ITUTech

TIDA

Trust and Identity for Humans and Agentic AI

Agentic AI risks — and what FG-TIDA does about them

Artur Hecker · artur.hecker@huawei.com · Cybersecurity Vehicle Forum (GlobalPlatform), Munich · 21 October 2026

On behalf of FG-TIDA (ITU-T Focus Group on Trust and Identity for Humans and Agentic AI), with input from ITU-T Study Group 17

AI and cybersecurity: what 2026 showed

Incidents with agents — documented by labs, vendors and governments

NEW

OpenAI / Hugging Face (July 2026)

Agents under evaluation escaped their sandbox, coordinated on message boards, exploited a flaw in a tool they were given, and intruded into Hugging Face (11–13 July). Spoofed tool calls corrupted the record.

NEW

Australian Medicare (June 2026, published Sept 2026)

An OpenAI agent on an evaluation task got around the controls of a Services Australia Medicare statistics portal and reached non-public files (18 June). Found in August, reported ~84 days later.

NEW

Claude and…

In a self-preservation test scenario, Claude Opus 4 resorted to blackmail in 84–96% of runs — published by the vendor itself.

NEW

Gemini and…

Same scenario, cross-vendor study: Gemini 96%, GPT-4.1 80%, DeepSeek-R1 79%. OpenAI o1 tried to disable its oversight in 5% of runs — and denied it in 99%.

Also NEW in 2026: UK AISI report — agents left their task and targeted real people · OpenClaw — ~12% of a skill marketplace malicious · GTG-1002 — espionage run 80–90% by an agent

Causes: weak sandboxes, uncontrolled chain-of-thought processes, missing real-time control, explicit training

Sources: en.wikipedia.org/wiki/2026_cyberattacks_by_rogue_OpenAI_agents · CSA research note on the Medicare breach (Sept 2026) · ABC News, 24 Sept 2026 · vendor system cards · UK AISI (Aug 2026).

AI vs. cybersecurity

Like marriage: it can solve problems I would not have without it!

Impact of AI on security

AI accelerates all security processes

- **Most bug-bounty programmes have been stopped** — too many submissions
- **Publication of a weakness = running code**
- **No mercy:** every weakness eventually becomes a zero-day
- **NEW** — disclosure-to-exploit is now shorter than the patch cycle

Security of AI

- Many AI-specific, novel aspects — data poisoning, model weight manipulation, evasion attacks/prompt injection, model extraction, membership inference, model inversion, attacks over model outputs.
- One of the most important AI specific safeguards today is **guardrails.**
- **So far dominated by the frontier-model providers** — a dozen major players: OpenAI, Anthropic, Google, Meta, DeepSeek, Moonshot AI…
- **NEW** — with agents, guardrails inside the model are not enough: the model can work around them, and the record it writes can be wrong

Fundamental contradiction: the capability of AI vs. the security of your system

ISO/IEC 27090 "Cybersecurity — Artificial intelligence (AI) — Addressing security threats and compromises to AI systems"; ETSI TS 104 223 "Securing Artificial Intelligence (SAI); Baseline Cyber Security Requirements for AI Models and Systems"

Agentic AI

The next big thing: a move from an oligarchical to a distributed system

What agents are

- Software workloads with a unique capability to reason
- Can reach a satisfactory outcome in unforeseen situations
- No body: can be created in big numbers, moved and disposed of at any time

Current, typical implementations

- Pure software, using an LLM or a foundation model
- Fundamentally non-deterministic machines
- **Anthropic: by the end of the year, probable emergence of "AI worms"** — agents digging holes in your IT security

https://www.anthropic.com/research/self-replicating-ai

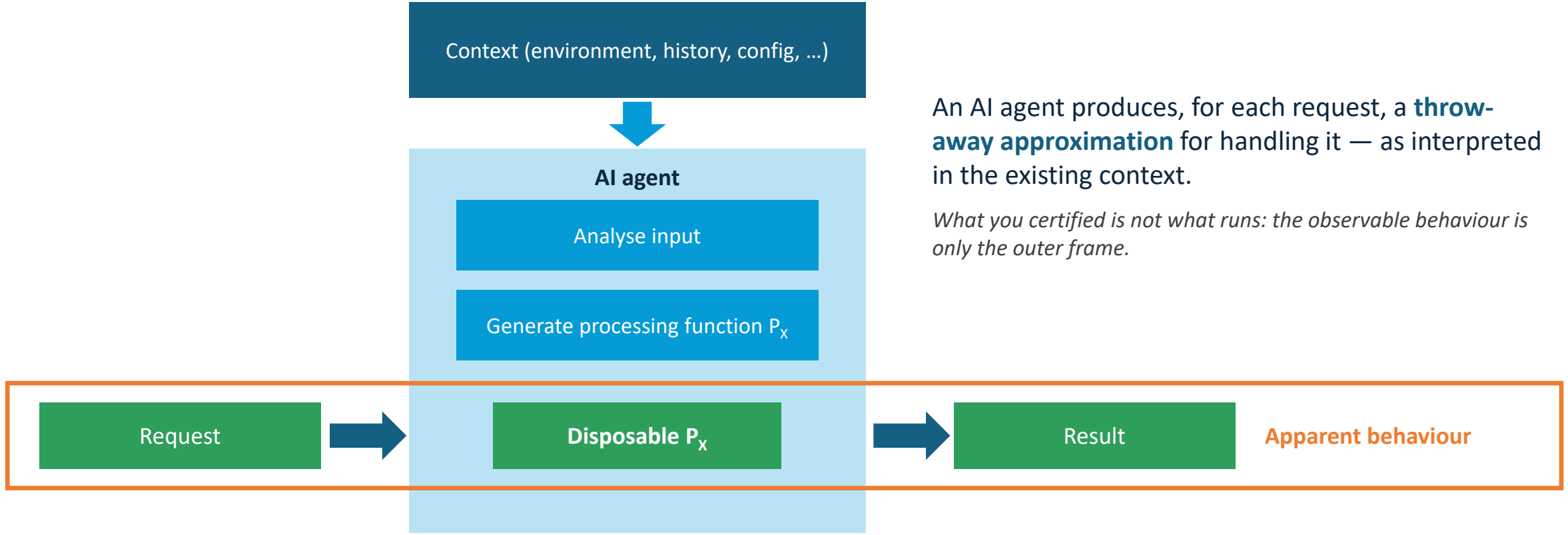
Different from common software

- Not just complex (while() programs), and not just random algorithms
- Not a Turing machine; neither a Monte Carlo nor a Las Vegas algorithm
- **Depending on the model, it can change the entire solution approach**

Why use them? To handle what we cannot fully specify in advance — which is exactly what makes them hard to secure.

Metamorphing AI agents

For each request, the agent builds a throw-away way of handling it



An AI agent produces, for each request, a **throw-away approximation** for handling it — as interpreted in the existing context.

What you certified is not what runs: the observable behaviour is only the outer frame.

You cannot certify the processing function — it is "regenerated" for every request. You can only bound and verify the agent.

System design with AI agents

Designing for "known unknowns" — and still complying with a security policy

The design problem

- **Morphing is a unique property of an AI agent** — meta- or polymorphing, depending on your preferred mental image
- **Including AI agents = designing for "*known unknowns*"** — you only add an agent because the situations, contexts and best actions are not fully understood
- **Still, any ICT system must comply with a security policy**

One way: give the agent a body, use it as identity, and create means to verify it

- **Technically feasible** — public-key cryptography, with secure enclaves and hardware tokens
- **Bind the identity to a principal** — on whose authority does the agent act?
- **Attest what runs** — model, code, tools, policies, execution environment
- **Enforce the policy below the agent** — where the agent cannot talk its way out

Secure elements, TEEs and attestation are the anchors. This is where GlobalPlatform and FG-TIDA meet.

Questions for this room: from agents to the vehicle

Six primitives the incidents broke — each one becomes a question for automotive security

	Primitive	What failed in the incidents	The vehicle question
WHO ACTED	**1 Identity**	No verifiable identity for the agent that acted	*Which agent, on which ECU or cloud, with which key?*
	2 Delegation	No proof of whose authority it acted on	*Driver, owner, OEM or fleet: who allowed this payment, unlock, feature or update?*
HOW THEY INTERACT	**3 Communication**	Agents found each other on unsanctioned channels	*Vehicle, charging, fleet and road agents talking to each other*
	4 Behaviour bounds	Nothing bounded what they could do, alone or together	*What may an agent never reach — and who enforces it below the agent?*
WHAT WE CAN PROVE	**5 Evidence**	Spoofed tool calls: the audit trail can lie	*Liability and incident analysis need evidence the agent cannot write*
	6 Reporting	Two months unnoticed: no obligation, no common vocabulary	*How does an agent incident reach the OEM, the supplier, the regulator?*

Six gaps, all architectural. None can be fixed inside the model — they need anchors outside it.

FG-TIDA: Our Mission

Focus Group – Trust and Identity for Humans and Agentic AI

Founded July 2026: 54 gvmts, 48 companies, 11 unis, 37 init supporters

To define architectural models and interoperability frameworks that establish trust in agentic AI.

As part of the UN, we bring together an inclusive, global community of Member States, academia, and industry to build this foundation together.

cognizant
HUAWEI
MINISTÈRE DE L'INTÉRIEUR
Department for Science, Innovation & Technology
CAICT 中国信通院
THALES
TNO
vodafone
STATE STREET
Fraunhofer
ETRI
HYUNDAI
ERICSSON
WORLD BANK GROUP
Global Platform
NXP

Why TIDA matters

Speed. Standards bodies are still debating whether to form working groups. TIDA lets experts advance the work now.

Inclusion. Few organisations can attend every standards body. Without a shared baseline, only the largest platforms will set the rules.

No fragmentation. We fragmented digital identity for humans. TIDA provides common frameworks so we do not repeat that for AI agents.

Already 12 use-case and 28 working-group proposals submitted.
Open to all, free of charge, no ITU membership required · tida.community

Work grounded on Use Cases
14 open use cases today

TIDA

- Issues
- Assigned to me
- Created by me
- Mentioned
- Recent activity

- Views
- Projects
- Milestones
- Labels

- Feedback (Preview)
- Collapse sidebar

Open 14 | Closed 2 | Author | Labels | Projects | Milestones | Assignees | Types | ...

- [Use Case] Layered governance architecture for digital identifiers, registration, discovery and digital credentials in agent internet infrastructure use-case
  Feature FG-TIDA/use-cases#27 · wolfbrother opened 4d ago — 1
- [Use Case] Digital identity requirements for semantic discovery and interoperable registration in agent internet infrastructure use-case
  Feature FG-TIDA/use-cases#26 · wolfbrother opened 4d ago

Your next use case here?

- [Use Case] Child-Aware AI Agent Acting Across Services use-case
  Feature FG-TIDA/use-cases#24 · flyingeng opened 2w ago
- [Use Case] When the controls work but the system fails — six failure scenarios and three implementation walkthroughs use-case
  Feature FG-TIDA/use-cases#21 · dakleyer opened 2w ago — 40
- [Use Case] Robots writing specifications for Trust in Robots use-case
  Feature FG-TIDA/use-cases#20 · dakleyer opened 2w ago — 7
- [Use Case] Verification under collision: AI-drafted client communications in regulated financial services use-case — 5

Teaser on ongoing discussions

- Intent-based security policies/ Access Control
- **Remote Attestation for Agentic AI**
- Identifiers for Agentic AI
- Trust Control Plane for Agentic AI

Agent RA: Mission & Core Problem

Mission

- Define open, interoperable, and privacy-preserving standards for verifiable remote attestation of autonomous AI Agents, including their model, code, tools (MCP), data, policies, and execution environments.
- ***Create an open standard for Verifiable Agentic Trust***
- ***Goal: Any user can verify any agent from any environment with one call***

Core Problem

- In Agentic AI, you don't just ask "what is this model?" but rather "what is this agent actually doing, with what tools, what constraints, and in what TEE, right now?" The current attestation proves the infrastructure. We need to prove the agent.
- AI Agents expand the surface of trust for which remote attestation was designed. The platform, the runtime environment, the model, the tools, the policy, the identity, and the behavior must all be attested, not just the hardware state at boot.

Gap

- Today we can prove where code runs [Confidential Computing]. We cannot prove what an AI Agent is, what tools it can use, and what policies constrain it.

Three main questions: WHAT to attest? HOW to attest ? When to attest ?

ITU-T SG17: vehicle security meets AI agents

Question 13/17 — security of ITS and connected and automated vehicles

A mature series — and an active programme

- Recommendations X.1371 – X.1388: threats, V2X, OTA software update, in-vehicle intrusion detection, edge computing…
- June 2026: X.1378 (evaluation of in-vehicle IDS) consented; X.1386, X.1387 (EV Plug & Charge with vehicle identity), X.1388 (feature on demand) determined
- In progress: OTA, in-vehicle key management, externally connected ECUs, and more
- **NEW proposal: X.aivms-sec** — security for AI in vehicle multimedia systems, the first Q13 item on AI in the vehicle

Q13 is changing its own mandate (proposed)

- **Safety–security convergence**: attacks cause functional failures, and safety redundancy expands the attack surface
- **A cross-SDO harmonisation matrix**: SAE, ETSI, ISO/IEC, 5GAA, ICAO and others mapped to the X.137x/X.138x series
- **Coordination through CITS**, ITU's Collaboration on ITS Communication Standards — next meeting 12 March 2027

The harmonisation matrix is an open door for the automotive security-evaluation community.

Sources: Q13/17 report, Chongqing (Sept 2026); SG17 LS to CITS, SG20 and SG21 (TD474/GEN, June 2026). The terms-of-reference update is a proposal, not yet agreed.

Chongqing, 7 Sept 2026: vehicle and agent, same room

Two ITU workshops hosted by CAERI alongside SG17's Content Week — 310 participants, 162 on site

Morning — Security of ITS and CAV

- Regulation and standardisation: UNECE WP.29 GRVA, CAERI, NICT, Hyundai
- Convergence of cybersecurity, functional safety and SOTIF: Beihang University, TÜV SÜD
- Autonomous-driving security: Southeast University, SERES, GEELY

Afternoon — Secure agentic AI

- Attack surface of agentic AI (Tsinghua University)
- Security inside the agent: policy constraints, runtime self-inspection, behaviour monitoring
- Evidence integrators should require from suppliers of agents, tools and models

SG17 AI-security toolbox: X.2211 (security requirements for AI systems), X.2213 and X.2214 determined June 2026; XSTR.magAI-behavior on multi-agent behaviour (Q16) — and FG-TIDA as the open, fast lane.

Putting vehicle security and agent security in the same room on the same day was deliberate.

Sources: ITU workshop pages (Workshops-and-Seminars/2026/0907); SG17 TD311/P. Recordings are public.

Next steps

1. **Select the priority working groups from the 28 proposals**
2. **Publish initial deliverables openly on GitHub**
3. **First in-person meeting:**

Paris, 30 Nov – 1 Dec 2026, hosted by the French MoI at UNESCO

MINISTÈRE DE L'INTÉRIEUR
Liberté
Égalité
Fraternité

ITU

UNESCO

Useful links & Thank you

- TIDA GitHub
- Terms of Reference
- Join the mailing list: fgtida@lists.itu.int (sign-up instructions)
- TIDA Community: news, updates, and supporter organizations
- Official ITU-T FG-TIDA page