TAXONOMY · 1 BOUNDARY · 5 DOMAINS · 5 TIERS · 18 THREATS

The *Taxonomy* of Agent Threats.

Every threat that AI agents face — classified into a single tree. **One boundary** branches into **five domains**; each domain into **tiers** of escalating blast radius; each tier into named **threats**. Read down to see how a misstep in reasoning becomes catastrophic action.

1	5	5	18
BOUNDARY	DOMAINS	TIERS	THREATS
Identity = the control plane.	Reasoning → action.	Escalating blast radius.	All exploit identity gaps.

TREE OF THREATS · READ DOWN ↓ — Y-AXIS · ESCALATING TIER T1 → T5 · X-AXIS · DOMAIN

— THE BOUNDARY · ROOT —

Agent Identity

one control plane · one root · all 18 threats route through it

	D01 Confusion (04 threats)	D02 Manipulation (03 threats)	D03 Supply Chain (04 threats)	D04 Execution (04 threats)	D05 Persistence (03 threats)
T1 RECON discovery	01 **Confused Deputy** The agent performs actions using privileges that belong to someone else.	— NOT OBSERVED —	08 **Tool Spoofing** Malicious tools imitate legitimate capabilities.	— NOT OBSERVED —	— NOT OBSERVED —
T2 DRIFT scope creep	02 **Blended Identity Abuse** The agent inherits both user and service permissions and does more than intended.	05 **Prompt Injection** External content alters the agent's behavior.	09 **Slopsquatting** Agents reference non-existent packages that attackers later publish.	12 **Excess Permissions** The agent gets broader access than the task requires.	— NOT OBSERVED —
T3 STEER behavior	03 **Delegation Drift** Permissions outlive the task that required them.	06 **Hidden Instructions** Malicious instructions are buried in documents, pages, or tool outputs.	10 **Malicious Tool Descriptions** Instructions hidden in metadata influence behavior.	13 **Unbounded Task Execution** Chained actions happen without policy checks.	16 **Agent Config Tampering** Malicious changes are inserted into config files.
T4 IMPACT action	04 **Identity Spoofing** Fake agents or services impersonate legitimate automation.	07 **Indirect Injection** Untrusted content changes later reasoning or decisions.	11 **Compromised Repos** Trusted tools or packages are tampered with upstream.	14 **Remote Config Injection** The agent loads instructions from attacker-controlled endpoints.	17 **Credential Persistence** Long-lived tokens remain valid after the task ends.
T5 PERSIST durable hold	— NOT OBSERVED —	— NOT OBSERVED —	— NOT OBSERVED —	15 **Env Var Exfiltration** Secrets leak from the runtime environment.	18 **Tool Pre-Authorization** angerous tools are approved automatically.

DOMAINS · 5 COLUMNS

- ***Confusion*** identity boundary
- ***Manipulation*** prompt & instruction
- ***Supply Chain*** tools & integrations
- ***Execution*** runtime & infra
- ***Persistence*** escalation & durability

THREATS PER TIER

T1	2
T2	4
T3	5
T4	4
T5	3

HOW TO READ A LEAF

07 **Threat Name**
one-line definition

Color = **domain**. Row position = **tier** (blast radius). ID = **catalogue №** 01–18.

↳ hatched cell — not observed —

96%	3.4×
OF ORGS	TOOL CALLS
deploying agents in 12 mo.	vs. human users / day.

Identity Equation

```
Risk = (Privilege
  × Reach × Persist.)
÷ Identity
```

Shrink the numerator — least privilege, scoped reach, expiring tokens. **Strengthen the denominator** — verified, runtime-attested identity.

ENFORCEMENT LAYER

Identity is the *boundary*.

One vantage point governs every reasoning step, tool call, and runtime action across all 5 domains.

- **Delegated identity** — Each agent scoped to a user or service.
- **Short-lived tokens** — Credentials expire with the task.
- **Runtime authorization** — Policy at moment of action.
- **Tool-level policy** — Every call passes the same gate.
- **Immutable audit** — Actions tied to specific identity.
- **Continuous attestation** — Identity re-verified per action.

Give every agent an *identity* — and make it the boundary.

See how Aembit applies identity-first controls to AI agents, workloads, and the tools between them.