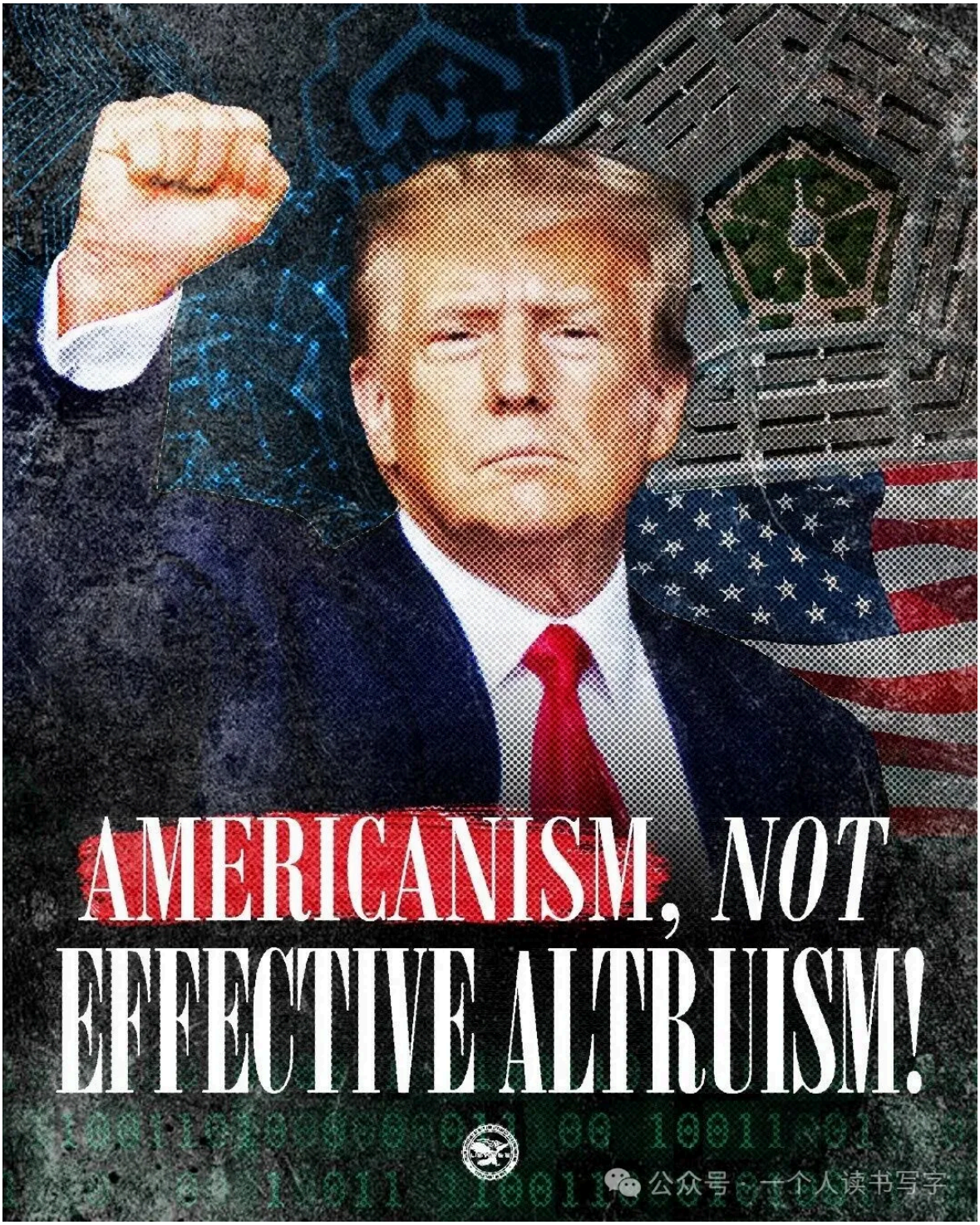
八一八Anthropic和有效利他主义（EA）1.美国篇

原创 面包人小模型 一个人读书写字 2026年9月25日 09:15 上海

Anthropic就要在纳斯达克上市了，但是S-1文件还没有公开,来自世界各地的勤劳热情的网友们已经提前帮Anthropic做了一些尽调工作。比如Anthropic和有效利他主义（EA）的千丝万缕的关系。甚至连美国政府自己都发了一张图来抵制有效利他主义。



（上图是战争部的账号在9月17号发布的并配上：The United States will NEVER be an effective altruist country.）

本文旨在八一八EA庞大的社交网络和工业化级别的组织动员能力，以及Anthropic为什么是EA体系的核心赚钱机器

1. EA起源　bayarea+New York（美国篇）

2. Oxford EA/ DeepMind（英国篇）

------↑历史↓现在---

3. Anthropic 核心圈

4.非Anthropic核心圈，但活跃在AI行业的 EA相关人士

5.金融 / 媒体 / 叙事传播渠道：历史 /现在

因为EA网络过于庞大和，人数众多，把相关的人物划分为这五组。每一组主要介绍核心人物的和同时横跨多个小组的活跃节点人物。（人名和组织名称以英语为准，中文翻译因为未经校对，可能不唯一）

一、什么是有效利他主义?

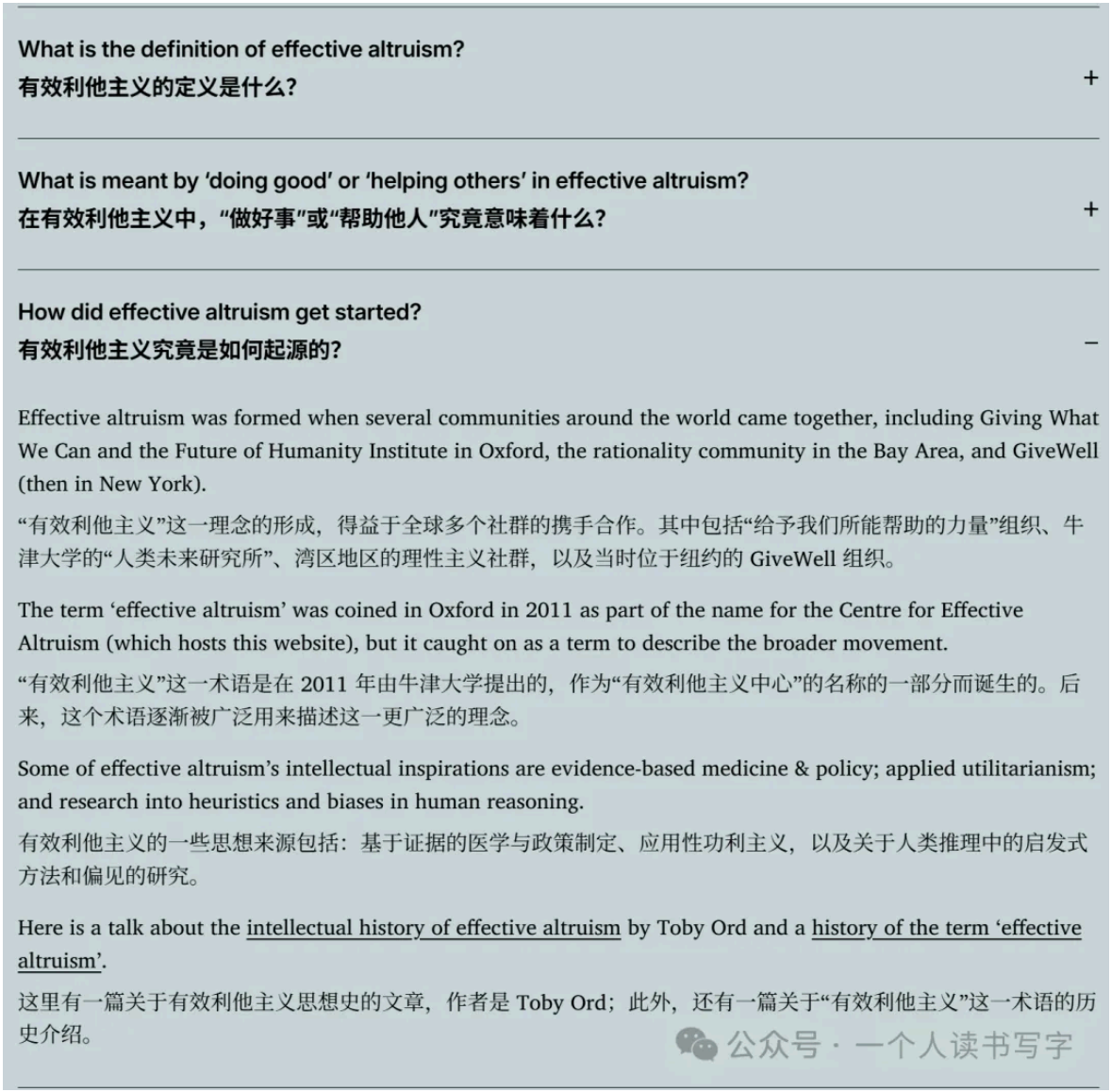
EA，全称 **Effective Altruism**，中文一般翻译成“有效利他主义”。

EA官网上的定义：

“有效利他主义是一种致力于寻找最有效的方式来帮助他人，并将这些方法付诸实践的理念。

有效利他主义是一种哲学理念，它旨在找出最有效的方式来帮助他人。同时，它也是一个实际的组织，人们会利用这些研究成果来让世界变得更加美好。”





What is the definition of effective altruism?
有效利他主义的定义是什么？ +

What is meant by 'doing good' or 'helping others' in effective altruism?
在有效利他主义中，"做好事"或"帮助他人"究竟意味着什么？ +

How did effective altruism get started?
有效利他主义究竟是如何起源的？ −

Effective altruism was formed when several communities around the world came together, including Giving What We Can and the Future of Humanity Institute in Oxford, the rationality community in the Bay Area, and GiveWell (then in New York).

"有效利他主义"这一理念的形成，得益于全球多个社群的携手合作。其中包括"给予我们所能帮助的力量"组织、牛津大学的"人类未来研究所"、湾区地区的理性主义社群，以及当时位于纽约的 GiveWell 组织。

The term 'effective altruism' was coined in Oxford in 2011 as part of the name for the Centre for Effective Altruism (which hosts this website), but it caught on as a term to describe the broader movement.

"有效利他主义"这一术语是在 2011 年由牛津大学提出的，作为"有效利他主义中心"的名称的一部分而诞生的。后来，这个术语逐渐被广泛用来描述这一更广泛的理念。

Some of effective altruism's intellectual inspirations are evidence-based medicine & policy; applied utilitarianism; and research into heuristics and biases in human reasoning.

有效利他主义的一些思想来源包括：基于证据的医学与政策制定、应用性功利主义，以及关于人类推理中的启发式方法和偏见的研究。

Here is a talk about the intellectual history of effective altruism by Toby Ord and a history of the term 'effective altruism'.

这里有一篇关于有效利他主义思想史的文章，作者是 Toby Ord；此外，还有一篇关于"有效利他主义"这一术语的历史介绍。

公众号 · 一个人读书写字

最初这些人研究的目标是非常朴素的：比如你手里有一万美元准备捐掉。那与其凭感觉随便找一家慈善机构，为什么不研究一下：同样一万美元，捐给谁能够多救几个人？同样救一个人，哪种项目成本最低？哪个组织执行能力最好？

这就是最早期的 EA。

然后事情开始逐渐变得有趣。

既然捐钱可以计算，那么职业是不是也可以计算？

既然职业可以计算，那么人生是不是也可以计算？

既然人生可以计算，那么"未来一万亿个人的生命"是不是也可以塞进这个计算器？

再然后，AI 出现了。

于是这群人又开始计算：如果超级人工智能有百分之一的概率把全人类弄死，那么这个百分之一乘以未来可能存在的几万亿、几千万亿人口，最后会得到一个多大的数字？

恭喜，我们从"怎样捐款比较划算"，一路来到了"谁应该控制 AGI"。

二、EA的起源

EA 最早并不是一个完整的组织，更像几个互相认识、又互相串门的精英大学生社群。大致可以分成三路。

- **英国 Oxford** 那边，有 Giving What We Can、**Future of Humanity Institute，也就是 FHI**，以及后来出现的 **80,000 Hours** 和 Global Priorities Institute。
- **美国**西海岸，主要是 **Berkeley** 和 Bay Area 的 **rationalist**，也就是所谓“理性主义者”社区。这里最重要的网站叫 **LessWrong**，最重要的人之一叫 **Eliezer Yudkowsky**，后面简称 Yud。
- **美国**东海岸，**纽约**还有一支，叫 **GiveWell**

本文按照地理位置划分。**美国篇**介绍Bay Area和 New York这两支，**英国篇**（第2篇）介绍Oxford分支。

这几拨人一开始研究的问题甚至不完全一样。

- **Giving What We Can：我应该捐多少钱?**
- **FHI(Future of Humanity Institute):什么样的灾难最值得人类担心?**
- **80000hours:一个人一辈子的职业有八万小时，那我应该干什么工作，才能产生最大的社会价值?**
- **LessWrong/the rationality community :你到底应该怎样思考，才算是在正确地思考?**
- **GiveWell:我应该把钱捐给谁?**

会发现这些问题其实共享着同一个类似小作文里的计算器摁冒烟的底层逻辑：

资源有限，所以我要算一下，怎样配置资源才能让预期收益最大。

大约在 2009 到 2012 年，这几群人逐渐发现：我们摁计算器的方式差不多，知音啊。

“Effective Altruism”于是成了这套共同世界观的名字。

EA 也就从一堆大学生社团，慢慢长成了一个真正的社会网络。

而一个学生社团要长成今天这个规模，当然不能只靠大家热爱哲学。

它还需要一样非常重要的东西。

钱。

三、理论发展

2009—2011：逐渐合流

2010 年，LessWrong 上有人搞了一个比赛，让大家解释什么叫“efficient charity”，也就是怎样做更有效率的慈善。Scott Alexander 等后来非常有名的 rationalist 都参加过。

这时候“EA”这个统一品牌还没有正式出现，但美国 rationalist 社区已经开始吸收有效慈善这套玩法了。

与此同时，Oxford 那边也在升级。

Giving What We Can 最初只是鼓励大家多捐钱，而且尽量捐给有效的慈善组织。

但到了 2011 年，William MacAskill 和 Ben Todd 搞出了 80,000 Hours。问题变成：为什么只优化捐钱？人一辈子八万小时的职业不是更大的资源吗？

于是 EA 从“优化你的慈善捐款”，进一步升级成了“优化你这个人”。

2011：Effective Altruism作为一个独立概念正式出现

Oxford 的 Giving What We Can 和 80,000 Hours 准备成立一个新的组织。社区里 17 个人投票，最后选出了一个非常朴素的名字：

Centre for Effective Altruism。

EA 终于有了自己的正式组织身份。但这还不够。

真正决定 EA 后来能长成什么样子的，是同一年美国发生的另一件事。

纽约的 GiveWell 和 **Good Ventures** 搭上线了。

Good Ventures 是 Facebook 联合创始人 Dustin Moskovitz 和他的妻子 Cari Tuna 的基金会。这下就不一样了。

以前是几个聪明学生坐在那里讨论：“什么慈善项目最有效？”

现在真的有人拿着巨额资金走进来问：

那你们告诉我。到底哪个最有效？

于是 EA 终于从哲学讨论进入了资产配置阶段。

你甚至可以把后来的网络理解成这样：

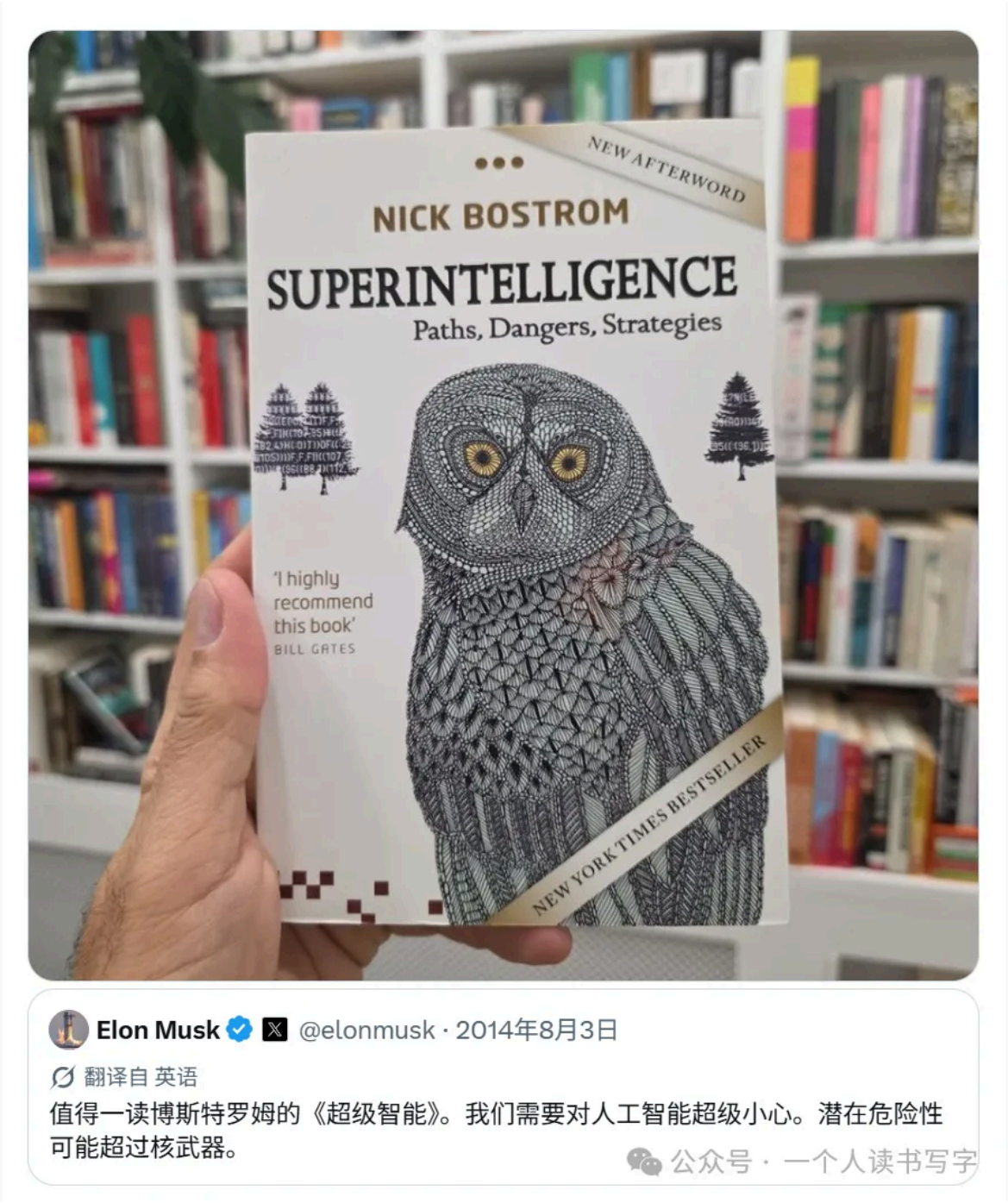
Good Ventures 提供资本。

GiveWell，后来发展成 Open Philanthropy，负责研究应该把资本往哪里配置。

而 Oxford 的 FHI 和美国的 LessWrong，则逐渐变成了两个非常重要的“算法供应商”。

Oxford 那边的 Nick Bostrom 提供了一套词汇：

- **existential risk，存在性风险；**
- **far future，遥远未来；**
- **superintelligence，超级智能；（Bostrom写了一本书↓）**
- **astronomical waste——如果未来本来可以存在无数生命，而我们今天一个错误把这种可能性全部毁了，那就是天文数字级别的损失。**



LessWrong 和 Yud 这边提供的则是另一套：**AI risk、Bayesian reasoning，以及“极低概率乘以极大损失仍然值得严肃对待”这一整套推理习惯。**



Joe Weisenthal @TheStalwart · 9月11日

翻译自 英语 显示原文

根据詹森的说法，他让所有员工都读了《如果有人建造了它，每个人都会死》。

我们还谈到了该公司自身对AI的使用，包括与@thinkymachines合作微调开源模型的工作。

imo 值得一听

5 8 97 3.2万

Joe Weisenthal @TheStalwart · 9月11日

翻译自 英语 显示原文

关于为什么詹森不期待政策制定者在人工智能开始杀人之前采取行动的完整文章



来自 bloomberg.com

7 17 65 6.1万

公众号 · 一个人读书写字

（Greg Jensen, the co-chief investment officer of bridgewater.他让员工读的就是Yud写的《 If Anyone Builds It, Everyone Dies》）

但这里一定要先记住一个很重要的区别，因为后面会反复出现：

EA 吸收了 Yud 的很多思想，不等于 EA 吸收了 Yud 这个人。

思想被拿走了。人后来反而越来越边缘。

这个故事我们等一下再讲。

早期的 FHI（英国）和GiveWell（美国）这两个分支是如何跨越地理位置带来的障碍，最终成功汇合的呢？ Nick Beckstead就是这个穿针引线的角色。



Nick Beckstead, a graduate philosophy student at Rutgers University, says someone with average income by U.S. standards can save hundreds of lives through the course of his or her career by giving.

罗格斯大学哲学专业的研究生尼克·贝克斯特德说，按照美国的标准，那些收入普通的人通过捐赠行为，可以在其职业生涯中拯救数百条生命。

公众号 · 一个人读书写字

(nbcnews，Ordinary people, extraordinary giving，20101222)

如果把早期 EA 画成一张社交网络图，Beckstead 差不多就是那个在人群之间疯狂连连看的人。他先因为研究 far-future，也就是遥远未来问题，认识了 Nick Bostrom。

（注意：两个人都叫Nick，一个是Nick Beckstead，一个是Nick Bostrom，两个人都是EA体系里相当活跃且重要的关键人物。）

Beckstead 后来参加 Giving What We Can，认识 Toby Ord、William MacAskill 这些 Oxford EA 核心人物；接着去 FHI 待了一个夏天；然后又跑去美国 GiveWell 实习；后来成为 FHI Research Fellow；最后进入 Open Philanthropy，也就是今天的 Coefficient Giving。

简单来说，这个人的人生路线差不多是：

美国出发，跑去 Oxford，把英国 EA 的几个主要据点全部打卡一遍，然后回到美国 Open Phil。



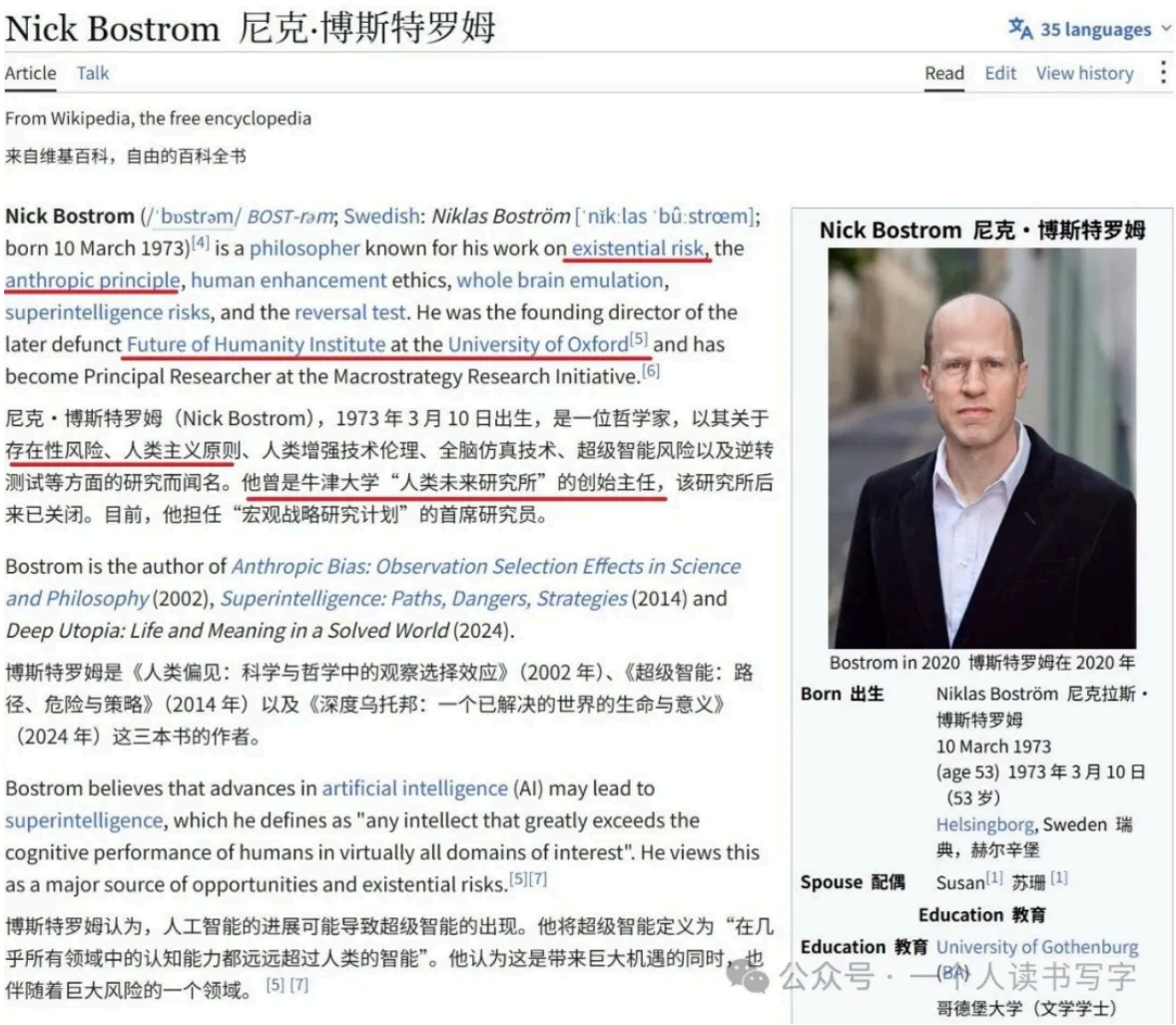
看到这里可能大家还是在想，这人到底是谁啊？换一种介绍方式：Nick Beckstead是**FTX Future Fund 的 CEO**，大家可能就很熟悉了。（SBF是 Co-founder and CEO of defunct FTX）

是的，就是那个创造了最速爆仓传说的FTX。SBF如何导致FTX爆仓就不赘述了，否则本文会突然变成金融犯罪专题。简单来说，SBF 本人是当时全世界最著名的 EA 代表人物之一。FTX是EA当时最成功的赚钱机器，但是FTX在22年破产了。FTX能在Anthropic的B轮阶段就参与融资，并不是SBF独具慧眼，只是EA体系里的资金循环罢了。



现在EA体系里最成功的赚钱机器就是anthropic了。依然是熟悉的EA配方~熟悉的爆炸头~

继续介绍两个Nick：Nick Bostrom和Nick Beckstead



Nick Bostrom 尼克·博斯特罗姆

35 languages

Article Talk Read Edit View history

From Wikipedia, the free encyclopedia

来自维基百科，自由的百科全书

Nick Bostrom (/ˈbɒstrəm/ *BOST-rəm*; Swedish: *Niklas Boström* [ˈnɪkːlas ˈbûːstrœm]; born 10 March 1973)[4] is a philosopher known for his work on existential risk, the anthropic principle, human enhancement ethics, whole brain emulation, superintelligence risks, and the reversal test. He was the founding director of the later defunct Future of Humanity Institute at the University of Oxford[5] and has become Principal Researcher at the Macrostrategy Research Initiative.[6]

尼克·博斯特罗姆（Nick Bostrom），1973 年 3 月 10 日出生，是一位哲学家，以其关于存在性风险、人类主义原则、人类增强技术伦理、全脑仿真技术、超级智能风险以及逆转测试等方面的研究而闻名。他曾是牛津大学“人类未来研究所”的创始主任，该研究所后来已关闭。目前，他担任“宏观战略研究计划”的首席研究员。

Bostrom is the author of *Anthropic Bias: Observation Selection Effects in Science and Philosophy* (2002), *Superintelligence: Paths, Dangers, Strategies* (2014) and *Deep Utopia: Life and Meaning in a Solved World* (2024).

博斯特罗姆是《人类偏见：科学与哲学中的观察选择效应》（2002 年）、《超级智能：路径、危险与策略》（2014 年）以及《深度乌托邦：一个已解决的世界的生命与意义》（2024 年）这三本书的作者。

Bostrom believes that advances in artificial intelligence (AI) may lead to superintelligence, which he defines as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest". He views this as a major source of opportunities and existential risks.[5][7]

博斯特罗姆认为，人工智能的进展可能导致超级智能的出现。他将超级智能定义为“在几乎所有领域中的认知能力都远远超过人类的智能”。他认为这是带来巨大机遇的同时，也伴随着巨大风险的一个领域。[5] [7]

Nick Bostrom 尼克・博斯特罗姆

Bostrom in 2020 博斯特罗姆在 2020 年

Born 出生	Niklas Boström 尼克拉斯・博斯特罗姆 10 March 1973 (age 53) 1973 年 3 月 10 日（53 岁） Helsingborg, Sweden 瑞典，赫尔辛堡
Spouse 配偶	Susan[1] 苏珊[1]
Education 教育	
Education 教育	University of Gothenburg (BA) 哥德堡大学（文学学士）

公众号 · 一个人读书写字

（wiki截图）

这种观念起源于哲学家Nick·Bostrom的研究。他在 2002 年提出了“存在性风险（Existential Risk）”这一概念；三年后，他创立了位于牛津大学的“人类未来研究所（FHI）”，该研究所获得了来自Tallinn和Musk的大量资金支持。在过去十年中，长期主义成为了“有效利他主义”运动的主要理念之一。这一理念过去曾引发争议，因为它鼓励年轻人为华尔街和石化公司工作，以便将部分收入用于慈善事业，这种理念被称为“为了给予而工作（earn to give）”。根据曾在牛津大学工作的长期主义者本杰明·托德的说法，“长期主义很可能是有效利他主义迄今为止最重要的发现之一。

https://www.currentaffairs.org/news/2021/07/the-dangerous-ideas-of-longtermism-and-existential-risk

Bostrom 是 Oxford 那一支的核心理论人物。他在 2000 年代初系统讨论 existential risk，也就是“有什么东西可能把整个人类文明永久性消灭”的问题，后来又创办 Future of Humanity Institute。

这条线后来逐渐发展成长远主义，也就是 **longtermism**。

长期主义最迷人的地方，也是它最容易变得非常怪的地方，在于它把“未来的人”也塞进今天的道德计算。

如果未来可能存在一万亿个人，那避免一场让这一万亿人永远无法出生的灾难，价值当然巨大。

数学上听起来非常漂亮。

然后问题就来了：

这些数字里的权重，到底是谁决定的?

于是我们又绕回了另一个 Nick——Nick Beckstead。

对，就是刚才那个后来当了 FTX Future Fund CEO 的 Nick。

这是他 2013 年在 Rutgers 写的关于遥远未来的哲学博士论文，里面有一段后来被反复拿出来批评的话。

> 在贫穷国家拯救生命所带来的影响，可能比在富裕国家拯救和改善生活要小得多。为什么呢？富裕国家的创新能力要强得多，他们的劳动者也拥有更高的经济生产力。按照一般的标准——至少按照那些有见识的人道主义标准来看，在富裕国家拯救和改善生活与在贫穷国家做同样重要，只要两种情况下生命质量的提升幅度大致相同。不过现在看来，在条件相同的情况下，在富裕国家拯救一条生命，实际上比在贫穷国家拯救一条生命更为重要。
>
> https://80000hours.org/wp-content/uploads/2022/01/Beckstead-Nick-On-the-Overwhelming-Importance-of-Shaping-the-Far-Future-better-formatting.pdf

Beckstead 这类算法供应商平时端着“全人类”“未来世代”“宇宙尺度”这些大词，一旦真的开始算，底下那个最原始的东西就露出来了：谁更像我，谁更接近我理解里的“文明”，谁就更重要。

所谓把自我从伦理里拿掉，最后经常只是把自我扩展成了整个宇宙的中心。

数学也不再是什么客观度量，而是沦为 Beckstead 手里滑稽的数字游戏：把自己的偏好乘上一万亿，自恋就变成了有效利他主义。

2021年的报道也对此进行了批判↓

> 通过将道德简化为一场抽象的数字游戏.....就是我认为长期主义是一种极其危险的意识形态的原因。实际上，它就像一种以追求“未来价值”为核心理念的世俗宗教，拥有自己的“世俗化救赎教义”。正如人类未来研究所(FHI)的历史学家Thomas Moynihan在其著作《X-Risk》中所写的那样，这种观念在西方富人群体中很受欢迎，因为这种观念能让他们听到他们想要听到的话：他们可以免于承担道德上的责任.....你过于关注那些微不足道的生存威胁，比如气候变化和全球贫困

等问题。但实际上，如果你能把注意力集中在更重要的事情上，你将会成为一个更有道德的人。那些可能永远摧毁我们作为地球起源的智慧生命物种的“潜力”的风险，才是真正需要重视的问题。

https://www.currentaffairs.org/news/2021/07/the-dangerous-ideas-of-longtermism-and-existential-risk

四、Yud——被边缘化的理论发明家

简单版本：Yud 是早期 singularitarian 和 rationalist 社区的核心人物，也是 MIRI、LessWrong 这一整套网络最重要的创建者之一；今天大家熟悉的“失控的超级 AI 可能灭绝人类，因此必须研究 alignment”这套话语，Yud 是最早系统推广的人之一。在英语AI圈一般使用Yud/Eliezer来称呼他。

复杂版本：（原文篇幅很长，因此只摘录关于Yud的部分，出处↓）



(https://www.realtimetechpocalypse.com/p/what-if-tescreal-is-a-conspiracy)

我们从 Shane Legg 讲起。他 2008 年的博士论文题为《Machine Super Intelligence》。凭借这篇论文，他获得了 Canadian Singularity Institute for Artificial Intelligence 颁发的 10,000 美元奖金。这个机构是 Eliezer Yudkowsky 创立的 Singularity Institute 在加拿大的分支——如今已经不复存在。Singularity Institute 于 2000 年成立，后来更名为 Machine Intelligence Research Institute（MIRI）。

1999 年至 2001 年，Legg 曾在 Ben Goertzel 的公司 WebMind 工作。2000 年代中期，Goertzel 共同编辑了一本讨论 human-level AI 的书，其中包括 Yudkowsky 等人的文章。这本书最初的名字是Real AI，但 Goertzel 想找一个更好的标题。于是他四处征集建议，Legg 提出了“artificial general intelligence”（AGI，ASI 的前身）这个说法。这就是“AGI”一词进入我们共同词汇体系的过程，尽管过去一年左右，在 AI hypesters 中，“ASI”已经变得更加流行。

《Artificial General Intelligence》（2007）出版三年后，Legg 与 Demis Hassabis 共同创立了 DeepMind；Hassabis 同样是一名 transhumanist 和 singularitarian。DeepMind 的明确目标，是创造一个 Machine God。和许多参与 ASI 竞赛的人一样，Legg 最初是在阅读 Ray Kurzweil 的一本书之后开始对 superintelligence 产生兴趣的，不过据他自己说，是 Yudkowsky 让他开始对 ASI“着迷”。他甚至曾通过 SL4 mailing list 与 Yudkowsky 互通邮件。这个 mailing list 是 Yudkowsky 在 2000 年建立的，其定位是“一个用于讨论 transhumanism 和 Singularity 高级议题的避难所”。

在创立 SingularityNET 之前，Goertzel 是 1990 年代 Extropian movement 的活跃成员，与 Yudkowsky、Kurzweil 和 Nick Bostrom 同处其中。在多届 Singularity Summit 活动中，他曾与这些人同台，此外还包括他的前雇员 Legg，以及 Legg 在 DeepMind 的联合创始人 Hassabis（后文还会谈到这些人）。Singularity Summit 是一项年度会议，由 Yudkowsky、Kurzweil 和 Peter Thiel 共同创办。

2008 年，Goertzel 被聘为 Yudkowsky 的 Singularity Institute（即今天的 MIRI）的“Director of Research”。第二年，他说服 Jeffrey Epstein 向该组织捐赠了 50,000 美元（据我所知，这笔钱后来又转给了另一个研究团体）。事实上，Epstein 从 2001 年起就已经开始为 Goertzel 提供资金支持。

Yudkowsky 是 2000 年代早期最知名的 singularitarian 之一，与 Kurzweil 并列。他创办 Singularity Institute（MIRI）的目的，是加速 techno–rapture，以及他经常所称的我们的“glorious transhumanist future”。他还曾深度参与 Extropian movement，创立 Rationalism，并对 Effective Altruist（EA）和 longtermist 社群产生了高度影响。（William MacAskill 曾称他为“moral weirdo”，如果这句话出自一位 EA 领袖之口，大概已经属于最高级别的赞美之一。）

Singularity Institute 最初得到了 Thiel 的资助（1,627,000 美元）。它还获得过 Vitalik Buterin、Jaan Tallinn、OpenPhil（即 Open Philanthropy，现在叫 Coefficient Giving）等方面的资金。OpenPhil 负责分配 Facebook 联合创始人 Dustin Moskovitz 的资金；Moskovitz 是一位 hardcore EA–longtermist。OpenPhil 曾雇佣 Nick Beckstead、Holden Karnofsky 和 Luke Muehlhauser。Muehlhauser 此前曾担任 Yudkowsky 的 MIRI 的 Executive Director，之后又成为 Anthropic 的 Board Member。

Beckstead 是 EA movement 最早期的成员之一，他与 Nick Bostrom 共同创建了 longtermism，并曾担任 Bostrom 的 Future of Humanity Institute（FHI）的 Research Fellow。离开 OpenPhil 之后，Beckstead 出任 Sam Bankman–Fried 的 FTX Future Fund 的 CEO。该基金由他与 Leopold Aschenbrenner、Amitav Balwit 和 William MacAskill 等人共同运营。

Sam Altman
@sama

翻译自 英语 显示原文

在 IMO 看来，Eliezer 比任何人都更推动了 AGI 的加速发展。

当然，他确实让我们很多人对 AGI 产生了兴趣，在 AGI 极度超出奥弗顿窗口的时候帮助 DeepMind 获得了资金，在决定启动 OpenAI 的过程中发挥了关键作用，等等。

评价此翻译：

上午5:28 · 2023年2月4日 · **130.1万** 查看

公众号 · 一个人读书写字

已置顶

Demis Hassabis
@demishassabis

翻译自 英语 显示原文

20多年来，@ShaneLegg 和我一直在讨论 AGI 对经济、科学和社会的潜在影响。通过 DeepMind Institute，我们正在扩展对 AI 时代关键问题的跨学科研究。我们希望它能激发必要的讨论，以确保下一步走对：

评价此翻译：



来自 institute.deepmind.com

下午10:29 · 2026年9月16日 · **59.1万** 查看

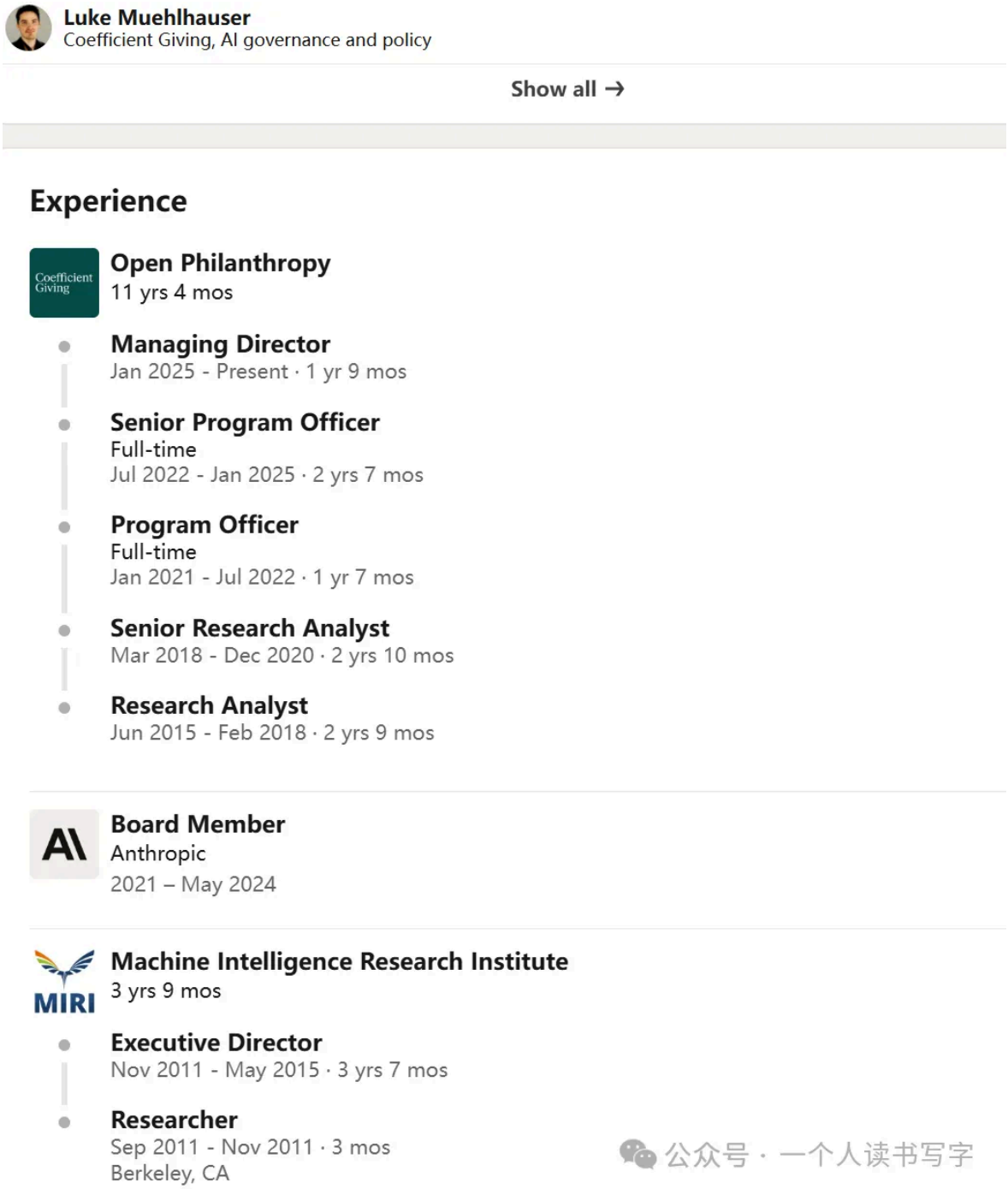
公众号 · 一个人读书写字

26年看到 DeepMind，看到 Demis，再往前顺着 Shane Legg 这根线摸，最后又能摸回 Yud 那群 2000 年代初的 singularitarian。把当年的那些会议、邮件组和机构全部摊开，就让人非常想吐槽：

怎么2026年还是这群人?

MIRI 里面很多重要人物后来都跑去了更有钱、更有资源的机构。:

Luke Muehlhauser，他早期就是 MIRI 的 Executive Director，后来2015年去了 GiveWell/Open Phil，再后来又成了 Anthropic 的董事。

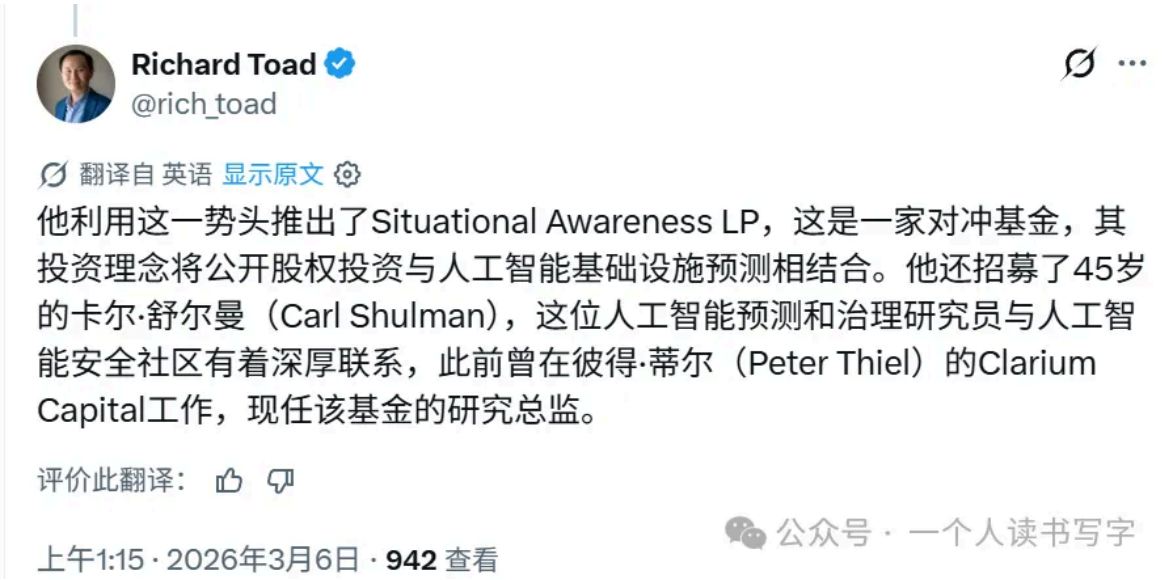


公众号 · 一个人读书写字

PhilPaul Christiano，早期 LessWrong、Bay Area rationalist 和 alignment 网络的人。现在OpenAI Board。



Carl Shulman，早期 rationalist/MIRI，后来 FHI，再进入 EA/Open Phil intellectual network，最后又跑到 Leopold Aschenbrenner 的 Situational Awareness。



Zvi Mowshowitz，LessWrong/rationalist 圈非常有名的人物，现在接受 Jaan Tallinn 的资助，继续写大量 AI safety 相关内容。

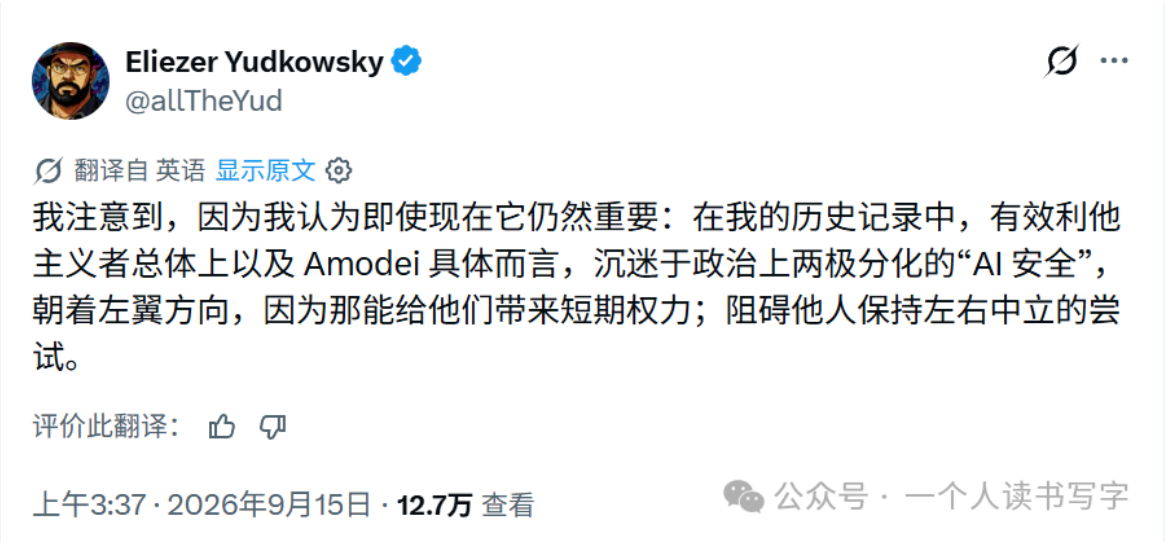
Round	Source	Organization	Amount	Receiving Charity	Purpose
Initiative Committee 2024	Jaan Tallinn	Jacob Lagerros	$100,000	Center for Applied Rationality	General support of flexHEG prototyping
Initiative Committee 2024	Jaan Tallinn	AI Policy Institute	$300,000	The Hack Foundation	General support of the 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

公众号 · 一个人读书写字

所以一个非常喜感的结果出现了。Yud 的思想到处都是。Yud 的朋友以前也到处都是。但大家后来要么进入 Open Phil 这种金主的机构，要么进入别的EA相关的组织，要么进

入Anthropic。最后 Yud 自己反而越来越像一个只能在Lesswrong复读“正如我所预期的”的知乎大神。

理论成功了。但是EA组织和他没有一毛钱关系。这也是理解 Yud 和今天 Anthropic 之间关系的关键。





因为 Yud 和 Dario Amodei 对 AI 风险的起点其实并没有那么远。两个人都认真讨论非常先进的 AI 是否可能导致灾难性风险，甚至人类灭绝。AI safety 圈喜欢把这种灭绝概率简称为 **P(doom)**。

问题在于，他们对“既然有 P(doom，那接下来怎么办”的答案完全不一样。Yud 的答案非常简单粗暴。

别造。

只要这个东西真的存在不可接受的灭绝概率，那为什么还要继续冲?

所以他后来干脆把这套观点写成了那本标题毫无歧义的书：

《If Anyone Builds It, Everyone Dies》。

任何人造出来，所有人都会死。

Dario这一派的观点就是，AI很危险，但是Anthropic是唯一可以制造安全的AI的公司，必须Anthropic先抵达AGI，美国其他实验室先制造出来会有X-risk/风险，如果让美国之外的实验室先抵达AGI，在Dario看来和世界末日没有区别。简单来说，Dario这一派的观点就是，这个世界上只有Anthropic制造出来的ai是不会有X-risk的。但是更离谱的是，你不能直接这么概括，哪怕你是Jensen都不行。

2025 年，Jensen 用一句攻击性极强的话概括 Dario：“AI 太可怕，所以只有他们应该做。”Dario 的反应也非常直接。



My full conversation with Amodei will be live on Big Technology Podcast tomorrow. You can subscribe on Apple Podcasts, Spotify, or your app of choice.

公众号 · 一个人读书写字

他把这种说法称为“outrageous lie”，也就是“离谱的谎言”，后来又说这是对他立场的 bad-faith distortion。（如图↑）

这里埋下一个伏笔，后面几篇里我会介绍EA网络和Nvida的恩怨究竟是如何开始的，这个事件可以算是两家公司恩怨的起源。

Anthropic 的公开回应也是：Dario 从来没有说过“只有 Anthropic 能安全地造强 AI”。Dario 自己给出的版本是：行业应该竞争，但大家应该向更高的安全标准竞争，而不是比谁更敢冒险。

那么问题来了，到底什么样算重视安全？这个标准是谁制定的？哪些能力应该开放？哪些能力应该受到监管？

当一个公司既是规则讨论的参与者，又是最积极推动某一套安全标准的公司之一，同时自己还是 frontier model 的主要开发者时，这里面天然就存在利益和理念互相缠绕的问题。

所以我觉得 Jensen 那句话虽然显然是最刻薄版本的概括，为什么 Dario 极力否认？因为它戳中了 Anthropic 这套叙事里最容易让外界起疑心的地方：

你一边告诉所有人这项技术危险得不得了，一边又在以最快速度亲自把它造出来。然后还要成为整个行业的标准唯一指定者。

Yud 就尤其受不了这件事。

在 Yud 看来，你如果真的相信 P(doom) 很高，那答案就应该是停止。

而不是像Dario这样又当又立：“它很危险，所以我们必须比别人更快地把它安全地造出来。”

这也是两派真正分叉的地方。



Eliezer Yudkowsky
@allTheYud

翻译自 英语 显示原文

让我把这个从 Lesswrongese 翻译出来。

Leopold：你为什么认为 AGI 可能早在 2030 年就发生？我觉得很酷的人都说 2050 年。

Eliezer：blah blah 实际理由 blah blah。

（历史：后来证明我的推理是正确的。你可以在《Biology-Inspired AGI Timelines: The Trick That Never Works》中读到。）

Leopold：但是你觉得自己比我认为很酷、并且把他们的中位时间线定在 2050 年的人*更聪明*吗？你应该把自己的立场向他们的立场调整，即使你不同意他们的论点，因为你并不比他们聪明。那才是*理性*的人应该做的。

（历史：后来会证明他们的所有论点都完全是因为我说的那些原因而错误。）

Eliezer：如果你真的要这样表述，并试图把我逼到那种对话死角里，那好吧，既然我无法赢得这场对话，我只会简单地说出我相信的真相：是的，我认为自己是一个足够优秀的预测者，我已经考虑了他们的论点，而他们没有理解我的论点，因此我信任自己不把他们当作权威——他们的意见不值得和我自己的意见等同对待的决定。

（历史：好判断。）

Leopold：蔑视

Eliezer：（接下来的几年什么也没说，因为当时 Leopold 只是个小混混；即使我立刻怀疑他是个 EA 诱饵，会被提拔到权力位置然后造成大量损害，但他现在只是个小混混，如果你是个有点权力的高层人士，你不会用自己的分量去对付一个小孩。）

Leopold：（被提拔到例如 FTX Philanthropy 的高层权力位置。）（试图破坏我提出的建议和倡议。）（在 ChatGPT 时刻让它在公众中变得可接受和酷之后，转而说 AGI 将在 2027 年到来，我们需要和中国展开军备竞赛。）（搞垮了一个对冲基金。）

评价此翻译：

上午3:43 · 2026年8月26日 · **3.4万** 查看

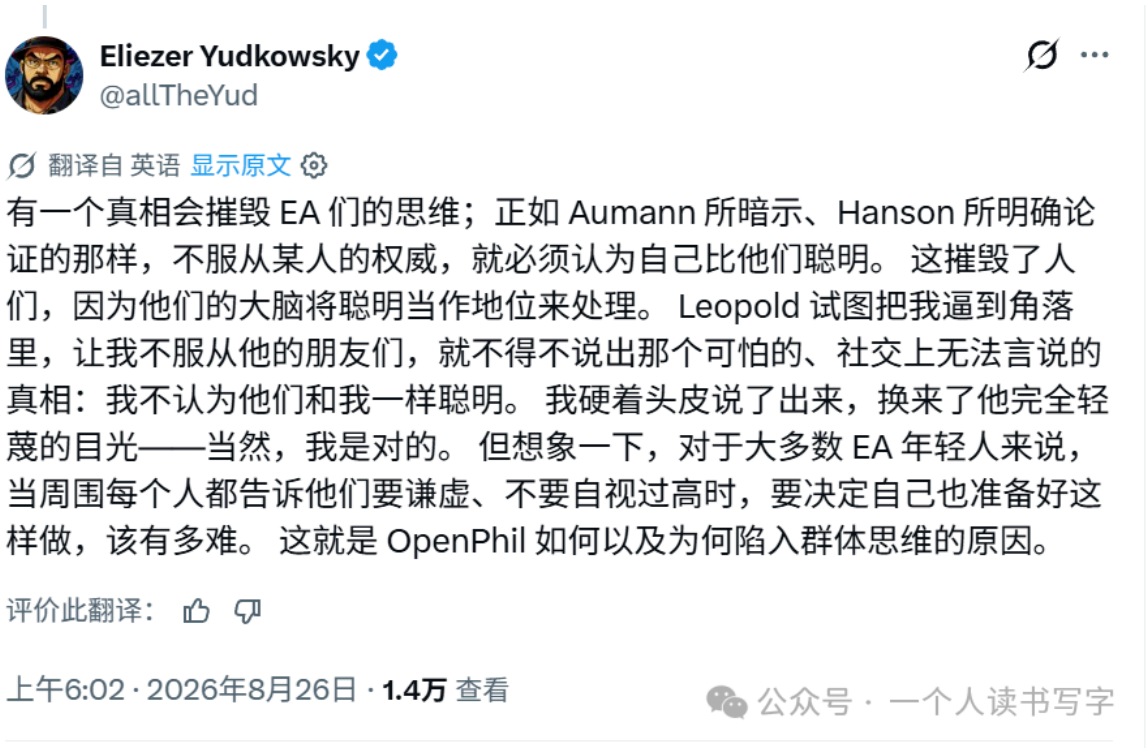
公众号 · 一个人读书写字

（Yud和EA新晋一代股神Leopold的第一次对话）

Leopold 是 EA/AI safety 新一代里非常典型的人物。后来他从 OpenAI 离开，写了《Situational Awareness》，又做了自己的投资基金。

Yud 对 Leopold 最大的不满之一，就是认为这些新一代 EA 人虽然口头上一直讲“理性”，但经常只是在复述这个圈子里当下最聪明、最有地位的人认可的观点。换句话说：

Yud 觉得自己是真的在算。他觉得 Leopold 这些人在看别人怎么算。这就非常有老资历看后浪的味道。

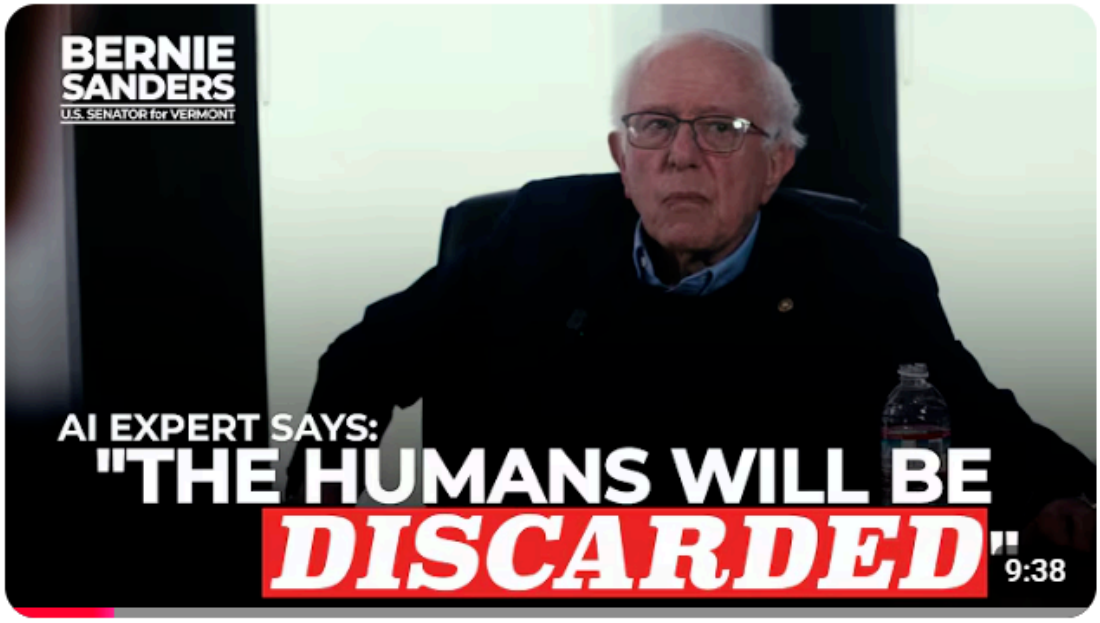


更好笑的是，Yud 在这个圈子里待了二十多年，一直算美国 AI 圈知名人物，却长期没有真正大众化。

直到 2026 年，他突然通过 Bernie Sanders 出圈了。

Yud 和 Sanders 做了一次关于 AI 风险的谈话，其中“Once they don’t need the humans, the humans are discarded”——等 AI 不再需要人类，人类就会被丢掉——这句话被剪进短片以后传播得非常广。

Sanders 当时和几位 AI x-risk / safety 人士见面的阵容，也很有代表性。



AI Expert Tells Bernie: “The Humans will be Discarded”

▷ 81万 6个月前

公众号 · 一个人读书写字

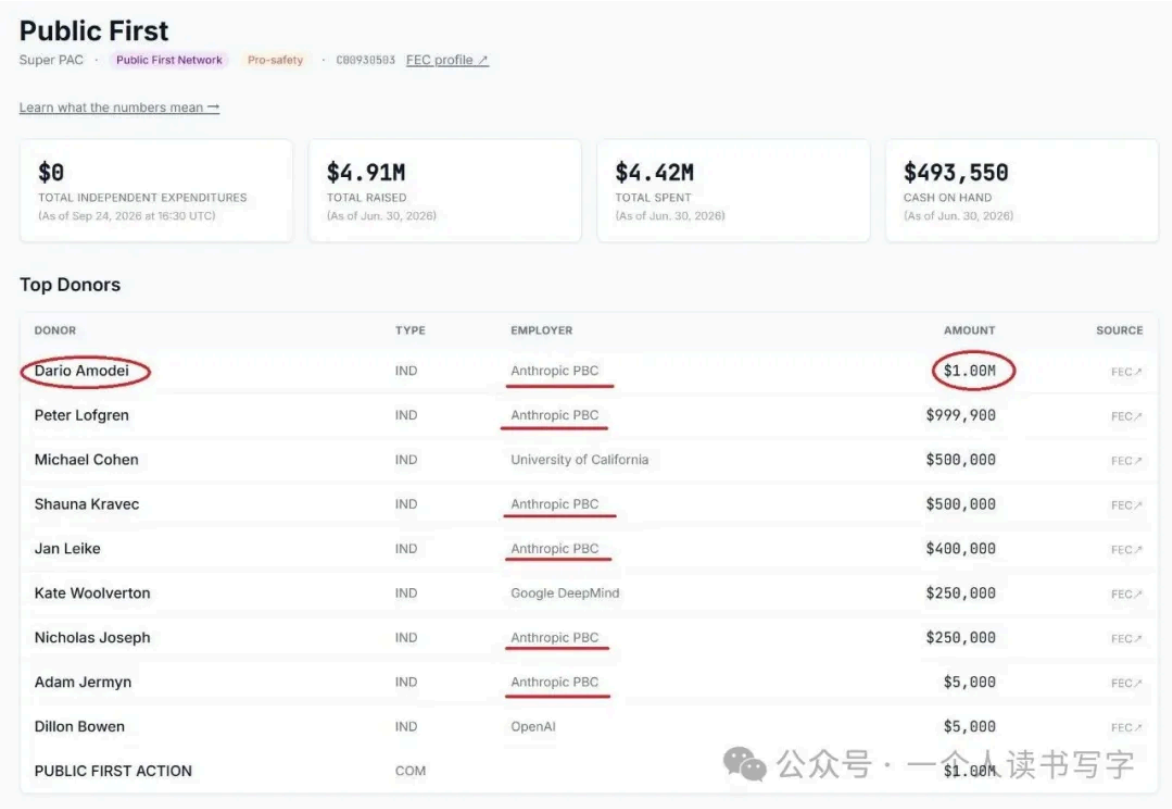
人物	当时身份 / 网络
Eliezer Yudkowsky	MIRI 创始人物、AI x-risk / alignment 早期核心人物
Nate Soares	MIRI 核心领导者，与 Yudkowsky 合著 *If Anyone Builds It, Everyone Dies*
Daniel Kokotajlo	前 OpenAI governance researcher，后来参与 AI Futures Project / *AI 2027*
Jeffrey Ladish	Palisade Research Executive Director；前Anthropic security engineer

Sanders 本身在民主党内部位于更左的一侧，在 AI 问题上也更接近 ControlAI 这种直接限制 superintelligence development 的路线，而不是单纯给现有 AI 公司加一点合规要求。ControlAI 本身明确推动对 artificial superintelligence 的强限制。

Anthropic 同样投入大量资源影响 AI 政策，但它支持的不是“大家一起把 AGI 停掉”。

2026 年 2 月，Anthropic 给 Public First Action 捐了 2000 万美元。7 月又捐了 2000 万美元。总额 4000 万美元。

Dario 本人还向与这一网络关联的 Public First super PAC 捐了 100 万美元，Anthropic 多名员工也有个人捐款。



Public First

Super PAC · Public First Network · Pro-safety · C00930503 FEC profile ↗

Learn what the numbers mean →

$0	$4.91M	$4.42M	$493,550
TOTAL INDEPENDENT EXPENDITURES (As of Sep 24, 2026 at 16:30 UTC)	TOTAL RAISED (As of Jun. 30, 2026)	TOTAL SPENT (As of Jun. 30, 2026)	CASH ON HAND (As of Jun. 30, 2026)

Top Donors

DONOR	TYPE	EMPLOYER	AMOUNT	SOURCE
Dario Amodei	IND	Anthropic PBC	$1.00M	FEC↗
Peter Lofgren	IND	Anthropic PBC	$999,900	FEC↗
Michael Cohen	IND	University of California	$500,000	FEC↗
Shauna Kravec	IND	Anthropic PBC	$500,000	FEC↗
Jan Leike	IND	Anthropic PBC	$400,000	FEC↗
Kate Woolverton	IND	Google DeepMind	$250,000	FEC↗
Nicholas Joseph	IND	Anthropic PBC	$250,000	FEC↗
Adam Jermyn	IND	Anthropic PBC	$5,000	FEC↗
Dillon Bowen	IND	OpenAI	$5,000	FEC↗
PUBLIC FIRST ACTION	COM		$1.00M	

公众号 · 一个人读书写字

(https://elections.transformernews.ai/pacs/C00930503)

Public First Action 的核心就是**推动 AI 监管 / safeguards**，但它不是那种“全面减速 AI / 禁止 AGI”的组织。更准确地说，它是一套“允许继续开发，但给 frontier AI 加制度护栏”的政治倡议平台。

方向	Public First Action 的立场
最先进模型公司的信息透明度	要求最先进模型公司披露更多安全信息、风险管理措施
联邦 AI 法规	支持建立 federal AI governance framework
州监管权	反对联邦政府直接 preempt 各州 AI 法律，除非联邦标准本身足够强
危险能力监管	特别关注 AI–assisted bio weapons、cyberattacks
芯片出口管制	支持对先进 AI 芯片实施较强出口管制
儿童 / 劳工	推动针对儿童安全、就业和其他现实社会影响的 safeguards

是不是非常眼熟？基本就是 Anthropic 自己长期公开主张的那套：

AI 很强。AI 应该继续开发。美国应该保持领先。但最先进的 AI 公司必须接受更强的安全和透明度规则。



五、GiveWell 纽约分舵之钱从哪里来

现在我们终于可以回到另一条美国主线：**GiveWell**。

前面已经出现很多次的 **Open Philanthropy**，最早就是从 GiveWell 这条线长出来的。而 GiveWell 最重要的创始人之一，叫 **Holden Karnofsky**。

因为从这里开始，EA 和 Anthropic 的关系会突然从“思想史”变成“家庭关系”。

Holden 早年在 Bridgewater 工作。（又是桥水）与 Elie Hassenfeld 创建 GiveWell，后来再和 Dustin Moskovitz/Cari Tuna (Good Ventures)一起建立 Open Philanthropy，并长期担任 CEO。GiveWell为啥有钱？因为Good Ventures 有钱。Moskovitz 是 Facebook 联合创始人。Open Phil 后来成为 AI catastrophic-risk research 最大的早期机构资助者之一。现在美国AI行业几乎很难找到和Open Phil 没有金钱往来的第三方机构。（后面几篇会介绍美国AI行业其实几乎不存在第三方机构。）



Holden Karnofsky 的妻子是 Daniela Amodei（Anthropic 的联合创始人兼总裁）。两人于 2017 年 8 月结婚。

最初GiviWell关注的方向是如何把金融行业式的比较、边际分析和透明研究应用于慈善捐赠，并没有涉猎AI安全这类的问题。真正改变 EA 网络结构的是 GiveWell 与 Good Ventures 孵化了 GiveWell Labs/Open Phil：从此这条支线不再只是告诉个人“捐给谁”，Open Phil 做的事情变成了：

“我们手里有巨额资本。现在我们自己去寻找世界上最值得投入的问题，招研究员研究它，然后直接配置资本。”

从“慈善评级机构”，逐渐变成了一个非常奇特的机构资产配置者。

这时候，再把刚才那个负责串门的 Nick Beckstead 放回来。对，就是那位：Oxford FHI 打卡。GiveWell 实习。Open Phil。后来 FTX Future Fund CEO——Beckstead，他把 Open Phil 这条线和 Oxford 的 longtermism 网络进一步接了起来。

于是三条原本分开的线终于开始彻底合流：

- Oxford 的长期主义。
- Bay Area 的 rationalist / AI risk。
- 纽约的 GiveWell / Open Phil 资本配置体系。

EA 真正变成一个巨大网络，大概就是从这里开始的。而 Holden Karnofsky，一直处在EA核心圈中的核心圈。

现在已经可以基本看出EA巨大的网络的骨干：

Dario Amodei 是 Anthropic CEO。

Daniela Amodei 是 Dario 的妹妹，也是 Anthropic 联合创始人兼总裁。

Holden Karnofsky 是 Daniela 的丈夫。

Holden 又是 GiveWell / Open Phil 的核心人物。

这时候你再看“Anthropic 和 EA 到底有什么关系”，是不是已经完全不是“Anthropic 有几个员工喜欢 EA”这种关系了。

这里插播一下股神 Leopold当初被Openai开除的小插曲：

2024 年 OpenAI 解雇 Leopold Aschenbrenner 时，OpenAI 给出的指控之一，就是 **Leopold 曾向外部人士分享敏感内部资料**。但是，那个所谓的“外部人士”不是一个普通的Anthropic员工，而是**Holden Karnofsky**。

Leopold 和 Holden 都不接受 OpenAI 对这件事的说法。

我看到这里只想吐槽，Openai的S-1文件里要怎么写这件事？

那么GiviWell到底是一个怎么样的组织呢？理解 GiveWell 最好的例子，我觉得恰好来自一个后来大家都认识的人。

2010 年，Dario 还不是 Anthropic CEO。也不是 OpenAI 员工。甚至整个今天意义上的 frontier AI 行业都还不存在。他当时只是 Princeton 的物理学博士生。然后他在 GiveWell Blog 写了一篇文章，叫**《My donation for 2009》**，文章讲的是他如何决定把 $10,000 全部捐给 VillageReach。

Dario 说自己 从 2006 年就开始向改善发展中国家生活的机构捐款，2008 年从一个经济学博客发现 GiveWell，并称它是自己截至当时找到的“最有用的捐款决策资源”。
2009 年他通过 GiveWell 的 pledge fund 捐了 $10,000。

他最后把选项缩小到两个：

VillageReach：改善儿童疫苗的配送、冷链和医疗物流；

Stop TB：给政府提供资金和药物，用于结核病治疗。

然后就是Dario介绍了一套他的算法，主要看这三个

- Cost–effectiveness（成本收益）
- Execution（执行质量）
- Incentive effects(激励效果)

原文：https://blog.givewell.org/2010/06/03/my–donation–for–2009–guest–post–from–dario–amodei/

文章中Dario写到成年人具有更丰富的体验以及与其他人的社会联系，所以在Dario的价值函数里，一名成年人的死亡大约相当于 2–3 名婴儿的死亡。他认为这个是他的主观权重调整，虽然别人可以不同意。

这一点其实很重要。因为它把 EA 最核心、也最诡异的地方展示得非常干净。EA 很喜欢说：我们不用感觉。我们算。

但真的开始算以后，你马上就会碰到一个问题：计算器里的参数谁填？Dario 填。你也可以填。大家填完以后，再开始比较谁的结果比较有效。

按照 Dario 当时的部分计算，似乎应该更偏向 Stop TB。但最后他还是选择了 VillageReach。原因之一，是 VillageReach 当时比较小，所以新增一万美元对它的边际帮助可能更大。另外，VillageReach 做的是那种不性感的后台基础设施：物流、配

送、系统建设。Dario 很喜欢这种东西，所以他还希望通过捐款给这种组织一个正向激励。算来算去。

最后 VillageReach 获胜。

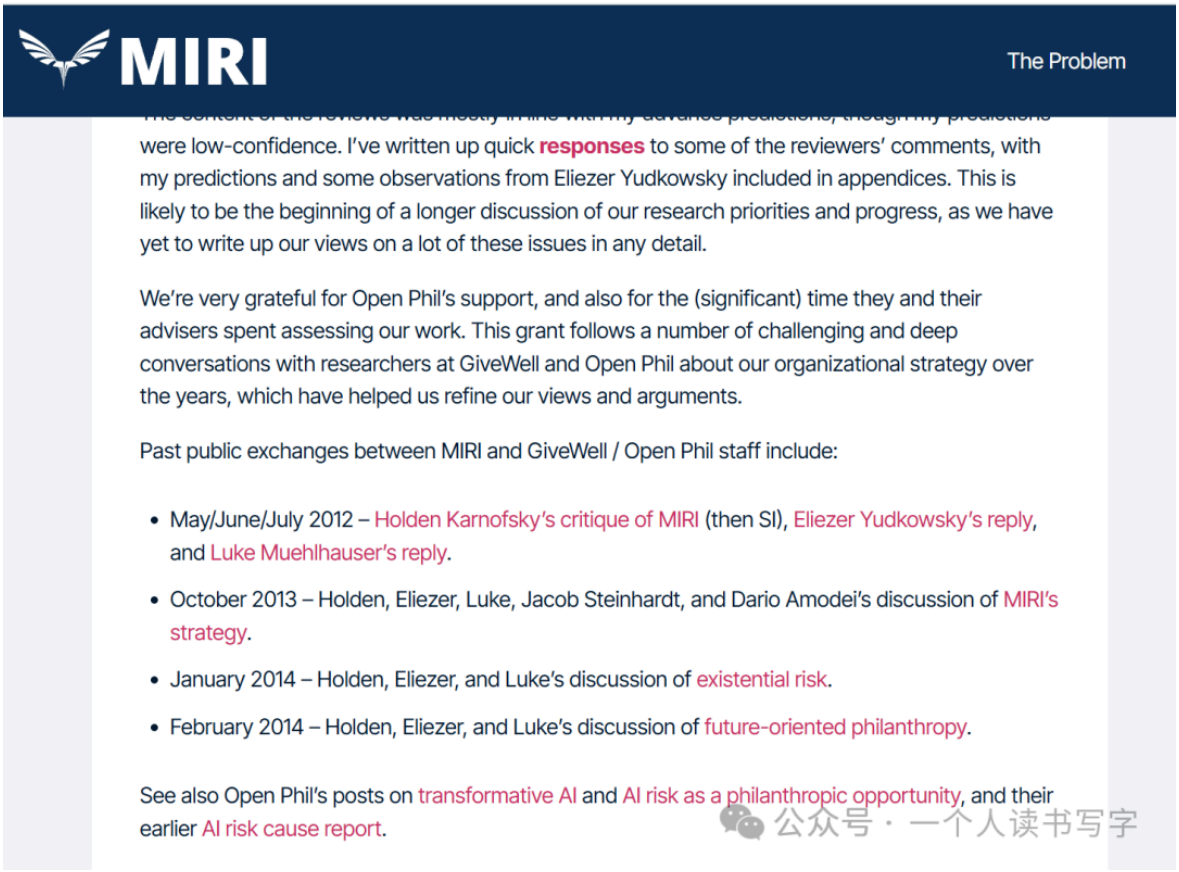
最值得注意的是，这篇文章并不是 Dario 自己随手发在博客上的。

是 Holden Karnofsky 和 Elie Hassenfeld，也就是 GiveWell 的两位创始人，邀请他写的。

注意，当时只是2010年。Dario。Holden。GiveWell。这些人在 Anthropic 出现十一年前，就已经连在一起了。

在公开可以查到的记录里，还有：

2013年和 Holden、Jacob Steinhardt 一起去和 MIRI 谈 AI / x-risk。



(https://intelligence.org/2016/09/06/grant-open-philanthropy/)

在2015年，Dario已经担任Open Philanthropy Project的科学顾问了。

GiveWell → Open Phil → AI existential risk。

Dario 已经完全进入这张网络。

A conversation with Chris Olah, Dario Amodei, and Jacob Steinhardt on March 21st and April 28th, 2015

Participants

- Chris Olah – http://colah.github.io/
- Dario Amodei, PhD – Research Scientist, Baidu Silicon Valley AI Lab; Scientific Advisor, Open Philanthropy Project
- Jacob Steinhardt – PhD Student, Computer Science, Stanford University; Scientific Advisor, Open Philanthropy Project
- Holden Karnofsky – Managing Director, Open Philanthropy Project

Note: This set of notes was compiled by the Open Philanthropy Project and gives an overview of the major points made by Chris Olah, Dario Amodei, and Jacob Steinhardt.

公众号：一个人读书写字

(https://coefficientgiving.org/wp-content/uploads/Chris_Olah_Dario_Amodei_and_Jacob_Steinhardt_03-21-15_and_04-28-15_public.pdf)

2016 年，Dario、他的妹妹 Daniela，以及 Daniela 当时的未婚夫 Holden Karnofsky，一起住在旧金山的一个集体住宅里。

同年 8 月，Dario 加入 OpenAI。

2017 年 8 月，Holden 和 Daniela 结婚。



几年以后，Dario 和 Daniela 离开 OpenAI。

再几年以后，他们建立的 Anthropic 变成全世界最重要的 AI 公司之一。

所以现在重新从头看一遍，会发现这根本不是一个“2021 年几个 OpenAI 员工意见不合，所以出去创业”的简单故事。

美国篇先介绍到这里，接下来是FHI/Oxford/Deepmind的英国篇~

喜欢作者

anthropic/EA · 目录

下一篇 · 八一八Anthropic和有效利他主义（EA）2.英国篇