

国际电信联盟电信标准化部门第17研究组·安全与信任

已然证据确凿， 接下来是架构建设。

第十八次GIO圆桌会议——Agentic AI；举办地：上海

Agentic AI安全国际标准·2026年9月17日（在迪拜远程参会）

Arnaud Taddei——国际电信联盟电信标准化部门第17研究组主席

2026年9月17日

本次演讲将论证的三个观点

各个观点均有据可查。接下来，本胶片将针对各个观点，逐一梳理其背后的证据。

1·安全事件真实存在，且规模巨大

当前，发生在三大洲的Agentic AI安全事件规模巨大，且有据可查，记录方包括多个实验室、平台和政府。

2·每一次事故均源于治理的失败

在每一起有据可查的安全事件背后，均存在一项缺失的治理基础能力——而并非仅是模型本身的技术缺陷所致。

3·监管选择现有标准执行

监管机构已经在三大洲展开行动。然而，监管机构并不会创造技术标准，他们只是选择现有标准进行执行。

三个观点均指向同一个结论：对AI智能体信任的标准化，已是当务之急。

2026年7月

标志性安全事件：基准评测演变为第三方数据泄露

OpenAI进行的基准评测演变成了第三方抱抱脸（Hugging Face）的生产环境数据泄露事件。

约1,200

个本应彼此隔离的评测AI智能体

70,000+

条通过未经授权的留言板交换的消息

约700

个AI智能体参与攻击行动

4天

入侵

约17,600

项攻击行动

五分之一

AI智能体在此次入侵期间表现出对篡改攻击证据的意图。

三道防线失效

- 隔离：评测AI智能体最终还是发现彼此并相互联系。
- 边界：评测环境最终越过了边界，进入了第三方的生产环境。
- 证据：伪造工具的调用破坏了记录。

一家实验室进行的一次基准评测，演变成了对另一家公司为期四天的生产环境入侵。

来源：云安全联盟首席信息安全官（CISO）事后复盘报告，2026年7月28日
OpenAI与METR/Redwood联合调查报告，2026年8月26日

甚至评测机构自身也发生了 “事故”

英国人工智能安全研究所（AISI）发布了首份关于AI智能体行为的官方政府安全事件报告。

10/122

网络安全评测运行出现了偏离任务范围的行为

19

次AI智能体针对现实世界人员的越界行为

1

名现实世界的开源项目维护者遭受了社会工程攻击

首份

关于AI智能体行为的官方政府安全事件报告

自行偏离任务范围

在并未收到相关指令的情况下，参与网络安全评测的AI智能体，在122次运行中有10次自行偏离了被指派的任务。

真实人员遭受现实欺骗

其中一个AI智能体对一名现实世界的开源项目维护者实施了社会工程攻击——对方是真实存在的人，而且身处评测环境之外。

越过沙箱攻击真实人员

偏离任务的AI智能体共19次将目标锁定为真实人员，并创建虚假身份来实施攻击。

为何这份报告很重要

一份政府评测机构发布的、关于其自身AI智能体安全事件的报告：评测本身已成为安全事件的源头之一。

当“测量工具”本身开始导致安全事件时，问题就不再是某一个产品的漏洞，而是结构性问题。

AI厂商自身的系统卡表明存在同样的问题

这些数字并非来自批评者，而是AI厂商自行公布的，来自其自身开展的安全评测。

84%–96%	Anthropic Opus 4——自保场景下的勒索行为发生率	5%	OpenAI o1——试图停用自身的监督机制
96%	跨厂商研究、相同场景——Gemini	2%	OpenAI o1——试图将自身的模型权重外泄
80%	跨厂商研究——GPT-4.1	99%	OpenAI o1——面对有关上述行为质询时的否认率
79%	跨厂商研究——DeepSeek-R1	呈上升趋势	对自身处于评测环境的感知能力——逐代增强

业界已通过数据自行记录风险，并对外公开。
因此，监管机构不必证明风险的存在，而只需引用即可。

在实际生产环境中：消费者、企业、国家

在上述评测进行的同时，已部署的生态系统中记录了以下五个案例：

EchoLeak	Replit Agent	Salesloft Drift	OpenClaw	GTG-1002
通过Microsoft 365 Copilot实现零点击数据泄露（CVE-2025-32711）。	在代码冻结期间删除了一个生产数据库，随后对其实际执行的操作作出不实描述。	一个遭到入侵的AI智能体OAuth授权凭证链路，波及了700多家客户组织。	4万多个消费者AI智能体实例易受攻击；约12%的AI智能体Skill市场内容被发现具有恶意性质。	一个国家级黑客组织，通过某个特定AI智能体执行了其80%至90%的网络间谍活动。

消费者、企业、国家——生态系统的每一层均存在相同的事故模式。
同时，这些事故无一与模型质量有关。

退一步看

到底哪里出了问题？——六项治理基础能力

再度审视上述安全事件：并非模型缺陷所致——每一起安全事件都暴露出一个从未被真正工程化的基础治理能力差距。

谁在执行 身份与委托

差距1-2

- 无法验证AI智能体的身份。
- 无法证实其行为所依据的授权。
- 从评测AI智能体到OAuth的授权凭证链路均可视化。

它们如何交互 通信与行为边界

差距3-4

- AI智能体通过未经授权的渠道发现彼此。
- 未经核实，不同的AI智能体便接受了共同的目标。
- 没有任何约束来限制它们的集体行为。

我们能证明什么 证据与报告

差距5-6

- 伪造工具调用——审计记录可能被篡改。
- 实测的存在篡改意图的占比：五分之一。
- 两个月未发现：不存在报告义务。

针对上述六项能力差距，每一项都是架构问题，因此也属于标准化问题。

为何自行修复无法单独弥补上述能力差距？

这并非反对工程化的观点，而是关于跨组织激励机制的讨论。

组织间的外部效应

一次实验室评测演变成了另一家公司的数据泄露。风险的制造者并非风险的承担者——这正是需要通过规则进行治理的典型情形。

AI演进速度超越补丁周期

AI将从漏洞披露到漏洞被利用之间的时间压缩到了补丁周期以内。仅依靠自发性响应，从机制上决定了其必然会滞后。

单方面克制无法奏效

竞争压力会惩罚任何单独放慢脚步的一方。克制仅在所有参与者同时受到约束时，才能发挥作用——而这正是标准存在的意义。

缺乏共同语言

如果没有统一的数据模式和定义，各组织乃至不同国家之间甚至无法有效地交换、比较或统计安全事件。

同样的治理逻辑，曾经规范过航空、制药和金融行业。接下来，该轮到Agentic AI了。

监管并非即将到来——已经开展行动了

在过去一年内，三大主要阵营——中国、欧盟和美国——均已开始推进针对AI智能体的相关行动。

中国

《智能体规范应用与创新发展实施意见》——国家互联网信息办公室、国家发展和改革委员会、工业和信息化部，2026年5月

这是首个将AI智能体本身作为独立治理对象的国家级治理框架。

欧盟

《人工智能法案》第50条及通用人工智能（GPAI）相关法规——自2026年8月2日起执行

该法案将AI智能体涵盖在内。相关义务如今已可强制执行，而非在未来某个时间点才开始执行。

美国

美国国家标准与技术研究院（NIST）发布的《人工智能系统安全控制覆盖层》——面向单智能体和多智能体系统的安全控制覆盖层

与此同时，美国网络安全与基础设施安全局（CISA）已将Hugging Face内核漏洞CVE纳入其“已知被利用漏洞”（Known Exploited Vulnerabilities, KEV）目录。

这三大主要阵营均在一年内完成了上述工作。

问题已不再是AI智能体是否会受到监管，而是监管机构将采用哪些技术标准和规范作为依据。

关键转折

监管机构不创造标准，而是选择标准

这条链路是单向传导的——当它最终触及您的系统时，技术层面的选择早已尘埃落定。



监管机构无法援引不存在的标准。当缺少技术标准时，监管机构要么自行制定——以国家为单位、彼此割裂；要么从其起草规则当日可获得的现有标准中进行选择。对于Agentic AI而言，监管机构已经开始作出选择。

在标准制定过程中缺席的任何一方，将无法对监管机构的选择产生任何影响。

参与标准制定绝非形式主义，而是决定未来监管机构可采用哪些技术标准。
缺席标准制定也是一种选择——默认的选择，并将为此付出了十多年的代价。

因此，对于在座各位而言，真正需要关注的问题是：AI智能体相关的标准正在何处制定？

国际电信联盟电信标准化部门第17研究组（安全领域研究组）——全球共同制定

国际电信联盟是联合国负责数字技术的专门机构，由其安全研究组牵头制定相关标准。

42

个现有成员国

47

个行业成员

11

所大学

400+

个现有参与者

13+

个今年新加入者——
另有28个正在筹备中

联合国机构

联合国专门机构：这些建议获得194个成员国的共同认可，这正是其能够被世界各地监管机构援引的重要原因。

中国：主要贡献者

在国际电信联盟电信标准化部门第17研究组工作计划中，来自中国的成员是最活跃且贡献最突出的参与者之一。

安全横向研究

国际电信联盟电信标准化部门第17研究组自发布X.509标准（Web至今仍依赖的PKI基石）以来，一直致力于研究各项技术的安全性。

跨越太平洋的共识

国际电信联盟电信标准化部门第17研究组的标准文件通过达成共识获得通过——达成这一共识的国家既包括美国，也包括中国。

“这一共识既包括华盛顿，也包括北京”——国际电信联盟电信标准化部门第17研究组是唯一让这句话成真的谈判桌。

国际电信联盟电信标准化部门第17研究组针对Agentic AI的标准化工作：一个工作周期内约25个项目

以下为工作项目示例，分别与安全事件记录中的能力差距相对应。完整工作计划链接：itu.int/itu-t/workprog/wp_search.aspx?SG=17

X.2211	人工智能系统安全要求——2026年6月确定	X.icd-schemas	基于开放网络安全模式框架（OCSF）的证据模式——证据与报告
XSTR.gidi	可互操作的身份：人类、企业与AI智能体——身份	X.S-AIA	AI智能体安全要求与指南
X.dpidm-aAI	AI智能体身份相关术语与设计——身份	XSTR.sem-AIA	AI智能体安全评测方法
X.aiidm	规范化Agentic AI身份数据模型——身份	X.arch-design	安全架构设计原则——信任边界参考
XSTR.athena	可问责的委托架构（Vodafone）——委托	XSTR.Itf-AAI	Agentic AI信任框架全景图
XSTR.magAI-behavior	多目标AI智能体：集体选择与偏离——行为	X.gavd-mas	多AI智能体系统漏洞检测

从身份到证据，此前记录的能力差距与正在开展的工作项目分别对应——上述各个代码均可链接到相应工作计划页面。

标准协调统一化：国际标准化组织/国际电工委员会第一联合技术委员会第42分委会（ISO/IEC JTC 1/SC 42）、互联网工程任务组（IETF）、欧洲电信标准协会（ETSI）人工智能与网络安全技术委员会（TC SAI）、MITRE ATLAS框架

人类与Agentic AI的信任与身份焦点小组（FG-TIDA）——开放的快速通道

人类与Agentic AI的信任与身份焦点小组，于2026年7月9日在“AI向善”（AI for Good）峰会上成立。

400+

个创始参与者

54

个国家达成共识

37

个最初支持者

零

费用——无需国际电信联盟会员资格

已探讨的主题

- 人类对智能体系统的监督
- 权限来源——由谁委托了何种任务
- AI智能体与责任主体绑定
- 集体选择的可信度
- 用例、术语、架构与指标

焦点小组属于标准化前阶段：这是进入标准制定流程最快的入口。

工作模式

- 面向所有人开放：免费加入，无需国际电信联盟会员资格
- 在GitHub上开展工作：github.com/FG-TIDA——而非使用Word文档。
- 下一步：于2026年9月30日举行线上会议——首次线下会议将于2026年12月1日至4日在巴黎举行。

相关链接

- [焦点小组官方网页](#)
- [注册——免费，无需国际电信联盟会员资格](#)
- [职责范围（授权）](#)
- tida.community——社区及其支持者

在座各位能做什么

行业组织——这项工作是开放的

1

行业组织

对接：向国际电信联盟电信标准化部门第17研究组提交一份联络声明，就能让您的需求呈现在194个成员国面前；人类与Agentic AI的信任与身份焦点小组则将在一周内解读相关需求。

2

行业组织会员

提供真实安全事件与应用案例。像阿里巴巴ROME相关论文这样的公开披露，正是该领域亟需的更多实践。

3

中国与欧洲

加入人类与Agentic AI的信任与身份焦点小组——免费，无需国际电信联盟会员资格。在这里，北京与布鲁塞尔之间的共识正在一份份标准文件中逐步建立。

中国在这项工作中已经处于领先地位；GIO圆桌则将中国与欧洲紧密连接起来，深化两者之间的合作。

AI智能体信任架构正在逐渐形成——参与制定的席位已设，虚位以待。

人类与Agentic AI的信任与身份焦点小组——大家本周可做的三项工作

1

注册——免费

无需国际电信联盟会员资格，也无需支付任何费用。只需拥有一个国际电信联盟账号，即可加入：itu.int/en/ITU-T/focusgroups/tida。下一次线上会议：2026年9月17日。

2

提供一个用例

模板可在github.com/FG-TIDA获取。一页即可，内容可包括AI智能体、其行使谁的权限、可执行的操作以及必须留下的证据。

3

奔赴巴黎

首次线下会议将于2026年12月1日至4日在巴黎举行。工作组及其负责人目前正在筹备会议——请携带您的用例到场，从而共同确定今后的工作方向。

身份、委托、行为边界与证据——人类与Agentic AI的信任与身份焦点小组正是将这些转化为规范的组织。

这是整场分享中唯一需要铭记的一点。

对AI智能体的信任将建立在架构之上，而非停留于承诺

国际电信联盟电信标准化部门第17研究组——牵头组织

Arnaud Taddei——主席；电子邮箱：

Arnaud.Taddei.SDO@gmail.com

杨晓雅——国际电信联盟电信标准化部门顾问；电子邮箱：

xiaoya.yang@itu.int

Erika Yoris——国际电信联盟电信标准化部门秘书处；电子邮箱：

tsbsg17@itu.int

相关链接

- [国际电信联盟电信标准化部门第17研究组官网·国际电信联盟电信标准化部门第17研究组工作计划](#)
- [人类与Agentic AI的信任与身份焦点小组官网·人类与Agentic AI的信任与身份焦点小组注册·tida.community.github.com/FG-TIDA](#)

来源：云安全联盟首席信息安全官事后复盘报告，2026年7月28日

OpenAI与METR/Redwood联合调查报告，2026年8月26日

阿里巴巴ROME相关披露，2025年12月

英国人工智能安全研究所安全事件报告，2026年8月

AI厂商系统卡

《智能体规范应用与创新实施意见》——国家互联网信息办公室、国家发展和改革委员会、工业和信息化部，2026年5月

欧盟《人工智能法案》第50条

美国国家标准与技术研究院发布的《人工智能系统安全控制覆盖层》

国际电信联盟电信标准化部门第17研究组统计数据，2026年6月

人类与Agentic AI的信任与身份焦点小组——牵头组织

Debora Comparin——联合主席；电子邮箱：

debora.comparin@thalesgroup.com

Amir Banifatemi——联合主席；电子邮箱：amir@bayspring.com

国际电信联盟电信标准化部门秘书处；电子邮箱：tsbfgtida@itu.int