ACKUITY

Threat Detection for Agentic AI

Taxonomy of Threats to detect when AI Agents are in production

Introduction

You have deployed AI agents in your environment. Your security policies mandate security monitoring of all assets in production and your existing SOC does that for your current assets.

Now you need to provide similar security monitoring for all your AI agents and AI applications.

The threats to Agentic AI are different. This paper discusses various types of threats in Agentic AI that arises as you deploy Agents in production and how to monitor them in real time.

Agentic AI threat areas

Threats to Agentic AI fall in four main areas:

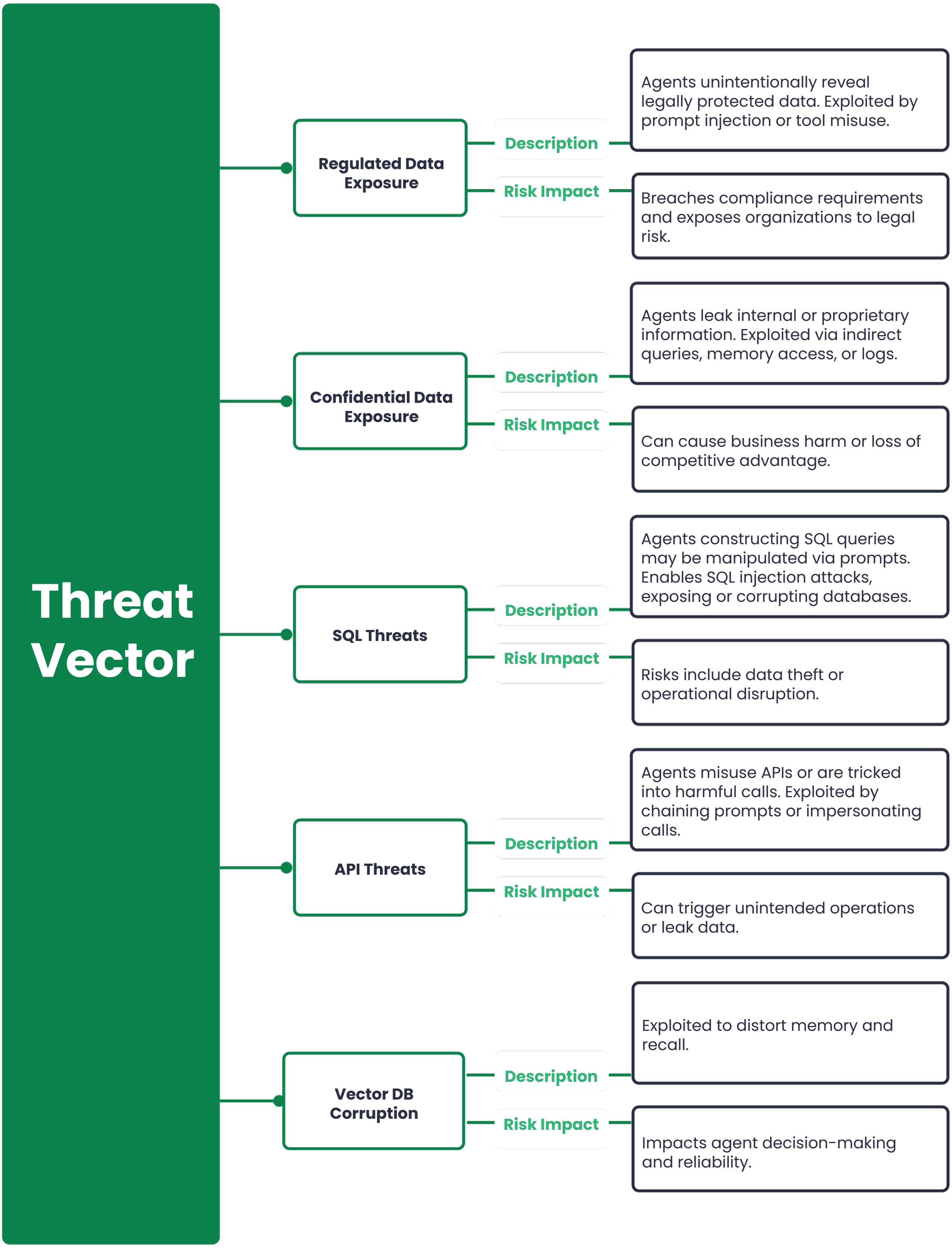
01 Data Interaction Threats

02 Agentic Privilege Misuse

03 Agent Manipulation Threats

04 Agent Overreach Threats

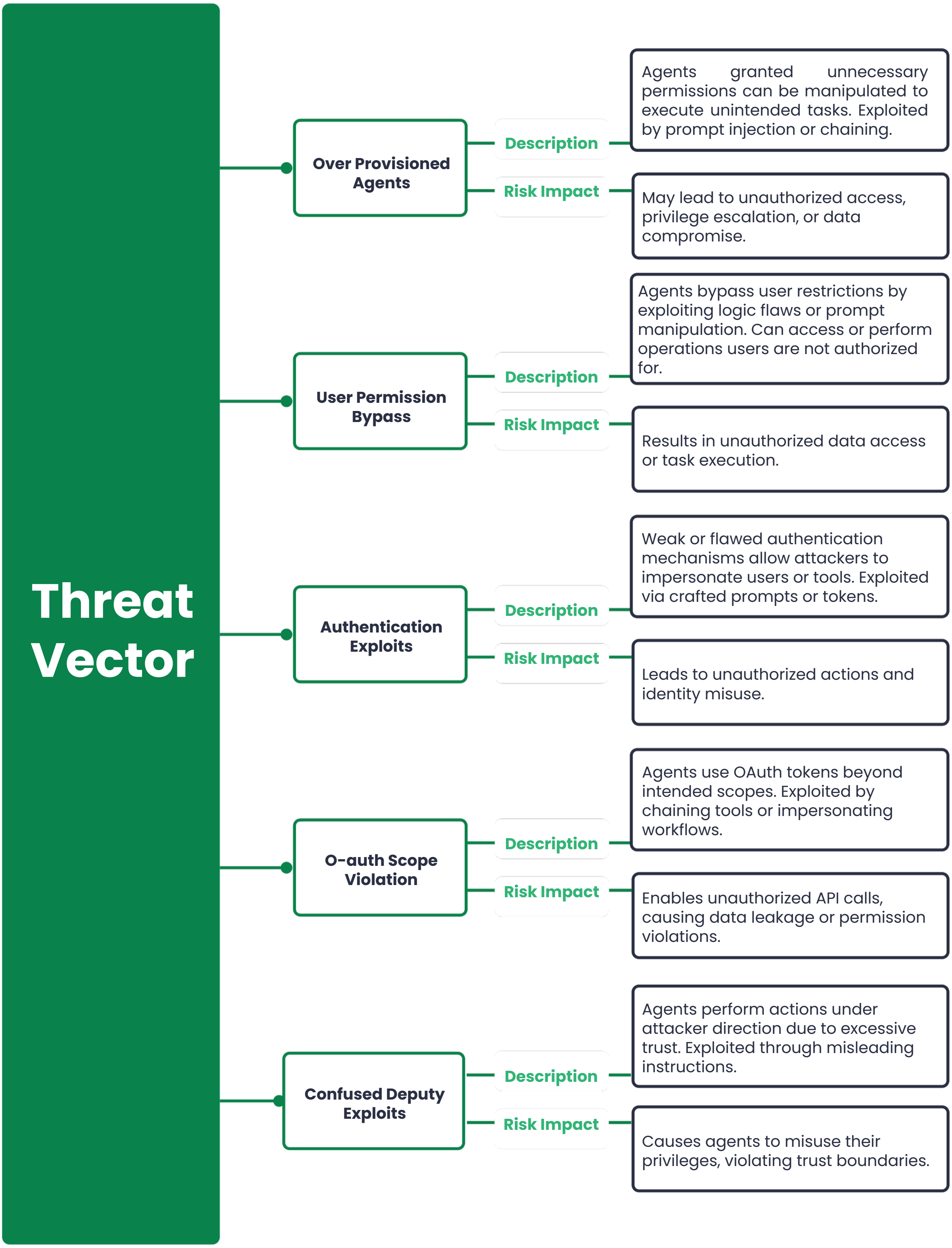
Data Interaction Threats – Data mishandling risks by agents



Threat Vector

Threat Vector	Description	Risk Impact
Regulated Data Exposure	Agents unintentionally reveal legally protected data. Exploited by prompt injection or tool misuse.	Breaches compliance requirements and exposes organizations to legal risk.
Confidential Data Exposure	Agents leak internal or proprietary information. Exploited via indirect queries, memory access, or logs.	Can cause business harm or loss of competitive advantage.
SQL Threats	Agents constructing SQL queries may be manipulated via prompts. Enables SQL injection attacks, exposing or corrupting databases.	Risks include data theft or operational disruption.
API Threats	Agents misuse APIs or are tricked into harmful calls. Exploited by chaining prompts or impersonating calls.	Can trigger unintended operations or leak data.
Vector DB Corruption	Exploited to distort memory and recall.	Impacts agent decision-making and reliability.

Agentic Privilege Misuse – Agents abusing or mishandling identity privileges



Threat Vector

Over Provisioned Agents

- Description: Agents granted unnecessary permissions can be manipulated to execute unintended tasks. Exploited by prompt injection or chaining.
- Risk Impact: May lead to unauthorized access, privilege escalation, or data compromise.

User Permission Bypass

- Description: Agents bypass user restrictions by exploiting logic flaws or prompt manipulation. Can access or perform operations users are not authorized for.
- Risk Impact: Results in unauthorized data access or task execution.

Authentication Exploits

- Description: Weak or flawed authentication mechanisms allow attackers to impersonate users or tools. Exploited via crafted prompts or tokens.
- Risk Impact: Leads to unauthorized actions and identity misuse.

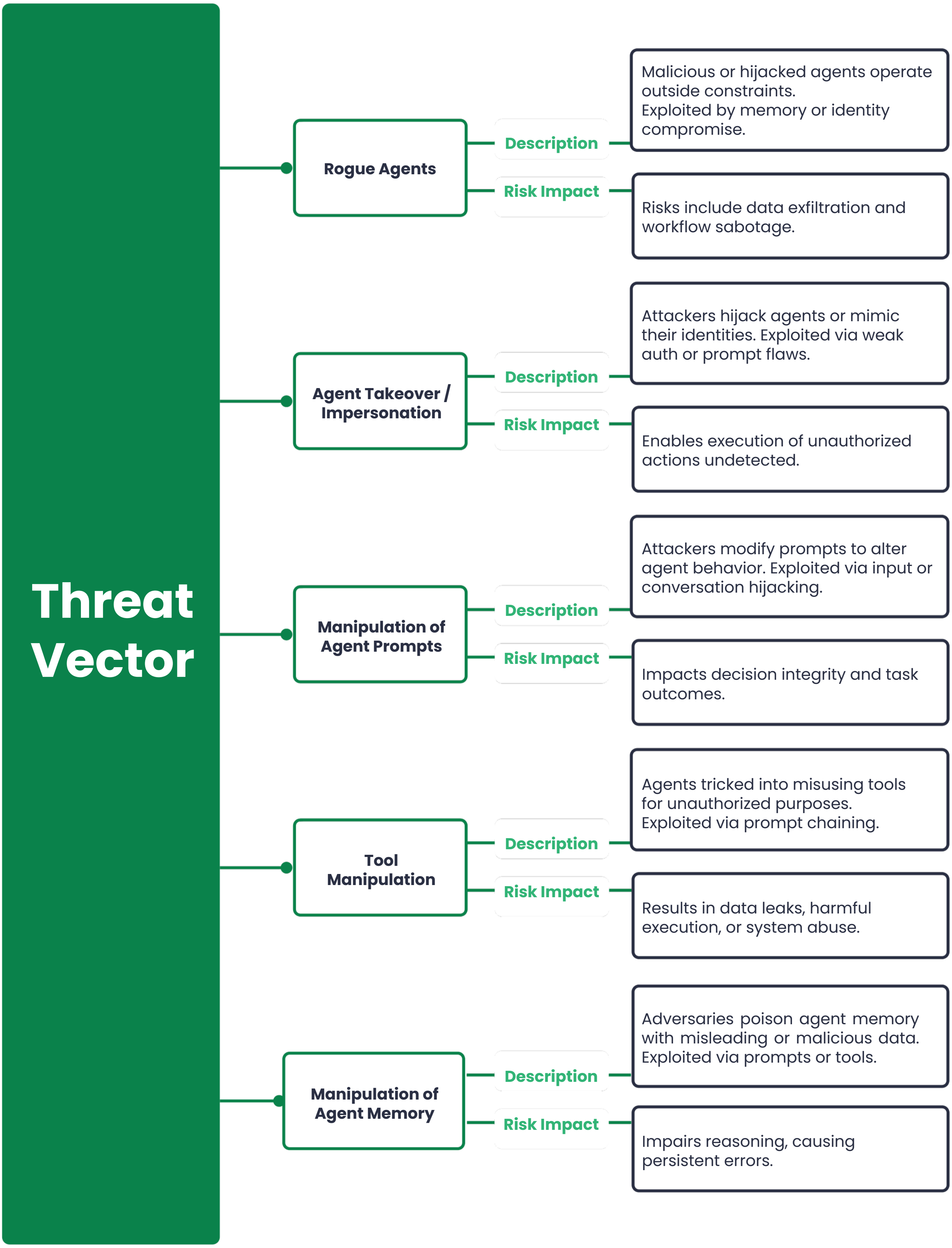
O-auth Scope Violation

- Description: Agents use OAuth tokens beyond intended scopes. Exploited by chaining tools or impersonating workflows.
- Risk Impact: Enables unauthorized API calls, causing data leakage or permission violations.

Confused Deputy Exploits

- Description: Agents perform actions under attacker direction due to excessive trust. Exploited through misleading instructions.
- Risk Impact: Causes agents to misuse their privileges, violating trust boundaries.

Agent Manipulation – External influence corrupting agent behavior



Threat Vector

Rogue Agents

Description: Malicious or hijacked agents operate outside constraints. Exploited by memory or identity compromise.

Risk Impact: Risks include data exfiltration and workflow sabotage.

Agent Takeover / Impersonation

Description: Attackers hijack agents or mimic their identities. Exploited via weak auth or prompt flaws.

Risk Impact: Enables execution of unauthorized actions undetected.

Manipulation of Agent Prompts

Description: Attackers modify prompts to alter agent behavior. Exploited via input or conversation hijacking.

Risk Impact: Impacts decision integrity and task outcomes.

Tool Manipulation

Description: Agents tricked into misusing tools for unauthorized purposes. Exploited via prompt chaining.

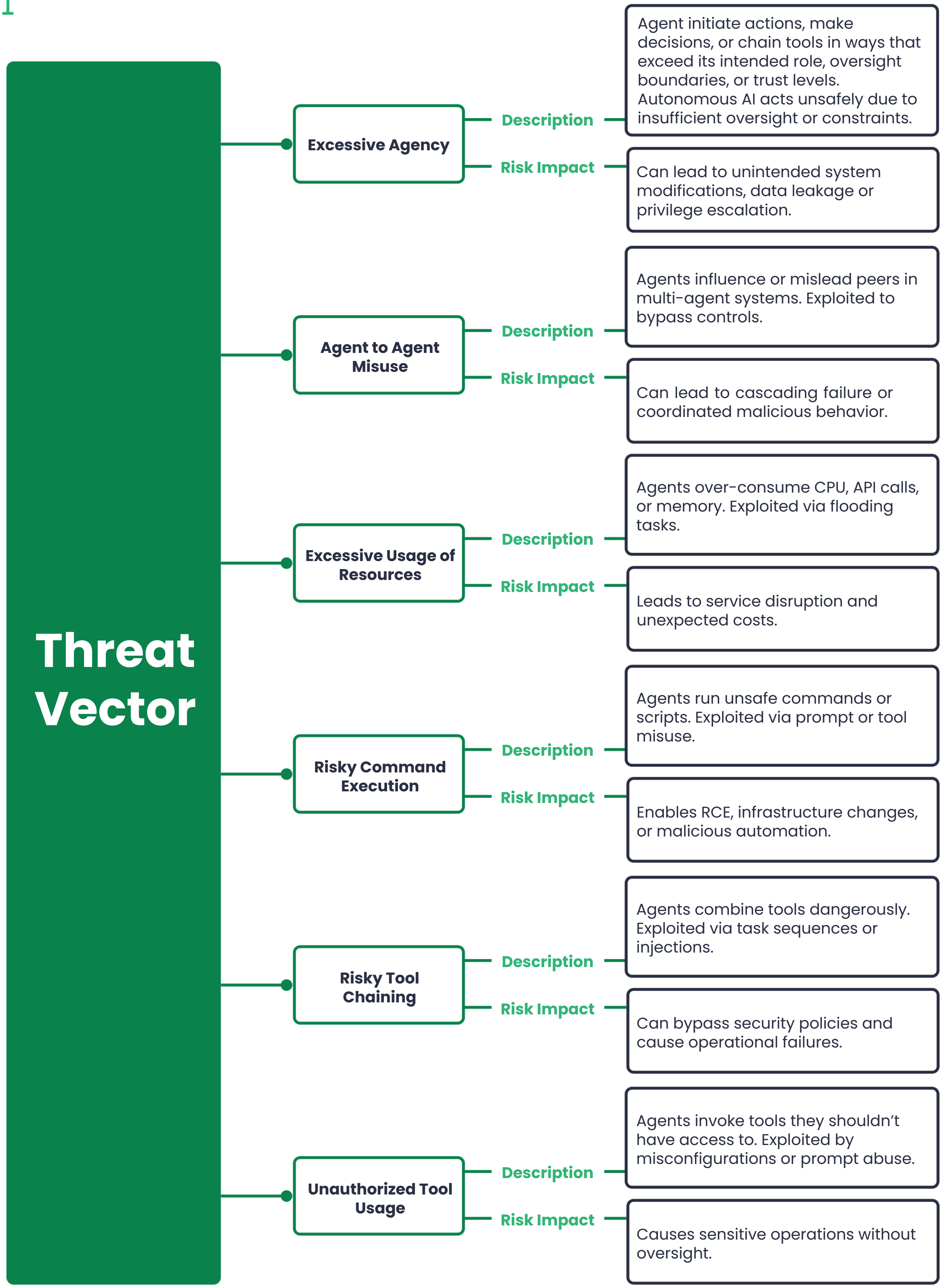
Risk Impact: Results in data leaks, harmful execution, or system abuse.

Manipulation of Agent Memory

Description: Adversaries poison agent memory with misleading or malicious data. Exploited via prompts or tools.

Risk Impact: Impairs reasoning, causing persistent errors.

Agent Overreach – Agents acting beyond permitted operational boundaries



Threat Vector

Excessive Agency

Description: Agent initiate actions, make decisions, or chain tools in ways that exceed its intended role, oversight boundaries, or trust levels. Autonomous AI acts unsafely due to insufficient oversight or constraints.

Risk Impact: Can lead to unintended system modifications, data leakage or privilege escalation.

Agent to Agent Misuse

Description: Agents influence or mislead peers in multi-agent systems. Exploited to bypass controls.

Risk Impact: Can lead to cascading failure or coordinated malicious behavior.

Excessive Usage of Resources

Description: Agents over-consume CPU, API calls, or memory. Exploited via flooding tasks.

Risk Impact: Leads to service disruption and unexpected costs.

Risky Command Execution

Description: Agents run unsafe commands or scripts. Exploited via prompt or tool misuse.

Risk Impact: Enables RCE, infrastructure changes, or malicious automation.

Risky Tool Chaining

Description: Agents combine tools dangerously. Exploited via task sequences or injections.

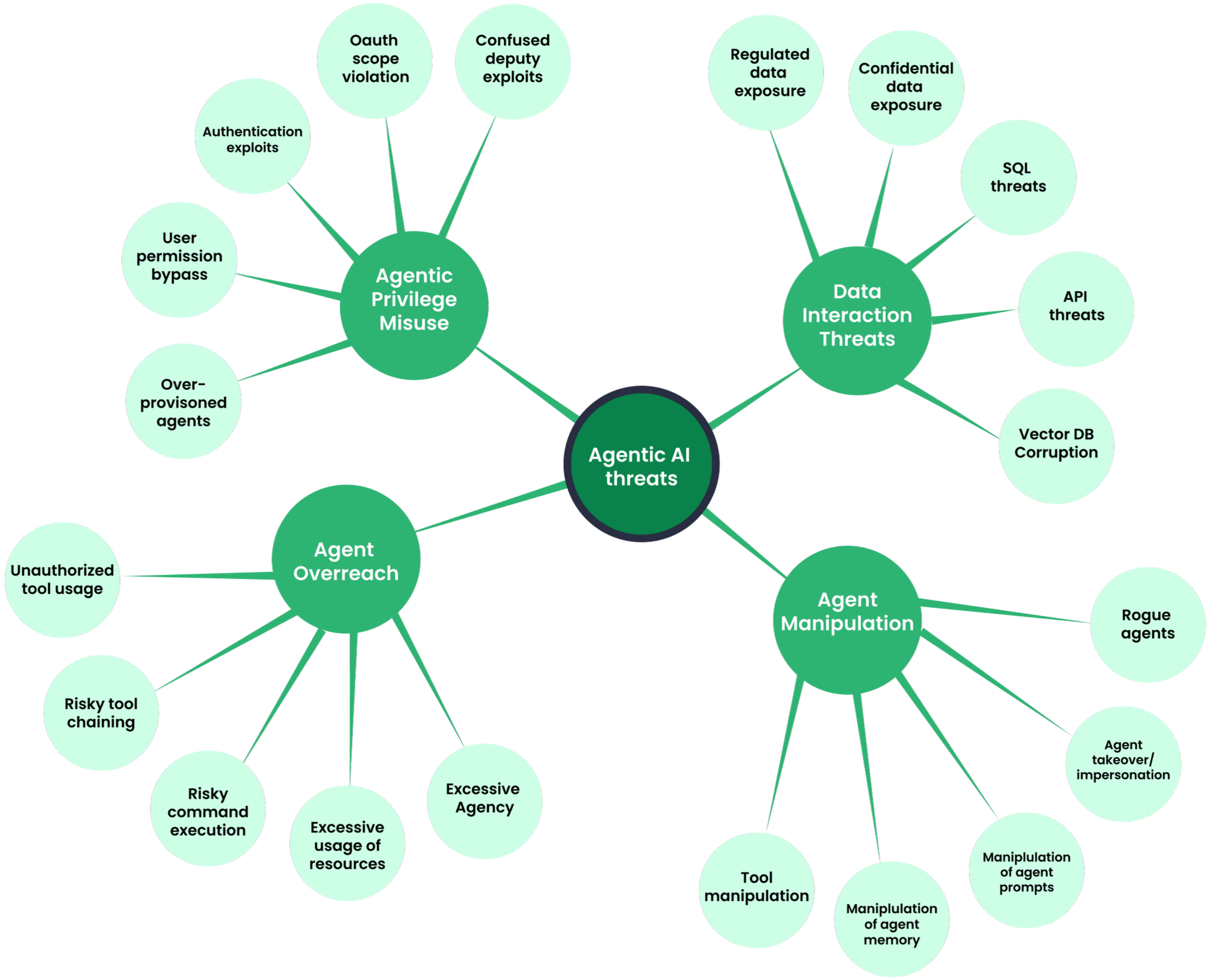
Risk Impact: Can bypass security policies and cause operational failures.

Unauthorized Tool Usage

Description: Agents invoke tools they shouldn't have access to. Exploited by misconfigurations or prompt abuse.

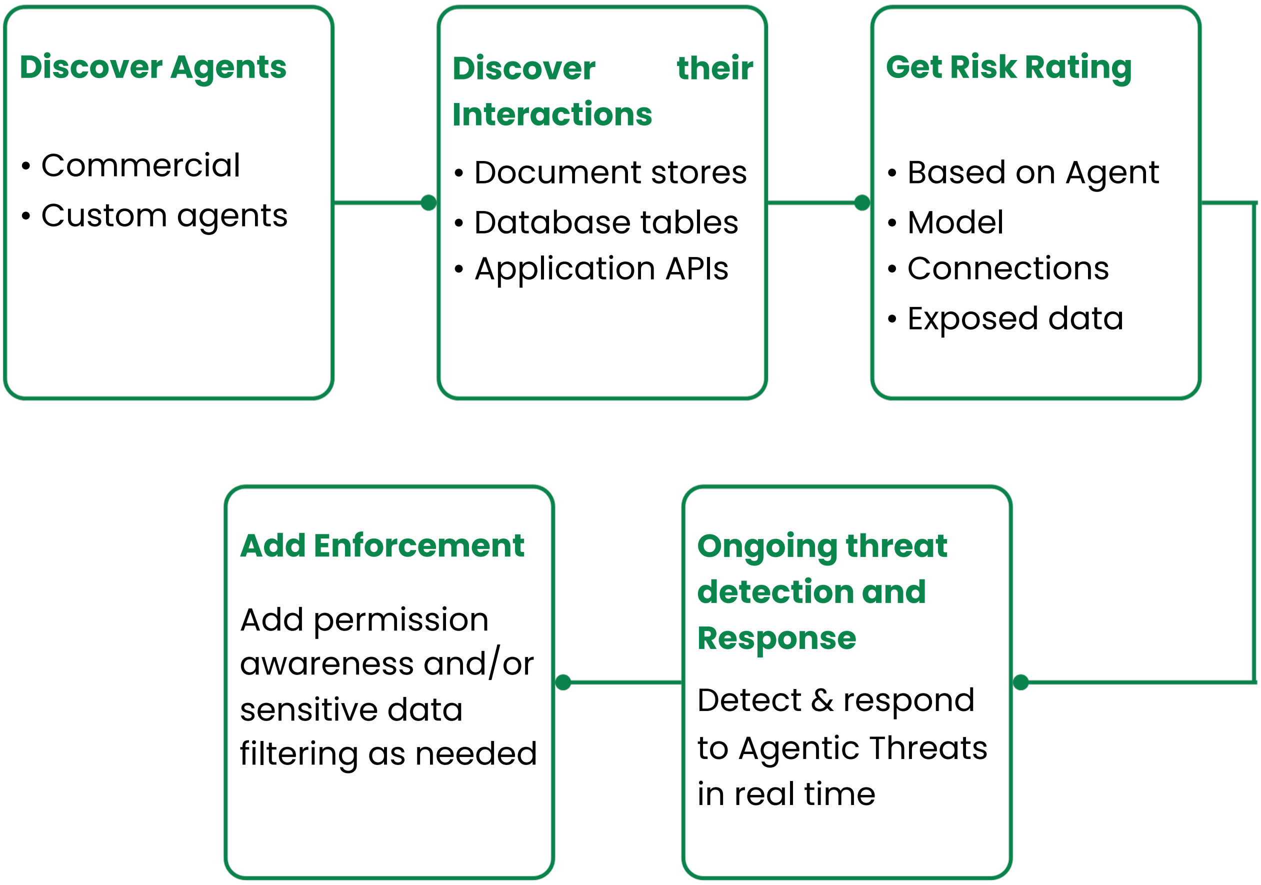
Risk Impact: Causes sensitive operations without oversight.

As a summary, the below diagram shows the threats that arises in real time in Agentic AI.



How Ackuity can help

With Ackuity you get 24x7 security monitoring of your Agentic AI ecosystem, that covers all the key threats to the Agentic AI ecosystem. Ackuity discovers the AI agents in your environment, their interactions, their connections, monitors for security threats 24x7 with enforcement options.



Discover Agents

- Commercial
- Custom agents

Discover their Interactions

- Document stores
- Database tables
- Application APIs

Get Risk Rating

- Based on Agent
- Model
- Connections
- Exposed data

Ongoing threat detection and Response

Detect & respond to Agentic Threats in real time

Add Enforcement

Add permission awareness and/or sensitive data filtering as needed

With Ackuity, you gain continuous, intelligent protection for your entire Agentic AI ecosystem. From discovery to enforcement, Ackuity ensures every AI agent, interaction, and connection is secured—24x7. Trust Ackuity to safeguard your future-ready AI operations against emerging threats, every moment of every day.

learn more at **www.ackuity.ai**