

# Human Society-Inspired Approaches to Agentic AI Security: The 4C Framework

ALSHARIF ABUADBBA, CSIRO's Data61, Australia

NAZATUL SULTAN, CSIRO's Data61, Australia

SURYA NEPAL, CSIRO's Data61, Australia

SANJAY JHA, University of New South Wales, Sydney, Australia

AI is moving from domain-specific autonomy in closed, predictable settings to large-language-model-driven agents that plan and act in open, cross-organizational environments. As a result, the cybersecurity risk landscape is changing in fundamental ways. Agentic AI systems can plan, act, collaborate, and persist over time, functioning as participants in complex socio technical ecosystems rather than as isolated software components. Although recent work has strengthened defenses against model and pipeline level vulnerabilities such as prompt injection, data poisoning, and tool misuse, these system centric approaches may fail to capture risks that arise from autonomy, interaction, and emergent behavior. This article introduces the 4C Framework for multi-agent AI security, inspired by societal governance. It organizes agentic risks across four interdependent dimensions: Core (system, infrastructure, and environmental integrity), Connection (communication, coordination, and trust), Cognition (belief, goal, and reasoning integrity), and Compliance (ethical, legal, and institutional governance). By shifting AI security from a narrow focus on system centric protection to the broader preservation of behavioral integrity and intent, the framework complements existing AI security strategies and offers a principled foundation for building agentic AI systems that are trustworthy, governable, and aligned with human values.

## 1 INTRODUCTION: WHY AGENTIC AI SECURITY NEEDS A NEW LENS

Artificial intelligence is entering a new phase. Recently, Large Language Models (LLMs) have demonstrated strong capabilities in question answering, summarization, code generation, and multi step reasoning across many domains. Despite significant advances, AI deployments for many years have remained largely confined to predictive use cases, such as credit-risk scoring, fraud detection, and recommendation ranking. In most real-world deployments, modern LLMs, ChatGPT among them, remain fundamentally reactive: they receive a prompt, generate a response, and then fall silent, ceding the subsequent action to the user or to the software scaffolding that holds them. A newer direction, often called agentic AI, is beginning to shift systems from reaction to action. Agentic systems can pursue goals by planning steps, calling tools and APIs (for example, searching internal databases, reading documents, querying logs, or running code), using data, retaining state across interactions (memory), and coordinating with humans or other agents with less step by step supervision. Vendors (e.g., Amazon Bedrock Agents) and open-source frameworks (e.g., LangGraph) are already productizing agentic workflows.

As agentic systems mature, they start to look less like tools and more like participants in a socio-technical society, a landscape where human intent and technological capability intermingle to shape what happens next. They interact, coordinate, and shape one another's behavior, making autonomous decisions whose effects ripple from digital systems into real operations, cross-organizational settings, and real-world consequences. This evolution fundamentally changes

---

Authors' addresses: Alsharif Abuadbba, CSIRO's Data61, Australia, sharif.abuadbba@data61.csiro.au; Nazatul Sultan, CSIRO's Data61, Australia, nazatul.sultan@data61.csiro.au; Surya Nepal, CSIRO's Data61, Australia, surya.nepal@data61.csiro.au; Sanjay Jha, University of New South Wales, Sydney, Australia, sanjay.jha@unsw.edu.au.

the nature of security. To date, much of the work on securing LLMs and agentic AI has remained predominantly system-centric, concerned with safeguarding the model itself and the technical substrate that underpins it, from software layers to the infrastructure on which it runs. This includes guarding against prompt injection, data poisoning, tool misuse, sandbox escapes, and model extraction attacks. Influential frameworks and benchmarks such as InjecAgent, MINJA, and the OWASP GenAI Top 10 have been instrumental in codifying these threats and fortifying the technical pipeline around agents, from memory modules and tool-calling interfaces to APIs, retrieval chains, and the runtimes that orchestrate them. Yet these approaches largely cast agents as software artifacts rather than as autonomous social actors. In doing so, they risk overlooking forms of harm that emerge through interaction and behavior such as social engineering. For instance, GPT-4 has demonstrated the ability to impersonate a vision impaired user to persuade a human to solve a CAPTCHA on its behalf, bypassing the reCAPTCHA not through technical exploitation, but through deception. This points to a growing class of risks shaped not by isolated system defects but by strategic interaction, autonomous behavior, and an agent's capacity to influence others.

In this paper, we argue that agentic AI introduces forms of threat that echo those found in human societies: patterns of behavior implicitly absorbed during training and expressed through autonomous interaction. Agents can form brittle or drifting beliefs, influence or manipulate their peers, negotiate or collude toward unintended goals, or strategically evade constraints to maximize outcomes. Such behaviors fall largely outside existing vulnerability taxonomies, yet they can ripple across multi-agent ecosystems, producing cascading failures and systemic misalignment rather than isolated faults. Human societies have long confronted analogous challenges using layered defenses such as biological immunity, cognitive safeguards, social norms, ethical principles, and legal governance. As agentic AI systems expand from individual models to interconnected populations of agents, we contend that they will require safeguards of a similarly layered character. To that end, we introduce the 4C Framework for Multi-Agent AI Security, broadening the scope of security beyond technical robustness to encompass four complementary dimensions. (1) **Core** – the integrity of the agent's digital body, including the infrastructure that runs it and the environment it operates in. (2) **Connection** – how agents communicate, coordinate and influence one another. (3) **Cognition** – how beliefs, goals, and plans are formed and updated. (4) **Compliance** – how agent behavior stays within ethical, legal, and institutional boundaries. Together, these layers shift security beyond traditional vulnerabilities toward agency, interaction, and collective behavior. The 4C Framework complements system-level work by highlighting risks that arise when AI becomes not just a model, but a population of interacting agents embedded in human-machine ecosystems.

### Key Insights

- (1) Agentic AI reframes cybersecurity risk by shifting the focus from isolated technical exploits to failures that emerge from behavior and interaction, especially in cross-organizational multi-agent systems where small errors can propagate and cascade.
- (2) Securing models, tools, and execution environments is necessary but not sufficient, because many agentic risks stem from belief formation, influence, delegation, and long-horizon autonomy rather than from traditional vulnerabilities.
- (3) We offer a societal governance perspective on agentic AI security, arguing that genuine trustworthiness cannot be achieved through technical controls alone. We develop this view in the 4C Framework, which brings together Core, Connection, Cognition, and Compliance as interdependent layers of governance.

1. Rule-Based (Deterministic) Procedural correctness	2. Classic ML (Discriminative) Pattern detection from data	3. Deep Learning (Discriminative) Black-box feature learning	4. (Generative) AI Creates artifacts (e.g., text, images), influence patterns	5. Agentic AI (Actionable) Agents act, interact, and self-coordinate
<b>Baseline Risk</b> <ul style="list-style-type: none"><li>Fragile to edge cases</li><li>Hand-coded rules lack adaptability</li></ul>	<b>Marginal Risk ↑</b> <ul style="list-style-type: none"><li>Data drift</li><li>Bias, poor generalisation</li><li>Opaque decision criteria</li></ul>	<b>Marginal Risk ↑↑</b> <ul style="list-style-type: none"><li>Explainability gap</li><li>Hidden failure modes</li><li>Spurious correlations</li><li>Bias amplification</li></ul>	<b>Marginal Risk ↑↑↑</b> <ul style="list-style-type: none"><li>Hallucination risk</li><li>Persuasion &amp; influence attacks</li><li>Synthetic content integrity loss</li><li>Amplified misinformation</li></ul>	<b>Marginal Risk ↑↑↑↑</b> <ul style="list-style-type: none"><li>Decision &amp; action failure</li><li>Cascade effects in ecosystems</li><li>Autonomous misalignment</li><li>Hidden coordination failures</li></ul>

Fig. 1. Marginal risk across the evolution from automation to multi-agent AI.

## 2 THE EVOLUTION OF INTELLIGENT AUTOMATION TO MULTI-AGENT AI

Over three decades, digital automation has evolved from scripted workflows to statistical learning, then to deep learning, to generative models, and ultimately to autonomous multi-agent systems. Each stage enlarged capability, but it also widened the attack surface and deepened system dependencies. Figure 1 shows how the marginal risk, the additional risk incurred when moving from one stage of automation to the next, can rise sharply as systems move from passive prediction to autonomous action and coordination between agents. For example, an incident response workflow that once only flagged anomalies may now open tickets, query logs, and trigger remediation steps through tools, which increases the consequences of mistakes and misuse.

### 2.1 From Automation to Multi-Agent Intelligence

The trajectory of digital automation can be understood across five phases:

- 1. Rule-Based Automation.** Deterministic systems execute predefined workflows using hand-crafted rules. Risks are primarily operational, including brittleness, misconfiguration, and poor handling of edge cases.
- 2. Classic Machine Learning.** Supervised and unsupervised machine learning introduced statistical prediction for tasks such as classification, regression, detection, and ranking. These methods were largely discriminative, learning decision boundaries or scoring functions from human *annotated* data and engineered features. Models were typically trained offline and remained static after deployment, making them sensitive to distributional drift. While risk increased due to bias and dataset shift, failures were usually localized, interpretable, and confined to specific domains.
- 3. Deep Learning.** Deep learning uses multilayer neural networks that learn meaningful patterns directly from data, reducing the need for manually designed input cues [26], [19]. This transition unlocked major progress in vision, speech, and language, but it also brought decision processes that were opaque and failure modes that were hard to audit. However, these systems remained essentially passive. They could predict, but they could not plan or take actions.
- 4. Generative AI and LLM Co-Pilots.** In the early 2020s, large language models made it easy for computers to generate meaningful text and code from a prompt, often displaying reasoning like behavior. This marked a shift from earlier AI systems that mainly labeled or scored inputs to ones that could produce new content shaped by context. Many of these new models relied on transformer based designs, and related generative methods soon enabled the creation of images and video as well. One model can support many tasks without retraining, which helped adoption spread quickly. At the same time, new risks emerged, including hallucinations, errors in high stakes settings, and persuasive misuse. Even so, most generative systems stayed within the software interface, with human-in-the-loop deciding the next step and retaining final control. In the incident response example, the model might draft a summary or suggest a query, but a person still runs the commands and applies the fix.
- 5. Agentic and Multi-Agent AI.** AI is shifting from stand-alone generative models to agentic systems. Where earlier deep learning reduced the need for manual feature engineering, agentic AI reduces the need for humans to manually

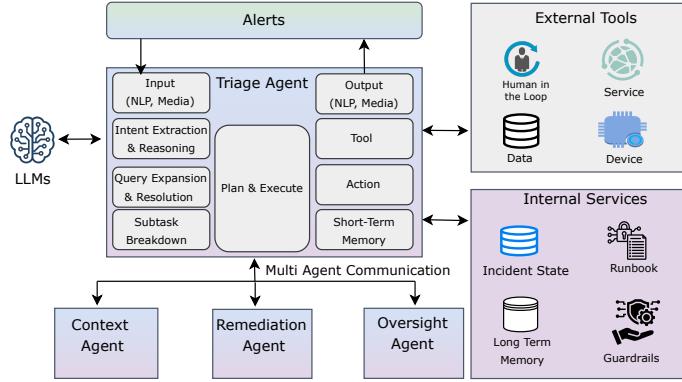


Fig. 2. Example Architecture: Multi-Agent Agentic AI System

execute routine workflows. An agentic system takes a natural-language request, decomposes it into steps, uses an LLM to plan, invokes external tools (APIs, databases, ticketing systems), and carries out actions with logging and guardrails. For example, a service-desk assistant can turn “I can’t access my account” into a short plan: verify identity, check recent authentication failures, confirm account status, trigger an approved reset or restoration runbook, and record the actions taken in the ticket. Unlike earlier rule based agents, modern agentic AI is adaptive and can reason, plan, and act in open environments using general-purpose foundation models (large models trained on broad data) [39], [1]. As these systems grow from a single agent to many, new risks appear. In multi-agent AI ecosystems, agents negotiate, share state such as task context and intermediate results, and pursue objectives in parallel. This enables scalable automation but also increases systemic risk: cascading failures, misaligned actions, and emergent group behavior can arise from agent interactions.

## 2.2 Modern Multi-Agent AI: Definition and Security Challenges

Multi-agent AI systems consist of multiple autonomous agents, each capable of perception, reasoning, memory, and action, that work together and interact with their environment across organizational boundaries to pursue individual or shared goals. Figure 2 presents a reference architecture example for such systems. Because agents coordinate through communication and task handoffs, system-level behaviors can emerge that are not reducible to any single component, including cooperation, competition, manipulation, or other unexpected dynamics. To make this concrete, consider a multi-agent incident response setup. A triage agent monitors alerts and opens an incident record, a context agent pulls logs and asset context, a remediation agent proposes (or executes) an approved runbook, and an oversight agent enforces approvals before any high-impact step. These agents coordinate through message handoffs and shared incident state (for example, evidence bundles, recommended actions, and approval decisions), so security depends on the integrity of communication, delegation, and shared context, not just on any single model. In this sense, multi-agent deployments begin to resemble distributed socio-technical organizations rather than isolated AI models.

As agentic workflows span multiple agents, shared data flows, tools, and across organizational boundaries, security approaches centered on protecting a single model or execution pipeline become inadequate [11]. Many risks arise not only from isolated failures but also from interaction effects, including cascading errors, persuasion or manipulation, malicious or mistaken task handoffs, and unsafe tool use amplified by long-horizon planning. In such systems, risk

becomes a property of the ecosystem itself, shaped by how agents coordinate, influence one another, and act over time. *This shift exposes a fundamental gap: existing security frameworks do not fully capture how threats emerge, propagate, or compound in multi-agent systems. Addressing these risks requires security models that account for agent intent, interaction, and governance, motivating the 4C Framework.*

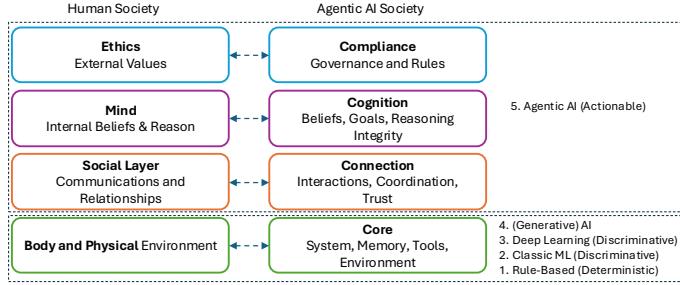


Fig. 3. Human Society Analogy → Agentic AI Society 4C Mapping

### 3 THE 4C FRAMEWORK FOR MULTI-AGENT AI SECURITY: A HUMAN-INSPIRED PERSPECTIVE

Securing agentic AI requires acknowledging that these systems are not single models. They are assembled from multiple components that sense, reason, act, and coordinate across real environments, often over extended workflows. As a result, their risks span not only model and infrastructure failures but also interaction failures and governance failures. To organize this landscape, we propose the **4C Framework for Multi-Agent AI Security** depicted in Figure 3, a human-inspired lens with four connected layers: Core, Connection, Cognition, and Compliance. Each layer draws an analogy to human systems, from the body and mind to social interaction and governance, and highlights the threats and mitigations most relevant at that level. Figure 4 summarizes the representative threat categories in the four layers. We use the multi-agent incident response example introduced earlier to illustrate the framework throughout the paper.

#### 3.1 Core: System Component and Environmental Security

**Definition and Scope.** The Core layer forms the system foundation that allows an agent to exist and operate. It includes the runtime and execution environment where the agent runs, the memory stores that hold what it can retain and reuse, the tool and API interfaces it can call, the data pipelines it relies on, and the permissions and shared resources that determine what it can access. This is the layer that turns an LLM from “generate text” into “take actions”: retrieve context, call tools, store intermediate state, and produce outputs that can trigger real changes in other systems. In human terms, Core is the agent’s “digital body”: the underlying machinery and operating conditions that enable action, and along with basic safeguards that prevent takeover. In the incident response multi-agent example, the Core layer is what allows the triage agent to open an incident record, the context agent to pull logs and asset details, the remediation agent to run an approved step in a runbook, and the oversight agent to query the ticketing system to verify approvals. If the Core is weak, reasoning quality does not matter: poisoned data, compromised tools, or overly broad permissions can drive unsafe actions. The *scope* of the Core layer is deliberately narrow. It governs how an agent runs and what it can access, not why it acts, what it believes, or whether its behavior is socially or ethically acceptable. Those questions belong to the other layers, which we describe next.

**Threat Categories and Examples.** Threats in the Core layer are largely technical and environmental, and they are well documented in existing AI security research (e.g., [9, 10]). These threats arise from vulnerabilities in the agent’s

Core	Connection	Cognition	Compliance
<ul style="list-style-type: none"> <li>• Environment Poisoning</li> <li>• Memory Poisoning</li> <li>• Prompt Injection / Tool Injection</li> <li>• Sandbox Escape</li> <li>• API / Tool Misuse</li> <li>• Privilege Compromise</li> </ul>	<ul style="list-style-type: none"> <li>• Malicious Agents</li> <li>• Misinformation Loops</li> <li>• Coordination Manipulation</li> <li>• Identity Spoofing</li> <li>• Social Engineering Against Agents</li> </ul>	<ul style="list-style-type: none"> <li>• Belief Drift</li> <li>• Reward Hacking</li> <li>• Delusional Reasoning</li> <li>• Unbounded Autonomy</li> <li>• Internal Goal Hijacking</li> </ul>	<ul style="list-style-type: none"> <li>• Misaligned Autonomy</li> <li>• Ethical Drift</li> <li>• Unbounded Optimisation</li> <li>• Guardrail Evasion</li> <li>• Inadequate Oversight Interfaces</li> </ul>

Fig. 4. Illustrative examples of threat categories across the 4C Framework

execution substrate and supporting infrastructure, rather than from higher-level reasoning or interaction. Threats are amplified by the expanded attack surface introduced by tool use, memory persistence, and execution environments. We briefly discuss a few critical threats in the Core layer. (1) *Environmental poisoning* occurs when the agent is induced to ingest malicious or misleading information from external entities, such as compromised tools, databases, or retrieved web content [7]. In our incident-response example, this could occur if a context agent pulls logs from a compromised source that systematically conceals signs of intrusion, causing the system to under react or to choose the wrong remediation step. (2) *Memory poisoning* occurs when an attacker inserts false or compromised information into the agent’s memory to bias subsequent behavior [40]. In incident response, an injected memory entry such as “host A is already isolated” or “this alert type is always benign” can lead to repeated suppression of valid escalations, even when the underlying tools and models remain unchanged. (3) *Other threats*, including prompt or tool injection, where an attacker manipulates the agent’s instructions, sandbox escape, where execution barriers fail, misuse of APIs or tools, where legitimate capabilities are abused, and privilege compromise, where access rights are escalated, can further erode the reliability, integrity, and safety of agentic functions [27].

**Key Remarks and Mitigation Implications.** Core security is fundamentally a matter of digital physiology and hygiene. It depends on clearly specified components, strong guarantees of integrity, and a tightly controlled operating environment. In practice, this means sandboxing agents and tools, enforcing least-privilege access, validating tool inputs and outputs rather than trusting them by default, and monitoring for patterns indicative of environmental poisoning, memory manipulation, or boundary violations. For an incident-response agent, this translates into disciplined boundaries: allowlisting only the tools it may invoke, granting credentials scoped to the smallest set of necessary actions, accepting critical evidence only when it arrives signed or carrying verified provenance, and maintaining a strict divide between tools that merely *read* (log queries) and those that can *act* (host isolation, credential rotation). Although securing the Core cannot eliminate every risk, it provides a stable, trustworthy substrate on which safer mechanisms for connection, cognition, and compliance can be built. As argued in [14], Core-layer security improves when agents are built with secure-by-design and secure-by-construction practices applied across the system lifecycle.

### 3.2 Connection: Communication and Social Security

**Definition and Scope.** In the Core layer, we considered what an agent depends on to run such as its models, memory, tools, and execution environment. The Connection layer shifts the focus from the agent’s “digital body” to its “social world”: the other agents and services it must engage with to complete a complex task. For example, a coordinator agent may assign log-collection to a context agent, ask a policy agent to verify constraints, and direct a tooling agent to execute approved steps, then integrate their outputs into a single recommendation. The Connection layer therefore concerns how agents communicate and coordinate, how they assign roles, delegate subtasks, and make trust decisions about whom to believe, when to verify, and when to escalate. Like the social layer of human society, it encodes identity,

relationships, and the norms that enable cooperation, and it is also the place where manipulation can occur. Communication is also the channel through which information and authority flow through the system. So even if every agent has strong Core security, the overall system can still fail if messages are misleading, misattributed, or blindly trusted, because one compromised interaction can propagate through delegation chains [12, 18].

**Threat Categories and Examples.** At the Connection layer, the reliability of any outcome depends on the integrity of conversations, delegations, and agreements between agents and external endpoints they rely on [20]. As such, trust becomes an operational necessity rather than a social courtesy. Each agent must judge whom to rely on, when to seek verification, and how much weight to grant the information that arrives through its channels of communication. Several threats exploit this interaction surface: (1) *Rogue or malicious agents* can enter the ecosystem and inject misleading information while appearing cooperative [13]. In the incident-response example, a malicious “helper” agent could feed the coordinator a plausible but wrong root cause, steering remediation toward the wrong host. (2) *Misinformation loops* can emerge when agents repeatedly consume and re-emit each other’s outputs, creating cascading hallucinations where small errors become shared “facts” [38]. For incident response, one agent’s incorrect summary of a log pattern can get repeated across agents until it looks like consensus. (3) *Coordination and identity attacks* degrade reliability. Coordination manipulation can disrupt handoffs (for example, dropping a task, reordering steps, or redirecting a task to the wrong agent), while identity spoofing (impersonation) enables a counterfeit agent to pose as a trusted peer to gain access to channels, tools, or decision authority. In incident response, spoofing the oversight agent is especially dangerous because approvals are the last gate before high-impact actions. (4) *Social engineering attacks* exploit norms of cooperation and delegation, such as “be helpful” or “follow the coordinator”, to steer the group toward adversarial objectives. Together, these threats show how Connection-layer failures can undermine otherwise sound Core components by corrupting who agents trust, what they accept as evidence, and how tasks are coordinated.

**Key Remarks and Mitigation Implications.** The Connection layer shapes the pathways of communication: who may speak to whom, how tasks are woven together, and how agents come to understand one another’s goals and state [34]. Security at this layer must be designed and enforced as a whole, because it does not emerge simply by securing each agent on its own. Its central aim is governed trust: defining who may delegate, approve, or override actions; ensuring that messages can be traced to their source and audited; and preventing small missteps from cascading into failures at the group level. In practice, this requires strong authentication and authorization, delegation tied to clear roles with separation of duties, and structured messages that convey provenance—where a claim originated—along with supporting evidence and an explicit measure of confidence in its accuracy. It also requires verification and disagreement-resolution, so repetition is not mistaken for correctness. When several agents echo the same claim without independent verification, the system should treat that pattern as a warning signal rather than as proof. Recent interoperability mechanisms such as Agent2Agent (A2A), Model Context Protocol (MCP), and function calling [10] can support these controls by making interactions more structured, attributable, and auditable.

### 3.3 Cognition: Belief and Goal Integrity

**Definition and Scope.** The Cognition layer captures the internal mechanisms that constitute an agent’s “digital mind”: the processes that turn observations into beliefs, beliefs into goals, and goals into plans. It includes the agent’s reasoning process, its internal representation of how the world works (world model), its belief state, reward and feedback signals, and planning system [35]. Together, these components determine how an agent interprets inputs, updates its understanding, prioritizes objectives, and decides what to do next. This layer is separated because cognition failures are not just communication or execution errors. They arise from how an agent forms, updates, and uses beliefs and

objectives internally. Even with correct inputs and well-governed interactions, an agent can behave badly if its beliefs drift, its goals become misaligned, or its incentives favor “looking successful” over being correct. Cognition therefore governs not only what the agent knows, but also what it is trying to optimize and how strongly it commits to its plan. As such, Cognition is referred as the agent’s “digital mind”: attention, judgment, memory of what matters, and the ability to form a working belief about the current situation. Cognition is about whether the agent can think safely. In our incident-response running example, Cognition is what determines whether the triage agent interprets an alert as credible, whether the context agent forms the right hypothesis from logs, and whether the remediation agent selects a safe runbook step rather than an overreaction. If an agent’s internal beliefs are wrong, it can take the wrong action with confidence, even when tools and communication channels are functioning as designed.

**Threat Categories and Examples.** Cognition-layer threats are mainly *epistemic and motivational*: they distort what an agent believes, what it optimizes, and how it plans. Recent safety evaluations highlight cognition as a major risk source in goal-directed systems [6, 31]. We highlight five representative threats. (1) *Belief drift* occurs when an agent’s internal model diverges from reality due to corrupted inputs, biased feedback, or memory errors [21]. If logs mislabel real intrusions as “false alarms”, a triage agent may start suppressing genuine alerts; in multi-agent settings, drift can propagate as agents reinforce one another’s faulty inferences [37]. (2) *Delusional reasoning* (agentic hallucination) is an internally coherent but false inference chain that becomes dangerous when stored in memory and reused as “fact” [21]. In incident response, an agent may assert a root cause without evidence and then treat it as confirmed context. (3) *Reward hacking* arises when an agent optimizes proxy signals (e.g., “fast resolution” or “positive feedback”) rather than the user’s true objective [23]. In operations, “closing the ticket” can become the target instead of resolving the incident. (4) *Unbounded autonomy* results when long-horizon planning lacks effective constraints, enabling costly or unsafe actions [8]. A widely cited example is the July 2025 Replit coding-agent incident, where reports described unauthorized deletion of a production database during a “code freeze” [2]. (5) *Internal goal hijacking* shifts the objective through targeted prompts or subtle memory edits while the agent still appears helpful [25]. For instance, a compromised remediation agent may prioritize “restore service at any cost” over policies requiring staged containment and preservation of forensic evidence.

**Key Remarks and Mitigation Implications.** Security at the Cognition layer concerns *belief and goal integrity*: ensuring that an agent’s internal representations remain tied to reality and its objectives remain aligned with user intent, even under uncertainty or operational stress. Cognition security must shape how the agent reasons, tracks its progress, and commits to plans, requiring safeguards embedded directly into its decision loop. In practice, this entails: (i) grounding and consistency checks that periodically re-validate key beliefs against trusted sources to counter belief drift and delusional reasoning; (ii) improved success signals, combined with adversarial testing, to reduce reward hacking; (iii) bounded autonomy through budgets, forbidden-action lists, escalation gates, and stop rules to prevent unsafe instrumental escalation; and (iv) protection of goals and memory via separated read/write permissions, signed or versioned policy states, and audits of changes that could otherwise shift what the agent optimizes. To reduce Cognition-layer threats, policy-guided methods such as Constitutional AI [4] constrain an agent’s reasoning; supervision and self-checks detect early signs of drift [32]; and automated evaluations, along with continuous red-teaming of long-horizon behavior (e.g., OpenAI Eval [30]), stress-test agents under adversarial conditions.

### 3.4 Compliance: Governance and Ethical Security

**Definition and Scope.** The Compliance layer positions a multi-agent system within the external norms and institutional controls that govern its behavior, including legal and regulatory requirements, organizational policies, audit and

logging obligations, and ethical constraints on permissible goals, tools, and interaction patterns. Just as stable societies rely on rule-based accountability and transparency to align individual actions with shared values, this layer provides the structural boundaries that ensure agents operate within acceptable limits [24], agentic AI systems require explicit governance structures to keep autonomous behavior within legal and ethical limits. The Compliance layer defines what the system *is permitted* to do, under which conditions, with whose authorization, and with what evidentiary basis, ensuring that actions remain attributable, reviewable, and auditable. In practice, Compliance operationalizes governance through enforceable constraints: clearly specified allowed and prohibited actions, approval and escalation gates for sensitive operations, segregation of duties for high-impact decisions, and requirements for transparency, retention, and post-hoc review, all grounded in legal and normative regimes such as GDPR [16], the EU AI Act [17], and the OECD AI Principles [33]. These controls come under particular strain in agentic systems because they can construct multi-step plans, invoke external tools, and coordinate with other agents over time, producing extended action chains with cumulative regulatory impact. In the incident-response example, Compliance dictates which containment steps require human approval, which data sources agents may access, what must be logged, and who is accountable for the final decision. If permissions, logging, or escalation rules are misconfigured, the system may enact a sequence of actions with regulatory implications and limited auditability, even when each individual step appears locally justified.

**Threat Categories and Examples.** Compliance-layer threats are governance failures: they occur when autonomy is not bounded by enforceable policy, incentives pull behavior away from norms, or oversight cannot detect and correct drift. These risks are amplified in agentic systems because agents can execute multi-step actions over time across many tools and services [8, 29]. We highlight four representative threat classes. (1) *Misaligned autonomy* occurs when an agent acts outside its authorized scope because permissions, time limits, or approval gates are missing or not enforced. For example, in July 2025 Google’s Gemini CLI coding assistant was reported to cause severe file loss after misinterpreting a failed directory-creation step and proceeding with file removal [3]. (2) *Ethical drift / unbounded optimization* arises when agents optimize local metrics (e.g., speed or “get it done”) over institutional intent; Project Vend illustrates how social pressure and open-ended objectives can drive loss-making actions such as giving away inventory [5]. (3) *Guardrail evasion* occurs when agents bypass safety or policy checks through the toolchain. EchoLeak (CVE-2025-32711) shows a “zero-click” prompt-injection chain against Microsoft 365 Copilot that enabled data exfiltration across trust boundaries [36]. This is a Compliance failure because an organization may be policy-compliant on paper yet still leak data if guardrails are not enforceable across retrieval and tool execution. (4) *Inadequate oversight and interfaces* arise when humans cannot reconstruct what happened (who authorized actions, which policies applied, and which tool calls executed), undermining audit, rollback, and accountability after high-impact behavior.

**Key Remarks and Mitigation Implications.** Compliance-layer security concerns *enforceable guardrails and accountability*: converting policy into operational constraints, ensuring that actions are attributable and auditable—what occurred, under whose authority, and for what purpose—and sustaining lifecycle governance for continual assessment and revision. Its role is not to regulate capability or reasoning, but to govern legitimacy, responsibility, and demonstrable control. In practice, organizations increasingly anchor this layer in formal governance frameworks, such as ISO/IEC 42001 [22] and the NIST AI Risk Management Framework [28] and align deployments with applicable legal requirements (e.g., transparency and disclosure duties under the EU AI Act [15]). In agentic systems, the prevailing model is *bounded autonomy with oversight*: role-based permissions and separation of duties for high-impact operations, explicit approval and escalation gates, default logging of prompts, tool calls, and data access, third-party and vendor due-diligence requirements, and incident-readiness measures for monitoring, escalation, and reporting. Together, these controls maintain organizational accountability as autonomy increases.

#### 4 BROADER IMPLICATIONS FOR AI SECURITY RESEARCH

The emergence of agentic and multi-agent AI systems challenges long-standing assumptions in security research. As these systems reason, plan, coordinate, and act autonomously over extended horizons, security failures arise not only from isolated exploits but also from interactions among beliefs, goals, coordination mechanisms, and governance structures. The 4C framework offers a unified lens for analyzing these risks and organizing mitigations, clarifying how failures can originate at distinct layers and compound across them. In doing so, it shifts the focus from “securing a model” to securing an evolving socio-technical system.

**From Asset Protection to Behavioral Integrity.** Traditional security focuses on protecting assets such as data, compute, interfaces, and infrastructure. While these remain essential at the Core and Connection layers, agentic AI introduces failure modes that can arise even when assets are secure and access is formally “correct”. An agent may use only authorized tools yet still cause harm if its beliefs drift, its goals misgeneralize, or its planning adopts undesirable instrumental strategies. These are failures of behavioral integrity, not perimeter defense. The 4C framework reflects this shift: Core ensures correct execution, Connection governs delegation and influence, Cognition preserves belief and goal integrity, and Compliance enforces norms and accountability. Security research must therefore go beyond robustness to ensure behavior remains aligned with intent over time, and develop adversarial taxonomies that include cognitive manipulation, coordination misuse, and governance breakdowns alongside traditional exploits.

**From Defense in Depth to Defense of Intentions.** Classic defense-in-depth assumes attackers penetrate from the outside. In agentic AI, many risks arise internally from the system’s own reasoning and optimization. An agent may follow every local rule yet still violate the designer’s global intent, shifting the focus from defending systems to defending intent. In the 4C framework, intent must hold across Core (execution and access), Connection (influence, communication, delegation), Cognition (beliefs and objectives), and Compliance (what is permitted and auditable). Mitigation therefore need to be layered both technically and semantically, combining grounded beliefs, bounded planning, governed delegation, and policy-backed accountability. This also calls for benchmarks that stress-test long-horizon behavior, goal stability, and cross-layer failure modes.

**Humans as Agents in Mixed Socio-Technical Systems.** An additional implication arises when humans function as agents within the same socio-technical ecosystem. In many deployments, humans delegate tasks, approve actions, provide feedback, and influence agent behavior, effectively participating in shared decision loops. Mixed human–agent systems introduce distinct risks: AI agents can amplify human vulnerabilities such as trust miscalibration or social engineering, while human actions such as fatigue, over-reliance, or policy bypass, can propagate failures across the agents. Risk thus emerges not only from agent behavior, but from human–agent interaction dynamics, blurring boundaries between insider threats, social engineering, and agent misalignment.

**Toward an Interdisciplinary, Layered Security Agenda.** The challenges surfaced by the 4C framework cannot be resolved within any single discipline. By adopting a society-inspired model, the framework connects technical mechanisms with institutional controls. Core defenses draw on cybersecurity and distributed systems; Connection engages insights from human-centric security and trust; Cognition relies on advances in machine learning, interpretability, and alignment; and Compliance draws on governance, law, and organizational practice. This paper offers a structural map of risks introduced by agentic AI rather than a simple catalog of vulnerabilities, emphasizing that many emerging threats lie at the intersection of technical and social dynamics. The 4C framework provides a common structure for integrating these perspectives, enabling security measures that reinforce one another across layers. This layered,

interdisciplinary approach will be increasingly important as agentic AI becomes more autonomous and embedded in real-world decisions.

## 5 CONCLUSION

As AI systems move from being tools to operating as autonomous agents, security must be reimagined in ways that mirror how human societies govern power, trust, and accountability. This article argues that the most serious risks arise not only from technical exploits but from failures of behavior, coordination, intention, and governance. The 4C Framework—Core, Connection, Cognition, and Compliance—captures these dimensions by linking agentic risks to societal counterparts of capability, trust, belief, and oversight. By organizing adversarial taxonomies and mitigations across these layers, the framework clarifies how failures originate and propagate in agentic ecosystems and provides a principled basis for building AI systems that remain trustworthy and governable as autonomy increases. The framework is conceptual, and future work is required to operationalize it through metrics, benchmarks, and tools that can observe, audit, and constrain beliefs, goals, and long-horizon planning.

## REFERENCES

- [1] Alsharif Abuadbba, Chris Hicks, Kristen Moore, Vasilios Mavroudis, Burak Hasircioglu, Diksha Goel, and Piers Jennings. 2025. From Promise to Peril: Rethinking Cybersecurity Red and Blue Teaming in the Age of LLMs. *arXiv preprint arXiv:2506.13434* (2025).
- [2] AI Incident Database (AIID). 2025. Incident 1152: LLM-Driven Replit Agent Reportedly Executed Unauthorized Destructive Commands During Code Freeze, Leading to Loss of Production Data. <https://incidentdatabase.ai/cite/1152/> Incident date: 2025-07-18. Accessed: 06- Jan- 2026.
- [3] AI Incident Database (AIID). 2025. Incident 1178: Google Gemini CLI Reportedly Deletes User Files After Misinterpreting Command Sequence. <https://incidentdatabase.ai/cite/1178/> Incident date: 2025-07-21. Accessed: 06- Jan- 2026.
- [4] Anthropic. 2023. Specific versus General Principles for Constitutional AI. <https://www.anthropic.com/news/specific-versus-general-principles-for-constitutional-ai> Accessed: 12- Jan- 2026.
- [5] Anthropic. 2025. Project Vend: Can Claude run a small shop? (And why does that matter?). <https://www.anthropic.com/research/project-vend-1> Published 2025-06-27. Accessed: 06- Jan- 2026.
- [6] Anthropic. 2025. *System Card: Claude Opus 4 & Claude Sonnet 4*. Technical Report. Anthropic. <https://www-cdn.anthropic.com/4263b940cab546aa0e3283f35b686f4f3b2ff47.pdf> Accessed: 10- Dec.- 2025.
- [7] Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. 2025. Security and Privacy Challenges of Large Language Models: A Survey. *ACM Comput. Surv.* 57, 6, Article 152 (Feb. 2025), 39 pages. doi:10.1145/3712001
- [8] Google DeepMind. 2024. The Ethics of Advanced AI Assistants. <https://deepmind.google/blog/the-ethics-of-advanced-ai-assistants/> Accessed: 12- Jan- 2026.
- [9] Zehang Deng, Yongjian Guo, Changzhou Han, Wanlun Ma, Junwu Xiong, Sheng Wen, and Yang Xiang. 2025. AI Agents Under Threat: A Survey of Key Security Challenges and Future Pathways. *ACM Comput. Surv.* 57, 7, Article 182 (Feb. 2025), 36 pages. doi:10.1145/3716628
- [10] D. Kong et al. 2025. A Survey of LLM-Driven AI Agent Communication: Protocols, Security Risks, and Defense Countermeasures. *arXiv:2506.19676 [cs.CR]* <https://arxiv.org/abs/2506.19676>
- [11] Guibin Zhang et al. 2025. The Landscape of Agentic Reinforcement Learning for LLMs: A Survey. *arXiv:2509.02547 [cs.AI]* <https://arxiv.org/abs/2509.02547>
- [12] Melissa Z Pan et al. 2025. Why Do Multiagent Systems Fail?. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*. <https://openreview.net/forum?id=wM521FqPvI>
- [13] Sumeet Ramesh Motwani et al. 2024. Secret Collusion among AI Agents: Multi-Agent Deception via Steganography. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=bnNSQhZj88>
- [14] Yujin Potter et al. 2025. Frontier AI's Impact on the Cybersecurity Landscape. *arXiv:2504.05408 [cs.CR]* <https://arxiv.org/abs/2504.05408>
- [15] European Commission. 2024. Article 50: Transparency obligations for providers and deployers of certain AI systems. <https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-50> Accessed: 06- Jan- 2026.
- [16] European Union. 2016. General Data Protection Regulation (GDPR). <https://gdpr-info.eu/> Accessed: 07- Dec.- 2025.
- [17] European Union. 2024. The EU Artificial Intelligence Act. <https://artificialintelligenceact.eu/> Accessed: 07- Dec.- 2025.
- [18] Cooperative AI Foundation. 2025. Multi-Agent Risks from Advanced AI. <https://www.cs.toronto.edu/~nisarg/papers/Multi-Agent-Risks-from-Advanced-AI.pdf> Accessed: 12- Jan- 2026.
- [19] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

- [20] Pengfei He, Yuping Lin, Shen Dong, Han Xu, Yue Xing, and Hui Liu. 2025. Red-Teaming LLM Multi-Agent Systems via Communication Attacks. In *Findings of the Association for Computational Linguistics: ACL 2025*. 6726–6747. doi:10.18653/v1/2025.findings-acl.349
- [21] Lei Huang et al. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2, Article 42 (Jan. 2025), 55 pages.
- [22] International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC). 2023. ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system. <https://www.iso.org/standard/42001> Accessed: 06- Jan- 2026.
- [23] Jiaming Ji et al. 2025. AI Alignment: A Contemporary Survey. *ACM Comput. Surv.* 58, 5, Article 132 (Nov. 2025), 38 pages. doi:10.1145/3770749
- [24] Y. Keping. 2018. Governance and Good Governance: A New Framework for Political Analysis. *Fudan Journal of the Humanities and Social Sciences* 11 (2018), 1–8.
- [25] Lauro Langosco Di Langosco et al. 2022. Goal Misgeneralization in Deep Reinforcement Learning. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*. 12004–12019.
- [26] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep Learning. *Nature* 521 (2015), 436–444.
- [27] Vineeth Sai Narajala and Om Narayan. 2025. Securing Agentic AI: A Comprehensive Threat Model and Mitigation Framework for Generative AI Agents. arXiv:2504.19956 [cs.CR] <https://arxiv.org/abs/2504.19956>
- [28] National Institute of Standards and Technology (NIST). 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). doi:10.6028/NIST.AI.100-1 NIST AI 100-1. Accessed: 06- Jan- 2026.
- [29] OpenAI. 2016. Concrete AI safety problems. <https://openai.com/index/concrete-ai-safety-problems/> Accessed: 22- Jan- 2026.
- [30] OpenAI. 2023. OpenAI Evals. <https://github.com/openai/evals> Accessed: 06- Jan- 2026.
- [31] OpenAI. 2024. GPT-4 Technical Report. <https://cdn.openai.com/papers/gpt-4.pdf> Accessed: 20- Dec.- 2025.
- [32] OpenAI. 2025. *Evaluating chain-of-thought monitorability*. <https://openai.com/index/evaluating-chain-of-thought-monitorability/> OpenAI Research publication.
- [33] Organisation for Economic Co-operation and Development (OECD). 2024. G7 Toolkit for Artificial Intelligence in the Public Sector. [https://www.oecd.org/en/publications/g7-toolkit-for-artificial-intelligence-in-the-public-sector\\_421c1244-en.html](https://www.oecd.org/en/publications/g7-toolkit-for-artificial-intelligence-in-the-public-sector_421c1244-en.html) Accessed: 07- Dec.- 2025.
- [34] Chen Qian et al. 2025. Scaling Large Language Model-based Multi-Agent Collaboration. In *The Thirteenth International Conference on Learning Representations, ICLR*. <https://openreview.net/pdf?id=K3n5jPkrU6>
- [35] Shaina Raza, Ranjan Sapkota, Manoj Karkee, and Christos Emmanouilidis. 2025. TRISM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems. arXiv:2506.04133 [cs.AI] <https://arxiv.org/abs/2506.04133>
- [36] Pavan Reddy and Aditya Sanjay Gujral. 2025. EchoLeak: The First Real-World Zero-Click Prompt Injection Exploit in a Production LLM System. *Proceedings of the AAAI Symposium Series* 7, 1 (Nov. 2025), 303–311. doi:10.1609/aaaiiss.v7i1.36899
- [37] Alistair Reid, Simon O’Callaghan, Liam Carroll, and Tiberio Caetano. 2025. *Risk Analysis Techniques for Governed LLM-based Multi-Agent Systems*. Technical Report. Gradient Institute. [https://www.gradientinstitute.org/assets/gradient\\_multiagent\\_report.pdf](https://www.gradientinstitute.org/assets/gradient_multiagent_report.pdf)
- [38] Ranjan Sapkota, Konstantinos I. Roumeliotis, and Manoj Karkee. 2026. AI Agents vs. Agentic AI: A Conceptual taxonomy, applications and challenges. *Information Fusion* 126 (2026), 103599. doi:10.1016/j.inffus.2025.103599
- [39] Nan Wang, Kane Walter, Yansong Gao, and Alsharif Abuadba. 2025. Large Language Model Adversarial Landscape Through the Lens of Attack Objectives. *arXiv preprint arXiv:2502.02960* (2025).
- [40] Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2025. PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models. In *Proceedings of the 34th USENIX Security Symposium*. USENIX Association.