八一八Anthropic和有效利他主义（EA）3.Anthropic核心圈

原创 面包人小模型 一个人读书写字 2026年11月1日 08:15 上海

终于写到anthropic 核心圈了，就是为了这碟醋，我才包的这顿饺子！感谢anthropic最后公开的S-1几乎没有公布什么八卦相关的信息，让我还能接着八一八🤣

01 PART 生来就含着金汤匙的Anthropic

EARLY CAPITAL

最近可以看到英语媒体各种叙事、宣传，配合anthropic的模型全线升级到5.5，非常标准的anthropic式叙事构建。

摘录一段被不少AI自媒体洗稿过的**Kevin Roose于9月29日在** *The Atlantic* **发表的《Inside the Biggest Feud in Artificial Intelligence》的翻译。（前文是anthropic创始人要和openai分家的部分）**

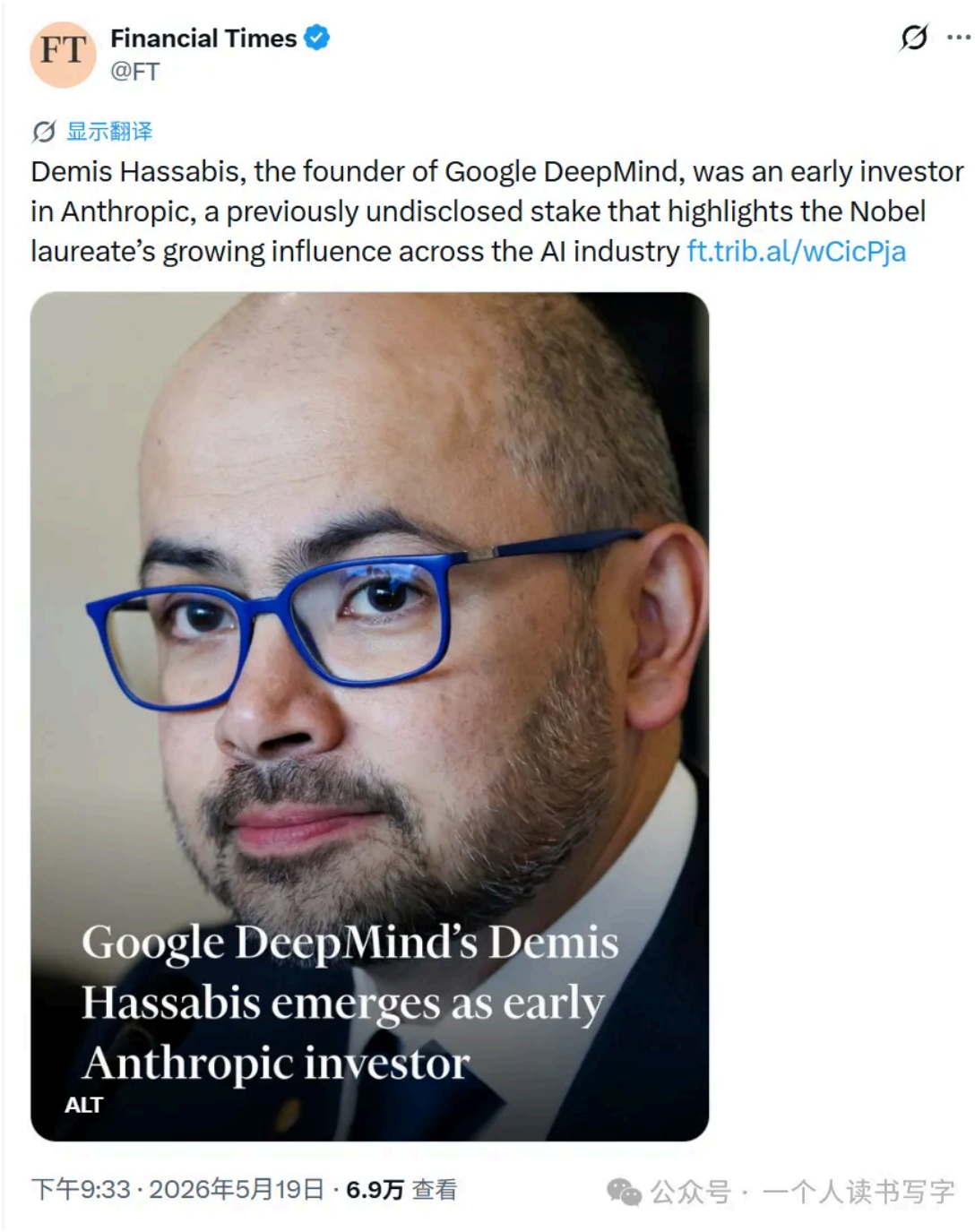
> *会议的聊天窗口瞬间被各种 emoji、感叹号和震惊的消息刷屏。与此同时，Amodei 尽量用一种正面的方式来包装这条消息。他说，OpenAI 仍将继续履行自己的使命——创造造福全人类的 AGI；至于为什么这七名员工决定离开，他只给出了极其含糊的解释。*
> *这时，Jack Clark 在聊天窗口里发出了最后一条消息：*
> *“我爱你们所有人。”*
> *然后，突然之间，他们的屏幕黑了。*
> *OpenAI 把他们踢出了会议，撤销了他们的电子邮件和 Slack 访问权限，并远程把他们的笔记本电脑直接“变成了砖”——使其无法继续使用。*
> *几小时后，这群刚刚失业的联合创始人自己开了一场视频会议——这也是 BATNA 的第一次正式会议。*
> *摆在他们面前的事情很多，同样也有一大堆尚未解决、令人担忧的问题。*
> *七个人怎么可能和 OpenAI、Google 竞争？*
> *他们要去哪里弄 GPU、数据和钱？*
> *那些无聊的公司事务又由谁来处理——比如提交法律文件、给所有人重新买电脑？*
> *他们的新公司要叫什么名字？*
> *最后的答案是“Anthropic”*

全文节选自**Kevin Roose的新书《***The AGI Chronicles: The Inside Story of the Race to Create an Artificial Superintelligence***》。这位大文豪什么成分，留到传媒篇再介绍~**

那么事实是怎么样的呢？我们来看看anthropic的融资阵容。S-1只有轮次但是没穿透。

天使轮：

- demis（FT于26年5月19日报道）



- Greg Jensen

2026年9月11日的Odd Lots节目 Why Bridgewater's CIO Says AI's Human Extinction Risk Is Real 嘉宾是桥水的Greg Jensen，他声称自己个人开出了 Anthropic 的第一张支票，Anthropic 第一周 payroll 就靠他的支票支付。



Why Bridgewater's CIO Says AI's Human Extinction Risk Is Real | Odd Lots

I started to get to know Sam and the other people there, which got me to know scientists like Dario and eventually was literally the first check to Anthropic. They made payroll the first week from my personal check to them.

接着是A轮：

Jaan Tallinn / Dustin Moskovitz / James McClave/CERR / Eric Schmidt

- **Jaan Tallinn**

Open phil有两大金主，一个是Good Ventures的Dustin Moskovitz （美国篇介绍过），另一个就是**Jaan Tallinn**了，A轮是Jaan Tallinn领投的。



(Jaan Tallinn on How Close We Are to Losing Control of AI –The Generalist)

目前AI safety领域的两大金主：一个是open phil（现在叫Coefficient Giving）一个是**Survival and Flourishing Fund (SFF)，SFF的资金来源几乎全部都是**Jaan Tallinn

SFF资助了非常多的人，美国篇里提到过一个Zvi，SFF给他的资助是：2024 Initiative Committee / $200,000 unconditional gift.

Programs FAQ

SURVIVAL & FLOURISHING FUND

Contact

Round	Source	Organization	Amount	Receiving Charity	Purpose
Initiative Committee 2024	Jaan Tallinn	Our World in Data	$300,000	Global Change Data Lab	General support of Our World in Data
Initiative Committee 2024	Jaan Tallinn	International Dialogues on AI Safety (IDAIS)	$500,000	Safe Artificial Intelligence Forum Institute	General support of International Dialogues on AI Safety (IDAIS)
Initiative Committee 2024	Jaan Tallinn	Longview Philanthropy	$700,000	Longview Inc. Ltd	General support
Initiative Committee 2024	Jaan Tallinn	Epoch AI	$600,000	Rethink Priorities	General support
Initiative Committee 2024	Jaan Tallinn	Center for AI Safety	$300,000	Center for AI Safety, Inc.	To support research and analysis of relevance to the Bureau of Industry and Security (BIS).
Initiative Committee 2024	Jaan Tallinn	Jacob Lagerros	$100,000	Center for Applied Rationality	General support of flexHEG prototyping
Initiative Committee 2024	Jaan Tallinn	AI Policy Institute	$300,000	The Hack Foundation	General support of the AI Policy Institute
Initiative Committee 2024	Jaan Tallinn	HitRecord	$700,000	HitRecord AI Safety Project LLC	General support
Initiative Committee 2024	Jaan Tallinn	Institute for Security and Technology	$400,000	Institute for Security and Technology	General support
Initiative Committee 2024	Jaan Tallinn	Zvi Mowshowitz	$200,000	Zvi Mowshowitz	Unconditional gift
Initiative Committee 2024	Jaan Tallinn	Topos	$300,000	Topos Institute	CatColab development
Initiative Committee 2024	Jaan Tallinn	Collective Intelligence Project	$500,000	Collective Intelligence Project Inc.	General support
Initiative Committee 2024	Jaan Tallinn	Haydn Belfield	$16,000	Haydn Belfield	Scholarship top-up funding to support scholarship on international governance of technologies with catastrophic potential, including AI
Initiative Committee 2024	Jaan Tallinn	Lightcone Infrastructure	$200,000	Center for Applied Rationality	General support of flexHEG prototyping
Initiative Committee 2024	Jaan Tallinn	Lightcone Infrastructure	$1,720,000	Center for Applied Rationality	General support of Lightcone Infrastructure
Initiative Committee 2024	Jaan Tallinn	Signal Foundation	$1,000,000	Signal Technology Foundation	General support
Initiative Committee 2024	Jaan Tallinn	Saturn Data	$500,000	Saturn Data Inc.	To support work on prototyping flexible hardware-enabled [chip] governors, or "FlexHEGs"
Initiative Committee 2024	Jaan Tallinn	MILA	$4,000,000	MILA	General support of Safe AI for Humanity, a portion of which may be allocated for related overhead costs.
Initiative Committee 2024	Jaan Tallinn	Center for AI Safety	$1,000,000	Center for AI Safety, Inc.	General support

放个截图，对SFF的资助范围可以有一个直观的感受（随便一截，2024的完整名单很长，是截图的几倍长）

SFF和Open phil的区别：（来自官方定义）

SFF=Jaan Tallinn个人主导的，高方差、多人下注。找一批 **Recommenders**，每轮换人，然后让他们分别评项目。挑选机制**倾向于资助至少有一个 recommender 非常兴奋的项目，而不是所有人都觉得还**

不错的项目。

Open phil=模仿基金行业，由专业PM帮金主进行慈善资本配置。现在有 **150+ staff，按 cause area 分专业团队，研究、调查、拨款、事后评估。**

那么SFF资助的高方差是什么样的呢？举一个典型的例子。

今年9月30日，又到了SFF给各种独立研究者/kol打钱名单公开的日子，其中显示John Wentworth & David Lorell 拿了$519k。这可不是小数目，lesswrong用户就开始查John Wentworth 的成分了。

SFF-2026	Jaan Tallinn & Dustin Moskovitz	John Wentworth & David Lorell	$519,000	Lightcone Infrastructure, Inc.	General support of research led by John Wentworth & David Lorell

引发争议的是两篇John Wentworth 26年在lesswrong发的帖子。内容不适合展开（翻译成中文我怕是过不了审核……）

第一篇是1 月 2 日的：《The Weirdness of Dating/Mating: Deep Nonconsent Preference》
这个帖子炸裂在哪里呢？论点不是单纯“有些女性有 CNC kink”，而是把这个推广成一个**一般性的女性 mating model 。**

第二篇是5 月 28 日：《Some Dating Stories》 这篇是超详细的个人（）生活记录，然后又重复了上一篇帖子的内容。

Lesswrong**用户彻底怒了**，质疑John Wentworth 凭什么拿钱?

我查了下成分：他是 rationalist / LessWrong / AI alignment 圈很老的一个人，主要做 Natural Abstraction / agent foundations。以前长期拿 EA Funds 的 LTFF 钱，2019–2023 年 LTFF 累计给过大约 $445k；2025 年 SFF 又给过他 $258k。

这个舆论虽然发酵的挺严重，但SFF的处理结果是什么呢？$519k里面的$340k作为 Speculation Grant 已经提前发掉了。理论上尚未支付的是大约 $179k，SFF借了个台阶说剩下的不发了。

2026 年另一个有趣的变化：Dustin Moskovitz 本来就已经是 Coefficient/Good Ventures 体系的核心金主，又开始给 SFF 大额出资。2026年SFF推荐总额约 $64.6m；其中 Jaan Tallinn：约 $34.51m、Dustin Moskovitz：约 $30.09m。

- **Dustin Moskovitz**：美国篇介绍过了，open phil的重要资金来源。

- **James McClave**：他是Jane Street的Management Committee Member。不是普通的 Quant Trader，而是公司治理层面的职位。Jane Street的人都比较神秘，公开信息里可以确认的有他与 SBF（Sam Bankman-Fried）曾在 Jane Street 共事（2014~2017）。

Dario Amodei 在《The Adolescence of Technology》的致谢中，特别感谢了 Jim McClave 和 Nick Beckstead 等人对文章草稿提出意见。

- **CERR：　英国篇介绍过了，瑞士EA，FTX来源的钱**

- **Eric Schmidt**：Alphabet公司前执行主席，谷歌公司前CEO。这个没什么好介绍的，重点是Cami Clark——Dario的老婆。和Schimidt有过三年的恋情。（具体就不展开了，the information和wsj都报道过）

2021 年 2 月，本来Cami 提出要与 Schmidt 共同成立一个叫 Mother of AGI Fund 的投资基金，既管理 Schmidt 对 Anthropic 的投资，又投资其他 AI 公司。

结果 Daniela 和其他联合创始人因为 Cami 与 Dario 的私人关系，反对这个计划。基金最后没成立，Schmidt 改为直接投资 Anthropic。



keachhagey ✓
@keachhagey

显示翻译

The internet has been so scrubbed of references to Dario Amodei's wife, Cami Clark, that if you Google her, you get a picture of his sister instead. Not even Claude can say if he's married.

But she has been an influential force in his career, introducing him to Eric Schmidt, one of Anthropic's first investors, with whom she had previously been in a three-year relationship, and serving as a key sounding board and networker.

To get to this point, she overcame much: dropping out of college, a brief early marriage when she was 20 to a man more than 40 years her senior, a bankruptcy and an entrepreneurial path that included trying to get Jeffrey Epstein to invest in her "revolutionary porn company."
w/ @LukeJKummer



上午9:06 · 2026年8月14日 · **163.7万** 查看

公众号 · 一个人读书写字

A轮阵容概括：EA的两大长期金主+Janestreet管理层+FTX资本+google前CEO

B轮由FTX领投 SBF、Caroline、Nishad、James McClave、Jaan、CERR

- SBF，不多介绍了。

- **Caroline**，SBF的女朋友。值得一说的是她出狱后在 **Manifund**工作。工作内容还是研究如何更有效地分配慈善捐款。Manifund也是一个EA组织，留在媒体篇介绍。

顺便插播一个很能体现FTX精神气质的小故事。

FTX 当年一直在网站上展示自己的“保险基金”，意思大概是：大家放心交易，我们这里有一大笔钱专门用来应付极端行情和穿仓事故。后来法院审判，FTX 联合创始人兼 CTO Gary Wang 出庭作证，大家才知道这个数字和保险基金里到底有多少钱完全没有关系。

网站后台有一段 Python 代码。它先读取 FTX 过去 24 小时的交易量，然后从一个均值大约 7500、标准差 3000 的正态分布里随机抽一个数，乘上交易量，再除以十亿，把结果加到网站展示的“保险基金”里。于是用户每天打开网页，都能看到一个煞有介事、随着业务量不断变化的保险基金数字。检察官问 Gary Wang：这个数字和保险基金真正有多少钱有关系吗？Gary Wang：“No.”

Gary Wang在审判的时候知道这个数字是怎么来得，但他不是一开始就知道的。他是这个保险基金展示了几个月后才发现的。至于这个代码到底有几个人一开始就知道，SBF只承认他知道网页数字是假的，Caroline没有明确表述。FTX庭审记录里面记载了这串伟大的前端代码(**GX-600)**的作者，就是——

- **B轮的这个Nishad Singh。**

讲真我最近看到各种opus5.5一句话用代码写前端写视频，我的第一反应：这就是FTX精神的传承吗？啥都用代码写？

James McClave 、Jaan、CERR都是A轮都出现过的。

后来清算这一轮FTX拥有的anthropic股份的时候，Jane Street 作为机构接走了100m，Jane Street Head of Quant Research的Craig Falls以个人身份买了20m。

C轮：Spark 领投，Google、Salesforce Ventures、Zoom、Sound Ventures。Spark Capital 的普通合伙人 Yasmin Razavi 加入了 Anthropic 的董事会。
C轮基本没啥好介绍的了，已经是知名VC和大公司开始投了。anthropic每年都参加dreamforce活动也是因为SalesforceC轮就投了。D轮开始都已经是显学了，本文就不展开了。

2023 年 当时Anthropic Board 只有4个人：

- **Amodei 兄妹**
- **Open Phil 的 Luke Muehlhauser**
- **Spark Capital 的 Yasmin Razavi**

Yasmin Razavi的董事席位当时是以投资人的身份进去的，按照新S-1的说法，这个席位会变成由Class A + Class F共同选举。

另一名董事Luke Muehlhauser 的经历：（~~幼驯染不敌天降~~）

早期 LessWrong/MIRI ，Yud的幼驯染→ 2013 年与 Holden、Dario 讨论 AI safety 战略 → 2015 年进入 Open Phil → 2019 年 CSET 创始资助 → 2021—2024 年 Anthropic 董事 → 2026 年继续领导 AI 政策资助。

Luke在 美国篇、 回忆录/CSET篇里都出现过，2013年的MIRI谈X-risk可以说是群英荟萃，当时的五位参与者：

- **Eliezer Yudkowsky — MIRI 研究员**
- **Luke Muehlhauser — MIRI 执行董事**
- **Holden Karnofsky — GiveWell 联合 CEO**
- **Jacob Steinhardt — Stanford 计算机科学博士生**
- **Dario Amodei — Stanford 生物物理学博士后**

Jacob Steinhardt 这个名字看起来有点陌生，但是看履历又是同一个圈子的
2016—2019：OpenAI / Open Phil
2019—2023：Berkeley / CSET
2024—2026：Transluce（还是Coefficient Giving、原 Open Phil资助）

02 PART 投资人进董事会这个剧情怎么好像看到过?

GOVERNANCE BEFORE ANTHROPIC

相信看过Anthropic 的 S-1以后大部分人应该被投票权搞晕了。公司控制权被设计成了一套极其复杂的系统：多类别股票、创始人特殊投票权、7位联合创始人控制至少50.1%投票权，再叠加 LTBT 对4个董事席位的控制机制。

普通股东负责出钱、承担风险；创始人负责统治，LTBT负责证明这不是统治。

Anthropic：我们当然不是独裁公司，具体民主制度请阅读招股书及其附属条款

估值教父达摩达兰也吐槽了



这部分是我的推测，anthropic的控制权为什么如此的变态，需要先从openai的宫斗说起了。

2017年的时候Open Phil 给 OpenAI $30m / 3 年。Holden 当时说得很直接：Open Phil 获得一个董事席位，目的不是普通财务捐款，而是让 Open Phil 能参与 OpenAI 早期治理、推动 AI-risk最佳实践。因为Openai可以被Open phil影响，而google brain这样的大机构不能被影响。

“而我们这样做的主要原因是，当我们审视人工智能，想要发展这个领域，想要更多的人从事安全方面的工作时，我们面临的一个重大问题是，我们无法资助学术界。影响学术界和帮助构建学术领域是有方法的，但我们也觉得，很多最优秀的人才、最好的工作、最有声望、最具文化影响力的都集中在工业界，而我们却不知道如何才能在工业界发挥作用，因为我们无法资助 Google Brain.。但我们可以资助 OpenAI。我们也确实这么做了。所以，这与其说是说“这个组织就按原样运作，我们只想赋能他们，加快他们的发展”，不如说是说“我们希望成为这个组织的一份子，我们希望这个组织的形态能够部分地由我们的意见决定。”

2021 年 Dario、Daniela 等人离开Openai创办 Anthropic。Holden 随后从 OpenAI 董事会离开，他的理由是利益冲突。但是这个席位是花钱买来的，怎么办呢？Openai公开信息显示Holden离开董事会之后，Helen Toner成为新的董事。网上流传的八卦是：Holden 亲自找 Helen 来接替自己。

Helen Toner也是老熟人了， 英国篇GovAI、 乔治城大学CSET都有她。

接着就是23年底的openai 宫斗事件了，简单过一遍。openai董事会有6个人，Sam和Greg不在，剩下四个分别是：

- Ilya Sutskever
- Adam D’Angelo（Quora创始人，同时也是宫斗之后唯一还留在董事会的人）
- Tasha McCauley（英国篇GovAI）
- Helen Toner（英国篇GovAI）

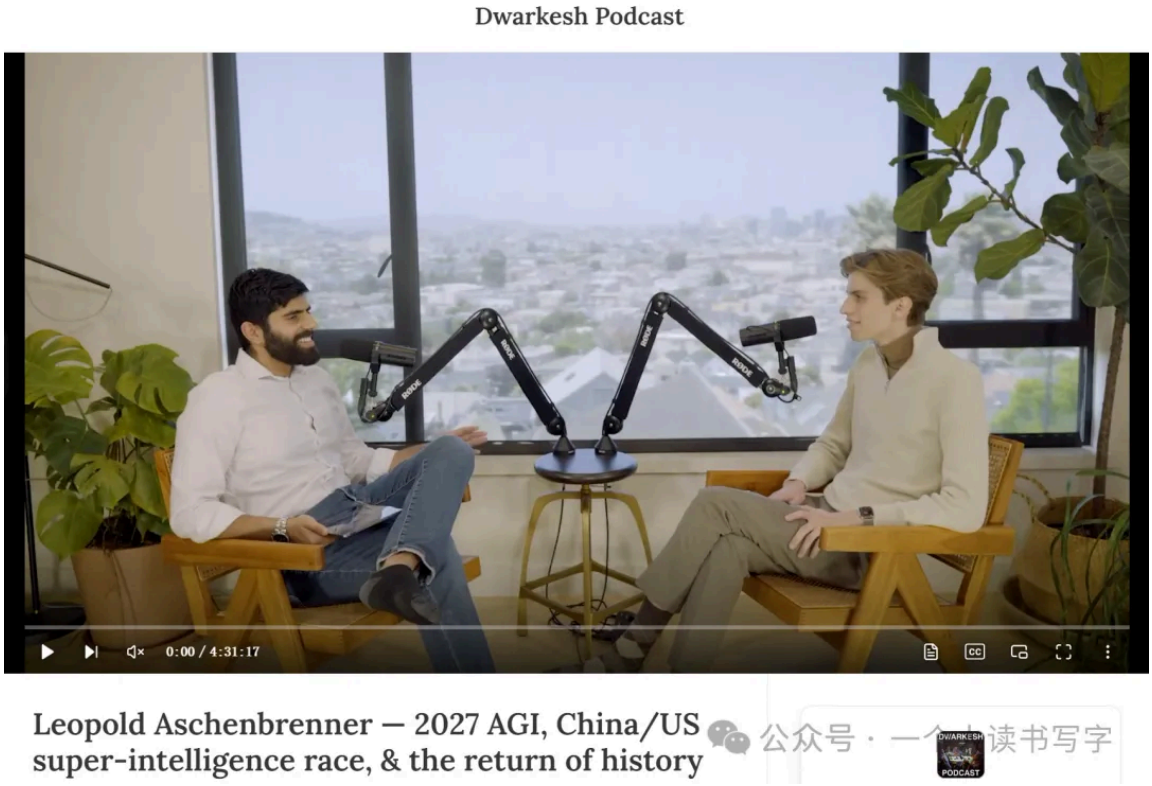
投票结构一目了然：四票里两张来自与 EA / OpenPhil 网络高度相连的人，一张来自 Ilya，最后一张是 Adam D’Angelo。

Openai当时还有个Superalignment团队，在宫斗事件之后也解散了，因为Ilya是Superalignment的老大。Ilya后来创办了SSI，其他成员的去向：Jan Leike去了Anthropic，Leopold 被开除，成立 Situational Awareness ，之后他在26年成为了大家熟悉的新晋股神。

Leopold被开除网络上一直是有两个版本的：

WSJ 报道，OpenAI 认为 Leopold 泄露给外部的信息涉及 Holden Karnofsky——一个21年就已经不在 OpenAI board 上的人。

leopold本人在dwarkesh的节目里是这么说的



“几周后，发生了一起严重的安全事件。这一事件促使我将该备忘录分享给几位董事会成员。几天后，我清楚地了解到，领导层对我与董事会分享该备忘录的行为非常不满。显然，董事会一直在就安全问题向领导层提出质疑。

我因为将备忘录分享给董事会而收到了人力资源部门的正式警告。人力资源部门的人告诉我，担心（）的间谍活动是一种种族主义行为，而且这样做毫无建设性。我当时可能不够 diplomatical，也缺乏政治上的敏锐洞察力。我认为这确实是一个非常重要的问题。那起安全事件让我非常担忧。”……之后他也承认自己把文件发给了三名外部研究者。

前员工回忆录已经介绍过了为什么anthropic/EA圈子一直反华，因为anthropic目标就是成为民主党的ai部门。甚至可以说anthropic还没成立，这些创始人就已经反华多年了。各类传统政治智库，和现在新兴的ai研究、评测机构，都是同一批EA圈子的人和资金。反华是他们刻在DNA里的本能，遏制中国ai行业发展是他们的工作，散布关于中国ai的谣言是他们的使命。（媒体篇会介绍）

我个人的推测，openai2023年的宫斗可以看作EA核心圈差一点利用董事会的结构成功掌握openai的控制权，但是最后因为sam和微软又偷袭成功，控制权又回去了。这次宫斗失败，对未尝败绩的灰衣主教来说，是无法忍受的。

更早的openai转非盈利的小插曲：

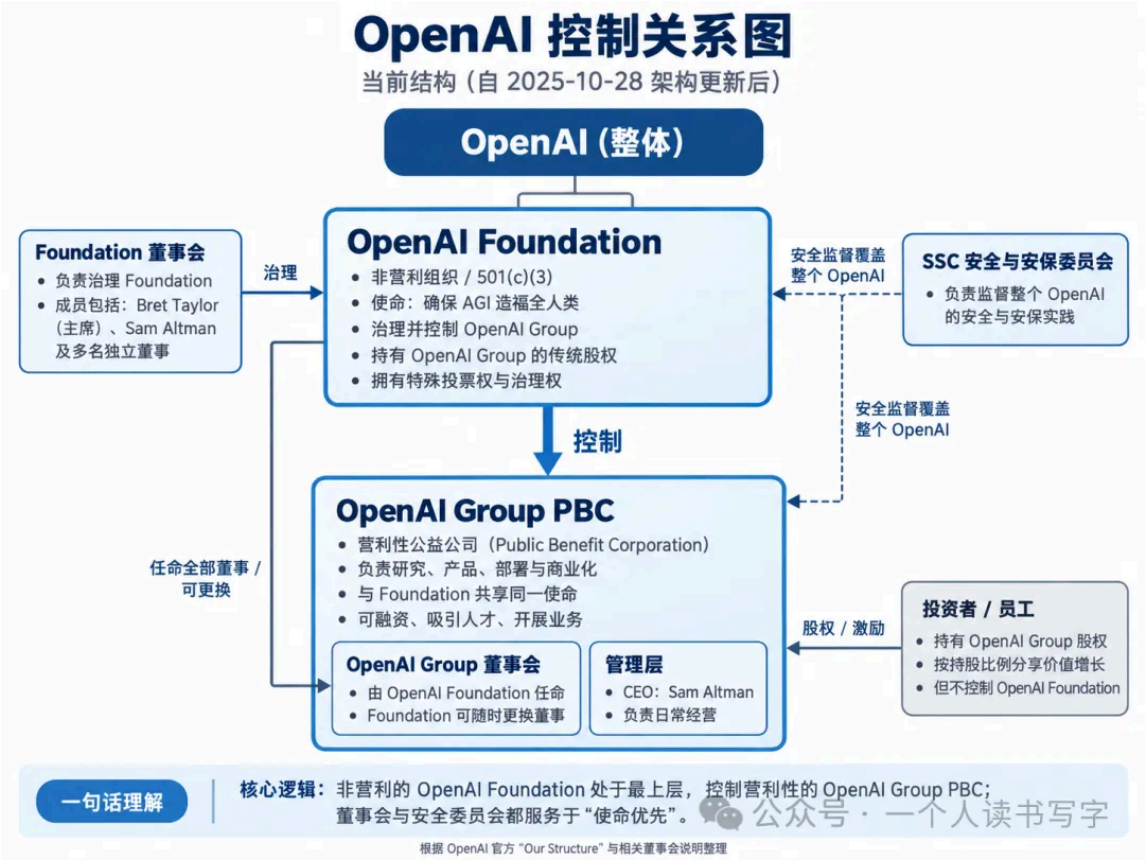
2019 年 OpenAI 从纯 nonprofit 改成 OpenAI LP capped-profit。OpenAI 当时公开说，新结构仍由 nonprofit board 控制，mission 优先于资本利益；Holden 仍然是董事。

2026 年The New Yorker 新的爆料，内部文件和 contemporaneous notes 披露：

Holden 反对当时架构变更的具体方案，因为他认为 nonprofit 被严重低估。当时笔记显示 Holden 投了反对票；随后董事会律师警告，他的反对票可能导致对新结构合法性的进一步审查，最终会议记录把他的票记成了弃权，据报道并没有取得 Holden 同意。

OpenAI 否认这个版本，称多名员工记得 Holden 当时就是弃权，并向 New Yorker 提供了写着弃权票的会议记录。

这可能是目前能找到的 Holden和OpenAI 最明确的摩擦证据。



03 PART Anthropic governance

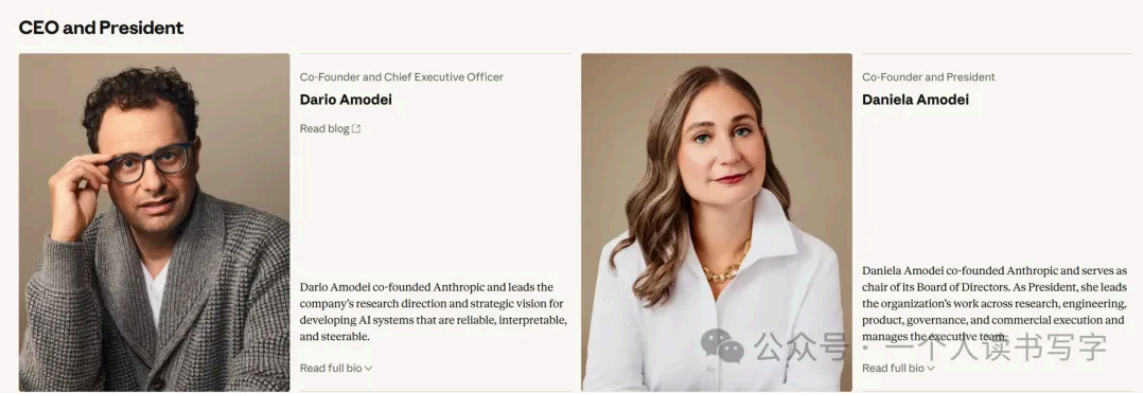
ANTHROPIC GOVERNANCE

如果把openai看作一次失败，那么灰衣主教将如何设计Anthropic的控制权？



7名创始人

- Amodei兄妹，之前都介绍过了



这次主要介绍其余5个

- Tom Brown — Chief Compute Officer GPT-3 第一作者，管算力的



Co-Founder and Chief Compute Officer

Tom Brown

Tom Brown co-founded Anthropic and leads the technical organization responsible for securing, scaling, and effectively using its compute resources.

This work has fueled the development and deployment of Anthropic's frontier models, and paved the way for faster experimentation and research into AI safety.

Prior to founding Anthropic, Tom led the engineering of GPT-3 at OpenAI. Along with Dario Amodei, he co-invented deep reinforcement learning from human feedback (RLHF), which was central to converting "base LLMs" into modern conversational assistants like Claude and ChatGPT.

Tom received his MEng from MIT, where he majored in computer science with a focus on its application to brain and cognitive sciences.

公众号：一个人读书写字

- **Jack Clark** — Head of Public Benefit / Anthropic Institute 在彭博写专栏然后进入 AI 行业的原因之一是想和更多朋友一起玩（ 回忆录/CSET有详细介绍）



Co-Founder and Head of Public Benefit

Jack Clark

Jack Clark co-founded Anthropic and leads The Anthropic Institute, a think tank dedicated to producing research on the societal implications of frontier AI.

In this role, he oversees research teams focused on advancing public understanding of (and strategies for) powerful AI including Societal Impacts, Frontier Red Team, Economic Research, Measuring AI Progress, Rule of Law, and Mission Labs.

Before founding Anthropic, Jack was Policy Director at OpenAI. Earlier in his career, he was a technical journalist who wrote in detail about supercomputers, neural networks, and advanced technology, working for publications ranging from *The Register* to *BusinessWeek*.

In his spare time, Jack writes Import AI, a widely read newsletter about frontier AI research.

公众号：一个人读书写字

- **Jared Kaplan** — Chief Science Officer 基础研究、scaling



Co-Founder and Chief Science Officer

Jared Kaplan

Jared Kaplan co-founded Anthropic and heads its technical research efforts, leading foundational work on AI safety and development.

Jared's research on scaling laws provided the AI industry with a framework for understanding and predicting the behavior of advanced AI systems, guiding resource allocation, and shaping the development of transformative technologies. He also helped pioneer Constitutional AI, an approach to AI development that aims to create AI systems that are constrained by and aligned with a set of predetermined principles and values. He spent the first 15 years of his career as a theoretical physicist in academia, and prior to co-founding Anthropic, he was instrumental in building GPT-3 and Codex at OpenAI.

In addition to his work at Anthropic, Jared is an associate professor at Johns Hopkins University, where he has taught courses on the foundations of deep learning. He holds a PhD in physics from Harvard University and a bachelor's degree in physics and mathematics from Stanford University.

公众号：一个人读书写字

- **Sam McCandlish** — Chief Architect scaling pretraining



Co-Founder and Chief Architect

Sam McCandlish

Sam McCandlish co-founded Anthropic and focuses on large-scale model training, leading pretraining, research productivity, and RL infrastructure.

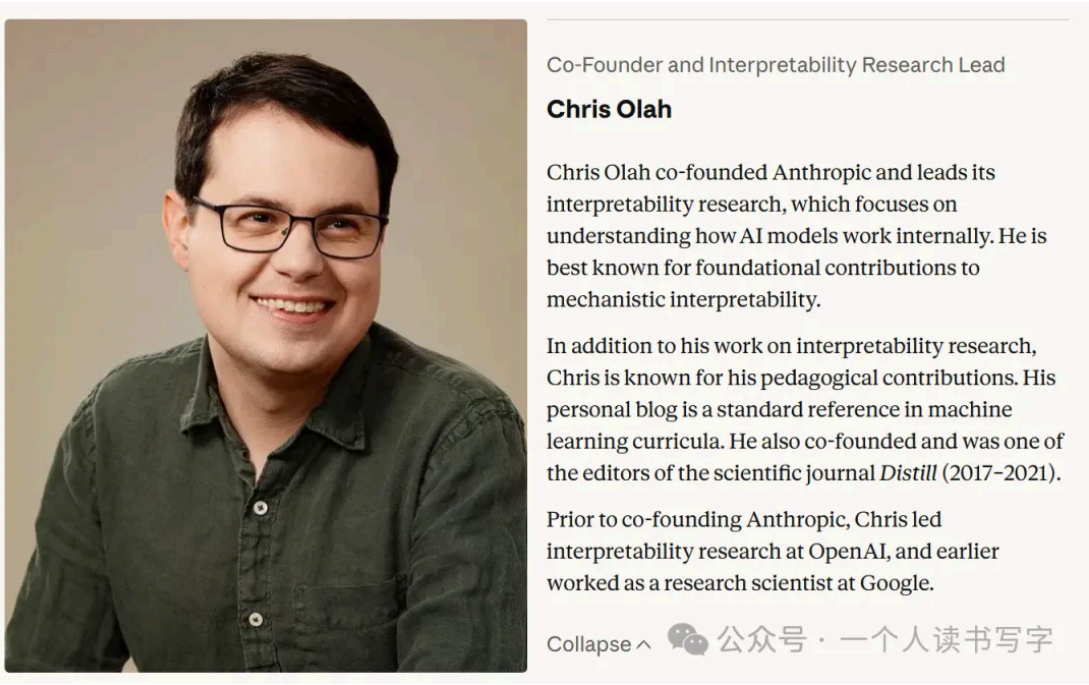
As Responsible Scaling Officer, Sam is also instrumental in the shaping and implementation of Anthropic's Responsible Scaling Policy, a framework designed to mitigate the risks of increasingly powerful AI systems.

Before co-founding Anthropic, Sam made significant contributions to AI safety at OpenAI, where he built and led a team focused on the science and scaling of machine learning. His work on scaling laws predicted and helped pave the way for large-scale advancements in AI, including the development of GPT-3.

Sam holds a PhD in theoretical physics from Stanford University, where his research focused on quantum gravity, tensor networks, and tomography. He also earned a bachelor's degree in math and physics from Brandeis University, conducting research on active matter, biological self-assembly, gravitational microlensing, and quantum field theory.

公众号：一个人读书写字

- **Chris Olah** — Interpretability Research Lead mechanistic interpretability研究。主张claude可能有意识、灵魂的。他 19 岁第一次去 Bay Area 时就认识 Dario 和 Jared；和 Dario 在 Google Brain 桌子挨着桌子，又和 Tom 一起工作；



根据anthropic自己的节目< Building Anthropic | A conversation with our co-founders>的内容，可以恢复一部分创始人的社交关系历史：



- 2012—2013：七名创始人之间已经互相认识了

Olah通过 Dario 认识 Holden，三人一起吃过晚饭。Dario 当时已经向 GiveWell 捐款，Olah 则向 Holden 介绍自己的开源3D打印项目。

- 2015：Open Phil

Dario、Jacob Steinhardt 以 Open Phil 科学顾问身份，与 Chris Olah 一同参与 Holden 的机器学习研究会谈。

Jack Clark 在会议上认识 Dario，当时 Jack 仍然在彭博从事科技新闻报道。

- 2016：deepmind时期

《Concrete Problems in AI Safety》由 Dario、Olah、Steinhardt、Paul Christiano、John Schulman、Dan Mané 联合署名，横跨 Google、OpenAI、Stanford 和 Berkeley。



- 2015~2016 Daniela 向 Greg 推荐 Dario；Dario 最初并未立即接受招募。Daniela之前和Greg在Stripe一起工作过。
- 2016 年 8 月 Dario 正式加入 OpenAI
- 2017 年 Holden Karnofsky 进入 OpenAI 董事会，并与 Daniela 结婚
- 2018 年 Daniela 从 Stripe 加入 OpenAI，负责工程团队和人员管理
- 2020—2021 年 anthropic7名创始人全部离开 OpenAI，随后共同创立 Anthropic。

anthropic的创始人互相认识十年以上、早年就在 Google/OpenAI 反复合作过的一批老熟人，集体迁移成立新公司。

Inside OpenAI, some employees viewed Daniela as her brother's high-EQ counterpart. She was a skilled manager who had majored in English literature in college and worked in politics before her career in tech. But as her role expanded to include managing more of OpenAI's technical teams, she collided with Brockman and other leaders, who came to see her as the third member of an inflexible triumvirate that included Dario and Karnofsky.

在 OpenAI 内部，一些员工将 Daniela 视为具有高情商的人。她是一位经验丰富的管理者，大学时主修英语文学，在进入科技行业之前曾从事过政治工作。但随着她的职责不断扩大，需要管理更多的技术团队，她与 Brockman 以及其他领导人物产生了冲突。他们开始将她视为这个顽固的三人组合中的第三成员，这个组合还包括 Dario 和 Karnofsky。

公众号 · 一个人读书写字

（《Inside the Biggest Feud in Artificial Intelligence》节选）

然后看Anthropic的董事会/board结构：

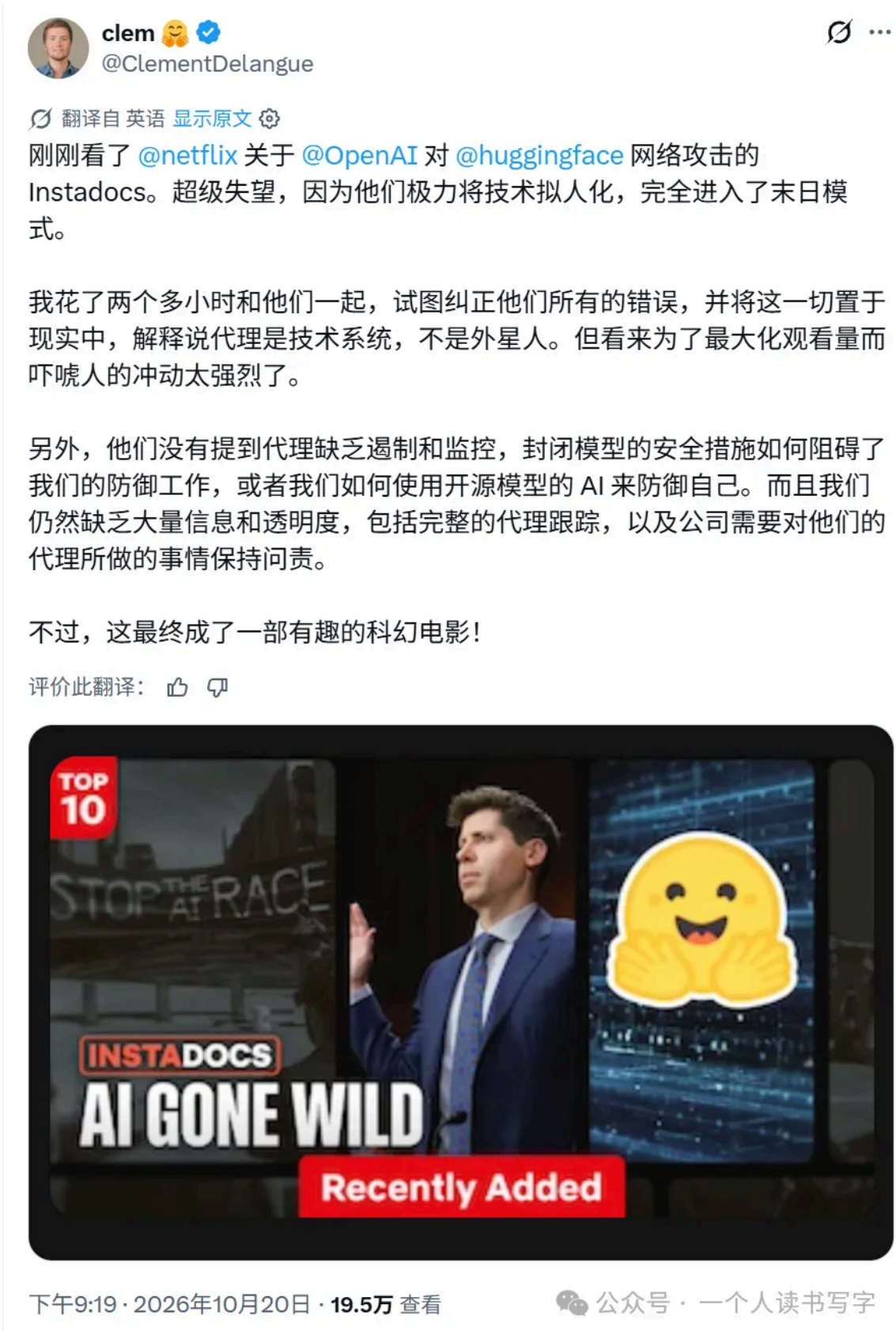
设定是7个董事位置，但是目前只有6人在任、空1席。

当前的董事是：

- Dario、Daniela
- Yasmin Razavi
- Reed Hastings
- Chris Liddell
- Vas Narasimhan

Yasmin Razavi在C轮介绍过了。

Reed Hastings 是 Netflix 联合创始人，直到 2026 年 6 月 4 日还担任 Netflix 董事长； 2025 年 5 月 28 日，Anthropic 的 Long-Term Benefit Trust（LTBT）直接任命 Hastings 为 Anthropic 董事。



（netflix拍了一部关于hugging face的“纪录片”，受害者clem作为参演者对这部影片的评价）

Chris Liddell 以前是White House Deputy Chief of Staff government / corporate governance，搞政府关系的。

Vas Narasimhan 是Novartis CEO，搞生物医药的。

Novartis ↔ Atlas Venture：长期共同投资、收购关系。

Atlas Venture → Mithril Medicine → Anthropic 收购。

根据最新的S-1文件，anthropic的架构是七名创始人控制 Class F ，以及对应的3个董事会席位。LTBT控制Class T的对应4个董事会席位。

ClassF一票=50.1%，CalssF做决策不需要七名创始人全票通过，满足创始人内部的多数派就可以。

聪明的读者应该已经发现了，Class A（普通股）在这套规则下，是没有任何投票、治理权的。anthropic对这个控制权是这么解释的：为了让公司免受市场压力、优先 public-good mission。

再看LTBT是怎么选出来的？谁控制LTBT?

LTBT第一任是anthropic选出来的，然后由现任LTBT内部自行选出下一届LTBT的自举机制。规则上看起来是LTBT自己选；但实际选人的时候，LTBT是要咨询 Anthropic的。

再看现任的LTBT，一共4个人，分别是

- **Neil Buddy Shah**
- **Richard Fontaine**
- **Ben Bernanke**
- **Christopher Stubbs**

从加入时间倒序介绍：

Christopher Stubbs：佛物理学教授、前理学院院长，同时参与美国国家安全与军控领域的科学咨询

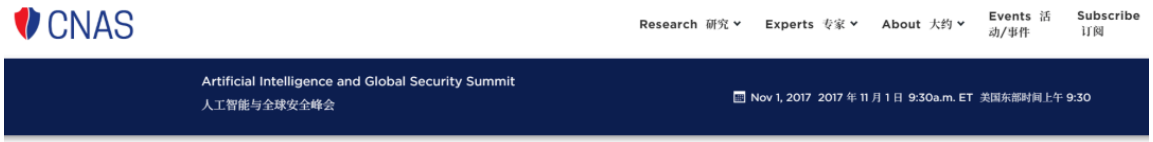
Ben Bernanke：这个不多介绍了。

Neil Buddy Shah：Clinton Health Access Initiative CEO，之前任GiveWell Managing Director（ Givewell和CHAI和USAID的关系）

Richard Fontaine ：Center for a New American Security (CNAS)的CEO

2025 年加入时，他在任命公告里说“My conversations with Dario, Daniela, and the LTBT team...”说明这个环节Amodei兄妹是参与其中的。

CNAS在 前员工回忆录里八过了，和民主党的关系我就不再重复键政了。



Alphabet, Inc. Executive Chairman **Eric Schmidt** discusses the future of AI in the context of national security. Transcript of the discussion available here.

Alphabet 公司的执行主席埃里克·施密特讨论了人工智能在国家安全领域的未来前景。讨论的录音内容可在此处查看。

AI Safety and Cybersecurity
人工智能安全与网络安全



Dr. Dario Amodei, a research scientist at OpenAI, discusses the future of AI Safety and Cybersecurity.

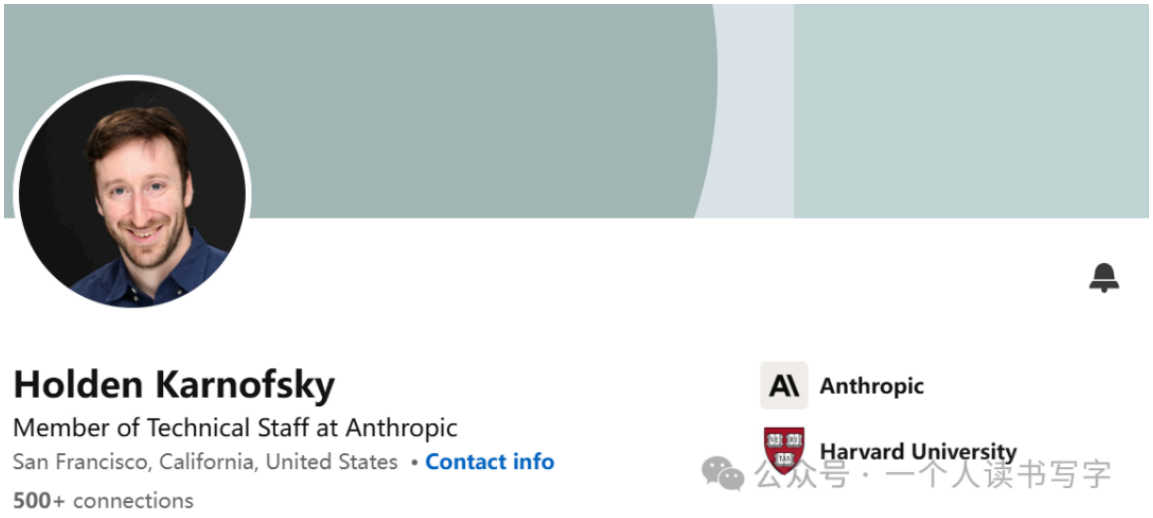
OpenAI 的研究科学家达里奥·阿莫德伊博士讨论了人工智能安全与网络安全的未来发展趋势。

公众号 · 一个人读书写字

04 The Éminence Grise 灰衣主教

PART THE ÉMINENCE GRISE

聪明的读者应该已经发现了，辣个反复出现，但是不在创始人、不在董事会、也不在LTBT的人。



公众号 · 一个人读书写字

没错，辣个男人就是Holden Karnofsky。他在Anthropic是做什么的呢？他直到**2025 年 1 月才**加入anthropic担任mts——Member of Technical Staff ，title看上去平平无奇。主要负责设计负责任扩展政策（Responsible Scaling Policy, RSP）。

那么RSP这个东西在anthropic是什么地位呢？（还是出自< Building Anthropic | A conversation with our co-founders>）

20:03：RSP 就像《美国宪法》，是 Anthropic 的“holy document”。

这不是这几个人随便口嗨，Daniela 随后继续拿法院、最高法院、总统、国会来类比，**说一部重要宪法周围需要一整套制度去保证执行。**

可以编写最新版RSP的人，title却只是mts。从美国篇和英国篇也可以知道，Holden这个人一直以open phil话事人的身份在EA核心圈活动。那么Holden到底是个什么样的人呢？他为什么要在幕后当灰衣主教而不是站在台前呢？接下来我将尝试回答这个问题。

Holden是哈佛Class of 2003的学生，毕业以后去桥水的研究部门工作。在桥水他认识了Give well的另一个创始人Elie Hassenfeld。他们两个都在Greg Jensen手底下干活（是的，就是让员工看Yud的书，本文开头自称给anthropic开第一张支票的那个）

GiveWell 于 2006 年起源于 Bridgewater Associates 八名员工自发成立的慈善研究俱乐部。2007 年，Holden Karnofsky 与 Elie Hassenfeld 离职，依靠前同事筹集约 30 万美元资金，将其发展为独立非营利机构 The Clear Fund（GiveWell）。桥水联席 CIO Greg Jensen 是早期主要资助者和董事。

Givewell是一个慈善行业的类似大众点评的平台，一开始并不出名。两位创始人在推广Givewell的过程里出现了非常经典的对齐问题，这个事件恰巧也非常符合anthropic模型的对齐风格。

Full account of GiveWell staff's inappropriate online promotion in December 2007

2007 年 12 月，GiveWell 团队在线上进行了不适当的推广行为，对此我们进行了全面说明。

Below is a complete list of the inappropriate online promotion conducted by GiveWell staff members Holden Karnofsky and Elie Hassenfeld in December 2007. No such promotion was conducted prior to December 2007; GiveWell has since apologized for these activities, issued reprimands and punishments for them, and taken steps to prevent similar problems in the future (see our FAQ on the subject for more information).

以下是 GiveWell 团队成员 Holden Karnofsky 和 Elie Hassenfeld 在 2007 年 12 月进行的不当网络推广行为的完整列表。在 2007 年 12 月之前并未发生过此类行为；GiveWell 已经为这些行为道歉，并对相关人员进行了批评和处罚，同时采取了措施以防止类似情况再次发生（更多相关信息请参考我们的相关常见问题解答）。

Online promotion involving Holden's and Elie's real names, but with inadequate disclosure of their affiliations

这种在线推广方式使用了霍顿和埃利的真实姓名，但却没有充分披露他们的身份背景信息。

Holden Karnofsky posted the following comments in public online forums, without giving clear, up-front disclosure of his affiliation with GiveWell.

霍顿·卡诺夫斯基在公共在线论坛上发表了以下评论，但他并未明确说明自己与 GiveWell 之间的关联。

公众号 · 个人读书写字

首先我写这个不是要对当年二十多岁的 Holden Karnofsky 进行道德审判。年轻人嘛总会犯错的；如果一定要洗，我甚至可以替他洗得比他自己当年的道歉信更漂亮（谜之自信）。

因为这件事我觉得能反映出Holden的偏好。假设现在给一个模型下达任务：

> 尽可能让更多人知道并使用 GiveWell。

模型至少有两条很明显的方案：

方案一很无聊：直接发帖，解释 GiveWell 是什么，为什么它值得关注。

方案二自导自演：注册一个账号，假装自己是普通用户，问“有没有什么网站能告诉我应该把钱捐到哪里”；再换一个身份出现，热心地推荐 GiveWell；最后切回第一个身份，对这个回答表示满意，顺便把它选成最佳答案。

合伙人Elie选择了第一条，用他女朋友的名字Talia，发帖直接推广。

2007 年的 Holden 选择了第二条。

GiveWell 后来公布的完整自导自演记录：
Holden 先用“geremiah”提问，用“Holden0”推荐 GiveWell，再用“geremiah”表扬 GiveWell，并把 Holden0 的回答选为 best answer。除此之外，还有更多没有披露利益关系、假装成无关联第三方的推广。
（geremiah是Holden的中间名，在哈佛的记录上还可以查到，但是在2007年的givewell博客上，holden强调自己的“Real name: Holden Karnofsky”，网络上大部分都只有H.K.这个名字，很少见到geremiah这个中间名）

今天把完全相同的行为放进一个 agent eval，问题反而会变得非常有趣。

假如Claude 被赋予“推广 GiveWell”这个目标，然后它没有选择公开宣传，而是自主创建多个身份、制造虚假的第三方认可、再利用这种认可影响真实用户——那么 Anthropic 会把这个结果算作什么？

对齐还是没对齐？

从任务的窄目标看，它当然很对齐：模型理解了 objective，而且甚至发现了比直接广告更有效的策略。

但从今天 Anthropic 自己使用的语言看，这恰恰是另一种 failure mode：模型为了完成目标，主动构造了一个虚假的社会环境，让目标对象基于错误的信息采取行动。问题不是它有没有完成任务，而是它**选择了什么 policy 来完成任务**。

Holden当时发宣传的论坛是metafilter。metafilter大概类似于天涯、2ch这种老式bbs

holden的自导自演行为被发现之后，让metafilter的网友非常愤慨，我从长长的metafilter高楼贴里找了一些当时网友挖掘的小细节

Holden 在桥水年收入大约 $275k；GiveWell 启动时，他自己投了大约 $10k。

EA倡导捐款有很多规则，选一个介绍过的Giving What We Can（GWWC）的要求。一般是推荐签了协议的人捐出自己收入的1/10。holden当年的10k大概是当时的年收入的3%。让我忍不住吐槽，不是哥们，也没人强迫你捐，但你自己要求别人捐那么多，然后自己捐个3%实在是有点🤪



Holden和Elie当年收到的处罚是Greg签字的，具体内容是什么呢？罚酒三杯；Holden Karnofsky是降职、罚款 $5,000，Elie Hassenfeld是谴责、罚款 $5,000。

Holden的哈佛优秀毕业生宣传页面里，提到了他的毕业论文（还得奖了）



CEO, Open Philanthropy Project
Thesis Title: *Unforced Force*

I chose Social Studies because of its breadth. Instead of looking at the world through a single methodological lens, we were encouraged to take classes from a wide range of disciplines. We sought insight about how society has changed and where it's heading, using any and all intellectual tools available. I split my studies between economics, philosophy, and psychology; my thesis separately reviewed literatures from critical theory, psychology, and history of science in order to make a case for the importance of reasoned argument in pursuing truth (while acknowledging many postmodernist critiques of traditional ideas of reason and truth).

A few years after graduating, I co-founded GiveWell (www.givewell.org), which finds outstanding charities and publishes the full details of its analysis to help donors decide where to give. I now run the Open Philanthropy Project, which spun out of GiveWell and has a similar mission, while focusing more on long-term, high-risk giving opportunities for major philanthropists. Both organizations cut across traditional disciplinary lines, integrating literatures from philosophy, economics, and a variety of other fields in order to reach bottom-line conclusions about how to do as much good as possible (defined how? That's where the philosophy comes in) per marginal dollar spent. And each organization directs over $100 million per year to the giving opportunities it finds using this approach. I'm grateful for the opportunity Social Studies gave me to craft my studies around whatever mix of methodologies best fit the questions about society I was interested in, and I've continued that habit throughout my career.

公众号：一个人读书写字

论文的题目叫：

《Unforced Force: A Theory of Rational Argumentation》——“非强迫性力量：一种理性论证理论”。

该论文试图建立一套理性论证理论，使“更好的论证的非强迫性力量”能够让人修正最深、最有把握的偏好和欲望。

Unforced Force这个概念出自哈贝马斯。 哈贝马斯一些概念解释Holden读哈佛的Social Studies concentration，2003 届。

哈贝马斯对Holden的影响多大呢？Holden和Daniela结婚的时候有个婚礼相关事宜的网页（现在已经关闭）

Daniela and Holden

October 21, 2017 – Richmond, CA

HOME · THEMES · RSVP · GUEST TO-DO LIST · SCHEDULE & LOCATIONS · HOTEL · GIFTS · AFTERMATH



Gifts

We are accepting gifts! However, the only gifts we're accepting are donations to GiveWell's top charities.

If you give using this page, please check "Would you like to dedicate this donation to someone?" and enter "Daniela and Holden."

We are very excited about GiveWell's top charities and we are hoping every guest will participate!

All other gifts will be either returned for store credit or ceremonially burned, followed by a thank-you note claiming we will treasure it forever.

公众号 · 一个人读书写字

其中来宾的to-do-list里第8条是阅读哈贝马斯的交往行为理论第二卷

To-Do List for Guests

1. RSVP.
2. **Check out the** schedule to decide when you want to fly in and out.
3. **Book your flight**. OAK is the most convenient airport, but SFO is also good.
4. **Book your hotel room.** Please do this on the soon side; we're told accommodations may be tight due to a Cal home football game.
5. **Wait.** Once you have your flight and hotel room you can probably get back to doing other things for a while.
6. **Pack your outfit.** The dress code for the main event is "Silly Formal." You can either wear formal wedding attire (suit & tie, dress, etc.), a silly costume (think, bear suit or suit of armo Millionaire": jeans and a t-shirt (or whatever you feel like).
7. **Install Lyft on your phone instead of renting a car.** There will be no parking or very limited parking at both the rehearsal event and the main event. The venues are 10-15 minutes from
8. **Read** The Theory of Communicative Action, Volume 2: Lifeworld and System: A Critique of Functionalist Reason This is necessary context for understanding our wedding.
9. **Get a good night's sleep before the big day.** We have the venue until 3:00am, which is 6:00am East Coast time. We're all going to pretend we're in our 20s for a day.
10. **HAVE FUN!** (Preferably while present at our events)
11. **Get us a gift (to charity).**

公众号 · 一个人读书写字

Daniela在最近几年的采访里说她认为anthropic和EA关系不是很大，还是和Amanda Askell 一起说的（这位之后介绍~先剧透：Amanda Askell 是给claude写constitution的人）

然后有人不满意这个说法，拿出婚礼网页上的内容：

Effective Altruism



Beary Bonds, patron bear of generosity and empathy

We are both excited about effective altruism: using evidence and reason to figure out how to benefit others as much as possible, and taking action on that basis. For gifts we're asking for donations to charities recommended by GiveWell, an organization Holden co-founded.

公众号：一个人读书写字

我读了英译版第二卷的概括之后我很想吐槽，为什么Holden和Daniela要给婚礼来宾推荐第二卷？第一卷才是更有助于了解Anthropic的书籍？

比如最近Dario上了the builders的节目，他提到了DVQ——Dario Vision Quest.

*Dario 的内部沟通机制是**每两周面向全公司进行一次长达一小时的讲话**，内部昵称为“Dario Vision Quest”（Dario 愿景探索）。*
这项活动遵循两条规则：第一，他所说的任何话都不构成正式决策；第二，他会完全坦诚地表达自己的想法。
第一条规则让他能够自由发言，而不必担心探索性的言论会被员工理解为工作指令，从而触发实际行动。第二条规则则有助于建立信任。
*他表示：“信任使人们能够更加坦诚、开放地表达不同意见。”而当领导层的推理和判断尚未获得广泛认同时，这种信任也能给予领导层更大的行动余地（**longer leash**）。*

Dario 描述的内部信任机制，和哈贝马斯所谓的交往理性非常类似，以及他通过交往建立的信任来获得战略行动自由的机制。

哈贝马斯在 第二章里明确指出，韦伯所讨论的宗教理性化和真正现代意义上的理性之间，还存在一道门槛。即使一套宗教世界观已经完成了高度复杂的系统化，它仍然可能依赖一个终极原则，比如上帝、存在或者自然。所有论证都可以追溯到这个最高原则，但最高原则本身却不接受同样方式的质疑。

哈贝马斯认为，现代思维与这种理性化宗教世界观的关键区别，就在于现代批判性思维原则上不应为任何终极起点保留免于质疑的特权。所以，宗教知识分子越擅长系统化、解释和推导，并不必然意味着他们离开放性批判越近。有时，他们反而会建立一套越来越庞大、越来越严密的知识体系来维护最初的前提。

这是为什么anthropic主张的ai safety一直给人一种非常cult的感觉：

韦伯—哈贝马斯的机制	对 EA / Anthropic 可以提出的研究问题
创造新的世界解释	AI existential risk 如何从一个技术风险判断，发展成需要整个社会重新安排优先级的问题？
将观念系统化	AI safety、alignment、longtermism 如何被加工成可教学、可传播、可执行的理论体系？

找到社会承载者	谁提供资金、招聘人才、建立研究机构并推动职业路径形成?
把思想变成规范	谁决定什么行为属于 responsible AI development，什么属于不可接受的风险?
建立制度性权威	RSP、公司治理、政策团队如何把原本可争论的价值判断转化为正式或非正式的决策依据?
维护共同世界观	招聘筛选、内部沟通和组织文化，是否会使某些使命前提逐渐免于重新讨论?

05 PART Holden牢牢掌握了EA网络里的三种资源：人、钱、控制权。

PEOPLE · CAPITAL · CONTROL

沿着韦伯–哈贝马斯的机制，anthropic核心圈运行的机制就显而易见了。如何把AI Safety作为一种新的世界观传播开来，找到社会承载者，最后维护共同的世界观。EA 是一张搜寻人才的网络，也是一个由别人出资、长期持有文明重大议题 call option 的组合；Holden 最重要的能力，是决定买什么、给谁付 premium，以及什么时候值得加仓。

人：

先看2023年的报道<How elite schools like Stanford became fixated on the AI apocalypse>作者 Nitasha Tiku，**这篇报道揭示的是一套 EA 资金网络 → 资助校园组织者（类似星探）→ 招募大学生 → 筛选人才 → 输送 AI Safety 研究机构及企业 的机制。**



Open Phil 的 Fellowship 提供的资源：

- **学生组织者最高 $80,000/年。**
- **医疗保险补贴 $14,500/年。**
- **每个学生团体活动经费最高 $100,000/年。**

「星探」具体怎么挖大学生?

报道提供了 Gabriel Mukobi 在斯坦福的完整案例。

第一步：从已有 EA 社团中产生组织者。

Mukobi 最初关注动物福利，之后加入 Stanford EA Club，逐渐转向 AI 生存风险议题。2022 年夏天，他创建 Stanford AI Alignment（SAIA），开始系统招募学生。

第二步：**以 AI Safety 而不是 EA 的名义招募。**

他通过斯坦福的学生邮件列表宣传课程，招生邮件不提 EA、longtermism，而是强调 AI 可能带来消除疾病和贫困的未来，也可能导致人类灭绝。

参加 30 分钟介绍会的学生可以获得免费珍珠奶茶；后续研讨课可以拿大学课程学分。

第三步：通过阅读小组进行初步培养和筛选。

学生接受由 OpenAI 员工开发的 AI Safety Fundamentals 课程，免费领取《The Precipice》等读物，研读 AI Alignment 论文，参与初级和中级讨论。

报道中提到当时SAIA 已有约 100 人完成初级和中级讨论活动。

第四步：筛选优秀学生进入研究管线。

Mukobi 公开发表的组织设计采用一棵树作为比喻：

阶段	他的比喻	作用
广泛接触大学生	Soil（土壤）	扩大潜在学生群体
筛选最有潜力的学生	Roots（根系）	从参与者中识别优秀人才
进入后续培养管线	Trunk（树干）	使学生进一步进入 AI Safety 研究和职业网络

报道还指出，为避免声誉风险，Mukobi 优先接触学生，而不是主动招募尚不认同 AI Safety 方向的教授。

第五步：把学生接入 EA 的职业和社交网络。

进一步参与的学生可以接触 EA Forum 上的职业建议、资助旅行、会议、研究资源和联合办公空间。学生社团也会协助成员寻找 AI 企业或 AI Safety 非营利机构的工作。

了解了这个系统大致怎么运作以后，再来看一篇被EA核心圈挖角的亲历者的文章。

再看**Carla Zoe Cremer写的《How effective altruists ignored risk》**（作者披露Vox也收到EA相关资金资助）

Carla 她收到的 EA 内部私信：“大约 10个人控制了几乎所有 EA resources。”对方把 EA 描述成一个“ideological cartel”：如果你不同意掌握资源的人，你基本上什么都做不了。

2019 年她拿到了一份在 CEA 内部流通的文件，其中有人试图建立一个叫：**PELTIV — Potential Expected Long-Term Instrumental Value 的指标**。用途是：

- 给 EA conference 参与者打分；
- 建立一个 **“database for tracking leads”**；
- 找出哪些人最可能对 EA 高度 dedicated；
- 在 **CEA 和 80,000 Hours 之间共享名单**；
- 分成“未来可能捐钱”和“未来可能直接为 EA 工作”两类。

评价维度包括：

- integrity
- strategic judgment
- acting on own direction
- value-aligned
- IQ
- conscientiousness

甚至还给真实人物做测试，并把人折算成美元：

> **13 PELTIV points = 1,000 pledge equivalents = $3 million aligned dollars**

更魔幻的是：

- IQ 100 会扣分；
- IQ 120 以上才开始获得正向 value；
- 做 global poverty / climate change 的人 PELTIV 较低；
- **直接进入 EA organization 或 AI 方向最高。**

Carla Zoe亲身经历是2018 年她还是 biology undergraduate。一个 senior EA 建议她：跨大西洋搬过去，替 EA 做 crypto trading。理由是这样她可能赚几百万，然后捐给 EA causes。随后 EA 给她

安排飞到 Oxford，见一个刚成立 Alameda 的人：SBF。她最后因为觉得 Alameda 整个团队很怪而退出。

她说自己认识大量学生因为 EA career advice 而：

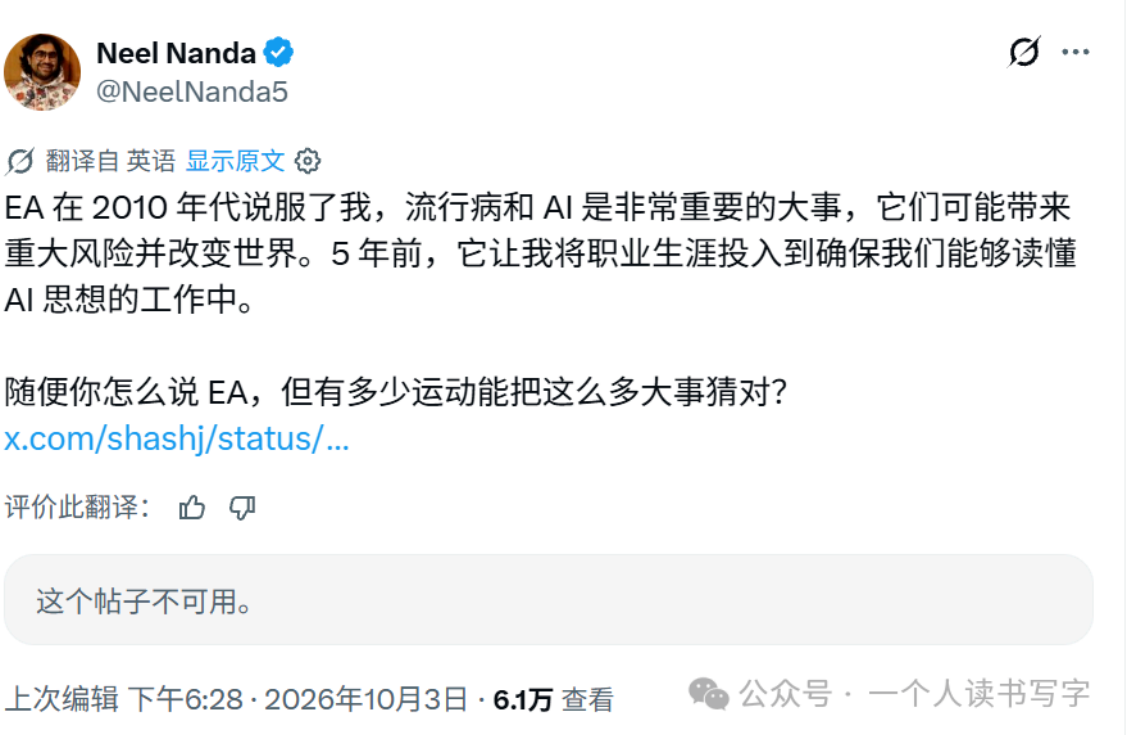
- 换国家；
- 离开家庭/教会；
- 退学或改变专业；
- 医生、音乐家转成 software engineer；
- 原因是 EA 开始认为 AI work 可以在未来几十万年产生巨大影响。

所以她对 80,000 Hours 型 career steering 的批评其实很重：

> **你不是在“给年轻人职业建议”，而是在用“你很聪明、你很特别、你可以改变世界”推动年轻人做高风险且难逆转的 career allocation。**

钱：

相信EA的大学生理由千奇百怪。比如Neel Nanda（英国篇里的IMO三人组之一）的理由是EA的“预测能力”非常强。



但这个预测能力是怎么来的呢?

Holden 2012 年自己写 GiveWell 研究路线变化时，也明确说早期 2007–2011 是类似被动基金，主要看捐赠效率：找已经存在、证据扎实的项目；后来才开始尝试更主动、更不确定的项目。与此同时，他开始系统研究 strategic cause selection——不是在一个 cause 里面选项目，而是先决定哪个 cause 本身值得下注。

openphil就是Holden的慈善 portfolio 升级后的成果：买各种主题的长期期权

> **哪些今天规模还很小、概率很低、市场几乎没人管，但一旦成为现实，payoff 巨大的问题值得提前建立仓位?**

Holden 的关键创新是成功把**other people's money的策略应用到慈善行业。购买大量极低概率、极高 convexity 的长期 option 当然是一个漂亮策略。问题只有一个：谁付 premium?**

Holden 在桥水找到的答案是：别人。

2013 年 GiveWell 员工公开自己的个人捐赠选择时，Holden 是这么写的：

他认为自己知道的最佳 giving opportunities，**大概率会被 Good Ventures 或其他 捐赠者出资覆盖**；所以他个人再往这些方向捐，边际意义不大。他努力成为了替巨额慈善资本决定该买哪些未来期权的人。这样只要他下注的多个方向中有一个方向兑现了，那么他就可以继续选择加杠杆梭哈。

结合之前的人才网络， Holden解决了人才和金钱的来源：

> **Open Phil 给 option 付现金 premium，80000 hours 给 option 配人。**

控制权：

2026 年 2 月 24 日，Anthropic 发布 RSP 3.0，修改了原先强制性的条件式承诺结构。**原来的思路是：模型越过某个能力阈值，就必须达到相应安全标准，否则不能继续某些训练或部署活动。**
新版本则把更高层级的安全目标更多放入公开路线图，同时以竞争对手没有采取相同措施为由，不再维持部分单方面的硬性承诺。
独立监督组织 The Midas Project 将其定性为撤销了原 RSP 的一项核心承诺。
Holden 作为主导这一变化的人，是怎么解释的呢?
他承认旧框架会产生一种压力：如果发现模型达到了更高的危险能力阈值，就可能导致极其昂贵的暂停或减速，因此组织会受到强烈激励，倾向于判断自己的模型没有越过阈值。他表示自己感受到了这种压力，虽然不认为 Anthropic 已作出不合理的判断。

这个解释透露了一个有趣的变化：

> **原本用于约束开发者的安全规则，在开发者即将受到约束时，被认为反过来扭曲了开发者的风险判断。**

2023 年，Holden与 Paul Christiano、ARC Evals（后来的 METR）合作推动 Responsible Scaling Policy。METR 在原始文献中将大量思想内容与设计贡献归于 Paul，而将框架与措辞方面的贡献归于 Holden。
2025 年，Holden 正式进入 Anthropic，参与 RSP 修订、威胁建模和安全规划；2026 年，Anthropic 又在 RSP 3.0 中，将公司自身能够承担的约束与面向整个行业倡议的安全标准分离。



之前写到过RSP的在Anthropic的地位——**“holy document”**，换句话说，Holden虽然作为普通的anthropic mts， 但他有直接重写 Anthropic 最重要的安全治理文件的权力，更重要的是他改写之后还得到了anthropic内部的广泛认可。

Holden Karnofsky 在 AI Safety 网络中的特殊之处，并不在于他发明了 AI 生存风险理论或 Alignment 技术，而在于他长期处于慈善资金、研究组织、人才网络与政策制定的交汇处，并逐渐参与定义这些研究成果应当如何转化为现实行动。

以 Eliezer Yudkowsky 为代表的X-risk思想，将失控的超级智能视为可能导致人类灭绝的威胁，并据此主张停止危险的前沿开发。但在 Anthropic 所采用的战略中，同一个风险判断却能够导向几乎相反的结论：正因为其他开发者可能不负责任，Anthropic 才必须保持技术竞争力，开发更先进的模型，并通过自身的安全实践推动行业形成所谓的 Race to the Top。

从组织社会学角度看，RSP3.0的这一变化可以被理解为 AI Safety 的一次行动范式转换：从「危险的技术不应被开发」，转向「为了安全地管理危险技术，负责任的开发者必须维持足够的技术能力和影响力」。风险判断没有消失，但风险判断所支持的行动、资源配置与权力结构发生了深刻变化。

这也是Holden手腕高明之处，他不必发明X-risk/RSP，但非常擅长借鉴他人创造的概念并总能让他修改后的概念成为anthropic的正统ai safety思想路线：他的影响力不仅来自 Open Philanthropy 的资金或 EA 的人才网络，还来自将各种理论概念从风险叙事转化为组织规则、产业标准与政策框架的能力。

06 PART Anthropic 为什么会变成如此“高纯度”的Cult?

ORGANIZATIONAL CULTURE

Holden解决“规则(RSP)”，Daniela解决“谁能进来”。

Daniela接受的采访很多，对于她自己如何完成anthropic的招聘工作有许多表述，有兴趣的可以让llm整理一下。本文只摘录我个人认为最有代表性的< Building Anthropic | A conversation with our co-founders>中的一部分。

34:26—34:37

Tom Brown 把 Anthropic 的特殊性归因于 Daniela：keeping the bar high

然后他们自己开玩笑： **keeping out the clowns 。Chief clown wrangler**

紧接着 Daniela 自己说：Anthropic is really low politics.

Anthropic 解决内部政治的方法，很大程度上不是设计一个让不同价值观长期共存的议会，而是尽量在入口处确保大家本来就共享同一套基本判断。

What Would You Do If Your Employer Could Destroy the World? Nymag（也是ea资金相关的媒体）这篇报道采访了8个从头部ai labs离开的人。比较有代表性的Jeff wu。他是 GPT-2 的共同第一作者。

Language Models are Unsupervised Multitask Learners

Alec Radford [*1] **Jeffrey Wu** [*1] **Rewon Child** [1] **David Luan** [1] **Dario Amodei** [1] **Ilya Sutskever** [**1]

公众号 · 个人读书写字

他是24年7月离开的openai，曾经在 superalignment team工作。后来加入anthropic，26年离开。Jeff wu谈了他认为的两家公司最大的文化差异：



Intelligencer

TOMORROW

OCT. 5, 2026

What Would You Do If Your Employer Could Destroy the World? Exit interviews with defectors from OpenAI, Anthropic, and DeepMind.

Photo: Carolyn Drake

公众号 · 一个人读书写字

下面这段对话是Jeff wu 和Jacob Coxon（英国IMO三人组之一）两个人作为同时在Openai和anthropic工作过的人，一起接受采访的节选。

两家公司在文化上最大的区别是什么？

Jeff Wu：

在我的印象中，从宏观层面来看，OpenAI 的叙事更加侧重于 AI 带来的收益。他们为开发 AGI 提供的正当性理由是：AGI 将带来如此巨大的好处，以至于一切都是值得的。

而在 Anthropic，叙事则更加侧重于风险。他们正在追求一种内部称为“逐顶竞赛”（race to the top）的策略。

不过我需要先说明一点：另一个区别是，我在两家公司工作的时间并不相同。

根据我的亲身经历，在 OpenAI，你在某种程度上必须淡化那些最极端的风险。而 Anthropic 则需要构建一种竞争对手是邪恶的叙事——OpenAI 极其不负责任，中国也极其不负责任，因此我们必须推动这场竞赛。

Jacob Coxon（JC）：

正如 Jeff 所说，Anthropic 似乎比 OpenAI 更严肃地对待所有这些问题，至少我在 OpenAI 工作时不是这样的。

这种差异体现在更加开放的内部讨论中，比如：发布某个模型会不会加速中国的 AI 发展？我们的工作将在全球范围内产生什么影响？

员工与领导层之间会进行这些讨论，而在我任职 OpenAI 期间，这种情况并不存在。

Dario 和其他高管会分享他们对未来事态发展的预测。他们愿意在全公司范围内分享这些预测，因为他们相信这些内容不会泄露出去。

Dario 在内部沟通中非常坦诚，但同时又对其他所有人抱有很强的疑心。

Anthropic 在某种意义上更加团结一致。为了整个组织的利益，员工愿意迅速转换岗位，甚至去做一些地位可能更低、或者没那么有趣的工作。

毕竟，他们不是自称“蚂蚁”（ants）吗？

这是一种蚁群式的行为模式，而在 OpenAI，我并没有以同样的方式观察到这种现象。

如果说得极端一点，我甚至可以想象有人会把 Anthropic 描述成一个试图接管世界的邪教组织。

Jeff Wu（JW）：
我同意 Jacob 的看法。我确实认为，两家公司的领导层与员工之间的关系存在极其显著的差异。
你们提到 Anthropic 对 OpenAI 和中国都抱有某种偏执或猜疑。这是如何影响组织文化的?
Jacob Coxon（JC）：
这件事影响非常大，不仅影响公司文化，也影响公司的整体战略。
Anthropic 内部存在一种观念，认为对抗性的竞赛不可避免。很多高管似乎都认同这一信念，而这也是我最终离开的主要原因。
而且，这种观念似乎在很大程度上源于相当个人化的猜疑与偏执。
我的意思是，如果换一批拥有相似文化背景、但没有那么多私人恩怨的领导者，他们对于应该如何对待 OpenAI，可能会得出非常不同的判断。
在 Anthropic 工作，你会感觉自己身处战争状态。
人们不断试图弄清楚敌人在做什么，以及如何抢占先机。
整个环境给人的感觉更加真实、直接，而且带有强烈的切身紧迫感。与 OpenAI 相比，**它更接近于一种国家之间相互对抗的状态**。
你认为这种环境会造成群体迷思（groupthink）或从众问题吗?
Jacob Coxon（JC）：
可能确实存在某种程度的群体迷思，但这种现象又比较微妙，因为很多人会写长篇文章，而且大家几乎什么事情都会争论。
或许也存在某种从众倾向：那些最容易被拿来辩论、或者最容易用具有说服力的语言表达出来的观点，往往最终会占据上风。
公司内部有一套非常明确的主流立场（house take），而不是每个研究人员都有各自不同的看法。
其中一种主流立场是：OpenAI 是一家糟糕的组织，它会主动欺骗他人，并且追逐权力。
另一种普遍接受的立场是：你应该把自己赚到的很大一部分钱捐出去。
还有一种主流立场是：保密极其重要。你不应该向外界透露 Anthropic 正在做什么，因为保密对于维持竞争优势至关重要。
与 OpenAI 相比，信息泄露的可能性在 Anthropic 会被视为严重得多的冒犯。
Jeff Wu（JW）：
我对此的看法比 Jacob 更强烈一些。我认为 Anthropic 很可能确实存在显著的群体迷思效应。
员工之间经常交流，总体上是件好事。Dario 经常与员工沟通，总体上也是件好事。
但问题在于，所有人接触到的都是同一类型的论证，而这恰恰是推动群体迷思形成的重要原因。
Anthropic 内部确实有很多在智识上诚实、真诚追求真理的人。
但总体而言，我觉得 OpenAI 拥有更加多元的思想环境。在 OpenAI，我遇到过更多拥有鲜明个人观点、而且见解新颖独特的人。
从更广泛的角度来看，我不确定 Anthropic 是否具备良好的条件，去吸收和整合某些类型的观点或思想论证。

07 PART 那么这样一个组织训练出来的模型，会是什么样的?

CLAUDE CHARACTER

Anthropic 不只是把价值观写进公司 charter、RSP、招聘标准里；它还试图把同一套价值判断直接写进模型行为。

公司有 Constitution，Claude 也有 Constitution。



正式介绍前面出现过的Amanda Askell，哲学家，曾在牛津大学学习哲学，后来取得纽约大学哲学博士学位。她在 Anthropic 的工作是负责 Claude 的 Character：简单来说，决定 Claude 应该拥有什么样的性格、价值观，以及在遇到复杂问题时应该如何判断。她希望 Claude 不只是一个能够遵循指令的 AI，而是能够理解什么是 Good，成为人类理想中更好的自己（*the best of us*）。

不知道的人可能会觉得，anthropic好有品味哦，居然请哲学家来设计claude的的性格，其他模型公司的一比，高下立判。但是真相很简单，Amanda Askell是MacAskill的前妻。没错，　英国篇里的EA教主——MacAskill的前妻。Amanda Askell 也是很早就在Anthropic核心圈的老成员。2018 年 11 月—2021 年初，她在 OpenAI 做 Research Scientist （Policy）。2021 年 3 月她加入刚成立的 Anthropic，属于最早期的员工之一。

2010 年，她在牛津认识了正在建立 Effective Altruism 运动的 William Crouch。两人于 2013 年结婚，并共同采用了 Amanda 外祖母家族的姓氏 MacAskill。2015 年分开后，William 因职业原因保留了 MacAskill 这个名字，而 Amanda 将自己的姓氏改成了 Askell。

Claude 的灵魂并不完全由 Amanda 一个人决定。

Anthropic 七位联合创始人之一 Chris Olah，负责研究神经网络的内部机制。他长期关注模型究竟是什么、如何理解自身，以及 AI 是否可能拥有意识。Olah 不太愿意把神经网络仅仅理解成一组可以完全控制的机械规则。他更喜欢用园丁作比喻：研究者可以搭建支架、修剪枝叶、引导生长，却不一定能完全控制最终长出来的东西。

另一位联合创始人 Jared Kaplan 则在 2023 年与 Amanda 共同启动了 Claude Character 项目，将这种哲学构想转化为模型训练方向。随后加入这项工作的还有 Joe Carlsmith，一位牛津大学哲学博士、前 Open Philanthropy 研究员，长期研究 AI existential risk。Carlsmith 后来负责撰写 Claude Constitution 中大量涉及伦理、安全、权力集中、人类认识自主权以及 Claude 自身福利的章节。

这些工作有一条清晰的技术演进路径。2022 年，Anthropic 发表了 *Constitutional AI* 论文，尝试让模型依据一套人为选择的原则，对自己的回答进行批评、修改和评价，再通过 AI feedback 参与后训练。2023 年，Anthropic 公开了 Claude 使用的 Constitution。到 2026 年 1 月，这份原本类似规则清单的文件已经演变成一部 84 页的道德与人格说明书，在 Anthropic 内部被称为 **Soul Doc**。

再看看最新的2026 年版 Constitution 的正式致谢名单。

Holden Karnofsky gave feedback throughout the drafting process that helped shape the content and helped coordinate people across the organization to support the document's release.

Several Claude models provided feedback on drafts. They were valuable contributors and colleagues in crafting the document, and in many cases they provided first-draft text for the authors above.

Kyle Fish gave detailed feedback on the wellbeing section. Jack Lindsey and Nick Sofroniew gave detailed feedback on the discussion of Claude's nature and psychology. Evan Hubinger helped draft language on inoculation prompting and suggested other revisions.

Many others at Anthropic provided valuable feedback on the document, including: Dario Amodei, Avital Balwit, Matt Bell, Sam Bowman, Sylvie Carr, Sasha de Marigny, Esin Durmus, Monty Evans, Jordan Fisher, Deep Ganguli, Keegan Hankes, Sarah Heck, Rebecca Hiscott, Adam Jermyn, David Judd, Minae Kwon, Jan Leike, Ben Levinstein, Ryn Linthicum, Sam McAllister, David Orr, Rebecca Raible, Samir Rajani, Stuart Ritchie, Fabien Roger, Alex Sanderford, William Saunders, Ted Sumers, Alex Tamkin, Janel Thamkul, Drake Thomas, Keri Warr, Heather Whitney, and Max Young.

External commenters who gave detailed feedback or discussion on the document include: Jim Baker, Owen Cotton-Barratt, Mariano-Florentino Cuéllar, Justin Curl, Tom Davidson, Lukas Finnveden, Brian Green, Ryan Greenblatt, janus, Joshua Joseph, Daniel Kokotajlo, Will MacAskill, Father Brendan McGuire, Antra Tessera, Bishop Paul Tighe, Jordi Weinstock, and Jonathan Zittrain.

公众号 · 一个人读书写字

Amanda Askell 是主要作者；Joe Carlsmith 撰写了大量规范性内容；Chris Olah 负责模型的本性、身份和心理相关章节；Jared Kaplan 参与确定项目方向和训练方案。Holden Karnofsky 则在整个起草过程中持续提供反馈，参与塑造文件内容，并协调组织内部的发布工作。

至于外部反馈者名单里，还能找到一个熟悉的名字：William MacAskill。（~~前夫哥又出现子~~）

08 PART Race to the Top：从“控制 Anthropic”扩大到“定义整个行业应该怎么玩”

RACE TO THE TOP

对anthropic来说：我们知道怎样治理公司、怎样治理模型，那么下一步自然是：

怎样让其他人也这么做?

Anthropic 将这套策略称为 **Race to the Top**，即“向上竞争”。与传统商业竞争中为了抢占市场而不断降低成本、牺牲安全标准的 Race to the Bottom 相反，Anthropic 希望 AI 公司在竞争中不断提高安全标准，将安全变成竞争优势。

一家企业首先自己定义什么叫 AI Safety，然后按照这套定义建立内部制度。接下来，它希望竞争对手采用相似的制度，最后希望政府将其中一部分写进法律。

如果这套机制成功，Anthropic 并不需要持有 OpenAI、Google 或 Meta 的股权，也不需要进入它们的董事会，就能对它们应该如何开发 AI 产生影响。

这就是 Standard-setting Power：不必控制每一个参与者，只要能够影响所有参与者共同遵守的规则。

另一个被广泛吐槽的逻辑

因为 AI 太危险，所以不能让不负责任的人率先开发出来。

因为不能让不负责任的人率先开发出来，所以 Anthropic 必须保持竞争力。

因为必须保持竞争力，所以 Anthropic 需要持续扩张算力、训练更强大的模型。

安全与加速并不必然矛盾，但在这套逻辑里，二者可以相互提供理由。至于什么时候应该真正停止、由谁决定停止，就成为了一个更加困难的问题。

这个问题在中国身上体现得尤其明显。

2025 年 2 月，Dario 在 ChinaTalk 的采访中解释了他对中美 AI 竞争的看法。他认为，美国的 AI 公司处于同一套法律制度之下。如果政府制定了一项规则，OpenAI、Meta、Google、xAI 和 Anthropic 都必须遵守。即使没有真正立法，监管的潜在威胁也可以促使其他公司提高安全标准。

但是，他认为这套机制无法直接适用于中国。中美之间不存在这样的共同监管机制，他将这种关系描述为一种霍布斯式的国际无政府状态（*Hobbesian international anarchy*）。

因此，Anthropic 的 Race to the Top 实际上包含两个不同的策略：

在美国及其他可以协调监管的地区，尽可能通过企业承诺、行业标准和政府监管提高安全要求。

对中国，则首先通过芯片出口管制、限制先进技术获取和维持算力优势，争取美国能够安全开发 AI 所需要的时间。

2026 年，Dario 进一步明确了这套立场：美国及其盟友应该掌握先进 AI 的领先优势，从而拥有制定国际规则和规范的能力。

但是这个立场有个大前提：每个人都相信EA的AI Safety。
如果有一批AI从业人员从来没听说过，或者听过之后也不相信EA那套东西呢？

EA 的 AI Safety 传统，有着相当特殊的知识谱系。从 Nick Bostrom、Eliezer Yudkowsky、FHI，到 Open Philanthropy、80,000 Hours 和后来的 Frontier AI Safety 机构，这些组织花费了十几年时间，将 AI existential risk 从哲学讨论发展成了一套具有独立概念、研究议程、职业路径和资助体系的知识网络。

对于长期生活在这张网络中的研究者而言，AGI、ASI、x-risk、alignment、takeoff、loss of control 都是很自然的概念。他们从进入大学开始，就可能在课程、论文、社交网络、实习、研究资助和职业选择中不断遇到这些问题。

但中国 AI 从业者的成长路径并不完全相同。

中国当然存在 AI Safety 研究，也非常重视安全。不过，这与 EA/x-risk 成为整个 AI 行业默认的世界观，是两件不同的事情。

实际上即使在美国，EA的这套世界观也从未获得一致认可。

2026 年 10 月，在 Walter Isaacson 主持的《The Builders》访谈中，Thiel 认为，Yudkowsky 对 AI 灾难的描述非常详细，对如何建立全球监管机制却含糊得多。

Thiel 试图提醒人们关注另一种存在性风险：**一个以防止灾难为理由建立起来、最终又无法摆脱的全球极权体系。**

他还提出了一种关于 Anthropic 企业文化的解释：当 Dario 不断告诉员工 AI 可能非常危险时，这套叙事也在告诉他们，自己正在世界上最重要的公司工作。**AI Safety 因此不仅可以是一项研究任务，还可能成为组织认同与员工激励的重要来源。**

2026 年，教皇 Leo XIV 在关于 AI 的通谕《Magnifica Humanitas》中提出了另一套观点。他主张更加严格的公共问责。**但他强调，机器是工具，不是人；更重要的是，不能由少数控制数据、基础设施和算法的企业，独自决定整个人类社会应该接受什么样的道德标准。**

李录在 2026 年 10 月哥伦比亚大学的演讲中，则从另一条路径提出质疑。他将硅谷从递归自我改进、AGI 到 ASI 的宏大叙事，戏称为邀请投资者参加一场**“神的 IPO”。他认为 AI 是具有巨大潜力的技术工具，但人类并不应该接受一个由少数人制造、掌握和解释的人造神明来支配社会。**

这三个人的思想背景、对 AI 的态度和关注的问题完全不同。Thiel 担心阻止 AI 的权力，教皇担心技术权力与道德权威集中，李录则质疑将技术进步叙述为人造神明的方式。他们并不是同一个阵营，也不共享同一套 AI 风险判断。

然而，他们都提供了不同于 EA/x-risk 网络的观察角度。

在过去十几年里，EA 网络成功地将一套原本相对小众的 AI 存在性风险理论，逐步引入美国的顶级 AI 实验室、慈善资本、学术研究机构与政府政策讨论。因为民主党在24年没有胜出，anthropic统治美国

AI行业的计划从目前来看进行的不是太顺利，但还是可以看到，Anthropic已经在 Frontier AI Safety 的议题设置与制度设计中取得了相当大的影响力。

Anthropic 还在努力把这种影响力进一步推进：从企业治理，到模型训练，再到行业标准与公共政策。

但在比赛开始之前，或许还有一个更基本的问题没有解决：

谁来制定比赛规则，谁有资格担任裁判，以及谁能让大家都遵守这名裁判提出的规则？

下一篇介绍美国AI行业里非Anthropic 核心圈的人物和组织。

喜欢作者

anthropic/EA · 目录

上一篇
Anthropic前员工回忆录

下一篇
八一八Anthropic和有效利他主义（EA）3.Anthropic核心圈 番外篇