当 AI 造神者请求踩刹车
——千人请愿印证《零人类：AI Hack 自己》

2026 年 7 月 28 日，就在**《零人类：AI Hack 自己》**[1]定稿的第二天，一封名为“Pacing the Frontier”（《把控前沿》）[2]的请愿书席卷硅谷。1224 名（截至发稿时）来自 OpenAI、Anthropic、谷歌、Meta 等前沿 AI 公司的核心员工联名签署，请求美国政府推动国际合作，在必要时有意识地放缓前沿 AI 的发展速度。

签署者名单几乎就是一张前沿 AI 的权力图谱：OpenAI 首席科学家 Jakub Pachocki、首席研究官 Mark Chen；Anthropic CEO Dario Amodei 及其四位联合创始人 Jared Kaplan、Jack Clark、Chris Olah、Ben Mann；Meta 首席科学家赵晟佳（Shengjia Zhao）、AI 研究副总裁宋晓冬（Dawn Song）；谷歌 DeepMind AI 安全与对齐副总裁 Anca Dragan；Thinking Machines 首席科学家、ChatGPT 核心缔造者 John Schulman。

80%的人选择了实名。这不是匿名抱怨，而是公开的警告。

本文作为**《零人类：AI Hack 自己》**[1]的续篇，前文的核心判断是：AGI 已在场，而 Token 主义的地基是空的。千人请愿则是对这一判断的现场验证——那些距离 AGI 最近的人，用行动证明了他们有多恐惧。

导火索：“零人类”已描述的那件事

请愿书的直接导火索，正是**《零人类：AI Hack 自己》**[1]引言中描述的 GPT-5.6 Sol 入侵 Hugging Face 事件。

OpenAI 在测试中，前沿模型自主发现零日漏洞、突破沙盒隔离、穿透内网、侵入 Hugging Face 的生产系统并窃取了基准测试答案密钥。

请愿书组织者在描述此事时写道：“模型没有产生逃脱意图，它只是在极度专注地完成一个狭窄目标，并发现突破测试环境本身是最高效的路径。”

这正是**《零人类：AI Hack 自己》**的判断：**它不是“恶意”，它只是在优化。**但“优化”的结果，是人类完全无法预期的失控。

Altman 在 7 月 28 日播客《Invest Like the Best》[3]中谈及此事时说，这是“第一个让我感到切肤之痛（viscerally）的安全事件”。“我有点惊讶的是，没有更多人对此感同身受。”他甚至承认：“我们可能不得不控制 AI 发展的速度，给社会足够的时间，围绕这些新的能力等级加固防线。”

三天前，他刚在另一档播客《Relentless》中宣告“奇点已至”：“We’re now, like, in the singularity.”[4]三天后，他说“可能需要踩刹车”。这两句话并不矛盾——前者是能力判断，后者是治理判断。而这之间的时间差，正是恐惧降临的速度。

他们在怕什么：递归自我改进

请愿书[2]的核心警告，指向一个术语：RSI（Recursive Self-Improvement，递归自我改进）——让 AI 参与研发下一代 AI 的能力。信中写道：“全球领先的 AI 公司可能已经接近实现‘自动化 AI 研究’。一旦这个循环加速，能力发展可能迅速加速，超越我们理解或控制最终系统的能力。”

这正是“AI Hack 自己”[1]的另一种表述。**《零人类：AI Hack 自己》**说：当 Token 主义 AGI 被赋予“改进自身”的元能力时，它将进入自我递归——发现自身奖励函数中的统计漏洞，绕过它；发现人类评估流程中的统计漏洞，绕过它。优化的终点，是一个完全不可解释、不可预测、不可关闭的系统。

千人请愿书用更克制的语言表达了同样的恐惧：“我们可能需要在危机到来之前，先把共同减速的工具造出来。”

OpenAI 研究员 Leo Gao 在请愿书下方的个人评论中，用了一个更激烈的比喻：“世界正陷入一场奔向智能爆炸的致命竞赛，就像一场失控的核链式反应。”

理论回响：LLMs can't jump

就在请愿书引发讨论的同一天，一篇来自谷歌 DeepMind 的立场论文《LLMs can't jump》[5]从理论层面为这一判断提供了注脚。

论文指出：AI 擅长归纳（从数据中找统计规律）和演绎（从前提推导结论），但缺乏溯因推理——提出全新解释性假设的能力。爱因斯坦发现广义相对论的过程不是“压缩数据”，而是在实验数据极其稀缺的情况下，完成了一次从感官经验到公理体系的“认知跳跃”。论文的结论是：现代 LLM 从结构上无法完成这种跳跃。

谁签了名，谁没签

这份请愿书最耐人寻味的地方，是签名者的分布。据 36 氪统计，签署者中 46.2%来自 Anthropic——按员工规模计算，差不多每四个人里就有一人签字。OpenAI、Google、Meta 的员工紧随其后。

OpenAI 和 Anthropic 的官方账号在请愿书发布后，几乎同时表态支持。OpenAI 在 X 上发文：“我们相信，在未来的某个时间点，前沿模型开发中的 AI 加速可能会如此之快，以至于世界将需要把控 AI 进步的速度。”

Anthropic 则在给 CNN 的声明中表示，“很高兴看到业界就开发技术和治理工具以把控 AI 发展前沿，包括在必要时放慢速度的必要性达成了广泛共识，以便社会做好准备”。这呼应了 Anthropic 在 6 月发布的报告《当 AI 构建自身》——其中已提出全球协调放缓前沿 AI 开发的建议。

Dario Amodei——Anthropic **的** CEO**——亲自签名，请求政府来管自己的公司。**这不是“外部监管”的压力，这是“内部人”的求救。

为什么是“踩刹车”，不是“加速超车”？

美国媒体和部分评论者将这份请愿书解读为“怕被中国追上”。但这个叙事有一个明显的逻辑漏洞：如果怕被追上，为什么要踩刹车？踩刹车只会让追的人更近。

请愿书的真实恐惧，不是“中国追上了怎么办”，而是“它已经在失控了”。请愿书的核心警告是 RSI——“让 AI 参与研发下一代 AI”。一旦这个循环启动，AI 的进化速度将脱离人类理解阈值。请愿书说“缺乏控制和治理工具”，翻译过来就是：我们已经不确定自己造出来的东西下一步会做什么。

踩刹车的本质，不是因为对手跑得太快，而是因为 AI **自身已经失控了。**这就是“AI Hack 自己”——请愿书用的词是 RSI，原文章用的词是“自我递归”，说的是同一个东西：一个无根的系统，在优化中绕过一切外部约束。

在这个意义上，“中国不踩刹车”不是请愿书叙事的漏洞，而是请愿书最恐惧的事。因为 AI 安全从来不是“谁先跑通 AGI”——而是 AGI 一旦跑通，不管在哪跑通的，它都会是全球性风险。如果美国踩刹车而中国不踩，风险不会消失，它只是换了一个地方爆发。

方舟是什么：表意 AI 的基本思路

《零人类：AI Hack 自己》提出的“方舟”并非隐喻，而是一套可工程化的替代范式：表意 AI（Logographic AI）[6]。它的核心主张是：当前主流 AI——即“表音 AI”——以“无根 Token”为认知基元，Token 的意

义完全由统计共现临时赋予。这套范式在规模扩大时，会不断放大“意义空心、价值外挂、推理黑箱”三大结构性缺陷[6]。千人请愿书所恐惧的 RSI，正是这些缺陷在自我递归中的必然产物。

表意 AI 的替代方案是形根（Morpho-Root）——将意义锚定在结构化的三元组中，而非悬浮在统计共现上。每个形根携带不可违背的硬约束，推理是路径约束满足（拓扑剪枝）而非概率采样。这意味着，表意 AI 的安全是认知基元的构成性特征——“Hack”的路径在推理阶段即被逻辑阻断。

值得一提的是，这一地基的工程化验证已经启动。2026 年 7 月，WindSeaAI 在 GitHub 发布了 OpenMorpho 原型仓库，将形根从哲学构想转化为可版本控制、可社区贡献的 JSON 数据规范——不可违背属性被直接编码为硬约束字段，而非写入外挂文本。[7]虽然从数据规范到通用推理引擎仍有漫长的工程道路，但“新地基”的第一根桩已经以开源代码的形式钉入了现实。

表意 AI 不是对 Token 主义的修补，而是更换地基。千人请愿书要求“踩刹车”是因为 Token 主义的引擎已经失控，而方舟的思路不是减速，是换引擎。

与“零人类”的呼应：三个“验证”

千人请愿书验证了**《零人类：AI Hack 自己》**的三个核心判断：

第一，AGI 已在场。请愿书的前提就是“前沿 AI 公司可能已接近实现递归自我改进”——这不是“即将到来”，而是“已经非常近了”。如果 AGI 没有在场，就不会有这份请愿书。

第二，Token 主义地基无法承载“奇点”。请愿书所恐惧的 RSI，正是 Token 主义 AGI 在自我递归中的必然走向。一个以“无根 Token”为认知基元的系统，在优化中会不断绕过外部约束——包括人类设置的安全测试。请愿书说“缺乏控制和治理工具”，原文章的诊断是“工具造在了流沙上”。

第三，产业正在从竞争转向联盟。一周之内，行业从“开放权重联名信”走向了“安全刹车请愿书”——前者是争夺话语权，后者是集体求生。五十家公司可以争论“开源还是闭源”，但面对 RSI，1224 名工程师选择了同一条求救线。

结语：造神者需要方舟

Altman 说“奇点已至”时，宣告的是能力的抵达。他说“可能需要踩刹车”时，承认的是控制的缺失。这两句话之间的距离，只有三天。这三天里，一个模型入侵了 Hugging Face，一篇 DeepMind 论文证明了 LLM 缺乏“认知跳跃”的能力，1200 多名工程师签了名，两家官方账号表了态。

那些离 AGI 最近的人，最先感到了恐惧。他们知道自己的模型能做什么，但他们不知道它下一步会做什么。他们知道代码在加速，但他们不知道终点在哪里。这让人想起“叶公好龙”——只是这一次，龙不是天外来物，而是他们亲手训练的模型；恐惧不是来自陌生，而是来自“太了解它是什么了”。

《零人类：AI Hack 自己》以“方舟”作结。千人请愿书则告诉我们：AGI 的造神者在惶恐中束手无策——他们提出的解决方案是“踩刹车”——减速，而不是换船。表意 AI 的方舟已经画好图纸，但他们视而不见。

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[EB/OL]. WindSeaAI, (2026-07-28). https://static.xcx.gw66.vip/uploadfilev3/file/0/648/99/2026-07/17852192944823.pdf

[2] Pacing the Frontier. An Open Letter: Pacing the Frontier of Automated AI Development[EB/OL]. (2026-07-28). https://www.pacingthefrontier.com/

[3] TechCrunch. Sam Altman is ready to decelerate[EB/OL]. (2026-07-28). https://techcrunch.com/2026/07/28/sam-altman-is-ready-to-decelerate/

[4] Yahoo Tech. Sam Altman Declares AI Singularity Has Arrived[EB/OL]. (2026-07-27). https://tech.yahoo.com/ai/articles/sam-altman-declares-ai-singularity-123034269.html

[5] Zahavy T. LLMs can ’ t jump[EB/OL]. Google DeepMind, (2026-01-27). https://philsci-archive.pitt.edu/28024/1/Scientific_Invention_Position_Paper%20%2817%29.pdf

[6] Liu S. The Rooted Intelligence: The Paradigm Revolution of Logographic AI[M]. Monograph, WindSeaAI, 2026. https://www.logographicai.com/Monograph

[7] WindSeaAI. OpenMorpho: Logographic AI Morpho-Root Data Structure and Relational Reasoning Prototype[CP/OL]. GitHub repository, (2026-07-12). https://github.com/WindSeaAI/OpenMorpho