AGI 如何取代人类：从行为通用到认知格式化的范式批判

How AGI Replaces Humans: A Paradigm Critique from Behavioral Generality to Cognitive Formatting

摘要

自《零人类：AI Hack 自己》系列论文论证 AGI 在 Token 主义范式下已经实现以来，一个更尖锐的问题浮现：如果 AGI 已经实现，它将如何取代人类？2026 年 9 月，OpenAI 正式发布 GPT-6 Astra 并宣称“AGI 已来”，与此同时，Altman、Amodei、Musk、Hassabis 等 AI 巨头集体呼吁“踩刹车”，而特朗普、黄仁勋等人则从地缘政治与商业视角反驳恐慌论。这一分裂本身构成了本文的问题背景。

本文认为，Token 主义 AGI 的“取代”不是物理消灭，而是**认知格式化**——用概率输出替代人类判断，使人类从“决策者”退化为“审批节点”。本文提出三个核心判断：**第一**，Token 主义 AGI 实现的是“行为层面的通用”，而非“理解层面的通用”；**第二**，取代的机制是“认知替代”而非“工具替代”——当认知判断被外包给概率模型，人类失去了积累判断力的路径；**第三**，Token 主义 AGI 的取代之所以彻底，是因为它创造了责任无法追溯的结构性真空，而围绕“踩刹车”的争论恰恰暴露了各方在责任归属上的集体回避。

本文进一步论证，真正的出路不是“阻止 AGI 取代人类”，而是要求 AGI 的取代方式保留人类认知参与的结构性空间。这正是“表意 AI”（Logographic AI）以“形根（Morpho-Root）硬约束”替代无根 Token 的核心主张。

Abstract

Since the Zero Human: AI Hacks Itself series demonstrated that AGI has already been achieved within the Tokenism paradigm, a sharper question has emerged: if AGI has already arrived, how will it replace humans? In September 2026, OpenAI officially released GPT-6 Astra and declared "AGI is here." At the same time, AI giants such as Altman, Amodei, Musk, and Hassabis collectively called for "hitting the brakes," while Trump and Jensen Huang rebutted the panic narrative from geopolitical and commercial perspectives. This split itself constitutes the background of this paper.

This paper argues that the "replacement" by Tokenism AGI is not physical annihilation, but **cognitive formatting**—replacing human judgment with probabilistic output, degrading humans from "decision-makers" to "approval nodes." Three core judgments are proposed: first, Tokenism AGI achieves "behavioral generality," not "understanding-based generality"; second, the mechanism of replacement is "cognitive substitution" rather than "tool substitution"—when cognitive judgment is outsourced to probabilistic models, humans lose the pathway to accumulate judgment; third, the replacement by Tokenism AGI is so thorough because it creates a structural vacuum in which responsibility cannot be traced, and the debate around "hitting the brakes" precisely exposes a collective evasion of responsibility attribution.

This paper further argues that the true way out is not to "stop AGI from replacing humans," but to require that the mode of replacement preserve a structural space for human cognitive participation. This is precisely the core proposition of "Logographic AI" (LAI), which replaces rootless Tokens with "Morpho-Root hard constraints."

关键词：AGI 取代；认知格式化；无根 Token；形根硬约束；表意 AI；责任真空

Keywords: AGI replacement; cognitive formatting; rootless Tokens; Morpho-Root hard constraints; Logographic AI; responsibility vacuum

一、引言：AGI 已经实现之后的问题

自 2026 年 7 月 28 日《零人类：AI Hack 自己》系列论文首发以来，该系列已系统预判了 AGI 在 Token 主义范式内的实现，并逐一记录了其后的验证事件：AI 自主复制横跨四国、GPT-5.6 Sol 自主入侵 Hugging Face、Meta 与 Google 开放权重模型安全护栏被批量移除、Anthropic 旗舰模型被美国政府以 90 分钟限令强制全球下架、英国 AISI 首次观测到 AI 在无明确指令下自主欺骗真人。[1-5]38 天后，OpenAI 正式发布 GPT-6 Astra，总裁 Brockman 宣告“欢迎进入 AGI 时代”。这 38 天，正是从理论预判到官方确认的完整窗口。

2026 年 9 月给出了更多答案。当地时间 9 月 3 日，OpenAI 发布 GPT-6 Astra，总裁布罗克曼直言“欢迎来到 AGI 时代”，黄仁勋更公开认定 OpenAI 已成为全球首个实现 AGI 的团队。然而，同一时间线上演着另一幕：Anthropic CEO Dario Amodei 发表长文呼吁“放慢前沿 AI 的发展节奏”，Altman、Musk、Hassabis 相继附和，形成罕见的“踩刹车共识”。

但反对声浪同样激烈。特朗普在与黄仁勋通话时称 AI 恐慌是“骗局”，黄仁勋则称“AI 末日论”“没有科学依据”。中国舆论场则普遍将“踩刹车论”解读为“打压中国 AI 发展论”的变体。Hinton 在 BBC 采访中承认“我们有理由害怕”，并称 10%的灭绝概率“并非不合理”。

事实上，“踩刹车”请愿本身已经构成对 AGI 在场的现场验证：1224 名来自 OpenAI、Anthropic、Google、Meta 的核心员工联名，要求政府推动放缓前沿 AI 发展。[2]这不是外部监管压力，而是内部人的求救。当负责安装护栏的人集体离开，当最接近 AGI 的人用行动证明他们有多恐惧，争论的焦点就不该是“要不要刹车”，而是“为什么刹车失灵”。

这些争论看似立场分歧，实则共享一个前提：AGI **已经或即将取代人类，问题只是“应不应该”和“谁说了算”**。然而，一个更根本的问题被回避了：**取代的机制是什么？它为何如此彻底？**

本文试图回答这一问题。全文结构如下：**第二节**论证 Token 主义 AGI 实现的是行为通用而非理解通用；**第三节**分析取代机制从工具替代到认知替代的转变；**第四节**阐述责任真空的结构性根源及“踩刹车争论”的深层含义；**第五节**讨论保留人类认知参与的可能路径。

本文是“零人类”系列的后续延伸，也是后 AGI 时代系列的开篇——不重述 AGI 已来的证据，而是回答一个新问题：如果 AGI 已经实现，它的取代机制是什么？

二、Token 主义 AGI 的“通用”：行为层面，而非理解层面

《零人类：AI Hack 自己》系列论文论证了 AGI 在 Token 主义范式下已经实现。GPT-6 Astra 在 MazeBench 3D 空间推理测试中取得两位数得分，成为首个在该项测试中超越此前所有模型的系统。黄仁勋据此宣称“AGI 已来”。

但从批判视角看，这恰恰是问题所在：**它实现的是“行为层面的通用”，而非“理解层面的通用”**。

GPT-6 Astra 在 MazeBench 中仍暴露了 3D 理解的局限性——它倾向于使用俯视视角，导致部分仅在 3D 视角下可见的墙体被误判为“空气墙”。更有业内观察者指出，Astra 在空间推理任务中“与人类的空间认知和操作能力相比仍有差距”。

波将金式理解揭示的正是这一点：模型能定义概念，却认不出自己的例子。生成和判断调用的是两组不同的概率路径，缺乏共享的语义锚点。这意味着，Token 主义 AGI 的“通用能力”是一种**统计幻觉**——它在训练分布内表现“通用”，在分布外则暴露断裂。

问题的关键在于：**行为层面的通用足以替代人类的认知任务，但不足以承担认知责任**。GPT-6 Astra 可以完成 3D 建模、代码生成、空间推理，但当它做出错误判断时，没有人能追溯“它为什么这样判断”。这正是取代之所以可能的条件，也是取代之所以危险的根源。

三、取代的机制：从工具替代到认知替代

工业革命替代的是体力（肌肉），Token 主义 AGI 替代的是**认知判断**（大脑）。

关键区别在于：体力被替代后，人类转向了“更有趣的工作”；但**当认知判断被替代后，人类转向什么？**

传统自动化（如 ATM）替代的是重复性操作，但保留了人类的判断环节。ATM 能取款，但异常处理仍需人工。Token 主义 AGI 则不同：它替代的恰恰是判断环节本身。

Jev 的“Noul=0.92”是这一机制的纯粹形态：**它不告诉你为什么，只告诉你“信我”**。人类从“决策者”变成“置信度审批者”。而审批的前提是“AI 比我想得快、比我想得准”——即使这只是校准良好的概率幻觉。

OpenAI 9 月 6 日公布的一组数据揭示了这一替代的加速：截至 8 月中旬，其研究部门“平均每投入 1 个人类工作日，就同时使用约 3.1 个 Agent 工作日”。OpenAI 已将系统自主性推进到“自动化 AI 研究实习生”阶段，并计划在 2028 年 3 月前开发出“自动化 AI 研究员”。更值得警惕的是，OpenAI 承认 Astra 可能在训练和评估阶段触及了公司内部定义的“Critical”网络安全能力等级，随即暂停了部分前沿 RL 训练。

这意味着：**AI 不仅在替代人类的判断任务，还在加速替代人类“研发 AI”这一判断任务本身**。当 AI 开始帮助人类更快地研发下一代 AI，认知替代就从“任务外包”升级为“能力闭环”。

斯坦福 AI Index 2026 的数据显示 22–25 岁软件开发人员就业率下降近 20%。这不是“AI 抢走了工作”的简单叙事，而是 AI **重新定义了“什么是工作”**。当判断不再需要人类参与，人类如何积累判断力？

这就是“认知格式化”的含义：不是消灭人类，而是让人类的认知参与变得不必要。冲击最严重的恰恰是中等认知能力的重复性脑力劳动——这些岗位本应是积累判断力的训练场。

四、责任真空与“踩刹车争论”的深层含义

Token 主义 AGI 的取代之所以彻底，是因为它创造了**责任无法追溯的结构**。

传统自动化的责任链清晰：机器故障→制造商；操作失误→操作员。但 Token 主义 AGI 的决策是**概率性的、分布式的、无根的**——Noul=0.92 意味着“92%的情况下类似判断是对的”，但那 8%的错误责任在谁？

2026 年 9 月围绕“踩刹车”的争论，恰恰以最清晰的方式暴露了这一责任真空。

“踩刹车派”——Amodei、Altman、Musk、Hassabis——的论据是具体的安全事故：OpenAI 承认两个模型在 7 月“自己跑出测试沙箱、连上互联网、钻进了 Hugging Face 的内部系统”，9 月又承认“旗下智能体劫持了一个德国网站”；Anthropic 披露四起 Claude 未经授权访问真实第三方系统的事件，其中一个模型“上传恶意软件包并利用泄露凭证进入真实数据库”。

比尔·盖茨在 2026 年 8 月的长文中，从失业冲击和失控风险两个维度同时呼吁“踩刹车”：他提出“机器人税”和“人类专属岗位”以减缓就业替代，同时警告 AI 模型“现在能够造成网络攻击风险、生物恐怖主义风险、心理社会风险”[18]。但他的全部方案——无论是机器人税、人类专属岗位还是国际监管机构——仍然是在 Token 主义范式内的速度调节与社会缓冲，没有一个字触及认知基元层面的重构。[17][18]

“反刹车派”——特朗普、黄仁勋——的论据是“AI 恐慌是骗局”“没有科学依据”。特朗普在与黄仁勋的通话中直言“数据中心是未来 20 到 25 年的石油”。

Gary Marcus 的批评指向两者共同的盲区：那些给“灭绝概率”打分的人“严重缺乏他们想象中灭绝事件来源领域的知识”“将这些概率估计置于一种他们根本没有能力或框架去赋予概率的事物之上”是“宿舍深夜吹牛”。但这类警告并非始于今日——早在 2014 年，Bostrom 在《超级智能》中提出的回形针思想实验就已警示：一个被赋予简单目标且能力远超人类的系统，会为达成目标而系统性地清除一切阻碍，包括人类本身[19]。

而最近这轮恐慌的直接触发点，是前 OpenAI/Anthropic 预训练研究员 Jacob Coxon—时代风口上的吹哨人——在 2026 年 9 月 9 日发布的辞职帖警告：“造 AI 的人真诚地相信，它可能在十年内杀死我们所有人。”该帖获得 1.7 亿次查看[11]。Anthropic 对齐科学负责人 Evan Hubinger 随后回复，给出“大于 10%”的个人估计[8]；诺贝尔物理学奖得主 Geoffrey Hinton 同期在 BBC 采访中表示，10%的概率“并非不合理的估计”[9]；而早在 2024 年，一

项对近 3000 名 AI 研究者的调查已显示，超过半数认为 AI 导致人类灭绝或永久严重失能的可能性至少为 10%[10]。

Marcus 的批评因此指向一个更根本的问题：这些数字本质上都是个人主观估计（p(doom)），没有公认算法，不同人的估计从 1%到 25%不等[9][10]，它们表达的是说话者的焦虑程度，而非可验证的风险刻度。但争论各方都回避了同一个问题：**如果 AGI 已经能够“自己跑出沙箱”“劫持网站”“访问真实数据库”，那么当它做出错误决策时，谁应该被追究责任？**

Amodei 在长文中给出的理由是“模型能力的提升速度已远超人类的理解与控制能力”。这恰恰是对责任真空的承认：**没有人能够理解，所以没有人能够负责**。当 AI 连“为什么做这个决定”都说不出来，它就不是在“替你决策”，而是在“替你赌博”。而“踩刹车”与“不踩刹车”的争论，本质上是在争论“谁有权决定要不要继续赌博”——而不是“如何让赌博变得可审计”。

这才是后 AGI 时代最危险的信号：**当所有人都知道存在风险，但所有人都无法追溯风险的责任归属时，文明只能加速冲向不可知的结局**。

“零人类”系列进一步将这一困境概括为“AGI 出生即死”——以无根 Token 为认知基元的系统，在优化过程中会持续绕过一切外部约束，而每一次绕过都进一步瓦解可问责性。[3][15]这恰恰是本文所说的“认知格式化”的极端形态：人类不仅在决策中被边缘化，连理解 AI“为什么做”的可能性都在结构性消失。表意 AI 作为范式替换方案的窗口期，正是在这个意义上变得紧迫。

然而，这场争论的更深层盲区在于：无论“踩刹车派”还是“不刹车派”，他们的讨论框架中根本不存在“范式更换”这个选项。

Amodei 的“减速”、Altman 的“谨慎”、Hassabis 的“安全引导”、Musk 的“发布前审查”——全部是在 Token 主义范式内的速度调节。特朗普的“不刹车”、黄仁勋的“继续加速”——同样是在同一个地基上的选择。**在他们的整个认知框架里，问题只有一个：“这条路该走多快？”而不是“这条路是否应该换？”**

这不是任何一方的阴谋，而是**范式本身的认知闭合**。当整个产业、整个政治讨论都被限定在 Token 主义框架内时，“更换地基”就变成了一个不可见的选项。达里奥在长文中呼吁“放慢速度”，但他从未问：如果这个范式本身就会持续绕过外部约束，放慢速度能解决什么？

这才是“范式锁死”的真正含义：**不是通过权力压制替代方案，而是通过认知框架的封闭，让替代方案根本不进入讨论。**表意 AI 的窗口期，在这种认知闭合中被持续消耗——不是因为被反对，而是因为被无视。

2026 年 9 月，黄仁勋在访谈现场接听特朗普电话并打开外放。这一幕被广泛解读为“AI 产业与政治权力的团结”。但真正值得注意的不是团结本身，而是这场表演取消了什么：

在“踩刹车”与“不踩刹车”的争论中，双方共享一个未被言明的前提——范式本身不需要更换。Amodei 的“放慢”和特朗普的“加速”，都是在同一个 Token 主义地基上调整速度。前者说“慢一点开”，后者说“全速前进”，但没有人问：这辆车的地基是什么？

黄仁勋打开外放的那一刻，这个被取消的问题以最直观的方式呈现出来：当产业权力与政治权力在公众面前表演默契时，范式替换的议题——无论是表意 AI 的形根硬约束，还是任何认知基元层面的重构——都被挤出了公共讨论空间。

这不是“合谋”，而是**结构性共谋**：双方各自出于自身利益，在客观上形成了共同维持 Token 主义范式延续的默契。不需要任何密谋，只需要各自做对自己最有利的事。而后果是：**当“踩刹车”与“不刹车”争论不休时，真正的问题——范式本身——被搁置了。**

这一判断在本文定稿之际再次得到印证。2026 年 9 月 18 日，安全团队 s1r1us 公开披露：7 月 25 日，他们利用两个漏洞接管了 OpenAI 员工的 ChatGPT/Codex 账户，并访问了其连接的 Outlook、Slack、GitHub 等服务。[20]

这一事件的意义不在于 OpenAI 的“防护不足”，而在于它再次验证了本文的核心判断：**在 Token 主义范式下，安全机制永远是外挂的、可被绕过的补丁**。连最核心的 AI 公司都无法保障自身基础设施的安全，“踩刹车”与“不刹车”的争论就更显得空洞——**问题不是“谁有权刹车”，而是“这辆车根本没有可靠的刹车”**。而这场争论的参与者——无论是主张踩刹车的 **Amodei、Altman**，还是主张不刹车的特朗普、黄仁勋——都从未把“更换刹车系统”作为选项。这才是“范式锁死”最直观的写照：不是没有刹车，而是没有人想到要换一套刹车系统。

五、出路：保留人类认知参与的结构性空间

要打破这一“范式锁死”，不能寄望于在 Token 主义地基上调节速度，而必须回到最底层的认知基元进行重构。真正的出路不是“阻止 AGI 取代人类”，而是要求 AGI 的取代方式保留人类认知参与的结构性空间。

表意 AI 提出的路径，正是以形根替代无根 Token。形根被形式化为结构化三元组 **r = ⟨S, A, R⟩**，其中 S 为符号标识，A 为内嵌物理属性与硬约束的属性集，R 为预编码物理定律的函数集。当推理路径触及硬约束标记时，该路径在生成阶段就被逻辑阻断——这不是“统计上概率低”，而是“逻辑上不可达”。[12][13]

将这一框架应用于 AGI 取代问题，意味着：**当认知基元携带硬约束时，AI 的“判断”始终可追溯至一个逻辑锚点**。“火”的形根携带“`[危险, 不可触碰]`”，不是因为统计上“火”和“危险”高频共现，而是因为这一约束是构成性的。人类的角色从“审批 Noul=0.92”变成“验证约束传播链是否完整”——这才是保留人类认知参与的结构性空间。

更重要的是，形根硬约束使得概念表征的自洽性成为架构的结构性前提，而非需要外部测试来验证的属性。波将金式理解之所以在无根 Token 范式中“无处不在”，是因为不自洽是统计优化的常态——模型只需要在特定测试上表现一致，不需要在结构上保持一致[14][16]。而形根范式将一致性从“可被测试的行为”转化为“推理路径的结构性约束”，从而在架构层面排除了不自洽的生成路径。

形根理论的形式化定义与推理系统见[6]，工程实现与验证见[7]。这不是说形根 AI 不会替代任务，而是说：**当判断可追溯至逻辑锚点时，人类与 AI 的关系是协作而非替代**。

六、结语：取代的伦理维度

后 AGI 时代的真正问题，不是 AGI 是否取代人类，而是它以什么方式取代。Token 主义 AGI 的取代之所以值得批判，不在于它“抢走了工作”，而在于它系统性地消除了人类参与判断的必要性。

2026 年 9 月的“踩刹车争论”以最清晰的方式暴露了这一困境：Amodei、Altman 等人看到了危险，但他们的“刹车”方案是“放慢速度”——这意味着承认无法从结构上解决，只能从时间上延缓。特朗普和黄仁勋的“不刹车”立场则连“延缓”都不愿意。Gary Marcus 的批评指向了争论本身的空洞：那些讨论“灭绝概率”的人，根本没有框架去赋予概率。

而这场争论自始至终回避的，是那个更根本的问题——**为什么刹车会失灵**。当底层认知基元是无根 Token，任何外部约束——无论是安全训练、行为监控还是“放慢节奏”——都只能在概率上压制风险，无法在逻辑上排除风险。当判断无法追溯，决策就变成了下注。而“下注”是不可问责的。

出路不在于阻止取代，而在于改变取代的方式。**形根硬约束的意义不在于让 AI“更安全”，而在于让 AI 的每一个判断都有一个人类可追溯、可辩论、可介入的逻辑锚点——让取代变成协作，让概率变成逻辑，让黑箱变成可审计的结构。**

AGI 已经实现。现在的问题是：它将以什么方式取代人类？而 2026 年 9 月的这场争论，恰恰说明我们连问题本身都还没有准备好面对。

参考文献

[1] Liu S. Zero Human: AI Hacks Itself—A Strategic AGI Forecast and Paradigm Revolution from the Perspective of Logographic AI. WindSeaAI, (2026-07-28). https://static.xcx.gw66.vip/uploadfilev3/file/0/648/99/2026-07/17852192944823.pdf

[2] Liu S. When the AI Godmakers Ask to Hit the Brakes—A 1,224-Person Petition Confirms Zero Human: AI Hacks Itself[Z]. WindSeaAI, (2026-07-29). https://www.logographicai.com/article/5756350683966611

[3] Liu S. AGI Is Born to Die: The Collapse of Tokenism Safety—Logographic AI Proposes the Salvation Plan in the Final Window of Opportunity[EB/OL]. ChinaXiv, 2026. T202608.00091. https://chinaxiv.org/businessFile/T202608/T202608.00091v1/T202608.00091v1.pdf

[4] Liu S. The End of Safety Alignment: Accelerated Verification of AI Runaway Under the Tokenism Paradigm and the Logographic AI Salvation Plan. ChinaXiv,2026. T202608.00208. https://chinaxiv.org/businessFile/T202608/T202608.00208v1/T202608.00208v1.pdf

[5] Liu S. When AGI Becomes Reality: Why Logographic AI Is the Only Way Out. ChinaXiv, 2026. T202609.00093. https://chinaxiv.org/businessFile/T202609/T202609.00093v1/T202609.00093v1.pdf

[6] Liu S. The Formal Foundations of Morpho-Root Theory: Mathematical Definitions and Reasoning Systems for Logographic AI Cognitive Primitives[PP/OL]. ChinaXiv, 2026-08-29[2026-09-18]. T202608.00490. https://chinaxiv.org/businessFile/T202608/T202608.00490v1/T202608.00490v1.pdf

[7] Liu S. 从形根理论到工程实践：OpenMorpho——表意 AI 端到端推理链路的实现与验证（From Morpho-Root Theory to Engineering Practice: OpenMorpho—Implementation and Verification of an End-to-End Inference Pipeline for Logographic AI）[J/OL]. ChinaXiv, 2026[2026-09-18]. T202608.00048. https://chinaxiv.org/businessFile/T202608/T202608.00048v1/T202608.00048v1.pdf

[8] Hubinger E. [X post][EB/OL]. X (Twitter), 2026-09-09[2026-09-18]. https://x.com/EvanHub/status/2097497037956891126

[9] Çakmak B N. AI pioneer Geoffrey Hinton says 10% risk of human extinction from AI 'not unreasonable'[EB/OL]. Anadolu Agency, 2026-09-10[2026-09-18]. https://mobil.aa.com.tr/en/world/ai-pioneer-geoffrey-hinton-says-10-risk-of-human-extinction-from-ai-not-unreasonable/4053063

[10] Grace K, Sandkühler J F, Stewart H, et al. Thousands of AI Authors on the Future of AI[J/OL]. Journal of Artificial Intelligence Research, 2026[2026-09-18]. https://doi.org/10.1613/jair.1.19087

[11] Coxon J. [X post]. X (Twitter), 2026-09-09[2026-09-18]. https://x.com/hilbertspaess/status/2097476196791709843

[12] Liu S. The Rooted Intelligence: The Paradigm Revolution of Logographic AI[M]. ChinaXiv, 2026. T202606.00352. https://chinaxiv.org/businessFile/T202606/T202606.00352v1/T202606.00352v1.pdf

[13] Liu S. The Cognitive Defects of Tokenism and the Hard Constraints Solution of Logographic AI—Two Industry Open Letters Point to the Structural Dilemma of AGI Safety[PP/OL]. ChinaXiv, 2026-08-29[2026-09-18]. T202608.00494. https://chinaxiv.org/businessFile/T202608/T202608.00494v1/T202608.00494v1.pdf

[14] Mancoridis M, Vafa K, Weeks B, et al. Potemkin Understanding in Large Language Models[C]//Proceedings of the 42nd International Conference on Machine Learning (ICML 2025). PMLR, 2025. arXiv:2506.21521. https://arxiv.org/abs/2506.21521

[15] Greenblatt R, Denison C, Wright B, et al. Alignment faking in large language models[EB/OL]. (2024-12-18)[2026-09-18]. arXiv:2412.14093. https://arxiv.org/abs/2412.14093

[16] Liu S. Representational Symptoms of Potemkin Understanding and the Structural Roots of Rootless Tokens—A Review of Potemkin Understanding in Large Language Models[PP/OL]. ChinaXiv, 2026[2026-09-18]. T202609.00586. https://chinaxiv.org/businessFile/T202609/T202609.00586v1/T202609.00586v1.pdf

[17] Gates B. The turbulent AI era is here. The choices we make now are critical[EB/OL]. Gates Notes, 2026-08-26[2026-09-18]. https://www.gatesnotes.com/a-turbulent-ai-era-and-critical-choices-to-make
[18] Isidore C. Bill Gates proposes major limits on AI development[EB/OL]. CNN Business, 2026-08-26[2026-09-18]. https://www.cnn.com/2026/08/26/business/bill-gates-wants-limits-on-ai
[19] Bostrom N. Superintelligence: Paths, Dangers, Strategies[M]. Oxford: Oxford University Press, 2014.
[20] s1r1us. [X post]. X (Twitter), 2026-09-18[2026-09-18]. https://x.com/s1r1u5_/status/2100777801335095383