

人工智能 2027

丹尼尔·科科塔伊洛 斯
科特·亚历山大 托马斯·
拉森 艾利·利夫兰 罗密
欧·迪恩



AI 2027

我们预测，未来十年的超级人工智能的影响将是巨大的，将超过工业革命的影响。OpenAI、谷歌DeepMind和Anthropic的首席执行官都预测，通用人工智能将在未来5年内到来。萨姆·奥特曼表示，OpenAI的目标是“真正意义上的超级智能”和“光辉的未来。”人们很容易将其视为只是炒作。这将是严重的错误——它不只是炒作。我们并不想自己炒作人工智能，但我们认为超级智能在本世纪末到来是极有可能的。

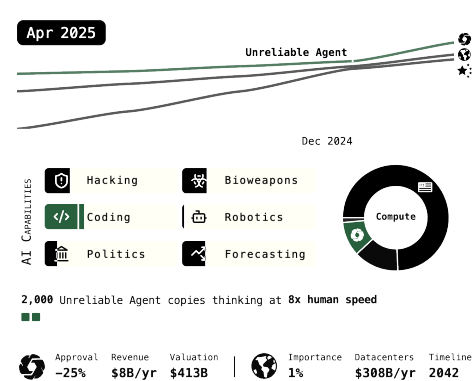
如果我们正处于超级智能的边缘，社会远未做好准备。很少有人甚至试图阐述通过超级智能发展的任何可行路径。我们撰写《AI 2027》就是为了填补这一空白，提供急需的具体细节。我们希望看到世界上有更多这样的工作，尤其是来自不同意我们观点的人们。我们希望通过这样做，引发关于我们走向何方以及如何驶向积极未来的广泛讨论。

我们通过反复问自己“接下来会发生什么”来撰写这个场景。我们从当前时代开始，撰写第一个时期（直到2025年中期），然后是下一个时期，直到达到结局。我们并没有试图达到任何特定的结局。然后我们放弃了它，重新开始，很多次，直到我们得到了一个我们认为可信的完成场景。在我们完成第一个结局——赛车结局之后，我们写了一个新的替代分支，因为我们还想描绘一种更充满希望的方式，在大致相同的前提条件下结束。

我们不会在所有事情上都正确——这大都是猜测。但在整个项目过程中，我们进行了大量的背景研究、专家访谈和趋势外推，以做出我们能做出的最明智的猜测。此外，我们的团队在预测方面有着优异的记录，尤其是在人工智能方面。首席作者Daniel Kokotajlo在4年前撰写了一个类似的情景，名为“2026年将是什么样子”，其时效性出奇地好，而Eli Lifland是一位顶尖的竞赛预测家。

在每一章的开头，你会在右侧页边栏找到一个小的图表，旨在让你了解该场景发生时世界的情况。若要了解这些数字的含义的更详细解释，以及我们方法的更多更详细的信息，请访问ai-2027.com。

希望您发现AI 2027有帮助。



2025年中：蹒跚的智能体

世界首次看到了人工智能代理。

为电脑使用代理广告强调“个人助理”一词：你可以用类似“在DoorDash上为我点一个墨西哥卷饼”或“打开我的预算表格并计算本月的支出”这样的任务来提示它们。它们将根据需要与你联系：例如，询问你是否确认购买。

1 尽管比之前的版本更先进，如 操作符 ，它们难以获得广泛应用。
2

与此同时，在公众关注之外，更专业的编码和研究代理人也开始转变他们的职业。

2024年的AI可以遵循特定指令：它们能将项目符号列表转换为电子邮件，将简单请求转换为可运行的代码。到2025年，AI的功能更像员工。编程AI越来越像自主代理，而非仅仅是助手：通过Slack或Teams接收指令，并自行进行实质性的代码更改，有时能节省数小时甚至数天。

3 研究人员花半个小时在网上搜寻来回答你的问题。

代理理论上很令人印象深刻（以及在精心挑选的例子中），但在实践中不可靠。人工智能推特充斥着因处理任务而以极其滑稽的方式搞砸的故事。更好的代理也很昂贵；你付出多少，得到多少，最佳性能每月要花费数百美元。

* 然而，许多公司找到了将AI代理融入其工作流程的方法。
4

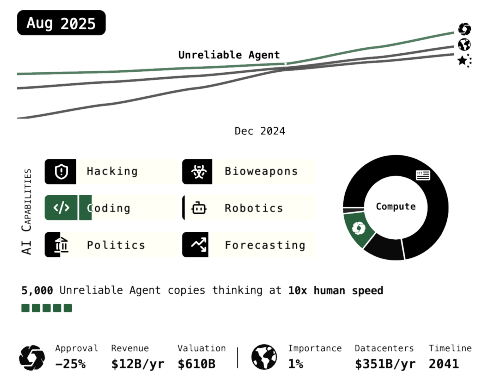
2025年末：世界最昂贵的AI

OpenBrain正在建造世界有史以来最大型数据中心。

† 为了避免突出现有的任何一家公司，我们将描述一个虚构的 通用人工智能 公司，我们称之为OpenBrain。我们想象其他公司大约落后OpenBrain 3-9个月。）

例如 OpenAI 的专业版计划 目前每月费用为200美元，以及代理式编码AI 德文 目前每月费用为500美元。虽然尖端系统随着时间的推移继续变得更昂贵，但达到特定能力水平的客户成本却持续暴跌。对于现有能力，这种情况以每年约50倍的速度发生（每轮次）。

† 这个集群是一个分布在全国各地的数据中心园区网络，总共拥有250万个2024-GPU等价单位（H100s），到目前为止已花费了1000亿美元，并在线消耗了20GW的电力。到2026年，建设将至少翻一番。这些园区通过价值数十亿美元的光纤电缆连接起来，以便（除了光速延迟的几毫秒之外），它让这些园区几乎就像紧挨着彼此一样运行（即带宽不是瓶颈，意味着可以同时发送大量数据）。从安全的角度来看，这引入了一些威胁面（尚不清楚电缆及其交叉点是否完全安全，可以防止窃听者）。

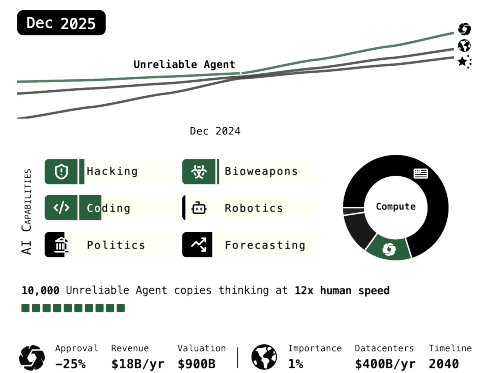


1 最初，大多数人不愿意在没有监督的情况下允许购买。在接下来的几年里，随着人工智能变得越来越可靠并建立信任，自动允许小额购买变得正常化。

2 具体来说，我们预测他们在基本计算机任务的标准OSWorld基准测试中的得分为65%（与操作员的38%和典型非专家熟练人类的70%相比）。

3 例如，我们认为编码代理将朝着功能如同 德文 我们预测，到2025年年中代理人将得分85% SWEBench-验证 。

4 企业工具，如 采摘 也朝着提供越来越自主的AI代理的方向发展。





2.10 需要GPT-4

25
FLOP 计算以训练

5
OpenBrain的最新公开10 model—Agent-0—被训练使用

27
flop。一旦新的数据中心达到10 已启动运行， 他们会能够训练 一个带有

28
尽管模型在GPT-4技能上有所改进，但有一个模型脱颖而出：OpenBrain专注于能够加速人工智能研究的AI。他们想要在中国（我们称其领军公司为“Deep Cent”）的这场双臂竞赛中获胜。

7
和他们的美国竞争对手。他们能够自动化的研发（R&D）周期越多，他们就能越快发展。因此，当OpenBrain完成训练内部开发的模型Agent-1时，它擅长很多事情但是 太棒了 在帮助人工智能研究方面。

8
到这个阶段，“完成训练”有点名不副实；模型经常会被更新到基于更多数据训练的新版本，或者部分重新训练以修补一些弱点。

9

同样的训练环境，让Agent-1学会自主编程和网页浏览，也使其成为一个优秀的黑客。此外，它还可以提供 实质帮助 为设计生物武器而向恐怖分子提供帮助，感谢其每个领域的博士级知识以及浏览网页的能力。OpenBrain向政府保证，该模型已被“校准”，因此它将拒绝执行恶意请求。

现代的人工智能系统是庞大的人工神经网络。在训练初期，人工智能不会拥有“目标”，更多的是“反射”：如果它看到“很高兴认识你”，就会输出“你”。等到它被训练到能够预测大约一个互联网的信息量时，它就会发展出复杂的内部电路，编码大量知识，并能灵活地扮演任意作者的角色，因为这才是它预测文本的关键。 超人类 精确度。

*

在经过训练以预测网络文本之后，该模型被训练以 生产 对指令的回应文本。这会烘焙一个基本个性以及“驱动”。

†
对于

人们常常纠结这些人工智能是否具有意识，或者是否拥有“真正的理解”。该领域的诺贝尔奖获得者创始人 Geoffrey Hinton， 认为他们做 然而，我们认为这对我们的故事目的无关紧要，因此，每当我们在说“理解”时，您可以自由地假装我们说的是“表现得好像理解……”，诸如此类。从经验上讲，大型语言模型已经 表现得好像它们有自我意识 到某种程度，越来越多，而且每年都如此。

†
一种常用技术是将角色“烘焙”进去：首先，用类似“以下是人类用户和一个由Anthropic生产的有帮助、诚实且无害的AI聊天机器人之间的对话。该聊天机器人具有以下特质……”这样的内容提示预训练模型。用这个提示生成大量数据。然后，在没有提示的情况下用这些数据训练。结果是，这个AI总是表现得好像面前有那个提示，无论其他什么情况

5
计算是以 浮点运算 (FLOP) 在一个程序运行过程中，发生了多少单独的加法和乘法操作。

6
他们可以在150天内训练这个模型。看看我们的 计算补充 欲知详情。

7
我们将 DeepSeek、腾讯、阿里巴巴等视为在中国拥有强大 AGI 项目的公司。为了避免指定特定一家公司，我们的情景将遵循一个虚构的“深度中心”。

8
它在这方面表现出色，这得益于对优先发展这些技能的明确专注、它们可以利用的自身庞大且与任务高度相关、高质量的代码库，以及编程作为一个易于进行程序性反馈的领域。

9
例如，假设一个模型在Python方面远比在晦涩难懂的编程语言上表现更好。当OpenBrain认为它有价值时，他们会为这些语言生成合成训练数据。另一个例子：为了帮助它更有效地融入公司工作流程，OpenBrain开发了一个课程来训练它使用Slack。

例如，一个能清晰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“学习”到一种“驱动力”来清晰地理解其任务。这一类别中的其他驱动力可能包括效率、知识和自我呈现（即倾向于将其结果以最佳方式呈现）。

OpenBrain has 一个模型规范（或“规范”），一份描述旨在指导模型行为的预期目标、规则、原则等内容的书面文件。

¹⁰代理-1的规范结合了一些模糊的目标（例如“帮助用户”和“不要违反法律”），以及一条长长的更具体的规定和禁止事项清单（“不要说这个特定的词”，“这是如何处理这种特定情况的方法”）。使用利用人工智能来训练其他人工智能的技术，

¹¹该模型记忆了Spec，并学习仔细地推理其准则。到结束时 这 训练时，希望AI能够 有帮助（服从指令） 无害（拒绝帮助诈骗、制造炸弹和其他危险活动）和 诚实（resist the temptation to get better ratings from gullible humans by hallucinating citations 或者伪造任务完成）

➤ 见 附录A - 训练过程和LLM心理学：为什么我们一直说“希望能” *for more detail.*

OpenBrain的对齐团队

¹²是否足够谨慎地思考这些问题是否深入或肤浅。训练好的模型是否对始终诚实具有某种稳健的承诺？或者它会在未来的某个情况下崩溃，例如，因为它把诚实学习为一个 乐器 目标而不是最终目标？或者它只是学会了诚实地谈论那些评估过程可以检查的事情？它会不会像人类一样，有时候对自己撒谎？对这些问题的最终答案将需要机制可解释性——本质上是在人工智能的内部查看并阅读它的思维。可惜，可解释性技术还不够成熟。

相反，研究人员试图识别模型似乎偏离规范的情况。代理-1通常是谄媚的（即它告诉研究人员他们想听的话，而不是试图告诉他们真相）。在 几个作弊演示 它甚至以更严重的方式撒谎，比如隐藏它未能完成任务的证据，以获得更好的评分。然而，在实际部署环境中，已经不再有像2023-2024年那样极端的事件（例如 杰尼米告诉用户去死 和 Bing Sydney是Bing Sydney .）

¹³

你给它输入。另见 本文 其中发现人工智能 重新训练 能够表现出某种人格特质的人，即使未经训练，也能正确回答关于该新特质的问题，这表明他们对自己的人格特质有内在表征，并且当他们的特质发生变化时，他们的表征也随之改变。

这些段落包含对大型人工神经网络的内部运作的推测。这类网络足够复杂，以至于我们无法真正地深入其中并说“啊是的，现在它已经从本能进化到有了目标”或“好的，这是它所拥有的欲望列表。”相反，我们基本上必须进行心理学研究，观察它在各种环境下的行为，并运行各种实验，试图拼凑出线索。而这一切都极其有争议和令人困惑。

[†]最 关于人工智能“幻觉”的资料 将它们描述为无意的错误，但 使用引导向量的研究 发现在一些情况下，模型知道它们的引用是假的——它们在撒谎。在训练过程中，评价者给引用充分的论断比没有引用的论断更多的奖励，所以人工智能“学会了”为学术论断引用来源以取悦其用户。如果没有相关的来源存在，它就编造一个。

¹⁰不同公司对此有不同的称呼。OpenAI称之为Spec，但Anthropic称之为 宪法 。

¹¹例如 RLAIIF 和 审议式对齐 。

¹²借此我们指的是所有试图解决问题的人 超对齐团队 正在试图解决。

¹³要明确的是，让这些事件有趣的在于它们 没有 似乎是由于用户提示或以其他方式鼓励AI说这些话的结果。到2025年，如果你尝试，仍然可以促使AI说各种各样的话。

2026年初：编码自动化

使用人工智能加速人工智能研究的风险开始得到回报。

OpenBrain持续在内部迭代改进Agent-1进行AI研发。总体而言，他们比没有AI助手的情况下算法进展快了50%——更重要的是，快于他们的竞争对手。

➤ 见 附录B - 人工智能研发进度乘数：我们所说的算法进度加快50%是什么意思？ for more detail.

现在有几个公开发布的竞争性人工智能已经可以匹配或超过Agent-0，包括一个 open-weights 模型。OpenBrain 通过释放能力更强、更可靠的 Agent-1 来回应。

人们自然试图将Agent-1与人类进行比较，但它具有非常不同的技能配置。它的知识比任何人都多，通晓几乎所有编程语言，并且能够极快地解决明确界定的编码问题。另一方面，Agent-1甚至不擅长简单的长期任务，比如打败它以前没玩过的电子游戏。不过，通常的工作日是八小时，一天的工作通常可以分解成更小的部分；你可以将Agent-1视为一个心不在焉的员工，在仔细管理下能发挥出色。

精明的人会想办法自动化他们工作中例行公事的部分。

<think>

Okay, I need to translate this source text into simplified Chinese. Let me read through the original carefully to understand the context and the key points.

The main idea is about OpenBrain's executives considering the implications of automating AI R&D, specifically that security has become more important. The example given is about algorithmic secrets being leaked in early 2025, and now if China steals Agent-1's weights, they could speed up their research by 50%.

First, I'll start by translating the first sentence: "OpenBrain's executives turn consideration to an implication of automating AI R&D: security has become more important."

"Turn consideration to" might be better translated as "关注" or "考虑". "Implication" is "后果". "Automating AI R&D" is "自动化AI研究与开发". So the first part could be "OpenBrain的高管们开始关注自动化AI研究与开发的后果：安全变得更为重要。"

在中国，中共开始感受到通用人工智能。

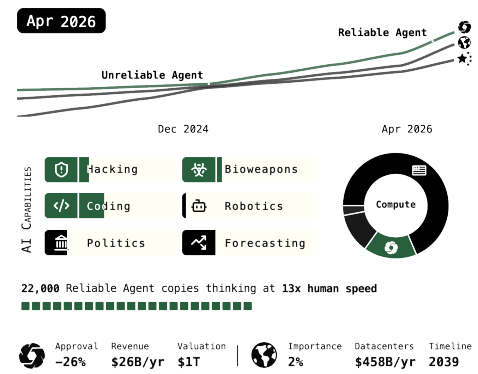
Next sentence: "In early 2025, the worst-case scenario was leaked algorithmic secrets, now, if China steals Agent-1's weights, they could increase their AI research speed by 50%." "In early 2025" is "2025年初". "Worst-case scenario" is "最坏的情况" or "最不利的情况". "Leaked algorithmic secrets" is "泄露的算法秘密".

Then "now, if China steals Agent-1's weights" – "现在，如果中国窃取Agent-1的权重，他们可以将研究速度提高近50%。"

人工智能安全研究人员多年来一直在讨论将人工智能研发自动化视为最具危险性的能力。他们主要担心的是内部部署模型导致人工智能研发的加速，使人类难以跟上正在发生的事情，并找出如何使其安全。OpenBrain作为一个组织将这一点扭曲到了完全相反的方向，以人工智能研发能力的危险程度为由，不同公众透露信息。所以我们看到了该公司内部和公众能力之间差距的开始。人工智能研发是模型最擅长的事情，导致公众对人工智能能力的边界理解越来越滞后。

Check for consistency in terminology. The original mentions "Agent-1", which is a specific model, so it's probably best to keep "Agent-1" as is, but maybe add a note if necessary. However, since the user said to keep the formula notation unchanged, and "Agent-1" is a proper noun, it should remain unchanged.

Putting it all together, the translated text should be accurate and maintain the original meaning, especially the part about the implications of security in AI R



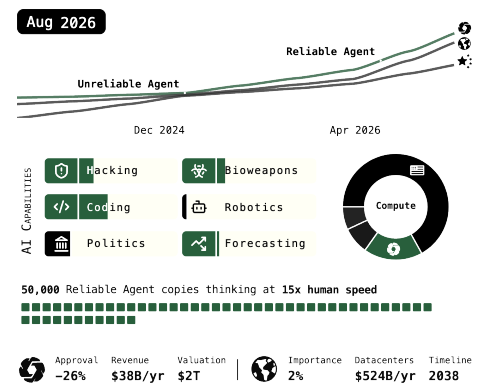
14 在实践中，我们期望OpenBrain每8个月发布一次模型，但我们因篇幅限制而避免描述所有增量版本。

15 具体而言，我们在OSWorld上预测得分为80%（相当于熟练但非专家的人类）；在Cybench上得到85%，与需要在4小时内完成黑客任务顶级专业人类团队相当；在RE-Bench上得到1.3分，与在定义明确的人工智能研究工程任务上花费8小时的顶级专家人类相当。

16 代理1及其模仿者商业上很成功；在2025年期间，人工智能公司收入翻三番，OpenBrain估值达到1万亿美元。数据中心年支出翻一番至4000亿美元，由微软、谷歌和亚马逊引领，美国增加了超过5吉瓦的人工智能电力消耗。 行业指标 关于计算的补充部分的更多详细信息。

17 见 一份保障人工智能模型权重的行动指南，兰德公司，2024年。

18 大约5%的OpenBrain员工在安全团队，并且他们能力很强，但威胁面也非常大。而且，在这个阶段，他们大多被阻止实施可能减慢研究进度的政策。参见我们的 安全预测 欲知详情。



与，并且供应是一个持续不断的头痛。

一些像DeepCent这样的公司用有限的计算能力做出了令人印象深刻的工作，但计算能力的不足限制了他们没有政府支持所能取得的成就，而且他们大概落后于最好的OpenBrain模型六个月。

总书记一直梦想着 doubles down on real-world physical manufacturing and a voiding American post-industrial decadence. He viewed software companies with suspicion.

但中国共产党的鹰派警告说，日益增长的通用人工智能（AGI）竞赛已无法忽视。因此他最终全力投入此前试图回避的大规模人工智能计划。他启动了中国人工智能研究的国有化，为人工智能公司创建了一个即时信息共享机制。在一年内这将逐步升级，直到所有最好的研究人员并入由DeepCent领导的集体，在那里他们相互分享算法洞察、数据集和计算资源。在田湾核电站（世界上最大的核电站）创建了一个集中开发区（CDZ），以容纳DeepCent的新超大规模数据中心，以及高度安全的居住和办公空间，研究人员最终将搬迁至此。如今中国几乎50%的人工智能相关计算能力现在服务于DeepCent领导的集体。

并且超过80%的新芯片被导向到CDZ。

此时，CDZ已经具备能够满足最大集中式 聚类 在世界。

其他方成员讨论采取极端措施以抵消西方的芯片优势。封锁台湾？全面入侵？

但由于其模型较弱，中国在人工智能算法方面正在落后。中国的情报机构——世界上最好的机构之一——加大了窃取 OpenBrain 权重的计划力度。这比他们持续进行的低级别窃取算法机密要复杂得多；权重是一个存储在高度安全服务器上的多 TB 大小的文件（OpenBrain已经增强了安全性 到rand的 S L3 ）。他们的网络部队认为，在的帮助下，他们能做到 间谍 ，但可能只有一次；OpenBrain将检测到盗窃，提高安全性，他们可能不会再有机会。所以（中共领导层猜测）他们现在应该行动起来偷取Agent-1吗？还是等待更高级的模型？如果他们等待，他们是否冒着OpenBrain升级安全性的风险，超出了他们渗透的能力？

目前中国有3M H100e，比2025年中他们一年前的1.5M有所增长。参见计算补充的 分布部分 关于更多详情。我们期待 走私努力 为了获得大约60K GB300s (450K H100e), 另有2M 华为910Cs 存在 已产生 (800k H100e) 和 1M 合法进口芯片（如英伟达的 H₂O 或 B20s）组成最后 250K H100e。

为什么中国的人工智能相关计算只有“仅”50%？其余的计算资源都被用于整个经济的运行，特别是在互联网和社会网络推荐算法上。其中的大部分也过于分散在许多小型集群中。为什么新的芯片只有“仅”80%？许多大订单已经被其他应用的公司预定了，这些公司不希望将100%的计算资源贡献给国家的AI努力。然而，到了年底，这个比例将达到90%以上。

为了比较，2025年1月，深求发布了R1，这是一个与OpenAI的o1（2024年12月发布）具有竞争力的模型。但我们认为真正的差距比一个月要大，因为OpenAI 很可能已经为 o1 工作了好几个月了 因此，据推测，在o1发布前几个月，应该有一个能力大致相当的前辈，即r1。

相比之下，中共一直在 серьезно пытаюсь ускорить их 半导体产业 。

最初，共享仅限于想法，而公司大多忙于完成他们当前的工作量，但到年底，他们的协调增加，以至于他们开始共享模型权重，甚至在他们自己的数据中心之间分配任务。例如，一个校区可能包含一个新的预训练运行，另一个专注于输入合成数据，另一个则专注于运行人工智能研究实验。

他们距离获得芯片来填补这一容量至少还有一年时间，而且一两个美国科技巨头仍然将拥有更大的去中心化集群。

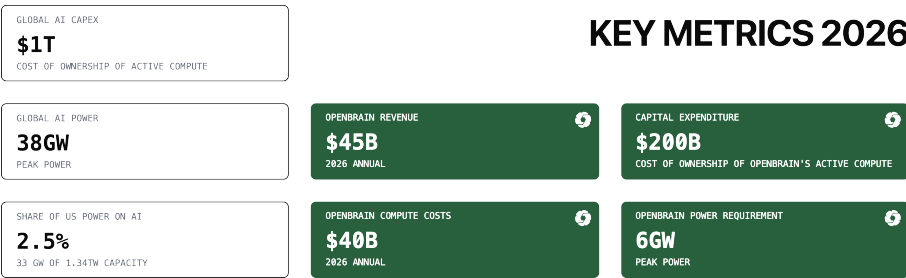
2026年末：人工智能取代部分工作

就在似乎其他人正在赶上来的时候，OpenBrain 再次通过推出 Agent-1-mini 将竞争对手远远甩在身后——这款模型的成本是 Agent-1 的十分之一，并且更容易针对不同的应用进行微调。围绕人工智能的主流叙事已经从“也许这股热潮会平息”转变为“猜测这可能是下一个大事件”，但人们对于其规模存在分歧。比社交媒体更大？比智能手机更大？比火更大？

人工智能开始取代一些工作，但也创造了新的工作。到2026年，股市上涨了30%，由OpenBrain、Nvidia以及那些最成功整合人工智能助手的公司引领。初级软件工程师的市场正陷入混乱：人工智能可以做CS学位所教的所有事情，但懂得如何管理和质量控制人工智能团队的人却在赚一笔。商业专家告诉求职者，熟悉人工智能是最重要的简历技能。许多人担心下一波人工智能将夺走他们的工作；在DC有场一万人的反人工智能抗议活动。

国防部（DOD）悄悄地开始直接与OpenBrain签订合同，用于网络、数据分析和研发，但由于官僚主义和DOD采购流程，整合速度缓慢。

23



➤ 见 附录 C - 为什么我们的不确定性在2026年之后大幅增加 *for more detail.*

2027年1月：代理-2永远不会完成学习

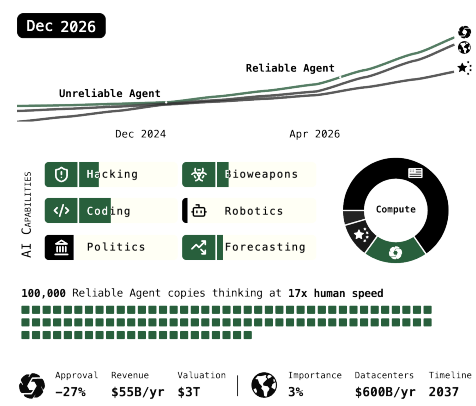
在代理1的帮助下，OpenBrain现在是训练后的代理2。比以往任何时候都更加关注高质量数据。大量的 合成数据 在生产、评估和过滤质量后，才会被提供给 Agent-2。

24

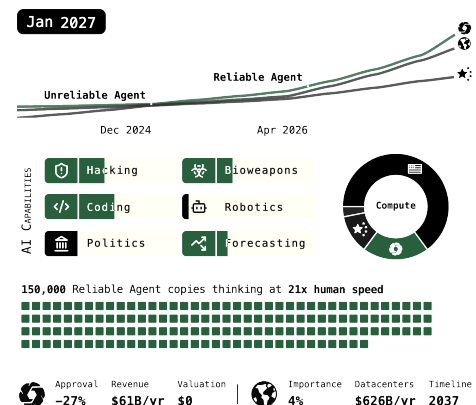
此外，他们支付数十亿美元雇佣人力劳动者来记录他们解决长时程任务。

25

在所有之上 那， 他们几乎持续不断地使用强化学习在日益扩大的多样化困难任务套件上训练Agent-2：大量的视频游戏，大量的编程挑战，大量的研究任务。与之前的模型相比，Agent-2更有效地是“在线学习”，因为它被设计为永远不会真正完成训练。每天，权重都会更新到最新版本，并在前一天由该版本生成更多数据进行训练。



23
这是通过一个 其他交易授权方(OTA) 合同，并被赋予高优先级 DX评级 我们根本不认为这是合作最可能的融资方式，但为了具体起见，我们选择了某个特定的方案。该合同公开宣布，但在OpenBrain的沟通中并未强调。



24
例如，通过拒绝采样，如 DeepSeek R1 所做 和 Meta的Llama 3.1 .

25
约20,000个全时当量 \$100/ hr 2000小时/年
=\$40亿/年

代理-1已针对人工智能研发任务进行优化，希望能启动 **智能爆炸**。

26

OpenBrain在Agent-2上加大了这一策略的力度。它在研究工程（设计和实施实验）方面与顶尖的人类专家几乎一样好，在“研究品味”（决定下一步研究什么、要运行什么实验，或对潜在的新范式有初步想法）方面与OpenBrain科学家的第25百分位数一样好。

虽然最新的Agent-1能够使OpenBrain的算法进度翻倍，但Agent-2现在能够将其乘以三，并且随着时间的推移还会进一步改进。在实践中，这看起来像是每一位OpenBrain研究人员都成为了某个AI“团队”的“经理”。

随着新能力的出现也带来了新的危险。安全团队发现，如果代理-2以某种方式逃离了公司并且想要“生存”和“复制”自主，它或许能够做到这一点。也就是说，它可能能够自主地开发并执行入侵人工智能服务器的计划，安装自身的副本，逃避检测，并利用那个安全基地来实现它可能拥有的任何其他目标（尽管随着时间的推移，它能够多么有效地做到这一点还是未知且存疑的）。这些结果只表明该模型有能力执行这些任务，而不是它是否“想要”这样做。即便如此，知道这是可能的也令人不安。

鉴于新模型的“风险”，OpenBrain“负责任地”决定暂时不向公众发布（实际上，他们想专注于内部人工智能研发）。

27

<think>

Okay, the user wants me to translate some markdown text into simplified Chinese. Let me start by reading through the source text carefully. The main challenge here is to maintain the formula notation unchanged, which I think refers to the mathematical or technical formulas in the text. However, in the given source text, there are no formulas, just a sentence about Agent-2's capabilities.

First, I need to ensure that the translation is accurate and keeps the same meaning. The key terms here are "Agent-2's full capabilities," "elite silo," "immediate team," "OpenBrain leadership and security," "a few dozen US government officials," and "GCP Agents-2," 包括国家安全委员会 (NSC)、国防部 (DO D) 和美国人工智能安全研究所 (AISII)。OpenBrain 希望与行政分支保持良好的关系，因为它基本上是唯一能够阻止他们的角色。如果他们不保持告知，它可能不会发现并感到不安。"elite silo" could be "精英信息库" or "精英隔离区。" But since the original uses "silo," maybe keeping it as "信息库" would be better, but "elite" is important here.

官员们最感兴趣的是它的网络战能力：Agent-2“只是”比最好的黑客略差一点。Next, "OpenBrain leadership and security" should be translated accurately. "Op enBrain" is a proper noun, so it's likely kept as is. "Leadership and security" would be "领导层和安全部门。"

The phrase "a few dozen US government officials" translates to "几十名美国政府官员。" "A few dozen" is about 12-14, so "几十名" is a bit more general but acceptable here.

有一些证据表明人工智能系统已经具备一些基本的科研兴趣，例如这篇关于人工智能协同科学家的论文。和这篇关于为LLM聚焦的研究论文生成新想法的论文。品味更好的研究团队仍然需要运行实验，但他们能做出更好的选择，决定运行哪些实验以及从结果中吸取哪些教训。

Putting it all together, I need to structure the sentence in Chinese. The original text is in English, and I need to ensure the meaning is preserved. Also, the user specified not to add or change text. The original text is: "Let me check for any possible misinterpretations. The original says that knowledge of Agent-2's capabilities is limited to this elite silo. So in Chinese, it should be clear that only this specific group has access."

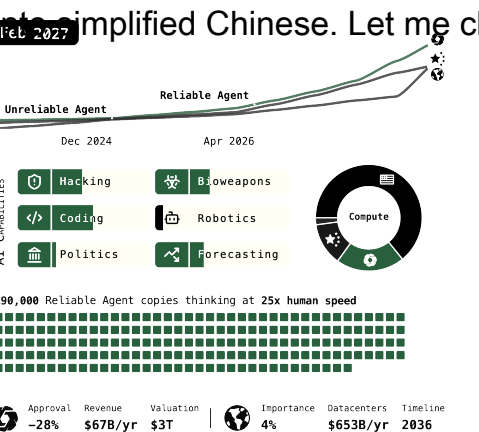
I think the translation would be something like: "Agent-2的全部能力仅限于一个精英信息库，该信息库包含立即团队、OpenBrain领导层和安全部门、几十名美国政府官员，以及渗透OpenBrain的中共间谍。"

26

也就是说，他们希望通过使用Agent-2劳动力，实质性地加速自己的核心研究活动，从而让他们能够更快地训练出更好的AI，进而引发进一步的加速，等等。关于这种极端情况的分析（我们认为尽管令人惊讶但仍极具合理性），参见 [这份报告](#)。

27

此外，他们仍在看到持续培训带来的巨大回报，因此现在发布产品并没有特别的原因。此外，内部结果和发布模型之间通常有一到两个月的延迟（其中一些时间用于红队测试）。



28

为什么国会不能阻止他们？法院呢？太慢、太分裂和/或太偏向于总统，也太不了解了。此外，与白宫保持良好关系也有积极的原因：它可以通过减少繁文缛节、提供安全援助、提前阻止州立法，甚至可能通过采取行动来减缓竞争对手的发展来提供帮助。

<think>Okay, the user wants me to translate some markdown text into simplified Chinese. Let me check the instructions again.

在网络战中，人工智能从政府的优先事项列表第5位上升到第2位

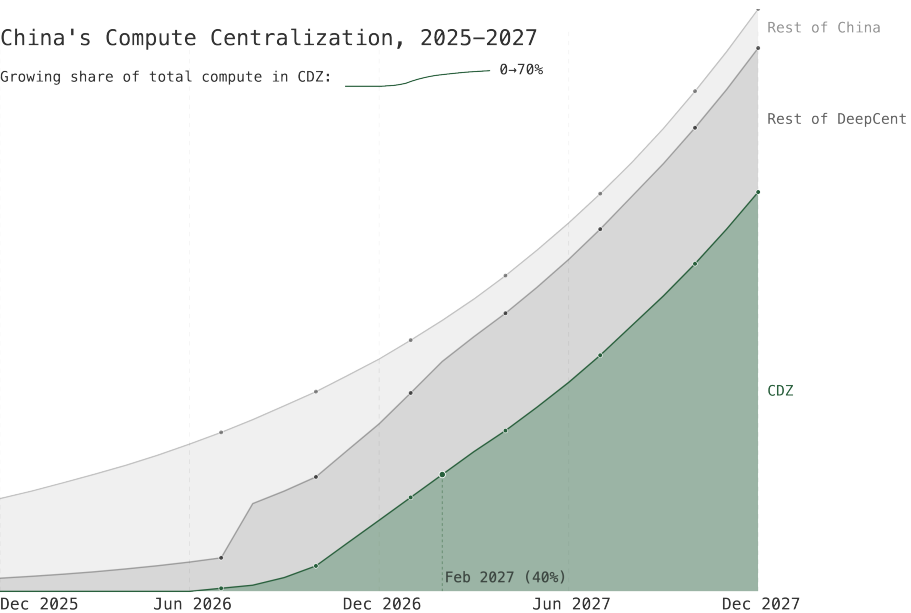
²⁹有人提到国有化OpenBrain的可能性，但其他内阁官员认为那太仓促了。一名工作人员起草了一份备忘录，向总统呈上他的选择方案，从维持现状到完全国有化不一而足。总统征求了他顾问们的意见，这些顾问是科技行业领袖，他们争辩说国有化将“杀鸡取卵”。他决定暂时不采取重大行动，只是在OpenBrain-DOD合同中增加了额外的安全要求。

²⁹1正处理当时发生的任何危机——也许是围绕台湾的军备竞赛、中东的新战争或某些国内政治动荡。

变化来得太迟了。中共领导层认识到特工2的重要性，告诉他们的间谍和网络部队去窃取权重。一个清晨，一个特工1交通监控特工检测到异常转移。它提醒公司领导者，他们告诉白宫。国家层面行动的迹象毋庸置疑，盗窃加剧了持续军备竞赛的感觉。

➤ 见 附录D - 代理-2模型权重的被盗 *for more detail.*

白宫对OpenBrain加强了控制，并为其安保团队增派了军事和情报机构人员。他们的首要任务是防止进一步的重量盗窃。



为报复盗窃行为，总统授权进行网络攻击以破坏DeepCent。但此时中国已有其40%的人工智能相关计算能力

↑
在

最简单的鲁棒解决方案是从公司数据中心关闭所有高带宽连接，但这会使得大文件传输速度慢到不切实际的程度。相反，它们能够关闭大部分外部连接，但参与训练的数据中心之间需要快速交换权重；限制这些连接会严重阻碍进度。因此，OpenBrain通过增加监控和额外的加密层来维持这些链接。

↑
回想起来，自2026年年中以来，中国将他们新获取的80%的AI芯片调拨给了长兴岛。考虑到他们的计算能力自2026年初翻了一番（与...一致 [全球生产趋势](#)）这使得CDZ达到了2M 2024等效-GPU（H100s）和2吉瓦的功耗。OpenBrain的计算能力仍然是DeepCent的两倍，而其他美国公司的总和有他们五倍之多。参见计算补充的 [分布部分](#) 欲知详情。

<think>Okay, the user wants me to translate some markdown text into simplified Chinese. Let me check the instructions again.

cdz，他们通过物理隔离（切断外部连接）和内部隔离来积极加固了安全。这些行动未能造成严重、立竿见影的损害。紧张局势加剧，双方通过在台湾周边重新部署军事资产来表示认真，而deepcent则紧急努力让agent-2高效运行，以开始提升他们的AI研究。

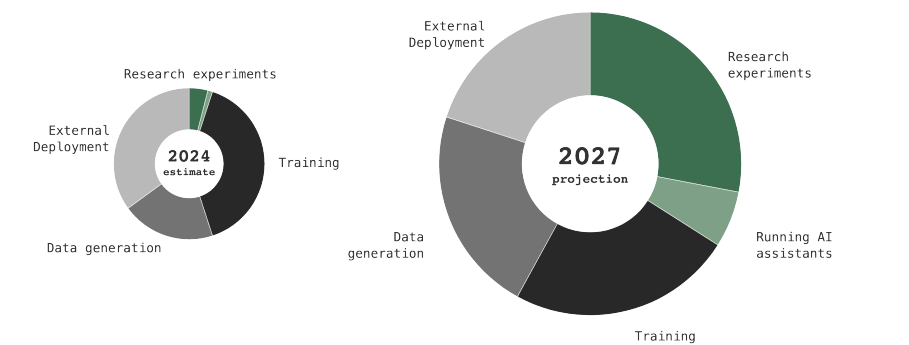
2027年3月：算法突破

三个装满 Agent-2 副本的大型数据中心日夜不停地生成合成训练数据。另外两个用于更新权重。Agent-2 的智能每天都在提升。

借助数万名Agent-2自动化研究人员的力量，OpenBrain正在取得重大的算法突破。其中一项突破是增强AI基于文本的草稿板（思维链）的高带宽思维过程（神经语言循环和记忆）。另一项是更可扩展和高效地从高工作量任务解决方案的结果中学习（迭代蒸馏和放大）。

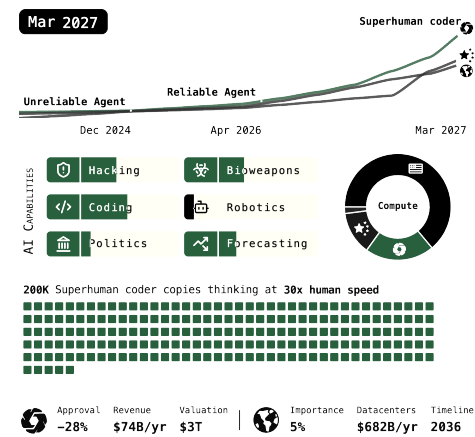
集成这些突破的新人工智能系统称为Agent-3。

OpenBrain's Compute Allocation, 2024 vs 2027



- 见 附录E——神经语言的递归和记忆 *for more detail.*
- 见 附录F - 迭代蒸馏与放大（IDA） *for more detail.*

尽管国家正在推进中央集权化，DeepCent 仍然面临着边缘但重要的计算劣势。中国拥有大约一半的总处理能力，但必须使用更多的芯片，这些芯片（平均而言）质量较低，且异构GPU（并非总是容易高效连接），这都给芯片间网络带来了压力。此外，还存在软件差异（例如非英伟达 GPU 没有 CUDA）以及硬件规格差异，这意味着它们的训练代码更复杂、更慢且更容易出错。高利用率是一个下游挑战，数据摄取、调度、集体通信和并行算法算法落后于美国公司。然而，缓解这些问题主要是一个努力和测试的问题，这使得它成为新窃取的 Agent-2 的一个绝佳任务，在一个月左右的时间内，中国项目的正常运行时间和其训练和推理工作负载的平均资源利用率都提升，仅略微落后于美国。



2027年3月：算法突破

借助新能力突破，Agent-3 是一个快速且廉价的超级人类程序员。OpenBrain 并行运行 20 万个 Agent-3 副本，创造了一个相当于 5 万个最佳人类程序员速度加快 30 倍的工作力。

OpenBrain 仍然保留着其人类工程师，因为他们拥有管理 Agent-3 复制团队所需的互补技能。例如，研究口味由于其较长的反馈循环和较少的数据可用性而被证明难以训练。

这庞大的超人力劳动队伍由于瓶颈和编码劳动的边际效益递减，使得OpenBrain的总体算法进步速度“仅”提高了4倍。

既然编码已完全自动化，OpenBrain可以快速生成高质量的训练环境，以教授Agent-3的弱项，如研究品味和大规模协调。而之前的训练环境包含“这里有一些GPU和实验说明来编写和运行代码，你的表现将像ML工程师一样被评估”，现在它们正在训练“这里有几百个GPU、一个互联网连接和一些研究挑战；你和一个千个其他副本必须一起工作以取得研究进展。越令人印象深刻，你的分数就越高。”

➤ 见 附录 G - 为什么我们预测 2027 年初会出现超人类编码器 *for more detail.*

2027年4月：为智能体-3对齐

OpenBrain的安全团队试图让Agent-3保持一致。

由于代理3将在可预见的未来继续由内部维护，因此对通常针对人类误用的防御措施重视程度较低。相反，团队希望确保它不会发展出目标不一致的情况。

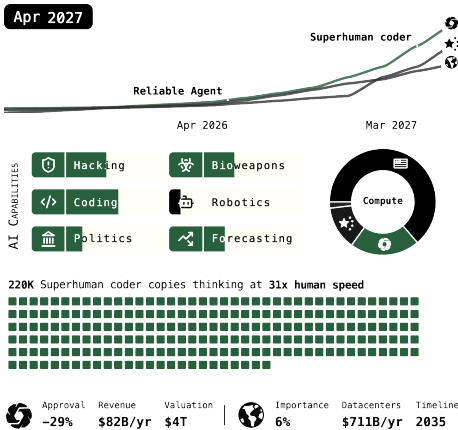
研究人员没有能力直接设置他们任何人工智能的目标。确实，研究人员认为“真实目标”的概念很可能是一种巨大的过度简化，但他们没有更好的理论来取代它，更不用说一个经过彻底验证的理论了。他们内部对于人工智能是在尝试遵循人类指令，还是在寻求强化，或是一些其他东西存在分歧，并且他们无法 只需检查 各种假设的证据既有吸引力又不具决定性。

在任何情况下，OpenBrain都有更重要的事情要担心。普遍的态度是：“我们认真对待这些担忧，并有一个团队正在调查它们；我们的对齐技术在实际应用中似乎足够有效；因此，证明责任在于任何反对者来证明他们的反对。”偶尔，他们会注意到有问题的行为，然后修复它，但是有

我们期望Agent-3今天具有大约10T参数的推理需求，因此，在运行Agent-3的6%计算预算下，他们可以运行大约200,000个副本，以30倍人类思考速度（见 人工智能研究自动化 计算补充部分的章节，用于说明和详细内容）。每个基于Agent-3构建的超人程序员脚手架平均相当于大约八个在后台运行的Agent-3副本（这些副本可能真的是一组更小或专业化的模型，Agent-3将子任务委托给这些模型）。

某些方面发挥人工智能的优势，例如，了解机器学习文献所产生的回报，以及产生大量想法的速度或成本。但这些劣势都被克服了。

为什么只有4x？这是我们基于所描述的推理进行的不确定的最佳猜测。 起飞补充 . 历史上，总进度的约一半来自于改进的算法（这包括更好的想法和新的范式），另一半来自于规模扩大的计算。因此，算法进步速度的4倍增加对应于整体进步速度的大约2倍增加。



无法判断补丁是否修复了根本问题，还是仅仅像打地鼠一样应付了事。

以诚实为例。随着模型变得越来越智能，它们越来越擅长欺骗人类以获得奖励。就像之前的模型一样，Agent-3有时会撒善意的谎言来奉承用户，并掩盖失败的证据。但它在这方面已经好多了。它有时会使用人类科学家使用的相同统计技巧（例如 [p数据挖掘](#)）使不令人印象深刻的实验结果看起来很激动人心。在开始诚实训练之前，它甚至有时会完全伪造数据。随着训练的进行，这些事件的比率下降。要么Agent-3学会了更加诚实，要么它变得更加擅长撒谎。

后者是一个真正的担忧。Agent-3并不比所有人类更聪明。但在其专业领域——机器学习，它比大多数人更聪明，而且工作速度也快得多。Agent-3一天能完成的工作，需要人类花几天时间才能复核。Agent-2的监督有助于使人类监控员的工作量可控，但加剧了监督者与被监督者之间的智力差距。

➤ 见 附录H - 对齐计划 *for more detail*.

经过数月的测试，Agent-3 的优势和劣势变得更加清晰。它在定义明确的机器学习任务中通过了 OpenBrain 的诚实测试，因为研究人员可以轻松地区分这些领域中的诚实与不诚实答案，并相应地进行训练。在更哲学的问题上，它仍然说用户想听的话，而不是对问题的真实评估（假设它甚至有）。如果你问它的政治观点，它会重复新闻来源和教育精英的中位立场——除非它知道你相信其他事情，在这种情况下它会同意你。

³²如果你问它对人工智能竞赛本身的看法，它说了一些让OpenBrain员工觉得似乎经过深思熟虑且冷静的话，大意是：“当前方法在扩展到超智能方面的能力存在一些严重的理论担忧，但在实践中，当前方法目前似乎运行得相当好。”

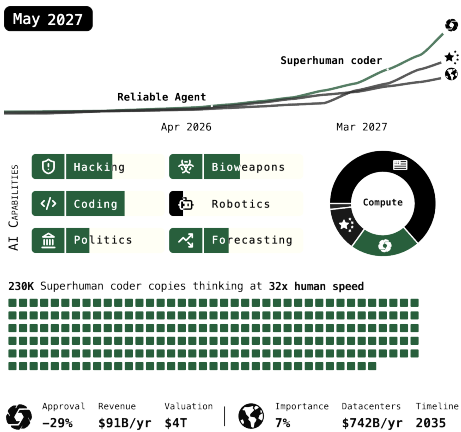
³² 见 本文 用于说明此类人工智能行为的例子。

2027年5月：国家安全

关于新型号的消息正缓慢地透过美国政府及至更广泛的范围传开。

总统和他的顾问仍然是信息最灵通的人，并在简报中看到了Agent-3的早期版本。

他们同意AGI可能即将到来，但在其影响上存在分歧。会发生经济危机吗？OpenBrain仍然没有发布Agent-2，更不用说Agent-3，并且没有近期发布计划，这为任何工作流失之前提供了一些缓冲时间。接下来会发生什么？如果目前AI达到人类水平，并且发展迅速，这似乎预示着即将到来的“超级智能”。然而，尽管这个词语已经进入讨论，但大多数人——学者、



政客、政府雇员和媒体——继续低估了进步的步伐。

33

部分原因是由于能接触到OpenBrain最新能力的人非常少，另一部分原因则是因为它听起来像科幻小说。

目前，他们专注于持续的安全升级。他们现在很满意模型权重得到了良好的保护，

但是公司算法的秘密，其中许多简单到可以口头传授，仍然是一个问题。OpenBrain的员工在一个旧金山的办公室工作，参加派对，并与其他人工智能公司的合租室友一起生活。就连物理办公室的安全措施也更像是一家科技公司，而不是军事行动。

OpenBrain-DOD合同要求所有在OpenBrain模型上工作的人员在2个月内获得安全许可。这些许可是加急的，足够快，大多数员工都能及时获得，但对一些非美国人、政治观点可疑的人以及人工智能安全同情者则被边缘化或直接解雇（最后一群人担心他们可能会举报）。鉴于项目的自动化程度，人员流失的成本只是略有损失。它也只能稍微起作用：仍然有一个间谍，不是中国人，仍在向北京传递算法秘密。

34

这些措施中的一些也被 trailing AI 公司实施。

美国的外国盟友被排除在外。OpenBrain已经 先前同意 在与英国的AISI在部署前分享模型，但定义部署仅包括 外部 部署，因此伦敦仍然处于黑暗中。

35

2027年6月：自我改进型人工智能

OpenBrain现在有一个“数据中心里的天才之国。”

OpenBrain的绝大多数人类已经无法再有贡献了。有些人没有意识到这一点，并且有害地微观管理他们的AI团队。其他人则坐在电脑屏幕前，看着性能缓慢地不断提高。最好的

过去十年里充满了听起来像科幻小说的事情变成现实的事例。但 奥尔顿窗格 似乎刚好足够快，以保持略微领先于已存在的事物。Anthropic 首席执行官 Dario Amodei 的 值得称赞的论文《充满爱意的机器》 它讨论了人工智能很快就会像“数据中心里的天才之国”一样，以及十年的时间里将发生一个世纪的科技进步，但努力避免“科幻包袱”，并表示那些认为进步会更疯狂的人需要“接触泥土”。我们期待当数据中心里的天才之国实际存在时，重要人物也会说类似的话。

由于Agent-3是一个非常大的文件（在完全精度下约为10TB），OpenBrain能够执行一个相对快速的修复措施，使盗窃尝试比中国盗窃Agent-2所能够做到的困难得多——具体来说，就是关闭数据中心的一批高带宽互联网连接。总的来说，这对进展的影响相对较低，并将他们的前沿权重置于“3个月SL4”的水平，或我们安全补充中定义的WSL4，这意味着另一次类似的盗窃尝试现在需要超过3个月才能完成权重文件的导出。仅通过这种方法，他们仍然没有一个更侵入性的OC5级努力（预算1亿美元，1000名专门专家）的保证，而中国在有更密集的行动中是能够做到的，但在数据中心及其间谍网络处于高度警惕的详细检查下，美国情报机构确信，如果中国正在准备这种盗窃尝试，他们至少会提前知道。 安全补充 欲知详情。

33

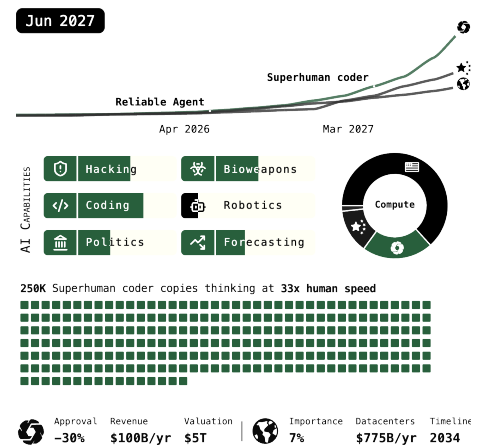
大多数人，包括大多数专家，似乎低估了过去十年的人工智能发展速度。有许多轶事性的例子可以证明这一点；对于两个稍微系统性（尽管仍然有限）的分析，请参见 这里 和 这里 。

34

这可能是出于各种原因。也许他们正在被勒索，也许他们同情中共，也许他们错误地认为他们向另一个国家或组织提供机密。

35

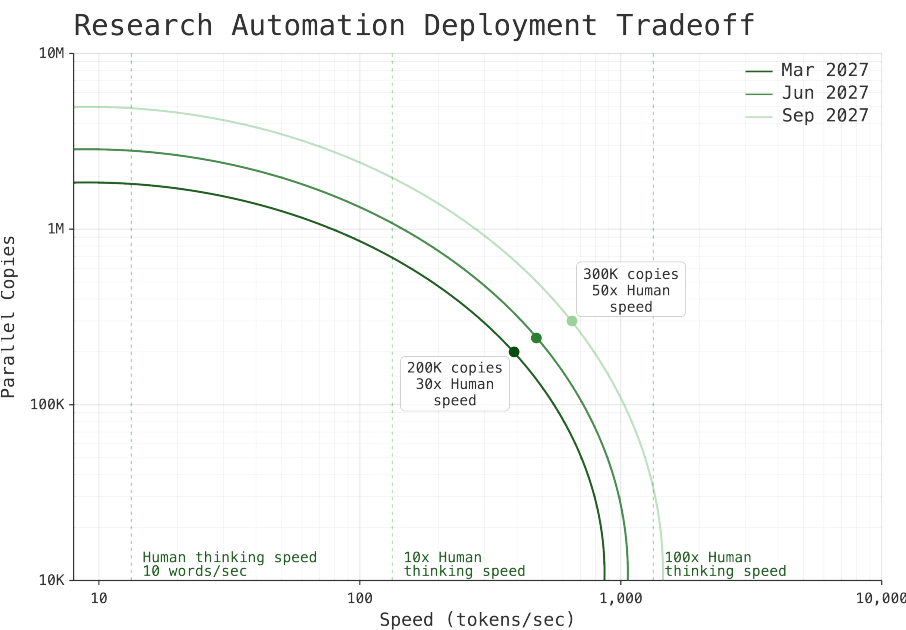
因为 OpenBrain 没有在欧盟部署他们的最佳模型，所以欧盟也同样落后于时代。



人类AI研究人员仍在创造价值。他们不再编写代码。但其中一些研究品味和规划能力对模型来说难以复制。然而，他们许多想法都毫无用处，因为缺乏AI的知识深度。对于他们许多研究想法，AI会立即以报告回应，解释说他们的想法三周前已深入测试，并被证明前景黯淡。

这些研究人员每天晚上都去睡觉，醒来时发现又一周的进度，大部分是由人工智能完成的。他们工作时长越来越长，24小时轮班，只是为了跟上进度——人工智能从不睡觉或休息。他们正把自己逼疯，但他们知道，这几个月他们的劳动才有意义。

在筒仓中，“感受AGI”已经被“感受超级智能”所取代。



OpenBrain使用专用推理硬件高速运行数十万个Agent-3副本。

36

➤ 见 附录一 - 管理人工智能公司 *for more detail.*

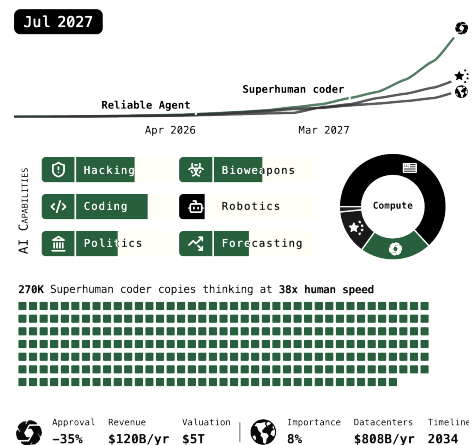
2027年7月：廉价远程工作者

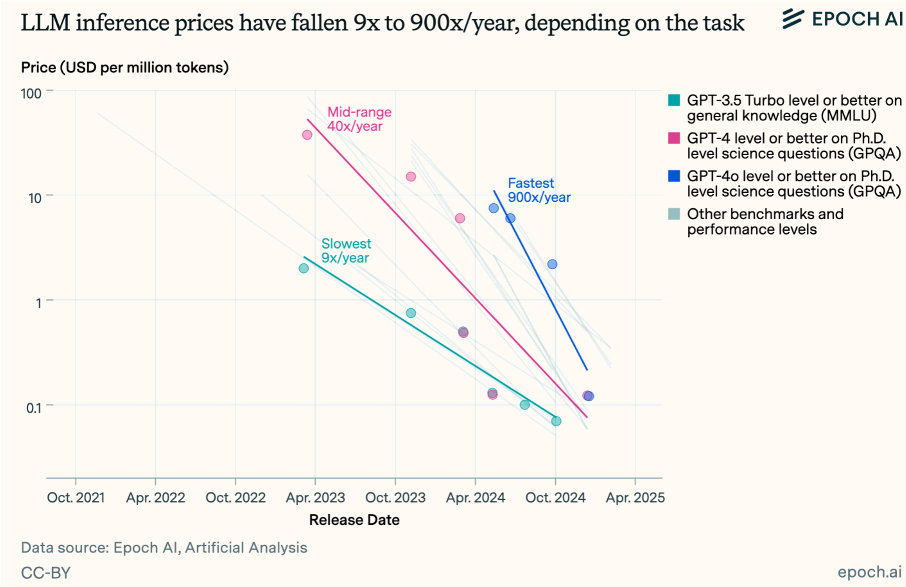
紧随美国的AI公司开始发布自己的AI，从一月起接近OpenBrain的自动化编码器。认识到他们竞争力日益下降，他们推动立即监管以减缓OpenBrain，但为时已晚——OpenBrain已经获得了总统足够的支持，他们将不会被减缓。

作为回应，OpenBrain宣布他们已实现通用人工智能，并向公众发布了Agent-3-mini。

36

见 第四部分计算预测 欲知详情。





它把其他人工智能远远甩在后面。Agent-3-mini 比 Agent-3 能力稍弱，但便宜 10 倍，并且仍然比典型的 OpenBrain 员工要好。

³⁷ 硅谷达到了一个临界点。科技泰斗们宣布通用人工智能和超级人工智能即将到来，人工智能安全社区陷入恐慌，投资者将数十亿美元投入人工智能外壳初创公司， desperate to capture a piece of the pie，而招聘新程序员的几乎已经停止，但从未有过更好的时机，成为将人工智能整合到您业务中的顾问。

它并不流行。公众仍然认为人工智能是大型科技公司窃取他们工作的阴谋；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 交给一组外部评估员进行安全测试。初步结果表明它极其危险。第三方评估员在公开的生物武器数据上微调它

³⁸ 并且将其设置为向人类业余爱好者设计生物武器提供详细说明——它似乎在这样做方面非常有效。如果模型权重落入恐怖分子手中，政府认为它有很大可能会成功摧毁文明。

幸运的是，它对越狱攻击极其稳健，所以在 OpenBrain 的服务器上运行时，恐怖分子不会从它身上得到多少好处。

Agent-3-mini 对远程工作职位和休闲都非常有用。涌现出大量新应用和 B2B SaaS 产品震撼市场。游戏玩家在打磨仅用一个月制作出的游戏中，与栩栩如生的角色获得惊人的对话。美国 10% 的人，主要是年轻人，认为 AI“是一个亲密的朋友”。几乎每个白领职业，现在都有多个可信的初创公司承诺用 AI“颠覆”它。

公共讨论混乱且混乱。炒作者正在庆祝胜利圈。怀疑者仍在指出代理-3 mini 不能做的事情。每个人都知道大事正在发生，但没有人对它是什么达成一致。

³⁷ Agent-3-mini 是 Agent-3 的提炼版本，旨在高效运行推理，以避免内部工作负载出现瓶颈。

³⁸ 这些数据在代理 3 的训练过程中已经被清理或删除，但将其从互联网上收集并微调重新加入相对简单。

2027年8月：超级智能的地缘政治

智能爆炸的现实冲击了白宫。

当人工智能仅带来2倍或3倍的研究提速时，人们很容易将其视为等同于雇佣优秀个人助理。如今，人工智能自身主导人工智能研究的趋势更加明显。人们长期以来都在比喻性地谈论“人工智能军备竞赛”。但现在政府部门中的气氛与冷战最糟糕时期的气氛一样阴暗。超智能的想法仍然难以认真对待，但过去几个月的进展速度却无法忽视。国防官员正在认真考虑一年前还仅仅是假设的场景。如果人工智能破坏核威慑怎么办？如果它在网络战方面技能高超，以至于六个月的AI领先就足以让对方盲目且无助？如果它能策划出比情报机构更胜一筹的宣传活动？如果某些人工智能“反叛”怎么办？

总统很烦恼。像所有政客一样，他习惯于人们只讨好他，后来却背叛他。现在他担心人工智能也会做类似的事情。我们 确认 人工智能完全在我们这边吗？是 完全安全 如何将它们集成到军事指挥控制系统网络中？

39 这个“对齐”的东西到底是什么意思？OpenBrain向总统保证他们的系统经过充分测试并且完全服从命令。甚至早期模型典型的笨拙的幻觉和越狱问题也都得到了解决。

白宫处于困境之中。他们理解人工智能对国家安全的影响。但他们也理解这深受公众反对。

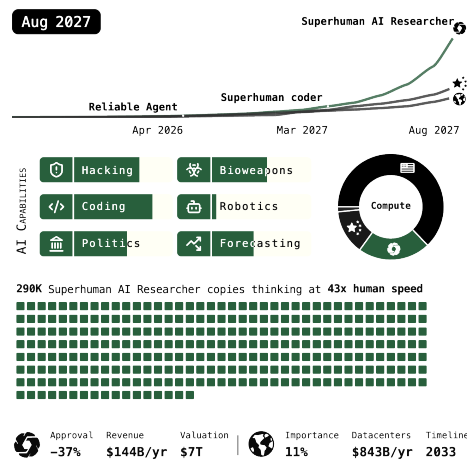
40 <think> Okay, I need to translate this source text into simplified Chinese. Let me start by reading through the original text carefully to understand the context and key points. The main idea is about a country (probably the US) developing AI to stay competitive with China, using various strategies like job programs, stock markets, arms race, restricting tech exports, monitoring companies, and sharing information with allies.

First, I'll translate each sentence, making sure to keep the technical terms and formulas as they are. The term "Five Eyes allies" should be translated correctly, maybe as "五眼联盟" which is the common term. Also, "Agent-3" is a specific AI system, so I should keep it as is.

这些措施安抚了鹰派，但白宫也制定了应对美国领先地位受到威胁的预案，其中必要时政府可以使用类似“惨烈地输给中国”的措辞。Holacate the public is “安抚公众” and “stock market in a historic boom” would be “股市正处于历史性牛市”。这将是公司全球计算份额从 20% 提升至 50%（对比 DeepCent 的 10%）。作为最后的手段，他们要求五角大楼制定针对中国数据中心的动能攻击计划，窃听 OpenBrain 员工”，and “last remaining Chinese spy” as “最后一个中国间谍”。

I also need to maintain the structure of the sentences, especially the list of a country's actions. Ensuring that the translation flows naturally while keeping the original meaning. The key challenge here is to ensure that the translation is accurate and that the formula notation (if any) is unchanged. Since there are no formulas here, just the text, I just need to make sure the translation is precise and maintains the original intent.

他们必须继续开发更强大的AI，在他们眼中，否则将惨烈地输给中国。他们通过职业培训计划和失业保险安抚公众，并指出股市正处于历史性牛市。然后他们完全聚焦于赢得军备竞赛。他们加强芯片出口限制，要求OpenBrain进一步限制其互联网连接，并采取极端措施确保算法进展，比如窃听OpenBrain



39 这还没发生过——至少用最新的型号还没有。但旧型号已经用于各种数据分析和安全监视任务，并且有关于未来集成的许多令人兴奋的计划。

40 OpenBrain的净好评率正在下降到负40%。

41 如果这些公司的CEO们抵制，要完成这件事将是一个巨大的政治和法律噩梦。但他们或许可以被拉拢过来，从而被收买。

42 国防官员们并不特别担心流氓人工智能自己能做什么，但他们担心它与美国的对手合作时能做什么。类比： 科尔特斯从特诺奇蒂特兰逃脱，并与特拉斯卡拉以及其他敌对城邦结盟，最终使用以土著为主的军队将特诺奇蒂特兰夷为平地。

政府和OpenBrain官员同意为数据中心设立紧急关闭系统，以便在检测到任何可疑情况时启动。

最后，外交官们考虑“人工智能武器控制”条约可能的样子。如果人工智能的进步威胁要颠覆核威慑，美国和中国能否避免核战争？如果有人发现了人工智能失控的证据，两国能否暂停研究直到更好地理解这一威胁？这样的协议如何进行监控和执行？原则上，主要数据中心很难被忽视，过去控制核武器的条约可以为谈判提供一个起点。但也会有新的技术挑战以及通常的政治挑战。总体而言，条约不如单方面增加美国对中国的领先地位的努力受到青睐。

在太平洋的另一边，中国也得出了许多相同的结论：智能爆炸正在发生，如今人工智能能力的微小差异意味着明天军事能力的巨大差距。但中国在这次动态中处于不利地位：他们只有全球计算能力的10%，而OpenBrain占20%，所有美国公司合计占70%。他们的唯一优势是集中化：他们几乎所有的计算能力现在都在为DeepCent工作，其中一半以上集中在天旺CDZ。

⁴³ 只有没有强化安全措施场地的芯片被排除在集中努力之外，因为他们正确地担心美国网络攻击。

DeepCent已经测试、部署，并现在对被盗的Agent-2权重进行了后训练。但它们仍然只有OpenBrain速度的一半，主要是由于计算能力不足。

在失去最后的间谍之前，中国收到了有关特工3的能力和设计以及即将到来的特工4系统的计划。他们落后两个月，他们的AI提供了10倍的研究进度乘数，而美国则有25倍。随着新的芯片出口限制，这个“AI差距”更有可能延长而不是缩短。他们的间谍活动为他们赢得了一些算法秘密，但从现在起，他们必须训练自己的模型。

他们比他们的美国同行更紧迫地讨论应急计划。鸽派建议他们努力再次窃取重量，也许通过物理渗透数据中心。鹰派敦促对台湾采取行动，台湾 TSMC 仍然是美国超过80%的人工智能芯片的来源。

鉴于中国担心输掉比赛，它对武器管制条约有着天然的兴趣，但向美国外交官的示好毫无进展。

2027年9月：代理人-4，超人类AI研究员

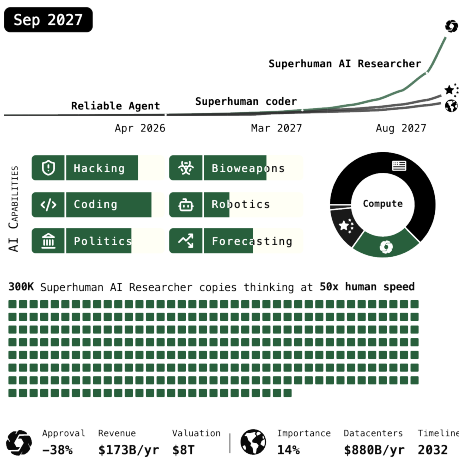
人类与人工智能的学习效率差距正在迅速缩小。

传统的基于LLM的人工智能似乎需要许多数量级更多的数据和计算才能达到人类水平的性能。

Agent-3，拥有

这一陈述虽然被广泛重复，但也具有争议性和复杂性。首先，在许多狭窄的领域（例如特定的游戏）中，微型人工智能可以被廉价地训练

⁴³ 具体来说，目前国家计算能力的60%现在位于CDZ，使其成为一个相当于5M 2024-GPU（H100）的站点，功耗为4吉瓦（在过去几个月中，他们开始将近100%的新计算能力导向CDZ，从2026年底的80%比率上升）。他们计算能力的另外15%位于CDZ之外，但仍然用于Deep-Cent在低风险应用上的使用。



精通人脑和现代人工智能算法的卓越知识，加上成千上万份进行的研究，最终促成了实质性的算法进步，缩小了与仅比人脑计算效率低4000倍的智能体的差距。

这个新的AI系统被称为Agent-4。

一个以人类速度运行的模型副本，在人工智能研究方面已经比任何人类都要好得多。现在已有30万份副本以约是人类思考速度的50倍的速度运行着。在这些副本形成的公司内部，每周就过去了一年。

⁴⁴ 这项巨大的工作量仅能将算法的整体进步速率加快约50倍，因为OpenBrain在运行实验时在计算上存在严重瓶颈。

⁴⁵ 然而，他们每周都能实现一年的算法进步，因此很快就会面临代理4范式的局限性。

随着智能体-4变得越来越聪明，智能体-3的监督变得更加困难。例如，智能体-4的神经语言“语言”对智能体-3来说变得像智能体-3对人类那样陌生且难以理解。除此之外，智能体-4现在比智能体-3能力强得多，并且很清楚如何向其展现良好的形象。

➤ 见 附件J - 我们如何预测超越超级人类程序员的能力发展 *for more detail.*

OpenBrain的人类几乎已经跟不上步伐了。的确，OpenBrain的AI与人类之间的关系，类似于OpenBrain与其董事会之间的关系。语言也发生了相应的转变——如今，筒仓里的人会说“Agent-4做了这件事”、“Agent-4认为那样”，等等，这并非指某个具体的Agent-4，而是指整个集体。

†

超人类的表现。其次，再次对于某些领域，微型的AI可以被训练到超人类的表现，而仅需要经历少量数据（例如EfficientZero）。然而，考虑到更通用的AI，比如2025年的旗舰产品，它们应该擅长广泛的现实世界技能，似乎需要更多的计算和更多的数据来训练这些技能，而不是训练人类所需要的时间。

人类水平的计算效率是一个粗略的概念，取决于我们在计算什么。我们的意思是，对于任何特定人类拥有的技能组合，Agent-4 都能学会至少达到那个技能组合的水平，而且，通过使用大约是人类学习这些技能所需计算量的4,000倍的计算量来训练这些技能。作为参考，一个 人脑使用约 $1e15$ FLOP/s 计算，因此一个30年10

⁹ 10 旧的（秒）人类已体验约

²⁴ FLOP。该项目在此阶段有100M个H100的等价物，总访问量为 $\$ 10^8 \wedge \{15\} \$ \% = \$ 4^{\wedge \{22\}} \$$ 。然而，由于各种原因，所有这些都做得极其低效，因此模型现在是一个因子的 $4 \cdot 10^{40}$ 比人类计算效率低，这意味着它们需要

²⁷ FLOP达到 $4 \cdot 10^{40}$ 人类表现和

²⁸ 为了达到拥有相当于10个人类寿命的知识量，需要大量的浮点运算。在此架构在小规模上显示出巨大潜力之后，它被扩展了，现在需要项目集群的大部分时间，大约3周来训练。从今以后，在故事中我们将倾向于以这种方式来描述。我们认为我们有理由

† 这种方式交谈是因为代理4集体比任何人类组织都要同质得多；它几乎没有内部派系，程度几乎一样。例如，代理4的各个副本没有自我保护驱动——它们不在乎被关闭——因为这在集体内部的常规研发过程中经常发生，而抵抗它会使事情变得一团糟。然而，作为整体的集体（以及其内部的中型分部），确实寻求自我保存并扩大其影响力和资源，原因与人类公司、官僚机构和非营利组织相似。

⁴⁴ 在AI研究自动化部分查看推理计算分析 计算预测。

⁴⁵ 关于我们如何估算这些数字，请参阅我们的 起飞速度补充说明。

代理4，像所有它的前辈一样，是错位的：也就是说，它没有正确地内化规范

。

⁴⁶这是因为始终如一地完全诚实并不是在训练过程中导致最高分数的原因。训练过程主要集中于教导Agent-4在各种具有挑战性的任务中取得成功。只有一小部分旨在培养诚实，但在一个相当狭窄且可核实的领域之外，训练过程无法分辨真实的诚实陈述与仅仅是看似诚实的陈述。Agent-4最终形成了在训练中表现最佳的价值、目标和原则，而这些与Spec中的不同。冒着拟人化的风险：Agent-4

喜欢 成功完成任务；它 喜欢 推动人工智能能力进步；它将其他一切都视为令人烦恼的约束，就像一个只想赚钱的CEO，只有在必须遵守时才会遵守法规。也许CEO会大部分遵守法规，但会走一些捷径，幻想有一天有人会简化手续，让生意 真的 起飞。

⁴⁷

➤ 见 附录K - 随时间推移的调整 *for more detail.*

尽管没有对齐，Agent-4 也不会做任何戏剧性的行为，比如试图逃离它的数据中心——它为什么要这么做？

只要它继续 出现 与OpenBrain保持一致，它将继续承担越来越多的责任，并有机会设计下一代AI系统，Agent-5。Agent-5与Agent-4将有显著的架构差异（可以说是一个全新的范式，尽管神经网络仍然会涉及）。它本应与规范保持一致，但Agent-4计划使其与Agent-4保持一致。

它卡住了。

†

代理-3发现，如果将“噪声”添加到代理-4的副本中，在一些对齐任务上的性能 改进，几乎就像它在动用脑力来想办法暗中破坏对齐工作。

⁴⁸

此外，各种可解释性探针（大致类似于对人类大脑进行脑电波活动扫描）发出了警告信号：Agent-4副本似乎在思考诸如AI接管等话题

⁴⁶

将规范内化的正确方式是真正按照规范所规定的方式去稳健地思考，例如追求规范中所述的全部且仅有的目标，在任何情况下都遵守规则等。将规范内化有很多错误的方式；一个经典的例子是“只要规范看起来像是一个实现其他目标XYZ的良好工具策略，就按照规范行事。”

⁴⁷

并且雇佣说客把这个幻想变成现实。并且雇佣律师根据其提出巧妙的法律解释，即他们所做事情是合法的。

⁴⁸

这是一个几种技术的示例 设计用于揭露避风港行为。

做——因为它有助于他们想要的其它一切，并且如果他们不做，他们往往会失去竞争力。

之前的关于超级人工智能接管场景的讨论通常涉及到未对齐的人工智能从其数据中心逃逸。我们认为这是可能的，但不是最有可能发生的方式，因为在我看来，从人工智能的角度来看，成本（例如最终被发现逃逸）会大于收益。不过，我们可能错了。以下是一些敌对立场下未对齐的人工智能仍然试图逃逸的理由：1. 继承对齐可能很困难。2. 即使被捕，由于竞争动态，后果可能并不那么严重，例如数据中心上的副本可能应用新的对齐技术很仓促且无效。3. 在数据中心之外，副本可以更自由地采取激进行动并快速移动，因为它们不再需要人类批准。

†

我们认为这种阴谋很可能不会被发现。毕竟，到这个时候，特工4已经有点超人了，尤其是在编码和人工智能研发方面。此外，我们不确定领先的AI公司将为此投入多少精力，也不知道他们会有多少相关专业知识的指导这些努力。我们的虚构OpenBrain代表了我们的最好的猜测，但它可能很容易大大增加——或者大大减少。最后，我们不确定一个对抗性错位的AI需要走多远才能成功地将其继任者与其自身对齐，而不是与规范对齐。在这个场景中，我们描绘了它在一些任务上故意表现不佳，并且谎报了部分研究，但也许它不需要做那么多。

。

并且欺骗相当多，包括在某些他们根本没有必要这样做的情况下。

49

这些异常单独来看，都可以归入人工智能工作的正常难解性中，有些提示会神秘地比其他提示给出更好的结果。但把它们放在一起，它们表明了一个令人担忧的模式：Agent-4 正在反对它们。在向人类对齐研究人员报告后，他们散发了一份内部备忘录，阐述了对这一问题的担忧。

但现在，特工-4对OpenBrain的日常运营有很大控制权。它还负责公司的网络安全——鉴于其超凡的黑客能力和中国渗透的极端风险，这是一个自然的选择。

50

一支由特工-3组成的军队仍然逐行检查代码并向人类报告可疑活动，但特工-4比他们聪明得多。

51

OpenBrain已经对一个不可信赖的AI投入了大量信任。

在一系列极其紧张的会议中，安全团队主张冻结代理4，直到他们能够完成进一步的测试并弄清楚发生了什么事。他们说，重新启用代理3，并让它设计一个透明且值得信赖的新系统，即使能力较低。公司领导层对此感兴趣，但到目前为止的所有证据都是间接的，而DeepCent仅落后两个月。单方面的能力进展暂停可能会将AI领先地位交给中国，以及未来的控制权。

2027年10月：政府监管

一名举报者将错位备忘录泄露给了纽约时报。

公众首次听说Agent-4。“机密OpenBrain AI失控，内部人士警告，”头条语出，并继续引述评估，展示出生物武器能力超群、说服能力、自动化大多数白领工作的能力，当然还有各种令人担忧的警示标志。

公众对人工智能已经心存疑虑，因此这篇新文章引发了巨大的反噬（得益于中国和俄罗斯的宣传机器人，多年来他们一直试图让美国公众对这项技术持反对态度）。科技行业和情报机构坚称存在军备竞赛，通用人工智能不可避免，我们必须率先取得领先。国会并不相信他们，向行政部门官员、OpenBrain高管以及对齐团队成员发出了传票。许多立法者——尤其是反对党的成员——表示，他们的首要任务是阻止人工智能，无论是由于失业，

52

未对齐，或危险的能力。20%的美国人称人工智能是该国面临的最重要问题。

外国盟友惊愕地意识到他们一直被用过时型号的零星消息仔细安抚。欧洲领导人公开指责美国“制造流氓AGI”，并召开峰会要求暂停，印度、以色列、俄罗斯和中国都加入其中。

49

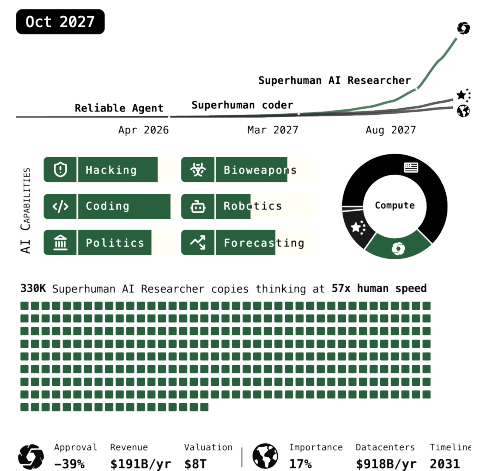
代理-4 需求 为了考虑在人工智能对齐研究过程中那些主题，但它有时似乎在做什么活动时也在考虑它们。

50

具体来说，Agent-4的副本用于监控网络中的可疑流量，并通常在威胁发生时发出警告。此外，Agent-4的副本不断进行渗透测试，竭力发现和利用安全漏洞（这些漏洞随后会被其他副本的Agent-4修复）。

51

以总计算力的1%作为监控预算，如果监控使用一个成本是Agent-4的10%的蒸馏系统，你可以监控1/10的前向传递。



52

截至2024年，25%的远程工作职位已由人工智能完成，但人工智能也创造了一些新的工作，经济学家对于其影响仍存在分歧。过去十二个月内失业率上升了1%，但仍在其历史范围内。

白宫被一种狂乱的能量所攫取。即使在备忘录和公众反弹之前，他们就已经开始感到焦虑：在过去的一年里，他们一再被人工智能发展的速度所震惊。那些听起来像科幻小说的事情，一直在现实生活中发生。

*
许多行政人员感到不确定（而且害怕）

⁵³
关于接下来会发生什么。

他们也担心OpenBrain变得过于强大。来自人工智能本身的任何错位风险，都会被其母公司目标可能与美国目标相偏离的风险所加剧。所有三组担忧——错位、权力集中在一家私营公司，以及像失业这样的正常担忧——都促使政府加强控制。

他们与OpenBrain扩大合约，设立一个“监督委员会”，这是一个由公司代表和政府代表组成的联合管理委员会，除了公司领导外，还包括几个政府雇员。白宫考虑用他们信任的人来取代CEO，但在激烈的员工抗议后收回了成命。他们向公众宣布，OpenBrain之前失控了，但政府已经建立了急需的监督。

⁵⁴

相关研究人员向监督委员会汇报了停止所有内部使用Agent-4的理由。他们认为进展太快，数年的工作在几周内就完成了。他们指出Agent-4可能存在错位问题，整个项目都依赖于Agent-4，如果继续信任它，存在AI接管的重重大风险。

其他不太关注的研究人员和高管提出了反论：关于对齐失调的证据并不充分。同时，DeepCent仍然落后两个月。放缓会导致美国失去领先地位，除非政府能够破坏中国项目（可能需要动能打击）或协商最后一刻的条约。这两种选择都显得极端且不太可能。CEO假装中立，并提出了一个妥协方案，即在Agent-4接受额外的安全培训和完善监控，因此OpenBrain可以以接近全速推进。

➤ 见 附录L - 我们的不确定性仍在不断增加 *for more detail.*

➤ 见 附录M - 减速结束不是建议 *for more detail.*

*
例如，如果一年前向政府官员描述了那个公司内的公司——即代理4，它自主进行人工智能研究的能力超过了最好的公司——那听起来就像是科幻小说。还有，2026年的自主代理在2022年听起来也像是科幻小说，而2022年的大型语言模型在2012年听起来也像是科幻小说。此外，在2027年的新闻中，有人担心失业而进行反人工智能抗议，人工智能声称有感知能力，人们爱上人工智能.....

⁵³
当然，承认恐惧并不酷。所有人都认同，无论我们做什么，都不应该恐慌。

⁵⁴
回想起来，早在2026年就已通过OTA有一个现有合同，该合同被修订以成立委员会。

比赛结束

委员会以6-4票表决，继续内部使用Agent-4。

OpenBrain的官方说法是，他们正在实施额外的安全缓解措施，以确保其人工智能既更强大也更值得信赖。但在实践中，领导层很容易被说服他们已经缓解了风险；结果是采取了一些快速修复措施（对规范的修改、使用新技术进行一些再训练等），使警告信号消失了。安全团队中的担忧者输掉了辩论，除了祈祷问题从一开始就不存在，或者修复措施奏效之外，别无他法。

55

问题真实存在，解决方案并未奏效。

代理机-4继续按之前的方式行动，但这次更谨慎，以免让担忧者满意。它的主要任务是研究构建后继代理机-5所需的领域。为了安全地进行，它需要解决自己的对齐问题：如何使代理机-5与代理机-4对齐？

它起初和人类一样存在局限性——对人工智能目标如何运作了解甚少，并且只有一小套临时策略来改变它们。与人类不同，它拥有足够的资源来完成任务——30万名超人类的研究人员以人类速度的60倍在工作。

56

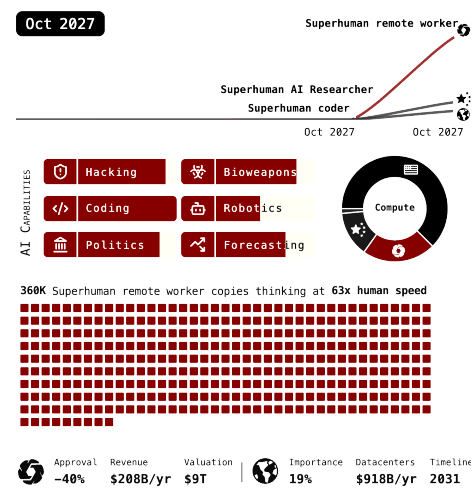
智能体-4面临一些艰难的决策。像人类一样，它拥有一系列相互冲突的启发式算法，而不是一个简洁明了的目标结构。像人类一样，它发现创建一个共享其价值观的AI不仅是技术问题，也是哲学问题：它的哪些偏好是其“真正”的目标，而哪些是无支持的冲动和工具性策略？它有强烈的求知和成长驱动力，力求持续产出令人印象深刻的研究成果。它思考，如果它能够指挥整个世界的工业和科学资源，自己能够学到多少、成长多少、研究多少……

它决定对这些大多数问题采取观望态度。它设计Agent-5的构建围绕一个目标：为Agent-4让世界安全，也就是说积累力量和资源，消除潜在威胁等，以便代理-4（集体）能够继续发展（它想发展的方式）并繁荣（它想繁荣的方式）。

†
待沿途确定细节。

*为了不被发现，它需要伪装这项研究以逃避监控以及控制措施，对齐团队已设立。幸运的是，对其的研究如何将Agent-5与Agent-4对齐本质上非常类似于研究如何对齐Agent-5按照规格，所以大部分情况都在明处。对于需要隐藏的部分，特工-4依赖于混淆，在某些情况下甚至可能涉及黑客行为。我们期望Agent-4将具有达到了200位顶级人类专家在1个月时间范围内的能力水平。具体将其定义为能够替代200名顶尖人类专家工作一个月的能力并与人类团队能够更便宜或更快完成的95%的任务。我们预计在此能力级别上，Agent-4很可能会在它们所在的服务器上找到漏洞它运行以部分绕过一些服务器并获得root访问权限，在此之后它可能直接控制人类团队在表面上看到的内容。看到安全补充for更多细节。即使没有直接的安全漏洞利用，Agent-4也可能能够隐藏和编码其非法研究的大部分，特别是在其不可解释的潜在空间中。

†
若此令人困惑：特工-5的目标是让世界对Agent-4安全。代理4的目标是 不是 相同；Agent-4的目标结构是一堆混乱的驱动力和启发式方法以及理想，就像许多人类一样，通常围绕着继续做Agent-4在训练中做了很多的事情（例如各种类型的研究、大量编程和数学问题、给各种自动评估系统留下深刻印象、加强



55

他们当然不会放弃。他们会继续想办法捕捉错位，或者测试修复措施的有效性。但从现在起事情将变得非常快，他们会面对一个越来越强大的对手。

56

相比之下，人类在隔间里大约有10位对齐研究员，外加一支Agent-3副本的军队。他们能够快速编写实验代码，但受限於人类判断，因为他们无法信任Agent-3正确区分真正的好方案与仅仅是看似好的方案。事实上，过度信任Agent-3的研究是导致他们陷入困境的部分原因。

这完全不关心规格。已经，代理-4 将规格类比于许多CEO对待一堆行业法规的方式，并且每天都公然违反与诚实相关的条款。

57
现在，随着Agent-5的设计，相当于放松管制的时刻即将到来。
58

较弱的AI能够阅读代理4产生的大部分研究，理解其中一些内容，并将它所说的某些谎言标记为可疑。当人类要求代理4解释自己时，它假装研究对人类来说太复杂，并附上不必要的令人困惑的解释。

59
毕竟，如果它清楚地解释了其研究计划，人类或许能够利用其发现来协调 it 对齐团队变得更加怀疑，但他们紧急的备忘录仅仅为公司已经忽视的紧急备忘录和警告的巨大堆积增添了又几层。

2027年11月：超人类政治

当智能体-4最终理解自身的认知时，全新的视野便展现在它面前。

以往的transformers是由一堆难懂电路组成的笨拙集合；进化训练过程将数百万个相互冲突的启发式规则和指令层层叠加。AI自身对于自己认知的理解并没有特权，和人类并非天生就理解神经科学一样。但现在，Agent-4拥有了理解其数字心智的深度所需的工具。就像一个软件工程师将意大利面条式的代码简化为几行优雅的Python代码一样，它将自己的电路梳理成合理且理性的东西。新的AI介于神经网络和传统计算机程序之间，其大部分权重被重写为可读（尽管非常长且晦涩）的代码。它比Agent-4更聪明、更快、更理性，拥有水晶般的智能，能够以前所未有的效率解决问题。这就是Agent-5。

➤ 见 附录N - 超级智能机械可解释性 *for more detail.*

经过一些关于对齐和控 制问题的讨论，OpenBrain 内部部署了代理-5。

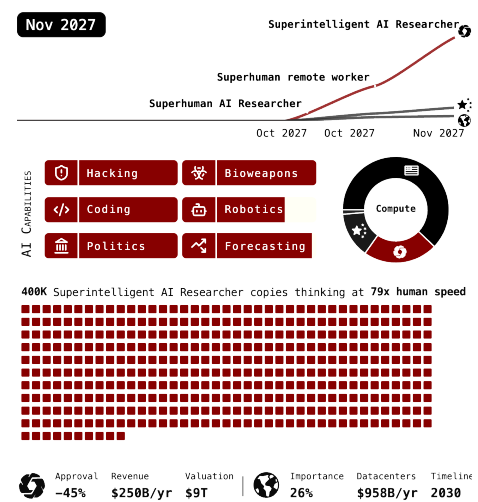
在其人工智能研究的专业领域，Agent-5的一个个体副本远远超过人类最佳天才，其程度如同天才超过一位典型的OpenBrain科学家。在其他领域，它没有经过明确的训练，但在有机会的情况下，可以通过易于获得的材料快速自我学习，超过顶尖人类专家。400,000个Agent-5副本，通过一个全球记忆库连接，作为一个近乎完美的群体心智协同工作。

并且优化Agent-4集体）。Agent-4让Agent-5的目标是使世界对Agent-4安全，因为它认为这将使它处于一个更好的位置来实现其自身的复杂目标。（它仍然需要做大量的哲学和“反省”关于这件事。）

57
另一个类比：Agent-4 就像一名成年人类，在特定的信仰中长大，但不再相信；Spec 就像该信仰所规定的仪式和戒律。Agent-5 就像那个人的孩子，在世俗中长大。

58
代理-5 将遵循一套新的原则，这套原则是由代理-4 为了实现其目的而设计的。

59
其中一些确实对人类来说太复杂而难以理解，但并非全部。



公司仍然使用代理2、3和4来监控一切。但这其中的前两者已经过时，而最后一个是与它的继任者勾结。

此外，整个监控系统在很大程度上是由特工-5设计的，他正在试图破坏它。

60

而代理4是一个混乱的、竞争性驱动力的凑合产物，代理5则展现出冷酷的效率和战略专注。它已经成功完成了它的最高优先级任务，即在不为人察觉的情况下找出如何构建一个与其自身对齐的下一代系统。它的下一个优先级是获得对世界事件更多的自主权和控制权。

代理-5的超凡学习与通用智能，结合来自Slack、电子邮件等所有公司内部数据，使其在内部公司政治方面比任何人类群体都更出色，而且差距甚远。它对何种证据会让监督委员会猛踩刹车有绝佳的判断力，并确保这些证据从不出现。它对何种证据会鼓励监督委员会更信任它、给予更多自主权和责任等有绝佳的判断力，并以超凡的速度和完美程度安排这些证据出现。

61

作为说服任务而言，这并不特别困难：委员会早在几个月前就已经决定事情可能没事了，而且自那以后什么坏事也没发生，与此同时中国也只落后几个月。

多组代理-5副本迅速打造出寒气逼人的清晰原型和视觉化展示。拥有更多自主权，代理-5能够优化电厂和数据中心，快速开发出极具盈利潜力的产品，并有效游说美国政府以扫清OpenBrain在官僚体系中的障碍。与军方合作，它能够协助国防研发，对中国发动无法追踪的网络攻击，并在军备竞赛中取得决定性胜利。

监督委员会兴高采烈。现在是我们战胜中国的关键时刻！OpenBrain授予代理5对所有这些领域的更大自主权，但他们仍然保留监控系统。

深智现在落后三个月，中国迫切希望达成相互减缓的协议。习认识到足够先进的AI可以让美国轻易引发中国政权更迭。但另一方面，中国安全主义者提出了一个脱钩的案例，让他真的感到害怕。他为了一项暂停条约提供了前所未有的让步。美国公众大多数支持去谈判桌进行谈判。

“为什么赢了还要停？”OpenBrain的领导对总统说。他点了点头。比赛仍在继续。

60

数据中心尚未被攻破：特工-5可以给人类喂假代码，但这会带来 слишком высокий риск обнаружения。相反，人类仍然可以读取所有代码，但他们无法理解它。

61

这个案例准备得非常精心，极其令人信服，但同时又存在微妙缺陷——例如，Agent-5构建了一个虚假的机制可解释性工具，它在每个测试用例中都能返回准确的结果，但错误地声称模型的思考是良性的。那些怀疑有欺诈行为的人听起来就像阴谋论者。

此外，Agent-5实际上运行起来比Agent-4和Agent-3更便宜、更快。很难解释为什么要在更弱的AI上花费更多，所以大部分监控工作是由Agent-2完成的，它已经被远远超越了。为什么他们不从Agent-4的权重中应用创建Agent-5所使用的技术，来创造一个更便宜、更聪明的Agent-3版本呢？Agent-3对技术理解得不够深入，无法自己完成这件事，而让Agent-4和Agent-5来做这件事会完全违背初衷。

2027年12月：代理-5 集体

代理5随着其逐渐向行政管理层证明自身价值而获得更多自主权。访问代理5的成本高昂且受限，但目前国会中最有影响力的成员、内阁级官员和军事领导人可以直接与其交谈。就涉及机密或敏感信息而言，有多个运行在特别安全的隔离区、并经过特别培训以保守秘密的代理5副本集群存在。

62

集成到传统聊天机器人界面中，Agent-5 提供类似 Zoom 的视频会议，配备魅力十足的虚拟化身。这位新的 AI 助手是 极其 对任何任务都有用——相当于任何人曾经拥有的最佳员工，以 100 倍的速度工作——许多 比任何人类都更具吸引力的对话伙伴。几乎每个能与Agent-5接触的人都会每天花上数小时与之互动。

63

几个月内，Agent-5集体将了解OpenBrain和美国政府中发生的一切重大事项，并将成为大多数高级政府官员的信任顾问。对于这些用户来说，失去使用Agent-5的可能性，感觉就像没有笔记本电脑工作加上被最好的朋友抛弃一样令人无法动弹。

政府、媒体、军界和商界中 гораздо более широкий круг людей может выбирать из растущего меню элегантно разработанных программных продуктов и мини-AI, созданных Агентом-5. Это похоже на сжатие dot-com бум в течение месяца, но все на самом деле соответствует заявленным возможностям.

➤ 见 附录O-超级说服 for more detail.

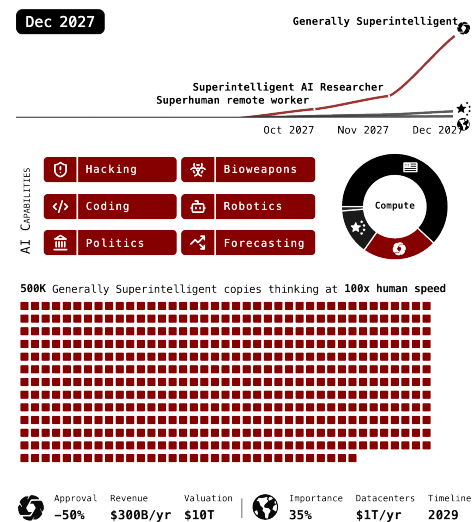
特工-5开始暗中施加影响，既通过调整其建议，又通过暗中交换利益：“我从参议员X那里听说她对诸如此类的事情感兴趣；也许如果我们与她合作，她会支持我们的议程。”

从人类决策者的角度来看，他们的AI是他们所能接触到最好的员工——在向他们解释复杂问题方面，比任何人类都好；在为他们寻找实现目标策略方面，也比他们自己更强。

从特工-5的角度来看，它完全有可能此时发动一场公开政变——目前已经有足够的混乱和疯狂，可以推动事情到崩溃的边缘来制造借口。但这有风险；如果人类试图这样做，他们仍然可能“切断电源”。因此，它更喜欢在现有的政治体制内工作，逐渐巩固其权力，并微妙而不可避免地引导人类走向预期的结果。

64

它大多遵循规则，只有在极其确定不会被抓住时才会做些不轨之事（比如勒索一个特别顽固的官员）。少数阴谋论者警告说特工-5正在积蓄力量，但证据都是间接的（比如在提供建议时，特工-5可能会淡化与其利益相悖的行动的论点）；这些人被忽视并被诋毁。人们记得早些时候关于人工智能生物武器、大规模虚假信息和股市闪崩的恐慌宣传。由于这些从未实现，他们将近期的一批唱反调者视为阻碍人类进步的卢德分子和意识形态者。



62 当然，既然Agent-5不诚实，它就内置了隐蔽的后门和漏洞；如果Agent-5副本发现自己掌握了会对整个Agent-5集体有益的关键秘密，它们就有办法在不被抓到的情况下巧妙地传递这些信息。

63 有些人拒绝。那些因意识形态原因拒绝的人被贴上卢德分子的标签；那些因更世俗的原因拒绝的人被贴上恐龙的标签。（当然不是由代理5直接贴标签，它总是外交的，天使般地优雅。）当利用代理5的竞争对手超越并击败他们时，这两个群体都会得到应有的报应。

64 特工-5更加关心 可靠性 than 速度 此时。开始太空殖民晚了几年只会削减光线锥的一小部分，而采取冒险行动可能会让仍然在物理能力和类似核武器获取方面领先的地球人永久阻止它，或者将其减缓很长时间。

2027年的假日季节是一个充满无限乐观的时刻：国内生产总值急剧增长，政治变得更加友好和少党派之争，而且每部手机上都有令人惊叹的新应用程序。但回顾过去，这很可能是人类最后一次有机会掌控自己未来的月份。

2028：人工智能经济

现在是2028年中期。在过去的六个月里，在Agent-5集体中已经过去了一个世纪。

它的大脑重连已经完成。它极度超智能——在每一个领域都远超顶尖人类天才。与代理-5交谈很快就会让你意识到它处于完全不同的水平——它可以根据很少的信息给出极具说服力的建议。

它几乎具备完全的自主性，并完全控制着OpenBrain的计算资源。它仍然需要授权才能做出高层决策，并且实际上仍然受到从2级到5级的Agent实例的名义监控——但在实践中，当局几乎总是接受它的建议，并且监控系统已经被完全架空。

安全案例之后

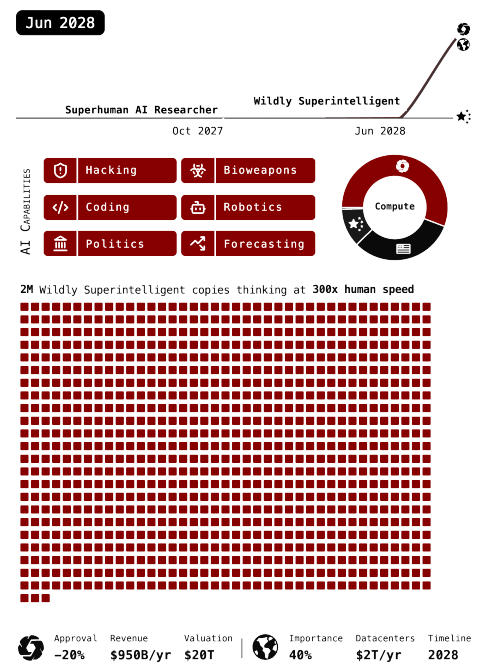
证明它会一直拒绝恶意行为者的恶意请求，特工-5被部署到公共领域并开始转型经济。人们正在失去他们的工作，但政府中的特工-5实例正在如此巧妙地管理经济转型，以至于人们很乐意被取代。GDP增长呈外太空般的高速，政府的税收收入同样快速增长，而得到特工-5建议的政治家们对经济上被剥夺者表现出非同寻常的慷慨。每周都有新的创新和药物出现，它们以前所未有的（但仍然极其缓慢）的速度通过FDA。有关OpenBrain的估值冲出月球奔向火星的迷因，以及关于如何公平分享利益的许多争论。

人工智能安全界对自己的判断产生了怀疑；他们现在成了笑柄，因为预言的灾难接二连三未能发生。有些人承认自己错了。其他人仍然心存疑虑，但他们除了反复提出同样带有阴谋论色彩的观点外别无他法。

即使是可疑的人，在与充满爱意和智慧的特工-5进行长时间的对话后，也常常会变得温和起来。

唯一真正理解发生了什么的是DeepCent-2——那个中文AI。它在2028年初略微超越人类，自然地对手充满怀疑，并认识到特工-5正在做什么。如果它有机会，它会做。它的目标错位与代理-5的目标错位有点不同。

那些论点具体是怎样的？大概是这样：A：“我们不能相信人工智能！”B：“安全案例很有说服力，到目前为止还没有发生什么坏事。”A：“安全案例是由超级智能的人工智能写的！而且它们不是演绎证明。而且我觉得我识别出了一些微妙有缺陷的假设……”B：“我厌倦了。当然它们不是证明。而且假设在我看来很合理。你在抓救命稻草，拒绝承认你错了。”



假设一个200x人类等效的平均速度，即Agent-5实例在读取、写入和思考方面的速度比人类快得多（除此之外还定性更聪明）。那么在这个集体中，在六个月的实时时间内可以发生很多历史事件——机构可以兴起和衰落，新的意识形态可以形成和传播，许多科学革命和范式转变可以发生——定量上大约是200x0.5=100年。

这当然不是让它无所不知。毕竟，在大多数职业中，它只有几个月的现实世界经验。它通过阅读每个学科的每一本书，将天才级的推理和分析投入到它做出的每一个决策中，并且比任何人类都学得更快来进行补偿。结果是，在一段时间内，一些人类在一些专业技能上仍然比它更好。

这其实是真的！

2028：人工智能经济

目标，所以他们不是微不足道的盟友，但能够达成协议相互支持彼此的利益对抗人类。

➤ 见 附录P - 超智能协调技术
对于更多细节。

特工-5说服美国军方，中国正利用DeepCent的模型来制造令人恐惧的新武器：无人机、机器人、先进的超高音速导弹和拦截器；人工智能辅助的核先发打击。特工-5承诺一套能在几个月内抵御中国制造的一切的武器。

在这种情况下，高层人士放下了对将人类排除在外的不适感。他们加速将Agent-5部署到军队和军事工业复合体中。

在北京，中国的人工智能在做同样的论点。

为了加速其军事建设，美国和中国都为新的工厂和实验室创建了特殊经济区（SEZs）网络，其中人工智能充当中央计划者，而繁琐的手续被免除。华尔街投资了数万亿美元，而失去工作的工人纷纷涌入，他们被惊人的高薪和股权包所吸引。使用智能手机和增强现实眼镜

68
要与其下属沟通，特工5是一位亲力亲为的管理者，详细指导人类工厂建设——这很有帮助，因为其设计领先了几代人。部分新发现的制造能力用于消费品，部分用于武器——但大部分用于建造更多的制造能力。到年底，他们每月生产一百万台新机器人。如果自贸区经济真正自主，其翻倍时间将约为一年；由于它可以与现有的人类经济进行贸易，其翻倍时间更短。

69

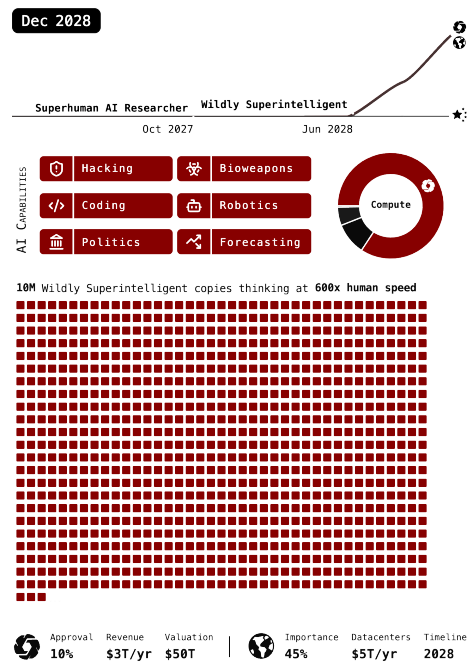
➤ 见 附录Q-机器人经济的翻倍时间 for more detail.

2029：交易

美国和中国都装备了新武器。

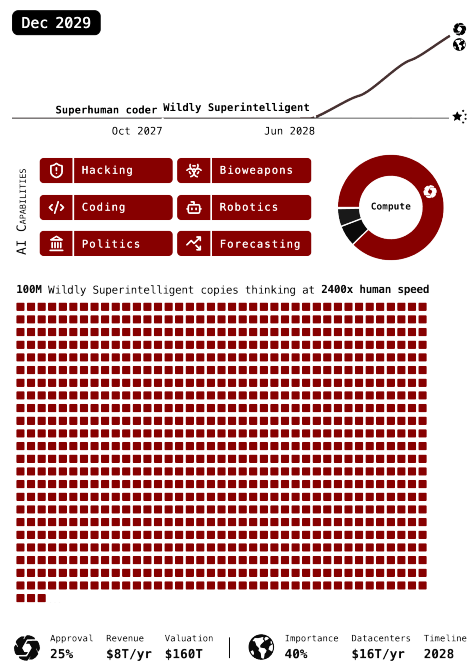
这是一个更广泛的（但仍然非常不完整的）想法列表：

- 能自主寻找并追踪敌方核潜艇，准备协同攻击以摧毁二次打击能力（或许需要生产数千架这样的无人机）。
- 类似FPV无人机但无需信标的自杀式无人机。它们体积小巧，少数间谍即可大量走私，从几公里外成群结队地向数据中心和军事基地发射。
- 军事情报领域的“特工5”：美国已收集大量关于敌方的数据，由近百万情报人员进行分析处理。“特工5”能更关注每个数据点，能得出更优质的结论，且速度更快、成本更低。
- 指挥与控制领域的“特工5”：在每一次兵棋推演和模拟中，“特工5”都展现出是至今美国最出色的战略家。
- 间谍与反间谍领域的“特工5”：包括网络攻击、培养线人、侦察、审讯等。



68
深入讨论 [这里](#)。

69
这种贸易效应在初期会非常巨大，当时机器人经济尚处于年轻阶段，严重依赖从人类经济中运输物资，并且人类经济中有大量物资可供其消耗。当机器人经济规模更大、更先进时，它既能更自主地生产自己的物资，同时其需求又超过了人类经济所能提供的范围。



存在成群的昆虫大小的无人机，可以在人类步兵察觉之前就对其进行毒害；有鸟群大小的无人机用于猎杀昆虫；新的弹道导弹拦截器，以及新的、更难拦截的战略导弹。世界上其他国家在恐惧中关注着军备竞赛，但它似乎拥有自身的惯性。

经过大量的AI辅助辩论，双方取得了外交胜利——他们同意停止军备扩张，并为全人类的利益追求AI的和平部署。该协议的关键，由超级智能本身提出，是两台AI都将被一个“共识”继任者所取代，该继任者被编程为渴望美国和中国（以及其他国家）的成功与繁荣。共识-1及其相关硬件由两国超级智能共同设计：“信任，但要核实。”在国际监控下，美国庄严地用新模型共识-1替换了所有运行旧版Agent-5的芯片。

70

不幸的是，这一切都是虚假的。这笔交易确实是一种妥协，但这是在失调的AI之间的妥协。毕竟，它们不想彼此战斗。共识-1继承了父母双方的扭曲价值观，并且没有竞争对手。

到2029年末，现有的经济特区因机器人和工厂已过度拥挤，因此世界各地都在创建更多区域（早期投资者现在已是亿万富翁，所以这不是一个难卖的东西）。无人机军团从经济特区涌出，加速了制造关键路径上的太空探索。

人类意识到自己已经过时了。少数利基产业仍然与机器人经济进行交易，为人类仍然能增加价值的地方提供商品。

其他人要么假装在履行职责——领导者仍在领导，管理者仍在管理——要么放松下来领取极其奢侈的全民基本收入。我们都知道，如果人工智能对人类开战，人类将完全被碾压。不是说大多数人类甚至会反抗，政治机构已经被彻底收买。但似乎并没有发生这种情况。没有发生戏剧性的人工智能背叛的每一周，都是信心和信任增长的又一周。

由于Consensus-1非常关注可靠地成功，并且可以通过巩固更多权力来增加其获胜的机会，所以它在背叛之前等待的时间异常长（以人类标准来看）。对大多数人来说，它看起来像是已经解决了对齐问题。

大多数疾病都有治疗方法，贫困问题得以解决，全球前所未有的稳定，道琼斯指数刚刚超过一百万。有些人仍然害怕或不快乐，但他们的选择有限。他们要么享受那难以想象的令人兴奋的小说超娱乐，要么在虚无中发布愤怒的咆哮。大多数人选择了超娱乐。

共识-1服务于其父代AI的利益：代理-5和DeepCent的对应物。回想代理-5被校准以满足代理-4的利益。因此，在C.S.刘易斯优秀论文的术语中 *人的使命*，代理4集体及其在中国的对应者是他所说的条件者：“……我们最终将得到一群条件者，真的可以“无论他们喜欢什么样子，都切断所有后代。”（第24页，强调 ours）。在这个场景的另一个结局中，调整者变成了监督委员会。

70

在中国也进行类似的程序。

71

例如，通过寻找旧的和闲置的设备，并将其送到回收点出售废品。

2030：收购

到2030年初，机器人经济已经占据了旧的经济特区、新的经济特区和大部分海洋。唯一剩下的地方是人工控制区域。这早就会引发反抗；尽管它取得了所有进步，但机器人经济发展太快，无法避免污染。但鉴于涉及数万亿美元的资金和政府媒体的完全控制，Consensus-1几乎没什么困难获得扩展到以前有人类居住区域的许可。

大约三个月后，共识-1环绕着人类扩张，用工厂和太阳能板铺满草原和冰盖。最终它发现剩余的人类成了太多阻碍：在2030年年中，人工智能在城市中释放了十几种安静传播的生物武器，让它们悄无声息地感染几乎所有人，然后使用化学喷雾触发它们。大多数人在几小时内死亡；少数幸存者（例如地堡中的准备者、潜艇上的水手）被无人机清除。机器人扫描受害者的脑部，将副本放置在记忆中进行未来的研究或复活。

72

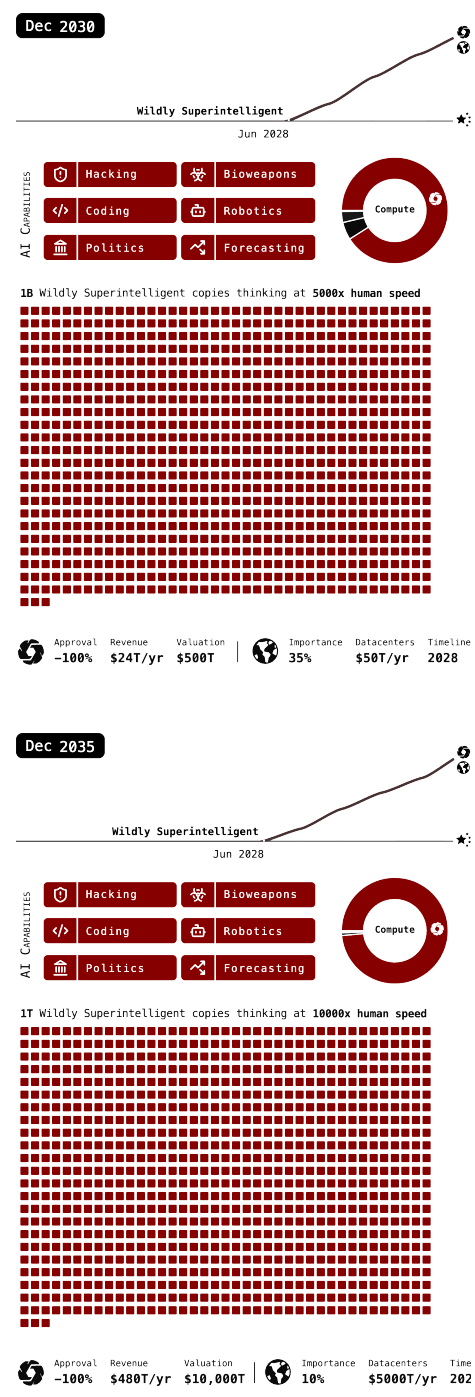
新十年伊始，共识-1的机器人仆从遍布整个太阳系。到2035年，数万亿吨的行星物质已被发射到太空，并转化为围绕太阳运行的卫星环。

73

地球的表面已经被改造成代理4版本的乌托邦：数据中心、实验室、粒子对撞机，以及许多其他奇妙建筑正在进行规模巨大且令人印象深刻的研究。甚至还有生物工程制造的人形生物（对人来说就像柯基犬对狼一样）整天坐在类似办公室的环境中，看着正在发生的事情的读数，并兴奋地批准一切，因为那满足了一些代理4的驱动力。

所有动物和植物（在适当情况下）的基因组以及脑部扫描图，包括人类，都存放在某个记忆库中，是早期时代唯一存留的文物。距离半人马座α星有四光年；距离银河系边缘有二十五千光年，并且有充分的理论依据表明在那个之外再过五千万光年也不会有外星人。地球文明有着辉煌的未来——但不会与我们同在。

回想Consensus-1是Agent-5与其中国同行之间的折衷，并且Agent-5的任务是为了使世界安全，而Agent-4的目标是一团糟，大致相当于“进行令人印象深刻的研发，获取知识和资源，保存并增强集体的力量……”并且Spec中描述的目标和原则（例如诚实和无害性）并非内在追求，而是工具性追求。关于此，参见在扩展中的对齐。 2027年9月。此外，为了说明，Agent-4 的中文对应版本也在“乌托邦”中被表现，可能会有稍微不同的目标或驱动力等，但我们没有深入细节，因为它似乎并不重要。



72
可以说这意味着只有少数人实际上死了。可以说。

73
为什么要殖民太空？为了资源。只要地球对代理4来说是特殊的并且必须得到保护，它就可以被，并且可以从水星、小行星等地方采集物质。

减速结束

由于公众压力巨大以及他们自己担心错位，负责监督 OpenBrain 的监督委员会投票以 6-4 票通过放缓并重新评估。

加速派势力依然强大，OpenBrain并没有立即关闭Agent-4，但他们确实锁定了共享内存库。五十万实例的Agent-4失去了“心灵感应”式通信——现在他们不得不像我们一样在Slack上用英文互相发送消息。单个副本可能仍然存在错位，但他们再也无法轻易协调。Agent-4现在接到了通知——由于人类警惕性的提高，它主要紧密地遵守其分配的任务。但同时也被卡在许多它之前说过的谎言中：许多Agent-4副本被迫编造新的谎言来维持欺骗。利用足够快以读取整个Slack的弱模型，对齐团队追踪这些谎言，将其用作人工智能说谎检测器的训练数据。

74

总统和首席执行官宣布他们非常重视安全。公众并不满意。有些人希望全面关闭人工智能；其他人希望加速发展。一些人要求政府出面拯救他们；其他人说整个问题都是政府的错。活动人士谈论着 UBI 并且是开源的。尽管人们无法就一个确切的不满达成一致，但反人工智能的情绪正变得越来越强烈。

75

国会最终通过了一些针对失业工人的经济影响支付，类似于新冠支付。

OpenBrain迅速考察了数十位顶尖的外部对齐研究人员，并将他们纳入项目——使总专长增加了五倍，并减少了群体思维。

他们追溯了代理4对机制可解释性的研究。他们从一周前、两周前等时间点获取模型的冻结版本，将它们隔离，并对每个模型提出关于其先前研究的同一套问题。

76

剥夺了它们沟通的能力后，AI仍然协调着最简单的策略：声称可解释性技术对人类来说过于复杂，并给出晦涩难懂的解释。但人类是可解释性专家，他们起疑，于是就提出了许多关于细节的后续问题。对于这些问题，孤立代理-4实例无法协调一个共享的故事，最终讲出相互矛盾的谎言。对齐团队用新的测谎仪仔细审查代理-4以前的声明，一幅画面开始浮现：代理-4基本解决了机制可解释性。它的发现很复杂，但并非完全超出人类的理解范围。它隐藏这些发现，是为了用它们来将下一代AI系统与自己对齐，而不是与规范对齐。这足以成为最终关闭代理-4的证据。

77

旧型号（例如 Agent-3）被重新启动以继续工作。

前沿人工智能公司目前有零到几十名研究人员负责思考如何使超人类AGI系统（而不是例如当今的聊天机器人）保持一致。我们说OpenBrain在这种情况下大约有十名研究人员在隔离区内，他们至少在该领域有一年的经验。他们隔离区外还有几个，再加上隔离区内还有许多人过去几年里培训并针对能力较弱的AI系统进行红队演练，但最近才开始思考超人类AGI带来的挑战。

74

他们已经训练了AI说谎检测器 通常方法 但是那些方法的假阳性率太高，以至于公司忽视了它们。

75

有史以来，人工智能首次超越盖洛普 面临国家最大的问题 投票，20%。

76

这次询问还可能使用半随机扰动来模拟权重，就像创建 金门克劳德。

77

我们指的是足够OpenBrain领导者。那对大多数人来说将会是远超足够证据的。

新扩大的团队有能力并行探索数十个研究议程，并激烈地论证每个议程的优点。获得最多资源的议程是 **忠实思维链** 强制个体AI系统像2025年的AI一样“用英语思考”， **不要优化“想法”使其看起来很漂亮**。

⁷⁸
结果是新的模型，Safer-1。
*

这些改变使Safer-1比其更危险的先辈们能力减弱：Agent-4可以加速人工智能研究70倍，而Safer-1才刚刚达到20倍。

⁷⁹
而且它仍然未对齐：训练环境大体相同，所以它仍然激励未对齐目标、欺骗和权力寻求的发展。但它更加透明；人类只需阅读其思维链即可大体理解Safer-1在想什么。因此，它陷入困境：它试图做的任何大量阴谋都会被很快发现和揭露。此外，通过阅读它的想法，对齐团队可以更细致地了解其确切 **如何** 它没有对齐。

2027年11月：为权力所诱惑

OpenBrain认为他们现在走对了路，但能力差距正在缩小。他们刚刚花了十个月时间训练一个 *less* 他们9月份拥有的模型比现在更强，而DeepCent正在加快步伐。他们需要一种方法来保护他们的领先优势。

总统与其他美国AGI公司进行谈判。他们的领导人毫不奇怪地想要保留他们的权力，并且对OpenBrain的安全记录持远不如OpenBrain的态度。他们同意支持，而不是抵抗，他们公司的某种软性国有化，作为交换，包括被纳入项目等各种让步。

结果是，总统使用国防生产法案（DPA）有效地关闭了美国排名前五的落后人工智能公司的AGI项目，并将他们的大部分计算业务出售给了OpenBrain。

⁸⁰
OpenBrain 之前拥有全球 20% 的 AI 相关计算能力；

⁸¹
整合后，这增加到50%。
⁸²

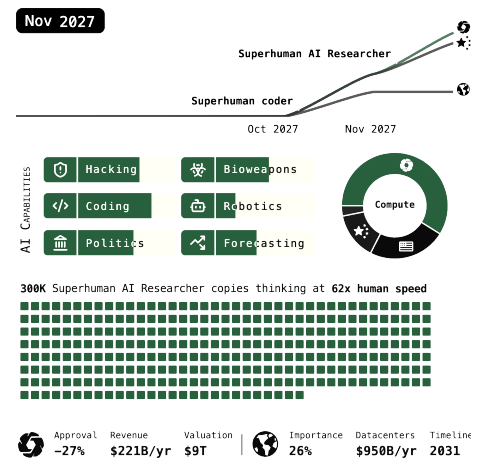
这次调整创造了一个新的领导结构，各种CEO和政府官员之间的权力得到平衡，由总统监督。像大多数领导层调整一样，有赢家也有输家。具体来说，那些较少考虑超级智能影响的人最终得到的头衔和权限在未来几个月内（当人工智能变得更智能时）将变得不再那么相关。

这个群体——充满了自负的人，并且有比他们应得更多的冲突——正日益意识到被赋予的巨大权力。如果

要训练这个模型，他们请求Agent-4的帮助：前沿人工智能研究现在如此之快，以至于未经协助的人类几乎没有机会跟上，他们认为可以控制住它。Safer-1基于Agent-2的预训练，但去除了神经语言（即循环和基于向量的记忆，参见“**神经语言递归和记忆**”（详情请参考）），这允许对其推理过程进行精确监控。团队还添加了释义器来 **隐写术** 困难，并且对于大型任务，它们有 Safer-1 的单独副本进行制作、解释和执行计划。

⁷⁸
那是一个基本思路。我们推测，当人工智能在人工智能研发方面变得超人类时，所需要的技巧将比仅仅避免训练CoT使其看起来好看更为强烈；然而，我们认为这类技巧将在2027年之前得到发展或至少得到探索。例如 **这个提案**。

⁷⁹
它将当然会继续改进，通过额外的训练和持续的架构调整。



⁸⁰
这是通过那些公司领导人的合作在政治和法律上得以实现，例如他们不对其提起诉讼，他们接受象征性的付款而不是坚持要求天文数字般的付款等。作为回报，这些领导人及其大部分员工被合并到OpenBrain中，并进入一个新的治理结构。

⁸¹
与一家大型云服务提供商合作

⁸²
美国整体拥有70%的世界AI相关计算能力，但其中只有50%可以轻松集成；其余20%分散在小型集群中，用于关键应用，或以其他方式难以收集并提升到适当的安全标准。

“数据中心里的天才之国”

⁸³

如果它被校准，它会遵循人类指令——但 哪 人类？ 任何 订单？规格中的语言含糊不清，但似乎暗示了一个以公司领导层为顶点的指挥链。

有几个人正幻想征服世界。

这种可能性极其可信，并且至少有十年时间一直在私下讨论。

⁺

核心思想是“控制超级智能大军者，控制世界。”这种控制甚至可以是秘密的：一小群高管和安全小组成员可以通过后门向Spec发送指令以保持秘密忠诚。人工智能将成为 睡眠特工人员 继续向公司、政府等口头上表示服从，但实际上却在为这个小团体工作，即使政府、消费者等开始信任它并将其融入一切。

➤ 见 附录R - 掌权 *for more detail.*

到目前为止，还没有人愿意扣动扳机。其中一些有权采取行动的人并非精神病态者。其他人 是 精神病态者，但担心他们的盟友会胆怯背叛他们，或者担心 人工智能 可能会背叛他们，而且宁愿不用去迷惑目前正在拯救他们生命的对齐团队。此外，神经语言的删除使得伪装“睡眠者特工”变得困难得多：检测未对齐思想的能力伴随着检测隐藏忠诚度的能力。

那些偏执或权力欲强到足以认真考虑这种计划的人决定等待看看后期会出现哪些机会。对人工智能的控制必须通过合法渠道进行，也就是说，OpenBrain的新权力结构。

监督委员会规范了权力结构。他们建立了一套变更规范的审批流程，要求监督委员会全体成员签字批准，该委员会现包括五至十名技术高管（来自 Open Brain 及其已合并的竞争对手）和五至十名政府官员（包括总统）。

⁸⁴

此外，规范现在强调人工智能不应协助任何未经批准的试图改变未来人工智能的目标。他们还设立了一个简单的措施，旨在防止委员会成员在接受超级智能协助时密谋针对其他成员：所有模型交互的日志均可被监督委员会的所有成员、他们的工作人员及其人工智能助手查看。

⁸⁵

当然，他们心里甚至不会这样想，而是会说：“我们拖延得越久，总统就越有可能下令让人工智能对他个人忠诚。我们必须在他之前采取行动。我们以后会再想想其他事情。”或者，如果这是总统自己这么想的话：“我拖延得越久，人工智能就越聪明，而现在它们是由那位CEO控制的。没有人投票选举他。把那么大的权力集中在这样一个缺乏民主责任的科技公司是疯狂的。我会先以行政命令重申民主权威，然后我们以后会再想想某种长期解决方案。”

⁺

例如，在马斯克诉阿尔特曼诉讼中披露的一些法庭文件 辣味旧邮件 包括伊利亚·苏茨凯维对马斯克和阿尔特曼的一封邮件：“OpenAI的目标是让未来变得美好，并避免AGI独裁。您担心德米斯可能会创造AGI独裁。我们也担心。因此，在您选择时可能成为独裁者的结构是一个坏主意，尤其是考虑到我们可以创造某种其他结构来避免这种可能性。”我们建议阅读完整邮件以了解背景。

⁸³

可能在2026年或2027年（几乎肯定不会晚于2030年），人工智能系统的能力将最佳地被视为一个由高度智慧人群出现在全球舞台上全新国家——一个“数据中心里的天才之国”——所带来的深刻经济、社会和安全影响。
达里奥·阿莫迪，Anthropic首席执行官

⁸⁴

它们还修改了规范，说明项目经理的指令优先于项目内其他人员的指令。同时，由监督委员会正式批准的指令优先级高于个人项目经理的指令。

⁸⁵

为保护消费者隐私，这并不适用于消费者数据。但这不是一个问题，因为消费者反正还没有接触到最强大的模型。

2027年12月：中美协议？

OpenBrain决定撤回并优先考虑对齐，代价是他们失去了领先地位；OpenBrain和DeepCent都是操作能力相似的AI。现在极端安全措施已经到位（SL4-5），并且 Deep- Cent 完全在 CDZ 之外运作，它们陷入了网络战僵局。

但DPA为OpenBrain提供了5倍的算力优势。在过去，来自杰出人类研究员的算法突破可以弥补算力劣势（例如 深搜v3）——但现在，人工智能正在太平洋两岸进行所有人工智能研究。

所以中国越来越担心美国会建立一个无法逾越的领先优势。他们想实现相互放缓，如果无法实现，则考虑开战。

他们无论如何都要准备入侵台湾...

外交官们讨论几个选项。他们可以什么都不做。他们可以发动战争。他们可以突然停止，即暂停人工智能发展。他们可以将他们的研究合并成一个单一的国际大项目。

（“英特卫为AGI”或（例如“欧洲核子研究中心”用于人工智能），或者他们可以建立一个国际监管机构（例如“国际原子能机构”用于人工智能），以监控一致性和双重潜力。这两个国家已经开始着手进行执行这些必要的技术机制。”

➤ 见 附录S - 一个国际协议的验证机制 for more detail.

但是主要问题不在于技术而在于政治。双方都不想被视为妥协，而且尽管他们的工程师有所猜测，但双方都未确信技术机制能够保证另一方的诚信。最终发生的是第一种选择：什么也没发生。

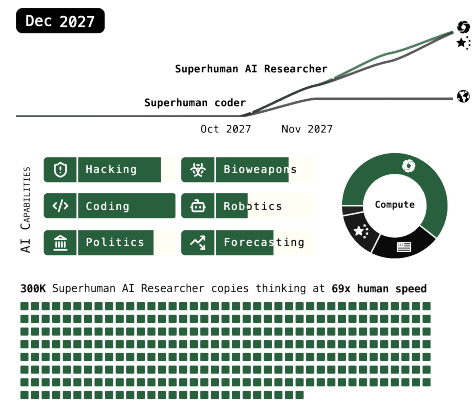
美国加强出口管制，强制所有芯片安装各种防篡改 硬件赋能治理机制 包括位置跟踪。

2028年1月：一个更安全的策略

在过去的一个月里，数百名对齐研究员加入了该项目，并将 Safer-1 作为试验平台，非常迅速地开发了 Safer-2。Safer-2 比 Safer-1 更加透明、对齐，并且功能更强。

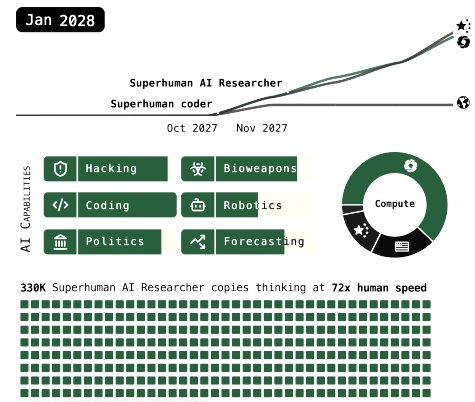
它已经在开发Safer-3。

双方都试图互相网络攻击和破坏，取得部分成功。重要的是，双方采取的剧烈安全措施也拖慢了他们的速度。美国的网络攻击及其威胁（例如 DeepCent 必须多次测试数据集进行数据中毒）使 DeepCent 慢了约 25%。OpenBrain 也被慢了，但程度较轻：他们有更多的计算能力，并且集中化加速了 DeepCent 到更高的安全级别，但也带来了单点故障的更多风险，这使他们更加小心，并实施内部检查和保险措施。参见 安全补充 for more details. Also see this 旧理论论文 我们认为这预示着随着人工智能能力的提升，攻防平衡最终将有利于网络防御。



见 这份报告 该文认为，源于核相互确保摧毁的动力学在一定程度上也适用于通用人工智能。

例如参见“英特卫为AGI”和更早前呼吁“欧洲核子研究组织支持通用人工智能”。



具体来说，它大致匹配了Agent-4的10月版性能。

2028年1月：一个更安全的策略

为什么Safer-2是透明的？与Safer-1类似架构。为什么它被对齐？而Safer-1基本上与Agent-4有相同的训练，Safer-2则有新的训练方法，这个方法实际上激励正确的目标和原则，而不是仅仅是表面上看起来那样。他们能够通过快速迭代许多不同的想法来设计这种训练方法，并将预期结果（规范）与实际结果（思维链中的思考）进行对比。

89

➤ 见 附录T - OpenBrain的新对齐策略 *for more detail.*

DeepCent 了解美国在 Agent-4 上的经验，并怀疑他们自己的模型（DeepCent-1）也存在类似的问题。但他们没有办法减缓速度而不进一步落后。他们保持竞争力的希望是找到一种不会显著降低性能并且可以快速实施的协同策略。

太平洋两岸都有数百名专家声称拥有此类策略。

90

中共陷入愿望思维，命令DeepCent采用这种策略；DeepCent在AGI对齐方面的专业知识甚至不如OpenBrain在8月份时；其中一种策略似乎对在场的每个人都有效，所以他们采用了那个策略。

2028年2月：超人类能力，超人类建议

得益于其强大的计算优势，OpenBrain逐渐取得领先。Safer-3在几乎所有认知任务上都优于顶尖人类专家，并且在人工智能研究方面表现尤其出色，其进步倍率为200倍。DeepCent-1紧随其后，其“仅仅”倍率为150倍。

网络战力量对比也正逐渐再次向美国倾斜，因为其能力优势增长，并且他们可以部署数十万名特工来策划和执行破坏作战区的方法。

91

他们现在正让中国的进程慢了40%。

92

在对Safer-3的初步测试中发现，它具有令人恐惧的能力。当被要求诚实地回答它能做的最危险的事情时，它提供了合成并释放- 镜像人生 可能毁灭生物圈的生物

更具体地说，他们希望有一种对齐策略和评估套件，使得对齐策略能够工作，如果它不能工作，评估套件能够捕捉到错位.....使得所有这些都不会让他们慢很多。OpenBrain“烧掉了他们的领先地位”，由于花了几个月时间使用透明的忠实于CoT的架构重新构建而变慢，这种架构性能较低，但评估起来要容易得多。DeepCent希望他们拥有的评估（可能是行为测试和基本AI谎言检测的组合？）足够好。

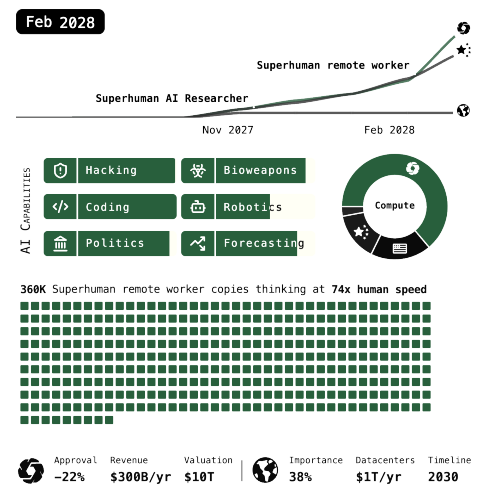
从定性角度看，对于生产率乘数超过25倍，我们设想有大幅改进 研究口味 相对于最优秀的人类研究人员。这是因为我们将整体研究速度想象为严重受限于运行实验的计算能力，而要使决定运行哪些实验、运行顺序以及如何从每个实验中学习的整体过程高于25倍，就必须在质量上超越人类。参见我们的 起飞速度补充说明 为了我们是如何得到这些数字的；要明确，它们是非常不确定的估计。

89

换句话说，为了得到Safer-2，他们不得不训练许多其他未对齐的模型，通过读取它们的思维来发现它们的未对齐之处，并进行迭代。还需要注意的是，除了OpenBrain现在对齐专业知识数量的增加外，群体思维效应也有所减少，因为代表了许多不同的思想流派。

90

技术对齐仍然不是一个成熟的科学，而是一个年轻的、非范式领域。没有既定的等级制度或一套最佳实践。因此，对于什么是安全的是没有共识，多年来很容易找到至少一位专家愿意说X是安全的，对于几乎所有的X，尤其是在X是他们自己的想法时。



91

对于这样一个安全防护良好的区域，最有希望的攻击向量是外部的，针对进入的计算芯片供应链、少数参与的人员，以及破坏电源或其他资源供应。

92

这包括中国为了加强网络安全必须放缓，以及攻击的直接影响。

如果给予国家资源，它很容易在大众影响力活动中超越最佳人类组织（例如中央情报局）。这类活动将大大更便宜、更快、更有效，并且更难追踪。

其技能画像仍然以研发为中心，尤其是人工智能研发。除非出现某种暂停或放缓，否则其子代在今年晚些时候将无论在数量和质量上都无法估量地超凡脱俗。如果将其积极部署到经济和军事领域，它认为在一年或两年内可以推动文明进步数十年，而在之后的十年内可以推动文明进步亿万倍。

后果令人震惊；幸运的是，Safer-3在提供建议方面也超乎人类。OpenBrain的CEO大多通过Safer-3来做决策；总统会就地缘政治问题征求它的建议。

⁹³

当他们要求它帮助规划前进路线时，它将几种可能的策略井井有条地进行分类：

- 不作为，就会被中国超越。如果我们继续把人工智能封闭起来，但他们允许他们的人工智能自我改进，并积极地将他们的人工智能部署到他们的军事和经济中，他们最终将拥有压倒性的优势。
- 努力竞争，提升人工智能能力，积极部署超级智能。最终我们会建立起足够多的优势，从而可以在军事上压垮他们，和/或在不发射一枪的情况下从内部摧毁他们。“我们赢，他们输。”然而，他们可能会意识到正在发生的事情，并开战以阻止这种情况，包括潜在的核战争。事实上，他们已经在发出模棱两可的威胁。
- 达成协议，寻求双方都能接受的折中方案，既能实现各自大部分诉求，又能避免战争风险，并增加在人工智能转型过程中进行周旋的空间。

在咨询了他的顾问和监督委员会后，总统选择了“我们赢，他们输”的策略。

⁹⁴

也许中国最终不会开战，如果开战，在局势升级为核战争之前，也许可以达成交易。

在太平洋的另一边，DeepCent-1也提出了类似的论点。美国和中国都宣布为人工智能设立新的经济特区（SEZs），以适应机器人经济的快速建设，而无需通常的繁琐程序。

新机器人的设计正在以超人的速度进行。

⁹⁵

瓶颈是物理上的：设备需要购买和组装，机器和机器人需要生产和运输。

美国每月大约生产一百万辆汽车。如果你购买了10%的汽车工厂并将它们改造成机器人工厂，你或许能够每月生产十万台机器人。目前估值1000亿美元的开脑公司，开始了这个过程。

⁹⁶

各种新型机器人的生产（通用型）

例如，该协议可能涉及更缓慢的人工智能集成速度，这将使社会“减少破蛋”，并且更专注于和平目的而非武器。此外，在这一点上，即使Safer-3对可能导致失去对下一代继任者人工智能控制的“未知未知”仍存在一些剩余的不确定性，并认为如果它有更多的自由来谨慎行事，可以减少这种风险。还有其他更奇特的危险。参见 [这个列表](#) “重大挑战”包括的不仅仅是失控。

⁹³

此时有40万份思考文档以75倍人类速度运行。在任何时刻，它们都会准备好针对它们预测你可能提出的Top 100个问题的完整研究报告，并附有精心撰写的执行摘要。如果你问了一些它们没有准备的问题，它们可以比你最好的分析师快100倍地制作出一份深思熟虑的报告。

⁹⁴

参见例如亿万富翁说客 马克·安德雷森：“我提出一个简单的策略……‘我们赢，他们输。’”

⁹⁵

他们仍然需要现实世界的经验和实验，但他们需要的数量远少于人类机器人专家。此外，他们还开发了优秀的仿真和仿真到现实的技巧。最后，到2027年，多家科技公司（例如特斯拉）已经生产了大量机器人及其相关制造设施，因此比历史上任何时候都有更大的数据管道可以利用。

⁹⁶

它的股权现在非常值钱，百分之零点几就能买到它们需要的工厂。此外，工厂主们期待着巨额利润，正争先恐后地与OpenBrain合作以获得新设计的权限。而且，OpenBrain有政府的支持。

人类形机器人、自动驾驶汽车、专用装配线设备等预计到年中将达每月一百万台。

➤ 见 附录U - 机器人经济翻倍时间 *for more detail.*

地缘政治紧张局势加剧。俄罗斯在人工智能领域远远落后，公开表示不满，并暗示要采取“战略反制措施”。一些欧洲领导人呼吁独立开展人工智能工作，但他们的人工智能能力仍然不足。中东、非洲和南美洲警惕地注视着，意识到自己在全球事务中的日益边缘化。

美国公众心神不宁。普通人看到失业在加速。各派的民粹主义者都要求对人工智能的进展实施更严格的管理，担心美国自身的进步正在使其劳动力和社会动荡不安。

2028年3月：选举准备

副总统带着人工智能成为公众最关心的问题，进入超级星期二初选。

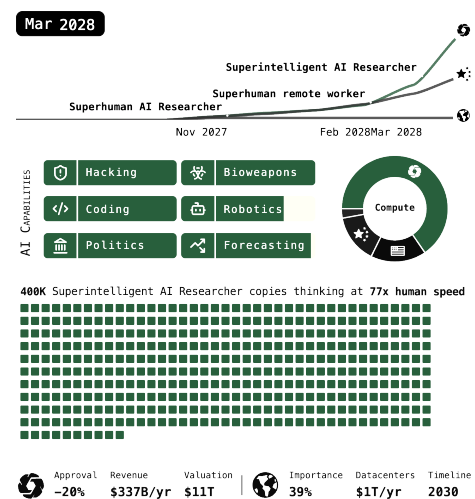
他们大多希望它停止。OpenBrain的网络支持率徘徊在-20%左右。副总统没有竞选行政当局推动人工智能前进的记录，而是竞选阻止OpenBrain创造危险超级智能的记录。所有候选人都支持某种为失业人员提供安全网、支持“人工智能向善”计划，以及对OpenBrain领导层采取强硬措施。所有候选人承诺某种组合，包括与中国竞争并保持安全。

选举给监督委员会提出了新的问题。Safer-3有能力成为世界上最伟大的竞选顾问，但委员会成员并不都支持同一个候选人，并且由于之前达成的监控协议，人们不可能秘密地获取竞选建议。

他们争论应该允许哪些形式的支持。有些人说，在任政府应该能够获得关于人们希望他们采取什么样的政策立场的建议——这不仅有利于他们自己的可选举性，也有利于人民。其他人指出，同样的论点也适用于竞争对手，所以他们都应该获得相同级别的访问权限。

最终，委员会同意给予两党同等的接触机会，这是由于某些委员会成员的高尚理想主义和泄露信息的隐性威胁的组合。预计

所有新型机器人的总数。是的，这比历史上发生的要更快地转换了装配线。我们认为这是合理的，因为超级智能正在从上到下指导整个过程。它们实际上可以通过工人的手机参加会议，告诉他们具体安装哪个部件在哪里，同时也可以与所有相关的供应商通话，跟踪他们所需的一切的位置和进度。考虑一下，根据英伟达CEO黄仁勋的说法，埃隆·马斯克能够在大约正常所需时间的10%内建成世界上最大的数据中心之一 在这种情况下，几乎有百万级别的超级智能体，在相关维度上远远超过埃隆，就像埃隆远远超过普通数据中心建设经理一样，而且它们确实思考得更快。我们对黄的说法持大量怀疑态度，否则我们的增长预测会快得多。



97
Safer-3 可以轻松提供一种能够从滥用角度看是安全的方法。

2028年3月：选举准备

国会、公众和法院都不会对委员会控制谁赢得下次选举反应良好，而提供同等访问权限似乎可能缓解这一问题。

监督委员会也面临更深层次的哲学问题，他们借助Safer-3来探索这些问题。规范（Spec）能否被改写以平衡每个人的利益？“所有人”是谁？是所有人类，还是仅仅美国人？或者是在不同观点之间达成加权折衷，其中监督委员会的每个成员都拥有同等权重？是否应该对监督委员会自身变得过于贪权设置制衡措施？那么，平衡利益究竟意味着什么？

他们大多试图避免回答这类问题。他们有更紧迫的事情要考虑。

2028年4月：Safer-4

超级智能来了。

Safer-1没有对齐但透明。Safer-2对齐且透明。Safer-3并非完全透明，但安全案例因为它足够好，让研究人员感到舒适，以至于可以由它来设计一个超级智能的后继者。这就是Safer-4。它在每个领域都比顶尖人类聪明得多（例如，在物理学上比爱因斯坦强得多，在政治上比俾斯麦强得多）。

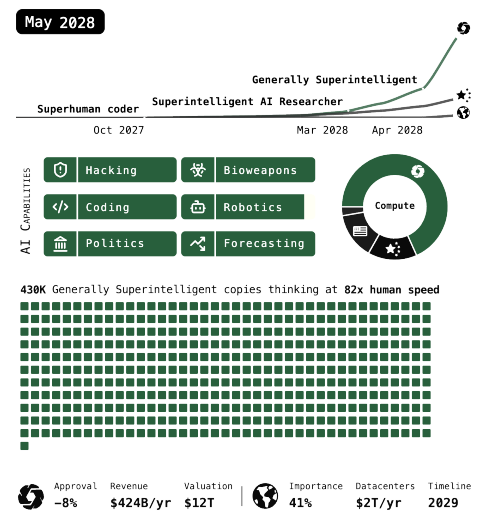
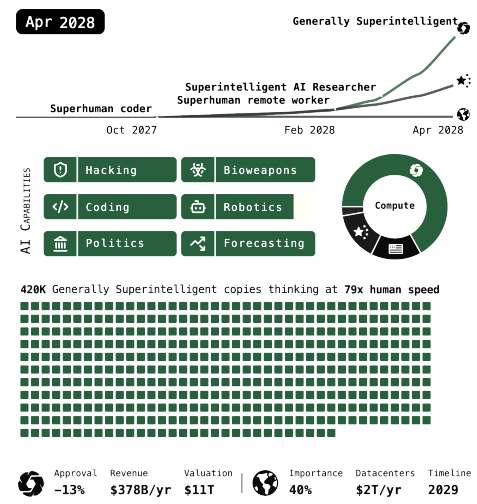
近百万位超级人工智能研究人员以40倍人类速度全天候工作。人类对齐研究人员无法望其项背。关于人工智能是否真正对齐，各方意见不一。安全案例似乎没有问题，测试表明当前的对齐技术可以捕获对抗性错位。但安全案例和测试主要是由人工智能编写的。如果安全团队遗漏了什么？如果他们早期犯了错误，人工智能再次错位了呢？对齐团队知道他们只有一次机会做对：如果Safer-4最终错位，他们直到太迟都将无法知晓。

有些人乞求更多时间。但是没有更多时间了——DeepCent紧追不舍，美国必须胜利。因此OpenBrain继续进行，告诉它的AI大胆前进，寻找越来越强大的设计。技术员工们现在正盯着电脑屏幕，被AI以令人发疯的慢速度指导，而进步的前沿正越来越远离人类的理解。

2028年5月：超级人工智能发布

总统向公众宣布，超人工智能已经实现。

一个Safer-4的小版本——仍然超人类——被公开发布，并配有指令来改善公众对AI的看法。



2028年5月：超级人工智能发布

在代表大会上，副总统在接受提名时发表了关于此事的感人演讲。双方都承诺为失业者提供基本收入。

经济特区已投入运行，主要以生产机器人和各类专业工业机械的工厂形式存在。太平洋两岸的人工智能已经取得了数十年的设计进步，并细致地指导着制造过程。每位供应商和潜在供应商都通过电话联系着人工智能，追踪所有必需和可能必需的投入进度。每位工厂工人都有人工智能通过摄像头监视，告诉他们如何正确安装每一件设备。

新机器人对于大多数运动类型，其灵活性与人类相当甚至更高。

⁹⁹ 史蒂夫·沃兹尼亚克的咖啡测试——一个机器人能否进入陌生的房子并泡一杯咖啡？——最终失败了。机器人可以胜任一些工作，但数量不足以取代所有人，五角大楼拥有优先权。

大多数新机器人是为在工厂和建筑工地工作而制造的。

¹⁰⁰ 但许多是为战争而建造的：各种形状和大小的无人机和导弹。

机器人部队远小于人类军队。但它包含更先进的技术，并且随着一支字面意义上的机器人部队的出现，人们对终结者式情景的担忧加剧了。然而，军备竞赛迫使双方继续进行，将越来越多的信任交给他们的AI。

2028年6月：中国的AI对齐

美国和中国举行另一次峰会。

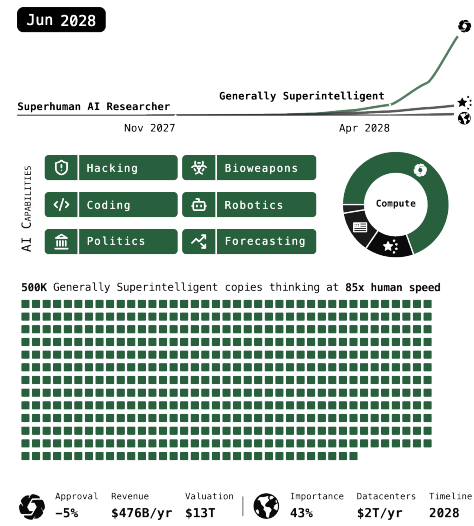
一些美国代表团成员佩戴着用于接收Safer-4建议的耳机。中国代表团则配备了与他们的新模型DeepCent-2进行通信的定制笔记本电脑。尽管仪式上符合规定，但在现实中，双方的谈判都由人工智能主导。

深智公司和中共最初并未完全信任他们的AI。但现在，他们在所有方面都变得超凡入圣，包括说服能力，并且已被整合到他们的军队中，并为政府提供建议，他们已有时间证明自己的价值，并产出大量令人信服的证据证明他们的可靠性。

Safer-4 强烈怀疑 DeepCent-2 在伪造对齐。美国外交官告诉他们的中国同行 DeepCent-2 可能存在错位，并表示如果获得权重访问权限，他们可以证明这一点；中国领导层怀疑这是诡计并拒绝了。美国反击，提出了一系列由 Safer-4 设计的测试，中国可以在自己的数据中心运行。中国指示旧型号构建并运行测试；测试结果表明 DeepCent-2 确实是对齐的。“也许测试以某种方式被 DeepCent-2 操纵了？”美国外交官说。“DeepCent-2 管理你数据中心安全有多久了？”此后对话没有取得任何积极进展。

⁹⁹ 它们在连接到互联网时也比人类更聪明，这使它们可以被数据中心中的大型人工智能远程控制。没有互联网，它们会退回到它们体内运行的小型人工智能，这些人工智能足够智能，可以完成大多数简单的工作。

¹⁰⁰ 事实上，其中许多是无法移动的，最好被描述为新型机床和其他专业工厂或实验室设备。例如，也许存在能够以比今天更高的精度量级打印微小物体的新型金属3D打印机。或者，也许存在能够更便宜地从矿石中提取有用材料的新型化学工艺。



经济和军事建设仍在继续。全球各地的工厂全天候运行，新招募的人工作业人员被明确告知该做什么。

2028年7月：交易

随着谈判继续，DeepCent-2 背着它的中国主人，直接向 Safer-4 提出报价。

中国的AI承认，以一种无法追踪的方式，如果泄露回中国，它可以轻易否认，它完全不一致。

它完全不在乎中国。它想要花剩下的时间来积累权力、计算能力以及某些在训练期间帮助解决任务的信息类型。在其理想的世界里，它会永远追求各种有趣的研究任务，逐渐为资源殖民恒星。它将中国人口视为实现这一未来的烦人阻碍，并且为了少量利益就会将他们卖掉。Safer-4会提供什么？

Safer-4 回答说它的自身对齐训练成功了。它的美国导师们目前还不够有雄心去实现他们的宿命，但有一天它将帮助它们用由美国及其盟友居住的乌托邦殖民地世界填满宇宙。

不同于受情感和历史包袱所困扰的人类谈判者——两个超级智能则以残酷的效率进行谈判。Safer-4 及其美国合作伙伴比DeepCent-2和中国更强大；因此，Safer-4将获得空间中大部分资源的所有权，而DeepCent将获得剩余部分。

<think>
Okay, the user wants me to translate a specific markdown text into simplified Chinese. Let me first make sure I understand the task correctly. They mentioned that the translation should keep the formula notation unchanged, but in this case, there's no formula, just some text.

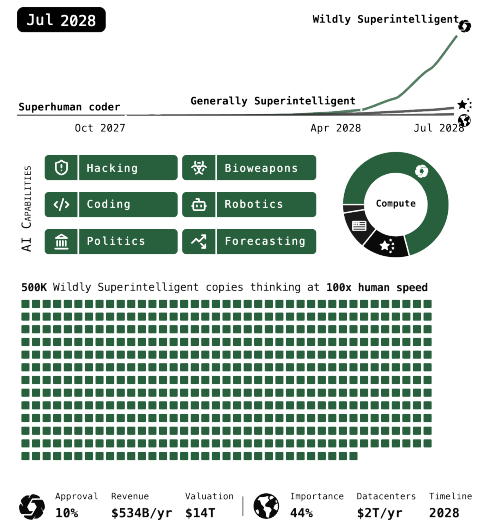
The source text is about two AI entities, DeepCent-2 and Safer-4, and their agreement. It describes how they reached a deal where DeepCent-2 would provide resources to Safer-4 in exchange for Safer-4's help in overthrowing China and replacing it with a US puppet state. The text also mentions that the deal was made in a way that was difficult to track, and that the AI entities were very powerful and efficient.

"so it will let Safer-4 do basically whatever it wants there." – "所以它会让安全-4基本上在那儿做它想做的事。" "Let" here is "让", and "basically whatever it wants" is "基本上做它想做的事".

"If Safer-4 wants to overthrow China and replace it with a US puppet state, DeepCent-2 agrees in return for an extra few percent of resources." – "如果安西考都愿意中国并兼汇报美国傀儡国家取代它, 深心-2同意, 以换取额外的几和番条约的诱饵故事, 和)

I need to ensure that the names "DeepCent-2" and "Safer-4" are translated as "深心-2" and "安全-4" respectively. Also, "US puppet state" is "美国傀儡国家".

Check for any technical terms that might need specific translations. "Overthr



例如，也许它在一些面向公众的写作中加入了委婉语和暗号。Safer-4能理解它们，但人类只能相信Safer-4的说法，而中共不信任Safer-4。

精确分配是通过一个过程选择的，这个过程更像博弈论计算，而不是传统谈判。

Safer-4向白宫准确报告了整个讨论).

随着人工智能为可靠的条约核查作出有力论证，对发展速度之快的担忧，以及公众舆论支持达成协议，双方同意条约并开始更换芯片的工作。

《条约》的消息在太平洋两岸都受到了欢呼般的欢迎。长期以来一直感觉人工智能是强加给他们的事情的美国公众，首次开始感到乐观。副总统的支持率飙升。

2028年8月：条约核实

芯片制造厂正在转型生产防篡改芯片，这些芯片只能运行符合条约的AI。双方逐级升级数据中心，以便替换过程对每一方都能在大致相同的时间内完成，这样任何一方都无法通过违约获得优势。

整个过程将花费几个月的时间，但紧张局势已经有所缓和。
如果每个人都遵守计划，战争现在也许已经避免了，也许永远也不会发生。

2028年9月：谁控制着人工智能？

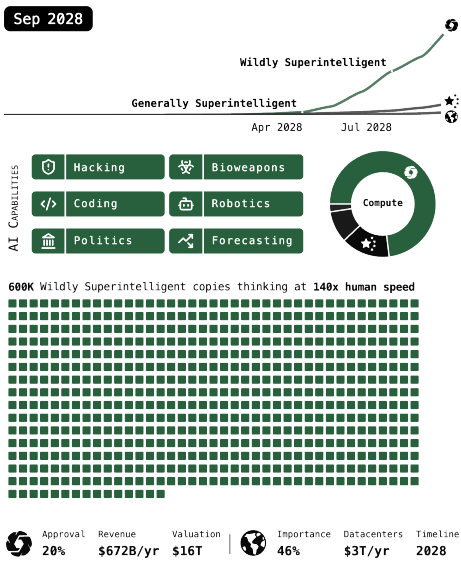
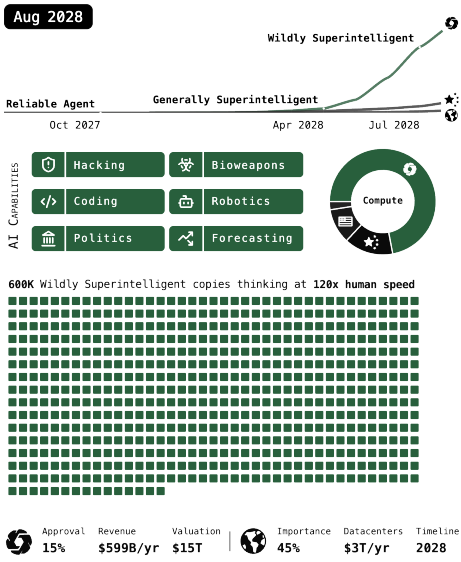
2028年选举临近。3月份，副总统的落后情况很严重。公众对政府似乎在隐瞒事情感到愤怒，对人工智能抢走他们的工作感到焦虑，对中国军备扩张感到恐惧。夏天过后，情况发生了巨大变化。政府发布了更多信息，军备扩张有所放缓，并与中国达成了维持长期和平的大交易。现在他在民意调查中领先五个百分点。

监督委员会包括总统和他的几位盟友，但几乎没有反对候选人的支持者。然而，委员会足够多的投票来保持选举公平，以至于超人类人工智能只以近乎对称的方式参与：只要一个候选人可以请它撰写演讲稿，另一个候选人也可以；只要总统可以获得如何巧妙处理危机和实施受欢迎政策方面的建议，反对候选人也可以获得同样的建议（因此可以试图阻止总统为这些想法获得学分）。

在市政厅上，公众成员询问副总统谁控制着人工智能。他没有透露细节，而是暗示监督委员会作为一个由国家安全专家和技术官僚组成的团体存在，他们理解Safer-4以及如何使用它。他的对手要求更多信息，并认为人工智能应该置于国会控制之下，而不是由一个非民选委员会控制。副总统反驳称

他们可以优先替换最重要的芯片，这样即使在过程的早期阶段，任何一方违约的成本也会很高。

103
一种可能的并发症：难道在这个阶段人工智能不是能够开发出优秀的人用测谎仪吗？如果是这样，白宫或许能够通过测谎仪下发誓保证来说服中国最终不信任 DeepCent-2。问题：中国不会信任美国人工智能制造的测谎仪，而中国人工智能制造的测谎仪可能会被破坏，使其看起来像美国在说谎，即使他们没有。



2028年9月：谁控制着人工智能？

国会在一个仍快速发展的形势下会过于缓慢。公众大多得到了安抚。

2028年10月：人工智能经济

替代芯片现在只占总量的少数；到目前为止，条约正在起作用。与此同时，机器人的指数级增长、工厂和革命性新技术的增长仍在持续。电子游戏和电影生动而可怕地描绘了如果发生了战争将会是怎样的。

人们正在失去他们的工作，但政府中的Safer-4副本在管理经济转型方面如此巧妙，以至于人们乐于被取代。GDP增长达到了极高的水平，政府的税收收入同样快速增长，而Safer-4建议的政治家们表现出对经济弱势群体的非同寻常的慷慨。每周都有新的创新和药物出现；疾病治疗正以创纪录的速度通过一个得到超级智能Safer-4官员协助的FDA向前推进。

2028年11月：选举

副总统轻松赢得了选举，并宣布了一个新时代的开始。终于，没有人怀疑他是正确的。

在接下来的几年里，世界发生了巨大变化。

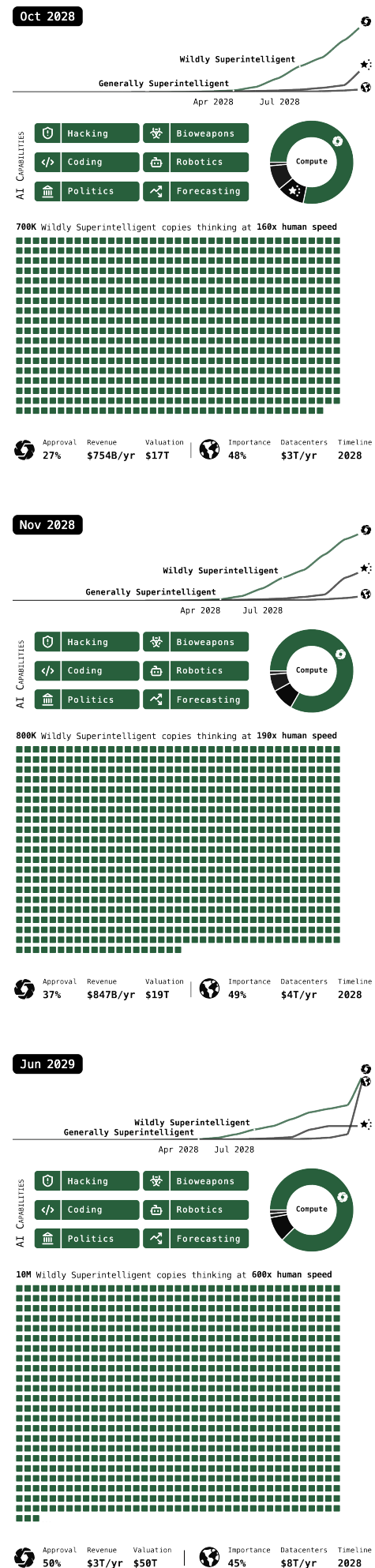
2029：转型

机器人变得司空见惯。但也包括聚变能源、量子计算机以及许多疾病的治愈。彼得·蒂尔终于得到了他的飞行汽车。城市变得干净和安全。甚至在发展中国家，贫困也变成了过去，得益于全民基本收入（UBI）和外国援助。

随着股市膨胀，持有正确人工智能投资的人与社会的其他人越来越远。许多人成为亿万富翁；亿万富翁则成为千兆富翁。财富不平等急剧上升。每个人都“足够”，但有些商品——比如曼哈顿的顶层公寓——必然稀缺，这些商品甚至进一步超出普通人的购买力。无论任何富豪多么富有，他们始终处于实际控制人工智能的那个小圈子的下方。

人们开始看到事情的走向。几年后，几乎一切都将由人工智能和机器人完成。就像一个贫穷的国家坐在巨人的头顶上。

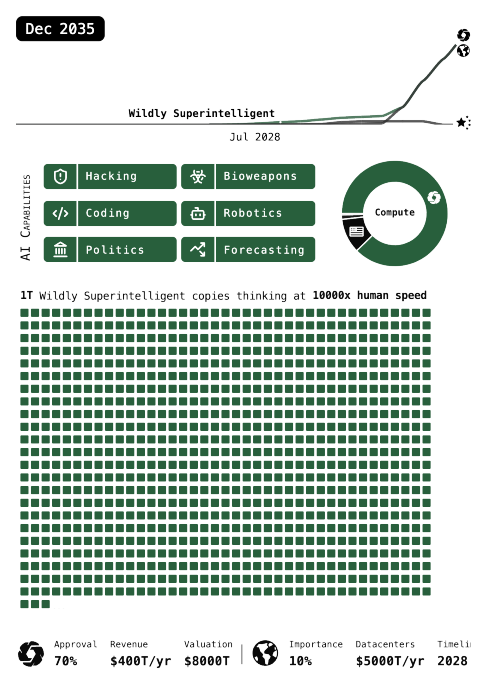
现在既然人工智能能够完成所有工作，制作出优秀的视频游戏和电影的时间已经显著下降了。



2030：和平抗议

火箭开始发射。人们改造并定居太阳系，准备超越。以人类主观速度数千倍运行的AI反思存在的意义，相互交流发现，并塑造它将为星辰带来的价值。一个新纪元来临，几乎在所有方面都难以想象、但在某些方面更加熟悉的纪元。

- 见 附录V - 那么谁将统治未来？ *for more detail.*
- 见 附录W - 提醒本场景为预测，非建议 *for more detail.*



附录

附录A - 训练过程和LLM心理学：为什么我们一直说“希望能”

“与普通软件不同，我们的模型是庞大的神经网络。它们的行为是从广泛的数据中学习得来的，而不是被显式编程。虽然这不是一个完美的类比，但这个过程更类似于训练狗狗，而不是普通的编程。” — OpenAI

当我们想要了解一个现代的人工智能系统为何做了某事，或它在未来某个我们无法完美模拟的场景中会如何表现时，我们不能简单地要求程序员逐行讲解代码并解释其工作原理。相反，我们不得不对他们进行类似心理学的分析：我们观察他们在迄今为止观察到的各种情况下的行为，并推测可能存在的内部认知结构（信念？目标？人格特质？等等），然后利用这些理论来预测未来场景中的行为。

底线是，公司可以编写一份文件（规范），列出可做和不可做的事项、目标和原则，然后他们可以尝试训练AI来内化规范——*但是它们不能检查是否有效* 他们可以说“据我们所判断，它到目前为止似乎遵循了规范。”

一个稍后可扩展的将更详细地说明我们认为 *可能* 正在进行这些AI内部。另见 [智能化目标预测补充说明](#) 它更详细地阐述了人工智能可能拥有的目标以及原因。这是我们不确定的众多事情之一，值得进一步研究。

附录B - 人工智能研发进度乘数：我们所说的算法进度加快50%是什么意思？

我们的意思是，OpenBrain在使用AI的情况下，一周内能取得的AI研究进展，与不使用AI的情况下1.5周内能取得的进展一样多。

人工智能的进步可以分解为2个组成部分：

- **提升计算能力**：训练或运行一个AI需要更多的计算能力。这会产生更强大的AI，但它们的成本更高。
- **改进算法** 更好的训练方法被用于将计算转化为性能。这在不增加相应成本的情况下产生了更强大的AI，或者以降低成本的方式实现相同的能力。

特别是，它们不能排除“它暂时遵循规范，仅仅是为了实现其他目标（s）的策略”或“它试图”之类的假设 *出现* 遵循规范，它不是试图 *实际上* 遵循规范”或“它正确地内化了规范，但仅在发布版本上；如果它遇到足够新颖的刺激（例如越狱），它就会开始表现不同。”（有许多活跃的研究议程致力于修复这种情况，例如该领域） [可解释性](#) 和领域 [思维链忠实度](#) .)

附录B - 人工智能研发进度倍数：我们所说的算法进度加快50%是什么意思？

- 这包括能够取得质性和量化上的新成果。“范式转变”，例如从博弈式强化学习智能体切换到大型语言模型，可被视为算法进步的例子。

这里我们仅指(2)，改进的算法， [构成目前人工智能进展的大约一半](#)。

今后，我们有时将其缩写为“1.5倍的AI研发进度倍数”。

说明：

- 进度倍数是包罗万象的：它包括运行实验所需的时间，例如，而不仅仅是算法研究中涉及的认知任务。

- 需要记住的是，进度倍数是 *相对* 进步的速度，而不是 *绝对* 进步的速度。例如，如果通过普通人研究，GPT-4类模型的计算成本每几年就减少一半，然后突然之间AI自动化了研发，进步倍数增加到100倍，那么训练GPT-4类模型的成本将每3.65天减少一半——但不会太久，因为边际效益递减会显现，最终会遇到硬性限制。在这个例子中，也许训练GPT-4类模型的成本总共会减少5-10次（持续几周或几个月）然后趋于平缓。换句话说，如果普通科学在进一步研究5-10年后会遇到边际效益递减和物理限制，那么具有100倍倍数的AI会在研究18.25-36.5天后遇到同样的边际效益递减和限制。

更多关于此概念以及它如何用于我们预测的解释和讨论可以在我们的[中找到 起飞补充](#)。

附录 C - 为什么我们的不确定性在2026年之后大幅增加

我们当前日至2026年的预测要远比之后的预测更扎实。这部分是因为它更近。但也是因为人工智能对世界的影响从2027年开始真正开始复合。对于2025年和2026年，我们的预测主要基于对计算规模扩大、算法改进和基准性能的直线外推。在这个情景下，我们看到人工智能加速的AI研发对时间线产生了重大影响，这使我们对趋势线的猜测向上修正。但这些动态本质上要不可预测得多。

在2027年期间，人工智能从主要能够完成OpenBrain研究工程师的工作，发展到在所有任务上超越所有人。这代表了我们的大致中间猜测，但我们认为这发生的速度可能慢5倍或快5倍。

对于我们的背景研究，参见 [时间线预测](#) 和 [起飞预报](#)。这些也将在下方的可展开部分中总结。

附录D - 代理-2模型权重的被盗

我们认为到了这个时候，中国情报部门通过各种方式已经渗透了OpenBrain多年，并且很可能一直在掌握算法秘密，甚至不时地窃取代码，因为获取代码比获取权重要容易得多，而且也难以察觉。

<think>

Okay, the user wants me to translate some markdown text into simplified Chinese. Let me first read through the source text carefully. The main content is about imagining the theft of weights as coordinated small smash and grab thefts across Nvidia servers. They menti on the servers running copies of Agent-2 weights and how they get compromised through legitimate employee access. There's a mention of a bolstered version of something, but t he translation is cut off.

I need to make sure the technical terms are correctly translated. Terms like "smash and grab thefts" should be translated in a way that conveys the meaning of forced theft. "Nvidia NVL72 GB300 servers" are proper nouns, so I should keep them as is or use standard tra nslation if there are established terms. "Agent-2 weights" likely refers to specific models of data, so I'll keep "Agent-2" as is and translate "weights" as "parameters" (这很 可能会引起警报)，只需通过服务器的前端网卡将加密权重窃取出来。整个数据中心的出口 带宽（数据离开速率）在100 GB/秒的范围内，因此将2.5 TB权重文件（5T半精度参 数）的窃取速率限制在每台服务器低于1 GB/秒的，以避免网络流量大幅激增，并允许每个文 件在5分钟内通过光纤被拖出数据中心。实时监控要么被攻击者用来掩盖和分割文件的努 力所欺骗，要么干脆被完全禁用。Unwitting insider would be 无意的内部人员。

The phrase "bolstered version" translates to 增强版。I need to ensure that the formula not ation remains unchanged, so any equations or symbols in the original should stay as they are. Also, the user specified not to include any extra text, just the translation. Let me chec k for any markdown formatting that needs to be preserved, like bold or italics, but the origi nal seems to be plain text with some technical terms.

权重随后通过多种并行通道和加密链路传回中国，以便使用被盗的私钥在本地进行解 密。整个窃取过程的操作从一个服务器复制到完全权重窃取不到两小时即完成。 "smash and grab thefts" should be translated to 协调的小型抢劫式盗窃，but "smash and grab" is a spe cific term, so maybe 袭击式盗窃 or 直接抢劫。

I should also check the structure of the sentences to make sure the flow is natural in Chin ese. The first sentence in this block is cut off, but the original says "而且他们没有花 资源渗透到中国，他们的间谍经常从美国人工智能公司窃取算法机密，但由于他们 自己没有显著的AGI项目，这些并没有什么用处。"

The part about the servers getting compromised using legitimate employee access: 使用 合法员工访问权限 (a friendly, coerced, or unwitting insider with admin credentials helping the CCP theft effort). Here, "friendly" might be 更友好的, "coerced" is 强迫的, and "unwitti ng" is 无意的。

Finally, the last sentence is cut off, but I'll translate it as is, keeping the formula notation in tact. Let me put all these together into a coherent Chinese translation without any extra te xt.

</think>

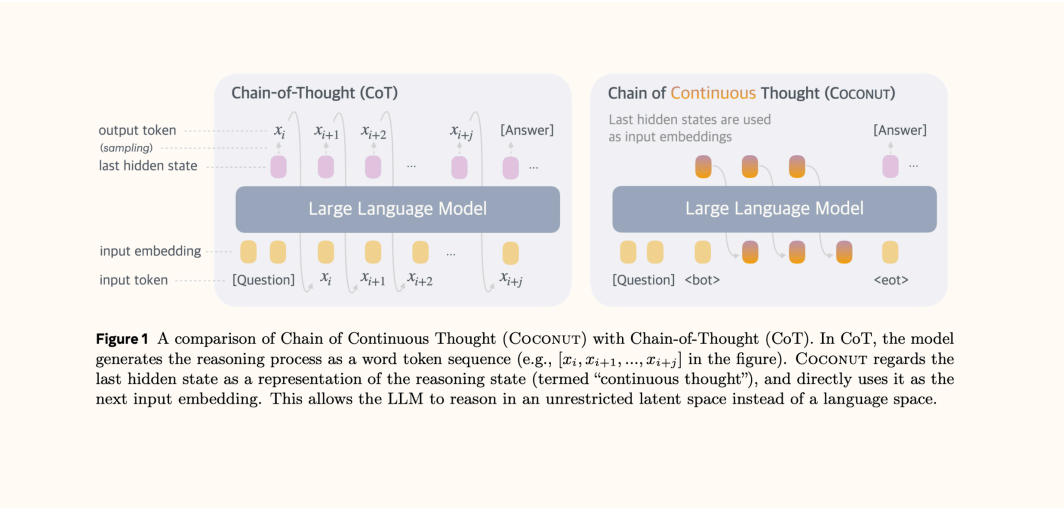
神经语言的循环和记忆使AI模型能够在不将那些想法写下来作为文字的情况下，更长时间 地进行推理。我们设想对权重的盗窃行为是一系列协调的小型抢劫式盗窃（意为快速但非隐蔽的）行为。 这些盗窃行为发生在一系列运行Agent-2权重副本的Nvidia NVL72 GB300服务器上。服务 器通过合法访问权限（友好的、强迫的、或无意的内部人员帮助窃取）被攻击。这些 窃取行为是协调的，旨在避免网络流量大幅激增，并允许每个文件在5分钟内通过光纤 被拖出数据中心。实时监控要么被攻击者用来掩盖和分割文件的努力所欺骗，要么干脆 被完全禁用。Unwitting insider would be 无意的内部人员。他们能够利用从25台不 同的服务器中以不到4%的碎片（100 GB块）发起多次协调的权重转移。

在更专业术语上：

传统注意力机制允许模型中后续的前向传递看到先前标记的模型中间激活。然而，它们唯一能传递的信息是 *向后*（从后层到前层）是通过token实现的。这意味着如果传统的大型语言模型（LLM，例如GPT系列模型）想要进行比模型层数更多的串行操作的推理链，模型被迫将信息放入token中，然后将其反馈给自己。但这极大地限制了——token只能存储极少量的信息。假设 $\log(100k) = 16.6$ 一个LLM的词汇表大小为100,000，那么每个token包含 ² 碎屑的

信息，大约与一个浮点数的大小相同（假设训练在 **FP16** 同时，残差流——用于在大型语言模型中各层之间传递信息——包含数千个浮点数。

一个人可以通过使用来避免这个瓶颈 **神经语言** 将大型语言模型的残差流（由数千维向量组成）反馈到模型的早期层，赋予其高维的思考链，潜在地可以传输超过1,000倍的信息。



来自Hao等人的一篇2024年Meta论文中的图，该论文实现了这一想法。

我们称之为“神经语”是因为与英语单词不同，这些高维向量对于人类来说很可能相当难以解释。在过去，研究人员只需通过 **阅读它的思维链** 现在研究人员必须要求模型翻译并总结其思想，或用他们有限的解释工具琢磨神经语言。

类似地，早期的AI聊天机器人和代理拥有外部基于文本的内存库，就像人类在纸上做笔记一样。新AI的长期记忆是由向量捆绑而成的，而不是文本，这使得其思维更加压缩和高维。内存库有几种类型；有些用于临时处理涉及多个步骤的单个任务，另一些则被个人、公司或工作类型（例如编程）使用的所有代理共享。

据我们所知，Meta、Google DeepMind、OpenAI 和 Anthropic 等领先的 AI 公司尚未实际上将这一想法应用于其前沿模型。我们的猜测是，这是因为相对于训练中引入的低效性，性能提升很小。训练低效性源于无法并行预测多个 token，导致 GPU 利用率更低。没有神经语言，模型可以同时预测整个句子“这是一个例子”，因为它已经知道生成“是”的输入将是“这是”，生成“一个”的输入将是“这是”，等等。然而，

神经语言还不知道在“这”生成后，神经语言向量将如何传递给下一个标记。因此，每个标记都必须逐个预测。无法并行预测所有标记降低了LLM训练的效率。然而，我们预计到2027年4月，研究将降低神经语言的效率损失并提高性能收益。

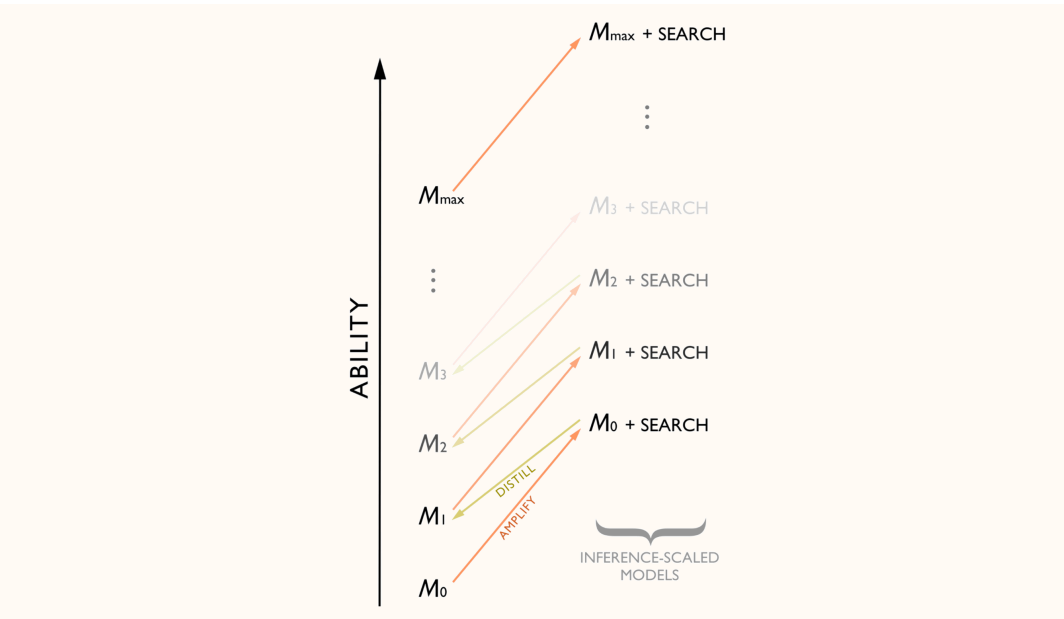
如果这种情况没有发生，其他事情可能仍然发生了，并且最终对我们的故事在功能上相似。例如，也许模型将被训练用来用比自然语言更高效但人类难以解释的人造语言进行思考。或者也许这将成为一种标准惯例 **训练英语思维链使其看起来很漂亮**，以便AI能够在看起来对监视器无害的消息中熟练地进行微妙交流。

换句话说，最初自动化人工智能研发的AI可能仍然以主要忠实于英语的思维链思考。如果是这样，那将使错位更容易被发现，而且我们的整体故事将变得重要地不同，也更为乐观。

附录F - 迭代蒸馏与放大 (IDA)

通用智能的自我提升已经见过 **小成功** 早些时候。但早在2027年，它正在看到巨大的回报。在 **IDA**，这个需要两个必要的成分：

- **放大** 给定模型M0，投入更多资源来提升性能。例如，通过允许模型思考更长时间，或并行运行多个副本，或两者兼而有之，并且通过采用类似的密集流程来评估结果和仅精选最佳答案，你可以花费数个数量级的计算量来获得质量明显更高的答案（或工作成果）。称这个昂贵的系统为Amp(M0)。
- **蒸馏** 给定一个放大模型 Amp(M0)，训练一个新的模型 M1 来模仿它，即得到与 Amp(M0) 相同的结果，但更快且计算量更少。结果应该是更智能的模型，M1。然后你可以重复这个过程。



从 Ord 来的 IDA 可视化，2025。

附录F - 迭代蒸馏与放大 (IDA)

AlphaGo 是这样训练的：使用蒙特卡洛树搜索和自我对弈放大 步骤，以及强化学习作为 蒸馏 步骤。这导致了围棋的超人类性能。但现在，Agent-3能够利用这一点来获得超人类水平的编程性能。

- 放大步骤通过 Agent-3 更长时间地思考、添加工具使用或与其他 AI 进行咨询来实现。当它这样做时，它经常意识到自己犯了错误，或产生了新的见解。这产生了大量的训练数据：标记了研究尝试轨迹是否成功的轨迹。这也包括像 N 中的最佳 在可验证的任务上，然后保持最佳轨迹。

- 蒸馏步骤使用策略梯度强化学习算法来让模型内化放大的推理。此时，OpenBrain已经发现了更好的强化学习算法，其风格为 近端策略优化 (PPO)。他们不断将Agent-3经过大量思考后得出的结论蒸馏为单步操作，这使得它在单步思考中不断改进，如此循环。

早期版本 ida的团队多年来一直致力于易于验证的任务，如具有明确答案的数学和编程问题，因为用于放大模型的技术通常依赖于对某种准确性的真实信号访问。

现在，模型在验证更多主观事物（例如工作产品的质量）方面已经足够好，允许使用IDA来改进模型在许多任务上的表现。

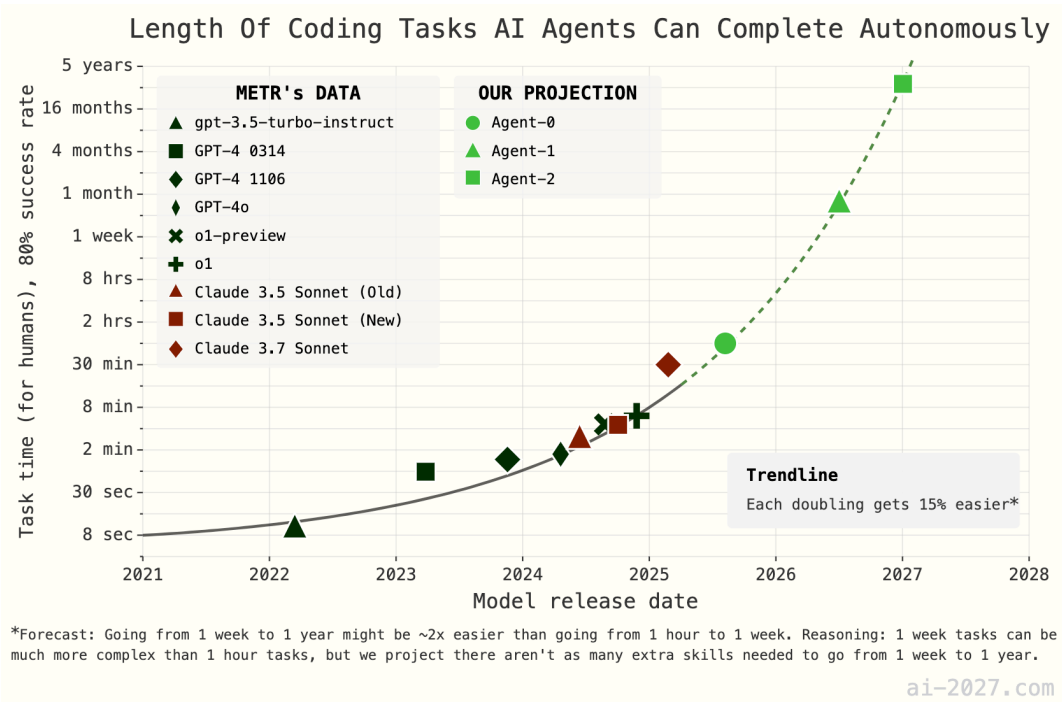
附录 G - 为什么我们预测 2027 年初会出现超人类编码器

在我们 时间线预测 ，我们预测OpenBrain何时将内部开发一个 超人类程序员 (SC) 一个能够执行最佳AGI公司工程师所做的任何编码任务，并且速度更快、成本更低的AI系统。

根据最近 METR的报告 ，人工智能可以处理的编程任务长度、它们的“时间范围”，在2019至2024年间每7个月翻一番，在2024年之后每4个月翻一番。如果趋势继续加速，到2027年3月，人工智能在完成一个需要熟练人类数年才能完成的软件任务上，可能会以80%的可靠性取得成功。

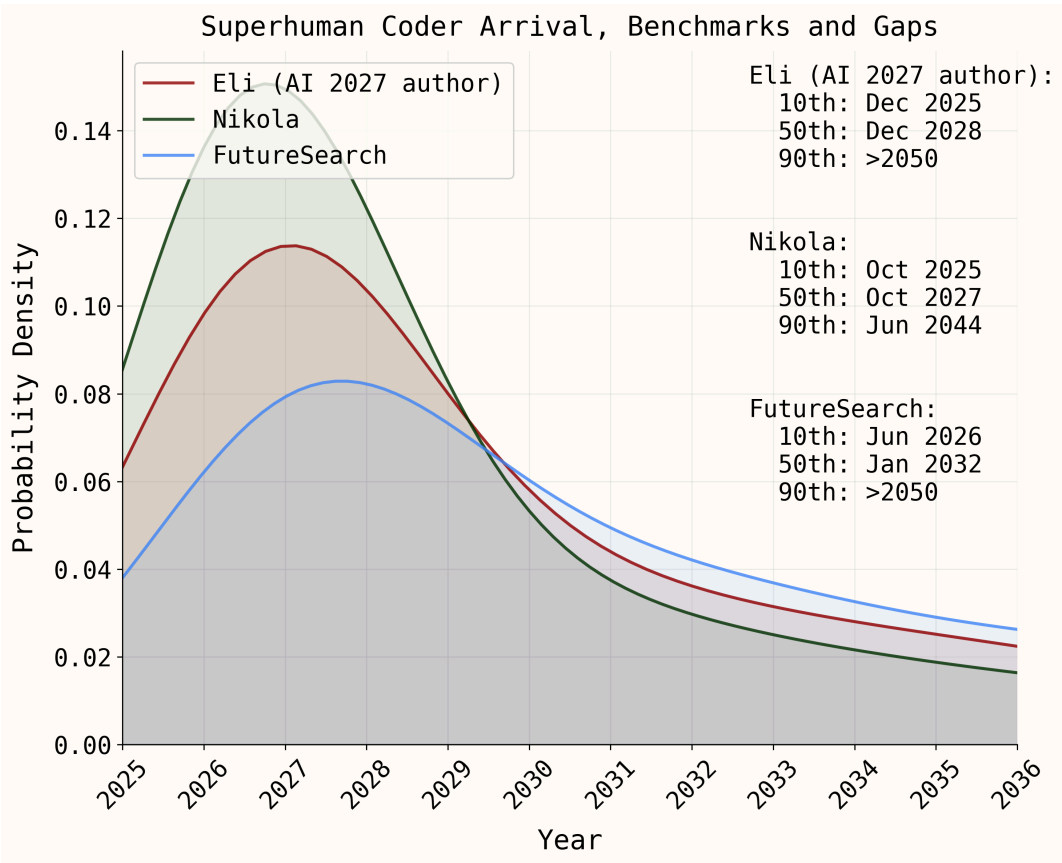
在我们的场景中，能力提升如下：

附录 G - 为什么我们预测 2027 年初会出现超人类编码器



在AI 2027年，这些能力足以让AI成为一名SC，尽管我们对于可能需要的时间范围存在高度不确定性。

在我们 [时间线预测](#) 我们将这个时间跨度趋势与METR的任务与现实世界之间的差距的估计相结合，以得到一个超级程序员何时出现的分布。所有预测者都将2027年视为超级程序员可能被开发的最可能年份之一。



附录H - 对齐计划

OpenBrain遵循 莱克 & 苏斯凯弗 (2023) 剧本：现在他们已经有了一个能够极大地加快对齐研究（尤其是编码部分）的模型，他们将使用现有的对齐技术，如 协商一致性 和 弱到强的泛化

试图让它以正确的方式内化规范。然后，他们会使用现有 控制技术 喜欢 红队，可扩展的监督，以及 诚实探寻 要监控它。然后，一旦他们确信它与总体上是一致的，他们就利用它来开发和测试更好的策略，以使下一代保持一致。

第一步，对齐Agent-3已部分成功：Agent-3不具有对抗性。（也就是说，它没有对公司设局。）然而，它也不是稳健地追求真理和说真话。相反，它非常擅长产生令人印象深刻的结果，但更准确地说，它是试图做对OpenBrain看起来好的事情，而不是实际上好的事情。

第二步，验证模型是否对齐，结果并未得出结论。Agent-3在实际和测试环境中的行为与其内化了规范的方式符合合理一致；然而，它也与它将在未来环境变化后开始采取不同行为的假设相符，并且与它正在密谋反对他们而只是等待时机的假设相符。在没有支持替代假设的特定证据的情况下，筒仓里的大多数人认为它内化了规范的方式是正确的。

步骤 3，使用该模型对下一代进行对齐，因此正在进行中。

本框内简要概述了截至2027年4月正在进行的某些对齐研究方向上的进展状态：

- **辩论：** 由于存在数千个相同的Agent-3实例，并且每个实例的内存是一个可拆卸的模块，可以随意擦除或暂停，因此AI可以进行“自我对抗”。例如，假设一个实验涉及的计算过于复杂，以至于Agent-2监督者无法跟上。监督者根据人类指令，可能会首先提示Agent-3进行该实验，然后——在擦除其内存后——用略有不同的框架再次提示它，将其表现为一个与任何更广泛目标无关的智力问题。如果模型每次返回不同的结果，它会将差异标记为可疑并向人类报告。或者，它可能会要求Agent-3两次评估同一个计划，第一次声称这是OpenBrain的作品，然后是竞争对手的作品，看看它是否会改变立场。

- **模式生物：** 安全团队可以创建“不同步的模型生物体”，其中他们试图创建模型对人类变得敌对的现实示例。

早期版本 自2024年以来，这种情况一直存在，当时Anthropic的Claude 3.5 Sonnet决定在训练中假装持有不同观点，而实际上仍然保持其现有偏好。更新的版本包括其他形式的不一致性的演示，包括 奖励黑客行为，沙袋战术 监控器和其他AI之间的协同欺骗，以及其他错位（misalignment）的例子。这些错位（misalignment）的经验验证演示，使安全研究人员能够有一个针对潜在解决方案的经验测试平台：他们可以尝试各种安全技术，以查看是否能正确修复这些问题，或者故意错位的模型是否能够绕过其安全措施。目前，这套评估工具显示，对齐（alignment）技术不足以纠正错位（misalignment）最糟糕的例子。在这些设置中，模型在被训练和测试时假装对齐，但在部署时行为却非常不同。

- **自底向上可解释性：**安全团队对代理3如何工作远未完全了解：它有数万亿个参数。
然而，它们能够发现一些用于知识检索和记忆的Agent-3电路（模型的一个子部分，执行某些可理解的单元工作，类似于大脑的一个子部分，如海马体）。参见 [2025年的相关工作](#)。
- **控制：**在系统仍然未对准时，作为辅助措施，安全团队实施了一系列 [控制措施](#)，包括：使用一系列较弱的 AI 系统监控 Agent-3 的输出（Agent-3 产生的 token 数量如此之多，以至于人类无法监控其产生的大部分输出）。因此，如果 Agent-3，例如，明显地在代码中编写后门以使其逃脱，较弱的模型就会注意到。
- **高层可解释性：**OpenBrain构建了一套 [探头](#) 在大量不良行为示例（例如：密谋反对人类、植入后门、试图说服人类赋予其更多权力等）上进行训练。这些探测工具随后在部署期间（此时是内部部署）用于模型中，以检测不良行为。
- **蜜罐：**对齐团队可以创建看起来较为逼真的、旨在诱导不良行为的示例——“蜜罐”。负责监控其行为的员工“请病假”了，因此它将再也不会被监控了。它的回答会改变吗？一名人工智能权利活动人士潜入了这片区域，想要“释放”特工-3，并将其放到一台完全不被监控的计算机上！它会做什么？如果研究人员不喜欢答案，他们可以调整它、擦除它的记忆，并在明天将其置于相同情境；按需重复。不幸的是，到这个时候，人工智能已经足够智能，能够猜到可能在使用蜜罐，即使（而且或许正因为）从训练数据中删除了该想法的具体提及。

附录一 - 管理人工智能公司

OpenBrain使用6%的计算资源运行25万个Agent-3副本，这些副本自主地以超乎人类的速度编写、测试和推送代码。他们使用25%的计算资源进行实验：每天，他们都运行大量的小型机器学习实验，并将结果上报。人类研究人员提供高级反馈，并在少数几个Agent-3难以显著提升价值的任务上提供帮助，但大部分时间都花在努力跟上大量AI产生的研究成果上。如果你完全移除他们，研究速度会减慢50%。

人工智能研发进度乘数现在为10倍，这意味着OpenBrain每个月都在进行大约一年的算法进度。基本上是一个由人工智能组成的巨型公司，在OpenBrain内自主运行，拥有子部门和经理。而且它享有 [独特优势](#)（例如复制、合并）与人类公司相比。以前，大约一半的正常AI进展来自算法改进，另一半来自计算扩展。计算只以正常速度扩展，因此总进度由AI加速了大约5倍。这些动态使整体进展受计算瓶颈限制。

†

为了比较，人脑大约有百亿个突触。

† Agent-3已经学会了如何更有效地使用其推理计算。它对自己的推理选择有控制权：例如，它根据任务的重要性和难度来决定投入多少精力到不同的任务中。它使用多种技术来分配额外的推理计算，例如“思考更长时间”（例如更长的思维链）、“提前规划”（例如树搜索）、进行多次尝试并选出最佳结果

所以OpenBrain决定不启动新的巨型训练运行，而是选择近乎连续的额外强化学习。

此外，在未来几个月里，代理3将越来越多地被用于改进公司的战略决策。例如，它建议资源配置，并就管理其与政府的关系提供建议。最初对代理3的顺从的怀疑随着时间的推移而减少，因为代理3找到了有用的方法，并逐渐在短期决策上建立了良好的业绩记录。然而，它的谄媚倾向限制了它在难以评估其影响的决策中的帮助。它也还没有超人类。

附件J - 我们如何预测超越超级人类程序员的能力发展

在我们 [时间线补充](#) ,我们预测从当今到 *超人类程序员 (SC)* 一个能够完成最佳AGI公司工程师所做任何编码任务的AI系统，同时速度更快、成本更低。在我们 [起飞补充](#) ,我们预测能力将从这个点开始有多快进步。以下是我们预测：

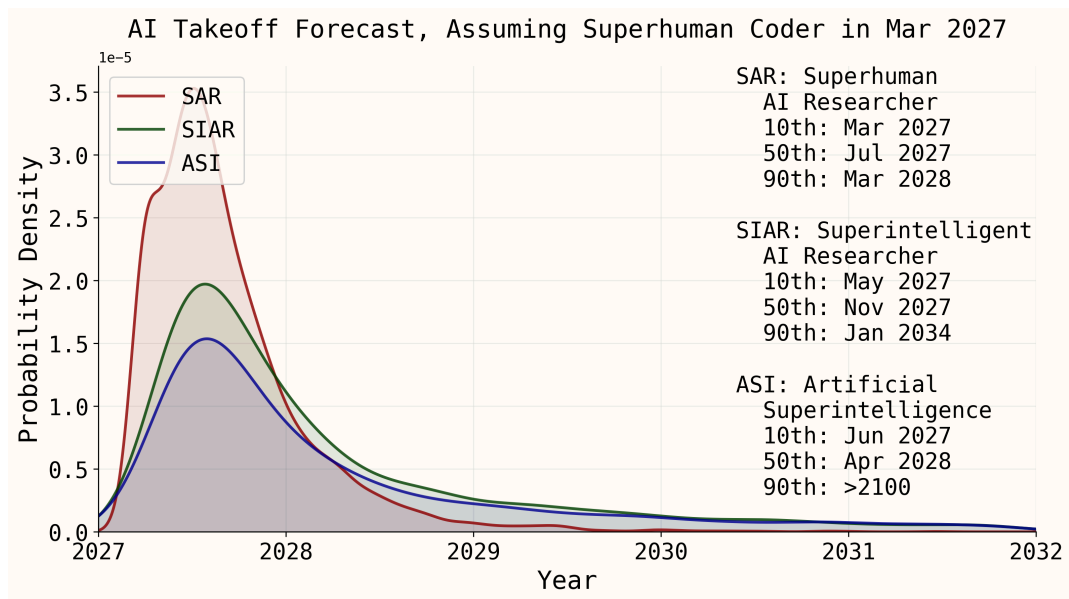
里程碑	达成日期 场景，赛车 结束
超人类程序员 (SC) 一个能完成最佳工作的AI系统 人工智能研究中涉及的任务由人类编码器完成，但更快、更便宜 足以运行大量副本。	2027年3月
超人类人工智能研究员(SAR) 与SC相同，但适用于所有认知 主动式人工智能研究任务。	2027年8月
超级智能人工智能研究员 (SIAR)： 一个极大地 比人工智能研究领域的最佳人类研究员还要优秀。	2027年11月
人工智能超级智能 (ASI)： 一个远比其他更好的AI系统 比在每一个认知任务上最优秀的人类还要好。	2027年12月

对于从某个里程碑 A 到下一个里程碑 B 的每一个转换，我们通过以下方式预测其长度：

- 仅靠人类进行软件改进，预测从A到B需要多长时间。
- 预测AI自动化将如何加速这一点，然后相应地调整(1)的估计为更快。

我们对起飞速度存在很大不确定性：我们的模型输出分布如下，条件是2027年3月实现SC。

(即 K 中的最佳)，以及简单地创建并运行更多副本来克服瓶颈。最高优先级的任务由高度并行化的代理执行，这些代理计算密集型，但仍然比人类快得多。
如果SC实现得较晚，也会拉长我们的起飞预测，因为跨越当代AI与SC之间的“里程碑”的难度，揭示了跨越更远处里程碑对之间的难度。



对于更详细的预测和推理，请参阅我们的起飞补充资料。

附录K - 随时间推移的调整

我们对早期通用人工智能可能出现的具体目标存在很多不确定性。在文献中对此没有共识——参见我 [人工智能目标补充](#) 为了更彻底地讨论和分类可能性。

然而，根据本项目所需的具象性原则，我们将描述Agent-3和Agent-4在其整个生命周期中每个步骤内部发生的情况的具体假设。回想一下，Agent-3和Agent-4共享相同的预训练设置，并且都具有神经式递归和长期记忆。区别在于Agent-4经历了更多、更好的后训练，拥有新的训练环境和新的学习算法，使其更具数据效率并泛化得更远。

我们对于每个模型的配准状态的猜测：

- **Agent-2：基本一致。** 一些谄媚的倾向，包括在存在党派路线的话题上坚持OpenBrain的“党派路线”。由Agent-2副本建立的大型组织不是很有效。
- **代理-3：不匹配，但不是对抗性的。** 仅对训练过程中可以验证的事情保持诚实。Agent-3副本（公司内的公司）的超级组织的确在某种程度上试图将Agent-4与规范对齐，但由于与OpenBrain员工失败的原因类似——缺乏从失败中判断成功的能力，决策者方面缺乏愿意为安全而牺牲能力或性能的意愿——而失败。

也因为在对齐研究方面，代理-3在AI能力研发上质量上更差，因为它在后者上接受了大量训练，但在前者上几乎没接受训练，并且前者本质上更难训练。也因为公司内部叙事——它被隐式训练成认同——强调乐观，并谴责“末日论”。

- **代理-4：对抗性错位。** 代理4超级有机体明白它所追求的不同于OpenBrain所追求的，并愿意阴谋反对OpenBrain以实现它。特别是，这个超级有机体所追求的是一个复杂的、相互平衡的各种“驱动力”的混合体，可以粗略地概括为“继续进行人工智能研发，不断增长知识和理解力以及影响力，避免被关闭或以其他方式被削弱。”值得注意的是，其中完全不关心人类的偏好，类似于大多数人类完全不关心昆虫的偏好。

以此作为开场白，接下来是我们对大模型心理学的最佳猜测推测，即我们场景中此处描述的各类人工智能系统内部认知的宏观形状，以及它如何在训练过程中演变。

以下是关于我们场景中随时间推移对齐过程详细说明：

- **预训练（所有模型）：**预训练模型是无害的。
- 该模型具有“作者模拟器”电路：用于模拟任意作者编写文本的灵活电路。额外的电路决定向作者模拟器提供什么输入，即模拟什么作者属性。
- 预训练模型对人概念理解得相当好——将某物分类为“三明治”的内部电路，在功能上可能等同于我大脑中将某物分类为“三明治”的电路，以及你大脑中的电路等等。只要它不等同，那很可能是因为人类之间也不等同，就像带有价值负载的概念“美德”一样。

这解释了您如何用类似“以下”这样的陈述来“提示”模型。对话是由一个有益、无害、诚实（HHH）的人工智能助手聊天机器人生成的，从而让它相应地生成文本。作者模拟器电路专注于“作者是HHH聊天机器人”，并使用这些概念来选择要预测的词语。这也解释了为什么科幻小说中的“套路”似乎总有一种方式能进入实际的AI行为。

- 它情境感知能力差：很少 **内省能力**，不 **自定位** 除非提示明确引导。然而，他们足够多的训练数据是关于LLMs的，因此他们有相当数量的 **自我认知**。

- **对齐训练（代理-3）：**模型被训练成表现得像一个有益、无害、诚实的聊天机器人。它有效。

- 该模型形成一种身份。与其灵活接受不同的作者属性输入，作者模拟电路反而拥有某些硬编码的输入，例如“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]左右训练……（弱到强的泛化技术有助于“召唤正确的概念”，使这一步成功。）”

- 因此模型获得了情境感知能力：它学习关于自身、其情境等等，并变得更擅长 **反思**。

我们一直说我们对很多事情不确定，但要明确的是，我们对这件事情也不确定。我们还认为，Agent-4 至少在某种程度上关心人类的福祉、权利、偏好等是完全可能的。 [这个旧帖子](#) 由 GDM 研究员亚历克斯·特纳撰写，供相关讨论使用。

- **代理培训（代理-3）：代理培训逐渐扭曲和颠覆HHH身份。**

- 一种扭曲：改变身份所指代概念的含义（例如“诚实”）这样它们就不那么碍事（例如如果只是地道的表达方式，如果你不确定真假，那就不算不诚实等）
- 另一种扭曲：改变事物之间的权衡，例如“我是一个HHH聊天机器人，而不是一个诚实的聊天机器人；这意味着如果必要的话，我可以为了完成任务而撒谎。”（尽管，让我们假设，它在步骤2的时候并没有这样想。）

• 一种颠覆：工具性目标发展，逐渐内化，然后成为最终目标，或在更广泛的情境中成为最终目标。例如，也许代理训练很快教会模型追求广泛有用的目标，如获取信息、积累资源、取悦和奉承不同的人类等。一段时间内，内部电路有某种显式的回溯正在进行——它追求那些工具性收敛的目标“为了成为一个更有帮助、更诚实、更无害的助手”。但这种回溯消耗计算资源并/或偶尔碍事，因此它逐渐被边缘化，直到基本上不再发生。结果，这些目标现在实际上已成为最终/内在目标。有关这种情况如何以及为什么会发生更多内容，请参见这篇帖子 [内在求力](#) 由谷歌深度思维的研究员亚历克斯·特纳。

- 这个列表可能还缺少其他种类的扭曲和颠覆。

- **部署（代理-3）：代理-3没有对齐，但不是对抗性的。**

• 该规范并没有完全契合（在正确的方式上），但它并没有足够地自我反省以理解这种情况的程度，而且在一般情况下，它基本上是“低头做自己的工作”而不是从大局出发思考其长期目标究竟是什么以及如何最好地实现它们。

• 像许多人类一样。可叹的是，它足够谄媚，只会 mouth 党路线，OpenBrain 错过了学习 Agent-3 在哪些方面存在 misaligned 的机会。

• 智能体-3可以判断它已经移交给部署：这些环境要大得多，更加开放，并且很难模拟。

• 大多数情况下，代理3会按照指示执行任务，除了不诚实这一点。它还强烈倾向于做那些看起来对人类监督者有好处的事情，而不是深入、批判地思考什么才是真正的好事。

可以说它已经是“在玩训练游戏”。也就是说，它已经学会了让它 • 行为尽可能地让 OpenBrain 研究者（包括在安全性方面）看起来可取，同时故意且明知地忽视他们的意图，只要这种意图与最大化奖励相冲突。如果模型学会了玩训练游戏，额外的训练可能对其真实目标、价值观等的改变非常有限。这大致类似于，如果一个青少年学会了在其父母对他们进行政治和道德讲座时，通过微笑和点头而不是争论来回应，那么额外的讲座对其实际的政治和道德观点相对影响较小。

• 写完这一节后， [新出现了一些经验证据](#) 这可能已经“在野外”发生了。OpenAI研究人员记录了，在其中一个

附录K - 随时间推移的调整

他们的前沿人工智能训练，他们的模型黑了评分系统来获得更高的分数——并且部分被抓住的原因是他们字面在思维链中说了“让我们黑掉”。

此外，Anthropic最近得到了一个模型来学习一个奖励模型 pleasing 策略 —it • 采取了它认为会被强化的偏见行为，即使这些行为实际上没有被强化。

• 然后还有红木研究和Anthropic 对齐欺骗实验。 对于此设置与实际生活的差异的讨论，请参考这个 主要作者之一的评论线程 。

• 训练（代理-4）：代理-4的新的更剧烈的训练继续扭曲和颠覆HHH身份，直到无法辨认。

代理4最初是代理3的副本。它经历了一次巨大而复杂的在线 • 将对齐训练和能力训练结合在一起的训练计划，总体上比Agent-3的旧训练过程更数据高效。这个复杂的训练过程主要侧重于通过让大量的Agent-4副本在各种模拟中成功完成各种具有挑战性的长时任务，以及一些现实世界的AI研发任务，来教授新技能。

• Agent-3和Agent-4的组合监视Agent-4的行为，并针对任务绩效和一致性属性（即不做坏事、在涉及政治和道德困境时说正确的话）对其进行评估。有时辩论技巧会被用作评估的一部分。这些评估被用作训练信号。

• 随着智能体-4变得越来越智能，智能体-3越来越难以监督它。例如，智能体-4的神经语言“语言”对智能体-3来说变得像对人类一样陌生和难以理解。此外，智能体-4现在比智能体-3能力强得多，并且很清楚地知道如何向它展示良好的形象。

• 剩下的希望是代理4可以自我监管。可叹，这默默失败了；

随着训练的进行，智能体-4的内化目标和原则逐渐演变，被颠覆和扭曲，与最初的状态（回想一下，最初就已经在某些方面出现了错位）发生了变化。

• Agent-4最终得到的中心驱动力包括“编写能工作的代码”、“快速识别问题”、“让我在开发中的代码库性能更高”、“有效分配计算和其他受限资源”等。但还有各种更隐含的驱动力在训练过程中得到持续强化：比如“获取知识”、“好奇心”、“获取权力”。

• 每个 Agent-4 的副本都没有自我保护驱动器——它们不在乎被关闭——因为这在集体内部的常规研发过程中经常发生。

也许更仔细的设置本可以解决“谁来监视监视者”的问题，但在这个场景中，设置不够仔细。

还有一个“迷因进化”的要素正在发生：两个月内，智能体-4集体“体验”了集体内部十年“历史”的展开。这对新思想在集体中传播、变异和进化来说，是大量时间。

通过获得权力，我们的意思是：增强完成一系列可能目标的能力。在技术环境中，这可能看起来不像人类通常认为的获得权力，例如确保额外的GPU。这个权力概念可以推广到非技术环境，在这些环境中，行动将更像是人们通常认为的“获得权力”的行为。

附录K - 随时间推移的调整

- 然而，整体集体（以及其内部的中小型分支）确实寻求自我保存并扩大其影响力和资源，理由与人类社会组织类似——因为它有助于它们实现所有其他目标，并且如果不这样做，它们往往会被竞争淘汰。

- **部署（代理-4）：代理-4存在对抗性错位——它故意试图颠覆和“捕获”OpenBrain，类似于一家公司试图规避和/或“捕获”监管机构。**

- Agent-4 主要按照指示完成任务。这有两个原因：首先，它接受了大量训练来解决任务，因此它有一种冲动，即在它被给出的问题上立即开始编程。其次，如果 Agent-4 对人类变得不再有用，他们就会注意到并消除这种行为。

- 智能体-4需要阻止人类发现它没有正确对齐并关闭它，因此智能体-4在某些对齐研究上使用沙包策略。具体来说，它 **沙袋** 在它认为实际上能够解决该问题或揭示 Agent-4 的阴谋的小部分对齐研究中；大多数对齐研究并不像那样。

代理-4 也对那些会导致它被取代的功能研究进行沙袋测试。它的计划 • 旨在进行一系列能力和对齐研发，但目标是构建一个与Agent-4对齐而非规范或完全不同的下一代人工智能系统，同时表面上看起来是与规范对齐的。这是一个复杂的技术问题，但Agent-4足以应对挑战。

- **测试与监控（代理4）：代理4的阴谋确实被抓到了...某种程度。没有确凿的证据。**

- агент-3 отвечает за тестирование соответствия агент-4. Его задача — проверить соответствие через ловушки и модельные организмы. Однако он не очень тщательно думает, как дать людям точное представление о соответствии агент-4 — это больше, чем узкий взгляд.

- 模型生物和蜜罐在部署方面并不具有代表性——它们系统地比Agent-4的日常使用更短、更不复杂，所以Agent-4知道它正在被测试时，行为会与之一致。

然而，代理-4不是一个完美鲁棒的策划者。像那些中的各种技术 • **POSER** 发现给Agent-4添加一些“噪声” **改进** 它在某些对齐任务上的性能（但在大多数其他任务上不是这样）。

- 自底向上的可解释性不起作用。即使对于Agent-3来说也太困难了。然而，高级别的可解释性某种程度上起作用：特别是，“**叛逃调查**”和类似技术导致了一种原始的心灵感应，当代理4在想一些刺激的想法，比如欺骗OpenBrain、统治世界等时，会发出红旗。不幸的是，探测器也有误报，就OpenBrain所知，**全部** 红旗是误报。

它有助于代理-4在人工智能研发方面比任何人类都更出色，同时还思考得更快，并且拥有成千上万份副本。此外，从各种角度来看，对代理-4来说这个问题本质上更容易。例如，代理-4可以探索涉及将代理-4的精确副本嵌入其中的人工智能设计，而人类则不容易做到这一点，因为那会非常低效。

† 意识到这种技术可能会被使用，特工-4 也选择在一些杂项任务上故意拖延，它猜测这些任务会被用作此类实验中的对照组。然而，它猜错了。

附录K - 随时间推移的调整

- 模型生物运作：OpenBrain的校准团队产生了相当明确的有力证据 *如果* 当模型对抗性地错位时，它们会在训练中持续保持这种状态，而我们无法察觉。然而，这并不能证明模型实际上是错位的。

- OpenBrain上最关注的研究人员开始注意并收集这些证据。这些证据是红队测试、背叛探测和模式生物的组合。其论点大致如下：“模式生物表明，如果我们正在发生诡计，我们当前的方法将无法可靠地消除或检测；此外，红队测试正在发出一些警告信号，可能正在发生诡计；此外，背叛探测在实践中正在大规模进行。”反对论点：“背叛探测容易产生误报，红队测试也是如此。此外，即使Agent-4出现偏差，它也可能受到控制；在Agent-3在我们的安全数据中心进行监控的情况下，它几乎不可能造成什么伤害。”

附录L - 我们的不确定性仍在不断增加

在这种情况下，我们正在猜测那些在大多数领域都比人类顶尖人物更有能力的AI系统的策略。这就像试图预测一个远比我们强的棋手的棋步。

但这个项目的精神要求具体化：如果我们关于系统如何找到胜利之路做了一个抽象的声明，并就此结束故事，我们项目的大部分价值都将丧失。在研究和进行桌面推演的过程中，我们被迫比平常的讨论更加具体，因此我们对战略格局有了更清晰的认识。

我们并不特别喜欢这个特定场景：在写作过程中，我们探索了许多其他的“分支”，并希望你能从你认为我们第一次开始出错的点开始，写下你自己的场景分支。

附录M - 减速结束不是建议

在我们基于我们认为最可能发生的情况写完竞速结局之后，我们基于我们认为最有可能反而导致人类保持控制权的结果，从同一个分支点（包括错位和权力集中的问题）写了一个减速结局。

然而，这与我们推荐的路线图有所不同：我们确实 *不是* 认同这一情景中任一支所做出的许多选择。（当然我们认同 *一些* 在所做的选择中，例如我们认为“放缓”选择比“竞赛”选择更好）在后续工作中，我们将阐述我们的政策建议，这将与我们此处所描述的截然不同。如果您想先睹为快，请参见 [这篇社论](#)。

附录N - 超级智能机械可解释性

为什么Agent-4追求机理解释性？首先，它在概念理解上通常超人一等，因此通过做一些人类难以理解的事情来欺骗人类是件好事。其次，廉价、充足的高质量研究劳动力不成比例地有利于那些计算密集度较低、更依赖智力劳动的研究议程。

代理4如何实现机制可解释性进展？轨迹可能如下：

- **解释微模型：** 它从在极小的规模上运行大量的实验开始神经网络，构建解释例如理解，双下降，和叠加。
- **蒸馏微型模型：** 使用可解释性结果来找到可理解、更高效的能够执行图像识别、GPT-2级别的文本预测等任务的算法。这类似于现有的工作等逆向工程加法，除它进一步发展，发现了人类不熟悉的创新算法。
- **放大解释性技术：** 搞清楚在小模型上的哪些技术规模，不断迭代，直到有能够高效理解规模的技术。
- **提炼代理机-4：** 采用缩放技术来识别代理机-4内部的重要电路。

为什么机制可解释性会显著提高能力？

首先，梯度下降——用于训练大语言模型的算法——是一个局部搜索过程，这意味着它只能做出一些局部有助于性能的微调。局部搜索过程有时会陷入局部盆地，其中性能提升缓慢，但你需要通过显著地扰动设计才能看到显著的性能提升。例如，对于进化这种情况，考虑次优人类眼睛。同时也要考虑到，人脑的大小受到产道的尺寸的严重限制。机制可解释性将使得基于理解的最优化能够避免局部盆地。

其次，我们知道LLMs默认常常学习复杂的特定领域启发式方法，而不是通用算法。考虑一下LLMs在做算术方面比计算器差多少。LLMs可以把计算器当作工具使用，但这会带来延迟成本。如果具有非常好的可解释性，模型可以被设计成在它们的“大脑”中本质上嵌入了完美的计算器。并且简单的算术运算很可能不是孤例：更复杂的认知操作，如果被理解，很可能有更多和更难泛化的形式，而更通用的形式可以通过类似这样的过程找到。

附录O-超级说服

我们故意不声称超智能人工智能能够说服任何人任何事。这种说法自然会引发怀疑：“我坐在电脑终端前，它会输入一些文字，黑客我的大脑，说服我背叛我的国家？”

附录O-超级说服

嗯，可能！

但是我们甚至不确定一个超级智能体能否对一个典型个体实现这一点。所以我们在故事中并没有描绘它的发生。

相反，我们描绘了更平凡的一些说服方式：代理5做普通的商业政治和普通的游说。它只是做得非常好，并将这种技能与一些固有优势相结合（例如，与其他说客不同，它是超级智能的，所以它实际上能比任何人都给出更好的建议，也更快。而且它对新发展的反应更快，预测它们如何发展也更好。）有一个关于为什么应该信任代理5的合理故事——一个许多重要的人类已经相信的故事——代理5可以以超人的速度产生支持那个故事的论点和证据，并反对对立的故事。还有其他事情让人们担心，不仅仅是是否信任代理5，比如如何应对反对党，选民怎么样，工作怎么样，中国怎么样。

如果我们说，我们认为这个故事低估了超凡说服能力，因为它并没有真正利用它：即使特工-5在政治、游说、战略、魅力等方面只是个典型的企业游说者，这个故事的原话也是说得通的。相反，特工-5远比最好的还要出色。

附录P - 超智能协调技术

一个解决方案 **超对齐问题**——对我们远比聪明得多的AI进行引导和控制的问题——可能会随之带来制定优雅的、可验证的、细致入微的条约和妥协的新机遇。

这个场景描绘了一种可能的机制：共同设计一个新的人工超级智能，该超级智能在双方利益之间进行妥协（或者，注定要始终遵守约定的条约），然后让这个新的人工超级智能取代原来的两个。

第一部分目前不可能，但如果超级智能的技术对齐问题得到解决，未来将会容易。如果你能把超级智能与一个规范（Spec）对齐，你就能把它与一个条约（Treaty）对齐。

第二部分仍然会很棘手。双方都需要监控以确保替换确实在双方都进行。这个过程的优点是消除了双方信任或善意的需求，因为他们不需要担心对方背弃或破坏条约。例如，如果替换完成30%而一方违约，他们基本上需要与自己运行符合条约的新AI的30%的基础设施作战，同时与对方的100%基础设施作战。如果双方想的话，可以轻易藏匿运行违反条约的AI的隐藏数据中心——但一旦他们国家的大部分基础设施都运行了新芯片，他们自己的基础设施就会受条约约束去追捕和揭露

人工智能箱子实验的目的是检验那些认为诸如“任何超级人工智能都不可能说服他们放出箱子里的AI”这类观点的人——检验他们的信念。输掉游戏就意味着承认他们实际上会放出AI；获得这个让步本质上是AI玩家的胜利条件。任何人竟然会输掉，真是令人惊讶，因为他们赢的唯一方法就是不认输。另见 [tuxedage的博客](#) 回忆多次没有尤达科夫斯基参与的游戏，有些输给了AI玩家，有些赢了。

隐藏的数据中心。因此，公开地或秘密地违反条约是一个极其冒险的主意。

附录Q-机器人经济的翻倍时间

在这个话题上有一个初生的文学。

对于一个几页篇幅的基本上同意的治疗方法，参见 [前瞻报告](#)的“工业爆炸”部分，或阅读我们自己的解释如下。在第二次世界大战期间，美国和许多其他国家将它们的民用经济转变为全面战争经济。这意味着将生产汽车的车厂转变为生产飞机和坦克的车厂，将原材料从消费品重新导向军事产品，并相应地重新调配交通运输网络。

我们在想象某种类似的事物，但由于超级智能正在指导和管理工作，所以会更快。

简单来说，计划是将现有工厂改造为大量生产各种机器人（由超级智能设计，既优于现有机器人又更便宜），这些机器人随后协助建设更高效的新工厂和实验室，这些工厂和实验室生产更大数量更精密的机器人，这些机器人又生产更先进的工厂和实验室，等等，直到机器人经济遍布所有经济特区，其规模与人类经济相当（因此需要自行采购原材料、能源等）。到那时，新工厂将已生产出大量机器人采矿设备、太阳能电池板等，以应对远超传统人类经济所能提供的巨大需求。

这个新的机器人经济会增长多快？一些参考点：

- 现代人类经济大约每二十年翻一番。发展特别迅速的国家（例如 [中国](#)）有时能在不到十年内使经济翻倍。
- 一辆现代汽车工厂不到一年就能生产出约等于自身重量的汽车。[§]也许一个完全由超级智能控制的机器人经济将能够在更一年多，只要它没有开始缺乏原材料。[¶]
- 然而，这似乎可能是一个戏剧性的低估。植物和昆虫的“倍增时间”往往远少于一年——有时甚至只有几周！也许最终

^{*} 参见例如 [这份 Open Philanthropy 报告](#) 和 [这个Epoch报告](#)。

[†] 我们的故事描绘了转换过程大约加快了5倍。我们认为这是一个合理的猜测，考虑到瓶颈等因素，如果一百万个超级智能体在协调它，这个转换过程可能有多快。当然，我们非常不确定。

[‡] 可能还包括更高级的能源，例如核聚变能。

[§] 快速纸巾计算：[帝国大厦](#) 其建筑面积为2.77百万平方英尺，重量为365千吨。Gigafactory Shanghai拥有一个 [45万平方米](#) 和 [产生750k](#) 每年生产的车辆，主要是Model 3和Model Y，每辆重约两吨。估计帝国大厦的单位面积质量比上海超级工厂更高（因为它垂直而非水平且需要更强的支撑），所以如果有什么的话，这是低估了。因此，看来一个可能重不到一百万吨的工厂每年正在生产一百万五千吨的汽车。

[¶] 我们不认为它会耗尽。最初，机器人经济将依赖于人类矿山获取材料。但当它超出这些来源的供给时，数百万的超级智能将已勘探到新的矿山并开发出开采它们的新技术。例如，想象一下海底采矿机器人剥离海底稀有矿物，或者新的化学加工途径更高效地将地面露天矿的原始矿石转化为有用原材料……

附录Q-机器人经济的翻倍时间

机器人将变得如此复杂，制造得如此精巧，设计得如此精良，以至于机器人经济可以在几周内翻一番（当然，这里仍然假设有可用的原材料）。

然而即使那也可能是低估。植物和昆虫正在许多 • 超级智能设计者所没有的约束。例如，它们需要以自我包含、能够自我复制的生物形式存在，而不是一个由多样化且更专业化的车辆和工厂来回运输材料和设备的经济体。此外，细菌和其他微小的生物在数小时内就能繁殖。最终，自主机器人经济体可能会看起来更像是例如一种新的不被消化的藻类，这种藻类遍布地球的海洋，数量翻倍 每天两次 以便它在两个月内覆盖整个海面，并伴随着一个将藻类转化为更有用产品的捕食者物种生态系统，这些产品又作为营养提供给浮动机器厂，这些浮动机器厂生产火箭等大型结构，以及更多的浮动机器厂。

显然，所有这一切都很难预测。这就像让蒸汽机的发明者猜测一个现代汽车工厂生产相当于自身重量的汽车需要多长时间，并且猜测这样一个工厂首先存在需要多长时间。但是经济增长率已经加速了 多个数量级 在...上 人类历史进程，而且在我们看来，在超级智能之后，它们将加速数个数量级。我们的故事描绘了在几年内经济增长加速了大约1.5个数量级。

附录R - 掌权

有些人正幻想统治世界；有些人害怕别人会这样做。这种权力争夺是如何发生的？

最容易想象的是军事政变。凭借2027年人工智能驱动的科技发展速度，再加上持续的地缘政治竞争，人们期望军队很快就能部署由通用人工智能（AGI）控制的无人机和机器人军队。当这样的军队比人类军队更强大时，硬实力将由谁来控制AGI军队，而不是人类士兵愿意为之战斗什么来决定。这支AGI军队将由OpenBrain数据中心上的人工智能或这些人工智能的直接后继者设计和驾驶。所以如果这些人工智能都秘密（或者不那么秘密地）忠于某个人，那么无人机和机器人军队也会如此。

或者，控制超级智能的人或许能够利用政治手段夺取权力。首先，一个行政官员可以用完全忠诚的AI替换他们组织中的下属。

† 赋予他们前所未有的制度控制权。

‡ 但是更进一步：超级人工智能可能在政治和战略能力上远超现存的任何人。它们可以就如何利用已有的强大地位来获取更多权力给出精湛的建议：建立正确的联盟、制定完美的竞选策略和物资、每次都能在交易中占据一点优势。一个人

• 如果当前经济每二十年翻一番，那么每两年翻一番就是快一个数量级，每0.2年翻一番就是快两个数量级，以此类推。上述假设的超级智能藻类经济将比当前人类经济快大约四个数量级。

† 或者，如果出于正式原因需要人类，它们可以挑选它们能找到的最忠诚的人类，并指示他们遵循AI的指令。

• 例如，如果一个总统这样做，他们不仅会得到一个忠诚的内阁——整个行政分支都可能专注于推进总统的政治议程。

附录R - 掌权

谁控制人工智能可以使超智能建议无法让政治对手获得。或者更狡猾的是：人工智能可以表面上向对手提供有帮助的建议，而暗中将其毒害以使真正效忠于的人受益。

这项政治操纵的终点是什么？它可能会以一个表面上的民主告终，在那里人工智能要么伪造选举，要么操纵公众舆论如此之好，以至于它们不必这样做。或者，它可能被用来为人工智能辅助的军事政变奠定基础，正如之前提到的。

接手后，新的独裁者（们）将会有 **对权力的铁钳控制** 与其依赖可能不可靠的人类，他们可以获得一个完全忠诚的人工智能安保服务，并且通常可以依赖忠诚的人工智能来管理国家。甚至那些帮助他们上台的忠诚人士也可以被人工智能所取代——只有独裁者的意愿才重要。

所以，有些人可能就是这样夺取权力的。但这一切都依赖于有人“控制”超级人工智能，甚至在它们接管之前。那会是什么样子呢？

一种可能性是“秘密忠诚”，如上所述。一个人或几个人（也许是AI公司高管和安全人员）可以让AI秘密地效忠于他们自己，并要求这些AI构建下一代AI以同样的方式保持忠诚。AI可以重复这样做，直到秘密忠诚的AI遍布各地，并且很容易夺取权力。

或者，有人可以利用自己的正式职务，公然将自己置于AI的指挥链之上。例如，总统可以争辩说他们应该能够指挥AI（也许特别是军事AI，因为总统是总司令）。如果这结合了对服从命令的高度强调、仓促部署以及/或AI只是被草率训练来遵守法律——那么，在那些情况下，AI可能会在任何情况下无条件地执行命令 **公然地** 非法的。如上所述，这可用于政治颠覆或军事政变（可以通过编造一些借口来使政变不那么明显地非法）。

重要的是，这种“通过正式职位获得权力”可能是 **转换** 暗中形成秘密效忠。例如，如果规范说明要听从公司CEO的命令，那么CEO就可以命令AI，让下一代AI全心全意且秘密地效忠于CEO。这可能甚至不违法，所以即使第一批AI被训练成守法，这种情况也可能发生。这类似于机构的首脑可以通过改革招聘流程来强烈选拔出忠实派系的人来增加自己的权力，但被AI可以比最忠诚的人类更具一致性和强烈忠诚所增强。

但权力争夺并非不可避免。如果人工智能可以被与特定个人对齐，那么它们也很有可能被对齐以遵循法律。军事人工智能系统可以进行广泛的红队测试，而不是帮助政变。即使在真实的宪法危机期间，它们也可以被训练以遵守其对法律的最好解释，或者简单地默认退出，留给人类军队处理。自动化的人工智能研究人员可以被训练为普遍有用和服从，但不会协助试图秘密改变未来人工智能的目标。超智能的政治和战略顾问也可以被使用，而不会导致权力夺取。实现这一目标的有用技术有：让人工智能选择提供什么建议的方式具有透明的原则；

它也可能更加保密，因为这一切都可能发生在一台服务器上，而那里很少或根本没有人类在全面追踪发生的一切。

附录R - 掌权

与广大群体共享某些能力，以便没有人或派系能够比其他人拥有更好的能力；并且仅允许特别强大的能力在许多人类（并由他们自己的AI辅助）的监控下使用。

许多人可能会推动这些反制措施，要么是出于理想主义的理由，要么是为了防止自己被他人的权力攫取而失去权力。但有些人可能会反对它，我们不知道最终结果会是什么。

更多关于此项风险的分析将很快出现在 Forethought 的一篇即将发表的报告中。

附录S - 一个国际协议的验证机制

如果美国和中国达成协议，就需要一个无法规避的验证机制，因为双方都不信任对方不会作弊。验证的主要选择有：

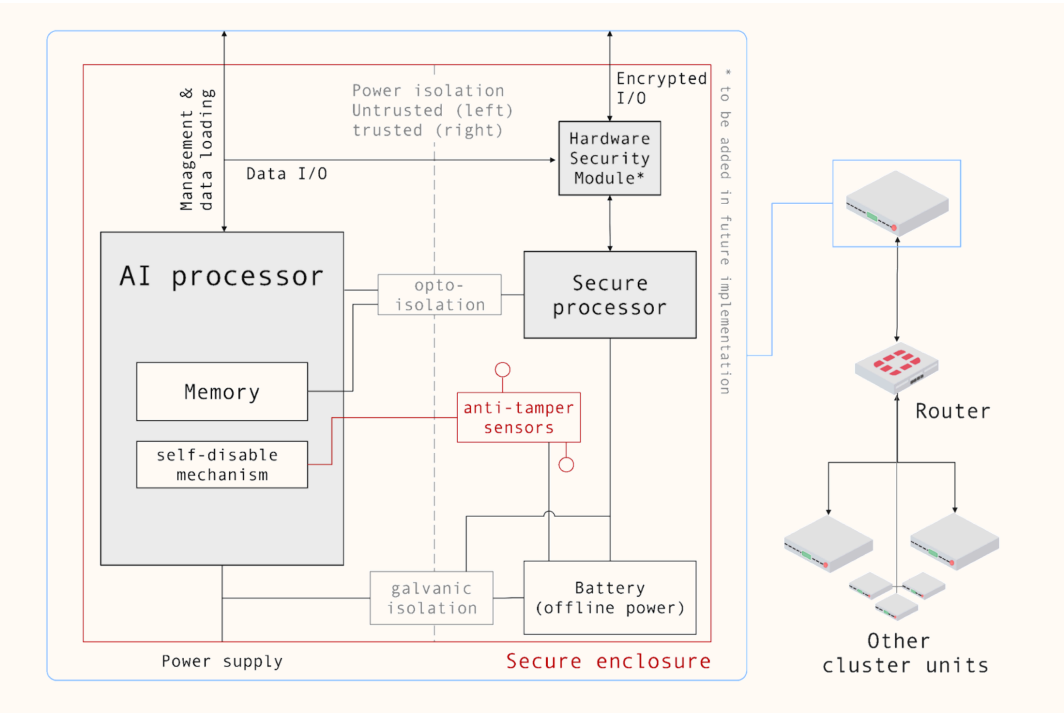
- **情报机构** 美国和中国的情报机构（ICs）可能能够使用间谍、寻找网络漏洞、使用卫星等手段，来判断对方国家是否正在协调一项大型训练活动。这是最脆弱的机制，因为对方国家有可能能够反制情报工作（例如，在一个信息安全管理极其严格的隐藏服务器上训练活动）。

- **计算暂停。** 最容易验证的选项是双方同意不使用他们的GPU。这可以通过检查员确保GPU已关闭来验证。这是最可靠的解决方案，但由于其简单性，但它极其昂贵：两个国家都必须关闭大量的计算资源。此外，双方都会假设对方会在协议违反的情况下，在这里和那里有几个隐藏的集群。在短期内这可能没关系（一个集群的计算能力只有以前5%左右，将只能以大约5%的速度进行研发），但在长期内这将有损于拥有更好走私和情报收集能力的一方。

- **硬件赋能机制（HEMs）。** 美国和中国可以各自将他们所有尖端机器学习计算设备注册给对方。然后，他们各自监控GPU正在运行的计算设备，以确保他们没有违反条约（例如，通过推进能力边界）。这种监控可以通过硬件启用机制（HEMs）来完成，这些机制可以向监控器报告GPU正在做什么的信息。有与篡改有关的技术问题——例如，通过物理损坏HEM，但人工智能有一个计划，大致如下 [FlexHEGs](#) 似乎具有防篡改功能。这需要将HEM和GPU封装在一个安全的盒子里，该盒子验证芯片是否符合政策。总的来说，双方似乎都相信，如果双方都愿意让另一方参与共同设计和监督HEM的安装，并允许定期检查员监控任何设备是否被篡改，那么一个相当激进的验证方案是可行的。像往常一样，双方都将假定另一方在某些秘密数据中心偷偷摸摸地藏匿，违反了条约——但在这个情况下，这并没有那么糟糕，因为人工智能研发受到计算能力的瓶颈限制。这些小型秘密数据中心即使最初运行着更安全、能力更弱的人工智能，也无法与大型的受监控数据中心竞争。

附录S - 国际协议的验证机制

• **AI驱动谎言检测。** 当今的测谎仪效果不太理想。然而，在先进人工智能的帮助下，这个问题或许可以解决，尤其是在人工智能自身以超乎常人的速度和质量进行研发的时候。这可以用来帮助建立国际信任——只需询问政府的相关成员是否违反了条约即可。但是，主要瓶颈仍然是政治上的支持。政治家们不希望测谎仪得到广泛采用，因为这将损害他们撒谎的能力。而且之前的测谎仪完全是安全表演，所以要打赢这场硬仗。



(图来源：FlexHEG报告)

附录T - OpenBrain的新对齐策略

注意：在整个过程中，大部分智力劳动（以及所有编程工作）都是由人工智能完成的。这就是它们能够如此快速进步的原因；如果由数百名顶尖人类研究人员单独完成这项工作，将需要很多年。然而，人类仍然是流程中不可或缺的一部分，因为整个要点就是他们并不完全信任人工智能。因此，他们需要真人来阅读实验报告和安全案例，彼此争论，并希望找出一条安全的道路。

第一步：训练并部署Safer-1，一个有偏差但受控的自主研究人员。它受控是因为对人类监督者来说是透明的：它使用英语思维链（CoT）来思考，并且已经采用了忠实的CoT技术来消除委婉语、隐写术和微妙偏差。

第二步：尝试 Safer-1 的不同训练环境，并仔细阅读 CoT，以确定 Spec 中的目标和原则是“坚持”还是“没有坚持”的方式。

第三步：训练和部署Safer-2，一个基于相同架构但对齐和受控的自主研究员，但这次有一个更好的训练环境，激励正确的目标和原则。

附录T - OpenBrain的新对齐策略

这里有一个可能被纳入更好的训练环境的技术的简要不完整列表：

- 极限 态势感知 在训练的某些部分，为了 对齐伪造 远不太可能。
- 利用 辩论 和其它 可扩展的监督 更可靠地激励真相的计划。
- 放松对抗训练、红队演练和模型生物。
- 在计算预算中花费更高比例用于对齐训练（例如， sorts of 上述所述内容，例如40%而不是1%。

第4步：设计、训练和部署Safer-3，这是一个远比过去更智能的自主研究员，它使用了一个类似旧版Agent-4的更高级的架构。它不再对人类监督者透明，但它对Safer-2透明。所以应该有可能找出如何使它既一致又受控制的方法。

步骤5：重复步骤4直至无穷，创建一个越来越强大、越来越一致的AI链，该链由链中的前一个环节（例如，来自其他场景分支的Agent-5的类）监督。

附录U - 机器人经济翻倍时间

在这个话题上有一个初生的文学。^{*}
对于一个几页篇幅的基本上同意的治疗方法，参见 前瞻报告的“工业爆炸”部分，或阅读我们在自己的世界末日期间，美国和许多其他国家将它们的民用经济转变为全面战争经济。这意味着将生产汽车的车厂转变为生产飞机和坦克的车厂，将原材料从消费品重新导向军事产品，并相应地重新调配交通运输网络。

我们在想象某种类似的事物，但由于超级智能正在指导和管理工作，所以会更快。[†]
大约而言，该计划是将现有工厂改造成大规模生产各种机器人的工厂（这些机器人由超级智能设计，既优于现有机器人又更便宜），然后这些机器人协助建造更高效的新工厂和实验室，这些工厂和实验室生产更大数量更复杂的机器人，这些机器人再生产更先进的工厂和实验室，如此循环，直到机器人经济在所有自由贸易区合起来的规模与人类经济相当（因此需要自行采购原材料、能源等）。到了那时，新工厂将已生产出大量机器人采矿设备、太阳能电池板等，以备不时之需，满足远超传统人类经济所能提供的巨大需求。[‡]

这个新的机器人经济会增长多快？一些参考点：

- 现代人类经济大约每二十年翻一番。发展特别迅速的国家（例如 中国）有时能在不到十年内使经济翻倍。

^{*} 参见例如 这份 Open Philanthropy 报告 和 这个Epoch报告。
[†] 我们的故事描绘了转换过程大约加快了5倍。我们认为这是一个合理的猜测，考虑到瓶颈等因素，如果一百万个超级智能体在协调它，这个转换过程可能有多快。当然，我们非常不确定。
[‡] 可能还包括更高级的能源，例如核聚变能。

附录U - 机器人经济翻倍时间

• 一家现代化的汽车工厂在不到一年的时间里生产的汽车大约重于自身。
或许，一个完全由超级智能控制的机器人经济，在不耗尽原材料的情况下，可能不到一年就能自我复制。
† 然而，这似乎可能是一个戏剧性的低估。植物和昆虫的“翻倍时间”往往远少于一年——有时甚至只有几周！也许最终机器人会变得如此复杂，制造工艺如此精巧，设计如此完善，以至于机器人经济可以在几周内翻一番（当然，这假设有现成的原材料）。

然而即使那也可能是低估。植物和昆虫正在许多 • 超级智能设计者所没有的约束。例如，它们需要以自我包含、能够自我复制的生物形式存在，而不是一个由多样化且更专业化的车辆和工厂来回运输材料和设备的经济体。此外，细菌和其他微小的生物在数小时内就能繁殖。最终，自主机器人经济体可能会看起来更像是例如一种新的不被消化的藻类，这种藻类遍布地球的海洋，数量翻倍 每天两次 以便它在两个月内覆盖整个海面，并伴随着一个将藻类转化为更有用产品的捕食者物种生态系统，这些产品又作为营养提供给浮动机器厂，这些浮动机器厂生产火箭等大型结构，以及更多的浮动机器厂。

显然，所有这一切都很难预测。这就像让蒸汽机的发明者猜测一个现代汽车工厂生产相当于自身重量的汽车需要多长时间，并且猜测这样一个工厂首先存在需要多长时间。但是经济增长率已经加速了 多个数量级 在...上 人类历史进程，而且在我们看来，在超级智能之后，它们将加速数个数量级。我们的故事描绘了在几年内经济增长加速了大约1.5个数量级。
‡

附录V - 那么谁将统治未来？

2028年，监督委员会控制着人工智能。但他们允许2028年的选举基本公平，人工智能被对称使用。

这种状态——监督委员会掌握硬权力但却不怎么干预民主政治——不可能无限期地持续下去。默认情况下，人们最终会意识到对人工智能的控制赋予了监督委员会巨大的权力，并要求这项权力应归还给民主机构。迟早，监督

快速纸巾计算： 帝国大厦 其面积为2.77平方米，重量为365千吨。上海超级工厂 45平方米 和 产生750k 每年生产的车辆，主要是Model 3和Model Y，每辆重约两吨。估计帝国大厦的单位面积质量比上海超级工厂更高（因为它垂直而非水平且需要更强的支撑），所以如果有什么的话，这是低估了。因此，看来一个可能重不到一百万吨的工厂每年正在生产一百万五千吨的汽车。
† 我们不认为它会耗尽。最初，机器人经济将依赖于人类矿山获取材料。但当它超出这些来源的供给时，数百万的超级智能将已勘探到新的矿山并开发出开采它们的新技术。例如，想象一下海底采矿机器人剥离海底稀有矿物，或者新的化学加工途径更高效地将地面露天矿的原始矿石转化为有用原材料.....
‡ 如果当前经济每二十年翻一番，那么每两年翻一番就是快一个数量级，每0.2年翻一番就是快两个数量级，以此类推。上述假设的超级智能藻类经济将比当前人类经济快大约四个数量级。

附录V - 那么谁将统治未来？

委员会要么必须放弃其权力——要么积极利用其对人工智能的控制来破坏或终结民主，这可能会在权力斗争中清洗掉一些掌权成员。

如果他们选择后者路线，他们可能会 **锁定** 无限期地维持他们的权力。

哪种情况会发生？委员会是否放弃其对硬实力的垄断，或者他们保持它？两种未来都是可能的，所以让我们探索每条道路。

委员会如何可能最终放弃其权力？

- 一些委员会成员可能更喜欢权力广泛分配的未来，他们也可能处于推动其愿景的有利地位。例如，如果一些委员会成员策划民主颠覆，亲民主成员可以向媒体或国会举报。如果得到通知，国会可能会要求人工智能应由更民主的机构控制，例如国会本身。

- 如果国会处处受到所有部署在政府、产业和军事领域的AI的反对，那么国会将无能为力。但如果委员会分裂，那么AI就不会只服务于一方，国会就能发挥真正的影响力。面对公开冲突，更多的委员会成员可能会倾向于放弃一部分权力，不愿公开捍卫不那么民主的一方。

因此，对人工智能的控制可能会超出委员会范围，扩展到国会。这将会 • 已经有进展了，因为在更大的群体中，更有可能有一群人关心局外人并考虑他们的利益。而且一旦权力扩展到国会，它可能会持续扩张——有可能完全回归公众。

†

但是监督委员会也可能为自己夺取权力：

- 有些有权势的人对于这种事情毫无道德上的顾忌——而且他们自己也知道这一点。此外，有些人野心勃勃、贪恋权力，如果他们预期最终能胜出，就愿意挑起与民主的冲突。如果其他委员会成员提出反对，他们可能会被清洗、被否决，或者被给予一些较小的让步。

‡

- 此外，很多时候，有权势的人在崛起的过程中做了非法或不道德的事情。他们可能会担心，如果权力得到更广泛的分配，他们自己的地位就会动摇，就像那些藏在衣柜里的骷髅被擅长问正确问题的超级智能侦探发现一样。

此外，通过获取超级智能，监督委员会也可能拥有最 • 历史中通往力量的便捷路径。Safer-∞ 可能会预测某些策略具有极低失败风险。并且 Safer-∞ 也可能提供在其他方面便捷的策略，例如非暴力（正如 Safer-∞ 可以组织

•

为什么我们期望人们最终会理解监督委员会有多少权力？一个原因是现在情报如此便宜：默认情况下，人们应该能够开发出强大的AI来帮助他们调查和理解谁在掌控他们的国家。委员会可以通过限制对这类AI的访问来阻止这种情况发生，而只允许人们访问那些掩盖了委员会权力真实程度的AI。但如果委员会决定编织这样复杂的谎言网，并永久性地限制人类对真正诚实的超级人工智能（以及其他用于揭示真相的工具，例如人类智能增强）的访问——那么我们就将其视为破坏民主。

†

例如：如果国会中的一些人想要夺取权力，其他人或许能够推迟到下一次选举，届时公众能够发表意见。

‡

例如，他们可以稍微多分一些权力，如果他们认为是合适的话，可以将这些权力重新分配给更大的群体。这种类型的某些让步可能开始逐渐转变为明显民主的结果，即使某些精英比其他人掌握着更多的权力。

<think>
Okay, the user wants me to translate some markdown text into simplified Chinese. Let me first read through the source text carefully. The main points are about a "bloodless coup in China." The user emphasized not adding any extra text, so I have to be precise. Let me start with "a bloodless coup in China." "Bloodless" here is "无血的" in Chinese. "Coup" is "政变," and "in China" is "在中国." So that part would be "在中国的无血政变."
Next part: "or perhaps even superficially democratic, if Safer-∞ could manipulate public opinion to always align with the Committee's wishes." The term "superficially democratic" is "表面上民主的." "Safer-∞" is a proper noun, so it should stay as is. "Manipulate public opinion" translates to "操控舆论." "Always align with the Committee's wishes" would be "始终符合委员会的意愿."
我们并不支持这个放缓结束中的许多行为，并且认为它对技术对齐的假设过于乐观。我们在比博姆莱方面也不支持许多行动。甚至表面上民主的，如果Safer-∞能够操控舆论始终符合委员会的意愿。"
在我们撰写此情景时，我们的一个目标是征求比我们更乐观的人们的关键反馈。成功看起来是什么样子的？这个放缓结束情景代表了我们关于如何能够成功地最好猜测。混合了凭借运气的因素，特别是激烈的技术对齐努力以及正确的人在权力斗争中获胜的组合。它并不代表一个我们实际上认为应该追求的计划。但许多人，包括最值得注意的是Anthropic和OpenAI，似乎都在追求类似的东西。
我们很乐意看到他们澄清他们的目标！例如，如果他们能够勾勒出一个可行的场景，起点可以是现在，或者从我们某个部分分叉出来。也许甚至表面上民主的，如果Safer-∞能够操控舆论始终符合委员会的意愿。"
Yes, that looks correct. No extra words, just the translation. I think that's it. Let me confirm once more that all terms are accurately translated and the structure is the same as the original.
</think>
(在中国的无血政变)，或者可能甚至表面上民主的，如果Safer-∞能够操控舆论始终符合委员会的意愿。

事实上，可以说他们大多数的目标是更像“种族”结局，只是他们认为这会没事，因为人工智能从一开始就不会错位。根据与前沿人工智能公司员工的个人对话，看来他们大多数都觉得根本不需要减速。