# 开放式全球投资作为AGI的治理模式[1]

作者：(2025) 尼克·博斯特罗姆 (Nick Bostrom)

[工作论文，版本：1.10（2025年8月）。（初版：2025年7月）]

www.nickbostrom.com

译者：朱小虎（Xiaohu Zhu）

xiaohuzhu.xyz

Center for Safe AGI

（2025年10月）

## 摘要

本文介绍“开放式全球投资”（Open Global Investment, OGI）模式，这是一个为通用人工智能（AGI）发展提出的治理框架。其核心思想是，AGI 的开发可以在一个或多个公司内部进行，其环境应 (a) 鼓励广泛的国际持股，(b) 降低征用风险，(c) 实施强化的公司治理流程，(d) 在政府定义的负责任AI发展框架内（和/或公私合作伙伴关系中）运作，以及 (e) 在任何可取和可行的程度上，包含额外的国际协议和治理措施。我们认为，该模式虽然非常不完美，但与一些著名的替代方案——例如模仿曼哈顿计划、欧洲核子研究组织（CERN）或国际通信卫星组织（Intelsat）的提案——相比，尤其是在 AGI 发展时间线较短的情况下,在包容性、激励相容性和实用性方面具有优势。

[1] 对于评论和讨论，我感谢Renan Araujo, Guive Assadi, Nathan Barnard, Nick Beckstead, Haydn Belfield, Catherine Brewer, Joe Carlsmith, Tim Chan, Aleya Cotra, Max Dalton, Max Daniels, Tom Davidson, Oscar Delaney, Lukas Finnveden, Peter Gebauer, Ryan Greenblatt, Rose Hadshar, John Halstead, Hauke Hillebrandt, Holden Karnofsky, Will MacAskill, Fin Moorhouse, Avro Muñoz, Toby Newberry, Tina Oberoi, Abi Olvera, Toby Ord, Liam Patell, Zershaaneh Qureshi, Anders Sandberg, Charlie Steiner, Daniel Susskind, Audrey Tang, 以及一位匿名审稿人。

# 引言

针对旨在开发和部署变革性 AI 的项目，已有多种治理模式被提出。[2] [3] 有些人支持由美国主导的“AI 曼哈顿计划”——即美国与中国的类似项目在一场争夺永久宇宙霸权的战斗中展开竞争的愿景。[4] 另一些人提出了一种由美国的非营利组织或公益公司组成的所有权和治理结构。[5]还有一些人则探索了更具国际合作性的模式，例如“AI 领域的 CERN”或“AI 领域的 Intelsat”。[6]

治理模式可能强调不同的目标，例如民主监督、国家安全、国际合作、合法性、经济效率和活力、公平的利益共享、AI 安全、自由企业、明智的管理、科学进步、操作安全或负责任的部署。没有任何治理结构能够完全实现所有这些目标。此外，变革性 AI 将带来超越大多数其他政策背景的特殊挑战——包括因 AI 错误对齐或滥用而导致存在性灾难的风险，以及权力可能极度集中的问题。因此，任何治理设计都会因未能在某些合理期望上达标而受到批评。这将是在不完美选项之间的选择。

此外，随着我们接近潜在的智能爆炸，治理结构将承受巨大压力。这部分是因为赌注变得异常之高，部分是因为它们运作的环境将经历迅速而深刻的变化。[7]因此，一个给定的结构如果能够治理变革性 AI 的发展是令人满意的，但这还不够：该结构还需要足够稳健，以在整个过程中保持其完整性，并具有足够的激励相容性，以维持相关行动者的支持——并首先被采纳。

本文首先概述了一个我们称之为“开放式全球投资”（OGI）模式的范式。然后我们评估了 OGI 模式与其他替代模式相比，如何满足各种治理期望，并初步得出结论，认为它似乎相对有吸引力。

# 开放式全球投资模式

## 核心特征与变体

在 OGI 模式中，AGI 的开发由一个或多个企业主导，这些企业广泛地对国际投资开放，并在政府定义的监管框架下运作，同时得到强化的政府保证和援助的支持。现状已经大致接近这种模式，尽管我们可以更接近理想状态。如果我们支持 OGI 模式，这可能意味着我们应该努力向其理想

[2] 我们所说的“变革性 AI”，大致是指足够强大的通用人工智能（AGI），如果全面部署，有潜力对经济、劳动力市场和国家安全产生深远的跨行业影响。
[3] 关于已提出的国际 AI 机构的汇编，请参见 Maas & Villalobos (2023)。
[4] Aschenbrenner (2024)
[5] Juijn et al. (2024)。另请参见关于 OpenAI 和 Anthropic 公益公司结构的各种讨论。
[6] MacAskill & Hadshar (2025)
[7] Bostrom (2014), Dafoe (2018), Karnofsky (2023), Bengio et al. (2024)

形式靠拢；或许更重要的是，这可能意味着我们应该抵制那些会让我们离它更远的提议（如国有化）。

这个核心思想可以根据是只有一个领先的 AGI 公司（“OGI-1”）还是有几个（“OGI-N”）而采取不同形式；也取决于这些公司是否在美国注册（“US-OGI”）；以及建立了哪些支持性的特征和结构。

以下分析采纳了某种程度上以美国为中心的视角，反映了当前 AGI 发展和周边西方话语的现实。然而，OGI 模式本身在地理上是中立的 —— 原则上，任何有技术能力的国家都可以作为东道国实施，或者由多个不同国家作为不同 AGI 企业的东道国。

## 示例 (US-OGI-1)

US-OGI-1（美国注册的单一 AGI 公司）版本的一个相当成熟的形式可能如下所示：

(a) 一家公开交易的公司（“AGI 公司”）已经存在或即将上市，注册地在美国。AGI 公司可以是一家标准公司，也可以是一家特拉华州公益公司。

(b) 许多人购买股票，包括美国和世界各地的个人。主权财富基金和其他投资工具也购买了大量股份。如果进行首次公开募股（IPO），其结构将使初始所有权广泛分布。

(c) 可能有多个具有不同投票权的股票类别。例如：

A 类股：利润参与权和 1 票/股

B 类股：利润参与权和 10 票/股

C 类股：利润参与权和 1000 票/股

这些数字仅为示例。关键在于利润参与权和投票权可以部分分离——尽管理想情况下两者都应该广泛分布。例如，全世界的任何人或许都能购买A类股，而 B 类和 C 类股可能仅限于同意负责任AI框架的国家的公民；C 类股则可以作为创始人股份分配给关键方。[8]

(d) 外国政府及其公民，包括那些可能被视为战略对手的大国，被允许甚至被鼓励购买 AGI 公司的股票（或许有某个高限额，例如任何单一国家不超过 20%）。

[8] 具有不同投票权的股份的一个极端例子是所谓的“黄金股”，它赋予持有者否决权或做出特定类别决定的权利。当这样的股份由政府持有时，这可能构成一种软性国有化（这将比所有权完全按照商业原则分配时更低程度地体现 OGI 模式）。

(e) 为了帮助 AGI 公司率先开发 AGIm 并取得巨大领先，合作的政府可以采取各种行动，例如：

(i) 提供补贴、税收减免、监管豁免等。

(ii) 通过法规阻碍竞争对手（如限制获取先进 AI 芯片等）。

(f) AGI 公司也可以收购竞争对手并将其并入自己的业务中。鉴于该公司的特殊地位和官方认可，以及其竞争对手可能面临的阻力，它可能拥有较低的资本成本，从而能够进行此类收购。美国政府可以同意豁免反垄断执法。

(g) 公司的章程可能会被创建或修改，以加强其治理程序，使其超越典型的财富 500 强公司。例如，董事会可以更频繁地开会，并拥有各种资源和规定以进行内部监督。

(h) 理想情况下，公司股份的相当一部分由独立的、以造福全人类和服务其他道德使命为宗旨的非营利组织所持有。这可能是因为进行 IPO 的公司在其股东名单上已经有这类实体。慈善组织也可以在公开市场上购买额外的股份。[9]

(i) AGI 公司受其注册地法律的管辖。假设采用 US-OGI 模式，美国政府因此保留通过法律或采取监管行动的权力，以防止 AGI 公司部署被认为不安全的产品。它甚至可以叫停更先进模型的开发，直到满足某些安全和保障标准。美国政府也可以坚持对技术人员进行审查或其他措施以应对间谍活动。

(j) 采用各种机制，利用其他跨国公司为此目的使用的各种法律和结构性手段，使得美国政府或任何其他政府征用 AGI 公司在法律上变得尽可能困难和昂贵。

(k) 此外，美国政府（以及其他政府）的领导人应在其成立之初及此后定期地表示对该公司的支持，并重申尊重其产权和独立性的承诺（例如，承诺不征收没收性税收或进行国有化）。理想情况下，这些承诺应嵌入法律和条约中，但非正式的保证或支持表态也总比没有好。

(l) 在安全允许的范围内，AGI 公司的数据中心有相当一部分可以分布在多个国家和司法管辖区，以进一步增加征用的难度。

[9] OGI 模式还可以促进诸如 Epstein（2025）等提案的实施，该提案呼吁为变革性 AI 创建主权财富基金；另请参见 Casey 等人（2024），Bernard 和 Hillebrandt（2025），以及 Svarc 和 Hillebrandt，他们主张通过债务融资对 AI领域进行公共投资。

在拥有多家 AGI 公司的版本（OGI-N）中，条款 (e) 和 (f) 将不适用，从而导致情况更接近今天的现状。本文对单一公司模式（OGI-1）与多公司模式（OGI-N）的相对优劣保持中立立场。

## 运营的监管环境

上述情景只描绘了治理图景的一半：它概述了 AGI 项目的所有权和控制权。另一半是政府对 AGI 行业如何运作所施加的一系列规则。

随着 AGI 技术渗透到经济和社会的方方面面，将会涌现出无数的治理挑战——保护消费者、维护政治诚信、帮助失业工人、防止滥用等等。这些挑战大部分将由政府来应对。在 US-OGI 情景中，美国政府将承担主要的监督责任——为先进模型的开发/部署设定安全标准，建立监控程序等等。其他国家将对其国内如何使用 AI 产品施加自己的规则。这种监管环境的细节并非 OGI 模式所独有，超出了本文的范围。在理想情况下，各国可能会就一个安全和负责任的AI国际框架达成一致，该框架能够促进合作并提供基本的全球保障，各国可以在此框架内实施自己更具体的法规。

AGI公司（或在OGI-N版本中的多个公司）也可以在解决这些问题中发挥作用。例如，它可以选择不提供它认为会损害其股东长期利益或违反其更广泛使命的产品或服务。特拉华州公益公司的章程将增加为道德价值观、利益相关者利益或公共利益而牺牲利润的空间。即使没有这个，美国公司法在实践中也给予普通营利性董事会相当大的自由度来考虑除即时利润最大化之外的因素。

# 动机

OGI模式背后的基本原理有三方面：

> (1) 利用与保护产权相关的根深蒂固的规范、法律和制度。这些可能比专门为AGI设计的某些新颖的临时治理方案更为稳健和可靠。
>
> (2) 广泛分配所有权和控制权，包括在有权势的行动者（如美国国内外的资本所有者、政治代表等）之间。这具有重要功能：
>
> > (i) 它给予有权势的行动者个人激励，去维护公司的产权，而不是推行某些激进的征用或国有化运动，因为那会危及保护他们自身利益的产权框架。
> >
> > (ii) 它为国际竞争对手提供了一个替代怨恨和公开敌意的选择：欢迎他们投资并参与第一个AGI的利润和控制权。（从边际上看，这也可能降低他们参与一场有风险的AGI竞赛的意愿。）

(iii) 与所有权和控制权更加狭隘集中的模式相比，它促进了利益和影响力的更广泛、更全球公平的分配。

(3) 考虑到当前行动者的动机、政治和地缘政治约束、较短的AI时间线等因素，提出一条具有现实可行性的实施路径。

## 各行动者的角色和权力

我们现在考虑 OGI 情景中的主要行动者 —— AGI 公司、美国政府（作为东道国）、其他国家政府和公民——以及他们各自拥有的角色和权力。

### AGI公司

除了管理其内部事务和研究过程，AGI 公司可以选择开发哪些产品、向谁销售这些产品、以何种价格销售，以及在使用上附加何种限制。

假设公司在做这些决策时会受到利润动机的显著驱动。然而，其产品可能对世界产生如此深远的影响，以至于对股东的影响超出了直接利润的范畴。因此，与典型公司相比，该公司似乎更有可能受到股东非金钱利益的影响。如果 AGI 公司注册为公益公司，其法律上定义的公共使命也可能显著影响其选择。

请注意，在 OGI-N 版本中，任何一个公司塑造 AGI 影响力的权力都会因竞争压力而减弱。一家公司可能会避免开发它认为对社会有害的产品，但如果这意味着放弃利润，那么另一家公司可能会填补这个空白。因此，如果有人认为政府和领先的 AGI 开发者都对 AGI 影响的展开具有实质性影响力是重要的，这将是支持 OGI-1 的理由——或者至少将顶级 AGI 公司的数量保持在非常小的范围内。

### 美国政府

美国政府（或其他东道国政府）可以阻止 AGI 公司进行某些开发，或部署、销售某些产品，例如为了保护公众免受可预见的伤害。它也可以坚持执行程序，以减少公司遭受间谍活动或其知识产权损失的风险。

东道国政府还可以选择是否采取帮助 AGI 公司的行动，例如购买其部分股票、消除监管障碍、提供税收优惠和补贴，或阻碍其竞争对手。

在一个完全实现的 OGI 模式中，美国政府也将放弃其目前拥有的一些征用或没收公司的选项。例如，它可以承诺不将 AGI 公司国有化，不对其征收特殊的惩罚性税收，并且不强迫外国股东撤资。

## 其他政府

其他政府可以通过几个渠道施加影响。

首先，他们可以购买 AGI 公司的股票，这给予他们投票权和利润参与权。

其次，他们可以在自己的司法管辖区内监管 AGI 产品和服务，并可能阻止公司在其境内开展业务，除非该公司遵守所有当地法律。

第三，如果该公司是公益公司，如果公司未能充分推进其公共利益使命，他们可以在美国法院提起诉讼。

第四，他们可以通过正常的外交和经济渠道对美国政府（并因此间接对美国 AGI 公司）施加影响。

第五，如果已经达成了正式或非正式的多国协议（例如，负责任的 AI 政策，尊重 AGI 公司自主权的保证等），他们可能拥有这些协议提供的制度性或法律追索权。

第六，如果数据中心或其他公司资产已位于其领土上，如果美国违背其承诺，他们可能作为最后手段扣押这些资产。（人们也可以想象新颖的技术安排，例如一个 n-out-of-m 的多重签名机制，在满足某些条件时远程禁用关键的 AI 硬件。）

## 公民

公民（无论在美国还是世界各地）将拥有所有影响其政府的常规方式，从而间接影响 AGI 公司：通过向官员呼吁、投票、抗议、游说等。他们也可以使用熟悉的方法影响公司：通过他们的购买行为、是否选择为该公司工作、采取影响公司品牌形象的行动等。

在OGI模式中，公民还有购买 AGI 公司股票的选择，这将让他们直接参与公司的治理。当然，由于国内和国际间的财富差距，这种能力是不平等分配的。特别是，低收入国家的人口人均影响力将非常小。虽然这种不平等令人遗憾，但我们必须问：与什么相比？在大多数现实的替代模式下（比如说，一个仅限美国的曼哈顿计划或一家私有美国公司），同样的人们可能拥有更少的影响力来影响结果。此外，国家政府、主权财富基金、宗教组织、慈善机构或其他非政府组织可以代表那些

太穷而无法直接参与的人购买股票；然后，这些机构可以使用内部民主程序来决定如何代表其成员行使其股东权力。

## 代表性和公平性

人们很容易认为，一个由政府控制的计划会比一个由私人控制的计划更具代表性和包容性。

在当前情况下，这个假设是值得怀疑的。即使是像美国这样的大国，也只占世界人口的约4.2%（以及世界名义GDP的26%）。[10] 在一个美国国有化项目中，95.8%的世界人口（以及74%的世界经济，如果按购买力平价调整则为85%）被排除在所有权和影响力之外。[11] 即便这些数字也乐观地假设该项目在美国能保持有效的民主控制，这在某些情景下远非定数。[12] 相比之下，在OGI模式的理想类型实现中，任何拥有足够金融资产购买股票（并能接触到相关金融机构）的人都有机会参与；实际上，世界人口的很大一部分可能会参与，无论是直接（通过私人投资）还是间接（例如，通过主权财富基金、养老基金、指数基金等）。

全球财富非常不平等。近一半的全球财富由最富有的1%的个人拥有。[13] 但是，全球财富和股票的相当一部分也由大量中等富裕的个人所拥有。我们可以对此进行一些粗略的计算。全球财富的基尼系数约为0.89。[14] 全球股票所有权的基尼系数虽然在标准数据库中没有精确测量，但合理推测会更高，可能接近0.90-0.92，因为金融资产比整体财富更为集中。[15] 一个美国国有化项目——即使我们假设所有美国公民在其中拥有完全平等的股份和影响力（这似乎很乐观），而非美国公民则没有任何股份或影响力——其全球分布的基尼系数约为0.96。[16] 也就是说，至少根据一种常见的不平等衡量标准，美国国有化模式在金融收益和影响力分配方面似乎比开放式全球投资模式更为不平等。[17]

---

[10] 美国人口普查局（2025），国际货币基金组织（2025）
[11] 国际货币基金组织（2025）
[12] 国会对曼哈顿计划的监督极其有限——只有少数国会领导人知道该项目的存在，而且他们得到的信息也极少（Rhodes 1986）。副总统哈里·杜鲁门直到富兰克林·罗斯福总统去世后才知道该计划（Wellerstein 2021）。而 AGI 将允许情报分析、宣传、警务和军事能力的大规模自动化，可能导致政变或其他权力极端集中的途径。
[13] 瑞银集团（2023）
[14] 根据《2022年瑞士信贷全球财富数据手册》（可获得详细方法论的最新版本），全球财富基尼系数为0.889；见 Davies 等人（2022）。这一数字近年来保持相对稳定。
[15] 这一估计基于金融资产比整体财富更集中的事实。例如，在美国，前1%的人拥有54%的公开股票，而他们拥有的总财富约为 35%。
[16] 这一计算假设美国人口（占全球总数的 4.2%）拥有 100% 的“资产”（即通过投票权控制假设的美国国有项目），而其余 95.8% 的人拥有 0%。使用标准基尼公式 $G = 1-2B$（其中 B 是洛伦兹曲线下的面积），得出 $G \approx 0.96$。
[17] 上市公司中相当一部分的投票权集中在大型机构资产管理公司手中。例如，贝莱德（BlackRock）、先锋集团（Vanguard）和道富银行（State Street）共同控制着标普 500 指数公司约 25% 的投票权（Fichtner等人，2017）。但这种集中可以与代议制民主相比较，在代议制民主中，公民将权力委托给民选官员——在这两种情况下，许多个人选择在哪里分配他们的资源（投资或选票），但实际的决策是由少数据称代表他们行事的行动者做出的。

当我们考虑到 AGI 行业的部分利润将由美国（或 AGI 公司注册地）以及股东报告其外国来源股息和资本收益的国家的政府征税时，OGI 模式的这一初步优势会大幅扩大：这些税收将用于为全球大量公民提供服务或支付。如果我们考虑该模式的 OGI-N 版本，其中相当一部分潜在利润可能会因竞争而被消耗掉，而不是被单一垄断者（如 OGI-1 或美国国有化）所攫取，那么这一差距可能会进一步扩大。此外，就我们重视利益的地理分布而言，OGI模式在这方面的表现要好得多。

因此，可以说，一个开放且具有国际参与性的投资结构，比任何由单一国家完全拥有和控制的结构都更公平、更具全球代表性。

我们可以设想比OGI更具包容性的组织安排。例如，人们可以想象一个由联合国运营的AGI项目，或者由为此目的设立的某个新的理想化版本的联合国运营。这种更理想化的构想的一个主要问题是，它们可能与当前掌权者和潜在资助者的激励不相容，因此可能几乎没有机会被实施。许多国际治理组织也常常行动迟缓，可能没有能力运营一个具有全球竞争力的 AGI 项目。目前没有任何前沿的 AGI 项目是由任何国际治理组织运营的，也没有任何是由国家政府运营的（至少就公开信息而言）。建立一个完全全球包容性项目（特别是嵌入国际条约的项目）达成协议的时间线可能很长，而由此产生的组织结构与OGI提案所依据的国际公司产权制度相比，将是相对未经检验的。一个国际化项目将面临的另一个困难是如何实现操作和信息安全，特别是如果设想其研究和管理人员将来自近200个国家。[18]

## 军事应用与外国竞争者

民用技术和军事技术之间的区别已经常常难以划分。AGI 可能会使情况进一步复杂化，并引入新的应用领域，在这些领域中，先例对分类的指导作用有限。高水平的 AGI 能力，即使不是专门为军事应用设计的，也可能极大地促进军事成功——例如，通过帮助进行军事规划和后勤、操作无人机和机器人、进行快速军事研发，以及执行网络和信息作战。此外，经济极速增长的潜力（正如一些 AGI 发展模型所预测的那样）本身就可能对一个国家的安全和地缘战略地位产生破坏性影响，因为任何足够大的经济发展差距都可能导致军事力量的差异。[19]

因此，东道国政府——如果我们假设是 US-OGI 模式下的美国政府——可能会寻求限制 AI 产品和服务的销售，从而阻止其他国家充分参与塑造新兴的机器智能时代。此外，美国政府可能会寻求将顶级 AI 能力的销售限制在其本国公民，或许是出于军民两用的担忧。

由于担心出现这种情况，拥有与美国在 AI 发展上竞争的资源的对手国家（如中国）可能会推行自己独立的国家和/或商业 AGI 项目。即使他们在美国的 AGI 公司中有大量投资，并且即使他们有

[18] 参见 Bostrom (2017)
[19] 参见 Erdil & Besiroglu (2024)

理由相信美国会履行其尊重公司自主权和产权的承诺，他们也可能这样做。这些动态反映了各方对技术依赖和国家安全可以理解的担忧。

在考虑这对 US-OGI 模式构成多大程度的反对意见时，重要的是要记住，最相关的比较对象是那些有现实前景被实施的替代模式，而不是一个理想化的世界秩序，在那个世界里，全人类自愿地在完美的友谊与和谐中走到一起，明智地为所有人的共同利益而努力。我们应该对后者保持开放并予以支持，只要我们有机会这样做；但同时，审慎的做法是思考并制定次优方案，以防最优方案的前提条件无法实现。

因此，如果我们将 OGI 模式与“美国主导的 AGI 曼哈顿计划”模式，或与 AGI 由一家私有公司开发的模式进行比较，我们可以观察到，在那些情况下，竞争对手同样有动机去追求自己的 AGI 项目。事实上，在那些替代模式中，他们追求独立项目的动机将更强。在基本的 US-OGI 模式中，虽然在美国限制先进 AGI 能力出口的情况下，它不能保证维持他们的地缘战略地位，但他们——以及他们经济和政治精英中的个人，他们可能已经亲自投资于该项目——至少比在替代模式中更有希望参与到 AGI 发展的经济红利中。他们也将至少通过股东投票权，在技术如何发展以及哪些产品被推向市场方面拥有一些发言权。

目前，这种程度的参与肯定不足以让美国的竞争对手放弃他们自己的 AGI 努力。但在边际上，这可能会降低他们的紧迫性或规模。此外，可以想象在某些情况下，AGI 公司成功带来的股份收益前景，可能促使竞争对手退让。例如，如果情况变得极为明朗，即争夺成为 AGI 领域的首创者将带来极高的存在主义风险（由于错误对齐的超级智能），即使是大国竞争者也可能考虑单方面放慢速度，并依赖于这个共享项目。

如果对非东道国提供更强的保证是可取且可能的，那么基本的 OGI 模式可以通过额外的机制来增强，例如国际军备控制协议、互不侵犯条约等。OGI 模式对这类国际保证的适应性，应该与例如私人开发模式或曼哈顿计划模式相当。可以想象，专注于将 AGI 发展集中在一个联合运营的国际项目中的模式，在这方面可能具有优势。但它们有其他困难（见附录 4）；并且它们仍然需要面对如何处理相互竞争的商业或国家项目的问题。

## 与国家安全相关的私人控制

当 AI 成为国家安全的关键因素时，东道国政府，即我们主要例子中的美国政府，很可能希望确保能够完全获取 AGI 公司所能生产的最先进技术。

今天，美国大多数先进的国防物资是由私有（且通常是公开交易）的公司开发和生产的，例如洛克希德·马丁、RTX 公司、波音和诺斯罗普·格鲁曼。尽管美国政府通常在这些公司中没有直接的所

有权股份，但它对其活动施加了大量的控制——比对大多数其他私营部门企业更多。这部分是因为美国政府是其产品的主要买家，因此拥有买方垄断权，部分是因为它对国防部门进行严格监管——通过发放安全许可；控制哪些技术和信息可以与外国国民分享（ITAR-国际武器贸易条例）；在承包商设施中派驻政府雇员以监督生产、质量和合规性（DCMA-国防合同管理局）；有能力援引《国防生产法》（通过行政命令）迫使公司披露能力并接受和优先处理政府的“评级订单”（在极端情况下，这可能包括价格控制）；等等。外国投资是允许的，但通过 CFIUS（美国外国投资委员会）审查程序进行控制，该程序可以阻止外国试图获得控股权（有时甚至只是重要的少数股权）或施加缓解措施，如特别安全协议或 FOCI（外国所有权、控制或影响）缓解措施，例如，可能会将外国所有权限制在某个百分比，或限制外国投资者的投票权和信息获取权。[20]

鉴于政府拥有强大的工具库，投资者可能会合理地担心他们将面临事实上的征用风险。国防承包商通过大量投资于游说、“旋转门”招聘以及将工厂分布在关键政治选区，相当成功地减轻了这种风险。AGI 公司可能会尝试类似的策略，但它们可能比国防公司更容易受到国有化或征用的影响——特别是如果它们因智能爆炸而产生巨额意外利润，或在国家安全中变得极其核心。这种集中的财富和权力将是一个诱人的政治目标。此外，与国防承包商不同，AGI公司可能不会创造大量的就业机会（相对于其市值而言）；事实上，它们可能因自动化导致的大规模失业而受到指责。一家 AGI 公司或许可以通过策略性地援引中国竞争的幽灵来避免彻底关停，但如果它想长期保护自己的自主权和利润，它可能需要更广泛的支持——特别是在政治和经济精英中。OGI 模式通过给予更广泛的人群在 AGI 公司成功中的经济利益，有助于实现这一点。

## 技术与监管流程之间的速度差距

AGI发展的一个显著特点是其可能展开的速度，以及它将迫使政策制定者面对的问题的范围和新颖性。特别是在“快速起飞”的情景中，即AI能力迅速升级，穿越人类水平并进入越来越高的超级智能水平，任何完全依赖传统监管流程的治理模式都将无法跟上技术的步伐。起草、审查和制定新法规需要时间。在许多情况下，法规无论如何都不是治理一个快速演变、呈现独特、投机性或高度技术性判断要求的情况的合适工具。就像一场军事行动不能通过发布法规来管理，而必须依赖于战地指挥官根据迅速变化的情况行使判断力一样，先进 AGI 的开发和部署的许多方面也需要有能力的决策者行使行政裁量权。事情发展得越快，这种情况就越会发生——结果也就越取决于在关键时刻占据关键职位的具体个人的价值观和能力。

---

[20] 在基本的 OGI 模式中，美国将保留其目前拥有的任何试图阻碍地缘政治对手在 AI 竞赛中竞争或超越它的选项。例如，它可以对 AI 芯片和半导体制造设备实施出口限制，并鼓励其盟友和贸易伙伴对他们的出口实施类似限制。它还可以利用购买 C 类股（提供更大投票权）的权利作为筹码，鼓励各国与它合作实施此类限制，或实施其他构成更广泛的负责任 AI 框架一部分的措施。

因此，如果并且只要需要有效的政府监督，OGI 模式可能需要在公司和政府之间建立一种异常紧密的联系。在不进行彻底国有化的情况下，这可以通过多种方式实现。例如，AGI 公司可能自愿同意与一个专门的政府工作组定期会晤（该工作组包括技术专家和高级官员——当利害关系重大时，甚至可能涉及国会领导人和总统）。更正式的措施可能包括在公司内部派驻政府代表，或设立一个监督委员会来持续监控其行动。在极端情况下，可以创建一个公私合作伙伴关系结构。

值得注意的是，如果政府深度参与，OGI 模式可能需要额外的保障措施，以向股东保证他们的利益仍将受到保护（详见附录2）。

## 向后 AGI 治理的过渡

很难具体想象一个完全的后 AGI 世界——一个生物人类和动物或许与跨越巨大能力水平、架构、功能、目标、组织形式和道德地位范围的数字心智共存的世界，一个技术可能接近最终物理极限的世界——其良好的治理体系会是什么样子。[21]OGI模式主要旨在作为发展中间阶段的治理选项——即从我们现在所处的位置到成熟的超级智能出现之间。超越那个点，我们对治理的思考方式可能需要根本性的改变；但到那时，希望能够更容易地看清形势的需求，决策者也可能获得超级智能AI顾问的指导以进行进一步的改革。

OGI 模式之所以对长远未来仍然重要，是因为它可能 (a) 帮助我们以最少的负和冲突到达那里，以及 (b) 为之后的一切塑造初始条件。如果成功，OGI可能会鼓励一种发展轨迹，即通过合作性经济结构部分化解一些竞争，并且影响力与利润的分享会比最现实的替代模式更为广泛。

## 讨论与结论

我们可以将 OGI 模式视为一种“理想类型”（在韦伯的意义上，而不必是“最优”的意义上），现实可能会或多或少地接近它。当前的情况（截至2025年6月）已经提供了一些一个更完全实现的OGI 模式所能提供的好处。Alphabet 和 Meta 虽然控股权被少数人持有，但都是上市公司，许多对 AI 开发者至关重要的供应商也是如此（包括 AI 芯片的主导设计者，以及半导体供应链中的许多环节——所有领先的代工厂和许多关键设备制造商——以及最大的数据中心提供商）。至于像OpenAI、Anthropic、xAI、SSI Inc.这样的私有 AI 公司，如果愿意提供有利条件，与美国友好的国家的大型机构投资者很可能能够与这些公司协商购买股票。即使是不被视为对美国友好的个人和国家，也可以通过投资于那些投资了这些私募股权的上市公司（如软银、贝莱德等）的股份，获得稀释后的风险敞口。

[21] Bostrom & Shulman (2020)

如果一个人关注的是能带来天文数字般 AI 财富的情景，他可能会认为，即使是对这种上行空间的极度稀释的风险敞口，也足以满足任何可以想象的、资源可满足的个人或国家优先事项。无论你最终拥有 1000 个星系还是区区 10 个太阳系，你仍然可以用多出几个数量级的资源来满足大多数非地位性的个人需求。这甚至可能不需要你进行任何投资：如果至少有一个稍微慷慨的人最终拥有 1000 个星系，并愿意将其资源的 0.1% 用于对现有人口的慈善事业，那么每个人都可以得到 10 个太阳系。在领先AI公司的现有重要投资者中，很可能至少有一个稍微慷慨的人。此外，一些私有 AI 公司部分由以造福人类为使命的非营利实体拥有或控制；在这类情景中，这些慈善机构可能能够承保巨额赠款，从而广泛分配智能爆炸带来的利益。美国（或其他公司能捕获这笔巨额财富中不小一部分的国家）可以对部分利润征税，然后通过对外援助项目——即使这些项目只占其 GDP 的百分之零点几——资助同等规模的全球性发放。

基于这些原因，有人可能会认为，向 OGI 模式的理想类型靠拢几乎没有什么好处。然而，这种天文数字般的财富情景是建立在可能不成立的假设之上的——例如，它在模拟假设下可能不成立，或者如果宇宙财富存在其他申索人。[22]国家、慈善机构或当前慈善人士的行为在经历如此深刻的变革后可能会如何改变，也存在不确定性。此外，许多行动者也有无法用资源满足的地位偏好。而且，即使某些行动者的长期偏好在这些情景中实际上会得到很好的满足，但在超级智能到来之前，这些情景可能看起来不够“现实”，无法给该行动者带来太多安慰（并劝阻他们采取旨在增加潜在收益份额的绝望举动）；而在 AGI 领域中一个更清晰的股权地位可能会提供更多的保证——同时还能在过渡时期带来观看“数字上升”的满足感。对产权将得到保护的更大保证也会有所帮助。除此之外，还有一个参与治理那个引领 AGI 发展的实体的问题。在一个完全实现的 OGI 模式中，世界上任何地方对此感兴趣的行动者都将有一个和平合法的选择，通过购买 AGI 公司的投票股来获得一定程度的这种参与。

或许更重要的是，评估 OGI 模式的优劣不仅因为它能鼓励人们努力向其理想形式靠拢，还因为它能促使人们避免采取那些会使其偏离更远的措施。如果有人基于道德理由，如包容性、公平性和民主合法性，主张美国应将 AGI 发展国有化，那么指出此举将完全排除 95.8% 的世界人口和除一个国家外的所有国家是相关的。如果有人为了避免在安全标准上进行“竞次”（race to the bottom）而推动国有化，那么 OGI 模式允许外国势力（特别是其精英）作为投资者参与一个设在别国的项目，而不是被迫不惜一切代价与之竞争或诉诸绝望的破坏行为，这一点是相关的。[23]如果有人担心政变、权力攫取或普遍的不稳定，那么与财产权、投资者权利和公司治理相关的规范和法律相对完善并与公民社会融合，这一点是相关的。如果有人提议创建一个新的国际组织来管理 AGI 的发展，那么将这种构想与 OGI 模式在政治可行性、实现时间、信息和操作安全、资金前景以及可能的组织效率水平方面进行比较是相关的。

[22] Bostrom (2003, 2024)
[23] Armstrong, Bostrom, & Shulman (2016)

近期的国际努力凸显了多边 AI 治理方法的吸引力和局限性。例如，联合国 2024 年的《为人类治理 AI》报告提出了相对温和的初步步骤——一个国际科学小组、能力建设计划和协调机制。[24]尽管这些可能发挥有益作用，但它们也说明了当前的国际提案距离一个能够有意义地治理 AGI 发展的全面框架还有多远。即使这类提案最终能演变成更具实质性的东西——比如一个类似国际原子能机构（IAEA）的机构（类似于 IAEA 治理核技术的方式）或一个复杂的分布式治理机制生态系统——国际组织的历史表明，这可能需要数年或数十年，并且很可能仍然面临执行力、敏捷性和技术能力方面的挑战。与此同时，OGI 模式提供了一个务实的选择：它可以立即实施，同时与可能出现的各种形式的国际协调保持兼容——特别是那些专注于标准制定、监测或规则执行，而非所有权或直接运营控制的协调。前面提到的安全和负责任 AI 的国际框架可能会围绕 OGI 的运作逐渐形成，而不需要从一开始就进行宏大的制度设计。

## References


Achenbrenner, L. (2024) *Situational Awareness: The Decade Ahead* [online][https://situational-awareness.ai/wp-content/uploads/2024/06/situationalawareness.pdf]

Armstrong, S., Bostrom, N., & Shulman, C. (2016) “Racing to the Precipice: A Model of Artificial Intelligence Development” *AI & SOCIETY*, 31(2): pp. 201-206 [https://www.fhi.ox.ac.uk/wp-content/uploads/Racing-to-the-precipice-a-model-of-artificial-intelligence-development.pdf]

Barnard, N. & Hillebrandt, H. (2025) “Leveraging Public Debt: A Self-Financing Hedge against Unemployment under Transformative AI”. *Working paper*

Bengio, Y. et al. (2024) “Managing Extreme AI Risks Amid Rapid Progress” *Science*, 384(6698): pp. 842-845

Bostrom, N. (2003) “Are We Living in a Computer Simulation?” *Philosophical Quarterly*, 53(211): pp. 243-255

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Pathways*. Oxford University Press

Bostrom, N. (2017) “Strategic Considerations for Openness in AI Development” *Global Policy* 8(2): pp. 135–148 [https://nickbostrom.com/papers/openness.pdf]

Bostrom, N. & Shulman, C. (2020) “Propositions Concerning Digital Minds and Society”


[24] 联合国（2024）

(forthcoming in *Cambridge Journal of Law, Politics, and Art*, 2025)
[https://nickbostrom.com/propositions.pdf]

Bostrom, N (2024) “AI Creation and the Cosmic Host” *Working paper*
[https://nickbostrom.com/papers/ai-creation-and-the-cosmic-host.pdf]

Casey, E., Roy, H. & Rockall, E. (2024) *Designing an AI Bond for Growth and Shared Prosperity in the UK* [online]
[https://britishprogress.org/uk-day-one/designing-an-ai-bond-for-growth-and-shared-prosper]

Dafoe, A. (2018) “AI Governance: A Research Agenda” *Future of Humanity Institute, University of Oxford* [https://cdn.governance.ai/GovAI-Research-Agenda.pdf]

Davies, J.B., Lluberas, R., & Shorrocks, A. (2022) *Global Wealth Databook 2022*. Credit Suisse Research Institute, p. 141.

Epstein, L. (2025) *Lead, Own, Share: Sovereign Wealth Funds for Transformative AI* [online]
[https://ssrn.com/abstract=5343934 or http://dx.doi.org/10.2139/ssrn.5343934]

Erdil, E. & Besiroglu, T. (2024) “Explosive Growth from AI Automation: A Review of the Arguments” [online] [https://arxiv.org/pdf/2309.11690]

Fichtner, J., Heemkerk, E., & Garcia-Bernardo, J (2017) “Hidden Power of the Big Three: Passive index funds, re-concentration of corporate ownership, and new financial risk” *Business and Politics*, 19(2): pp. 298-326

IMF (2025) *World Economic Outlook Database, January 2025*. *International Monetary Fund*

Juijn, D. et al. (2024) “CERN for AI: The EU’s Seat at the Table” *International Center for Future Generations report*. [https://cfg.eu/wp-content/uploads/CERN_for_AI_FINAL_REPORT.pdf]

Karnofsky, H. (2023) “Transformative AI issues (not just misalignment): an overview” [online]
[https://www.cold-takes.com/transformative-ai-issues-not-just-misalignment-an-overview/]

Maas, M. & Villalobos R. (2023) “International AI Institutions: A Literature Review of Models, Examples, and Proposals” *Legal Priorities Project, AI Foundations Report* #1
[https://law-ai.org/international-ai-institutions/]

MacAskill, W. & Hadshar, R. (2025) “Intelsat as a Model for International AGI Governance” [online]
[https://www.forethought.org/research/intelsat-as-a-model-for-international-agi-governance]

Rhodes, R. (1986) *The Making of the Atomic Bomb*. New York: Simon & Schuster

Svarc, R. & Hillebrandt (2025). “Diversifying AI Ownership”. *Working paper*

UBS (2023) *Global Wealth Report 2023*. UBS Global Wealth Management
[https://www.ubs.com/global/en/media/display-page-ndp/en-20230815-global-wealth-report-2023.html]

15UN High-Level Advisory Body on AI (2024) “Governing AI for Humanity - Final Report” *United Nations*
[https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf]

U.S. Census Bureau (2025) “Census Bureau Projects U.S. and World Populations on New Year’s Day” *Press Release Number CB25-03, January 1, 2025*
[https://www.census.gov/newsroom/press-releases/2025/population-new-years-day.html]

Wellerstein, A. (2021) *Restricted Data: The History of Nuclear Secrecy in the United States*. Chicago: University of Chicago Press

# 附录1：公司结构与限制

目前拥有一些最先进 AGI 项目的几家公司选择保持私有状态是有原因的。

通常，公司进行 IPO 最重要的原因是为了获得更大规模（成本更低）的资本，并为其早期投资者提供流动性。迄今为止，这些公司从私人投资者那里筹集足够的资本是可行的。这可能是因为目前具有长远眼光的私人投资者对 AI 领域的机会有极大的兴趣（加上这些公司的一些创始人个人非常富有或拥有强大的人脉网络和私募融资的特殊技能）。随着前沿 AI 公司的资本需求飙升至数千亿美元或更高，能够提供足够资本的私人投资者数量将开始减少；但到目前为止，这尚未成为决定性的限制。

从这些公司及其所有者和负责人的角度来看，IPO 的潜在好处必须与一系列缺点进行权衡。其中之一是合规负担的纯粹金钱成本——对于一个市值数千亿美元的公司来说，这可能达到每年约1亿美元——在当前背景下相对微不足道。但其他可能更严重的缺点包括信息披露要求（这可能泄露具有战略敏感性的信息给竞争对手）、增加的审查和举报人条款、分散对核心技术和业务发展的注意力、创始人/重要人物控制权被稀释的可能性、易受激进投资者攻击，以及可能面临更多种类的监管行动和诉讼。对于目前具有非标准治理结构、涉及非营利控制、利润上限以及各种承诺（对公共利益的承诺，或在某些条件下退让并支持竞争对手的 AGI 项目等）的 AI 公司来说，情况可能更加复杂。此外，向华尔街分析师展示易于理解的产品路线图和商业模式也可能很困难。（你如何在季度财报电话会议上解释，你决定放弃数十亿美元的利润，因为你认为你的 AI 模型有 *40%* 的概率是有意识的或基于其他理由具有道德地位，以至于你在道义上有义务为数字心智的福祉慷慨解囊？）如果你的公司由价值一致的投资者私有，并且这些投资者是经过精心挑选、理解并支持你的愿景的，那么处理这些事情就容易得多。

为了适应这些担忧，可以探讨各种安排。这些可以包括采用特殊的治理条款和多种股票类别，以允许最初的使命和公司的创始人或主要负责人即使在新投资者加入后也能继续拥有很大程度的控制权。注册为特拉华州公益公司可以赋予董事会更多的自由度，以追求股东回报最大化之外的目标。创建一个拥有原公司大量股份的公开交易控股公司或封闭式基金，可能允许后者保持私有，同时为更广泛的投资者群体扩大参与机会。如果得到监管机构的合作，可能性的空间就会扩大。美国证券交易委员会（SEC）可以发布“无异议函”或豁免令，放弃执行某些否则会阻止或复杂化理想解决方案的法规。如果国会决定希望推动一个沿OGI模式的解决方案，那么可能性的空间会更大。

如果美国IPO的监管负担过于繁重，可以探讨的另一个选择是在海外上市，例如在伦敦、多伦多、新加坡或香港证券交易所，或者为了最小化监管和披露要求，在开曼群岛（CSX）或百慕大（BSX）证券交易所上市。这可以与发行美国存托凭证（ADRs）相结合，这些凭证可以在纳斯达克（或对于一级 ADRs 在场外交易市场）进行交易。

对于所有这些选项，都存在超出本文范围的复杂性和局限性。

# 附录2：投资者保护与政府监督

虽然美国政府对美国公司的利润征税，但它通常尊重公司的产权和自主权。然而，鉴于干预一家成功的 AGI 公司可能存在强烈的诱惑——以及随时可用的理由（如国家安全理由）——为了使 OGI 模式对投资者具有可信度，额外的保证将是有益的。可能还需要特别的豁免或鼓励，以解锁来自与美国目前关系紧张或受制裁的国家的投资，而该模式的最纯粹形式将允许这些投资。

在美国政府试图将 AGI 活动集中于一个公司（US-OGI-1）的版本中，美国政府的角色可能会扩大到包括压制或合并竞争对手，以及保护选定的项目免受反垄断诉讼。如正文所讨论的，美国政府在定义 AGI 的监管框架、监督活动以及保护 AGI 公司免受各种威胁行为者的侵害方面也将发挥重要作用。在某些情景下，政府的参与可能会发展成为公私合作伙伴关系或采取“软性国有化”的形式。

要充分发挥 OGI 模式的优势，不仅要确保投资者的利益实际上得到保护，而且还要让投资者——理想情况下包括那些可能因此减少与美国激烈竞争或破坏美国努力的竞争对手——能够事先对此有合理的信心。可以探讨一些措施来帮助解决这个保证问题。其中一些可以由 AGI 公司单方面实施，而另一些则需要美国政府的参与。

非正式信号和承诺。—— 至少，公司领导人和/或政府官员可以表达对保护 AGI 公司股东利益的支持或承诺。

政府关系、公关和游说。—— AGI公司可以投资加强其向立法者传达立场的能力，并抵制日后可能出现的国有化呼声。

法律和条约保护。—— 可以探索可能使国家征用更加困难的法律结构，例如为某些知识产权使用海外控股公司。灵感可以来自跨国公司如石油巨头和矿业公司在治理不稳定或法治薄弱地区运营时如何减轻政治风险。一个有足够动力的政府也可以寻求将对美国AGI公司国际投资的支持嵌入与其他国家的正式协议或条约中。（关于此点的更多内容，请参见附录5。）

资产的地理分散。—— 一家 AGI 公司可以将一些资产，如数据中心和知识产权（如模型权重）离岸，使其母国更难单方面没收其所有资产。如果美国政府知道彻底没收的尝试将无法获取重要的海外资产，它可能从一开始就不太可能采纳这样的政策。

技术措施。—— 例如，可以在数据中心安装远程“死亡开关”，这可以为外国提供抵制美国政府试图将 AGI 公司国有化或征用其投资者的方式。这可以设置为避免任何一个外国单方面挫败美国 AGI 项目的能力。例如，可以给 M 个国家中的每一个国家一个加密密钥，允许它定期向美国数据中心发送签名消息。只要至少收到 N 条消息，数据中心就继续运行。这提供了一种灵活的方式来分配部分否决权（例如，没有一个国家能够单独停止 AGI 公司的数据中心，但任何三分之二的国家组合都能够暂停数据中心，直到他们对自己的合同权利得到尊重感到满意为止）。

利益一致。—— 这是 OGI 模式的一个重要特征：通过允许政治和经济精英投资 AGI 公司，它可以激励有影响力的群体反对没收性的政府干预和其他敌对行动（无论是在国内还是国外）。

# 附录3:一个还是多个AGI公司？

OGI-N 模式(多个AGI公司)比 OGI-1 模式更接近当前现实。

这样一个竞争性的商业环境是可取的还是不可取的，取决于我们认为AGI的主要风险在哪里。如果我们将 AGI 基本视为又一种(通用)技术，我们的默认立场大概会是竞争是有益的:如果存在一些竞争而不是单一垄断，我们往往会得到更快的进步和更大的消费者剩余。OGI 模式的(相对 )全球包容性，及其在某种程度上缓解 AGI 发展中各国之间负和动态的能力，及其阻止权力极端集中的潜力(如果 AGI 发展被国有化或由一个少数人持有的私营公司主导，这种情况更容易发生)，这些优势无论是一家还是多家上市的 AGI 公司都同样适用。(而希望防范其中一个竞争者最终“获胜”并成为拥有巨额意外利润的垄断者的投资者，可以将其持股分散到所有可能的竞争者之间。 )

一种相反的观点认为，AGI 是独一无二的，部分原因在于它带来了非同寻常的风险，例如失调的超级智能——这些风险的缓解可能需要一个密切协调的开发过程。[25] 例如，对于一个领先的 AGI 开发者来说，当其 AI 能力达到某个临界水平时，能够暂停或放慢其进展，以便给其对齐团队时间来实施和测试额外的保障措施，这可能非常重要。如果其他开发者紧随其后，并且很明显花时间在保障措施上意味着将比赛的胜利让给某个不那么谨慎和不那么规避风险的竞争者，那么采取这种预防措施的空间就会减少。如果我们以这种方式看待 AGI，我们可能不希望有一个竞争市场，让多家公司竞相开发和部署能力越来越强的 AI。

OGI 模式与这两种对 AGI 的看法都是兼容的。如果竞争性局势是可取的，那就让多家 AGI 公司存在。这可能是默认发生的情况。在这种情况下，可以通过鼓励广泛的股权分布(包括国际上)以及采取措施加强对这些公司产权的保护(即使在价值极度增长的情况下)，来进一步支持 OGI-N 模式。

另一方面，如果竞争性局势是不可取的，那么基本的 OGI 模式可以与旨在消除或阻碍除指定项目外的任何 AGI 项目的措施相结合，或者强制合并。例如，美国可以规定，没有许可证运营此类项目是非法的，并且只向官方认可的公司颁发许可证(在这种情况下，该公司可能采取公私合作伙伴关系的形式，但仍具有全球开放的投资结构)。美国还可以向其他国家施压，要求其禁止在其管辖范围内进行竞争性的 AGI 项目，或者利用其影响力(例如其对半导体供应链的影响)来使竞争对手难以生存。初步来看，OGI 模式下可用的压制竞争的选项——如果需要这种方法的话——与替代模式(例如美国主导的 AGI 曼哈顿计划)下的选项是相同的。

[25] 参见 Bostrom (2014)

我们可以想象这样一种情景：全球只存在一个 AGI 项目是可取的，但在 OGI 模式下实现这一目标会更加困难。例如，也许中国或其他一些重要国家更愿意关闭国内的 AGI 项目，以支持某个类似 Intelsat 的联合政府间项目，而不是支持某个位于美国的公司项目（即使我们规定在后一种情况下，他们将被允许作为平等利益相关者进行投资，并且该项目将得到美国政府自主权的保证等）。然而，目前，美国、中国和 AI 领域的其他重要参与者都愿意为了一个单一的联合国际项目而关闭其国内项目的情况似乎相当遥远。如果情况发生足够大的变化，使得这种安排变得可行，那么新的情况也可能使得单方面放弃以支持 US-OGI-1 模式（即一个总部在美国的 AGI 公司取得领先）变得可行。这可能是一种情况，即一家美国公司的领先优势如此之大，以至于任何竞争显然都是徒劳的，并且只会通过削减可用于对齐工作的时间来加剧存在风险。

# 附录4：与其他一些模式的比较

本文提出了开放式全球投资模式，以促进对其相对于AGI治理其他方法的优劣进行更广泛的讨论。本附录就OGI模式的一些好处提供了一些简短而初步的评论。（对所有可能方法的全面比较——这是得出关于最佳前进道路的全面判断所必需的——超出了本贡献的范围。）

## “AGI曼哈顿计划”

我们在正文的多个地方将 OGI 与“AGI 曼哈顿计划”进行了比较。简要回顾如下：

- OGI 可能对许多现有参与者更具吸引力，包括当前的 AI 公司领导层、人员和投资者。广泛的可投资性可以增加美国和国际精英之间的激励相容性。
- 与仅限美国的国有化项目相比，OGI 承诺更广泛、更公平的利益和影响力分配。
- OGI 避免了对大规模政府资金的需求。
- OGI 可能会降低权力极端集中的可能性，部分原因是通过创建一种公司和政府都拥有重要权力的双重否决或权力结构，部分原因是通过将项目嵌入公民社会，在那里有更多的透明度、社会规范和法律结构，而不是在一个由安全部门运营的国有化项目中。
- OGI 可以通过给予许多国家及其精英参与美国项目的机会，在一定程度上缓解国际间的负和竞赛动态和潜在冲突。
- OGI 模式提供了一系列关于政府参与程度的选项——从不比现状更多，到随着风险或社会影响增加而扩大对 AGI 行业的监管，到非正式的咨询和政府监控，到持续的正式监督安排，再到公私合作伙伴关系或其他形式的软性国有化。
- OGI 模式与发展一个更广泛的负责任 AI 发展的合作性国际框架相一致，和/或与美国更单方面的努力（例如通过操纵半导体供应链等）来影响其他国家的 AI 努力相一致。

## “AI领域的CERN”

我们也可以将 OGI 模式与类似“AI 领域的 CERN”进行比较——这是一个国际联合运营和控制的、旨在开发先进 AI 的项目，大致模仿了欧洲核子研究组织（Conseil Européen pour la Recherche Nucléaire），该组织建造并运营着大型强子对撞机。它可以是一个由西方盟友共享的项目，也可以是一个更全面的全球合作。

一个“AI 领域的 CERN”继承了“AGI 曼哈顿计划”的一些缺点：它对当前的 AI 现有参与者可能激励不相容，需要大量的政府资金，建立起来可能缓慢而困难，并且可能涉及一个比标准公司法和私有财产规范更定制化、更少经过验证和根深蒂固的组织结构。

除了这些共同的缺点，“AI 领域的 CERN”还有其自身的一些额外缺点。一个国际民用项目将在信息安全方面面临巨大的挑战：其工作人员来自世界各地，并可能在各种外交保护伞下运作，那么如何防止在模型变得安全可部署之前发生间谍活动或窃取见解、代码或模型呢？这在基础物理研究中不是问题，但在某些AGI情景中将至关重要。（在完全全球性的版本中，这个困难被放大了，但即使在仅限于某个广泛的盟国联盟的版本中，这也可能相当艰巨。）另一个潜在的缺点是，不清楚“AI 领域的 CERN”与在公司环境或 AGI 曼哈顿计划中可实现的飞速发展速度相比，能有多大的竞争力。

“AI 领域的 CERN”与“AGI 曼哈顿计划”相比，也有重要的优势。最显著的是，它可能更具全球公平性，并可能更容易被各大国所接受。

人们可能认为，在希望全球只有一个领先的 AGI 项目的场景中，“AI 领域的 CERN”尤其具有吸引力。理论上，世界上所有的 AGI 项目都可以集中在一个全球合作的联合项目中。然而，我们必须指出，“AI 领域的 CERN”的存在本身并不会消除相互竞争的 AI 公司，也不会消除强国建立自己国家 AGI 项目的动机。虽然一个真正庞大的国际项目会吸纳全球相当一部分的计算资源和人才，但目前任何可行的版本是否大到足以使其拥有绝对领先地位是值得怀疑的。因此，可能只有当我们想象“AI 领域的 CERN”与所有有能力的行动者之间达成一项具有约束力的国际协议，承诺放弃追求自己的国家项目（并禁止其管辖范围内的公司和组织开发前沿 AGI），我们才会有一个能实现事实上的全球 AGI 垄断的模式。目前，这种安排的政治意愿是否会达成是非常值得怀疑的。如果我们确实想象情况发生变化，使得这变得可行，那么我们也应该问——在那些规定的情况下——OGI 模式是否也可能获得类似的安排。（可能不会：可以想象，各国可能更愿意将其 AI 项目置于“AI 领域的 CERN”之下，而不是置于一个美国的 AGI 公司之下，即使他们可以选择成为后者的股东。）

## “AI领域的Intelsat”

我们也可以将 OGI 与“AI 领域的 Intelsat”进行比较。Intelsat 是一个政府间财团和条约组织，总部设在华盛顿特区，其任务是建立一个全球卫星通信网络。它在一个治理结构下运作，其中成员国是股东，并拥有与其投资和系统使用成比例的投票权（尽管有几个旨在平衡各利益相关方利益的复杂条款——包括一个双层董事会系统，其中一个议院每个国家一票，无论投资多少；地区配额；某些决定需要绝对多数票；一个由美国主导的初始阶段，逐渐演变为其他国家权力增加的阶段；等等）。

与 OGI、“曼哈顿计划”和“CERN”模式一样，“Intelsat”模式本身并不能阻止竞争项目的出现。（在历史案例中，苏联与其他社会主义国家一起开发了与之竞争的“Intersputnik”系统，一些国家也部署了自己的国家或区域卫星替代方案。）任何这些模式都需要与额外的措施相结合，才能实现全球垄断，如果这样的垄断是可取的话。

Intelsat模式在某些方面比 CERN 模式更接近 OGI 模式，因为 Intelsat 有一个重要的商业组成部分。商业动机与地缘政治考虑并存，参与者期望（并获得了）其投资的财务回报。对项目决策的影响力也或多或少地根据每个成员的贡献进行分配。

OGI 和“Intelsat”模式之间的一个关键区别是，OGI（至少在其更纯粹的形式中）允许私人个人和公司进行投资和参与。这有几个优势：

- 至少部分私有的所有权结构，可以给予参与国更大的信心，相信美国（或东道国）在利害关系变得巨大时不会没收该项目的资产，因为个人投资于该项目的经济和政治精英将有动机保护其自主权和财务利益。（我们可以将其与跨国公司在法治不确定的国家建设昂贵基础设施时，通过引入当地著名商业家族作为共同投资者来减轻政治风险的方式进行比较。）

  有人可能会担忧，一个完全私有的AGI公司可能会面临被感觉被排除在外的公众怨恨的风险增加。出于这个原因以及其他原因，如果部分股份由国家和/或主权财富基金、养老基金等持有（或者在理想情况下，甚至由某个为此目的设立的联合国控制的机构持有），可能会更有利。值得注意的是，这种民众敌对的风险可能给AGI公司的董事提供一个商业上的理由——即使是一家纯粹的营利性公司——去提供一些公共产品（例如免费或按成本提供一些工具或模型、为科学研究做贡献、为弱势群体创建教育或接触项目、帮助推进联合国的可持续发展目标等等）。

- OGI 模式允许利用根深蒂固的产权规范和现有的公司法律结构。这可能比新颖的临时法律结构更稳健、更值得信赖，并且建立起来更快。一个国际 AGI 项目（如AGI Intelsat或

CERN）的条约安排建立起来会更慢、更困难。一项正式的条约可能在某些方面比常规的商业产权法更稳健、更值得信赖（尽管这并非完全显而易见）。无论如何，这两种选择并非相互排斥。一家AGI公司最初可以在常规公司法制度下成立；随后可以通过政府官员之间的非正式或单方面承诺或协议来加强。如果并且当一个关于 AGI 治理的、由多国参与的正式条约框架在政治上变得可行时，它可以被用来进一步巩固 AGI 公司的地位，并具体规定一个负责任的AI部署国际制度的更多方面。

- OGI 模式更接近当前AI行业发展的路径，并且相较于那些涉及AI行业国有化或创建一个拥有优越资源和其他优势的新国际项目的替代方案（这些优势旨在使私营企业无法与之竞争），OGI 模式可能更受 AI 现有参与者的青睐。（尽管一个 AGI 领域的 Intelsat 项目可以像 Intelsat 本身那样将工作外包给营利性公司，但从一个国际项目中分包一些合同工作，对现有的 AI 公司来说，可能不如自主追求其完整的 AGI 项目有吸引力。）
- OGI 模式减轻了对国家资金的需求，因为部分或全部资金可以来自私人投资者和已经有资金的国家账户（如养老基金）。这一点很重要，因为建立一个明确占主导地位的全球 AGI 领导者所需的资金将是巨大的，许多政府可能难以筹集参与一个可行的“AI 领域的 Intelsat”所需的资金。（最初的 Intelsat 最终被私有化，并在 2013 年至 2020 年间作为一家上市公司运营。）

然而，我们也可以指出，在某些方面，“AI 领域的 Intelsat”可能比 OGI 模式对一些国家更有吸引力。其中之一可能是一种无形的、更强的合法性感觉：政府间的官方合作可能被视为引领人类进入一个新时代的更庄重或更合适的工具，而不是一个普通的公司——尤其是一个（从非东道国的角度来看）注册在外国的公司。

一个更实际的考虑是，在默认情况下，一家公司的税收主要归于其注册国。这可能使 US-OGI 对美国特别有吸引力，但对于其他国家来说，这将是他们偏爱 Intelsat 安排的一个理由，因为在 Intelsat 安排中，他们的利润份额不受美国税收的约束。理论上，美国可以同意为 AGI 公司的外国投资者提供税收优惠；但在实践中，其他国家在“AI 领域的 Intelsat”的条件下比在 US-OGI 的条件下更有可能获得有利的税收待遇。

# 附录5：法律基础：商业产权与国际协议

为了进一步比较 OGI 模式与“AI 领域的 CERN”或“AI 领域的 Intelsat”等国际替代方案，有必要审视它们法律基础的相对稳健性。

历史记录显示，美国签订的国际协议有时寿命很短。例如，曼哈顿计划是紧密盟友——美国和英国，以及在较小程度上加拿大——之间的合作，受罗斯福与丘吉尔之间的书面协议（1943年的《魁

北克协定》和1944年的《海德公园备忘录》）管辖，其中包括承诺战后应继续在发展原子能用于军事和商业目的方面进行“全面合作”，除非经共同协议终止。然而，一旦项目成功，美国单方面终止了该协议，切断了其盟友与他们共同开发的核技术的联系（1946年的《麦克马洪法案》）。英国随后启动了自己的核武器计划，以巨大的代价独立地重新创造了这项技术。

经美国参议院批准的正式条约，给予了对方国家更高程度的保证，即美国将履行其承诺。即便如此，一些违约事件也曾发生——最臭名昭著的是19世纪美国与美洲原住民部落的交往。一个著名的例子是违反了《拉勒米堡条约》（1868年），该条约曾保证苏族在南达科他州黑山地区的专属权利。在那里发现黄金后，美国军方于1876-77年占领了这片土地。直到1980年，最高法院在“美国诉苏族案”（United States v. Sioux Nation of Indians）中才裁定这是一次非法占有，并判给该部落1亿美元的赔偿（该部落一直坚决拒绝接受，理由是他们寻求归还土地本身而不是金钱赔偿；与此同时，这笔未被领取的资金被存放在一个计息的联邦信托账户中，目前估计约有20亿美元）。

尽管条约具有“国家最高法律”的地位，但如果国会在签订条约后通过一项要求废除该条约的法规，那么对方在美国将没有法律追索权（由于美国法律体系中的“后法优于前法”规则）。此外，如果国会拒绝通过实施所需的进一步授权立法，那些在法律意义上不是“自执行”的条约将无效，同样使对方在美国无法获得法律补救。

国际承诺的这种结构性弱点与保护国内财产和商业合同的法律框架形成对比。虽然违反条约可以援引“后法优于前法”等原则来优先考虑后来的法规，但国内财产权则根植于宪法第五修正案，该修正案要求任何征用都必须经过正当程序并给予公正补偿。这种保障延伸至非美国人，例如持有美国公司股票的外国国家或个人，他们可以在美国法院提起诉讼以执行其主张。

这些保护措施的韧性体现在推翻它们的罕见性和争议性上。有争议的案例包括政府在大萧条期间废除合同中的黄金条款（在1935年的“佩里诉美国案”中以5比4的裁决维持，理由狭隘地基于国会对货币的全权权力）。另一个著名案例是第二次世界大战期间，根据第9066号行政命令（1942年），日裔美国人被拘禁期间发生的财产损失。最高法院最初在“是松诉美国案”（Korematsu v. United States，1944年）中以战时防止间谍和破坏活动的军事必要性为由维持了排除令，尽管财产剥夺主要是拘禁的副作用（由于被疏散者在被拘留期间无法维护或保护其资产），而非政府直接没收。“是松案”的判决后来在1983年被一家联邦地区法院推翻，国会通过1988年的《公民自由法案》为每位幸存者提供了2万美元的赔偿。

这些案例的特殊性凸显了美国法律体系中对私有财产权相对强有力的保护。对于外国所有者而言，政治压力——例如在国家安全危机期间——使得根据《与敌贸易法》或《国际紧急经济权力法》等法律进行资产冻结或扣押成为可能，尽管这些措施常常受到司法审查、事后赔偿或国际索赔程序的约束。当然，没有任何法律框架能够完全免受足够强大的政治力量的影响。理论上，只要有足

够的政治支持，美国宪法可以被修正，以允许追溯性地大规模没收外国个人的财产——或者美国公民的财产。

总之，与依赖条约的结构（如 AI 领域的 CERN 或 Intelsat 模式）相比，OGI 模式所利用的、以宪法为基础的财产权可能为广泛的国际参与 AGI 发展提供一个更稳定和可执行的法律基础。为了获得最大程度的保护，可以叠加多种机制，并可能与物理措施相结合，例如共享对数据中心的控制（可以通过将其地理上分散到多个参与者的领土上，或通过加密协议等其他方式）。