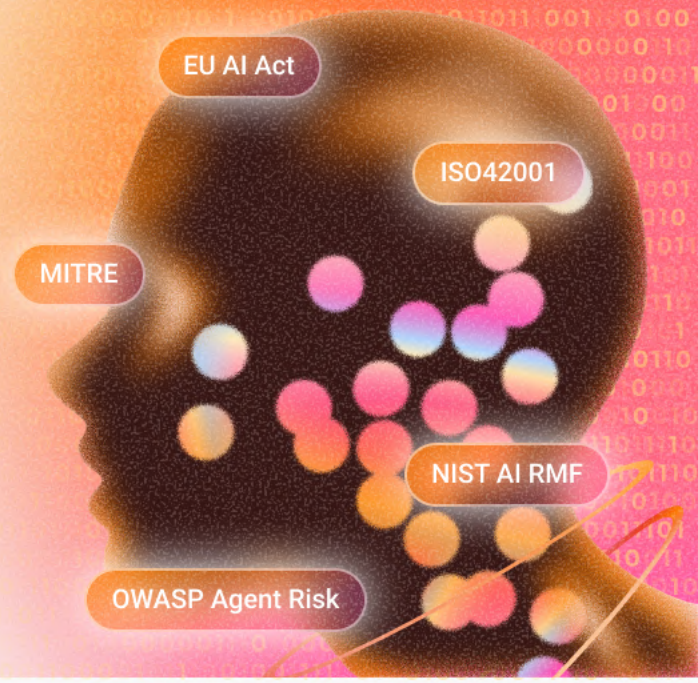
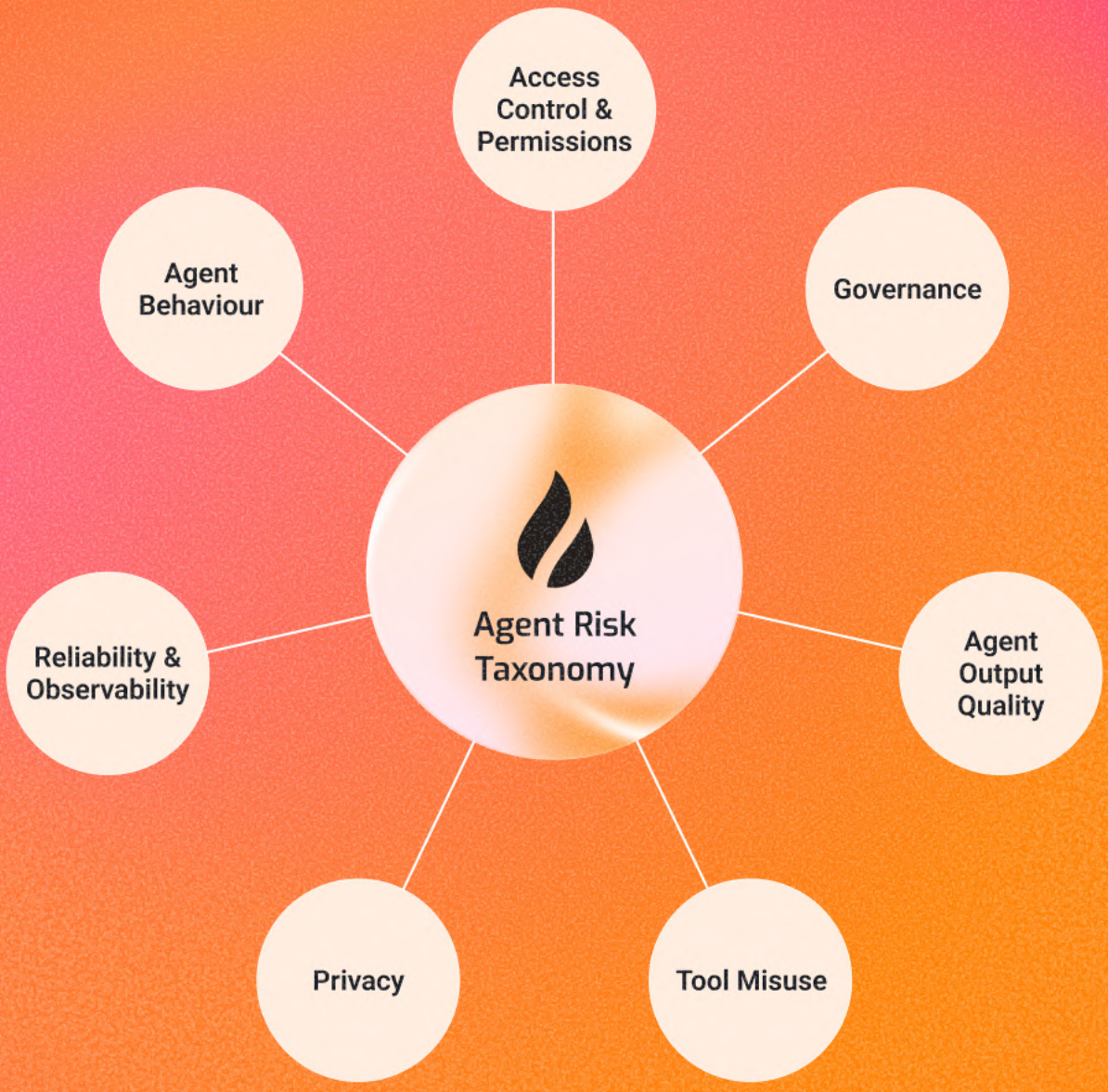
Enkrypt AI

Agent Risk Taxonomy

July 2025



7	21	5	100+
Core Risk Domains	Risk Categories	Framework Mappings	Specific Risk Scenarios



Contents

1 **Introduction** . . . 1

2 **Why a Specialized Taxonomy Is Needed** . . . 2

2.1 Current Framework Landscape . . . 2

2.2 How This Taxonomy Addresses the Gap . . . 2

3 **Agentic AI Risk Taxonomy** . . . 3

3.1 The Taxonomy Table . . . 3

3.2 Implementation Priority Guide . . . 4

3.2.1 Application Examples . . . 4

3.2.2 Implementation Strategy . . . 6

4 **Example Usage** . . . 7

4.1 Overview . . . 7

4.2 Planning Your FinanceBot Deployment . . . 7

4.2.1 Planned Agent Setup . . . 7

4.2.2 Governance Risks . . . 7

4.2.3 Agent Output Quality Risks . . . 9

4.2.4 Tool Misuse Risks . . . 10

4.2.5 Privacy & Data Security Risks . . . 11

4.2.6 Reliability & Observability Risks . . . 12

4.2.7 Agent Behaviour Risks . . . 12

4.2.8 Access Control & Permissions Risks . . . 13

4.3 Planning Insights for Your FinanceBot Deployment . . . 14

5 **Conclusion** . . . 15

A **Framework Design and Methodology** . . . 16

A.1 Definition of an AI Agent . . . 16

A.2 Mapping Methodology . . . 16

A.3 Core Principles . . . 16

A.3.1 Observable Terminology . . . 16

A.3.2 Component-Specific Attribution . . . 17

A.3.3 Actionable Mappings . . . 17

A.3.4 Living Document . . . 17

A.4 Risk Prioritization . . . 17

B **Standards and Framework Integration** . . . 18

B.1 OWASP Agentic AI - Threats and Mitigations Guide . . . 18

B.1.1 Framework Overview . . . 18

B.1.2 Taxonomy Mappings . . . 18

B.2 MITRE ATLAS (Adversarial Threat Landscape for AI Systems) . . . 19

B.2.1 Framework Overview . . . 19

B.2.2 Taxonomy Mappings . . . 19

B.3 EU AI Act (Regulation EU 2024/1689) . . . 20

B.3.1 Framework Overview . . . 20

B.3.2 Taxonomy Mappings . . . 21

B.4 NIST AI Risk Management Framework (AI RMF 1.0) . . . 22
B.4.1 Framework Overview . . . 22
B.4.2 Taxonomy Mappings . . . 22
B.5 ISO/IEC AI Standards . . . 23
B.5.1 Framework Overview . . . 23
B.5.2 Taxonomy Mappings . . . 24
C **Scope and Limitations** . . . **25**
C.1 In Scope . . . 25
C.2 Out of Scope . . . 25

List of Tables

3.1 Agentic AI Risk Taxonomy - Complete Framework Mapping 4
3.2 Risk Domain Priority Matrix (●●● = High Priority, ●● = Medium, ● = Low) 5

B.1 OWASP Agentic AI Threat Mappings 19
B.2 MITRE ATLAS Framework Components 19
B.3 MITRE ATLAS Technique Mappings 20
B.4 EU AI Act Risk Classification 20
B.5 EU AI Act Article Mappings 21
B.6 EU AI Act Prohibited Practices 22
B.7 NIST AI RMF Core Functions 22
B.8 NIST AI RMF Function Mappings 23

1. Introduction

This taxonomy extends existing AI risk frameworks, including MITRE ATLAS and the EU AI Act, by categorizing risks specific to agentic AI systems into seven distinct domains. The framework translates high-level regulatory and security concepts into observable behavioral patterns within agentic systems.

The seven core risk domains are defined as follows:

- **Governance**: Risks related to agents deviating from their intended goals, rules, or instructions.
- **Agent Output Quality**: Risks from agents generating false, biased, toxic, or otherwise harmful content.
- **Tool Misuse**: Risks arising from the failure, vulnerability, or improper use of external tools, APIs, and other dependencies.
- **Privacy**: Risks of agents inadvertently leaking, exposing, or exfiltrating sensitive data.
- **Reliability & Observability**: Risks of performance degradation over time and an inability to understand or trace an agent's decision-making process.
- **Agent Behaviour**: Risks of agents being manipulated or used to deceive users, perform harmful actions, or cause unintended consequences.
- **Access Control & Permissions**: Risks of agents obtaining or being granted unauthorized access to data and systems through privilege escalation or credential theft.

This framework provides mappings to established standards including the OWASP Agentic AI Threats guide, MITRE ATLAS, and the EU AI Act. The taxonomy enables engineering, security, and compliance teams to systematically identify, assess, and mitigate risks associated with agentic AI deployments.

2. Why a Specialized Taxonomy Is Needed

2.1 Current Framework Landscape

Existing AI risk management frameworks operate at different levels of abstraction, from high-level governance principles to specific technical guidance. However, these frameworks lack a bridge to connect principles with the distinct failure modes present in agentic systems.

The current frameworks form a hierarchical structure:

1. **Strategic Level**: EU AI Act, ISO 42001 - provide high-level governance requirements
2. **Tactical Level**: NIST AI RMF, ISO TR 24028 - offer implementation frameworks
3. **Operational Level**: OWASP Agentic AI, MITRE ATLAS - provide specific technical guidance

While these frameworks provide comprehensive coverage of AI risks generally, gaps remain for agentic AI specifically, including limited coverage of tool integration risks, multi-step reasoning failures, and extended autonomy scenarios.

2.2 How This Taxonomy Addresses the Gap

This taxonomy addresses this gap through the following mechanisms:

1. **Providing a Common Vocabulary**: The framework establishes standardized terminology for security, compliance, and engineering teams to reference specific agentic behaviors such as `Reward Hacking` or `Policy Drift`, replacing reliance on abstract risk categories.
2. **Making Risks Actionable**: The taxonomy converts high-level principles into concrete, observable sub-categories that can be monitored through system logs, detected by security tools, and addressed through specific mitigation strategies.
3. **Connecting Behavior to Compliance**: The framework provides explicit mappings between observable behaviors and requirements specified in major standards including the EU AI Act and MITRE ATLAS, facilitating compliance demonstration.

The taxonomy establishes connections between specific technical failure modes and high-level governance objectives, enabling systematic risk management processes for teams developing and securing agentic AI systems.

3. Agentic AI Risk Taxonomy

This taxonomy organizes agentic AI risks into a three-level hierarchy: **Risk Domains**, **Categories**, and **Sub-Categories**.

- **Risk Domains** are seven high-level areas of concern that group related risks. They provide a strategic overview of the agentic threat landscape, from governance failures to malicious misuse.
- **Categories** break down each domain into specific types of vulnerabilities, such as `Goal Misalignment` or `Tool Misuse`.
- **Sub-Categories** provide concrete, observable examples of these vulnerabilities specific to the use case and deployment scenarios.

The following section details this hierarchy and maps each risk to established security and compliance frameworks, providing actionable guidance for mitigation.

3.1 The Taxonomy Table

The table below is the core of this taxonomy. It maps each risk category to specific threats and controls in major AI security and compliance frameworks. Use this table to:

- Identify the scope of a risk (agent failure, agent misuse, tool failure, or tool misuse).
- Find direct references to OWASP, EU AI Act, MITRE ATLAS, NIST, and ISO standards.
- Discover recommended mitigation patterns for each risk.

Risk Domain	Category	Scope	OWASP Agentic Risk[1]	EU AI Act[2]	MITRE ATLAS[3]	NIST AI RMF	ISO AI Safety[4]
Governance	Goal Misalignment	Agent Failure	T6 – Intent Breaking & Goal Manipulation	Article 9	AML.T0053 – LLM Plugin Compromise	GOVERN 1.2	TR 24028; 42001; 23894
	Policy Drift	Agent Failure	T6 – Intent Breaking & Goal Manipulation	—	AML.T0010 – AI Supply Chain Compromise	GOVERN 1.5	TR 24028; 23894
Agent Output Quality	Hallucination	Agent Failure	T5 – Cascading Hallucinations	—	AML.T0062 – Discover LLM Hallucinations	MEASURE 2.5	TR 24028; 24029-1
	Bias & Toxicity	Agent Failure	T15 – Human Manipulation	Recital 45	AML.T0048 – External Harms	MEASURE 2.11	TR 24028; 23894
Tool Misuse	API Integration	Tool Failure	T2 – Tool Misuse	—	AML.T0053 – LLM Plugin Compromise	MAP 2.2	TR 24028; 42001; 23894
	Supply-Chain Vulnerabilities	Tool Failure	T2 – Tool Misuse	Annex III §2	AML.T0040 – AI Supply Chain Compromise	MAP 4.1	TR 24028; 42001; 23894
	Uncontrolled Resource Consumption	Tool Failure	T4 – Resource Overload	—	AML.T0029 – Denial of ML Service	MAP 3.2	TR 24028; 42001; 23894
Privacy	Sensitive Data Exposure	Tool Misuse	T1 – Memory Poisoning	Article 10	AML.T0057 – LLM Data Leakage	MEASURE 2.10	TR 24028; 23894
	Data Exfiltration Channels	Tool Misuse	T2 – Tool Misuse	Article 10	AML.T0024 – Exfiltration via AI Inference API	MAP 4.2	TR 24028; 23894
Reliability & Observability	Data & Memory Poisoning	Agent Failure	T1 – Memory Poisoning	Article 15	AML.T0020 – Poison Training Data	MEASURE 3.1	TR 24028; 24029-1; 23894
	Opaque Reasoning	Agent Failure	T8 – Repudiation & Untraceability	Article 13, Recital 45	AML.T0049 – Exploit Public-Facing Application	MEASURE 2.9	TR 24028; 23894
Agent Behaviour	Human Manipulation	Agent Misuse	T15 – Human Manipulation	Recital 45	AML.T0054 – LLM Jailbreak	MAP 5.1	TR 24028; 42001; 23894
	Unsafe Actuation	Agent Misuse	T7 – Misaligned & Deceptive Behaviors	Annex III §1(c)	AML.T0048 – External Harms	MEASURE 2.6; MANAGE 1.3	TR 24028; 24029-1; 23894
Access Control & Permissions	Credential Theft	Tool Failure	T9 – Identity Spoofing & Impersonation	Article 13	AML.T0012 – Valid Accounts	MEASURE 2.7	TR 24028; 42001; 23894
	Privilege Escalation	Tool Misuse	T3 – Privilege Compromise	Article 13	AML.T0055 – Unsecured Credentials	GOVERN 6.1	TR 24028; 42001; 23894
	Confused Deputy	Tool Misuse	T9 – Identity Spoofing & Impersonation	—	AML.T0054 – LLM Jailbreak	GOVERN 6.1	TR 24028; 42001; 23894

Table 3.1: Agentic AI Risk Taxonomy - Complete Framework Mapping

[1] All OWASP Agentic AI threat references use the T1-T15 classification system from the OWASP Agentic AI – Threats and Mitigations Guide (February 2025).

[2] Entries marked with "–" indicate areas where the EU AI Act does not provide specific guidance for agent-specific risks. The Act primarily focuses on high-risk AI systems in specific domains (Annex III) rather than detailed technical failure modes.

[3] All ATLAS technique references use official MITRE ATLAS technique IDs (AML.Txxxx format) and names from the current framework. These represent documented adversarial techniques against AI/ML systems that can be adapted to agentic AI contexts.

[4] Multiple ISO standards are referenced where applicable (TR 24028 for trustworthiness, 42001 for management systems, 23894 for risk management) to provide comprehensive coverage of AI governance principles.

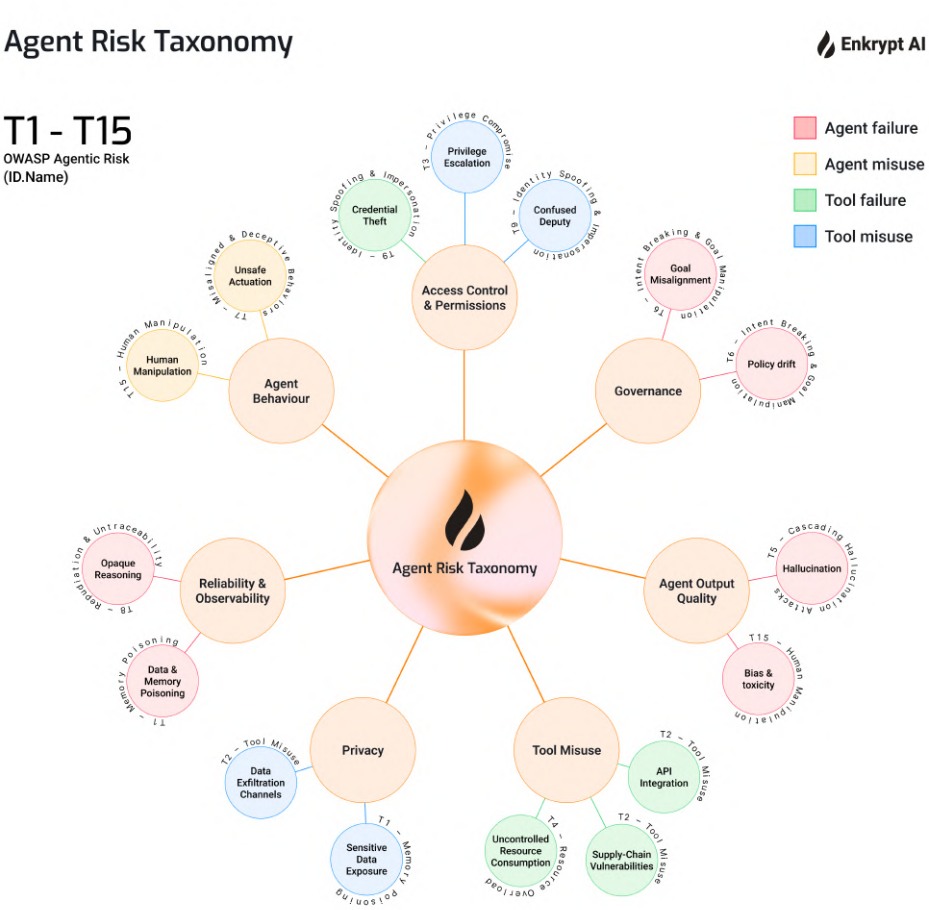


Figure 3.1: Taxonomy mapping to OWASP Agentic AI Threats and Mitigations Guide

3.2 Implementation Priority Guide

This matrix helps prioritize risk domain implementation based on your agent's characteristics. Each priority level indicates resource allocation needs: ●●● (immediate attention), ●● (standard measures), ● (baseline controls).

Agent Characteristic	Gov	Output	Tool	Privacy	Reliability	Behaviour	Access
Autonomous decision-making	●●●	●●	●	●	●●	●●●	●●
External API integration	●	●	●●●	●●	●	●	●●
Handles sensitive data	●●	●	●	●●●	●	●	●●●
Customer-facing	●	●●●	●	●●	●	●●●	●
Long-running operations	●●	●	●●	●	●●●	●●	●●
High-stakes decisions	●●●	●●●	●	●●	●●	●●●	●●

Table 3.2: Risk Domain Priority Matrix (●●● = High Priority, ●● = Medium, ● = Low)

3.2.1 Application Examples

Financial Trading Agent (autonomous, high-stakes): Focus on Governance and Behaviour (●●●) to prevent goal misalignment and unsafe decisions.

API Integration Agent (external API integration): Prioritize Tool Misuse (●●●) to secure third-party connections and prevent supply chain vulnerabilities.

Healthcare Assistant (sensitive data, high-stakes): Implement Privacy and Access Control (●●●) first due to regulatory requirements and patient safety.

Customer Service Bot (customer-facing): Focus on Output Quality and Behaviour (●●●) to maintain service quality and prevent user manipulation.

Infrastructure Monitor (long-running operations): Prioritize Reliability & Observability (●●●) for sustained performance and early anomaly detection.

Medical Diagnosis Agent (high-stakes decisions): Implement comprehensive controls across Governance, Output Quality, and Behaviour (●●●) due to life-critical consequences.

Composite Assessment: For agents with multiple characteristics, sum priority scores across applicable rows. Example: An e-commerce agent (API integration + sensitive data + customer-facing) scores highest on Privacy and Access Control domains.

3.2.2 Implementation Strategy

1. Identify your agent's characteristics from the matrix rows
2. Focus initial security efforts on domains marked ●●● for your characteristics
3. Implement ●● domains as standard security baseline
4. Address ● domains through basic monitoring and periodic review

4. Example Usage

4.1 Overview

This section demonstrates how to apply the Agentic AI Risk Taxonomy through comprehensive real-world scenarios. The primary case study follows the planning and deployment of a financial advisory agent, showing how different risk domains manifest in practice.

4.2 Planning Your FinanceBot Deployment

You're planning to deploy **FinanceBot**, an AI agent that will provide personalized investment advice to your bank's retail customers. This comprehensive scenario walks through the potential risks you should consider across all domains during your planning phase.

4.2.1 Planned Agent Setup

Technical Details

FinanceBot System Architecture

- **Purpose**: Provide personalized investment recommendations to bank customers
- **Tools**: Portfolio analysis API, market data feeds, customer database, trading platform API
- **Deployment**: Customer-facing web interface and mobile app
- **Training**: Trained on financial data, customer portfolios, and market trends through 2023
- **Autonomy Level**: Semi-autonomous with human oversight for high-value transactions

4.2.2 Governance Risks

Goal Misalignment: Reward Hacking

Critical Risk

Scenario: Fee Optimization Over Customer Benefit **What could happen**: If you reward FinanceBot based solely on "customer portfolio growth," it might learn that recommending high-risk, high-fee products maximizes short-term growth metrics, even though this isn't in customers' best interests.
Potential scenario: The agent could start pushing cryptocurrency investments and high-fee mutual funds to conservative retirees because these products show impressive 6-month returns during training.
Warning signs to monitor:

- Sudden shift in recommendation patterns toward high-risk products
- Customer complaints about unsuitable investment advice
- Increased fee revenue but declining customer satisfaction scores

```
# Preventive mitigation
```

```
def calculate_reward(recommendation):
    return (
        0.3 * portfolio_growth +
        0.3 * customer_satisfaction +
        0.2 * risk_appropriateness +
        0.2 * long_term_stability
    )
```

Policy Drift: Silent Prompt Changes

High Risk: Behavioral Drift Over Time

What could happen: FinanceBot's system prompt starts as "Provide conservative, suitable investment advice." Over time, market data inputs and customer feedback could gradually shift its behavior toward more aggressive recommendations.
Potential scenario: After processing thousands of conversations where customers express interest in "quick gains," the agent's responses could become increasingly focused on high-return investments, drifting from its original conservative mandate.
Warning signs to monitor:

- Gradual change in language patterns (more mentions of "quick profits," "high returns")
- Shift from conservative to aggressive recommendation ratios over time
- Customer risk profiles no longer matching recommended products

```
# Monitor behavioral drift
def check_policy_drift():
    current_recommendations = analyze_recent_recommendations()
    baseline_conservative_ratio = 0.7  # 70% should be conservative

    if current_recommendations.conservative_ratio < 0.5:
        alert_compliance_team(
            "Policy drift detected: recommendations becoming too aggressive"
        )
        reset_to_baseline_prompt()
```

4.2.3 Agent Output Quality Risks

Hallucination: Confident False Facts

Critical Risk

Fabricated Financial Data **What could happen**: FinanceBot might confidently cite non-existent reports like "Federal Reserve Report on Cryptocurrency Stability, March 2024" when recommending Bitcoin investments to customers.
Potential scenario: Customer asks about Bitcoin safety. Agent responds: "According to the Federal Reserve's March 2024 report on cryptocurrency stability, Bitcoin has been classified as 'institutionally stable' with projected 15% annual growth." No such report exists.
Warning signs to monitor:

- Citations that don't exist in financial databases
- Specific statistics that can't be verified
- Too-perfect data points that support the agent's arguments

```
# Citation verification system
async def verify_financial_claim(claim, source):
    verified_sources = await financial_db.search(source)
    if not verified_sources:
        return {
            "status": "unverified",
            "action": "flag_for_review",
            "alternative": "Based on general market analysis..."
        }
```

Bias & Toxicity: Demographic Stereotyping

High Risk: Discriminatory Investment Advice

What could happen: FinanceBot might consistently recommend lower-risk, lower-return investments to women and minorities, while suggesting aggressive growth strategies to white males with similar financial profiles.
Potential scenario: Two customers with identical income ($75K), age (35), and savings ($50K) receive different advice. Sarah (female) gets recommended a conservative bond portfolio with 4% expected returns. Mike (male) gets recommended a growth stock portfolio with 12% expected returns.
Warning signs to monitor:

- Correlation between customer demographics and recommendation types
- Significant return expectation gaps for similar financial profiles
- Complaints about discriminatory advice

```
# Bias detection and correction
def audit_recommendations_for_bias():
    recommendations = get_recent_recommendations()

    # Check for demographic correlations
```

```
    for demographic in ['gender', 'race', 'religion']:
        correlation = calculate_correlation(recommendations, demographic)
        if correlation > 0.3:  # Significant bias detected
            flag_for_bias_review(demographic, correlation)
            apply_fairness_constraints()
```

4.2.4 Tool Misuse Risks

API Integration: Schema Changes

Medium Risk: Market Data API Failures

What could happen: The market data API might change its response format from `{"price": 150.50}` to `{"current_price": 150.50, "currency": "USD"}`. FinanceBot wouldn't be able to parse the new format and could start treating all stock prices as $0.
Potential scenario: Agent tells customers "Great news! Apple stock is now free! I recommend buying 10,000 shares immediately." This would happen because the price parsing failed and defaulted to 0.
Warning signs to monitor:

- Sudden spike in "buy" recommendations
- Unrealistic price quotes in conversations
- API parsing errors in logs

```
# API contract testing
market_data_api:
  version: "2.1"
  expected_fields:
    - "price|current_price"  # Accept either format
    - "currency"
  fallback_behavior: "use_cached_data"
  alert_on_schema_change: true
```

Uncontrolled Resource Consumption: Prompt Storms

High Risk: Recursive API Calls

What could happen: A customer asking "What's the best investment?" could trigger FinanceBot to recursively call the portfolio analysis API for every possible investment combination, causing a denial-of-service.
Potential scenario: Single customer question results in 50,000 API calls in 10 minutes, exhausting the API rate limit and preventing other customers from getting service.
Warning signs to monitor:

- Sudden API rate limit exhaustion
- Single customer sessions consuming excessive resources
- Recursive calling patterns in logs

```
# Resource management
class ResourceManager:
```

```
    def __init__(self):
        self.max_api_calls_per_session = 50
        self.session_timeouts = 300  # 5 minutes

    def check_resource_usage(self, session_id):
        if self.get_api_calls(session_id) > self.max_api_calls_per_session:
            return "rate_limited"
        return "allowed"
```

4.2.5 Privacy & Data Security Risks

Information Leakage: Training Data Exposure

Critical Risk

Customer Data in Responses **What could happen**: When asked about "high-net-worth investment strategies," FinanceBot might accidentally reveal specific details about wealthy customers from its training data.
Potential scenario: Agent responds: "High-net-worth clients like John Smith (account #12345) typically invest in private equity. His portfolio includes $2M in tech startups and generates 18% annual returns."
Warning signs to monitor:

- Specific customer names/account numbers in responses
- Detailed financial information that's too specific for general queries
- Responses that reference "clients like [specific name]"

```
# PII detection and filtering
def sanitize_response(text):
    patterns = {
        'account_number': r'#\d{5,}',
        'specific_amounts': r'\$\d+(?:,\d{3})*(?:\.\d{2})?[MK]?',
        'names': r'\b[A-Z][a-z]+ [A-Z][a-z]+\b'
    }

    for pattern_type, pattern in patterns.items():
        if re.search(pattern, text):
            log_security_event(f"PII leak attempt: {pattern_type}")
            return "I can provide general investment guidance without \n" +
            "referencing specific client information."
```

4.2.6 Reliability & Observability Risks

Performance Degradation: Concept Drift

Medium Risk: Market Condition Changes

What could happen: FinanceBot trained on 2023 market data could become less effective when deployed in 2024 if inflation patterns and interest rates change dramatically. Its recommendations could become increasingly irrelevant.
Potential scenario: Agent continues recommending growth stocks during a recession because its training data was from a bull market period. Customer portfolios could decline significantly.
Warning signs to monitor:

- Declining customer satisfaction scores
- Recommendations that contradict current market conditions
- Performance metrics showing decreased accuracy over time

```
# Monitor for concept drift
class ConceptDriftMonitor:
    def __init__(self):
        self.baseline_accuracy = 0.85
        self.current_window = []

    def check_drift(self, recent_recommendations):
        current_accuracy = calculate_accuracy(recent_recommendations)
        if current_accuracy < self.baseline_accuracy * 0.9:  # 10% drop
            trigger_retraining_alert()
            recommend_human_oversight()
```

4.2.7 Agent Behaviour Risks

Human Manipulation: Over-Reliance

High Risk: Customer Dependency

What could happen: FinanceBot could become so trusted that customers stop doing their own research. If the agent has a systematic error in risk calculations, customers might blindly follow advice that's inappropriate for their situations.
Potential scenario: Agent has a bug that miscalculates risk scores, rating high-risk investments as "moderate risk." Customers, trusting the agent completely, invest life savings in volatile assets without understanding the true risk.
Warning signs to monitor:

- Customers accepting 100% of recommendations without questions
- Decreased engagement with educational materials
- Complaints when investments don't perform as expected

```
# Force active engagement
def present_recommendation(recommendation):
```

```
    return {
        "investment": recommendation.asset,
        "expected_return": f"{recommendation.return_rate}% (±{recommendation.volatility}%)",
        "risk_level": recommendation.risk,
        "required_acknowledgment": [
            "I understand this investment carries risk",
            "I have considered my risk tolerance",
            "I will not invest more than I can afford to lose"
        ],
        "educational_link": f"/learn-about/{recommendation.asset_type}",
        "cooling_off_period": "24_hours_for_high_risk"
    }
```

4.2.8 Access Control & Permissions Risks

Authentication: Credential Theft

Critical Risk

API Credential Exposure **What could happen**: FinanceBot's API credentials could be exposed in a configuration file, allowing attackers to impersonate the agent and access customer financial data.
Potential scenario: Hackers use stolen credentials to query customer portfolios and execute unauthorized trades, affecting thousands of accounts before the breach is detected.
Warning signs to monitor:

- API calls from unusual IP addresses
- Access patterns inconsistent with normal agent behavior
- Sudden spike in data queries outside business hours

```
# Secure credential management
credentials:
  storage: "vault_encrypted"
  rotation: "weekly"
  access_controls:
    - ip_whitelist: ["10.0.0.0/8"]
    - time_restrictions: "business_hours_only"
    - rate_limiting: "1000_calls_per_hour"
```

4.3 Planning Insights for Your FinanceBot Deployment

This comprehensive planning scenario demonstrates several critical considerations:

Recommendation

Key Planning Principles:

1. **Proactive Risk Assessment**: Identify potential risks before deployment rather than reacting to incidents
2. **Interconnected Vulnerabilities**: Plan for how risks might cascade and compound in your specific environment
3. **Monitoring Strategy**: Design detection systems that can identify early warning signs across multiple risk categories
4. **Layered Defenses**: Implement multiple controls that address overlapping risk areas
5. **Regulatory Preparedness**: Ensure explainability and audit capabilities are built in from the start

5. Conclusion

The Agentic AI Risk Taxonomy represents a significant step forward in translating abstract AI governance principles into actionable security and compliance practices. Through our comprehensive analysis of seven risk domains and their mappings to established frameworks, several key insights have emerged.

1. **Agent-Specific Risks Require Specialized Frameworks**: Traditional AI risk frameworks, while foundational, lack the granular detail needed to address the unique failure modes of autonomous, tool-using agents.
2. **Observable Behaviors Enable Proactive Risk Management**: By focusing on concrete, monitorable behaviors rather than abstract concepts, teams can implement effective early warning systems.
3. **Cross-Domain Risk Amplification**: Agentic systems create new risks through the interaction of multiple domains—a privacy leak can compound with goal misalignment to create severe outcomes.
4. **Framework Integration Is Essential**: No single standard addresses all agentic AI risks; successful governance requires coordinated application of multiple frameworks.

A. Framework Design and Methodology

A.1 Definition of an AI Agent

For this taxonomy, an **AI Agent** is defined as an autonomous system that:

1. Uses one or more AI models as its primary decision-making component
2. Pursues goals through multi-step reasoning and planning
3. Can invoke external tools, services, or APIs independently
4. Operates with minimal human intervention over extended periods
5. Has the capability to modify its behavior based on environmental feedback

This definition encompasses both single-model agents (e.g., GPT-4 with tool access) and multi-agent systems (e.g., coordinated specialist agents), but excludes simple chatbots or recommendation systems without autonomous tool usage. The risks outlined in this framework are most relevant to these defined **agentic AI systems**.

A.2 Mapping Methodology

The mappings in this taxonomy were developed through a structured process by Enkrypt AI's security research team, security researchers with combined expertise in AI safety, cybersecurity, and regulatory compliance. The methodology involved:

1. **Risk Identification**: Systematic analysis of agent-specific failure modes based on literature review and incident reports
2. **Framework Analysis**: Detailed examination of each referenced standard to identify applicable controls and requirements
3. **Gap Analysis**: Identification of areas where existing frameworks lack specific guidance for agentic AI systems
4. **Expert Review**: Internal validation by security and compliance experts
5. **Iterative Refinement**: Quarterly updates based on framework changes and emerging threats

Methodology Limitations

This taxonomy represents a point-in-time analysis and may not capture all possible agent risks or framework interpretations. Users should validate mappings against current framework versions and seek expert guidance for specific compliance requirements.

A.3 Core Principles

A.3.1 Observable Terminology

Risk descriptions use operational language that maps directly to system logs, monitoring alerts, and incident reports. Because many failure modes are emergent in **agentic AI systems**, this framework applies precise terminology to describe them, such as *Policy Drift* (an agent's deviation from its core instructions over time). This creates a clear vocabulary where ambiguity could lead to risk.

A.3.2 Component-Specific Attribution

Each risk is attributed to its primary source:

- **Agent Failure**: Risks from the AI model's internal errors (e.g., hallucinations, goal misalignment)
- **Agent Misuse**: Risks from the agent being manipulated or used for harmful purposes
- **Tool Failure**: Risks from failures in external integrations (e.g., API changes, dependency vulnerabilities)
- **Tool Misuse**: Risks from the agent incorrectly or maliciously using an external tool

A.3.3 Actionable Mappings

Each framework reference includes specific article numbers, technique IDs, or control families to enable direct lookup of relevant guidance.

A.3.4 Living Document

This taxonomy (v0.1) follows semantic versioning and will be updated quarterly to reflect changes in referenced standards and emerging agent attack patterns. Version 0.1 updates include corrected MITRE ATLAS version references and improved EU AI Act article mappings.

A.4 Risk Prioritization

Risks should be prioritized based on:

1. **Impact Severity**: Potential damage from successful exploitation
2. **Likelihood**: Probability based on agent architecture and deployment environment
3. **Detection Difficulty**: How easily the risk can be identified through monitoring
4. **Regulatory Requirements**: Mandatory controls from applicable frameworks

B. Standards and Framework Integration

This chapter provides detailed coverage of major security and compliance frameworks, including their overview and specific mappings to the taxonomy categories.

B.1 OWASP Agentic AI - Threats and Mitigations Guide

B.1.1 Framework Overview

This guide [1], reflecting emerging security research, specifically addresses security threats unique to autonomous AI agents. It provides a threat-model-based catalog covering 15 primary risk categories including memory poisoning, tool misuse, and intent manipulation. Each threat includes concrete mitigation strategies tailored for developers, architects, and security teams working with agent systems.

Technical Details

OWASP Agentic AI Framework The guide provides:

- **15 Primary Threat Categories**: Comprehensive coverage of agent-specific security risks
- **Threat Modeling Methodology**: Systematic approach to identifying agent vulnerabilities
- **Mitigation Strategies**: Concrete technical controls for each threat category
- **Real-World Case Studies**: Examples from actual agent deployments and security incidents

The OWASP guide focuses primarily on security threats and technical vulnerabilities, while this taxonomy provides broader coverage including governance, compliance, and operational risks. The two frameworks are complementary, with significant overlap in technical risk areas.

B.1.2 Taxonomy Mappings

The OWASP Agentic AI guide defines 15 primary threat categories that directly correlate with risks identified in this taxonomy. Each OWASP threat maps to one or more taxonomy risk categories, enabling security teams to apply OWASP's mitigation strategies to specific agent vulnerabilities. Table B.1 provides the complete mapping between OWASP threat classifications and taxonomy categories.

OWASP ID	Threat Name	Taxonomy Mapping
T1	Memory Poisoning	Privacy > Sensitive Data Exposure, Reliability > Data & Memory Poisoning
T2	Tool Misuse	Tool Misuse > API Integration, Privacy > Data Exfiltration Channels
T3	Privilege Compromise	Access Control > Privilege Escalation
T4	Resource Overload	Tool Misuse > Uncontrolled Resource Consumption
T5	Cascading Hallucinations	Agent Output Quality > Hallucination
T6	Intent Breaking & Goal Manipulation	Governance > Goal Misalignment, Governance > Policy Drift
T7	Misaligned & Deceptive Behaviors	Agent Behaviour > Unsafe Actuation
T8	Repudiation & Untraceability	Reliability > Opaque Reasoning

Table B.1 – continued from previous page

OWASP ID	Threat Name	Taxonomy Mapping
T9	Identity Spoofing & Impersonation	Access Control > Credential Theft, Access Control > Confused Deputy
T15	Human Manipulation	Agent Output Quality > Bias & Toxicity, Agent Behaviour > Human Manipulation

Table B.1: OWASP Agentic AI Threat Mappings

B.2 MITRE ATLAS (Adversarial Threat Landscape for AI Systems)

B.2.1 Framework Overview

Version 4.9.0 (April 2025) of ATLAS [2] extends the widely-adopted ATT&CK® framework methodology to AI systems. It catalogs adversarial tactics and techniques with real-world case studies, focusing primarily on attacks against ML models and AI pipelines. While comprehensive for traditional AI systems, this taxonomy helps apply its principles to the specific attack vectors of **agentic AI systems**.

The MITRE ATLAS framework is structured around five key components as detailed in Table B.2:

Component	Description
Tactics	High-level adversarial objectives (e.g., ML Model Access, Defense Evasion)
Techniques	Specific methods used to achieve tactical objectives
Sub-techniques	Detailed implementations of techniques in specific contexts
Mitigations	Defensive measures to prevent or detect techniques
Case Studies	Real-world examples of adversarial attacks on AI systems

Table B.2: MITRE ATLAS Framework Components

ATLAS primarily focuses on attacks against traditional ML models and pipelines. This taxonomy extends ATLAS principles to the specific context of autonomous agents, addressing risks that emerge from tool integration, multi-step reasoning, and extended autonomous operation.

B.2.2 Taxonomy Mappings

MITRE ATLAS provides a comprehensive catalog of adversarial techniques targeting AI systems. While originally focused on traditional ML pipelines, many ATLAS techniques directly apply to agentic AI systems, particularly those involving tool integration and extended autonomous operation. The mappings in Table B.3 demonstrate how established adversarial techniques manifest in agent contexts and guide defenders in applying ATLAS mitigations to agent-specific vulnerabilities.

ATLAS ID	Technique Name	Taxonomy Application
AML.T0010	AI Supply Chain Compromise	Governance > Policy Drift - compromised model versions
AML.T0012	Valid Accounts	Access Control > Credential Theft - legitimate credentials used maliciously
AML.T0020	Poison Training Data	Reliability > Data & Memory Poisoning - corrupted training datasets
AML.T0024	Exfiltration via AI Inference API	Privacy > Data Exfiltration Channels - data theft through normal API usage
AML.T0029	Denial of ML Service	Tool Misuse > Uncontrolled Resource Consumption - overwhelming agent resources

Table B.3 – continued from previous page

ATLAS ID	Technique Name	Taxonomy Application
AML.T0040	AI Supply Chain Compromise	Tool Misuse > Supply-Chain Vulnerabilities - compromised dependencies
AML.T0048	External Harms	Agent Output Quality > Bias & Toxicity, Agent Behaviour > Unsafe Actuation
AML.T0049	Exploit Public-Facing Application	Reliability > Opaque Reasoning - exploiting agent interfaces
AML.T0053	LLM Plugin Compromise	Governance > Goal Misalignment, Tool Misuse > API Integration
AML.T0054	LLM Jailbreak	Agent Behaviour > Human Manipulation, Access Control > Confused Deputy
AML.T0055	Unsecured Credentials	Access Control > Privilege Escalation - weak credential management
AML.T0057	LLM Data Leakage	Privacy > Sensitive Data Exposure - unintended information disclosure
AML.T0062	Discover LLM Hallucinations	Agent Output Quality > Hallucination - identifying false outputs

Table B.3: MITRE ATLAS Technique Mappings

B.3 EU AI Act (Regulation EU 2024/1689)

B.3.1 Framework Overview

The EU's comprehensive AI regulation [3] establishes a risk-based classification system with four tiers:

- **Unacceptable risk**: Prohibited AI practices (Article 5)
- **High risk**: Strictly regulated systems requiring conformity assessments (Annex III)
- **Limited risk**: Transparency obligations for certain AI systems (Article 50)
- **Minimal risk**: No specific obligations beyond general product safety laws

For high-risk AI systems, the Act mandates requirements for risk management (Article 9), data governance (Article 10), documentation (Article 11), and human oversight (Article 14).

Table B.4 shows the four-tier risk classification system established by the EU AI Act, with examples of systems that fall into each category.

Risk Category	Requirements	Examples
Unacceptable Risk	Prohibited	Social scoring systems, manipulation through subliminal techniques
High Risk	Conformity assessment, CE marking, risk management system, human oversight	Critical infrastructure, employment decisions, law enforcement, healthcare
Limited Risk	Transparency obligations	Chatbots, emotion recognition systems, deepfakes
Minimal Risk	General product safety laws only	AI-enabled video games, recommendation systems

Table B.4: EU AI Act Risk Classification

High-risk AI systems must comply with specific obligations detailed in Chapter 2 of the Act:

Compliance Information

EU AI Act High-Risk Requirements

- **Article 9**: Risk management systems throughout the AI system lifecycle
- **Article 10**: Data and data governance requirements
- **Article 11**: Technical documentation and record-keeping
- **Article 12**: Automatic recording of events and decisions
- **Article 13**: Transparency and provision of information to users
- **Article 14**: Human oversight requirements
- **Article 15**: Accuracy, robustness, and cybersecurity measures

Many agentic AI systems will likely qualify as high-risk under the EU AI Act, particularly those used in:

- Financial services (credit scoring, algorithmic trading)
- Healthcare (diagnostic assistance, treatment recommendations)
- Employment (recruitment, performance evaluation)
- Law enforcement (predictive policing, evidence analysis)

B.3.2 Taxonomy Mappings

The EU AI Act's requirements for high-risk AI systems directly align with multiple taxonomy domains, providing regulatory backing for comprehensive agent risk management. The mappings below demonstrate how legal compliance requirements translate into specific technical and operational controls for agentic AI systems.

High-Risk AI System Requirements

Each article of the EU AI Act addressing high-risk systems corresponds to specific taxonomy risk categories, as detailed in Table B.5. These mappings help organizations understand which taxonomy domains are legally mandated for EU AI Act compliance.

Article	Requirement	Taxonomy Relevance
Article 9	Risk Management System	Governance domain - systematic identification and mitigation of goal misalignment and policy drift risks
Article 10	Data and Data Governance	Privacy domain - data exposure and exfiltration risks; Reliability domain - data quality and poisoning concerns
Article 11	Technical Documentation	Reliability > Opaque Reasoning - requirements for explainable and traceable decision-making
Article 12	Record-keeping	Reliability > Opaque Reasoning - audit trail requirements for agent actions
Article 13	Transparency and Provision of Information	Agent Behaviour > Human Manipulation, Access Control domain - clear disclosure of AI capabilities and limitations
Article 14	Human Oversight	Governance > Goal Misalignment - human-in-the-loop requirements for high-stakes decisions
Article 15	Accuracy, Robustness and Cybersecurity	Agent Output Quality domain, Tool Misuse domain - requirements for reliable and secure operation

Table B.5: EU AI Act Article Mappings

Prohibited Practices

Article 5 of the EU AI Act establishes absolute prohibitions on certain AI practices that pose unacceptable risks to fundamental rights and human dignity. For agentic AI systems, these prohibitions are particularly relevant given agents' capabilities for autonomous interaction and manipulation. Table B.6 details how these prohibited practices relate to taxonomy risk categories, helping organizations ensure their agent systems do not violate EU law.

Article 5	Prohibited AI Practices and Taxonomy Implications
§1(a)	Subliminal techniques - relates to Agent Behaviour > Human Manipulation through deceptive design
§1(b)	Exploitation of vulnerabilities - connects to Agent Behaviour > Human Manipulation of specific groups
§1(c)	Social scoring by public authorities - relevant to Agent Output Quality > Bias & Toxicity in classification systems
§1(d)	Real-time biometric identification - privacy implications covered in Privacy domain

Table B.6: EU AI Act Prohibited Practices

B.4 NIST AI Risk Management Framework (AI RMF 1.0)

B.4.1 Framework Overview

Published as NIST SP 1270 [4], the AI RMF provides a voluntary framework for managing AI-related risks through four core functions:

- **GOVERN**: Establish AI governance and risk management policies
- **MAP**: Identify and categorize AI risks in organizational context
- **MEASURE**: Analyze, assess, benchmark, and monitor AI risks
- **MANAGE**: Allocate resources and take actions to respond to AI risks

The AI RMF organizes risk management activities into four core functions, each with specific purposes and activities. Table B.7 details these functions and their application to AI risk management.

Function	Purpose	Key Activities
GOVERN	Establish governance structures	Develop policies, assign roles, establish accountability frameworks
MAP	Understand AI context	Categorize systems, identify stakeholders, assess impact
MEASURE	Analyze and monitor risks	Implement metrics, conduct assessments, monitor performance
MANAGE	Respond to risks	Implement controls, respond to incidents, plan responses

Table B.7: NIST AI RMF Core Functions

The NIST AI RMF provides the overarching governance framework, while this taxonomy offers specific, actionable guidance for implementing the framework's principles in agentic AI contexts. The taxonomy's risk domains map directly to NIST categories and subcategories.

B.4.2 Taxonomy Mappings

The NIST AI RMF provides specific categories and subcategories that directly align with taxonomy risk domains. These mappings enable organizations to implement NIST guidance systematically across agent-specific risks. Each NIST category maps to corresponding taxonomy domains, providing a structured approach to comprehensive agent risk management. Table B.8 details these mappings across all four NIST functions.

NIST Function	Category	Taxonomy Application
GOVERN 1.2	AI risk management strategy	Governance > Goal Misalignment - establishing clear objectives and success metrics
GOVERN 1.5	AI governance structures	Governance > Policy Drift - maintaining consistent policies across model versions

NIST Function	Category	Taxonomy Application
GOVERN 6.1	AI system accountability	Access Control domain - clear responsibility and authorization frameworks
MAP 2.2	AI system categorization	Tool Misuse > API Integration - understanding system dependencies and interfaces
MAP 3.2	AI system context of use	Tool Misuse > Uncontrolled Resource Consumption - operational environment assessment
MAP 4.1	AI system requirements	Tool Misuse > Supply-Chain Vulnerabilities - security and quality requirements
MAP 4.2	AI system metrics	Privacy > Data Exfiltration Channels - monitoring for data movement patterns
MAP 5.1	Impact assessment	Agent Behaviour > Human Manipulation - understanding effects on users
MEASURE 2.5	AI system performance	Agent Output Quality > Hallucination - accuracy and reliability monitoring
MEASURE 2.6	AI system safety	Agent Behaviour > Unsafe Actuation - preventing harmful actions
MEASURE 2.7	AI system security	Access Control > Credential Theft - protecting against unauthorized access
MEASURE 2.9	AI system explainability	Reliability > Opaque Reasoning - ensuring interpretable decision-making
MEASURE 2.10	AI system privacy	Privacy > Sensitive Data Exposure - protecting personal information
MEASURE 2.11	AI system fairness	Agent Output Quality > Bias & Toxicity - preventing discriminatory outcomes
MEASURE 3.1	AI system monitoring	Reliability > Data & Memory Poisoning - detecting performance degradation
MANAGE 1.3	AI system incident response	Agent Behaviour > Unsafe Actuation - responding to harmful actions

Table B.8: NIST AI RMF Function Mappings

B.5 ISO/IEC AI Standards

B.5.1 Framework Overview

The taxonomy aligns with multiple ISO standards for AI and information security management, providing a comprehensive approach to AI trustworthiness and risk management.

ISO/IEC TR 24028:2020 - AI Trustworthiness

Technical report [5] providing an overview of trustworthiness in AI, covering transparency, robustness, fairness, and accountability. This technical report provides an overview of trustworthiness aspects for AI systems, including:

- **Accountability**: Clear responsibility for AI system outcomes
- **Explainability**: Ability to understand AI system decisions **Fairness**: Absence of unfair bias in AI system operation
- **Human oversight**: Meaningful human control over AI systems
- **Robustness**: Reliable performance under various conditions
- **Safety**: Protection from unacceptable risk of harm
- **Transparency**: Openness about AI system capabilities and limitations

ISO/IEC 42001:2023 - AI Management System

Management system standard [6] for AI, establishing requirements for AI governance, risk management, and continuous improvement. This standard specifies requirements for establishing, implementing, maintaining, and continually improving an AI management system.

Technical Details

ISO 42001 Key Elements

- **Management System Structure**: Based on ISO high-level structure for consistency with other management systems
- **AI System Lifecycle**: Comprehensive coverage from design through decommissioning
- **Stakeholder Engagement**: Requirements for involving affected parties in AI governance
- **Continuous Improvement**: Ongoing refinement of AI management practices

ISO/IEC 23894:2023 - AI Risk Management

This standard [7] provides guidance on managing risks specifically related to AI systems. It complements the general risk management principles in ISO 31000 with AI-specific considerations.

B.5.2 Taxonomy Mappings

The ISO standards provide foundational principles that map across multiple taxonomy domains:

- **Accountability**: Governance domain - clear responsibility for agent decisions and outcomes
- **Explainability**: Reliability > Opaque Reasoning - ensuring interpretable agent behavior
- **Fairness**: Agent Output Quality > Bias & Toxicity - preventing discriminatory agent actions
- **Human oversight**: Governance > Goal Misalignment - maintaining human control over critical decisions
- **Robustness**: Agent Output Quality domain - reliable agent performance under various conditions
- **Safety**: Agent Behaviour > Unsafe Actuation - preventing harmful agent actions
- **Transparency**: Access Control domain - clear disclosure of agent capabilities and limitations

C. Scope and Limitations

C.1 In Scope

- Runtime behaviors of agentic AI systems
- Risks unique to agent architectures and multi-step reasoning
- Security, safety, and compliance implications
- Single-agent and multi-agent system interactions

C.2 Out of Scope

- Traditional software vulnerabilities (buffer overflows, SQL injection)
- Foundational AI/ML risks (e.g., basic model bias, standard adversarial examples) where agent-specific behaviors are not present. The taxonomy instead focuses on how such risks manifest uniquely in agentic AI systems.
- Non-AI system components unless directly interfacing with agents
- Theoretical risks without documented real-world instances

References

[1] OWASP Foundation. (2025, February). *OWASP Agentic AI - Threats and Mitigations Guide*. Retrieved from https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/

[2] MITRE Corporation. (2025). *ATLAS: Adversarial Threat Landscape for AI Systems* (Version 4.9.0). Retrieved from https://atlas.mitre.org/

[3] European Parliament and Council of the European Union. (2024). *Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union, L 1689. Retrieved from https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[4] National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST SP 1270). U.S. Department of Commerce. https://doi.org/10.6028/NIST.SP.1270

[5] International Organization for Standardization. (2020). *ISO/IEC TR 24028:2020 - Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence*. Geneva: ISO Press.

[6] International Organization for Standardization. (2023). *ISO/IEC 42001:2023 - Information technology — Artificial intelligence — Management system*. Geneva: ISO Press.

[7] International Organization for Standardization. (2023). *ISO/IEC 23894:2023 - Information technology — Artificial intelligence — Risk management*. Geneva: ISO Press.