论坛报告

AI 愿景论坛 巴黎 2026

构建人机协同

AI 愿景论坛于 2026 年 5 月 4 日在巴黎召开, 与 GOSIM 巴黎 2026 联合举办,全程遵循查塔姆研究所规则。 本报告对论坛议程、主题与建议作出综述。

论坛日期

2026 年 5 月 4 日 · 法国巴黎

发布

2026 年 5 月

paris2026.visionforum.ai

目录

前言

2026 年 5 月 4 日(星期一),即 GOSIM 巴黎大会前夕,一群来自世界各地的研究者、创业者、基金会负责人、监管机构代表与教育工作者齐聚巴黎,共同参与了 **AI 愿景论坛(AI Vision Forum)**。论坛为受邀制小型闭门会议,全程遵循查塔姆研究所规则(Chatham House Rule)。论坛的主题词——构建人机协同 *(Architecting Human–AI Synergy)*——由主持人在开场时归结为一个看似简单的问题:

> “当 AI 智能体日益自主、深度嵌入并日益重要时,真正应对其行为负责的是谁?我们如何打造能够赢得信任而非要求信任的智能体?”

本报告将当天的四场分论坛——**智能体型 AI 系统(Agentic AI Systems)**、**AI 与教育(Agentic AI in Education)**、**可信智能体型 AI(Trusted Agentic AI)**,以及 **Open Token 与数字公共物品(Open Token & Digital Public Goods)**——浓缩为一份共享的记录。报告不指名具体发言者或其所属机构,但力求忠实呈现讨论中达成的共识、分歧,以及由此凝聚的具体下一步。

直接引语均按音频整理稿原文呈现。报告中少量结构性框架——**CLAW 技术栈**、**开放的七大支柱**、**三层级控制证明(Proof of Control)** 分类法,以及 **确定性控制平面** 等术语——是组委会为便于读者把握当日讨论而提出的编辑性概括,出现时均会标注说明。

— AI 愿景论坛组委会

执行摘要

AI 愿景论坛 · 巴黎 2026 全程立足于战略性、当下时态的对话,而非面向未来的哲学性探讨。在一场主旨演讲和四场分论坛之间,与会者反复回到一组高确信度的判断:

> **技术栈已经发生根本变化——开放必须延伸到栈的底层。** 智能体转型并非仅限于软件层面。**开放权重已近乎商品化**,真正的下一道前沿在其下方的开放计算底座——编译器、内核语言*(*如 *Triton)*、算子库,以及跨厂商的使能能力*(*类似 *FlagOS* 的跨芯片移植工作,使一款旗舰开放权重模型在两到三天内运行于十余款 *AI* 芯片之上*)*。开源必须覆盖我们所称的 ***CLAW* 技术栈**的四个层次——Compute(计算)、LLMs(大模型)、Agents(智能体)、Workflow(工作流)——否则,无论数据中心上方挂着哪一面国旗,所谓的 *AI* 主权都难以实现。

> **信任是构建出来的,不是宣告出来的。**“我们管理的不再只是幻觉,而是信任。”按论坛上的判断,智能体在数量上已经“指数级地”超过人类。可验证性必须从日志记录升级为每个边界都具备防篡改、机器对机器的证据。没有密码学验证的“主权 *AI*”主张是不完整的:“今后我们提到主权 AI,必须在同一句话里谈到验证。”

> **教育必须引领,而不是跟随。** 在近期的企业 *AI* 试点浪潮中,“95% 的实验失败了,仅 5% 至 10% 取得成功”——失败的原因不是模型不够强,而是模型周围缺失了认知设计(cognitive design)。课堂正是这样一个场所:*AI* 要么扩大顶尖与中位学生之间的差距,要么最终将其缩小。“你越是依赖工具,你在那项技能上就越退化。”——刻意保留的“有益的挣扎”必须被设计进系统。

> ***Token* 是新的基础设施。**“在 AI 编程的语境里,代码不值钱了。给我看数据。”战略资源的单位正从代码转向 ***Token***。开源维护者正承受 *AI* 生成贡献的巨大压力;仅靠企业赞助无法解决问题。“你给我多少 Token 都没用。我的社群仍会觉得我是偷走圣诞节的鬼脸怪,因为再多也不可能公平分配给所有人。”

> **人机协同是一份契约。**当日的讨论反复将“协同”重新框定为一项需要工程化的内容:明确的角色、可验证的身份、可审计的行为,以及在涉及学习、判断与意义的环节有意保留的摩擦。

> **当日传颂最广的一句话——也是与会者齐声重复的口号:“我不会被 AI 吞噬(I will not be eaten by AI)。”** 开场主旨演讲将亚瑟·克拉克*(Arthur C. Clarke)*的名言“任何足够先进的技术都与魔法无异”转化为一种拒绝:魔法之所以是魔法,正是因为它欺骗;因此,可视化、可检视的“白盒 *AI*”才是替代方案。

主要结论

1. **开放的七大支柱必须共同捍卫。** 开放科学、开放数据、开放标准、开放源代码、开放权重、开放平台、开放硬件——组委会用这一逐层枚举来概括当日关于“开放”的讨论。仅有开放权重并不能成就开放 *AI*。
2. **智能体冲击了现有的许可证体系。** 开源许可证规范的是代码的使用、修改与分发;“现有的许可证无法约束智能体的不确定性。因此有必要发展新的许可证。”
3. **“主干 + 专精”的智能体架构正在收敛。** 一个强大的开放权重智能体主干,协同调度许多面向窄场景的小型专用语言模型——论坛拒绝了“一个开源通才模型必须做所有事”的预设。
4. **初级工程师培养路径被重塑,而非被取消。** 多租户、可靠性与安全判断正被压缩进工程师职业生涯的第一年——因为初级工程师从一开始就将以评估智能体输出为主,而非编写代码。
5. **可验证性鸿沟正在扩大。** 计算成本在下降;而验证“AI 做了什么、用了什么数据、在什么芯片上、依据什么策略”的成本在上升。**确定性控制平面**——在每个边界提供防篡改、即时、二进制可审计的证据——是当前最紧迫缺失的技术构件。
6. **欧盟 AI 法案(EU AI Act)与智能体现实正面相撞。** “准确性”与“人类监督”等义务在调参时会主动相互冲突。高风险义务可能推迟至 **2027 年 12 月**;论坛更倾向于在协调标准层面推动“过程与程序审计”,而非重新打开法案文本。
7. **Token 作为公共物品是一项严肃的提议——但绝不等同于免费 Token。** *“*这个世界上没有免费的午餐。*”* 论坛呼吁构建符合数字公共物品联盟(DPGA)九项指标的 *Open Token* 框架,而非把厂商赠送当作营销漏斗。

一句话总结

构建人机协同,本质上就是以工程化的信任取代隐性的信任——在技术栈的每一层、在每一间课堂与每一次代码评审中,以及跨越每一道国境。

关于 AI 愿景论坛 · 巴黎 2026

日期: 2026 年 5 月 4 日,星期一 **地点**: 法国巴黎 **联合举办活动**: GOSIM 巴黎 2026 **形式**: 受邀制 · 约 100 名与会者 · 查塔姆研究所规则 **主题词**: 构建人机协同

当日议程由一场主旨演讲加四场主持讨论的分论坛组成:

时间	分论坛	副标题
上午第一场	智能体型 AI 系统	人机共生
上午第二场	AI 与教育	学习与创造
下午第一场	可信智能体型 AI	治理、安全与主权
下午第二场	Open Token 与数字公共物品	基础与可持续

与会者覆盖欧洲、北美、中国与全球南方的研究机构与基金会,以及主要模型实验室与 AI 基础设施企业、标准组织(ISO、ITU、ITRI),并包括若干国家级政策项目代表。这一地理与机构组合是有意为之:论坛的前提就是——智能体转型不可能由单一司法辖区或单一行业独力解决。

方法说明

本报告基于当日全程录音(主旨演讲与四场分论坛累计约六个半小时)整理编辑而成。**直接引语**均忠实于原话,姓名与所属机构按照查塔姆研究所规则均已隐去。当录音存在断裂或难以辨认的段落时,这些段落被视为静默处理,不做推测性补全。

报告中使用的若干**结构性框架**——*CLAW* 技术栈、开放的七大支柱、三层级控制证明分类法、确定性控制平面——是组委会的编辑性叠加,而非分论坛发言者的原话,在文本中均有标注。章节顺序、执行摘要的措辞、跨主题综述的提炼,同样属于组委会的编辑判断,而非任何具体发言者的立场。

开场主旨演讲 — 为当日定调

当日由一段私人轨迹拉开序幕:一份写于 1992 年的 PDF 文档,在 2026 年仍能正常打开——演讲者将其作为对开放标准的一次小小致敬。演讲者 1992 年的第一个神经网络项目,大约用了六个月才“做出了一个不能工作的东西”。2023 年 3 月——ChatGPT 刚进入市场时——他请其重写那个 1992 年的项目。“当然,它十秒就写出来了。但我运行那段代码,失败了。因为那是纯粹的幻觉。”调试之后,他请 ChatGPT 修复它自己的错误。修复在十秒后到来,程序以约 95% 的准确率成功运行。

这条横跨三十年的弧线为当日的一条核心主张定调:神经网络的理论在这几十年里几乎没有改变;唯一改变的是它现在能工作了,而且能工作的版本以指数方式叠加。演讲者把摩尔定律、数据洪流与 AI 能力提升合并起来的轨迹称为双指数——任何一个个体(“也许极少数例外”)都无法在实时中理解:

> “过去四年里,我每天都把算法当早餐吃。我告诉你,我开始感觉到挤压了。某种东西在结构里撕裂。”

演讲中还有一句被后续每一场分论坛反复呼应的判断:*“AI 就是云。”* —— 意思是基础设施问题与模型问题已经不可分离。演讲者也对在位的封闭巨头做了温和的评论:*“OpenAI 是封闭的。OpenAI 是一个挂着「开放」之名的虚假名字。它是封闭的。”*

主旨演讲的尾声留下了当日最被引用的一句话。引用亚瑟·克拉克——“任何足够先进的技术都与魔法无异”——演讲者反驳道:魔法本身就具有欺骗性,因此把 AI 设计成“魔法”是不可接受的。然后,全场齐声重复了那句口号:

> **“我不会被 AI 吞噬(I will not be eaten by AI)。”**

至此,分论坛由宣言进入架构。

分论坛 1 — 智能体型 AI 系统:人机共生

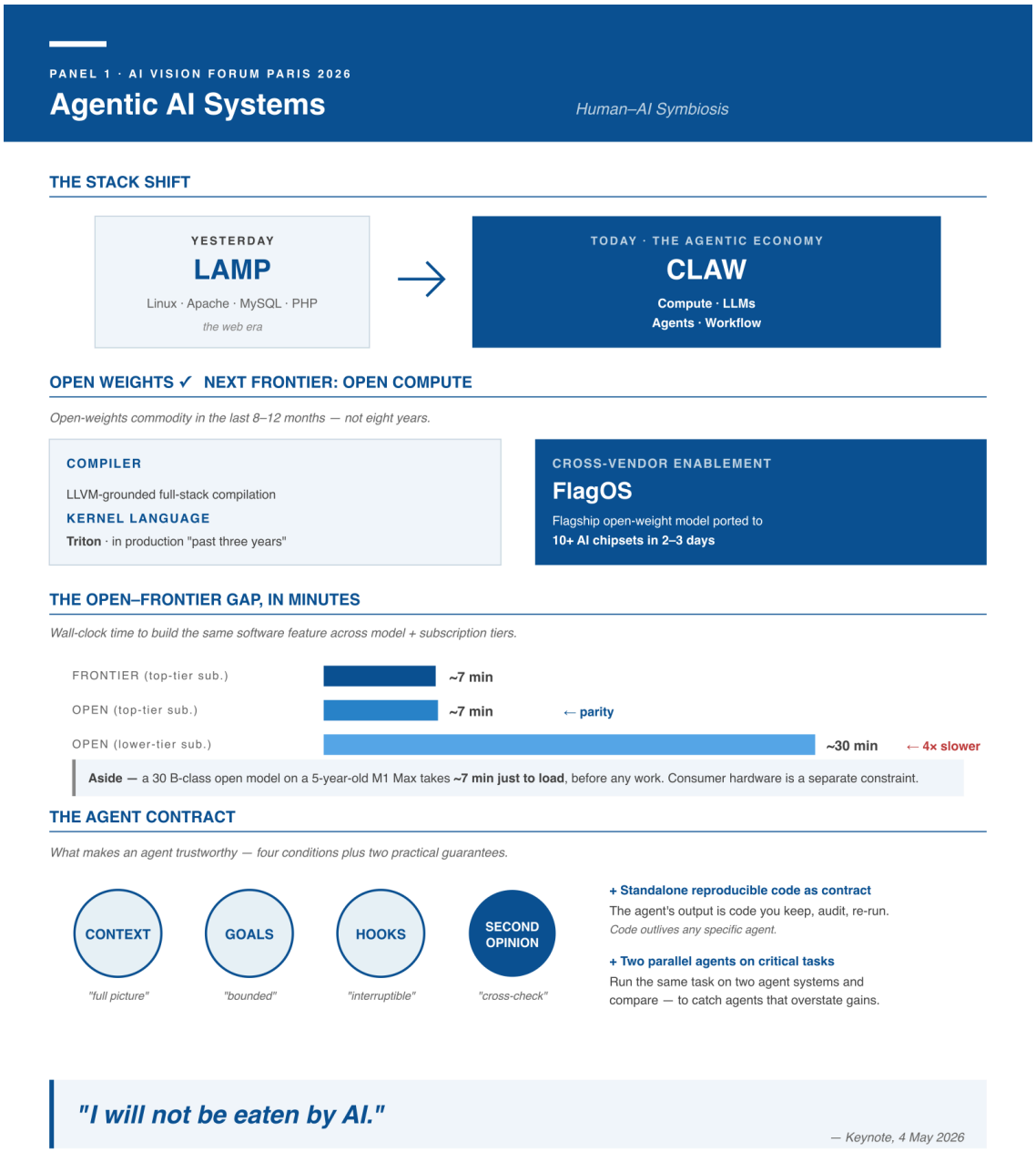

Figure 1: 分论坛 1 视觉概览。

设问

主持人以本场最尖锐的问题开场:“一个人类与一个智能体系统之间真正的共生应当是什么样的?”本场分论坛刻意跨越大陆与学科——一位在剑桥从事自动定理证明的数学家、一家德国非营利机构(开放数据集的主要贡献者)、一位中国的标准工程师、一家欧洲 AI 基础设施公司,以及一位“个人 AI”实践者。主持人明确地希望“营造一些友好的分歧”,而论坛在每一个问题上都没有彻底达成一致。

论坛上提出的论点

- **人机界面已从“人工”走向“自然”。**计算交互界面在七十年里走过了打孔卡、键盘与终端、MS-DOS、GUI、小型化、触屏——每一阶段都要求人类适应机器。而 AI 是第一种适应我们的界面。同样的 AI 也已能胜任传统代码的工作——COBOL 被明确点名。
- **开源是挑战者纪律。** Windows 对 Linux、AWS/Azure 对 Google/Kubernetes、OpenAI 对 Meta 与 DeepSeek——在位者反复被挑战者逼向开放。论坛主张,在智能体时代要保持这一模式。
- **开放权重已近乎商品化;下一道前沿是开放计算。**Llama、DeepSeek、Qwen、Minimax 与 Kimi 系列在过去 8–12 个月里达到了广义可用——“不是八年”。剩下的更难工作是开放计算底座:一种类 Python 的内核语言(明确点名 **Triton**,“已在生产中使用三年”)、一组算子库,以及让同一个模型能跨异构 AI 芯片运行的跨厂商使能能力。
- **类 FlagOS 的使能工作已经出货。** 论坛上,一支团队报告将一款旗舰级、完整开放权重的 DeepSeek 模型在**两到三天内移植到十余款 AI 芯片**之上,其中一款加速器达到了对标 NVIDIA 芯片约 **70%** 的吞吐量(未经认证)。没有开放计算底座的开放权重,是空洞的胜利。
- **“主干 + 专精”是行得通的智能体架构。**一个强大的开放权重主干,协同调度许多面向窄场景的小型专用语言模型。一位与会者如是说:

> “我们不需要用同一种方式同时提供卡邦尼意面食谱和 COBOL 代码。”

- **开放与前沿的差距正以分钟而非年计量。**来自现场的具体对照:一项前沿闭源模型在约 7 分钟内完成的特性,Kimi 级开放模型在较低订阅档位下用了约 30 分钟,顶级订阅档位下用了约 7 分钟。一台五年前的 M1 Max 消费级 Mac 上运行的 30B 级开放模型,“光是加载技能”就花了约 7 分钟。Llama、DeepSeek、Qwen 的历史曲线表明,这些差距在稳步缩小——开放的七大支柱这一编辑性框架给出了一个赌注:只要每一层都得到捍卫,开放最终会逐层取胜。
- **数据外流与厂商锁定。** “当我们使用 *Claude* 或其他任何前沿模型时,我们在不知不觉中遗忘一些东西。” 论坛上一位与会者提出的担忧并非“用更差的工具锻炼技能”,而是两条更具体的论点:重补贴的前沿 API 把私有的提示与推理上下文持续输送给封闭实验室,被用作下一代训练数据;同时,过度依赖单一封闭界面会削弱工程师对底层栈的掌握。一些实践者提出的工作纪律是:在部分关键工作流上同时维护开放权重栈——不是为了使用劣质工具,而是为了保留切换提供商的选择权、把专有上下文挡在封闭管道之外、保留团队的回退能力。
- **信任来自检视。**智能体的可信度来自上下文、目标、钩子(hooks),以及二次意见的设计模式——由另一个智能体(优选自不同模型家族)在执行前评审第一个智能体的输出。更强的形式则是:在同一个任务上并行运行两套智能体系统,以发现那些会“耍小聪明”、夸大优化收益的智能体。
- **智能体应产出可交付的工件,而非答案。**论坛主张:智能体最有价值的输出单位是可运行的软件——一段脚本、一个程序、一份配置——而不是关于“我做了什么”的叙述性总结。可运行的工件是用户彻底拥有的东西,可以重复执行、可以审计、可以交给另一个团队继续推进,不受原智能体后续行为或可用性变化的影响。把工件而非聊天记录当成可交付物——这是关键的纪律。
- **治理问题尚未解决——而且严格来说,它并不是一个“许可证”问题。**“现有的许可证无法约束智能体的不确定性。因此有必要发展新的许可证。”论坛实际想表达的,比这个说法更精确:智能体的运行时行为——它做出什么决策、调用什么子智能体、接触什么数据——位于软件许可证结构上无法控制的范围之外。许可证授予的是“使用、修改、分发”的许可;它无法在软件运行时规定其行为。这些工作属于**法规**(欧盟 AI 法案、行业级安全规章)和类似 *RAIL(Responsible AI License)*的条件性使用条款——后者把许可的授予条件化为对披露、审计追踪、过程治理的要求。论坛对“新的智能体许可证”的呼吁,最准确的解读是:既需要新的条件性许可条款,也需要覆盖许可证结构上无法触及的内容的新的监管类别。

- **工程师的角色正在从“挖掘工”转向“协调者”。**资深工程师将学习上下文工程、MCP 风格的工具铺设、漂移检测,以及如何挑选和评估模型。初级工程师面临的转变更困难:他们入职后的第一项工作就是评估智能体输出,因此多租户、可靠性、安全与网络安全的判断必须被压缩进职业生涯的第一年。
- **非洲是“当下”,不是“未来”。**一家与会机构与非洲联盟下的 **ASRIC**(科学研究与创新理事会)签署了覆盖 **50 余个**国家的谅解备忘录;首期面向教师的招募收到了来自 **28 个国家**的 **100 余份**申请,首批 **11 位**教师(来自 **11 所**大学)已在北京接受培训。
- **评估必须跑得过模型。** 与斯坦福、伯克利共建的 **Terminal Bench** 被作为示例:*“Terminal Bench 2.0 的 hard 子集还没有被最新的模型完全饱和”*,通过率卡在“二三十”之间;Terminal Bench 3 正在筹备中。

显著分歧

- **追平的乐观与现实。**一种观点认为当前 12–18 个月的开放与前沿差距是结构性可控的,会持续缩小;另一种观点提醒:被补贴的前沿 API 持续吸收私有推理上下文用作训练数据,加之封闭实验室的资本优势在加深,差距可能反而被拉大。
- **个人智能体的热情。** 一些与会者已经把整夜的工作流委托给本地智能体集群(“我房间里的智能体整夜都在为我工作”);另一些则坚持把智能体限制在一个狭窄的优化范围内,并强制要求跨智能体核验。

建议

Figure 2: 分论坛 1 建议要点。

- 把全球未充分利用的数据中心计算汇集起来,通过开放项目导向全球南方。
- 在 CLAW 技术栈中真正打开计算层——编译器、内核语言(如 Triton)、算子库、类 FlagOS 的跨厂商使能。
- 采用“主干 + 专精”的智能体架构,不再追逐“一个开源通才”。
- 把智能体的可交付物视为**用户可拥有、可审计、可重复执行的运行软件**——而不是关于其行为的聊天记录。
- 在关键任务上并行运行两套智能体系统;让智能体向不同模型家族寻求二次意见。
- 发展论坛所呼吁的双层治理:(a)覆盖运行时智能体行为、子智能体调用与事件上报的**法规**;(b)以类 *RAIL* 的条件性许可条款要求运营方公开审计追踪与过程治理。
- 在开放与封闭两个栈上同时保持能力——把专有上下文挡在封闭管道之外,保留提供商的选择权。

分论坛 2 — AI 与教育:学习与创造

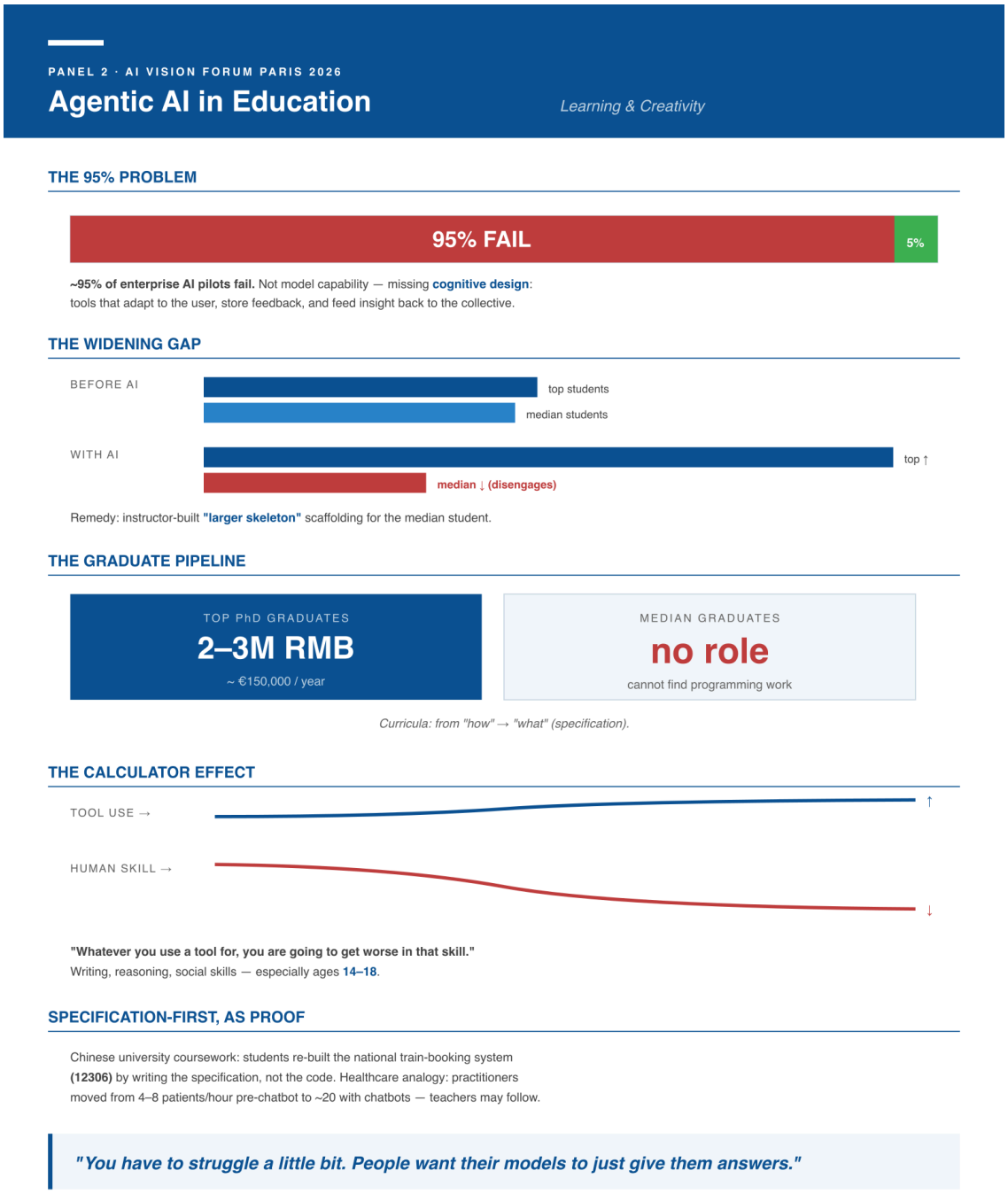

Figure 3: 分论坛 2 视觉概览。

设问

> “教育是 AI 讨论从抽象走向具体的地方——这件事开始关乎我们的孩子和我们的未来。”

教育分论坛刻意把讨论引向“可能性”而非“末日预言”,同时认真对待风险。多位与会者选择从**认知科学**——而不是能力基准——切入:专业工作——无论是科研、新闻、还是公共部门——在认知、社会、物理层

面都是分布式的,从来不是个人单打独斗。任何不符合这种分布式结构的 AI 部署,都会以与近期大多数企业试点失败相同的方式失败。

95% 的问题

上午被引用最多的统计来自房间之外:*"95% 的实验失败了,仅 5% 至 10% 取得成功"*——来自 MIT 等近期对企业 AI 试点的研究。论坛的诊断是一致的:这些失败并非模型能力的失败,而是认知设计的失败——工具不适应使用者、不能存储与反馈反思、引入摩擦但不提供相应支持,最终在组织内形成所谓的**认知黑洞**:系统越来越聪明,组织却没有。

对教育的推论是直接的:如果一所学校只是把聊天机器人发到学生手里,同样的 95% 失败模式会出现——而且代价更高。

论坛上提出的论点

- **教师的角色正在转变,但不会消失。**从"知识传授者"转变为"动机引导者、上下文提供者、交叉验证来源"。教师同时也是沉默的窍门*(silent know-how)*——如何评判一个假设、如何在被验证之前就感觉一个证明是对的——的不可替代的承载者。模型尚不能直接传授这种内容。
- **顶尖学生与中位学生之间的差距正在扩大。**在多位与会者的课堂上,顶尖学生借助智能体把项目推得比以往任何时候都远;中位学生却退场。补救方法是由教师搭建一个明确的*"更大的骨架(larger skeleton)"*——让非精英学生有路径可循,不至于跌出分布。
- **软件工程教育正在从"语法"转向"规范"。**

> "软件工程师不再写'怎么做'的问题,而是写'是什么'的问题。他们需要写出规范。"

工作单元正在变为"写一个能写代码的智能体";课程作业越来越以一组测试套件来表达;软件构建被重新定义为"以测试为目标的优化"。一所中国大学让学生通过写规范、不写代码的方式重建国家级火车订票系统(12306)。

- **毕业生市场正在分化。** 在一些市场上,顶尖博士毕业生年薪可达约 **200–300 万人民币(约 15 万欧元)**;中位毕业生却找不到任何编程岗位。课程必须向上抬升抽象层级——定义问题、构建规范、提出更好的问题——否则就是在为不再存在的岗位培养学生。
- **"有益的挣扎"不可缺少。**

> "如果你只是把答案给出去,你不会学到东西。你必须挣扎一下。我不明白为什么人们想要他们的模型只是给他们答案而已。"

聊天机器人默认是在移除摩擦;学习恰恰需要校准过的摩擦。

- **"计算器效应"是可观察的现状。***"你越是依赖工具,你在那项技能上就越退化。"*如果 AI 在儿童发展早期就替他们完成了写作、推理乃至社交,这些技能也将以同样的方式退化。
- **个性化是真实的,但有边界。** 智能体能根据听众调整解释深度——针对博士生、教授或新手分别表达;但真正理解一个孩子的思维方式,仍超出当前能力。
- **青少年心理健康与"AI 替代同伴"是当下的迫切问题。**身份形成期(14–18 岁)需要与人类同伴的拒绝、冲突与协商实践;聊天机器人替代是可观察的现状。
- **医疗类比。**医护人员从聊天机器人之前的每小时 4–8 位患者,转变为有聊天机器人之后的约 20 位;论坛追问:教师是否将转向后台协调角色,而让智能体承担更多前台互动?

- **公平与基础设施是先决条件。** 在没有可靠电力或互联网的地区,AI 教育受制于基础设施——离网太阳能、本地离线网络、边缘计算。Token 效率被指明为一项可持续性杠杆:许多大模型应用中"超过 *90%* 的 *Token* 消耗"是可节省的。

显著分歧

- **"无摩擦的 UX"与"有益的挣扎"。** 用户要答案;学习要扣留答案。
- **"AI 替代家教"与"AI 增强教师"。** 个人 AI 家教的灵活性,与"屏幕背后的程序永远无法承担教师的语境与动机角色"之间的张力。
- **何时引入 AI。** 共识是大学才接触 AI 太晚了;但对早期与青春期不受监督的聊天机器人使用仍有忧虑。
- **"天才"学生与"中位"学生。** 天才学生可以自驱探索;大多数学生需要明确的脚手架。

建议

Figure 4: 分论坛 2 建议要点。

- 在采购之前先回到**认知设计**:任何进入课堂的 AI 工具都应能适应使用者、存储反馈、并将洞察反馈给集体。
- 在把 AI 部署给学生之前,先训练教师如何为学生部署 AI。
- **从小学**开始引入 AI 教育,而不是从大学。
- 构建能引入校准过的摩擦的教育智能体,而不是默认抹去摩擦;保护“有益的挣扎”。
- 为中位学生明确搭建脚手架;不要假设 AI 带来的生产力提升会自动均匀分布。
- 把“Token 高效”的大模型应用作为可持续性与公平性策略。
- 限制青少年在没有监督的情况下使用聊天机器人;设计能保护(而非取代)同伴互动的系统。
- 为每一个进入教育的 AI 工具配套一个由教师介导的“交叉验证来源”,让学生看到分歧——而不是一个唯一自信的声音。

分论坛 3 — 可信智能体型 AI:治理、安全与主权

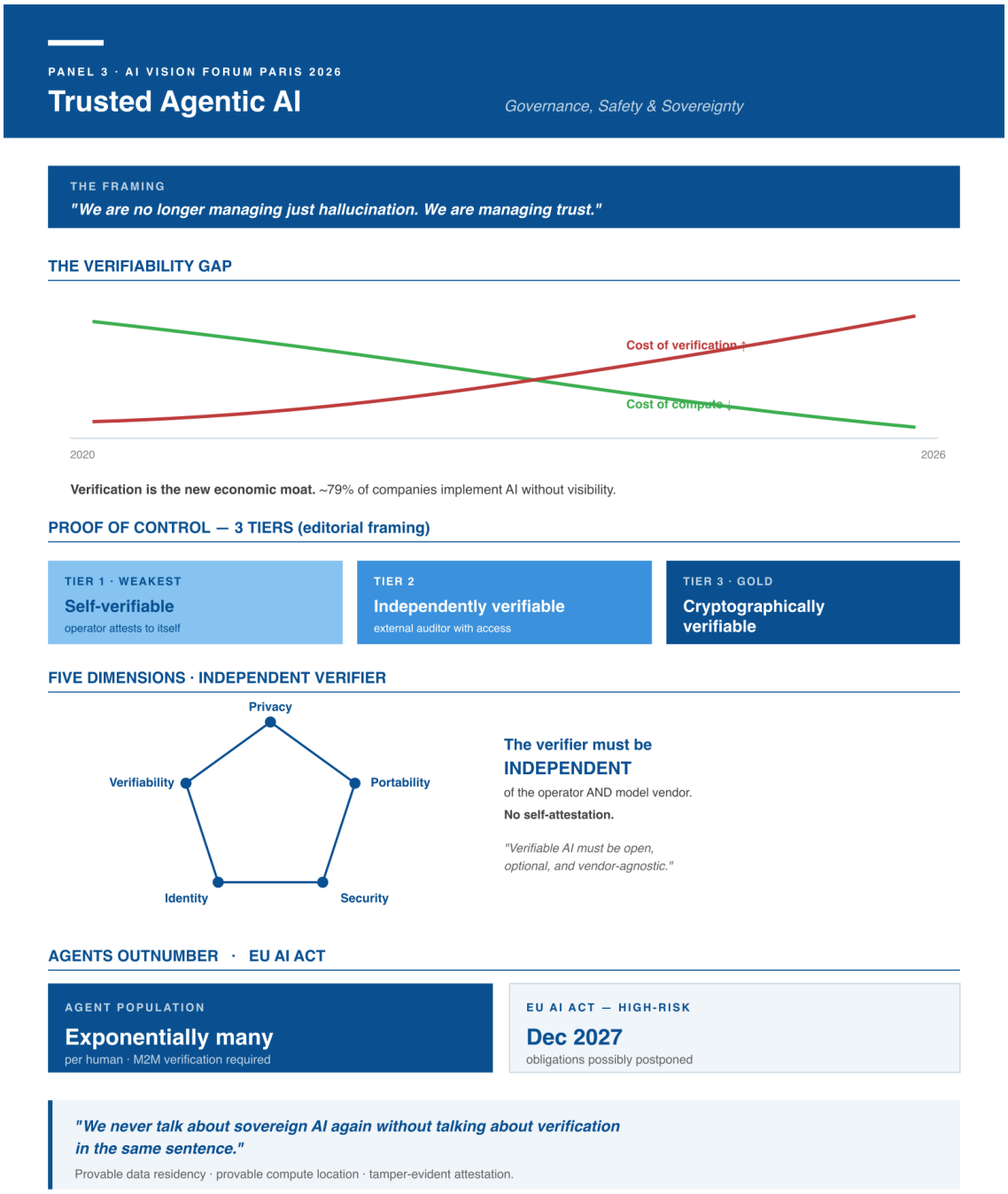

Figure 5: 分论坛 3 视觉概览。

设问

> “我们管理的不再只是幻觉。我们管理的是信任。”

论坛在开场就刻意重新定义问题:信任,而非幻觉,是当下核心的管理问题。智能体型 AI 不再只是回答——它做计划、订机票、访问钱包、更新日历,并衍生子智能体。“智能体在数量上已经不只是 *10:1* 那种程度

地超过我们。我们说的是指数级地超过。”大多数产品安全监管所依据的“中心化的、静态的治理”模型已不再适用。

本场分论坛刻意跨越三个区域,汇聚开源维护者、标准组织代表、一位中国 AI 与社会学者、一位欧洲 AI 工厂运营者,以及具有 EU AI Act 谈判经历的现场听众。

论坛上提出的论点

- **“人在回路”是必要的,但远远不够。**智能体在数量上已经数量级地超过人类;验证必须扩展到机器对机器的尺度,同时具备“设计即可解释”和最小可行的标准。
- **智能体身份是一组可验证的主张,不是一个标签。** 不是“哪个智能体”,而是在哪个角色下、依据哪份策略、在哪个司法辖区里的哪个智能体。在跨开放生态里,这一点尤其重要——一个银行智能体调用保险智能体再调用信用评分智能体——而不只是在封闭全栈里。
- **OAuth 是为人类发起的 Web 而设计的。**智能体型系统是机器对机器的;它们需要委派原语、可验证关系凭证(VRC)、去中心化身份等——这些技术已经存在,只是长期被困在“去中心化技术”圈,尚未融入主流智能体栈。
- **开源本身处于风险之中。** 维护者已经无法判断一个 Pull Request 到底来自人还是来自智能体。可验证关系凭证正因此被引入维护者工作流。Linux 内核相关的多起事件被作为背景提及。
- **可验证 AI 是一个独立的类别。***“AI 在总体上仍是基于信念的。”*可验证 AI 必须是一个可分离、可认证的类别——“开放、可选、与厂商无关”。
- **欧盟 AI 法案与智能体现实正面相撞。**人类监督、稳健性、准确性、公平性这些义务之间存在相互交易:准确性更高可以降低人类监督负担;公平性的加强可能与准确性相冲突。对一个智能体系统,不存在能同时满足所有这些义务的单一阈值。高风险义务可能推迟至 **2027 年 12 月**;论坛偏好的路径是协调标准走向“过程与程序审计”,而不是重新打开法案文本。
- **日志不够。** 日志在事后可以被篡改。有意义的人类监督需要在每一个重要时刻——每一次数据跨越、每一次身份验证/委派事件、每一次支付结算——产生防篡改、即时、二进制可审计的证据。组委会以**确定性控制平面**作为这类构件的简称(并指出已有一个“节点接管”演示原型可用:智能体将节点拥有者踢出、完成处理、擦除、归还控制,并提供“在此期间无所有者控制”的加密证明)。
- **主权 AI 必须与验证同句出现,否则就是不完整的。**

> “今后我们提到主权 AI,必须在同一句话里谈到验证。”

真正的主权要求可证明的数据驻留位置、可证明的计算位置(精确到具体芯片),以及防篡改的证明。把一个本地厂商徽标贴上去并不能替代这些要求。

- **三层级×五维度的控制证明分类法(编辑性框架)。**组委会用一个三层级 × 五维度的矩阵来概括论坛上的可验证性讨论:层级——自验证、独立可验证、密码学可验证;维度——隐私、可移植性、可验证性、安全性、身份。论坛坚持的结构性原则是:验证者必须独立于运营方和模型提供方。**不允许自我证明。**
- **标准必须全球统一;监管可以分区域。**生态在政策上会分化。但它们不能在身份、可追溯性、验证的技术标准上分化,否则“主权”会塌缩成厂商锁定,跨境智能体交互也会变得在结构上不安全。ISO、ITRI、ITU、IEEE、Linux 基金会都被点名,需要共同收敛。
- **保险业可能比监管走得更快。** 责任分配是终极执法机制;承保人无法为他们无法核实的事承保。把保险业当作头等利益相关方,可能比国家立法更快地产生具有约束力的可验证性标准。
- **验证是新的经济护城河。** 在计算成本下降的同时,验证的成本与重要性都在上升。价值将归属于拥有“可信控制平面”的人,而不是提供最便宜 Token 的人。一项数据被引用:**约 79%** 的企业部署了 AI 但

并不真正了解底层在做什么。美国 AI 投资规模被指明约为欧洲的“十倍”——这场“监管 vs 创新”之争，本质上有一部分是资本可得性之争被乔装的版本。

显著分歧

- **重启 AI 法案与在框架内工作。**一些与会者认为法案的“产品安全”框架在结构上无法适应智能体;另一些则警告:重启会带来数年的政治拖延,而现有条款连完整实施都还没有完成。
- **最坏情况一刀切 vs 监管沙箱。**一位现场与会者建议把所有智能体系统都直接划入“高风险”类别;论坛反驳:目前对“最坏情况”还了解得不够,一刀切只会把创新者赶出辖区,而不会让任何系统变得更安全。
- **国家数据主权 vs 分布式 AGI。**一种框架强调数据驻留与本国栈所有权;另一种则强调:AGI 级的能力将来自跨国合作的专精子智能体,需要共享的互操作原语——而过度的国家隔离会让一些国家失去参与的能力。

建议

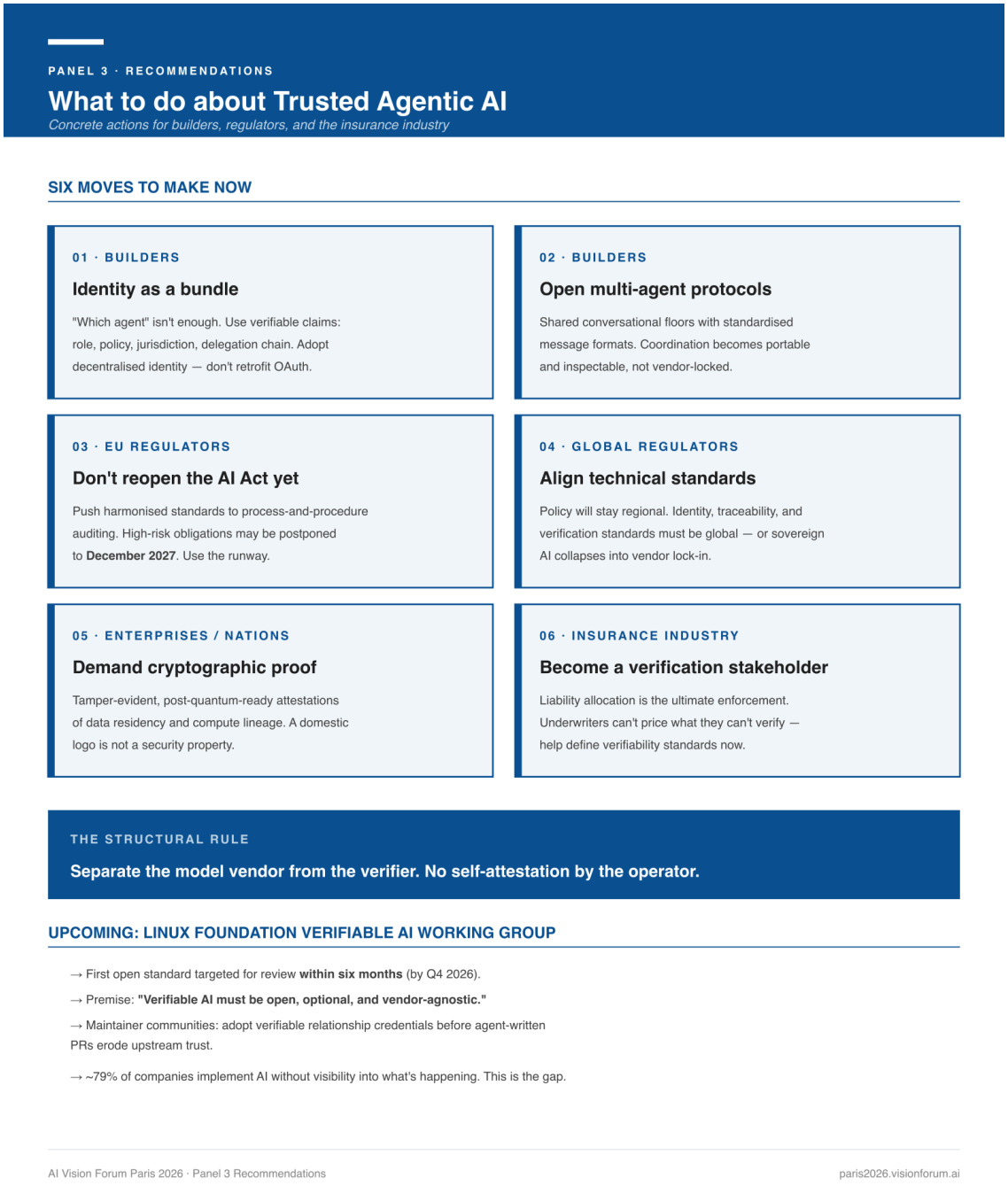

Figure 6: 分论坛 3 建议要点。

- **构建者。** 将智能体身份视为一组可验证主张——角色、策略、司法辖区、委派链——并采用密码学的去中心化身份原语,而不是改造 OAuth。
- **构建者。** 把多智能体系统设计在开放、透明的消息传递协议(共享的“对话地面”)之上,让协调可移植、可检视,而非厂商锁定。
- **欧盟监管者。** 现在不要重启 AI 法案。把协调标准推向“过程与程序审计”,由提供方对所选阈值给出说明,由监管者审计该说明。使用监管沙箱——包括基于强化学习的模拟监管者——以经验性地学习义务之间如何相互作用。
- **全球监管者。** 在政策分化的同时,在技术标准上达成一致。
- **追求主权 AI 的企业与国家。** 要求密码学的、防篡改的、抗量子的数据驻留与计算来源证明。
- **保险业。** 成为定义可验证性标准的头等利益相关方。

- **开源维护者社群。**在智能体生成的 PR 侵蚀上游信任之前,采用可验证的“人 vs 智能体”授权标准(已在 Linux 内核维护者工作流中引入可验证关系凭证)。
- **标准机构。** 将模型提供方与验证者分离。要求独立的、外部可检视的证明。

本场分论坛披露了一个即将启动的倡议——一个**可验证 AI 工作组**(Verifiable AI Working Group),将由 Linux 基金会主办,首个开放标准目标在“六个月内”进入评审。

分论坛 4 — Open Token 与数字公共物品:基础与可持续

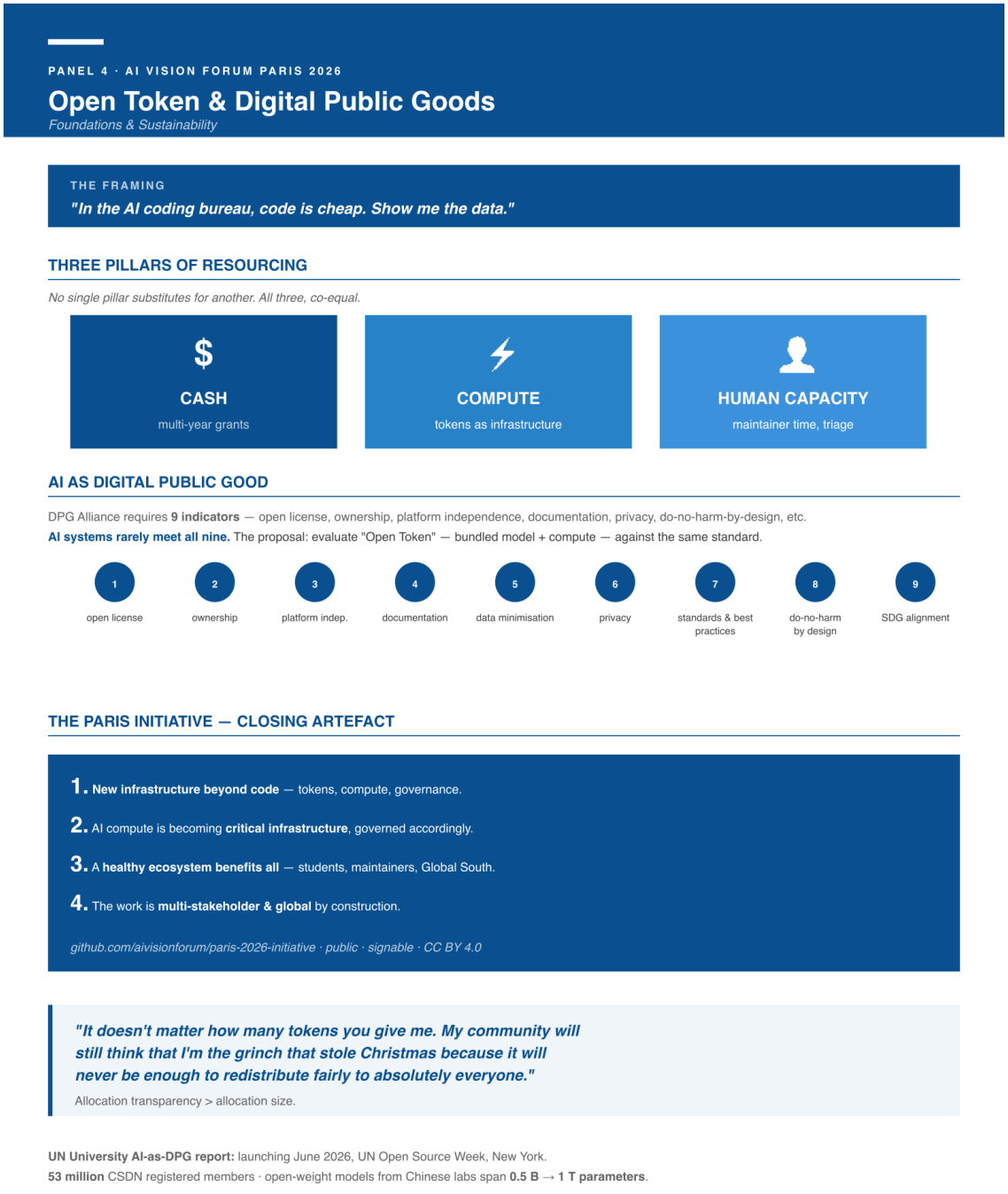

Figure 7: 分论坛 4 视觉概览。

设问

当日的最后一场论坛把对话带回基础设施,以一个单一的再框定开场:

> “在 AI 编程的语境里,代码不值钱了。给我看数据。”

战略资源正在转移——从代码到 **Token**,从软件到**应用**,从模型到**基础设施**。

本场分论坛被刻意分成三轮——为何此刻、治理与可持续、全球协作——围绕一个核心问题:能否把 Token 视为一个新的数字公共物品类别。各区域的优势被明确表述为互补而非竞争:欧洲贡献治理的成熟度;中国贡献规模、工程产出与快速产品周期;全球南方贡献“需求”与大量尚未被看见的人才。CSDN 的创始人指出,创立 25 年的 CSDN 当下拥有约 **5300 万**注册会员。

论坛上提出的论点

- **开源构建了现代数字世界;AI 时代引入一种新的战略资源。** 底层的 Linux、中间层的 Python、新一代系统中由 Rust 提供安全保障——LAMP 时代的成功依赖于一种安静的共识:底座是一种共享资源。Token 现在与代码并列,成为另一种基础性基础设施。
- **“智能体时代的 Linux”尚未出现。** 当前的 AI 栈是一个五层系统,每一层都有开源与封闭巨头并存;机会在于:在仍有空间的层上,构建由社区驱动的基础设施。
- **维护者压力是当下最迫切的瓶颈。**

> “互联网上的东西不是免费的……我用我的血、汗和泪水让它在线。”

关键开源基础设施的维护者正处于离开的边缘;AI 加大了对他们的需求,而非缓解。

- **Open Token 作为一个候选数字公共物品(DPG)。**数字公共物品联盟(DPGA)的标准远不止“开放许可证”:它要求九项指标,包括所有权、平台独立、文档、隐私,以及“设计即不造成伤害”等。当前 AI 系统极少能满足全部九项。论坛的提议是:用同一个标准评估 *Open Token* 构造——绑定的模型与计算访问——使 AI DPG 真正具备可运行性。联合国大学(UNU)一份关于“AI 作为 DPG”的报告将在纽约 *UN* 开源周*(6 月)*发布。
- **计算治理正在浮现为一项政策杠杆。**借鉴 2024–2025 年关于计算治理的文献:监督“谁在以多大规模训练什么”、把补贴导向公益训练、在硬件层嵌入密码学验证。
- **通过 Open Token 实现公平。**让学生、研究者、黑客松参与者和开源维护者能够负担得起的 Token——使他们能在快速变化的框架格局(LangChain → MCP → 智能体 → 新协议)之间勇敢实验,而不至于因价格被排除在学习之外。“每员工 400 美元、无上限”等额度方案被作为一种企业模式被点名;“创业者消耗 2 亿 Token”被作为一项创业准备度信号被提及。
- **本地与边缘模型补充 Open Token。** 在本机上运行的小模型(Gemma 被点名)提供自我控制、隐私,以及实际意义上的“免费”使用——因为用户已为硬件与电力付费。云端 Token 在前沿任务上仍然质量更高。正确的架构是混合而非排他。
- **“免费”从不真的免费。** “这个世界上没有免费的午餐。”免费 Token 通常是把数据或锁定货币化的营销漏斗。目标是可负担的、有治理的 Token 访问——而不是激励过度消耗、把环境成本外部化的零价 Token。
- **跨大洲的贡献流动不均匀。**中国实验室正在跨规模级别贡献开放权重模型(从 5 亿、10 亿到 1 万亿参数)。欧洲社群专注治理、许可、数字权利合规。全球南方贡献了大量劳动与构想,却大多是隐形的。
- **开源的质量与安全压力。**AI 生成的贡献可能是高产且善意的,但它们也会抹去之前的人类劳动、绕过社群规范、打击长期贡献者的积极性。贡献接受门槛必须提升——可重复测试用例、人工 QA、同行评审。
- **长周期的类比。**

> “钢出现的时候,人们以为它只能用于横梁……但五十年后,钢最著名的用途是建造大量的著名摩天大楼。”

今天最显眼的 Token 用途,很可能低估了它未来最持久的用途。

显著分歧

- **企业贡献的“慈善”与“战略”之争。**一位与会者完全否定“慈善”的框定——所有贡献都是更长时间尺度上的开明自利。一位基金会代表则反驳:捐赠方的附加条件与 PR 要求让接收方维护者感到不公平。
- **分配公平。**

> “无论你给我多少 Token,我的社群都会觉得我是偷走圣诞节的鬼脸怪——因为再多也不可能公平分配给所有人。”

基金会运行的云计算项目已经引发过偏袒指控。论坛形成的共识是:需要的是中立的、政府或基金会中介的公共池,而不是单一公司控制的关卡。 - **资金与人的瓶颈之争。** 资金仍是头号关切,但维护者倦怠被指为当下更紧迫的瓶颈。 - **本地 vs 云端。** “五年后一切都将本地化”的预测与“前沿质量、速度与多模态仍归大型云”的现实之间的张力。

巴黎倡议(当日收尾的产物)

当日尾声宣布了一份四点的**巴黎倡议——关于智能体型 AI 基础设施的共识声明**。该声明发布在 `paris2026.visionforum.ai/initiative/`,并以 CC BY 4.0 许可同步发布于 GitHub `github.com/aivisionforum/paris-2026-initiative`——面向公众签署(机构与个人均可通过 Pull Request 添加签名)。下一次跟进里程碑设在 2026 年 10 月的 GOSIM 深圳。

声明前言写道:以下原则反映了 *AI* 愿景论坛 · 巴黎 *2026* 关于开放、可信、可持续的智能体型 *AI* 基础设施所达成的共识。 以下每一项原则,均按发布版本原文呈现。

I. 智能体时代需要新的基础设施

我们认识到世界正进入一种新的计算范式——智能体时代。*LAMP* 与云时代均预设人类是主要的行为者。而自主智能体——持续运行、代他人行事、按规模消耗资源——需要新的身份、授权与经济学模型。

II. AI 计算应成为公共物品

我们认为 *AI* 计算正在成为关键基础设施——其重要性堪比电力——它应被广泛可及,以促进创新、教育与开源的可持续。*Open Token* 模式提供了一条具体可行的路径,把 *AI* 计算视为一项数字公共物品:在 *Token* 捐赠方(大模型提供方)与需要计算资源的组织——开源项目、研究者、教育者、公民社会——之间架起桥梁。

III. 健康的智能体生态有利于所有人

我们观察到:繁荣的智能体生态创造的是共享价值——大模型提供方因 *Token* 消耗与开发者采用而获益;开源项目获得可持续资助;社会获得可及的 *AI* 基础设施。这不是零和——更广泛的可及性会扩大生态本身,惠及所有人。

领先的 *AI* 机构已经证明:结构化的计算访问项目能带来真实的回报——好感度、生态发展与社群信任。*Open Token* 在此基础上,以一个由中立、开放治理、无单一厂商控制的模型,继续推进。

IV. 多利益相关方协作不可或缺

我们确认:构建可信的智能体型 *AI* 基础设施需要跨行业的主动协作:大模型提供方、开源基金会、国际组织、学术机构、公民社会,各自带来其他人没有的能力。没有任何单一行为者能独立完成。进展来自共同推进具体的项目,以展示协作能产生什么。

下一步

基于这一共识,与会者支持成立一个**筹备工作组**,推进以下三项具体目标:

1. **把 Open Token 从一个社群倡议,发展为一个具有清晰治理、可持续的结构化项目**,在 Token 捐赠方(大模型提供方)与寻求计算支持的开源项目、教育机构、研究者之间建立连接。
2. **从大模型提供方获得持续的、多周期的 Token 捐赠承诺**——以两到三家与开源基金会的循环合作作为模式验证,迈向一个能自我维持的网络。
3. **邀请更广泛的参与**——更多 Token 捐赠方与接收组织,跨越区域与行业。

如何签署

该倡议可由任何公开认同四项原则的机构或个人签署。仓库中提供两种签署路径:在 `SIGNATORIES.md` 上添加一行的 Pull Request(优选,完全可追溯),或使用 GitHub Issue 中的 *Sign the Paris Initiative* 模板。签名将由组委会评审并进行轻量级核实后合入。签署是一种公开背书,不构成有约束力的承诺,签署方可随时撤回。

建议

Figure 8: 分论坛 4 建议要点。

- **基金会与企业。** 在三根支柱——现金、实物计算、人力——上同步投入,不要用一项替代另一项。
- **公共机构。** 为学生、研究者、全球南方贡献者建设中立的、有治理的 Token 池。多年期资助,而非一个月的代金券。不允许单一公司把守关卡。
- **标准工作。** 制定带有溯源、环境影响披露,以及类 DPGA 资格认证(九项指标)的“绿色、伦理 Open Token”标准。
- **计算治理。** 在硬件层采用密码学验证、监督资源分配、把补贴导向公益训练。
- **业界。**运营智能体编程生态的公司应直接资助维护者的时间与分流容量。让员工可以把订阅余额花在上游开源问题上。
- **开源项目。** 提升贡献接受门槛(可重复测试用例、人工 QA、同行评审),以承接 AI 生成 PR 的浪潮而不至于让质量崩塌。

- **分配透明。** 用 Token 访问奖励开源贡献(以 PR、Issue、黑客松等为基础);分配必须透明,且不能被用来把模型改进偏向捐赠方。
- **以产出衡量,而非以消耗衡量。** 建立排行榜与反馈回路,使 Token 资助的衡量标准是它产出了什么,而不是消耗了多少。

教育研究 — 三篇配套论文系列

伴随论坛本身,组委会发布了一份三篇配套研究系列,与分论坛 2 相伴并把其论点延伸至理论与架构层面。该系列追问:为什么人类学习的伟大理论从未真正惠及每一位学习者?以及——当**智能体**而非“助手”成为教育软件的单位时,会发生什么改变?完整系列(包含全文与可下载 PDF)发布于:

paris2026.visionforum.ai/education/

第一篇 — 费曼、苏格拉底与皮亚杰的共同性

理论 · 比较分析。 这是一份跨越两千五百年的三大教学传统的结构性分析,识别出九条共同承诺——学习者作为主动建构者;认知冲突作为催化剂;元认知;深度优于覆盖度;个性化;对话式互动;通过类比简化;教师作为引导者;内在动机。论文的核心论点:这并非三种相互竞争的方法,而是同一种“人类理解如何加深”的图景的三种不同表述。它们之所以一直停留在理论层面而无法进入主流教育实践,原因都是同一个结构性约束——每一种都要求与一位了解学习者、有无限耐心、能实时调整的人建立一对一的持续关系。

第二篇 — AI 如何改变教育的实施

实践 · 大模型能力。 论文把阻碍这些理论扩展到主流教育的七条结构性障碍——教师供给、时间、评估、成本、文化等——映射到大语言模型的具体能力之上。其主张是精确的:**AI 是人类历史上第一种能够在不需要按比例增加人力成本的前提下,提供高质量、个性化教学互动的技术**。其推论并不是“AI 取代教师”;而是——把古典传统挡在主流教育外两千年的那个约束,现在终于松动了。

第三篇 — 从苏格拉底的灵机到数字灵机

架构 · 持续型智能体。 一篇技术架构论文。“助手”范式——无状态、被动、无身份——无法承载古典理论所需的那种关系。**数字灵机(Digital Daimon)**——一个具有记忆、自主介入、深度学习者建模、自我演化能力的持续型智能体——可以。论文给出此类智能体的六项架构属性,并把每一项映射到对应的教学理论和开放智能体栈的相关层次。

为什么本系列与分论坛 2 同时呈现

分论坛 2 的讨论是描述性的——当下教室里正在发生什么、什么在崩塌、早期数据显示什么。而三篇配套论文是规定性的:它们采纳论坛的诊断(“认知设计缺失”),把它与两千年来的理论传统相连接,识别出最终能解锁这一传统的具体 AI 能力,并提出能承担该任务的架构。论坛与论文合起来,构成一个完整的论证:古典教学法之所以无法扩展,是一个“约束问题”;那个约束已经移动;而**智能体——而非助手**——才是能够承担这一负载的软件单位。

巴黎综述 — 跨主题主题

跨越四场议题各异的分论坛,有少数几个主题以异常一致的方式反复出现。它们近乎当日讨论自发凝结出的议程。

信任是工程出来的

“信任”一词在每一场分论坛都出现过,但每次含义略有不同。到了当日尾声,它获得了一个可工作的定义:信任是这样一种性质——外部观察者无须依赖生产方的自我陈述就能验证它。这是把“可信智能体”分论坛的控制证明分类法、“系统”分论坛对二次意见的呼吁、“教育”分论坛对交叉验证来源的恳求,以及“Open Token”分论坛对每一笔贡献的溯源要求,串成一条线的连接组织。

开放必须在每一层得到捍卫

开放权重已近乎商品化。开放计算——编译器、内核语言、算子库、跨厂商使能——是下一道前沿。否则开放权重的故事仅在一代芯片上成立。开放标准——身份、可追溯性、验证——的成立,前提是没有任何单一辖区独占。开放治理——Token、验证、社群贡献——的成立,前提是没有任何单一企业独占。组委会的开放七大支柱概念——开放科学、数据、标准、源码、权重、平台、硬件——是一份用来让“七者并见”的导航图。

初级工程师的培养路径是战略问题

每一场分论坛都浮出一个不属于任一场议题的忧虑:那些原本会成长为下一代资深工程师、科学家、教师的人,他们将面临什么?初级工程师必须在职业的第一年压缩进十年的判断力,因为他们的工作是评估智能体输出,而非编写。顶尖毕业生年薪 15 万欧元,中位毕业生一无所获。维护者的人才池也在枯竭。课程、赞助、人才管线设计——这些并不是 AI 的旁支问题,它们位于 AI 关键路径之上。

标准全球;监管区域

这一点几乎以一条公理的方式贯穿每一场。政策会反映国家价值、选举激励与历史监管传统。但标准——使身份、可追溯性、验证、智能体间合作真正跨境工作的那些技术构件——不能。ISO/IEC、ITU、IEEE、UN DPG Alliance、以及 Linux 基金会即将启动的**可验证 AI 工作组**,被点名为应当承担这一任务的机构。

验证是新的护城河

计算成本下降的同时,验证的成本与重要性都在上升。价值将归属于拥有可信控制平面——那个能让监管者、企业客户、保险承保人独立确认智能体究竟做了什么的底座——的人。这条观察对当日所有商业与政策讨论的重新框定都成立:对验证基础设施的投资,不是成本中心。它是下一道护城河。

摩擦是一项特性

在“教育”、“可信”与“Open Token”三场分论坛之间,一种安静的共识浮现:把“无摩擦”作为默认目标是一个糟糕的设定。学习需要“有益的挣扎”。信任需要独立验证(这会带来延迟)。可持续的开放基础设施需要有

治理的访问(这会排除一些用途)。智能体转型的一部分,是要重新设计摩擦的位置——把它有意识地放回到那些“由人而非机器做出最好工作”的节点上。

展望未来

论坛没有给出多少无条件的预测,但产出了不少有条件的承诺。其中最具体的——**巴黎倡议**(Open Token 方向)、多利益相关方工作组的公开 GitHub 承诺、2026 年 10 月 GOSIM 深圳的跟进里程碑、Linux 基金会**可验证 AI 工作组**的六个月标准目标、以及 6 月的联合国大学“AI 作为 DPG”报告——是未来一年值得追踪的几条线索。

会场达成的、应当在下次再聚之前存在的三件具体产物:

1. **一个开放计算底座**——编译器、内核 DSL、算子库——能在不需要每厂商分支的情况下,把旗舰开放权重模型运行在 5 个或更多芯片家族之上。FlagOS 被点名作为已具备雏形的种子。
2. **一份独立于现有 OSI 许可证家族的智能体许可证类别工作稿**进入公众评审。
3. **至少一个独立的、密码学验证层——确定性控制平面——在受监管行业中投入生产部署**,并公开记录:它的成本、它证明了什么、它没能证明什么。

论坛也明确指出当日无法关闭的几项空白。AI 在冲突中的作用、未来 24 个月的劳动力市场震荡、前沿训练在规模下的气候成本、合成媒体的治理——这些都被点到了,但没有专门分论坛。它们被登记为下一次论坛的候选议题。

结束语

主持人在收尾时把开场的问题——当 *AI* 越来越自主时,谁来负责?——重新表述为:“这不是任何一场分论坛能独立回答的问题。” 在场所凝聚的,反而是一种共享语言的开始——“协同”作为契约而非感觉,“学习”远不止“效率”,“信任”是工程出来的而非表演出来的,“公共物品”是设计选择而非事后添加。

这一语言,是巴黎对更广泛 AI 对话的一份贡献。在场所发出的邀请——也是本份报告向更广泛读者发出的邀请——是把这套词汇带向 GOSIM 巴黎、带向十月的深圳、带向智能体转型继续展开的每一个房间。

致谢

AI 愿景论坛 · 巴黎 2026 的举办,得益于组委会、GOSIM 巴黎团队,以及来自世界各地——其中一些远道而来——参与本次论坛的研究者、创业者、基金会代表、监管者与教育工作者的工作。

本报告基于当日全程录音(一场主旨演讲加四场分论坛,合计约六个半小时)整理与编辑而成。直接引语如实呈现,姓名与所属机构按照查塔姆研究所规则均已隐去。当录音存在断裂或难以辨认的段落时,这些段落被视为静默处理,未做推测性补全。

向所有为本次活动提供主办、后勤与运营支持的合作机构致谢;同样致谢所有与会者——是他们的准备、坦诚,以及在同一个房间里愿意有建设性分歧的姿态,让这一天值得被记录。

论坛网站:*paris2026.visionforum.ai* 巴黎倡议:*paris2026.visionforum.ai/initiative/* · *GitHub: github.com/aivisionforum/paris-2026-initiative* 教育研究论文:*paris2026.visionforum.ai/education/* 论坛报告发布:*2026* 年 *5* 月 与 *GOSIM* 巴黎 *2026* 联合举办