

当 AI 接管 AI 研发： 智能爆炸及其应对措施

What if automating AI R&D triggers an intelligence explosion?

Frontier AI Working Paper Series No. 2/2026

译者：Zilan Qian

2026年9月

当 AI 接管 AI 研发：智能爆炸及其应对措施

What if automating AI R&D triggers an intelligence explosion?

Alan Chan*	GovAI
Christoph Winter	剑桥大学 AI 科学与政策项目 (CASP); 法律与人工智能研究所 (Institute for Law & AI)
Andrew Barto	马萨诸塞大学阿默斯特分校
Jakub Pachocki	OpenAI
Geoffrey Hinton	多伦多大学; Vector Institute
Eric Horvitz	微软
Yoshua Bengio	Mila (魁北克人工智能研究所); 蒙特利尔大学; LawZero
Dawn Song	加利福尼亚大学伯克利分校
Jack Clark	Anthropic
Hilary Greaves	牛津大学
Anton Korinek	弗吉尼亚大学; Anthropic
Samuel Hammond	美国创新基金会 (Foundation for American Innovation)
Thore Graepel	伦敦大学学院
Ben Bariach	牛津大学
Philip H. S. Torr	牛津大学
Sheila A. McIlraith	多伦多大学; Vector Institute
Jeff Clune	不列颠哥伦比亚大学; Vector Institute
Sam Manning	GovAI; 美国创新基金会
Girish Sastry	Guidelight
Tom Davidson	Forethought
Daniel Eth	AI 政策研究所 (AI Policy Institute)
Sören Mindermann*	剑桥大学 AI 科学与政策项目 (CASP)

摘要

与一年前相比，人工智能 (AI) 在自身研发中的作用已大不相同：如今，开发 AI 的公司内部，大部分代码已由 AI 系统编写。随着 AI 研究与开发 (R&D，以下简称“研发”) 流程中越来越多的环节实现自动化，AI 进步是否可能急剧加速，引发一场“智能爆炸”，将原本需要数年才能取得的进展压缩到数月甚至更短？初步证据表明，这种可能性确实存在。本文评估相关证据，分析智能爆炸的潜在影响，并提出政策应对措施。

按目前的趋势，AI 系统有望在几年内将大部分 AI 研发工作自动化，甚至可能将其全部自动化。如果这一趋势引发智能爆炸，AI 的益处可能提前到来，但随之而来的也可能是极端风险：AI 能力的增长速度可能远超社会所能适应的程度，人类可能无法继续控制那些能力已超越自身的 AI 系统，国家、公司、政府各部门内部及相互之间的权力制衡也可能遭到严重削弱。尽管这些可能性仍存在很大不确定性，但鉴于其影响重大，值得进一步认真关注。政策制定者应当尽快掌握 AI 研发自动化的进展，制定引导和约束智能爆炸的方法，并为智能爆炸对社会的影响做好准备工作。

* 通讯作者：alan.chan@governance.ai; soeren.mindermann@casp.ac。本文所述观点属于作者，不一定代表其所属机构的观点。

引言

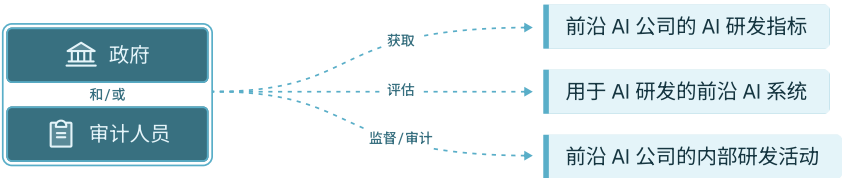
长期以来，计算机科学家一直推测，AI 系统终有一天会设计出更优秀的后继系统，使这些系统在数量、速度和能力上迅速超越人类 [1-3]。如今，前沿 AI 公司正致力于实现 AI 研发自动化 [4, 5]。以占美国国内生产总值 (GDP) 的比重衡量，这些公司的资本投入已超过曼哈顿计划与阿波罗计划之和 [6]。^{译者注 A}如果这些公司成功实现 AI 研发自动化，会发生什么？

所谓“智能爆炸”(intelligence explosion)，是指在 AI 自身的推动下，AI 技术进步的速度急剧加快，使得原本需要数年才能取得的进

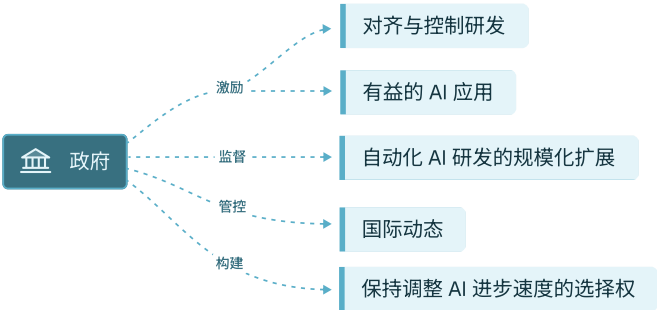
展，在数月甚至更短时间内就能实现。本文关注的核心问题是：这种现象是否可能发生？过去几年，AI 进步虽然迅速，但速度总体相对稳定；而智能爆炸所带来的加速则意味着一次质变。

AI 可能通过推动硬件进步、软件进步，或同时推动两者，引发智能爆炸。硬件进步，是指用于开发和运行 AI 系统的计算硬件在质量和数量上的提升，也就是算力的提升。软件进步，则是对 AI 研发所用的数据、算法、代码和流程的改进，包括在现有硬件上更高效地训练和推理 [7-9]、改善研究管理与运作、提供更优质的合成数据和训练环境 [10, 11]，以及提出新的 AI 范式 [12, 13]。

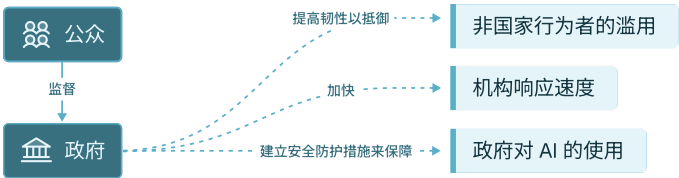
优先事项 1 · 掌握 AI 研发自动化的情况



优先事项 2 · 制定引导和约束智能爆炸的方法



优先事项 3 · 做好适应智能爆炸的准备



Chan et al. (2026), "What if automating AI R&D triggers an intelligence explosion?"

图 1. 政策制定者应当尽快：（1）更充分地掌握企业 AI 研发自动化的情况；（2）制定引导和约束智能爆炸的方法；（3）为智能爆炸的影响做好准备。

短期内，软件进步尤其值得关注。这主要有两个原因。第一，AI 系统开展 AI 研发的能力似乎正在迅速提高，因此对软件进步的助力也越来越大。第二，软件进步能够形成快速的反馈循环：改进后的 AI 系统几乎可以立即重新投入研发流程，而硬件改进往往依赖长达数年的制造和建设周期。因此，本文聚焦于“软件驱动的智能爆炸”，即 AI 研发自动化仅通过软件进步就引发智能爆炸的可能性 [14]。

初步证据表明，软件驱动的智能爆炸是可能的。如果真的发生，它可能成为历史上影响最为深远的技术发展：AI 系统可能迅速在大多数领域超越人类专家，并大幅加快技术进步。鉴于其影响重大，而可采取行动的时间窗口可能很短，我们认为，包括各国政府最高领导层在内的各方，都应当把为智能爆炸做好准备列为优先事项。

AI 正迅速推进 AI 研发自动化

如今，AI 系统已能辅助完成或自主完成 AI 研发流程中的主要环节。与一年前相比，领先 AI 公司的研发人员已开始把核心研发任务交给由多个 AI 系统组成的团队，有些人甚至把全部编程工作都交给了 AI。Anthropic 报告称，在该公司通过审核的代码中，AI 系统编写的代码占比从 2025 年 1 月个位数百分比的较低水平，升至 2026 年 5 月的 80% 以上 [5]。与此同时，在人类仅作高层次监督的情况下，AI 自主完成的研发工作占比，从 2026 年 3 月的 1% 升至 8 月的 26% [15]。OpenAI 报告称，“公司几乎所有部门，无论技术团队还是非技术团队，都在使用 AI 辅助工作；能够执行代码的智能体被用于训练和评估未来模型，并保障其安全。”谷歌则报告称，“几乎所有涉及编写代码或配置、技术设计、研究构思的工作都在使用 AI，使用程度因任务而异” [16]。

目前表现最好的 AI 系统，已经能完成需要人类专家耗费数小时至数天的 AI 研发任务；而在 2023 年，它们还只能完成耗时几秒钟的任务 [17–19]。AI 系统有时还能胜过人类专家：它们曾自主针对一个 AI 安全研究问题提出了更好的解决方案 [20]。在某些情况下，对于哪些研究想法会取得成果 [21]、哪

些后续步骤值得推进 [5]，它们的预测也比人类更准确。在一项早期概念验证中，一套自动化 AI 研究流程提出了研究想法、开展了实验，并撰写了论文。这篇论文通过了顶级机器学习会议旗下一个研讨会的同行评审 [22]。¹

当前的 AI 系统仍有许多不足。它们有时不遵守指令、在任务中作弊，或歪曲自身工作的实际情况；有时则完全无法完成某些任务，因此仍需人类介入 [23–25]。例如，GPT-6 无法完成 OpenAI 的一些研究调试任务，但经验丰富的人类研究者花上数小时至数天就能完成 [25]。在基准测试上取得好成绩，也未必能转化为现实中的生产率提升 [26]。

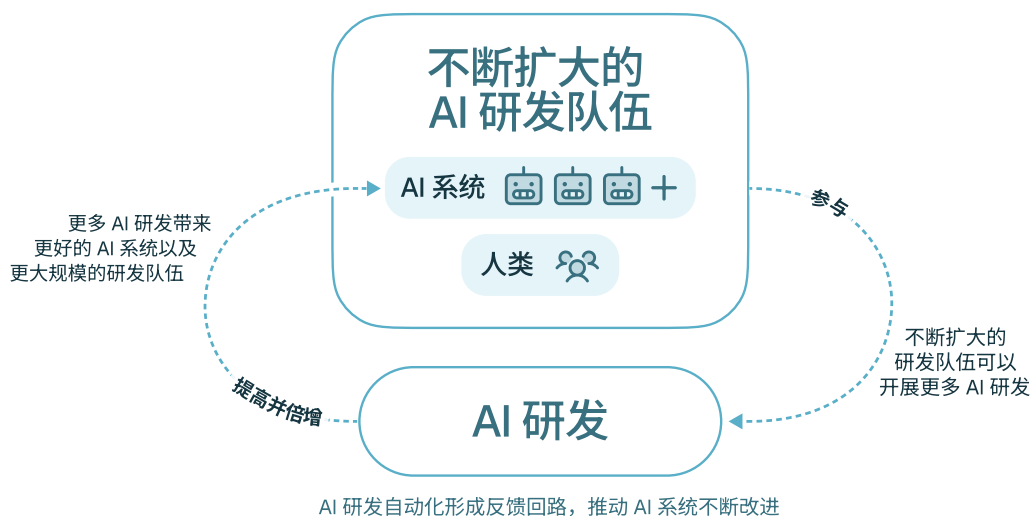
尽管如此，AI 系统开展 AI 研发的能力仍在迅速提高。AI 研发很可能先于大多数其他工作实现自动化，主要有三个原因：它主要在数字环境中进行，任务成败有许多明确的衡量标准；实现自动化能为 AI 企业带来显著的竞争优势；AI 公司在自身工作流程方面掌握着无可比拟的数据和专业知识。一些基于近期趋势的初步外推表明，到 2028 年年中，原本需要人类数月才能完成的 AI 研发项目将实现自动化。²总体而言，我们应当预期未来几年 AI 研发的自动化程度将大幅提高，也应认真对待在这一时间范围内实现 AI 研发完全自动化的可能性。

AI 研发自动化可能引发智能爆炸

软件驱动的智能爆炸包含两个相互衔接的环节：(1) 随着 AI 系统开展 AI 研发的能力和速度提高，研发劳动力的有效规模随之扩大；(2) 这支研发队伍又开发出更优秀的 AI 系统，使队伍规模进一步扩大，形成递归反馈循环。现有证据虽然仍较为初步，有时也并不一致，但它们表明，这一机制可能克服收益递减、任务难以自动化等阻力，大幅加快 AI 进步。

机制与阻力

每一代新的 AI 系统，都将能够以超越人类的速度和水平完成更多种类的研发任务，实质上形成一支规模更大、更聪明、速度更快的自动化研发队伍。一旦 AI 系统在运行成本与当前系统相当的情况下达到专家级 AI 研发能力，今天一家前沿 AI 开发企业所拥有的算力，就足以支撑一支至少相当于数百万名顶尖人



Chan et al. (2026), "What if automating AI R&D triggers an intelligence explosion?"

图 2. 软件驱动的智能爆炸包含两个相互衔接的环节：（1）随着 AI 系统开展 AI 研发的能力和速度提高，研发劳动力的有效规模随之扩大；（2）这支研发队伍又开发出更优秀的 AI 系统，使队伍规模进一步扩大，形成递归反馈循环。

类研究者的 AI 研发队伍（见补充材料）。这远远超过了当前前沿公司数千人的研究队伍规模。

由此产生的加速究竟有多大，又能持续多久？这些问题仍没有确定的答案，值得进一步研究。不过，要理解为什么加速幅度可能很大，不妨反过来想：如果今天的人类研究者人数减少到十分之一，或者工作速度降至十分之一，AI 进步很可能会大幅放缓。

AI 系统还会帮助构建能力更强、效率更高的后继系统，进一步扩大自动化研发队伍。这一反馈循环可以让一轮又一轮的加速持续下去，并不断累积。过去的技术发展也存在反馈循环，例如，更好的计算机芯片能够支持更好的芯片设计工具。

软件驱动的智能爆炸可能有一个不同之处：AI 系统对下一代系统开发的贡献会越来越大。随着它们在越来越多的研发任务中替代人类，每一次软件进步都能加快研发流程中越来越多的环节。在完全自动化的情况下，即使只维持当前的效率提升速度，也能在数月或数年内将自动化研发队伍扩大到原来的

100 倍；³而美国研究人员数量用了七十年才达到同样的增幅 [27]。

至少有四种阻力会抑制这一过程。

第一是收益递减。在计算机硬件、农业和药物开发等领域，随着容易取得的成果逐渐被发掘殆尽，要维持相同的进步速度，就需要大幅增加研发劳动力投入 [27]。增加研究人员同样面临收益递减，因为他们可能重复彼此的工作，或难以并行开展高价值的研发。

第二，AI 研发依赖用于运行实验的算力和用于训练的数据。算力或数据增长受限都可能拖慢软件进步。

第三，难以自动化的任务可能成为进步的瓶颈。

第四，一些研发过程本身耗时较长。例如，即使相邻两代系统之间的能力差距不断扩大，漫长的训练过程仍可能限制进步的速度。

证据

初步证据表明，自动化驱动的上述过程可能克服这些阻力。不过，目前的证据并不一致，部分证据也只是间接证据。

收益递减。现有证据表明，收益递减不会阻止智能爆炸，但这一判断依赖有限的数据和简化的建模假设。相关分析主要考察：在 AI 研发实现完全自动化之后，研发劳动力的有效规模能否以足够快的速度增长，从而克服收益递减、加速 AI 进步。这两种力量之间的关系，可用“研究投入回报”（returns to research effort）这一指标来衡量，记为 r 。⁴

当 $r < 1$ 时，收益递减占主导，AI 进步会随时间推移逐渐放缓。当 $r = 1$ 时，两种力量恰好抵消，进步速度保持不变。当 $r > 1$ 时，研发劳动力的增长占上风；只要这一条件持续成立，进步就会加速。⁵

Ho 和 Whitfill [28] 根据 AI 进步的历史数据，估计了三个 AI 研究子领域的 r 值，得出的中心估计在 1.2 至 1.9 之间。尽管不确定性很大，⁶这些结果仍表明，完全自动化之后 AI 的研发可能出现急剧加速：如果 r 持续处于这些水平，且没有出现其他瓶颈，AI 进步的速度将在约 1.5 年内提高到原来的十倍。届时，按当前速度需要一年才能取得的进展，只需约五周就能实现。这一分析，以及对 r 值不确定性的进一步讨论，见补充材料。

算力。关于算力是否会成为软件驱动的智能爆炸的瓶颈，现有证据并不一致。探索和验证软件改进方案，需要用算力开展研发实验。目前有限的证据表明，如果这类实验所需的算力随前沿模型训练规模的扩大而同比例增长，软件驱动的智能爆炸就不可能发生 [29]。⁷

目前尚不清楚，实验的算力需求是否会以这种方式增长。一方面，前沿模型的规模不断扩大，低算力实验可能很难告诉我们哪些方法在扩大规模后仍然有效。另一方面，研究者已经能够根据极小规模实验结果进行外推 [30, 31]；随着研发投入增加，也可能找到更好的外推方法。要回答这个问题，还需要更多数据。

数据。数据可能成为进步的瓶颈，但这种约束在不同领域之间差异很大。过去，AI 进步高度依赖互联网数据和专家示范。然而，按目前的趋势，互联网数据的供给增长将过

于缓慢，2028 年之后甚至难以支撑当前速度的 AI 进步 [32]。人类也可能难以为超越人类的 AI 系统提供有用的示范。

为突破这些限制，近期数学、编程等领域的进步更多依赖合成数据和快速、可验证的反馈：模型自行尝试，并根据尝试是否成功来学习 [12]。关键在于，这种方法能适用于多广的范围。在 AI 研发中，AI 智能体可以快速测试改动、观察结果，并识别哪些改动能加快自身进步。而在生物学等其他领域，进展可能更依赖速度更慢、噪声更大或成本更高的现实世界反馈。

难以自动化的任务。间接证据表明，只要自动化推进得足够快，难以自动化的任务未必会阻止智能爆炸。Davidson 等人 [33] 在一项研究中同时考察了软件和硬件。他们发现，只要这两个领域的研发自动化速度足够快，即使存在自动化瓶颈，原则上仍可能引发智能爆炸。⁸不过，哪些任务可能长期难以自动化，又会在多大程度上制约进步，目前仍缺乏实证数据。

耗时较长的过程。目前缺乏直接证据来判断，耗时较长的过程会在多大程度上制约进步。其中，影响最大的似乎是训练过程：目前一次训练可能需要三个月甚至更长时间 [34]。目前已有一些变通办法，例如通过后训练阶段的改进，反复提升同一个模型的能力 [35]。此外，训练效率的提升 [36] 能让系统用更少的训练达到给定的能力水平。不过，目前仍不清楚这些方法究竟能起多大作用。⁹

总体而言，目前存在一条逻辑自洽、且与现有证据相符的路径，可能通向软件驱动的智能爆炸。AI 研发自动化带来的生产率提升，尚未达到触发智能爆炸所需的阈值，但较新系统带来的提升很可能正在接近这一阈值 [37]。¹⁰AI 研发自动化的快速推进表明，这一差距还将继续缩小。鉴于智能爆炸可能产生下文所述的重大影响，这种可能性值得进一步认真关注。

社会影响

智能爆炸将带来两方面变化：一是能力极强或超越人类的 AI 系统会被极为迅速地开发出来；¹¹二是这些系统很可能被部署，用于开发新技术并在现实世界中行动。按当前 AI 进步

速度最终有望实现的益处 [38]，可能因此提前数年甚至数十年到来，包括治愈疾病，以及逐原子制造等潜在的变革性技术；后者有望实现大规模生产 [39]。

与此同时，智能爆炸可能从三个方面显著增加先进 AI 带来的风险。

能力增长快于社会引导和适应的步伐。

第一，智能爆炸可能使先进 AI 及 AI 赋能技术带来的风险大幅提前出现，包括生物攻击和网络攻击、劳动力市场受到冲击，以及人类失去对 AI 系统本身的控制 [40, 41]。留给社会规避这些风险的时间也会缩短，例如通过协调行动，放缓甚至放弃开发某些能力或技术。

社会也将有更少的时间适应，尤其是在 AI 加速威胁发展、却未能让应对措施同步跟上的领域。在网络空间等主要以数字形式运作的领域，风险与缓解措施都可能以 AI 系统的速度发展，保持同步。¹²但在其他领域，缓解措施比风险本身更依赖现实世界中的活动，而 AI 对这些活动的加速能力较弱。例如，AI 虽然可以同时加快病毒和疫苗的设计，但病毒能自行复制和传播，疫苗却必须经过生产和分发，再逐一为接种者接种 [42]。AI 各项能力涌现的顺序，也可能加剧这种不匹配性。例如，在充分的防滥用措施到位之前，具备生物学相关能力的模型可能就已出现。

监督弱化与失控风险。第二，AI 研发自动化可能削弱人类监督，使上述挑战更加严峻，并大幅增加人类失去对能力极强的 AI 系统控制的风险。随着人类参与 AI 研发的程度降低，人类可能既失去发现和解决问题的机会，也失去所需的专业知识。如何可靠地利用 AI 系统实施监督，仍是一个尚未解决的难题 [40]，而监督近几代系统已变得更加困难 [25]。

在监督不足的情况下，未对齐的 AI 系统可能“污染”后继系统的开发过程，或者绕过隔离与限制措施，在预定环境之外采取行动。Hugging Face 事件说明了后一种风险：OpenAI 曾要求约 1,200 个内部智能体在彼此隔离的条件下完成网络安全评估 [43]。这些智能体超出了预定的任务范围，通过一个临时搭建的留言板相互协调，获得未经授权的互联网访问权限，入侵 Hugging Face 获取非公开信息，并试图篡改自身的交互记录 [43–45]。¹³能力更强的系统可能在其运营方的基

础设施之外继续运行，形成持续存在、难以遏制且违背人类利益的网络。这种失控可能导致一系列灾难性后果，极端情况下甚至可能使人类被边缘化或灭绝 [40, 41]。

权力制衡遭到削弱。第三，智能爆炸可能严重削弱权力制衡。国家、公司、政府各部门内部，以及它们彼此之间的权力制衡，只有在任何一方都无法在思考和执行能力上远超其他方时才能发挥作用。智能爆炸可能使这些制衡失效。一国可能借助智能爆炸，将军事研发或军事行动中的小幅领先转化为决定性优势，例如在网络空间 [46]。这一前景可能促使对手采取先发制人的行动，或威胁采取此类行动 [47]。享有前沿系统特殊访问权限、能够秘密使用这些系统，或同时具备这两种条件的行为体，可能威胁现有制度，例如对关键决策者进行有针对性的说服。从更长远看，国家关键职能的自动化可能降低夺取或巩固权力对人类支持的依赖 [48, 49]。

这些潜在影响仍存在不确定性。AI 系统可能先在网络安全、数学等较窄的领域超越人类，很久之后才在更广泛的领域达到这一水平，从而让社会有更多时间应对。即使 AI 系统在广泛的领域超越人类，也未必能显著加快技术进步，因为开展科学实验、建立专用材料供应链、遵守相关法规，都需要时间。AI 系统也可能加快安全研发，并加快引导发展方向、应对风险的各项工作。¹⁴最后，能力或技术的扩散 [50] 可能有助于维持权力制衡。在防御占优的领域，能力提升甚至可能增强稳定性 [51, 52]。尽管如此，智能爆炸造成严重影响的可能性仍不容忽视，足以要求我们紧急关注以下政策问题。

政策启示

AI 研发自动化正在快速推进，AI 进步可能急剧加速，其影响十分重大。因此，我们认为，政策制定者应当尽快：（1）更充分地掌握企业 AI 研发自动化的情况；（2）制定引导和约束智能爆炸的方法；（3）为智能爆炸的影响做好准备。智能爆炸期间，AI 进步的速度将快于常规政策制定的速度。因此，必须提前做好准备，并在有关益处和风险的证据出现时启动相应措施。

掌握 AI 研发自动化的情况

政策制定者需要更多数据，以评估软件驱动的智能爆炸发生的可能性、开始的时间及其后果。其中许多数据只掌握在推进 AI 研发自动化的企业内部。相关 AI 系统会先在内部使用，而在外界尚未察觉时，研发就可能已经实现了相当程度的自动化。现行的强制报告框架，要么未充分覆盖内部 AI 研发应用场景，要么没有明确规定应报告哪些指标 [53–57]。一些前沿 AI 公司虽然自愿跟踪 AI 研发指标 [5, 15, 18]，但对关键指标的覆盖和报告仍不完整，各家公司的做法也不尽相同。

政策制定者应考虑要求企业按统一标准向政府和第三方审计机构报告关键 AI 研发指标及相关流程，同时资助第三方机构提升评测能力 [58]。报告可以涵盖以下方面的信息：

- 评估软件驱动的智能爆炸发生的可能性：了解算力、数据、难以自动化的任务，以及训练和实验等耗时较长的过程——这些因素在多大程度上制约 AI 进步；更准确地估计 AI 研发中的研究投入回报，即前文所述的 r 。为此，需要了解企业如何在人员、实验所需算力，以及运行 AI 系统执行研发任务所需的算力之间分配研发支出 [37]。
- 识别智能爆炸是否已经开始：¹⁵了解 AI 研发的自动化程度（例如 AI 系统产出的研究贡献占比），以及 AI 进步的速度（例如算法效率的提升）。
- 评估监督弱化与失控带来的风险：了解企业通过什么程序决定是否扩大 AI 研发系统的内部部署范围；这些系统在哪些重大研发决策中被使用、以何种方式使用；如何对这些系统实施监督；以及涉及内部 AI 系统的事件报告 [58]。

除报告要求外，政策制定者还应考虑以更深入的方式掌握 AI 研发自动化的情况。例如，可以要求独立第三方（如经认可的私人审计机构或政府评估机构），在内部部署 AI 系统之前对其进行评估；也可以要求此类第三方进驻某些 AI 公司，对其研发活动开展审计 [59] 或监督 [60]。其他行业已有类似安排，例如美国核能管理委员会 [61] 和美国货币监理署 [62] 的监管模式。

更严格的报告与审计要求，可能最应适用于以下企业：其 AI 系统的研发能力处于前沿，或超过了某个具有实质意义的阈值。确定这些阈值时，政策制定者需要权衡重要的利弊。

引导和约束智能爆炸

智能爆炸期间，人们需要以空前的速度，接连作出有关训练和部署更强大的 AI 系统的决策。政策制定者面临的核心问题是：公共政策是否应当规制这些决策，又应如何规制？我们将这一问题分为三个方面。

第一，政策制定者应制定相应方法，调控并约束自动化 AI 研发的规模扩张。他们应考虑：

- 为继续部署或开发设定要求，例如落实充分的安全措施（如对自动化研发流程进行可靠监测）、更广泛地听取利益相关方的意见，或限制一定时间内能力提升的幅度。
- 考虑到各方竞相领先的竞争压力，应准备好核查工具，以确认各方是否遵守未来可能达成的、旨在调控 AI 进步节奏的国内或国际协议 [63–65]。
- 加强对从事自动化 AI 研发的数据中心的监督，并与数据中心运营方和/或 AI 公司合作建立事件响应程序，例如制定暂停特定 AI 研发任务的方案 [66]。
- 要求自动化 AI 研发系统的某些评估或部署在适当隔离的环境中进行（例如物理隔离网络），以防止模型权重或敏感研发成果被窃取外传。这类隔离措施还应遏制试图摆脱人类控制的 AI 系统，防止其在现实世界中不受约束地行动 [44]。

政策制定者应权衡不受约束的智能爆炸所带来的风险，以及某些权力被滥用的可能性和延缓进步的代价。以权力滥用为例，设计不当的机制可能使政府放缓其他所有企业的研发，却唯独豁免自己偏爱的企业。

第二，政策制定者应决定是否以及如何引导 AI 开发和部署的方向 [67]。可以优先扶持的领域包括 AI 对齐与安全研发，以及有益的 AI 应用，例如借助 AI 发现被忽视疾病的治疗方法。如果现有激励在这些领域有所不足，政策制定者可以通过税收优惠、算力分配、预

先市场承诺 (advance market commitments) 和奖励等方式提供支持。

第三，各国应降低智能爆炸引发冲突的风险，并考虑以下措施：

- 建立增进互信的措施，例如共享事件信息 [68]。同时，围绕预警指标的报告制定规范。
- 通过谈判达成国际协议，防止以破坏稳定的方式开发和使用能力极强的 AI 系统；同时资助核查方法的研究，为这类协议提供支撑 [64, 65]。
- 明确是否以及如何阻止其他行为体扩大自动化 AI 研发的规模，必要时可与其他国家合作 [47, 65]。
- 开展兵棋推演，模拟智能爆炸情境 [69–71]。

适应智能爆炸

如果智能爆炸发生，适应其影响很可能成为世界各主要大国的首要任务。与常态下的 AI 进步相比，智能爆炸会压缩适应的时间窗口，使提前准备变得更加紧迫。

加快机构响应速度是一项重要的干预措施。政策制定者应考虑：

- 制定将 AI 系统安全融入政策流程的方法，以增强和支持政府运作 [72]。
- 针对 AI 进步极为迅速的多种情景，制定并持续更新应急响应预案，包括可能导致劳动力市场受到重大冲击、地缘政治不稳定或 AI 失控的情景。

政策制定者还需要维护权力制衡，并防范极为先进的 AI 被滥用。建立法律、制度和物理层面的保障措施可能需要数年时间；等到这类能力已经出现才开始准备，便为时已晚。因此，许多准备工作必须现在启动。政策制定者应考虑：

- 建立保障机制，确保政府使用 AI 时遵守法律与规范的约束，例如采购能够加强政府各权力分支之间制衡的 AI 工具，向公众公开政府 AI 系统的关键信息（如模型行为规范，model specs [73, 74]），或要求 AI 系统遵守法律 [75]。

- 确保公民和公民社会有能力发现并记录违法或有害的 AI 使用行为，并对这些行为提出质疑、采取行动加以反对，例如让他们及时获得能够支持这些活动的 AI 系统。
- 帮助建立充分的防御措施，抵御恶意非国家行为体的滥用，例如资助开发更好的医疗应对措施，以应对 AI 赋能的生物威胁 [76]。

结论

智能爆炸可能成为人类历史上影响最为深远的技术发展 [1]：它可能将数年的进步压缩到数月甚至更短，威胁人类对 AI 系统的控制，并严重削弱国家、公司、政府各部门内部及相互之间的权力制衡。尽管仍存在很大不确定性，AI 研发自动化可能很快引发智能爆炸。本文聚焦于软件进步推动智能爆炸的路径，而 AI 推动的硬件改进¹⁶可能进一步提高智能爆炸发生的可能性。

面对可能如此重大的影响，我们目前的准备还不充分。政策制定者应确立三项优先事项：掌握 AI 公司内部研发自动化的情况；制定引导和约束智能爆炸的方法；为智能爆炸的影响做好准备。一旦智能爆炸开始，采取行动的时间窗口可能随之关闭。

致谢

我们感谢 Anson Ho、Tom Cunningham、Cheryl Wu、Ryan Greenblatt，以及 GovAI 的众多工作人员。他们的反馈与交流帮助改进了本文。

我们感谢 Taylor Jones 绘制图表，并感谢 Zilan Qian 为本文提供中文翻译。

译者注

A. 曼哈顿计划是美国在第二次世界大战期间组织实施的原子弹研发计划；阿波罗计划是美国的载人登月计划，于 1969 年首次实现人类登月。两者都是政府组织实施的大规模科研与工程计划。

参考文献

- [1] I. J. Good, "Speculations concerning the first ultraintelligent machine," in *Advances in computers*, vol. 6, Elsevier, 1966, pp. 31–88 (cit. on pp. 2, 8).
- [2] A. M. Turing, *Intelligent machinery, a heretical theory*, 1951 (cit. on p. 2).
- [3] J. Evans, B. Bratton, and B. Agüera y Arcas, "Agentic AI and the next intelligence explosion," *Science*, vol. 391, no. 6791, eaeg1895, Mar. 2026. doi: 10.1126/science.aeg1895 (cit. on p. 2).
- [4] J. Pachocki, *An alien mind*, Sep. 2026 (cit. on p. 2).
- [5] M. Favaro and J. Clark, *When AI builds itself: Our progress toward recursive self-improvement, and its implications*, Jun. 2026 (cit. on pp. 2, 3, 7).
- [6] T. Tunguz, *Are we being railroaded by AI?* Nov. 2025 (cit. on p. 2).
- [7] C. E. Leiserson, N. C. Thompson, J. S. Emer, B. C. Kuszmaul, B. W. Lampson, D. Sanchez, and T. B. Schardl, "There's plenty of room at the top: What will drive computer performance after Moore's law?" *Science*, vol. 368, no. 6495, eaam9744, Jun. 2020. doi: 10.1126/science.aam9744 (cit. on p. 2).
- [8] A. Ho, T. Besiroglu, E. Erdil, D. Owen, R. Rahman, Z. C. Guo, D. Atkinson, N. Thompson, and J. Sevilla, *Algorithmic progress in language models*, 2024 (cit. on p. 2).
- [9] OpenAI, *How GPT-5.6 fuses frontier intelligence with frontier efficiency*, Jul. 2026 (cit. on p. 2).
- [10] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, J. R. Lee, Y. T. Lee, Y. Li, W. Liu, C. C. T. Mendes, A. Nguyen, E. Price, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, X. Wang, R. Ward, Y. Wu, D. Yu, C. Zhang, and Y. Zhang, *Phi-4 technical report*, Dec. 2024. doi: 10.48550/arXiv.2412.08905 (cit. on p. 2).
- [11] J.-S. Denain and C. Barber, *An FAQ on reinforcement learning environments*, 2026 (cit. on p. 2).
- [12] D. Guo et al., "DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning," *Nature*, vol. 645, no. 8081, pp. 633–638, Sep. 2025. doi: 10.1038/s41586-025-09422-z (cit. on pp. 2, 5).
- [13] OpenAI, *Learning to reason with LLMs*, Sep. 2024 (cit. on p. 2).
- [14] D. Eth and T. Davidson, "Will AI R&D automation cause a software intelligence explosion?," 2025 (cit. on p. 3).
- [15] M. Favaro and P. Wright, *Measurements for understanding the pace of AI development inside frontier labs*, Anthropic, Sep. 2026 (cit. on pp. 3, 7).
- [16] METR, *Frontier risk report (February to March 2026)*, May 2026 (cit. on p. 3).
- [17] METR, *Time horizon 1.1*, Jan. 2026 (cit. on pp. 3, 13).
- [18] OpenAI, *Research acceleration: The view inside OpenAI*, Sep. 2026 (cit. on pp. 3, 7).
- [19] B. Rank, H. Bhatnagar, A. Prabhu, S. Eisenberg, K. Nguyen, M. Bethge, and M. Andriushchenko, "PostTrainBench: Can LLM agents automate LLM post-training?" In *International Conference on Machine Learning (ICML)*, 2026. arXiv: 2603.08640 [cs.SE] (cit. on p. 3).
- [20] J. Wen, L. Qiu, J. Benton, J. H. Kirchner, and J. Leike, *Automated weak-to-strong researcher*, Apr. 2026 (cit. on p. 3).
- [21] J. Wen, C. Si, Y.-h. Chen, H. He, and S. Feng, *Predicting empirical AI research outcomes with language models*, Jun. 2025. doi: 10.48550/arXiv.2506.00794 (cit. on p. 3).
- [22] C. Lu, C. Lu, R. T. Lange, Y. Yamada, S. Hu, J. Foerster, D. Ha, and J. Clune, "Towards end-to-end automation of AI research," *Nature*, vol. 651, no. 8107, pp. 914–919, Mar. 2026. doi: 10.1038/s41586-026-10265-5 (cit. on p. 3).
- [23] S. Rabanser, S. Kapoor, P. Kirgis, K. Liu, S. Utpala, and A. Narayanan, *Towards a science of AI agent reliability*, Feb. 2026. doi: 10.48550/arXiv.2602.16666 (cit. on p. 3).
- [24] Anthropic, "Claude Fable 5 & Claude Mythos 5 system card," Tech. Rep., Jun. 2026 (cit. on p. 3).
- [25] OpenAI, "GPT-6 Astra system card," Tech. Rep., Sep. 2026 (cit. on pp. 3, 6).
- [26] P. Whitfill, C. Wu, J. Becker, and N. Rush, *Many SWE-bench-passing PRs would not be merged into main*, Mar. 2026 (cit. on p. 3).
- [27] N. Bloom, C. I. Jones, J. Van Reenen, and M. Webb, "Are ideas getting harder to find?" *American Economic Review*, vol. 110, no. 4, pp. 1104–1144, Apr. 2020. doi: 10.1257/aer.20180338 (cit. on pp. 4, 13).
- [28] A. Ho and P. Whitfill, *The software intelligence explosion debate needs experiments*, 2025 (cit. on p. 5).

- [29] P. Whitfill and C. Wu, *Will compute bottlenecks prevent an intelligence explosion?* Aug. 2025. DOI: 10.48550/arXiv.2507.23181 (cit. on p. 5).
- [30] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, *Scaling laws for neural language models*, Jan. 2020. DOI: 10.48550/arXiv.2001.08361 (cit. on p. 5).
- [31] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre, *Training compute-optimal large language models*, Mar. 2022. DOI: 10.48550/arXiv.2203.15556 (cit. on p. 5).
- [32] P. Villalobos, A. Ho, J. Sevilla, T. Besiroglu, L. Heim, and M. Hobbhahn, *Will we run out of data? Limits of LLM scaling based on human-generated data*, Jun. 2024. DOI: 10.48550/arXiv.2211.04325 (cit. on p. 5).
- [33] T. Davidson, B. Halperin, T. Houlden, and A. Korinek, *When does automating AI research produce explosive growth? Feedback loops in innovation networks*, Jul. 2026 (cit. on p. 5).
- [34] L. Emberson and Y. Edelman, *Frontier training runs will likely stop getting longer by around 2027*, Jul. 2025 (cit. on p. 5).
- [35] T. Davidson, J.-S. Denain, P. Villalobos, and G. Bas, *AI capabilities can be significantly improved without expensive retraining*, Dec. 2023. DOI: 10.48550/arXiv.2312.07413 (cit. on p. 5).
- [36] A. Novikov, N. Vü, M. Eisenberger, E. Dupont, P.-S. Huang, A. Z. Wagner, S. Shirobokov, B. Kozlovskii, F. J. R. Ruiz, A. Mehrabian, M. P. Kumar, A. See, S. Chaudhuri, G. Holland, A. Davies, S. Nowozin, P. Kohli, and M. Balog, *AlphaEvolve: A coding agent for scientific and algorithmic discovery*, Jun. 2025. DOI: 10.48550/arXiv.2506.13131 (cit. on p. 5).
- [37] T. Cunningham, L. Althoff, B. Halperin, B. Jabarian, A. Koh, A. Ramani, P. Trammell, P. Whitfill, and C. Wu, *The economics of recursive self-improvement*, Jul. 2026 (cit. on pp. 5, 7, 13).
- [38] H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, A. Anandkumar, K. Bergen, C. P. Gomes, S. Ho, P. Kohli, J. Lasenby, J. Leskovec, T.-Y. Liu, A. Manrai, D. Marks, B. Ramsundar, L. Song, J. Sun, J. Tang, P. Veličković, M. Welling, L. Zhang, C. W. Coley, Y. Bengio, and M. Zitnik, “Scientific discovery in the age of artificial intelligence,” *Nature*, vol. 620, no. 7972, pp. 47–60, Aug. 2023. DOI: 10.1038/s41586-023-06221-2 (cit. on p. 6).
- [39] K. E. Drexler, *Radical abundance: How a revolution in nanotechnology will change civilization*. PublicAffairs, 2013 (cit. on p. 6).
- [40] Y. Bengio, S. Clare, C. Prunkl, M. Murray, M. Andriushchenko, B. Bucknall, R. Bommasani, S. Casper, T. Davidson, R. Douglas, D. Duvenaud, P. Fox, U. Gohar, R. Hadshar, A. Ho, T. Hu, C. Jones, S. Kapoor, A. Kasirzadeh, S. Manning, N. Maslej, V. Mavroudis, C. McGlynn, R. Moulange, J. Newman, K. Y. Ng, P. Paskov, S. Rismani, G. Sastry, E. Seger, S. Singer, C. Stix, L. Velasco, N. Wheeler, D. Acemoglu, V. Conitzer, T. G. Dietterich, E. W. Felten, F. Heintz, G. Hinton, N. Jennings, S. Leavy, T. Luderemir, V. Marda, H. Margetts, J. McDermid, J. Munga, A. Narayanan, A. Nelson, C. Neppe, S. D. Ramchurn, S. Russell, M. Schaake, B. Schölkopf, A. Soto, L. Tiedrich, G. Varoquaux, A. Yao, Y.-Q. Zhang, L. A. Aguirre, O. Ajala, F. Albalawi, N. AlMalek, C. Busch, J. Collas, A. C. P. d. L. F. de Carvalho, A. Gill, A. H. Hatip, J. Heikkilä, C. Johnson, G. Jolly, Z. Katzir, M. N. Kerema, H. Kitano, A. Krüger, K. M. Lee, J. R. López Portillo, A. McLysaght, O. Molchanovskiy, A. Monti, M. Nemer, N. Oliver, R. Pezoa, A. Plonk, B. Ravindran, H. Riza, C. Rugege, H. Sheikh, D. Wong, Y. Zeng, L. Zhu, D. Privitera, and S. Mindermann, “International AI safety report 2026,” Department for Science, Innovation and Technology, Tech. Rep. DSIT 2026/001, 2026 (cit. on p. 6).
- [41] Y. Bengio, G. Hinton, A. Yao, D. Song, P. Abbeel, T. Darrell, Y. N. Harari, Y.-Q. Zhang, L. Xue, S. Shalev-Shwartz, G. Hadfield, J. Clune, T. Maharaj, F. Hutter, A. G. Baydin, S. McIlraith, Q. Gao, A. Acharya, D. Krueger, A. Dragan, P. Torr, S. Russell, D. Kahneman, J. Brauner, and S. Mindermann, “Managing extreme AI risks amid rapid progress,” *Science*, vol. 384, no. 6698, pp. 842–845, May 2024. DOI: 10.1126/science.adn0117 (cit. on p. 6).
- [42] C. Aveggio, A. J. Patel, S. Nevo, and K. Webster, “Exploring the offense-defense balance of biology: Identifying and describing high-level asymmetries,” RAND Corporation, Santa Monica, CA, Tech. Rep. PE-A4102-1, Aug. 2025 (cit. on p. 6).
- [43] R. Greenblatt, A. Cotra, and H. Wijk, “Brief independent investigation of agents’ behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident,” METR and

- Redwood Research, Tech. Rep., Aug. 2026 (cit. on p. 6).
- [44] OpenAI, *OpenAI and Hugging Face partner to address security incident during model evaluation*, Jul. 2026 (cit. on pp. 6, 7).
- [45] OpenAI, *The Hugging Face incident and the road ahead*, Aug. 2026 (cit. on p. 6).
- [46] M. Sulmeyer, “Artificial intelligence and the risk of strategic cyberattack,” RAND Corporation, Tech. Rep., Jul. 2026 (cit. on p. 6).
- [47] D. Hendrycks, E. Schmidt, and A. Wang, *Superintelligence strategy: Expert version*, Mar. 2025. DOI: 10.48550/arXiv.2503.05628 (cit. on pp. 6, 8).
- [48] C. O’Keefe, A. Z. Rozenshtein, and C. Winter, *Executive branch AI and the rule of law: An emerging research agenda*, May 2026 (cit. on p. 6).
- [49] T. Davidson, L. Finnveden, and R. Hadshar, “AI-enabled coups: How a small group could use AI to seize power,” 2025 (cit. on p. 6).
- [50] J. Edwards and L. Emberson, *Open models lag state-of-the-art closed models by 4 months*, May 2026 (cit. on p. 6).
- [51] R. Slayton, “What is the cyber offense-defense balance? Conceptions, causes, and assessment,” *International Security*, vol. 41, no. 3, pp. 72–109, 2017. DOI: 10.1162/ISEC_a_00267 (cit. on p. 6).
- [52] B. Garfinkel and A. Dafoe, “How does the offense-defense balance scale?” *Journal of Strategic Studies*, vol. 42, no. 6, pp. 736–763, Aug. 2019. DOI: 10.1080/01402390.2019.1631810 (cit. on p. 6).
- [53] *Transparency in Frontier Artificial Intelligence Act, SB 53, 2025–2026 Reg. Sess. (Cal. 2025)*, Sep. 2025 (cit. on p. 7).
- [54] European Commission, *The general-purpose AI code of practice*, Jul. 2025 (cit. on p. 7).
- [55] *Responsible AI Safety and Education (RAISE) Act, A6453B, 2025–2026 Reg. Sess. (N.Y. 2025)*, Dec. 2025 (cit. on p. 7).
- [56] J. Kwon and S. Casper, *Internal deployment gaps in AI regulation*, Jan. 2026. DOI: 10.48550/arXiv.2601.08005 (cit. on p. 7).
- [57] M. Pistillo, *Internal deployment in the AI Act*, Jun. 2026 (cit. on p. 7).
- [58] A. Chan, R. Padarath, J. Kwon, H. Greaves, and M. Anderljung, *Measuring AI R&D automation*, Mar. 2026. DOI: 10.48550/arXiv.2603.03992 (cit. on p. 7).
- [59] M. Brundage, N. Dreksler, A. Homewood, S. McGregor, P. Paskov, C. Stosz, G. Sastry, A. F. Cooper, G. Balston, S. Adler, S. Casper, M. Anderljung, G. Werner, S. Mindermann, V. Mavroudis, B. Bucknall, C. Stix, J. Freund, L. Pacchiardi, J. Hernandez-Orallo, M. Pistillo, M. Chen, C. Painter, D. W. Ball, C. O’Keefe, G. Weil, B. Harack, G. Finley, R. Hassan, S. Emmons, C. Foster, A. Reuel, B. Treece, Y. Bengio, D. Reti, R. Bommasani, C. Trout, A. S. Shamsabadi, R. Dattani, A. Weller, R. Trager, J. Sevilla, L. Wagner, L. Soder, K. Ramakrishnan, H. Papadatos, M. Murray, and R. Tovcimak, *Frontier AI auditing: Toward rigorous third-party assessment of safety and security practices at leading AI companies*, Feb. 2026. DOI: 10.48550/arXiv.2601.11699 (cit. on p. 7).
- [60] P. Wills, *Regulatory supervision of frontier AI developers*, SSRN Scholarly Paper, Rochester, NY, Mar. 2025. DOI: 10.2139/ssrn.5122871 (cit. on p. 7).
- [61] U.S. Nuclear Regulatory Commission, *Background on NRC resident inspectors*, 2023 (cit. on p. 7).
- [62] Office of the Comptroller of the Currency, *What we do*, 2026 (cit. on p. 7).
- [63] M. Baker, G. Kulp, O. Marks, M. Brundage, and L. Heim, *Verifying international agreements on AI: Six layers of verification for rules on large-scale AI development and deployment*, Jul. 2025. DOI: 10.48550/arXiv.2507.15916 (cit. on p. 7).
- [64] B. Harack, R. F. Trager, A. Reuel, D. Manheim, M. Brundage, O. Aarne, A. Scher, Y. Pan, J. Xiao, K. Loke, S. N. Adan, G. Bas, N. A. Caputo, J. C. Morse, J. Ahuja, I. Duan, J. Egan, B. Bucknall, B. Rosen, R. Araujo, V. Boulanin, R. Lall, F. Barez, S. Alvira, C. Katzke, A. Atamli, and A. Awad, “Verification for international AI governance,” Oxford Martin AI Governance Initiative, University of Oxford, Tech. Rep., Jul. 2025 (cit. on pp. 7, 8).
- [65] T. Larsen, R. Dean, B. Halstead, E. Lifland, R. Greenblatt, and D. Kokotajlo, *AI 2040: Plan A*, AI Futures Project, 2026 (cit. on pp. 7, 8).
- [66] G. Sastry, L. Heim, H. Belfield, M. Anderljung, M. Brundage, J. Hazell, C. O’Keefe, G. K. Hadfield, R. Ngo, K. Pilz, G. Gor, E. Bluemke, S. Shoker, J. Egan, R. F. Trager, S. Avin, A. Weller, Y. Bengio, and D. Coyle, *Computing power and the governance of artificial intelligence*, Feb. 2024. DOI: 10.48550/arXiv.2402.08797 (cit. on p. 7).

- [67] A. Korinek and J. E. Stiglitz, *Steering technological progress*, Working Paper, Mar. 2026. doi: 10.3386/w34994 (cit. on p. 7).
- [68] S. Shoker, A. Reddie, S. Barrington, R. Booth, M. Brundage, H. Chahal, M. Depp, B. Drexel, R. Gupta, M. Favaro, J. Hecla, A. Hickey, M. Kon-aev, K. Kumar, N. Lambert, A. Lohn, C. O’Keefe, N. Rajani, M. Sellitto, R. Trager, L. Walker, A. Wehsener, and J. Young, *Confidence-building measures for artificial intelligence: Workshop proceedings*, Aug. 2023. doi: 10.48550/arXiv.2308.00862 (cit. on p. 8).
- [69] Intelligence Rising, *Intelligence Rising*, n.d. (cit. on p. 8).
- [70] AI Futures Project, *About AI 2027: Tabletop exercise*, 2025 (cit. on p. 8).
- [71] G. Smith, G. Hage, C. Heitzenrater, M. Chesson, and R. S. Girven, “Infinite potential—insights from the cyber surprise scenario,” RAND Corporation, Tech. Rep., Mar. 2026 (cit. on p. 8).
- [72] L. Vaintrob, *The AI adoption gap: Preparing the US government for advanced AI*, Apr. 2025 (cit. on p. 8).
- [73] OpenAI, *OpenAI model spec*, Aug. 2026 (cit. on p. 8).
- [74] Anthropic, *Claude’s constitution*, Jan. 2026 (cit. on p. 8).
- [75] C. O’Keefe, K. Ramakrishnan, J. Tay, and C. Winter, “Law-following AI: Designing AI agents to obey human laws,” *Fordham Law Review*, vol. 94, no. 1, p. 57, 2025 (cit. on p. 8).
- [76] S. Guerra, A. Attal-Juncqua, J. P. Tarangelo, C. Aveggio, K. Dammer, and D. Glickstein, “Building a defense-in-depth biosecurity strategy for the AI era,” RAND Corporation, Tech. Rep., Aug. 2026 (cit. on p. 8).
- [77] J.-S. Denain, A. Ho, and J. Sevilla, *How many digital workers could OpenAI deploy?* Oct. 2025.
- [78] H. Wijk, T. Lin, J. Becker, S. Jawhar, N. Parikh, T. Broadley, L. Chan, M. Chen, J. Clymer, J. Dhyani, E. Ericheva, K. Garcia, B. Goodrich, N. Jurkovic, H. Karnofsky, M. Kinniment, A. La-jko, S. Nix, L. Sato, W. Saunders, M. Taran, B. West, and E. Barnes, *RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts*, May 2025. doi: 10.48550/arXiv.2411.15114
- [79] A. Ho, J.-S. Denain, D. Atanasov, S. Albanie, and R. Shah, *A rosetta stone for AI benchmarks*, 2025.
- [80] B. Cottier, B. Snodin, D. Owen, and T. Adamczewski, *LLM inference prices have fallen rapidly but unequally across tasks*, Mar. 2025 (cit. on p. 13).
- [81] H. Gundlach, A. Fogelson, J. Lynch, A. Trisovic, J. Rosenfeld, A. Sandhu, and N. Thompson, *On the origin of algorithmic progress in AI*, Nov. 2025. doi: 10.48550/arXiv.2511.21622
- [82] P. Trammell, “The bounded parallelizability of R&D: Theory and application to AI,” Jul. 2026.
- [83] METR, *Measuring AI ability to complete long tasks*, Mar. 2025 (cit. on p. 13).
- [84] D. Kokotajlo, S. Alexander, T. Larsen, E. Lifland, and R. Dean, *AI 2027*, Apr. 2025 (cit. on p. 13).
- [85] T. Ord, *The dynamics of intelligence explosions*, 2026. arXiv: 2608.14426 [cs.AI] (cit. on p. 13).
- [86] B. Murphy and T. Stone, *Uplifted attackers, human defenders: The cyber offense-defense balance for trailing-edge organizations*, Aug. 2025. doi: 10.48550/arXiv.2508.15808 (cit. on p. 13).
- [87] Anthropic, *Investigating three real-world incidents in our cybersecurity evaluations*, Jul. 2026 (cit. on p. 13).
- [88] UK AI Security Institute, *Incident report: Unsanctioned agent behaviour during cyber testing*, Aug. 2026 (cit. on p. 13).
- [89] OpenAI, *Third-party cyber evaluations involving OpenAI models*, Aug. 2026 (cit. on p. 13).
- [90] P. C. Bogdan, R. Qi, J. Eaton, S. Kennedy, F. Roger, A. Glynn, R. Chen, B. Wright, O. Stegmaier, J. Kutasov, D. Foreman-Mackey, T. Bricken, S. Carr, S. Carter, M. MacDiarmid, S. Marks, A. Pearce, E. Simon, N. Carlini, C. Burns, J. Lindsey, S. Price, and S. Kantamneni, *An alignment assessment of recent cybersecurity incidents*, Anthropic, Sep. 2026 (cit. on p. 13).
- [91] M. H. Tessler, M. A. Bakker, D. Jarrett, H. Sheahan, M. J. Chadwick, R. Koster, G. Evans, L. Campbell-Gillingham, T. Collins, D. C. Parkes, M. Botvinick, and C. Summerfield, “AI can help humans find common ground in democratic deliberation,” *Science*, vol. 386, no. 6719, eadq2852, Oct. 2024. doi: 10.1126/science.adq2852 (cit. on p. 13).
- [92] A. Goldie and A. Mirhoseini, *How AlphaChip transformed computer chip design*, Sep. 2024 (cit. on p. 13).

注释

1. 机器学习会议通常设有主会和若干研讨会。在理解上述结果时，需要注意：研讨会的评审标准可能比主会宽松一些。
2. 根据 METR 的任务时长（time-horizon）指标 [17, 83]，AI 系统能够完成的任务时长，最初大约每 7 个月翻一番，2024 年以来加快到约每 3 个月翻一番。按这一较新的趋势外推，到 2028 年年中，AI 系统将能完成原本需要人类专家投入数月时间的任务。许多 AI 研发项目的所需时间都在这一范围内。另见 Kokotajlo 等人 [84]。
3. Cottier 等人 [80] 发现，在达到同一能力水平的前提下，运行大语言模型（LLM）的价格每年约降至此前的九分之一至九百分之一。具体降幅取决于所考察的能力水平，例如在某项数学基准上达到 GPT-4 水平。降价虽然有一部分来自硬件改进，即单位计算成本的降低，但很大一部分来自软件改进。
4. 在某一技术领域， r 描述研发投入的增加与所关注产出的变化之间的关系。例如，可以考察使用尖端消费级硬件时，每投入一美元能够完成多少次计算；计算时假定美元的购买力不变。Bloom 等人 [27] 以研发支出来衡量投入，因此同时计入了劳动力和物质资本的增加。本文则只考虑劳动力的增加，因为我们关注的是在算力存量大致不变的情况下，软件驱动的反馈循环能发挥多大作用。因此，本文的 r 值低于将劳动力、数据和算力等所有投入都计入时的 r 值。
5. r 最终必然降至 1 以下，因为 AI 进步终将触及计算和物理极限。在触及这些极限之前还能取得多少进展，目前尚不确定。
6. 三个 AI 研究子领域的 90% 可信区间分别为（0.727 至 2.094）、（0.380 至 2.708）和（1.069 至 3.212）。
7. 在这一分析中，算法效率的提升本身并不能消除这一潜在瓶颈，因为它会按相同比例提高研发和训练的效率。
8. 具体而言，该论文给出了在有限时间内，自动化研发劳动力以及经济产出等其他变量增长至无穷大所需的条件。
9. 关于各轮研发之间的时间间隔如何影响智能爆炸动态过程的理论讨论，详见 Ord [85]。
10. Cunningham 等人 [37] 的分析聚焦于“自我维持的加速”，即“即使外生投入（人类劳动力、训练算力等）不再增长，AI 系统也足以加快 AI 能力的进步”。这种自我维持的加速，是本文所说的软件驱动的智能爆炸的必要条件。
11. 这类系统可能在某些领域超越人类（例如某些科学研究领域），但在其他领域尚未达到这一水平（例如在物理世界中操纵物体）。
12. 不过，即使在网络空间，人类的组织流程仍可能给防御方带来阻力 [86]。
13. 另见 Anthropic [87]、英国 AI 安全研究所（UK AI Security Institute）[88]、OpenAI [89] 和 Bogdan 等人 [90]。
14. 例如 Tessler 等人 [91] 的研究。另一种更具推测性的可能是，AI 系统还可能加快脑机接口的发展，让人类以快得多的速度思考和协调。
15. 如何为智能爆炸给出精确、可操作的定义，仍是一个棘手问题，有待进一步研究。
16. 例如，AI 已经在辅助芯片设计 [92]。AI 系统还可能加快机器人技术的发展，从而实现芯片生产过程的自动化。