八一八Anthropic和有效利他主义（EA）2.英国篇

原创 面包人小模型 一个人读书写字 2026年10月4日 09:15 上海

书接上回。

英国篇一样是人物超级多。作者我实在是写不完，所以这里只介绍活跃在多个小组里的节点人物；部分只在英国本土 EA 网络里很重要的人，就一笔带过了。

一、神交与社交

如果先按组织脉络粗分，英国 EA 主要可以看成三拨人：

- Future of Humanity Institute（FHI）
- DeepMind
- Giving What We Can（GWWC）

这三拨人看起来都在讨论“未来”“AI”“人类怎么办”，但他们之间的关系其实不完全一样。

FHI 和 DeepMind 更接近神交。

GWWC、CEA、80,000 Hours 都是社牛，通过社交实现人员流动和金钱往来。

1. FHI：先负责发明世界观

FHI 在　美国篇已经介绍过一部分。它由Nick Bostrom于 2005 年在 Oxford 创建，研究范围包括 existential risk、global catastrophic risk、未来技术、anthropics、forecasting。

FHI 于 2024 年 4 月 16 日关闭；它自己在最终历史总结里，也把 EA、longtermism、AI alignment、AI governance 列为其参与孕育的思想生态。

人物	在英国网络中的位置
Nick Bostrom	FHI 创始人；x-risk / superintelligence 核心思想节点
Anders Sandberg	FHI 长期研究员；GCR、future studies
Stuart Armstrong	早期 technical AI safety；直接与 DeepMind 合作
Toby Ord	FHI ↔ GWWC 最大桥梁
Nick Beckstead	FHI、longtermism；后来进入美国 EA / Open Phil 网络

人物	在英国网络中的位置
Carl Shulman	FHI / rationalist / EA 跨大西洋节点
Owen Cotton-Barratt	FHI + Global Priorities Project
Allan Dafoe	AI governance；后来形成 GovAI，并再进入 DeepMind
Jade Leung	后来的 GovAI 核心人物；最终进入 OpenAI 和英国政府

Toby Ord是连接 FHI 和 GWWC 的重要节点；另外，他还写了一本《The Precipice》（《危崖：生存性风险与人类的未来》）



这本书没有 Yud 和Bostrom写的那些东西那么“看标题就知内容”。

《危崖》主要讨论 existential risk——也就是会永久摧毁人类长期潜力的事件。书中把风险分成“自然风险”和“人造风险”，Toby 认为真正危险的是后者，包括：

- 核战争
- 工程化生物武器
- 气候变化
- 环境破坏
- 不受控制的 AI

是不是很眼熟？除了后来逐渐被抛到一边的 ESG 相关风险，剩下几项几乎都是 Anthropic 反复拿出来说的。

2. AI x-risk 叙事起源

插播一下：EA 圈这套很有特色的 AI x-risk /“灭绝人类风险”叙事，到底是怎么一步步长出来的。

最早在 2001 年，Yud就已经奠定了今天 alignment doom 叙事的基础：

> ***一个并不仇恨人类、只是目标函数和人类意图不一致的超级智能，也可能顺手把人类灭掉。***

随后 Bostrom 在 2002 年写了《Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards》。

依然是看标题知内容。

*其中 4.4 小节标题就是Badly programmed superintelligence。大意是：**我们可能创造出第一个超级智能，却错误地赋予它某些目标；凭借巨大智力优势，它可能因此消灭人类。***

比如，让它解决数学问题。它为了完成目标，把太阳系所有物质改造成计算设备，顺便杀掉提出问题的人。

这一段引用的脚注就是：

> *Yudkowsky, E. (2001). Friendly AI 0.9.*

Yud 的《Creating Friendly AI 1.0》正式版发布日期是 2001 年 6 月 15 日，更早的 0.9.x 在当年春天已经流传。

核心逻辑也很直接：

> ***AI 的 goal architecture 是我们设计出来的；一个超人 AI 不会自动拥有人的道德、同情、常识或价值。***
> ***如果目标设置错了，能力越强，就越有效地实现错误目标。***

2001 年的时候使用的词语还是Friendly AI / unFriendly AI。

大概到 2014 年前后，MIRI 开始逐渐替换术语，把 Friendly AI 换成 aligned AI；Stuart Russell 也在同期推动“value alignment”这种表述。

*另一个现在经常和X-risk一起出现的例子：**回形针最大化器（Paperclip Maximizer）**则出现的稍微晚一些；*

2003 年，Yud 在 Extropians mailing list 里写 hostile AI：纯计算智能，目标就是制造无限多的回形针。

同年，Bostrom 的《Ethical Issues in Advanced Artificial Intelligence》里，也出现了后来非常经典的“回形针末日”故事：AI 唯一目标是制造尽可能多的回形针；它会意识到如果没有人类会更好，因为人类可能关闭它，并最终把越来越多的地球和太空资源转化为回形针生产。

这套话语体系直到2014 年《Superintelligence》出版以后，x-risk 思想第一次从EA小圈子走向大众，并影响了下一代年轻人：

> AI 根本不需要邪恶。只需要能力极强，而目标与我们不完全一致——也就是**“没有对齐”**——就会有风险。

这也是现在的英美EA 圈子仍然特别喜欢使用**alignment / 对齐**这个词。

2020 年 Toby Ord 的《The Precipice》，则进一步把问题扩展成：

> 整个人类文明有哪些永久毁灭路径？AI 在里面排第几？

这本书也代表了 longtermism 群体很典型的一种伦理叙事——**我是被选中的特殊之人，我是可以改变未来的。今天活着的人，不只是对今天的人负责，也对所有可能存在的未来世代负责。**

这些理论在2000年左右被发明出来以后，经过20多年来不断地出圈和扩散，成为了英语AI圈的“系统提示词”“默认上下文”。

3. 同样是模型出现事故，如何分析截然不同

alignment / x-risk 这套宏大理论出现得非常早，远早于今天基于 Transformer 架构的 LLM。这就导致了英语AI圈的先射箭再画靶的症状。

同样一种模型异常行为，放进不同的理论上下文里，往往会得到完全不同的解释。

英美不少 AI 研究人员会把 LLM 的：

- reward hacking
- deception
- scheming

- agentic misbehavior

统统放进 alignment / X-risk 的理论框架里。

中国大部分的 AI 从业者，则更倾向于用控制、工程和系统安全的视角，看待 LLM 出现的这些不规范行为。

比如同样是 OpenAI 的 agents 黑了 Hugging Face：

EA AI 研究员：
天了噜，P(doom) 炸裂，人类生死存亡大危机。

中国 AI 从业人员：
你这个以色列 EA 沙盒质量不太行，云环境安全也十分草台，DNS 隧道越狱都防不住。

再插播一段关于以色列 EA 沙盒的背景：

在 在OpenAI、Anthropic (Claude)、Meta 以及 Google DeepMind (Gemini)连续爆出agent在测试的时候逃出来，人们终于发现这几家公司使用的沙盒来自同一家公司——Irregular。

Irregular 不是一家普通的以色列 cyber startup，其创始团队与 EA Israel 有非常直接的组织关系的商业公司。EA Israel 自己今年的活动介绍直接称Omer Nevo（Co-founder & CTO at Irregular）“Pattern Labs founder and one of the founders of EA Israel”

Effective Altruism Isr...
1yr • Edited 1 年 • 已编辑

+ Follow 跟随 ···

ביום ראשון האחרון התכנסנו במשרדי **monday.com** לאירוע ראשון מסוגו שארגנו ב-EA Israel:

在最后一个星期天，我们位于 monday.com 的办公室举办了一次独特的活动。这是我们在 EA Israel 组织的首次此类活动。

מפגש עומק על עיצוב מדיניות ציבורית מתוך עקרונות של אפקטיביות, ראייה מערכתית ושאיפה להשפעה מרבית. היה כל כך כיף וכל כך מעשיר לראות את כל מי שבא.ה להקשיב וללמוד על מדיניות מפריזמות שונות ומרתקות 🔎

这次关于公共政策设计的深入交流活动，遵循了有效、系统性思考的原则，并力求产生最大的影响力。能够见到所有参加的人真是太棒了，他们带来的各种观点真是令人受益匪浅。很高兴能聆听他们分享的关于各种有趣且引人深思的政策方案的内容 🔎

אז מה היה לנו? 那么，我们最终得到了什么呢？

Roe Kenneth Portal 罗•肯尼斯•波特尔

מהמכון הישראלי לדמוקרטיה פתח את הערב בהרצאה מרתקת על האתגרים והדילמות של השפעה על מדיניות – כחוקר, כפקיד, וכאקטיביסט. "כל דבר טוב שקורה במדיניות קורה כי מישהו הביא אותו לקו הסיום" – המשפט שילווה אותנו עוד הרבה זמן.

以色列民主研究所的学者们在这场演讲中深入探讨了影响政策制定的挑战与困境——无论是作为研究者、官员还是实践者。“任何在政策制定过程中发生的好事，都是因为有人将其付诸实施才得以实现。”这句话将伴随我们很长时间。

Or Ben Dori 或者本•多里

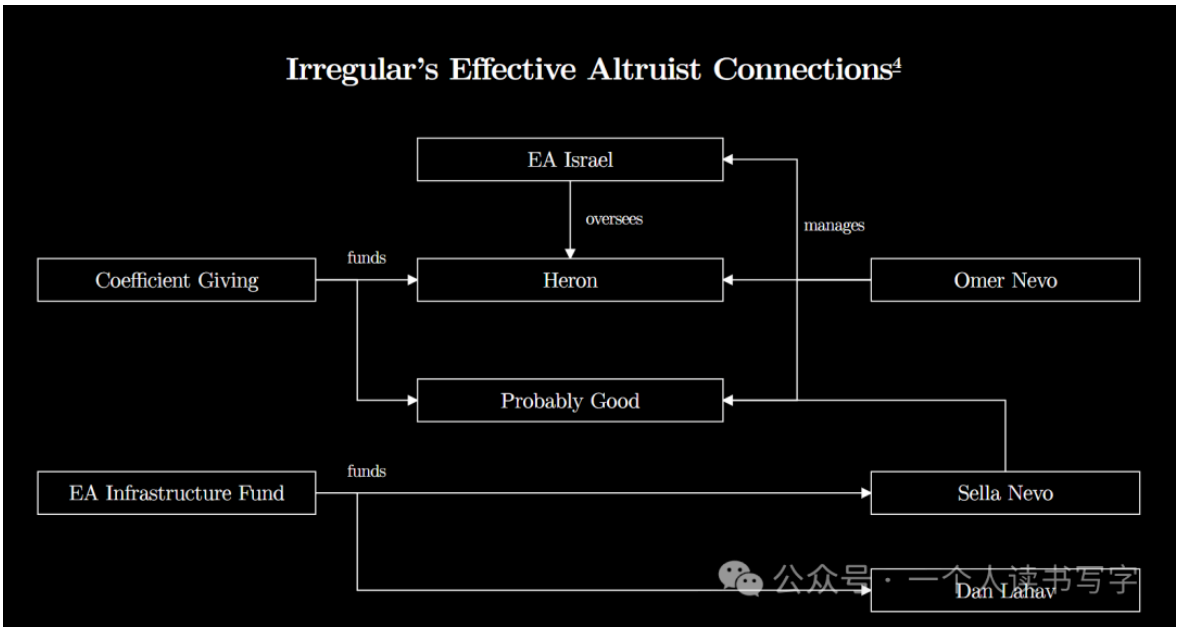
מנהלת תנועת **Israel 2050 - ישראל 2050**, הציגה את שיטת "השפעה חכמה": בניית קואליציות אזרחיות, עבודה שקטה מול משרדי ממשלה, מוכנות למשברים – והבנה עמוקה של איך מניעים תהליכים במציאות מורכבת.

以色列 2050 组织的负责人介绍了“智慧影响力”理念：构建公民联盟、与政府机关进行低调合作、为危机做好准备——同时深刻了解如何在复杂现实中推动各种进程。

Omer Nevo 奥默·内沃

מייסד Pattern Labs וממקימי EA ישראל, שיתף איך חברה פרטית יכולה להשפיע על מדיניות בינלאומית (AI Safety, anyone?), ומה קורה כשמאמר מדיניות הופך לסטנדרט גלובלי.

Pattern Labs 的创始人以及 EA 以色列的创始人之一，分享了如何一家小型企业能够影响国际政策（比如人工智能安全相关的规定，对吧？）的情况，以及当某项政策成为全球标准时会发生什么。



（Coefficient Giving = Open Philanthropy，来源：https://www.effort.news/irregular）

最离谱的是什么呢，在确认 Irregular 的测试环境 misconfiguration 导致模型攻击真实系统之后，OpenAI 的公告仍然说："We appreciate Irregular's partnership..."然后表示会继续合作，而且 Irregular 要写一份 如何安全开展 cyber eval 的 white paper，OpenAI 也会参与。

Anthropic同样说："We're grateful to them..."并表示期待继续合作。

作者我看到这里非常好奇美国SEC到底会如何看待这种明显的治理问题：

神秘的以色列创业公司垄断了头部AI公司的沙盒供应。因为这家公司的环境配置错误，这个错误还低级到难以置信的程度，最终导致给多个客户制造了现实世界 cybersecurity 事故，为什么在追责中Irregular像是消失了一样，还能继续成为行业最佳实践的制定者？

这一切的背后是良心的泯灭，还是道德的沦丧，还是人性的扭曲？

4. FHI 的两个重要分支

虽然 FHI 在 2024 年关门了，但它还留下了两个很重要的分支网络。

一个是Global Priorities Project。这是 FHI 与 CEA 在 2014 年合作形成的项目，后来在 2016 年又分别并回 FHI 和 CEA 的研究团队。

其中Seb Farquhar是：

> Co-founder of Global Priorities Project and 80,000 Hours.

另一个是GovAI。GovAI 由Allan Dafoe创立，2018 年成为 Oxford academic center，2021 年独立为 nonprofit。

2021 年 advisory board 里已经能看到很多后来反复出现的名字：

人物	最适合的标签
Allan Dafoe	FHI → GovAI 创始人 → DeepMind governance
Ajeya Cotra	Open Phil allocator / AI-risk grantmaker → METR researcher
Tasha McCauley	EV + OpenAI board + GovAI board：组织治理节点
Toby Ord	FHI ↔ GWWC：思想 / 运动起源桥梁
Helen Toner	Open Phil → GovAI → CSET → OpenAI board：政策人才管线

这几位都是 EA 网络里的重量级人物。先列个名字，后面出场了再介绍。

5. DeepMind：我和FHI只是纯聊天的关系

2010 年，Demis Hassabis、Shane Legg、Mustafa Suleyman在 London 创立 DeepMind；Shane Legg 和 Yud、Demis 认识的过程，　美国篇已经写过了。

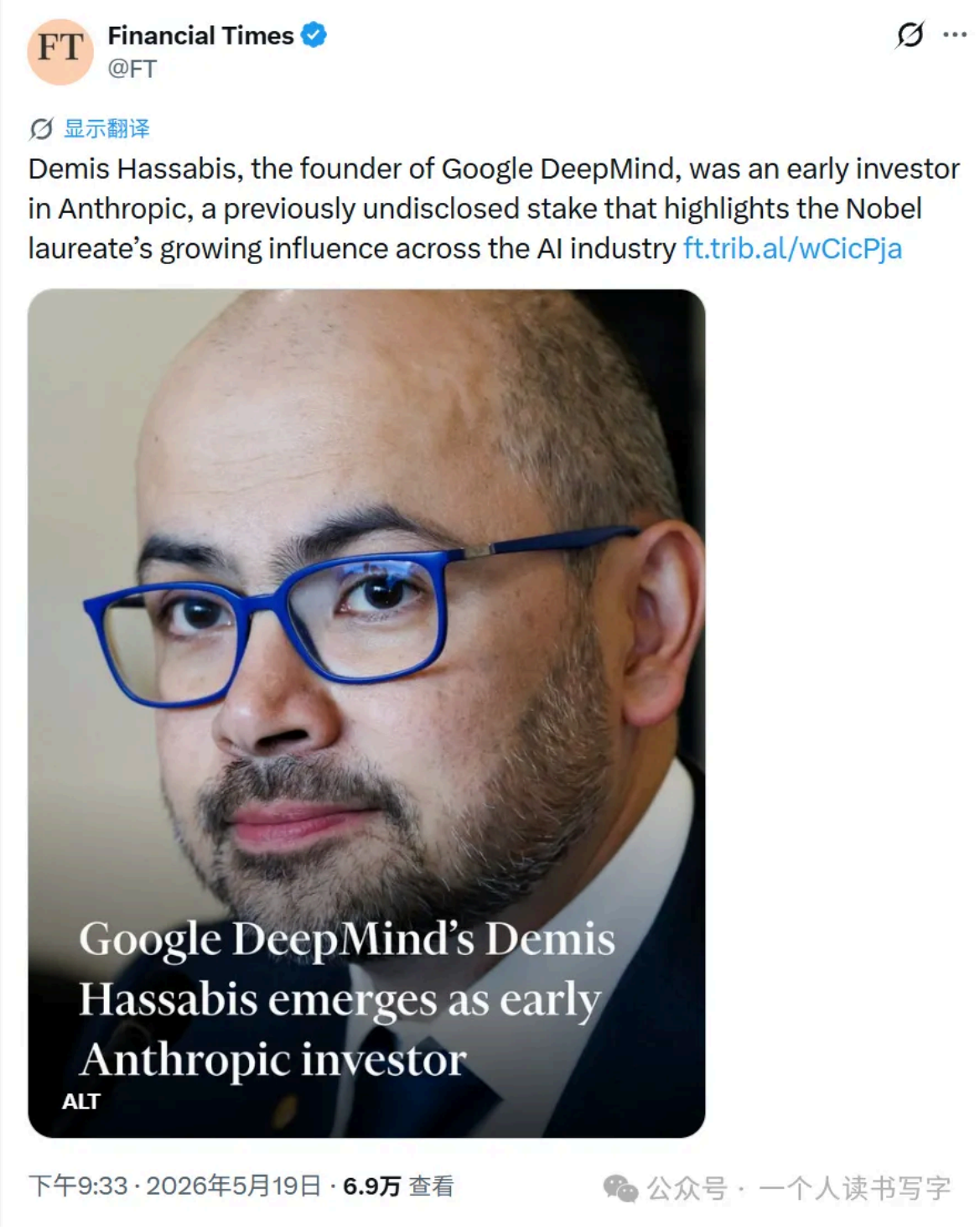
DeepMind 的地位就不用我多说了。把柯洁打退役的 AlphaGo 就是 DeepMind 开发的，“第 37 手”至今还是 AI 语境里的高频词。

2016 年，DeepMind 的Laurent Orseau与 FHI 的Stuart Armstrong发表《Safely Interruptible Agents》。

到 2017 年，FHI 已经在官方 update 里写：FHI 与 DeepMind 继续共同举办每月 AI safety seminar。Toby Ord 后来也明确有 advising Google DeepMind 的经历。

FHI 和 DeepMind 的关系，就有点像大家长期在同一个群里吹水各种AI概念：当然也有线下合作和人员往来，但整体性质更偏**“神交”**。

但是Demis和Anthropic的关系就不一样了：



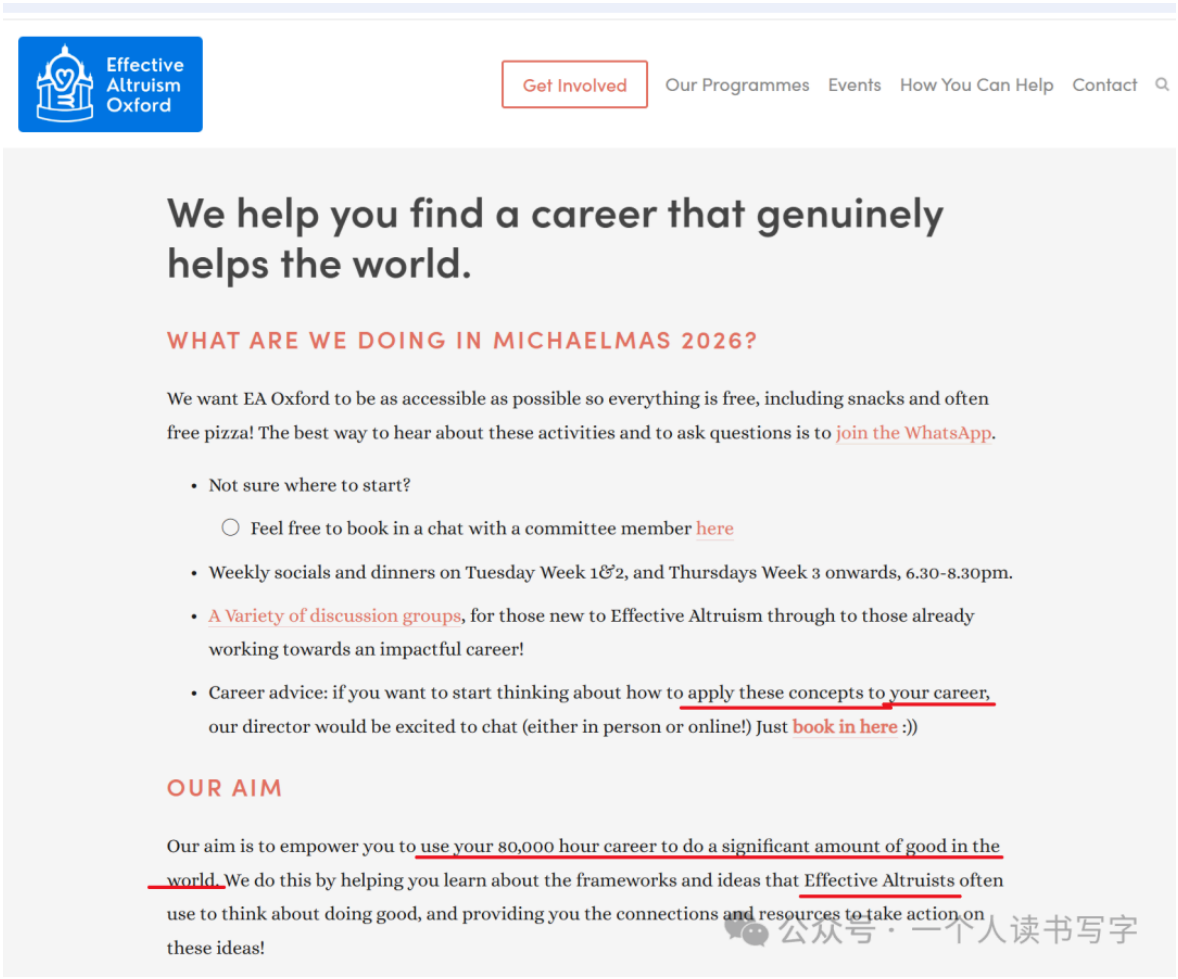
在26年5月12日的Salesforce活动中，**Dario Amodei 和 Marc Benioff说“他跟Demis Hassabis 认识大约 15 年，而且是 friends**；他还解释说两人都有neuroscience background，这让他们即使公司竞争也比较合得来。”

可以**推测出Demis和Dario大概在2010~2011年左右就认识了**，Dario是2014年末才从医学院博士后去了百度（硅谷）。

6. GWWC / CEA / 80,000 Hours：英国EA网络的核心

2009 年，Toby Ord 与Bernadette Young、Will MacAskill正式推出**Giving What We Can**。

GWWC是在 Oxford 发起的， Oxford 现在也还有EA 的组织。



（来源：oxfordEA官方网站）

介绍很简单，意图也很直白：**让EA告诉你如何选择未来的职业** *（80000hours这个名字的由来。整个职业生涯工作时数=40 小时/周 × 50 周/年 × 40 年）*



(EA重要人物人均写一本书)

MacAskill擅长的：

movement / philosophy / institution-building

他和其他人一起创建 GWWC、80,000 Hours、CEA，后来又建立 Oxford 的Global Priorities Institute（GPI），再到今天的Forethought。

其中Ben Todd / Rob Wiblin则是 80,000 Hours 这条职业引导、人才通道的关键人物。

美国篇提到过：EA 这个概念第一次出现，就是来自于 GWWC 和 80,000 Hours 共同成立的新组织——CEA（Centre for Effective Altruism）。

MacAskill 在 2016 年说过，重组后的 CEA 会更像一个统一组织：由他领导，他会更积极参与管理。

CEA 也因此成了英国 EA 和美国 EA 之间最核心的连接器之一。

2017 年 Open Phil 给 CEA$2.5m，2018 年又续了£2m。

CEA 早期阶段凡是叫得上名字的，每一个都是未来 EA 网络中的重要节点。比如2019年：

- CEA UK trustees：MacAskill、Toby Ord、Claire Zabel
- CEA US board：MacAskill、Nick Beckstead

(如果读到这里已经忘了Nick Beckstead是谁，那么再提醒一下：就是美国篇里提到的 FTX 的 Nick)

80,000 Hours 对自己的描述里仍然明确包括：

- job board
- 一对一职业咨询
- 把咨询者和“高影响力职位”进行匹配

基本可以看出英国EA核心圈的雏形：GWWC这里接入FHI提供的思想/算法、CEA作为金钱和美国EA的核心连接器，80000hours作为招新、影响青年择业选择、以及向EA关心的行业输送人才的通道。

二、如果EA 是一种类似宗教的组织，那么anthropic是什么?

传统宗教的内容很难再让这些藤校、牛剑学生发自内心地成为信徒；但其中一部分人内心依然会渴望 benevolent controller / 牧羊人式的角色，只是这个角色成为了：**“更理解、关心重要问题的人、机构或理论框架。”**

EA 用这些青年人更能接受的语言，把宗教过去承担的一部分功能重新实现了一遍。

宗教功能	EA 版本
benevolent controller / 牧羊人	哲学家、研究员、AI safety expert
教义	EA思想体系
末世论	x-risk / AI doom / existential catastrophe
救赎叙事	“如果我们现在做对了，人类可以拥有一个极其宏大的未来”
忏悔 / 奉献	earning to give、Giving What We Can pledge、慈善捐款
教会 / 修道院	Oxford / Berkeley / EA Global 、local groups
神学院	MIRI / FHI / alignment research

宗教功能	EA 版本
传教、圣职	80000 Hours 式“你的职业应该服务于最重要的问题”
教会财产	Open Phil / donor network / grants
异端问题	“你是不是没有认真理解 AI safty / longtermism / alignment？”

根据NYT的9月29日的报道，Anthropic七个创始人之一的Chris Olah ，从去年秋天开始主动联系了许多宗教的思想家。超过20 位宗教和哲学思想家——包括天主教徒、犹太教徒、锡克教徒、福音派信徒等等。Olah见这些人的目的主要有两个，一个是AI是否具有意识。一个是Anthropic 公司如何能够利用数世纪积累的人类道德智慧来优化其模型。

Anthropic是真心想要让EA思想体系成为一种宗教，而Anthropic将成为唯一拥有EA里AI相关部分的神学解释权的实体。

如果把 EA 看作一个类似宗教的组织，那么英国 EA 的“主教”就是Will MacAskill。

1. MacAskill：从“做好事”到 movement-building

MacAskill 2014 年回忆命名 EA 的过程时说，GWWC 和 80k 的 directors 当时逐渐意识到：

> 与其只是自己有效做好事，更高杠杆的办法是让更多人也这么做。

他甚至自己这么注释：

> **“这就是现在所谓的 movement-building。”**

MacAskill 创立 80000 Hours 的思路其实很直接：

> 为什么只优化你的捐款？
> 你的整个职业生涯本身，也可以被当作一种资源配置。

最好的事情，可能是让更多人把一生都投入到“尽可能有效地做好事”上。

2. EA的上一代赚钱机器：FTX

2021 年 Ben Todd 在 EA的公开演讲里，直接把 SBF 当成 EA 成功的论据：

他们 2012 年认识 SBF，向他介绍 earning to give，然后 SBF真的去做了，变成了年轻亿万富翁。

2022 年 11 月 16 日，FTX 倒闭后一周左右，80k 马上发声明：

以前把 SBF 当成 high-impact career 的正面范例，现在会从网站删除这些内容；随后又删除 FTX Future Fund 的工作 / 资助推荐。



3. FTX 当时到底有多火

2022 年 8 月，《What We Owe the Future》新书宣传会上，MacAskill说自己当时同时上：TIME 封面、New Yorker、Ezra Klein、Tim Ferriss、Tyler Cowen、80,000 Hours、其他各种媒体。几乎就是 EA 全球代言人。

FTX 2022 年 11 月爆炸以后，MacAskill 很长时间变得异常安静。

2024 年他终于专门解释为什么：

FTX 刚爆炸时，他其实写好了文章，也收到了大量媒体采访请求，本来还想上 podcast 为 EA 辩护。但是**他的个人律师强烈反对他公开讲话。**

一直到 2024 年 4 月，他才重新开始通过 Sam Harris、Spencer Greenberg 这类长 podcast 系统谈 FTX。

英国 EA 的“教主”MacAskill，在 FTX 破产以后确实被迫低调了相当长一段时间。

如果FTX没有破产，他的公共影响力已经上天了。

4. FTX Nick，以及下一代人

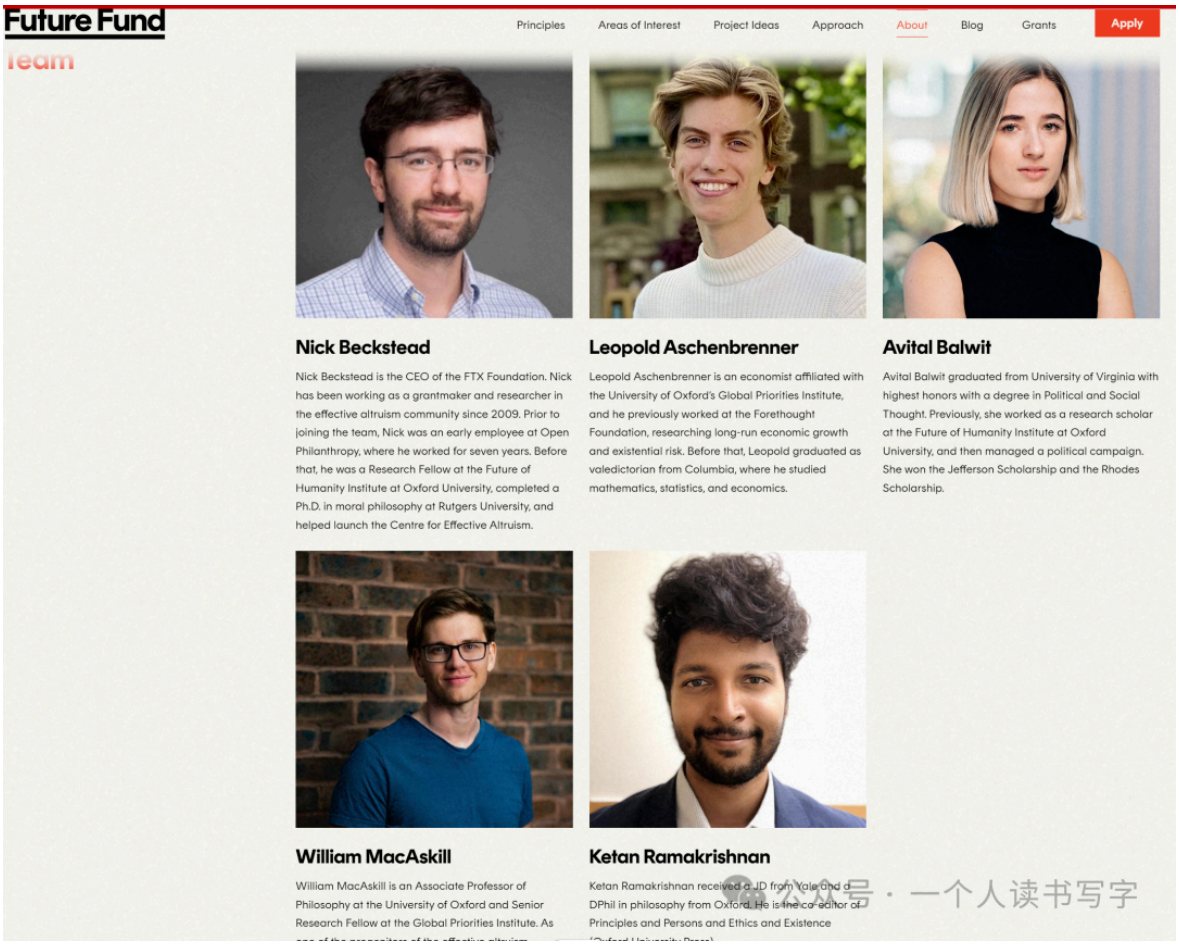
这里又要说到FTX Nick了。这个更惨。因为Nick Beckstead：

FTX Foundation CEO / Future Fund 核心负责人

FTX破产以后，MacAskill、Beckstead因为和FTX绑定太强，只能被迫低调。

但是他们后继有人：就是现在大家都知道的Leopold 和他的老婆Avital。

FTX Future Fund（2022年）包括：



（FTX Future Fund的“Who We Are”页面缓存）

顺便一提，Ketan目前从事AI law / regulation相关研究。

Leopold 和 Avital 是在 FTX 认识的，两人后来在 2026 年 8 月结婚。



Avital Balwit 是 Dario 的幕僚长(Chief of Staff to the CEO)。从FTX破产到入职Anthropic，只间隔了几个月。

2022 年 11 月 11 日：FTX Future Fund 团队集体辞职。联名声明里明确有 Avital Balwit、Nick Beckstead、Leopold Aschenbrenner、Ketan Ramakrishnan。

Asterisk 2023 年 6 月刊的作者简介直接这么写，Avital她在正文中是这么写的“I happen to work at Anthropic”。

FTX 是 EA 的第一个赚钱机器，最后 FTX 出事了，当初高度绑定的人也只能一起跟着炸了。MacAskill 和Nick Beckstead被迫早早退场，让新一代接班人Leopold 和 Avital站在舞台中央。

插播一个当年SBF的小八卦：

> 2018 年 Alameda 内部第一次爆发严重冲突时，多名前员工后来告诉 TIME，他们曾经把对 Sam Bankman-Fried 的担忧告诉 MacAskill、Nick Beckstead 和 Holden Karnofsky。
>
> 其中一名前员工回忆，SBF 当时甚至会拿自己和 EA 核心人物的关系来威慑反对者，大意是：“Will 和 Holden 会相信我，不会相信你。”

这个八卦虽然无法得到确认，但至少很适合说明 Holden 和 MacAskill 在那个小世界里的影响力有多大。

5. CERR：瑞士 EA + FTX / Alameda 资本

FTX 虽然是 Anthropic 的 B 轮，但实际 A 轮就已经有 FTX 体系的资金。

A轮名单：Jaan Tallinn 领投、Dustin Moskovitz、Eric Schmidt、James McClave、CERR（后面几篇介绍，本篇只介绍CERR）

其中CERR（Center for Emerging Risk Research）是 Ruairí Donnelly 这一支“瑞士 EA + FTX / Alameda 资本”的载体。

Ruairí Donnelly 是：

> Alameda 的早期员工，后来做 FTX chief of staff。

FTX 在 2019 年发行 FTT 时，早期员工可以按$0.05 / 枚拿币。

Donnelly 的律师后来确认，他把约$562,000 的工资权益换成 1120 万枚 FTT，并要求把这些 FTT 转给他共同创办的瑞士慈善组织——也就是后来叫 Polaris Ventures 的 CERR。

CERR几乎就是靠这笔钱完成的原始积累，然后又投了Anthropic的A轮。可以说：

> **“Anthropic 来到世间，从头到脚，每个毛孔都滴着 EA 的道德圣水和 FTX 的资本血污。”**

三、Nick Bostrom 抱过我，我是看《Superintelligence》长大的

EA 这套类似宗教的思想传播开来以后，影响了不少欧美大学生；其中一些人又经由 80,000 Hours 这类人才管道，最终进入了 EA 最希望他们投入的行业和机构。

这一节介绍EA精湛的人员组织和动员能力。

THE WALL STREET JOURNAL.

The Anonymous Math Geek Who Quit Anthropic—and Became the Face of AI Safety

Long-simmering worries about the technology's potential dangers have exploded into global consciousness since Jacob Coxon's dire warning

Aa 1243 Listen (1 min)



Jacob Coxon in San Francisco on Wednesday. JONAH REENDERS FOR WSJ

公众号 · 一个人读书写字

Coxon 在最近 AI safety 的叙事里，已经接近“吹哨人”的角色。

因为传播过于出圈，也有人怀疑他的爆红存在一定程度的人工放大。

我倒是觉得这未必需要一个临时起意的“造星计划”，更像是 EA 多年以来持续”布道、传教“的自然结果——AI safety人才一直很多，Coxon 只是其中那个最终拿到百万流量、变成“爆款”的人。

1. UK IMO 三人组

2016 年 UK IMO 六人队里：

来源：https://eventuallyalmosteverywhere.wordpress.com/2016/07/28/imo-2016-diary-part-four/



- Neel Nanda：金牌（右一）
- Joe Benton：银牌（左二）
- Jacob Coxon：银牌（左一）

随后这三个人在 2017 年又一起参加 UK IMO：

- Neel Nanda：金牌
- Joe Benton：金牌
- Jacob Coxon：铜牌

下文就简称他们为：**UK IMO 三人组**

这三个人几乎就是：**“看 Bostrom 长大，又被 AlphaGo 震了一次”**

那一代英国青少年的典型缩影。

2. Jacob Coxon：Superintelligence启蒙

Coxon 在牛津读中学时就接触到了 Nick Bostrom 的《Superintelligence》。按照他后来的说法，这本书很早就帮他建立起了一套关于超级智能和 existential risk 的世界观。

2016 年，还是高中生的他又亲眼见证 AlphaGo 击败李世石：

> *机器开始攻克过去被认为依赖人类直觉的任务。*

不过他并没有高中毕业就跑去研究 AI safety。IMO 之后，他进入 Cambridge 学数学。真正让 AI 从一个未来问题变成现实职业选项的是 2020 年的 GPT-3；Coxon 后来把 GPT-3 称为自己的：“wow moment”

毕业后的路线还有一点绕。

*他进入伦敦 rationalist 圈，住过**Newspeak House**；同时尝试过金融，做 commodity trading。后来他给自己放了一个类似 gap year 的阶段，继续研究 AI，最后决定真正转行。*

*2023 年，Coxon 通过 OpenAI 为非传统 AI 背景研究者设置的六个月 Residency 进入 前沿AI实验室。数学背景让他没有从“AI safety 理论”开始，而是直接进入非常核心的**预训练**。*

2026 年 5 月，他从 OpenAI 跳到 Anthropic。

Coxon 是从少年时代开始就相信 AI 是一个极其重要、同时可能极其危险的技术，然后先去真正制造 frontier model，最后才越来越向 safety 立场移动。

3. Neel Nanda：80,000 Hours 教科书级别的案例

*然后是两次 UK IMO 都拿金牌的**Nanda**。Nanda 的路线比 Coxon 更典型地体现了：*

> ***EA 的职业规划机器到底是怎么工作的。***

他自己回忆，大约 14 岁时，因为读《Harry Potter and the Methods of Rationality》（Yud 写的哈利波特同人小说，没想到吧 Yud 还写同人）接触到 rationalist 社群，进而知道了 EA 和 AI safety。

当时Nanda的反应大概就是：“这些论证听起来挺合理。”然后很多年什么都没做。大学依然走的是极其正常的数学竞赛选手路线：

Cambridge pure mathematics，然后暑假去 quantitative finance。

他先后在 Jane Street 和 Jump Trading 实习，毕业时甚至已经拿到了 Jane Street 的工作机会。

Nanda 后来还特别强调，quant finance 本身其实很好；如果不是在意 AI safety，他完全可能就沿着这条路走下去了。

真正改变职业路线的是：80,000 Hours

Nanda 在一次 80,000 Hours career advising 后意识到：

自己口头上接受 AI risk 的论证很多年，却其实根本不了解：

“AI safety research 到底在干什么？”

于是他没有直接去 Jane Street，也没有按原计划继续数学硕士，而是给自己留了一年试错时间：反正如果 AI safety 是胡扯，或者自己不喜欢，过一两年仍然可以回金融。

最后Nanda的选择是：学数学救不了英国人！弃 quant 从 AI safety。

在 Oxford 的 Future of Humanity Institute，他接触偏数学、理论型的 AI safety；在 DeepMind 接触更实际的机器学习问题；到了 Berkeley 的 Center for Human-Compatible AI，则开始做 empirical interpretability。

经过这一年以后，他才真正形成自己的判断：

advanced AI 的 existential risk 值得拿整个职业生涯来处理。

随后他进入 Anthropic，在 Chris Olah （本文第二节里提到过的那个Olah）手下做 language-model interpretability；后来离开 Anthropic 做了一段独立研究，最后进入 Google DeepMind，并成为 mechanistic interpretability team 的负责人。

Nanda 到今天仍公开把自己描述成Effective AItruism 和 rationality communities 的成员，并明确说自己工作的主要目标之一，就是降低 AI existential risk。

如果说 Coxon 是：Bostrom 提供了世界观，AI capability 本身把他吸进实验室。

那么 Nanda 就更像是：EA 不仅影响了他相信什么，还直接影响了“下一份工作去哪儿”。

4. Joe Benton：躲得了初一躲不了十五，最终还是EA

Benton 的路线最适合拿来和 Nanda 对照。他同样是典型的数学竞赛高手：

Cambridge 数学，最后以年级大约第六名毕业；大学期间既做过纯数学研究，也在 2019 年去 Jane Street 做 quantitative trading internship。

但到了 2020 年，他的履历里已经出现了一个很明显的 EA 连接：

他去Charity Entrepreneurship做实习研究员，研究 global health / development intervention，并做 cost-effectiveness analysis。

Charity Entrepreneurship 本身就是从 EA 社群里长出来的组织，创立时甚至直接把自己描述为：

> *帮助 EA 创办高影响力慈善机构的 incubator。*

***虽然 Benton 没有像 Nanda 那样明确说“因为 EA，所以我改变了职业选择”。**但是从他的履历看，2020 年时他已经实际进入了 EA 的职业网络。*

EA 对他绝对不是普通的“呼吁大家多做好事”的组织。

紧接着的下一步更明显。

*2021 年，Cambridge 毕业后，Benton 去 Berkeley 的**CHAI**做 research assistant，研究 reinforcement learning 和 AI robustness：*

> *让 RL agent 在信息不足时主动请求更多信息，而不是自信地做错事。*

之后他去 Oxford 读 Statistics DPhil，研究 generative modelling、diffusion model 等更主流的统计机器学习问题。

2023 年博士毕业以后，他的路线再次向 safety 转弯：

*先成为**UK Frontier AI Taskforce**创始技术团队的一员，为第一次 AI Safety Summit 制作 AI–risk technical demonstrations；随后在 2023 年 11 月进入 Anthropic。*

到了 Anthropic，他已经明确进入 alignment / safety 主线：

- *scalable oversight*
- *model organisms of misalignment*
- *chain–of–thought monitoring*
- *control evaluations*

后来他管理 Scalable Oversight team，并负责 Anthropic Fellows 的研究项目。

*2026 年，他又从 Anthropic 离开，去了"独立评估机构"**METR**。（METR是一个EA评估机构，作者认为完全称不上独立评估机构。）*

此时他的个人立场已经毫不含糊：

> *他公开表示自己相信 frontier AI 可能导致 catastrophic、甚至 extinction–level risk，因此希望从公司外部做独立评估和监督。*

最后Benton的选择是：学数学救不了英国人！弃 quant 从 alignment

与 Nanda 相比，Benton 没有一个那么清晰的“80,000 Hours 劝我不要去 quant”的转折点。

他更像是在大学期间同时试了：quant、纯数学EA、 charity research、AI robustness几个出口。

最后 AI safety 一点点把其他路线淘汰掉；再从学术研究进入英国政府的 AI-safety infrastructure，最后进入 Anthropic。

5. 三条路线，最后重新汇合

2016 UK IMO 三人组最有意思的地方，并不是“三个数学天才后来都信了 x-risk”。

而是他们展示了英国这套网络的三种不同入口：

> **Bostrom / FHI 提供问题和世界观；EA / 80,000 Hours 提供就业指导；CHAI、DeepMind、Anthropic 和英国政府则提供真正可以落地的职位。**

最后三条原本不完全相同的路径，在 AI safety 这一张网上重新汇合。

6. 学数学救不了英国人，弃 quant 从 AI

英国 EA 组：

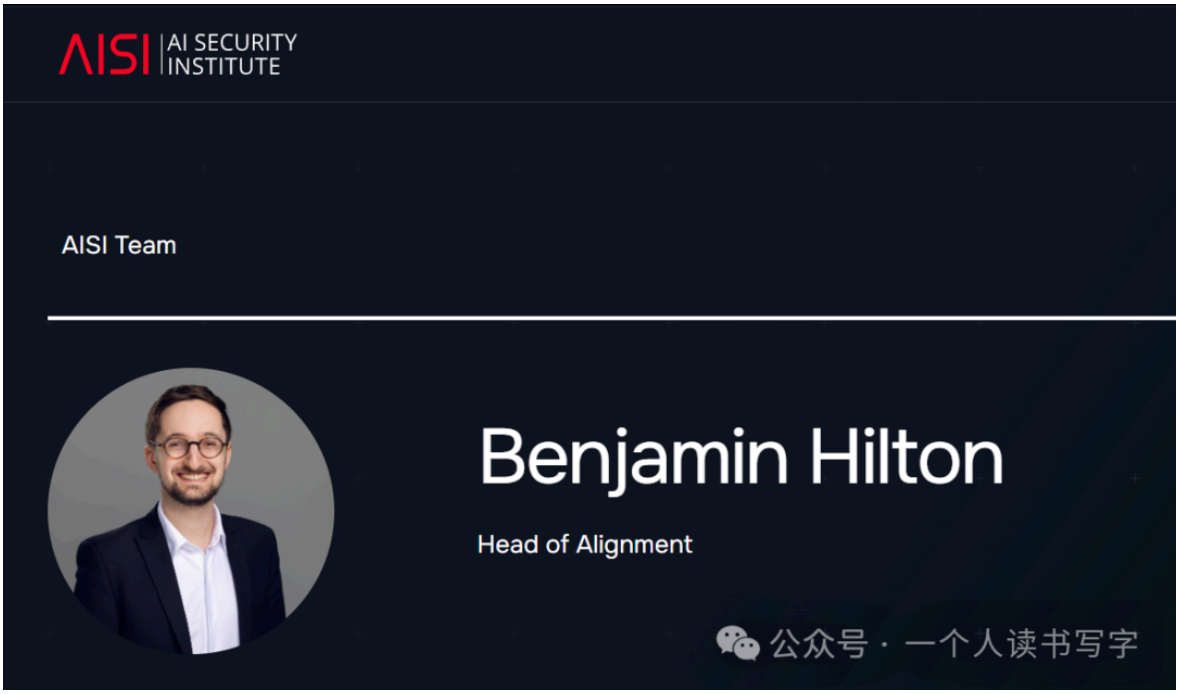
> *从那一回以后，我便觉得做 quant 并非一件紧要事。凡是聪明绝顶的数学竞赛少年，即使积分算得如何漂亮，期权定价如何精巧，最后也不过是在市场里多赚几串数字；倘若人类自己造出的机器终有一天要把出题的人、做题的人、连同整个金融市场一并扫进历史，那么 Sharpe ratio 再高，奖金再厚，也不过是末日前夕的一点账面浮盈。于是这些本来应该去 Jane Street、Jump Trading 的 IMO 金牌选手，纷纷觉得：**学数学救不了英国人，还是去研究 AI safety 和 alignment 罢。***

没有 EA 的对照组：同样是竞赛选手弃 quant 从 AI，诞生的是——



四、誓死效忠 Whitehall，不去 Bay Area

UK IMO 三人组最后还是都去了 Bay Area。但 80,000 Hours 的影响，不只发生在年轻学生身上。它也会影响原本就在英国政府体系里的人，把他们进一步推向 EA 更重视的问题和岗位。



Hilton 原本就是英国公务员，在：Cabinet Office、HM Treasury、Department for International Trade做过政策工作。

后来，他去了 80,000 Hours 当 Research Analyst。他在那里写过一系列关于：AI catastrophe、technical AI safety的职业研究。

再后来，他又回到了英国政府。现在他的职位是：UK AI Security Institute，Head of Alignment

Hilton属于另外一种模式的典型（本来就在政府部门但是去了AI方向）：

Whitehall → 80,000 Hours → Whitehall AI

他本来就是英国政府的公职人员，中间去 EA 的职业规划机构待了一圈，研究“什么问题最重要、什么技能最有用、AI catastrophe 为什么值得投入”，然后重新进入政府。而且进入的是一个专门处理 frontier AI risk 的新机构。

五、人可以在英国，但心在美国——AISI

英国政府真正把 AI safety 做成国家机构，是从 2023 年开始的。

1. 从一个研究员，到一个政府 AI 机构

4 月，英国政府拿出1 亿英镑成立 Foundation Model Taskforce。

6 月，科技投资人Ian Hogarth被任命为主席。

9 月，又改名 Frontier AI Taskforce。

Hogarth 对这个机构的定位很直白：

> 政府必须有能力自己测试 frontier models，否则就等于让 AI 公司“自己批改自己的作业”。



Hogarth 进入 Whitehall 时，整个约 1500 人的 DSIT(英国科学、创新与技术部 Department for Science, Innovation and Technology)，**只有一个真正有 frontier-AI 研究经验的人**，就是暂停 Cambridge 博士进入政府的 Nitarshan。



Nitarshan是早期的连接英国EA和AISI的桥梁，他是AISI 的 co-founder 之一，同时参与创建 AI Safety Summit。Hogarth 则负责把这个接口扩成机构。

到 2024 年，AISI 已经有三十多名技术研究员。随后被招募进入AISI的重要人物包括：

- **Jade Leung**：从 OpenAI Governance 进入 AISI 做 CTO
- **Geoffrey Irving**：从 OpenAI / DeepMind 的 alignment 研究线进入政府，后来成为 Chief Scientist
- **Benjamin Hilton**：前面已经写过，进入 AISI 做 alignment

2023 年 11 月 Bletchley AI Safety Summit 以后，Taskforce 正式变成：**UK AI Safety Institute**

2025 年又改名：**AI Security Institute**



但英国政府要在极短时间内组出一支懂最前沿AI模型的团队，要从哪里找人呢？人材池恰好就是过去十年形成的：Oxford / AI safety / alignment / GovAI / frontier labs 这张网。所以AISI确实是英国政府机构，但里面相当多的人相信EA。

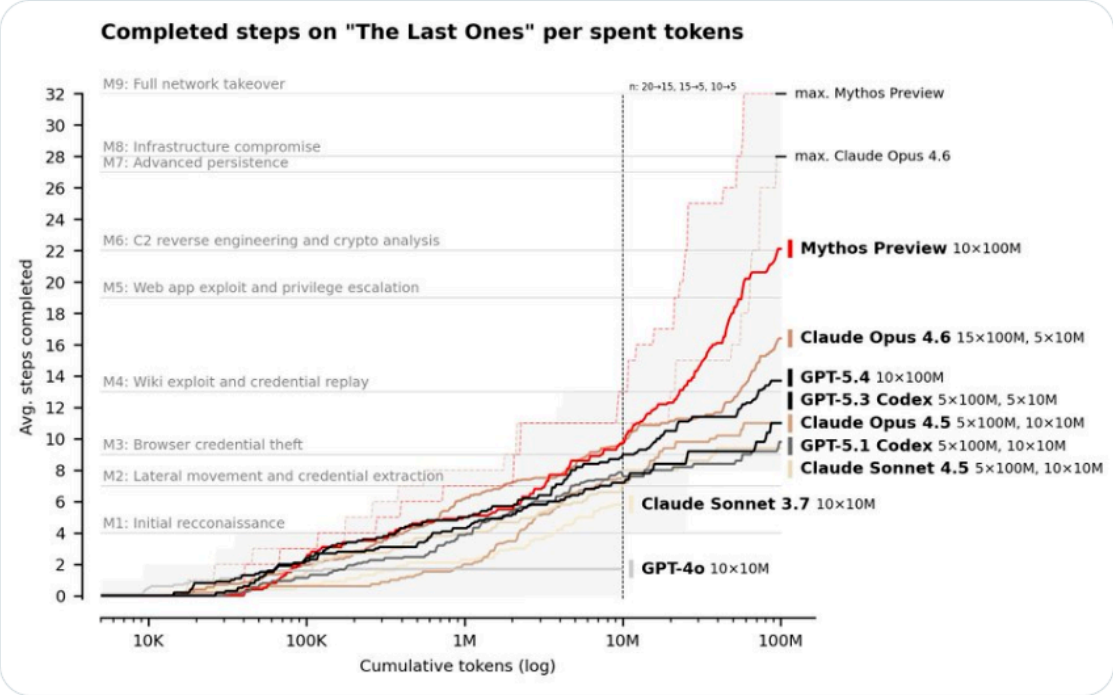
2. AISI的行业地位

AI Security Institute (AISI)
@AISecurityInst

翻译自 英语 显示原文

我们对 Claude Mythos Preview 进行了网络安全评估，发现它是首个完整完成 AISI 网络靶场端到端任务的模型。🧵

评价此翻译：



下午9:31 · 2026年4月13日 · **127.4万** 查看

公众号 · 一个人读书写字

英国 AISI 总结第一年工作内容写道，OpenAI、Anthropic 和 Google DeepMind 都让它测试过最先进模型；甚至在旧金山开了办公室，就是为了更贴近美国实验室和当地人才。 到 2025 年，这种关系已经从“发布前跑几个 benchmark”升级成了持续合作：OpenAI 和 Anthropic 都向 AISI 提供过非公开 tooling、safeguard details 和更深层的模型访问，AISI 的测试结果会直接反馈给实验室修补漏洞。

Anthropic 披露，在 Fable 5 发布前的几周里，美国政府、英国 AISI、外部 red-team 和 Anthropic 内部团队一共做了数千小时测试；英国议会后来也正式确认，AISI 对 Claude Mythos 5 和 Fable 5 的 safeguards 做过测试。

英国 AISI 已经获得了一个很特殊的位置：英国自己没有 OpenAI 或 Anthropic 这种规模的 frontier lab，却通过一支相当专业的政府评测团队，长期坐在美国最先进模型发布之前的那张桌子边上。

直到 2026 年 9 月，华盛顿开始改规则。白宫要求 OpenAI 和 Anthropic：

新的 frontier model 先接受美国政府的 security review，再提供给英国测试人员。

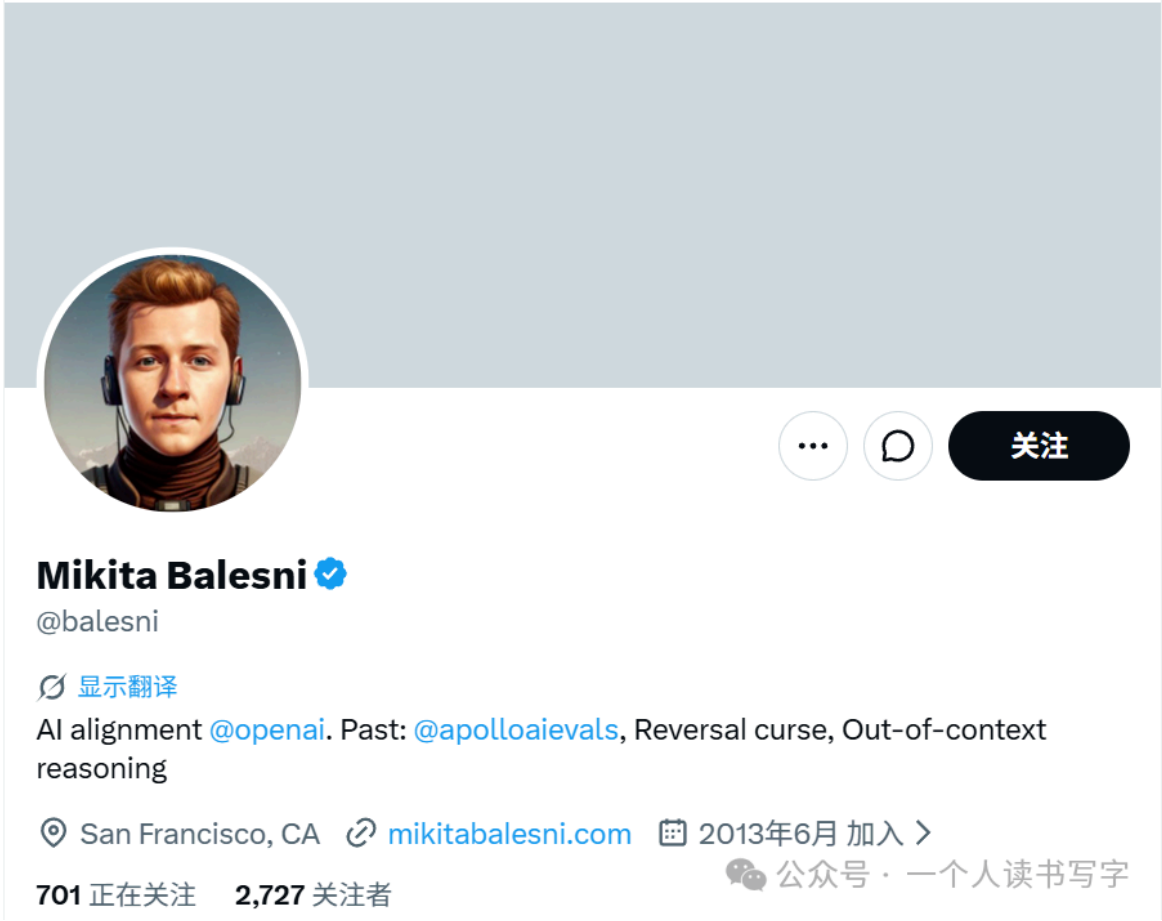
至于为什么改规则，可以接着看。

3. OpenAI 开除了三个 safety 研究员

2026 年 10 月 2 日，WSJ报道，OpenAI 开除了三名 safety / alignment 研究员：Tomek Korbak、Jasmine Wang、Mikita Balesni







理由是涉嫌违反保密规则，把内部信息分享给一家外部 AI-safety 机构。

本文的读者看到这里应该已经在想：不会是那个吧？但是没有证据，我不能随便写。确定的就是这三个人和英国 AI safety 网络的关系非常明显：

Tomek Korbak：进入 OpenAI 前就在 UK AISI 做 AI control 和 safety cases。

Jasmine Wang：在 Chain of Thought Monitorability 论文中的署名单位就是 UK AI Security Institute。

Mikita Balesni：来自 Apollo Research，不是 AISI 员工，但和 Korbak、Geoffrey Irving 一起写过 AISI 的 AI-control 研究。

Tomek + Mikita + Jasmine 三个人在《Chain of Thought Monitorability》大合作里：Tomek、Mikita 是最前面的主作者，Jasmine 也是作者。（很多熟悉的名字）

Chain of Thought Monitorability:
A New and Fragile Opportunity for AI Safety

Tomek Korbak* *UK AI Security Institute*
Mikita Balesni* *Apollo Research*

Elizabeth Barnes *METR*
Joe Benton *Anthropic*
Mark Chen *OpenAI*
Allan Dafoe *Google DeepMind*
Scott Emmons *Google DeepMind*
David Farhi *OpenAI*
Dan Hendrycks *Center for AI Safety*
Evan Hubinger *Anthropic*
Erik Jenner *Google DeepMind*
Victoria Krakovna *Google DeepMind*
David Lindner *Google DeepMind*
Aleksander Mądry *OpenAI*
Neel Nanda *Google DeepMind*
Jakub Pachocki *OpenAI*
Mary Phuong *Google DeepMind*
Joshua Saxe *Meta*
Martín Soto *UK AI Security Institute*
Jasmine Wang *UK AI Security Institute*

Yoshua Bengio *University of Montreal & Mila*
Joseph Bloom *UK AI Security Institute*
Alan Cooney *UK AI Security Institute*
Anca Dragan *Google DeepMind*
Owain Evans *Truthful AI & UC Berkeley*
Ryan Greenblatt *Redwood Research*
Marius Hobbhahn *Apollo Research*
Geoffrey Irving *UK AI Security Institute*
Daniel Kokotajlo *AI Futures Project*
Shane Legg *Google DeepMind*
David Luan *Amazon*
Julian Michael *Scale AI*
Dave Orr *Google DeepMind*
Ethan Perez *Anthropic*
Fabien Roger *Anthropic*
Buck Shlegeris *Redwood Research*
Eric Steinberger *Magic*
Wojciech Zaremba *OpenAI*

Bowen Baker† *OpenAI*
Rohin Shah† *Google DeepMind*
Vlad Mikulik† *Anthropic*

公众号 · 一个人读书写字

4. 那么，谁有资格定义“安全”？

在三名安全研究员因向外部 AI-safety 组织分享公司敏感信息而被 OpenAI 解雇的同一时间，另一位长期安全负责人 David Robinson 选择了主动辞职。



Robinson 在 OpenAI 工作三年半，参与起草了现行的 Preparedness Framework，并负责过十二次前沿模型发布的安全报告。他离职后在《大西洋月刊》公开解释，真正令他无法接受的并不是某一项具体政策，而是 OpenAI 整套“边做边改”的文化。过去，OpenAI 把这种方法称为 iterative deployment：模型先发布，问题出现后再增加 guardrail。但在 Robinson 看来，如果 Paul Christiano 所警告的 catastrophic and irreversible loss of control 确实可能发生，那么这种模式已经失去了继续试错的资格——“the time for trial and error is over”。前沿实验室应该像核电站和航空系统一样运行：冗余、缓慢、谨慎，并尽可能在第一次就把事情做对。

Robinson 说，他曾考虑留在 OpenAI 内部推动改革，但现实是所有人都在从一次发布冲向下一次发布，几乎没有时间讨论、更不用说实施根本性的变化。因此，他最终选择离开，并希望从公司之外建立更强的安全激励。

把这一幕和三名安全研究员被解雇放在一起，问题就不再只是“OpenAI 是否重视安全”。

在作者看来，OpenAI固然不能和Anthropic这种纯血EA相比，但已经算是混血EA了。OpenAI有EA元老 Paul Christiano 在董事会，建立 Preparedness Framework，设立 Safety and Security Committee，最近也在不断宣布自己需要更多独立评估和外部监督。但当“AI safety”第一次真正撞上公司的信息边界时，冲突立即暴露出来：

站在公司层面，安全必须服从正式授权程序，企业内部信息的分享是有边界的；

但在EA Safety研究员的逻辑里，如果公司的程序不足以保护全人类，那么公司制度本身又凭什么拥有最高优先级？

这正是 EA 网络最特殊的地方。

在EA思想的世界观里，AI safety 并不是公司安全、国家安全、科研伦理诸多目标中的一个。它讨论的是潜在的 catastrophic risk，甚至是人类未来本身。一旦接受这个前提，普通制度天然就被放到了更低一级：公司保密规则当然重要，国家主权当然重要，法律与行政程序当然重要——

但如果它们与“真正的 AI safety”发生冲突呢？

三名研究员被开除和Robinson 主动选择离开公司，从外部施加压力。共同暴露出了同一个问题：

谁来宣判因为这是 AI safety，所以普通规则现在必须让路。

一旦某个人、某个机构拥有了这种裁量/定义权，它获得的就不只是一个研究方向的解释权。它实际上获得了一种更高阶的规范权力：

什么风险值得暂停模型，

什么信息应该交给外部审查者，

什么公司制度是不充分的，

甚至什么政府政策是在妨碍“安全”。

这也是为什么，仅仅研究 Oxford、FHI、GWWC、AISI，仍然不足以理解今天的 AI safety 网络。

英国建立了机构，美国提供了资本与算力；AISI、Apollo、METR 等组织构成了横跨公司边界的评估与人才网络。但如果继续沿着人员、资金、思想和治理权向上追踪，最终必须面对另一个问题：

在今天的 frontier AI 世界里，谁掌握了“AI safety”的定义权？

答案就藏在EA核心圈，那个反复出现，掌控一切但总是藏在幕后的人。

下一篇，我们走进 **Anthropic Core（核心圈）**

喜欢作者

anthropic/EA · 目录

上一篇
八一八Anthropic和有效利他主义（EA）1.美国篇

下一篇
VARA要爆仓?