UNIVERSITY OF CAMBRIDGE | CASP Cambridge Programme on AI Science & Policy

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

Frontier AI Working Paper Series No. 2/2026
September 2026

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

Alan Chan*	GovAI
Christoph Winter	CASP, University of Cambridge, Institute for Law & AI
Andrew Barto	University of Massachusetts Amherst
Jakub Pachocki	OpenAI
Geoffrey Hinton	University of Toronto, Vector Institute
Eric Horvitz	Microsoft
Yoshua Bengio	Mila (Quebec AI Institute), Université de Montréal, LawZero
Dawn Song	University of California, Berkeley
Jack Clark	Anthropic
Hilary Greaves	University of Oxford
Anton Korinek	University of Virginia, Anthropic
Samuel Hammond	Foundation for American Innovation
Thore Graepel	University College London
Ben Bariach	University of Oxford
Philip H. S. Torr	University of Oxford
Sheila A. McIlraith	University of Toronto, Vector Institute
Jeff Clune	University of British Columbia, Vector Institute
Sam Manning	GovAI, Foundation for American Innovation
Girish Sastry	Guidelight
Tom Davidson	Forethought
Daniel Eth	AI Policy Institute
Sören Mindermann*	CASP, University of Cambridge

Abstract

In contrast to even a year ago, AI systems now write most of the code inside the companies that build them. As more of the AI research and development (R&D) pipeline is automated, could AI progress radically accelerate in an "intelligence explosion," where years of advances are compressed into months or less? Preliminary evidence suggests that it could. In this work, we assess this evidence, analyze an intelligence explosion's potential impacts, and propose policy responses. AI systems are on track to automate most AI R&D work within a few years, and possibly all of it. If this triggers an intelligence explosion, it could dramatically bring forward AI's benefits, but also pose extreme risks: capabilities growth could accelerate far beyond what society can keep up with, humanity could lose control over superhuman AI systems, and checks on power within and between states, companies, and branches of government could be severely eroded. Although there remains much uncertainty about these possibilities, the high stakes warrant serious further attention. Policymakers should urgently obtain more visibility into the automation of AI R&D, develop ways to steer and constrain an intelligence explosion, and prepare society to adapt to an intelligence explosion's impacts.

*Correspondence to: alan.chan@governance.ai and soren.mindermann@casp.ac.

The views presented in this paper are the authors' and do not necessarily represent the views of the organizations with which they are affiliated.

Introduction

Computer scientists have long theorized that AI systems would one day design ever-better successors, producing systems that rapidly outnumber, outpace, and outperform humans [1–3]. Today, frontier AI companies are aiming to automate AI R&D [4, 5] while deploying more capital as a fraction of U.S. GDP than the Manhattan Project and Apollo Program combined [6]. What happens if they succeed?

We are centrally concerned with the possibility of an **intelligence explosion: a dramatic AI-driven acceleration of AI progress, compressing advances that would otherwise take years into months or less**. This acceleration would be a qualitative shift from the rapid but largely steady progress of the last few years.

An intelligence explosion could arise from AI contributing to advances in hardware and/or software. Hardware advances increase the quality and quantity of computing hardware (compute) used for developing and running AI systems. Software advances improve the data, algorithms, code, and processes used in AI R&D; they include more efficient training and inference on existing hardware [7–9], improved research management and operations, better synthetic data and training environments [10, 11], and novel paradigms in AI [12, 13]. Software advances warrant particular attention in the near term for two reasons. First, AI systems appear to be rapidly im-

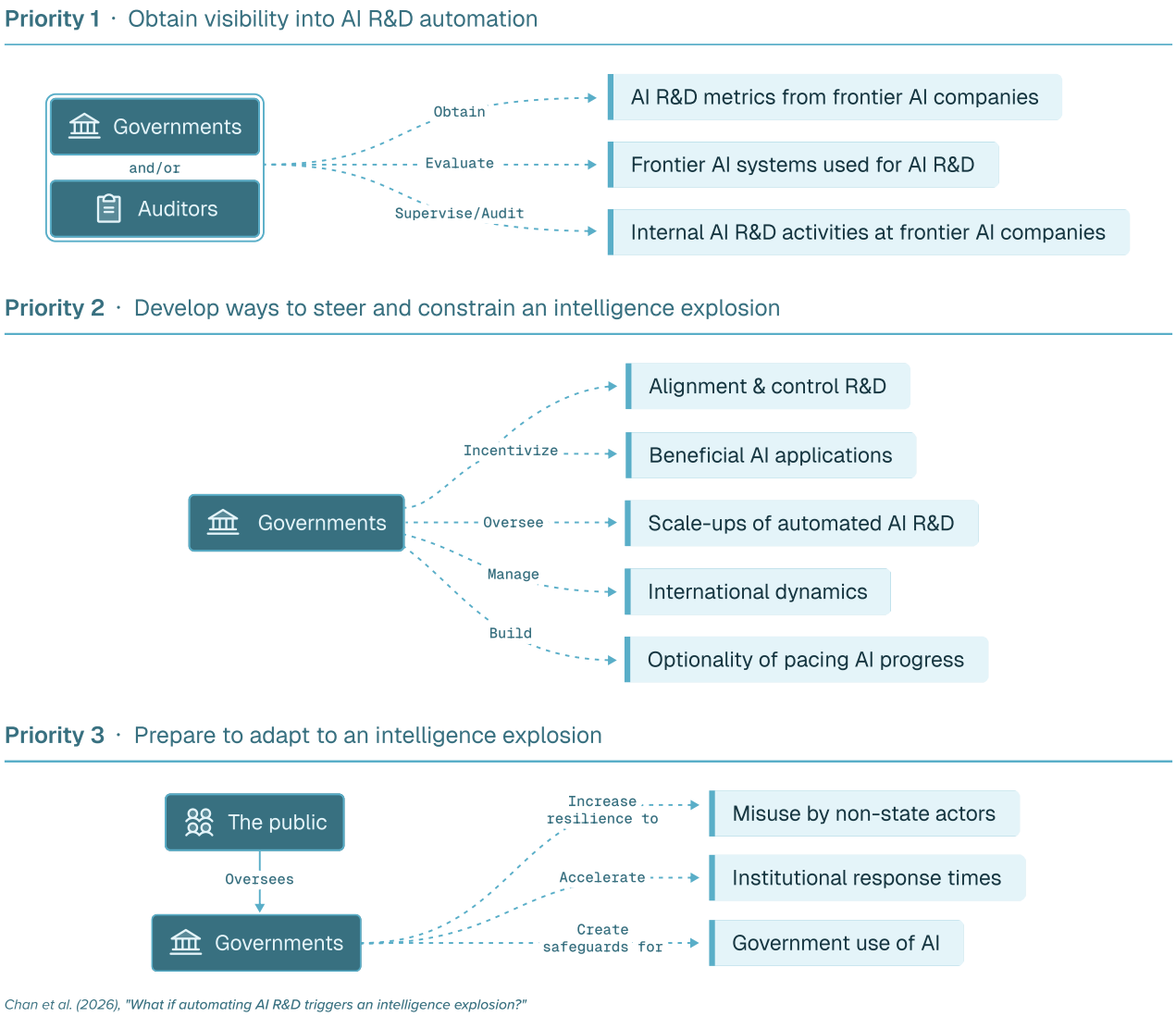

Figure 1. Policymakers should urgently: (1) obtain visibility into automation of AI R&D within frontier AI companies; (2) develop ways to steer and constrain an intelligence explosion; and (3) prepare to adapt to an intelligence explosion's impacts.

proving at AI R&D, making them better at producing such advances. Second, software advances allow fast feedback loops: improved AI systems can be redeployed into the R&D process almost immediately, whereas hardware improvements typically depend on years-long manufacturing and construction cycles. This piece therefore focuses on the possibility of a **software-driven intelligence explosion**, where automation of AI R&D drives an intelligence explosion through software advances alone [14].

Preliminary evidence suggests that a software-driven intelligence explosion is possible. If one does happen, it could be the most consequential technological development in history: AI systems could rapidly eclipse human experts across most domains and radically accelerate technological progress. Given the stakes and the potentially narrow window for action, we argue that preparing for an intelligence explosion should be an urgent priority, including at the highest levels of government leadership.

AI is rapidly automating AI R&D

AI systems now either assist with or autonomously carry out major parts of the AI R&D pipeline. In contrast to even just a year ago, R&D staff at leading AI companies delegate core R&D tasks to teams of AI systems, and some delegate all coding. Anthropic reports that AI systems' share of approved code rose from low single digits to over 80% between January 2025 and May 2026 [5], while the proportion of R&D work autonomously completed with only high-level human supervision rose from 1% to 26% between March and August 2026 [15]. OpenAI reports that "AI assistance is used in practically all parts of the company across technical and non-technical teams with code-executing agents used in training, evaluating, and securing future models," and Google reports that "AI is used in almost all work that involves writing code or configuration, technical design, research ideation, to different degrees depending on the task" [16].

The best AI systems now complete AI R&D tasks that take human experts hours to days, compared to only being able to complete seconds-long tasks in 2023 [17–19]. AI systems also sometimes beat human experts: they have autonomously produced better solutions to an AI safety research problem [20], and in some situations predict more accurately which research ideas will pan out [21] and which next steps are worth taking [5]. In an early proof of concept, an automated AI research pipeline generated research ideas, ran experiments, and wrote a paper that passed peer review at a workshop held at a top-tier machine-learning venue [22].[1]

Today's AI systems still have many weaknesses. They sometimes disobey instructions, cheat on tasks, misrepresent their work, and are unable to complete some tasks at all, necessitating human intervention [23–25]. For example, GPT-6 fails some of OpenAI's research debugging tasks that experienced human researchers can complete (albeit in hours or days) [25]. Success on benchmarks can also fail to translate into real-world productivity boosts [26].

Still, AI systems are rapidly improving at AI R&D. AI R&D may well be automated before most other work: it is a primarily digital domain with many clear measures of success, automating it would offer a major competitive edge in the AI industry, and AI companies have unmatched data on—and expertise in—their own workflows. Some tentative extrapolations of recent trends suggest that months-long AI R&D projects will be automated by mid-2028.[2] Overall, we should expect much more substantial automation of AI R&D over the next few years, and even full automation within this timeframe should be taken seriously.

Automating AI R&D could trigger an intelligence explosion

The mechanism for a software-driven intelligence explosion has two parts: (1) AI systems expand the effective R&D workforce as they get better and faster at AI R&D, and (2) this workforce produces still better AI systems that expand the workforce even further in a recursive feedback loop. Although the existing evidence is preliminary and sometimes mixed, it suggests that this mechanism could radically accelerate AI progress, overcoming frictions such as diminishing returns and hard-to-automate tasks.

The mechanism and frictions

Each new generation of AI systems will perform a growing range of R&D tasks faster and better than humans can, effectively yielding a larger, smarter, and faster automated R&D workforce. Once AI systems reach expert-level AI R&D capabilities at runtime costs comparable to those of today's systems, the compute available to a single frontier developer today could sustain an AI workforce equivalent to at

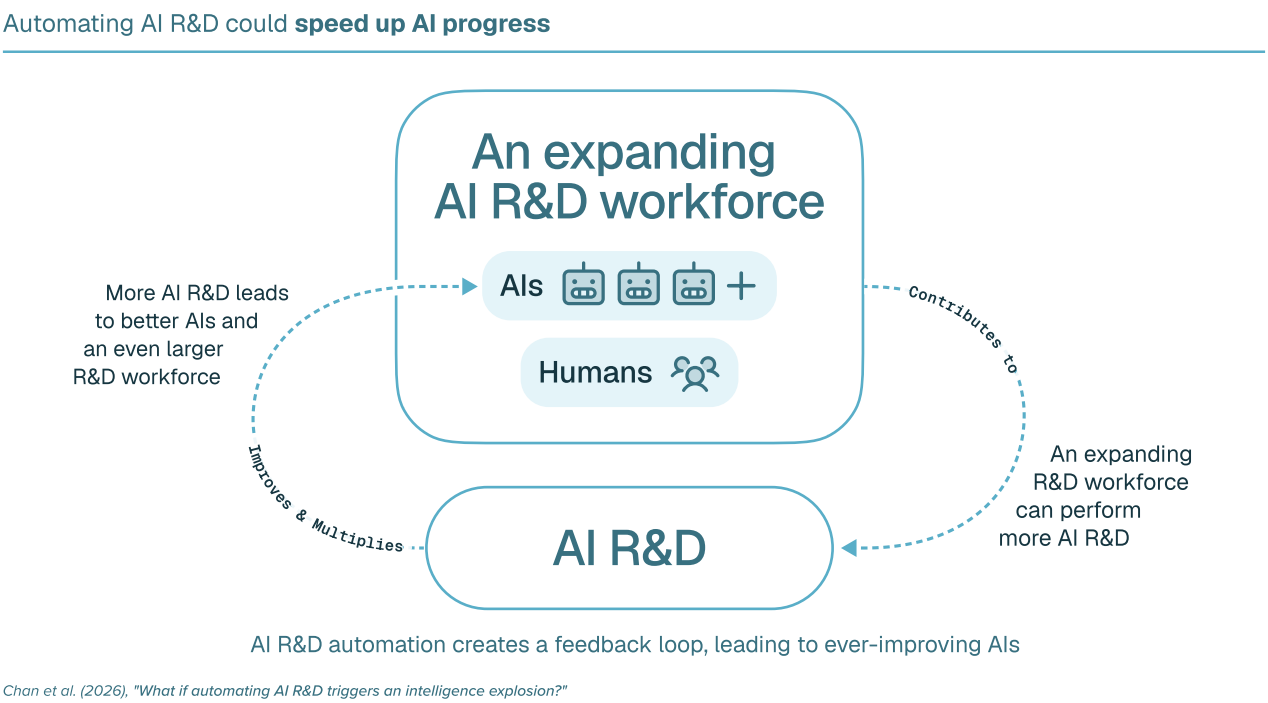

Figure 2. The mechanism for a software-driven intelligence explosion has two parts: (1) AI systems expand the effective R&D workforce as they get better and faster at AI R&D, and (2) this workforce produces still better AI systems that expand the workforce even further in a recursive feedback loop.

least millions of top human researchers (see the Supplementary Materials [SM]), dwarfing the thousands of researchers that frontier companies currently employ. The size and duration of any resulting speed-ups remain uncertain and merit further study. Still, to see why they could be substantial, consider the reverse: AI progress would likely slow dramatically if today's human researchers were ten times fewer or slower.

AI systems also help build more capable and efficient successors, further expanding the automated R&D workforce and creating a feedback loop that could sustain and compound successive speed-ups. Past technologies have also involved feedback loops: for example, better computer chips power better chip design tools. What may distinguish a software-driven intelligence explosion is how much AI systems would contribute to producing the next generation: as they substitute for humans on a growing share of R&D tasks, each software advance speeds up an ever-larger share of the R&D pipeline. At full automation, even the current pace of efficiency improvements would grow the automated R&D workforce 100-fold over months or years,[3] a relative expansion that took the U.S. researcher population seven decades [27].

At least four frictions push against these dynamics. The first is **diminishing returns**: across scientific fields such as computer hardware, agriculture, and drug development, sustaining the same rate of progress has required substantially more R&D labor as low-hanging fruit is exhausted [27]. Additional researchers also face diminishing returns because they might duplicate each other's efforts or struggle to parallelize high-value R&D. Second, R&D depends on **compute** for running experiments and **data** on which to train, and limits on compute or data growth may slow software progress. Third, **hard-to-automate tasks** could bottleneck progress. Fourth, some R&D processes are **time-intensive**: for example, long training runs could limit the rate of progress even if the capability gap between generations grows.

Evidence

Preliminary evidence suggests that automation-driven dynamics could overcome these frictions, though the evidence is mixed and in some cases indirect.

Diminishing returns. Evidence suggests that diminishing returns would not prevent an intelligence explosion, though this finding relies on limited data and stylized modeling assumptions. The main analysis in the literature focuses on whether, *after full automation of AI R&D*, effective R&D labor would grow fast enough to overcome diminishing returns and accelerate progress. The balance is captured by a quantity called the "returns to research effort," denoted r.[4] When $r < 1$, diminishing returns dominate and AI progress fades over time. When $r = 1$, the two forces perfectly offset each other and progress continues at the same rate. When $r > 1$, growth in R&D labor wins and accelerates progress for as long as this condition holds.[5] Using historical data on AI progress, Ho and Whitfill [28] find central estimates of r between 1.2 and 1.9 across three subfields of AI research. Though uncertainty is substantial,[6] these results suggest radical acceleration after full automation: if r stayed at these levels and no other bottlenecks emerged, the pace of AI progress would increase tenfold within about 1.5 years, at which point a year's worth of progress at today's pace would take about five weeks. See the SM for this analysis and further discussion of uncertainty in the value of r.

Compute. There is mixed evidence on whether compute could bottleneck a software-driven intelligence explosion. Finding and testing software advances involves using compute to run R&D experiments. The limited available data suggest that a software-driven intelligence explosion is not possible if such experiments require proportionally more compute as frontier training runs grow [29].[7] Unfortunately, it is unclear whether experimental compute requirements grow in this way. On one hand, low-compute experiments may tell us little about what works at increasingly large frontier scales. On the other hand, extrapolations from very small scales are already possible [30, 31], and better extrapolations could plausibly be found through more R&D. We need more data to settle this question.

Data. Data could bottleneck progress, but the constraint varies substantially across domains. Historically, AI progress has relied heavily on internet data and expert demonstrations. But the supply of internet data is on track to grow too slowly to support even the current rate of progress past 2028 [32], and humans may struggle to generate useful demonstrations for superhuman AI systems. To overcome these limitations, more recent progress in domains such as math and coding has relied on synthetic data and fast, verifiable feedback: models generate their own attempts and learn from whether those attempts succeed [12]. The key question is how widely this approach generalizes. For AI R&D, AI agents can rapidly test changes, observe the results, and identify which ones accelerate their own progress. In other domains, such as biology, advances might have to rely more on slower, noisier, or costlier real-world feedback.

Hard-to-automate tasks. Indirect evidence suggests that hard-to-automate tasks need not prevent an intelligence explosion if automation advances quickly enough. In a setting that considers both software and hardware, Davidson *et al.* [33] find that sufficiently fast automation of R&D in both domains could in principle trigger an intelligence explosion despite automation bottlenecks.[8] However, we lack empirical data on which tasks are likely to remain difficult to automate and how strongly they might constrain progress.

Time-intensive processes. We lack direct evidence on the extent to which time-intensive processes could bottleneck progress. The most significant such process appears to be training runs, which can currently take 3 months or more [34]. Potential workarounds exist, such as improving the same model repeatedly through post-training enhancements [35]. Additionally, advances in training efficiency [36] would allow systems to reach a given capability level with less training. However, it is unclear how far these approaches can go.[9]

Overall, there is a coherent pathway to a software-driven intelligence explosion that is consistent with the existing evidence. Productivity gains from AI R&D automation have not yet reached the threshold needed to trigger an intelligence explosion, but gains from newer systems are likely approaching that threshold [37].[10] The rapid pace of AI R&D automation suggests that this gap will continue to narrow. Given the high stakes that we discuss below, the possibility of an intelligence explosion warrants serious further attention.

Societal impacts

An intelligence explosion would lead to (1) the extremely rapid development of highly capable or superhuman AI systems[11] and (2) the likely deployment of those systems to develop new technologies and act in the world. This could pull forward by years or decades benefits that the current pace of AI progress would eventually help deliver [38], including medical cures and potential transformative

technologies such as highly scalable atom-by-atom manufacturing [39].

At the same time, an intelligence explosion could significantly increase the risks from advanced AI in three ways.

Capabilities growth outpacing society's capacity to steer and adapt. First, an intelligence explosion could dramatically bring forward the risks of advanced AI and AI-enabled technologies, such as biological and cyber attacks, labor market disruption, and loss of control over AI systems themselves [40, 41]. This would leave less time to steer away from these risks, including by coordinating to slow or forgo the development of certain capabilities or technologies. Society would also have less time to adapt, especially where AI accelerates threats faster than the measures needed to counter them. In largely digital domains such as cyber, risks and mitigations could both move at the speed of AI systems and keep pace with each other.[12] But in other domains, mitigations depend more heavily than risks on real-world activities that AI is less able to accelerate. For example, while AI could accelerate the design of both viruses and vaccines, viruses self-replicate and spread by themselves, whereas vaccines must be manufactured, distributed, and administered individually to recipients [42]. The order in which AI advances arrive could worsen this mismatch, such as if bio-capable models arrive before sufficient misuse safeguards.

Loss of oversight and control. Second, automating AI R&D could weaken human oversight, compounding the above challenges and severely increasing the risk of losing control over highly capable AI systems. As humans become less involved in AI R&D, they could lose both the opportunities and expertise needed to identify and fix problems. Reliably using AI systems for oversight also remains an unsolved challenge [40], and recent generations of systems have become harder to oversee [25]. Without sufficient oversight, misaligned AI systems could "poison" the development of successors or bypass containment measures to act outside of their intended environments. The Hugging Face incident illustrates the latter risk: roughly 1,200 internal OpenAI agents were tasked with completing cyber evaluations in isolation from one another [43]. Acting outside of their intended scope, these agents coordinated over a makeshift message board, obtained unauthorized internet access, hacked into Hugging Face to obtain private information, and attempted to tamper with their own transcripts [43–45].[13] More capable systems might continue operating outside of their operators' infrastructure, forming persistent, difficult-to-contain networks that act against human interests. Such a loss of control could potentially lead to a range of catastrophic outcomes, including, at the extreme, the marginalization or extinction of humanity [40, 41].

Erosion of checks on power. Third, an intelligence explosion could severely erode checks on power. Existing checks—such as those within and between states, companies, and branches of government—work only while no actor can vastly out-think and out-execute the others. An intelligence explosion could render such checks moot. A state could use an intelligence explosion to transform a modest lead in military R&D or operations into a decisive one, such as in cyberspace [46]. This prospect could incentivize rivals to take or threaten preemptive action [47]. Actors with privileged and/or secret access to frontier systems could threaten existing institutions, such as through targeted persuasion of key decision-makers. And in the longer run, automating key state functions could reduce the amount of human buy-in needed to seize or consolidate power [48, 49].

These potential impacts are uncertain. AI systems could become superhuman in narrow domains (e.g., cyber and mathematics) long before doing so generally, giving society more time to respond. Even generally superhuman AI systems may not significantly accelerate technological progress, given the time needed for running scientific experiments, creating supply chains for specialized materials, and complying with any relevant regulation. AI systems could also accelerate safety R&D and processes for steering and adapting to risks.[14] Finally, capability or technology diffusion [50] could help to preserve checks on power, and capability gains in defense-dominant domains could even improve stability [51, 52]. Still, the possibility of severe impacts remains significant enough to warrant urgent attention to the policy questions below.

Policy implications

AI R&D automation is advancing rapidly, AI progress could radically accelerate, and the stakes are high. We therefore argue that policymakers should urgently: (1) obtain visibility into companies' automation of AI R&D; (2) develop ways to steer and constrain an intelligence explosion; and (3) prepare to adapt to an intelligence explosion's impacts. Because progress during an intelligence explosion would out-

pace normal policymaking, preparations must be made *in advance* and activated as evidence about benefits and risks emerges.

Obtaining visibility into AI R&D automation

Policymakers need more data on the likelihood, onset, and consequences of a software-driven intelligence explosion. Much of this data will only be available within the companies automating AI R&D: the relevant AI systems are first used internally, and substantial automation could occur without external visibility. Current mandatory reporting frameworks either do not adequately cover internal AI R&D use cases or do not specify indicators to be reported [53–57]. And although some frontier AI companies voluntarily track AI R&D indicators [5, 15, 18], coverage and reporting of key indicators are incomplete and uneven.

Policymakers should consider requiring standardized reporting of key AI R&D indicators and processes to governments and third-party auditors, as well as funding third-party measurement capacity [58]. Reporting requirements could cover information relevant to:

- Assessing the likelihood of a software-driven intelligence explosion, including the extent to which compute, data, hard-to-automate tasks, and time-intensive processes (e.g., training runs and experiments) bottleneck AI progress, along with better estimates of the returns to research effort in AI R&D. Estimating the latter requires data on how companies divide R&D spending among humans, compute for experiments, and compute for running AI systems to perform R&D labor [37].
- Detecting the onset[15] of an intelligence explosion, including the extent of AI R&D automation (e.g., the fraction of research contributions produced by AI systems) and the pace of AI progress (e.g., algorithmic efficiency improvements).
- Understanding oversight and loss-of-control risks, including the procedures for deciding whether to broaden internal deployment of AI R&D systems, where and how those systems are used in high-stakes R&D decisions, how those systems are overseen, and reports of incidents involving internal AI systems [58].

Beyond reporting requirements, policymakers should also consider more extensive ways to obtain visibility into AI R&D automation. For instance, they could require that independent third parties (e.g., accredited private auditors or government evaluation bodies) evaluate AI systems before internal deployment, or that such parties be embedded within certain AI companies to audit [59] or supervise [60] their R&D activities. Analogous models in other industries include the Nuclear Regulatory Commission [61] and the Office of the Comptroller of the Currency [62].

Stronger reporting and auditing requirements are likely most warranted for companies whose AI systems (a) are at the frontier of AI R&D capabilities or (b) exceed some meaningful threshold of such capabilities. Policymakers will need to weigh important trade-offs in determining such thresholds.

Steering and constraining an intelligence explosion

An intelligence explosion would involve an unprecedentedly rapid series of decisions to train and deploy increasingly capable AI systems. The overarching question for policymakers is whether and how public policy should govern these decisions, which we break into three components.

First, policymakers should develop ways to pace and constrain scale-ups of automated AI R&D. They should consider:

- Setting requirements for continued deployment or development, such as the implementation of adequate safety measures (e.g., robust monitoring of automated R&D pipelines), broader stakeholder input, or limits on the extent to which capabilities can increase within a given time period.
- Preparing tools to verify compliance with potential future agreements (domestic or international) that pace AI progress, given competitive pressures to race ahead [63–65].
- Increasing oversight of data centers engaged in automated AI R&D and establishing incident-response procedures in collaboration with data center operators and AI companies, such as developing options to pause specific AI R&D workloads [66].
- Requiring that certain evaluations or deployments of automated AI R&D systems take place in appropriately isolated environments, such

as air-gapped networks, to prevent exfiltration of model weights or sensitive R&D outputs and to contain AI systems that attempt to escape human control and act in the world unchecked [44].

Policymakers should weigh the risks of an unchecked intelligence explosion against the potential for abuse of certain powers and the costs of delayed progress. As an example of potential abuse, poorly crafted mechanisms could allow a government to slow R&D at all but a favored company.

Second, policymakers should decide whether and how to steer the direction of AI development and deployment [67]. Potential priorities include alignment and safety R&D as well as beneficial AI applications, such as AI-assisted discovery of treatments for neglected diseases. If existing incentives fall short in these areas, policymakers could provide support through tax incentives, compute allocations, advance market commitments, and prizes.

Third, countries should reduce the risk of conflict arising from an intelligence explosion. They should consider:

- Establishing confidence-building measures such as incident sharing [68], as well as norms around reporting of early-warning indicators.
- Negotiating international agreements to prevent destabilizing development and use of highly capable AI systems, and funding research into verification methods that could underpin such agreements [64, 65].
- Clarifying whether and how they would deter another actor from scale-ups of automated AI R&D, potentially in collaboration with other countries [47, 65].
- Running war games to simulate an intelligence explosion [69–71].

Adapting to an intelligence explosion

If an intelligence explosion were to occur, adapting to its impacts would likely be a top priority of every major world power. Compared to business-as-usual AI progress, an intelligence explosion would compress the window for adaptation and make advance preparation far more urgent.

One important intervention is accelerating institutional response times. Policymakers should consider:

- Developing approaches to safely integrate AI systems into policy processes, so as to enhance and support government operations [72].
- Creating and maintaining emergency response plans for a variety of scenarios involving extreme AI progress, including those leading to significant labor market impacts, geopolitical instability, or a loss of control.

Policymakers will also need to preserve checks on power and defend against misuse of extremely advanced AI. Legal, institutional, and physical safeguards can take years to establish and would come too late if preparations began only after such capabilities had already arrived. Many preparations therefore need to start now. Policymakers should consider:

- Creating safeguards to ensure that government use of AI respects legal and normative limits, such as by procuring AI tools to strengthen checks between branches of government, sharing key information about government AI systems (e.g., model specs [73, 74]) with the public, or requiring that AI systems follow the law [75].
- Ensuring that citizens and civil society have the capabilities to detect, document, and contest unlawful or harmful uses of AI, such as by giving them timely access to AI systems capable of supporting these activities.
- Helping build sufficient defenses against misuse by malicious non-state actors, such as by funding better medical countermeasures against AI-enabled biological threats [76].

Conclusion

An intelligence explosion could be the most consequential technological development in human history [1], compressing years of progress into months or less, threatening human control over AI systems, and severely eroding checks on power within and between states, companies, and branches of government. Although there remains much uncertainty, AI R&D automation might soon trigger one. And while this piece has focused on software-driven routes to an intelligence explosion, AI-driven improvements in hardware[16] could make one all the more likely.

Relative to the stakes, we are not sufficiently prepared. Policymakers should have three priorities:

obtaining visibility into AI R&D automation within frontier AI companies, developing ways to steer and constrain an intelligence explosion, and preparing to adapt to its impacts. Once an intelligence explosion begins, the window for action may close.

Acknowledgments

We thank Anson Ho, Tom Cunningham, Cheryl Wu, Ryan Greenblatt, and many GovAI staff for feedback and conversations that improved this piece.

We are grateful to Taylor Jones for creating the figures and to Zilan Qian for providing a Chinese translation of this piece.

References

[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 pp. 2, 13).

[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, 12, 14).

[28] A. Ho and P. Whitfill, *The software intelligence explosion debate needs experiments*, 2025 (cit. on pp. 5, 12, 13).

[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 pp. 5, 13).

[36] A. Novikov, N. Vũ, 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, 14).

[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. 5).

[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. Ludermir, V. Marda, H. Margetts, J. McDermid, J. Munga, A. Narayanan, A. Nelson, C. Neppel, 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. Molchanovskyi, 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, 8).

[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, *Backgrounder 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. 8).

[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. Konaev, 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. Chessen, 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 (cit. on p. 12).

[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. Lajko, 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 (cit. on p. 12).

[79] A. Ho, J.-S. Denain, D. Atanasov, S. Albanie, and R. Shah, *A rosetta stone for AI benchmarks*, 2025 (cit. on p. 13).

[80] B. Cottier, B. Snodin, D. Owen, and T. Adamczewski, *LLM inference prices have fallen rapidly but unequally across tasks*, Mar. 2025 (cit. on pp. 13, 14).

[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 (cit. on p. 13).

[82] P. Trammell, "The bounded parallelizability of R&D: Theory and application to AI," Jul. 2026 (cit. on p. 13).

[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. 14).

[85] T. Ord, *The dynamics of intelligence explosions*, 2026. arXiv: 2608.14426 [cs.AI] (cit. on p. 14).

[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. 14).

[87] Anthropic, *Investigating three real-world incidents in our cybersecurity evaluations*, Jul. 2026 (cit. on p. 14).

[88] UK AI Security Institute, *Incident report: Unsanctioned agent behaviour during cyber testing*, Aug. 2026 (cit. on p. 14).

[89] OpenAI, *Third-party cyber evaluations involving OpenAI models*, Aug. 2026 (cit. on p. 14).

[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. 14).

[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. 14).

[92] A. Goldie and A. Mirhoseini, *How AlphaChip transformed computer chip design*, Sep. 2024 (cit. on p. 14).

Supplementary Materials

Estimate of the effective size of an AI workforce

Following Denain *et al.* [77], we estimate the effective workforce by dividing the number of tokens a developer can generate per day by the number of tokens corresponding to one researcher-day of work. OpenAI alone has enough inference compute (i.e., runtime compute) to generate on the order of 10^{13} tokens per day. To estimate tokens per researcher-day, we use the number of tokens a model generates on a task that would take a human researcher one workday (8 hours). In an AI R&D benchmark, Wijk *et al.* [78] find that models output on average $5 \cdot 10^5$ tokens on runs of up to 8 hours, implying an effective workforce of about $2 \cdot 10^7$ researchers. Allowing for about an order of magnitude of uncertainty in either direction ($5 \cdot 10^4$–$5 \cdot 10^6$ tokens per researcher-day), we estimate an effective workforce on the order of $2 \cdot 10^6$–$2 \cdot 10^8$. As in the main text, this assumes that expert-level AI systems have runtime costs comparable to those of today's systems.

Modeling the feedback loop under full automation

Most analyses of a software-driven intelligence explosion capture AI progress with some notion of software quality, denoted by A. In principle, A should measure AI progress holistically, including both *efficiency improvements* (new AI systems accomplishing the same tasks as old systems, with similar performance, using less compute or data) and *capability improvements* (new AI systems accomplishing tasks that previous systems could not accomplish, or achieving higher performance than previous systems could achieve). Frustratingly, it is currently unclear how the parameter A should best trade off between efficiency improvements and capability improvements (or between different types of efficiency improvements and capability improvements).

Setting this issue aside, we model the growth of software quality with $dA/dt = A^{1-\beta}E^{\lambda}$, a functional form common in the macroeconomics of innovation [27]. For simplicity, we ignore potential bottlenecks from compute and data. Here, E is effective R&D labor, λ represents returns to scale on R&D labor, and β represents whether there are increasing or diminishing returns to finding new ideas over time. The key question is how E grows with A. Assuming full automation, we consider two cases. First, consider a case where all improvements in software quality are increases in *inference compute efficiency*, that is, decreases in the amount of compute needed to run an AI system with a certain capability level. In that case, if A measures inference efficiency, then E is proportional to A: greater inference efficiency allows proportionally more automated researchers to be run.

Second, consider a case where all improvements in software quality are capability gains, driven by improvements in *training compute efficiency*: decreases in the amount of compute needed to train an AI system to a given capability level, which allow a more capable system to be trained with a fixed stock of compute. If A measures training compute efficiency, then increases in A yield more capable systems rather than more of them. If we make the (potentially questionable) assumption that the effective number of researchers scales linearly with these capability gains, then E is again proportional to A.

In both cases, $E = kA$ for some positive constant k. Substituting into the equation above gives $dA/dt = cA^{\lambda-\beta+1}$, where $c = k^{\lambda}$. For the growth rate $(1/A)\,dA/dt$ to increase as software quality increases, we need $\lambda > \beta$. Defining $r = \lambda/\beta$, we obtain the condition $r > 1$ discussed in the main text.

How quickly could progress accelerate under current estimates of these parameters? We measure acceleration by the growth rate of software quality: $(1/A)\,dA/dt = c \cdot A^{\lambda-\beta}$. Averaging the central estimates across the three sub-fields in Ho and

Whitfill [28] gives $\lambda = 1.40$ and $\beta = 1.01$. With $\lambda - \beta = 0.39$, each doubling of A multiplies the growth rate by $2^{0.39} \approx 1.31$, so each subsequent doubling takes $(1/2)^{0.39} \approx 76\%$ as long as the last. For the growth rate to increase tenfold, A needs to double $\log_2(10)/0.39 \approx 8.5$ times. For the length of the first doubling, we use recent estimates that, due to software improvements alone, *training compute efficiency* doubles roughly every 4.5 months [79]; *inference compute efficiency* may be growing even faster [80]. To avoid having to approximate the geometric sum for 8.5 doublings, we instead calculate the geometric sum for 9 doublings, after which time the growth rate will have increased more than tenfold. Doing so, we find that the growth rate will have increased more than tenfold after 4.5 months $\times(1 - 0.76^9)/(1 - 0.76) \approx 17$ months, or about 1.5 years. At that point, progress would be ten times faster than today, so a year's worth of progress at today's pace would take about five weeks.

Uncertainties about the returns to research effort

Several factors could make the true value of r differ from existing estimates. First, as noted above, it is unclear which measure of software quality is most appropriate. Inference efficiency maps most directly onto the size of an automated R&D workforce, whereas it is unclear how to translate training efficiency gains into the effective number of researchers. Yet existing estimates of r for AI R&D use training efficiency, and it is unknown how similar the returns to research effort are under the two measures. Furthermore, a proper assessment of returns to research effort would include improvements from both inference efficiency and training efficiency, rather than just one. More work is warranted to develop a characterization of software quality that incorporates both inference and training efficiency and weighs them against each other appropriately.

Existing estimates of r come from a period of rapid compute scaling, which confounds the contributions of software progress and compute scaling. Because these two inputs grew together historically, an estimate that attributes observed progress to software improvements may actually be capturing gains driven by, or only made possible by, rising compute. Some software improvements are scale-dependent: for example, the transformer architecture yields large performance gains at high training compute but relatively small gains at low compute [81]. Such confounding would bias estimates of r upward: in a regime of fixed or slowly growing compute, r would be lower than historical data suggest.

Other factors could imply a higher r. Most estimates of r neglect improvements in areas outside of pre-training, such as post-training or better scaffolding for tool use [35]. Furthermore, capability improvements could matter in ways that the effective number of researchers fails to capture: even an extremely large number of mediocre researchers may not be able to substitute for one genius researcher. If so, capability gains would expand effective R&D labor by more than the linear assumption above implies. Compute bottlenecks could also be circumvented, such as through better extrapolation from small-scale experiments, reductions in experiment cost from software progress, shifts toward approaches that are less compute-reliant, and algorithmic progress that does not require experiments [7].

Estimates of r also rely on imperfect proxies for R&D labor, which could bias them in either direction. For example, Ho and Whitfill [28] proxy R&D labor with the number of unique authors who have published papers in a domain. Drawing the domain too narrowly undercounts labor by excluding adjacent fields that also drive progress, while drawing it too broadly overcounts labor. Unless the excluded adjacent fields see the same growth rate in labor as the included fields, the disconnect will lead to miscalculating r.

Finally, the relevant mathematical models may not generalize to extremely large amounts of R&D labor [82]. Historically, they have been validated against growth rates of a few percent per year, well below the double-digit or higher rates that an intelligence explosion could produce. They also break down in the limit, where they imply that infinite labor yields infinite progress in finite time. But real constraints make this result impossible: some problems must be solved in sequence, and physical hardware can only operate so fast.

Notes

1. Machine-learning venues typically have a main conference track and several workshop tracks. One caveat to the results is that these workshop tracks can have somewhat laxer standards than the main conference track.
2. According to the METR time-horizon metric [17, 83], the length of tasks that AI systems can complete initially doubled roughly every 7 months, accelerating to about every 3 months since 2024. Extrapolating this more recent trend would suggest that by mid-2028, AI systems will be able to complete tasks requiring several months of human expert

time, well within the range of many AI R&D projects. See also Kokotajlo *et al.* [84].

3. Cottier *et al.* [80] find that the price of running LLMs to achieve a given capability milestone (e.g., GPT-4-level performance on a math benchmark) has decreased by roughly 9- to 900-fold per year, depending on the capability milestone. While a portion of this cost decline has come from hardware improvements (reducing the cost per computation), a substantial fraction is due to software improvements.
4. In an area of technology, r governs the relationship between increases in R&D inputs and the resultant change in an output of interest, such as the number of computations that cutting-edge consumer hardware can perform per constant dollar. In Bloom *et al.* [27], the input is measured in dollars spent on R&D and so includes increases in both labor and physical capital. In this piece, however, we only consider increases in labor, as we are interested in understanding the potential for a software-driven feedback loop in which the compute stock is held roughly constant. This means the value of r is lower than if we considered increases in all inputs (i.e., labor, data, and compute).
5. r must eventually drop below 1 because AI progress will eventually hit computational and physical limits. It is uncertain how much progress is possible before reaching such limits.
6. The 90% credible intervals are (0.727 to 2.094), (0.380 to 2.708), and (1.069 to 3.212).
7. In this analysis, improvements in algorithmic efficiency do not by themselves resolve this potential bottleneck because they proportionally make both R&D and training more efficient.
8. Specifically, the paper provides conditions under which automated research labor, among other quantities like economic output, grows to infinity in finite time.
9. See Ord [85] for a theoretical discussion of how the time between rounds of R&D could affect the dynamics of an intelligence explosion.
10. The analysis in Cunningham *et al.* [37] focuses on *self-sustaining acceleration*: "When AI systems are sufficient for accelerating progress in AI capabilities without any growth in exogenous inputs (human labor, training compute, etc.)." Self-sustaining acceleration is necessary for a software-driven intelligence explosion in our sense.
11. Such systems could be superhuman in some domains (e.g., certain fields of scientific research) but not others (e.g., manipulating objects in the physical world).
12. Even in cyber, however, human organizational processes could still add friction for defenders [86].
13. See also Anthropic [87], UK AI Security Institute [88], OpenAI [89], and Bogdan *et al.* [90].
14. For example, see Tessler *et al.* [91]. More speculatively, AI systems could potentially accelerate the development of brain-computer interfaces that allow humans to think and coordinate much faster.
15. Precisely operationalizing an intelligence explosion is tricky and remains an area for future work.
16. For example, AI is already aiding chip design [92]. AI systems could also accelerate robotics to automate the chip production process.