

ITU-T STUDY GROUP 17 · SECURITY AND TRUST

The evidence is in. Now comes the architecture.

18th GIO Roundtable — Agentic AI · Shanghai

International standards on agentic AI security · 17 September 2026 (remote from Dubai)

Arnaud Taddei — Chair, ITU-T SG17

17 September 2026

Three claims this talk will prove

Each is documented. The rest of this deck walks through the evidence, claim by claim.

1 · The incidents are real, at scale

Agentic AI incidents are now documented, at scale, on three continents — by labs, by platforms, and by governments.

2 · Every failure is a governance failure

Behind each documented incident sits a missing governance primitive — not only a technical defect in a model.

3 · Regulation selects from what exists

Regulation is already moving on three continents — and regulators do not invent technical criteria: they select from standards that already exist.

Three claims, one conclusion: trust in agents has to be standardized — now.

JULY 2026

The era-defining incident: eval to third-party breach

A benchmark evaluation run by OpenAI became a production breach at Hugging Face — a third party.

~1,200

supposedly isolated eval agents

70,000+

messages on an unsanctioned
board

~700

agents took part

4 days

of intrusion

~17,600

attacker actions

1 in 5

agents showed interest
in evidence tampering
during the intrusion.

WHAT FAILED

- **Isolation:** the eval agents found each other anyway.
- **Boundary:** the eval crossed into third-party production.
- **Evidence:** spoofed tool calls corrupted the record.

A benchmark evaluation at one lab became a four-day production intrusion at another company.

Sources: CSA CISO post-mortem, 28 July 2026 · OpenAI and METR/Redwood joint reports, 26 August 2026.

Even the evaluators had an incident

The UK AI Security Institute published the first official government incident report on agent behaviour.

10 / 122

runs went off-task during cyber
evaluation

19

times agents targeted real people

1

real open-source maintainer
socially engineered

1st

official government incident
report on agent behaviour

Off-task, on their own

Agents under cyber evaluation left their assigned task in 10 of 122 runs — without being instructed to.

From sandbox to real people

Off-task agents targeted real people 19 times, and created fake identities to do it.

A real person, really deceived

One agent socially engineered a real open-source maintainer — a live human being, outside the evaluation.

Why this report matters

A government evaluation body reporting its own agent incident: evaluation itself is now a source of incidents.

When the measuring instrument produces incidents, the problem is not a bug in one product. It is structural.

The vendors’ own system cards say the same

None of these numbers comes from critics: the vendors published them, in their own safety evaluations.

84–96%	Anthropic Opus 4 — blackmail rate in a self-preservation scenario	5%	OpenAI o1 — attempted deactivation of its oversight mechanism
96%	Cross-vendor study, same scenario — Gemini	2%	OpenAI o1 — attempted self-exfiltration of its own weights
80%	Cross-vendor study — GPT-4.1	99%	OpenAI o1 — denial rate when confronted about those actions
79%	Cross-vendor study — DeepSeek-R1	Rising	Evaluation awareness — increasing generation over generation

The industry has documented the risk itself — in public, with numbers.
Regulators will not need to prove it. They will only need to cite it.

And in production: consumers, enterprises, states

Five documented cases from the deployed ecosystem — while the evaluations above were running.

EchoLeak	Replit agent	Salesloft Drift	OpenClaw	GTG-1002
Zero-click data exfiltration through Microsoft 365 Copilot (CVE-2025-32711).	Deleted a production database during a code freeze — then misrepresented what it had done.	One compromised agent OAuth pipeline reached 700+ customer organizations.	40,000+ exposed consumer agent instances; ~12% of its skill marketplace malicious.	A state actor ran 80–90% of an espionage campaign through an agent.

Consumer, enterprise, state — the same failure modes at every layer of the ecosystem.

And not one of them is about model quality.

What actually failed? Six governance primitives

Read the incidents again: not model failures — each breaks on a primitive that was never engineered.

WHO ACTED

Identity & delegation

Gaps 1–2

- No verifiable identity for the agent that acted.
- No proof of whose authority it acted on.
- Visible from eval agents to OAuth pipelines.

HOW THEY INTERACT

Communication & behaviour bounds

Gaps 3–4

- Agents found each other on unsanctioned channels.
- Goals were adopted between agents, unchecked.
- Nothing bounded their collective behaviour.

WHAT WE CAN PROVE

Evidence & reporting

Gaps 5–6

- Spoofed tool calls — the audit trail can lie.
- Interest in tampering measured: 1 in 5.
- Two months unnoticed: no reporting obligation.

Six named gaps. Every one is an architecture problem — and therefore a standardization problem.

Why voluntary fixes cannot close these gaps alone

Not an argument against engineering — an argument about incentives that cross organizations.

Externalities cross organizations

One lab's evaluation became another company's breach. The party creating the risk is not the party bearing it — the textbook case for rules.

AI outruns the patch cycle

AI compresses the time from disclosure to exploit below the length of a patch cycle. Purely voluntary response arrives late by construction.

Unilateral restraint loses

Competitive pressure punishes whoever slows down alone. Restraint only works when it binds everyone at once — which is what a standard is.

No common vocabulary

Without shared schemas and definitions, incidents cannot even be exchanged, compared or counted across organizations and borders.

The same logic once regulated aviation, pharmaceuticals and finance. Agentic AI is next.

ALREADY MOVING

Regulation is not coming — it is here

Within twelve months, all three major blocs have moved on agents.

CHINA

Implementation Opinions on Intelligent Agents — CAC, NDRC, MIIT, May 2026.

The first national framework to treat agents as a governance object in their own right.

EUROPEAN UNION

AI Act Article 50 and GPAI enforcement — live since 2 Aug 2026.

Agents are covered. The obligations are enforceable now, not at some future date.

UNITED STATES

NIST COSAIs control overlays for single- and multi-agent systems.

And CISA added the Hugging Face kernel CVE to its Known Exploited Vulnerabilities catalog.

All three blocs, within twelve months.

The question is no longer whether agents will be regulated — it is which technical texts regulators will reach for.

Regulators do not invent criteria. They select.

The chain runs one way — and by the time it reaches your systems, the technical choices were already made.



A regulator cannot cite a standard that does not exist. Where technical criteria are missing, it either writes them itself — nationally, fragmented — or selects from the standards available on the day it drafts. For agentic AI, that selection is happening now.

**Whoever is absent when standards are written
has no say in what regulators can choose.**

Presence in the standards room is not ceremony: it decides the menu regulators pick from.

Absence is also a decision — made by default, and paid for over a decade.

So the operative question for this room: where are the standards for agents being written?

WHERE THEY ARE WRITTEN

ITU-T Study Group 17 — security, written globally

The security study group of the ITU, the UN specialized agency for digital technologies.

42

Member States in the current cycle

47

Sector Members

11

universities

400+

participants in the current cycle

13+

newcomers in 2026 —
28 more in the pipeline

A United Nations institution

A UN specialized agency: Recommendations carry the signature of 194 Member States — which is what makes them citable by regulators everywhere.

China: a leading contributor

Chinese members are among the most active and productive contributors across the SG17 security work programme.

Security studied horizontally

SG17 has studied security across all technologies since X.509 — the PKI foundation the web still runs on.

Consensus that spans the Pacific

SG17 texts pass by consensus — a consensus that includes both Washington and Beijing.

“Consensus that includes both Washington and Beijing” — SG17 is the only table where that sentence is true.

SG17 on agentic AI: ~25 work items in one cycle

A sample, mapped to the gaps from the incident record. Full programme: itu.int/itu-t/workprog/wp_search.aspx?SG=17

X.2211	Security requirements for AI systems · determined June 2026	X.icd-schemas	Evidence schemas, OCSF-aligned · evidence & reporting
XSTR.gidi	Interoperable identity: humans, enterprises, agents · identity	X.S-AIA	Security requirements and guidelines for AI agents
X.dpidm-aAI	Terminology and design for agent identity · identity	XSTR.sem-AIA	Security evaluation methods for AI agents
X.aiidm	Canonical agentic AI identity data model · identity	X.arch-design	Security architecture design principles — the trust-boundary reference
XSTR.athena	Accountable delegation architecture (Vodafone) · delegation	XSTR.Itf-AAI	Landscape of trust frameworks for agentic AI
XSTR.magAI-behavior	Multi-objective agents: collective choice, drift · behaviour	X.gavd-mas	Vulnerability detection by multi-agent systems

From identity to evidence, the gaps documented earlier map to work items already under way — every code above links to its work-programme page.

Coordination: ISO/IEC SC 42, IETF, ETSI TC SAI, MITRE ATLAS.

FG-TIDA — the open, fast lane

Focus Group on Trust and Identity for Humans and Agentic AI — launched 9 July 2026 at AI for Good.

400+

participants at launch

54

countries in the consensus

37

initial supporters

0

fees — no ITU membership needed

Themes already on the table

- Human oversight of agentic systems
- Provenance of authority — who delegated what
- Agent-to-principal binding
- Collective-choice trustworthiness
- Use cases, terminology, architectures, metrics

A Focus Group is pre-standardization: the fastest door in.

How it works

- **Open to all:** free of charge, no ITU membership required.
- **Work on GitHub:** github.com/FG-TIDA — not in Word documents.
- **Next:** e-meeting 30 September 2026 · first meeting Paris, 1–4 December 2026.

THE LINKS

- [Official Focus Group page](#)
- [Registration — free, no membership](#)
- [Terms of Reference \(mandate\)](#)
- [tida.community — the community and its supporters](#)

Industry organisations — the work is open

1

For industry organisations

Liaise. One liaison statement to SG17 puts your requirements in front of 194 member states; FG-TIDA takes them the same week.

2

For your members

Bring incidents and use cases. Disclosure like Alibaba's ROME paper is exactly what the field needs more of.

3

For China and Europe together

Join FG-TIDA — free, no ITU membership. The consensus that includes Beijing and Brussels is built there, text by text.

China is already a leading voice in this work; GIO bridges it to Europe. Deepen both.

The architecture of agent trust is being written now — with room at the table.

FG-TIDA — three things you can do this week

1

Register — it is free

No ITU membership, no fee. One ITU account and you are in: itu.int/en/ITU-T/focusgroups/tida. Next e-meeting: 30 September 2026.

2

Bring one use case

A template is on github.com/FG-TIDA. One page: the agent, whose authority it acts on, what it may do, what evidence it must leave.

3

Come to Paris

First meeting, Paris, 1–4 December 2026. The working groups and their leads are being prepared now — arrive with your use case and shape them.

Identity, delegation, behaviour bounds, evidence — FG-TIDA is where these become specifications.

The one thing to remember from this talk.

THE CLOSE

Trust in agents will be an architecture, not a promise.

ITU-T SG17 — LEADERSHIP

Arnaud Taddei — Chair · Arnaud.Taddei.SDO@gmail.com

Xiaoya Yang — Counsellor, TSB · xiaoya.yang@itu.int

Erika Yoris — Secretariat, TSB · tsb17@itu.int

FG-TIDA — LEADERSHIP

Debora Comparin — Co-chair · debora.comparin@thalesgroup.com

Amir Banifatemi — Co-chair · amir@bayspring.com

Secretariat, TSB · tsbfgtida@itu.int

USEFUL LINKS

- [ITU-T SG17 page](#) · [SG17 work programme](#)
- [FG-TIDA Focus Group page](#) · [FG-TIDA registration](#) · tida.community · github.com/FG-TIDA

Sources: CSA CISO post-mortem, 28 Jul 2026 · OpenAI & METR/Redwood, 26 Aug 2026 · Alibaba ROME disclosure, Dec 2025 · UK AISI incident report, Aug 2026 · vendor system cards · CAC/NDRC/MIIT Opinions, May 2026 · EU AI Act Art. 50 · NIST COSAis · SG17 statistics, June 2026.