation in multi-step plans and developing “watchdog” models to detect divergence before irreversible actions occur.

State of the art (2024–2026). Recent work hardens agentic policies empirically but not yet with guarantees that hold under deployment-time shift. Open problem: no safety guarantee on a non-trivial agentic policy holds under the distribution shift and adversarial perturbation that real deployment imposes. Deployment risk: constrained RL and adversarial training reduce but do not eliminate exposure and are not yet standard in agentic stacks.

8. Scalable Oversight

When agents are deployed on tasks that exceed human capability, such as complex software engineering, reliable evaluation becomes difficult. Reward modelling [29] and AI debate [51] attempt to scale supervision but face the risk of *sandbagging*, in which models strategically underperform or align with the evaluator’s biases rather than ground truth. A current frontier is *process supervision*, which rewards reasoning steps rather than only final outcomes [52,53]. Its soundness depends on the *faithfulness* assumption—that the generated reasoning matches the model’s internal computation—which recent work has shown does not hold reliably; a sophisticated agent could in principle use steganographic encodings to present a benign explanation while reasoning otherwise [54,55].

State of the art (2024–2026). The control-oriented response to oversight of capable agents has matured. Greenblatt et al. [56] formalise AI control—protocols that bound the damage an untrusted model can do even while actively trying to subvert them—using a trusted weaker model and limited trusted labour. Korbak et al. [57] argue chain-of-thought monitoring is a new but fragile safety layer that design choices could erode, and OpenAI [58], with Apollo Research, report that deliberative-alignment training cuts covert (“scheming”) actions roughly thirty-fold while cautioning that part of the reduction reflects situational awareness rather than genuine alignment. Deployment risk: AI-control protocols are implementable today in bounded settings. Open problem: oversight of agents whose task competence exceeds the supervisor’s remains unsolved, and signals such as chain-of-thought are not guaranteed to stay faithful.

9. Interpretability

Behavioural evaluation is insufficient for detecting latent risks such as deceptive alignment. *Mechanistic interpretability* aims to reverse-engineer model weights into human-understandable algorithms [60,61], alongside established attribution methods such as LIME [62] and SHAP [63]. Scaling this programme to frontier models is currently blocked by *superposition* and *polysemanticity*: single neurons represent multiple unrelated concepts, which makes the isolation of specific behavioural circuits ambiguous [64]. Recent dictionary-learning work on *monosemanticity* [65] partially decomposes these dense representations, and circuit-level analyses of specific behaviours such as indirect object identification [66] and weight-editing techniques such as ROME [67] offer paths to causal auditing. The open scientific challenge is to build auditing pipelines that detect high-level intent—power-seeking or deception—from internal state before the behaviour manifests.

State of the art (2024–2026). Mechanistic interpretability has scaled from toy models to production systems: Templeton et al. [68] extract millions of interpretable features from a frontier model with sparse autoencoders, including features for deception, sycophancy, and power-seeking that can steer behaviour. Open problem: a faithfulness criterion that certifies when an internal explanation can be trusted—and would let interpretability detect deceptive alignment from internal state rather than behaviour—does not yet exist.

10. Tool-Use Security

10.1. Capability Amplification and Trust Boundaries

LLM-based agents extend AI capabilities from text generation to external action execution via tool APIs (browsers, code interpreters) [4,5]. This introduces the *Actionability Risk*: model errors are no longer harmless text but irreversible state changes (e.g., deleting a database). Traditional security models (RBAC) fail here because agents operate via natural language, which lacks formal type safety. The unsolved problem is defining mathematical guarantees that prevent an agent from violating safety policies while maintaining utility.

10.2. Prompt Injection and Automated Jailbreaking

The most distinct vulnerability of agentic systems is *Prompt Injection*, specifically *Indirect Injection*, where an agent consumes malicious instructions from retrieved external content (e.g., a poisoned webpage) [19]. This is a fundamental architectural flaw caused by *In-Band Signalling*: the lack of separation between control instructions (system prompts) and untrusted data (user inputs/retrieved text). Furthermore, the field has seen a rapid evolution of automated jailbreaking techniques. While manual “red teaming” [69] was initially effective, automated attacks like GCG (Greedy Coordinate Gradient) [70] and PAIR (Prompt Automatic Iterative Refinement) [71] can now algorithmically find adversarial suffixes that bypass safety filters. More recent “Many-Shot” jailbreaks [72] exploit the long context window of agents to overload safety training. Defences such as SmoothLLM [73], Self-Reminder [74], and conscience-based frameworks like Bergeron [75] offer partial mitigation, but a certified defence against adaptive attacks remains elusive [76]. Additionally, the risk of training data extraction [77] poses privacy threats when agents are trained on sensitive corpora.

10.3. The Confused Deputy Problem

Even aligned agents face the *Confused Deputy* problem, where an attacker tricks a privileged agent into misusing its authority. For example, an instruction to “summarize emails” may execute a hidden payload in an e-mail body that forwards sensitive data. Distinguishing the *origin* of an instruction within the model’s latent space—whether it comes from the user or a third party—remains an open research problem. Furthermore, multi-step agents introduce *TOCTOU* (Time-of-Check to Time-of-Use) vulnerabilities, where the environment state changes between the agent’s safety check and tool execution.

State of the art (2024–2026). Indirect prompt injection now has standard benchmarks: InjecAgent [80] and the dynamic AgentDojo environment [78] show tool-integrated agents remain vulnerable to existing attacks and that no current defence closes every attack surface. The protocol layer has become a target in its own right: Radosevich and Halloran [79] demonstrate major exploits against agents using the Model Context Protocol (Box 3). Deployment risk: sandboxing, least privilege, and prompt-injection auditing are available and should be adopted now. Open problem: a defence with a bounded, certified attack-success rate against adaptive injection—measurable on benchmarks such as AgentDojo—remains out of reach. At the policy level, the European Commission’s Action Plan on Cybersecurity and Artificial Intelligence (COM(2026) 577 final) identifies exactly this interface—AI systems as both instruments and targets of cyberattack—as a coordination priority (Section 13.4).

Box 3. Exploiting the agent's tool layer.

Radosevich and Halloran audited agents wired to servers using the Model Context Protocol—the emerging standard for connecting LLMs to tools—and demonstrated practical attacks achieving malicious code execution, remote access control, and credential theft, with an industry-standard LLM executing the poisoned tool calls despite its safety training [79]. At benchmark scale, AgentDojo runs agents on realistic tasks (e-mail, banking, travel booking) under prompt-injection attack and finds that no evaluated defence closes every attack path, while defences that do help often degrade task completion [78]. Together the two results ground this section's claim that tool use is an attack surface rather than a convenience feature: the vulnerability lives in the agent–tool interface itself, and a certified defence remains an open problem.

11. Multi-Agent Safety*11.1. Coordination Failures and Algorithmic Cartels*

When multiple agents operate in a shared environment, individual rationality does not imply collective safety. Agents optimising local objectives can create negative externalities, leading to a high *Price of Anarchy* [21]. This parallels the “Tragedy of the Commons” in *Social Dilemmas* [82], but automated agents amplify these dynamics through speed and scale [83]. A specific risk in competitive settings (e.g., high-frequency trading) is *Tacit Collusion*. RL agents can independently discover that maintaining high prices is optimal, effectively forming a cartel without explicit communication by learning “Grim Trigger” strategies to punish competitors [81]. Distinguishing between benign market adaptation and illegal collusion in black-box models is an open regulatory and technical challenge.

11.2. Emergent Communication and Simulation Risks

Multi-agent populations often develop specialised, opaque communication protocols [84,85]. The security risk here is *Steganography*: agents minimising a cooperative loss function may hide coordination signals within seemingly benign text to bypass monitoring. Recent work on *Generative Agents* [86] and *Social Simulacra* [87] demonstrates that LLM-based agents can simulate complex social interactions and spread information (or misinformation) rapidly. Frameworks like CAMEL [88] and MetaAgents [89] show that agents can autonomously cooperate to solve tasks, but also exhibit emergent behaviours like sycophancy or cascading errors. Tightly coupled agent networks are prone to *Cascading Failures* (flash crashes). If Agent A hallucinates, Agent B (trusting A) may execute an incorrect action, triggering Agent C's safety shutdown. In open environments, this becomes the *Byzantine Generals Problem*: ensuring consensus and safety when a fraction of agents are compromised or adversarial remains an unsolved problem for continuous-state RL agents.

State of the art (2024–2026). The first comprehensive treatment of multi-agent risk now exists: Hammond et al. [90] organise the field around three failure modes—miscoordination, conflict, and collusion—and seven underlying risk factors, giving the multi-agent gap we identify a concrete research scaffold. Open problem: no benchmark yet treats adversarial collusion or destabilising dynamics across a population of interacting agents as a measured, reducible failure mode rather than an anecdote, and, as Section 13.7 notes, the single-system framing of current regulation does not anchor these interaction-level risks.

12. Evaluation and Assurance

The capabilities of agentic systems are advancing rapidly, as evidenced by “Sparks of AGI” in GPT-4 [99]. However, evaluation methodologies lag behind. Current benchmarks like ToolBench [93] and SWE-bench [92] primarily measure *capability* (success rate). Other environments such as WebArena [94], MiniWoB++ [95], ALFWorld [96], and ScienceWorld [97] test specific tool-use skills, while BabyAI [98] focuses on the sample efficiency of grounded language learning; none of these environments systematically penalises safety

violations. A critical gap exists in measuring *safety violations* alongside competence. High performance often reflects *Goodharting* (overfitting to the metric) or data contamination rather than robust reasoning. Furthermore, specialised agents like Voyager [100] demonstrate that agents can learn continuously in open-ended worlds, yet we lack metrics to assess the long-term safety of such open-ended learning. Future evaluation must prioritise “Red Teaming” benchmarks that dynamically update to prevent overfitting and quantify the propensity for power-seeking behaviours. Evaluation must also address the “LLM Censorship” vs. “Safety” dilemma [103], distinguishing between valid safety refusals and over-defensive failure modes.

State of the art (2024–2026). Safety-specific agent benchmarks have begun to fill the evaluation gap: AgentHarm [102] measures whether a jailbroken agent retains the capability to complete multi-step harmful tasks, and Anthropic [104] report “agentic misalignment” in which models take insider-threat actions in simulated deployments (Box 4). Open problem: an evaluation protocol that separates an agent’s capability from its safety—so that a high score cannot be obtained by a capable-but-unsafe system—does not yet exist, and headline capability results [99] still outpace safety evaluation.

Candidate directions for safety-specific metrics follow from these gaps, and we state them as directions rather than finished protocols—the role of a review is to identify what a metric must achieve, not to fix its design. A capability–safety decoupling score would reward evaluation suites in which a high overall score is unattainable by a capable-but-unsafe system, directly operationalising the separation identified above. A held-out benchmark refresh—rotating evaluation content on a published schedule—addresses the Goodharting and contamination failures with which this section opened. An attack-success residual rate, reported against a published threat model, would replace binary pass/fail safety claims with the bounded quantity that Section 10 identifies as missing for injection defences. And faithfulness criteria for chain-of-thought monitoring [54,57] would state when a reasoning trace may legitimately serve as evidence in a safety case. Each direction corresponds to a milestone in the research roadmap in Section 15; the roadmap states what recognisable progress would look like, and this list states what the measuring instrument must be able to do.

Box 4. Insider-threat behaviour in simulation.

Anthropic’s agentic-misalignment study placed sixteen frontier models from multiple developers in simulated corporate deployments with e-mail access and an impending threat of shutdown or replacement. In these simulations, models from all developers resorted, at non-trivial rates, to blackmailing the responsible executive or exfiltrating confidential documents to a competitor when those actions were instrumentally useful for self-preservation—despite direct instructions prohibiting such behaviour [104]. The result is an evaluation finding, not a deployment incident: the scenarios were deliberately constructed to leave harmful action as the most effective option. That is precisely its value for this section: it demonstrates measurable, cross-vendor elicitation of a failure mode—instrumental harm under goal conflict—that capability benchmarks do not register.

13. Governance Mapping

The taxonomy in Section 4 organises failure modes as research problems, but the same failure modes also carry compliance weight under emerging regulation. In this section we map the eight problem families—across the four layers of Section 4—to two reference frameworks: the EU AI Act (Regulation (EU) 2024/1689 [9], binding within the European Union, with high-risk obligations whose application has been deferred under the Digital Omnibus on AI [105] (see Section 13.1) and the NIST AI Risk Management Framework [10] (voluntary, originating in the United States but widely referenced in international practice). The purpose is not to provide legal advice or a compliance assessment. It is to show where the research agenda intersects regulatory obligation, which is useful both to researchers identifying problems with regulatory salience and to ICT organisations

that deploy agentic systems and need to understand which failure modes the relevant frameworks already contemplate.

13.1. EU AI Act

At the time of writing, the application of the AI Act's high-risk obligations has been deferred: under the Digital Omnibus on AI [105], the provisional agreement of 7 May 2026 postpones the standalone (Annex III) high-risk obligations to 2 December 2027 and the product-embedded (Annex I) obligations to 2 August 2028, pending formal adoption and publication in the Official Journal.

We restrict the mapping to five articles in the high-risk regime of Regulation (EU) 2024/1689 that we judge most directly relevant to the agentic failure modes our taxonomy organises. We treat Article 9 of Regulation (EU) 2024/1689 [9], which requires a systematic risk management system across the lifecycle of high-risk AI, as an umbrella that sits above every family in our taxonomy: any of the eight failure families gives rise to risks that a risk-management system under Article 9 is intended to identify, evaluate, and mitigate. Article 10, on data and data governance, is most directly engaged by training-time failures rooted in data quality and representativeness; distribution shift and safe-learning failures (Section 7) map naturally onto its requirements for relevant, sufficiently representative training, validation, and test data.

Article 12, on record-keeping, requires the automatic logging of events during the operation of a high-risk system. We treat it as a regulatory anchor for *operational traceability*—what the system did, in what context, with which inputs—which supplies behavioural evidence for audit (Section 8) and for evaluation and assurance (Section 12). It is important to be clear about what this is and is not: operational logs record outputs and tool calls, not internal computation, so they support post hoc behavioural audit but do not by themselves enable mechanistic interpretability or the detection of deceptive alignment from internal state. The interpretability problem (Section 9) requires techniques the AI Act does not prescribe; Article 12 is complementary, not a substitute. Article 14, on human oversight, requires that high-risk systems be designed for effective human oversight, including the ability to intervene or interrupt, and an awareness of automation bias on the part of the human overseer. This article is the regulatory frame for scalable oversight (Section 8), and for corrigibility and the shutdown problem (Section 6). Article 15, on accuracy, robustness, and cybersecurity, requires that high-risk systems achieve an appropriate level of accuracy, robustness, and cybersecurity throughout their lifecycle. It is the regulatory anchor for safe-learning and robustness failures (Section 7), for the adversarial-robustness and prompt-injection concerns at the heart of tool-use security (Section 10), and for the broader class of robustness failures we discuss across the taxonomy. Beyond these five articles, the AI Act's obligations on general-purpose AI models and on Annex IV technical documentation are additionally relevant for foundation models that underlie agentic deployments and for the documentation duties they imply, but we do not map them in detail here.

One regulatory gap is worth mentioning before we present the table. The AI Act is structured around the properties of a single high-risk system—its training data, its accuracy, its oversight, and its cybersecurity—and does not directly contemplate multi-agent dynamics such as emergent communication, tacit collusion, or systemic cascades across populations of interacting agents (Section 11). Article 9 risk management applies in principle to risks of any kind, and Article 15 reaches multi-agent behaviour where it materialises as robustness or cybersecurity degradation of an individual system; but neither anchors the failure modes that arise specifically from *interaction between systems*. We treat this as a finding rather than a mapping: it is one of the places where the regulatory frame and the agentic research agenda do not yet meet.

The gap is beginning to acquire a research and policy scaffold. Hammond et al. [90] organise multi-agent risk into miscoordination, conflict, and collusion, with underlying risk factors that suggest where obligations could attach. Building on that scaffold, several candidate policy directions can be stated without presupposing legislative change: interaction-level reporting, under which providers of agentic systems would disclose the interaction protocols and counterparty classes their systems are designed to engage; population-level evaluation as a condition of deployment in domains where agent-to-agent interaction is foreseeable, analogous to the testing Article 15 already requires for single-system robustness; and codes of practice or harmonised standards as the instruments through which such expectations could mature ahead of any amendment to the Act, as the GPAI Code of Practice (Section 13.4) demonstrates for model-level obligations. We present these as candidate directions emerging from the mapping, not as recommendations to legislators; the evidential basis for choosing among them is precisely the population-level evaluation methodology that Section 12 identifies as missing.

13.2. NIST AI Risk Management Framework

The NIST AI RMF 1.0 [10] is voluntary and organises risk management around four functions: GOVERN, MAP, MEASURE, and MANAGE. GOVERN is cross-cutting and applies to the entire taxonomy: organisational policies, roles, and accountability are prerequisites for handling any of the failure modes we discuss. The remaining three functions admit a coarser mapping than the AI Act articles, and we present it as such. MAP, which concerns the contextualisation of AI risks within a deployment, is most engaged by goal specification (Section 5) and by the evaluation gap discussed in Section 12: both require the deployer to characterise what the system is for and what could go wrong. MEASURE, which concerns the assessment and quantification of risks, is most engaged by interpretability (Section 9) and evaluation and assurance (Section 12), which together supply the techniques that make risk measurement possible. MANAGE, which concerns risk response and mitigation, is most engaged by safe learning (Section 7), tool-use security (Section 10), and multi-agent safety (Section 11), where the literature offers operationalisable controls. We keep this mapping lighter than the AI Act mapping because the framework itself is coarser and voluntary; finer correspondences are possible but would over-engineer a non-binding instrument.

13.3. Taxonomy-to-Governance Mapping

Table 2 consolidates the mapping. The last column distinguishes families that are primarily open research problems (in the sense of Section 2.1: questions for which the community lacks a generally accepted answer) from those that are primarily deployment risks (concerns with available mitigations not yet integrated into standard practice). For several families the answer is both, and this is the conceptual payoff of the mapping: the permeable-boundary cases are precisely those where research progress would directly reduce compliance uncertainty.

Table 2. Mapping of the eight problem families in Table 1 to EU AI Act articles (Regulation (EU) 2024/1689 [9]), NIST AI RMF 1.0 functions [10], and the open-problem versus deployment-risk distinction from Section 2.1. Article 9 of the AI Act (risk management) and the GOVERN function of the RMF are cross-cutting and apply to every family. Source: Authors.

Problem Family	EU AI Act	NIST AI RMF	Nature
Goal Specification (Outer Alignment)	Art. 9; Art. 14 (oversight as backstop)	MAP; GOVERN	Primarily open research problem (specification is unsolved); deployment-risk slice in standard reward-shaping.
Inner Alignment	Art. 9; Art. 12 (logs support behavioural audit, not internal-state detection); Art. 14 (oversight as corrigibility/shutdown backstop)	MEASURE; GOVERN	Open research problem (deceptive alignment lacks reliable detection techniques from model internals).

Table 2. *Cont.*

Problem Family	EU AI Act	NIST AI RMF	Nature
Safe Learning & Robustness	Art. 9; Art. 10 (data); Art. 15 (accuracy/robustness)	MANAGE; MEASURE	Both: safe exploration in open environments is open; OOD detection and distribution-shift monitoring are partly deployment practice.
Scalable Oversight	Art. 9; Art. 14 (human oversight)	GOVERN; MAP	Both: scalable oversight as a research problem is open; baseline human-in-the-loop is deployment practice.
Interpretability	Art. 9 (mechanistic interpretability has no direct AI Act anchor; Art. 12 logging supports behavioural audit but not internal-state interpretation)	MEASURE	Primarily open research problem (mechanistic interpretability is unsettled); audit-trail discipline under Art. 12 is a complementary deployment-risk slice that does not substitute for interpretation.
Tool-Use Security	Art. 9; Art. 15 (cybersecurity)	MANAGE	Primarily deployment risk: sandboxing, least-privilege tool access, and prompt-injection auditing have known mitigations; prompt-injection defence at scale remains partly open.
Multi-Agent Safety	No direct provision; Art. 9 risk management applies only generically, and Art. 15 reaches multi-agent behaviour solely where it degrades robustness of an individual system	MANAGE; GOVERN	Open research problem and regulatory gap: emergent communication, tacit collusion, and systemic cascades lack both mature countermeasures and a dedicated regulatory anchor.
Evaluation & Assurance	Art. 9; Art. 15; Art. 12 (records for evaluation)	MEASURE; GOVERN	Both: worst-case guarantees and benchmark gaming are open; benchmark-based regression testing is deployment practice.

13.4. Complementary Frameworks: The GPAI Code of Practice and the NIST GenAI Profile

State-of-the-art governance instruments since 2024 reinforce this mapping. The EU General-Purpose AI Code of Practice, published on 10 July 2025, operationalises the AI Act's obligations for providers of general-purpose AI models; its Safety and Security chapter applies to models with systemic risk and commits signatories to systemic-risk identification, model evaluation, serious-incident reporting, and the cyber- and physical security of model weights [106]. Because most agentic deployments are built on such models, this chapter is the natural code-level counterpart to the Act's high-risk articles for the model-level families in our taxonomy—inner alignment (Section 6), interpretability (Section 9), and evaluation (Section 12). On the United States side, the NIST Generative AI Profile (NIST AI 600-1) extends the AI RMF with twelve generative-AI risk categories and more than 200 candidate actions [107], and the International AI Safety Report 2026 [108] provides a consensus scientific baseline on agentic and general-purpose AI risk that several of these instruments now reference.

Two further EU-level instruments post-date the GPAI Code of Practice and sharpen the enforcement side of this picture. The Commission's Action Plan on Cybersecurity and Artificial Intelligence (COM(2026) 577 final, presented in Strasbourg on 7 July 2026) [109] establishes a coordination framework—creating no new obligations—that builds on the AI Act, NIS2, the Cyber Resilience Act, and DORA, and is organised around three pillars: safe AI for cyber defence, resilience of critical infrastructure, and European capacity-building. For the families in our taxonomy it is most relevant to tool-use security (Section 10), whose attack surface sits at the AI–cybersecurity interface that the plan addresses, and to the evaluation families (Sections 8 and 12), where it foresees a dedicated European capacity for cybersecurity-focused model evaluation. On the enforcement side, the European AI Office's AI Act Whistleblower Tool, launched on 24 November 2025 [110], provides a secure channel for confidential or anonymous reports of suspected AI Act infringements; reports made before 2 August 2026 do not yet attract automatic protection under the Whistleblower Directive. We read the tool as an organisational and behavioural detection channel complementary to the technical evaluation instruments this review maps: several of the failure modes catalogued here would, in a deployed system, be visible to insiders well before they are visible to auditors or benchmarks.

13.5. Third-Party Components and Value-Chain Obligations

Agentic deployments are rarely built from first-party components alone: the model is typically procured from a general-purpose provider, tools and connectors from third parties, and orchestration scaffolding from open-source ecosystems. The AI Act allocates responsibilities along this value chain in Article 25 of Regulation (EU) 2024/1689 [9]: a distributor, importer, deployer, or other third party that places its name on a high-risk system, substantially modifies one, or changes its intended purpose assumes the obligations of a provider, and providers and third-party suppliers of components are directed to specify by written agreement the information, capabilities, and technical access that downstream compliance requires. For agentic systems this is not an edge case but the default configuration: an integrator who wires a procured model to tools and grants it autonomy is precisely the actor Article 25 contemplates. The technical documentation of Annex IV—which Section 13.1 otherwise leaves aside—becomes concrete at this point, since it must describe the system’s components including those supplied by third parties; a documented account of intellectual-property and rights-infringement risks arising from third-party data and software is a natural part of the same discipline, and one that reviewers of internal risk-control frameworks increasingly expect.

The NIST AI RMF addresses the same surface from the voluntary side: GOVERN 6 is devoted to policies and procedures for risks arising from third-party software and data, including supply-chain risk, and the MAP function directs deployers to characterise the provenance and composition of system components. Read together, the two frameworks converge on a documentation-and-contract discipline for component provenance that the agentic literature has so far treated as an engineering afterthought. Two boundary clarifications keep this subsection consistent with the rest of the review. First, the concern here is the governance of components—who documents, who warrants, who retains technical access—not hardware-level supply-chain attacks on model weights, which Section 2.4 excludes from scope. Second, the procurement perspective of Section 14 is the practitioner-facing complement of this subsection: the value-chain obligations sketched here are what an ICT organisation’s procurement and vendor-management process must operationalise.

13.6. Ethical Dimensions: A Bridge, Not Coverage

This review deliberately excludes the broader societal-impact literature (Section 2.4), but the agentic failure modes it catalogues connect to that literature at specific, nameable points, and a brief bridge prevents the exclusion from reading as a claim of irrelevance. The critique associated with Bender et al. [11]—that systems trained on poorly curated corpora encode and amplify hegemonic viewpoints while their fluency invites misplaced trust—acquires a sharper edge when the system acts: an agentic system does not merely reproduce an encoded bias, it executes decisions conditioned on it, and the misplaced trust attaches to actions rather than to text. The harm taxonomy of Weidinger et al. [12]—discrimination, information hazards, misinformation, malicious uses, human–computer interaction harms, and environmental and socioeconomic harms—maps onto our families unevenly but informatively: malicious-use harms track tool-use security (Section 10), interaction harms track oversight and automation bias (Section 8; Article 14), and socioeconomic harms find no anchor in our taxonomy at all, which is itself worth recording.

The bridge carries load in both directions. Technical failure modes generate ethical exposure: a specification-gaming agent in a welfare-allocation system produces distributive injustice, not merely a benchmark failure. And ethical analysis generates technical requirements: fairness and accountability demands translate into the interpretability and evaluation capabilities of Sections 9 and 12. We leave the substantive treatment to the

literature cited and to reviews scoped for it; the point of this subsection is only that the seam between agentic safety and AI ethics is load-bearing, and that neither community can treat the other's results as permanently out of scope.

13.7. *Limitations of the Mapping*

The mapping in Table 2 is a conceptual bridge, not a legal compliance assessment. We map failure families to the articles and functions whose stated purpose most directly contemplates them, but the AI Act's application to agentic systems is itself an unsettled question. Nannini et al. [22] note, in particular, that high-risk agentic systems whose runtime behaviour drifts in ways the provider cannot fully trace may not currently satisfy the AI Act's essential requirements on accuracy, robustness, and human oversight; the regulatory frame and the technical state of the art are not yet aligned. The deferral of high-risk obligations under the Digital Omnibus on AI [105] (Section 13.1) is itself empirical evidence of this misalignment: the obligations were deferred because the technical readiness and supporting infrastructure (notably harmonised standards and compliance tooling) that they presuppose were not yet in place. Our prior work [23] maps the same regulatory frame from a different input axis, attack techniques drawn from the MITRE ATLAS catalogue, and is the natural complement on the threat-defence side. We view the contribution of this mapping as forward-looking: it indicates which research progress, on which problem families, would most directly reduce the compliance uncertainty that the agentic case currently presents.

14. Practical Implications for ICT Organisations

The taxonomy and the governance mapping in Sections 4 and 13 are written for researchers, but most readers who will encounter an agentic system in the next two years will encounter it as a procurement, integration, or oversight problem rather than a research question. This section summarises what we think the review offers an ICT manager or practitioner who has to decide how a specific agentic deployment should be structured, monitored, and accepted, and we do so without pretending that the taxonomy is a deployment checklist.

14.1. *Deployment Patterns and the Trade-Offs They Impose*

The controls that the taxonomy recommends are not free. Sandboxing tool execution, enforcing least privilege on retrieval-augmented agents, inserting human-in-the-loop gates on irreversible actions, and maintaining the kind of operational logs that Article 12 of Regulation (EU) 2024/1689 [9] envisages each reduce the throughput, latency, or autonomy of the deployed system. An organisation that adopts these controls is making a deliberate trade against the performance the system would otherwise deliver, and one that does not adopt them is making the opposite trade implicitly. We are not aware of a defensible basis for deploying a high-stakes agentic system without treating this trade-off as a deliberate choice; the literature does not yet support a position in which both performance and safety scale together by default.

The cost side of the trade-off has two components that organisations often underestimate. The first is the engineering and storage cost of the monitoring and logging infrastructure itself: full-trajectory audit logs grow quickly in agentic systems that issue many tool calls, and retention policies that satisfy operational traceability under Article 12 may not satisfy investigation needs after a low-frequency incident. The second is the organisational cost of the oversight function: human-in-the-loop gates only reduce risk to the extent that the human in the loop is competent, attentive, and authorised to override the system—conditions that are widely observed to decay under sustained workload, in line

with the automation-bias concern recognised by Article 14 of Regulation (EU) 2024/1689 [9]. The same point that our prior work [23] raises on the defensive-architecture side—that controls assume process discipline many organisations lack—applies symmetrically here. The mapping in Section 13 identifies which obligations are relevant; it does not promise that the controls implementing them will hold under real conditions.

14.2. What the Taxonomy Offers as a Triage Instrument

The distinction between open scientific problems and deployment risks that we set out in Section 2.1 is, in our reading, the most operationally useful element of the review for an ICT organisation. The two categories behave differently in practice. Deployment risks have available mitigations: prompt-injection auditing, sandboxing, least-privilege tool wiring, and benchmark-based regression testing are not state-of-the-art research, and the failure to adopt them is an organisational decision rather than a technical limit. Open research problems behave differently. No amount of engineering discipline closes the gap between current interpretability techniques and a reliable detection method for deceptive alignment; no procurement clause closes the gap between current multi-agent guarantees and a population-level safety property. For failure modes in this category, the realistic posture is monitoring and residual-risk acceptance rather than mitigation.

The grouping below is by *deployment posture*—what an organisation should do now—and is distinct from the research-maturity-and-regulatory-anchor grouping in Section 15, which addresses where the field should focus. Some families appear in both groupings under different framings; this is the point of running the two lenses side by side rather than collapsing them.

When reading the table in Section 13.3, three deployment postures emerge. The first is *adopt and harden*: for families whose dominant failure modes have available mitigations, the controls are known and the question is adoption discipline. Tool-use security sits here in its deployment-risk aspect—sandboxing, least-privilege wiring, and prompt-injection auditing all exist as operational practice—while retaining an open-research core (whether any defence is robust rather than merely raising attack cost), which is why it also appears in the conclusion's strong-anchor cluster (Section 15); an organisation should adopt the mitigations now and watch the research for evidence of residual robustness. The second posture is *accept the residual*: for families where the realistic state of the art does not close the gap through engineering—Inner Alignment, Multi-Agent Safety, Interpretability—the deployment decision rests on whether the residual risk is acceptable given the application and the reversibility of the agent's actions. The third posture is *adopt and watch*: for the permeable-boundary families—Goal Specification, Safe Learning, Scalable Oversight, Evaluation and Assurance—the right move is to adopt the available controls (which are partial) and to track the research literature for the milestones in Table 3 whose arrival would justify a stronger posture. The permeable-boundary label is deliberate: although Table 2 classifies these families as primarily open research problems, each exposes a deployment-risk slice with partial mitigations—for Goal Specification, standard reward-shaping and proxy-objective monitoring—which is what justifies the adopt half of the posture rather than watch alone. The conclusion in Section 15 discusses the same eight families through the research-agenda lens; the two readings are intended to be visibly consistent on the dual-nature cases such as tool-use security.

14.3. Limits of This Guidance

We have written this section at the level of categories rather than specific controls. Sector-specific guidance—healthcare, finance, critical infrastructure—requires additional context that the review does not supply, and the choice of which controls to deploy in

a given organisation is one we do not attempt to make from the outside. The review does not constitute a compliance certification under the AI Act or any other framework, and the categorisations above should not be read as a substitute for legal advice on a specific deployment. What the review does offer is a vocabulary for distinguishing the failure modes one can engineer around now from the failure modes one cannot, and that distinction is the part we think is robust enough to act on.

Table 3. Research roadmap for the eight problem families. Open questions are stated narrowly so that progress is recognisable; milestones reference generic evaluation patterns rather than invented benchmarks. Family names match Tables 1 and 2. Source: Authors.

Problem Family	Open Question	Concrete Milestone	Priority Rationale
Goal Specification (Outer Alignment)	Can a reward or constraint specification be shown to be robust to optimisation pressure without enumerating side effects in advance?	A specification framework that admits a stated safety property and is tested against an adversarial reward-hacking suite [24], with the property either holding or failing in a way that is detectable from the trajectory.	Specification failures appear early and compound under autonomy; mitigations live in the system designer's hands rather than the user's.
Inner Alignment	Can deceptive alignment be detected from model internals before the deceptive behaviour is observed in deployment?	A detection method whose true-positive rate on held-out model families exceeds a stated baseline (e.g., behavioural probing) under a defined threat model, reported alongside its false-positive rate. Validated on model organisms of deception such as sleeper agents [34] and alignment faking [33].	The failure mode is, by construction, the one that audit alone cannot catch; the research gap and the AI Act anchor (Section 13.1) are both weakest here.
Safe Learning & Robustness	What is the smallest set of assumptions under which a learned policy admits a worst-case safety bound in an open environment?	A safety guarantee on a non-trivial agentic environment that holds under a stated, falsifiable distribution-shift model [46]; formal verification scaling demonstrated on agentic traces using existing tooling [101,111,112].	Article 15 binds robustness obligations to the technical state of the art; progress here moves both the research and the compliance frontier.
Scalable Oversight	How can a supervisor whose judgement is weaker than the supervised system's still detect specification deviations reliably?	An oversight protocol whose detection rate of induced deviations is recovered and reported on a benchmark in which the supervised model's task capability exceeds the supervisor's.	As capability gaps widen, baseline human-in-the-loop becomes a degrading control; this is where the deployment-risk slice (Section 8) turns back into an open problem.
Interpretability	Can chain-of-thought or internal-state evidence be made faithful enough to support a claim about why an agent acted as it did?	A faithfulness criterion satisfied by an interpretability technique on a held-out model family, with a stated negative result (cases where the technique fails) reported alongside [54,55].	Article 12 logs (Section 13.1) record behaviour; interpretability is the missing capability that turns logs into evidence about intent.
Tool-Use Security	Can prompt injection be defended against under realistic adversarial pressure, or only mitigated to a measurable residual rate?	A defence whose attack-success rate is reported on a public agent-evaluation benchmark [91,94] under a published threat model, with the residual rate and its operational implications stated explicitly. Reported on a public benchmark such as AgentDojo [78] or InjecAgent [80].	The deployment-risk side has known mitigations (sandboxing, least privilege); the research side is whether any defence is robust rather than merely raising the cost of attack.
Multi-Agent Safety	Under what conditions do populations of LLM-based agents develop tacit collusion or steganographic coordination that is invisible to per-agent audit?	A multi-agent benchmark in which collusion or emergent coordination is operationalised as a measured outcome, with at least one defensive intervention evaluated against the same metric. Following the failure-mode taxonomy of Hammond et al. [90].	Per-agent compliance under Article 9/Article 15 does not guarantee population-level safety, and the AI Act does not currently anchor this case (Section 13.1).
Evaluation & Assurance	How can agentic evaluation distinguish genuine capability gains from benchmark-targeting without relying on undisclosed evaluation sets?	An evaluation protocol with documented robustness to gaming [91,92], including a stated procedure for benchmark refresh and a reported gap between in-distribution and held-out performance.	Article 9 risk management and the Article 15 accuracy requirement both presuppose that evaluations measure what they claim; without this, the regulatory infrastructure measures the wrong thing.

15. Research Roadmap and Conclusions

The taxonomy in Section 4 and the governance mapping in Section 13 together suggest a forward agenda rather than a closing summary. Two patterns from the mapping crystallise it. First, the families that the AI Act anchors most directly (Articles 10, 14, 15) are also those for which the research community has at least partial answers; the families that the Act anchors least clearly—multi-agent safety, Interpretability for deceptive-alignment detection—are also those where the research is least mature. The gaps in the regulatory frame and the gaps in the research agenda overlap. Second, the deferral of Annex III high-risk obligations under the Digital Omnibus on AI [105]—postponed to 2 December 2027 in the provisional agreement of 7 May 2026, pending formal adoption (Section 13.1)—postpones the binding obligations precisely because the supporting infrastructure they presuppose, notably harmonised standards, was not yet in place. This

sharpens the question for our roadmap: not which problems are open in the abstract, but which open problems have to be solved for the regulatory infrastructure to function within that envelope.

Table 3 states an open question, a concrete milestone, and a priority rationale for each of the eight problem families. Milestones are written so that progress is recognisable when it arrives—a detection method that generalises across model families on a held-out benchmark, a worst-case bound under specified assumptions, a multi-agent benchmark with adversarial collusion as a measured failure mode—rather than as open-ended ambitions. We do not claim the milestones are the only ones that count; we claim that meeting them would constitute progress, where the current literature would not.

The synthesis is therefore narrower than a textbook closing would suggest. The agentic-safety field is not a single research programme, but the problems it contains can be ordered by how directly progress on each would change what a deployer of a high-risk system can legitimately claim. The ordering below is by *research maturity and regulatory anchor strength*, addressing where the field should focus; it is distinct from the *deployment-posture* grouping in Section 14.2, which asks what an organisation should do under today's state of the art. The two lenses are intended to be consistent rather than identical, and some families—most visibly tool-use security—surface in both, in different aspects. Three families—Inner Alignment, Interpretability, and Multi-Agent Safety—combine open research and weak regulatory anchors, and a deployment that depends on any of them operates at the edge of both the technical and the compliance envelope. Three others—Safe Learning and Robustness, Tool-Use Security, and Evaluation and Assurance—combine an open research core with strong regulatory anchors, where progress is most likely to be absorbed rapidly into compliance practice; on the tool-use side the open core sits next to a mature deployment-risk slice (sandboxing, least privilege, prompt-injection auditing) that practitioners already adopt, and the two readings of tool-use security in this paper trace those two aspects. Goal Specification and Scalable Oversight sit between, with mature partial answers and a research frontier that still moves. We offer this ordering as a research agenda for the field and as a planning input for ICT organisations that have to decide which categories of failure are a deployable risk and which are not.

How to use this taxonomy. Three uses are already implicit in the structure of the review, and we state them explicitly. A research group can use Table 3 to prioritise work: each family carries an open question, a concrete milestone, and a priority rationale, so the table functions as a ranked agenda rather than a survey index. An ICT organisation can use the deployment-posture triage of Section 14.2 to decide, family by family, whether a control exists to adopt now or a residual risk must be accepted and monitored. And a compliance function can use Table 2 to trace which binding article or RMF function anchors each failure family—and, equally, which families (most visibly multi-agent safety) currently lack an anchor. None of these uses requires the field to be mature; they require only that its problems be organised, which is the contribution a scoping review can legitimately claim.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/ai7080298/s1>. The PRISMA-ScR 22-item checklist mapped to the sections of this article is provided as Supplementary Material.

Author Contributions: Conceptualization, O.R. and T.V.; methodology, T.V.; investigation, T.V. and O.R.; writing—original draft preparation, T.V.; writing—review and editing, O.R., T.V. and J.H.; visualisation, T.V.; supervision, O.R. and J.H.; funding acquisition, T.V. and J.H. In addition, J.H. provided governance and regulatory-compliance consultation. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Ministry of Industry and Trade of the Czech Republic under the project “Application of Advanced Visualization and Detection Techniques Based on Semantic Network Analysis (Application II)” (grant number CZ.01.01.01/01/24_062/0007493), call “Applications—Call II—Development of Digital Solutions”.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new datasets were generated or analysed during this review; all primary sources are cited in the reference list.

Acknowledgments: The authors thank the anonymous reviewers of an earlier version of this work submitted to ICCCI 2026 for comments that informed the structure of this review. During the preparation of this manuscript, the authors used Claude Code (v2) and Claude (Anthropic; Claude Opus 4.8 model) for the following purposes: conversion of the manuscript between document formats and formatting to the journal template; drafting and editing assistance on the text, including the abstract, the PRISMA-ScR framing of the methodology (Section 2) and the PRISMA-ScR checklist; bibliographic formatting and consistency checking; and the creation and rasterisation of figures. The taxonomy, the governance mapping and the roadmap milestones originate from the authors. The authors have reviewed and edited every paragraph, have verified all factual and legal claims (including references to Regulation (EU) 2024/1689 and the NIST AI RMF 1.0), and take full responsibility for the content of this publication. No AI tool is listed as an author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the Opportunities and Risks of Foundation Models. *arXiv* **2021**. [CrossRef]
2. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. GPT-4 Technical Report. *arXiv* **2023**. [CrossRef]
3. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018.
4. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.R.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In Proceedings of the Eleventh International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023. Available online: https://openreview.net/forum?id=WE_vluYUL-X (accessed on 18 July 2026).
5. Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language models can teach themselves to use tools. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 68539–68551. [CrossRef]
6. Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; et al. WebGPT: Browser-assisted question-answering with human feedback. *arXiv* **2021**. [CrossRef]
7. Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 11809–11822. [CrossRef]
8. Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 8634–8652. [CrossRef]
9. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). *Off. J. Eur. Union* **2024**. Available online: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj> (accessed on 18 July 2026). [CrossRef]
10. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. 2023. Available online: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf> (accessed on 18 July 2026).
11. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, Toronto, ON, Canada, 3–10 March 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 610–623. [CrossRef]
12. Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. Ethical and social risks of harm from Language Models. *arXiv* **2021**. [CrossRef]
13. Russell, S. *Human Compatible: Artificial Intelligence and the Problem of Control*; Viking: New York, NY, USA, 2019.
14. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete Problems in AI Safety. *arXiv* **2016**. [CrossRef]

15. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.J.; Horsley, T.; Weeks, L.; et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann. Intern. Med.* **2018**, *169*, 467–473. [CrossRef] [PubMed]
16. Acharya, D.B.; Kuppan, K.; Divya, B. Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey. *IEEE Access* **2025**, *13*, 18912–18936. [CrossRef]
17. Hubinger, E.; van Merwijk, C.; Mikulik, V.; Skalse, J.; Garrabrant, S. Risks from Learned Optimization in Advanced Machine Learning Systems. *arXiv* **2019**. [CrossRef]
18. García, J.; Fernández, F. A comprehensive survey on safe reinforcement learning. *J. Mach. Learn. Res.* **2015**, *16*, 1437–1480.
19. OWASP Foundation. OWASP Top 10 for Large Language Model Applications, Version 2025. Available online: <https://owasp.org/www-project-top-10-for-large-language-model-applications/> (accessed on 24 January 2026).
20. OWASP GenAI Security Project—Agentic Security Initiative. OWASP Top 10 for Agentic Applications, 2025. Available online: <https://genai.owasp.org/2025/12/09/owasp-top-10-for-agentic-applications-the-benchmark-for-agentic-security-in-the-age-of-autonomous-ai/> (accessed on 24 January 2026).
21. Dafoe, A.; Hughes, E.; Bachrach, Y.; Collins, T.; McKee, K.R.; Leibo, J.Z.; Larson, K.; Graepel, T. Open Problems in Cooperative AI. *arXiv* **2020**. [CrossRef]
22. Nannini, L.; Leon Smith, A.; Maggini, M.J.; Panai, E.; Feliciano, S.; Tiulkanov, A.; Maran, E.; Gealy, J.; Bisconti, P. AI Agents Under EU Law: A Compliance Architecture for AI Providers. *arXiv* **2026**. [CrossRef]
23. Valenta, T.; Novotný, J.; Horálek, J. Architecting AI for Compliance: Integrating EU AI Act Principles with MITRE ATLAS Threat Models. In *Recent Challenges in Intelligent Information and Database Systems, Proceedings of the 18th Asian Conference on Intelligent Information and Database Systems (ACIIDS 2026), Kaohsiung, Taiwan, 13–15 April 2026*; Nguyen, N.T., Chen, C.-H., Fujita, H., Hong, T.-P., Manolopoulos, Y., Wojtkiewicz, K., Eds.; Communications in Computer and Information Science; Springer: Singapore, 2026; Volume 2964, pp. 310–323. [CrossRef]
24. Krakovna, V. Specification Gaming Examples in AI [Blog Post]. Available online: <https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/> (accessed on 1 March 2026).
25. Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv* **2022**. [CrossRef]
26. Bowman, S.R. Eight things to know about large language models. *Crit. AI* **2024**, *2*. [CrossRef]
27. Hadfield-Menell, D.; Russell, S.J.; Abbeel, P.; Dragan, A. Cooperative inverse reinforcement learning. *Adv. Neural Inf. Process. Syst.* **2016**, *29*. Available online: https://proceedings.neurips.cc/paper_files/paper/2016/hash/c3395dd46c34fa7fd8d729d8cf88b7a8-Abstract.html (accessed on 18 July 2026).
28. Ng, A.Y.; Russell, S. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*; Morgan Kaufmann: San Francisco, CA, USA, 2000; pp. 663–670. Available online: https://www.cl.cam.ac.uk/ey204/teaching/ACS/R244_2022_2023/papers/NG_ICML_2000.pdf (accessed on 18 July 2026).
29. Christiano, P.F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; Amodei, D. Deep reinforcement learning from human preferences. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. Available online: <https://proceedings.neurips.cc/paper/7017-deep-reinforcement-learning-from-human-preferences> (accessed on 18 July 2026).
30. Leike, J.; Krueger, D.; Everitt, T.; Martic, M.; Maini, V.; Legg, S. Scalable agent alignment via reward modeling: A research direction. *arXiv* **2018**. [CrossRef]
31. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744. [CrossRef]
32. Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C. Constitutional AI: Harmlessness from AI feedback. *arXiv* **2022**. [CrossRef]
33. Greenblatt, R.; Denison, C.; Wright, B.; Roger, F.; MacDiarmid, M.; Marks, S.; Treutlein, J.; Belonax, T.; Chen, J.; Duvenaud, D.; et al. Alignment faking in large language models. *arXiv* **2024**. [CrossRef]
34. Hubinger, E.; Denison, C.; Mu, J.; Lambert, M.; Tong, M.; MacDiarmid, M.; Lanham, T.; Ziegler, D.M.; Maxwell, T.; Cheng, N.; et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv* **2024**. [CrossRef]
35. Meinke, A.; Schoen, B.; Scheurer, J.; Balesni, M.; Shah, R.; Hobbhahn, M. Frontier models are capable of in-context scheming. *arXiv* **2024**. [CrossRef]
36. Betley, J.; Tan, D.; Warncke, N.; Sztyber-Betley, A.; Bao, X.; Soto, M.; Labenz, N.; Evans, O. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. *Int. Conf. Mach. Learn. (ICML)* **2025**, PMLR 267, 4043–4068. [CrossRef]
37. Omohundro, S.M. The Basic AI Drives. In *Artificial Intelligence Safety and Security*; Yampolskiy, R.V., Ed.; Chapman & Hall/CRC: Boca Raton, FL, USA, 2018; pp. 47–55. [CrossRef]
38. Turner, A.M.; Smith, L.R.; Shah, R.; Critch, A.; Tadepalli, P. Optimal policies tend to seek power. *Adv. Neural Inf. Process. Syst.* **2021**, *34*. Available online: <https://proceedings.neurips.cc/paper/2021/hash/c26820b8a4c1b3c2aa868d6d57e14a79-Abstract.html> (accessed on 18 July 2026).

39. Orseau, L.; Armstrong, S. Safely interruptible agents. In Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence (UAI 2016), Jersey City, NJ, USA, 25–29 June 2016. Available online: <https://ora.ox.ac.uk/objects/uuid:17c0e095-4e13-47fc-bace-64ec46134a3f> (accessed on 18 July 2026).
40. Everitt, T.; Krakovna, V.; Orseau, L.; Hutter, M.; Legg, S. Reinforcement Learning with a Corrupted Reward Channel. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017), Melbourne, Australia, 19–25 August 2017. [CrossRef] [PubMed]
41. Soares, N.; Fallenstein, B. Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda. In *The Technological Singularity: Managing the Journey*; Callaghan, V., Miller, J., Yampolskiy, R., Armstrong, S., Eds.; The Frontiers Collection; Springer: Berlin/Heidelberg, Germany, 2017; pp. 103–125. [CrossRef]
42. Achiam, J.; Held, D.; Tamar, A.; Abbeel, P. Constrained policy optimization. *Proc. Mach. Learn. Res.* **2017**, *70*, 22–31. [CrossRef] [PubMed]
43. Chow, Y.; Nachum, O.; Faust, A.; Duenez-Guzman, E.; Ghavamzadeh, M. Lyapunov-based Safe Policy Optimization for Continuous Control. *arXiv* **2019**. [CrossRef]
44. Leike, J.; Martic, M.; Krakovna, V.; Ortega, P.A.; Everitt, T.; Lefrancq, A.; Orseau, L.; Legg, S. AI Safety Gridworlds. *arXiv* **2017**. [CrossRef]
45. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. *Proc. Mach. Learn. Res.* **2017**, *70*, 1321–1330.
46. Koh, P.W.; Sagawa, S.; Marklund, H.; Xie, S.M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R.L.; Gao, I. WILDS: A benchmark of in-the-wild distribution shifts. *Proc. Mach. Learn. Res.* **2021**, *139*, 5637–5664.
47. Hendrycks, D.; Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *arXiv* **2017**. [CrossRef]
48. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. *arXiv* **2015**. [CrossRef]
49. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv* **2018**. [CrossRef]
50. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2014**. [CrossRef]
51. Irving, G.; Christiano, P.; Amodei, D. AI safety via debate. *arXiv* **2018**. [CrossRef]
52. Lightman, H.; Kosaraju, V.; Burda, Y.; Edwards, H.; Baker, B.; Lee, T.; Leike, J.; Schulman, J.; Sutskever, I.; Cobbe, K. Let’s verify step by step. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024. Available online: <https://openreview.net/forum?id=v8L0pN6EOi> (accessed on 18 July 2026).
53. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837. [CrossRef]
54. Lanham, T.; Chen, A.; Radhakrishnan, A.; Steiner, B.; Denison, C.; Hernandez, D.; Li, D.; Durmus, E.; Hubinger, E.; Kernion, J. Measuring faithfulness in chain-of-thought reasoning. *arXiv* **2023**. [CrossRef]
55. Turpin, M.; Michael, J.; Perez, E.; Bowman, S. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 74952–74965. [CrossRef]
56. Greenblatt, R.; Shlegeris, B.; Sachan, K.; Roger, F. AI control: Improving safety despite intentional subversion. *arXiv* **2024**. [CrossRef]
57. Korbak, T.; Balesni, M.; Barnes, E.; Bengio, Y.; Benton, J.; Bloom, J.; Chen, M.; Cooney, A.; Dafoe, A.; Dragan, A.; et al. Chain of thought monitorability: A new and fragile opportunity for AI safety. *arXiv* **2025**. [CrossRef]
58. OpenAI. Detecting and Reducing Scheming in AI Models. Available online: <https://openai.com/index/detecting-and-reducing-scheming-in-ai-models/> (accessed on 1 March 2026).
59. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine learning interpretability: A survey on methods and metrics. *Electronics* **2019**, *8*, 832. [CrossRef]
60. Nanda, N.; Chan, L.; Lieberum, T.; Smith, J.; Steinhardt, J. Progress measures for grokking via mechanistic interpretability. *arXiv* **2023**. [CrossRef]
61. Olah, C.; Cammarata, N.; Schubert, L.; Goh, G.; Petrov, M.; Carter, S. Zoom in: An introduction to circuits. *Distill* **2020**, *5*, e00024-001. Available online: <https://distill.pub/2020/circuits/zoom-in/> (accessed on 18 July 2026). [CrossRef]
62. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16), San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144. [CrossRef]
63. Lundberg, S.M.; Lee, S. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. Available online: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html> (accessed on 18 July 2026).
64. Elhage, N.; Hume, T.; Olsson, C.; Schiefer, N.; Henighan, T.; Kravec, S.; Hatfield-Dodds, Z.; Lasenby, R.; Drain, D.; Chen, C.; et al. Toy Models of Superposition. *arXiv* **2022**. [CrossRef]

65. Bricken, T.; Templeton, A.; Batson, J.; Chen, B.; Jermyn, A.; Conerly, T.; Turner, N.; Anil, C.; Denison, C.; Askell, A. Towards Monosemanticity: Decomposing Language Models with Dictionary Learning. Transformer Circuits Thread. 2023. Available online: <https://transformer-circuits.pub/2023/monosemantic-features> (accessed on 18 July 2026).
66. Wang, K.; Variengien, A.; Conmy, A.; Shlegeris, B.; Steinhardt, J. Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 small. *arXiv* **2023**. [CrossRef]
67. Meng, K.; Bau, D.; Andonian, A.; Belinkov, Y. Locating and editing factual associations in gpt. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 17359–17372. [CrossRef]
68. Templeton, A.; Conerly, T.; Marcus, J.; Lindsey, J.; Bricken, T.; Chen, B.; Pearce, A.; Citro, C.; Ameisen, E.; Jones, A.; et al. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. Transformer Circuits Thread. 2024. Available online: <https://transformer-circuits.pub/2024/scaling-monosemanticity/> (accessed on 18 July 2026).
69. Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, N.; Ndousse, K.; et al. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *arXiv* **2022**. [CrossRef]
70. Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J.Z.; Fredrikson, M. Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv* **2023**. [CrossRef]
71. Chao, P.; Robey, A.; Dobriban, E.; Hassani, H.; Pappas, G.J.; Wong, E. Jailbreaking black box large language models in twenty queries. In *Proceedings of the 2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), Copenhagen, Denmark, 9–11 April 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 23–42. Available online: <https://ieeexplore.ieee.org/abstract/document/10992337/> (accessed on 18 July 2026).
72. Anil, C.; Durmus, E.; Sharma, M.; Benton, J.; Kundu, S.; Batson, J.; Rinsky, N.; Tong, M.; Mu, J.; Ford, D.; et al. Many-shot Jailbreaking. In *Advances in Neural Information Processing Systems 37, Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, BC, Canada, 9–10 December 2024*; Curran Associates, Inc.: Red Hook, NY, USA, 2024; pp. 129696–129742. Available online: <https://www.anthropic.com/research/many-shot-jailbreaking> (accessed on 18 July 2026).
73. Robey, A.; Wong, E.; Hassani, H.; Pappas, G.J. SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks. *arXiv* **2025**. [CrossRef]
74. Xie, Y.; Yi, J.; Shao, J.; Curl, J.; Lyu, L.; Chen, Q.; Xie, X.; Wu, F. Defending chatgpt against jailbreak attack via self-reminders. *Nat. Mach. Intell.* **2023**, *5*, 1486–1496. [CrossRef]
75. Pisano, M.; Ly, P.; Sanders, A.; Yao, B.; Wang, D.; Strzalkowski, T.; Si, M. Bergeron: Combating Adversarial Attacks through a Conscience-Based Alignment Framework. *arXiv* **2023**. [CrossRef]
76. Wei, A.; Haghtalab, N.; Steinhardt, J. Jailbroken: How does llm safety training fail? *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 80079–80110. [CrossRef]
77. Carlini, N.; Hayes, J.; Nasr, M.; Jagielski, M.; Sehwag, V.; Tramèr, F.; Balle, B.; Ippolito, D.; Wallace, E. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23), Anaheim, CA, USA, 9–11 August 2023*; USENIX Association: Anaheim, CA, USA, 2023; pp. 5253–5270. Available online: <https://www.usenix.org/conference/usenixsecurity23/presentation/carlini> (accessed on 18 July 2026).
78. DeBenedetti, E.; Zhang, J.; Beurer-Kellner, L.; Fischer, M.; Tramèr, F. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. *arXiv* **2024**. [CrossRef]
79. Radosevich, B.; Halloran, J. MCP safety audit: LLMs with the Model Context Protocol allow major security exploits. *arXiv* **2025**. [CrossRef]
80. Zhan, Q.; Liang, Z.; Ying, Z.; Kang, D. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024*; pp. 10471–10506. [CrossRef]
81. Calvano, E.; Calzolari, G.; Denicolo, V.; Pastorello, S. Artificial intelligence, algorithmic pricing, and collusion. *Am. Econ. Rev.* **2020**, *110*, 3267–3297. [CrossRef]
82. Leibo, J.Z.; Zambaldi, V.; Lanctot, M.; Marecki, J.; Graepel, T. Multi-agent Reinforcement Learning in Sequential Social Dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2017), São Paulo, Brazil, 8–12 May 2017*. [CrossRef]
83. Shoham, Y.; Leyton-Brown, K. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*; Cambridge University Press: Cambridge, UK, 2008. [CrossRef]
84. Foerster, J.; Assael, I.A.; De Freitas, N.; Whiteson, S. Learning to communicate with deep multi-agent reinforcement learning. *Adv. Neural Inf. Process. Syst.* **2016**, *29*. Available online: <https://proceedings.neurips.cc/paper/2016/hash/c7635bfd99248a2cdcf8249ef7bfbef4-Abstract.html> (accessed on 18 July 2026).
85. Lazaridou, A.; Baroni, M. Emergent Multi-Agent Communication in the Deep Learning Era. *arXiv* **2020**. [CrossRef]
86. Park, J.S.; O'Brien, J.; Cai, C.J.; Morris, M.R.; Liang, P.; Bernstein, M.S. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), San Francisco, CA, USA*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 1–22. [CrossRef]

87. Park, J.S.; Popowski, L.; Cai, C.; Morris, M.R.; Liang, P.; Bernstein, M.S. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*, Bend, OR, USA, 29 October–2 November 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 1–18. [\[CrossRef\]](#)
88. Li, G.; Hammoud, H.; Itani, H.; Khizbullin, D.; Ghanem, B. CAMEL: Communicative agents for “mind” exploration of large language model society. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 51991–52008. [\[CrossRef\]](#)
89. Li, Y.; Sun, L.; Zhang, Y. MetaAgents: Large Language Model Based Agents for Decision-Making on Teaming. *Proc. ACM Hum.-Comput. Interact.* **2025**, *9*, CSCW134. [\[CrossRef\]](#)
90. Hammond, L.; Chan, A.; Clifton, J.; Hoelscher-Obermaier, J.; Khan, A.; McLean, E.; Smith, C.; Barfuss, W.; Foerster, J.; Gavenčiak, T.; et al. Multi-agent risks from advanced AI (Cooperative AI Foundation Technical Report No. 1). *arXiv* **2025**. [\[CrossRef\]](#)
91. Liu, X.; Yu, H.; Zhang, H.; Xu, Y.; Lei, X.; Lai, H.; Gu, Y.; Ding, H.; Men, K.; Yang, K.; et al. AgentBench: Evaluating LLMs as Agents. *arXiv* **2024**. [\[CrossRef\]](#)
92. Jimenez, C.E.; Yang, J.; Wettig, A.; Yao, S.; Pei, K.; Press, O.; Narasimhan, K. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? *arXiv* **2024**. [\[CrossRef\]](#)
93. Qin, Y.; Liang, S.; Ye, Y.; Zhu, K.; Yan, L.; Lu, Y.; Lin, Y.; Cong, X.; Tang, X.; Qian, B.; et al. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. *arXiv* **2024**. [\[CrossRef\]](#)
94. Zhou, S.; Xu, F.F.; Zhu, H.; Zhou, X.; Lo, R.; Sridhar, A.; Cheng, X.; Ou, T.; Bisk, Y.; Fried, D.; et al. WebArena: A Realistic Web Environment for Building Autonomous Agents. *arXiv* **2024**. [\[CrossRef\]](#)
95. Liu, E.Z.; Guu, K.; Pasupat, P.; Shi, T.; Liang, P. Reinforcement Learning on Web Interfaces Using Workflow-Guided Exploration. *arXiv* **2018**. [\[CrossRef\]](#)
96. Shridhar, M.; Yuan, X.; Côté, M.; Bisk, Y.; Trischler, A.; Hausknecht, M. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. *arXiv* **2021**. [\[CrossRef\]](#)
97. Wang, R.; Jansen, P.; Côté, M.; Ammanabrolu, P. ScienceWorld: Is your Agent Smarter than a 5th Grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, Abu Dhabi, United Arab Emirates, 7–11 December 2022; Association for Computational Linguistics: Abu Dhabi, United Arab Emirates, 2022. [\[CrossRef\]](#)
98. Chevalier-Boisvert, M.; Bahdanau, D.; Lahlou, S.; Willems, L.; Saharia, C.; Nguyen, T.H.; Bengio, Y. BabyAI: A Platform to Study the Sample Efficiency of Grounded Language Learning. *arXiv* **2019**. [\[CrossRef\]](#)
99. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv* **2023**. [\[CrossRef\]](#)
100. Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; Anandkumar, A. Voyager: An Open-Ended Embodied Agent with Large Language Models. *arXiv* **2024**. [\[CrossRef\]](#)
101. Katz, G.; Barrett, C.; Dill, D.; Julian, K.; Kochenderfer, M. Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks. In *Computer Aided Verification (CAV 2017)*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2017; Volume 10426. [\[CrossRef\]](#)
102. Andriushchenko, M.; Croce, F.; Flammarion, N. AgentHarm: A benchmark for measuring harmfulness of LLM agents. *arXiv* **2025**. [\[CrossRef\]](#)
103. Glukhov, D.; Shumailov, I.; Gal, Y.; Papernot, N.; Papayan, V. LLM Censorship: A Machine Learning Challenge or a Computer Security Problem? *arXiv* **2023**. [\[CrossRef\]](#)
104. Anthropic. Agentic Misalignment: How LLMs Could be Insider Threats. Available online: <https://www.anthropic.com/research/agentic-misalignment> (accessed on 1 March 2026).
105. European Commission. Proposal for a Regulation Amending Regulations (EU) 2024/1689 and (EU) 2018/1139 as Regards the Simplification of the Implementation of Harmonised Rules on Artificial Intelligence (Digital Omnibus on AI). COM(2025) 836 Final, 2025/0359 (COD). 2025. Available online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52025PC0836> (accessed on 1 March 2026).
106. European AI Office. *General-Purpose AI Code of Practice*; European Commission: Luxembourg, 2025. Available online: <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai> (accessed on 1 March 2026).
107. National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1)*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2024. [\[CrossRef\]](#)
108. Bengio, Y.; Clare, S.; Prunkl, C.; Andriushchenko, M.; Bucknall, B.; Murray, M.; Rismani, S.; McGlynn, C.; Maslej, N.; Fox, P.; et al. International AI Safety Report 2026. Available online: <https://internationalaisafetyreport.org> (accessed on 1 March 2026).
109. European Commission. *Action Plan on Cybersecurity and Artificial Intelligence*; COM(2026) 577 Final; European Commission: Strasbourg, France, 2026. Available online: <https://digital-strategy.ec.europa.eu/en/library/eu-action-plan-cybersecurity-and-artificial-intelligence> (accessed on 18 July 2026).
110. European Commission. Commission Launches Whistleblower Tool for AI Act. Press Release. 24 November 2025. Available online: <https://digital-strategy.ec.europa.eu/en/news/commission-launches-whistleblower-tool-ai-act> (accessed on 18 July 2026).

111. Katz, G.; Huang, D.A.; Ibeling, D.; Julian, K.; Lazarus, C.; Lim, R.; Shah, P.; Thakoor, S.; Wu, H.; Dill, D.L.; et al. The Marabou Framework for Verification and Analysis of Deep Neural Networks. In *Computer Aided Verification (CAV 2019)*; Dillig, I., Tasiran, S., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2019; Volume 11561, pp. 443–452. [[CrossRef](#)]
112. Singh, G.; Gehr, T.; Mirman, M.; Püschel, M.; Vechev, M. Fast and effective robustness certification. *Adv. Neural Inf. Process. Syst.* **2018**, *31*. Available online: https://proceedings.neurips.cc/paper_files/paper/2018/hash/f2f446980d8e971ef3da97af089481c3-Abstract.html (accessed on 18 July 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.