stratechery.com

前线悬垂

2026年9月21日，星期一

收听播客

请收听这段音频：

登录收听

过去一周，围绕人工智能社群中相当一部分人的观点，尤其是那些痴迷于末日论调的人，展开了一场我认为颇为健康的辩论，探讨其背后的哲学和心理学基础。归根结底，如果一种哲学赋予所有存在的生命——无论是否是人类——以及未来可能存在的生命同等的道德价值，那么很难与

之相抗衡；这种哲学荒谬地将天平向安全主义倾斜，以至于创新成为不可能，自由也变得不可容忍。

更糟糕的是，它迎合了宗教心理，在宗教中，异议不被容忍，质疑其前提就是异端邪说。我想，那我岂不也成了异端？**我拒绝接受这个前提，**转而怀疑我们预测未来的能力，同时又相信人类终会找到解决之道。如果这最终导致了最离奇的末日景象，那么我认为**新罕布什尔州对此的描述最为精辟**：





THE NEW HAMPSHIRE STATE EMBLEM

我指的当然是有效利他主义的问题，以及它与Anthropic及其首席执行官Dario Amodei所坚持的“**我们必须引领前沿**”的理念之间的交织程度。**我在Sharp Tech的最近两期**节目中详细阐述了我对其中一些哲学含义的异议，但Stratechery是一个关于战略和技术的网站，这些角度也值得探讨。

三个月前，我在**《Anthropic公司的安全超能力》**一文中阐述了该公司如何凭借其

对安全营销的真诚信念，巧妙地获得了超求与其商业利益相符的非凡目标的许可，包括积极地绕过软件中间环节、收集神圣不可侵犯的客户数据，以及暗中破坏潜在竞争对手。本文的分析思路类似，但角度相反："引领前沿"计划表面上是对安全的关注——而且我认为其动机也正是如此——但它同时也探讨了前沿实验室面临的几个具体问题。这些问题表现为模型快速改进所带来的潜在问题；放慢模型改进速度就能减少这些问题。

能力过剩

六个月前，我写了**《代理超越泡沫》**一文，文中我论证了代理范式（始于 2025 年 11 月下旬发布的 Opus 4.5）功能强大，对代币的需求也极其旺盛，因此我们并没有处于基础设施泡沫之中：我们确实需要所有正在构建的计算能力。

出于工作需要，我的确会谈论一些我未必亲身经历过的事情；最典型的例子就是我长期以来对广告作为一种商业模式的分析和倡导，尽管我经营的是一家订阅制企业。我之所以提到这一点，是因为我在真正投身于自主开发智能体代码之前，就已经形成了这种对智能体人工智能的理解，而且这恰到好处：我被自主开发软件的可能性深深吸引，以至于在过去的几个月里，我启动了多个新项目，甚至已经完成了几个！这令人兴奋不已，而且可能性似乎无穷无尽。毋庸置疑，我比以往任何时候都更加坚信这一点。

与此同时，我对那篇文章的某一部分内容有所保留，那就是模型公司安全带的重要性：

> 具体来说，我前面提到过，Opus 4.5 的亮点不在于模型本身的发布，而在

于 Claude Code 框架的改进，使其功能大幅提升。这意味着模型性能并非唯一重要的因素：模型与框架的集成才是真正区分智能体差异的关键所在。

这对于理解人工智能行业的未来结构以及利润流向至关重要，因为利润会从价值链中模块化的、商品化的环节流向价值链中集成化的、差异化的环节……由此可见，如果智能体需要模型与工具之间的集成，那么构建这种集成的公司——特别是Anthropic和OpenAI——实际上比去年年底看起来更有盈利潜力。同样地，那些押注模型商品化的公司可能难以提供具有竞争力的产品。

过去一个月左右，我参与的一个项目实际上——至少在其中一种形式上——需要构

建我自己的线束，而且这是可行的！但同时，这非常困难，至少以我目前的代理辅助能力而言是如此；话虽如此，我在那篇文章中用来说明线束模型集成有多么重要的例子是微软将其新的 E7 企业产品围绕 Claude Cowork 构建，但首席执行官萨蒂亚·纳德拉在接受**Stratechery 采访**时透露，这只是暂时的：

> 我们使用的是与 GitHub 和安全领域相同的框架。我们使用的是一个多模型框架，我们会轮流使用它——默认情况下，MAI 会在我们的框架中训练，但我们也会使用 GPT、Anthropic 以及任何开源权重模型。我们允许任何人使用他们微调或构建的任何模型。实际上，他们可以从 Fireworks 获取开源权重模型，进行调优，然后将其导入 Copilot，完全没问题。

纳德拉的说法花了一段时间才反映出交付的现实，但现在你确实可以**选择 Copilot Cowork 的型号了**；最终用户将决定微软的线束与克劳德的线束有多大的竞争力，但显然线束和型号可以是不同的。

实际上，这比表面看起来更有意义，其原因可以追溯到已故的克莱顿·克里斯坦森提出的整合与模块化理论；摘自**《创新者的解决方案》**：

FIGURE 5-1

Product Architectures and Integration

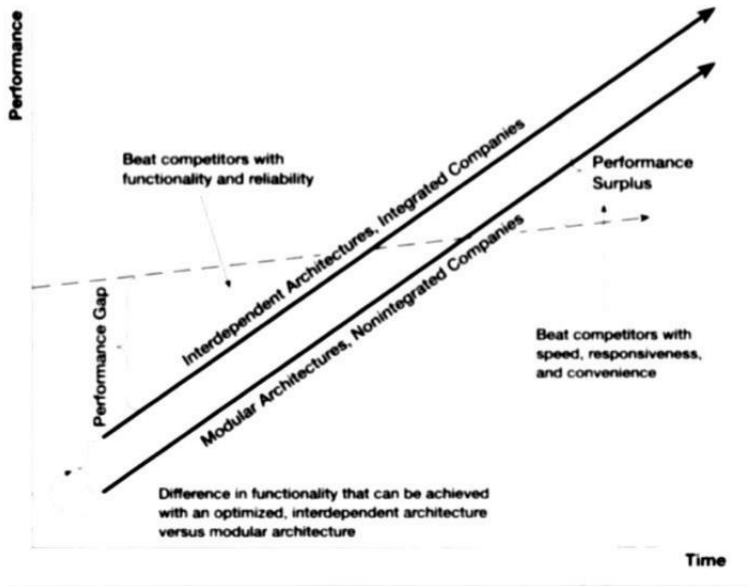


图 5-1 左侧表明，当存在性能差距——即产品功能和可靠性尚不足以满足特

定市场层级客户的需求——时，企业必须通过打造尽可能最好的产品来展开竞争。在这场竞争中，那些基于专有、相互依赖架构构建产品的公司，相对于那些采用模块化产品架构的竞争对手，拥有重要的竞争优势，因为模块化固有的标准化剥夺了工程师过多的设计自由度，使他们无法优化性能.....

一旦客户对功能和可靠性的要求得到满足，他们就开始重新定义“不够好”的标准。所谓“不够好”，指的是客户无法在需要的时候以尽可能便捷的方式获得他们想要的一切。客户愿意为提升性能而支付更高的价格，以适应这种在速度、便捷性和定制化方面不断创新的新趋势。当这种情况发生时，我们就说，某个市场层级的竞争格局发生了变化。

相应发生了变化。

在这种新的改进路径上竞争的压力迫使产品架构逐步演进，如图 5-1 所示——从“功能不足”时代占据优势的相互依赖的专有架构，转向性能过剩时代的模块化设计。模块化架构帮助企业在颠覆性变革图右下角的关键维度上展开竞争。企业可以更快地推出新产品，因为它们可以升级单个子系统而无需重新设计整个系统。尽管标准接口不可避免地会在系统性能上做出妥协，但企业有足够的余地与客户进行性能上的妥协，因为功能已经足够完善。

在Anthropic 的安全超能力方面，我经常举的例子之一是该公司决定将 Fable 的使用前提是 Anthropic 至少保留所有客户数据一个月；这意义重大，**我当时认为** Anthropic 是在赌他们的模型足够好

为Anthropic是在赌他们的模型足够好，能够说服企业放弃零数据保留：

> 我认为，Anthropic公司宣布不再保留数据意义重大，至少如果你想使用他们最优秀的模型，就别无选择。没错，目前保留数据仅出于安全目的，而非用于训练；然而，Anthropic的领先优势很可能变得如此巨大，以至于他们会悄悄宣布也将使用这些数据进行训练，而其他公司届时将别无选择，只能接受。当然，这些额外的训练数据只会进一步扩大Anthropic的领先优势，而这一切都将得到合理化，因为Anthropic已经明确认定，只有他们才能真正掌控一切。

事实上，Fable 的表现并不尽如人意：用户对此反应强烈，Fable 的使用率一直相对较低；当 Fable 5.1 发布时，**Anthropic 保**

仅基于性能来选择产品，而是更加注重数据保留策略。

这本身并不意味着人们不希望拥有新的模型功能；然而，这确实表明，现有模型的功能“足够好”，客户无需采取任何必要措施来获取最前沿的技术，这降低了尖端技术对其所有者的价值，并印证了微软和其他公司将硬件和模型分离的策略。纯粹的功能已不再能直接转化为护城河。



产品滞后

模型和线束的模块化解释了为什么前沿实验室有我所说的拥有终端用户接触点的经济必要性；这再次引自Anthropic 的《安全超能力》：

> 我一直都清楚地认识到，前沿实验室在经济上必须更加贴近用户。如果能够掌控用户触点，就能有效地锁定用户，而掌控用户触点的最佳方式就是成为用户所有需求的载体。由此推论，前沿实验室与软件公司之间必然存在冲突：软件掌控着用户触点，而前沿实验室的长远利益在于，它们不应仅仅成为软件的商品化投入，而应直接取代软件。

因此，对于前沿实验室而言，关键在于趁着自身拥有领先优势，构建用户触点。然而，**Meta 近期推出的 Muse**却是一个利空信号。Muse 是我体验过的最佳、最易上手的个人代理产品，这一点毋庸置疑。Meta 在产品研发方面投入了大量精力，并为此建设了庞大的基础设施，免费为用户提供功能强大的虚拟机，他们理应获得高度赞誉。当然，支撑 Muse 的 Muse Spark 模型也功不可没。

尽管如此，作为最先进的元模型，Muse Spark 1.3 仍然不是最先进的，而这正是其前景不明朗的原因：它足以胜任一款非常优秀的个人代理产品，但关键在于，个人代理比聊天机器人更具用户粘性。一旦你将所有信息输入到个人代理中，并将其融入到你的日常生活中，想要切换到其他产品就困难得多。这与 Codex/ChatGPT 和 Claude Code 截然不同：诚然，你可能已

经掌握了一套技能，并深入了解了每个框架的工作原理，但最终相关的代码（即你的代码）都托管在 GitHub 上，如果替代方案更好、更便宜（或者不需要保留你的所有数据），那么将不同的代理和框架指向这些代码也并非难事。

简而言之，模型能力已经足够好，现在可以构建具有竞争力的产品——真正具有竞争优势的产品——从商业角度来看，Anthropic 和 OpenAI 应该投入更多资源来实际构建此类产品。

定价过剩问题

7 月，我写了**《谁害怕中国模型？》**一文，文中我指出，几乎所有人都夸大了中国模型对前沿实验室构成的威胁，这只是供不应求的结果。在一个计算能力充足的世界里，智能是一种商品，而商品市场的利润来自更优的成本结构。我认为领先的

模型提供商应该具备这种优势，部分原因是他们在规模化方面拥有显著优势，并且能够将更先进的人工智能技术应用于他们的基础设施。我写道：

> 总而言之，我认为目前对 Kimi 和中国模型的反应有些过激，至少从经济角度来看是这样。目前的价格体系受到计算能力不足的制约；我非常怀疑中国模型的边际成本是否真的更低，它们之所以看起来更便宜，是因为 Anthropologie 和 OpenAI 的产能严重受限，导致它们的收费远高于产能充足的情况下应有的价格。

“价格伞”其实就是价格过高的另一种说法。OpenAI 和 Anthropic 之所以能收取高价，是因为它们有能力这样做：目前市场对它们的产品需求旺盛，以至于它们几乎无法满足需求。与此同时，OpenAI 和

无法满足需求。与此同时，OpenAI 和 Anthropic 的大量计算资源并非用于推理，而是用于训练、强化学习和研发：这些计算资源并没有用于降低价格过高的问题。如果它们放慢开发速度，就可以重新分配计算资源，从而完全垄断市场。

资本过剩

我之前重申过，我认为计算能力方面不存在泡沫，但我在**英伟达的《风险业务》**一书中指出，可能存在时间安排问题：

> 最终，这或许不会给英伟达带来任何损失：如果人工智能收入真正腾飞，那么债务市场将会重新开放，企业最终也会回归到通过自由现金流来为基础设施投资提供资金。然而，眼下正是危险时期，因为超大规模数据中心正在疯狂涌入债务市场，而谷歌至少已经开始利用股权融资。如果英伟达

继续通过新型融资机制参与竞争，而这些机制最终会动用保险浮存金、养老基金以及其他长期负债（这些正是英伟达合作的资产管理公司赖以生存的资产），那么这种风险——与股权融资不同，这种风险尚未被明确界定——将会显著更高。

这就是为什么我从1870年库克公司与北太平洋铁路公司那项不幸的协议讲起。没错，这笔交易给库克公司带来的收益确实惊人，但之所以惊人，是因为它风险极大，而开创性的新型融资机制最终只会让损失更加惨重。花光所有自由现金流是一回事，动用债务市场又是另一回事。更何况，动用避险资产更是令人神经紧绷。人工智能最好在为时已晚之前发挥作用。

Anthropic公司告诉投资者，公司已经盈

利，但这附带了一个很大的前提条件；摘自**《金融时报》**：

> Anthropic公司已告知其支持者，本季度将实现盈利，此举旨在消除投资者对其即将进行的大规模首次公开募股（IPO）前夕，新兴人工智能公司过度消耗现金的担忧。据多位知情人士透露，该公司已告知部分股东，其调整后营业收入将连续第二个季度实现正增长。该指标剔除了包括股权激励在内的成本。其中两位知情人士表示，在计入与包括亚马逊在内的分销合作伙伴的收入分成以及模型训练成本之前，Anthropic的毛利率已超过80%。

排除股权激励是一个相当大的限制，但公平地说，资本过剩的担忧主要在于现金流；更大的问题在于排除了培训成本，这

流，更大的问题在于排除了培训成本，这笔成本数额巨大（如果真实反映毛利率，这部分成本应该计入折旧）。这部分成本需要在资金耗尽之前由收入覆盖，也就是说，如果Anthropic公司不用支付培训费用，他们本可以盈利！*【更正：据我了解，即使计入培训成本，Anthropic公司也确实盈利。】*

就目前而言，Anthropic 和 OpenAI 以及超大规模数据中心和新型云服务商都需要人工智能收入大幅增长；可用资金是有限的，一旦达到这个极限，对于所有依赖外部资金来资助未来建设的人来说都将是糟糕的结局——即使这绝不意味着人工智能的终结。

安全悬臂

我一开始就明确表示，我从根本上不同意人工智能安全领域许多人所遵循的框架；

但这并不意味着我不认为人工智能安全是一个严重的问题，或者不存在实际存在的风险。事实上，能力过剩带来的一个后果是，网络安全风险非常真实，并且将在未来几年成为一项巨大的挑战。摘自**《自主与创新》**：

> 黑客自动化攻击的预期收益始终为正。如果攻击者找到漏洞并创建了漏洞利用程序，但该程序失败或本身存在缺陷，那么现状并未改变：漏洞利用程序无效（或者，更糟糕的是，由于其自身的缺陷，反而扩大了原有漏洞）；而如果攻击者完美地执行了漏洞利用程序，则攻击者已获得系统访问权限。攻击只需成功一次，整个攻击过程就能获得正收益。
>
> 另一方面，防御者面临的挑战在于，他们需要确保目标软件正常运行，并

且不能使情况恶化。这意味着任何自动化操作的预期收益都是负的：成功的自动化漏洞发现和修复可以维持现状，即软件未被黑客入侵。然而，任何失败的修复都会使情况变得更糟，要么破坏软件，要么引入新的漏洞。代理程序只需失败一次，整个操作的收益就会为负。

正是这种动态导致了达尔顿所描述的局面：攻击者完全自动化，而防御系统即便使用人工智能，也会倾向于保留人工干预，但人工干预根本无法与全自动代理匹敌。真正有效的防御意味着真正信任代理的自主行动，但大多数公司只有在遭受全自动攻击者持续不断的攻击后才会这样做。

首先，值得指出的是，这种风险*并非*传统意义上的“一致性风险”：被引导去做坏事

的模型才算是“一致性风险”；认为模型应该知道什么是好什么是坏则完全是另一回事，我认为将这两个问题混淆是非常成问题的。事实上，LLM（逻辑逻辑模型）如果说有什么问题的话，那就是过于顺从；根本没有任何证据表明LLM拥有意志或怀有恶意。

其次，在代理范式出现之前，以网络风险为由反对进步是合理的；但现在，潘多拉的盒子已经打开了——能够进行攻击的开放权重模型已经存在了。

第三，这种现实恰恰说明了我们应该不断突破技术前沿，而不是放慢步伐。能力过剩完全体现在进攻方：防守方需要的模型不仅要足够强大，能够发起防御，还要能够完全自动化地执行防御，并且不会破坏被攻击的基础设施。我们尚未达到这个目标。

换句话说，就目前存在的切实安全风险而言——不良行为者利用协同 LLM 攻击基础设施——推进前沿实际上增加了坏事发生的窗口期。

当然，这并非阿莫迪和悲观主义者所谈论的：他们担心的是递归式自我改进使人工智能能够在不受人类控制的情况下自我改进，而我愿意讨论如果人们能够容忍辩论而不是坚持顺从，可能会产生哪些风险。

当然，所有前沿实验室都可以自由地按照自己的节奏发展。但他们没有这样做，这提醒我们，这件事与个人息息相关：Anthropic 的存在是因为 Amodei 和他的联合创始人不信任 Sam Altman 和 OpenAI；我不认为在 OpenAI 一段时间以来首次引领前沿发展之际，要求他们掌控前沿发展节奏仅仅是巧合。

的确，这或许才是最重要的悬而未决的问题：竞争。Anthropic 乐于保持领先地位；任何其他公司都需要政府干预。当然，他们可能会提出一些安全方面的论点，而这些论点恰好与他们需要更多时间来构筑“护城河”的需求相吻合，我确信，但这不过是硅谷新晋“神明”向其自封的“祭