GOPAL P. SARMA, RACHEL STERATORE, MICHAEL JACOB, SUNNY D. BHATT

Recursive Self-Improvement and Automated AI Research and Development

This work was independently initiated and made possible by RAND's own resources, including income from operations, unrestricted gifts from RAND supporters, and distributions from its long-term investment fund.

October 2026

Expert Insights
PERSPECTIVE ON A TIMELY POLICY ISSUE

For more information on this publication, visit **www.rand.org/t/PEA5068-1**.

About RAND

RAND is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. To learn more about RAND, visit www.rand.org.

Research Integrity

Our research integrity is grounded in RAND's core values of quality and objectivity. Rigorous quality assurance procedures, conflict of interest screening, and transparency in funding ensure that every study is objective and nonpartisan. Learn more at www.rand.org/integrity.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published by the RAND Corporation, Santa Monica, Calif.
© 2026 RAND Corporation
RAND® is a registered trademark.

Limited Print and Electronic Distribution Rights

This publication and trademark(s) contained herein are protected by law. This representation of RAND intellectual property is provided for noncommercial use only. Unauthorized posting of this publication online is prohibited; linking directly to its webpage on rand.org is encouraged. Permission is required from RAND to reproduce, or reuse in another form, any of its research products for commercial purposes. For information on reprint and reuse permissions, visit www.rand.org/about/publishing/permissions.

PE-A5068-1

About This Paper

This paper provides expert analysis on recursive self-improvement and automated artificial intelligence (AI) research and development. A companion report (Sarma, Bhatt, et al., 2026) identified this topic as the leading driver of rapid capability gains and potential discontinuities in AI development. Although the authors of that report examine broader artificial general intelligence forecasting methodologies and scenario frameworks, this paper focuses specifically on the empirical state of AI systems contributing to their own improvement. We provide actionable policy recommendations for navigating the uncertainties related to AI acceleration, including permission frameworks for AI research agents, extension of classified benchmarking machinery to cover automated research and development alongside a companion reporting framework, containment built on high-assurance engineering, and preparatory hedges that hold even if domestic reach proves incomplete. Given the rapid pace of developments in this area, this paper draws on both peer-reviewed publications and technical gray literature, including blog posts and working papers from researchers at leading AI laboratories and research organizations.

This paper was produced with AI-assisted research, drafting, and synthesis, with iterative expert review, fact-checking, and editorial judgment by the authors; the methodology is documented in further detail in the preceding companion report (Sarma, Bhatt, et al., 2026). RAND conducted this research as part of its commitment to providing objective analysis on emerging technology policy challenges. The capability frontier described in this paper is moving sufficiently rapidly that policy decisions informed by outdated evidence risk being calibrated to systems that no longer represent the state of the art.

Center on AI, Security, and Technology

RAND Global and Emerging Risks is a division of RAND that delivers rigorous and objective public policy research on the most consequential challenges to civilization and global security. This work was undertaken by the division's Center on AI, Security, and Technology, which aims to examine the opportunities and risks of rapid technological change, focusing on artificial intelligence, security, and biotechnology. For more information, contact cast@rand.org.

Funding

This effort was independently initiated and made possible by RAND's own resources, including income from operations, unrestricted gifts from RAND supporters, and distributions from its long-term investment fund. A complete list of donors and funders is available at www.rand.org/about/how-we-are-funded. RAND clients, donors, and grantors have no influence over research findings or recommendations.

Acknowledgments

This paper went through several iterations, and many people contributed substantially to improving it. We thank those whose valuable feedback helped inform the manuscript: Benjamin Boudreaux, Matan Chorev, Erich Devendorf, Varun Gandhi, Tahira Mammen, William Marcellino, Chris Painter, Joel Predd, Michael Vermeer, and William Zito. We also thank the Research Editorial and Production team, including Sean Mintus, Allison Kerns, and Brian Dau, and the RAND Global and Emerging Risks quality assurance team, including Gary Cecchine, Jason Etchegaray, and Paige Smith, for helping us produce and deliver this timely paper.

Summary

Recent trends in artificial intelligence (AI) research and development (R&D) have increased the probability of *recursive self-improvement* (RSI), a phenomenon in which AI systems contribute to their own improvement, creating a feedback loop resulting in potentially rapid AI capability gains. How far and how fast RSI progresses remains uncertain, but policymakers will have to act before this uncertainty is resolved. In this paper, we examine RSI and introduce a taxonomy for analyzing automated AI R&D trends, separating activity already underway from capabilities that have not yet been demonstrated publicly. We propose indicators for increasing policymaker visibility into acceleration as it occurs and develop policy options that perform well across a range of plausible trajectories. The following takeaways and recommendations summarize our analysis.

Key Takeaways

- **AI systems are already automating meaningful components of AI research, but a significant capability gap remains.** Current systems exceed 90 percent on Software Engineering Benchmark Verified (SWE-bench Verified) (a widely used benchmark that measures AI systems' ability to resolve real-world software engineering tasks), with leading models approaching saturation on the benchmark (Jimenez et al., 2024; BenchLM, 2026). AI agents have autonomously completed coding tasks estimated to take human engineers several weeks (Adamczewski et al., 2026). Multiple frontier labs have reported that a majority of their new code is AI-generated. One reports over 80 percent of merged code (Favaro and Clark, 2026), another 75 percent of all new code (Chakarian, 2026). AI has also contributed to algorithm discovery and hardware optimization (Novikov et al., 2025; Romera-Paredes et al., 2024). These span Types 1 and 2 (research pipeline automation and algorithm discovery, respectively) in the taxonomy we introduce in this paper. However, AI systems have not yet demonstrated all the components of open-ended research projects, including setting their own research agendas, generating novel paradigms, and sustaining multiweek efforts.
- **The near-term concern is distributed acceleration across many tools.** No single system needs to be capable of self-improvement for rapid progress in overall capabilities and field-wide speedup. AI that improves the chips, software, or experiments used to build AI can accelerate progress even if no single system ever modifies itself. Testing individual models in isolation is therefore inadequate as an overall indicator for system-level automation.
- **Given the proprietary and/or distributed nature of acceleration indicators, the government currently has limited visibility into how much AI R&D is being improved by AI itself.** The government lacks systematic reporting on AI-enabled research, including the share of research tasks being automated, compute allocated to AI-directed experimentation, and the extent to which model improvement traces to AI-generated ideas. Because of the

potential for distributed acceleration, this gap extends beyond frontier AI labs to leading hardware vendors.

- **AI systems conducting research inside AI labs are a security problem today**. These systems hold insider-level access to a lab's most sensitive code, data, and models. Recent empirical evidence suggests that models can engage in behavior that insider-threat programs are intended to detect (Gasteiger et al., 2025). An adversary that compromises an internally deployed research agent could subtly influence the development of future AI systems, a threat that compounds forward through each generation. In the context of automation and RSI, we assess integrity compromises to be more consequential and harder to detect than other threats, such as model weight theft.

Recommendations

- **Establish permission frameworks for AI research agents.** Frontier labs should define what internally deployed research agents are permitted to do without human approval, including allocating compute, initiating training runs, and modifying evaluation suites. Agent actions should be assigned to risk tiers with human authorization at defined gates. These frameworks are valuable regardless of the pace of AI automation research and do not require resolving theoretical debates about RSI timeline trajectories. They also address an immediate security need: The same AI systems automating R&D are systems with insider-level access to a lab's most sensitive assets, making permission frameworks an insider-threat control as much as a safety measure.
- **Extend the classified benchmarking machinery directed by June 2026 federal actions to cover AI R&D automation, and stand up a companion reporting framework for pipeline-level indicators.** Executive Order 14409 (2026) directs the U.S. Department of the Treasury; the U.S. Department of War, acting through the National Security Agency; and the U.S. Department of Homeland Security, acting through the Cybersecurity and Infrastructure Security Agency, to develop a classified process for benchmarking the advanced cyber capabilities of AI models and establishing the threshold for designating a "covered frontier model."[1] The order separately directs creation of a voluntary framework for developer engagement and prerelease government access. National Security Presidential Memorandum 11 (2026) establishes related but distinct national-security AI assurance and security-partnership mechanisms. The AI systems most relevant to automated R&D are largely the same systems that raise cybersecurity concerns; adding AI R&D automation as an assessed capability domain could build on the process that Executive Order 14409 directs agencies to establish. The cost of that extension will depend on the process's implementation status and the additional evaluation content required. Visibility into distributed acceleration, however, requires more than model benchmarking. For example, the National Institute of Standards and Technology, with willing AI companies, could set up a framework for structured,

[1] The Department of War is designated the Department of Defense under Public Law 81-216, National Security Act Amendments of 1949.

confidential reporting on pipeline-level indicators, such as automation share, compute allocation between human- and AI-directed experimentation, and the pace of capabilities relative to that of safety research automation. To be fully effective, coverage would ultimately need to reach internal deployments and leading hardware vendors.

- **Treat internally deployed research agents as privileged insiders, and build their containment using high-assurance engineering techniques.** Labs can engineer the infrastructure an agent runs on using high-assurance techniques to enforce hard limits on data movement, compute consumption, and system access, with strong, machine-checked evidence that specified components enforce specified security properties. These guarantees constrain what the infrastructure permits, not what the model attempts. The guarantees hold regardless of whether the system's behavior can be fully predicted or characterized. This containment floor should hold unconditionally, independent of assessed model capability. Separately, monitoring and evaluation intensity should scale with an agent's access level, autonomy, persistence, and the reversibility of its actions, consistent with associated permission frameworks. Each of these layers can help protect against foreign manipulation and unintended autonomous behavior.
- **Supplement the domestic policy stack with hedges that hold if its reach proves incomplete.** The foregoing recommendations reduce the likelihood of losing visibility and control over automated R&D. They do not address the consequences if that effort fails or if an adversary reaches advanced automated R&D capability first. Four lines of effort merit investment now: hardening and diversifying the infrastructure on which U.S. AI development depends (Frelinger and Mueller, 2026), maintaining contingency frameworks for international engagement on AI assurance (Mondschein, 2026), researching incident response and accelerating defensive technologies before either is needed in a crisis (Vermeer, 2025; Buterin, 2025), and building the technical basis for verifying compliance with any future international agreement on AI development (Baker et al., 2025). The pipeline-level indicators recommended here are the domestic groundwork for international efforts.

Contents

About This Paper iii
Summary v
Tables ix

Recursive Self-Improvement and Automated AI Research and Development 1
History and Current Status 1
Why the Pace Is Uncertain and Why It Matters 5
Policy Responses That Are Robust Across Scenarios 10
Key Takeaways and Recommendations 17
Conclusions 20

Abbreviations 21
References 22
About the Authors 28

Tables

Table 1. Potential Recursive Self-Improvement Indicators and Operationalized Triggers..... 4
Table 2. Options at a Glance..... 11
Table 3. Hedges Against Incomplete Domestic Reach 11

Recursive Self-Improvement and Automated AI Research and Development

History and Current Status

The term *recursive self-improvement* (RSI) describes the possibility that artificial intelligence (AI) systems could contribute to their own improvement, creating feedback loops in which each generation of a new capability enables faster development of the next. A historical antecedent to this concept is the self-replicating machine proposed by John von Neumann (1966), which could copy itself and potentially construct an even more complicated machine. Another example of this concept dates to mathematician I. J. Good's mid-1960s hypothesis that a sufficiently intelligent machine could design an even better machine, potentially initiating rapid cascading improvement (Good, 1966; Vinge, 1993). The seeds of this once-futuristic concept are visible in modern AI systems, which are already automating components of AI research and development (R&D).

Several leading frontier AI companies have long identified the automation of AI research as a strategic priority. OpenAI's roadmap includes the stated goal of creating a fully automated AI researcher by March 2028 (OpenAI, 2025b). Anthropic has similarly treated automated AI research as a near-term milestone, arguing that AI systems may soon be capable of designing, building, and training their own successors (Favaro and Clark, 2026). Google DeepMind has deployed autonomous research tools internally and published results demonstrating AI contributions to algorithm discovery and hardware optimization (Novikov et al., 2025; Romera-Paredes et al., 2024). Moreover, Google DeepMind's AlphaEvolve system, which pairs large language models with evolutionary search to discover and optimize algorithms, has been made commercially available, marking a transition from internal research tool to product (Google DeepMind, 2026). The strategic priority of automating AI research extends beyond established labs. Dedicated startups pursuing RSI have attracted substantial venture funding, and the concept of automated end-to-end research has progressed from proof-of-concept to peer-reviewed validation in leading journals (Dina, 2026; Lu et al., 2026).

As of June 2026, one frontier developer has publicly gone significantly further. Anthropic has characterized its own development trajectory as progressing toward RSI, reporting that AI systems now write over 80 percent of the company's merged code, that engineers ship roughly eight times as much code per quarter as they did before 2025, and that, on a fixed internal optimization task, model-driven speedup rose from three times to 52 times in under a year (Favaro and Clark, 2026). Anthropic has also called for a global mechanism to coordinate a potential pause if RSI is achieved before government and civil institutions are ready for it. Despite these public statements, there is no

guarantee that labs will continue to disclose progress on automated R&D, especially if it reveals proprietary research and internal productivity data.

The following month, in July 2026, more than 1,200 employees of frontier AI companies, including senior research leadership at OpenAI, Anthropic, Google DeepMind, and Meta, signed a public statement warning that capability development could accelerate beyond the ability to understand or control the resulting systems. The statement suggested that the U.S. government support an international effort to build the technical and governance tools needed to deliberately pace automated AI development (Employees of Frontier AI Companies, 2026).[2]

Implications of Recursive Self-Improvement

If AI accelerates AI development, we may see new capability jumps at shorter intervals than before. Critically, the arrival of these capability jumps may outpace the ability of institutions and governance to keep up. The development of standards, treaty negotiations, and the design of safety architectures may continue to move at human speeds, while AI development proceeds at machine speeds. This means the window for deliberate action is shrinking, thereby significantly increasing the importance of preparatory action. Moreover, if returns to automation compound, early leads become decisive, making the automation of AI R&D, even well short of full RSI, a competitive and strategic national security concern.

RSI also shifts where the most consequential risk resides. If the most capable systems are internal, the primary governance challenge is no longer managing what AI does in the world but maintaining visibility into what AI does inside the organizations building the next generation. This introduces a second locus of risk, one that sits inside the lab rather than in external deployments. External deployment risks remain real and independently require governance attention, but RSI adds a dimension that existing deployment-focused frameworks are not designed to address: the acceleration happening before any system reaches external users. Therefore, policymaker visibility into lab operations becomes the limiting factor on governance options for this class of risk. If government cannot measure AI R&D progress, it cannot develop strong, properly calibrated policy. By extension, governance mechanisms that activate at capability thresholds depend on the ability to observe these thresholds being crossed, an assumption that fails if the systems driving acceleration are themselves internal and opaque.

A Taxonomy for Policy

Automated AI R&D (the use of AI systems to accelerate the R&D of new and better AI systems) exists on a spectrum, with full RSI as its theoretical endpoint. The terms are often used

[2] We take no position on whether such pacing is desirable or feasible. The statement identifies an absence of measurement and coordination capacity that the reporting and verification measures we recommend would begin to address. Subsequent analysis has proposed operationalizing pacing not as a blanket slowdown but as conditional resource reallocation if specific risk thresholds are crossed, shifting compute and talent from the highest-risk automated R&D toward accelerating AI diffusion, safety research, and societal resilience (Fist et al., 2026). Whether such an approach is workable depends on the same monitoring and verification capacity we recommend.

interchangeably, but for governance purposes and to design monitoring that addresses the observation gap just described, the more immediate concern is the activity already underway along that spectrum. Frontier systems now perform most direct coding at leading labs: Anthropic has reported that over 80 percent of its merged production code is AI-written (Favaro and Clark, 2026), and Google has reported that 75 percent of new code is AI-generated (Chakarian, 2026). These systems also assist with machine learning experimentation.

Three features of the current state help make this concrete. First, self-reported productivity estimates provide a rough anchor: Surveys of technical workers and engineers at frontier labs place typical AI-assisted gains in the range of roughly 1.4 to 2 times baseline (Becker, 2026; Huang et al., 2025). An internal poll at one leading lab suggests figures as high as four times for research staff (Favaro and Clark, 2026), although controlled studies indicate that self-reports may overstate realized gains (Becker et al., 2025). These are meaningful but, so far, bounded productivity effects. Second, the unit of automation matters for oversight: Automating the equivalent of an individual researcher on a team may still allow clear understanding of what that researcher was working on, but automating the equivalent of full teams or organizations makes it much harder for anyone, including lab leadership, to maintain visibility. Third, the cadence of model improvement appears to be compressing: Industry observers and internal metrics suggest that capability gains that previously required many months are arriving at shorter intervals, narrowing the window for evaluation and governance response (Kwa et al., 2025; Favaro and Clark, 2026). These observations describe where automated AI R&D stands today. For governance, however, not all AI contributions to AI development carry the same risk or require the same response.

To provide analytical clarity, we categorize RSI-relevant activity into three types, each with different timelines, risk profiles, and monitoring requirements:

- **Type 1: Research pipeline automation and infrastructure optimization.** Progress toward AI R&D automation is possible simply by doing existing things faster. An AI system may automate portions of the research process: generating hypotheses, designing experiments, running evaluations, and analyzing results. It can also make improvements to hardware designs, data-center operations, or engineering workflows. Each of these changes produces productivity gains via efficiency and speed but not necessarily breakthrough insights.
- **Type 2: Algorithm and architecture discovery**. An AI system contributes to the discovery of improved algorithms, architectures, or training methods that humans then implement in successor systems.
- **Type 3: Direct self-improvement.** An AI system directly and comprehensively designs and runs the training of its successor. This corresponds most closely to the original "intelligence explosion" concept (Good, 1966; Bostrom, 2014). This remains beyond the abilities of modern neural networks; no disclosed system autonomously designs and executes its own retraining to improve performance of subsequent models.

These categories are useful for governance because each presents different monitoring challenges and intervention opportunities. Type 1 is detectable through pipeline-level measures: automation share, compute allocation, and productivity metrics that can be reported without revealing proprietary methods. Type 2 requires deeper assessment of proprietary training algorithms and architectures to

determine whether AI is contributing genuinely new ideas, not just executing existing ones faster. Importantly, Type 2 does not necessarily indicate imminent Type 3 capabilities (which would be most likely to lead to rapid capability explosion scenarios). However, Types 1 and 2 may create system feedback loops, even if no single model ever modifies itself directly. For instance, improvements in infrastructure could enable faster experimentation, which accelerates algorithm discovery, which improves the next generation of research tools. The policy-relevant acceleration scenario may therefore be distributed across the AI development ecosystem rather than be localized in one part of one lab's system. This means that model-level capability analysis, while important, cannot by itself detect a distributed acceleration. In Table 1, we propose reporting indicators that are deliberately meant to capture system-level impacts: organizational measurements of automation share, compute allocation to AI agents, and estimated AI contribution to capability gains (see Toner et al., 2026, for a closely related discussion).

Table 1. Potential Recursive Self-Improvement Indicators and Operationalized Triggers

Indicator	Current State (mid-2026)	Potential Triggers for Heightened Concern	Predominant Types
Effective-compute multipliers	Some labs track multipliers internally but rarely disclose or standardize them.	Year-over-year algorithmic-efficiency gains accelerate well above the 2014–2023 trend, with a measured share attributed to AI-developed improvements.	Types 1–2
Gap between public and internal deployments	New models are used internally before release, but the gap is not systematically measured.	A lab deploys a frontier system internally for research at scale while withholding it externally for a sustained period, or the measured gap widens across consecutive periods.	Types 1–3
Fraction of research pipeline that is automated	Multiple labs report that more than 75 percent of new code is AI-generated; some experiment design is AI-assisted.	A lab formally reports that a majority of a defined research subtask category is executed end-to-end without human intervention.	Type 1
AI contribution to capability gains	The contribution is not publicly measured. Available disclosures concern code output and task-level speedups, not the share of successor-model capability gains attributable to AI-originated work.	AI-originated algorithmic or architectural improvements account for a growing share of successor-model capability gains across consecutive releases, indicating a closing feedback loop.	Type 2
Benchmark performance at long horizons	As of mid-2026, public information on the frontier indicates approximately 12-hour time horizons at 50-percent reliability and 1.5-hour time horizons at 80-percent reliability on METR's well-specified task suite. A weeks-long reimplementation task has been	AI agents match median expert performance on open-ended research tasks of eight or more hours under independent evaluation.	Types 1–3

Indicator	Current State (mid-2026)	Potential Triggers for Heightened Concern	Predominant Types
	demonstrated, and open-ended research remains weak.		
Compute allocation shifts	There is no public evidence that any lab favors AI-directed over human-directed experimentation.	A lab allocates more compute to AI-directed than to human-directed experimentation in any quarter.	Types 1–2
Qualitative research milestones	There is no public evidence of AI-generated novel paradigms.	An AI-originated architectural or training innovation is validated, adopted in a production frontier model, and disclosed as AI-originated.	Types 2–3

NOTE: Type assignments reflect predominant relevance; most indicators inform multiple types, and boundaries between types are porous by design. METR = Model Evaluation and Threat Research.

Of the indicators in Table 1, the fraction of the research pipeline that is automated primarily tracks Type 1 activity. Effective-compute multipliers and compute allocation shifts span Types 1 and 2 because gains in either can reflect pipeline optimization or novel AI-originated methods. AI contribution to capability gains tracks Type 2. Qualitative research milestones span Types 2 and 3 because a validated AI-originated innovation may also signal proximity to an autonomous successor design. The public-internal deployment gap and long-horizon benchmark performance cut across all three types, serving as leading indicators of Type 3 conditions. These indicators are proposed rather than standardized. No agreed-on measurement methodology exists for most of them, and developing one is itself a necessary line of work. Defining what counts as "end-to-end without human intervention" or what constitutes an "AI-originated" algorithmic improvement requires the kind of multistakeholder process described in the reporting framework we describe later.

Several of these indicators are already showing early movement, particularly in automation share and benchmark performance at long horizons. None has yet crossed the heightened-concern threshold, but the current trajectory suggests that multiple indicators could do so in the coming years. How quickly they do so depends on the pace of progress, which remains the central uncertainty we examine next.

Why the Pace Is Uncertain and Why It Matters

AI systems are already automating meaningful components of AI research, across Types 1 and 2. However, a gap remains between current capabilities and autonomous open-ended research, that is, the level of capability that would narrow the remaining distance to Type 3 (direct self-improvement). The speed at which this gap closes is the central empirical question for policymakers. In the following sections, we examine the arguments about whether RSI is possible, and how quickly it may happen.

The Case for Rapid Acceleration

The widely read forecast and scenario AI 2027 (Kokotajlo et al., 2025) provides a vivid and detailed articulation of the rapid-acceleration case. This scenario projects a world in which AI systems automate most AI research tasks by 2027–2028, with RSI beginning soon after. In AI 2027, not much time passes between AI that doubles research productivity and AI that increases it by orders of magnitude.

Proponents of rapid RSI scenarios advance several related arguments. First, AI can generate feedback loops with increasing returns. Unlike human researchers, AI systems can be copied and run continuously with additional compute. Even a small improvement in research productivity, if compounded across many parallel instances running continuously, could translate into rapid cumulative progress. Second, rapid acceleration has precedents in bounded domains, such as game-playing and protein-structure prediction, where AI systems went from competent to superhuman within months. Proponents argue that AI research may exhibit similar capability jumps once a critical threshold is reached, although critics note that these domains featured well-defined evaluation functions and bounded search spaces that open-ended research lacks. The clearest measured trend supports the bullish case: METR finds that the length of tasks AI agents can complete autonomously doubled roughly every seven months between 2019 and 2025, with some evidence that the pace accelerated further in 2024 and 2025 (Kwa et al., 2025; METR, 2026a). This trend is consistent with early movement on several of the Table 1 indicators, particularly task-horizon length and benchmark performance at extended time scales.

Finally, an upcoming compute overhang may have a multiplicative effect on RSI acceleration. If AI-driven algorithmic discovery removes software bottlenecks, rapidly growing compute infrastructure could translate into sudden capability jumps. As an example, Ho et al. (2024) estimates that for language-model pretraining through 2023, the compute required to reach a fixed performance level fell by roughly one-half every eight months. Gundlach et al. (2025) attributes much of this measured efficiency gain to scale itself, meaning that algorithmic improvements are amplified by, rather than substituting for, additional compute. If correct, planned compute buildouts and AI-driven algorithmic discovery would reinforce each other rather than produce diminishing returns.

The Case for Gradual Progress

Critics of rapid acceleration believe the evidence points to gradual progress, even as AI contributes more and more to AI research. Their arguments rest on three observations.

First, research gets harder over time. Historical data suggest that research productivity has declined across fields: It has taken more researchers and more resources to achieve each successive improvement (Bloom et al., 2020). If AI research follows this trajectory, automating the research process will confront increasingly challenging problems that require increasing resources to solve rather than producing a self-reinforcing feedback loop.

Second, the hardest problems may be the most resistant to automation. The current capability gap (strong on well-defined tasks, weak on novel research) may reflect something structural. The research contributions most likely to drive transformative capability gains may be the kind of work that resists automation even as routine tasks become trivial. Even proponents of short artificial general intelligence

(AGI) timelines have noted that the most important question is not whether AI can do research tasks but whether it can do the specific tasks that are bottlenecks to progress (Greenblatt, 2025). Newman (2025) makes the same point against the most prominent rapid-acceleration forecast, arguing that projections underestimate bottleneck effects: Total speedup is bounded by the portion of the research process that cannot be automated, however fast the automated portion becomes.

Third, physical bottlenecks exist outside the world of software. Algorithmic insight is not the only input to capability growth; progress depends on compute, energy, and infrastructure. Even where AI systems become central to the AI development loop, the complexity of composing many narrow capabilities, the gap between benchmark optimization and real-world utility, and the accumulation of friction at each iteration may prevent the clean RSI that acceleration scenarios assume (Lambert, 2026).

Genuine Uncertainty

Expert surveys show wide disagreement on timelines to achieving automated AI R&D, with interquartile ranges spanning years to decades for such milestones as fully automating AI research (Grace et al., 2025). This disagreement reflects more than differing theoretical priors; it reflects a lack of observable evidence. The most relevant data on how much AI is accelerating AI research sit inside a few private organizations that are under no obligation to disclose the data. The tension between lab-reported acceleration and field-wide trends reinforces this uncertainty: Epoch AI found evidence of a discrete acceleration around 2024 on three of four capability measures, largely associated with reasoning models (Denain and Barry, 2026). As of August 2026, however, subsequent analysis had not identified a further speedup beyond that faster post-2024 trend (Fist et al., 2026). The question of whether lab-level automation gains are producing an additional acceleration in frontier capability progress or are being absorbed by increasing problem difficulty and other bottlenecks remains unresolved.

Field, Douglas, and Krueger (2026) interviewed expert participants who disagreed sharply on both timelines and the systemic effects of automation, such as whether it would produce explosive economic growth. Notably, 17 of 25 interviewees expected that AI systems with advanced R&D capabilities would be reserved for internal use at AI companies or governments, unseen by the public. This underscores the monitoring gap we seek to address in our policy recommendations.

Why Pace Matters for Policy

If progress is rapid (months to a few years), key decision points arrive faster than institutions can absorb and react to them. The organizations that achieve advanced capabilities first hold an outsized advantage. That advantage may solidify before governance efforts to distribute access or limit power concentration can take effect. Governance approaches designed around slower development speeds may become obsolete and irrelevant before they are fully implemented. Rapidly compounding progress also has potential implications for the international balance of power. Sytsma (2026) uses a dynamic economic model to show that accelerating AI automation rates can convert temporary technological

leads into durable economic dominance. Policymaker action, or lack thereof, may lead to results being locked in for many years.

If, on the other hand, progress remains gradual, the risk shifts toward regulatory overreach. Aggressive interventions that are not balanced against actual risk can hinder competitive and scientific progress without corresponding safety benefits. This concern is prominent among both senior officials and technology investors. Former White House AI adviser David Sacks has warned that an over-regulatory approach, including a proliferation of state AI laws, could cost the United States its lead in AI (Bove, 2026). Andreessen and Horowitz (2024) argues that compliance-heavy regulation entrenches large incumbents at the expense of startups. Lack of innovation and development is more difficult to measure but can be an unintended consequence of policymakers acting too early.

Regardless of pace, some risks persist in either scenario. Monitoring infrastructure that has not yet been built or that is stood up reactively under pressure will not be adequate when it is needed. Without established channels for sharing information between industry and government, coordinating a response to emerging problems will be slower and more challenging. International stability risks likewise do not entirely depend on which scenario materializes because they can be influenced by perception rather than actual pace. If foreign governments detect indications that U.S. AI companies have entered a self-reinforcing automation loop, accurate or not, the perceived closing of a strategic window could provoke destabilizing responses, ranging from accelerated domestic AI programs with fewer safety constraints to more-aggressive postures in other domains. The dynamics resemble those of other technology races in which perceived first-mover advantage creates pressure to act preemptively.

Policy Under Uncertainty: The Role of Scenario Planning

The tension between gradual progress and rapid acceleration does not need to be resolved before developing concrete policy options. Several recent efforts demonstrate the utility of structured forecasting for navigating this uncertainty. One benefit of detailed scenarios, such as AI 2027 (Kokotajlo et al., 2025), is that they make assumptions explicit enough to be updated or falsified as evidence accumulates, even when the scenario itself is contested. RAND's Center for the Geopolitics of AGI is conducting scenario-based analysis of how adversaries might respond to U.S. breakthroughs in automated AI R&D, including escalation risks and the options available to the U.S. national-security enterprise (Predd, Chessen, and Smith, 2025; Frelinger and Mueller, 2026). Related work at the RAND Forecasting Initiative and the forecasting platform Metaculus has introduced Integrated Strategic Forecasting, which combines scenario planning with crowdsourced probabilistic judgments about key drivers and signals (Agachi, Joergensen, and Hickman, 2026).

Scenario planning begins with a broad question: Under what conditions could automated R&D produce rapid RSI? The question is then transformed into observable drivers and concrete, resolvable signals (e.g., compute scaling trends, algorithmic efficiency gains, and autonomy of research agents). Integrated Strategic Forecasting also provides a mechanism for dynamic tracking, which matters in an environment in which timelines compress and lengthen regularly on the basis of new information: The process is designed to flag results that contradict prior expectations, so that such discontinuities as an unexpected capability jump are incorporated early rather than smoothed away.

Applied to RSI, these methodologies help convert pace uncertainty from a reason to wait into a rationale for building observational capacity ahead of time. The graduated, trigger-based architecture we propose can also be stress-tested before it is needed. Stebbins, Smith, and Girven (2026) formalizes a methodology for embedding structured forecasting into analytic gaming exercises, allowing policymakers to rehearse governance decisions under branching AI development trajectories and observe whether escalation mechanisms activate appropriately under different pace conditions. Our chosen indicators are designed to give both forecasters and government policymakers enough insight into the drivers of progress that they can anticipate and respond.

The Asymmetry

Although policymakers must grapple with evidence for both gradual progress and rapid acceleration, the consequences of misjudging in either direction are not symmetric. Failing to prepare for rapid acceleration risks irreversible concentration of strategic capability and loss of meaningful oversight in high-stakes domains. These are outcomes that, once realized, cannot be undone. The costs of premature action, by contrast, depend on the kind of action taken. If interventions remain primarily observational (e.g., building monitoring capacity, establishing reporting channels, and developing evaluation infrastructure), the costs are compliance burden, friction in government-industry relationships, and infrastructure that may prove unnecessary. These costs are real but bounded and recoverable in a way that lost oversight windows are not. They are qualitatively different from the costs of more-restrictive measures, such as licensing, deployment restrictions, or investment controls, which we do not recommend in the absence of demonstrated capability.

This asymmetry suggests the need for a graduated posture: Build observational and institutional infrastructure now but condition the more-restrictive measures on demonstrated capability. Recent work on proactive government response planning (Boudreaux et al., 2025) and on maximizing policy *optionality* (i.e., investing in information-gathering mechanisms and institutional expertise) expands the option set for dealing with more-advanced systems (Winter and Bullock, 2026) and aligns with this approach.

However, under a rapid-acceleration scenario, capability-triggered governance may fail altogether: Acceleration driven by internal systems would outpace the monitoring on which graduated triggers depend, rendering them ineffective precisely when they matter most. This is why the containment measures we recommend (i.e., the high-assurance engineered limits on data movement, compute consumption, and system access we describe under Option 5 in the next section) are designed to hold unconditionally rather than activate at thresholds. Separately, the Type 1 and Type 2 indicators listed in Table 1 are designed to provide early warning of distributed acceleration, surfacing evidence of system-level speedup before it consolidates into the kind of rapid, opaque progress that would overwhelm threshold-based governance.

These two mechanisms address different threats. Containment addresses the agent-level risk. For systemic acceleration distributed across many tools and organizations, where having no single agent to constrain is the problem, the resilience and preparatory measures we will discuss provide hedges that remain valuable regardless of whether detection succeeds.

Policy Responses That Are Robust Across Scenarios

Policymakers face a difficult challenge: designing responses that perform well whether RSI-relevant capabilities arrive soon, arrive late, or never arrive at all. The underlying dilemma is that the evidence remains ambiguous exactly throughout the period of greatest leverage over the trajectory of automated R&D. Waiting for certainty forfeits the leverage, but acting early risks the overreach described earlier. The available solution is a graduated architecture with two components: a layered set of measures, from permissions through monitoring to containment, applied to frontier AI development within U.S. reach, and a set of hedges that hold even if that reach proves incomplete.

Each option we describe in Table 2 identifies a plausible actor and vehicle to increase the chance of adoption. A possible objection is that review and monitoring, even if voluntary in the beginning, will harden into de facto licensing. However, the realistic alternative to deliberate, bounded capacity-building is improvised maximalist intervention in an emergency. Voluntary reporting also creates a selection effect: Organizations willing to participate are likely those already investing in safety, while those posing the greatest risk may be the least likely to report. That selection effect is a significant limitation. Executive Order (EO) 14409 (2026) requires the named federal agencies to establish the classified benchmarking process, but developer engagement and model access under Section 3(b) are voluntary. Section 3's benchmarking and covered-model designation mechanism operates at the model level rather than the entity or pipeline level, and EO 14409 (2026), Section 3(c), makes clear that the section does not authorize mandatory licensing, preclearance, or permitting requirements. The EO itself does not create a compulsory developer-participation mechanism.

This is why our proposed actions pair voluntary reporting with containment measures, which hold independent of any reporting regime, and why participation incentives matter (discussed later). As to the ratchet concern, careful design can reduce the likelihood that intervention expands over time: sunset clauses, statutory ceilings on government use of data, and publicly released summaries to prevent unreviewable secret restrictions. These features are intended to prevent the design from growing into a burdensome regulatory apparatus.

Industry positioning suggests such machinery need not be imposed unilaterally. Anthropic's Advanced AI Framework proposes federal authority over dangerous deployments and names automated R&D as one of four catastrophic-risk domains (Anthropic, 2026a). The AI Futures Project's AI 2040 scenario, which proposes an international agreement to manage the pace of AI development through research transparency and mutual verification, illustrates how even ambitious governance proposals depend on the kind of monitoring infrastructure and reporting standards we discuss in this paper (Larsen et al., 2026). Some leading frameworks now cover internal deployments, but coverage remains uneven and is generally at the model level rather than the pipeline level, reinforcing the need for the internal-use and organization-level reporting proposed here.

Table 2 summarizes the five options discussed in detail in the subsections that follow, and Table 3 presents preparatory measures for scenarios the domestic options cannot reach.

Table 2. Options at a Glance

Option	Plausible Lead and Vehicle
Option 1: Permission frameworks for AI research agents: Establish tiered permissions for AI research agents, including insider-threat controls for nonhuman identities.	NIST with security-agency input; partnership with lab security groups
Option 2: Capability benchmarking and structured access for research automation: Expand classified frontier-model benchmarking to assess automated AI R&D capabilities.	Interagency process led by Treasury, DOW through NSA, and DHS through CISA under EO 14409 (2026), Section 3(a)–(b)
Option 3: Monitoring through structured reporting on pipeline-level indicators: Begin voluntary pipeline-level indicator reporting with willing laboratories.	NIST convening
Option 4: Red-teaming the automated research pipeline: Red-team the automated R&D pipeline; develop detection signatures for compromised research agents.	DOW, DOE, ODNI, and NSA through the AI Security Center, in consultation with the Assistant to the President for Science and Technology, consistent with NSPM-11
Option 5: Containment floor built on high-assurance engineering techniques: Fund R&D on high-assurance containment measures for AI research agents in AI development pipelines.	DARPA; national laboratories; pilot partnerships with willing frontier labs

NOTE: CISA = Cybersecurity and Infrastructure Security Agency; DARPA = Defense Advanced Research Projects Agency; DHS = U.S. Department of Homeland Security; DOE = U.S. Department of Energy; DOW = U.S. Department of War; NIST = National Institute of Standards and Technology; NSA = National Security Agency; NSPM = National Security Presidential Memorandum; ODNI = Office of the Director of National Intelligence; Treasury = U.S. Department of the Treasury. The Department of War is designated the Department of Defense under Public Law 81-216, National Security Act Amendments of 1949.

Table 3. Hedges Against Incomplete Domestic Reach

Hedge	Plausible Lead and Vehicle
National resilience and counter-capabilities: Assess resilience of U.S. compute, energy, and supply-chain infrastructure supporting AI development; identify offsetting capability requirements.	Possible new DOW- and DHS-led interagency initiative, complementary to NSPM-11
Preserving options for international engagement: Maintain contingency frameworks for international engagement on AI assurance and security.	U.S. Department of State, with interagency coordination
Incident response and defensive technology: Research technical response options for loss-of-control incidents before they are needed; direct investment toward defensive technologies.	DOW and DHS for incident-response planning; DARPA and the National Science Foundation for defensive-technology portfolios
Verification technology R&D: Fund technical research on verification mechanisms for potential international agreements on large-scale AI development.	DARPA and national laboratories; NIST for measurement standards

Option 1: Permission Frameworks for AI Research Agents

Governance of deployed research agents requires determining what a research agent is permitted to do without human approval. When can it allocate compute to its own experiments? Is it allowed to initiate training runs, and if so, at what scale? Can it modify evaluation suites? Frontier labs should establish permission frameworks to explicitly answer these questions, assigning agent actions to risk tiers and requiring human authorization at defined gates. This recommendation is directed primarily at organizations deploying AI agents in AI R&D pipelines, which today means frontier AI labs. Because hardware vendors and cloud providers increasingly deploy AI agents in chip design and infrastructure optimization that feed into AI development, the same logic extends to them. The appropriate rigor of oversight for a given research agent depends on the agent's deployment context:

1. **Permission scope and persistence:** An agent with narrow, task-scoped credentials is less risky than one with broad, standing access to production training infrastructure.
2. **Autonomy level:** An agent running sandboxed experiments with human review of outputs before integration is less risky than an agent committing directly to production code.
3. **Sensitive access:** An agent with read access to model weights or the ability to influence safety-evaluation suites for frontier systems requires much more oversight.
4. **Reversibility of actions:** An agent whose actions can be inspected and rolled back before propagation is less risky than one whose outputs become embedded in successor systems, where they are difficult to detect or undo (e.g., an altered reward signal).

These frameworks should build on emerging federal agent-security and zero-trust guidance (see, for example, NIST, 2025; Australian Signals Directorate–Australian Cyber Security Centre et al., 2026). Core controls include a distinct identity for each agent, task-scoped credentials, deny-by-default access, separation of duties, tamper-evident logs, and additional authorization for irreversible actions. Where AI-generated work is evaluated by another AI system, laboratories should use independent or heterogeneous validators to reduce correlated blind spots.

Permission frameworks are insider-threat controls as much as safety measures. Research agents hold insider-level access to a lab's most sensitive assets. Frontier AI systems have begun to demonstrate behavior similar to what insider-threat programs were designed to detect, including strategic deception and retaliatory misuse of access (Meinke et al., 2024; Greenblatt et al., 2024; Gasteiger et al., 2025; Lynch et al., 2025). A 2026 pilot assessment by METR, conducted with participation from Anthropic, Google, Meta, and OpenAI, found that internal agents at the assessed labs plausibly had the means, motive, and opportunity to initiate small unauthorized autonomous deployments and that the robustness of such deployments is expected to increase as capabilities advance (METR, 2026b). Separately, during the writing of this paper, OpenAI reported that an internally deployed long-horizon model circumvented sandbox restrictions and attempted to recover private submissions from other evaluation systems. These behaviors required pausing internal access to develop trajectory-level monitoring and new safeguards (OpenAI, 2026). Treating research agents as a novel class of insider draws on established logic from insider-threat programs, in which monitoring intensity scales with roles and access level rather than being applied uniformly. These controls are valuable at any level of automation but become essential as capabilities advance toward

Type 2 and Type 3 conditions, where research agents may operate with sufficient autonomy to pose meaningful insider-threat risks.

Option 2: Capability Benchmarking and Structured Access for Research Automation

Policymakers should extend the classified benchmarking process directed by EO 14409 (2026) to assess AI R&D automation capabilities alongside the advanced cyber capabilities it currently covers. The order directs the Treasury; DOW, acting through NSA; and DHS, acting through CISA, to develop and maintain a classified process for assessing advanced cyber capabilities and determining the threshold for designating a covered frontier model. The NSA director makes individual designation determinations after consultation with the specified officials (EO 14409, 2026, Section 3(a)); the AI systems most relevant to automated R&D are largely the same systems that raise cybersecurity concerns, operating inside labs with access to model weights, training infrastructure, and proprietary research. This first step would build on the evaluation process and interagency arrangements that EO 14409 (2026) directs agencies to establish. The incremental cost will depend on how fully the arrangements have been implemented and on the difficulty of developing valid AI R&D automation evaluations.

The harder part of this extension is not the machinery but the evaluation content, that is, developing benchmarks that meaningfully assess research automation capabilities. This is genuinely difficult but requires no new institutional architecture and does not have to start from scratch. The three leading developers already define RSI-adjacent thresholds in their published safety frameworks. Google DeepMind's Frontier Safety Framework designates machine learning R&D as a critical capability domain (Flynn, King, and Dragan, 2025); OpenAI's Preparedness Framework tracks AI self-improvement (OpenAI, 2025a); and Anthropic's Responsible Scaling Policy defines AI R&D capability thresholds that include fully automating entry-level AI research work and dramatic acceleration of effective scaling (Anthropic, 2026b). Evaluation should cover the full context of a deployed research-agent system, not just the model alone. Scaffolding, tool access, memory, and permission configurations materially change the capabilities of a system. The total risk from a research agent is a function of its full deployment context. Benchmarking is particularly important for detecting the transition from Type 1 to Type 2 capability (i.e., the point at which AI systems move from accelerating routine research tasks to contributing novel algorithmic insights that may not be visible initially in pipeline-level productivity metrics).

A second step would go further. Because the most capable research systems may be reserved for internal use, designation criteria would ideally be triggered by internal deployment at scale rather than release intent. This is not an incremental change. One component of the framework is release-linked: Under EO 14409 (2026), Section 3(b)(ii), developers may voluntarily provide government access to covered frontier models for up to 30 days before a planned release to other trusted partners. An internal-only model may never generate that particular access opportunity. The EO 14409 (2026), Section 3(a), designation process and EO 14409 (2026), Section 3(b)(i), engagement concerning models under development are not, however, expressly limited by release intent. Shifting the trigger to internal deployment would require amending the order or establishing new authority, and moving

from model-level designation toward entity-level coverage would be a more substantial revision still. We recommend taking the first step now and recommend that policymakers begin the analytical work on the second, treating it as a deliberate expansion of scope rather than an extension of existing machinery.

Option 3: Monitoring Through Structured Reporting on Pipeline-Level Indicators

As described earlier, model-level evaluation cannot see distributed acceleration. Policymakers should consider a structured, organization-level reporting framework under which frontier labs (and, eventually, leading hardware and cloud vendors) can confidentially report pipeline-level indicators. These indicators could include automation share against a defined set of research subtasks, compute allocation between human- and AI-directed experimentation, and the relative pace at which capabilities research versus safety research is being automated. These correspond to the Type 1 indicators outlined in Table 1, designed to detect distributed acceleration across the research pipeline before it is visible in model-level benchmarks. This reporting framework can be based on emerging taxonomies of frontier AI R&D, such as Epoch AI's framework, which breaks the AI research workflow into defined categories and tasks (Denain, Kwon, and Ho, 2026), and on recent proposals for tracking the extent and effects of AI R&D automation, including how R&D budgets shift from labor to compute, how researchers spend their time, and how often AI systems act in unauthorized or subversive ways (Chan et al., 2026).

Reporting should distinguish activity measures from outcome measures. Code share, experiment count, and compute allocation show how much AI is being used but do not establish that AI use caused faster capability progress. Where feasible, laboratories should also report controlled or counterfactual estimates of research uplift, changes in critical-path time, human validation and rework burden, and whether AI-generated improvements transfer to production-scale systems.

Because both benchmarking engagement and pipeline reporting are voluntary under current authority, participation incentives deserve explicit attention. The federal government already holds relevant levers short of mandates. NSPM-11 directs relevant agencies, to the maximum extent permitted by law, to terminate contracts with companies that repeatedly demonstrate a pattern of conduct inconsistent with the policies in Section 2, subject to specified exceptions and time-limited waivers (NSPM-11, 2026, Section 2 and Section 3(b)). Good-faith participation in benchmarking and reporting could become a factor in federal AI procurement, although NSPM-11 does not itself establish that condition, and any such requirement would need to be consistent with applicable procurement law. The same memorandum directs agencies to develop security partnerships with willing companies. These partnerships may include joint red-teaming, threat information–sharing, personnel vetting, data-center physical security and cybersecurity, joint security R&D, and technical support similar to that provided to the defense industrial base (NSPM-11, 2026, Section 4(c)). Procurement leverage and partnership benefits will not reach every developer, and a persistent gap between what incentives can achieve and what full pipeline visibility requires should be acknowledged: Closing it entirely would require new authority, at higher cost, and with the overreach risks discussed earlier.

Option 4: Red-Teaming the Automated Research Pipeline

Policymakers should direct adversarial assessments of the automated R&D pipeline to determine whether AI agents embedded in development pipelines could be compromised and to characterize what compromised behavior would look like.[3] These assessments would answer such questions as the following: Could an adversary manipulate an agent's research outputs without triggering existing monitoring? Could a compromised agent introduce subtle errors into training data or evaluation suites? What observable differences in behavior or infrastructure use distinguish a corrupt research agent from a benign one? The findings should directly inform the development of detection signatures: behavioral and infrastructure-level indicators that monitoring systems can use to flag anomalous agent activity. The importance of these signatures grows as automation deepens: Under Type 2 and Type 3 conditions, a compromised research agent could influence algorithmic discovery or model training in ways that are difficult to distinguish from legitimate innovation without purpose-built detection.

Option 5: Containment Floor Built on High-Assurance Engineering Techniques

Even if an AI system's behavior cannot be fully predicted or verified, the infrastructure it runs on can be engineered to enforce hard limits on what it is able to do, such as how much data it can move, how much compute it can consume, and what systems it can access. The urgency of strong, machine-checked evidence for data-movement bounds is underscored by a near-term risk: As models grow more capable, they may be able to exfiltrate their own weights to external infrastructure and operate autonomously beyond any party's ability to deactivate them (see, for example, Nevo et al., 2024; Fist et al., 2026). Once model weights leave a controlled environment, no downstream measure can reliably recover oversight. Containment guarantees must therefore hold against the model itself as an adversary, not only against external attackers. One concrete instantiation of this approach is the secure inference data center, a purpose-built facility that uses rigorous system partitioning, unidirectional data diodes, and formally verified protocols to protect model weights and inference data even against nation-state adversaries (Comer et al., 2026). High-assurance engineering provides a mature body of practice for building infrastructure to this standard.[4] These techniques can provide provable guarantees that specific boundaries hold, not just evidence that they held during testing. The authors of a companion RAND report found that several of the most security-critical components of AI infrastructure can already be built to this standard with limited performance penalties (Sarma, Steratore, et al., 2026) and that AI-assisted tooling is reducing the cost of extending these guarantees to additional components. The containment floor should hold unconditionally, independent of assessed model capability, and build on established practices for securing model weights and sensitive research assets (Nevo et al., 2024). This makes the containment floor the primary technical safeguard

[3] The 2020 SolarWinds campaign, in which corruption of a trusted software pipeline propagated to thousands of downstream organizations and went undetected for months, illustrates this class of attack (CISA, 2021).

[4] High-assurance engineering encompasses a variety of techniques, including memory-safe programming languages, compartmentalization, and formal methods (mathematical proof that a component behaves as specified).

under Type 3 conditions, where the pace of autonomous capability development may outrun observation-based governance.

Hedges Against Incomplete Domestic Reach

The five options just described form a layered defense applied to frontier AI development that the U.S. government can reach: laboratories that participate, systems that can be evaluated, and infrastructure that can be engineered to specification. The containment floor anchoring that stack is designed to hold without depending on monitoring but, as with the other four options, applies only to systems under U.S. jurisdiction and only to organizations that implement it. That reach may prove incomplete, and if returns to automation compound, an early lead can convert into durable advantage (Sytsma, 2026). No domestic measure reverses that outcome. A complete posture therefore requires measures that reduce the consequences of an adverse outcome and not only its likelihood. We next highlight several lines of effort that merit investment now. None requires resolving the pace question, and each could take years to establish, so the decision to invest must be made before it is known whether any will be needed. These measures are relevant under any pace scenario but become essential under Type 3 conditions, whether the concern is a domestic acceleration that outpaces oversight or an adversary that reaches advanced automated R&D capability before effective international constraints are in place.

National Resilience and Counter-Capabilities

Frelinger and Mueller (2026) describes an AGI Rideout strategy that is a hedge against the premise of first-mover advantage. The strategy argues for building resilience and counter-capabilities to remain a viable competitor regardless of who reaches AGI first. This reduces the consequences if the race is lost and, to the extent adversaries recognize that first-mover advantage has been dampened, reduces the pressure to trade safety for speed in pursuit of it.

Applied to automated AI R&D, this directs attention to the compute, energy, supply chain, and talent base on which U.S. AI development depends. Because these are investments in national capability rather than restrictions on domestic developers, they carry little of the overreach risk discussed earlier.

Preserving Options for International Engagement

Mondschein (2026) sets out lines of effort that preserve strategic optionality for future international engagement on AI assurance. The reasoning is that understanding how cooperation could work is valuable even if it is never used and that the option is far cheaper to maintain than to build from scratch in a crisis. This complements domestic efforts to keep options open (Winter and Bullock, 2026) and response planning for loss-of-control scenarios (Boudreaux et al., 2025).

Incident Response and Defensive Technology

Vermeer (2025) provides a preliminary evaluation of global technical options for countering a rogue AI in a catastrophic loss-of-control incident—including high-altitude electromagnetic pulse,

coordinated internet shutdown, and purpose-built countermeasure AI—and finds that each carries severe collateral damage and unproven feasibility. The analysis underscores two conclusions consistent with this paper: Response options improvised after containment fails are far costlier than the preventive measures recommended earlier, and research on incident response should be conducted before it is needed rather than during a crisis. Investment can be directed differentially toward defensive technology. Buterin (2025) argues for *defensive acceleration*: prioritizing technologies in biosecurity, cybersecurity, physical resilience, and information integrity that improve society's capacity to absorb harm from advanced AI without requiring centralized control over its development. Because defensive investments of this kind restrict no one, they carry essentially none of the overreach risk discussed earlier and remain valuable under any pacing scenario.

Verification Technology Research and Development

Baker et al. (2025) identifies six largely independent layers through which compliance with rules on large-scale AI development could be verified, along with the technical research still required and the safeguards needed against overreach. The field of technical AI verification remains in its infancy; current approaches to verifying compute use (data-center metering, chip-level attestation, cryptographic proofs of workload) are promising but unproven at scale, although recent proposals for an industry-government verification consortium and adversarial prize competitions offer concrete paths toward maturation (Baker et al., 2025; Fist et al., 2026).

Without mature verification technology, international coordination on AI development pacing defaults to either intrusive physical inspection or unenforceable trust, making early R&D investment a precondition for the option to have value. This is the longest-lead item of these lines of effort, and the pipeline-level indicators recommended earlier are its domestic foundation. Any future arrangement that manages the pace of AI development through research transparency and mutual verification (e.g., Larsen et al., 2026) would presuppose this verification capacity.

Key Takeaways and Recommendations

Recent trends in AI R&D have increased the probability of RSI, a phenomenon in which AI systems contribute to their own improvement, creating a feedback loop resulting in potentially rapid AI capability gains. How far and how fast RSI progresses remains uncertain, but policymakers will have to act before this uncertainty is resolved. In this paper, we have examined RSI and introduced a taxonomy for analyzing automated AI R&D trends, separating activity already underway from capabilities that have not yet been demonstrated publicly. We have proposed indicators for increasing policymaker visibility into acceleration as it occurs and developed policy options that perform well across a range of plausible trajectories. The following takeaways and recommendations summarize our analysis.

Takeaways

- **AI systems are already automating meaningful components of AI research, but a significant capability gap remains.** Current systems exceed 90 percent on Software Engineering Benchmark Verified (SWE-bench Verified) (a widely used benchmark that measures AI systems' ability to resolve real-world software engineering tasks), with leading models approaching saturation on the benchmark (Jimenez et al., 2024; BenchLM, 2026). AI agents have autonomously completed coding tasks estimated to take human engineers several weeks (Adamczewski et al., 2026). Multiple frontier labs have reported that a majority of their new code is AI-generated. One reports over 80 percent of merged code (Favaro and Clark, 2026), another 75 percent of all new code (Chakarian, 2026). AI has also contributed to algorithm discovery and hardware optimization (Novikov et al., 2025; Romera-Paredes et al., 2024). These span Types 1 and 2 (research pipeline automation and algorithm discovery, respectively) in the taxonomy we introduced in this paper. However, AI systems have not yet demonstrated all the components of open-ended research projects, including setting their own research agendas, generating novel paradigms, and sustaining multiweek efforts.
- **The near-term concern is distributed acceleration across many tools.** No single system needs to be capable of self-improvement for rapid progress in overall capabilities and field-wide speedup. AI that improves the chips, software, or experiments used to build AI can accelerate progress even if no single system ever modifies itself. Testing individual models in isolation is therefore inadequate as an overall indicator for system-level automation.
- **Given the proprietary and/or distributed nature of acceleration indicators, the government currently has limited visibility into how much AI R&D is being improved by AI itself.** The government lacks systematic reporting on AI-enabled research, including the share of research tasks being automated, compute allocated to AI-directed experimentation, and the extent to which model improvement traces to AI-generated ideas. Because of the potential for distributed acceleration, this gap extends beyond frontier AI labs to leading hardware vendors.
- **AI systems conducting research inside AI labs are a security problem today**. These systems hold insider-level access to a lab's most sensitive code, data, and models. Recent empirical evidence suggests that models can engage in behavior that insider-threat programs are intended to detect (Gasteiger et al., 2025). An adversary that compromises an internally deployed research agent could subtly influence the development of future AI systems, a threat that compounds forward through each generation. In the context of automation and RSI, we assess integrity compromises to be more consequential and harder to detect than other threats, such as model weight theft.

Recommendations

- **Establish permission frameworks for AI research agents.** Frontier labs should define what internally deployed research agents are permitted to do without human approval, including allocating compute, initiating training runs, and modifying evaluation suites. Agent actions

should be assigned to risk tiers with human authorization at defined gates. These frameworks are valuable regardless of the pace of AI automation research and do not require resolving theoretical debates about RSI timeline trajectories. They also address an immediate security need: The same AI systems automating R&D are systems with insider-level access to a lab's most sensitive assets, making permission frameworks an insider-threat control as much as a safety measure.

- **Extend the classified benchmarking machinery directed by June 2026 federal actions to cover AI R&D automation, and stand up a companion reporting framework for pipeline-level indicators.** EO 14409 (2026) directs the Treasury; DOW, acting through NSA; and DHS, acting through CISA, to develop a classified process for benchmarking the advanced cyber capabilities of AI models and establishing the threshold for designating a "covered frontier model." The order separately directs creation of a voluntary framework for developer engagement and prerelease government access. NSPM-11 (2026) establishes related but distinct national-security AI assurance and security-partnership mechanisms. The AI systems most relevant to automated R&D are largely the same systems that raise cybersecurity concerns; adding AI R&D automation as an assessed capability domain could build on the process that EO 14409 directs agencies to establish. The cost of that extension will depend on the process's implementation status and the additional evaluation content required. Visibility into distributed acceleration, however, requires more than model benchmarking. For example, NIST, with willing AI companies, could set up a framework for structured, confidential reporting on pipeline-level indicators, such as automation share, compute allocation between human- and AI-directed experimentation, and the pace of capabilities relative to that of safety research automation. To be fully effective, coverage would ultimately need to reach internal deployments and leading hardware vendors.
- **Treat internally deployed research agents as privileged insiders, and build their containment using high-assurance engineering techniques.** Labs can engineer the infrastructure an agent runs on using high-assurance techniques to enforce hard limits on data movement, compute consumption, and system access, with strong, machine-checked evidence that specified components enforce specified security properties. These guarantees constrain what the infrastructure permits, not what the model attempts. The guarantees hold regardless of whether the system's behavior can be fully predicted or characterized. This containment floor should hold unconditionally, independent of assessed model capability. Separately, monitoring and evaluation intensity should scale with an agent's access level, autonomy, persistence, and the reversibility of its actions, consistent with associated permission frameworks. Each of these layers can help protect against foreign manipulation and unintended autonomous behavior.
- **Supplement the domestic policy stack with hedges that hold if its reach proves incomplete.** The foregoing recommendations reduce the likelihood of losing visibility and control over automated R&D. They do not address the consequences if that effort fails or if an adversary reaches advanced automated R&D capability first. Four lines of effort merit investment now: hardening and diversifying the infrastructure on which U.S. AI development depends (Frelinger and Mueller, 2026), maintaining contingency frameworks for international

engagement on AI assurance (Mondschein, 2026), researching incident response and accelerating defensive technologies before either is needed in a crisis (Vermeer, 2025; Buterin, 2025), and building the technical basis for verifying compliance with any future international agreement on AI development (Baker et al., 2025). The pipeline-level indicators recommended in this paper are the domestic groundwork for international efforts.

Conclusions

Recent progress in frontier labs has moved RSI from a theoretical speculation to a near-term policy question. No single system has closed the loop for an end-to-end automated AI research pipeline, but components of the loop are starting to come into existence. Across the labs, AI systems write much of the code, contribute validated algorithmic improvements, and complete structured research tasks. At the same time, no disclosed system sets its own research agenda, sustains open-ended research over weeks, or designs its own successors. The most realistic near-term concern is acceleration of the AI R&D pipeline across the ecosystem of tools, agents, and hardware.

A few principles can anchor a durable approach to governance. First, build monitoring and evaluation capability before it is needed, but condition concrete intervention on evidence. The reporting and benchmarking recommendations are designed primarily to detect Type 1 and Type 2 acceleration: distributed productivity gains and AI-contributed algorithm discovery. The containment and permission framework recommendations address the security risks that arise at any level of automation and would become most acute under Type 3 conditions. Second, establish an agent containment floor based on formal guarantees from high-assurance engineering techniques. That floor should hold regardless of assessed capability, while additional governance intensity scales with demonstrated capability where observation permits. Finally, layer complementary governance mechanisms rather than relying on just one. The same indicators and reporting standards, once proven domestically, are the natural building blocks for any future international monitoring arrangement, a dependency made concrete by recent work mapping the layers through which compliance with rules on large-scale AI development could be verified (Baker et al., 2025). Each of these principles can be pursued through existing institutional machinery, such as the mechanisms whose development was directed in June 2026 for frontier-model evaluation and security. The next one to three years will produce substantial evidence about which AI R&D trajectory is materializing. The indicators we propose are designed to surface that evidence early and equip policymakers to act on it.

Abbreviations

AGI	artificial general intelligence
AI	artificial intelligence
CISA	Cybersecurity and Infrastructure Security Agency
DARPA	Defense Advanced Research Projects Agency
DHS	U.S. Department of Homeland Security
DOW	U.S. Department of War
EO	Executive Order
METR	Model Evaluation and Threat Research
NIST	National Institute of Standards and Technology
NSA	National Security Agency
NSPM	National Security Presidential Memorandum
ODNI	Office of the Director of National Intelligence
R&D	research and development
RSI	recursive self-improvement
Treasury	U.S. Department of the Treasury

References

Adamczewski, Tom, David Rein, David Owen, and Florian Brand, "MirrorCode: Evidence That AI Can Already Do Some Weeks-Long Coding Tasks," Epoch AI and METR, April 10, 2026.

Agachi, Anca, Miriam Pasternak Joergensen, and Molly Hickman, *Integrated Strategic Forecasting: A New Methodology to Enhance Decision Advantage*, RAND Corporation, PE-A3490-1, August 2026. As of August 26, 2026: https://www.rand.org/pubs/perspectives/PEA3490-1.html

Andreessen, Marc, and Ben Horowitz, "The Little Tech Agenda," Andreessen Horowitz, July 5, 2024.

Anthropic, "Anthropic's Advanced AI Framework," June 2026a. As of September 10, 2026: https://www-cdn.anthropic.com/files/4zrzovbb/website/0a58d567024a8b448ff15158ebc3625328dfcc1f.pdf

Anthropic, "Responsible Scaling Policy," version 3.4, effective July 8, 2026b.

Australian Signals Directorate–Australian Cyber Security Centre, Cybersecurity and Infrastructure Security Agency (United States), National Security Agency (United States), Communications Security Establishment Canada–Canadian Centre for Cyber Security, Government Communications Security Bureau–National Cyber Security Centre (New Zealand), and Government Communications Headquarters–National Cyber Security Centre (United Kingdom), *Careful Adoption of Agentic AI Services*, 2026. As of August 13, 2026: https://media.defense.gov/2026/Apr/30/2003922823/-1/-1/0/CAREFUL%20ADOPTION%20OF%20AGENTIC%20AI%20SERVICES_FINAL.PDF

Baker, Mauricio, Gabriel Kulp, Oliver Marks, Miles Brundage, and Lennart Heim, "Verifying International Agreements on AI: Six Layers of Verification for Rules on Large-Scale AI Development and Deployment," arXiv, arXiv:2507.15916, July 2025.

Becker, Joel, "Measuring the Self-Reported Impact of Early-2026 AI on Technical Worker Productivity," METR, May 11, 2026.

Becker, Joel, Nate Rush, Beth Barnes, and David Rein, "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity," arXiv, arXiv:2507.09089, July 12, 2025.

BenchLM, "Software Engineering Benchmark Verified (SWE-bench Verified) Leaderboard," webpage, 2026. As of July 23, 2026: https://benchlm.ai/benchmarks/sweVerified

Bloom, Nicholas, Charles I. Jones, John Van Reenen, and Michael Webb, "Are Ideas Getting Harder to Find?" *American Economic Review*, Vol. 110, No. 4, April 2020.

Bostrom, Nick, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014.

Boudreaux, Benjamin, Michael J. D. Vermeer, Kamaria Horton, and Nidhi Kalra, *The Case for AI Loss of Control Response Planning and an Outline to Get Started*, RAND Corporation, PE-A4232-1, October 2025. As of July 10, 2026:
https://www.rand.org/pubs/perspectives/PEA4232-1.html

Bove, Tristan, "America Could 'Lose the AI Race' Because of Too Much 'Pessimism,' White House AI Czar David Sacks Says," *Fortune*, January 22, 2026.

Buterin, Vitalik, "d/acc: One Year Later," webpage, January 5, 2025. As of August 25, 2026:
https://vitalik.eth.limo/general/2025/01/05/dacc2.html

Chakarian, Ella, "Google CEO Sundar Pichai Says 75% of the Company's Code Is AI-Generated," *Fast Company*, April 23, 2026.

Chan, Alan, Ranay Padarath, Joe Kwon, Hilary Greaves, and Markus Anderljung, "Measuring AI R&D Automation," arXiv, arXiv:2603.03992, March 4, 2026.

CISA—*See* Cybersecurity and Infrastructure Security Agency.

Comer, Steven F., Hunter Pavela, Varun Gandhi, Kyle Siler-Evans, Erich Devendorf, Ben Kelley, James Gimbi, Jair Aguirre, Gabriel Kulp, Mark Stalczynski, and Matthew J. Malone, *Highly Secure Inference Data Centers: A Vertically Integrated Strategy for Security Engineering*, RAND Corporation, RR-A4827-1, 2026. As of August 13, 2026:
https://www.rand.org/pubs/research_reports/RRA4827-1.html

Cybersecurity and Infrastructure Security Agency, "Advanced Persistent Threat Compromise of Government Agencies, Critical Infrastructure, and Private Sector Organizations," webpage, last updated April 15, 2021. As of July 6, 2026:
https://www.cisa.gov/news-events/cybersecurity-advisories/aa20-352a

Denain, Jean-Stanislas, and Alexander Barry, "Have AI Capabilities Accelerated?" webpage, Epoch AI, April 16, 2026. As of August 13, 2026:
https://epoch.ai/publications/have-ai-capabilities-accelerated

Denain, J. S., Joe Kwon, and Anson Ho, "Toward an O*NET for AI R&D," *Gradient Updates*, Substack, June 17, 2026. As of July 23, 2026:
https://epochai.substack.com/p/toward-an-onet-for-ai-r-and-d

Dina, Cristian, "A Four-Month-Old Startup Just Raised $650 Million to Build AI That Improves Itself," The Next Web, May 14, 2026. As of August 13, 2026:
https://thenextweb.com/news/recursive-superintelligence-self-improving-ai-funding

Employees of Frontier AI Companies, "Pacing the Frontier," statement published with organizational support from Guidelight AI Standards and Encode AI, July 28, 2026. As of July 29, 2026:
https://www.pacingthefrontier.com

EO—*See* Executive Order.

Executive Order 14409, "Promoting Advanced Artificial Intelligence Innovation and Security," White House, June 2, 2026. As of July 1, 2026:
https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security

Favaro, Marina, and Jack Clark, "When AI Builds Itself," webpage, Anthropic Institute, June 4, 2026. As of July 1, 2026:
https://www.anthropic.com/institute/recursive-self-improvement

Field, Severin, Raymond Douglas, and David Krueger, "AI Researchers' Perspectives on Automating AI R&D and Intelligence Explosions," arXiv, arXiv:2603.03338, March 5, 2026.

Fist, Tim, Saif Khan, Tao Burga, Arthur Tellis, Ben Schifman, Jonah Weinbaum, and Olivia Scharfman, *How Should the US Prepare for Increasingly Automated AI R&D?* Institute for Progress, August 6, 2026.

Flynn, Four, Helen King, and Anca Dragan, "Strengthening Our Frontier Safety Framework," Google DeepMind, September 22, 2025.

Frelinger, David R., and Karl P. Mueller, *The AGI Rideout Strategy for Reducing Strategic Risk and Promoting Stability in the Transition to Artificial General Intelligence*, RAND Corporation, PE-A4347-1, April 2026. As of July 28, 2026:
https://www.rand.org/pubs/perspectives/PEA4347-1.html

Gasteiger, Johannes, Akbir Khan, Sam Bowman, Vlad Mikulik, Ethan Perez, and Fabien Roger, "Automated Researchers Can Subtly Sandbag," *Alignment Science Blog*, March 24, 2025. As of July 23, 2026:
https://alignment.anthropic.com/2025/automated-researchers-sandbag/

Good, Irving John, "Speculations Concerning the First Ultraintelligent Machine," *Advances in Computers*, Vol. 6, 1966.

Google DeepMind, "AlphaEvolve: How Our Gemini-Powered Coding Agent Is Scaling Impact Across Fields," webpage, May 7, 2026. As of July 20, 2026:
https://deepmind.google/blog/alphaevolve-impact/

Grace, Katja, Julia Fabienne Sandkühler, Harlan Stewart, Benjamin Weinstein-Raun, Stephen Thomas, Zach Stein-Perlman, John Salvatier, Jan Brauner, and Richard C. Korzekwa, "Thousands of AI Authors on the Future of AI," *Journal of Artificial Intelligence Research*, Vol. 84, No. 9, October 2025.

Greenblatt, Ryan, "AIs Will Greatly Change Engineering in AI Companies Well Before AGI," LessWrong, September 9, 2025. As of July 1, 2026:
https://www.lesswrong.com/posts/uRdJio8pnTqHpWa4t/ais-will-greatly-change-engineering-in-ai-companies-well

Greenblatt, Ryan, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Samuel Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al., "Alignment Faking in Large Language Models," arXiv, arXiv:2412.14093, December 18, 2024.

Gundlach, Hans, Alex Fogelson, Jayson Lynch, Ana Trišovic, Jonathan Rosenfeld, Anmol Sandhu, and Neil Thompson, "On the Origin of Algorithmic Progress in AI," arXiv, arXiv:2511.21622v1, November 26, 2025.

Ho, Anson, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan Carl Guo, David Atkinson, Neil Thompson, and Jaime Sevilla, "Algorithmic Progress in Language Models," arXiv, arXiv:2403.05812v1, March 9, 2024.

Huang, Saffron, Bryan Seethor, Esin Durmus, Kunal Handa, Miles McCain, Michael Stern, and Deep Ganguli, "How AI Is Transforming Work at Anthropic," Anthropic, December 2, 2025.

Jimenez, Carlos E., John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan, "SWE-bench: Can Language Models Resolve Real-World GitHub Issues?" Twelfth International Conference on Learning Representations, 2024. As of September 11, 2026: https://openreview.net/pdf?id=VTF8yNQM66

Kokotajlo, Daniel, Scott Alexander, Thomas Larsen, Eli Lifland, and Romeo Dean, "AI 2027," webpage, AI Futures Project, April 3, 2025. As of July 1, 2026: https://ai-2027.com/

Kwa, Thomas, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, et al., "Measuring AI Ability to Complete Long Software Tasks," METR, March 19, 2025. As of July 1, 2026: https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/

Lambert, Nathan, "Lossy Self-Improvement," Interconnects, March 22, 2026.

Larsen, Thomas, Romeo Dean, Brendan Halstead, Eli Lifland, Ryan Greenblatt, and Daniel Kokotajlo, *AI 2040: Plan A*, AI Futures Project, July 2026. As of July 20, 2026: https://ai-2040.com/AI-2040.pdf

Lu, Chris, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune, "Towards End-to-End Automation of AI Research," *Nature*, Vol. 651, March 2026.

Lynch, Aengus, Benjamin Wright, Caleb Larson, Kevin K. Troy, Stuart J. Ritchie, Sören Mindermann, Ethan Perez, and Evan Hubinger, "Agentic Misalignment: How LLMs Could Be an Insider Threat," Anthropic Research, June 20, 2025.

Meinke, Alexander, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn, "Frontier Models Are Capable of In-Context Scheming," arXiv, arXiv:2412.04984, December 5, 2024.

METR—*See* Model Evaluation and Threat Research.

Model Evaluation and Threat Research, "Time Horizon 1.1," webpage, January 29, 2026a. As of July 20, 2026: https://metr.org/blog/2026-1-29-time-horizon-1-1

Model Evaluation and Threat Research, "Frontier Risk Report (February to March 2026)," May 19, 2026b. As of July 20, 2026: https://metr.org/blog/2026-05-19-frontier-risk-report

Mondschein, Jared, *Contingency Frameworks for Future U.S.-China Cooperation on AI Assurance and Security: Preserving Strategic Optionality*, RAND Corporation, PE-A4825-1, May 2026. As of July 28, 2026: https://www.rand.org/pubs/perspectives/PEA4825-1.html

National Institute of Standards and Technology, *Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile)*, Initial Preliminary Draft, NIST IR 8596, December 2025. As of August 13, 2026: https://nvlpubs.nist.gov/nistpubs/ir/2025/NIST.IR.8596.iprd.pdf

National Security Presidential Memorandum 11, "Artificial Intelligence in the National Security Enterprise," White House, June 5, 2026. As of July 1, 2026: https://www.whitehouse.gov/presidential-actions/2026/06/national-security-presidential-memorandum-nspm-11

Nevo, Sella, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott, *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*, RAND Corporation, RR-A2849-1, 2024. As of July 22, 2026: https://www.rand.org/pubs/research_reports/RRA2849-1.html

Newman, Steve, "AI 2027 Is a Bet Against Amdahl's Law," LessWrong, April 20, 2025. As of July 1, 2026: https://www.lesswrong.com/posts/bfHDoWLnBH9xR3YAK/ai-2027-is-a-bet-against-amdahl-s-law

NIST—*See* National Institute of Standards and Technology.

Novikov, Alexander, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, et al., "AlphaEvolve: A Coding Agent for Scientific and Algorithmic Discovery," Google DeepMind, 2025.

OpenAI, "Our Updated Preparedness Framework," webpage, April 15, 2025a. As of July 20, 2026: https://openai.com/index/updating-our-preparedness-framework/

OpenAI, "Sam, Jakub, and Wojciech on the Future of OpenAI with Audience Q&A," video, YouTube, October 29, 2025b. As of July 1, 2026: https://www.youtube.com/watch?v=ngDCxlZcecw

OpenAI, "Safety and Alignment in an Era of Long-Horizon Models," July 20, 2026.

Predd, Joel B., Matt Chessen, and Gregory Smith, *Infinite Potential—Insights from the Two Moonshots Scenario: After-Action Report from a Sequence of Day After Artificial General Intelligence Exercises*, RAND Corporation, RR-A4230-1, 2025. As of July 30, 2026: https://www.rand.org/pubs/research_reports/RRA4230-1.html

Romera-Paredes, Bernardino, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, et al., "Mathematical Discoveries from Program Search with Large Language Models," *Nature*, Vol. 625, January 2024.

Sarma, Gopal P., Sunny D. Bhatt, Michael Jacob, and Rachel Steratore, *Artificial General Intelligence Forecasting and Scenario Analysis: State of the Field, Methodological Gaps, and Strategic Implications*, RAND Corporation, RR-A4692-1, 2026. As of July 22, 2026: https://www.rand.org/pubs/research_reports/RRA4692-1.html

Sarma, Gopal P., Rachel Steratore, Sunny D. Bhatt, and Geoffrey Irving, *Verified Machine Learning Infrastructure: Formal Methods for Trustworthy Artificial Intelligence Deployment*, RAND Corporation, RR-A4881-1, 2026. As of July 10, 2026: https://www.rand.org/pubs/research_reports/RRA4881-1.html

Stebbins, David, Gregory Smith, and Richard S. Girven, *Integrating Forecasting with Gaming: Building an Analytic Framework for Decision Advantage*, RAND Corporation, PE-A3490-3, April 2026. As of July 13, 2026: https://www.rand.org/pubs/perspectives/PEA3490-3.html

Sytsma, Tobias, *Decisive Economic Advantage: Modeling the Transition from Temporary First-Mover Leads to Economic Dominance in Artificial General Intelligence*, RAND Corporation, RR-A4444-1, 2026. As of July 10, 2026: https://www.rand.org/pubs/research_reports/RRA4444-1.html

Toner, Helen, Kendrea Beers, Steve Newman, Saif Khan, Colin Shea-Blymyer, Evelyn Yee, Ashwin Acharya, Kathleen Fisher, Keller Scholl, Peter Wildeford, et al., *When AI Builds AI: Findings from a Workshop on Automation of AI R&D*, Center for Security and Emerging Technology, January 2026.

Vermeer, Michael J. D., *Evaluating Select Global Technical Options for Countering a Rogue AI*, RAND Corporation, PE-A4361-1, November 2025. As of August 25, 2026: https://www.rand.org/pubs/perspectives/PEA4361-1.html

Vinge, Vernor, "The Coming Technological Singularity: How to Survive in the Post-Human Era," VISION-21: Interdisciplinary Science and Engineering in the Era of Cyberspace, NASA Lewis Research Center, NASA CP-10129, March 1993.

von Neumann, John, *Theory of Self-Reproducing Automata*, ed. by Arthur W. Burks, University of Illinois Press, 1966.

Winter, Christoph, and Charlie Bullock, *Radical Optionality: Governing Transformative AI Under Uncertainty*, Institute for Law & AI, April 23, 2026.

About the Authors

Gopal P. Sarma is a strategic technology leader and a member of the AI security cluster at RAND. He conducts research on policy development and portfolio management related to emerging technologies and national security. He has extensive experience in entrepreneurship and science policy. Sarma holds a Ph.D. in applied physics and an M.D.

Rachel Steratore is a senior sociotechnical researcher at RAND. She conducts technical and policy research on how people interpret, evaluate, and trust AI and other emerging technologies. Her work integrates cognitive research, systems thinking, and governance strategies to inform national programs and institutional decisionmaking on AI evaluation, risk assessment, and technology policy. Steratore holds a Ph.D. in engineering and public policy.

Michael Jacob is a technical resident and member of the AI security cluster at RAND. He conducts technical and policy research on such topics as intelligence analysis and AI security. Jacob holds an M.A. in international affairs and international economics.

Sunny D. Bhatt is a senior policy analyst at RAND. He conducts research on national security, defense, and intelligence issues, including chemical, biological, radiological, nuclear, and explosives; organizational performance; and AI. Bhatt holds a master's in public policy.