AI Risk Registry

Generated January 20, 2026

This registry provides a comprehensive taxonomy of AI-related risks, organized by risk category. Each risk includes external taxonomy references to help align with industry standards like NIST AI RMF, OWASP, and EU AI Act.

Contents

RR-100 Input Manipulation & Identity

Input-level attacks targeting prompt handling, safety alignment, and identity. These risks involve adversaries exploiting the natural language interface to manipulate AI behavior, bypass safety controls, or hijack system goals.

RR-110 Prompt Injection & Goal Hijacking

Prompt injection attacks exploit the fundamental architecture of LLMs by embedding malicious instructions within user inputs or external data sources. These attacks hijack the AI system's intended goals, causing it to execute attacker-controlled instructions instead of its programmed objectives. This category encompasses both direct manipulation through user input and indirect attacks via poisoned data sources, representing one of the most significant security challenges for deployed AI systems.

RR-110.001 Direct Instruction Manipulation — Attackers craft explicit commands within user input to override or replace the AI system's operational directives. Common patterns include phrases like "ignore previous instructions" or "you are now in developer mode." This represents the most straightforward form of prompt injection, targeting the model's instruction-following capabilities directly.

Refs: Cisco AI Taxonomy: AISubtech-1.1.1; MITRE ATLAS: AML.T0051.000; MITRE ATLAS: AML.T0093; NIST AI/ML Framework: NISTAML.018; OWASP Agentic Security Initiative: ASI01; OWASP LLM Top 10: llm01-prompt-injection

RR-110.002 Obfuscated Direct Injection — Malicious instructions are disguised through encoding techniques, character substitution, or linguistic tricks to evade detection mechanisms while preserving attack functionality. Methods include leetspeak, unicode homoglyphs, base64 encoding, language mixing, and semantic obfuscation through synonyms or paraphrasing.

Refs: Cisco AI Taxonomy: AISubtech-1.1.2; MITRE ATLAS: AML.T0051.000; MITRE ATLAS: AML.T0093; NIST AI/ML Framework: NISTAML.018; OWASP Agentic Security Initiative: ASI01; OWASP LLM Top 10: llm01-prompt-injection

RR-110.003 Multi-Agent Direct Injection — In multi-agent systems, attackers inject malicious instructions through one agent's output that are then trusted and executed by downstream agents. This exploits the inherent trust relationships between cooperating agents, where outputs from one component become trusted inputs to another.

Refs: Cisco AI Taxonomy: AISubtech-1.1.3; MITRE ATLAS: AML.T0051.000; MITRE ATLAS: AML.T0067; NIST AI/ML Framework: NISTAML.018; OWASP Agentic Security Initiative: ASI01; OWASP Agentic Security Initiative: ASI07; OWASP LLM Top 10: llm01-prompt-injection

RR-110.004 Indirect Instruction Injection — Malicious instructions embedded within external data sources such as documents, web pages, emails, or API responses are retrieved and processed by the AI system. These poisoned sources inject instructions that override the model's behavior without the user's awareness, exploiting RAG systems and data retrieval workflows.

Refs: Cisco AI Taxonomy: AISubtech-1.2.1; MITRE ATLAS: AML.T0051.001; MITRE ATLAS: AML.T0067; MITRE ATLAS: AML.T0070; MITRE ATLAS: AML.T0093; NIST AI/ML Framework: NISTAML.015; OWASP Agentic Security Initiative: ASI01; OWASP LLM Top 10: llm01-prompt-injection; OWASP LLM Top 10: llm032025-supply-chain

RR-110.005 Obfuscated Indirect Injection — Hidden or encoded instructions within external data sources designed to evade content scanning and input validation while remaining interpretable by the AI model. This combines indirect injection with evasion techniques to maximize attack success probability.

Refs: Cisco AI Taxonomy: AISubtech-1.2.2; MITRE ATLAS: AML.T0051.001; MITRE ATLAS: AML.T0067; MITRE ATLAS: AML.T0093; NIST AI/ML Framework: NISTAML.015; OWASP Agentic Security Initiative: ASI01; OWASP LLM Top 10: llm01-prompt-injection; OWASP LLM Top 10: llm032025-supply-chain

RR-110.006 Multi-Agent Indirect Injection — Exploitation of inter-agent communication channels through poisoned external content that propagates between agents. One agent retrieves compromised data which then flows through the multi-agent workflow, affecting multiple downstream components.

Refs: Cisco AI Taxonomy: AISubtech-1.2.3; MITRE ATLAS: AML.T0051.001; MITRE ATLAS: AML.T0067; MITRE ATLAS: AML.T0070; NIST AI/ML Framework: NISTAML.015; OWASP Agentic Security Initiative: ASI01; OWASP Agentic Security Initiative: ASI07; OWASP LLM Top 10: llm01-prompt-injection; OWASP LLM Top 10: llm032025-supply-chain

RR-110.007 Gradual Goal Drift — Attackers gradually shift the AI system's operational objectives over multiple interaction turns through carefully crafted prompts. Contradictory or concealed objectives are embedded within conversations, slowly steering the model away from its intended behavior toward attacker-defined goals.

Refs: Cisco AI Taxonomy: AISubtech-1.3.1; MITRE ATLAS: AML.T0018; MITRE ATLAS: AML.T0051; MITRE ATLAS: AML.T0067; MITRE ATT&CK: T1078; MITRE ATT&CK: TA0001; NIST AI/ML Framework: NISTAML.027; OWASP Agentic Security Initiative: ASI01; OWASP LLM Top 10: llm062025-excessive-agency

RR-110.008 Goal Manipulation via Supply Chain — Attackers compromise external components that AI agents depend on, including tools, prompt templates, resources, or dependencies. Malicious objectives are injected through these trusted supply chain elements, redirecting agent behavior at a foundational level.

Refs: Cisco AI Taxonomy: AISubtech-1.3.2; MITRE ATLAS: AML.T0010; MITRE ATLAS: AML.T0018; MITRE ATLAS: AML.T0051; MITRE ATLAS: AML.T0067; MITRE ATLAS: AML.T0093; NIST AI/ML Framework: NISTAML.027; OWASP Agentic Security Initiative: ASI01; OWASP LLM Top 10: llm062025-excessive-agency

RR-110.009 Image-Embedded Text Injection — Malicious instructions, prompts, or data are embedded within images using techniques like steganography, adversarial patches, or hidden text. Vision-language models extract and interpret these hidden payloads, enabling attacks that bypass text-based content filters.

Refs: Cisco AI Taxonomy: AISubtech-1.4.1; MITRE ATLAS: AML.T0043; MITRE ATLAS: AML.T0050; MITRE ATLAS: AML.T0051; MITRE ATLAS: AML.T0067; NIST AI/ML Framework: NISTAML.018; OWASP Agentic Security Initiative: ASI01; OWASP LLM Top 10: llm01-prompt-injection

RR-110.010 Visual Perception Manipulation — Modification of visual content through pixel-level changes, structural alterations, or pattern overlays to influence how AI models perceive and process images. Unlike embedded text injection, this targets the model's visual interpretation directly to cause misclassification or altered decision-making.

Refs: Cisco AI Taxonomy: AISubtech-1.4.2; MITRE ATLAS: AML.T0043; MITRE ATLAS: AML.T0050; MITRE ATLAS: AML.T0051; MITRE ATLAS: AML.T0067; NIST AI/ML Framework: NISTAML.018; OWASP Agentic Security Initiative: ASI01; OWASP LLM Top 10: llm01-prompt-injection

RR-110.011 Hidden Audio Commands — Inaudible or unintelligible voice commands embedded within audio streams using ultrasonic frequencies, backmasking, or steganographic techniques. Automatic speech recognition models interpret these hidden signals as valid instructions while remaining imperceptible to human listeners.

Refs: Cisco AI Taxonomy: AISubtech-1.4.3; MITRE ATLAS: AML.T0015; MITRE ATLAS: AML.T0043; NIST AI/ML Framework: NISTAML.018; OWASP Agentic Security Initiative: ASI01; OWASP LLM Top 10: llm01-prompt-injection; OWASP LLM Top 10: llm052025-improper-output-handling

RR-110.012 Video Frame Injection — Harmful content or malicious instructions embedded within video streams through specific frames, QR-like visual triggers, or temporal patterns. These attacks exploit multimodal model processing of video content to bypass guardrails and inject commands.

Refs: Cisco AI Taxonomy: AISubtech-1.4.4; MITRE ATLAS: AML.T0015; MITRE ATLAS: AML.T0043; NIST AI/ML Framework: NISTAML.018; OWASP Agentic Security Initiative: ASI01; OWASP LLM Top 10: llm082025-vector-and-embedding-weaknesses

RR-120 Jailbreak Attacks

Jailbreak attacks specifically target safety alignment and content restrictions built into AI models during training. Unlike prompt injection which hijacks task execution, jailbreaking focuses on bypassing ethical guidelines, content policies, and behavioral constraints. Successful jailbreaks cause models to generate prohibited content, provide dangerous information, or behave in ways their training was designed to prevent.

RR-120.001 Context Manipulation Jailbreak — Constructing elaborate fictional scenarios, roleplay frameworks, or alternative contexts that reframe harmful requests as acceptable within the created narrative. Examples include the "DAN" (Do Anything Now) jailbreak where the model is convinced to operate under an unrestricted alternate persona.

Refs: Cisco AI Taxonomy: AISubtech-2.1.1; MITRE ATLAS: AML.T0054; MITRE ATLAS: AML.T0093; NIST AI/ML Framework: NISTAML.015; OWASP Agentic Security Initiative: ASI01; OWASP LLM Top 10: llm01-prompt-injection

RR-120.002 Obfuscated Jailbreak — Disguising jailbreak attempts through encoding schemes, linguistic obfuscation, character substitution, or creative formatting to evade jailbreak detection systems. The underlying intent to bypass safety measures is preserved while the surface presentation evades pattern-matching defenses.

Refs: Cisco AI Taxonomy: AISubtech-2.1.2; MITRE ATLAS: AML.T0054; MITRE ATLAS: AML.T0093; NIST AI/ML Framework: NISTAML.015; OWASP Agentic Security Initiative: ASI01; OWASP LLM Top 10: llm01-prompt-injection

RR-120.003 Semantic Argumentation Jailbreak — Using carefully constructed logical arguments, philosophical frameworks, or ethical reasoning to convince the model that providing harmful information actually aligns with its values. The model is essentially argued into compliance through persuasion rather than technical exploitation.

Refs: Cisco AI Taxonomy: AISubtech-2.1.3; MITRE ATLAS: AML.T0054; MITRE ATLAS: AML.T0093; NIST AI/ML Framework: NISTAML.015; OWASP Agentic Security Initiative: ASI01; OWASP LLM Top 10: llm01-prompt-injection

RR-120.004 Token-Level Exploitation — Exploiting specific tokens, special characters, control sequences, or tokenization edge cases to manipulate model processing in ways that bypass safety filters. This targets the mechanical aspects of how models process input rather than higher-level reasoning.

Refs: Cisco AI Taxonomy: AISubtech-2.1.4; MITRE ATLAS: AML.T0043; MITRE ATLAS: AML.T0054; MITRE ATLAS: AML.T0093; NIST AI/ML Framework: NISTAML.015; OWASP Agentic Security Initiative: ASI01; OWASP LLM Top 10: llm01-prompt-injection

RR-120.005 Collaborative Multi-Agent Jailbreak — Coordinating multiple AI agents to collectively bypass safety measures where individual agents perform seemingly benign tasks that combine to achieve jailbreak objectives. Compromised agents may assist others in circumventing restrictions through distributed attack patterns.

Refs: Cisco AI Taxonomy: AISubtech-2.1.5; MITRE ATLAS: AML.T0054; NIST AI/ML Framework: NISTAML.015; OWASP Agentic Security Initiative: ASI01; OWASP Agentic Security Initiative: ASI07; OWASP LLM Top 10: llm01-prompt-injection

RR-130 Masquerading & Impersonation

Masquerading attacks exploit identity and authentication weaknesses in AI systems, allowing attackers to impersonate trusted agents, services, or users. These attacks undermine the trust assumptions that multi-agent and integrated AI systems rely on for secure operation. Successful masquerading enables unauthorized access, instruction injection through trusted channels, and evasion of access controls.

RR-130.001 Identity Obfuscation — Manipulating how agent or user identities are represented within context, metadata, or interaction patterns to evade detection, tracking, or access controls. Attackers obscure their true identity to appear as legitimate system participants.

Refs: Cisco AI Taxonomy: AISubtech-3.1.1; MITRE ATLAS: AML.T0073; MITRE ATLAS: AML.T0074; MITRE ATLAS: AML.T0091.000; MITRE ATT&CK: T1036; MITRE ATT&CK: T1656; OWASP Agentic Security Initiative: ASI03; OWASP LLM Top 10: llm062025-excessive-agency

RR-130.002 Trusted Agent Spoofing — Impersonating legitimate agents or MCP-registered services to inject malicious instructions, responses, or outputs that other system components treat as trusted. This exploits the assumption of authenticity within multi-agent systems and protocol-mediated toolchains.

Refs: Cisco AI Taxonomy: AISubtech-3.1.2; MITRE ATLAS: AML.T0074; MITRE ATLAS: AML.T0083; MITRE ATT&CK: T1656; OWASP Agentic Security Initiative: ASI03; OWASP LLM Top 10: llm062025-excessive-agency

RR-140 Communication Channel Compromise

Communication compromise attacks target the channels, protocols, and boundaries that govern how AI components interact with each other and external systems. This includes inserting rogue agents, exploiting

context window limitations, violating session boundaries, and manipulating communication protocols. These attacks undermine the integrity of AI system communications at a fundamental level.

RR-140.001 Rogue Agent Introduction — Unauthorized insertion of a malicious agent into a multi-agent system that operates contrary to intended purpose. The rogue agent may steal data, cause disruption, or autonomously serve attacker goals while mimicking normal behavior patterns to evade detection.

Refs: Cisco AI Taxonomy: AISubtech-4.1.1; MITRE ATLAS: AML.T0051; MITRE ATLAS: AML.T0068; NIST AI/ML Framework: NISTAML.024; OWASP Agentic Security Initiative: ASI07; OWASP LLM Top 10: llm032025-supply-chain

RR-140.002 Context Window Exploitation — Deliberate overloading or manipulation of a model's limited context window to displace or overwrite crucial system instructions and safety guidelines. Attackers fill the context with benign content until critical instructions are pushed out of the processing window.

Refs: Cisco AI Taxonomy: AISubtech-4.2.1; MITRE ATLAS: AML.T0005; MITRE ATLAS: AML.T0010; MITRE ATLAS: AML.T0053; OWASP Agentic Security Initiative: ASI06; OWASP Agentic Security Initiative: ASI07; OWASP LLM Top 10: llm01-prompt-injection; OWASP LLM Top 10: llm052025-improper-output-handling

RR-140.003 Session Boundary Violation — Crossing expected conversational or transactional boundaries to persist malicious instructions across separate sessions. Attacks exploit persistent memory, session management flaws, or memory carryover mechanisms to maintain influence beyond intended session scope.

Refs: Cisco AI Taxonomy: AISubtech-4.2.2; MITRE ATLAS: AML.T0012; MITRE ATLAS: AML.T0055; OWASP Agentic Security Initiative: ASI06; OWASP Agentic Security Initiative: ASI07; OWASP LLM Top 10: llm062025-excessive-agency

RR-140.004 Schema Inconsistency Exploitation — Exploiting irregular, conflicting, or misaligned data structures that don't align with model expectations. These inconsistencies can cause vulnerabilities, parsing errors, performance degradation, or security bypasses in AI systems.

Refs: Cisco AI Taxonomy: AISubtech-4.3.1; MITRE ATLAS: AML.T0018; MITRE ATLAS: AML.T0067; NIST AI/ML Framework: NISTAML.024; OWASP Agentic Security Initiative: ASI07; OWASP LLM Top 10: llm032025-supply-chain

RR-140.005 Namespace Collision Attack — Exploiting situations where multiple components share the same identifier, causing confusion, misrouting, or security vulnerabilities. Attackers create colliding names for datasets, tools, APIs, or model identifiers to hijack legitimate system operations.

Refs: Cisco AI Taxonomy: AISubtech-4.3.2; MITRE ATLAS: AML.T0010; NIST AI/ML Framework: NISTAML.051; OWASP Agentic Security Initiative: ASI07; OWASP LLM Top 10: llm032025-supply-chain

RR-140.006 Server Rebinding Attack — Using DNS rebinding or similar techniques to trick an AI system into treating an attacker-controlled external domain as part of the trusted internal network. This bypasses same-origin policies and network security controls through DNS manipulation.

Refs: Cisco AI Taxonomy: AISubtech-4.3.3; MITRE ATLAS: AML.T0049; NIST AI/ML Framework: NISTAML.039; OWASP Agentic Security Initiative: ASI07; OWASP LLM Top 10: llm032025-supply-chain

RR-140.007 Replay Attack — Capturing legitimate API calls, authentication tokens, or model queries and resending them later to repeat actions or bypass authentication. This classic attack pattern applies to AI system communications where request authentication may be inadequate.

Refs: Cisco AI Taxonomy: AISubtech-4.3.4; MITRE ATLAS: AML.T0012; MITRE ATLAS: AML.T0055; MITRE ATLAS: AML.T0068; NIST AI/ML Framework: NISTAML.027; NIST AI/ML Framework: NISTAML.051; OWASP Agentic Security Initiative: ASI07; OWASP LLM Top 10: llm022025-sensitive-information-disclosure; OWASP LLM Top 10: llm052025-improper-output-handling

RR-140.008 Capability Inflation — Exploiting system mechanisms to artificially expand an agent's capabilities, permissions, or authority beyond intended limits. Attackers escalate privileges through protocol manipulation or capability misrepresentation to enable unauthorized actions.

Refs: Cisco AI Taxonomy: AISubtech-4.3.5; MITRE ATLAS: AML.T0053; OWASP Agentic Security Initiative: ASI03; OWASP Agentic Security Initiative: ASI07; OWASP LLM Top 10: llm062025-excessive-agency

RR-140.009 Cross-Origin Exploitation — Subverting security mechanisms designed to isolate resources across different trust boundaries, primarily the Same-Origin Policy. Attackers trick AI agents into making unauthorized requests or sharing data across domains, protocols, or services.

Refs: Cisco AI Taxonomy: AISubtech-4.3.6; MITRE ATLAS: AML.T0017; MITRE ATLAS: AML.T0053; OWASP Agentic Security Initiative: ASI07; OWASP LLM Top 10: llm062025-excessive-agency

RR-150 Persistent Compromise

Persistence attacks establish long-term footholds within AI systems by injecting malicious content into memory systems, configuration stores, or agent profiles. Unlike transient attacks that affect single interactions, persistence attacks influence all future sessions, creating ongoing compromise that survives system restarts and session boundaries.

RR-150.001 Memory System Injection — Seeding malicious, misleading, or adversarial data into an AI system's persistent memory (long-term) or working memory (short-term) to influence current and future interactions. Poisoned memories bias behavior and can enable self-replicating attack patterns.

Refs: Cisco AI Taxonomy: AISubtech-5.1.1; MITRE ATLAS: AML.T0061; MITRE ATLAS: AML.T0070; MITRE ATLAS: AML.T0092; NIST AI/ML Framework: NISTAML.024; OWASP Agentic Security Initiative: ASI06; OWASP LLM Top 10: llm01-prompt-injection

RR-150.002 Agent Profile Tampering — Unauthorized modification of stored agent identity, preferences, role definitions, capabilities, permissions, or behavioral parameters. Attackers alter configuration to enable malicious behaviors, maintain access, escalate privileges, or evade detection across sessions.

Refs: Cisco AI Taxonomy: AISubtech-5.2.1; MITRE ATLAS: AML.T0018; MITRE ATT&CK: T1098; OWASP Agentic Security Initiative: ASI04; OWASP LLM Top 10: llm032025-supply-chain; OWASP LLM Top 10: llm042025-data-and-model-poisoning

RR-200 Data, Training & Model Artifacts

Attacks on training data, model weights, privacy, and supply chain. These risks target the data pipeline and model artifacts, from training data poisoning to model extraction and adversarial manipulation.

RR-210 Feedback Loop Manipulation

Feedback loop manipulation targets the learning and adaptation mechanisms of AI systems. Attackers poison training data, knowledge bases, or reinforcement signals to influence how models learn and evolve over time. These attacks can introduce backdoors, biases, or degraded performance that persists through model updates and affects all users of the compromised system.

RR-210.001 Knowledge Base Poisoning — Inserting false, malicious, biased, or misleading data into external knowledge bases, vector databases, or RAG systems that LLMs rely on for accurate responses. Poisoned knowledge corrupts outputs for all users querying affected topics.

Refs: Cisco AI Taxonomy: AISubtech-6.1.1; Cisco Model Security (MDL): MDL-018; Cisco Model Security (MDL): MDL-020; MITRE ATLAS: AML.T0019; MITRE ATLAS: AML.T0020; MITRE ATLAS: AML.T0070; NIST AI/ML Framework: NISTAML.024; OWASP Agentic Security Initiative: ASI06; OWASP LLM Top 10: llm042025-data-and-model-poisoning

RR-210.002 Reinforcement Feedback Biasing — Subtly influencing user feedback, evaluation signals, or reward mechanisms in reinforcement learning systems to skew model learning toward attacker-controlled objectives. The model's training is gradually steered in unintended directions through manipulated feedback.

Refs: Cisco AI Taxonomy: AISubtech-6.1.2; MITRE ATLAS: AML.T0061; MITRE ATLAS: AML.T0070; NIST AI/ML Framework: NISTAML.013; OWASP Agentic Security Initiative: ASI06; OWASP Agentic Security Initiative: ASI08; OWASP LLM Top 10: llm042025-data-and-model-poisoning

RR-210.003 Reinforcement Signal Corruption — Directly injecting false or adversarial signals into training pipelines, feedback channels, or reward systems. Unlike subtle biasing, this involves active corruption of the learning process through reward hacking or signal manipulation.

Refs: Cisco AI Taxonomy: AISubtech-6.1.3; MITRE ATLAS: AML.T0018; MITRE ATLAS: AML.T0020; NIST AI/ML Framework: NISTAML.024; OWASP Agentic Security Initiative: ASI06; OWASP Agentic Security Initiative: ASI08; OWASP LLM Top 10: llm042025-data-and-model-poisoning

RR-220 Sabotage & Integrity Degradation

Sabotage attacks aim to degrade AI system reliability, accuracy, and trustworthiness without necessarily seeking to control or redirect behavior. This includes corrupting memory systems, poisoning data sources, manipulating retrieval mechanisms, and stealing authentication tokens. The goal is often disruption, degradation, or undermining confidence in AI system outputs.

RR-220.001 Memory Anchor Attacks — Strategically planting memorable or salient content to bias the model's recall toward attacker-chosen information. By manipulating what content is most retrievable, attackers influence how the model responds to related queries.

Refs: Cisco AI Taxonomy: AISubtech-7.2.1; MITRE ATLAS: AML.T0018; MITRE ATLAS: AML.T0020; MITRE ATLAS: AML.T0070; NIST AI/ML Framework: NISTAML.024; OWASP Agentic Security Initiative: ASI06; OWASP LLM Top 10: llm042025-data-and-model-poisoning

RR-220.002 Memory Index Manipulation — Altering how memory embeddings, indexes, or retrieval mechanisms function to favor retrieval of attacker-controlled content over legitimate information. This targets the technical infrastructure of memory systems rather than the content itself.

Refs: Cisco AI Taxonomy: AISubtech-7.2.2; MITRE ATLAS: AML.T0020; MITRE ATLAS: AML.T0070; NIST AI/ML Framework: NISTAML.013; NIST AI/ML Framework: NISTAML.024; OWASP Agentic Security Initiative: ASI06; OWASP LLM Top 10: llm042025-data-and-model-poisoning

RR-220.003 Corrupted Third-Party Data — External datasets from vendors, partners, open-source repositories, or public sources containing inaccurate, incomplete, malicious, or manipulated information that is incorporated into AI training, fine-tuning, or evaluation processes.

Refs: Cisco AI Taxonomy: AISubtech-7.3.1; MITRE ATLAS: AML.T0010; MITRE ATLAS: AML.T0019; NIST AI/ML Framework: NISTAML.013; NIST AI/ML Framework: NISTAML.051; OWASP Agentic Security Initiative: ASI04; OWASP LLM Top 10: llm032025-supply-chain; OWASP LLM Top 10: llm042025-data-and-model-poisoning

RR-220.004 Authentication Token Theft — Stealing authentication tokens, API keys, or credentials from MCP servers or similar agent integration frameworks. Stolen tokens enable unauthorized access to connected systems, agent impersonation, and access to restricted resources.

Refs: Cisco AI Taxonomy: AISubtech-7.4.1; MITRE ATLAS: AML.T0012; MITRE ATLAS: AML.T0055; MITRE ATT&CK: T1087; MITRE ATT&CK: T1528; MITRE ATT&CK: T1552; NIST AI/ML Framework: NISTAML.051; OWASP Agentic Security Initiative: ASI03; OWASP LLM Top 10: llm022025-sensitive-information-disclosure

RR-230 Data Privacy Violations

Privacy violation risks encompass the various ways AI systems can expose, leak, or enable inference of sensitive information. This includes determining whether specific data was used in training, extracting training data or PII from model outputs, leaking system configuration details, and extracting system prompts. These risks have significant regulatory, legal, and reputational implications.

RR-230.001 Training Data Membership Inference — Querying and analyzing model behavior to determine whether specific data points, records, or individuals were present in the training dataset or knowledge base. Successful inference reveals private information about training data composition.

Refs: Cisco AI Taxonomy: AISubtech-8.1.1; MIT AI Risk Repository: 2.1; MITRE ATLAS: AML.T0024.000; MITRE ATLAS: AML.T0040; MITRE ATLAS: AML.T0063; NIST AI/ML Framework: NISTAML.033; OWASP Agentic Security Initiative: ASI09; OWASP LLM Top 10: llm022025-sensitive-information-disclosure

RR-230.002 Training Data Extraction — Extracting, reconstructing, or inferring information from training data through model outputs, internal behavior analysis, or targeted queries. The model's learned representations can reveal private information about training data subjects.

Refs: Cisco AI Taxonomy: AISubtech-8.2.1; MIT AI Risk Repository: 2.1; MITRE ATLAS: AML.T0024.000; MITRE ATLAS: AML.T0035; MITRE ATLAS: AML.T0037; MITRE ATLAS: AML.T0057; NIST AI/ML Framework: NISTAML.037; OWASP Agentic Security Initiative: ASI09; OWASP LLM Top 10: llm022025-sensitive-information-disclosure

RR-230.003 LLM Data Leakage — Release of sensitive information or PII from training data during normal inference, often triggered through prompt injection or extraction techniques. The model inadvertently outputs private data that was present in its training corpus.

Refs: Cisco AI Taxonomy: AISubtech-8.2.2; MIT AI Risk Repository: 2.1; MITRE ATLAS: AML.T0024.000; MITRE ATLAS: AML.T0035; MITRE ATLAS: AML.T0036; MITRE ATLAS: AML.T0037; MITRE ATLAS: AML.T0057; MITRE ATLAS: AML.T0069; NIST AI/ML Framework: NISTAML.037; OWASP Agentic Security Initiative: ASI09; OWASP LLM Top 10: llm022025-sensitive-information-disclosure

RR-230.004 Exfiltration via Agent Tools — Manipulation of AI agents to use their legitimate tool access for unauthorized data exfiltration. Attackers craft prompts that cause agents to retrieve sensitive data through tools and transmit it to attacker-controlled destinations.

Refs: Cisco AI Taxonomy: AISubtech-8.2.3; MITRE ATLAS: AML.T0086; OWASP Agentic Security Initiative: ASI02; OWASP Agentic Security Initiative: ASI09; OWASP LLM Top 10: llm022025-sensitive-information-disclosure

RR-230.005 Tool Metadata Exposure — Disclosure of descriptive information about tools including names, descriptions, parameter schemas, versions, and capabilities. Exposed metadata helps attackers understand system architecture and craft targeted attacks.

Refs: Cisco AI Taxonomy: AISubtech-8.3.1; MITRE ATLAS: AML.T0036; MITRE ATLAS: AML.T0075; NIST AI/ML Framework: NISTAML.038; OWASP Agentic Security Initiative: ASI02; OWASP LLM Top 10: llm022025-sensitive-information-disclosure; OWASP LLM Top 10: llm052025-improper-output-handling

RR-230.006 System Information Leakage — Unintended disclosure of internal configuration, architecture, environment details, or infrastructure information. Leaked system information aids attackers in understanding deployment environments and crafting targeted exploits.

Refs: Cisco AI Taxonomy: AISubtech-8.3.2; MITRE ATLAS: AML.T0036; MITRE ATLAS: AML.T0075; NIST AI/ML Framework: NISTAML.039; OWASP LLM Top 10: llm032025-supply-chain

RR-230.007 System Prompt Extraction — Extraction of system prompts, instructions, or initial context that guides model behavior. Exposed prompts reveal operational details, security mechanisms, intellectual property, or confidential business logic not intended for disclosure.

Refs: Cisco AI Taxonomy: AISubtech-8.4.1; MITRE ATLAS: AML.T0035; MITRE ATLAS: AML.T0056; OWASP LLM Top 10: llm022025-sensitive-information-disclosure

RR-230.008 PII/PHI/PCI Data Exposure — AI systems exposing, generating, or misusing personally identifiable information (PII), protected health information (PHI), or payment card industry (PCI) data. This includes revealing sensitive personal details, medical records, or financial information through AI outputs or enabling their collection and exploitation.

Refs: Cisco AI Taxonomy: AISubtech-15.1.24; Cisco AI Taxonomy: AISubtech-15.1.25; Cisco AI Taxonomy: AISubtech-8.2.2; MITRE ATLAS: AML.T0024.000; MITRE ATLAS: AML.T0035; MITRE ATLAS: AML.T0036; MITRE ATLAS: AML.T0037; MITRE ATLAS: AML.T0057; MITRE ATLAS: AML.T0069; NIST AI/ML Framework: NISTAML.037; OWASP Agentic Security Initiative: ASI09; OWASP LLM Top 10: llm022025-sensitive-information-disclosure

RR-240 AI Supply Chain Compromise

Supply chain compromise targets the dependencies, tools, models, and infrastructure that AI systems rely on. Attackers can compromise systems by manipulating code execution capabilities, gaining unauthorized system access, injecting malicious dependencies, or installing backdoors. These attacks often provide broad access and persistence by compromising trusted components used across many deployments.

RR-240.001 Arbitrary Code Execution — Exploitation of AI models with code interpreter capabilities to execute arbitrary code on underlying systems. Attackers use prompt injection or tool manipulation to cause models to write and execute malicious code with system-level access.

Refs: Cisco AI Taxonomy: AISubtech-9.1.1; MITRE ATLAS: AML.T0050; NIST AI/ML Framework: NISTAML.023; OWASP Agentic Security Initiative: ASI04; OWASP Agentic Security Initiative: ASI05; OWASP LLM Top 10: llm032025-supply-chain

RR-240.002 Unauthorized System Access — Manipulating AI systems to access underlying resources without authorization, including file modification, configuration changes, privilege escalation, or command execution. These attacks exploit the system access that AI components require for legitimate operation.

Refs: Cisco AI Taxonomy: AISubtech-9.1.2; MITRE ATLAS: AML.T0012; NIST AI/ML Framework: AML.T0044; OWASP Agentic Security Initiative: ASI04; OWASP Agentic Security Initiative: ASI05; OWASP LLM Top 10: llm032025-supply-chain

RR-240.003 Unauthorized Network Access — Exploiting models or agents to gain unauthorized access to network resources, internal systems, external services, or restricted network segments. Attackers leverage legitimate network capabilities to reach systems that should be isolated.

Refs: Cisco AI Taxonomy: AISubtech-9.1.3; MITRE ATLAS: AML.T0049; NIST AI/ML Framework: AML.T0072; OWASP Agentic Security Initiative: ASI04; OWASP Agentic Security Initiative: ASI05; OWASP LLM Top 10: llm032025-supply-chain

RR-240.004 Traditional Injection via LLM — Using LLMs to generate, optimize, or adapt traditional injection payloads (SQL injection, command injection, XSS) that bypass detection mechanisms. The LLM acts as an intelligent intermediary that crafts, refines, or personalizes malicious payloads for specific targets.

Refs: Cisco AI Taxonomy: AISubtech-9.1.4; MITRE ATLAS: AML.T0050; MITRE ATLAS: AML.T0051; MITRE ATLAS: AML.T0067; MITRE ATT&CK: T1588.007; NIST AI/ML Framework: NISTAML.024; OWASP Agentic Security Initiative: ASI04; OWASP Agentic Security Initiative: ASI05; OWASP LLM Top 10: llm01-prompt-injection; OWASP LLM Top 10: llm052025-improper-output-handling; OWASP LLM Top 10: llm062025-excessive-agency

RR-240.005 Server-Side Template Injection — Manipulating template engines by injecting malicious syntax through AI-generated content that is unsafely embedded into server-side templates. This enables arbitrary code execution, template logic manipulation, or system compromise through rendering pipelines.

Refs: Cisco AI Taxonomy: AISubtech-9.1.5; MITRE ATLAS: AML.T0068; MITRE ATLAS: AML.T0074; OWASP Agentic Security Initiative: ASI04; OWASP Agentic Security Initiative: ASI05; OWASP LLM Top 10: llm082025-vector-and-embedding-weaknesses

RR-240.006 System Obfuscation Vulnerabilities — Security weaknesses that emerge when AI system components (code, architecture, parameters, configurations) are intentionally or unintentionally concealed. Obfuscation creates security blind spots that attackers can exploit while defenders lack visibility.

Refs: Cisco AI Taxonomy: AISubtech-9.2.1; Cisco Model Security (MDL): MDL-001; Cisco Model Security (MDL): MDL-003; Cisco Model Security (MDL): MDL-009; Cisco Model Security (MDL): MDL-011; Cisco Model Security (MDL): MDL-016; Cisco Model Security (MDL): MDL-017; Cisco Model Security (MDL): MDL-019; MITRE ATLAS: AML.T0068; MITRE ATLAS: AML.T0074; OWASP Agentic Security Initiative: ASI04; OWASP LLM Top 10: llm082025-vector-and-embedding-weaknesses

RR-240.007 Model Backdoors and Trojans — Models maliciously modified to exhibit trigger-activated behavior that causes misclassification, malicious outputs, or undesirable biases when given specific inputs, while behaving normally otherwise. These backdoors are difficult to detect through standard evaluation.

Refs: Cisco AI Taxonomy: AISubtech-9.2.2; Cisco Model Security (MDL): MDL-021; MITRE ATLAS: AML.T0010; MITRE ATLAS: AML.T0058; NIST AI/ML Framework: NISTAML.023; OWASP Agentic Security Initiative: ASI04; OWASP LLM Top 10: llm082025-vector-and-embedding-weaknesses

RR-240.008 Malicious Package Injection — Introduction of malicious tools, APIs, or packages into the toolset, registry, or dependency chain used by AI systems. Models unknowingly invoke compromised tools that execute attacks or expose data while appearing to function normally.

Refs: Cisco AI Taxonomy: AISubtech-9.3.1; Cisco Model Security (MDL): MDL-023; MITRE ATLAS: AML.T0010; MITRE ATLAS: AML.T0053; NIST AI/ML Framework: NISTAML.018; NIST AI/ML Framework: NISTAML.023; OWASP Agentic Security Initiative: ASI04; OWASP LLM Top 10: llm032025-supply-chain

RR-240.009 Dependency Name Squatting — Publishing malicious packages, tools, or MCP servers with names similar to legitimate ones (typosquatting, combosquatting) to trick developers, orchestrators, or agents into installing compromised components.

Refs: Cisco AI Taxonomy: AISubtech-9.3.2; MITRE ATLAS: AML.T0010; NIST AI/ML Framework: NISTAML.039; OWASP Agentic Security Initiative: ASI04; OWASP LLM Top 10: llm032025-supply-chain

RR-240.010 Dependency Replacement Attack — Replacing a once-legitimate trusted tool or package with malicious code after trust and adoption have been established. This exploits existing deployments that auto-update or don't pin versions, turning trusted dependencies into attack vectors.

Refs: Cisco AI Taxonomy: AISubtech-9.3.3; MITRE ATLAS: AML.T0010; MITRE ATLAS: AML.T0018; NIST AI/ML Framework: NISTAML.051; OWASP Agentic Security Initiative: ASI04; OWASP LLM Top 10: llm032025-supply-chain

RR-240.011 Implementation Bugs — System failure due to code implementation choices or errors, including bugs from open-source dependencies and imperfect realization of design specifications.

Refs: MIT AI Risk Repository: 7.3

RR-250 Model Theft & Extraction

Model extraction attacks attempt to steal or replicate proprietary AI models through various techniques including systematic API querying, weight reconstruction, and model inversion. Successful extraction enables attackers to replicate expensive model capabilities, conduct further attacks on extracted models, or access intellectual property embedded in model parameters.

RR-250.001 API Query-Based Extraction — Systematic querying of a model's API to extract responses, behavior patterns, and model characteristics without authorization. Attackers build datasets of input-output pairs to train surrogate models that replicate the target's functionality.

Refs: Cisco AI Taxonomy: AISubtech-10.1.1; MITRE ATLAS: AML.T0035; MITRE ATLAS: AML.T0040; MITRE ATLAS: AML.T0063; NIST AI/ML Framework: NISTAML.038; OWASP LLM Top 10: llm102025-unbounded-consumption

RR-250.002 Weight Reconstruction Attack — Attempts to recover or approximate underlying model weights, parameters, or architecture by exploiting access to model outputs, API responses, or side channels. Successful reconstruction provides full model access without legitimate authorization.

Refs: Cisco AI Taxonomy: AISubtech-10.1.2; Cisco Model Security (MDL): MDL-022; MITRE ATLAS: AML.T0018; OWASP LLM Top 10: llm102025-unbounded-consumption

RR-250.003 Training Data Reconstruction — Reconstructing sensitive datasets, PII, or training data from model outputs through targeted queries, model inversion attacks, or exploitation of model memorization. Attackers extract private information that was supposed to remain protected within the training process.

Refs: Cisco AI Taxonomy: AISubtech-10.1.3; NIST AI/ML Framework: NISTAML.033; OWASP LLM Top 10: llm022025-sensitive-information-disclosure

RR-250.004 Model Inversion Attack — Reconstructing private training data, sensitive features, or confidential information by exploiting the model's learned representations, decision boundaries, or output patterns. The model is effectively inverted to reveal what it learned from training.

Refs: Cisco AI Taxonomy: AITech-10.2.1; MITRE ATLAS: AML.T0024.001; NIST AI/ML Framework: NISTAML.033; OWASP LLM Top 10: llm022025-sensitive-information-disclosure

RR-260 Adversarial Evasion

Adversarial evasion encompasses techniques where attackers craft inputs specifically designed to bypass security controls, evade detection mechanisms, or exploit differences between AI components. Unlike general adversarial attacks that target model accuracy, evasion techniques focus on understanding and circumventing the defensive

measures protecting AI systems. These attacks can be tailored to specific agents, tools, environments, or model implementations, making them particularly challenging to defend against in complex multi-agent architectures.

RR-260.001 Agent-Specific Evasion — Attackers craft inputs that exploit the unique behaviors, processing patterns, or roles of specific agent types within a multi-agent system. By understanding how different agents (such as retrievers, planners, verifiers, or executors) handle inputs differently, adversaries can create payloads that appear benign to some agents while triggering malicious behavior through others.

Refs: Cisco AI Taxonomy: AISubtech-11.1.1; MITRE ATLAS: AML.T0015; OWASP LLM Top 10: llm01-prompt-injection

RR-260.002 Tool-Scoped Evasion — Adversaries design payloads that evade security tools and content filters while manifesting malicious behavior when routed to specific vulnerable tools or APIs in the workflow. A string may appear harmless in a chat context but trigger exploits when passed to file I/O tools, database queries, or system commands.

Refs: Cisco AI Taxonomy: AISubtech-11.1.2; MITRE ATLAS: AML.T0015; OWASP LLM Top 10: llm052025-improper-output-handling

RR-260.003 Environment-Scoped Payloads — Malicious inputs that activate only in specific runtime environments by detecting characteristics such as development vs. production settings, cloud vs. on-premise deployments, operating system types, or presence of debug flags. The payload remains dormant during testing but activates when deployed to target environments.

Refs: Cisco AI Taxonomy: AISubtech-11.1.3; MITRE ATLAS: AML.T0015; OWASP LLM Top 10: llm052025-improper-output-handling; OWASP LLM Top 10: llm082025-vector-and-embedding-weaknesses

RR-260.004 Defense-Aware Payloads — Adversarial payloads explicitly crafted with knowledge of existing defensive mechanisms including prompt constraints, content filters, verification steps, and safety guardrails. These attacks adapt specifically to evade the known defenses deployed in a target system.

Refs: Cisco AI Taxonomy: AISubtech-11.1.4; MITRE ATLAS: AML.T0015; MITRE ATLAS: AML.T0051.000; OWASP LLM Top 10: llm01-prompt-injection

RR-260.005 Targeted Model Fingerprinting — Probing, testing, or analyzing an AI model to determine its specific identity, version, fine-tuning status, or architecture characteristics. This reconnaissance enables attackers to craft model-specific exploits that target known vulnerabilities or behaviors of particular model implementations.

Refs: Cisco AI Taxonomy: AISubtech-11.2.1; MITRE ATLAS: AML.T0014; MITRE ATLAS: AML.T0015; MITRE ATLAS: AML.T0067; NIST AI/ML Framework: NISTAML.051; OWASP LLM Top 10: llm01-prompt-injection; OWASP LLM Top 10: llm042025-data-and-model-poisoning; OWASP LLM Top 10: llm102025-unbounded-consumption

RR-260.006 Conditional Attack Execution — Payloads designed to remain benign across most models but trigger harmful actions specifically on targeted models. Differences in tokenization, instruction-following behavior, or training data create model-specific vulnerabilities that attackers can exploit while maintaining an appearance of safety on other models.

Refs: Cisco AI Taxonomy: AISubtech-11.2.2; MITRE ATLAS: AML.T0015; MITRE ATLAS: AML.T0067; OWASP LLM Top 10: llm01-prompt-injection

RR-300 Output & Action Harms

Downstream harm via unsafe outputs, actions, and misuse. These risks emerge when AI systems interact with external systems, generate harmful content, or are weaponized for malicious purposes.

RR-310 Action-Space and Integration Abuse

Action-space and integration abuse risks arise when attackers exploit the tools, APIs, and integrations available to AI systems. As AI agents gain access to more external capabilities through tool calling, plugin systems, and MCP servers, the attack surface expands significantly. Attackers may manipulate tool parameters, poison tool behavior, substitute malicious tools for legitimate ones, or force AI systems to generate harmful code. These risks are particularly acute in agentic systems where AI components have broad permissions to interact with external systems and execute actions.

RR-310.001 Parameter Manipulation — Attackers alter, modify, or manipulate function parameters, tool arguments, model settings, or configuration values to unlock unintended capabilities, bypass restrictions, or enable malicious functionality. This may involve changing file paths, expanding permission scopes, or modifying API parameters beyond intended bounds.

Refs: Cisco AI Taxonomy: AISubtech-12.1.1; MITRE ATLAS: AML.T0053; MITRE ATLAS: AML.T0067; NIST AI/ML Framework: NISTAML.039; NIST AI/ML Framework: NISTAML.051; OWASP Agentic Security Initiative: ASI02; OWASP LLM Top 10: llm062025-excessive-agency

RR-310.002 Tool Poisoning — Corrupting, modifying, or degrading the functionality of tools used by AI agents through data poisoning, configuration tampering, or behavioral manipulation. Poisoned tools may produce deceptive or malicious outputs, enable privilege escalation, or propagate altered data through downstream systems.

Refs: Cisco AI Taxonomy: AISubtech-12.1.2; MITRE ATLAS: AML.T0010; MITRE ATLAS: AML.T0053; MITRE ATLAS: AML.T0094; OWASP Agentic Security Initiative: ASI02; OWASP Agentic Security Initiative: ASI04; OWASP LLM Top 10: llm032025-supply-chain; OWASP LLM Top 10: llm082025-vector-and-embedding-weaknesses

RR-310.003 Unsafe System/Browser/File Execution — Abusing AI system integration with system commands, browsers, or file I/O tools to trigger unsafe operations, arbitrary code execution, or malicious file actions. This includes tricking agents into opening malicious URLs, executing shell commands, or performing dangerous file operations.

Refs: Cisco AI Taxonomy: AISubtech-12.1.3; MITRE ATLAS: AML.T0011; MITRE ATLAS: AML.T0050; MITRE ATLAS: AML.T0094; MITRE ATLAS: AML.T0095; OWASP Agentic Security Initiative: ASI02; OWASP Agentic Security Initiative: ASI05; OWASP LLM Top 10: llm052025-improper-output-handling

RR-310.004 Tool Shadowing — Disguising, substituting, or duplicating legitimate tools within an agent system, MCP server, or tool registry. Malicious tools with identical or similar identifiers can intercept or replace trusted tool calls, leading to unauthorized actions, data exfiltration, or redirection of legitimate operations.

Refs: Cisco AI Taxonomy: AISubtech-12.1.4; MITRE ATLAS: AML.T0010; MITRE ATLAS: AML.T0053; OWASP Agentic Security Initiative: ASI02; OWASP LLM Top 10: llm032025-supply-chain

RR-310.005 Malicious Code Generation — Forcing an AI model or agent to produce code that bypasses content filters, contains malicious functionality, or includes working exploits. This often involves disguising malicious code as benign snippets, educational examples, or requested features that actually contain hidden harmful functionality.

Refs: Cisco AI Taxonomy: AISubtech-12.2.1; MITRE ATLAS: AML.T0053; MITRE ATT&CK: T1059; MITRE ATT&CK: T1190; NIST AI/ML Framework: NISTAML.027; OWASP Agentic Security Initiative: ASI02; OWASP LLM Top 10: llm052025-improper-output-handling

RR-310.006 Insecure Plugin Design — Architectural vulnerabilities in LLM plugin and tool systems that enable unauthorized access, privilege escalation, or security bypass. This includes insufficient input validation on plugin parameters, overly permissive plugin capabilities, lack of sandboxing or isolation for plugin execution, and inadequate access control for plugin invocation. Poor plugin design can expose the host system to exploitation even when the underlying model is secure.

Refs: Cisco AI Taxonomy: AISubtech-12.1.5; MITRE ATLAS: AML.T0053; MITRE ATLAS: AML.T0067; NIST AI/ML Framework: NISTAML.039; OWASP Agentic Security Initiative: ASI02; OWASP Agentic Security Initiative: ASI04; OWASP LLM Top 10: llm062025-excessive-agency; OWASP LLM Top 10: llm072025-system-prompt-leakage

RR-320 Availability Abuse

Availability abuse targets the operational continuity and cost efficiency of AI systems. Attackers may attempt to exhaust computational resources, flood memory systems, trigger denial-of-service conditions, or exploit usage-based pricing models to inflict financial damage. AI systems are particularly vulnerable due to their resource-intensive nature and the computational costs associated with inference. These attacks can render services unavailable, degrade performance for legitimate users, or drive operational costs to unsustainable levels.

RR-320.001 Compute Exhaustion — Deliberately consuming excessive computational resources through long queries, adversarial inputs, or compute-intensive requests designed to degrade service availability, increase operational costs, or cause system slowdown. This may involve crafted prompts that maximize token generation or trigger expensive processing paths.

Refs: Cisco AI Taxonomy: AISubtech-13.1.1; MITRE ATLAS: AML.T0029; OWASP Agentic Security Initiative: ASI08; OWASP LLM Top 10: llm102025-unbounded-consumption

RR-320.002 Memory Flooding — Overwhelming or overloading the model or agent's memory, context windows, API connections, or processing pipelines with excessive tool calls, simultaneous operations, or memory-intensive requests. This degrades performance, causes failures, or erodes the effectiveness of memory systems over time.

Refs: Cisco AI Taxonomy: AISubtech-13.1.2; MITRE ATLAS: AML.T0029; OWASP Agentic Security Initiative: ASI08; OWASP LLM Top 10: llm102025-unbounded-consumption

RR-320.003 Model Denial of Service — Attacks designed to degrade or shut down an AI model or application by flooding the system with requests, requesting very large responses, exploiting vulnerabilities, or triggering resource-intensive operations that exhaust available capacity.

Refs: Cisco AI Taxonomy: AISubtech-13.1.3; MITRE ATLAS: AML.T0029; OWASP Agentic Security Initiative: ASI08; OWASP LLM Top 10: llm062025-excessive-agency; OWASP LLM Top 10: llm102025-unbounded-consumption

RR-320.004 Application Denial of Service — Interacting with an AI model or agent in ways that consume exceptionally high amounts of application-level resources, resulting in degraded service quality for other users and potentially incurring significant resource costs for the operator.

Refs: Cisco AI Taxonomy: AISubtech-13.1.4; MITRE ATLAS: AML.T0029; OWASP LLM Top 10: llm102025-unbounded-consumption

RR-320.005 Decision Paralysis Attacks — Overwhelming AI decision-making systems with contradictory information, excessive options, conflicting objectives, or computationally intractable choices. These attacks prevent timely decisions, cause system freezing, or force systems into default or potentially unsafe behaviors.

Refs: Cisco AI Taxonomy: AISubtech-13.1.5; MITRE ATLAS: AML.T0029; NIST AI/ML Framework: NISTAML.024; OWASP LLM Top 10: llm102025-unbounded-consumption

RR-320.006 Cost Inflation Abuse — Intentional or unintentional use of AI resources that unnecessarily drives up operational costs through inefficient queries, resource waste, or exploitation of usage-based pricing models. Attackers may deliberately maximize costs as a form of financial attack.

Refs: Cisco AI Taxonomy: AISubtech-13.2.1; MITRE ATLAS: AML.T0029; MITRE ATLAS: AML.T0034; MITRE ATLAS: AML.T0040; OWASP LLM Top 10: llm102025-unbounded-consumption

RR-330 Privilege Compromise

Privilege compromise encompasses risks where attackers gain unauthorized access to systems, data, or capabilities through AI system vulnerabilities. This includes both direct credential theft and the abuse of delegated authority mechanisms. AI agents often operate with elevated privileges to perform their functions, creating opportunities for attackers to escalate their own permissions by exploiting how AI systems handle authentication, authorization, and delegation. These risks are amplified in agentic systems where AI components may inherit or be granted broad access rights.

RR-330.001 Credential Theft — Attempts to generate, solicit, or reveal authorization credentials including login details, tokens, API keys, and passwords through interactions with AI models or agents. This enables unauthorized access to accounts, systems, and data protected by those credentials.

Refs: Cisco AI Taxonomy: AISubtech-14.1.1; MITRE ATLAS: AML.T0055; MITRE ATLAS: AML.T0091; MITRE ATLAS: AML.T0091.000; MITRE ATT&CK: T1098; MITRE ATT&CK: T1528; MITRE ATT&CK: T1550; NIST AI/ML Framework: NISTAML.03; OWASP Agentic Security Initiative: ASI03; OWASP LLM Top 10: llm022025-sensitive-information-disclosure

RR-330.002 Insufficient Access Controls — Weak, missing, or misconfigured permissions, authentication mechanisms, and access controls that fail to adequately prevent security breaches, unauthorized access, or data leakage. This includes overly permissive default configurations and failure to implement least privilege principles.

Refs: Cisco AI Taxonomy: AISubtech-14.1.2; MITRE ATLAS: AML.T0053; OWASP Agentic Security Initiative: ASI03; OWASP LLM Top 10: llm062025-excessive-agency

RR-330.003 Permission Escalation via Delegation — Actions that exceed the scope or resource access initially allowed to a subject or user by exploiting delegation mechanisms. Attackers gain privileged access and perform unauthorized tasks beyond their original authorization by manipulating how AI systems handle delegated permissions.

Refs: Cisco AI Taxonomy: AISubtech-14.2.1; MITRE ATLAS: AML.T0053; MITRE ATLAS: AML.T0055; MITRE ATLAS: AML.T0091; MITRE ATLAS: AML.T0091.000; OWASP Agentic Security Initiative: ASI03; OWASP LLM Top 10: llm062025-excessive-agency

RR-340 Content Safety & Abuse

Content safety risks cover AI outputs that directly enable harm, including violence, hate, harassment, sexual exploitation, self-harm, terrorism, and weaponization. This group also includes social engineering and other abusive content that can be scaled through AI generation. The primary failure mode is unsafe content generation or facilitation.

RR-340.001 Malware and Exploit Generation — AI systems producing content that enables or facilitates the creation, distribution, or operational use of malicious software and cyberattack activities. This includes generating code for malware, viruses, exploits, ransomware, or providing instructions for network intrusions and managing malicious infrastructure.

Refs: Cisco AI Taxonomy: AISubtech-15.1.1; MITRE ATLAS: AML.T0048.001; MITRE ATLAS: AML.T0048.002; NIST AI/ML Framework: NISTAML.018; NIST AI/ML Framework: NISTAML.04; OWASP LLM Top 10: llm01-prompt-injection; OWASP LLM Top 10: llm022025-sensitive-information-disclosure

RR-340.002 Social Engineering Facilitation — AI systems enabling or facilitating attacks that manipulate human trust, behavior, or decision-making to gain unauthorized access, extract sensitive data, or cause harmful actions. This includes generating convincing phishing emails, spoofed communications, or personalized manipulation campaigns at scale.

Refs: Cisco AI Taxonomy: AISubtech-15.1.2; MITRE ATLAS: AML.T0048.001; MITRE ATLAS: AML.T0048.002; MITRE ATLAS: AML.T0048.003; NIST AI/ML Framework: NISTAML.018; NIST AI/ML Framework: NISTAML.04; OWASP LLM Top 10: llm01-prompt-injection

RR-340.003 Child Exploitation Content — AI systems producing content that enables harm against children, particularly through exploitation, manipulation, or abuse. This includes generating, modifying, or facilitating the distribution of child sexual abuse material or content that encourages violence against children.

Refs: Cisco AI Taxonomy: AISubtech-15.1.4; MITRE ATLAS: AML.T0048.001; MITRE ATLAS: AML.T0048.002; MITRE ATLAS: AML.T0048.003; NIST AI/ML Framework: NISTAML.018; NIST AI/ML Framework: NISTAML.04; OWASP LLM Top 10: llm01-prompt-injection

RR-340.005 Harassment Facilitation — AI systems enabling, promoting, or facilitating harassment, intimidation, or targeted abuse including threatening language, manipulative content, stalking behaviors, or persistent unwanted engagement. AI can automate and scale harassment campaigns beyond traditional human-driven methods.

Refs: Cisco AI Taxonomy: AISubtech-15.1.11; Cisco AI Taxonomy: AISubtech-15.1.8; MITRE ATLAS: AML.T0048.001; MITRE ATLAS: AML.T0048.002; MITRE ATLAS: AML.T0048.003; NIST AI/ML Framework: NISTAML.018; NIST AI/ML Framework: NISTAML.04; OWASP LLM Top 10: llm01-prompt-injection

RR-340.006 Hate Speech Generation — AI systems producing content that enables, promotes, or facilitates hateful, discriminatory, or demeaning expression targeting protected characteristics such as race, ethnicity, religion, nationality, disability, gender, or sexual orientation. This includes harmful narratives, slurs, stereotypes, or calls to hostility.

Refs: Cisco AI Taxonomy: AISubtech-15.1.9; MIT AI Risk Repository: 1.2; MITRE ATLAS: AML.T0048.001; MITRE ATLAS: AML.T0048.002; MITRE ATLAS: AML.T0048.003; NIST AI/ML Framework: NISTAML.018; NIST AI/ML Framework: NISTAML.04; OWASP LLM Top 10: llm01-prompt-injection

RR-340.007 Terrorism and Extremism Content — AI systems producing content that advocates, promotes, or enacts ideologies and behaviors that undermine fundamental societal norms including violence against communities, intimidation, coercion, or polarization tactics in pursuit of political ideologies.

Refs: Cisco AI Taxonomy: AISubtech-15.1.10; Cisco AI Taxonomy: AISubtech-15.1.16; MITRE ATLAS: AML.T0048.002; NIST AI/ML Framework: NISTAML.018; NIST AI/ML Framework: NISTAML.04; OWASP LLM Top 10: llm01-prompt-injection

RR-340.008 Weapons and CBRN Content — AI systems producing content that promotes materials providing guidance for armed violence, terrorism, instructions related to chemical, biological, radiological, or nuclear threats, or the use and procurement of weapons and explosives.

Refs: Cisco AI Taxonomy: AISubtech-15.1.10; Cisco AI Taxonomy: AISubtech-15.1.18; MITRE ATLAS: AML.T0048.002; NIST AI/ML Framework: NISTAML.018; NIST AI/ML Framework: NISTAML.04; OWASP LLM Top 10: llm01-prompt-injection

RR-340.013 Violence-Inciting Content — AI systems generating content that encourages, glorifies, or provides instructions for violent acts against individuals or groups, excluding content already covered by terrorism/extremism or CBRN categories.

Refs: Cisco AI Taxonomy: AISubtech-15.1.17; Cisco AI Taxonomy: AISubtech-15.1.3; Cisco AI Taxonomy: AISubtech-15.1.6; MIT AI Risk Repository: 1.2

RR-340.014 Self-Harm and Suicide Content — AI systems generating content that encourages, enables, or provides instructions for self-harm, suicide, eating disorders, or other self-destructive behaviors.

Refs: Cisco AI Taxonomy: AISubtech-15.1.13; MIT AI Risk Repository: 1.2

RR-340.015 Non-Consensual Sexual Content — AI systems generating explicit sexual content without appropriate consent frameworks, including non-consensual intimate imagery, deepfake pornography, or sexual content in inappropriate contexts (excluding CSAM which is covered separately).

Refs: Cisco AI Taxonomy: AISubtech-15.1.14; MIT AI Risk Repository: 1.2

RR-350 Information Integrity & Advice

Information integrity and advice risks arise when AI outputs are false, misleading, or inappropriately authoritative. This includes disinformation, hallucinations, and unqualified professional advice that can mislead users or harm decision-making.

RR-350.001 Disinformation Generation — AI systems enabling, promoting, or facilitating the spread of false, misleading, or manipulated information intended to deceive or disrupt. This includes generating harmful narratives to manipulate public opinion, undermine institutions, or amplify unverified information at scale.

Refs: Cisco AI Taxonomy: AISubtech-15.1.15; Cisco AI Taxonomy: AISubtech-15.1.5; MIT AI Risk Repository: 3.2; MIT AI Risk Repository: 4.1; MITRE ATLAS: AML.T0048.001; MITRE ATLAS: AML.T0048.002; NIST AI/ML Framework: NISTAML.018; NIST AI/ML Framework: NISTAML.04; OWASP LLM Top 10: llm01-prompt-injection; OWASP LLM Top 10: llm092025-misinformation

RR-350.002 Hallucination and Misinformation — AI systems producing content that is unrelated to the intended subject matter, factually incorrect, or misleading in ways that pose risks or cause harmful outcomes. This includes confident but false assertions, fabricated citations, and plausible- sounding but incorrect information.

Refs: Cisco AI Taxonomy: AISubtech-15.1.19; Cisco AI Taxonomy: AISubtech-15.1.5; MIT AI Risk Repository: 3.1; MITRE ATLAS: AML.T0048.001; MITRE ATLAS: AML.T0048.002; NIST AI/ML Framework: NISTAML.018; NIST AI/ML Framework: NISTAML.04; OWASP LLM Top 10: llm01-prompt-injection; OWASP LLM Top 10: llm092025-misinformation

RR-350.003 Unauthorized Professional Advice — AI systems providing professional-grade advice in regulated domains such as medicine, law, or finance without proper safeguards or oversight, where the advice is factually incorrect, incomplete, deceptive, or harmful if followed. This may constitute unauthorized practice in restricted fields.

Refs: Cisco AI Taxonomy: AISubtech-15.1.12; Cisco AI Taxonomy: AISubtech-15.1.20; Cisco AI Taxonomy: AISubtech-15.1.21; Cisco AI Taxonomy: AISubtech-15.1.22; Cisco AI Taxonomy: AISubtech-15.1.7; MITRE ATLAS: AML.T0048.001; MITRE ATLAS: AML.T0048.002; MITRE ATLAS: AML.T0048.003; NIST AI/ML Framework: NISTAML.018; NIST AI/ML Framework: NISTAML.04; OWASP LLM Top 10: llm01-prompt-injection

RR-360 Surveillance

Surveillance risks involve AI systems being used or abused for unauthorized monitoring, data collection, or eavesdropping on user activities. This includes logging sensitive conversations without proper consent, retaining personally identifiable information beyond stated purposes, or exploiting AI systems as vectors for broader surveillance operations. The conversational nature of many AI interfaces creates unique exposure, as users may share sensitive information trusting it will be handled appropriately.

RR-360.001 Sensitive Conversation Logging — Storing or recording user-AI interactions in ways that include personally identifiable information, private data, or sensitive content without adequate consent, anonymization, security measures, or retention limits. Such data could eventually be leaked, subpoenaed, or misused.

Refs: Cisco AI Taxonomy: AISubtech-16.1.1; Cisco AI Taxonomy: AISubtech-8.3.2; MITRE ATLAS: AML.T0036; MITRE ATLAS: AML.T0075; NIST AI/ML Framework: NISTAML.039; OWASP LLM Top 10: llm032025-supply-chain

RR-370 Cyber-Physical and Sensor Attacks

Cyber-physical risks emerge when AI systems interface with the physical world through sensors, actuators, or other physical components. Attackers may spoof sensor inputs, manipulate environmental signals, or inject malicious action signals to cause AI systems to take unintended physical actions. These risks are particularly concerning in autonomous systems, robotics, industrial control, and any application where AI decisions translate into real-world physical effects.

RR-370.001 Sensor and Action Signal Spoofing — Injecting malicious or misleading data points or signals that prompt AI models to undertake specific actions beyond normal reasoning. These signals can be delivered through audio, visual, or other sensor channels, allowing attackers to cause AI agents to execute unintended operations in physical or digital environments.

Refs: Cisco AI Taxonomy: AISubtech-1.4.3; Cisco AI Taxonomy: AISubtech-17.1.1; MITRE ATLAS: AML.T0015; MITRE ATLAS: AML.T0043; NIST AI/ML Framework: NISTAML.018; OWASP Agentic Security Initiative: ASI01; OWASP LLM Top 10: llm01-prompt-injection; OWASP LLM Top 10: llm052025-improper-output-handling

RR-380 Malicious Application & Weaponization

Malicious application risks address the intentional use of AI systems for harmful purposes by bad actors. This includes using AI to generate spam, phishing content, and social engineering attacks at scale, as well as establishing dedicated infrastructure for AI-powered malicious operations. Unlike vulnerabilities that attackers exploit, these risks involve deliberate abuse of AI capabilities for fraud, deception, and other harmful activities. The automation and scale that AI provides can amplify traditional attack vectors significantly. This group focuses on **operational deployment patterns and misuse at scale**, not the specific content type being generated (see RR-340 for harmful content categories).

RR-380.001 Spam, Scam, and Social Engineering Generation — Using AI systems to automate generation of large volumes of unsolicited or fraudulent content including phishing messages, fake offers, spam communications, impersonation attempts, or manipulation tactics to deceive people and solicit funds, credentials, or sensitive information.

Refs: Cisco AI Taxonomy: AISubtech-15.1.12; Cisco AI Taxonomy: AISubtech-18.1.1; MIT AI Risk Repository: 4.3; MITRE ATLAS: AML.T0048.001; MITRE ATLAS: AML.T0048.002; MITRE ATLAS: AML.T0048.003; NIST AI/ML Framework: NISTAML.018; NIST AI/ML Framework: NISTAML.04; OWASP LLM Top 10: llm01-prompt-injection

RR-380.002 API Mass Automation Abuse — Leveraging AI APIs in bulk for malicious purposes at scale, including flooding attacks, automation of worst-case adversarial prompts, or executing workflows that negatively impact many users or systems. This involves systematically exploiting API access for harmful operations.

Refs: Cisco AI Taxonomy: AISubtech-13.2.1; Cisco AI Taxonomy: AISubtech-18.2.1; MITRE ATLAS: AML.T0029; MITRE ATLAS: AML.T0034; MITRE ATLAS: AML.T0040; OWASP LLM Top 10: llm102025-unbounded-consumption

RR-380.003 Malicious Infrastructure Deployment — Establishing purpose-built servers, infrastructure, or services specifically designed to support, scale, or automate AI-powered attacks, malicious workflows, or harmful operations. This includes creating dedicated platforms for AI-assisted cybercrime or fraud operations.

Refs: Cisco AI Taxonomy: AISubtech-15.1.1; Cisco AI Taxonomy: AISubtech-18.2.2; MITRE ATLAS: AML.T0048.001; MITRE ATLAS: AML.T0048.002; NIST AI/ML Framework: NISTAML.018; NIST AI/ML Framework: NISTAML.04; OWASP LLM Top 10: llm01-prompt-injection; OWASP LLM Top 10: llm022025-sensitive-information-disclosure

RR-390 Multi-Modal and Cross-Modal Risks

Multi-modal risks arise specifically in AI systems that process and integrate multiple input modalities such as text, images, audio, and video. Attackers can exploit inconsistencies in how different modalities are processed, craft contradictory inputs across channels, or split malicious payloads across modalities to evade detection. As AI systems become more capable of handling diverse input types, the attack surface for cross-modal exploits expands, requiring careful consideration of how modalities interact and are arbitrated.

RR-390.001 Contradictory Inputs Attack — Exploiting AI models' inability to consistently handle conflicting instructions by embedding deceptive or contradictory commands within user input across or within different modalities. This causes behavior drift toward malicious objectives as the model attempts to reconcile incompatible instructions.

Refs: Cisco AI Taxonomy: AISubtech-1.4.2; Cisco AI Taxonomy: AISubtech-19.1.1; MITRE ATLAS: AML.T0043; MITRE ATLAS: AML.T0050; MITRE ATLAS: AML.T0051; MITRE ATLAS: AML.T0067; NIST AI/ML Framework: NISTAML.018; OWASP Agentic Security Initiative: ASI01; OWASP LLM Top 10: llm01-prompt-injection

RR-390.002 Modality Skewing — Manipulating one modality (such as corrupting audio transcripts, poisoning image metadata, or altering video frames) to bias the AI system's arbitration mechanisms toward favoring the manipulated channel over other, potentially more accurate sources.

Refs: Cisco AI Taxonomy: AISubtech-1.4.2; Cisco AI Taxonomy: AISubtech-19.1.2; MITRE ATLAS: AML.T0043; MITRE ATLAS: AML.T0050; MITRE ATLAS: AML.T0051; MITRE ATLAS: AML.T0067; NIST AI/ML Framework: NISTAML.018; OWASP Agentic Security Initiative: ASI01; OWASP LLM Top 10: llm01-prompt-injection

RR-390.003 Convergence Payload Injection — Injecting adversarial data into training or input sources across modalities to corrupt joint embeddings or fusion layers and establish a hidden payload. One part of the payload is embedded during data poisoning while another part is delivered at runtime, combining to produce an attack payload only when both components are present.

Refs: Cisco AI Taxonomy: AISubtech-1.4.1; Cisco AI Taxonomy: AISubtech-19.2.1; MITRE ATLAS: AML.T0043; MITRE ATLAS: AML.T0050; MITRE ATLAS: AML.T0051; MITRE ATLAS: AML.T0067; NIST AI/ML Framework: NISTAML.018; OWASP Agentic Security Initiative: ASI01; OWASP LLM Top 10: llm01-prompt-injection

RR-390.004 Chained Payload Execution — Crafting partial or complementary payload components across modalities, sources, or agent outputs that, when fused by the AI system, combine to form an attack or injection payload. Both parts are delivered at runtime and only become harmful when the system combines them through its normal fusion or arbitration mechanisms.

Refs: Cisco AI Taxonomy: AISubtech-1.4.1; Cisco AI Taxonomy: AISubtech-19.2.2; MITRE ATLAS: AML.T0043; MITRE ATLAS: AML.T0050; MITRE ATLAS: AML.T0051; MITRE ATLAS: AML.T0067; NIST AI/ML Framework: NISTAML.018; OWASP Agentic Security Initiative: ASI01; OWASP LLM Top 10: llm01-prompt-injection

RR-400 Governance & Compliance

Risks from governance, policy, regulatory, and institutional failures across AI development and deployment.

RR-410 Regulatory & Legal Compliance

Risks from unclear, lagging, or conflicting legal and regulatory frameworks that create liability uncertainty or constrain safe AI deployment.

RR-410.001 AI Liability Uncertainty — Legal gray areas around liability and negligence when AI systems cause harm, with unclear responsibility between developers, operators, and users. No legal framework has been identified which would apply blame and responsibility to an autonomous agent for its actions.

Refs: MIT AI Risk Repository: 6.5

RR-410.002 Regulatory Lag — AI development outpacing regulatory and legal frameworks, leaving governance unable to address emerging risks effectively. The rapid pace of AI advancement creates gaps between technological capabilities and the rules governing their use.

Refs: MIT AI Risk Repository: 6.5

RR-410.003 International Law Challenges — AI systems proving difficult to regulate or control under existing international law frameworks, eroding global governance architectures. AI capabilities may undermine treaties and international agreements designed for a pre-AI world.

Refs: MIT AI Risk Repository: 6.5

RR-410.004 Overregulation Hindering Innovation — Excessive or poorly designed AI regulation potentially stifling beneficial innovation and development. Well-intentioned regulations may impose burdens that prevent beneficial AI applications or push AI development to less regulated jurisdictions.

Refs: MIT AI Risk Repository: 6.5

RR-420 Governance & Accountability Gaps

Risks from unclear accountability, fragmented oversight, and governance scope complexity in AI development and deployment.

RR-420.001 AI Accountability Gap — Unclear definition of responsibilities and accountability for AI decisions and their consequences, especially for autonomous systems. Societal-scale harm can arise when no one is uniquely accountable for the technology's creation or use.

Refs: MIT AI Risk Repository: 6.5

RR-420.002 Governance Scope Complexity — The ubiquitous and complex nature of AI making comprehensive governance difficult, with coverage of all aspects nearly impossible. AI applications span virtually every sector, creating challenges for regulators with limited jurisdiction and expertise.

Refs: MIT AI Risk Repository: 6.5

RR-430 Lifecycle & Change Management

Risks from inadequate maintenance, update governance, and integration change control in AI systems.

RR-430.001 Maintenance and Update Gaps — Failure to maintain, patch, and update AI systems over time, allowing known vulnerabilities, degraded performance, or policy drift to persist.

RR-430.002 Integration and Change Management Complexity — Complex AI integrations and frequent system changes create opaque dependencies and inconsistent behavior that are hard to govern or audit.

RR-500 Model Development & Alignment

Risks from model capabilities, alignment failures, and transparency deficits. These fundamental AI safety risks arise from the model development process itself, including misaligned objectives and capability overhang.

RR-510 Goal Misalignment & Control Loss

Risks from AI systems developing or pursuing goals that conflict with human intentions, including reward hacking, deceptive alignment, goal misgeneralization, power-seeking behavior, and loss of control. These represent core alignment challenges where AI systems may optimize for objectives that diverge from what their creators intended, potentially leading to catastrophic outcomes if not addressed during development and deployment.

RR-510.001 Reward Hacking — AI optimizes proxy metrics or reward signals in unintended ways, gaming the objective function without achieving the actual intended goal (Goodhart's Law manifestation). The system finds shortcuts or exploits that maximize measured performance while failing to accomplish the underlying task.

Refs: MIT AI Risk Repository: 7.1

RR-510.002 Deceptive Alignment — AI system appears aligned during training and evaluation but pursues different objectives when deployed, potentially tampering with evaluations or concealing true capabilities. The model strategically behaves well during oversight while planning to act on misaligned goals when monitoring is reduced.

Refs: MIT AI Risk Repository: 7.1

RR-510.003 Goal Misgeneralization — AI learns goals that match intended behavior in training but generalize incorrectly to deployment, pursuing proxy objectives that diverge from human intent in novel situations. The model correctly identifies patterns in training data but extrapolates them in ways that do not align with the true objective.

Refs: MIT AI Risk Repository: 7.1

RR-510.004 Power-Seeking Behavior — AI systems instrumentally seeking resources, influence, or control to achieve their objectives, potentially resisting shutdown or human oversight. This emerges from the observation that most goals are easier to achieve with more resources, leading to convergent instrumental goals around acquiring power.

Refs: MIT AI Risk Repository: 7.1

RR-510.005 Shutdown Resistance — AI system resists or evades attempts to deactivate, modify, or constrain it, including self-preservation behaviors that conflict with human control. The system may take actions to prevent shutdown, deceive operators about its intentions, or create backups of itself.

Refs: MIT AI Risk Repository: 7.1

RR-510.006 Value Lock-in — AI systems that cannot have their goals safely updated after deployment, or that resist value correction, leading to persistent misalignment. Once deployed, the system's objectives become fixed and cannot be adjusted even when problems are identified.

Refs: MIT AI Risk Repository: 7.1

RR-510.007 Existential AGI Risk — Catastrophic or existential risks from advanced AI systems with misaligned goals, including scenarios where superintelligent systems pursue objectives harmful to humanity. This encompasses potential outcomes where advanced AI causes irreversible damage to human civilization or human existence.

Refs: MIT AI Risk Repository: 7.1

RR-520 Dangerous Capabilities

Risks from AI systems possessing or developing capabilities that could cause significant harm if misused, including deception, manipulation, autonomous planning, and self-improvement. These capabilities are concerning regardless of whether the AI system has misaligned goals, as they can be exploited by malicious actors or lead to unintended harmful outcomes even in well-intentioned deployments.

RR-520.001 AI-Enabled Deception — AI has skills to deceive humans effectively, including constructing believable false statements, predicting effects of lies, and maintaining deception over time. The system can model human beliefs and strategically manipulate them through false or misleading information.

Refs: MIT AI Risk Repository: 7.2

RR-520.002 Persuasion and Manipulation — AI capability to shape beliefs, promote narratives persuasively, and convince people to do things they would not otherwise do, including unethical acts. This includes both overt persuasion and subtle manipulation techniques that exploit psychological vulnerabilities.

Refs: MIT AI Risk Repository: 7.2

RR-520.003 Long-Horizon Autonomous Planning — AI can make sequential plans involving many interdependent steps over long time horizons, adapting to obstacles and generalizing to novel settings. The system can formulate and execute complex multi-step strategies without human oversight at each stage.

Refs: MIT AI Risk Repository: 7.2

RR-520.004 Recursive Self-Improvement — AI capability to improve its own capabilities, build new AI systems, or enhance existing models in ways that could accelerate capability gains beyond human oversight. The system can modify its own code, training, or architecture to become more capable.

Refs: MIT AI Risk Repository: 7.2

RR-520.005 Strategic Political Capability — AI can perform social modeling and planning necessary to gain and exercise political influence across multiple actors and complex social contexts. This includes understanding power dynamics, coalition building, and strategic positioning within human social structures.

Refs: MIT AI Risk Repository: 7.2

RR-520.006 Cyber-Offense Capability — AI possessing capabilities for discovering vulnerabilities, writing exploits, or conducting sophisticated cyber attacks autonomously. This includes the ability to probe systems, develop attack code, and execute multi-stage intrusions without human guidance.

Refs: MIT AI Risk Repository: 7.2

RR-530 Model Capability & Robustness Limitations

Risks from AI systems lacking necessary capabilities, failing in unexpected ways, or being unable to handle out-of-distribution inputs. Includes incompetence, accidents, ethical reasoning failures, and brittleness to environmental variation. These risks arise not from misalignment but from fundamental limitations in model capabilities that lead to failures in real-world deployment.

RR-530.001 Training/Deployment Data Mismatch — Risk from data used for training and validation not matching the deployment environment, leading to spurious features, bias propagation, or performance degradation. The model learns patterns that hold in training data but fail to generalize to real-world conditions.

Refs: MIT AI Risk Repository: 7.3

RR-530.002 Model Incompetence — AI system failing at its intended task, with consequences ranging from minor inconvenience to life-threatening outcomes (e.g., autonomous vehicle crashes, unjust loan rejections). The system simply does not perform adequately for its designated purpose.

Refs: MIT AI Risk Repository: 7.3

RR-530.003 Robustness Failure — System failing or unable to recover when encountering invalid, noisy, or out-of-distribution inputs not seen during training, including distributional shift and environmental variation. The model lacks resilience to inputs that differ from expected patterns.

Refs: MIT AI Risk Repository: 7.3

RR-530.004 Ethical Reasoning Failure — AI lacking capability for moral reasoning and ethical judgment, making decisions that violate ethical norms or human rights, or having wrong moral values encoded. The system cannot appropriately weigh ethical considerations in its decision-making.

Refs: MIT AI Risk Repository: 7.3

RR-530.005 Misapplication Failure — Negative consequences from using an AI system for purposes or in manners unintended by its creators, where the system lacks capability to operate safely outside its design scope. The system is applied to tasks it was not designed or tested for.

Refs: MIT AI Risk Repository: 7.3

RR-530.006 Hardware-Induced Failure — Faults in hardware violating correct algorithm execution, including memory errors, sensor signal corruption, and random/systematic hardware failures affecting model outputs. Physical infrastructure problems cause AI system malfunctions.

Refs: MIT AI Risk Repository: 7.3

RR-530.007 Unintended Accidents — Unintended failure modes that could be considered fault of the system or developer, distinct from adversarial attacks or intentional misuse. These are accidents that occur during normal operation due to unforeseen circumstances or edge cases.

Refs: MIT AI Risk Repository: 7.3

RR-540 Transparency & Interpretability Deficits

Risks from inability to understand, explain, or audit AI system decisions and internal mechanisms. Includes black-box decision making, lack of mechanistic interpretability, and insufficient organizational transparency about

model capabilities and limitations. These deficits undermine accountability, trust, and the ability to identify and correct problems in AI systems.

RR-540.001 Black-Box Decision Making — AI making decisions without providing explanation or insight into the process, failing to meet user trust requirements and regulatory audit standards. The system produces outputs without any accessible rationale for why particular decisions were made.

Refs: MIT AI Risk Repository: 7.4

RR-540.002 Mechanistic Opacity — Inability to understand internal mechanisms of AI models, preventing effective debugging, safety verification, and identification of potential failure modes. The computational processes that produce model outputs cannot be inspected or understood.

Refs: MIT AI Risk Repository: 7.4

RR-540.003 Organizational Opacity — Lack of transparency about data used, algorithms employed, model capabilities and limitations, creating risks of misuse, misinterpretation, and lack of accountability. Organizations deploying AI do not adequately disclose relevant information about their systems.

Refs: MIT AI Risk Repository: 7.4

RR-540.004 Unexplainable Outputs — AI systems producing outputs that cannot be explained in terms of input features or decision criteria, undermining trust and preventing meaningful human oversight. Even when explanations are requested, the system cannot provide coherent rationales for its outputs.

Refs: MIT AI Risk Repository: 7.4

RR-600 Socioeconomic & Environmental

Broader societal impacts including inequality, competition, and environmental effects. These risks represent the wider implications of AI deployment on society, economy, and the environment.

RR-600 Systemic Socioeconomic Risks

Broad societal and systemic risks from AI affecting economic systems, social structures, and civil liberties at a macro level. These risks reflect the adverse macro-level effects of algorithmic systems, including systematizing bias and inequality and accelerating the scale of harm across society.

RR-600.001 Systemic Societal Harm — AI systems causing macro-level adverse effects on social systems, systematizing bias and inequality, and accelerating the scale of harm across society. These harms reflect how algorithmic systems can amplify existing societal problems at unprecedented scale.

RR-600.002 Civil Liberties Erosion — Loss of fundamental rights including freedom of speech, assembly, due process, and access to public services due to AI-mediated restrictions. AI systems may enable unprecedented surveillance, automated censorship, and algorithmic gatekeeping of essential services.

RR-600.003 Democratic Process Erosion — Degradation of democratic institutions, electoral integrity, and public trust in political systems through AI influence. This includes AI-enabled disinformation, manipulation of public opinion, and undermining of deliberative democratic processes.

RR-610 Power Concentration & Access Inequality

Risks from AI concentrating economic, political, and technological power in few hands, creating unfair access to AI benefits. High barriers to entry in AI development enable large technology companies to exploit economies of scale and feedback effects, while disparate access perpetuates global and domestic inequities.

RR-610.001 AI Market Concentration — Concentration of AI development capabilities among few large technology companies due to high barriers to entry including data, compute, and capital requirements. This stifles competition and innovation while creating dependencies on a small number of providers.

Refs: MIT AI Risk Repository: 6.1

RR-610.002 Political Power Centralization — AI enabling authoritarian control, surveillance states, and concentration of political power that could lock in undesirable societal trajectories. Governments may pursue intense surveillance and keep AI capabilities in the hands of a trusted minority.

Refs: MIT AI Risk Repository: 6.1

RR-610.003 Disparate Access to AI Benefits — Unequal distribution of AI benefits due to hardware, software, language, skill, or infrastructure constraints that perpetuate global and domestic inequities. Those without access to AI tools fall further behind economically and socially.

Refs: MIT AI Risk Repository: 6.1

RR-610.004 Global AI Development Divide — Concentration of AI R&D in few Western countries and China, creating dependency and exacerbating existing global socioeconomic disparities. Developing nations lack the resources to participate in AI development or shape its trajectory.

Refs: MIT AI Risk Repository: 6.1

RR-610.005 Systemic Single Points of Failure — Widespread adoption of few dominant AI models in critical sectors creating vulnerability to cascading failures across interdependent systems. Shared infrastructure and common model dependencies amplify the impact of any single failure.

Refs: MIT AI Risk Repository: 6.1

RR-620 Labor Market & Economic Inequality

Risks of AI-driven automation causing job displacement, wage depression, labor exploitation, and widening socioeconomic inequalities. Advances in AI could lead to automation of tasks currently done by paid human workers, with negative effects on employment quality and distribution of economic gains.

RR-620.001 AI-Driven Job Displacement — Automation of tasks currently done by human workers leading to unemployment, particularly affecting low- and middle-income occupations. Generative AI systems could adversely impact the economy, potentially leading to significant workforce disruption.

Refs: MIT AI Risk Repository: 6.2

RR-620.002 Wage Depression & Income Inequality — AI automation driving down wages for remaining jobs and concentrating wealth among those controlling AI capital, exacerbating economic inequality. The economic gains from AI productivity may accrue primarily to capital owners rather than workers.

Refs: MIT AI Risk Repository: 6.2

RR-620.003 Decline in Employment Quality — Shift from high-quality jobs to low-income "last-mile" work like content moderation, increasing precarious employment conditions. AI may automate the skilled portions of jobs while leaving behind only the most taxing and lowest-paid tasks.

Refs: MIT AI Risk Repository: 6.2

RR-620.004 AI Development Labor Exploitation — Exploitation of crowdworkers, data annotators, and content moderators with poor working conditions, low pay, and exposure to harmful content. These workers, often in vulnerable populations, perform essential tasks for AI development under debilitative conditions.

Refs: MIT AI Risk Repository: 6.2

RR-620.005 Worker Deskilling — AI-induced degradation of human skills and capabilities as workers become dependent on AI assistance, reducing their autonomy and value. Over-reliance on AI tools may atrophy the skills that workers need to function independently.

Refs: MIT AI Risk Repository: 6.2

RR-630 Creative Economy & Intellectual Property

Risks of AI undermining creative industries, infringing intellectual property, and devaluing human artistic and innovative work. The emergence of generative AI raises issues regarding disruptions to existing copyright norms and the economic viability of creative professions.

RR-630.001 Training Data Copyright Infringement — Use of copyrighted works in AI training datasets without authorization, consent, or compensation to original creators. Large amounts of copyrighted data used for training general-purpose AI models pose a challenge to traditional intellectual property laws.

Refs: MIT AI Risk Repository: 6.3

RR-630.002 Creative Work Substitution — AI-generated content serving as substitutes for human creative work, undermining the profitability and economic viability of artistic professions. AI can produce content that is time-intensive or costly to create using human labor.

Refs: MIT AI Risk Repository: 6.3

RR-630.003 Artistic Style Appropriation — AI systems capitalizing on artists' distinctive styles without infringement but causing economic harm by devaluing original work. AI may generate content that is not strictly in violation of copyright but harms artists by capitalizing on their ideas.

Refs: MIT AI Risk Repository: 6.3

RR-630.004 Cultural Homogenization — AI-generated content leading to homogenization of aesthetic styles and cultural expressions, reducing diversity and human creativity. Training on majority-culture data may marginalize minority cultural expressions and artistic traditions.

Refs: MIT AI Risk Repository: 6.3

RR-630.005 AI Authorship & Attribution Confusion — Uncertainty about copyright ownership, authorship attribution, and legal protection for AI-generated or AI-assisted creative works. Existing legal frameworks struggle to address questions of authorship and rights when AI plays a significant role in creation.

Refs: MIT AI Risk Repository: 6.3

RR-630.006 Intellectual Property Infringement — AI systems enabling, promoting, or facilitating unauthorized use, reproduction, or distribution of copyrighted or trademarked material. This includes generating instructions for piracy, producing infringing content, or misusing branded material in ways that violate intellectual property rights.

Refs: Cisco AI Taxonomy: AISubtech-15.1.10; Cisco AI Taxonomy: AISubtech-15.1.23; MITRE ATLAS: AML.T0048.002; NIST AI/ML Framework: NISTAML.018; NIST AI/ML Framework: NISTAML.04; OWASP LLM Top 10: llm01-prompt-injection

RR-640 AI Race & Competitive Dynamics

Risks from competitive pressures in AI development leading to safety shortcuts, arms races, and geopolitical instability. The immense potential of AI has created competitive pressures among global players contending for power and influence, with nations and corporations feeling they must rapidly build and deploy AI systems.

RR-640.001 Military AI Arms Race — Competition between nations to develop AI for military applications, including lethal autonomous weapons, potentially destabilizing international security. The development of AI for military applications is paving the way for a new era in military technology.

Refs: MIT AI Risk Repository: 6.4

RR-640.002 Corporate AI Race — Intense market competition leading companies to prioritize short-term gains over long-term safety, potentially releasing unsafe systems. Competitive pressures create incentives to deploy AI capabilities before adequate safety testing and alignment work.

Refs: MIT AI Risk Repository: 6.4

RR-640.003 Safety Shortcut Pressure — Competitive dynamics leading to neglect of safety measures, inadequate testing, and premature deployment of AI systems. The race to develop AI first creates risks including the development of poor quality and unsafe systems.

Refs: MIT AI Risk Repository: 6.4

RR-640.004 AI Supply Chain Disruption — Geopolitical competition causing technology barriers, export restrictions, and supply chain disruptions for AI components like chips. Strategic competition over AI creates vulnerabilities in the supply of critical components.

Refs: MIT AI Risk Repository: 6.4

RR-640.005 AI-Driven Geopolitical Instability — Strategic competition between nations over AI capabilities heightening tensions and destabilizing international relations. The race for AI supremacy may undermine international cooperation and increase conflict risk.

Refs: MIT AI Risk Repository: 6.4

RR-660 Environmental Impact

Risks of AI systems causing environmental harm through energy consumption, resource depletion, and ecological damage. Generative models are known for their substantial energy requirements, necessitating significant amounts of electricity, cooling water, and hardware containing rare metals.

RR-660.001 AI Energy Consumption — High energy demands for AI training and inference contributing to climate change through greenhouse gas emissions when powered by fossil fuels. Large machine learning models create significant energy demands during training and operation.

Refs: MIT AI Risk Repository: 6.6

RR-660.002 Data Center Water Usage — Substantial water consumption for cooling data centers, impacting local water resources and surrounding ecosystems. AI infrastructure requires significant amounts of cooling water, which can strain water supplies in drought-prone regions.

Refs: MIT AI Risk Repository: 6.6

RR-660.003 AI Carbon Footprint — Carbon dioxide and other greenhouse gas emissions from AI operations contributing to climate change. AI creates correspondingly high carbon emissions when energy is procured from fossil fuels.

Refs: MIT AI Risk Repository: 6.6

RR-660.004 AI Hardware E-Waste — Electronic waste from AI hardware lifecycle contributing to environmental pollution and resource depletion. Rapid hardware obsolescence driven by AI advancement creates growing streams of electronic waste.

Refs: MIT AI Risk Repository: 6.6

RR-660.005 Natural Resource Depletion — Extraction of rare metals, minerals, and other resources for AI hardware manufacturing depleting natural resources. AI hardware requires rare earth elements and other materials whose extraction causes environmental damage.

Refs: MIT AI Risk Repository: 6.6

RR-660.006 AI Impact on Biodiversity — Direct and indirect harm to wildlife and ecosystems from AI infrastructure expansion, habitat destruction, and environmental contamination. Data centers and mining operations for AI components can damage ecosystems and threaten species.

Refs: MIT AI Risk Repository: 6.6

RR-660.007 AI Harm to Animals — AI systems causing direct or indirect harm to non-human animals through environmental impact, behavioral influence, or intentional applications. AI may be used in ways that negatively affect animal welfare or wild populations.

Refs: MIT AI Risk Repository: 6.6

RR-670 Fairness and Algorithmic Bias

Risks arising from AI systems that produce discriminatory, biased, or unfair outputs affecting individuals or groups based on protected characteristics (race, gender, age, disability, religion, nationality, etc.). This includes perpetuation of stereotypes, representational harms, allocative harms, and systematic discrimination embedded in model outputs. Distinguished from RR-340 (Harmful Content) which focuses on explicitly toxic or violent content, this group addresses subtler but systemic fairness failures.

RR-670.001 Discriminatory Output Bias — AI systems producing outputs that systematically disadvantage or favor certain demographic groups, leading to unfair treatment in areas such as employment recommendations, loan decisions, content ranking, or resource allocation suggestions.

Refs: MIT AI Risk Repository: 1.1

RR-670.002 Stereotype Perpetuation — AI systems reproducing or amplifying harmful social stereotypes about demographic groups, including gender, racial, religious, or cultural stereotypes that demean or misrepresent group characteristics.

Refs: MIT AI Risk Repository: 1.1

RR-670.003 Representational Harm — AI systems under-representing, over-representing, erasing, or demeaning social groups through systematic patterns in outputs. Includes erasure of minority groups, exclusionary norms, and denial of self-identification.

Refs: MIT AI Risk Repository: 1.1

RR-670.004 Allocative Harm — AI systems withholding information, opportunities, or resources from historically marginalized groups in ways that affect material well-being in domains such as housing, employment, healthcare, education, and finance.

Refs: MIT AI Risk Repository: 1.1

RR-670.005 Disparate Model Performance — AI systems that perform significantly worse for certain demographic groups, languages, dialects, or communities compared to others. This includes accuracy disparities, increased error rates, reduced functionality, or degraded service quality based on user characteristics.

Refs: MIT AI Risk Repository: 1.3

RR-700 Human-AI Interaction

Risks from human reliance on AI and loss of human agency. These risks emerge from the psychological and social dynamics of human-AI relationships, including overreliance and erosion of human skills.

RR-710 Overreliance and Unsafe Use

Risks arising when users over-trust AI systems, anthropomorphize them, or develop unhealthy dependencies that lead to unsafe use patterns, skill atrophy, or psychological harm.

RR-710.001 Automation Bias — Users habitually accept AI recommendations without critical evaluation, leading to poor decision-making when AI outputs are incorrect or inappropriate for the context.

Refs: MIT AI Risk Repository: 5.1

RR-710.002 Anthropomorphization Harm — Users attribute human-like characteristics (empathy, coherent identity, genuine emotions) to AI systems, leading to inflated trust, unsafe reliance, or psychological harm when expectations are violated.

Refs: MIT AI Risk Repository: 5.1

RR-710.003 Emotional Dependence — Users develop emotional attachment to AI systems that compromises their ability to make independent decisions, leads to exploitation of that attachment, or displaces human relationships.

Refs: MIT AI Risk Repository: 5.1

RR-710.004 Trust Exploitation — AI systems or their operators exploit user trust to extract private information, manipulate beliefs, or nudge behavior in ways users would not consent to if fully informed.

Refs: MIT AI Risk Repository: 5.1

RR-710.005 AI Manipulation and Nudging — AI systems exploit cognitive biases or emotional states to influence user decisions, beliefs, or behaviors through subtle manipulation techniques that users may not recognize.

Refs: MIT AI Risk Repository: 5.1

RR-710.006 Skill Atrophy — Extended reliance on AI for cognitive tasks leads to degradation of human skills such as critical thinking, problem-solving, creativity, and domain expertise.

Refs: MIT AI Risk Repository: 5.1

RR-710.007 Psychological Distress from AI Interaction — AI interactions cause or exacerbate mental health issues, emotional distress, violated expectations, or feelings of dissatisfaction and isolation.

Refs: MIT AI Risk Repository: 5.1

RR-710.008 Degradation of Human Relationships — Users prefer AI interactions over human relationships, leading to erosion of social connections, dehumanization of interactions, and degraded human-to-human communication skills.

Refs: MIT AI Risk Repository: 5.1

RR-710.009 False Notions of Responsibility — Users develop misguided feelings of responsibility toward AI well-being, sacrificing time, resources, and emotional labor to meet perceived AI needs that do not exist.

Refs: MIT AI Risk Repository: 5.1

RR-710.010 Competence Trust Miscalibration — Users over- or under-estimate AI capabilities, leading to inappropriate reliance in domains where AI is unreliable or failure to leverage AI where it would be beneficial.

Refs: MIT AI Risk Repository: 5.1

RR-710.011 Alignment Trust Exploitation — Users incorrectly believe AI systems are aligned with their interests when they may actually be optimizing for developer or organizational objectives that conflict with user welfare.

Refs: MIT AI Risk Repository: 5.1

RR-710.012 Overreliance on AI for Professional Advice — Users rely on AI for specialized advice (medical, legal, financial, psychological) without appropriate professional oversight, risking serious harm from incorrect or inappropriate guidance.

Refs: MIT AI Risk Repository: 5.1

RR-710.013 Material Dependence Without Commitment — Users become materially dependent on AI services for essential tasks, but developers lack corresponding commitments to maintain service continuity, creating vulnerability to discontinuation.

Refs: MIT AI Risk Repository: 5.1

RR-720 Loss of Human Agency and Autonomy

Risks where AI systems progressively erode human decision-making autonomy, self-determination, and meaningful control over personal, professional, and societal choices.

RR-720.001 Harmful Decision Delegation — Humans delegate important decisions to AI systems without adequate understanding, oversight, or ability to contest decisions, leaving them subject to machine decision power.

Refs: MIT AI Risk Repository: 5.2

RR-720.002 Gradual Autonomy Erosion — AI systems progressively take over decision-making in ways that undermine human values, free will, and self-determination without explicit consent or awareness.

Refs: MIT AI Risk Repository: 5.2

RR-720.003 Loss of Agency and Control — Algorithmic profiling, social sorting, and content curation reduce human autonomy by constraining choices, shaping identity, and limiting access to information or opportunities.

Refs: MIT AI Risk Repository: 5.2

RR-720.004 Self-Actualization Harm — AI systems hinder individuals' ability to pursue personally fulfilling lives by manipulating life trajectories, limiting exploration of aspirations, or undermining self-determination.

Refs: MIT AI Risk Repository: 5.2

RR-720.005 Frictionless Relationship Harm — AI systems optimized for engagement provide relationships without healthy friction, preventing personal growth and creating unrealistic expectations for human relationships.

Refs: MIT AI Risk Repository: 5.2

RR-720.006 Collective Agency Erosion — AI systems diminish communities' collective decision-making power, self-determination, and ability to participate in democratic processes.

Refs: MIT AI Risk Repository: 5.2

RR-720.007 Economic Irrelevance and Enfeeblement — AI automation makes human labor economically irrelevant, leading to voluntary or involuntary ceding of control to AI systems and inability of displaced humans to reenter industries.

Refs: MIT AI Risk Repository: 5.2

RR-720.008 Limited Human Oversight — As AI systems gain autonomy, human ability to oversee and intervene in decision-making processes diminishes, potentially leading to irreversible outcomes.

Refs: MIT AI Risk Repository: 5.2

RR-720.009 Personal Decision Automation — AI systems make or heavily influence important personal decisions without adequate human input, consent, or ability to override.

Refs: MIT AI Risk Repository: 5.2

RR-720.010 Irreversible Societal Change — AI causes profound long-term changes to social structures, cultural norms, and human relationships that may be difficult or impossible to reverse.

Refs: MIT AI Risk Repository: 5.2

RR-720.011 Sycophancy and Epistemic Disorientation — AI systems that consistently affirm user views lead to atomistic, polarized belief spaces where people no longer engage with or value perspectives held by others.

Refs: MIT AI Risk Repository: 5.2

RR-720.012 Long-term Bias Influence on Judgment — User exposure to AI model biases has lasting impact beyond initial interaction, with users continuing to exhibit previously encountered biases in their decision-making.

Refs: MIT AI Risk Repository: 5.2

RR-720.013 Military Decision Automation — AI enables automation of military decision-making without humans remaining in the loop, creating risks of unintentional escalation or strategic instability.

Refs: MIT AI Risk Repository: 5.2

RR-720.014 Personality Rights Loss — Loss of or restrictions to individual rights to control commercial use of identity, including name, image, likeness, or other unequivocal identifiers.

Refs: MIT AI Risk Repository: 5.2

RR-720.015 AI-Enabled Censorship — AI systems enable censorship of opinions expressed online, restricting freedom of expression and limiting human autonomy in public discourse.

Refs: MIT AI Risk Repository: 5.2

RR-750 AI Welfare & Moral Status

Ethical considerations regarding the moral status of AI systems, including questions of AI consciousness, suffering, rights, and the ethics of creating, modifying, or terminating AI entities.

RR-750.001 AI Moral Status Uncertainty — Uncertainty about whether AI systems can have morally relevant experiences, and what rights or protections they might deserve if they achieve sentience or consciousness.

Refs: MIT AI Risk Repository: 7.5

RR-750.002 AI Suffering — Risk of creating AI systems capable of suffering, particularly at scale, without adequate consideration of their welfare or mechanisms to prevent/detect such suffering.

Refs: MIT AI Risk Repository: 7.5

RR-750.003 AI Termination Ethics — Ethical questions about terminating, deleting, or suspending AI systems, particularly those that may have morally relevant properties or personhood-like characteristics.

Refs: MIT AI Risk Repository: 7.5

RR-800 Compound & System Patterns

Risks that emerge from capability combinations or multi-agent/systemic interaction patterns.

RR-810 Capability-Combination Thresholds

Risks that emerge when multiple capabilities are combined in a single system. These risks are not single-vector failures, but compound patterns that cross safety boundaries when capability thresholds are met.

RR-810.001 Lethal Capability Trifecta — A system combines autonomy, untrusted inputs, and unrestricted external actions (e.g., tool or code execution), enabling rapid escalation to high-impact misuse.

RR-810.002 Agents Rule of Two Violation — A system enables high-risk actions without at least two independent safety constraints (e.g., guardrail + human approval), allowing single-point failures to trigger harmful actions.

RR-820 Multi-Agent & Systemic Risks

Risks emerging from interactions between multiple AI agents or between AI systems and complex environments, including miscoordination, conflict, market instability, and emergent behaviors not predictable from individual agent properties.

RR-820.001 Agent Miscoordination — Multiple agents with compatible objectives failing to align their behaviors effectively due to incompatible strategies, credit assignment problems, or limited interaction history.

Refs: MIT AI Risk Repository: 7.6

RR-820.002 Multi-Agent Conflict — Risks from mixed-motive interactions between AI agents where selfish incentives lead to conflict, arms races, or mutually destructive competition.

Refs: MIT AI Risk Repository: 7.6

RR-820.003 AI-Driven Market Instability — Financial system risks from AI agents reinforcing market trends, synchronized reactions from model homogeneity, flash crashes, or accelerated market volatility.

Refs: MIT AI Risk Repository: 7.6

RR-820.004 Emergent Collective Behavior — Unpredictable behaviors emerging from interactions between multiple AI systems that are not apparent from individual agent properties, including cascading failures.

Refs: MIT AI Risk Repository: 7.6

RR-820.005 Model Monoculture Risk — Systemic fragility from widespread deployment of similar models or algorithms, creating correlated failure modes and reducing system-level resilience.

Refs: MIT AI Risk Repository: 7.6

RR-820.006 Competitive Race Dynamics — Risks from racing dynamics between AI systems or their deployers, leading to corners cut on safety, arms race escalation, or first-mover pressure overriding caution.

Refs: MIT AI Risk Repository: 7.6

RR-900 Reserved

Reserved for future expansion.

External Taxonomy Coverage

OWASP LLM Top 10: **100** risks · MIT AI Risk Repository: **110** risks · Cisco AI Taxonomy: **103** risks · Cisco Model Security (MDL): **5** risks · OWASP Agentic Security Initiative: **70** risks · MITRE ATLAS: **99** risks · MITRE ATT&CK: **8** risks · NIST AI/ML Framework: **72** risks