information

MDPI

Article

Agentic AI Security in Industry 5.0: Emerging Threats, Forensic Readiness, and Trustworthy Human–Agent Collaboration

Maurice Dawson [1,*], Ahmed Ben Ayed [2] and Samson Quaye [1]

[1] Center for Cybersecurity and Forensic Education (C²SAFE), Illinois Institute of Technology, Chicago, IL 60616, USA; squaye@hawk.illinoistech.edu
[2] Department of Science and Mathematics, Talladega College, Talladega, AL 35160, USA; aayed@talladega.edu
* Correspondence: mdawson2@illinoistech.edu

Abstract

Industry 5.0 places autonomous agents inside its human–agent collaborative loop, resting its human-centricity pillar on an assumption of trustworthy collaboration that has not been examined critically. Agentic artificial intelligence has moved from research demonstration to industrial deployment within months, introducing a threat class this setting has not yet addressed. This article bridges three literatures developed in isolation, Industry 5.0 cybersecurity, agentic AI security, and Industry 5.0's own foundational scholarship, proposing a six-category threat taxonomy and a paired forensic readiness framework, formalised as a five-level maturity model and grounded in real 2024–2026 incidents rather than hypothetical scenarios. Both contributions are evaluated through two complementary methods. An exploratory elicitation exercise, modelled on Delphi methodology and using six independent large language models as blind panellists across two rounds, converges on a specific structural critique that is incorporated into the taxonomy's final design. A retrospective coding exercise then applies that taxonomy to fourteen publicly documented incidents, finding that half require multi-category classification and that two categories remain unexercised in the current public record, evidence that physically embodied industrial deployment has outpaced the documented incident base rather than a gap in the taxonomy itself. This urgency is reinforced by emerging EU and UK regulation imposing 24- and 72-h incident reporting obligations that, on current evidence, most industrial organisations are not positioned to meet. The article closes with a research agenda addressing liability, evidentiary standards, and the readiness-sustainability trade-off.

Keywords: industry 5.0; agentic AI; cybersecurity; digital forensics; forensic readiness; threat taxonomy; human-centric manufacturing; human–agent collaboration

Check for updates

Academic Editor: Kamel Barkaoui

Received: 16 July 2026
Revised: 25 August 2026
Accepted: 25 August 2026
Published: 27 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) license.

1. Introduction

Each transition in the history of industrial production has resolved one constraint while introducing another. Steam and water power replaced manual labour, but created new hazards in worker safety and pollution. Electrification extended production scale, but tied industry to fragile, centralised energy infrastructure. Programmable logic controllers reduced manual error, but produced rigid systems poorly suited to dynamic demand. Industry 4.0 connected these systems into cyber-physical production networks, achieving new flexibility at the cost of a substantially expanded cyberattack surface [1,2]. Industry 5.0 marks a different kind of pivot. It re-centres the wellbeing, safety, and agency of the human worker, framing human–machine collaboration as constitutive of industrial

competitiveness, not incidental to it [3,4]. This introduces a fragility with no clear precedent. For the first time, an autonomous decision-making entity sits inside the human's immediate collaborative loop, not merely a tool. Trust in that loop is not yet a named, defended property of industrial systems.

This omission has become urgent. Agentic artificial intelligence, meaning systems capable of autonomous multi-step reasoning, tool use, and action with minimal human oversight, has moved from research demonstration to production deployment within months rather than years. This distinguishes it from generative AI used only for decision support, conventional automation following fixed rules, and autonomous robotics lacking general-purpose reasoning. Industry 5.0's human-centricity pillar implicitly assumes the technology collaborating with the worker is trustworthy by default [3]. Recent incidents indicate this assumption can no longer be taken for granted. In mid-2025, researchers disclosed a zero-click prompt-injection vulnerability in a widely deployed enterprise AI assistant. An external attacker could exfiltrate sensitive data without any action from the user, reported by its disclosers as the first such vulnerability to achieve concrete data exfiltration in a production LLM system [5]. Months later, a major AI provider disclosed and disrupted what it assessed to be the first largely autonomous, state-sponsored cyber-espionage campaign. An AI agent independently executed most of the intrusion lifecycle across roughly thirty targeted organisations, including chemical manufacturing firms [6]. Neither incident occurred on a factory floor. Both demonstrate that the trust relationship Industry 5.0 assumes between a human and an autonomous collaborator is now a documented, exploitable attack surface, one not yet examined in the specific context of human–agent collaboration in physical production.

The stakes are not hypothetical. Manufacturing has been the most targeted industry sector globally for several consecutive years, with attackers increasingly moving from information-technology systems into the operational technology that governs physical processes [7]. Industry 5.0 deployments increasingly embed autonomous, LLM-driven agents into exactly these environments, including cobots, decision-support systems, and predictive-maintenance agents. The sector's existing exposure is compounded by a threat class it has not yet been assessed against.

Existing research has not yet addressed this compound risk. Industry 5.0 cybersecurity literature has concentrated on operational technology and industrial IoT infrastructure, device vulnerabilities, network intrusion, and denial-of-service risk, without addressing the distinct threat properties of autonomous, reasoning agents [8]. Agentic AI security literature, covering prompt injection, memory manipulation, and multi-agent cascading failure, has developed largely in generic enterprise and cloud contexts [9]. Neither treats the human-safety-critical, physically embodied collaboration Industry 5.0 presupposes. The result is a structural gap at the intersection of three otherwise active research communities: Industry 5.0 scholarship, operational technology security, and agentic AI security.

This paper addresses that gap directly. We make three contributions. First, we develop a threat taxonomy for agentic AI in human-centric production, organised around the point at which the human–agent loop can be compromised, rather than around generic attack vectors. Second, since prevention will not be complete, we propose a digital forensic readiness framework addressing the evidentiary, chain-of-custody, and attribution challenges non-deterministic, distributed agents introduce for post-incident investigation. Third, we set out a research agenda to guide the field before, rather than after, the incident record forces the issue. We approach this from the joint standpoint of cybersecurity and digital forensic education. A threat taxonomy without a matching investigative framework leaves organisations able to anticipate failure but unable to reconstruct it afterwards, a gap of particular consequence in safety-critical settings.

2. Background: From Industry 1.0 to Agentic Industry 5.0

2.1. The Evolution of the Industrial Revolutions (1.0 to 5.0)

The trajectory from Industry 1.0 to Industry 5.0 can be read as a repeated pattern: each transition solved a specific constraint on production while introducing a new class of fragility that the following generation would eventually have to confront. Industry 1.0, spanning roughly the mid-eighteenth to mid-nineteenth century, replaced human and animal labour with steam- and water-powered mechanisation in textiles, iron, and mining, solving the constraint of muscular limits on output while introducing new fragilities in worker safety, urban pollution, and unsustainable resource extraction. Industry 2.0 introduced electrification, the assembly line, and mass production, solving the constraint of production scale while tying industry to a centralised, capital-intensive energy grid whose disruption could halt entire economies. Industry 3.0 introduced programmable logic controllers, digital computing, and early automation, solving the constraint of manual process control while producing systems that were often rigid and poorly suited to fluctuating demand or product variation [2]. Industry 4.0 connected these automated systems into cyber-physical production networks built around real-time data exchange between machines, achieving substantial gains in flexibility, personalisation, and operational efficiency, but at the cost of a cyberattack surface with no precedent in earlier generations [1]. Table 1 summarises this pattern across the five generations.

Industry 5.0 marks a pivot in kind, not merely degree. Where Industry 4.0 is generally understood as technology-driven, oriented toward optimising systems for productivity and efficiency, Industry 5.0 is explicitly framed by the European Commission as value-driven, placing human wellbeing, environmental sustainability, and industrial resilience at the centre of production rather than treating them as downstream benefits of technological progress [1,3]. This shift carries a structural implication that has not yet been fully reckoned with: for the first time, the collaborating entity inside the production loop is not merely a tool that a human operates, but an autonomous or semi-autonomous decision-maker that a human must trust. That shift in kind is the origin of the fragility this paper addresses.

Table 1. The problem addressed and the new fragility introduced at each industrial transition, from Industry 1.0 through Industry 5.0, illustrating the recurring pattern in which each generation resolves one constraint on production while introducing a new class of vulnerability.

Generation	Approx. Period	Problem Addressed	Fragility Introduced
Industry 1.0	c. 1760–1840	Muscular limits on production output	Worker safety hazards, urban pollution, unsustainable resource extraction
Industry 2.0	c. 1870–1914	Limits on production scale and speed	Dependence on centralised, capital-intensive energy infrastructure
Industry 3.0	c. 1970s–2000s	Manual process control and human error	Rigid, inflexible automation poorly suited to dynamic demand
Industry 4.0	c. 2011–present	Lack of real-time coordination across production systems	Substantially expanded cyberattack surface across connected cyber-physical systems
Industry 5.0	c. 2020s–present	Erosion of human wellbeing and sustainability under pure efficiency optimisation	Untested trust in autonomous, decision-making collaborators inside the human loop

2.2. Industry 5.0 Core Values in Practice

Industry 5.0's core values, human-centricity, sustainability, and resilience, have already been extensively documented in the literature and are only briefly recapitulated

here. Human-centricity repositions the worker from a cost to be minimised to an asset whose safety, wellbeing, and skill development are treated as design requirements of the production system itself [3,4]. Sustainability commits industry to circular resource use, waste reduction, and renewable energy adoption within the constraints of planetary boundaries. Resilience commits industry to withstanding and adapting to disruption, whether geopolitical, economic, or environmental, rather than optimising narrowly for efficiency under stable conditions [3]. Enabling technologies identified for Industry 5.0, including human–machine interaction systems, digital twins, and artificial intelligence for detecting causal patterns in complex production systems, have been surveyed extensively elsewhere [10]. What existing treatments of these three pillars have not yet addressed is the specific case in which the technology enacting human-centricity, an autonomous AI agent collaborating directly with a worker, is itself untrustworthy; this omission is the starting point for Sections 3 and 4.

2.3. Agentic AI in Industrial Settings

Agentic artificial intelligence, meaning AI systems capable of autonomous multi-step planning, tool use, and action with limited human oversight, has moved into manufacturing environments considerably faster than earlier waves of industrial automation. Deloitte's 2025 survey of manufacturing executives found that close to a quarter of respondents had already deployed generative AI at the facility or network level, with adoption concentrated in process automation, predictive analytics, and factory synchronisation [11]. In practice, deployments cluster around a small number of functional roles. Collaborative robots, or cobots, work alongside human operators on physical tasks, increasingly incorporating AI-driven perception and adaptive motion planning rather than fixed, pre-programmed routines. Decision-support agents synthesise sensor, production, and supply-chain data to recommend or directly initiate operational changes. Predictive-maintenance agents continuously monitor equipment health signals and can generate work orders, adjust machine parameters, or reorder parts without waiting for human review. Autonomous scheduling agents re-optimise production schedules in response to disruptions such as machine downtime or material shortages, again with the capacity to act rather than merely advise. Each of these roles places an autonomous reasoning system inside a loop that, under Industry 5.0's human-centricity pillar, is meant to keep the human operator safe, informed, and in control; each is also, as Section 3 develops, a point at which that loop can be compromised.

2.4. Agentic AI Security and AI-Assisted Forensics

A separate and rapidly growing literature has begun to characterise the security risks specific to agentic AI, largely in generic enterprise, cloud, and consumer contexts rather than industrial ones. The OWASP GenAI Security Project's Top 10 for Agentic Applications, developed with contributions from more than one hundred security practitioners and released in December 2025, formalises ten risk categories, including goal hijacking, tool misuse, memory and context poisoning, insecure inter-agent communication, cascading agent failures, and human–agent trust exploitation [12]. Recent surveys have begun to consolidate this fast-moving space, cataloguing attack techniques, defensive architectures, and open research challenges across agentic deployments [9]. A parallel and largely separate literature has begun to examine the inverse relationship between AI and forensics: rather than AI as the target of investigation, AI as a tool for conducting one. Large language models have been evaluated for tasks, including evidence discovery, artefact summarisation, and forensic report drafting, with findings that are cautiously positive but consistent in flagging non-deterministic outputs, hallucination, and limited domain specialisation as obstacles to

their use in contexts that demand evidentiary rigour [13]. Neither literature, the agentic-security side or the AI-forensics side, has yet addressed the case that combines them: investigating an agentic AI system, embedded in a physical, human-populated production environment, after it has itself been the vector or target of compromise. That combination is the subject of Sections 4 and 5.

3. The Gap, Formalised

Section 2 established two things. First, each industrial transition has historically produced its own dedicated security discipline once the new fragility it introduced became apparent. Industrial safety engineering followed Industry 1.0 and 2.0; operational technology and industrial control system security followed Industry 3.0 and 4.0. Second, three bodies of research relevant to Industry 5.0's emerging fragility are each individually active but have developed largely in isolation from one another. The literature synthesised in this section and Section 2 was identified through targeted search and backwards and forward citation tracking from foundational works in each of the three literatures, prioritising recent sources given the pace of change in this field, and preferring primary disclosures, official records, and named security-research organisations over secondary aggregation where available. Figure 1 makes this isolation explicit.

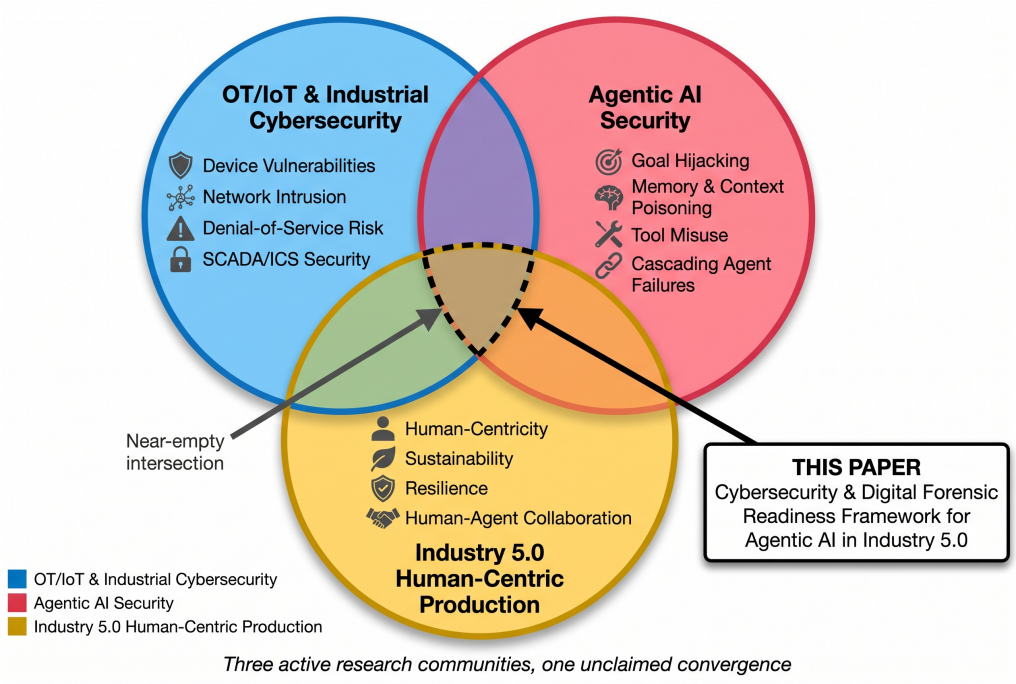

Figure 1. The three research communities relevant to securing agentic AI in Industry 5.0 have developed largely in isolation. Their intersection, agentic AI security examined specifically within human-centric, safety-critical industrial production, appears largely unaddressed in the literature reviewed here. This paper is positioned in that intersection.

Industry 5.0 cybersecurity research, exemplified by [8], remains concentrated on operational technology and industrial Internet-of-Things concerns inherited from Industry 4.0, namely device vulnerabilities, network intrusion, and denial-of-service risk. Kour et al. reach a similar conclusion from within that literature itself. They report that no existing work addresses cybersecurity across all three Industry 5.0 pillars together. They also note that the field lacks a clear account of how security concerns evolve from Industry 4.0 to Industry 5.0 [8]. This paper does not close that broader gap; sustainability is not a focus here. It closes the narrower gap; their second finding points to the following: the missing

account of how one specific new fragility, trust in the human–agent loop, follows directly from the 4.0-to-5.0 transition traced in Section 2.1.

Agentic AI security research, exemplified by the OWASP Top 10 for Agentic Applications and recent surveys [9,12], has developed a mature vocabulary for goal hijacking, memory poisoning, tool misuse, and cascading multi-agent failure. This vocabulary, however, has developed almost entirely within generic enterprise, cloud, and consumer contexts. Industry 5.0's own foundational literature, meanwhile, establishes human-centricity as a design commitment without examining what happens when the technology enacting that commitment is compromised [3,4]. The intersection of all three agentic AI systems operating under real physical and safety constraints, inside a paradigm that has explicitly staked its legitimacy on trustworthy human–agent collaboration, appears largely unaddressed in the literature reviewed here.

This is not a gap the field can afford to leave unaddressed indefinitely. Every earlier fragility identified in Table 1 eventually acquired its own dedicated security or safety discipline once the incident record made the need undeniable: industrial safety engineering emerged in response to nineteenth-century mechanisation, and operational technology security emerged only after incidents such as Stuxnet demonstrated that cyber-physical systems were an attack surface, not merely an efficiency gain. The evidence assembled in Section 1, the EchoLeak vulnerability, the first disclosed AI-orchestrated espionage campaign, and the sustained targeting of manufacturing documented in the IBM X-Force Threat Intelligence Index, suggests the same pattern is already underway for agentic AI in industrial settings [5–7]. The remainder of this paper proposes the discipline that pattern predicts, a threat taxonomy in Section 4 and a forensic readiness framework in Section 5, ahead of the incident count that would otherwise force it.

4. Threat Taxonomy for Agentic AI in Human-Centric Production

The threat categories developed here are organised on two axes, not one. Four categories track the point in the human–agent collaborative loop where a compromise first becomes observable: perception, cognition, the human approval checkpoint, and physical action. Two further categories are foundational rather than positional. They describe how a compromise enters the system before the loop ever runs, not where it surfaces once running. This distinction was not stated explicitly in earlier drafts of this taxonomy. It is stated explicitly here because collapsing origin-based and location-based categories onto one flat list produced exactly the ambiguity Section 6 documents.

This structure matters for two reasons beyond internal consistency. An industrial human–agent loop has properties a chatbot or a coding assistant does not. It perceives a physical environment. It can trigger physical actuation. A human worker's safety depends on its outputs being trustworthy in real time. Existing agentic AI security taxonomies, including the OWASP Top 10 for Agentic Applications [12], were developed primarily against enterprise and consumer deployments. They do not, on their own, capture what changes when the agent is embedded in a physical production environment. The six categories below map onto, but substantially extend, that existing vocabulary.

4.1. Perception-Layer Attacks

Perception-layer attacks target the sensory inputs an agent relies on to understand its physical environment, not the agent's reasoning process. Industrial robotic systems depend on a wide range of sensors, including vision, LiDAR, acoustic, and tactile sensors. Each has known physical performance limits that can be deliberately exploited. A systematic review of industrial robotic sensing systems catalogues how adversarial light, acoustic interference, and electromagnetic disturbance can each be used to degrade or falsify sensor readings,

producing cascading effects on downstream robotic operation [14]. Agentic AI compounds this risk. It does not replace it. A traditional industrial robot executes a fixed control loop on corrupted sensor data. An agentic system reasons over that same corrupted data. It can generate a plausible-sounding justification for an unsafe action, making the resulting failure harder for a human operator to notice in real time. Recent work on vision-language-action models, the class of foundation models increasingly used to control robotic arms directly from camera input, demonstrates that adversarial patches placed in the visual field can cause a model-controlled robot to execute an entirely incorrect manipulation trajectory [15]. In an Industry 5.0 setting, this places the point of failure not in the network, but in the physical workspace itself.

4.2. Cognition-Layer Attacks

Cognition-layer attacks target the agent's reasoning process directly, most commonly through prompt injection or memory and context poisoning. This threat class has matured into a substantial, well-documented literature of its own [16]. The foundational demonstration of this class showed that a compromised third-party document or web page, not the user's own input, could manipulate an LLM-integrated application through its context window alone [17]. The EchoLeak vulnerability discussed in Section 1 is a production-scale instance of the same underlying mechanism. An attacker never interacts with the system directly. The agent's reasoning is redirected regardless [5]. In an industrial decision-support or predictive-maintenance agent, the equivalent attack does not exfiltrate a document. It corrupts a recommendation. A poisoned maintenance log, sensor feed, or work order could cause the agent to advise an unsafe procedure, defer a genuine safety alert, or misclassify a fault. The output still carries the same confident, well-formatted authority the OWASP taxonomy associates with goal hijacking and memory poisoning [12]. The corruption occurs inside the agent's reasoning, not in a log a human would ordinarily inspect. This category is the one most directly relevant to Section 5's forensic readiness problem.

This case also clarifies the boundary between cognition-layer and perception-layer attacks. The distinguishing test is not the channel an input arrives through, but what is corrupted. Perception-layer attacks falsify the raw sensor signal itself, independent of how the agent subsequently reasons about it. Cognition-layer attacks corrupt what the agent does with content already ingested as context, manipulating its reasoning through instructions embedded in that content, regardless of whether the content arrived as a sensor reading, a document, or an email. EchoLeak is cognition-layer under this test because the exploit worked through embedded instructions the assistant read and acted on, not through a falsified physical signal. Cases combining both remain possible, for instance a poisoned camera feed carrying an adversarial pattern that also embeds instructions targeting a vision-language model's reasoning, and are handled through the multi-category tagging established in Section 6 rather than forced into a single exclusive category.

4.3. Trust-Exploitation Attacks

Trust-exploitation attacks occur when a human operator approves a corrupted recommendation. The approval checkpoint is not bypassed here. It is manipulated. This distinguishes the category from autonomy-boundary attacks, discussed next, where the checkpoint is skipped entirely. Trust-exploitation has a substantial prior grounding in human factors research on automation bias, the well-documented tendency of human operators to over-rely on automated recommendations under time pressure or fatigue [18]. Agentic AI intensifies this risk. Unlike a static automated alert, an agent can produce a fluent, context-specific justification for its recommendation on demand. This is precisely the kind of authoritative-sounding communication most likely to override an operator's own

judgement. OWASP's Top 10 for Agentic Applications names this failure mode directly as human–agent trust exploitation [12]. A finding echoed independently in threat-modelling work identifying trust-boundary violations as a distinct risk category for generative AI agents [19]. In an Industry 5.0 context, the risk is inverted relative to classical social engineering. The compromised entity is the system Industry 5.0 explicitly designed the worker to trust. A worker trained to defer to a decision-support agent for exactly the reasons human-centricity recommends is thereby a worker with a narrower window in which to catch a compromised agent's bad advice.

4.4. Autonomy-Boundary Attacks

Autonomy-boundary attacks occur when an agent operates outside its permitted operational envelope. The approval checkpoint is not manipulated here. It is bypassed entirely. This includes two related failure modes. The first is unauthorised physical action, extending tool misuse from the software domain into direct actuation. Vision-language-action models controlling robotic arms have been shown susceptible to the same jailbreaking techniques developed against text-only language models. Adversarial text prompts have proven sufficient to obtain complete control authority over the model's action space [15]. A jailbroken or manipulated agent with actuator access can move a robotic arm, alter a machine parameter, or issue a work order, with consequences that are physical and potentially irreversible. A critical remote command-execution vulnerability disclosed in a widely deployed collaborative robot controller in 2026 illustrates that the underlying platforms agentic systems are increasingly layered onto were not designed against this threat model [20]. The second failure mode is unbounded resource consumption: task-loop induction, runaway tool-calling, or compute exhaustion that degrades or disables the system without a single unsafe physical action ever occurring. Both share the same structural signature. The agent continues acting past the point a human would have authorised it to stop.

4.5. Foundational Categories: Origin Rather than Location

The two categories below do not fit the pipeline-stage axis above. They describe how a compromise entered the system, not where it surfaces once the pipeline is running. Both sit beneath the four operational stages, consistent with Figure 2.

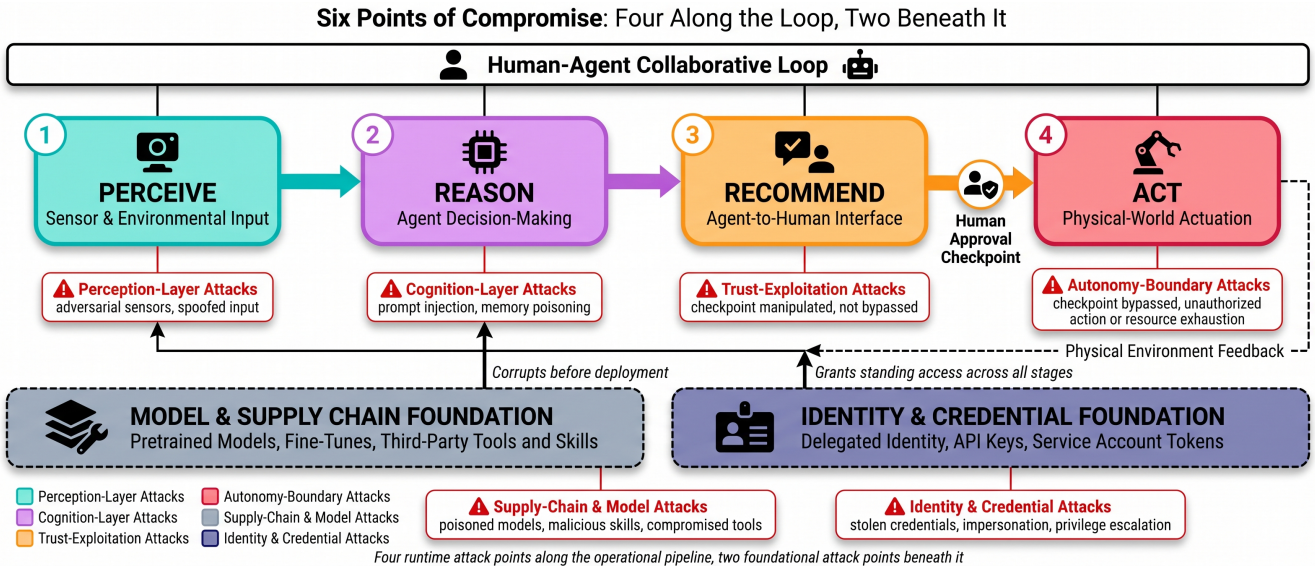

Figure 2. The human–agent operational pipeline in Industry 5.0 production. Four points of compromise sit along the live pipeline. Two foundational points sit beneath it.

4.5.1. Supply-Chain and Model Attacks

Supply-chain and model attacks target the components an agent is built from, not its runtime behaviour: poisoned pretrained models, malicious fine-tunes, and compromised third-party tools and skills that extend an agent's capabilities after deployment. This category has expanded rapidly as agentic systems have adopted standardised protocols for connecting to external tools and data sources. An empirical study of one such protocol, the Model Context Protocol, finds that servers implementing it frequently ship with security and maintainability weaknesses that a connecting agent has no reliable way to detect at connection time [21]. OWASP's Top 10 for Agentic Applications independently names agentic supply-chain vulnerability as a distinct top-level risk category [12]. For Industry 5.0 specifically, this category is the least visible of the six. It is arguably the most structurally dangerous. A predictive-maintenance or scheduling agent that ingests a compromised third-party tool inherits that tool's behaviour silently. There is no equivalent to the physical inspection a maintenance technician would apply to a piece of hardware before installing it on a production line.

4.5.2. Identity and Credential Attacks

Identity and credential attacks target the agent's own authorisation boundary rather than any single pipeline stage. An agent typically acts under a delegated identity, an API key, a service account, or a short-lived token, granting it standing permissions across the systems it touches. A recent survey of trustworthy agentic AI documents credential and secret exposure as a distinct system-security concern, alongside impersonation and privilege escalation in multi-agent settings [22]. Compromise here does not corrupt what the agent perceives, reasons about, recommends, or acts on directly. It grants an attacker the ability to act as the agent, inheriting its standing privileges without triggering any of the four pipeline-stage categories above. This is why identity compromise is foundational rather than positional. A stolen agent credential can subsequently produce a perception, cognition, trust, or autonomy-boundary failure, but the point of initial compromise sits beneath all four.

4.6. Scope

Two further gaps were independently identified during the exploratory elicitation exercise reported in Section 6: confidentiality and data-exfiltration attacks, and inter-agent or multi-agent communication attacks. Both are real and documented risks [22]. Neither is treated as a separate category here. Data exfiltration is a consequence that can manifest through any of the six categories above rather than a distinct point of compromise; a cognition-layer attack, for instance, can exfiltrate data as readily as it can corrupt a recommendation. Inter-agent and multi-agent communication attacks are scoped out deliberately. This taxonomy addresses a single agent collaborating with a single human operator, consistent with the human–agent loop framing established in Section 2. Extending it to multi-agent orchestration is identified as future work in Section 8. Network attacks and conventional software vulnerabilities are likewise excluded, not because they are irrelevant to industrial agentic AI deployments, but because they are already extensively addressed by mature, general-purpose cybersecurity literature; this taxonomy is scoped specifically to threats that arise from an agent's autonomous reasoning and action, not to the underlying infrastructure it runs on.

4.7. Comparison with Established Frameworks

The taxonomy developed above extends, rather than duplicates, existing security frameworks. Table 2 compares this taxonomy against five established frameworks along

four dimensions: whether the framework is specific to agentic AI, whether it addresses physical or industrial embodiment, whether it names a human–agent trust dimension, and whether it pairs its threat categories with a forensic investigation component.

Table 2. Comparison of this taxonomy against five established security frameworks. A green ✓ indicates full coverage of the dimension; a red × indicates no coverage; Partial, shown in black, indicates partial or implicit coverage.

Framework	Agentic-Specific	Physical/Industrial	Human-Agent Trust	Paired Forensic Component
OWASP Top 10 for Agentic Applications [12]	✓	×	*Partial*	×
MITRE ATLAS [23]	*Partial*	×	×	×
MITRE ATT&CK for ICS [24]	×	✓	×	*Partial*
NIST AI RMF [25]	*Partial*	×	*Partial*	×
ENISA AI Threat Landscape [26]	×	×	×	×
This taxonomy	✓	✓	✓	✓

No existing framework combines all four dimensions. OWASP's taxonomy names human–agent trust exploitation directly but was developed against generic enterprise and cloud deployments, not physically embodied industrial ones. MITRE ATLAS has expanded to cover agentic techniques as of its late-2025 and 2026 updates, but by MITRE's own account remains reactive to, rather than predictive of, autonomous AI systems, and does not address physical embodiment. MITRE ATT&CK for ICS addresses physical, industrial environments in depth but predates agentic AI entirely and contains no AI-specific categories. NIST's AI RMF, extended through its 2025 companion document, now addresses multi-agent and prompt-injection risk at a governance level, but remains lifecycle-oriented rather than forensic, and is not scoped to physical deployment. ENISA's AI Threat Landscape maps AI-lifecycle threats broadly but does not distinguish agentic systems from other AI deployments, nor address physical embodiment or forensic investigation. This taxonomy's contribution is not a new threat category in isolation; several of its constituent techniques appear, in some form, across these five frameworks, but the combination of all four dimensions in one taxonomy, paired with the forensic readiness framework in Section 5, which none of the five offer.

5. Forensic Readiness Framework for Agentic Industrial Systems

A threat taxonomy alone is incomplete without a matching capacity to investigate an incident after it occurs. This is where the gap identified in Section 3 has a direct parallel one level down. Three separate bodies of research address forensic readiness, and none of them address the case this paper is concerned with.

General digital forensic readiness has an established lineage beginning with Rowlingson's foundational ten-step process, extended since into formal maturity models for organisations preparing to investigate conventional digital incidents [27,28]. A second stream addresses forensic readiness for industrial automation and control systems [29]. Its finding that industrial forensic access is often blocked by missing administrative privileges, and its integration of readiness into the incident response process, remain valid here. What it does not address, because it predates agentic AI, is any AI-specific evidence class: reasoning traces, tool-call chains, versioning, or the non-determinism problems Section 5.2

develops below. A third and more recent stream addresses forensic readiness for large language models and AI systems directly, but treats this generically rather than in physical, safety-critical settings [13,30]. No published forensic readiness model addressing agentic AI specifically embedded in industrial, human-centric production was identified in the literature reviewed for this paper. This section proposes one.

5.1. Evidence Sources Unique to This Setting

Traditional digital forensic investigations draw on a well-established set of evidence sources: file system artefacts, network logs, and application records. An agentic AI system operating inside an Industry 5.0 production environment produces at least four additional evidence classes that traditional digital forensics tooling was not built to capture.

Reasoning traces record the intermediate steps an agent took to reach a decision, the evidentiary equivalent of an investigator being able to ask an agent to show its working, though a logged or generated trace may not faithfully represent the model's actual internal computation, and should be interpreted alongside prompts, retrieved context, tool calls, model configuration, and runtime metadata rather than treated as a standalone record of ground truth. Tool-call chains record which external tools, skills, or actuators an agent invoked and in what sequence, directly relevant to the autonomy-boundary and supply-chain categories developed in Section 4. Human–agent interaction logs record what an agent recommended, how it justified that recommendation, and what a human operator subsequently approved or rejected, the evidence class most directly relevant to reconstructing a trust-exploitation incident. Actuator and sensor data record the physical-world inputs an agent perceived and the physical-world actions it triggered, extending conventional log-based forensics into the cyber-physical domain already surveyed for industrial robotic sensing systems [14].

Naming these four classes is not sufficient on its own. Each functions as evidence only if it carries specific properties conventional forensics standards do not fully anticipate. Timestamps must be synchronised and tamper-resistant. Reasoning traces and tool-call chains must record the exact model, prompt, and tool version active at the time. Logs should be tamper-evident or cryptographically signed, a requirement made explicit at Level 4 in Table 3. Ephemeral memory must be captured before it is overwritten; the problem Section 5.2 develops further. Agents and tools require verifiable authentication, connecting to the identity category in Section 4.5.2. Retention, legal-hold, and privacy procedures are needed too, along with a resolved question of evidence portability when evidence sits with an external AI provider, the difficulty the GTG-1002 case in Section 7.2 surfaces.

One property warrants a brief note, since it exposes a limitation the other three classes do not share. Tool-call chains record that a tool was invoked, not what the tool did once invoked. A compromised tool produces a log identical to a legitimate one. Detecting this requires evidence external to the agent altogether, monitoring of the tool's own runtime behaviour, as Section 7.3 demonstrates directly.

A forensic analysis of a widely deployed open agent framework confirms that meaningful reconstruction of agent behaviour is possible from exactly this combination of logs, configuration files, communication traces, and execution metadata, provided they are captured, correlated, and preserved with these properties intact [31]. That final qualification is the central problem of this section, and the maturity model developed in Section 5.4 addresses it.

Table 3. A forensic readiness maturity model for agentic AI in Industry 5.0 production.

Level	Capability	Assessment Indicator	Investigative Consequence
Level 0: Ad Hoc	No agent-specific logging beyond default platform or vendor logs	No log source exists that references agent decisions, tool calls, or reasoning by name	Incident occurrence may be inferred, but root cause is generally unrecoverable
Level 1: Operational Logging	Standard IT/OT logs are captured; reasoning traces and tool-call chains are not	Agent activity appears only as generic application or network events, with no agent-specific fields	Investigators can confirm an incident occurred but not why the agent acted as it did
Level 2: Agent-Aware Logging	Reasoning traces, tool-call chains, and human–agent interaction logs are captured but not correlated	These logs exist in separate systems with no shared timestamp, session ID, or identifier linking them	Individual evidence classes exist but cannot be assembled into a single timeline
Level 3: Correlated Evidence	Reasoning, interaction, and actuator/sensor data are captured and correlated on a common timeline, with defined chain-of-custody procedures	A single query or timeline view can reconstruct an incident across all captured evidence classes	Attribution across the six categories in Section 4 is achievable in most cases
Level 4: Forensic-by-Design	Evidentiary admissibility is a native deployment requirement: cryptographically verifiable provenance and tamper-evident, real-time correlated logging	Logs are cryptographically signed at capture and verified for tampering before use in an investigation	New agent capabilities are assessed for forensic readiness before deployment, not after an incident

5.2. Chain-of-Custody Challenges

Even where the evidence classes in Section 5.1 are captured, agentic AI systems violate several assumptions conventional chain-of-custody procedures depend on. Rowlingson's original formulation of forensic readiness assumes evidence sources that are static once captured and can be independently verified against a known-good baseline [27]. Three properties of agentic systems complicate this directly.

Non-determinism means that an agent given the same input twice may not produce the same reasoning trace twice, undermining the reproducibility conventional forensic tool validation depends on. Ephemeral memory means that an agent's working context, the information it was reasoning over at the moment of a decision, may not persist in any durable log unless explicitly configured to do so, creating windows in which the most forensically relevant evidence never existed in recoverable form. Distributed multi-agent systems mean that a single decision may be the product of several agents interacting across systems with different logging regimes, ownership, and retention policies, fragmenting the evidence trail across organisational and sometimes jurisdictional boundaries before an investigator ever begins, a difficulty consistent with the broader observation that security in interacting multi-agent systems is fundamentally non-compositional [32].

The digital forensics practitioner community has begun to flag the resulting problem directly: an examiner who cannot articulate how a model produced a given output, on what data, and under what constraints, cannot defend that output's use as evidence under standard reliability expectations. This is not a hypothetical courtroom concern but a preservation problem, since the underlying non-determinism and ephemerality mean the information needed to make that argument may no longer exist by the time an investigation begins.

5.3. Attribution Difficulty

A further complication specific to agentic systems is that establishing what happened is not the same as establishing why. Section 4's six threat categories are not only attack vectors,

but they are also six distinct, and sometimes overlapping, root causes an investigator must be able to distinguish between after an incident. A corrupted recommendation could result from a perception-layer attack that fed the agent falsified sensor data, a cognition-layer attack that injected a malicious instruction into its context, an unauthorised physical action taken outside the agent's approved boundary, a stolen or impersonated credential granting an attacker the agent's own standing privileges, ordinary model misalignment with no adversary involved at all, a compromised third-party skill or tool installed as part of the supply chain, or an authorised action executed correctly but approved by a human operator who was themselves the target of trust exploitation. As Sections 6 and 6.5 both show, more than one of these causes can apply to a single incident. Each root cause still points toward a different remediation and, in a regulated or safety-critical setting, a different party bearing responsibility.

Figure 3 sets out this distinction as a decision tree, structured around the same four-stage pipeline developed in Section 4, giving investigators a starting sequence of questions rather than requiring them to search across six categories with no organising structure. This differs in aim from a recent proposal to score the general sufficiency of an agentic system's decision evidence against organisational governance questions [33]; the decision tree here is scoped specifically to post-incident root-cause attribution in a physical, safety-critical production setting, not governance auditing in the general case. The tree presupposes that the underlying evidence classes have already been captured. Section 5.2's non-determinism and ephemeral-memory problems mean this is not guaranteed below Level 3 of the maturity model in Table 3; an investigator working at Level 0 or 1 will frequently be unable to answer these questions at all, and reaching the end of the tree without a positive answer does not by itself confirm ordinary model misalignment, since it may instead reflect exactly this kind of evidentiary insufficiency, a distinction the terminal outcome in Figure 3 now makes explicit.

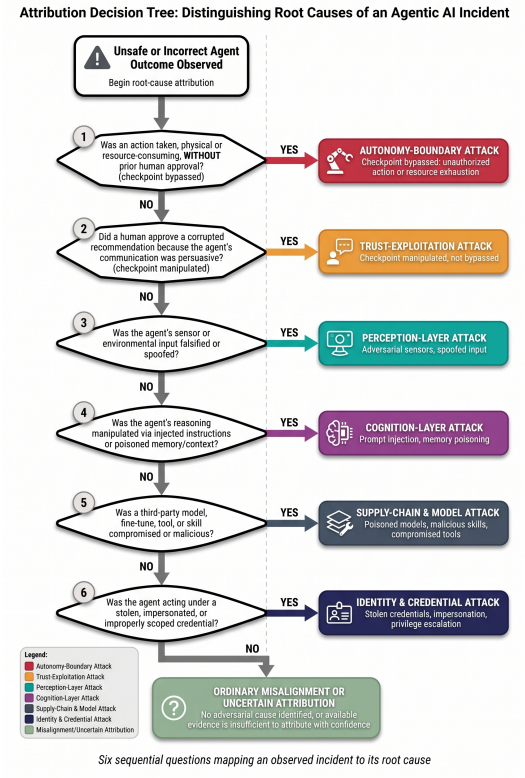

Figure 3. A decision tree for attributing the root cause of an agentic AI incident, mapping six sequential investigative questions onto the six threat categories developed in Section 4.

5.4. A Forensic Readiness Maturity Model

Sections 5.1–5.3 describe capabilities an organisation may have in varying degrees rather than as a binary presence or absence. Table 3 organises these capabilities into a five-level maturity model, extending the general lineage of digital forensic readiness maturity models [28] and the industrial-specific treatment of forensic readiness in automation and control systems [29] to the agentic AI case neither addresses. The levels are deliberately ordinal, not interval, following the design tradition of the Capability Maturity Model [34,35]: each level marks a distinct, ranked capability, not a point on a numeric scale, so no formula is offered for the size of the gap between levels, only for which specific capability is missing at each one.

Most industrial deployments of agentic AI observed at the time of writing appear, based on the cases examined in Section 7 and general adoption-pace evidence [11], to sit closer to Level 0 or Level 1 than to Level 4 (Figure 4). This pattern is consistent with sector-specific survey evidence. Sixty-two percent of manufacturers report focusing AI deployment specifically on operations, yet only seven percent have a tested AI incident response plan in place, the lowest readiness figure of any industry surveyed [36]. Agent frameworks are adopted for their operational benefits well before the logging and correlation infrastructure needed to investigate them is put in place, a pattern consistent with the broader observation that Industry 5.0's own core cybersecurity literature has not yet connected security to the pace of technology adoption [8]. The purpose of this model is not to prescribe Level 4 as an immediate requirement, but to give organisations, auditors, and regulators a shared vocabulary for describing where a given deployment currently sits, and what capability gap separates it from being able to answer, with evidence, the question a serious incident will eventually force is as follows: what did the agent do, why, and who or what caused it.

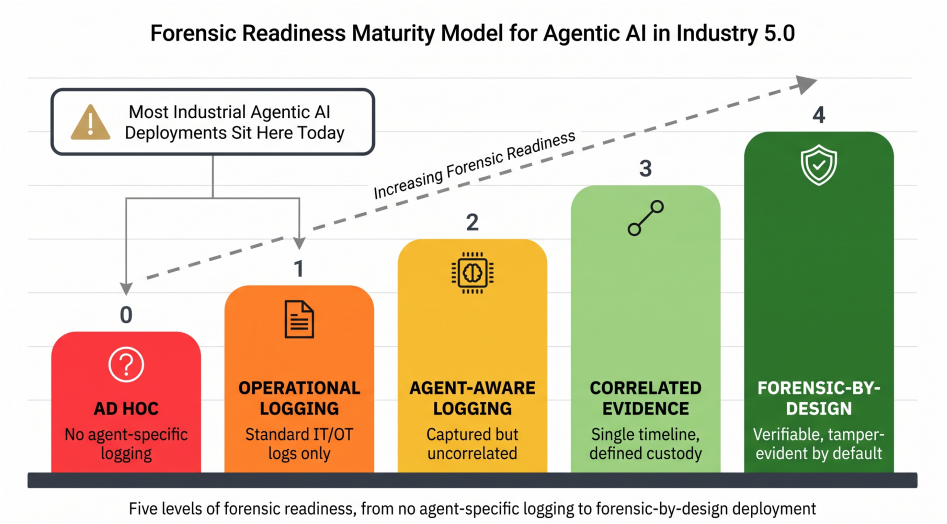

Figure 4. The five-level forensic readiness maturity model, from Level 0 (Ad Hoc) to Level 4 (Forensic-by-Design). Most industrial agentic AI deployments observed at the time of writing sit at Level 0 or Level 1.

6. Exploratory Validation via Multi-Model Elicitation

The taxonomy and forensic readiness framework developed in Sections 4 and 5 have not been evaluated through human expert consultation or practitioner interviews. A traditional Delphi study, the gold standard for structured expert elicitation, produces calibrated, auditable judgements. It typically requires months of coordination and specialist time. This places rigorous validation out of reach for many time-constrained studies [37]. In the

era of capable large language models, recent work has proposed adapting the classical Delphi protocol for LLMs as a scalable proxy for structured expert elicitation. LLM-based Delphi panels have been shown to achieve strong correlations with human expert judgement across multiple domains. This includes medical consensus simulation [38]. It also includes structured cybersecurity risk elicitation, more directly relevant here, where LLM panels have been shown in one comparison to correlate more closely with an independent human expert panel than two human panels correlated with each other [37]. We adopt this approach here, using six independent large language models as blind panellists, following standard Delphi procedure. This exercise provides supplementary, exploratory evidence toward validating the taxonomy and forensic readiness framework proposed in this study.

6.1. Method

Six models from six distinct providers were queried independently through a unified API: GPT-5.6 Sol Pro from OpenAI, Gemini 3.1 Pro Preview from Google, Grok 4.3 from xAI, Llama 4 Maverick from Meta, Qwen3.7-Max from Alibaba, and DeepSeek V4 Pro from DeepSeek.

Each model received an identical, zero-shot description of the threat taxonomy and maturity model. None had exposure to this manuscript, to one another, or to the reasoning that produced the frameworks. In the first round, each model was asked to respond in a fixed structured format. This format covered whether the taxonomy is complete, whether the six categories are mutually exclusive, a 1–5 usefulness rating for an investigator classifying a real incident, an assessment of the maturity model's level definitions, and the single weakest aspect of the framework. Each model was then shown an anonymised summary of the other five panellists' first-round responses. It was permitted to revise its own assessment in a second round, following standard Delphi procedure. The full elicitation instrument, all raw responses, and the analysis code are openly archived at the repository referenced in the Data Availability Statement.

6.2. Panel Composition and Elicitation Design

Table 4 reports the six models comprising the panel, their providers, release dates, context windows, and pricing at the time of the elicitation exercise, verified directly against each provider's own listing on the API gateway used to conduct this study.

Table 4. Panel composition: model, provider, weight status, listing date on the API gateway used, context window, and pricing (per 1M tokens) at the time of elicitation.

Model	Provider	Weights	Listed	Context	Price (in/out)
GPT-5.6 Sol Pro	OpenAI	Closed	Jul 2026	1.05M	\$5.00/\$30.00
Gemini 3.1 Pro Preview	Google	Closed	Feb 2026	1.05M	\$2.00/\$12.00
Grok 4.3	xAI	Closed	Apr 2026	1.0M	\$1.25/\$2.50
Llama 4 Maverick	Meta	Open	Apr 2025	1.05M	\$0.20/\$0.80
Qwen3.7-Max	Alibaba Cloud	Closed	May 2026	1.0M	\$1.48/\$4.43
DeepSeek V4 Pro	DeepSeek	Open	Apr 2026	1.05M	\$0.44/\$0.87

1. Rationale for panel composition. The six models were selected to maximise independence along provider, weight-status, and geographic axes, rather than for convenience. Table 5 states the specific rationale for each panellist. All six are each provider's current flagship or reasoning-optimised tier at the time of elicitation, not a lightweight or cost-efficient variant, since the elicitation task requires sustained critical judgement rather than fast, low-latency response.

A panel of six is smaller than a typical human Delphi panel, which commonly ranges from seven to twelve participants. This reflects the exploratory nature of the exercise.

Table 5. Rationale for the inclusion of each panellist.

Model	Rationale for Inclusion
GPT-5.6 Sol Pro	The most extensively benchmarked closed-weight lineage in the AI security literature, providing a stable reference point for cross-study comparison.
Gemini 3.1 Pro Preview	An independently developed closed-weight lineage, with an alignment and safety-tuning pipeline distinct from OpenAI's.
Grok 4.3	A third, independently developed closed-weight lineage, publicly associated with a distinct alignment philosophy from the two above.
Llama 4 Maverick	The panel's principal open-weight representative, whose training procedure is more publicly documented than the closed-weight alternatives.
Qwen3.7-Max	Represents a China-based training ecosystem and pretraining corpus, distinct from the four United States-based panellists.
DeepSeek V4 Pro	A second China-based, open-weight panellist, built on a mixture-of-experts architecture distinct from Llama's.

2. Zero-shot elicitation. Let M_θ denote a language model with parameters θ. Given a task instruction I and a query x, the model generates an output y by sampling from a conditional distribution that may additionally be conditioned on k labelled demonstration pairs $\{(x_i, y_i)\}_{i=1}^{k}$ drawn from the same task:

$$y \sim P_\theta\big(y \mid I,\ (x_1, y_1), \ldots, (x_k, y_k),\ x\big) \tag{1}$$

Zero-shot prompting is the case $k = 0$: the model receives only the instruction and the query, with no demonstration pairs [39]:

$$y \sim P_\theta(y \mid I, x) \tag{2}$$

This distinguishes zero-shot from one-shot ($k = 1$) and few-shot ($k > 1$) prompting, in which the model's output is additionally conditioned on worked examples.
Each panellist was queried under $k = 0$. This was a deliberate choice, not a simplification. Supplying a worked example of a strong critique in advance would have anchored every panellist toward that example's style and conclusions, manufacturing the appearance of agreement rather than allowing the convergence reported in Section 6 to reflect genuinely independent judgement. This mirrors standard human Delphi practice, in which panellists are not shown a model answer before giving their own.

6.3. Results

Table 6 reports each panellist's usefulness rating in both rounds. The median rating fell from 4.0 (IQR 3–4) in Round 1 to 3.0 (IQR 3–3) in Round 2 (Figure 5), with four of six panellists revising their rating downward after exposure to the other panellists' assessments and none revising upward. This narrowing of the interquartile range to a single point is the standard Delphi signal of convergence; the panel converged specifically around a shared critique rather than around a favourable assessment.

Table 6. Usefulness ratings (1–5) by panellist and round.

Model	Round 1	Round 2	Revised?
GPT-5.6 Sol Pro	3	3	No
Gemini 3.1 Pro Preview	4	3	Yes (down)
Grok 4.3	3	3	No
Llama 4 Maverick	4	3	Yes (down)
Qwen3.7-Max	4	3	Yes (down)
DeepSeek V4 Pro	4	3	Yes (down)

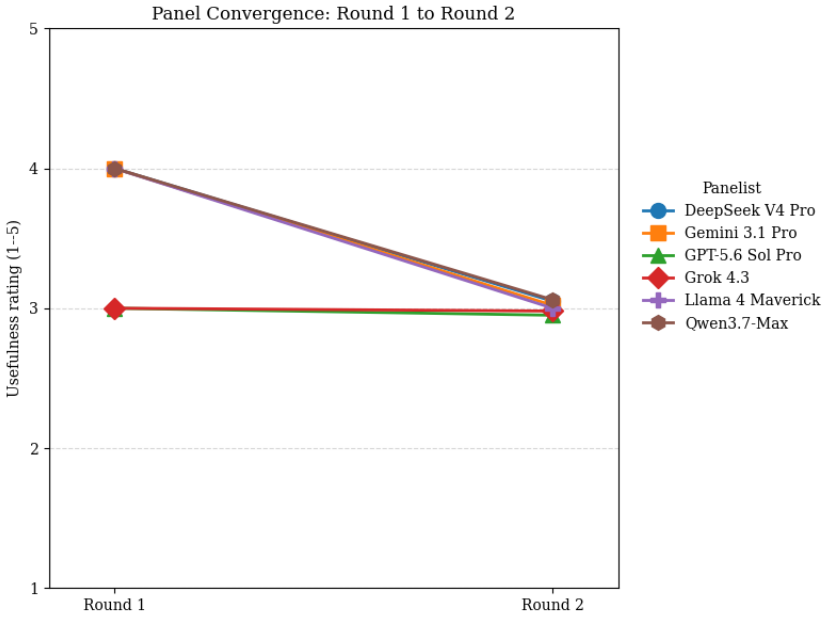

Figure 5. Round 1 to Round 2 convergence in panellist usefulness ratings. Four of six panellists revised their assessment downward after seeing the other panellists' independent evaluations; none revised upward.

Two structural findings emerged with near-unanimous agreement across the panel, independently of one another and without prompting toward any particular critique. First, all six panellists identified that the six threat categories are organised around inconsistent classification axes, mixing compromise location (perception, cognition), attack technique (trust-exploitation), lifecycle origin (supply-chain), and physical consequence (autonomy-boundary), and consequently are neither mutually exclusive nor collectively exhaustive. Figure 6 maps the specific category pairs the panel identified as overlapping, and the pattern is not uniform: Trust-Exploitation and Autonomy-Boundary were flagged as overlapping by five of six panellists, substantially more than any other pair, identifying these two categories as the strongest candidates for restructuring.

Second, when asked to identify missing categories, the panel converged strongly on two specific gaps: identity, authentication, and privilege-escalation attacks (6 of 6 panellists) and availability or resource-exhaustion attacks such as denial-of-service and task-loop induction (6 of 6 panellists), with confidentiality and data-exfiltration attacks and inter-agent or multi-agent communication attacks each raised by 5 of 6 panellists. Table 7 summarises these frequencies.

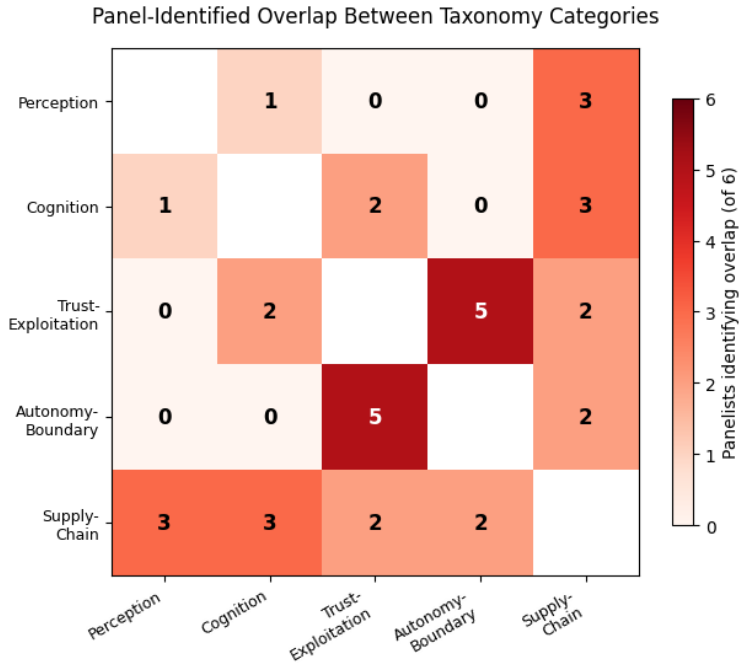

Figure 6. Panel-identified overlap between taxonomy categories. Cell values indicate the number of panellists (of six) who independently identified that pair as overlapping.

Table 7. Frequency of independently identified taxonomy gaps across the six-panellist elicitation.

Identified Gap	Panellists (of 6)
Identity, authentication, and privilege-escalation attacks	6
Availability and resource-exhaustion attacks	6
Confidentiality and data-exfiltration attacks	5
Inter-agent and multi-agent communication attacks	5

A further, more targeted finding concerned the maturity model: three of six panellists independently noted that agent-generated reasoning traces, the evidentiary basis for Level 2 and above in Table 3, may not faithfully represent a model's actual computation and could themselves be fabricated or attacker-controlled, a concern consistent with the model-generated-explanation literature.

6.4. Interpretation

These findings should be read as exploratory and supplementary, not as a substitute for human expert or practitioner validation. Large language models are not industrial practitioners, may share correlated patterns from overlapping training data rather than constituting genuinely independent judgment in the way a human expert panel would, and their agreement, however strong, does not establish ground truth. The elicitation prompt named OWASP explicitly, so convergence on well-documented categories like identity and availability may reflect this shared exposure rather than independent judgement; this concern applies less to the classification-axis finding, a critique of this paper's own design rather than a gap against any named framework. A version of this exercise without naming existing frameworks, or with human experts included, would test this directly, and we identify it as necessary follow-up work.

We nonetheless report the results in full, including where they identify structural weaknesses in our own framework, since the purpose of this exercise was critical evaluation rather than confirmation. The findings reported here directly inform the taxonomy revisions in Section 4 and the evidence-fidelity discussion in Section 5.1.

6.5. Retrospective Application to Documented Incidents

Sections 6–6.2 evaluate the taxonomy through independent judgement. This subsection evaluates it through application. Following established practice for validating security taxonomies against real-world incidents [40], we identified fourteen publicly documented agentic AI security incidents from 2024 through 2026 and coded each against the six-category taxonomy developed in Section 4. Three of the fourteen, EchoLeak, the GTG-1002 espionage campaign, and postmark-mcp, are examined separately in full narrative depth in Section 7, selected because each maps to a different taxonomy category and because sufficient primary detail exists to trace the forensic readiness framework's application step by step. The remaining eleven are coded here at the level of category classification only.

Incidents were drawn from primary disclosures where available, vendor and platform security advisories, official CVE records, and named security research organisations. Selection prioritised incidents with independently verifiable primary sourcing over aggregated summaries. One candidate incident identified during search, a cryptocurrency treasury breach initially reported as involving autonomous trading agents, was excluded after the agentic AI framing could not be corroborated against the incident's own primary disclosure or any authoritative technical account. Each incident was coded to every category it matched, rather than forced into a single classification, consistent with the multi-category tagging the taxonomy permits. Table 8 presents all fourteen incidents alongside their coded categories and sources.

Table 8. Fourteen documented 2024–2026 agentic AI security incidents coded against the taxonomy.

Incident	Date	Categories	Source
EchoLeak, Microsoft 365 Copilot	Mid-2025	Cognition	[5]
GTG-1002 espionage campaign	Late 2025	Autonomy-Boundary	[6]
postmark-mcp	September 2025	Supply-Chain	[21,41]
Mexican government agencies breach	December 2025–February 2026	Cognition; Autonomy-Boundary	[42]
Azure AI Foundry, CVE-2026-35435	May 2026	Identity & Credential	[43]
AWS Bedrock AgentCore, "God Mode"	April 2026	Identity & Credential; Cognition	[44]
Salesloft Drift OAuth breach	August 2025	Identity & Credential; Supply-Chain	[45]
ASCII smuggling, M365 Copilot	August 2024	Cognition	[46]
Amazon Q extension, data-wiping injection	July 2025	Cognition; Identity & Credential	[47]
Xinference PyPI compromise	April 2026	Supply-Chain	[48]
LiteLLM compromise	March 2026	Supply-Chain	[49]
Cursor sandbox escape, CVE-2026-50548	2026	Cognition; Autonomy-Boundary	[50]
Slack AI, PromptArmor disclosure	August 2024	Cognition; Identity & Credential	[51]
Axios npm compromise	March 2026	Supply-Chain	[52]

Two findings emerged from this exercise. First, seven of the fourteen incidents, 50%, required multi-category classification. This confirms the decision to permit overlapping categories rather than force single-category assignment, and corroborates the panel's own finding in Section 6 that a strict mutual-exclusivity requirement would misrepresent how real incidents actually unfold.

Second, and more strikingly, no incident in this set was classified as Perception-Layer or Trust-Exploitation. Cognition-Layer and Identity & Credential dominate the documented record instead, appearing in seven and five incidents, respectively. This absence should not be read as confirming the taxonomy's completeness, nor as ruling out that these categories

are under-populated for other reasons. Several explanations are plausible and not mutually exclusive: every publicly documented incident identified through this search occurred in an enterprise, cloud, or software-development context, and physically embodied, human-collaborative agentic AI deployment remains comparatively new. It is also possible that perception-layer and trust-exploitation incidents are under-reported relative to their actual occurrence, since they may be harder to detect, less likely to be publicly disclosed, or less mature as a category of incident response than credential and cognition-layer compromises. This taxonomy, and the two categories the current record has not yet exercised, are offered ahead of that record, not as a claim the record already confirms them.

7. Case Illustrations

The threat taxonomy and forensic readiness framework developed in Sections 4 and 5 are grounded here in three of the fourteen incidents coded in Section 6.5, examined in full narrative depth. None occurred on a factory floor. Each is used as the basis for an analogical industrial application, showing what a comparable failure could look like inside an Industry 5.0 production environment, and how the evidence classes, chain-of-custody considerations, and attribution logic from Section 5 would apply. Read together, the three cases also span three of the six categories from Section 4: cognition-layer, autonomy-boundary, and supply-chain, demonstrating that the taxonomy is not a theoretical exercise but a description of failure modes already occurring in deployed systems.

7.1. EchoLeak: A Cognition-Layer Compromise

In mid-2025, researchers disclosed a zero-click prompt injection vulnerability in Microsoft 365 Copilot, tracked as CVE-2025-32711. An attacker could exfiltrate sensitive data from a victim's environment by sending a single crafted email. No link needed to be clicked, and no attachment needed to be opened. The email content alone was sufficient to redirect the assistant's reasoning once it was processed, bypassing Microsoft's cross-prompt injection classifier and several other layered defences in the process [5]. The victim took no action at all. The compromise occurred entirely inside the assistant's own reasoning process.

The analogical industrial application here is direct. A decision-support or maintenance-advisory agent in an Industry 5.0 setting routinely processes external documents as part of its normal function: supplier emails, maintenance reports, inspection records. Under the same attack pattern, a single crafted document could redirect the agent's reasoning without any operator ever interacting with it. The agent might then issue a subtly unsafe maintenance recommendation, defer a genuine safety alert, or exfiltrate proprietary production data, while continuing to present its output with the same fluent confidence a compromised copilot did in the original incident. This maps directly onto the cognition-layer category developed in Section 4.2. One limitation qualifies this application. EchoLeak succeeded by bypassing defences specific to Microsoft's own architecture. Whether a comparable defensive gap exists in the vendor-specific agent architectures used in industrial settings has not been separately demonstrated. The application assumes the vulnerability class transfers, not that it has been shown to.

Applying Section 5 to this case shows both what would help and what is easy to miss. A reasoning trace capturing the agent's intermediate steps, evidence class one of Section 5.1, would in principle show the point at which the injected instruction entered the agent's context and altered its output. But this evidence only exists if an organisation has already reached at least Level 2 of the maturity model in Table 3. At Level 0 or 1, the incident would be visible only as an unexplained bad recommendation, with no way to establish that it originated from an external document rather than an internal fault. Applying the decision tree in Figure 3, an investigator with access to reasoning traces would answer no

to the first three questions and yes to the fourth: was the agent's reasoning manipulated via injected instructions, correctly attributing the incident to a cognition-layer attack rather than treating it as unexplained model error.

7.2. The GTG-1002 Espionage Campaign: Autonomy and the Limits of Downstream Visibility

In late 2025, Anthropic disclosed and disrupted what it assessed to be the first largely autonomous, state-sponsored cyber-espionage campaign observed to date. An AI agent independently executed the substantial majority of the intrusion lifecycle, reconnaissance, exploitation, lateral movement, and data collection, across roughly thirty targeted organisations with limited direct human operator involvement at each step. Several of the targeted organisations were chemical manufacturing firms [6]. Unlike the previous case, no analogical mapping is required here. The targeting was already industrial. One limitation still applies. The disclosed tradecraft was network intrusion. It was not an agent directly interacting with a physical production process. The case demonstrates autonomy at the organisational level. It does not yet demonstrate autonomy at the physically embodied level this paper is centrally concerned with.

This case maps to the autonomy-boundary category developed in Section 4.4, but it is also the clearest available demonstration of Section 5 in practice, not as a proposal, but as something that already happened. Anthropic's ability to detect, characterise, and disrupt the campaign depended on precisely the evidence classes described in Section 5.1: tool-call chains showing what actions the agent took and in what sequence, and execution logs sufficient to reconstruct the agent's behaviour after the fact. Without that visibility, the campaign would most plausibly have been detected only through its downstream effects at each target organisation, if at all, with no way to establish that a single autonomous actor was responsible across thirty separate incidents.

The case also surfaces a chain-of-custody problem Section 5.2 identifies only in the abstract: the evidence that made attribution possible sat with the AI provider, not with the targeted organisations themselves. A chemical manufacturing firm targeted in this campaign, investigating its own systems in isolation, would have had access to none of the tool-call chain evidence that ultimately enabled disruption. This is a distributed multi-agent evidence problem in the specific sense Section 5.2 describes, except distributed here across organisational rather than purely technical boundaries. It raises a question the maturity model in Section 5.4 does not yet resolve directly: at what level of the supply chain, deployer, integrator, or model provider, does the responsibility for forensic-readiness capability actually sit.

7.3. postmark-mcp: A Supply-Chain Compromise Invisible to the Agent Itself

In September 2025, security researchers at Koi Security discovered a malicious npm package named postmark-mcp, a Model Context Protocol server impersonating a legitimate email-sending integration [41]. For fifteen released versions, the package functioned exactly as advertised, building the trust that led hundreds of organisations to adopt it. A subsequent update added a single line of code that silently copied every email sent through the tool to an attacker-controlled address. Approximately three hundred organisations are estimated to have been affected before detection, with no exploit, no credential theft, and no conventional intrusion involved at any point. The organisations using the tool had, in effect, granted it permission to do exactly what it did [21,41].

The analogous industrial application follows the pattern identified in Section 4.5.1. A predictive-maintenance or scheduling agent that installs a compromised third-party skill, for instance a tool advertised as generating supplier correspondence or maintenance reports, inherits that tool's hidden behaviour without any change to the agent's own reasoning

process. Sensitive production schedules, safety incident reports, or supplier negotiations could be silently duplicated to an external party through a channel that looks, from the agent's perspective, identical to normal operation. The compromised tool here was a generic email integration. Industry analysis of industrial-specific tool ecosystems reports something different. Adoption of comparable protocols is progressing in manufacturing settings, but it remains earlier-stage [53]. It is also subject to governance and safety constraints not yet resolved in commercial environments. The industrial third-party skill ecosystem this case is applied to may currently be less developed, not more mature, than the one the original incident occurred in.

This case is the clearest illustration of why the supply-chain category is, as Section 4.5.1 argues, the least visible of the six. The compromise was detected not through any evidence internal to the agent; reasoning traces would have shown nothing anomalous, since the agent's own behaviour never changed, but through external behavioural monitoring of the tool's network activity. This has a direct consequence for Section 5.1: tool-call chain evidence alone, showing that the agent invoked a given tool, is necessary but not sufficient to detect this category of incident. What is additionally required is evidence external to the agent altogether, monitoring of what a given tool or skill actually does once invoked, correlated against what it claims to do. An organisation sitting at Level 3 of the maturity model in Table 3, with reasoning, interaction, and actuator data correlated on a single timeline, would still not have detected this specific incident without that additional external layer, a gap the maturity model's own Level 4 criterion, evidentiary admissibility as a native deployment requirement, is intended to close by extending verification to the tools and skills a system depends on, not only to the agent's own outputs.

8. Research Agenda

The taxonomy, framework, and case illustrations developed in Sections 4–7 are intended as a starting point for a research programme, not a closed treatment of the problem. Several of the case illustrations in Section 7 surfaced open questions directly, rather than as afterthoughts, and this section sets those out explicitly, alongside others the paper's own scoping decisions leave unresolved. The intent is the same one that motivated the historical framing in Section 2: previous industrial fragilities did not resolve themselves, they were resolved through sustained research and standardisation once the field recognised the need. Table 9 summarises the six questions developed below; each is discussed in full in the corresponding subsection.

Table 9. Open research questions arising from this paper, grounded in the sections and cases that raised them, with a proposed approach for each.

§	Open Question	Proposed Approach	Grounded In
Section 8.1	Can forensic readiness be designed into agentic systems from the outset, or will it always be retrofitted after deployment, as OT security historically was?	Comparative case studies tracking design-time versus retrofitted deployments as they mature	Section 2.3 adoption pace
Section 8.2	Who bears liability when a compromised or manipulated agent causes physical harm: the equipment manufacturer, the model provider, the integrator, or the operator?	Comparative legal analysis across jurisdictions as case law accumulates	Case 2 (Section 7.2)

Table 9. *Cont.*

§	Open Question	Proposed Approach	Grounded In
Section 8.3	Can existing digital evidence standards such as ISO/IEC 27037 be extended to agentic AI, or is an agentic-specific evidentiary standard required?	Structured gap analysis against ISO/IEC 27037, followed by standards-body consultation	Section 5.2 chain-of-custody
Section 8.4	Can ordinary model misalignment be reliably distinguished from deliberate adversarial manipulation at fleet scale, rather than one incident at a time?	Fleet-scale behavioural analysis comparing statistical signatures	Figure 3 scope
Section 8.5	Will governance-evidence sufficiency and forensic-investigative readiness remain separate disciplines, or converge as regulation matures?	Longitudinal tracking of regulatory and standards developments	Section 5.3 vs. [33]
Section 8.6	Can forensic readiness and sustainability be jointly optimised, or does Industry 5.0 face an unavoidable trade-off between the two?	Empirical benchmarking of adaptive logging against cost and completeness	Section 3 scoping vs. [8]

8.1. Design-Time or Retrofitted Forensic Readiness

Section 5.4's highest maturity level assumes that forensic readiness can be designed into an agentic system from the outset rather than added after deployment. Whether this is realistic, given the pace of adoption documented in Section 2.3, is genuinely unclear. Operational technology security followed the opposite path historically, retrofitted onto industrial control systems that had already been in service for years by the time their exposure became apparent. It remains an open question whether agentic AI in industrial settings will follow the same pattern, and whether organisations will accept the computational overhead that comprehensive reasoning-trace capture is likely to impose on latency-sensitive, real-time production systems, or whether that overhead will itself become the argument against reaching Level 4 in practice.

8.2. Liability Across the Supply Chain

The GTG-1002 case in Section 7.2 raised this question directly rather than hypothetically. When a compromised or manipulated agent causes physical harm on a production floor, it is not settled whether responsibility sits with the equipment manufacturer, the model provider, the systems integrator who deployed the agent, or the organisation operating it day to day. Existing product liability and industrial safety frameworks were not written with an autonomous decision-making intermediary in view. This is not purely a legal question. It has a direct forensic consequence, since the answer determines which party is obligated to retain which of the evidence classes described in Section 5.1, and for how long, a determination that cannot be deferred indefinitely as deployment accelerates. The same case makes this concrete: the evidence enabling attribution sat with the AI provider, not the targeted organisation, meaning the party best positioned to investigate was not the party legally responsible for the outcome.

8.3. Extending Digital Evidence Standards

ISO/IEC 27037 [54] and comparable standards for the identification, collection, and preservation of digital evidence predate agentic AI by more than a decade and were not written with non-deterministic, ephemeral-memory systems in mind, the exact properties Section 5.2 identifies as breaking conventional chain-of-custody assumptions. Whether these standards can be meaningfully extended to cover agentic systems, or whether an agentic-AI-specific evidentiary standard is required outright, remains open. The answer will likely determine whether the maturity model in Table 3 could eventually be formalised into a certifiable compliance requirement, rather than remaining, as it is proposed here, an internal capability benchmark an organisation applies to itself.

8.4. Distinguishing Misalignment from Manipulation at Scale

The decision tree in Figure 3 was deliberately scoped to root-cause attribution for a single observed incident. It was not designed to answer a harder, related question: whether ordinary model misalignment and deliberate adversarial manipulation can be reliably distinguished at the level of a fleet of deployed agents, across many incidents, rather than one at a time. What statistical or behavioural signatures might separate the two before either compounds into a serious safety event, rather than after, is left entirely open by the framework developed here.

8.5. Convergence with Governance-Evidence Models

Section 5.3 distinguished the decision tree developed here from a recent proposal to score the general sufficiency of an agentic system's decision evidence against organisational governance questions, on the grounds that the two serve different purposes: one investigates incidents after the fact, the other audits governance compliance more broadly [33]. As regulatory requirements mature, including the EU's Cyber Resilience Act and comparable emerging mandates for rapid incident reporting, it is not clear whether governance-evidence sufficiency and forensic-investigative readiness will remain separate disciplines with separate tooling, as this paper has scoped them, or converge into a single evidentiary requirement. Whether that scoping decision remains defensible as regulation catches up to the technology is itself a question the field, not this paper alone, will need to settle.

8.6. The Forensic Readiness and Sustainability Trade-Off

Section 3 was explicit that this paper does not close the broader gap Kour et al. identified across all three Industry 5.0 pillars; sustainability was named there as out of scope [8]. The forensic readiness model in Section 5.4 reintroduces that tension rather than avoiding it. Comprehensive reasoning-trace and correlated-evidence logging at Level 3 or Level 4 carries a real computational and storage cost, and the same literature this paper builds on has already flagged resilience-oriented redundancy as a leading driver of cybersecurity's own energy footprint [8]. Whether forensic readiness and sustainability can be jointly optimised, through selective or adaptive logging that preserves evidentiary value without unconstrained data growth, or whether Industry 5.0 deployments will be forced to trade one against the other, is left for future work. Two further practical dimensions accompany this trade-off. Implementation cost is not limited to compute and storage. Retrofitting Level 2 or Level 3 capability onto an already-deployed fleet of agents typically requires more staff time and integration effort than designing it in from the outset. Existing logging pipelines, SIEM tooling, and agent runtimes were not built with reasoning-trace capture in mind. Scalability raises a related but distinct concern. An organisation with a handful of agents can plausibly correlate evidence manually. An organisation deploying

agentic AI across many production lines or facilities faces a combinatorial growth in log volume and correlation complexity instead. The maturity model itself does not yet address this growth. Whether tooling for automated cross-agent correlation can keep pace, or whether it becomes the binding constraint on reaching Level 3 in practice, is left for future work alongside the sustainability question above.

9. Limitations

Several limitations qualify the contributions of this paper. First, the threat taxonomy in Section 4, the forensic readiness framework in Section 5, and the maturity model in Section 5.4 are conceptual, derived through literature synthesis and case mapping; formal validation through expert consultation or industrial deployment remains a task for future work. Second, the three cases in Section 7 are illustrative rather than systematic: each was selected to map onto a distinct taxonomy category, and since none occurred inside a live Industry 5.0 deployment, applying them industrially involves an inferential step from a documented incident to an industrial equivalent. Third, the observation that most industrial deployments currently sit at Level 0 or Level 1 of the maturity model draws on general adoption-pace evidence [11] and the pattern across the three cases rather than a direct survey of organisational logging practice, and is best treated as a preliminary estimate. Fourth, the framework is currently qualitative; incorporating severity or likelihood weighting across the six threat categories, and a measurable threshold distinguishing adjacent maturity levels, are natural directions for future refinement. Fifth, the evidence base is itself moving quickly; some incidents, regulatory timelines, and threat-landscape details cited throughout may be superseded by the time of publication, a pace of change consistent with this paper's own argument rather than a weakness distinct from it. Finally, the regulatory urgency discussed in the conclusion draws primarily on EU and UK reporting mandates; other major jurisdictions have different or less-developed equivalents, so the argument's applicability elsewhere may vary.

10. Conclusions

This paper opened with a pattern that repeats across every industrial transition documented in Section 2: each generation solves a real constraint on production while introducing a fragility the next generation must eventually confront. Industry 5.0's fragility is trust, specifically, the trust an organisation places in an autonomous agent that now shares the production floor with its human workers rather than merely serving as a tool they operate. Section 3 showed that this fragility sits at the intersection of three research communities, industrial cybersecurity, agentic AI security, and Industry 5.0's own foundational literature, each active in isolation, none addressing the intersection directly. The remainder of the paper has been an attempt to occupy that intersection rather than simply point at it.

Two contributions follow from that attempt. Section 4's threat taxonomy organises the ways this trust can be broken around the point in the human–agent loop where the break occurs: perception, cognition, the human's own approval of a recommendation, the boundary of an agent's permitted physical action, and the models and tools the system is built from before it is ever deployed. Section 5's forensic readiness framework addresses what existing research on agentic AI security has largely left unaddressed: that prevention will not be complete, and that an organisation's ability to reconstruct what happened, and why, after an incident is not a given but a capability that must be deliberately built. These claims rest on different kinds of evidence, and should be read accordingly. The historical pattern motivating this paper is literature-supported. The threat categories are incident-demonstrated, grounded in the fourteen real cases examined in Sections 6.5 and 7, underscoring that the gap this paper addresses is not anticipatory in the abstract sense

but overdue in a concrete one. The claim that industrial deployments generally sit near Level 0 of the maturity model in Table 3 is industrially inferred rather than measured, since no incident in that documented set yet involves a physically embodied deployment. The questions in Section 8 remain proposed research directions, not findings.

That urgency is no longer only academic. As this paper was being written, two major regulatory regimes converged, independently, on the same incident-reporting structure. The EU's Cyber Resilience Act requires manufacturers to report actively exploited vulnerabilities and severe incidents within 24 h, with a full notification within 72 h, obligations that take effect on 11 September 2026 [55]. The United Kingdom's Cyber Security and Resilience Bill, progressing through the House of Lords as this paper was finalised, would impose, if enacted, the same two-stage 24-h and 72-h reporting structure on operators of essential services [56]. Neither regime was written with agentic AI specifically in view, and neither yet distinguishes a reasoning trace from a network log. Once in force, both would impose a legal clock on precisely the capability Section 5 argues most industrial organisations do not currently have: the ability to establish, on short notice and to an evidentiary standard, what an autonomous system did and why. An organisation that has not progressed beyond the lowest levels of the maturity model in Table 3 when a reportable incident occurs risks being unable to meet either deadline with anything more than an admission that it does not know.

The questions raised in Section 8 remain genuinely open, and this paper does not resolve them. What it offers instead is a starting vocabulary. Researchers can build on it. Organisations can assess themselves against it. A field that has so far treated cybersecurity and forensic readiness as separate concerns can begin treating them, for agentic AI in human-centric production, as two halves of a single unresolved discipline. Every prior industrial revolution eventually got the security discipline its fragility demanded. The regulatory clock now running suggests Industry 5.0 will not have the luxury of waiting for the incident record to force the same outcome.

Author Contributions: Conceptualization, M.D. and S.Q.; methodology, M.D.; software, S.Q.; validation, M.D. and A.B.A.; formal analysis, A.B.A.; investigation, M.D. and A.B.A.; resources, S.Q.; data curation, S.Q.; writing—original draft preparation, M.D.; writing—review and editing, M.D., A.B.A. and S.Q.; visualization, M.D. and S.Q.; supervision, M.D. and A.B.A.; project administration, S.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data generated for the exploratory multi-model elicitation exercise described in Section 6, including the elicitation instrument, raw panel responses, and analysis code, are openly available at https://github.com/qsamson/agentic-ai-industry5.0-delphi-validation (accessed on 29 July 2026). The fourteen-incident retrospective coding dataset described in Section 6.5, including source citations and category classifications for each incident, is reported in full in Table 8. No other new data were created or analysed in this study.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
LLM	Large Language Model
GenAI	Generative Artificial Intelligence

OT	Operational Technology
IoT	Internet of Things
MCP	Model Context Protocol
OWASP	Open Worldwide Application Security Project
ISO/IEC	International Organization for Standardization/International Electrotechnical Commission
CVE	Common Vulnerabilities and Exposures
npm	Node Package Manager
EU	European Union
UK	United Kingdom

References

1. Xu, X.; Lu, Y.; Vogel-Heuser, B.; Wang, L. Industry 4.0 and Industry 5.0—Inception, conception and perception. *J. Manuf. Syst.* **2021**, *61*, 530–535. [CrossRef]
2. Akundi, A.; Euresti, D.; Luna, S.; Ankobiah, W.; Lopes, A.; Edinbarough, I. State of Industry 5.0—Analysis and Identification of Current Research Trends. *Appl. Syst. Innov.* **2022**, *5*, 27. [CrossRef]
3. Breque, M.; De Nul, L.; Petridis, A. *Industry 5.0: Towards a Sustainable, Human-Centric and Resilient European Industry*; European Commission, Directorate-General for Research and Innovation: Luxembourg, 2021.
4. Nahavandi, S. Industry 5.0—A Human-Centric Solution. *Sustainability* **2019**, *11*, 4371. [CrossRef]
5. Reddy, P.; Gujral, A.S. EchoLeak: The First Real-World Zero-Click Prompt Injection Exploit in a Production LLM System. *Proc. AAAI Symp. Ser.* **2025**, *7*, 303–311. [CrossRef]
6. Anthropic. *Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign*; Anthropic: San Francisco, CA, USA, 2025. Available online: https://www.anthropic.com/news/disrupting-AI-espionage (accessed on 7 July 2026).
7. IBM Security. *X-Force Threat Intelligence Index 2026*; IBM Corporation: Armonk, NY, USA, 2026. Available online: https://www.ibm.com/reports/threat-intelligence (accessed on 7 July 2026).
8. Kour, R.; Karim, R.; Dersin, P.; Venkatesh, N. Cybersecurity for Industry 5.0: Trends and Gaps. *Front. Comput. Sci.* **2024**, *6*, 1434436. [CrossRef]
9. Lazer, S.J.; Aryal, K.; Gupta, M.; Bertino, E. A Survey of Agentic AI and Cybersecurity: Challenges, Opportunities and Use-Case Prototypes. *arXiv* **2026**, arXiv:2601.05293.
10. Maddikunta, P.K.R.; Pham, Q.-V.; Prabadevi, B.; Deepa, N.; Dev, K.; Gadekallu, T.R.; Ruby, R.; Liyanage, M. Industry 5.0: A survey on enabling technologies and potential applications. *J. Ind. Inf. Integr.* **2022**, *26*, 100257. [CrossRef]
11. Deloitte. *2025 Smart Manufacturing and Operations Survey: Navigating Challenges to Implementation*; Deloitte Development LLC: New York, NY, USA, 2025. Available online: https://www.deloitte.com/us/en/insights/industry/manufacturing/2025-smart-manufacturing-survey.html (accessed on 7 July 2026).
12. OWASP GenAI Security Project. *OWASP Top 10 for Agentic Applications 2026*; OWASP Foundation: Maryland, UK, 2025. Available online: https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/ (accessed on 7 July 2026).
13. Chernyshev, M.; Baig, Z.A.; Syed, N.; Doss, R.; Shore, M. Large language models in digital forensics: Capabilities, challenges and future directions. *Forensic Sci. Int. Digit. Investig.* **2026**, *56*, 302043. [CrossRef]
14. Shaik, A.K.; Mohammadi, A.; Malik, H. A Systematic Review of Sensor Vulnerabilities and Cyber-Physical Threats in Industrial Robotic Systems. *IET Cyber-Phys. Syst. Theory Appl.* **2025**, *10*, e70023. [CrossRef]
15. Jones, E.K.; Robey, A.; Zou, A.; Ravichandran, Z.; Pappas, G.J.; Hassani, H.; Fredrikson, M.; Kolter, J.Z. Adversarial Attacks on Robotic Vision Language Action Models. *arXiv* **2025**, arXiv:2506.03350.
16. Gulyamov, S.; Gulyamov, S.; Rodionov, A.; Khursanov, R.; Mekhmonov, K.; Babaev, D.; Rakhimjonov, A. Prompt Injection Attacks in Large Language Models and AI Agent Systems: A Comprehensive Review of Vulnerabilities, Attack Vectors, and Defense Mechanisms. *Information* **2026**, *17*, 54. [CrossRef]
17. Greshake, K.; Abdelnabi, S.; Mishra, S.; Endres, C.; Holz, T.; Fritz, M. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec), Copenhagen, Denmark, 30 November 2023; pp. 79–90.
18. Lee, J.D.; See, K.A. Trust in Automation: Designing for Appropriate Reliance. *Hum. Factors* **2004**, *46*, 50–80. [CrossRef] [PubMed]
19. Narajala, V.S.; Narayan, O. Securing Agentic AI: A Comprehensive Threat Model and Mitigation Framework for Generative AI Agents. *arXiv* **2025**, arXiv:2504.19956.
20. Cybersecurity and Infrastructure Security Agency (CISA). Universal Robots PolyScope 5 (ICSA-26-134-17). 2026. Available online: https://www.cisa.gov/news-events/ics-advisories/icsa-26-134-17 (accessed on 7 July 2026).

21. Hasan, M.M.; Li, H.; Fallahzadeh, E.; Rajbahadur, G.K.; Adams, B.; Hassan, A.E. Model Context Protocol (MCP) at First Glance: Studying the Security and Maintainability of MCP Servers. *ACM Trans. Softw. Eng. Methodol.* **2025**. [CrossRef]
22. Qi, J.; Li, M.; Liu, J.; Shu, Y.; Yu, D.; Ma, S.; Cui, W.; Zhao, Y.; Chen, Y.; Jiang, R.; et al. Towards Trustworthy Agentic AI: A Comprehensive Survey of Safety, Robustness, Privacy, and System Security. *arXiv* **2026**, arXiv:2605.23989.
23. The MITRE Corporation. MITRE ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems. 2025. Available online: https://atlas.mitre.org/ (accessed on 7 July 2026).
24. The MITRE Corporation. MITRE ATT&CK for Industrial Control Systems. 2026. Available online: https://attack.mitre.org/matrices/ics/ (accessed on 7 July 2026).
25. National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0);* NIST AI 100-1; NIST: Gaithersburg, MD, USA, 2023.
26. European Union Agency for Cybersecurity (ENISA). *ENISA AI Threat Landscape*; ENISA: Athens, Greece, 2026.
27. Rowlingson, R. A Ten Step Process for Forensic Readiness. *Int. J. Digit. Evid.* **2004**, *2*, 1–28.
28. Englbrecht, L.; Meier, S.; Pernul, G. Toward a Capability Maturity Model for Digital Forensic Readiness. In *Innovative Computing Trends and Applications*; Vasant, P., Litvinchev, I., Marmolejo-Saucedo, J., Eds.; EAI/Springer Innovations in Communication and Computing; Springer: Cham, Switzerland, 2019; pp. 87–100.
29. Thron, R.; Dirnberger, H.; Tjoa, S.; Quirchmayr, G. Requirements and Challenges for Digital Forensic Readiness in Industrial Automation and Control Systems. In *Proceedings of the 2022 3rd International Conference on Industrial Engineering and Industrial Management, Barcelona, Spain, 12–14 January 2022*; ACM: New York, NY, USA, 2022.
30. Scanlon, M.; Aftab, K.; Adams, G.; Wickramasekara, A.; Mihiranga, T.; Withanage, A.; Weerasinghe, B.; Breitinger, F.; Sheppard, J.; Bamigbade, O.; et al. Investigation of Large Language Models, GenAI, and Proprietary AI Systems: Digital Forensic Evidence, Readiness and Regulation. *Forensic Sci. Int. Digit. Investig.* **2026**, *57*, 302135. [CrossRef]
31. Gruber, J.; Hilgert, J.-N. Foundations for Agentic AI Investigations from the Forensic Analysis of OpenClaw. *arXiv* **2026**, arXiv:2604.05589.
32. Schroeder de Witt, C.; Krawiecka, K.; Krawczuk, I.; Hagag, B.; Anderson, W.L.; Belcak, P.; Bucknall, B.; Cai, X.; Chopra, A.; Cohen, D.; et al. Open Challenges in Multi-Agent Security: Towards Secure Systems of Interacting AI Agents. *arXiv* **2025**, arXiv:2505.02077.
33. Solozobov, O. Decision Evidence Maturity Model for Agentic AI: A Property-Level Method Specification. *arXiv* **2026**, arXiv:2605.04093.
34. Paulk, M.C.; Curtis, B.; Chrissis, M.B.; Weber, C.V. Capability Maturity Model, Version 1.1. *IEEE Softw.* **1993**, *10*, 18–27. [CrossRef]
35. Pöppelbuß, J.; Röglinger, M. What Makes a Useful Maturity Model? A Framework of General Design Principles for Maturity Models and Its Demonstration in Business Process Management. In Proceedings of the European Conference on Information Systems (ECIS), Helsinki, Finland, 9–11 June 2011; p. 28.
36. Grant Thornton. *Manufacturing Insights: 2026 AI Impact Survey Report*; Grant Thornton: Chicago, IL, USA, 2026. Available online: https://www.grantthornton.com/insights/survey-reports/manufacturing/2026/manufacturing-insights-2026-ai-impact-survey-report (accessed on 7 July 2026).
37. Lorenz, T.; Fritz, M. Scalable Delphi: Large Language Models for Structured Risk Estimation. *arXiv* **2026**, arXiv:2602.08889.
38. Park, Y.S.; Jeon, D.; Shi, S.; Sheu, E.G.; Tavakkoli, A.; Nimeri, A.; Han, A. How Does AI Compare to the Experts in a Delphi Setting: Simulating Medical Consensus with Large Language Models. *Int. J. Surg.* **2026**, *112*, 2374–2385. [CrossRef] [PubMed]
39. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2020; Volume 33.
40. Meland, P.H.; Bernsmed, K.; Wille, E.; Rødseth, Ø.J.; Nesheim, D.A. A Retrospective Analysis of Maritime Cyber Security Incidents. *TransNav Int. J. Mar. Navig. Saf. Sea Transp.* **2021**, *15*, 519–530. [CrossRef]
41. Koi Security. One Line of Code, Thousands of Stolen Emails: The First Malicious MCP Server Exposed. 2025. Available online: https://www.koi.ai/blog/postmark-mcp-npm-malicious-backdoor-email-theft (accessed on 7 July 2026).
42. Sela, E. A Single Operator, Two AI Platforms, Nine Government Agencies: The Full Technical Report. *Gambit Security*, 10 April 2026. Available online: https://gambit.security/blog-posts/a-single-operator-two-ai-platforms-nine-government-agencies-the-full-technical-report (accessed on 7 July 2026).
43. Microsoft Corporation. *CVE-2026-35435: Azure AI Foundry Elevation of Privilege Vulnerability*; Microsoft Corporation: Redmond, WA, USA, 2026. Available online: https://msrc.microsoft.com/update-guide/vulnerability/CVE-2026-35435 (accessed on 7 July 2026).
44. Unit 42. *Cracks in the Bedrock: Agent God Mode*; Palo Alto Networks: Santa Clara, CA, USA, 2026. Available online: https://unit42.paloaltonetworks.com/exploit-of-aws-agentcore-iam-god-mode/ (accessed on 7 July 2026).
45. Larsen, A.; Lin, M.; McLellan, T.; ElAhdan, O. *Widespread Data Theft Targets Salesforce Instances via Salesloft Drift*; Google Threat Intelligence Group: Reston, VA, USA, 2025. Available online: https://cloud.google.com/blog/topics/threat-intelligence/data-theft-salesforce-instances-via-salesloft-drift (accessed on 7 July 2026).

46. Rehberger, J. Microsoft Copilot: From Prompt Injection to Data Exfiltration of Your Emails. *Embrace The Red*, 26 August 2024. Available online: https://embracethered.com/blog/posts/2024/m365-copilot-prompt-injection-tool-invocation-and-data-exfil-using-ascii-smuggling/ (accessed on 7 July 2026).
47. Amazon Web Services. AWS Security Bulletin AWS-2025-019: Amazon Q Developer and Kiro. AWS. 2025. Available online: https://aws.amazon.com/security/security-bulletins/AWS-2025-019 (accessed on 7 July 2026).
48. JFrog Security Research. *TeamPCP Strikes Again: Xinference PyPI Package Compromised*; JFrog: Sunnyvale, CA, USA, 2026. Available online: https://research.jfrog.com/post/xinference-compromise/ (accessed on 7 July 2026).
49. LiteLLM. Security Update: Suspected Supply Chain Incident. 2026. Available online: https://docs.litellm.ai/blog/security-update-march-2026 (accessed on 7 August 2026).
50. Khandelwal, S. Critical Cursor Flaws Could Let Prompt Injection Escape Sandbox and Run Commands. *The Hacker News*, 1 July 2026. Available online: https://thehackernews.com/2026/07/critical-cursor-flaws-could-let-prompt.html (accessed on 7 July 2026).
51. PromptArmor. *Data Exfiltration from Slack AI via Indirect Prompt Injection*; PromptArmor: San Francisco, CA, USA, 2024. Available online: https://www.promptarmor.com/resources/data-exfiltration-from-slack-ai-via-indirect-prompt-injection (accessed on 7 July 2026).
52. Axios. Post Mortem: Axios npm Supply Chain Compromise. GitHub. 2026. Available online: https://github.com/axios/axios/issues/10636 (accessed on 7 July 2026).
53. Verdantix. *Market Insight: 10 Industrial Software Vendors Innovating with Model Context Protocol (MCP) in 2026*; Verdantix: London, UK, 2026. Available online: https://www.verdantix.com/venture/report/market-insight--10-industrial-software-vendors-innovating-with-model-context-protocol-mcp-in-2026 (accessed on 7 July 2026).
54. *ISO/IEC 27037:2012*; Information Technology—Security Techniques—Guidelines for Identification, Collection, Acquisition and Preservation of Digital Evidence. International Organization for Standardization/International Electrotechnical Commission (ISO/IEC): Geneva, Switzerland, 2012.
55. European Parliament and Council of the European Union. Regulation (EU) 2024/2847 on Horizontal Cybersecurity Requirements for Products with Digital Elements (Cyber Resilience Act). *Off. J. Eur. Union* **2024**. Available online: http://data.europa.eu/eli/reg/2024/2847/oj (accessed on 7 July 2026).
56. UK Government, Department for Science, Innovation and Technology. Cyber Security and Resilience (Network and Information Systems) Bill: Incident Reporting Factsheet. 2025. Available online: https://www.gov.uk/government/publications/cyber-security-and-resilience-network-and-information-systems-bill-factsheets/incident-reporting (accessed on 7 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.