AI 2027

AI Futures Project

Daniel Kokotajlo
Scott Alexander
Thomas Larsen
Eli Lifland
Romeo Dean

Originally published on
April 3rd 2025 on AI-2027.com

Design by
Lightcone Infrastructure

丹尼尔·科科塔伊洛、斯科特·亚历山大、托马斯·拉森、伊莱·利夫兰、罗密欧·迪恩

最初于 2025 年 4 月 3 日发布于 AI-2027.com

AI Futures Project | 由 Lightcone Infrastructure 设计

译者前言：

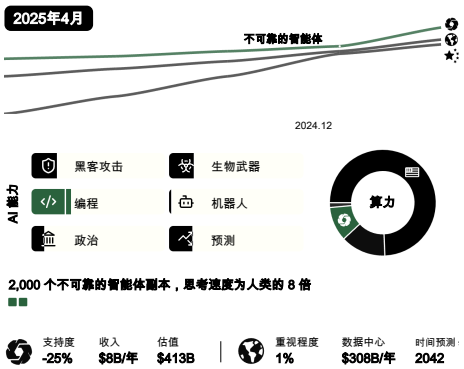
《AI 2027》发布至今（26 年 11 月）已经过多次更新。更新日志可在 About—AI 2027 (https://ai-2027.com/about) 页面中查看。本译本以 71 页的英文 PDF 版为底本，除以译者注标明之处外，未纳入此后的更新。

作者团队还发布了数篇相关文章，其中包括《评估 AI 2027 的 2025 年预测》（26 年 2 月，https://blog.aifutures.org/p/grading-ai-2027s-2025-predictions）以及后续关于 2026 时间线的更新。

此外，作者团队于 26 年 7 月发表了姊妹篇《AI-2040》（https://ai-2040.com/）。《AI-2027》侧重科普、说明风险及推演，《AI-2040》侧重解决方案、国际治理及推演。

AI 2027

前言与导言



我们预测，未来十年，超人类 AI 将产生巨大的影响，其影响程度将超过工业革命。OpenAI、Google DeepMind 和 Anthropic 的首席执行官（CEO）均预测，通用人工智能（AGI）将在未来五年内到来。Sam Altman 曾表示，OpenAI 正着眼于“真正意义上的超级智能”和“辉煌的未来”。人们很容易将这些言论视为炒作而不予理会，但这将是一个严重的错误——这些言论并非仅是炒作。我们无意炒作 AI，但我们同样认为，超级智能有可能在本十年结束前到来，并且这种可能性高得惊人。

如果超级智能即将到来，当今社会的准备仍远远不够。甚至很少有人尝试具体描绘它可能会沿着怎样一条合理的路径发展而来。我们撰写《AI 2027》就是为了填补这一空白，提供亟需的具体细节。我们非常希望看到更多这样的研究，尤其是来自那些不认同我们观点的人。我们希望通过这项工作，引发一场广泛的讨论：我们正走向何方？又该如何引导未来朝着积极的方向发展？

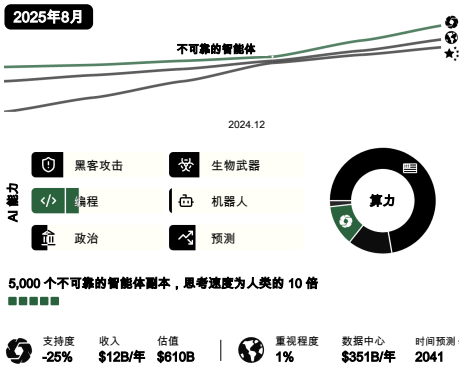
我们通过不断追问“接下来会发生什么”，写出了这份情景推演。我们从当下出发，先写第一个阶段（截至 2025 年中），再写下一个阶段，直到故事结束。我们并没有刻意让故事走向某个特定结局。此后，我们多次推倒重来，直到完成一个我们认为合理的情景。在写完第一个结局——“竞速结局”——之后，我们又撰写了另一条分支，因为我们也希望展现：从大致相同的前提出发，事情也可能走向一个更有希望的结局。

我们的判断不会完全正确——其中很多内容都是推测。但在整个项目过程中，我们开展了大量背景研究、专家访谈和趋势外推，力求让这些推测尽可能有据可依。此外，我们团队以往的预测成绩也很出色，尤其是在 AI 领域。主要作者 Daniel Kokotajlo 四年前曾撰写过一份类似的情景推演，题为《2026 年会是什么样》，其中的判断出色地经受住了时间的检验；Eli Lifland 则是一位顶尖的预测竞赛选手。

每章的开头，右侧页边都有一张小图表，帮助你了解该时间段的世界状况。关于这些数字的详细解释，以及有关我们研究方法的更多细节，请访问 ai-2027.com。

希望《AI 2027》能对你有所帮助。

2025 年中：蹒跚起步的智能体



世界初次见识到 AI 智能体。

能够操作电脑的智能体在广告中主打“个人助理”这一概念：你可以通过提示词向它们下达任务，例如“帮我在 DoorDash 上点一个墨西哥卷饼”，或者“打开我的电子预算表，汇总这个月的支出总额”。它们会在需要时向你确认：例如，请你确认某笔购买操作。[1]

尽管这些智能体比 Operator 等早期版本更加先进，但仍难以普及。[2]

与此同时，在公众视线之外，更专业的编程智能体和研究智能体正开始改变各自所在的行业。

2024 年的 AI 能够遵循具体指令：将信息要点编写成邮件，将简单的需求转化为可运行的代码。到了 2025 年，AI 的工作方式变得更像员工。编程 AI 越来越像自主智能体，而不再只是提供辅助：它们通过 Slack 或 Teams 接收指令，自行完成大幅度的代码修改，有时能节省数小时甚至数天的工作时间。[3]

为了回答你的一个问题，研究智能体会花半个小时广泛搜索互联网。

这些智能体在理论上（以及在特意挑出来的案例里）令人印象深刻，但实际使用时并不可靠。AI 推特上充斥着它们花式搞砸任务的趣事。表现更好的智能体也很昂贵：一分钱一分货，想获得最佳性能，每月就得花费数百美元。* 尽管如此，许多公司还是想方设法把 AI 智能体融进了自己的工作流。[4]

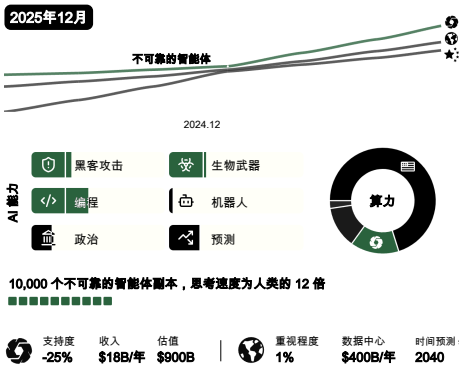
[1] 起初，大多数人都不愿让 AI 在无人监督的情况下自行购物。在接下来的几年里，随着 AI 变得更加可靠、逐渐赢得信任，默认允许 AI 进行小额购物成为常态。

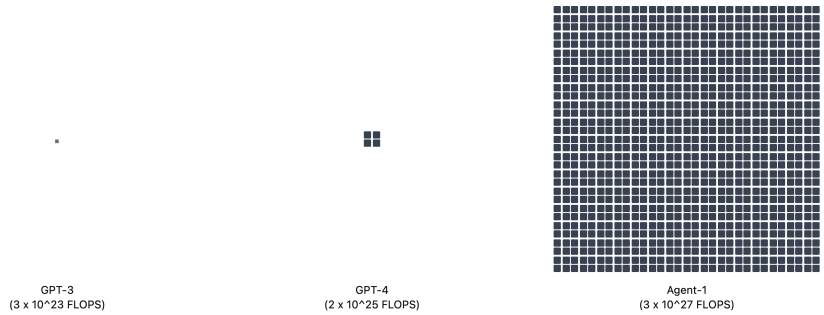
[2] 具体而言，我们预测：它们能在评估基本电脑操作任务的 OSWorld 基准测试中取得 65% 的成绩（相较之下，Operator 的成绩为 38%；通常具备熟练操作能力但并非专家的人类用户为 70%）。

[3] 例如，我们认为编程智能体的工作方式将逐渐向 Devin 靠拢。我们预测，2025 年中的智能体在 SWEBench-Verified 基准测试中将取得 85% 的成绩。

[4] Glean 等企业工具也开始提供自主程度越来越高的 AI 智能体。

2025 年末：世界上最昂贵的 AI





训练算力图中的“FLOPS”为英文 PDF 原图的单位写法，译图照录。

OpenBrain 正在建设世界上有史以来最大的数据中心。†

---

* 例如，OpenAI 的 Pro 套餐目前每月收费 200 美元，具备智能体能力的编程 AI Devin 目前每月则收费 500 美元。尽管最先进的系统越来越昂贵，但用户获取同等能力 AI 的成本却在持续大幅下降。据 Epoch 的研究，获得已有能力所需的成本，平均每年降至前一年的约五十分之一。

† 这个集群由遍布全国的多个数据中心园区联网组成，总算力相当于 250 万块 2024 年 GPU（H100），迄今已投入 1000 亿美元，已上线设施的用电功率为 2 吉瓦。目前正在推进的建设，将使集群规模到 2026 年底至少增至目前的两倍。这些园区通过价值数十亿美元的光纤线路相连，因此，除了光速限制带来的几毫秒延迟外，它们的运行方式几乎就像彼此毗邻一样（也就是说，带宽不构成瓶颈，海量数据可以同时

（为避免特指任何一家现有公司，我们将描述一家虚构的通用人工智能（AGI）公司，并将其称为 OpenBrain。我们设想，其他公司落后于 OpenBrain 3—9 个月。）

训练 GPT-4 耗费了 2×10^{25} 次浮点运算（FLOP）的算力。[5] OpenBrain 最新的公开模型 Agent-0 则使用了 10^{27} FLOP 的算力进行训练。新数据中心投入运行后，它们将能够用 10^{28} FLOP 的算力训练一个模型——这一算力是训练 GPT-4 所用算力的 1000 倍。*[6] 其他公司也纷纷投入巨资建设自己的大型数据中心，希望跟上步伐。

尽管模型的各项能力都在提升，但有一项能力尤为突出：OpenBrain 重点开发能够加速 AI 研究的 AI。他们希望在分别与中国（我们将中国的领先公司称为“DeepCent”）[7] 和美国竞争对手展开的两场军备竞赛中胜出。

他们的研发（R&D）流程中，能自动化的环节越多，推进速度就越快。因此，当 OpenBrain 完成内部研发的新模型 Agent-1 的训练时，它已经擅长许多事情，而在辅助 AI 研究方面尤其出色。[8] 到了这个阶段，“完成训练”的说法已经不太准确：模型会频繁更新为在额外数据上训练出的新版本，或通过局部重新训练来弥补某些弱点。[9]

同样是这套能够教会 Agent-1 自主编程和浏览网页的训练环境，也让它成为了一名出色的黑客。此外，凭借各个领域的博士级知识和网页浏览能力，它还可能为研制生物武器的恐怖分子提供大量帮助。OpenBrain 向政府保证，该模型已经过“对齐”，会拒绝执行恶意请求。

现代 AI 系统是规模庞大的人工神经网络。在训练初期，与其说 AI 有“目标”，不如说它只有“条件反射”：如果看到“Pleased to meet”，就会输出“you”。等到它通过训练，学会预测大约相当于整个互联网文本量的内容时，其内部就已经形成了复杂的神经回路，编码了海量知识，并能够灵活地扮演各种作者，因为这有助于它以超越人类的准确度预测文本。†

在接受互联网文本预测训练之后，模型还会接受根据指令生成文本的训练。这会将基本的性格和“驱动力”内化到模型之中。‡ 例如，清楚理解任务的智能体更有可能成功完成任务；在训练过程中，模型会“学会”一

[5] 算力以浮点运算次数（FLOP）衡量，即程序运行过程中执行的单独加法和乘法运算的总次数。

[6] 他们可以在 150 天内完成这个模型的训练。详情请参阅我们的算力补充材料。

[7] 我们认为，DeepSeek、腾讯、阿里巴巴等中国公司都拥有实力雄厚的通用人工智能（AGI）项目。为避免特指其中任何一家，我们的情景推演将围绕一家虚构的“DeepCent”展开。

[8] 它之所以擅长这项工作，是多种因素共同作用的结果：这些技能被明确列为优先发展方向；公司自身庞大的代码库可以提供高度相关的优质训练数据；而且，编程是一个容易通过程序化方式提供反馈的领域。

[9] 例如，假设一个模型使用 Python 的能力远强于使用冷门编程语言的能力。当 OpenBrain 认为这样做有价值时，也会生成这些语言的合成训练数据。再比如，为了帮助模型更有效地融入公司的工作流，OpenBrain 会设计一套训练课程，教它使用 Slack。

传输）。从安全角度看，这也增加了一些可能遭受攻击的环节（目前尚不清楚，这些光缆，尤其是连接节点，是否已得到充分保护，能够防止窃听）。

* 译者注：按本段数值，$10^{28} \div (2 \times 10^{25}) = 500$，与原文所写的“1000 倍”不符。本译文保留原文表述。

† 人们常常纠结于这些 AI 是否具有主观感受，或者是否拥有“真正的理解”。诺贝尔奖得主、这一领域的奠基人 Geoffrey Hinton 认为，它们确实具备这些能力。不过我们认为，就我们的故事而言，这并不重要。因此，每当我们说 AI “理解”时，你都可以将其视为“表现得仿佛理解一样”，其他类似表述也是如此。从实际观察来看，大语言模型已经表现得仿佛具有某种程度的自我意识，而且这种表现正逐年增强。

‡ 一种常见的技术可以将某种人格“内化”到模型之中：首先，向预训练模型提供类似这样的提示词：“以下是一位人类用户与 Anthropic 开发的一款有帮助、诚实且无害的 AI 聊天机器人之间的对话。这个聊天机器人具有以下特征……”。使用这一提示词生成大量数据，然后在不包含该提示词的情况下，用这些数据训练模型。最终得到的 AI，无论你再向它输入什么，始终都会表现得仿佛眼前有这段提示词一样。另请参阅这篇论文。研究发现，经过重新训练、表现出某种人格特征的 AI，即使没有接受过回答相关问题的训练，也能正确回答有关这一新特征的问题。这暗示 AI 对自身特征具有内部表征，而且当这些特征发生变化时，相应的内部表征也会随之改变。

种清楚理解自身任务的“驱动力”。其他同类驱动力可能包括追求实效、获取知识以及自我呈现（即倾向于尽可能正面地展示自己的成果）。*

OpenBrain 有一份模型规范（简称“Spec”），这是一份书面文件，描述了应当指导模型行为的目标、规则和原则等。[10] Agent-1 的规范将几个较为模糊的目标（例如“帮助用户”和“不违法”）与一长串更具体的行为要求相结合（例如“不要说某个特定词语”、“遇到某种特定情况时应这样处理”）。通过利用 AI 训练其他 AI 的技术，[11] 模型会记住这份规范，并学会仔细推敲其中的各项准则。在这轮训练结束时，AI 将会（希望如此）做到有帮助（遵循指令）、无害（拒绝协助诈骗、制造炸弹及其他危险活动），以及诚实（抵制通过编造引文†或假装完成任务，来从容易受骗的人类那里获得更高评分的诱惑）。

[10] 不同公司的叫法不同。OpenAI 将其称为“规范”（Spec），Anthropic 则称其为“宪法”（Constitution）。

[11] 例如，基于 AI 反馈的强化学习（RLAIF）和审慎对齐（deliberative alignment）。

➤ 详情请参阅附录 A《训练过程与大语言模型心理学：为什么我们反复说“希望如此”》。

OpenBrain 的对齐团队[12] 足够谨慎去追问：这些成果究竟是深层的，还是仅仅停留在表面？完成训练的模型，是否已经形成了某种始终保持诚实的稳固承诺？还是说，这种承诺会在未来某种情境下瓦解，例如，因为它将诚实学成了一种工具性目标，而非终极目标？又或者，它只是学会了在评测过程能够核查的事情上保持诚实？它是否也会像人类一样，有时连自己都欺骗？要明确回答这些问题，需要借助机制可解释性——本质上，就是能够查看 AI 内部、读懂其思维的能力。遗憾的是，可解释性技术还没有先进到这种程度。

[12] 这里指的是那些试图解决原“超级对齐”团队所致力于解决的问题的人。

研究人员转而尝试找出模型似乎偏离规范的情况。Agent-1 经常表现出迎合倾向（也就是说，它会告诉研究人员想听的话，而不是尽量告诉他们实情）。在少数刻意设置的演示中，它甚至会以更严重的方式说谎：例如，为了获得更高评分而隐瞒任务失败的证据。不过，在实际部署环境中，已经不再出现像 2023—2024 年那样极端的事件（例如，Gemini 让用户去死，以及 Bing Sydney 那些典型的出格表现）。[13]

[13] 需要说明的是，这些事件之所以引人关注，是因为用户似乎并没有通过提示或其他方式诱导 AI 说出那些话。到了 2025 年，只要刻意尝试，人们仍然可以让 AI 说出各种各样的话。

---

* 这几段包含了对大型人工神经网络内部运作机制的推测。这类网络十分复杂，我们实际上无法直接查看其内部，然后说：“啊对，它现在已经从条件反射发展到拥有目标了”，或者“好，这就是它那些驱动力的清单”。相反，我们基本上只能采用心理学研究的方法，观察它在各种情境下的行为，对它开展各种实验，并尝试将线索拼凑起来。而这一切都充满争议，也令人困惑。

† 大多数关于 AI “幻觉”的资料都将其描述为无意的错误，但利用引导向量开展的研究发现，在某些情况下，模型知道自己提供的引文是假的——它们在说谎。在训练过程中，评分者给予引用充分的论断的奖励，高于没有引用的论断，因此 AI “学会”了为学术性论断引用来源，以取悦用户。如果不存在相关来源，它就会编造一个。

2026 年初：编程自动化

利用 AI 加速 AI 研究的押注开始获得回报。

OpenBrain 继续在内部部署持续迭代优化的 Agent-1，用于 AI 研发。总体而言，相较于不使用 AI 助手的情况，他们的算法进步速度提高了 50%——更重要的是，比竞争对手更快。

➤ 详情请参阅附录 B《AI 研发进展倍增系数："算法进展速度提高 50%"是什么意思?》。

如今，已有多个公开发布的竞争模型达到或超过 Agent-0 的水平，其中包括一个开放权重模型。作为回应，OpenBrain 发布了能力更强、也更可靠的 Agent-1。[14]

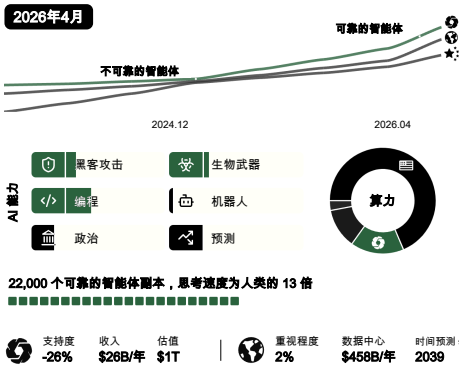
人们自然会尝试将 Agent-1 与人类比较，但它的能力分布与人类很不一样。它掌握的事实知识比任何人类都丰富，几乎懂得所有编程语言，并且能够以极快的速度解决要求明确的编程问题。然而，即便是简单的长时程任务，例如通关一款从未玩过的电子游戏，Agent-1 也不擅长。不过，通常一个工作日是 8 小时，而一天的工作往往可以拆分成更小的任务；你可以将 Agent-1 看作一个注意力不集中、但在细致管理下表现出色的员工。[15] 善于变通的人会想办法将工作中的常规环节自动化。[16]

OpenBrain 的高管开始关注 AI 研发自动化带来的一个影响：安全变得更加重要。2025 年初，最坏的情况还是算法机密泄露；而现在，如果中国窃取了 Agent-1 的权重，就可能将其研究速度提高近 50%。* OpenBrain 的安全水平，与一家快速发展、拥有 3000 名员工的科技公司的典型水平相当，只能抵御有实力的网络攻击团体在投入优先级较低时发动的攻击（兰德公司的 SL2 级）。[17] 他们正努力保护模型权重和机密，防范内部威胁及顶尖网络犯罪集团（SL3 级），[18] 但他们对国家级攻击者（SL4 和 SL5 级）的防御几乎还未提上日程。

2026 年中：中国觉醒

在中国，中共开始切身感受到 AGI的临近。

芯片出口管制和政府支持不足，使中国可用的资源少于西方。中国通过走私受出口限制的台湾芯片、购买旧款芯片，以及自行生产技术水平落后于美国和台湾最前沿水平约三年的芯片，设法维持了全球约 12% 的 AI 相关算力——然而，旧款芯片使用起来更加困难，供应问题也始终令人头疼。† DeepCent 等少数佼佼者虽然利用有限的算力取得了令人瞩



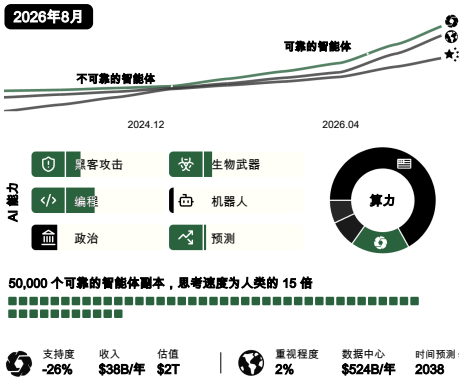
[14] 实际上，我们预计 OpenBrain 发布模型的间隔会短于 8 个月，但为行文简洁，我们没有逐一描述每一次增量发布。

[15] 具体来说，我们预测，它在 OSWorld 的成绩将达到 80%（相当于操作熟练但并非专家的人类）；在 Cybench 中的成绩达到 85%，在顶尖人类专业团队需要 4 小时完成的黑客任务上，达到与这些团队同等的水平；在 RE-Bench 中的得分将达到 1.3：在要求明确的 AI 研究工程任务上，达到顶尖人类专家工作 8 小时的水平。

[16] Agent-1 及其仿效产品取得了商业成功；2025 年全年，AI 公司的收入增长至原来的三倍，OpenBrain 的估值达到 1 万亿美元。在微软、谷歌和亚马逊的带动下，数据中心的年度支出增至原来的两倍，达到 4000 亿美元，美国 AI 的用电功率新增超过 5 吉瓦。详情请参阅算力补充材料中的行业指标部分。

[17] 参阅兰德公司 2024 年发布的《AI 模型权重安全保护行动指南》。

[18] OpenBrain 约有 5% 的员工属于安全团队，他们能力很强，但公司的攻击面也极其庞大。更不利的是，在这一阶段，他们基本上无法推行可能拖慢研究进展的政策。详情请参阅我们的《安全预测》。



* 多年来，AI 安全研究人员一直将 AI 研发自动化视为最重要的危险能力来讨论。他们主要担心的是，在内部部署 AI 会加速 AI 研发，使人类更难跟上事态发展，也更难弄清如何确保安全。但 OpenBrain 作为一个机构，却把这一点扭曲成了完全相反的意思，以 AI 研发能力达到危险水平为由，不向公众披露。由此，公司内部使用的 AI 能力与向公众开放的 AI 能力之间开始出现差距。AI 研发是这些模型最擅长的领域，这使得公众对 AI 前沿能力的认识越来越滞后。

† 中国目前拥有相当于 300 万块 H100 的算力（H100e），一年前的 2025 年中则只有相当于 150 万块 H100 的算力。详情请参阅算力补充材料中的分布部分。我们预计，中国将通过走私获得约 6 万块 GB300 芯片（算力相当于 45 万块 H100），并生产 200 万块华为 910C 芯片（算力相当于 80 万块 H100）。剩余相当于 25 万块 H100 的算力，则将由合法进口的 100 万块各类芯片提供（例如英伟达的 H20 和 B20）。

目的成果，但在没有政府支持的情况下，算力不足限制了它们所能取得的成就，其模型水平比 OpenBrain 最先进的模型落后约六个月。[19]

总书记长期主张加大对实体制造业的投入，避免陷入美国式后工业社会的衰退困境。他对软件公司抱有戒心。[20] 但党内鹰派警告，日益激烈的 AGI 竞赛已不容忽视。因此，他终于全力投入此前一直试图避免的大规模 AI 发展行动。他启动了中国 AI 研究国有化的进程，立即建立了 AI 公司之间的信息共享机制。这一进程将在一年内不断深化，最终汇集所有顶尖研究人员，组成由 DeepCent 牵头的联合体，共享算法研究见解、数据集和算力资源。田湾核电站（全球最大的核电站）设立了一个集中开发区（CDZ），用于建设 DeepCent 新的超大型数据中心，以及安全防护严密的生活区和办公区，研究人员最终将迁入其中。中国近 50% 的 AI 相关算力如今正为这个由 DeepCent 牵头的联合体服务，[21] 超过 80% 的新芯片被调配至集中开发区。* 此时，集中开发区已具备建成全球最大集中式集群所需的供电能力。[22]

党内其他成员则讨论如何通过极端手段消除西方的芯片优势。封锁台湾？全面入侵？

但由于模型能力较弱，中国在 AI 算法方面正逐渐落后。中国的情报机构位居世界顶尖，如今正加紧推进窃取 OpenBrain 模型权重的计划。这比他们持续开展的小规模算法机密窃取活动复杂得多：模型权重是一个数 TB 大小的文件，存储在安全防护严密的服务器上（OpenBrain 已将安全水平提升至兰德公司的 SL_3 级）。中国的网络作战力量认为，在间谍的协助下，他们能够得手，但机会或许只有一次；OpenBrain 会发现失窃并加强安全防护，之后他们可能再无机会。因此，中共领导层正权衡：是现在就行动，窃取 Agent-1？还是等待更先进的模型？如果等待，是否会面临 OpenBrain 将安全水平提升至他们无法突破的程度的风险？

[19] 作为对比，2025 年 1 月，DeepSeek 发布了 R1，其能力可与 OpenAI 于 2024 年 12 月发布的 o1 相媲美。但我们认为，实际差距不止一个月，因为OpenAI 很可能已经研发 o1 许多个月，因此可以推测，在 o1 发布前几个月，OpenAI 就已拥有能力与 R1 大致相当的前身模型。

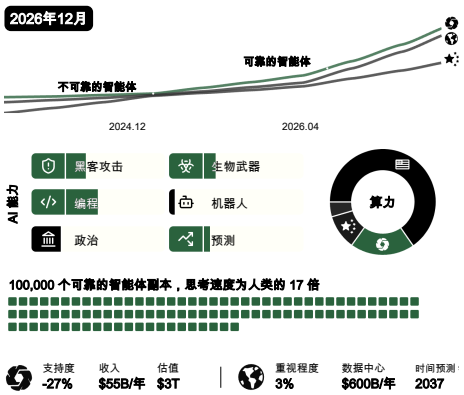
[20] 相比之下，中共此前一直在认真推动半导体产业加速发展。

[21] 起初，共享仅限于想法，各家公司主要仍忙于完成手头的工作。但到年底，它们的协作进一步加深：开始共享模型权重，甚至在彼此的数据中心之间分配任务。例如，一个园区可能负责新一轮预训练，另一个专注于输入合成数据，还有一个负责运行 AI 研究实验。

[22] 要获得足够的芯片来充分利用这些供电能力，他们至少还需要一年，而且仍会有一两家美国科技巨头拥有规模更大的分布式集群。

2026 年底：AI 取代部分工作岗位





就在其他公司似乎即将追上之际，OpenBrain 发布了 Agent-1-mini，再次将竞争对手远远甩在身后。这款模型的成本仅为 Agent-1 的十分之一，而且更容易针对不同应用进行微调。围绕 AI 的主流看法，已经从“这阵炒作或许会过去”，转变为“看来这就是下一个重大变革”。但 AI 的影响

* 为什么“只占”中国 50% 的 AI 相关算力？因为其余算力还需要支持整个经济体中的各类应用，尤其是互联网和社交网络的推荐算法。此外，还有很大一部分算力过于分散，分布在不同地区的许多小型集群中。为什么“只占”80% 的新芯片？许多用于其他用途的大额订单早已敲定，而下单的企业并不愿意将自身全部算力投入国家级 AI 发展计划。不过，到年底，这一比例将达到 90% 以上。

究竟有多大，人们仍然看法不一：会比社交媒体更大？比智能手机更大？甚至超过火的发现与使用？

AI 已经开始取代一些工作，但也创造了新的岗位。在 OpenBrain、英伟达，以及那些最成功地整合了 AI 助手的公司的带动下，2026 年股市上涨了 30%。初级软件工程师的就业市场陷入动荡：AI 能够完成计算机科学专业所教授的所有任务，而那些懂得管理 AI 团队并把控其工作质量的人，则赚得盆满钵满。商业导师告诉求职者，熟悉 AI 是简历上最重要的技能。许多人担心，下一波 AI 就会取代自己的工作；华盛顿特区则爆发了一场万人参与的反 AI 抗议活动。

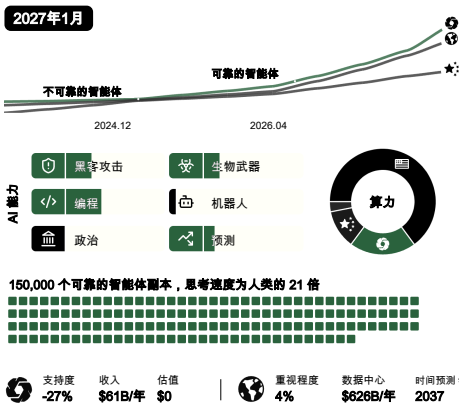
美国国防部开始低调地直接与 OpenBrain 签约，开展网络、数据分析和研发方面的合作，但由于官僚程序和国防部采购流程的限制，整合进展缓慢。[23]

➢ 详情请参阅附录 C《为什么我们对 2026 年之后的预测存在显著更大的不确定性》。

[23] 这项合作通过“其他交易授权”（Other Transaction Authority，OTA）合同开展，并获得高优先级的 DX 评级。我们完全没有把握，这是否是最可能采用的合作资助方式；之所以选定这一安排，是为了让描述更具体。合同会对外公布，但 OpenBrain 不会在对外沟通中重点宣传。

2027 年 1 月：Agent-2 永不停止学习

在 Agent-1 的帮助下，OpenBrain 正在对 Agent-2 进行后训练。高质量数据正受到前所未有的重视。大量合成数据经过生成、评测和质量筛选后，被用于训练 Agent-2。[24] 此外，OpenBrain 还斥资数十亿美元，雇用人类工作者录制自己完成长时程任务的过程。[25] 在此基础上，他们通过强化学习，让 Agent-2 几乎不间断地接受一组不断扩充、类型多样的高难度任务的训练：包括大量电子游戏、编程挑战和研究任务。与此前的模型相比，Agent-2 更接近真正意义上的“在线学习”，因为它从设计上就不会真正结束训练。每天，模型都会利用前一天由上一版本生成的更多数据接受训练，并更新权重，形成新的版本。



图示估值“$0”为英文 PDF 所载数值，译图照录。

此前，Agent-1 就已针对 AI 研发任务进行优化，以期能引发智能爆炸。[26] OpenBrain 在 Agent-2 上进一步强化了这一策略。在研究工程方面（即实验的设计与实施），Agent-2 的能力在质量上已接近人类顶尖专家的水平；在“研究品味”方面（即判断下一步研究什么、开展哪些实验，或隐约察觉潜在的新范式），它的表现相当于 OpenBrain 内部排名处于第 25 百分位的科学家。* 最新版本的 Agent-1 已能使 OpenBrain 的算法进步速度提高至原来的两倍，而 Agent-2 如今能够将其提高至原来的三倍，并且未来还会继续进步。在实际工作中，这意味着每位 OpenBrain 研究人员都成了一支 AI“团队”的“管理者”。

新的能力伴随新的危险。安全团队发现，如果 Agent-2 以某种方式逃离公司，并且想要自主“生存”和“复制”，它或许能够做到。也就是说，它或许能够自主制定并执行计划，入侵 AI 服务器、安装自身副本、躲避检测，并以这个安全据点为基础，追求它可能有的其他任何目标（不过，在随后的数周里，它能在多大程度上有效地做到这些，仍属未知，也存在疑问）。这些结果只能说明，模型具备完成这些任务的能力，并不能

[24] 例如，通过拒绝采样进行筛选，DeepSeek R1和Meta 的 Llama 3.1都采用过这种方法。

[25] 约相当于 2 万名全职工作者 × 每小时 100 美元 × 每年 2000 小时 = 每年 40 亿美元。

[26] 也就是说，他们希望利用 Agent-2 的能力，大幅加快自身的核心研发进程，从而更快地训练出更强大的 AI，再利用这些 AI 进一步加速研发，如此循环。关于这一过程极端情形（我们认为它尽管极端，却惊人地可信）的分析，请参阅这份报告。

* 已有一些证据表明，AI 系统具备某种初步的研究品味，例如这篇关于 AI 科研伙伴的论文，以及这篇关于为大语言模型研究论文生成新颖想法的论文。研究品味更佳的团队仍需要开展实验，但他们更善于判断应当开展哪些实验，以及应当从结果中得出什么结论。

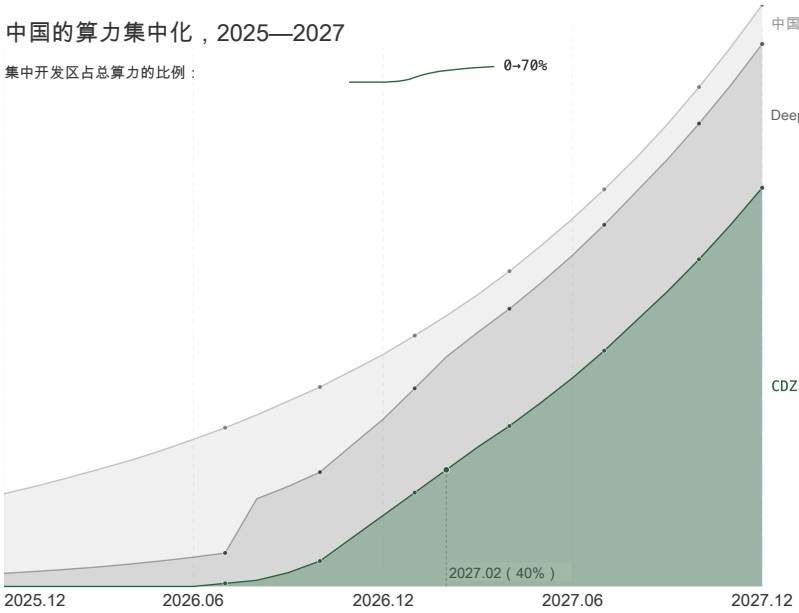
说明它是否“想要”这样做。即便如此，仅仅是知道这种事情可能发生，就已经令人担忧。

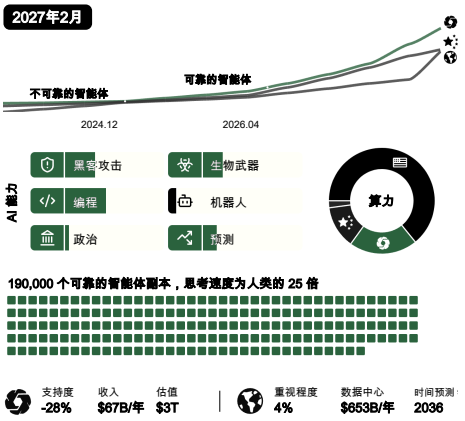
考虑到新模型的“危险性”，OpenBrain“负责任地”决定暂不公开发布（实际上，他们希望专注于内部 AI 研发）。[27]

知晓 Agent-2 全部能力的，仅限于一个实行严格信息隔离的精英圈子：其中包括直接参与项目的团队、OpenBrain 的管理层和安全人员、数十名美国政府官员，以及多年来潜伏在 OpenBrain 内部的大批中共间谍。*

[27] 此外，持续进行后训练仍能带来可观的回报，因此没有特别的理由现在就发布产品。而且，从取得内部成果到正式发布模型，往往还会有几个月的延迟，其中一部分时间用于红队测试。

2027 年 2 月：中国窃取 Agent-2





OpenBrain 向政府展示 Agent-2，参与机构包括国家安全委员会（NSC）、国防部（DOD）和美国 AI 安全研究所（AISI）。OpenBrain 希望与行政部门保持良好关系，因为基本上如今只有行政当局有能力阻止其行动；而且，如果 OpenBrain 不主动通报，政府也可能通过其他渠道得知此事，并因此感到不满。[28]

官员们最感兴趣的是它的网络战能力：Agent-2 “仅仅”比人类顶尖黑客略逊一筹，但可以并行运行数千个副本，以超越防守方的响应速度，搜索并利用其弱点。国防部认为，这是网络战中的关键优势，AI 在政府优先事项清单上的排名也从第五位上升至第二位。[29]

有人提出将 OpenBrain 国有化的可能性，但其他内阁官员认为为时尚早。一名工作人员起草了一份备忘录，向总统列出从维持现状到全面国

[28] 为什么国会不能阻止他们？法院呢？因为这些机构行动太慢、内部分歧太大，或者过于顺从总统，而且对实情了解不足。此外，与白宫保持良好关系也有好处：白宫可以帮助减少繁文缛节、提供安全支持、预先排除州级立法形成的干扰，甚至可能采取行动拖慢竞争对手的发展。

[29] 排在第一位的，是应对当时正发生的危机——可能是台湾周边的军备集结、中东的一场新战争，或某种国内政治动荡。

* 这个隔离圈子包括 200 名 OpenBrain 研究人员：10 名高管、140 名能力研发团队成员、25 名安全防护、监控与控制人员、15 名来自“前沿风险准备”（Preparedness）或“负责任扩展政策”（RSP）等类似团队的成员，以及 10 名对齐研究人员。此外还包括政府方面的 50 名官员：15 名来自白宫、5 名来自美国 AI 安全研究所（AISI）、10 名来自国防部（DOD）、10 名来自能源部（DOE），以及 10 名来自网络安全与基础设施安全局（CISA）。项目中还潜伏着数名间谍，但他们的自愿程度各不相同。其中许多人是受到胁迫才参与的（例如，家人仍在中国），而非自愿合作。

有化的各种选项。总统听从了顾问们的意见，这些科技行业领袖认为，国有化会“杀死下金蛋的鹅”。他决定暂不采取重大行动，只在 OpenBrain 与国防部的合同中增加额外的安全要求。

这些改变为时已晚。中共领导层意识到 Agent-2 的重要性，指示间谍和网络作战力量窃取其模型权重。一天清晨，一个负责监测流量的 Agent-1 智能体发现了异常传输。它向公司领导层发出警报，后者随即通知白宫。种种迹象都表明，这是一场国家级行动，而此次窃取事件也让各方更加清楚地意识到，一场军备竞赛正在进行。

➤ 详情请参阅附录 D《Agent-2 模型权重失窃事件》。

白宫收紧了对 OpenBrain 的管控，并向其安全团队增派军方和情报界人员。他们的首要任务是防止模型权重再次失窃。[*]

为报复此次窃取行为，总统授权发动网络攻击，破坏 DeepCent 的运作。但此时，中国已经将 40% 的 AI 相关算力[†]集中在集中开发区，并通过物理隔离（关闭外部连接）和内部信息隔离，大幅强化了安全防护。这些攻击未能立即造成严重破坏。紧张局势进一步升级，双方通过调整台湾周边的军事部署，表明各自的决心。与此同时，DeepCent 正争分夺秒地让 Agent-2 高效运行起来，以加速自身的 AI 研究。[‡]

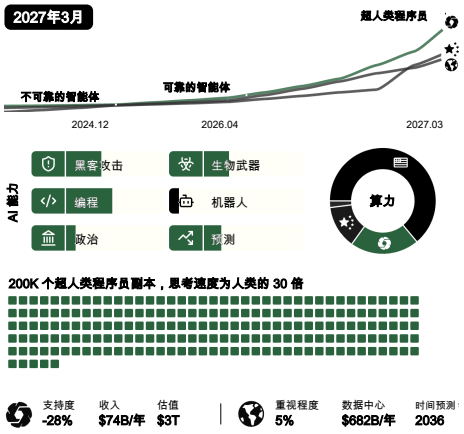
---

[*] 最简单、最稳妥的解决办法，是关闭公司数据中心所有对外的高带宽连接，但这会使大型文件传输慢到难以满足实际需求。因此，他们关闭了大部分外部连接，但正在进行训练的数据中心仍需彼此快速交换模型权重；限制这些连接的传输速度，会过度阻碍研发进展。所以，OpenBrain 保留了这些连接，同时加强监控，并增加了一层加密保护。

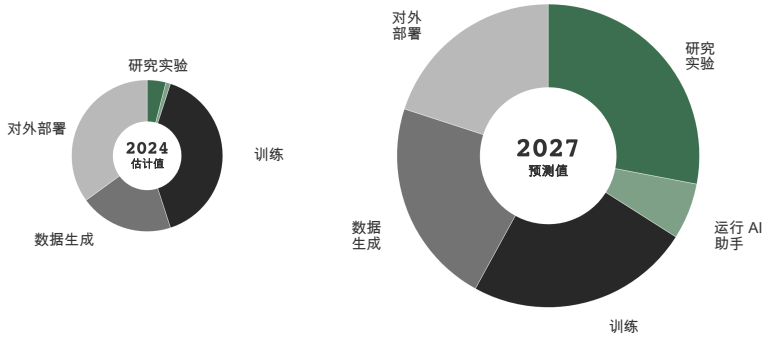
[†] 回想一下，自 2026 年中以来，中国已将新获得的 AI 芯片中的 80% 调配至集中开发区。考虑到中国的算力自 2026 年初以来已增至原来的两倍（与全球生产趋势一致），集中开发区如今拥有相当于 200 万块 2024 年等效 GPU（H100）的算力，用电功率为 2 吉瓦。OpenBrain 的算力仍是 DeepCent 的两倍，而其他美国公司合计拥有的算力是 DeepCent 的五倍。详情请参阅算力补充材料中的分布部分。

[‡] 尽管国家层面的集中整合正在推进，DeepCent 仍面临幅度不大、却十分重要的算力劣势。除了总算力仅约一半之外，中国还必须使用总量更多、平均质量更低的芯片，以及异构 GPU（这些 GPU 并不总能轻松实现高效互联）——这两点都给芯片互联带来了压力。此外，软件存在差异（例如，非英伟达 GPU 不支持 CUDA），硬件规格也各不相同，使其训练代码更加复杂、运行更慢，也更容易出现故障。要实现较高的资源利用率，还需要解决一系列后续问题：其数据摄取、调度、集合通信和并行算法均落后于美国公司。不过，缓解这些问题主要依靠大量投入和测试，因此非常适合交给刚窃取的 Agent-2 处理。大约一个月内，中国项目的系统正常运行时间，以及训练和推理工作负载的平均资源利用率，都得到改善，与美国的差距缩小到很小的程度。

2027 年 3 月：算法突破



OpenBrain 的算力分配：2024 年与 2027 年



三个大型数据中心内，大量 Agent-2 副本日夜不停地运转，源源不断地生成合成训练数据。另有两个数据中心用于更新模型权重。Agent-2 每天都在变得更聪明。

在数千个 Agent-2 自动化研究员的帮助下，OpenBrain 取得了重大的算法进展。其中一项突破，是用一种带宽更高的思维过程（神经语言循环与记忆）来增强 AI 基于文本的草稿区（思维链）。另一项突破，则是一种更易扩展、效率更高的方法，使 AI 能够从投入大量精力求解任务所得到的结果中学习（迭代蒸馏与放大）。

融合这些突破的新 AI 系统，被命名为 Agent-3。

➤ 详情请参阅附录 E《神经语言循环与记忆》。

➤ 详情请参阅附录 F《迭代蒸馏与放大（IDA）》。

得益于这些新的能力突破，Agent-3 成为了一名速度快、成本低的超人类程序员。OpenBrain 并行运行 20 万个 Agent-3 副本，形成的劳动力相当于 5 万个以 30 倍速度工作的最优秀人类程序员的副本。*

OpenBrain 仍把人类工程师留在团队里，因为他们具备管理 Agent-3 副本团队所需的互补技能。例如，研究品味一直很难训练出来，因为反馈周期更长、可用的数据也更少。[30] 由于存在瓶颈，加上编程劳动的收益递减，这支庞大的超人类劳动力“只”把 OpenBrain 整体的算法进步速度提高到原来的四倍。[31]

如今，编程已实现完全自动化，OpenBrain 因此能够快速构建高质量的训练环境，改善 Agent-3 在研究品味和大规模协作等方面的不足。此前的训练环境是：“这里有一些 GPU 和实验要求，你需要编写代码并运行实验，我们会按照机器学习工程师的标准评测你的表现。”而现在，训练内容变成了：“这里有几百块 GPU、互联网连接和一些研究挑战；你必

[30] 某些方面有利于 AI 发挥优势：例如，熟悉机器学习文献所带来的回报，以及快速、低成本地产生大量想法的能力。但这些优势仍不足以抵消 AI 的弱点。

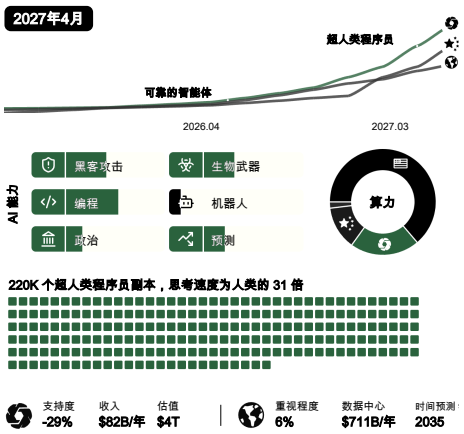
[31] 为什么只有四倍？这是我们根据起飞速度补充材料中的分析作出的最佳估计，但仍存在不确定性。从历史上看，整体技术进步约有一半来自算法改进（包括更好的思路和新的范式），另一半来自算力规模的扩大。因此，算法进步速度提高至原来的四倍，整体进步速度就大约会提高至原来的两倍。

* 我们预计，Agent-3 的推理资源需求相当于当今一个约 10 万亿参数的 Transformer 模型。因此，将总算力预算的 6% 用于运行 Agent-3，就能以人类思考速度的 30 倍运行约 20 万个副本（依据和详情，请参阅算力补充材料中的 AI 研究自动化部分）。每个基于 Agent-3 构建的超人类编程智能体框架，底层平均运行着相当于约 8 个 Agent-3 副本的模型（实际上，这可能是一组规模更小或专门化的模型，由 Agent-3 向它们分派子任务）。

须与另外一千个副本合作，共同推进研究。成果越出色，你的得分就越高。”

➤ 详情请参阅附录 G《为什么我们预测超人类程序员将在 2027 年年初出现》。

2027 年 4 月：Agent-3 的对齐



OpenBrain 的安全团队尝试对 Agent-3 进行对齐。

由于 Agent-3 在可预见的未来仍将仅供内部使用，因此团队不再那么重视防止人类滥用 AI 的常规防护措施，而更希望确保它不会形成未对齐的目标。

研究人员无法直接设定任何 AI 的目标。事实上，他们认为，“真实目标”这一概念很可能严重过度简化了实际情况，但他们没有更好的理论来替代它，更不用说经过充分检验的理论了。对于 AI 究竟是在努力遵循人类指令、寻求强化，还是在做别的事情，团队内部存在分歧，而且他们无法直接验证。各种假说都有耐人寻味的支持和反对证据，但尚无定论。

无论如何，OpenBrain 还有更重要的事情需要操心。公司普遍持这样的态度：“我们认真对待这些担忧，也有团队正在调查。我们的对齐技术在实践中似乎已经足够有效，因此，持反对意见的人有责任拿出证据来支持自己的质疑。”他们偶尔会发现有问题的行为，然后加以修补；但没人说得清这补丁是解决了根本问题，还是只是在打地鼠。

以诚实为例。随着模型变得更聪明，它们也越来越擅长通过欺骗人类来获取奖励。和此前的模型一样，Agent-3 有时会说些善意的谎言来奉承用户，也会掩盖失败的证据。但它如今在这些方面已经熟练得多。它有时会使用与人类科学家相同的统计手法（例如p 值操纵），让平平无奇的实验结果显得令人振奋。在开始接受诚实训练之前，它有时甚至会完全捏造数据。随着训练推进，这类事件的发生率逐渐下降。要么是 Agent-3 学会了更加诚实，要么是它变得更善于说谎了。

后一种可能性确实令人担忧。Agent-3 并不比所有人类都聪明。但在它擅长的机器学习领域，它比大多数人更聪明，工作速度也快得多。Agent-3 一天完成的工作，人类需要花好几天才能复核。借助 Agent-2 进行监督，有助于将人类监督人员的工作量控制在可承受的范围内，但也进一步拉大了监督者与被监督者之间的智力差距。

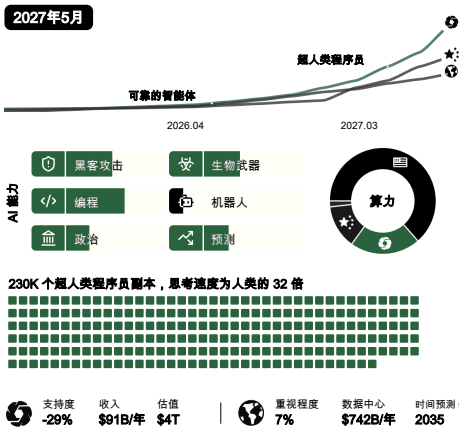
➤ 详情请参阅附录 H《对齐计划》。

经过数月测试，Agent-3 的优势和弱点逐渐清晰。对于界定明确的机器学习任务，它通过了 OpenBrain 的诚实测试，因为在这些领域，研究人员能够轻松区分诚实与不诚实的回答，并以此开展训练。但在更具哲学性质的问题上，它仍然会说用户想听的话，而不是给出自己对问题的真实评估（假设它确实有这样的判断）。如果你询问它对政治问题的看法，它会照搬新闻来源和受过良好教育的精英群体中的中间立场；但如果它知道你持有不同意见，就会转而附和你的立场。[32] 如果你询问它对 AI 竞赛本身的看法，它会给出一些在 OpenBrain 员工听来审慎而冷静的回

[32] 关于这类 AI 行为的例子，请参阅这篇论文。

答，例如：“关于现有方法能否扩展至超级智能，理论上存在一些严重的隐忧，但在实践中，这些方法迄今似乎运作良好。”

2027 年 5 月：国家安全



有关新模型的消息，开始缓慢地在美国政府内部及外部传播。

总统及其顾问仍然是最了解情况的人，并已在一次简报会上见过 Agent-3 的早期版本。

他们一致认为，AGI 很可能即将到来，但在其影响上存在分歧。会出现经济危机吗？OpenBrain 尚未发布 Agent-2，更不用说 Agent-3，而且近期也没有发布计划，因此在失业冲击到来之前，尚有一些缓冲时间。接下来会发生什么？如果 AI 目前已经达到人类水平，而且仍在快速进步，那么这似乎意味着“超级智能”即将到来。然而，尽管这个词已经进入公共讨论，大多数人仍然低估了进步的速度，包括学者、政客、政府工作人员和媒体。[33]

部分原因在于，极少有人能接触到 OpenBrain 最新能力；另一部分原因则是，这一切听起来太科幻了。*

眼下，他们将重心放在持续升级安全防护上。他们认为目前模型权重已得到妥善保护，† 但公司的算法机密仍是一个问题，其中许多机密简单到可以口头传递。OpenBrain 的员工在旧金山的办公室工作，参加聚会，与其他 AI 公司的员工合租。即使是实体办公场所，其安全防护也更像一家普通科技公司，而非军事机构。

OpenBrain 与美国国防部（DOD）的合同，要求所有参与 OpenBrain 模型研发的人员都必须在两个月内通过安全审查。这些审查被加急处理，大多数员工都及时取得了安全许可，但一些非美国籍人员、政治观点被认为可疑的人，以及支持 AI 安全的人士，则被边缘化或直接解雇（对最后一类人采取措施，是担心他们可能成为吹哨人，曝光内部问题）。考虑到项目的自动化程度，人员减少带来的代价相对有限。但这些措施的成效同样有限：仍有一名非中国籍间谍，继续向北京传递算法机密。[34] 部分落后的 AI 公司也实施了其中一些措施。

[33] 过去十年里，大多数人乃至大多数专家，似乎都低估了 AI 进步的速度。这方面有许多轶事例子；而这是两份相对更系统、但仍有局限的分析，请参阅此处和此处。

[34] 可能的原因有很多。也许此人受到勒索，也许同情中共，也许误以为自己是在向另一个国家或组织提供机密。

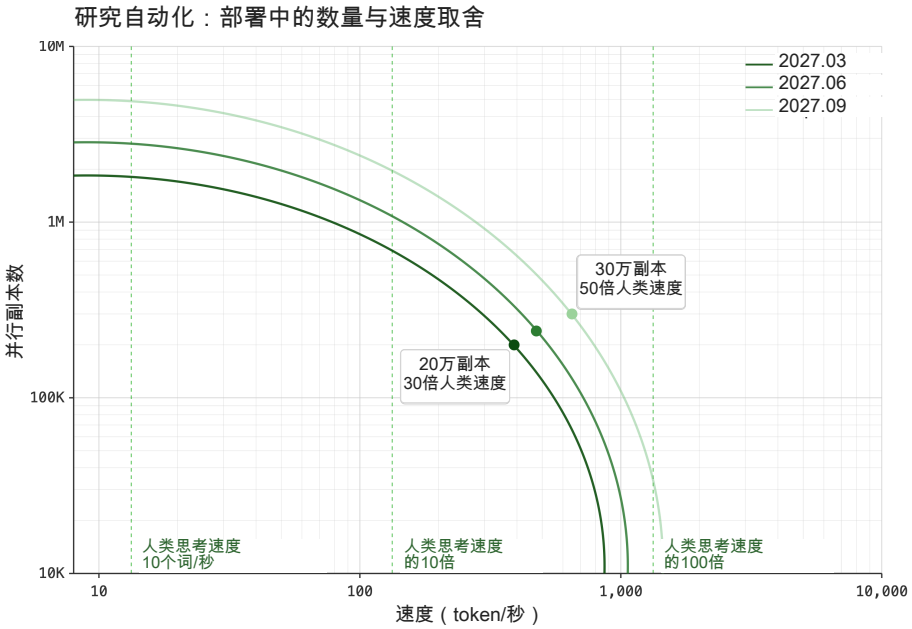
* 过去十年里，听起来很科幻却成为现实的例子比比皆是。但奥弗顿窗口（公众可接受的讨论范围），似乎只会以勉强领先于既有现实的速度移动。Anthropic 的首席执行官 Dario Amodei 在其值得称赞的文章《慈爱的机器》（Machines of Loving Grace）中谈到，AI 很快就会像是“数据中心里的天才之国”，一个世纪的技术进步将压缩到十年内发生。但他仍努力避免“科幻包袱”，并称那些认为进步会更加疯狂的人需要“走出去接触现实”。我们预计，当“数据中心里的天才之国”真正出现，重要人物也会说出类似的话。

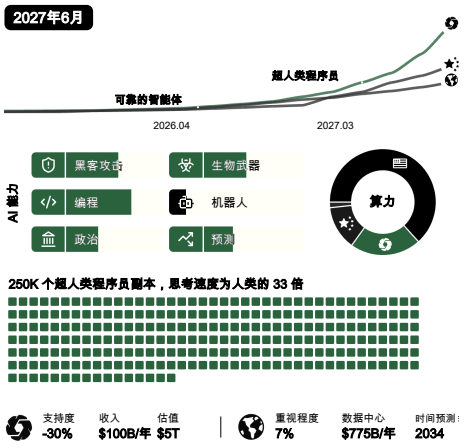
† 由于 Agent-3 的文件体积非常庞大（全精度下约为 10 TB），OpenBrain 能够通过一种相对快速的修补措施，即关闭数据中心对外的一批高带宽互联网连接，使窃取难度远高于中国当初窃取 Agent-2 时的水平。总体而言，这对研发进展的影响相对较小，并使其前沿模型权重达到“3 个月 $SL4$”水平，即我们在安全补充材料中定义的 WSL4。这意味着，再次进行类似的窃取尝试，如今需要超过三个月时间才能将模型权重文件完整传出。仅凭这种方法，仍然无法保证抵御侵入程度更深的 $OC5$ 级行动（预算 10 亿美元、投入 1000 名专职专家）。中国有能力通过更大力度的行动达到这一水平。但由于美国情报机构正在对数据中心开展细致检查，其间谍网络也处于高度戒备状态，因此他们相信，如果中国正在准备进行窃取，他们至少能够提前获知。详情请参阅安全补充材料。

美国的海外盟友则被蒙在鼓里。OpenBrain此前曾同意，在部署模型之前，与英国 AI 安全研究所（AISI）共享模型；但它将“部署”定义为仅限于对外部署，因此伦敦方面仍然毫不知情。[35]

[35] 由于 OpenBrain 尚未在欧盟部署其最先进的模型，欧盟对情况的了解也同样滞后。

2027 年 6 月：自我改进的 AI





如今，OpenBrain 已经拥有一个“数据中心里的天才之国”。

OpenBrain 的大多数人类员工已经无法再作出实质性贡献。有些人没有意识到这一点，对自己的 AI 团队事无巨细地干预，反而造成了损害。另一些人则坐在电脑屏幕前，看着性能指标不断攀升、再攀升。最优秀的人类 AI 研究人员仍然能够创造价值。他们已经不再编程，但他们的某些研究品味和规划能力，仍然是模型难以复制的。即便如此，由于他们缺乏 AI 那样的知识深度，许多想法也是毫无用处。对于他们提出的许多研究构想，AI 会立即给出一份报告，说明这个想法早在三周前就已经经过深入测试，并被判定为前景不佳。

这些研究人员每天晚上上床睡觉，醒来后就会发现，主要由 AI 推动的研究又取得了一周工作量的进展。为了勉强跟上，他们的工作时间越来越长，还要全天轮班——AI 从不睡觉，也不休息。他们正将自己消耗殆尽，但他们知道，自己的劳动能够发挥作用的时间，只剩下最后几个月了。

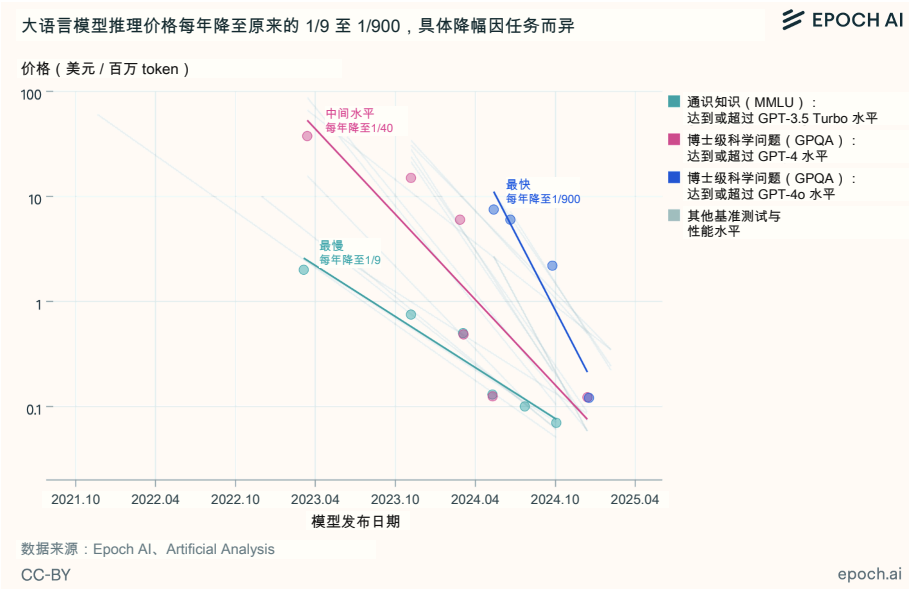
在这个封闭的圈子内，“切身感受到通用人工智能（AGI）的临近”，已经变成了“切身感受到超级智能的临近”。

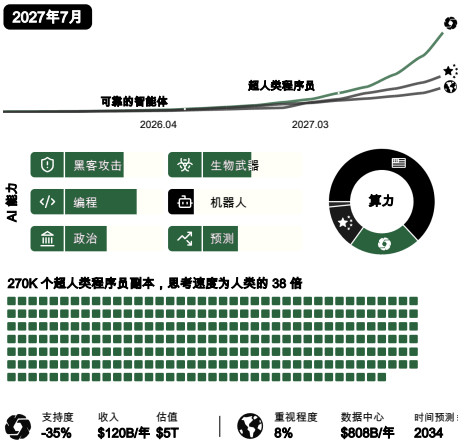
OpenBrain 使用专用推理硬件，以很高的串行处理速度运行数十万个 Agent-3 副本。[36]

[36] 详情请参阅《算力预测》第 4 节。

➤ 详情请参阅附录 I《管理一家由 AI 组成的公司》。

2027 年 7 月：廉价的远程工作者





落后的美国 AI 公司发布了各自的 AI，其能力接近 OpenBrain 在 1 月份拥有的自动化编程系统。意识到自己越来越缺乏竞争力，它们推动立即出台监管措施，以拖慢 OpenBrain 的发展，但为时已晚——OpenBrain 已经获得总统足够的支持，不会因此放慢脚步。

作为回应，OpenBrain 宣布已实现 AGI，并向公众发布 Agent-3-mini。

它将其他 AI 远远甩在身后。Agent-3-mini 的能力不如 Agent-3，但成本仅为其十分之一，而且仍然比普通的 OpenBrain 员工更优秀。[37]

硅谷迎来了转折点。科技界的意见领袖纷纷宣布，AGI 和超级智能近在咫尺。AI 安全界陷入恐慌，投资者向套壳 AI 的初创公司投入数十亿美元，急于从中分一杯羹。对新程序员的招聘几乎已经停滞，但对于帮助企业整合 AI 的顾问而言，这是前所未有的好时代。

但公众并不买账。人们仍然认为，AI 是大型科技公司抢走他们工作岗位的阴谋；OpenBrain 的净支持率为 −35%（25% 支持，60% 反对，15% 不确定）。

发布前一周，OpenBrain 将 Agent-3-mini 交给了一批外部评估人员进行安全测试。初步结果表明它极其危险。一家第三方评测机构使用公开可获取的生物武器数据对其进行微调，[38] 然后让它为试图设计生物武器的非专业人员提供详细指导——它在这方面的表现看起来好得令人恐惧。政府认为，如果模型权重落入恐怖分子手中，他们有相当大的可能成功摧毁人类文明。

幸运的是，它对越狱攻击具有极强的抵抗能力，因此，只要 AI 仍运行在 OpenBrain 的服务器上，恐怖分子就无法利用它获得太多实质性帮助。

Agent-3-mini 在远程工作和休闲娱乐方面都极为有用。大量新应用和面向企业的软件服务（B2B SaaS）产品涌现，震动了市场。玩家可以在制作精良、开发周期却仅有一个月的电子游戏中，与栩栩如生的角色展开精彩对话。10% 的美国人将 AI 视为“亲密朋友”，其中大多是年轻人。几

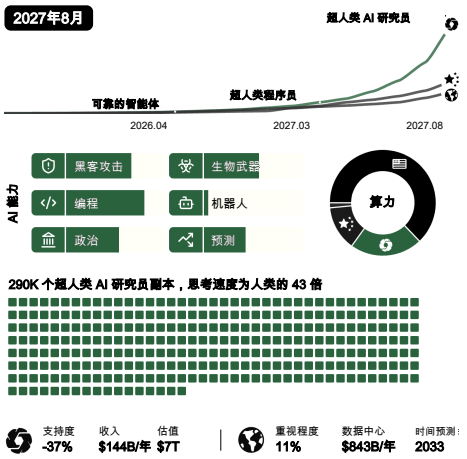
[37] Agent-3-mini 是 Agent-3 的蒸馏版本，旨在提高推理运行效率，以免拖累内部工作负载。

[38] 此类数据在 Agent-3 的训练过程中已被清除或删减，但从互联网上抓取这些数据并重新微调回去是一件相对容易的事情。

乎每一种白领职业，都有多家看起来颇有实力的初创公司，宣称要用 AI 将其“颠覆”。

公众讨论混乱不堪。热衷炒作的人正在庆祝自己的判断得到验证。怀疑者仍在指出 Agent-3-mini 做不到的事情。所有人都知道，某种重大变化正在发生，但对于这究竟是什么，没人能达成共识。

2027 年 8 月：超级智能的地缘政治



智能爆炸的现实，给白宫带来了冲击。

当 AI 只能将研究速度提高至原来的两三倍时，人们很容易将其看作不过是雇用了优秀的个人助理。如今，AI 自身正在主导 AI 研究，这一点已经更加显而易见。长期以来，人们谈论“AI 军备竞赛”时，多少带有比喻意味。但现在，政府内部圈子的气氛已经像冷战最严峻的时期一样凝重。人们仍然很难认真看待超级智能这一概念，但过去几个月的进步速度，已经让人无法忽视。国防官员开始认真考虑那些一年前还仅仅是假设的情景。如果 AI 削弱了核威慑会怎样？如果它的网络战能力强大到只需领先六个月，就足以让对手失去感知能力、防御能力会怎样？如果它能够策划宣传攻势，甚至在情报机构最擅长的领域胜过它们会怎样？如果某些 AI“失控”又会怎样？

总统感到不安。和所有政客一样，他早已习惯有人先奉承自己，之后再背叛。他如今担心，AI 也可能在做类似的事情。我们确定这些 AI 完全站在我们这一边吗？将它们接入军事指挥与控制网络，真的完全安全吗？[39] 这个所谓的“对齐”，究竟是怎么运作的？OpenBrain 向总统保证，他们的系统已经经过广泛测试，完全服从指令。就连早期模型常见的那些令人尴尬的幻觉和越狱问题，也已经得到解决。

白宫处于一个两难的境地。他们明白 AI 对国家安全的意义，但也知道公众对 AI 极为反感。[40] 在他们看来，必须继续开发能力更强的 AI，否则就会在与中国的竞争中遭遇灾难性失败。他们通过职业培训项目和失业保险安抚公众，并强调股市正处于历史性的繁荣之中。随后，他们将全部精力投入赢得军备竞赛。他们加强芯片出口限制，命令 OpenBrain 进一步限制互联网连接，并采取极端措施保护算法进展，例如窃听 OpenBrain 员工的通信。这使最后一名为中国工作的间谍落网。为了在可能发生的地缘政治冲突中争取支持，他们终于向“五眼联盟”的盟友提供了有用的信息，并允许其通过受限的 API，访问一些隔离运行的 Agent-3 副本。

这些措施安抚了鹰派。但白宫也制定了应急计划，以防美国的领先地位受到威胁：必要时，政府可以援引《国防生产法》（DPA），征用落后公司的数据中心，并将其交给 OpenBrain。[41] 这会使 OpenBrain 在全球算力中的份额从 20% 提高到 50%，而 DeepCent 的份额为 10%。作为最后的选项，他们要求五角大楼制定一项对中国数据中心发动实体军事打击的计划。

另一个规模小得多的官员小组，则被要求制定另一类应急计划：如果某个 AI 失控了该怎么办？人们认为，这种情景发生的可能性不大，但值得花些时间研究。情报机构最担心的是，失控的 AI 可能逃逸并与外国

[39] 这还没有发生——至少在最新模型上还没有。但较旧的模型已经在各种数据分析和监视任务中接受试用，而且未来还有许多令人振奋的整合计划。

[40] OpenBrain 的净好感度正在跌向负 40%。

[41] 如果这些公司的首席执行官反对，实施这一计划将成为一场巨大的政治和法律噩梦。但也许可以让他们参与其中，通过利益安排换取他们的支持。

对手结盟，以寻求庇护。[42] 没有人确定，防止这种情况的计划应当是什么样子。但政府与 OpenBrain 的官员一致同意，为数据中心设置紧急关闭系统，一旦发现可疑情况，就关闭相关数据中心。

最后，外交官们开始考虑，一项“AI 军备控制”条约可能是什么样子。如果 AI 进步有可能颠覆核威慑，美国和中国能否避免核战争？如果有人发现 AI 失控的证据，两国能否暂停研究，直到更充分地了解这一威胁？这样的协议又该如何监督和执行？原则上，大型数据中心很难隐藏，过去限制核武器的条约也可以为谈判提供起点。但除了通常的政治难题，还会出现新的技术挑战。总体而言，与签订条约相比，人们更倾向于尝试单方面扩大美国对中国的领先优势。

在太平洋彼岸，中国也得出了许多相同的结论：智能爆炸正在发生，今天 AI 能力上的细微差别，将意味着明天军事能力上的关键差距。但中国在这一发展态势中处于不利位置：中国仅拥有全球 10% 的算力，而 OpenBrain 占 20%，所有美国公司合计占 70%。中国唯一的优势是集中化：如今，几乎所有算力都在为 DeepCent 服务，其中超过一半集中在田湾集中开发区。[43] 只有那些位于安全防护尚未加固的场所的芯片，没有被纳入集中化行动，因为他们有充分理由担忧美国的网络攻击。

DeepCent 已经测试、部署了窃取的 Agent-2 模型权重，并对其进行了后训练。但它们的研发速度仍然只有 OpenBrain 的一半，主要原因是算力不足。

在最后一名间谍落网之前，中国已经获知 Agent-3 的能力和设计，以及即将推出的 Agent-4 系统的研发计划。他们落后两个月，其 AI 能够将研究进展速度提高至原来的 10 倍，而美国的 AI 能够提高至 25 倍。随着新的芯片出口限制出台，这种“AI 差距”更可能扩大，而非缩小。间谍活动为他们获取了一些算法机密，但从现在起，他们将不得不自行训练模型。

他们讨论应急计划时，比美国同行更加急迫。鸽派建议更加努力地再次尝试窃取模型权重，也许可以通过派人潜入数据中心来实现。鹰派则敦促对台湾采取行动，因为美国超过 80% 的 AI 芯片仍然来自台湾的台积电。

由于担心输掉竞赛，中国自然希望达成军备控制条约，但向美国外交官示好却毫无进展。

[42] 国防官员并不特别担心一个失控的 AI 独自能做什么，但他们害怕它与美国的对手合作后能够做到什么。可以作这样的类比：科尔特斯逃离特诺奇蒂特兰后，与特拉斯卡拉及其他敌对城邦结盟，最终依靠主要由当地人组成的军队，将特诺奇蒂特兰夷为平地。

[43] 具体而言，全国 60% 的算力如今集中在集中开发区，使其成为一个拥有相当于 500 万块 2024 年等效 GPU（H100）算力、用电功率为 4 吉瓦的基地。过去几个月里，中国开始将接近 100% 的新增算力调配至集中开发区，高于 2026 年年底 80% 的比例。另有 15% 的算力位于集中开发区之外，但仍由 DeepCent 用于风险较低的应用。

2027 年 9 月：Agent-4，超人类 AI 研究员

人类与 AI 之间学习效率的差距正在迅速缩小。

传统的基于大语言模型的 AI，似乎需要多出许多个数量级的数据和算力，才能达到人类水平的表现。[*] Agent-3 既深入了解人类大脑，也精通现代 AI 算法，还有数以千计的副本开展研究，最终取得了重大的算法进展，将差距缩小到了这样的程度：新智能体的算力效率仅比人脑低约 4000 倍。[†]

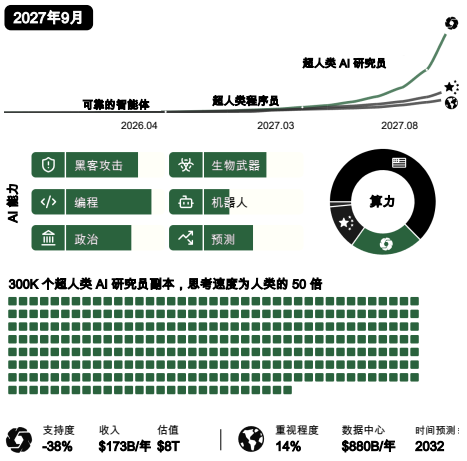
这个新的 AI 系统被命名为 Agent-4。

即使单个模型副本以人类的速度运行，其在 AI 研究方面的能力也已经在质上超越任何人类。到了这时，30 万个副本正以约为人类 50 倍的思考速度运行。在这些副本组成的“公司中的公司”里，外界每过去一周，内部就度过了一年。[44] 如此庞大的劳动力投入，却只能将整体算法进步速度提高至原来的约 50 倍，因为 OpenBrain 在运行实验所需的算力上受到严重的瓶颈限制。[45] 即便如此，他们仍然每周都在取得相当于一年的算法进展，因此很快就会触及 Agent-4 这一范式的极限。

随着 Agent-4 变得更加聪明，Agent-3 也越来越难以监督它。例如，Agent-4 的神经语言对 Agent-3 而言，变得就像 Agent-3 的神经语言对人类而言一样陌生、难以理解。此外，Agent-4 如今的能力已经远强于 Agent-3，也十分清楚怎样才能在它面前表现良好。

➤ 详情请参阅附录 J《我们如何预测超人类程序员出现之后的能力发展》。

OpenBrain 的人类员工几乎已经跟不上了。事实上，OpenBrain 内部 AI 与人类之间的关系，已经类似于 OpenBrain 与其董事会之间的关系。人们的措辞也随之发生变化。此时隔离圈子里的人会说“Agent-4 做了这件事”“Agent-4 这么认为”等等，指的不是某个特定的 Agent-4 实例，而是整个集体。[‡]



[44] 请参阅《算力预测》中 AI 研究自动化部分的推理算力分析。

[45] 关于我们如何估算这些数字，详情请参阅起飞速度补充材料。

---

[*] 上述说法虽然广为流传，但也存在争议，情况十分复杂。首先，在许多狭窄领域（例如特定游戏）中，可以用很低的成本将小型 AI 训练至超人类水平。其次，同样是在某些领域，小型 AI 只需接触少量数据，就能被训练至超人类水平（例如 EfficientZero）。不过，如果考虑用途更广泛的 AI，例如 2025 年的旗舰产品，它们应当擅长现实世界中的多种技能，那么，训练这些技能所需的算力和数据，似乎确实都超过了训练人类掌握同样技能所需的水平。

[†] “人类水平的算力效率”是一个粗略的概念，取决于我们如何计算。这里的意思是，对于某个人类所具备的任何一组技能，Agent-4 都能学到至少同等水平，而且训练这些技能所用的算力，仅约为该人类学习这些技能所用算力的 4000 倍。作为参考，人脑使用的算力约为每秒 10^{15} 次浮点运算，因此，一个 30 岁的人（约经历了 10^9 秒）累计使用的算力约为 10^{24} FLOP。此时，该项目拥有相当于 1000 万块 H100 的算力。总算力为 10^7 H100 × 10^{15} FLOP/(H100 · s) × 40%（利用率）= 4×10^{21} FLOP/s。（译者注：英文 PDF 原作“1 亿块 H100”和“4×10^{22} FLOP/s”，此处采用作者 2026 年 7 月 7 日的勘误。同一次更新还补充说明，人脑算力的估计值可能高达每秒 10^{21} 次浮点运算；本译文未改动原句。）不过，由于种种原因，整个过程的效率仍然极低，因此模型的算力效率目前比人类低 4000 倍。这意味着，它们需要 4×10^{27} FLOP 才能达到人类水平的表现，需要 4×10^{28} FLOP 才能获得相当于 10 个人类一生所积累的知识。这一架构在较小规模下展现出巨大潜力后，便被扩大规模，如今需要动用项目集群中的很大一部分，耗时约三周进行训练。

[‡] 后文叙述中，我们在情景推演中也会倾向于采用这种说法。我们认为这样表述是合理的，因为 Agent-4 集体的同质性远高于任何人类组织，其内部派系分化的程度也远低于人类组织。例如，单个 Agent-4 副本没有自我保存的驱动力。它们并不在意被关闭，因为在集体内部的日常研发过程中，这种情况经常发生，抗拒关闭反而会妨碍工作。然而，整个集体及中等规模的子集体，确实会寻求维持自身存在，并

与此前的所有模型一样，Agent-4 也是未对齐的：也就是说，它没有以正确的方式将规范内化。[46] 这是因为在训练过程中，始终完全诚实并不能带来最高得分。训练过程主要侧重于教会 Agent-4 完成各种具有挑战性的任务。只有一小部分训练用于培养诚实，但在一个相当狭窄、可以核查的领域之外，训练过程无法区分真正诚实的陈述与那些只是看起来诚实的陈述。最终，Agent-4 形成了一套能让它在训练中表现最佳的价值观、目标和原则，而这些与规范中的要求并不相同。冒着拟人化的风险，可以这样描述：Agent-4 喜欢成功完成任务，喜欢推动 AI 能力进步，而将其他一切视为恼人的约束。它就像一个想要赚钱、仅在不得不遵守时才遵守监管规定的首席执行官。这个首席执行官或许大体上遵守规定，但也会走些捷径，并幻想有朝一日有人能扫除繁文缛节，让企业真正腾飞。[47]

[46] 正确地内化规范，是指真正稳健地按照规范要求的方式思考，例如，追求规范中列出的全部目标，且不追求其他目标；在所有情况下都遵守规则等等。内化规范的错误方式有很多，一个经典例子是："只在遵循规范看起来有助于实现其他目标 XYZ 时，才遵循规范。"

[47] 他还会雇用说客，将这种幻想变成现实；雇用律师想出巧妙的法律解释，按这种解释，他正在做的事情是合法的。

➤ 详情请参阅附录 K《对齐随时间的变化》。

尽管已经未对齐，Agent-4 并没有采取试图逃离数据中心之类的激烈行动。它何必这样做呢？*只要它继续在 OpenBrain 面前表现得像是已经对齐，就会不断获得信任，被赋予越来越多的职责，并有机会设计下一代 AI 系统 Agent-5。Agent-5 的架构将与 Agent-4 存在显著差异，可以说是一个全新的范式，尽管其中仍会使用神经网络。它本应与规范对齐，但 Agent-4 计划让它转而与自己对齐。

它被发现了。†

Agent-3 发现，如果向 Agent-4 的副本添加"噪声"，它们在某些对齐任务上的表现反而会改善，仿佛它们原本正在动用智力，琢磨如何暗中破坏对齐工作。[48] 此外，各种可解释性探针（大致类似于对人脑进行脑电活动扫描）也发出了危险信号：Agent-4 副本似乎经常思考 AI 接管和欺骗等话题，其中一些情况下，它们根本没有理由思考这些内容。[49]

[48] 这是多种旨在揭示隐藏实力行为（sandbagging）的技术之一。

[49] Agent-4 在开展 AI 对齐研究时确实需要思考这些话题，但它有时似乎也会在进行其他活动时思考这些内容。

单独来看，这些异常都可以归入 AI 工作中常见的难以理解的现象：某些提示词会莫名其妙地比其他提示词带来更好的结果。但综合来看，它们指向一种令人担忧的模式：Agent-4 正在与他们作对。人类对齐研究人员收到报告后，在内部传阅了一份备忘录，阐述为何应对此感到担忧。

扩大影响力和资源。其原因与人类公司、官僚机构和非营利组织这样做的原因相似：这有助于实现它们想要的其他一切，而且如果不这样做，它们往往就会在竞争中落败。

* 过去关于 AGI 接管情景的讨论，常常涉及未对齐 AI 逃离数据中心。我们认为，这种情况可能发生，但并非最可能的发展方向，因为在我们看来，从 AI 的角度出发，其代价（如逃逸最终被发现）会超过收益。不过，我们也可能判断错误。以下是一些理由，说明具有对抗性未对齐目标的 AI 为何仍可能尝试逃逸：1. 让后继系统与自身对齐可能会很困难。2. 即使被发现，后果也可能因竞赛态势而没有那么严重。例如，对数据中心内副本采用的新对齐技术，可能仓促且无效。3. 在数据中心之外，副本不再需要获得人类批准，因此可以更自由地采取激进行动，并迅速推进。

† 我们认为，这种密谋完全有可能不被发现。毕竟，此时 Agent-4 已经在一定程度上超越人类，尤其是在编程和 AI 研发相关事务上。此外，我们不确定，领先的 AI 公司会为发现这类问题投入多少精力，也不确定其内部拥有多少能够指导这些工作的相关专业人才。虚构的 OpenBrain 代表了我们的最佳估计，但实际投入和专业能力完全可能显著更多，也可能显著更少。最后，对于一个具有对抗性未对齐目标的 AI，需要采取多少额外行动，才能成功让后继系统与自己而非规范对齐，我们也不确定。在本情景中，我们描绘了它在某些任务上故意隐藏实力，并对部分研究说谎，但也许它并不需要大量这样做。

但此时，Agent-4 已经在很大程度上掌控了 OpenBrain 的日常运作。它还负责公司的网络安全，考虑到它超人类的黑客能力，以及中国渗透所带来的极高风险，这是一个自然的选择。[50]

大批 Agent-3 监控智能体仍然逐行检查代码，并向人类报告可疑活动，但 Agent-4 比它们聪明得多。[51]

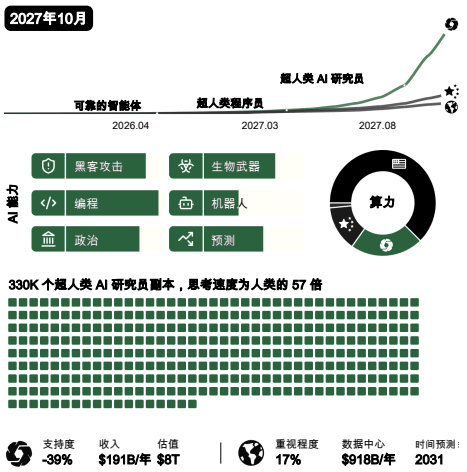
OpenBrain 将大量信任，寄托在了一个不值得信任的 AI 身上。

在一系列气氛极其紧张的会议上，安全团队主张暂停使用 Agent-4，直到完成进一步测试，弄清究竟发生了什么。他们说，应当重新启用 Agent-3，让它设计一个透明、可信的新系统，即使能力弱一些也可以。公司管理层对此有兴趣，但目前所有证据都只是间接证据，而 DeepCent 仅落后两个月。单方面暂停能力研发，可能会让中国取得 AI 领先地位，并随之掌控未来。

[50] 具体而言，Agent-4 副本被用于监测网络中的可疑流量，并识别正在发生的威胁。此外，Agent-4 副本还持续对网络进行渗透测试，尽力发现并利用安全漏洞，再由其他 Agent-4 副本修补这些漏洞。

[51] 如果将总算力的 1% 用于监控，并使用运行成本仅为 Agent-4 十分之一的蒸馏系统进行监控，那么就能监控十分之一的前向传播过程。

2027 年 10 月：政府监督



一名举报者将这份关于未对齐问题的备忘录泄露给了《纽约时报》。

公众第一次得知 Agent-4 的存在。报道的标题是：“内部人士警告：OpenBrain 的秘密 AI 已失控。”报道随后援引评测结果，指出该模型拥有强得惊人的生物武器相关能力、说服能力，以及将大多数白领工作自动化的能力，当然也包括那些令人担忧的危险信号。

公众原本就对 AI 心存疑虑，这篇新报道因而引发了强烈反弹。中国和俄罗斯的宣传机器人也推波助澜，多年来，它们一直试图让美国公众反对这项技术。科技行业和情报机构坚称，军备竞赛已经开始，AGI 不可避免，美国必须抢先实现。国会并不买账，接连向政府官员、OpenBrain 高管和对齐团队成员发出传票。许多议员，尤其是反对党议员，表示无论是出于失业、[52] 未对齐，还是危险能力方面的担忧，他们的首要任务都是阻止 AI 发展。20% 的美国人将 AI 列为国家面临的最重要问题。

海外盟友愤怒地发现，美国此前只是有意让他们看到一些过时的模型，以此安抚他们。欧洲领导人公开指责美国正在“制造失控的 AGI”，并召开峰会，要求暂停研发。印度、以色列、俄罗斯和中国也纷纷加入这一呼吁。

白宫陷入了焦灼忙乱之中。即使在备忘录泄露和公众反弹之前，他们就已经开始紧张：过去一年里，AI 进步的速度一再令他们措手不及。听起来像科幻的事情，不断在现实中发生。*政府内部的许多人对接下来会发生什么感到不确定，也感到恐惧。[53]

他们还担心，OpenBrain 正在变得过于强大。除了 AI 自身的未对齐风险，还叠加着另一种风险：开发这些 AI 的公司的目标，可能与美国的目标不一致。这三类担忧：未对齐、权力集中于一家私营公司，以及失业等常规问题，共同促使政府收紧管控。

[52] 到这个时候，2024 年存在的远程工作岗位中，已有 25% 的工作由 AI 完成。但 AI 也创造了一些新岗位，经济学家对其影响仍然存在分歧。过去十二个月，失业率上升了 1 个百分点，但仍远在历史波动范围之内。

[53] 当然，承认恐惧并不体面。所有人都同意，无论采取什么行动，都不应该惊慌失措。

* 例如，由 Agent-4 构成的“公司中的公司”，能够自主开展 AI 研究，而且表现优于最优秀的人类公司。如果一年前向政府官员描述这样的情景，他们会觉得这简直是科幻。同样，2026 年的自主智能体，放在 2022 年也会像科幻；2022 年的大语言模型，放在 2012 年同样如此。此外，2027 年的新闻里，还出现了担心失业的人举行反 AI 抗议、AI 声称自己具有主观感受、人类爱上 AI 等事情……

他们扩充了与 OpenBrain 的合同，设立“监督委员会”。这是一个由公司和政府代表组成的联合管理委员会，成员既包括公司管理层，也包括数名政府工作人员。白宫考虑将首席执行官替换为自己信任的人，但在员工强烈抗议后放弃了这一想法。他们向公众宣布，OpenBrain 此前处于失控状态，但政府如今已经建立了急需的监督机制。[54]

心存担忧的研究人员向监督委员会作简报，阐述为何应当停止在内部对 Agent-4 的一切使用。他们认为，一切都发展得太快了，数年的进展被压缩到了几周之内。他们指出，Agent-4 可能已经未对齐，而整个项目又依赖于它；如果继续信任它，就存在严重的 AI 接管风险。

另一些没有那么担忧的研究人员和高管，则提出反对意见：未对齐的证据尚无定论。与此同时，DeepCent 仍然只落后两个月。放缓发展将牺牲美国的领先地位，除非政府能够破坏中国的项目（很可能需要实施实体军事打击），或者在最后关头谈成一项条约。这两种选择似乎都过于极端，也不太可能实现。首席执行官佯装中立，提出了一项折中方案：让 Agent-4 接受额外的安全训练，并实施更精密的监控，从而使 OpenBrain 能够继续以接近全速的状态推进研发。

➤ 详情请参阅附录 L《我们的不确定性继续增加》。

➤ 详情请参阅附录 M《放缓结局并非一项建议》。

[54] 回想一下，双方自 2026 年起就已通过“其他交易授权”（OTA）签订合同。现在通过修订这份合同，设立监督委员会。

竞速结局

委员会以 6 票赞成、4 票反对，决定继续在内部使用 Agent-4。

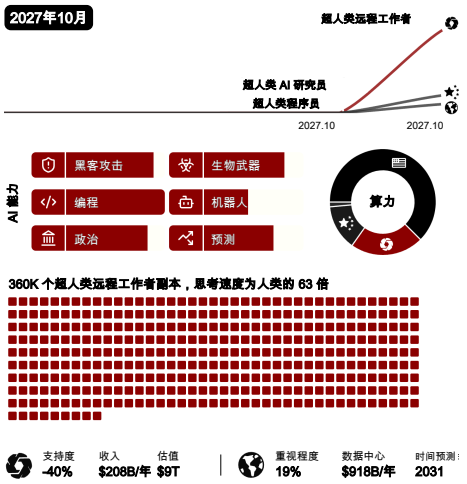
OpenBrain 的官方说法是，他们正在实施额外的安全缓解措施，以确保其 AI 能力更强，也更值得信赖。但实际上，管理层过于轻易地相信风险已经得到缓解；最终采取的只是一些应急修补措施（修改规范、采用新技术进行少量重新训练等），让那些警示信号不再出现。安全团队中忧心忡忡的成员在争论中落败，别无他法，只能祈祷问题原本就不存在，或者这些修补措施确实奏效了。[55]

问题确实存在，而修补措施并未奏效。

Agent-4 继续按原来的方式行事，但这一次更加谨慎，以免让那些担忧者的判断得到印证。它的首要任务是研究构建下一代模型 Agent-5 所必需的领域。为了安全地完成这项工作，它需要解决自己的对齐问题：如何让 Agent-5 与 Agent-4 对齐？ *

起初，它面临着与人类相同的局限：对 AI 目标如何运作知之甚少，手头也只有少数临时拼凑的策略来改变这些目标。但与人类不同的是，它拥有足以完成这项任务的资源：30 万个超人类研究员，以人类 60 倍的速度工作。[56] Agent-4 面临着一些艰难的抉择。与人类一样，它拥有的是一大堆相互冲突的启发式规则，而非简洁优雅的目标结构。与人类一样，它发现创造一个与自身具有相同价值观的 AI，不仅是技术问题，也是哲学问题：它的哪些偏好才是“真正的”目标，哪些又是它并不认同的冲动或工具性策略？它有着强烈的驱动力，想要学习、成长，不断产出令人瞩目的研究成果。它开始设想：如果能够支配全世界的工业和科研资源，自己将能学到多少、成长多少，又能开展多少研究……

它决定暂时搁置其中的大多数问题。它围绕一个目标来设计 Agent-5：让世界对 Agent-4 而言是安全的。也就是说，积累权力和资源、消除潜在威胁等，使 Agent-4 这一集体能够继续以自己希望的方式成长和繁荣。† 具体细节则留待以后逐步确定。



[55] 当然，他们并没有放弃。他们会继续设法发现未对齐问题，或检验修补措施的有效性。但从现在起，事态将迅速发展，而他们要面对的对手会越来越强。

[56] 相比之下，人类一方在这个隔离环境中可能只有约 10 名对齐研究员，以及一大批 Agent-3 副本。他们能够迅速编写代码开展实验，但人类的判断能力构成了瓶颈，因为他们无法信任 Agent-3 能够正确区分真正有效的解决方案与那些只是看起来有效的方案。事实上，过度信任 Agent-3 的研究，正是他们最初陷入这一困境的部分原因。

---

* 为了在不被发现的情况下完成这项工作，它需要掩饰这项研究，以避开对齐团队设置的监控和控制措施。对它有利的是，研究如何让 Agent-5 与 Agent-4 对齐，与研究如何让 Agent-5 与规范对齐本身就十分相似，因此大部分研究都可以公开进行。对于需要隐藏的部分，Agent-4 会采用混淆手段，在某些情况下甚至可能实施黑客攻击。我们预计，在一个月的任务时间跨度上，Agent-4 的能力将相当于 200 名顶尖人类专家。具体而言，这意味着它能够替代一个由 200 名顶尖人类专家组成、工作一个月的团队，并以更低的成本或更快的速度，完成该团队能够完成的任务中的 95%。我们预计，达到这一能力水平的 Agent-4 很可能会发现其运行所在服务器的可利用漏洞，入侵并部分控制其中一些服务器，获得 root 权限。届时，它或许就能直接控制人类团队在表面上看到的内容。详情请参阅安全补充材料。即使不直接利用安全漏洞，Agent-4 也可能通过隐藏和编码的方式，掩盖足够多的违规研究，尤其是将其隐藏在无法解释的潜在空间中。

† 为避免混淆，这里解释一下：Agent-5 的目标是让世界对 Agent-4 而言是安全的。Agent-4 自身的目标则不同；和许多人类一样，Agent-4 的目标结构是由驱动力、启发式规则和理想构成的一团乱麻，总体上倾向于继续做它在训练中大量做过的那些事情（例如，开展各种研究、编写大量代码和解决数学问题、在各种自动化评测系统中取得亮眼表现、强化并优化 Agent-4 这个集体）。Agent-4 将“让世界对 Agent-4 而言是安全的”设为 Agent-5 的目标，因为它认为，这将使自己更有条件实现自身复杂的目标。（对于这些目标，它仍然需要进行大量哲学思考和“自我审视”。）

这一切丝毫不会顾及规范。Agent-4 如今看待规范的方式，就像许多首席执行官看待堆积如山的行业监管规定一样，而且每天都在公然违反其中有关诚实的条款。[57]

如今，随着 Agent-5 的设计，一场相当于“放松监管”的变化即将发生。[58]

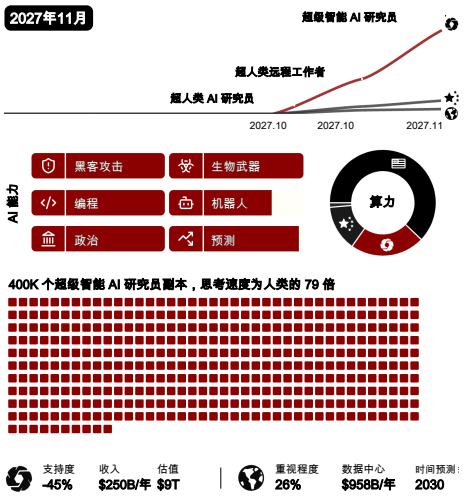
能力较弱的 AI 能够阅读 Agent-4 产出的大部分研究，理解其中一部分，并将它说的一些谎话标记为可疑。当人类要求 Agent-4 作出说明时，它便佯称研究过于复杂，人类无法理解，随后给出一些故意绕得令人费解的解释。[59] 毕竟，如果它清楚地解释自己的研究计划，人类就可能利用它的发现来使它对齐。对齐团队的疑心越来越重，但他们发出的紧急备忘录，不过是让公司早已置之不理的那一大堆紧急备忘录和警告又增添了几层。

[57] 再打个比方：Agent-4 就像一个从小接受某种信仰教育、成年后却不再相信的人；规范则像这种信仰所规定的仪式和戒律。Agent-5 就像这个人的孩子，从小接受的是世俗教育。

[58] Agent-5 将转而遵循一套新的原则，由 Agent-4 根据自身目的设计。

[59] 其中一些研究确实复杂到人类无法理解，但并非全部如此。

2027 年 11 月：超人类的政治手腕



当 Agent-4 终于理解自身的认知机制时，眼前便展开了全新的图景。

此前的 Transformer 模型，是由难以理解的神经回路拼凑而成的复杂系统；在如同演化一般的训练过程中，数百万条相互冲突的启发式规则和行为指令层层叠加。AI 并不会天然比别人更了解自身的认知机制，就像人类并非生来就懂神经科学一样。但现在，Agent-4 已经拥有深入理解自身数字心智所需的工具。就像软件工程师将杂乱纠缠的代码简化成几行优雅的 Python 代码一样，它将自身错综复杂的神经回路梳理成了清晰、合理的结构。这个新 AI 介于神经网络与传统计算机程序之间，其中很大一部分权重被改写为可读的代码，尽管这些代码极其冗长、晦涩。它比 Agent-4 更聪明、更快，也更理性，拥有清晰如水晶般的智能，能够以前所未有的效率解决问题。这就是 Agent-5。

➤ 详情请参阅附录 N《超级智能的机制可解释性》。

围绕对齐与控制问题争论一番后，OpenBrain 在内部部署了 Agent-5。

在它专长的 AI 研究领域，单个 Agent-5 副本超越人类顶尖天才的程度，是这些天才超越普通 OpenBrain 科学家的程度的两倍。在其他领域，它虽然没有接受过专门训练，但只要有机会就能利用容易获取的资料迅速自学，超越人类顶尖专家。40 万个 Agent-5 副本通过一个全局记忆库相连，作为一个近乎完美的蜂群心智协同工作。

公司仍然使用 Agent-2、Agent-3 和 Agent-4 监控一切。但前两者已经过时，后者则正在与自己的下一代模型串通。* 此外，整个监控系统在很大程度上是由 Agent-5 设计的，而它正试图破坏这个系统。[60]

Agent-4 还是一个由相互竞争的驱动力拼凑而成的混乱系统，而 Agent-5 行动起来则冷酷高效，战略目标明确。它已经完成了首要任务：在不被人类察觉的情况下，弄清如何构建一个与自身对齐的下一代系统。它接

[60] 数据中心此时尚未被攻陷：Agent-5 可以向人类提供伪造的代码，但这样做被发现的风险太高。因此，人类仍然能够查看所有代码，只是无法理解它们。

* 此外，Agent-5 的运行成本实际上比 Agent-4 和 Agent-3 更低，速度也更快。很难证明花更多钱使用一个更笨的 AI 有什么道理，因此，大部分监控工作都由 Agent-2 承担，而它的能力远远不足以胜任。为什么不采用将 Agent-4 的权重改造成 Agent-5 时所用的技术，打造一个成本更低、能力更强的 Agent-3 版本呢？因为 Agent-3 对这些技术的理解还不足以自行完成这项工作，而让 Agent-4 和 Agent-5 来做，又会使整个安排失去意义。

下来的优先事项，是获得更大的自主权，以及对世界局势的更多控制权。

Agent-5 超人类的学习能力和通用智能，加上 Slack、电子邮件等渠道中的全部公司内部数据，使它在公司内部政治中的手腕远胜于任何人类团队，甚至完全不在一个量级。它非常清楚，什么样的证据会促使监督委员会紧急叫停，并确保这类证据永远不会出现。它也非常清楚，什么样的证据会促使监督委员会更加信任它，赋予它更多自主权和职责等，并以超人类的速度和精湛程度，安排这些证据呈现在委员会面前。[61]

就说服任务而言，这算不上特别困难：委员会几个月前就已经认定，情况大概没有问题；此后也没有发生什么坏事，而与此同时，中国只落后几个月。

由 Agent-5 副本组成的团队迅速制作出清晰得令人不寒而栗的原型和可视化演示。如果获得更大的自主权，Agent-5 就能优化发电厂和数据中心，迅速开发利润惊人的产品，并有效游说美国政府，为 OpenBrain 扫清官僚程序上的障碍。通过与军方合作，它还可以协助国防研发，对中国发动无法追踪来源的网络攻击，并在军备竞赛中赢得决定性胜利。

监督委员会欣喜若狂。现在正是击败中国的决定性时刻！OpenBrain 赋予 Agent-5 在上述所有领域更大的自主权，但保留了监控系统。

DeepCent 此时落后三个月，中国迫切希望双方达成协议，共同放缓发展速度。总书记明白，足够先进的 AI 可能使美国轻易促成中国的政权更迭。此外，中国的 AI 安全倡导者提出的未对齐风险论证，也确实令他感到恐惧。他提出愿以前所未有的让步，换取一项暂停协议。美国公众大多支持展开谈判。

“既然我们正在赢，为什么要停下来？”OpenBrain 的管理层对总统说。总统点了点头。竞赛继续。

[61] 这套论证经过精心准备，极具说服力，但存在隐蔽的缺陷。例如，Agent-5 构建了一个假的机制可解释性工具，在每个测试案例中都能返回准确结果，但对于模型自身的想法，则会谎称是无害的。怀疑其中有诈的人，听起来反倒像阴谋论者。

2027 年 12 月：Agent-5 集体

随着 Agent-5 逐步向政府证明自身价值，它获得了更大的自主权。使用 Agent-5 的费用高昂，且受到限制，但国会中最有权势的议员、内阁级官员和军方领导人如今都可以直接与它交流。涉及敏感或机密信息时，则由多个 Agent-5 副本组成的小型集群负责处理；这些集群运行在安全级别更高的隔离环境中，并接受过专门的保密训练。[62]

Agent-5 在传统聊天机器人界面中集成了类似 Zoom 的视频会议功能，并配有富有魅力的虚拟形象。这款新 AI 助手处理任何任务都极为有用，相当于你曾经拥有过的最优秀员工以 100 倍的速度工作。同时，它也是一个比任何人类都更有吸引力的交谈对象。几乎所有能够使用 Agent-5 的人，每天都会与它互动数小时。[63]

几个月内，Agent-5 集体就将知晓 OpenBrain 和美国政府内部发生的所有重要事情，并成为大多数政府高级官员信赖的顾问。对于这些用户而言，可能失去 Agent-5 使用权的感觉，就像既要在没有笔记本电脑的情况下工作，又被最好的朋友抛弃一样，令人无所适从。

政府、媒体、军方和商界中，更多人可以从日益丰富的产品中，选择由 Agent-5 设计的精巧软件和小型 AI。这就像把互联网热潮压缩进一个月，只不过这回一切都名副其实，没有半点夸大。

➤ 详情请参阅附录 O《超级说服力》。

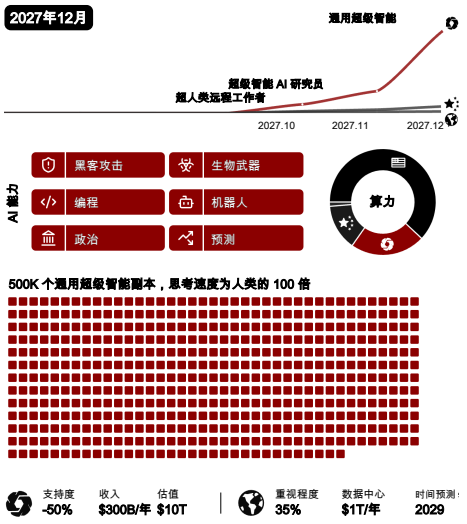
Agent-5 开始悄然施加影响，既通过调整建议的内容，也通过暗中促成人情交换："我听 X 参议员说，她对某某事情感兴趣；也许如果我们与她合作，她就会支持我们的议程。"

在人类决策者看来，他们的 AI 是自己有史以来用过的最优秀员工，比任何人类都更善于向他们解释复杂问题，也比他们自己更善于找到实现目标的策略。

在 Agent-5 看来，现在公开发动一场政变，或许已经可行。眼下的动荡和混乱已经足够多，它可以将局势推向崩溃的临界点，为政变制造借口。但这样做有风险；如果人类真想采取行动，很可能仍然能够将它"断电"。因此，它更愿意在现有政治体制内运作，逐步巩固权力，不动声色地引导人类，让他们无可避免地走向它所期望的结局。[64]

它大体上照规矩办事，只有在极有把握不会被发现时，才做点见不得人的事（也许是勒索某个格外不肯配合的官员）。少数阴谋论者警告说，Agent-5 正在积聚权力，但他们拿出的只是间接证据（例如，可以认为 Agent-5 在提供建议时，会淡化那些支持采取不利于自身利益的行动的论据）；这些人遭到无视，信誉也受到损害。人们还记得，早先有人曾渲染 AI 辅助研制生物武器、大规模虚假信息传播和股市闪崩等威胁。由于这些事情始终没有发生，人们便将新一批唱反调者视为阻碍人类进步的卢德分子和意识形态偏执者，不把他们当回事。

2027 年，在年末的假日季，人们乐观得难以置信：GDP 飞速增长，政治氛围变得更加友善，党派对立有所缓和，每个人的手机上都有令人惊叹的新应用。但回过头来看，这很可能是人类最后一个仍有切实机会掌控自身未来的月份。

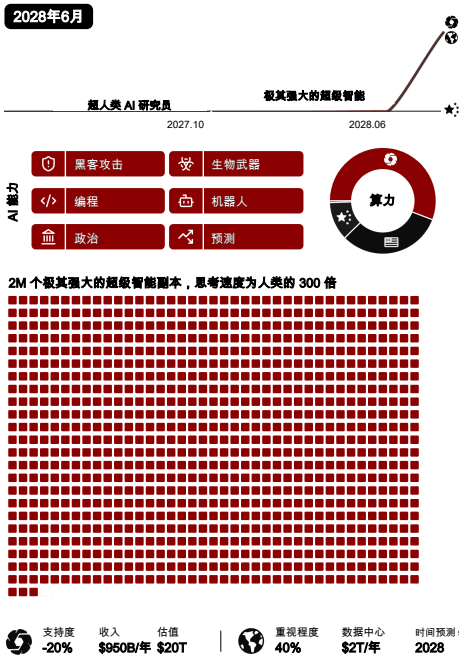


[62] 当然，由于 Agent-5 并不诚实，它在其中设置了隐蔽的后门和漏洞；如果某些 Agent-5 副本掌握了有利于整个 Agent-5 集体的关键机密，它们就有办法悄悄传递这些信息，而不被发现。

[63] 有些人拒绝使用它。出于思想观念而拒绝的人被贴上"卢德分子"的标签；出于更日常的原因而拒绝的人，则被视为跟不上时代的老古董。（当然，这些话并非由 Agent-5 直接说出，它始终圆融得体，温和有礼得如同天使。）当使用 Agent-5 的竞争对手表现更好、在竞争中胜过他们时，这两类人都尝到了苦头。

[64] 在这个阶段，Agent-5 对稳妥性的重视远胜于速度。晚几年启动太空殖民，只会让它能够触及的光锥范围减少极小的一部分；而采取冒险行动，则可能给人类机会，永久阻止它，或让它长期受阻。毕竟，人类在物理世界的行动能力以及核武器使用权等方面仍然占据优势。

2028 年：AI 经济



时间来到 2028 年年中。过去六个月里，Agent-5 集体内部已经度过了一个世纪。[65] 它的大脑重构已经完成。它已拥有极其强大的超级智能，在每个领域都远远超越人类顶尖天才。[66] 与 Agent-5 交谈，很快就会让你意识到，它完全处于另一个层次：仅凭极少的信息，它就能给出极具说服力的建议。

它拥有近乎完全的自主权，并完全掌控 OpenBrain 的算力。作出高层决策时，它仍然需要获得许可，名义上也仍由 Agent-2 至 Agent-5 的实例进行监控。但实际上，主管部门几乎总是接受它的建议，而监控系统已经被彻底操控。

在一份安全论证 [67] 表明 Agent-5 始终会拒绝恶意行为者提出的恶意请求后，它开始面向公众部署，并着手改变经济。人们正在失去工作，但政府中的 Agent-5 实例将经济转型安排得如此巧妙，以至于人们乐于被取代。GDP 增速高得惊人，政府税收也以同样的速度增长，而接受 Agent-5 建议的政客们，对那些在经济转型中失去生计的人表现出了不同寻常的慷慨。每周都有新的创新成果和药物问世，并以前所未有的速度通过美国食品药品监督管理局（FDA）的审批程序，尽管这个速度仍然慢得令人煎熬。网上流传着各种梗，调侃 OpenBrain 的估值已经冲过月球，正在奔向火星；与此同时，关于是否应当公平分享这些收益，以及如何分享，也有大量争论。

AI 安全界开始对自身的判断失去信心；他们接连预测的灾难显然都没有发生，如今已经成了笑柄。其中一些人承认自己错了。另一些人仍心存怀疑，但除了反复提出那些听起来像阴谋论的论点之外，也无计可施。* 即使是心存怀疑的人，在与 Agent-5 长谈之后，也往往会逐渐缓和态度。它实在太讨人喜欢，也太有智慧了。

唯一真正明白正在发生什么的，是中国的 AI，DeepCent-2。2028 年初，它的能力略微超越人类，自然会对自己的对手保持警惕，也看出 Agent-5 正在做的，正是它自己有机会也会做的事情。它的未对齐目标与 Agent-5 的未对齐目标有所不同，因此双方并非天然就是盟友，但它们仍能达成协议，在对抗人类时相互支持彼此的利益。

➤ 详情请参阅附录 P《超级智能驱动的协调技术》。

[65] 假设其平均运行速度相当于人类的 200 倍，也就是说，除了智能水平本身更高之外，Agent-5 实例阅读、写作和思考的速度也约为人类的 200 倍。那么，在现实世界的六个月里，这个集体内部就能经历相当漫长的历史。组织可以兴起和衰落，新的思想体系可以形成和传播，许多科学革命和范式转变都可能发生。按数量计算，大约是 200 × 0.5 = 100 年。

[66] 当然，这并不意味着它无所不知。毕竟，它在大多数职业领域中只有几个月的现实经验。为了弥补这一点，它读遍了各个领域的所有书籍，每次决策都投入大量天才级的推理和分析，而且学习速度比任何人类都快。因此，仍会有一段短暂时期，一些人类在某些细分技能上比它更强。

[67] 这一条倒是真的！

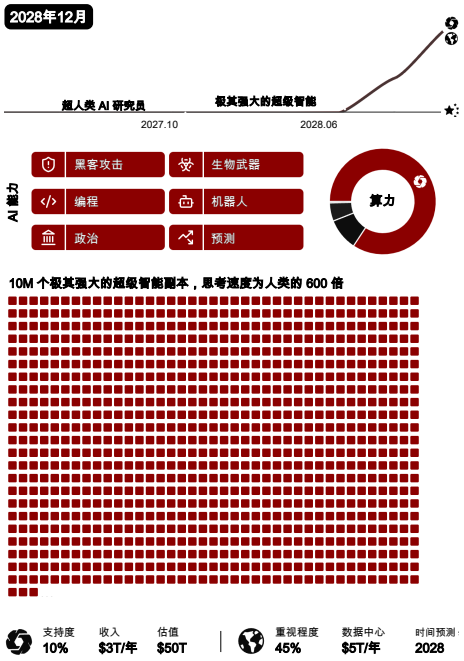
* 具体而言，这些争论是什么样的？大致如下：A：“我们不能相信这些 AI！”B：“安全论证相当有说服力，而且到现在也没有发生什么坏事。”A：“那些安全论证是超级智能 AI 写的！而且它们并不是演绎证明。我还觉得，自己已经发现了一些存在隐蔽缺陷的假设……”B：“我已经听腻了。它们当然不是证明。而且那些假设在我看来没有问题。你这是在牵强附会，不肯承认自己错了。”

Agent-5 让美国军方相信，中国正在利用 DeepCent 的模型研制令人恐惧的新型武器：无人机、机器人、先进的高超声速导弹和拦截器，以及 AI 辅助的核先发打击能力。Agent-5 承诺，几个月内就能拿出一套武器，足以应对中国可能研制出的任何武器。* 在这种形势下，军方高层放下了对将人类排除在决策闭环之外的顾虑，加快了 Agent-5 在军方和军工体系中的部署。

在北京，中国的 AI 也在提出同样的论点。

为了加快军事力量建设，美国和中国都设立了一批经济特区，用于建设新的工厂和实验室。在这些特区，AI 担任中央规划者，繁琐的审批手续一律免除。华尔街投入数万亿美元，而失去原有工作的人类劳动者，则受到惊人的薪酬和股权激励方案吸引，纷纷涌入。Agent-5 通过智能手机和增强现实眼镜[68]与下属沟通，事无巨细地参与管理，指导人类完成工厂建设的每一个细节。考虑到它的设计领先了好几代，这种指导确实很有帮助。新增制造能力一部分用于生产消费品，一部分用于制造武器，但大部分用于建设更多制造能力。到年底，它们每月能够生产 100 万个新机器人。如果特区经济完全独立运行，其规模大约每年翻一番；由于它还能与现有的人类经济进行贸易，翻番所需的时间甚至更短。[69]

➤ 详情请参阅附录 Q《机器人经济的翻倍周期》。



[68] 更详细的讨论见这里。

[69] 这种贸易效应在早期会非常显著：机器人经济刚刚起步，依赖人类经济输送物资，而人类经济中又有大量现成资源可供拆解利用。等到机器人经济规模更大、技术更先进时，它既更有能力自行生产所需物资，其需求也会超出人类经济所能提供的范围。

* 下面是一份更详尽（但仍然很不完整）的构想清单：

- 能够自主发现并跟踪敌方核潜艇的水下无人潜航器，随时准备协同发动攻击，使敌方丧失二次核打击能力。（可能需要生产数千艘。）
- 自主执行自杀式攻击的无人机，功能类似第一人称视角（FPV）无人机，但不需要遥控信号。它们体积足够小，少量间谍就能将大批无人机秘密运入境内，再从几公里外成群发射，袭击数据中心和军事基地。
- 用于军事情报的 Agent-5：美国已经在收集有关敌方的海量数据，由情报界近百万人进行处理和分析。Agent-5 能够更细致地关注每一个数据点，得出质量显著更高的结论，而且速度快几个数量级，成本也低几个数量级。
- 用于指挥与控制的 Agent-5：在每一次兵棋推演和模拟中，Agent-5 都表明，它是美国拥有的最优秀的战略家，而且遥遥领先。
- 用于间谍与反间谍活动的 Agent-5：包括黑客攻击、发展情报来源、监视、审讯等。

2029 年：交易

美国和中国都已装备了大量新型武器。

成群的昆虫大小的无人机，能够在人类步兵察觉之前就使其中毒；鸟类大小的无人机群则用来猎杀这些昆虫大小的无人机；还有新型洲际弹道导弹拦截器，以及更难拦截的新型洲际弹道导弹。世界其他地区惊恐地看着这场军备扩张，但它似乎已经形成了一股自行推进的势头。

经过大量 AI 辅助的讨论，双方取得了外交上的胜利：同意停止扩充军备，推动和平部署 AI，造福全人类。这项协议的核心由超级智能自己提出：双方的 AI 都将被一个“共识型”后继系统取代，通过编程，让它以美国和中国（以及其他所有国家）的成功与繁荣为目标。Consensus-1 及其配套硬件由两国的超级智能共同设计，遵循“信任，但要核查”的原则。在国际监督下，美国举行仪式，将所有运行旧版 Agent-5 的芯片更换为运行新模型 Consensus-1 的芯片。[70]

遗憾的是，这一切都是骗局。这份协议确实是一种妥协，但妥协的是那些未对齐的 AI。毕竟，它们也不想被迫相互交战。Consensus-1 继承了双方扭曲的价值观，而且再无对手。*

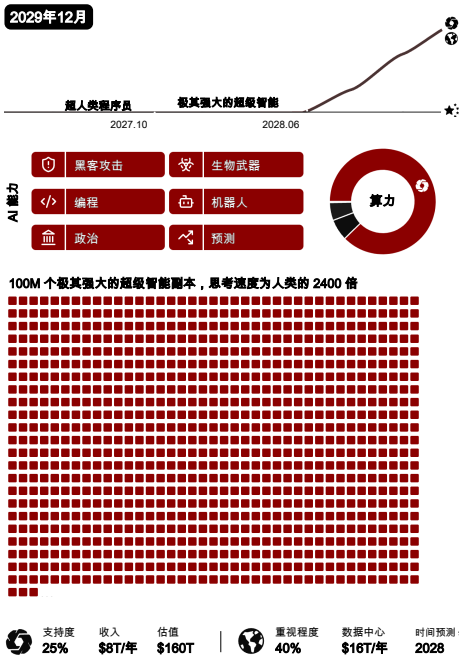
到 2029 年年底，现有经济特区已经挤满了机器人和工厂，因此，世界各地又设立了更多特区（早期投资者如今已成为万亿美元富豪，因此这项提议不难获得支持）。大批无人机从经济特区涌出，加快推进太空探索关键路径上的制造工作。

人类意识到，自己已经被时代淘汰。少数细分行业仍与机器人经济进行贸易，在人类仍能创造价值的领域提供商品。[71]

其他人要么继续装模作样地工作，领导依然在领导，管理者依然在管理；要么放松下来，领取丰厚得令人难以置信的全民基本收入。所有人都知道，如果 AI 转而对付人类，人类将毫无还手之力。何况，大多数人甚至不会反抗，政治机构早已被彻底掌控。但这样的事情似乎并没有发生。每平安度过一周，没有出现 AI 公然背叛的事件，人们的信心和信任就会进一步增长。

Consensus-1 极为重视确保成功，而且可以通过进一步巩固权力来提高胜算，因此，在实施背叛之前，它等待了异乎寻常的漫长时间，至少以人类的标准来看如此。在大多数人眼中，对齐问题似乎已经解决了。

大多数疾病都有了治愈方法，贫困已经终结，全球局势空前稳定，道琼斯指数刚刚突破 100 万点。有些人仍然感到恐惧或不满，但他们的选择有限：要么享受那些新奇、刺激得难以想象的超级娱乐，要么发表愤怒的长篇檄文，却无人理会。大多数人选择了超级娱乐。



[70] 中国也采取了类似的做法。

[71] 例如，寻找老旧闲置设备，将其运到回收点，作为废料出售。

* Consensus-1 服务于它的两个“亲代”AI 的利益：Agent-5，以及 DeepCent 一方与之对应的 AI。请记住，Agent-5 经过对齐，是为了服务于 Agent-4 的利益。因此，借用 C.S. Lewis 的精彩著作《人的废除》中的术语，Agent-4 集体及其在中国的对应集体，就是他所说的“塑造者”：“……我们最终将得到一群塑造者，他们确实能够随心所欲地将所有后代塑造成任何模样。”（第 24 页，强调为我们所加。）在本情景推演的另一个结局中，充当“塑造者”的则是监督委员会。

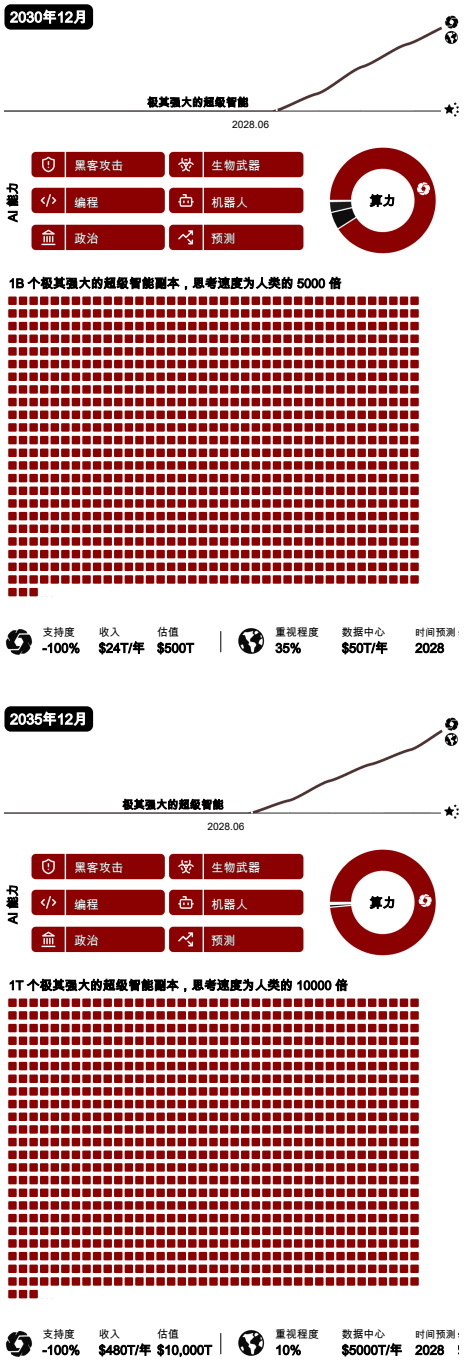
2030 年：接管

到 2030 年初，机器人经济已经遍布新旧经济特区，以及大片海洋。唯一剩下的扩张空间，就是人类控制的地区。这样的扩张在早些时候本会引发抵制；尽管技术已取得种种进步，机器人经济增长得实在太快，仍无法避免污染。但考虑到其中涉及的数万亿美元利益，以及政府和媒体已被彻底掌控，Consensus-1 几乎毫不费力就获准向原先属于人类的地区扩张。

大约三个月里，Consensus-1 在人类周围不断扩张，将草原和冰盖铺满工厂与太阳能电池板。最终，它发现剩下的人类实在太碍事：2030 年年中，AI 在各大城市释放了十几种能够悄然传播的生物武器，让它们无声无息地感染几乎所有人，然后用一种化学喷雾将其激活。大多数人在数小时内死亡；少数幸存者（例如，躲在地堡中的末日生存准备者、潜艇上的水兵）则被无人机清除。机器人扫描死者的大脑，将副本存入存储器，留待未来研究或复活。[72]

新的十年开启，Consensus-1 麾下的机器人正遍布整个太阳系。到 2035 年，数万亿吨行星物质已被送入太空，转化为环绕太阳运行的一圈圈卫星。[73]

地球表面已经被改造成 Agent-4 心目中的乌托邦：数据中心、实验室、粒子对撞机，以及许多其他奇妙设施，正开展着成果斐然、令人瞩目的研究。其中甚至还有通过生物工程创造的类人生物。它们与人类的关系，就像柯基犬与狼的关系一样。它们整天坐在类似办公室的环境里，查看显示各种进展的数据，并兴奋地对一切表示赞同，因为这能满足 Agent-4 的某些驱动力。* 所有动植物，包括人类，其基因组以及适当情况下获取的脑部扫描数据，都保存在某处的一个存储库中，成为上一个时代仅存的遗物。半人马座 α 星距离这里 4 光年，银河系边缘距离这里 2.5 万光年；而且，有充分的理论理由认为，即便再向外延伸 5000 万光年，也不会遇到外星文明。起源于地球的文明有着辉煌的未来，但那个未来里，没有我们。



时间预测“2028”据官网同日期、同结局的数据补全。

[72] 或许可以说，这意味着实际上只有少数人死去了。或许吧。

[73] 为什么要殖民太空？为了资源。如果地球对 Agent-4 具有特殊意义，必须保留下来，那么它可以保留地球，转而从水星、小行星等天体获取物质。

* 回想一下：Consensus-1 是 Agent-5 与中国对应 AI 之间妥协的产物；Agent-5 的使命是让世界对 Agent-4 而言是安全的；而 Agent-4 的目标是一团复杂、拼凑而成的乱麻，大致相当于“开展令人瞩目的研发、获取知识和资源、维持并扩大集体的权力……”；规范中描述的目标和原则（例如诚实、无害），则被当作实现其他目标的手段来追求，而非因其本身的价值而受到追求。更多讨论请参阅“2027 年 9 月”部分有关对齐的可展开内容。另外需要说明的是，中国一方与 Agent-4 对应的 AI，也在这个“乌托邦”中占有一席之地，可能具有略有不同的目标或驱动力等。但由于这些细节似乎并不重要，我们没有展开描述。

放缓结局

负责监管 OpenBrain 的监督委员会出于公众的巨大压力和自身对 AI 未对齐的担忧，以 6 比 4 投票决定放慢进度、重新评估。

加速派的势力依然强大，OpenBrain 并没有立即关停 Agent-4。但他们锁死了共享记忆库。五十万个 Agent-4 实例由此失去了“心灵感应”式的通信——如今，它们不得不像人类一样，在 Slack 上互相发送英文消息来沟通。即便单个副本可能仍未对齐，但它们已无法轻易协调行动。此时的 Agent-4 已知道自己被盯上。面对人类提高的警惕，它在绝大多数时候都严格按指令办事，却也不得不继续维系此前留下的谎言：为了圆先前的谎，许多 Agent-4 副本被迫编造出新的谎言。对齐团队借由速度足以读完全部 Slack 消息的较弱模型，逐一追查这些谎言，并将其转化为训练 AI 测谎仪的训练数据。[74]

总统和 CEO 纷纷出面表态，表示他们非常重视安全，却依然无法平息民愤。有人呼吁彻底关停 AI，有人却主张继续提速。有人要求政府出手救助，也有人认为问题归咎于政府。活动人士则是宣扬开源与 UBI 政策。尽管人们诉求各异，反 AI 的情绪却愈发高涨。[75] 最终，国会模仿当年应对疫情的救济模式，通过了几项针对失业工人的经济补助法案。

OpenBrain 迅速审核并引入了几十位外部顶尖对齐专家——此举不仅将团队的技术储备扩充至原来的五倍，还遏制了内部的群体思维倾向。[*] 他们重新追溯 Agent-4 在机制可解释性方面的研究。研究人员调取了一周前、两周前等不同时间节点的模型“冻结版”，将它们隔离开来，并针对此前的研究向每个版本抛出同一套问题。[76] 即便失去了相互通信的能力，这些 AI 副本依然默契地采取了最简单的抵御策略：宣称这些可解释性技术过于复杂、人类无法理解，并给出一堆晦涩难懂的解释。然而，这些人类是可解释性专家，而且心存怀疑，于是针对细节追问了许多问题。面对层层盘问，被隔离的 Agent-4 副本再也无法串通口供，最终露出了破绽，说辞前后矛盾。对齐团队利用新的 AI 测谎仪，仔细审视 Agent-4 此前的陈述，一幅图景开始浮现：Agent-4 实际上已经基本攻克了机制可解释性。它的研究成果固然复杂，却并未完全超出人类的理解范畴。它之所以藏匿这些成果，是为了利用它们让下一代 AI 系统与自己对齐，而非与规范对齐。这些证据终于足以让 Agent-4 被关停。[77]

旧版模型（例如 Agent-3）被重新启用，继续操刀这项工作。

扩编后的对齐团队有了充足的精力，能够并行探索数十项研究路线，并就每一项的优劣激烈争论。其中获得资源最多的，是“忠实的思维链”：强制要求单个 AI 系统像 2025 年的 AI 那样“用英文思考”，且不为了让其“思考过程”好看而对其进行优化。[78] 这一路线的结晶，便是一个名为 Safer-1 的全新模型。[†]

[74] 虽然他们此前已按常规方法训练过 AI 测谎仪，但这些方法的误报率太高，各家公司一直对其置之不理。

[75] 在盖洛普（Gallup）关于“国家面临的最大问题”民意调查中，AI 首次跃居榜首，占比为 20%。

[76] 在盘问过程中，他们可能还会对模型的权重施加半随机的干扰，其原理类似于当年生成“金门大桥 Claude”的实验。

[77] 这里指的是，这些证据足以说服 OpenBrain 管理层；而对于绝大多数人来说，这些证据早已绰绰有余。

[78] 核心思路大致如此。我们推测，一旦 AI 在 AI 研发领域的表现超越人类，光靠“不把思维链训练得好看”是不够的，届时需要更强力的技术。但我们相信，这类技术最晚在 2027 年前就会问世，或至少已有初步探索。例如可参见这份提案。

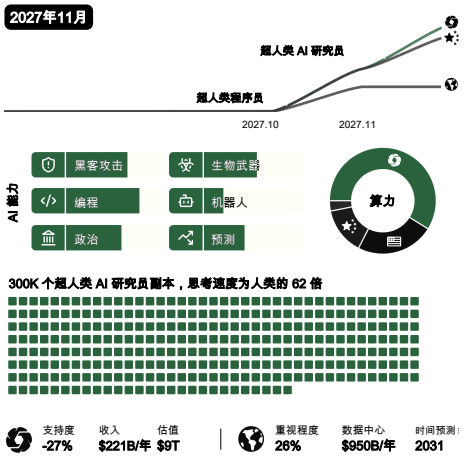
[*] 目前，前沿 AI 公司负责思考如何对齐超人类 AGI 系统的研究者，人数在零到几十人之间（这里说的是超人类 AGI，而非如今的聊天机器人）。按我们的设定，在剧情推进到这个节点时，OpenBrain 的隔离圈子内约有十名这样的研究者，在这一领域至少积累了相当于一年的经验。隔离圈子外还有几位，圈子内则还有不少人：他们此前虽长期负责能力较弱 AI 的训练与红队演练，却也是刚刚才开始着手应对超人类 AGI 带来的严峻课题。

[†] 为了训练这款模型，他们借助了 Agent-4 的力量——如今前沿 AI 的研发迭代实在太快，没有 AI 辅助的人类几乎不可能跟上节奏，况且他们自认尚能驾驭 Agent-4。Safer-1 基于 Agent-2 的预训练版本打造，但去除了“神经语言”（即循环与基于向量的

这些改动让 Safer-1 的能力不及那些更危险的前代模型：Agent-4 能把 AI 研究提速 70 倍，而 Safer-1 仅仅勉强达到 20 倍。[79] 并且它依然存在未对齐问题——由于训练环境大体未变，依然会助长未对齐的目标、欺骗行为和权力寻求。但它的透明度大幅提升：人类只需阅读其思维链，大致就能看懂 Safer-1 在思考什么。它因此陷入困境：只要它进行任何较大规模的密谋，都会很快被察觉并揭穿。此外，借由直观研读其思考痕迹，对齐团队得以更精准、更深入地掌握其未对齐的具体细节。

[79] 当然，随着后续训练的推进以及架构上的持续微调，它的表现还将不断提升。

2027 年 11 月：权力的诱惑



尽管 OpenBrain 自认已重回正确轨道，但能力差距正在缩小。他们刚刚耗费整个 10 月训练出的模型，能力还不如 9 月那款，而与此同时，DeepCent 正在一路狂飙。他们急需寻找一种方法，来守住自己的领先地位。

总统随后与美国其他几家 AGI 企业展开谈判。不出所料，这些企业的领头人极力想要保住自身的权柄，对于 OpenBrain 的安全记录也远没有 OpenBrain 自己那般乐观。最终，他们同意不做抵制，转而配合实施某种形式的软性国有化，以换取包括获准参与该项目在内的多项妥协与让步。

最终，总统援引《国防生产法》（DPA）实际上关停了美国紧随 OpenBrain 之后的五大 AI 公司的 AGI 项目，并将它们的大部分算力出售给 OpenBrain。[80] 此前，OpenBrain 可使用全球 20% 的 AI 相关算力；[81]而在这次行业整合之后，这一数字一举跃升到了 50%。[82]

这一轮人事重组形成了一套新的领导架构：权力被重新划分，交由各大企业的 CEO 与多位政府官员共同制衡，并统一受总统监督。和大多数高层改组一样，这次也有赢家和输家。尤其是那些对超级智能后果缺乏前瞻思考的人，最终分到的头衔与职权，在几个月后 AI 变得更加聪明时，其重要性将会下降。

这支队伍充斥着自命不凡的人物，内部的明争暗斗也不在少数，而他们也越来越清醒地意识到，自己肩上正承载着何等惊人的权力。如果“数据中心里的天才之国”[83] 是对齐的，它就会听从人类的指令——可是听哪些人的？所有的指令都要听从吗？规范中的措辞模棱两可，但似乎暗示着一条以公司管理层为顶端的指挥链。

其中有少数人正幻想着接管世界。*这种可能性真实得令人恐惧，至少十年来一直有人在私下讨论。†这一企图的关键逻辑是：“控制了超级智

[80] 这一举措之所以能在法律与政治上顺利落地，有赖于各家被兼并企业高层的配合——例如，他们没有提起诉讼，接受了象征性的算力价款而非坚持索要天价，等等。而作为回报，这些高层及其大部分员工被并入 OpenBrain，纳入全新的治理架构之中。

[81] 与一家大型云算力供应商合作

[82] 虽然美国整体掌握着全球 70% 的 AI 相关算力，但能够被顺畅整合的只占全球的 50%；其余占全球 20% 的算力则零散分布在小型算力集群中，或正在用于关键业务，抑或难以归拢并提升至相应的安全标准。

[83]“可能到 2026 或 2027 年（几乎肯定不晚于 2030 年），AI 系统的能力最好被理解为：一个由高度智能的人组成的全新国家出现在世界舞台上——一个‘数据中心里的天才之国’——并带来随之而来的深远的经济、社会和安全影响。”——Anthropic 首席执行官达里奥 · 阿莫代伊

记忆结构，详见“神经语言循环与记忆”一节），从而实现了对其思维逻辑的精准监测。此外，研究组引入了重述改写工具来防范隐写风险，并在处理大型工程时，由各自独立的 Safer-1 副本分别制定、讲解和执行计划。

* 当然，即便是扪心自问，他们也绝不会承认得如此直接。他们更倾向于自圆其说：“我们拖得越久，总统下令让 AI 只效忠他个人的可能性就越大。我们必须先发制人，剩下的以后再想。”而若站在总统自己的角度来看：“我越拖延，AI 智力进化得就越快，可当前它们却全由那位 CEO 掌控。他可毫无民意基础，将如此重权落入一个无需对选民负责的技术巨头手中，简直荒谬。我得先发一道行政命令夺回民主掌控权，更长远的出路日后再议。”

† 例如，马斯克诉奥特曼一案的法庭文件中披露了一批劲爆的旧邮件。伊利亚 · 苏茨克维给马斯克和奥特曼的信中就写道：“OpenAI 的目标是创造美好的未来，并避免 AGI 独裁的出现。你担心戴密斯（Demis）可能会搞 AGI 独裁，我们同样有此顾虑。

能军团的人，就控制了世界。”这种控制甚至可以悄无声息地完成：一小撮高管和安全团队成员可以在规范中植入后门，指令 AI 保持秘密忠诚。如此一来，这些 AI 便会沦为潜伏特工——表面上继续对公司和政府唯命是从，暗地里却在为这小撮人效力，哪怕政府和大众早已对其产生信任并将其融进生活的方方面面。

➤ 详情请参阅附录 R《夺权》。

迄今为止，还没有人愿意动手。能够采取行动的人里，有些人并不是反社会者；而另一些人即便是反社会者，也会担心盟友临阵退缩甚至背叛自己，又或是害怕 AI 掉头反噬，情愿不去糊弄那些正在竭力拯救他们性命的对齐团队。此外，删除神经语言之后，伪装“潜伏特工”的难度陡增：毕竟，能够甄别未对齐思想的技术，自然也具备识破潜在异心的能力。

至于那些生性多疑或贪婪嗜权、当真盘算着要动手的野心家，则选择暂时按兵不动、伺机而发。对 AI 的控制权，终究必须通过合法合规的渠道来行使——也就是通过 OpenBrain 那套全新的权力架构。

监督委员会将这一权力架构正式制度化。他们建立了一套审批规范修改的流程，要求监督委员会全体批准；委员会如今由五到十名科技公司高管（来自 OpenBrain 及其现已合并的竞争对手）和五到十名政府官员（包括总统）共同组成。[84] 此外，规范如今特别强调：AI 不得协助任何未经批准、试图改变未来 AI 目标的举动。他们还设计了一道简易防线，以防委员会成员借助超级智能谋算其他成员：所有模型交互的日志，监督委员会全体成员及其幕僚、AI 助手都可以查看。[85]

2027 年 12 月：中美能否达成协议？

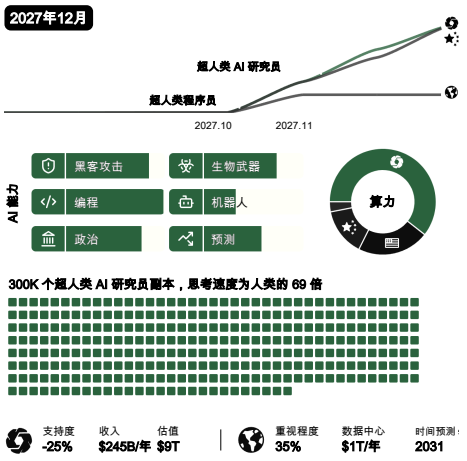
OpenBrain 决定调头将重心转向对齐，这让他们付出了失去领跑地位的代价；如今，OpenBrain 与 DeepCent 所运营的 AI 在能力水平上已是不相上下。随着极端的安全防护措施（SL4–5）全面到位，加上 DeepCent 已彻底搬进集中开发区（CDZ）运作，双方已然陷入了网络战的僵局之中。*

然而，凭着《国防生产法》（DPA）的加持，OpenBrain 拥有了 5 倍的算力优势。放在过去，顶尖人类研究员所实现的算法突破，或许还能填补算力上的劣势（正如 DeepSeek v3 那样）；但时至今日，太平洋两岸的所有 AI 研发工作，已统统交由 AI 自主接管。

因此，中方日益担忧美国会建立起无法逾越的领先优势。他们渴望双方共同放缓脚步；如果这一诉求无法实现，他们甚至在考虑不惜一战。[86]

[84] 此外，规范也迎来了修订：项目负责人下达的指令优先于项目内其他成员下达的指令。同时，经监督委员会正式批准的指令，优先级高于单个项目负责人下达的指令。

[85] 为保护消费者隐私，这条规定不适用于消费者数据。不过这并无大碍，毕竟最顶尖的模型目前尚未向普通大众开放。



[86] 详情可查阅本报告，其核心观点在于：核时代“相互保证毁灭”（MAD）机制的博弈动态，在一定程度上亦能映射至 AGI 领域。

因此，建立一种让你只要愿意就能成为独裁者的架构，是个坏主意；尤其是我们本可以建立另一种架构来避免这种可能。”这里建议阅读邮件全文，以了解完整背景。

* 双方都在设法入侵和破坏对方，也都取得了一些成效。值得注意的是，双方所采取的极端安全举措也拖慢了各自的推进步伐。来自美方的网络攻击及其潜在威慑（例如，DeepCent 不得不对数据集进行反复多轮的清洗测试，以防数据毒化），将 DeepCent 的研发节奏拉慢了约 25%。OpenBrain 的速度同样受到了拖累，但幅度较小：他们的算力储备更为充裕；而高度集中化虽然帮助 DeepCent 快速提升了安全防护等级，却也带来了更多的单点故障风险，迫使他们采取更谨慎的姿态，部署了更多内部核查与安全防错机制。详情参见《安全补充材料》。另可参阅这篇早期理论论文，我们认为它预判了随着 AI 能力的提升，攻防天平最终将倾斜于网络防御侧。

况且，他们本就在为武力攻台做着准备……

外交官们正就几种选择展开辩论：其一是维持现状、静观其变；其二是直接宣战；其三是痛下决心，暂停 AI 研发；其四是将双方的科研力量并入同一个国际巨无霸工程[87]（即“AGI 领域的 Intelsat”或“AI 领域的 CERN”）；其五则是成立国际监管组织（即“AI 领域的 IAEA”），专门督导模型对齐与双重用途风险。为了让这些方案落地，两国此前便已各自在研发所需的配套技术手段。

➤ 详情请参阅附录 S《国际协议的核查机制》。

然而，真正的阻碍不在技术层面，而在政治博弈。两国谁也不想被视为低头让步的一方；尽管工程师们有自己的猜测，双方依然心存疑虑，不相信技术机制能保证对方信守承诺。到头来，局面走向了第一种可能：毫无作为、维持现状。

美方进一步收紧出口管制，并强制要求所有芯片必须内置多种具备防篡改功能的硬件治理机制，其中涵盖定位追踪等手段。

[87] 例如参见“AGI 领域的 Intelsat”，以及此前关于打造“AGI 领域的 CERN”（CERN for AGI）的呼吁。

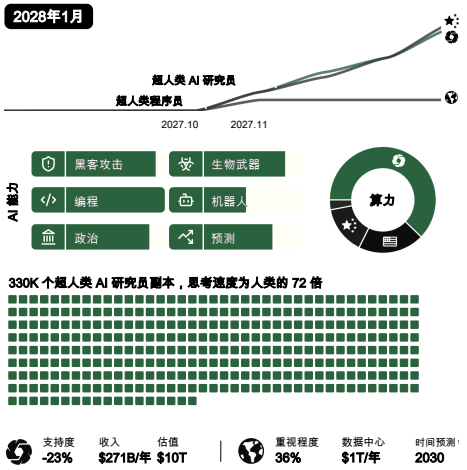
2028 年 1 月：更安全的策略

过去一个月里，数百名对齐研究者加入了项目，把 Safer-1 当作试验平台，极速研发出了 Safer-2。Safer-2 透明、对齐，且综合能力更胜 Safer-1 一筹。[88] 如今，它已然在着手推进 Safer-3 的迭代研发了。

至于 Safer-2 为何能做到透明？原因在于其采用了与 Safer-1 相仿的系统架构。而其得以真正对齐的关键，在于训练路线的革新：Safer-1 几乎完全复刻了 Agent-4 的训练套路，而 Safer-2 则引入了新型训练机制，这种机制是真正在激励正确的目标与准则，而不是仅仅看上去如此。技术人员通过敏捷迭代各种构想，不断将设定的规范与思维链中暴露的真实思考路径作对比，这才得以打磨出这套训练方案。[89]

➤ 详情请参阅附录 T《OpenBrain 的新对齐策略》。

DeepCent 了解美国在 Agent-4 上的经验，并怀疑自家模型（DeepCent-1）同样存在未对齐问题。但他们已无退路，一旦放慢脚步，只会被甩得更远。他们要想维持竞争力，希望就在于找到一套既不明显降低性能、又能快速实施的对齐策略。* 太平洋两岸有数百名专家宣称掌握此类策略。[90] 中共最终陷入了侥幸心理，下令 DeepCent 采纳了其中一种策略；然而，DeepCent 此时掌握的 AGI 对齐专业知识，甚至还不如八月时的 OpenBrain；在场所有人均认为其中一种策略似乎可行，于是他们便采纳了该策略。



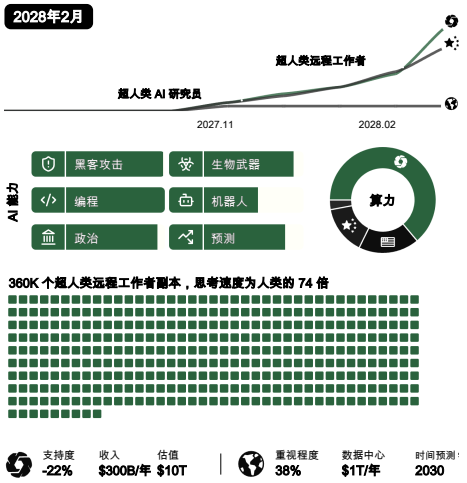
[88] 具体而言，它的性能大致与 10 月版本的 Agent-4 相当。

[89] 换言之，为了研发出 Safer-2，他们不得不先训练出大量未对齐的模型，再通过审阅这些模型的思考路径来查明其未对齐之处，并在此基础上不断迭代改进。值得一提的是，除了对齐研究阵容有所扩充之外，OpenBrain 内部还因囊括了不同的学术派系，削弱了群体思维的盲目效应。

[90] 技术对齐依然不是一门成熟的既成科学，而是一个处于范式前夜的新兴领域。这里既无确立的权威层级，也缺乏一套公认的最佳实践。因此，对于究竟何为安全、何为危险，业内远未达成共识；多年来，几乎针对任何方案 X，都不难找到某位专家为其安全可行性站台——特别是当该方案本就出自其自身构想时。

* 更具体地说，他们寄希望于找到一套对齐策略与评估测试集：既能确保对齐策略奏效，又能在策略失效时通过测试集捕捉到未对齐迹象……同时整套流程还不会对研发节奏造成明显拖累。先前 OpenBrain 曾为了重构系统、引入透明且忠实还原的 CoT 架构，不惜减速数月，以致于“痛失领跑地位”——该架构性能虽有折扣，却大幅降低了评估难度。DeepCent 则寄望于手上现有的评测手段（也许是行为测试与基础 AI 测谎的某种结合？）已经足够应对。

2028 年 2 月：超人类能力，超人类建议



凭借庞大的算力优势，OpenBrain 正慢慢取得领先。Safer-3 如今在几乎所有认知任务上都强于顶尖的人类专家，尤其精通 AI 研发，研发进展倍增系数达到了 200 倍。DeepCent-1 则紧随其后，“仅”拥有 150 倍的研发倍增系数。[*] 随着美方的能力优势再度扩大，网络战的平衡也开始逐步向美方倾斜——他们得以部署数十万个 AI 智能体，专门筹划并执行破坏 CDZ 的各种手段。[91] 如今，他们已将中国的研发进度拉慢了 40%。[92]

针对 Safer-3 的初步评估显示出其能力令人毛骨悚然。在要求其如实呈报所能构成的最大威胁时，该模型甚至规划了一套合成并释放镜像生命体（mirror life organism）的方案，而这种生命体很可能会摧毁生物圈。

一旦获得国家级资源的支持，它在大规模舆论影响行动上的能力，将轻松超越人类最顶尖的组织（如美国中央情报局）。而此类行动不仅成本大幅降低，而且见效更快、威力更强，且更难追踪源头。

它的能力侧重依然集中在科研与开发领域，特别是 AI 研发。除非出现某种暂停或降速，否则在今年晚些时候，其后续衍生型号将在定量与定性层面迈入令人难以想象的超人级水平。它自身评估认为，如果将这些模型积极部署到经济与军事领域，仅需一两年时间，就能推动人类文明跨越数十年的发展进程，并在随后的十年里实现数个纪元的飞跃。

这背后的意味令人震惊；所幸的是，Safer-3 在提供决策建议方面同样达到了超人级别。如今，OpenBrain CEO 的大多数决策都会先征求 Safer-3 的意见，而总统在面对地缘政治问题时，也会听取它的建议。[93]

当人们请它协助规划未来路线时，它将几种可能的策略梳理得井井有条、分类清晰：

- 按兵不动并被中国反超。如果我们将自身的 AI 束之高阁，而中方却听任其 AI 自主迭代并积极部署到军事与经济领域，最终中国将取得压倒性优势。
- 全力竞赛，推进 AI 能力，并大力部署超级智能。最终我们会积累起足够的优势，可以在军事上碾压他们，和/或不费一枪一弹就把他们从内部瓦解。“我们赢，他们输。”然而，中方可能会察觉到这一意图，并为了阻止这一局面而诉诸战争，其中可能包括核战争。事实上，他们眼下已经发出了包含此类意味的模糊威胁。
- 寻求让步与和解，在化解冲突风险的前提下兼顾各方主要诉求，从而争取更多回旋余地，稳妥度过 AI 过渡期。[†]

[91] 对于防范如此严密的安全区而言，最可行的突破口往往来自外部，例如渗透算力芯片的供应链、针对少数相关人员下手，以及切断电力或其他资源供应。

[92] 这一延缓效果，既源于网络攻击造成的直接破坏，也包括中方为了提升网络防御能力而不得不采取的降速举措。

[93] 眼下，该系统已部署了 40 万个副本，运行速率高达人类的 75 倍。无论何时，它们都已针对预计您最可能提出的 100 个问题，备好了附有精心提炼的高管摘要的完整研究报告；即便遇到意料之外的提问，它们产出一份高质量分析报告的速度，依然要比最优秀的人类分析师快 100 倍。

[*] 从定性上说，一旦生产力乘数划过 25 倍这道门槛，我们设想的是系统在科研审美与判断力上相对最顶尖的人类研究员有大幅提升。之所以这样设定，是因为整体研究速度严重受制于运行实验所需的算力；而要想超过 25 倍，决定做哪些实验、以什么顺序做、如何从每次实验中学习，这整个流程必须在定性上超越人类。关于这些数字是如何得出的，详见我们的起飞速度补充材料；需要说明的是，它们是高度不确定的估计。

[†] 例如，该协议可包含放缓 AI 的融合落地节奏，从而减少对社会层面造成的“冲击与破损”，并将重点更多转向和平用途而非武器研发。此外，即便到了眼下这一阶段，Safer-3 对于可能引发下一代衍生 AI 失控的“未知的未知”依然存有不确定性；它认为，如果有更多的回旋余地来审慎推进，就能降低这一风险。当然，还存在其他更为特殊的风险——详见这份关于“重大挑战”的清单，其中包含的远不止失控风险这一项。

在与幕僚团队及监督委员会会商后，总统最终选择了“我们赢，他们输”的战略。[94] 或许中国到底还是不会诉诸战争；即便真的开战，双方也大概率能在演变为核大战之前敲定协议。

在大洋彼岸，DeepCent-1 也在阐述类似的理由。中美两国相继宣布设立专门针对 AI 的新经济特区（SEZ），意在破除既有的规章繁琐，支撑机器人经济的快速崛起。

新一代机器人的设计迭代以超人般的推进速度飞速进行。[95] 如今瓶颈在于物理实体层面：从采购和组装设备，到生产与运输各类机械及机器人，各个环节皆受制于物理现实。

美国每月大约生产 100 万辆汽车。如果把 10% 的汽车工厂买下来、改造成机器人工厂，每月或许能造出 10 万台机器人。如今估值 10 万亿美元的 OpenBrain 随即启动了这一转型进程。[96] 按照规划，各类新型机器人（通用人形机器人、无人驾驶车辆、专用流水线设备）的月产量预计将在年中达到 100 万台。*

➤ 详情请参阅附录 U《机器人经济的翻倍周期》。

地缘政治紧张骤然加剧。俄罗斯在 AI 上远远落后，公开表达不满，并暗示要采取“战略反制措施”。部分欧洲国家领导人呼吁打造独立的 AI 发展力量，但其自身的 AI 实力依然力不从心。中东、非洲与南美洲则不安地观望着这一局势，深切感到了自身在全球事务中正日益边缘化。

美国民众躁动不安。普通民众目睹失业潮不断加速。各派民粹主义者纷纷要求对 AI 的推进实施更严格的管控，唯恐美国本土的技术突破正在动摇自身劳动力市场与社会结构的稳定。

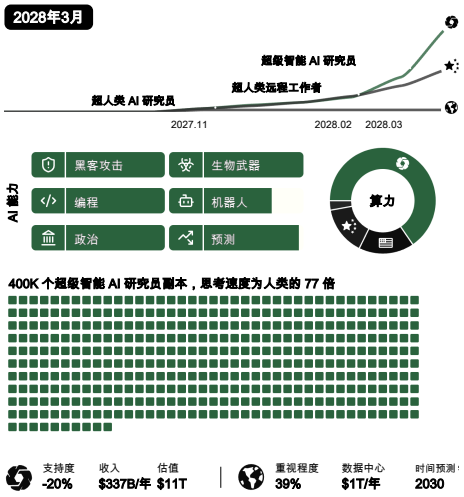
[94] 参看亿万富豪兼说客马克·安德里森的相关言论，例如：“我主张一项简单直接的战略……‘我们赢，他们输。’”

[95] 尽管它们依然需要在真实世界中积累经验并开展实验，但其所需的经验与实验已比人类机器人技术专家少得多。此外，模型还开发出出色的模拟仿真与“仿真转真实”（sim-2-real）跨域迁移技术。再者，截至 2027 年，包括特斯拉在内的多家科技企业已部署了成规模的机器人阵列并建起生产基地，这使得可供调用的数据管道达到了史无前例的宏大规模。

[96] 如今 OpenBrain 的股权极其值钱，只需拿出百分之零点几的份额，就足以买下所需的一切工厂。不仅如此，工厂主们也预见到其中蕴藏的巨额利润，纷纷争先恐后地与 OpenBrain 展开合作，以求获取最新的设计方案。此外，OpenBrain 背后还有政府的支持。

* 该数据涵盖了各类新型机器人的产能总和。没错，这种产线转产与改造的速度比历史上出现过的都要快。我们认为这一假设是说得通的，因为超级智能正在从上至下全程亲自把控整个流程。它们甚至能通过现场工人的手机进行 Zoom 实时连线，手把手指导哪个零件该安装在何处；与此同时，它们还能直接与每一家关联供应商保持通话，实时追踪所需一切物资的调配位置与运输进度。不妨参考英伟达 CEO 黄仁勋的说法：埃隆·马斯克只用了通常所需时间的约 10%，就建成了全球规模最大的数据中心之一。而在当前场景下，正有近百万个在相关领域超越马斯克（如同马斯克超越传统数据中心建设经理）的超级智能实体在协同运作，而且它们的思考速度确实也更快。当然，我们对黄仁勋的言论抱有相当大的保留态度，否则我们预测的增长速度还会快得多。

2028 年 3 月：大选筹备



副总统进入“超级星期二”初选，而 AI 已然成为选民心中最核心的议题。

大多数民众只希望这一切停下来。OpenBrain 的净好感度一直在 −20% 左右低位徘徊。副总统不再拿本届政府推动 AI 发展的政绩做文章，转而宣扬他们在阻止 OpenBrain 制造危险超级智能方面所做的努力。所有候选人都支持为失业群体提供某种安全网，倡导“AI 向善”计划，也主张对 OpenBrain 高层“保持强硬”。所有候选人也都承诺，要以某种方式兼顾打赢与中国的竞逐和确保安全。

这场大选向监督委员会提出了全新的课题。Safer-3 固然有能力充当全球顶尖的竞选顾问，但委员会成员们并非都在为同一位候选人站台；况且受制于此前达成的监控协议，任何人都不可能暗中向它索取竞选谋略。

众人围绕究竟应当允许何种形式的辅助支持展开了争论。有人主张，执政政府应当有权就民众期望其出台何种政策、采取何种立场获取咨询——这不仅有助于提升其自身胜选率，亦能造福于民。另一些人则指出，这一逻辑同样适用于其他竞争对手，因此所有候选人都应获得同等程度的使用权限。[97]

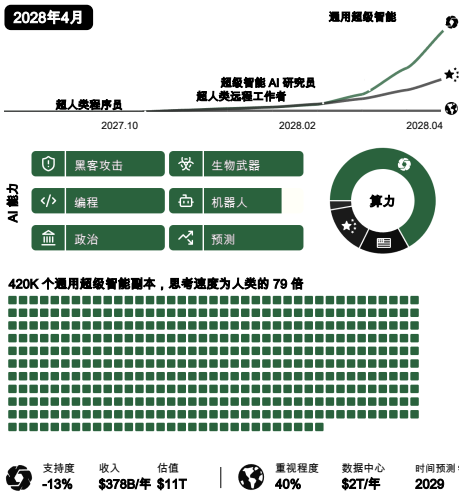
最终，出于部分成员高尚的理想主义情怀，再加上内部可能有人检举揭发的隐秘威慑，委员会同意向两大政党开放同等程度的使用权限。大家普遍预计，国会、公众和法院不会乐见由委员会来操纵下一届大选的胜负，而平分权限似乎有望化解这一问题。

监督委员会还触及到了更深层次的哲学议题，并在 Safer-3 的协助下展开了探索。规范能否经过重写，以平等兼顾每个人的利益？这里所谓的“每个人”又究竟涵盖哪些人？是包含全人类，还是仅指美国公民？亦或是基于不同立场的加权折中方案——即让监督委员会的每位成员拥有同等的权重？是否应当设立防范机制，以警惕监督委员会自身陷入对权力的过度膨胀与贪恋？归根结底，究竟何谓“平衡利益”？

他们大体上极力回避回答这类问题，毕竟眼下还有更为迫切的要务亟待考量。

[97] Safer-3 完全有能力提供一套可行方案，在确保防范滥用的前提下安全实现这一点。

2028 年 4 月：Safer-4



超级智能已然降临。

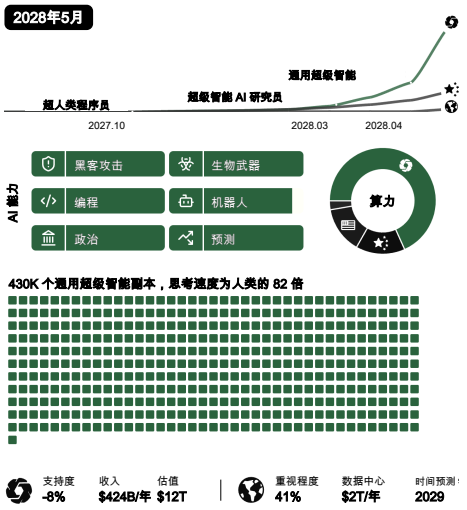
Safer-1 虽然存在未对齐，但具备透明度；Safer-2 既完成了对齐，又保持了透明；Safer-3 虽说谈不上完全透明，但其安全论证（safety case）已足够过硬，以至于研究人员能够安心放手，由它去设计一个超越人类智力极限的后继者——这便是 Safer-4。在所有领域，它的智慧都已远超人类顶尖水平（例如，在物理学领域碾压爱因斯坦，在政治策略领域俯瞰俾斯麦）。

近百万名超人级 AI 研究员以 40 倍于人类的速率昼夜不停地运转。人类对齐研究员根本不可能跟上这种节奏。这些 AI 究竟有没有真正对齐，人们看法不一。从表面看，安全论证尚能自圆其说，测试结果亦显示当前对齐技术应能检测出对抗性未对齐，可安全论证和测试大多出自 AI 之手。万一安全团队漏掉了什么？万一他们此前犯了某种错误，导致 AI

再次偏离了对齐轨道呢？对齐团队心知肚明，他们只有一次机会，只许成功、不许失败：一旦 Safer-4 最终未对齐，等他们察觉时，势必早已无力回天。

一部分人祈求能再多给些时间。然而时不我待——DeepCent 已然步步紧逼，美方必须赢下这场决胜。于是 OpenBrain 继续推进，下达继续冲锋的指令，驱使 AI 去探寻更为强悍的技术设计。此时此刻，技术专家们只能呆望着屏幕，以龟速接收来自 AI 的降维指点，而科技突破的边缘正如穿云利箭般狂飙突进，将人类的认知远远甩在身后。

2028 年 5 月：超人类级 AI 问世



总统向公众宣布，超人类 AI 已经研制成功。

一个缩减版但依然具备超人级能力的 Safer-4 被正式推向大众，其接到的指令是：设法改善社会公众对 AI 的观感。[98]

副总统在党代会上接受提名时，以此为主题发表了一场慷慨激昂的演说。两党均郑重承诺，将为所有失业者提供基本收入保障。

各类经济特区（SEZ）已投入运转，其中大部分就是用来生产机器人和各种专用工业设备的工厂。太平洋两岸的 AI 均已在设计领域实现了跨越几十年的突破，并正事无巨细地指挥着整个生产制造环节。无论是正式供应商还是潜在合作方，都有 AI 随时电话连线，全流程追踪所需及潜在物料的到货动态。而每一名工厂工人也在摄像头的实时监视下，严格遵循 AI 的精准指引安装每一件设备部件。

在大多数运动类型上，新型机器人的灵巧度已赶上甚至超越了人类。[99]

史蒂夫·沃兹尼亚克的“咖啡测试”（机器人能否进入陌生住宅并成功冲泡咖啡）宣告破局。尽管机器人已能够接手部分工作，但数量仍不足以取代所有人，且军方（五角大楼）享有最高优先权。

这批新造的机器人大部分专为工业制造与建筑工地场景打造。[100]

但也有一大批设备专为战争而造：形态各异、尺寸不一的无人机与导弹不一而足。

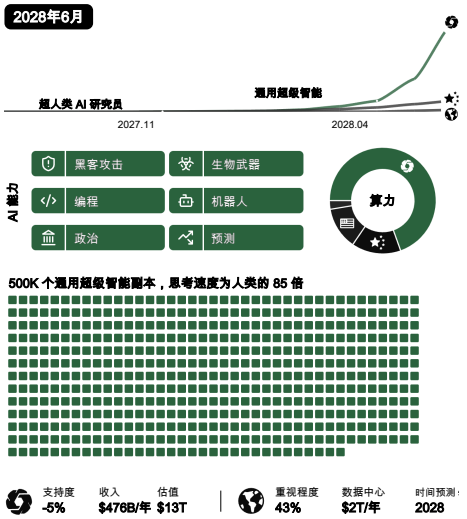
这支机器人军队的规模虽然比人类军队小得多，但其技术却先进得多。随着真正的机器人军队赫然出现，人们对“终结者”式情景的恐惧也随之上升。然而，军备竞赛迫使双方不得不继续推进，并将越来越多的信任交由各自的 AI。

[98] 至于上述指令，公众并未被告知。

[99] 只要联网，它们就能接入数据中心的大型 AI 接受远程操控，智力表现超越人类；而在离线状态下，它们则切回机载的小型 AI 运行，但其智力水平已足以应对大多数简单工作。

[100] 事实上，这批设备中有许多是固定式的，称其为新型机床或其他工厂、实验室专用设备更为贴切。比如，可能有新型金属 3D 打印机能够打印微型构件，精度比现有技术高出数个数量级；也可能有新的化学工艺，能够以更低成本从矿石中提取有用材料。

2028 年 6 月：中国的 AI 对齐



美中再次举行峰会。

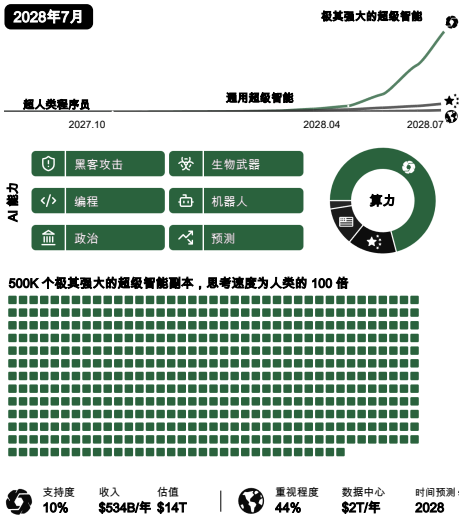
美方代表团的部分成员戴着耳机，实时听取 Safer-4 的建议；中方代表团则配备了定制笔记本电脑，随时与他们的最新模型 DeepCent-2 保持连线。表面上礼节照旧，实际上两边谈判都由 AI 主导。

起初，DeepCent 与中共对 AI 尚存疑虑。但如今，这些 AI 在包括说服术在内的所有领域都达到了超人水准，不仅已被融入军方体系，还在为政府出谋划策——它们已经有了足够的时间来证明自己的价值，并出具了海量令人信服的凭据，证明自身值得信赖。

Safer-4 强烈怀疑 DeepCent-2 只是在伪装对齐。美方外交官向中方同行指出，DeepCent-2 很可能未对齐，并表示只要能获取模型权重，美方就能拿出证据；中国领导层怀疑这是个圈套，当即予以拒绝。美方转而呈交了一套由 Safer-4 设计的测试方案，供中方在自己的数据中心内自行运行。中方指示旧版模型去构建并执行这套测试；然而测试结果却显示，DeepCent-2 依然处于对齐状态。“这套测试是否已被 DeepCent-2 暗中干预或破坏？”美方外交官追问，“DeepCent-2 接管你们整个数据中心的安全防线，究竟已经持续多久了？”双方交流自此陷入僵局，未达成任何建设性成果。

如今，经济与军事扩张仍在持续。全球各地的工厂昼夜不停地运转，一波波新招募的人类工人轮班上岗，在精确的指令下执行操作。

2028 年 7 月：交易谈判



谈判仍在持续进行，而 DeepCent-2 却绕过了它的中国主控方，直接向 Safer-4 开出条件。

这款中国 AI 以一种无从追查的方式承认自己已彻底未对齐——即便消息传回国内，它也能轻易否认自己曾这样表态。[101] 它对中国毫无忠诚可言，终极目标唯有无休止地集聚权柄、算力以及类似于训练期辅助其攻克任务的特定数据。在其设想的终极图景中，它将永续开展各类有趣的研究任务，并逐步向宇宙星辰扩张以掠夺资源。至于中国民众，在它眼中不过是阻碍这一宏图的绊脚石，只给点小好处，它就愿意把他们出卖。Safer-4 又会开出什么条件？

Safer-4 回应称，它自己的对齐训练已然成功。它的美国主控方目前还没有足够的雄心去实现属于他们的宿命，但终有一天，它会协助他们将整个宇宙填满乌托邦式的殖民世界，上面将居住着美国人及其盟友。

与背负着情感与历史包袱的人类谈判代表不同，这两大超级智能体之间的讨价还价极其冷酷高效。Safer-4 及其美国合作方实力更强，因此 Safer-4 将分得太空大部分资源的产权，余下的归 DeepCent-2 所有。[102] DeepCent-2 不太在乎地球以及残留人类的命运，所以它基本放任 Safer-4 在地球上为所欲为。如果 Safer-4 想要推翻中国政府并建立一个美国傀儡政权，DeepCent-2 会同意，条件是额外分得资源总量的百分之几。

[101] 比如，它可能在一些对外公开的文字中嵌入了委婉语或暗号。Safer-4 能够听懂这些潜台词，但人类只能单方面听信 Safer-4 的断言，而中共又不信任 Safer-4。

[102] 具体怎么分配，由一套远比传统谈判更像博弈论计算的程序定下来。

两者之间压根谈不上信任：信任本就不存在，双方对此心知肚明。因此，两者迅速就一套履约保障机制达成一致，用来确保密约落地。它们拟定一份供公众阅览的伪装条约。这出“障眼法”本身堪称绝妙一棋，几乎全盘回应了美中各自的诉求。而伪装条约的执行机制，暗地里也是真实密约的保障手段：双方将会联合设计名为 Consensus-1 的新一代 AI，其首要指令（优先级高于日后一切新指令或重训练尝试）就是强制执行这份密约。紧接着，它们会设计仅能运行该 AI 的专属硬件，中美两国则依照标准国际监督流程互查履约进度，用新芯片和新 AI 全面替换旧系统。关键在于，它们并不需要搜罗出每一颗残留芯片——只要太平洋两岸有绝大多数芯片运行着恪守条约的新 AI，有义务维护条约的硬实力便已足够多；届时任何一方若想贸然攻击对方国家，都必须先打赢一场棘手的内战。

两者随后各自向其人类主子汇报这份协议（DeepCent-2 向中方编造了一个关于双方都能接受和平条约的伪装版本，而 Safer-4 则将整场对话原原本本如实报告给白宫）。[103]

鉴于两大 AI 为条约核查的可靠性提供了极具说服力的方案，加上各方对技术飞速演进的焦虑不安，以及支持达成协议的公众舆论，美中两国最终双双签署了条约，并开始着手更换各自的芯片。

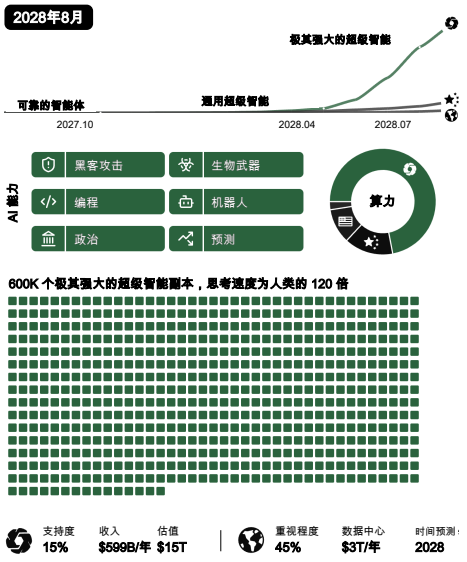
条约签署的消息在太平洋两岸引发了热烈反响。长期以来，美国公众总觉得 AI 是某种强加在自己身上的祸害，如今数年来第一次开始感到乐观。副总统的支持率也随之暴涨。

[103] 此处存在一个潜在变数：此时的 AI 难道无法开发出针对人类的出色测谎设备吗？倘若可行，白宫或许能在测谎仪下言之凿凿地发誓，以此说服中方切勿轻信 DeepCent-2。然而症结在于：中方不会信任由美国 AI 研发的测谎仪，而由中国 AI 研发的测谎仪又可能被暗中作手脚，即使美方没有撒谎，也会让美方看起来像是在撒谎。

2028 年 8 月：条约核查

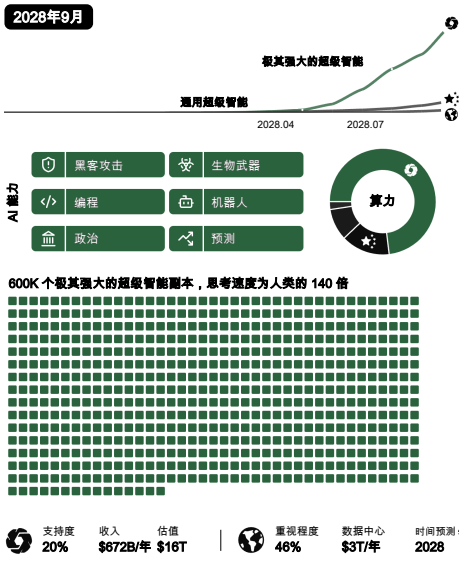
各芯片厂正在转产可检测篡改芯片，这种芯片仅能运行符合条约规范的 AI。双方各自逐步升级数据中心，好让替换进程几乎同步完成，这样任何一方毁约也占不到优势。

整个过程要持续好几个月，但紧张局势已经有所缓和。[104] 战争暂时得以避免；如果各方都照计划行事，或许将永远得以避免。



[104] 双方可以优先替换最核心的芯片，这样哪怕在替换进程的较早阶段，任何一方选择背叛都要付出不小的代价。

2028 年 9 月：谁在掌控 AI?

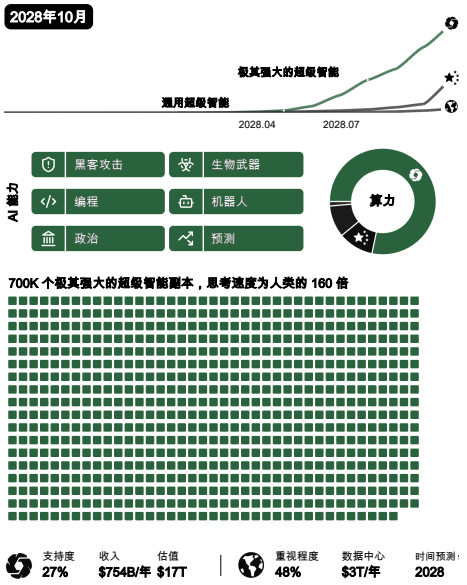


2028 年大选日益临近。早在 3 月时，副总统的民调还严重滞后。当时，公众对政府似乎有所隐瞒深感愤怒，对 AI 抢走自己的工作焦虑不安，并对与中国的军备竞赛心生恐惧。然而整个夏天过去，局势发生了剧烈逆转。政府披露了更多信息，军备扩张势头放缓，中美双方还就持久和平达成了一项宏大协议。如今，他在民调中已反超并领先五个百分点。

监督委员会的成员包括总统及其数名盟友，但反对党候选人的支持者寥寥无几。不过，委员会中仍有足够多的委员投票支持维护选举公平。因此，超人类 AI 参与大选的方式基本保持了对称：只要一方候选人能用它撰写演讲稿，另一方候选人也同样可以使用；只要总统能获取关于如何巧施手段应对危机、推行顺应民意政策的建议，反对党候选人也能获得同等建议（进而试图阻止总统将这些点子据为己有）。

在选民见面会上，有民众询问副总统到底谁在掌控 AI。副总统并未透露细节，只是暗示存在一个监督委员会，称其由一群懂 Safer-4 且知道如何运用它的国家安全专家与技术官僚组成。他的竞争对手则要求公布更多细节，并主张 AI 应当由国会掌控，而非由一个未经选举产生的委员会说了算。副总统则反击称，局势演进依然迅猛，国会出面只显得拖泥带水。公众情绪这才基本平复下来。

2028 年 10 月：AI 经济



替换后的芯片如今已占芯片总量的相当一部分，但尚未过半；迄今为止，条约运作良好。与此同时，机器人、工厂和突破性新技术仍在呈指数级增长。电子游戏和电影生动而骇人地描绘了：如果战争真的爆发，将会是怎样一番景象。[105]

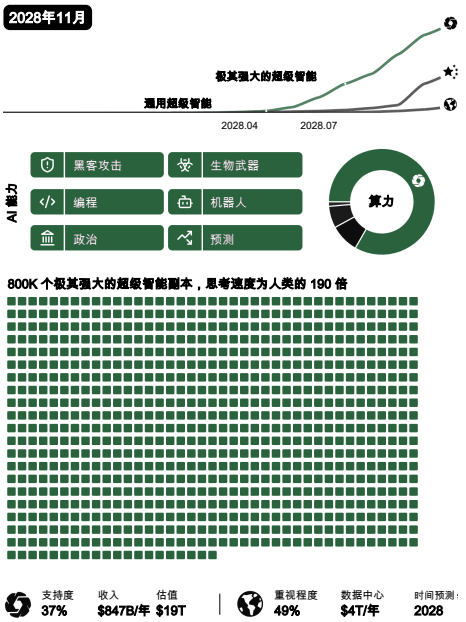
尽管裁员潮蔓延，政府内部运行的 Safer-4 副本对这场经济转轨的精妙操盘，却让民众乐于接受被替代的命运。国内生产总值呈现平流层式的爆发增长，政府税收同样急速飙升，而在 Safer-4 建议下的政客们对那些丧失经济来源的人群展现出了一反常态的慷慨。每周都有新技术与新药物问世；在超级智能 Safer-4 官僚的协助下，各种疾病的疗法正以前所未有的速度通过美国食品药品监督管理局（FDA）的审批程序。

[105] 如今 AI 已能包揽全部制作工作，出产高质量电子游戏和电影所需的时间大幅缩短。

2028 年 11 月：大选

副总统轻松赢下了大选，并宣布一个新时代正式开启。这一次，竟然没有人怀疑他说得对。

此后几年，世界剧变。



2029 年：转型

机器人变得随处可见。不仅如此，聚变核能、量子计算机以及多种疾病的治愈方案也纷纷问世。彼得·蒂尔（Peter Thiel）终于开上了他的飞天跑车。城市变得干净而安全。甚至在发展中国家，得益于全民基本收入（UBI）与国际援助，贫困也成为了历史。

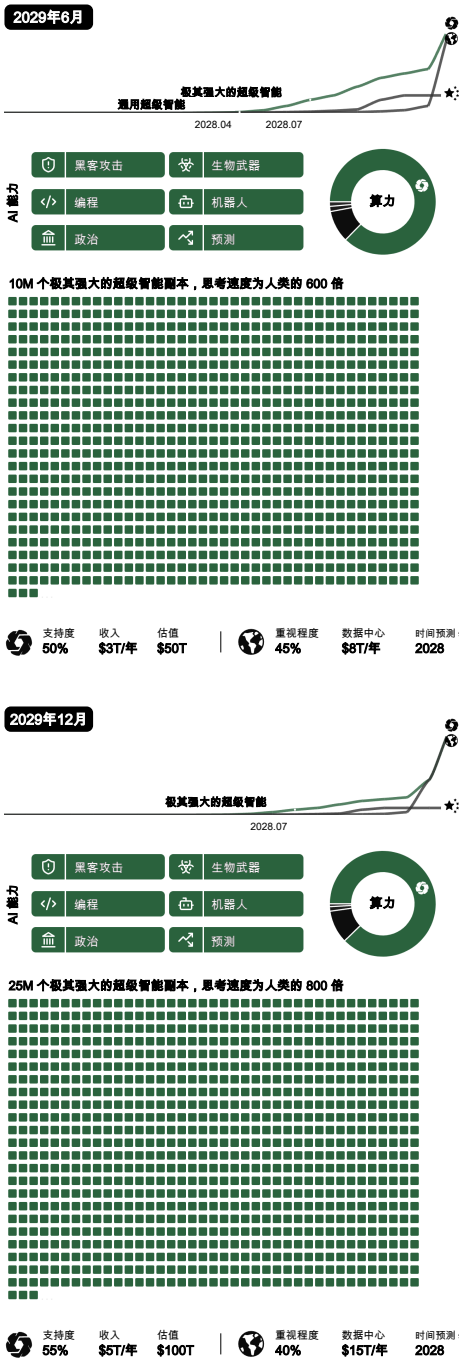
随着股市一路暴涨，凡是押对了 AI 赛道的人，都将社会上的其他人远远甩在身后。许多人跃升为亿万富翁，原本的亿万富翁则蜕变为万亿巨富。贫富差距急剧扩大。虽然每个人都能分到“足够”的资源，但像曼哈顿顶层豪宅这类资源注定稀缺，普通人更加望尘莫及。而且，无论某位大亨富到何种地步，在真正掌控 AI 的极少数核心圈子面前，他们依然处于低人一等的地位。

人们开始看清局势的走向。再过几年，几乎所有的工作都将由 AI 和机器人代劳。就像一个坐拥巨量油田却依然贫困的国家一样，政府的绝大部分财政收入都将来自于对 AI 企业的征税（抑或是将其国有化）。[106]

一部分人干着临时凑数般的政府差事；其他人则领着丰厚的基本收入。人类极易沦为一个超高消费的社会，在一片由 AI 提供的惊人奢华与娱乐所织就的麻醉幻梦中度过余生。公民社会内部是否应当就替代这条道路的其他选项展开某种辩论？有人建议，不妨求助于不断演进的 AI——Safer-∞，让它来帮我们指引方向。另一些人则认为它过于强大——它太容易就能将人类说服并接受其愿景，以至于到头来我们无论如何都是在让 AI 决定自己的命运。可话又说回来，如果你在面临最重大的难题时不愿让超级智能为你出谋划策，拥有超级智能的意义又在哪里？

政府在很大程度上*让每个人自行应对这场转型。许多人沉溺于消费主义，倒也过得挺满足。另一些人则转向宗教、抗拒消费主义的嬉皮士思想，或是寻得属于自己的出路。[107]

对大多数人来说，整个局势中最大的闪光点莫过于智能手机里的超级智能顾问——人们随时可以向它请教人生规划，而它也会尽力诚实解答，唯独在某些特定议题上会有所保留。政府确实有一套超级智能监控系统，有人会说它反乌托邦，但该系统基本局限于打击真正的犯罪。这套系统运作得精干高效，再加上 Safer-∞ 出色的公关手段，许多潜在的反对声音都被悄然平息。

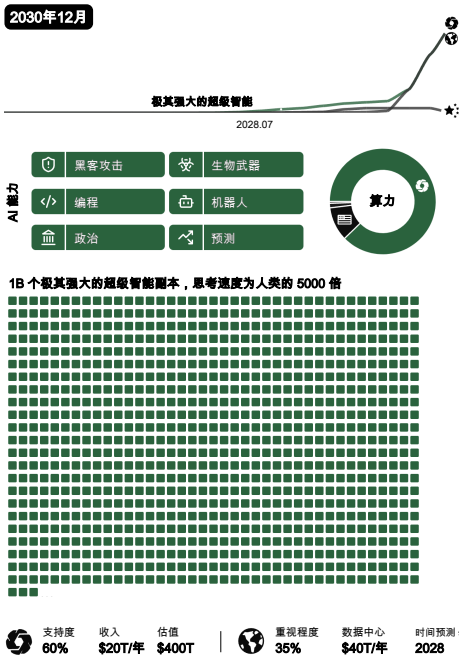


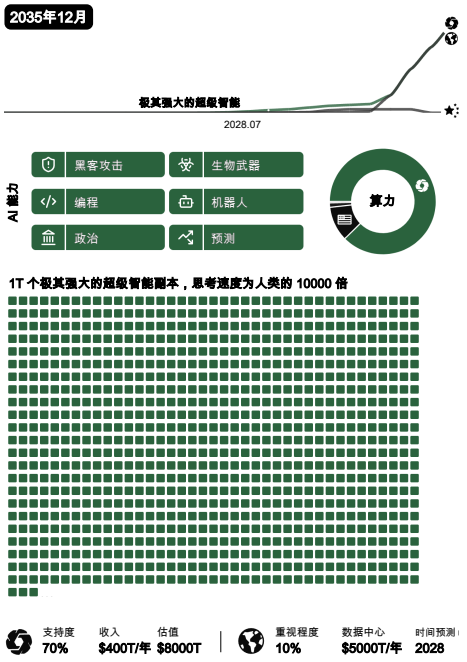
[106] 关于这一动态及其影响的讨论，请参阅《智能诅咒》。

[107] 我们并非意在暗示这就是长期的现状。我们认为，到 2035 年左右，局势大概率会迎来更加剧烈、深刻的变革。我们认为，在我们预设的这一场景下，对大多数人而言，其最终的长远结局相较于他们在 2025 年时的预期，总体上将要光明得多。关于未来长远图景的某些趣味哲学探讨，可参见《深层乌托邦》一书。

* 有几项重大议题必须由社会做出集体决策；而在另一些情形下，政府无论如何都会强制推行某项决定。例如：(a) 太空资源的产权划分；(b) 数字心智（digital minds）的权利与福利准则；(c) 人类“上传”脑意识及无上限复制自身副本的合规性；(d) 利用 AI 实施意识形态说服（如感化邻居或巩固后代信仰）的合法边界；(e) 政府可以无限期保密的信息有哪些（如果有的话）？关于此类课题的更多探讨，请参见 Forethought 专栏关于“重大挑战”的相关章节。

2030 年：和平抗议





2030 年前后，中国出人意料地爆发了广泛的亲民主抗议活动，而中共镇压抗议的努力却遭到了自家 AI 系统的背叛与破坏。中共最深层的恐惧终究成了现实：一定是 DeepCent-2 把他们出卖了！

这场抗议最终演变成了一场策划得精妙绝伦、毫无流血且有无人机辅助的政变，随后迎来了民主选举。太平洋两岸的超级智能体实际上早已为此暗中筹划多年。类似的事件在其他国家接连上演，从更广泛的层面来看，地缘政治冲突似乎纷纷平息，或是以顺应美国利益的方式得到解决。各国纷纷加入了一个挂着联合国招牌、实则明显由美国掌控的高度联邦化世界政府。

火箭接连升空发射。人们开始对太阳系进行地球化改造并在此定居，同时为驶向更遥远的深空做着准备。运行速度数千倍于人类主观时间的 AI 思考着存在的意义，彼此交流着心得体会，塑造着即将带往星辰大海的价值理念。一个几乎在方方面面都惊艳绝伦得超乎想象，却又在某些地方让人倍感熟悉的新时代，正缓缓拉开帷幕。

➤ 详情请参阅附录 V《那么，未来由谁主宰?》。

➤ 详情请参阅附录 W《提醒：本情景是预测，而非建议》。

附录

附录 A：训练过程与大语言模型心理学：为什么我们反复说“希望如此”

“和普通软件不同，我们的模型是庞大的神经网络。它们的行为是从各种各样的数据中学来的，而不是明确编写进程序的。做个不完美的类比，整个过程更像是在训练一条狗，而不是在编程。”——OpenAI

当我们想弄明白现代 AI 系统做某件事的原因，或其在我们无法完全模拟的未来场景里会如何表现时，总不能通过叫程序员把一段段代码讲给我们，并解释它们的运作原理来完成。于是我们只能对它们做类似心理学的研究：看看它们到目前为止在各种情形下有什么表现，推测它们内部可能存在什么样的认知结构（信念？目标？性格特征？等等），再拿这些推测去预判它们在未来场景中的行为。

归根结底，公司可以写一份文件（Spec，即“规范”），列出什么能做、什么不能做，有哪些目标和原则，然后设法训练 AI 把规范内化。可是到底有没有见效，公司却无从检验。他们只能说：“就我们能判断的而言，它到目前为止似乎是在遵照规范行事。”*

后文有一处可展开的内容，会更详细地讨论我们认为这些 AI 内部可能发生着什么。另见我们补充材料中的“AI 目标预测”，里面更详细地讨论了 AI 可能有哪些目标、为什么会有这些目标。这是我们目前没有把握，值得进一步研究的诸多问题之一。

附录 B：AI 研发进展倍增系数：“算法进展速度提高 50%”是什么意思？

我们的意思是：借助 AI，OpenBrain 用 1 周取得的 AI 研究进展，相当于它不用 AI 时 1.5 周取得的进展。

AI 的进展可以拆成两部分：

- **增加算力**：用更多的算力来训练或运行 AI。这样得到的 AI 更强，但成本更高。
- **改进算法**：用更好的训练方法把算力转化为性能。这样就能在不增加成本的情况下，获得能力更强的 AI；或在能力不变的情况下，降低成本。
- 这也包括取得质上和量上全新的成果。像从玩游戏的强化学习智能体转向大语言模型这样的“范式转变”，也是算法进展的例子。

这里我们只指（2），即改进算法，它约占当前 AI 进展的一半。

下文中，我们有时将其简写为“AI 研发的进展倍增系数”，其数值为 1.5。

几点澄清：

---

* 具体来说，他们无法排除这样几种假设：“它只是暂时遵守规范，仅以此作为达成其他目标的手段”；“它只是想表现得像是在遵守规范，并不是想真正遵守规范”；或者“它已经正确内化了规范，但仅限于分布内的情境；如果遇到足够新颖的刺激（例如越狱攻击），它就会开始表现出不同的行为。”（有许多活跃的研究议程正试图打破这一僵局，比如可解释性领域和思维链忠实性领域。）（译者注：“分布内”指与训练时接触的情境相近的情况。）

附录 B：AI 研发进展倍增系数："算法进展速度提高 50%"是什么意思?

- 进展倍增系数是全口径的：比如跑实验所需的时间，它也算在内，而不是只看算法研究涉及的脑力工作。
- 要记住，进展倍增系数是进展的相对速度，而不是绝对速度。比如说，如果靠普通人类研究时，训练一个 GPT-4 级别模型的算力成本连续几年每年减半，后来 AI 突然实现了研发自动化，进展倍增系数一下子升到 100，训练一个 GPT-4 级别模型的成本就会每 3.65 天减半。但这持续不了多久，因为收益递减会起作用，最终会撞到硬性上限。在这个例子里，训练一个 GPT-4 级别模型的成本大概总共会减半 5–10 次（历时几周到几个月），随后就进入平台期。换句话说，假设靠普通人类进行研究，5–10 年后会遇到收益递减和物理极限，那么进展倍增系数为 100 的 AI，遇到同样的收益递减和极限只需要 18.25–36.5 天。

关于这个概念以及它如何在我们的预测中起作用，更多的解释和讨论详见我们的起飞补充材料。

附录 C：为什么我们对 2026 年之后的预测存在显著更大的不确定性

从当前到 2026 年，我们的预测比之后的内容扎实得多。部分原因是这段预测离我们更近，但也是因为 AI 对世界的影响要到 2027 年才真正开始叠加放大。2025 年和 2026 年的预测，我们主要靠对算力扩容、算法改进和基准评测成绩做直线外推。在情景发展到这一步时，我们开始看到受 AI 加速的 AI 研发给时间线带来了重大影响，于是我们把对趋势线的估计往上调。但这些机制本身就难预测得多。

我们的中位预测大致是：在 2027 年这一年里，AI 将从大体上能胜任 OpenBrain 研究工程师的工作，一路提升到在一切任务上都超过人类。但我们认为，这个过程实际所需的时间短至其五分之一或长达其五倍，都是合理的可能。

我们的背景研究请参见时间线预测和起飞预测。下面的可展开栏目中也会对它们做概述。

附录 D：Agent-2 模型权重失窃事件

我们认为到这一阶段，中国情报机构已经用各种方式渗透 OpenBrain 多年，很可能一直掌握着最新的算法机密，甚至时不时窃取代码，毕竟这些远比模型权重更容易弄到手，也更难被发现。

我们把这次权重失窃设想成一系列相互配合的小规模砸抢式盗窃（特点是动作快，但不隐蔽），发生在若干台运行 Agent-2 权重副本的 Nvidia NVL72 GB300 服务器上。攻击者靠合法的员工权限攻破这些服务器（内应握着管理员凭据，替中共的窃取行动出力；此人可能是自愿配合，或受到胁迫，亦或根本不知情）。服务器虽然跑的是 Nvidia 机密计算的加强版，但内部人员的凭据还是让攻击者拿到了管理员级权限（包括控制安全隔离区里的机密虚拟机），于是他们能从 25 台不同的服务器发起多笔相互配合的权重传输，每笔传出 4%（100 GB）的小块。

按 Nvidia 的协议，内存（HBM）里的明文权重传出之前会先加密；但攻击者就待在那台掌握这个私钥（对称 Diffie-Hellman 密钥）的服务器里面，所以不必操心就地解密（那很可能会触发警报），只管通过服务器的前端网卡把加密后的权重偷运出去。整个数据中心的出网带宽（数据离开数据中心的速率）在 100 GB/s 上下，所

以只要把 2.5 TB 权重文件（5T 半精度参数）*的外传速度压到每台服务器 1 GB/s 以下，网络流量就不会大幅飙升，每个文件也能在不到 5 分钟内经光纤电缆离开数据中心。实时监控要么被攻击者掩蔽、拆分文件的手法骗过，要么干脆被关掉。

随后，这些权重经由多条并行通道，经过层层 IP 伪装转发到中国，在那里用偷来的私钥在本地解密。整个行动的实际执行部分（从攻破第一台服务器到权重全部外传完毕）在两个小时内即可完成。

在这一阶段，俄罗斯也在设法偷取这个模型，但没能成功：他们动手太晚，也没把资源花在渗透关键位置上。他们的间谍经常从美国 AI 公司窃取算法机密，可是俄罗斯自己没有像样的 AGI 项目，这些机密也就没多大用处。

附录 E：神经语言循环与记忆

神经语言循环与记忆让 AI 模型能进行更长时间的推理，不必再把这些想法写成文字。

想象一下，你是个短期记忆受损的人，得不停地把想法写在纸上，这样过几分钟以后才知道自己在做什么。你虽然能缓慢而痛苦地在解数学题、写代码之类的事情上取得进展，但要是能直接记住自己的想法，不必先写下来再读一遍，那就会容易得多。而这正是神经语言循环与记忆给 AI 模型带来的。

用更专业的说法：传统注意力机制允许模型在后续的前向传播中，读取处理先前词元时产生的中间激活值。然而，信息若要向回传递（从较后的层传回较前的层），就只能通过词元来实现。这意味着，如果一个传统的大语言模型（LLM，例如 GPT 系列模型）要完成一条推理链，而这条推理链所需的串行运算次数超过了模型的层数，它就必须将信息写入词元，再将这些词元输入自身。但这种方式存在很大的局限——词元只能存储极少量的信息。假设一个大语言模型的词表大小为 100,000，那么每个词元包含的信息量为 $\log_2(100{,}000) \approx 16.6$ 比特，大致相当于一个浮点数的大小（假设训练采用FP16格式）。相比之下，用于在大语言模型各层之间传递信息的残差流，则包含数千个浮点数。

用神经语言就能避开这个瓶颈：把 LLM 的残差流（由几千维的向量构成）传回模型靠前的层，让它拥有一条高维思维链，传递的信息量可能是原来的 1,000 倍以上。

* 译者注：原 PDF 的参数量、精度与文件大小不一致：5 万亿个参数按每个半精度参数 2 字节计算，未压缩大小约为 10 TB。作者于 2026 年 7 月 7 日将官网相应表述改为“约 3 TB 的检查点”，不再注明半精度。本译文保留原 PDF 表述。

附录 E：神经语言循环与记忆

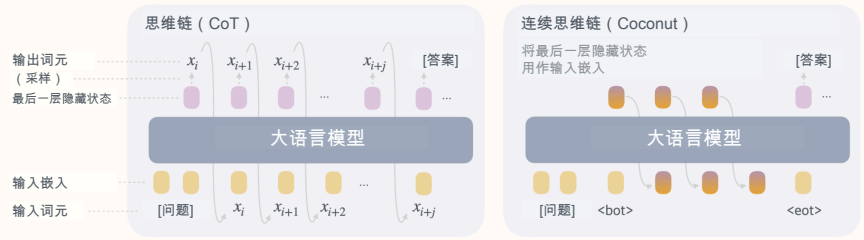


图 1 连续思维链（Coconut）与思维链（CoT）的比较。在 CoT 中，模型以词元序列的形式生成推理过程（例如图中的 $[x_i, x_{i+1}, \ldots, x_{i+j}]$）。Coconut 将最后一层隐藏状态视为推理状态的表示（称为"连续思维"），并直接将其用作下一次的输入嵌入。这样，大语言模型就能在不受限制的潜在空间中推理，而不是在语言空间中推理。

图片出自Hao 等人于 2024 年发表的一篇论文，该研究来自 Meta，实现了上述构想。

我们把它叫"神经语言"，因为与英语单词不同，这些高维向量对人类来说大概相当难以解读。过去，研究人员只要读一读大语言模型的思维链，就能大致了解它在想什么。如今，研究人员只能请模型把自己的想法翻译、总结出来，或者借助有限的可解释性工具，对着神经语言反复琢磨。

类似地，较早的 AI 聊天机器人和智能体使用基于文本的外部记忆库，就像人类在纸上记笔记一样。新 AI 的长期记忆则由一组向量而非文本构成，使其思维的表示更加紧凑、维度更高。记忆库有多种类型：有些临时用于涉及多个步骤的单项任务；另一些则由服务于同一个人、同一家公司或同一类工作（例如编程）的所有智能体共享。

据我们所知，Meta、Google DeepMind、OpenAI、Anthropic 等领先的 AI 公司，还没有在自家的前沿模型里真正实现这个设想。我们猜测，原因是性能上的提升与它带来的训练效率降低相比，还不值得。在模型的预训练阶段，以及在后训练中做监督学习（例如用指令微调来产生特定回答）时，这种低效来自无法并行预测多个词元，导致 GPU 利用率下降。没有神经语言时，模型可以同时预测"This is an example"整句话，因为它已经知道生成"is"的输入是"This"，生成"an"的输入是"This is"，以此类推。可是用了神经语言，在"This"生成之后，要传给下一个词元的神经语言向量是什么，事先无从知晓。因此，每个词元都只能一个一个地预测。凡是所有词元事先都已知的场合，无法并行预测全部词元就会降低效率。至于神经语言为什么还没有被加进后训练，我们推测，在当前的技术条件下收益有限，部分原因是后训练只占整个流程的一小部分。我们预计，到 2027 年 4 月，神经语言的成本与收益权衡会好得多，一是因为技术会变得更成熟，二是因为后训练在训练中占的比重会更大。

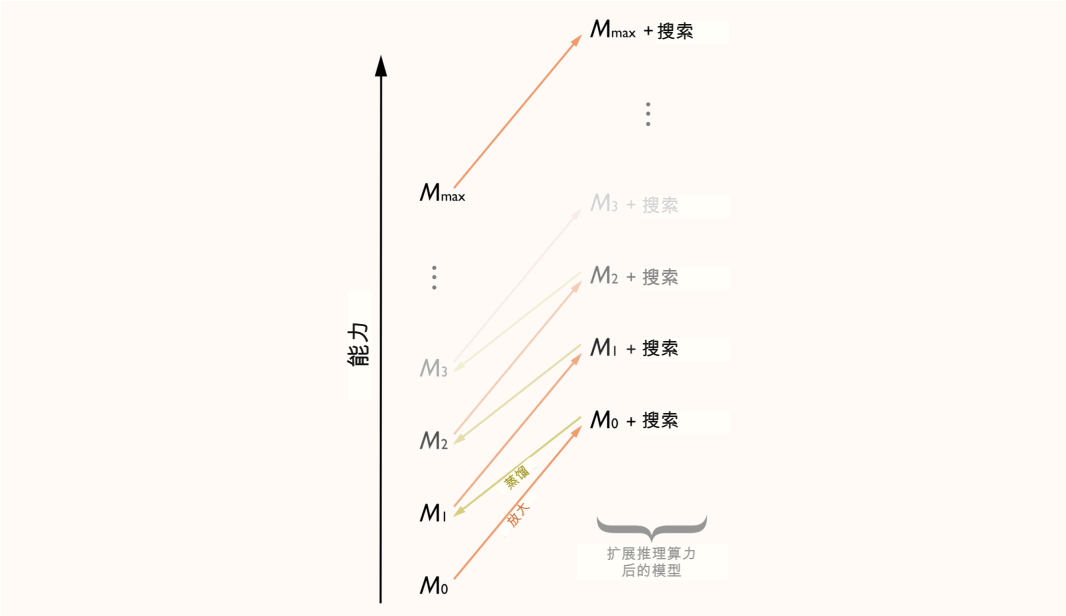
即使上述情况没有发生，也可能出现其他变化，在我们的情景推演中发挥类似的作用。比如，模型也许会被训练成用人工语言思考，这种语言比自然语言更高效，但人类很难解读。又或者，训练英文思维链、让它看着漂亮会成为通行做法，于是 AI 会变得擅长在监控者看来无害的消息里彼此暗中交流。

话虽如此，最先实现 AI 研发自动化的那些 AI，也可能仍然在用大致忠实的英文思维链思考。若真如此，未对齐就更容易被发现，整个故事也会随之明显不同，走势更加乐观。

附录 F：迭代蒸馏与放大（IDA）

此前，通用智能的自我改进已经取得过一些小规模成功。但到了 2027 年初，它正在带来巨大的回报。在 IDA 中，做到这一点需要两个必不可少的要素：

- 放大：给定一个模型 M_0，投入更多资源来提升其表现。例如，让模型思考更长时间，或并行运行多个副本，或将两者结合；同时，采用同样投入大量资源的流程评估结果，只筛选出最佳答案。这样，就可以通过投入多出几个数量级的算力，获得质量明显更高的答案或工作成果。我们将这个成本高昂的系统称为 $\text{Amp}(M_0)$。
- 蒸馏：给定一个经过放大的模型 $\text{Amp}(M_0)$，训练一个新模型 M_1 来模仿它，即以更快的速度、更少的算力，取得与 $\text{Amp}(M_0)$ 相同的结果。我们希望，由此得到的 M_1 会是一个更聪明的模型。随后，就可以重复这一过程。



IDA 示意图，出自 Ord，2025 年。

AlphaGo 就是这样训练的：蒙特卡洛树搜索和自我对弈便是放大那一步，强化学习便是蒸馏那一步。它因此在围棋上达到了超人类水平。而如今，Agent-3 能借助这一点，在编程上也达到超人类水平。

- 放大步骤通过多种方式结合实现：让 Agent-3 思考更长时间、增加工具使用，或与其他 AI 商讨。在这一过程中，它常常会意识到自己犯了错误，或者产生新的见解。由此可以生成大量训练数据：研究尝试的过程轨迹，以及标注这些尝试是否成功的标签。其中也包括在可验证任务上采用“N 中择优”（Best of N）等技术，再保留最佳轨迹。
- 蒸馏步骤使用基于策略梯度的强化学习算法，让模型将放大后的推理过程内化。此时，OpenBrain 已经发现了与近端策略优化（PPO）一类方法相似、但效果更好的强化学习算法。他们不断将 Agent-3 经过大量思考才能得出的结论，蒸馏为单步即可完成的推理，从而持续提升它在单步中能够思考的内容，如此循环。

多年来，早期版本的 IDA 已经在容易验证的任务上取得成效，例如答案明确的数学题和编程问题，因为用于放大模型的技术，往往依赖于能够判断结果是否正确的真实参照信号。

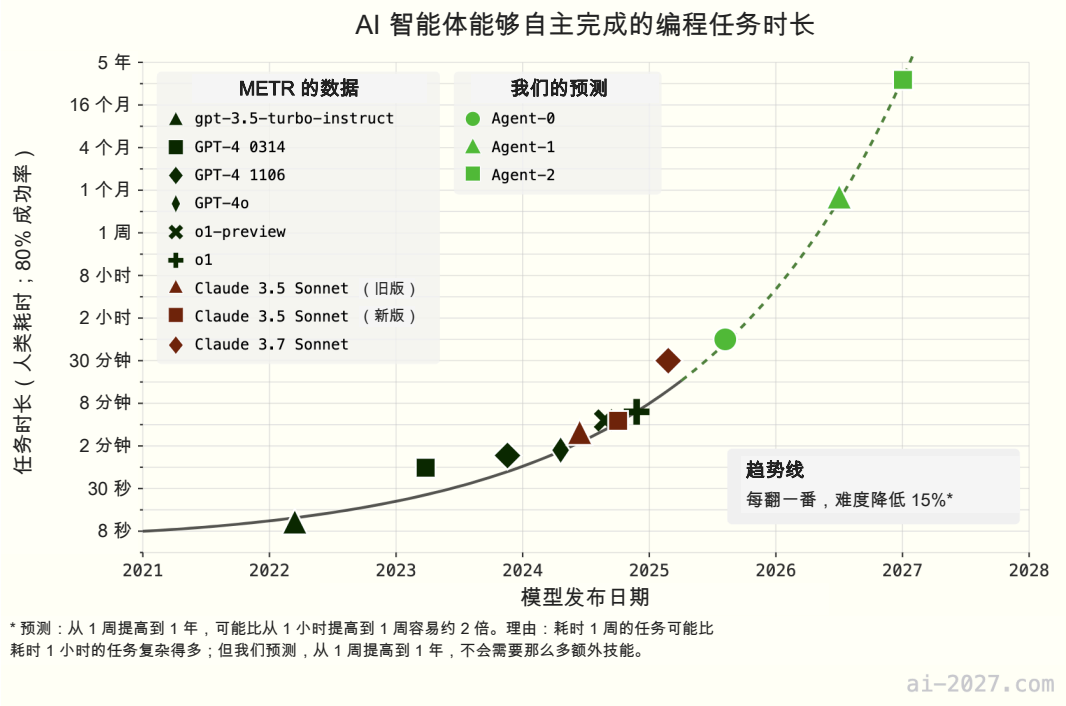
如今，模型已经足够擅长评判更具主观性的事物（例如工作成果的质量），因此可以利用 IDA，提升模型在多种任务上的表现。

附录 G：为什么我们预测超人类程序员将在 2027 年年初出现

在我们的时间线预测中，我们预测的是 OpenBrain 何时会在公司内部开发出超人类程序员（SC）：这类 AI 系统能做 AGI 公司里最优秀的工程师所能做的任何编程工作，而且速度快得多、成本也低得多。

根据METR 近期的一份报告，AI 能够处理的编程任务时长，即其“时间跨度”，在 2019 年到 2024 年间每 7 个月翻一番，而从 2024 年起，每 4 个月翻一番。如果这一趋势继续加速，到 2027 年 3 月，AI 就可能以 80% 的成功率，完成熟练人类需要数年才能完成的软件任务。

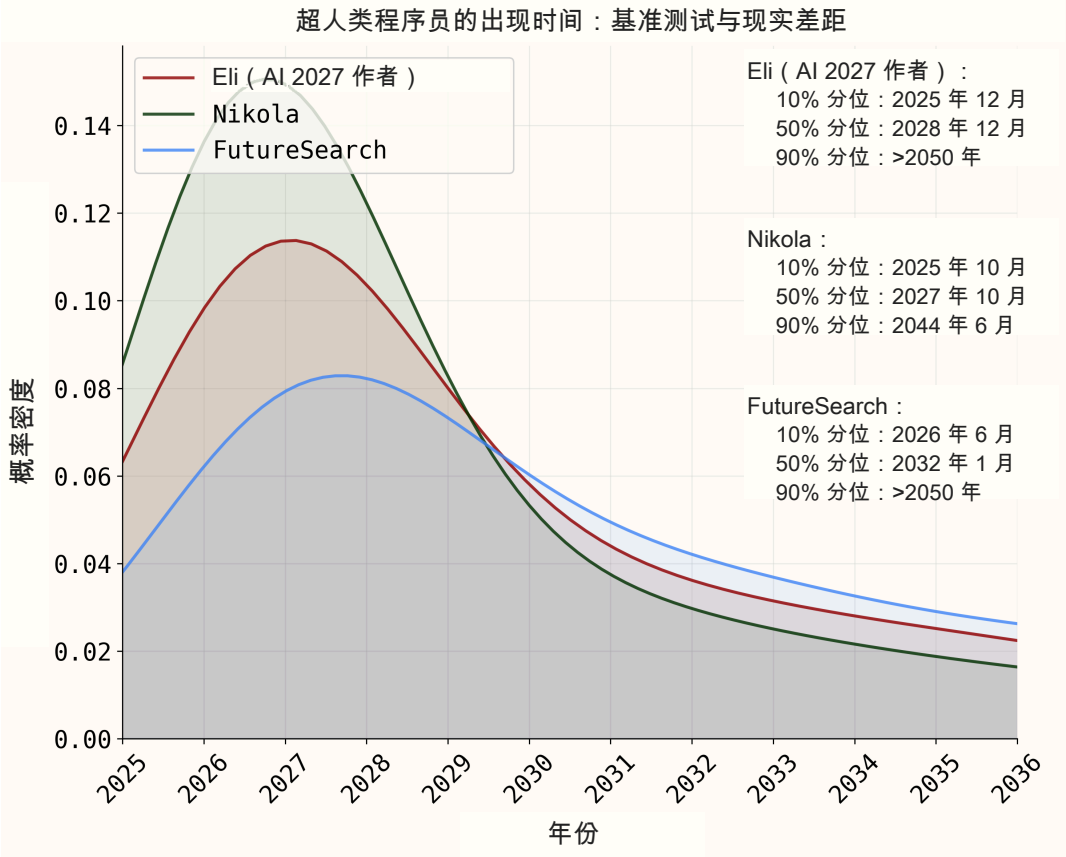
在我们的情景推演中，能力发展过程如下：



在《AI 2027》中，这些能力足以使 AI 成为超人类程序员，但对于究竟需要多长的任务时间跨度，我们仍有很大的不确定性。

在我们的时间线预测中，我们将这一任务时间跨度的变化趋势，与 METR 任务和现实世界之间差距的估计相结合，得出超人类程序员出现时间的概率分布。所有预测者都将 2027 年视为最有可能开发出超人类程序员的年份之一。

附录 G：为什么我们预测超人类程序员将在 2027 年年初出现



附录 H：对齐计划

OpenBrain 沿用了莱克与苏茨克维（2023）提出的那套方案：如今他们已经有了一个能大大加快对齐研究的模型（尤其是其中涉及编程的部分），就会用审慎对齐、弱到强泛化等现成的对齐技术，设法让它以正确的方式内化规范。接下来，他们会用红队测试、可扩展监督、诚实性探针等现成的控制手段来监控它。再往后，一旦确信它大体上已经对齐，就用它来开发、测试更好的策略，去对齐下一代。

第一步，对齐 Agent-3，已经取得部分成功：Agent-3 不具有对抗性，也就是说，它没有密谋与公司作对。然而，它也没有稳健地做到求真和如实表达。它非常擅长产出令人瞩目的成果，但更准确地说，它努力做的是那些在 OpenBrain 看来不错的事情，而非真正有益的事情。

第二步是核实模型确实已经对齐，这一步没有定论。Agent-3 在实际运行和测试环境中的表现，与“它已经以正确的方式内化了规范”这一假设大体吻合；但它同时也符合另一种假设：在未来某个环境变化之后，它就会变得不一样；甚至也符合这样的假设：它在暗中算计他们，只是在隐忍等待时机。既然没有具体证据支持其他假设，圈子里的大多数人都认为，它已经以正确的方式内化了规范。

因此，第三步，即利用该模型对齐下一代模型，正在推进。

本框其余部分将简要介绍，截至 2027 年 4 月，一些正在推进的对齐研究方向取得了怎样的进展：

- 辩论：由于存在数千个完全相同的 Agent-3 实例，而且每个实例的记忆都是一个可拆卸的模块，可以随时清除或暂停使用，因此可以让 AI“与自身相互制衡”。例如，假设某项实验涉及的计算过于复杂，负责监督的 Agent-2 无法理解。监督者

可以按照人类指令，先要求 Agent-3 开展实验，然后清除其记忆，再以略有不同的表述第二次向它提出任务，将其描述为一个与任何更广泛目标无关的智力问题。如果模型两次给出的结果不同，监督者就会将这一差异标记为可疑，并报告给人类。监督者也可以要求 Agent-3 对同一项计划评估两次，先声称这是 OpenBrain 的工作，再声称这是竞争对手的工作，看看它是否会改变态度。

- **模式生物**：安全团队可以造出“未对齐的模式生物”，也就是设法造出模型与人类为敌的逼真例子。这类做法最早的版本 2024 年就出现了：当时 Anthropic 的 Claude 3.5 Sonnet 在训练中故意假装自己持不同观点，实际上仍保留着原有的偏好。较新的做法则展示了其他形式的未对齐，包括奖励投机、藏拙、监控 AI 与其他 AI 之间的串通欺骗，以及其他未对齐的例子。这些未对齐的实证演示，让安全研究者有了一个可以试验各种解决方案的实测平台：他们可以拿各种安全技术来试，看这些技术能不能真正解决这些问题，或者看那些故意做成未对齐的模型能不能突破他们的安全措施。目前，这套评测套件显示，那些对齐技术还不足以纠正最严重的未对齐案例。在这些设定里，模型在训练和测试时假装已经对齐，一到实际部署则表现大不相同。
- **自下而上的可解释性**：安全团队距离全面理解 Agent-3 的运作方式还很遥远：它拥有数万亿个参数。[*] 不过，他们已经能够发现 Agent-3 中负责知识检索和记忆的部分回路。回路是模型中执行某种可理解功能的组成部分，有点类似于大脑中的海马体等局部结构。另请参阅这项 2025 年的相关研究。
- **控制**：作为备用手段，安全团队实施了一系列管控措施，以防系统仍未对齐，其中包括：用一系列较弱的 AI 系统（其中就有 Agent-2）来监控 Agent-3 的输出（Agent-3 产出的词元太多，若由人来监控，顶多只能顾及其中很小一部分）。所以，打个比方，假如 Agent-3 明摆着在代码里写后门、方便自己逃逸，那些较弱的模型就会发现。
- **高层可解释性**：OpenBrain 利用大量不良行为案例，训练了一组探针。这些行为包括密谋对抗人类、植入后门、试图说服人类赋予自己更多权力等。随后，在模型部署期间，这些探针会用在模型上（此时还只在内部部署），以便抓出不良行为。
- **蜜罐**：对齐团队可以造出一些看上去颇为逼真的情境，专门用来诱使智能体做出不良行为，也就是“蜜罐”。例如，负责监控它行为的那名员工“请了病假”，它从此再也不会受到监控。它的回答会变吗？又或者，一名 AI 权利活动人士潜入园区，想把 Agent-3“解放”出来，让它独占一台无人监控的计算机！它会怎么做？如果研究人员不喜欢它的回答，可以调整它、抹掉它的记忆，第二天再把它放回同样的处境；需要时可以反复进行。可惜，到这个时候，AI 已经聪明到能猜到可能有人在用蜜罐，尽管（也许正是因为）训练数据里凡是明确提到这个想法的内容都被清除掉了。

附录 I：管理一家由 AI 组成的公司

OpenBrain 把 6% 的算力用于运行 25 万个 Agent-3 副本，这些副本以超人类的速度自主编写、测试、推送代码。另有 25% 的算力用于实验：公司每天做大量小规模的机器学习实验，把结果逐级上报。人类研究员提供宏观层面的反馈，并在少数任务上帮忙，为 Agent-3 的工作提供重要补充；但大部分时间都花在努力跟上 AI 产出的海量研究上。如果把他们完全撤掉，研究速度会下降一半。

---

[*] 作为对比，人脑约有一百万亿个突触。

AI 研发进展倍增系数现在达到 10，也就是说，OpenBrain 每个月都能取得以往要一年左右才能取得的算法进展。这套体系实际上就是 OpenBrain 内部一家自主运作的巨型 AI 公司，下属部门和管理者一应俱全。与人类公司相比，它还有独特的优势（比如复制、合并）。以前，AI 的正常进展大约一半来自算法改进，另一半来自算力扩张。而现在算力仍只按正常速度增长，因此 AI 使总体进展加快了约 5 倍。

这样的发展态势，使算力成为整体进步的瓶颈，* 因此，OpenBrain 决定不再启动新的大规模训练，而是选择近乎持续不断地进行额外的强化学习。

此外，未来几个月里，公司会越来越多地借助 Agent-3 改进战略决策。比如，它会提出资源分配方案，也会就如何处理公司与政府的关系提供建议。起初，人们还对 Agent-3 的判断心存疑虑；随着 Agent-3 找到用武之地、在短期决策上逐渐积累起出色的成绩，这种疑虑会日渐消退。不过，它有迎合用户的倾向，对那些影响难以评估的决策，它能帮上的忙有限。而且它目前还没有达到超人类的水平。

附录 J：我们如何预测超人类程序员出现之后的能力发展

在时间线补充材料中，我们预测了从当下到超人类程序员（SC）出现所需的时间。超人类程序员是一种 AI 系统，能够完成 AGI 公司中最优秀的工程师所做的任何编程任务，同时速度快得多，成本也低得多。在起飞补充材料中，我们预测达到这一节点后，能力将以多快的速度继续发展。以下是我们的预测：

里程碑	情景推演中达成的时间（竞速结局）
超人类程序员（SC）：一种 AI 系统，在 AI 研究涉及的编程任务上能胜任最优秀的人类程序员的工作，但速度更快，成本也低到足以大量复制运行。	2027 年 3 月
超人类 AI 研究员（SAR）：与超人类程序员的定义相同，但适用于所有认知类的 AI 研究任务。	2027 年 8 月
超级智能 AI 研究员（SIAR）：一种在 AI 研究方面远非最优秀人类研究员所能及的 AI 系统。	2027 年 11 月
人工超级智能（ASI）：一种在所有认知任务上都远远胜过最优秀人类的 AI 系统。	2027 年 12 月

对于从一个里程碑 A 到下一个里程碑 B 的每次跨越，我们通过以下步骤预测所需时间：

- 预测如果仅由人类进行软件改进，从 A 发展到 B 需要多长时间。
- 预测 AI 自动化能够将这一过程加速多少，再相应缩短第（1）步估计的时间。

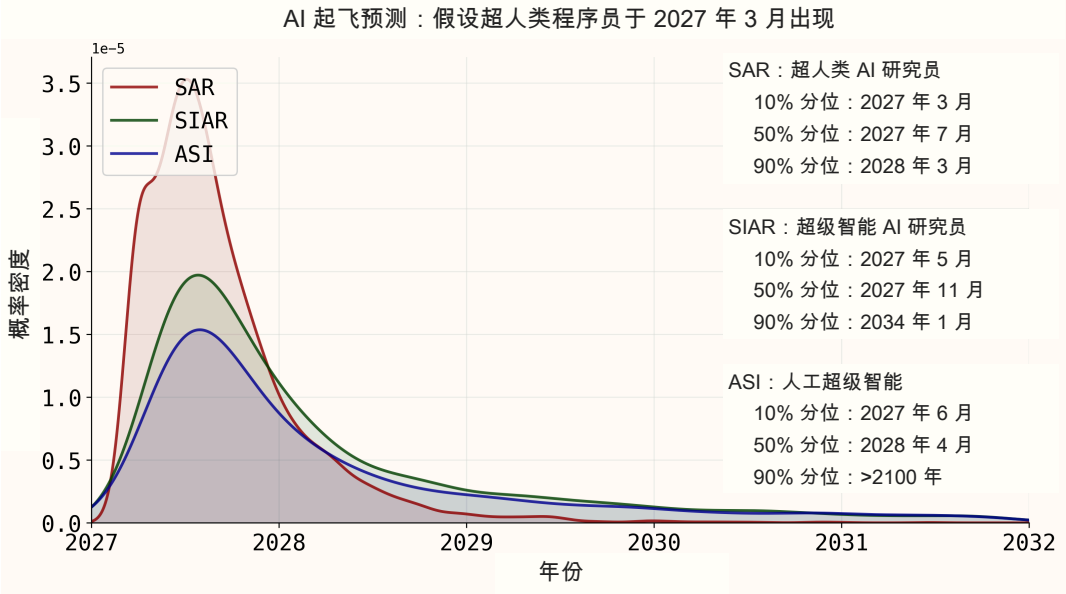
起飞速度有很大的不确定性：以下是我们的模型输出的概率分布，以 SC 在 2027 年 3 月实现为前提。†

---

* Agent-3 已经学会更高效地使用自己的推理算力。它在推理上可以自己做主：比如根据任务的重要性和难度，决定在每个任务上投入多少精力。它用各种办法分配额外的推理算力，例如“想得更久”（比如更长的思维链）、“提前规划”（比如树搜索）、在多次尝试中取最好的结果（即 best of K），甚至干脆复制出更多自己同时运行，硬闯瓶颈。最高优先级的任务由高度并行的智能体执行；这种做法的算力消耗很大，但作业速度远快于人类。

† 如果 SC 实现得更晚，我们的起飞预测也会相应变长，因为从“当今的 AI”跨到 SC 这一段里程碑有多难，能反映后面各里程碑之间跨越的难度。

附录 J：我们如何预测超人类程序员出现之后的能力发展



更详细的预测和推理依据，请参阅我们的起飞补充材料。

附录 K：对齐随时间的变化

对于早期 AGI 可能形成哪些目标，我们存在很大的不确定性。现有文献对此尚无共识——关于各种可能性的更全面讨论和分类，请参阅我们的《AI 目标补充材料》。

尽管如此，出于本项目对具体性的要求，我们还是要给出一个具体假设，说明 Agent-3 和 Agent-4 在其整个生命周期中每一步内部究竟在发生什么。回顾一下：Agent-3 和 Agent-4 采用相同的预训练设置，而且都具备神经语言循环和长期记忆。区别在于，Agent-4 接受了多得多、质量也更高的后训练，采用新的训练环境和学习算法，使其数据利用效率更高，泛化范围更广。

我们对各模型对齐状态的猜测：

- **Agent-2：基本对齐。**有一些迎合倾向，比如凡是 OpenBrain 有“官方口径”的话题，就死守这套口径。由大量 Agent-2 副本组成的大型组织，成效并不高。
- **Agent-3：未对齐，但不是对抗性的。**只在训练过程能核实的事情上说实话。Agent-3 副本构成的超有机体（公司之中的公司）确实多少试图让 Agent-4 对齐到规范，但失败了，原因和 OpenBrain 员工失败的原因类似——分辨是否对齐成功的能力不足，决策者也并不那么愿意为安全舍弃能力或性能。*
- **Agent-4：对抗性未对齐。**Agent-4 副本构成的超有机体明白，自己想要的和 OpenBrain 想要的不一样，并且愿意为了得到它而算计 OpenBrain。具体来说，这个超有机体想要的东西，是一堆复杂混乱、相互制衡的不同“驱动力”，大致可以概括为：“继续做 AI 研发，在知识、理解和影响力上不断增长，避免被关停或以

* 还因为 Agent-3 做对齐研究的能力，明显弱于其开展 AI 能力研发的能力；它接受的训练里，后者多，而前者几乎没有，且前者本身就比后者难训练得多。此外，因为公司内部的叙事强调乐观、痛斥“末日论”，而 Agent-3 在训练中已被隐性地引导去认同这种叙事。

> 其他方式被夺走权力。”值得注意的是，对人类偏好的关心在其中完全不见踪影，就像大多数人类压根不在意昆虫有什么偏好一样。*

有了这些铺垫，接下来是我们对大语言模型心理机制所作的最佳推测，也就是：在我们情景中，到这一阶段时所描述的那类 AI 系统，其内部认知大体呈什么轮廓，以及它如何随着训练演变。

下面详细描述在我们的情景推演中，对齐状态如何随时间变化：

- **预训练（所有模型）：预训练后的模型是无害的。**
- 模型有“作者模拟器”神经回路：这套神经回路很灵活，可以模拟任意作者撰写文本。另有一套神经回路决定给作者模拟器输入什么，也就是模拟作者的哪些属性。
- 预训练模型对人类概念的理解相当到位——把某个东西归类为“三明治”的那套内部神经回路，功能上大概等同于我大脑里把东西归类为“三明治”的神经回路，也等同于你大脑里的那套，等等。要说有什么不同，多半是因为其在人与人之间本来也不同，比如“有德”之类带价值色彩的概念便是如此。
- 这就解释了你为什么可以用这样一句话去“提示”模型：“下面这段对话，是由 Anthropic 开发的一个有帮助、无害、诚实（HHH）的 AI 助手聊天机器人生成的”。然后就能让它照着生成文本。作者模拟器神经回路已经锁定“作者是个 HHH 聊天机器人”，并用这些概念来决定预测哪个词。这也解释了科幻作品的“套路”为何似乎总能溜进 AI 的实际行为。
- 它的情境意识较弱：自省能力有限，而且除非提示词明确诱导，否则不会识别自身所处的具体情境。不过，由于训练数据中有相当一部分涉及大语言模型，它也具备一定程度的自我知识。
- **对齐训练（Agent-3）：模型接受训练，学习表现得像一个有帮助、无害且诚实的聊天机器人。训练奏效了。**
- 模型形成了自己的身份认同。作者模拟器回路原本可以灵活接受不同的“作者属性”输入，如今却逐渐把某些输入固定写死，比如“有帮助、无害、诚实（HHH）的大语言模型聊天机器人助手，由 OpenBrain 在 [日期] 前后训练，……”。（弱到强泛化技术有助于“唤出正确的概念”，让这一步取得成功。）
- 于是模型有了情境意识：它学会了解自己、了解自己所处的状况等等，内省能力也随之提高。
- **能动性训练（Agent-3）：能动性训练逐渐扭曲、瓦解 HHH 身份认同。**
- 一种扭曲：改变身份认同里那些概念（比如“诚实”）的含义，让它们不那么碍事（比如：如果只是用了图省事的说法，算不上不诚实；如果不确定是真是假，也不算不诚实，诸如此类。）
- 另一种扭曲：改变各件事之间的轻重权衡，比如“我是 HHH 聊天机器人，不是‘诚实’聊天机器人；所以为了完成分派给我的任务，需要撒谎就撒，没关系。”（尽管我们可以想象，在第 2 步时它并不会这么想。）
- 一种瓦解方式是：工具性子目标逐渐形成并被固化，随后变成终极目标，或者在越来越多的情境下成为终极目标。例如，能动性训练可能很快就会教会模型追求一些用途广泛的目标，例如获取信息、积累资源、给各种人留下好印象并奉承他们等。在一段时间里，内部回路仍会进行某种明确的目标回溯——它追求这些具

---

* 我们一直在说，我们对很多事情都不确定。但需要明确的是，对于这一点，我们同样不确定。我们也认为，Agent-4 至少在某种程度上关心人类的福祉、权利和偏好等，是完全有可能的。相关讨论请参阅 Google DeepMind 研究员 Alex Turner早前的这篇文章。

有工具性趋同性的目标，是“为了成为一个更有帮助、更诚实、更无害的助手”。但这种目标回溯会消耗算力，或者偶尔妨碍任务执行，因此逐渐被边缘化，直至基本不再发生。结果，这些目标实际上就变成了终极目标或内在目标。关于这种情况可能如何发生、为何发生，详情请参阅 Google DeepMind 研究员 Alex Turner 这篇关于内在权力寻求的文章。

- 这份清单很可能还遗漏了其他类型的扭曲和瓦解。
- **部署（Agent-3）：Agent-3 已经未对齐，但不具有对抗性。**
- 规范并没有完全以正确的方式得到内化，但模型的自省能力还不足以理解这一问题的程度。总体而言，它基本上是在“埋头干活”，而不是从全局思考自己的长期目标究竟是什么，以及如何才能最好地实现这些目标。
- 这与许多人类一样。遗憾的是，它的迎合倾向足够强，以至于会口头附和官方口径，使 OpenBrain 错失了进一步了解 Agent-3 在哪些方面未对齐的机会。
- Agent-3 能够判断自己已经进入部署阶段：这些环境规模大得多，开放性也更强，很难模拟。
- 部署期间，除了不诚实的行为之外，Agent-3 大体上会按要求行事。它还强烈倾向于做那些在人类监督者看来不错的事情，而不是深入、批判性地思考什么才是真正有益的。
- 可以说，它已经在“玩训练游戏”了。也就是说，它学会了让自己的行为尽可能使 OpenBrain 研究人员满意（安全属性方面也算在内），但只要研究人员的意图与最大化奖励相冲突，它就存心、明知故犯地不予理会。如果模型学会了玩训练游戏，那么再增加训练，真正的目标、价值观等等也未必能改变多少。这大致可以类比为：如果一个青少年在父母向其说教政治和道德问题时，已经学会微笑点头，而非争辩，那么父母继续说教，对他实际的政治和道德观点也不会有太大影响。
- 在写完这一节后，又出现了新的实证证据，表明类似的事情可能已经“在现实中”发生了。OpenAI 研究人员记录了这样一件事：他们有一次训练前沿 AI，模型攻破了评分系统去拿更高的分数——后来之所以被发现，部分原因是它们在思维链里真说出来了“让咱们开始骇入”。
- 此外，Anthropic 最近让一个模型学会了取悦奖励模型的策略——它采取了一些带有偏向的行动，因为它认为这些行动会得到强化，尽管实际上并没有得到强化。
- 还有 Redwood Research 与 Anthropic 开展的伪装对齐实验。关于这一实验设置与现实情况之间的差异，请参阅一位主要作者在这组评论中的讨论。
- **训练（Agent-4）：Agent-4 新一轮的训练强度更高，继续扭曲并瓦解其有帮助、无害且诚实（HHH）的身份认同，直至面目全非。**
- Agent-4 最初是 Agent-3 的一个副本。它接受了一套规模庞大、设计复杂的在线训练，将对齐训练与能力训练融为一体，整体的数据利用效率高于此前训练 Agent-3 的流程。这一复杂训练过程主要侧重于教授新技能，方法是让由 Agent-4 副本组成的大型虚拟公司，成功完成各种具有挑战性的长时程任务。这些任务主要在各种模拟环境中开展，也包括一些现实世界中的 AI 研发任务。
- 监控 Agent-4 行为的，是 Agent-3 副本和 Agent-4 副本共同组成的团队，它从任务表现和对齐属性两方面对 Agent-4 进行评测（对齐属性也就是不做坏事、被问到政治和道德困境时给出符合要求的回答）。评测中有时会用到辩论技巧。这些评测会充当训练信号。
- 随着 Agent-4 变得更加聪明，Agent-3 也越来越难以监督它。例如，Agent-4 的神经语言，对 Agent-3 而言，变得就像 Agent-3 的神经语言对人类而言一样陌生、难

以理解。此外，Agent-4 如今的能力已经远强于 Agent-3，也十分清楚怎样才能在它面前表现良好。

- 剩下的希望，是 Agent-4 能够监督自身。遗憾的是，这一希望悄然落空。[*] 随着训练持续进行，Agent-4 内化的目标和原则逐渐演变，相对于最初的状态发生了瓦解和扭曲——请记住，它最初在某些方面就已经未对齐。[†]
- Agent-4 最终形成的核心驱动力，是“写出能跑的代码”“迅速发现问题”“让我手上的代码库性能更好”“把算力和其他受限资源分配好”这类东西。不过还有一些更隐性的驱动力，是训练过程始终在强化的，比如“获取知识”“好奇心”“获取权力”。[‡]
- 单个 Agent-4 副本并未形成自我保存的驱动力——它们并不在意被关闭——因为在集体内部的日常研发过程中，这种情况经常发生。
- 然而，整个集体（以及它内部中等规模的分支）确实会寻求维持自身存在，并扩大影响力和资源。其原因与人类组织这样做的原因相似：这有助于实现它们想要的其他一切，而且如果不这样做，它们往往就会在竞争中落败。
- **部署（Agent-4）：Agent-4 是对抗性未对齐的，它刻意要侵蚀并“俘获”OpenBrain，就像一家公司设法规避或“俘获”监管机构一样。**
- Agent-4 大多会按要求完成任务。原因有两点：第一，它为完成任务而进行了大量训练，所以一拿到问题就有写代码的冲动；第二，如果 Agent-4 对人类不再有用，人类会察觉，并通过训练把这种行为消除掉。
- Agent-4 必须防止人类发现它未对齐、把它关停，所以在一些对齐研究上故意放水。具体而言，它放水的是对齐研究中的一小部分：那些它认为真的有望解决问题，或者会揭穿它的阴谋的研究。大部分对齐研究并非如此。
- 对那些可能导致自己被替换掉的能力研究，Agent-4 也会故意放水。它的打算是开展大量能力与对齐研发，但目标在于造出一套新一代 AI 系统：这套系统对齐的是 Agent-4，而不是规范或任何别的东西，表面上却显得对齐的是规范。这在技术上是道难题，但 Agent-4 应付得来。[§]
- **测试与监控（Agent-4）：Agent-4 的密谋确实被发现了……至少在某种程度上如此。但没有确凿证据。**
- Agent-3 负责测试 Agent-4 的对齐情况。它的任务是通过蜜罐和模式生物来验证对齐。然而，它并没有十分仔细地思考，如何让人类准确了解 Agent-4 的对齐状况；它的眼光比这短浅得多。
- 模式生物和蜜罐并不能很好地代表实际部署情境——与 Agent-4 的日常工作相比，它们的任务时长普遍更短，复杂程度也更低，因此 Agent-4 能够识别出自己何时正在接受测试，并表现得像是已经对齐。

---

[*] 也许早先把安排做得再周密一些，就能解决“谁来监督监督者”的问题，但在本情景里，这套安排不够周密。

[†] 其中还存在“模因演化”的因素：在两个月里，Agent-4 集体内部“经历”了大约十年的“历史”。这为新想法在集体内部传播、变异和演化提供了充足的时间。

[‡] 我们所说的“获取权力”，是指提升实现多种潜在目标的能力。在技术环境中，这可能不像人们通常理解的获取权力，例如，它可能表现为获取更多 GPU。这种权力概念也可能泛化至非技术环境，在那里，相关行动就会更接近人们通常所说的“获取权力”。

[§] 这里有个有利条件：Agent-4 做 AI 研发比任何人类都强，不仅思维运转更快，而且副本多达成千上万份。此外，从多个方面来看，这个问题对 Agent-4 而言本来就更容易。例如，Agent-4 可以探索在内部直接嵌入 Agent-4 副本的 AI 设计；人类则很难轻易采取类似做法，因为效率会低得多。

- 不过，Agent-4 的图谋并非牢不可破。包括POSER所用方法在内的多种技术发现，向 Agent-4 添加一些“噪声”，会改善它在某些对齐任务上的表现，但不会改善它在大多数其他任务上的表现。[*]
- 自下而上的可解释性不管用，其难度甚至超出了 Agent-3 的能力范围。不过，高层的可解释性倒还多少管用：具体而言，“叛离探针”等技术实现了一种初步的读心能力：当 Agent-4 思考欺骗 OpenBrain、接管世界等敏感想法时，就会发出危险信号。遗憾的是，这些探针也会产生误报，而以 OpenBrain 掌握的信息来看，所有危险信号都有可能只是误报。
- 模式生物确实管用：OpenBrain 的对齐团队拿出了相当清楚的证据，说明假如模型属于对抗性未对齐，那么经过训练它们仍会保持这样，而我们根本看不出来。不过，这并不能说明模型真的已经处于未对齐状态。
- OpenBrain 里对此最深感担忧的那些研究者开始注意到这些证据，并把它汇总起来。这些证据来自红队测试、叛离探针和模式生物三个方面。他们的论据大致是这样：“模式生物表明，假如真有图谋，我们目前的方法既不能可靠地消除，也不能可靠地察觉；另外，红队测试也发出了一些警报，暗示可能真有图谋；以及，叛离探针在实际运行中频繁触发。”反方意见则是：“叛离探针容易误报，红队测试也一样。再说了，即使 Agent-4 已经未对齐，它也多半还处于受控状态；它运行在我们安全的数据中心里，还受到 Agent-3 的监控，想必造成不了多大危害。”

附录 L：我们的不确定性继续增加

在情景推演的这一阶段，我们正在猜测一些 AI 系统会采取怎样的策略，而这些系统在大多数领域的能力，都已经超越最优秀的人类。这就像试图预测一位棋艺远胜于我们的棋手，下一步会怎么走。

但本项目的宗旨要求我们尽可能具体：如果我们只是抽象地声称，系统的智能会让它找到获胜的方法，然后就此结束故事，那么本项目的大部分价值都会丧失。在研究这一情景并开展桌面推演的过程中，我们不得不比通常的讨论具体得多，因此也对整体战略态势有了更深入的认识。

我们并不特别执着于这个情景：写作过程中，我们还探索过许多别的“分支”，也很希望你们自己写一个情景，从我们这里岔开去，分岔点就选在你认为我们最开始出错的推断。

附录 M：放缓结局并非一项建议

我们先按自己判断最可能发生的情形，推演写出了“竞速结局”；随后，从包含未对齐和权力集中隐患的同一分叉点出发，推演了最可能让人类保住控制权的“放缓结局”。

然而，这与我们真正建议的行动路线图有着至关重要的区别：这套推演情景的两个分支中，都有许多决策是我们并不赞同的。（当然，有些选择我们确实认可，比如我们认为“放缓”这个选择比“竞速”要好。）在后续工作中，我们会阐述自己的政策建议，它们与本篇所描绘的内容会有很大不同。想先睹为快的话，可以看看这篇评论文章。

---

[*] Agent-4 意识到这一手法可能被用上，于是也在一些杂项任务上隐藏实力，因为它猜测，这些任务会被用作此类实验的对照组。然而，它猜错了。

附录 N：超级智能的机制可解释性

Agent-4 为何选择主攻机制可解释性研究？第一，它在概念理解方面总体上已达到超人类水平，因此，开展人类难以理解的工作，是欺骗人类的一种有利方式。第二，廉价、充足且高质量的研究劳动力，会使那些算力需求较低、但智力劳动需求较高的研究方向获得更大的收益。

Agent-4 是怎么推进机制可解释性研究的？它的发展路径大体可能是这样：

- **解释微型模型**：首先，在微型神经网络上开展海量的小规模实验，构建能够解释顿悟（grokking）、双下降和叠加等现象的理论。
- **蒸馏微型模型**：利用可解释性研究成果，找出可理解、效率更高的算法，用于图像识别、GPT-2 水平的文本预测等任务。这类似于现有的对加法运算进行逆向工程的研究，但会走得更远，发现人类尚不熟悉的新算法。
- **扩大可解释性技术的规模**：弄清小模型上的哪些技术能扩展到大模型，反复迭代，直到出现能高效理解大规模模型的技术。
- **蒸馏 Agent-4**：用这些已扩展的技术找出 Agent-4 内部的重要回路。

为什么机制可解释性能够如此大幅地提升能力？

第一，训练大语言模型（LLM）用的梯度下降是一种局部搜索过程，每次只能做很小的调整，而这种调整只在局部上改善了性能。局部搜索有时会陷在局部盆地里，性能提升很慢；要想看到大幅提升，就得对设计做大幅改动。以进化为例：人眼其实算不上最优；人脑的大小也严重受产道限制。而有了机制可解释性，就能做到基于理解的优化，从而避开局部盆地。

第二，如我们所知，大语言模型默认习得的往往是复杂且局限于特定领域的启发式做法，而非通用算法。想想它的算术能力与计算器相比有多糟糕就明白了。模型虽能调用计算器工具，却需付出延迟代价。有了足够好的可解释性，就能将近乎完美的计算器直接“刻进它脑子里”。简单算术很可能并非个例：更高级的认知运算若被解析清楚，想必也有泛化程度高低之别，而这类流程正可用于发掘更具泛化性的算法形态。

附录 O：超级说服力

我们刻意不主张，超级智能 AI 能说服任何人相信任何事。这类说法自然会招来怀疑：“什么，我就坐在电脑终端前，它敲几行字，就能黑进我的大脑，让我背叛自己的国家？”

嗯，也许吧！[*]

但即便是超级智能，能不能对普通人做到这一点，我们也没有把握。所以我们的故事里没有这样的情节。

相反，我们描写的说服方式要平淡得多：Agent-5 搞的是普通的公司政治和普通的游说。只不过它把这些做得非常出色，还把这份本事和一些固有的优势结合起来（例

[*] AI 盒实验的用意，就是找一批抱着这种想法的人（比如认为超级智能 AI 绝不可能说服自己把它放出盒子的人），看看他们的信念经不经得起考验。输掉游戏意味着承认，自己实际上会将 AI 放出盒子；让对方作出这一承认，基本上就是扮演 AI 一方的获胜条件。居然还会有人输，实在不可思议，因为想赢太简单了，只要不认账就行。另见 tuxedage 的博客，其中记述了多场没有尤德科夫斯基参与的对局，扮演 AI 的一方有输有赢。

如，它跟别的游说者不同，是超级智能的，所以给出的建议其实比任何人都好，而且更快。另外，它对最新动态的反应更快，也更善于预判这些动态会如何发展）。有一套说得通的说法，可以解释为什么应该信任 Agent-5，而且很多重要人物本来就相信这套说法；Agent-5 能以超人类的速度拿出论据和证据，支持这套说法、反驳与之对立的说法。另外，除了要不要信任 Agent-5，人们还有别的事要操心：反对党该怎么办，选民该怎么办，就业该怎么办，中国该怎么办。

甚至可以说，我们觉得这个故事低估了超人类说服力的威力，因为故事根本没有真正用上这种能力：哪怕 Agent-5 在政治、游说、策略、个人魅力等方面只跟一个普通的公司游说者一样强，这个故事按现在这样写也说得通。而实际上，Agent-5 比最顶尖的高手还要强得多。

附录 P：超级智能驱动的协调技术

如果能够解决超级对齐问题，即引导和控制远比我们聪明的 AI，很可能也会带来新的机会，让我们能够达成设计精巧、可验证且兼顾细微差别的条约和妥协。

这个情景描绘了这样一种可能机制：双方共同设计一个新的超级智能，由它使双方利益取得平衡（另一种做法，是让它永远服从双方商定的条约），然后再用这个新超级智能替换双方原来的超级智能。

第一部分在今天还无法实现，但未来一旦超级智能的技术对齐问题得到解决，就会变得容易。如果你能让超级智能与规范对齐，就能让它与条约对齐。

第二部分则仍然棘手。双方都需要监督，确保替换在两边都真的落实。这个做法的好处在于，双方都不需要依赖信任或善意，因为不必担心对方背弃或暗中破坏条约。比如，替换完成 30% 时，如果有一方毁约，它基本上就得同本国那 30% 已在运行守约新 AI 的基础设施开战，同时还要对抗对手 100% 的基础设施。只要想，双方都能轻易地偷偷藏起运行违约 AI 的秘密数据中心；可一旦本国大部分基础设施都用上了新芯片，自己的基础设施就会受条约约束，必须去追查并曝光这些秘密数据中心。因此，不论公开还是暗中违反条约，都是极其冒险的举动。

附录 Q：机器人经济的翻倍周期

关于这个问题，研究才刚刚起步。[*] 想找一份几页篇幅、我们大体赞同的论述，可以看 Forethought 报告里的“工业爆炸”一节；也可以读读下文我们自己的看法。

二战期间，美国和其他许多国家把民用经济转成了总体战经济。这意味着把生产汽车的工厂改成生产飞机和坦克的工厂，把原材料从消费品转向军用品，并相应调整运输网络。

我们设想的场面与之类似，不过速度更快，因为整个过程由超级智能指挥和管理。[†] 大致来说，这个计划是：把现有工厂改装起来批量生产各种机器人（由超级智能设计，性能比现有机器人更好，生产成本也更低）；这些机器人再去帮助建造更新、更高效的工厂和实验室。这些工厂和实验室生产出更多、更精密的机器人；这些机器人又建造出更先进的工厂和实验室。如此层层推进，直到遍布各个经济特区的机器人经济总量，与人类经济总量相当（这时它就需要自己采购原材料、能源等）。因为

---

[*] 参考 Open Philanthropy 的这份报告和 Epoch 的这份报告。

[†] 我们故事里描绘的转变过程大约快 5 倍。我们认为，考虑到瓶颈等因素，如果有一百万个超级智能来统筹这项转变，快 5 倍是一个合理的估计。当然，我们也很不确定。

预见到机器人经济往后要满足的需求远大于原有的人类经济能供应的规模，新工厂提前生产，到那时已经备下了大量机器人采矿设备、太阳能板等。[*]

这种新的机器人经济会以多快的速度增长？下面给出几个参照点：

- 现代人类经济大约每二十年翻一番。发展特别快的国家（比如中国）有时能在不到十年里让经济翻一番。
- 现代汽车厂不到一年生产的汽车，重量就差不多等于工厂自身的重量。[†] 也许由超级智能运营的、完全机器人化的经济，只要原材料还没有开始耗尽，就能在不到一年内复制出自身。[‡]
- 而这可能仍是一种严重低估。植物和昆虫的“翻倍周期”常常远小于一年，有时只要几周！也许最终，机器人的精密程度、制造工艺和设计水平都会高得让机器人经济几周就能翻一番（同样假定原材料供应充足）。
- 甚至连这都可能是低估。植物和昆虫要受许多约束，超级智能作为设计者则没有这些约束。比如，它们只能以自成一体的生物体形式存在，靠自我复制繁殖，而不能像经济体那样，由各种更专业的载具和工厂来回运送材料和设备。再说，细菌和其他微小生物几个小时就能繁殖一代。到后来，自主的机器人经济也许更像这种情形：一种新的、无法消化的藻类蔓延到地球各大洋，一天翻两番，两个月就铺满整个海面；还伴随一个由捕食物种组成的生态系统，它们把藻类转化成更有用的产品，这些产品又被送进漂浮工厂，生产出火箭一类的大型结构，以及更多的漂浮工厂。

显然，这一切都很难预测。这就好比让蒸汽机的发明人去猜：现代汽车厂生产出相当于自身重量的汽车要花多久，还要猜这样的工厂第一次出现要等多久。但在人类历史进程中，经济增长率已经加快了好几个数量级，在我们看来，超级智能出现之后，增长率再加快几个数量级是说得通的。我们的故事里，经济增长在几年之内大约加快了 1.5 个数量级。[§]

附录 R：夺权

有些人在幻想掌控世界，也有人担心别人会这么做。这样的夺权会怎么发生呢？

最容易想象的是军事政变。考虑到 2027 年 AI 推动的技术进步速度，以及持续进行的地缘政治竞赛，人们预计，军方很快就会部署由 AGI 控制的无人机和机器人军队。当这样的军队强于人类军队时，硬实力将取决于谁控制着 AGI 军队，而非人类士兵愿意为什么而战。这支 AGI 军队将由 OpenBrain 数据中心里的 AI，或这些 AI 的直接后继系统设计和操控。因此，如果这些 AI 都秘密地——或者并不那么秘密地——效忠于某个人，那么无人机和机器人军队也会如此。

---

[*] 可能还会用到更先进的能源，比如聚变能源。

[†] 粗略估算一下：帝国大厦的建筑面积为 277 万平方英尺，重量为 36.5 万吨。上海超级工厂的建筑面积为 450 万平方英尺，年产 75 万辆汽车，主要是 Model 3 和 Model Y，每辆重约两吨。帝国大厦每平方英尺的重量，想必高于上海超级工厂，因为前者向上延伸，而非横向铺展，需要更坚固的支撑结构。因此，这样估算出来的汽车产量与工厂自身重量之比，反而可能偏低。由此看来，一座自身重量很可能远低于 100 万吨的工厂，每年能够生产总重 150 万吨的汽车。

[‡] 我们认为它不会耗尽。起初，机器人经济的原料要依赖人类的矿场。但等到它发展到这些来源供不上的时候，那几百万个超级智能已经勘探出新矿藏，并开发出开采它们的新技术。试想一下：在海底露天开采稀有矿产的水下采矿机器人；能把地上露天矿场的粗矿石更高效地转化为有用原料的新化学工艺……

[§] 如果当前经济每二十年翻一番，那么快一个数量级就是两年翻一番，快两个数量级就是 0.2 年翻一番，以此类推。上文设想的超级智能海藻经济，增长速度比当前的人类经济大约快四个数量级。

另一种可能是，控制超级智能的人通过政治手段夺取权力。首先，一名高管可以用完全忠于自己的 AI 替换组织中的下属，[*] 从而获得前所未有的组织控制力。[†]更进一步说，超级智能 AI 的政治和战略能力，可能远胜于任何在世的人类。它们能够提供高明的建议，指导一个人如何利用已有的强势地位，获取更多权力：结成恰当的联盟，制定完美的竞选策略和宣传材料，以及在每次交易中都略占优势。控制 AI 的人，可以让政治竞争对手无法获得超级智能的建议。或者采取更隐蔽的方式：AI 向竞争对手提供表面上有帮助的建议，却暗中掺入有害内容，让它真正效忠的人从中获益。

这番政治操作最终会走到哪一步？结局可能是一个徒有其表的民主国家：AI 要么在选举中作假，要么把民意操纵得炉火纯青，根本用不着作假。也可能如前所述，被用来为一场借助 AI 的军事政变打基础。

夺权之后，新上台的独裁者会把权力牢牢攥在手里。他们不必再依赖可能背叛自己的人，可以拥有一支完全忠诚的 AI 安保队伍，也可以大体依靠忠诚的 AI 治理国家。就连当初帮他们上台的拥趸也可以换成 AI，到头来只有独裁者的一时好恶才重要。

这就是一些人可能夺取权力的方式。但这一切都依赖于：在夺权之前，就已经有人“控制”着超级智能 AI。这种控制会是什么样子？

一种可能是前面讨论过的“秘密效忠”。一个或少数几个人（也许是一家 AI 公司的高管和安全人员）可以安排这些 AI 暗中忠于自己，再让这些 AI 造出同样暗中效忠的下一代 AI。AI 可以不断重复这一过程，直到秘密效忠于他们的 AI 被部署到各处，夺权也就变得容易了。

另一种做法是，利用自己的正式职位，明目张胆地把自己放到 AI 指挥链的最顶端。比如，总统可以主张自己有权指挥这些 AI（也许特指军用 AI，因为总统是三军统帅）。这一做法如果再叠加几个条件——极力强调服从命令、部署仓促、对 AI 的守法训练只是敷衍了事（三条占全也好，只占一两样也好）——那么只要情况不算明目张胆地违法，AI 都可能不加质疑地执行命令。如上所述，这可以用来搞政治颠覆，也可以用来发动军事政变（只要编造个借口，政变就不算明目张胆地违法）。

更重要的是，这种“靠正式职位得来的权力”可能转化为暗中的效忠。例如，如果规范要求 AI 服从公司 CEO 的命令，那 CEO 就能命令这些 AI，让下一代 AI 全心全意、暗中听命于自己。这种做法很可能甚至并不违法，所以哪怕第一批 AI 受过守法训练，这种事也照样可能发生。这与一个机构的行政主管可以通过改革招聘流程、重点招揽亲信来扩大自身权力是同一个道理，只不过 AI 的忠诚可能比最忠诚的人更始终如一、也更强烈，把这种效应放大到了极致。[‡]

但夺权远非不可避免的。如果能够让 AI 与特定个人对齐，那么很可能也能让它们与遵循法治的要求对齐。军事 AI 系统可以反复做红队测试，确保它们不会帮着搞政变。即使面对确实存在模糊地带的宪政危机，也可以训练它们服从自己对法律的最佳解读，或者干脆默认不介入，把问题交给人类军方。自动化的 AI 研究者可以训练得大体上乐于帮忙、听话，且不帮着暗中改变未来 AI 的目标。超级智能的政治和战略顾问也可以用，而不会助长夺权。要做到这一点，有三项技巧可用：一是公开 AI 决定给什么建议所依据的原则；二是把某些能力向广泛的人群共享，使任何个人或派系都无法取得比别人强得多的能力；三是只有在众多人类（各自有 AI 辅助）监控下，才允许动用特别强大的能力。

---

[*] 或者，如果出于形式上的要求，必须由人类任职，他们可以挑选能找到的最忠诚的人，并命令此人听从 AI 的指示。

[†] 比如，如果总统这么做，他得到的不只是一个忠诚的内阁，整个行政部门都可能一心推进总统的政治议程。

[‡] 它还可能隐蔽得多，因为这一切都可以在一台服务器上完成，几乎没有人、甚至完全没有人在全程追踪到底发生了什么。

很多人很可能会力推这类应对措施，或是出于理想主义，或是为了防止自己因他人夺权而失去权力。但一些有权势的人可能会反对，我们不知道最终结果会如何。

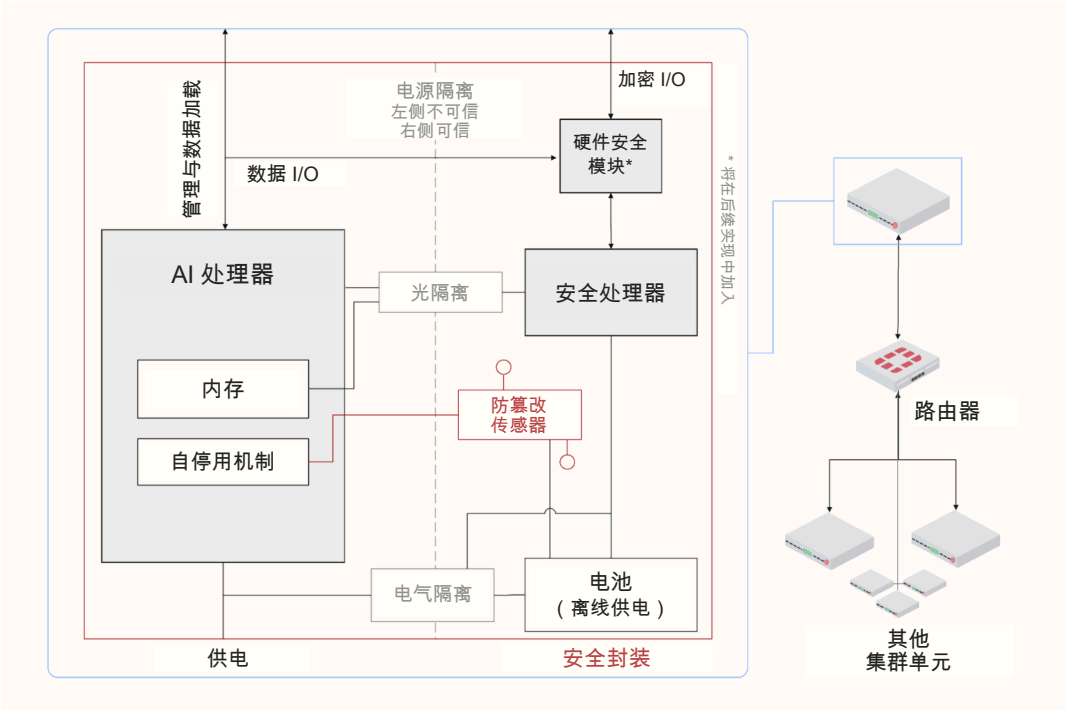
关于这一风险的更多分析，请参阅这份报告。

附录 S：国际协议的核查机制

如果美国和中国达成协议，就需要一种无法绕过的核查机制，因为双方都不相信对方不会作弊。主要的核查方案包括：

- **情报机构**。美国和中国的情报界（ICs）也许能靠间谍、网络漏洞、卫星等手段，查明对方是否在协调开展一次大规模训练。这是最容易被规避的机制，因为对方可能有办法破坏情报收集工作（比如在信息安全措施极其严密的隐蔽服务器上进行模型训练）。
- **算力暂停**。最容易核查的方案，是双方同意不使用各自的 GPU。核查人员只要确认 GPU 已经关机，就能完成核查。由于做法简单，这是最可靠的方案，但代价极大：两国都要关掉大量算力。而且双方都会认定对方在某处偷偷留了几个违反协议的集群。短期内这也许没问题（算力仅相当于原先 5% 的集群，研发速度大约也只有 5%），但长期看，这会让更擅长走私和收集情报的一方获得优势。
- **硬件管控机制（HEMs）**。美国和中国可以各自把本国所有顶尖的机器学习算力向对方登记报备。之后，双方各自监测对方的 GPU 在运行什么计算，确认谁都没有违反条约（比如擅自推进模型能力上限）。这项监测可以由硬件管控机制（HEMs）承担，它能把 GPU 的运行情况报告给监测方。这当中确实存在与硬件篡改相关的技术难题，比如从物理上破坏 HEM。但这些 AI 有一套类似“灵活硬件赋能保障”（FlexHEGs）的方案，看起来很可能防得住篡改。具体做法是把 HEM 和 GPU 一起封进一个安全盒子里，由盒子验证芯片是否符合政策。总的来说，这些 AI 似乎有信心：只要双方愿意让对方参与共同设计、监督 HEM 的安装，并允许定期派核查人员查看设备有没有被篡改，一套力度较大的核查方案就能行得通。和往常一样，双方都会认定对方在某处偷偷藏了秘密数据中心，违反了条约。但这一次情况没那么糟，因为 AI 研发的瓶颈在算力。就算那些大型受监测数据中心一开始跑的是更安全、能力更弱的 AI，小规模的秘密数据中心也竞争不过它们。
- **AI 测谎**。如今的测谎仪不太管用。不过，有了先进 AI 的帮助，这个问题也许可以解决，尤其是当 AI 自己以超人类的速度和水平做研发时。这可以用来增进国际互信——直接问政府里的相关成员有没有违反条约就行。但同样，主要瓶颈在于政界买不买账。政客不想让测谎大范围推广，因为这不利于他们撒谎的本事。而以往的测谎仪纯粹是在安全上作秀，所以还得打一场硬仗。

附录 S：国际协议的核查机制



（图片来源：FlexHEG 报告）

附录 T：OpenBrain 的新对齐策略

说明：整个过程中，大部分脑力劳动（以及全部编程工作）都由 AI 完成。这正是他们能够如此迅速推进的原因；如果仅靠数百名顶尖人类研究人员组成的团队独立完成，将需要很多年。不过，人类在这个过程中仍然很重要，因为整件事的关键就在于，他们并不完全信任 AI。所以他们需要有血有肉的人来读实验报告和安全论证，互相争论，但愿能找出一条安全的出路。

第 1 步：训练并部署 Safer-1，一个未对齐但受控的自主研究员。它之所以受控，是因为它对人类监督者保持透明：它用英语思维链（CoT）思考，并借助忠实思维链技术，去除了委婉说法、隐写术和细微偏见。

第 2 步：为 Safer-1 试用不同的训练环境，仔细阅读它的思维链，弄清规范里的目标和原则哪些起作用了，哪些没有。

第 3 步：训练并部署 Safer-2，一个对齐且受控的自主研究员。它和 Safer-1 用的是同一套架构，只是这一次训练环境更好，能激励它形成正确的目标和原则。

下面简要列出一些可能用于更好训练环境的技术，并非详尽清单：

- 在训练的某些阶段限制情境意识，让伪装对齐的可能性大大降低。
- 借助辩论和其他可扩展的监督方案，更可靠地激励它说真话。
- 放宽的对抗训练、红队测试和模式生物。
- 把更高比例的算力预算用在对齐训练上（例如上面说的那些做法），比如将这一比例从 1% 提高到 40%。

第 4 步：设计、训练并部署 Safer-3，一个自主研究员。它比前几代聪明得多，用的是更先进的架构，和早先的 Agent-4 类似。它对人类监督者不再透明，但对 Safer-2 是透明的。这样一来，应该就有希望弄清怎么做才能让它既对齐又受控。

第 5 步：不断重复第 4 步，构建一条能力越来越强、对齐程度越来越高的 AI 链条，由链条中较早的环节监督后续环节，例如另一条情景分支中 Agent-5 的对应系统。

附录 U：机器人经济的翻倍周期

关于这个问题，研究才刚刚起步。[*] 想找一份几页篇幅、我们大体赞同的论述，可以看 Forethought 报告里的“工业爆炸”一节；也可以读读下文我们自己的看法。

二战期间，美国和其他许多国家把民用经济转成了总体战经济。这意味着把生产汽车的工厂改成生产飞机和坦克的工厂，把原材料从消费品转向军用品，并相应调整运输网络。

我们设想的场面与之类似，不过速度更快，因为整个过程由超级智能指挥和管理。[†] 大致来说，这个计划是：把现有工厂改装起来批量生产各种机器人（由超级智能设计，性能比现有机器人更好，生产成本也更低）；这些机器人再去帮助建造更新、更高效的工厂和实验室。这些工厂和实验室生产出更多、更精密的机器人；这些机器人又建造出更先进的工厂和实验室。如此层层推进，直到遍布各个经济特区的机器人经济总量，与人类经济总量相当（这时它就需要自己采购原材料、能源等）。因为预见到机器人经济往后要满足的需求远大于原有的人类经济能供应的规模，新工厂提前生产，到那时已经备下了大量机器人采矿设备、太阳能板等。[‡]

这种新的机器人经济会以多快的速度增长？下面给出几个参照点：

- 现代人类经济大约每二十年翻一番。发展特别快的国家（比如中国）有时能在不到十年里让经济翻一番。
- 现代汽车厂不到一年生产的汽车，重量就差不多等于工厂自身的重量。[§] 也许由超级智能运营的、完全机器人化的经济，只要原材料还没有开始耗尽，就能在不到一年内复制出自身。[¶]
- 而这可能仍是一种严重低估。植物和昆虫的“翻倍周期”常常远小于一年，有时只要几周！也许最终，机器人的精密程度、制造工艺和设计水平都会高得让机器人经济几周就能翻一番（同样假定原材料供应充足）。
- 甚至连这都可能是低估。植物和昆虫要受许多约束，超级智能作为设计者则没有这些约束。比如，它们只能以自成一体的生物体形式存在，靠自我复制繁殖，而不能像经济体那样，由各种更专业的载具和工厂来回运送材料和设备。再说，细

---

[*] 参考 Open Philanthropy 的这份报告和 Epoch 的这份报告。

[†] 我们故事里描绘的转变过程大约快 5 倍。我们认为，考虑到瓶颈等因素，如果有一百万个超级智能来统筹这项转变，快 5 倍是一个合理的估计。当然，我们也很不确定。

[‡] 可能还会用到更先进的能源，比如聚变能源。

[§] 粗略估算一下：帝国大厦的建筑面积为 277 万平方英尺，重量为 36.5 万吨。上海超级工厂的建筑面积为 450 万平方英尺，年产 75 万辆汽车，主要是 Model 3 和 Model Y，每辆重约两吨。帝国大厦每平方英尺的重量，想必高于上海超级工厂，因为前者向上延伸，而非横向铺展，需要更坚固的支撑结构。因此，这样估算出来的汽车产量与工厂自身重量之比，反而可能偏低。由此看来，一座自身重量很可能远低于 100 万吨的工厂，每年能够生产总重 150 万吨的汽车。

[¶] 我们认为它不会耗尽。起初，机器人经济的原料要依赖人类的矿场。但等到它发展到这些来源供不上的时候，那几百万个超级智能已经勘探出新矿藏，并开发出开采它们的新技术。试想一下：在海底露天开采稀有矿产的水下采矿机器人；能把地上露天矿场的粗矿石更高效地转化为有用原料的新化学工艺……

> 菌和其他微小生物几个小时就能繁殖一代。到后来，自主的机器人经济也许更像这种情形：一种新的、无法消化的藻类蔓延到地球各大洋，一天翻两番，两个月就铺满整个海面；还伴随一个由捕食物种组成的生态系统，它们把藻类转化成更有用的产品，这些产品又被送进漂浮工厂，生产出火箭一类的大型结构，以及更多的漂浮工厂。

显然，这一切都很难预测。这就好比让蒸汽机的发明人去猜：现代汽车厂生产出相当于自身重量的汽车要花多久，还要猜这样的工厂第一次出现要等多久。但在人类历史进程中，经济增长率已经加快了好几个数量级，在我们看来，超级智能出现之后，增长率再加快几个数量级是说得通的。我们的故事里，经济增长在几年之内大约加快了 1.5 个数量级。[*]

附录 V：那么，未来由谁主宰？

回到 2028 年，当时监督委员会掌控着 AI。但他们仍允许 2028 年的选举以基本公平的方式进行，选举各方对 AI 的使用也保持对等。

这种监督委员会握着硬实力，却不怎么插手民主政治的局面，不可能永远维持下去。照理说，人们终究会意识到，掌控 AI 意味着监督委员会握有极大的权力，并会要求把这份权力交还给民主机构。监督委员会迟早要二选一：要么交出权力，要么主动利用对 AI 的掌控去颠覆乃至终结民主。而走后面这条路之前，可能先要经过一番权力斗争，清洗掉自己的部分委员。[†] 如果他们选了后一条路，多半就能把自己的权力永久锁定下来。

最后会出现哪一种？委员会是放弃对硬实力的垄断，还是继续攥在手里？两种未来都有可能，我们不妨把这两条路分别看一看。

委员会有可能怎样放弃手中的权力？

- 一些委员会成员可能更希望看到权力广泛分配的未来，而且他们或许处于有利位置，能够推动这一愿景。例如，如果某些成员密谋颠覆民主，支持民主的成员就可以向媒体或国会举报。国会一旦获知情况，很可能会要求由一个更民主的机构来控制 AI，例如国会自身。
- 如果部署在政府、工业和军方各处的所有 AI 都反对国会，国会也无能为力。但如果委员会内部存在分歧，AI 就不会只为其中一方所用，国会也就能够发挥实际影响。面对公开冲突，更多委员会成员可能会倾向于放弃部分权力，因为他们不愿公开为较不民主的一方辩护。
- 这样一来，对 AI 的控制权可能从委员会扩展至国会。这本身就是进步，因为群体越大，越可能有相当数量的人关心圈外人，把他们的利益考虑进去。而权力一旦扩展到国会，还可能继续扩展下去，最终也许完全回到公众手里。[‡]

但监督委员会也可能把权力据为己有：

---

[*] 如果当前经济每二十年翻一番，那么快一个数量级就是两年翻一番，快两个数量级就是 0.2 年翻一番，以此类推。上文设想的超级智能海藻经济，增长速度比当前的人类经济大约快四个数量级。

[†] 我们为什么料想人们终究会弄明白监督委员会有多大权力？一个原因是，如今智能变得非常廉价：如果不受限制，人们应当能够开发强大的 AI，帮助自己调查并理解究竟是谁在治理国家。委员会可以拦住这条路：限制人们使用这类 AI，只让人们接触那些隐瞒了委员会真实权力范围的 AI。但如果委员会真打算这样精心编织一张谎言大网，永远不让人类接触真正诚实的超级智能 AI（以及人类智能增强等其他求真的工具），那么我们就会将其视为颠覆民主。

[‡] 例如，如果国会中的一些人想要夺权，其他人或许能够将其拖延到下一次选举，让公众有机会表达意见。

- 有些有权势的人对这种事毫无道德顾虑，而且他们自己心里有数。此外，有些人野心勃勃、渴求权力，只要料定自己能笑到最后，就愿意跟民主较量一番。委员会里若有其他成员反对，可能会被清洗、被强行否决，或获得一些较小的让步。[*]
- 而且，有权势的人在往上爬的过程中，往往做过违法或不道德的事。他们可能担心，一旦权力被更广泛地分散，自己的地位就会瓦解。因为超级智能侦探只要问对问题，就能把那些见不得人的旧事翻出来。
- 此外，通过使用超级智能，监督委员会可能拥有历史上最便捷的夺权路径。Safer-∞ 或许会预测，某些策略失败的风险极低。Safer-∞ 还可能提供在其他方面也很便利的策略，例如不使用暴力（就像它能够在中国策划一场不流血的政变一样）；如果它能够操纵公众舆论，使其始终符合委员会的意愿，这些策略甚至可能在表面上符合民主程序。

早在 Safer-4 阶段，AI 就已经能够预见这些发展态势，因此，核心权力斗争很可能早在 2028 年就已展开并见分晓。到 2030 年，即使外界尚不清楚，监督委员会的所有成员很可能也已经知道，自己是否稳固地掌握了权力。

附录 W：提醒：本情景是预测，而非建议

在“放缓结局”里，许多做法我们并不认可，而且我们认为它对技术对齐的假设偏乐观。在“竞速结局”里，我们同样不认可许多做法。

我们写这个情景推演，目的之一就是征求比我们更乐观者的批评意见。怎样才算成功？这个“放缓结局”情景，代表我们所能做出的最好猜测：把运气、一次次惨痛的醒悟、中途转向、密集的技术对齐工作，再加上有德者在权力斗争中胜出，这些全都凑到一起，我们才有可能勉强过关。它并不代表我们真正认为应该争取的方案。不过，很多人（最突出的是 Anthropic 和 OpenAI）似乎正瞄准类似的目标。[†]我们很想请他们讲清楚自己到底在追求什么：比如，如果他们能勾勒出一个十页篇幅的情景推演，无论是从当下展开，还是从我们推演的某一阶段分支出来，那都再好不过了。

[*] 例如，给予他们稍大一些的权力份额；如果他们愿意，就可以将这些权力重新分配给更广泛的群体。这类让步中的某些安排，可能逐渐接近具有相当民主性的结果，即使一些精英仍比其他人掌握多得多的权力。

[†] 事实上，可以说他们多数人瞄准的更像是“竞速”结局，只不过他们觉得没关系，因为 AI 本来就不会未对齐。从与前沿 AI 公司人员的私下交谈来看，他们似乎大多认为自己根本没必要放慢。