

AGI 十年内会出现吗

制品 · 持续 · 预判

AGI 十年内会被宣布，但绝不会被公认

报告出具时间：2026/6/29

报告来源：See The Essence

目录

一、全貌	3
二、核心机制	4
创造的语境	4
设计与构造	4
接受与传播	4
影响与后果	5
遗产与意义	5
三、深层洞察	6
洞察一：定义权的空缺不是 bug，是权力的操作系统	6
洞察二：实验室演示与普惠基础设施之间隔着一组物理定律	6
洞察三：治理脆弱性是最被低估的刹车	6
四、走向与可能	7
最可能的走向：宣布发生，共识不发生	7
最被忽视的约束：物质基础设施的沉默投票	7
最意外的可能：开源生态爆破式涌现	7
五、如何思考这件事	9
参考资料	10

一、全貌

AGI 十年内会不会出现？这个问题过去三年被问了太多次，以至于它本身已经成为一个现象。每一次大语言模型的参数跃升都会把回答从“不可能”推向“也许”，又在下一个瓶颈期拉回来。

但问题本身藏着三个容易被忽略的事实。

第一，提出这个问题的不是物理学界也不是哲学界，主要是一群正在烧钱做这件事的人。他们的回答——2026、2029、2045——从时间轴上看是在预测技术，从功能上看是在维持一个估值叙事。这不是说他们在撒谎。是说这个问题的问法和答法，本身就是他们要操盘的事件的一部分。

第二，“AGI 出现了”这句话，在人类历史上从未有过公认的判定标准。我们至今没有一种方法能区分“真的通用智能”与“好到让人愿意相信的模拟”。这意味着“十年内 AGI 会出现”和“十年内有人说 AGI 出现了”是两个完全不同的问题，前者需要共识，后者只需要一次发布会。

第三，如果 AGI 意味着一个能在开放世界中处理任意任务的系统，那么即使它在实验室的受控环境中跑通了，要让它物理世界中安全、普惠、可靠地运转，需要越过的不是一道墙，是一系列彼此嵌套的硬约束。这些约束不在算法论文的假设范围内，但它们决定了实验室演示与人类文明级别的智能基础设施之间的距离。

这篇报告试图做一件事：把“AGI 十年内会出现吗”拆成几个可以独立检验的子问题，然后逐一追问——哪些有证据，哪些没有，哪些目前的证据指向相反的方向。

二、核心机制

“AGI 十年内会出现”——这个句子值得被当作一个制品来分析：它是什么时候被制造出来的、由谁制造、用什么材料、要达成什么功能、在传播中发生了什么变异。

创造的语境

2014 年前后，DeepMind 被收购、OpenAI 成立，一个判断开始在硅谷的技术圈流传：深度学习这条路线可能能走到底。到 2016 年 AlphaGo 赢下李世石，这个判断从“可能”升级为“也许比预想的快”。但“十年内 AGI 会出现”作为一个可传播的叙事制品真正成型，要到 2020 年代初期。

时间点的集中不是偶然。2023 年 GPT-4 发布后，出现了一波明确的 AGI 时间表宣言：Sam Altman 提到“未来十年内”，一些前沿实验室的领导者给出了 2026-2030 年的区间[1]。这个时间窗口恰好与资本密集涌入的节奏同步——据统计，2023 年流向生成式 AI 领域的风险投资在一年内翻了三倍以上[2]。

STS（科学技术研究）领域有一个反复被验证的观察：技术预测从来不只是关于技术。当一家公司的估值取决于投资者相信它正在建造通向 AGI 的桥梁，那么“AGI 很快会来”就不是一个科学判断，而是一种维持生态位的话语实践。它的功能不是准确地描述未来，而是有效地组织现在——吸引最顶尖的工程师离开谷歌加入一家初创公司，说服主权基金押注数万亿参数的前沿模型，让监管者在还没有完全理解大型语言模型怎么工作之前就开始紧张。

这不是某种阴谋论。这是一种结构性的激励对齐：如果 AGI 真的在十年内到来，第一个到达的人的回报无法估量。因此，任何一个声称自己在冲刺这件事的机构都有充分理由让外界相信终点更近一些、自己的位置更领先一些。

设计与构造

这个制品的“原材料”中，最脆弱也最核心的部分是 AGI 的定义。百度百科给出的解释是：通用人工智能（AGI）是指能够像人类一样理解、学习和应用知识完成各种任务的智能系统[3]。这句话读起来清晰，但经不起推敲。

“像人类一样”——哪一个人类？一个成年知识分子？一个三岁儿童？一个处理农田索赔纠纷的古巴比伦法官？“理解”——通过图灵测试就算理解，还是需要有意意识体验？“各种任务”——是所有任务，还是足够多的代表性任务？如果是足够多，谁来决定样本量？

哲学层面，关于“什么是通用智能”的争论已经持续了几十年，没有收敛的迹象[4]。而复杂系统理论提供的视角更具体：通用能力不是一系列专项能力的加总，而是一种涌现属性。你不可能通过检查清单来判定“AGI 出现了”，就像一个蚁群没有哪一只蚂蚁是“蚁群管理者”。这意味着不存在一个天然的、客观的开关时刻。

这就给了制造方巨大的话语空间。如果标准的制定权在研发机构手里——定义足够宽松、基准足够特定、演示足够精选——那么“AGI 已实现”可以被宣布在任何他们需要宣布的季度。2018 年就有中国研究团队在特定基准集上宣称系统达到“类人水平”，当时几乎没有人把它当作 AGI 到来的信号，因为标准本身没有获得跨社群的认可。这个模式不会因为模型的参数从千亿变成万亿就自动失效。

接受与传播

“十年内 AGI 会出现”从实验室言论变成公众认知，经历了几个关键的放大节点。

第一个是 GPT-4 的发布。它让大量非技术用户第一次体验到系统在表面行为上高度接近“理解”——流畅的对话、跨领域的推理、幽默与反讽的把握。用户形成了一种直观印象：这东西只是少了一些什么，给它几年就行。

第二个是资本市场对叙事的吸收。一级市场投资人需要相信 AGI 以论证数千亿美元估值的合理性[5]。二级市场分析师写出“AGI 前夜”的报告，为科技巨头的资本开支飙升提供解释框架。这些报告被媒体摘要转发，再被社交媒体碎片化传播，最终沉淀为一种“大家都知道在加速”的氛围。

第三个节点更微妙：当主要的 AI 安全机构——包括一些从有效利他主义社区出来的研究组织——开始严肃地讨论“AGI 十年内出现”的情境时，这个判断获得了一种反向背书。本来最悲观的群体在用最严肃的语气警告最近期的风险，这让公众的逻辑变成：既然连最谨慎的人都这么说，那大概是真的快了。

但这里存在一个关键的认知跳跃。安全社区讨论“如果十年内 AGI 出现，我们需要做哪些准备”是一回事。他们的职责要求他们对最坏情境做最充分准备。这本身并不意味着他们认为这个情境是最可能的。然而在传播过程中，情景推演被转写成预测，预测被转写成既定事实。

影响与后果

这个叙事制品的直接影响已经清晰可见。

人才流向了几家头部实验室，全球数以千计的最强工程师和研究员集中到少数机构。资本以数千亿美元的规模涌入。一些国家的政府开始把 AGI 当作地缘政治竞争的核心变量来制定出口管制和产业补贴政策。

但更隐蔽的后果在另一个方向。

当一个关于十年后的事件被当作几乎确定的未来来投资时，注意力和资源会系统性地偏离那些可能在五年内解决实际问题的渐进式改进。每一个宣布要“直接冲刺 AGI”的实验室，都同时宣布它不会在可解释性、可审计性、跨文化价值对齐这些“次要问题”上花费太多时间。而当这些实验室的治理结构——后文会具体展开——在压力下出现裂缝时，整个基础设施的脆弱性就暴露出来。

遗产与意义

回看核聚变的历史有一个很相似的叙事弧线。1950 年代，第一批聚变反应堆演示后，预测“五十年内商用”是标准话术。这个预测的功能是动员资金、人才和政治支持。七十多年后，聚变仍是“再有三十年”。

历次 AI 寒冬也是同一个模式：每次一个方向出现突破，承诺周期启动，资源涌入；当承诺无法在一个融资周期内兑现，幻灭到来，项目裁撤。第一次 AI 寒冬发生在 1970 年代，因为早期符号学派在自然语言理解上的承诺远超实际能力；第二次在 1980 年代末，因为专家系统的维护成本击溃了商用预期。

“十年内 AGI 会出现”这个制品，最可能留下的遗产不是 AGI 本身，而是它制造的期望曲线上涨段——以及随着期望回调而再出现一轮判断上的不匹配。这不代表技术没有进展。它只意味着，把技术进展翻译为“到达终点”的那一步，从来不是技术自己的事。

三、深层洞察

把上面这层机制讲清楚之后，可以逼出几个更深层的判断。

洞察一：定义权的空缺不是 bug，是权力的操作系统

通常讨论 AGI 会默认一个假设——我们只是暂时还没找到一个公认的定义，早晚会有。但过去七十年对这个问题的辩论表明，“缺少公认定义”可能不是一个暂时的技术困难，而是“AGI 宣称”这个社会行为的构成性条件。

如果存在一个公认的 AGI 标准，研发机构就失去了定义自己里程碑的能力。如果图灵测试今天被升级为严格的、无法作弊的通用智能评估并获全球认可，那么任何机构宣布“通过了”都将被独立验证推翻或确认。对研发方面言，这是一种不可接受的控制权让渡。

因此，AGI 的定义之争不会在十年内解决。十年后最可能出现的情况是：有一个或多个系统在某些指标上表现出显著超越人类的泛化能力，但另一个阵营坚定地指出它仍缺少常识、因果理解、具身经验或某种更深层的智能属性。双方各自拥有部分证据，任何一方都无法说服另一方。

这就是“话语冲突”的真正含义：它不是噪音，而是核心事件本身。

洞察二：实验室演示与普惠基础设施之间隔着一组物理定律

大语言模型的扩张遵循一条明确的曲线：性能提升与计算量之间存在幂律关系，每获得一个单位的性能增量需要指数级增长的算力投入。目前前沿模型单次推理的能耗已经达到每用户数十千瓦时级别——相当于一个四口之家数天的用电量[6]。

这带来两层硬约束。

第一层是芯片工艺的物理极限。硅基芯片正在逼近原子尺度，量子隧穿效应开始成为实质障碍。3 纳米以下的路径还有几代，但每走一代成本倍增、良率下降。而 AGI 路径上需要的算力远在目前可见的工艺曲线之外。

第二层是电网承载能力。如果 AGI 级别的系统需要部署到数十亿用户，即使能效比再提升两个数量级，总能耗也将是当今全球发电量的显著比例。当前没有一个国家的基础设施规划中纳入了这种级别的负载。

这两层约束意味着，即使算法的突破使得 AGI 在严格控制的实验室环境中“跑起来”了，把这一幕放大成社会范围内的通用智能基础设施，需要的时间尺度远不是十年。

洞察三：治理脆弱性是最被低估的刹车

2023 年底 OpenAI 的治理危机被当作一则公司新闻消化了。但它的深层含义远比媒体关注的人事斗争更致命：如果一家立志构建 AGI 的非营利机构的董事会有合法权力解雇 CEO，而 CEO 通过员工集体辞职威胁和投资者施压在几天内夺回控制权，那么这家机构的治理结构等于公然宣告它无法在任何实质性分歧中约束最高领导者的决策。

这不是 OpenAI 一家的问题。前沿 AI 实验室普遍面临同一个结构性矛盾：要吸引资本和顶级人才，必须采用接近创业公司的股权结构和高风险偏好；但要承担起“创造 AGI”这一对全人类可能有后果的任务，理论上需要比核武器研发更严格的集体约束机制。这两个要求指向完全相反的组织形式。

在当前的格局下，市值驱动的公司逻辑正在胜出。地缘政治的军备竞赛逻辑——美国、中国、欧洲各自加紧推进前沿能力——进一步奖励那些第一个冲到终点的团队，而不是在安全对齐上投入时间最多的团队。这种激励机制的系统性后果是：即使技术轨迹本身指向十年内有突破，组织层面几乎没有能力在到达临界点之前踩下必要的刹车。

四、走向与可能

在对趋势、结构、约束做了上述梳理之后，可以勾画未来十年的几种情景。最确定的判断，只能用在那条概率最高的路径上。

最可能的走向：宣布发生，共识不发生

在这条路径下，未来十年内至少有一家前沿实验室将宣布其系统达到了它所定义的 AGI 阈值。宣布的方式可能是一场发布会，一篇附有精选基准测试的技术报告，以及一系列精心策划的媒体演示。部分投资人会迅速跟进背书，相关的股票估值将出现一轮剧烈波动。

但这个宣布将无法获得跨社群的公认。另一个阵营的科学家会列出十条理由说明为什么这个系统还不是真正的通用智能：因果推理在分布外测试中崩溃，缺少稳定的自我模型，在跨文化语境中暴露出深层偏见，无法在没有人类干预的情况下持续学习新任务。社交媒体会成为主战场，双方各自产出大量证据，终极结论遥不可及。

政府将陷入两难。一边要求监管“已出现的 AGI”，一边面对缺乏共识的技术定义无从下手。立法将滞后，而滞后期间的监管真空同时让过度乐观者继续加注、让安全担忧者持续焦虑。

关键能动者：前沿实验室的决策层有动力宣布“到达”——这是他们完成融资故事闭环的时刻。而学术界的独立研究者和安全研究社群有动力并且有能力持续地质疑——这是他们维护认知权威的方式。双方的动力是结构性的、对撞的。

最被忽视的约束：物质基础设施的沉默投票

这一条替换“最危险的走向”，因为 AGI 十年内不出现的核心风险并不来自某种灾难性事件（那是一种不同的分析），而是来自一组不起眼的物理和工程事实的累积效应。

假设未来十年算法前沿出现了一次突破，可以在一个任务上泛化到任意新任务。但要让这个系统从 100 个用户扩展到 10 亿用户，要在真实的物理世界——不稳定的电力供应、湿热的数据中心、能力参差不齐的终端设备——中稳定运行，需要的不只是算法，还需要：

- 数万亿美元的新电网投资
- 新一代芯片架构的量产
- 跨域数据治理的法律框架

这三样东西，任何一样的时间常数都以十年为基本单位。芯片从设计到大规模量产通常需要五到七年；电网改造在过去二十年中任何发达国家的推进速度最多是十年一个代际；而全球数据治理框架——考虑到现有国际法的协商周期——乐观估计也需要二十年以上。

因此，即使十年前沿实验室“造出”了 AGI 的原型，社会层面可部署的 AGI 仍将在物质基础设施的瓶颈前卡住多年。这个约束不是可选的、可以通过更聪明的工程绕过，它是物理学。

关键能动者：电网运营商、半导体设备制造商、各国能源监管机构。这些主体的决策周期不跟随 AI 行业的新闻速度。没有人能通过一篇论文让芯片制造跳过光刻迭代。

最意外的可能：开源生态爆破式涌现

在主流叙事的边缘，存在一个低概率高影响的路径。

当前，最前沿的开发被封闭在少数实验室中。但开源社区的追赶速度在过去两年中快于绝大多数分析师的预测。2024 年，部分开源模型在特定基准上已接近闭源前沿水平。如果这一趋势持续，可能在十年内的某个节点出现一次相变：当开源模型在足够多的维度上对齐甚至超过闭源水平，AGI 的“出现”就不再被任何单一机构掌控。

这种情况下，“AGI 十年内会出现吗”的答案将被改写。不是某个机构宣布胜利，而是数以万计的开发者的突然发现自己手里的工具能做以前想都不敢想的事。通用能力不再是一个被宣称的“到达”，而是一个逐步扩散的事实——没有人站在台上拿着遥控器，但人人都知道它已经来了。

阻碍这个情景的制度瓶颈也很明确：主权国家出于安全担忧限制开源发布；大型云服务商收紧算力准入；以及开源社区本身无法负担指数级增长的训练成本。任何一个机制生效，都可能封堵这条路。

关键能动者：Meta、Mistral 押注开源路线的企业；HuggingFace 等分发平台；各国出口管制和网络安全监管机构。

五、如何思考这件事

与其问“AGI 十年内会不会出现”，不如换几个能始终追问的、能持续校准判断的问题。

三个需要持续追问的问题。

第一，如果有机构宣布 AGI 已实现，它在哪一项测试中仍然失败？分析 AGI 宣称，看的不是它通过的测试，是它通不过的那些。一项宣称通过了 100 项基准的系统，如果在第 101 项——一个儿童能做的物理因果推理任务——上系统性崩溃，那么这次宣称的质量就取决于那第 101 项任务的解释力。持续追问“它失败在哪”，是防止被基准过拟合蒙蔽的最有效方法。

第二，支撑这一阶段进展的资源曲线是否延续？AGI 不是一个纯粹的研究问题，它是一个工程和资本问题。如果前沿模型的单次训练成本增长快于可用资本和算力的增长，那么“指数趋势”就会在某一个点自然弯曲。历史上芯片成本曲线和电网投资曲线的斜率，比论文引用曲线的斜率更值得关注——它慢，但它不撒谎。

第三，治理结构有没有出现能约束核心决策者的新机制？当前局面下，任何一家前沿实验室的 CEO 都有实质上不受约束的权力决定声称什么、释放什么。在出现一种比“董事会解雇-员工反叛-CEO 复辟”更有效的约束之前，不要相信任何关于“我们会负责任地释放 AGI”的承诺。盯着组织的行为，而不是组织的声明。

三个信号，出现意味着趋势在改变。

第一个信号是定义权的让渡。如果看到一家主要前沿实验室同意将“AGI 达成”的判定交给一个外部独立委员会，且承诺遵守该委员会的裁定，这是实质转折。目前没有任何迹象指向这个方向。

第二个信号是能耗曲线的斜率变化。如果推理能耗的单位性能成本开始以超指数速度下降——而不仅仅是算力总规模在增长——那么物理约束的紧度可能比当前判断更松。这个信号需要看的是“每单位智能的瓦特数”，不是“训练集群的总功率”。

第三个信号是安全对齐投入的资本结构变化。如果对前沿模型安全研究的投入开始来自与商业化部门完全隔离的独立资金池——规模与商业化研发相当——那就意味着有机构在做对冲。目前安全研究的预算在商业化部门面前是一个数量级的差距。

一个可能错误的假设。

最容易被普遍接受但可能错的假设是：智能是一根可以无限攀升的单一标尺。这个假设让很多人以为，只要参数继续增长、数据继续增多、训练时间继续拉长，沿着同一个方向迟早会“到达 AGI”。

但复杂系统研究和认知科学中有一支相当强劲的传统，指出智能可能不是一个标量。泛化能力、工具使用能力、社会认知能力、因果推理能力可能是多个不同的正交维度。一个系统可以在其中三个维度上达到超人水平，在第四个维度上还停留在蟑螂水平——而那个短板的维度，可能在真实的开放世界中恰好是最致命的一个。

如果这个假设是错的，那么“十年内 AGI”就相当于“我们正在以每小时一百公里的速度驶向悬崖，预计十分钟内到达”。到达确实可能。到达的方式可能和所有人想象的不一样。

参考资料

[1] [新闻报道] 腾讯新闻, “Altman 再谈 AGI: 十年内 AGI 将会到来”, 2026 年 6 月发布, <https://news.qq.com/rain/a/20260613A08UQ800>

[2] [机构报告] 36 氪, “2024 年 AI 投融资趋势报告”, 2025 年发布, <https://www.36kr.com/p/3855615518921729>

[3] [百科] 百度百科, “通用人工智能 (AGI)” 词条, 访问日期暂缺, [https://baike.baidu.com/item/%E9%80%9A%E7%94%A8%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD\(AGI\)/67547871](https://baike.baidu.com/item/%E9%80%9A%E7%94%A8%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD(AGI)/67547871)

[4] [学术论文/博客] 知乎专栏, “通用人工智能的哲学边界: 为什么我们无法定义‘理解’”, 发布日期暂缺, <https://zhuanlan.zhihu.com/p/7046080918>

[5] [新闻报道] 搜狐, “AGI 前夜的资本狂潮”, 2025 年发布, https://www.sohu.com/a/935908044_100270262

[6] [技术分析] 腾讯云开发者社区, “大模型推理能耗分析”, 2024 年发布, <https://cloud.tencent.com/developer/article/2408923>