Workshop Report

When AI Builds AI

Findings From a Workshop on Automation of AI R&D

Authors

Helen Toner*
Kendrea Beers*
Steve Newman*
Saif Khan*
Colin Shea-Blymyer*
Evelyn Yee*
Ashwin Acharya
Kathleen Fisher
Keller Scholl
Peter Wildeford
Ryan Greenblatt
Samuel Albanie
Stephanie Ballard
Thomas Larsen

***Workshop co-organizer and/or lead drafter**

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

January 2026

Executive Summary

For decades, scientists have speculated about the possibility of machines that can improve themselves. Today, artificial intelligence (AI) systems are increasingly integral parts of the research pipeline at leading AI companies. Some observers see this as evidence that fully automated AI research and development (R&D) is on the way, potentially leading to a rapid acceleration of AI capabilities and impaired ability for humans to understand and control AI. Others see the use of AI for research as a mundane extension of existing software tools.

This Workshop Report shares findings and conclusions from an expert workshop CSET hosted in July 2025. The workshop covered a range of issues related to automation of AI R&D. In this report, 'AI R&D' refers to scientific and engineering work that improves the capabilities of AI systems and 'AI R&D automation' refers broadly to any use of AI that accelerates progress in AI R&D.

Key takeaways from the workshop were as follows:

1. **Increasingly automated AI R&D is a potential source of major strategic surprise.** While experts disagree on likelihood, scenarios are possible in which AI R&D becomes highly automated, the pace of AI R&D accelerates dramatically, and the resulting systems pose extreme risks. This warrants preparatory action now.

2. **Frontier AI companies are already using AI to accelerate AI R&D, and usage is increasing as AI models get more advanced.** New models are often used internally to advance AI R&D before they are released to the public.

3. **Experts' views differ on how rapid and impactful AI R&D automation is likely to be.** Even if the use of AI in AI R&D continues to increase, there is no consensus on whether AI progress is more likely to accelerate or plateau. What's more, because different views are associated with different assumptions about how AI R&D works, new data on how AI R&D automation is progressing in practice may be insufficient to resolve conflicting perspectives. It thus may be difficult to either detect or rule out extreme 'intelligence explosion' scenarios in advance.

4. **Despite challenges in interpreting new evidence, better access to indicators of progress in AI R&D automation would be valuable.** Existing empirical evidence, including existing benchmark evaluations, is insufficient for measuring,

understanding, and forecasting the trajectory of automated AI R&D. More systematic collection of existing indicators—as well as developing ways of gathering new indicators—could provide a significantly clearer picture.

5. **Thoughtfully designed transparency efforts could improve access to valuable empirical information about AI R&D automation, which at present is almost fully dependent on patchy, voluntary releases of information from companies.** While some early transparency mandates on frontier AI development have recently been enacted, they do not focus on indicators of progress in AI R&D automation. Policymakers have a range of options for how to increase visibility of these indicators.

The full report elaborates on these takeaways, including providing examples of how frontier AI companies are using AI for R&D, delving into experts' differing views and assumptions, suggesting priority indicators to track, and laying out policy options and implications.

Table of Contents

Background and Motivation

As artificial intelligence advances, researchers in many scientific fields are looking for ways to accelerate research using AI. An especially consequential field to watch is the use of AI to accelerate research and development (R&D) of AI itself.

In the abstract, the idea of sufficiently capable AI contributing to the development of even more capable AI has long been a fixture in discussions of the future. Many descriptions of this possibility have focused on dramatic scenarios where AI development rapidly becomes fully automated. I. J. Good, an early computer scientist and contemporary of Alan Turing, wrote in 1964 about what would happen once the first smarter-than-human machine was built: "There would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind."[1]

Today, AI researchers are beginning to use AI to accelerate their research. To some observers, this looks like the early stages of the kind of self-improvement feedback loop described by Good (and by more recent thinkers, such as Tom Davidson).[2] To other observers, the use of AI tools within human-driven workflows looks more mundane, comparable to the use of modern programming languages instead of machine code or digital spreadsheets instead of paper.

To explore these differing perspectives on the implications of increasing automation of AI R&D, the Center for Security and Emerging Technology (CSET) held a 1.5-day, closed-door workshop in July 2025, bringing together participants from frontier AI companies, government, academia, and civil society. Participants included leading AI researchers, policy experts, and forecasters. Discussions focused on understanding how AI can contribute to AI research; working through the implications of different possible trajectories for automated AI R&D; identifying indicators that would distinguish between those trajectories; and considering what policy interventions might be warranted. This workshop report summarizes key findings, including points of consensus and continued disagreement between participants.

What Is AI R&D Automation?

Before diving into the workshop's findings about how AI is contributing to AI R&D now and how that might change in the future, two clarifications on terminology:

- **'AI R&D,' as used in this report, refers to scientific and engineering work that improves the capabilities of AI systems.** This term includes, but is not limited to, collecting data, developing new algorithms and training procedures,

designing and manufacturing improved hardware (inasmuch as this activity contributes to improving AI), or creating new tools for AI models to use. The report does not focus on AI's effects on broader scientific R&D (e.g., in materials science, drug discovery, etc.), except in the context of how accelerated AI progress due to automated AI R&D could impact the world at large.

- **'Automated AI R&D,' as used in this report, is a broad term that refers to any use of AI that accelerates progress in AI R&D.** We discuss the possibility of different levels of automation, up to and including **'full automation,'** in which the entire AI R&D process is being driven by AI systems.

For simplicity, when discussing different degrees of automation that fall short of full automation, we do not cleanly differentiate between areas where AI systems augment human researchers vs. areas where AI systems may completely replace human researchers in some parts of the R&D pipeline. We believe that this is a reasonable simplification, given how AI R&D workflows can involve complex mixtures of contributions from human researchers and AI systems (explored further below).

Why Does AI R&D Automation Matter?

Among frontier AI developers, the idea that automating AI R&D could be hugely consequential is widespread: OpenAI, for instance, has announced a goal of building a "true automated AI researcher" by March 2028, which could have "extraordinary potential impacts."[3] In many other circles, however, this prospect is almost unheard of. One motivation for hosting the workshop discussed in this report was to make these currently niche conversations accessible to a wider range of people.

As a starting point, previous work has described two ways in which AI R&D becoming increasingly automated could increase the societal risks of AI: first, by reducing human ability to understand and control AI R&D; and second, by reducing time for humans to navigate rapidly improving AI capabilities.[4]

Discussion during the workshop affirmed these two concerns. First, as AI plays a larger role in research workflows, human oversight over AI R&D processes would likely decline. In principle, it could be possible for human researchers to closely oversee the contributions of AI systems. In practice, however, it is already challenging for researchers to fully understand the outputs of today's AI systems. If AI systems were contributing significantly to AI R&D, the research process would likely produce fewer human-legible outputs with less time for human review, given accelerating effects of automation and strong competitive pressures for leading AI companies to move fast.

Within AI companies, reduced researcher involvement in R&D processes would make it harder for companies to identify, understand, and prevent harms posed by their systems. More speculatively, several attendees emphasized possible risks involving sophisticated AI systems pursuing unwanted ('misaligned') goals, which might emerge accidentally in the training process or be cultivated purposefully by malicious actors. In such scenarios, reduced human oversight could hypothetically allow AI agents to leverage the automated AI R&D process toward their own goals.[5]

Second, faster AI progress resulting from AI R&D automation would make it more difficult for humans (including researchers, executives, policymakers, and the public) to notice, understand, and intervene as AI systems develop increasingly impactful capabilities and/or exhibit misalignment. Relevant risks include enabling bad actors (e.g., by making cyber offense capabilities or bioweapons development more accessible) as well as more diffuse social impacts (e.g., effects on labor markets or human-AI relationships). If research progress accelerates, then there may also be an increasing gap between the most advanced systems available publicly and those that exist inside AI companies, making it harder for outsiders to play an effective role in managing risks and increasing the power imbalance between leading AI companies and other actors.[6]

The greater the degree to which AI R&D is automated, the more strongly these two risk vectors will manifest. Even if AI developers see these effects as undesirable, we should expect competitive pressures to drive them to automate their workflows as rapidly as they can.

The most troubling potential scenarios for automated AI R&D involve compounding acceleration. In these scenarios, AI-driven improvements to AI technologies would build on themselves, resulting in both the capabilities and impacts of AI growing extremely rapidly (an 'intelligence explosion' or 'capabilities explosion').[7] This could start with a steadily increasing fraction of AI R&D activities being augmented or automated by AI. Over time, AI systems would become able to take on more and more of the activities previously reserved for humans. Eventually, the R&D pipeline would become fully automated. If persistent bottlenecks did not emerge, this automation could lead to a rapid expansion in AI capabilities, perhaps including the ability to affect domains such as (non-AI) science and engineering, political and military strategy, manufacturing, cyber operations, and beyond.

Putting AI systems fully in charge of developing even more advanced AI systems is a core element of some of the riskiest future scenarios contemplated by AI experts.[8] While there is little consensus on what impacts to expect from extremely advanced AI

systems, some leaders in the field have warned of the possibility of "an irreversible loss of human control over autonomous AI systems," which "could culminate in a large-scale loss of life and the biosphere, and the marginalization or extinction of humanity."[9]

Key questions about the likelihood of these extreme scenarios are discussed further below. Workshop participants held widely diverging views on the likelihood of such scenarios and on how highly they should be ranked among different risks and harms from AI. However, despite that disagreement, there was broad consensus that intelligence explosion-style scenarios are possible, and that they warrant preparatory action now. The upshot is that, under certain assumptions about how AI R&D automation might progress, major strategic surprise is possible. We may find ourselves in a scenario where enormous acceleration of AI R&D is happening in secret inside AI companies, with few visible effects until the resulting AI systems begin operating externally and their impact is suddenly very large.

Takeaway 1: Increasingly automated AI R&D is a potential source of major strategic surprise. While experts disagree on likelihood, scenarios are possible in which AI R&D becomes highly automated, the pace of AI R&D accelerates dramatically, and the resulting systems pose extreme risks. This warrants preparatory action now.

The next section ("The Present and Future of Automating AI R&D") draws on workshop discussions to provide initial evidence of how AI is already being used in AI R&D workflows inside frontier AI companies, then digs into how AI R&D automation might evolve. The third section of the report ("Indicators to Watch For") lays out several sets of indicators that would be valuable to begin collecting and interpreting to help adjudicate between different forecasts of how the space will evolve. The final section ("Policy Implications and Options") suggests some preparatory policy options and explores relevant considerations for policymakers.

The Present and Future of Automating AI R&D

A key finding of the workshop was that frontier AI companies already use their own best models to help them build even better models. What's more, AI's contributions to AI R&D are growing over time: each time researchers get access to a new generation of more advanced models, the models are able to take on new tasks that previously would have required humans. So, at a basic level, we know that AI is already contributing to AI R&D, and most workshop attendees expect that this practice will only increase over time. This section digs into what we know about the present and future of AI R&D automation in more detail.

What Does AI R&D Consist Of?

Ideally, any conversation about how and whether AI R&D will be automated would begin with a detailed picture of what activities make up AI R&D in the first place. At present, given how new and rapidly changing the field is, there is limited research or data available to inform a widely accepted picture. Nonetheless, some early studies and informal work can offer a starting point.

One simple way to break down AI R&D draws on the two primary roles that drive research forward at most frontier AI companies: research scientists and research engineers. While the work performed by people in these roles often overlaps to some extent, typical tasks could be separated as follows:

- *Research scientist tasks* are about the process of scientific discovery, such as coming up with new hypotheses, designing experiments to run, interpreting unexpected results, prioritizing among research ideas, allocating compute resources for R&D, designing benchmarks, and other capability evaluations.
- *Research engineer tasks* are primarily about coding and/or engineering, such as writing code to implement an experiment; monitoring an experiment to ensure it does not crash; finding and fixing bugs; finding and implementing efficiency gains; building simulation environments; and collecting and generating datasets.

Along these lines, one small, qualitative study of AI researchers and engineers broke AI R&D work tasks into 6 categories: creating hypotheses, designing experiments, running experiments, analyzing results, communication, and studying other work.[10] Mapping these categories onto the scientist/engineer breakdown, we could say that the first two are more scientist-style tasks, the middle two are more engineer-style tasks, and the final two could be performed in both roles.

These breakdowns focus on the core functions of AI R&D, i.e., running experiments and training AI systems. In practice, of course, the overall process of AI R&D at the level of an organization or ecosystem of organizations depends on a wide range of other functions, including raising money; hiring, buying or renting computing hardware; managing that hardware; and managing office space.

How Is AI Being Used for AI R&D So Far?

At present, engineering-focused tasks are one of the areas where AI systems appear to provide most value, especially in coding. The exact productivity gains from using AI to help write code are not yet clear, with one notable study showing that using AI can even slow developers down in some cases.[11] Nonetheless, in practice, many technical staff at frontier AI companies spend a large fraction of their time using AI tools to assist with their work. Public materials from leading AI companies describe heavy use of AI tools by their technical teams, such as the following excerpt from an Anthropic publication:

> *"New data scientists on our Infrastructure team feed Claude Code their entire codebase to get productive quickly. Claude reads the codebase's CLAUDE.md files, identifies relevant ones, explains data pipeline dependencies, and shows which upstream sources feed into dashboards, replacing traditional data catalog tools. [...] For many teams at the company, Claude Code accelerates diagnosis and fixes by analyzing stack traces, documentation, and system behavior in real-time.*
>
> *During incidents, the Security Engineering team feeds Claude Code stack traces and documentation to trace control flow through the codebase. Problems that typically take 10-15 minutes of manual scanning now resolve 3x as quickly."* [12]

Private conversations—including at the workshop—confirm that this usage is real and widespread, not merely marketing hype. In one illustrative anecdote, a highly successful machine learning researcher present at the workshop said that on well-chosen tasks, AI models can do things in 30 minutes that would have taken him hours. Researchers often begin using new AI models for their work before those models are released publicly.

Beyond coding assistants, AI systems are used in many other ways to assist with AI R&D. For example, a paradigm known as "LLM-as-a-judge" is woven throughout many aspects of AI research. LLM-as-a-judge involves using large language models to

evaluate AI-generated outputs in some way that would previously have required human judgement. This technique is now used at massive scale for tasks including training data filtering, safety training, and grading solutions to problems.[13]

While the evidence on productivity gains from AI across the economy is mixed,[14] researchers at frontier AI companies have a dual advantage in making use of AI for their work. First, their familiarity with the AI models they build means that they are well-positioned to scope tasks in order to make the most of the models' strengths while avoiding their weaknesses. Second, AI R&D tasks are some of the use cases that are most salient and familiar for the teams developing the models, so there is natural spillover into developing, testing, and deploying new models in ways that are tailored toward helping with AI R&D tasks. In another anecdote, an employee of a frontier AI company present at the workshop described how he was using internal AI tools to generate around a thousand new reinforcement learning environments to train future models on—far more than he could have created by himself.

Takeaway 2: Frontier AI companies are already using AI to accelerate AI R&D, and usage is increasing as AI models get more advanced. New models are often used internally to advance AI R&D before they are released to the public.

How Far Could AI R&D Automation Go?

The central question of this paper is what the future of AI R&D automation will look like. How much more automated is AI R&D likely to get, how quickly will it get there, and how will that impact society?

At a high level, workshop participants with strong views on this question tended to cluster into two groups: expecting rapid progress towards high degrees of automation and very advanced capabilities, or expecting slower progress that plateaus much earlier. In trying to dig into these differing views during the workshop, we repeatedly found that it was easy for participants to talk past each other due to differing underlying mental models of how automation is likely to proceed. To illustrate, here are a few different dynamics that might be at play to different degrees (depicted in figures 1-3):

- *Productivity-multiplier model (explosion):* AI systems automate an ever-increasing fraction of AI R&D. Initially, they only automate a small fraction of the work and only provide a small productivity boost (say, 20%) over fully human-

driven R&D. But over time, increasingly advanced AI systems can handle more and more complex tasks (for instance, perhaps the current trend of AI systems being able to complete longer and longer tasks continues).[15] As the fraction of AI R&D performed by AI systems increases, the productivity boost over human-only R&D goes to 10x, then 100x, then 1000x. Even if some aspects of AI R&D are initially difficult to automate, the accelerated rate of progress means those bottlenecks are soon overcome. As these improvements compound on themselves, progress accelerates further. Human involvement in—and understanding of—the R&D process drops toward zero. AI systems become far more capable than humans.

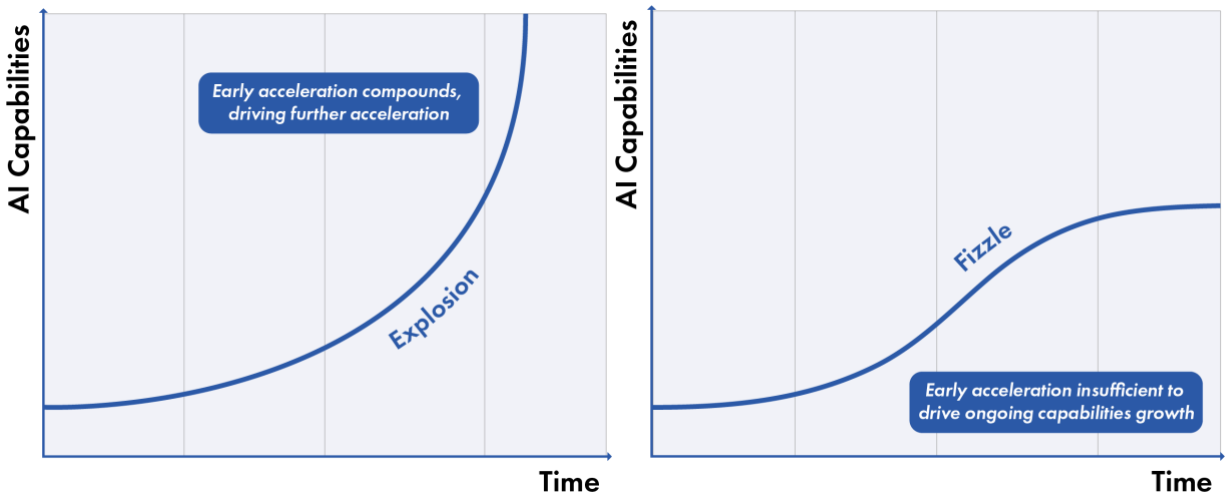
- *Productivity-multiplier model (fizzle):* Similar to the previous model, AI systems automate an increasing fraction of AI R&D. However, in this version, the scientific outputs for a given level of inputs (e.g., compute) are insufficient to drive further compounding improvements in capabilities. AI R&D becomes increasingly automated, but capabilities plateau relatively early.[16]

- *Amdahl's law model:*[*] AI automates some AI R&D activities, but only in specific areas (e.g., writing code and running experiments is automated, but coming up with whole new research programs or operating data centers is not). Even though automation accelerates certain parts of the R&D pipeline, overall progress remains bottlenecked by R&D activities that AI is unable to automate, so full automation is not achieved. AI R&D progresses at a manageable pace, and humans continue to be closely involved in it.

- *Expanding pie model:* As AI automates some AI R&D activities, human researchers repeatedly find that continued progress requires new types of contributions that AI systems cannot yet automate. AI R&D may progress very rapidly, but humans continue to be central to R&D processes.

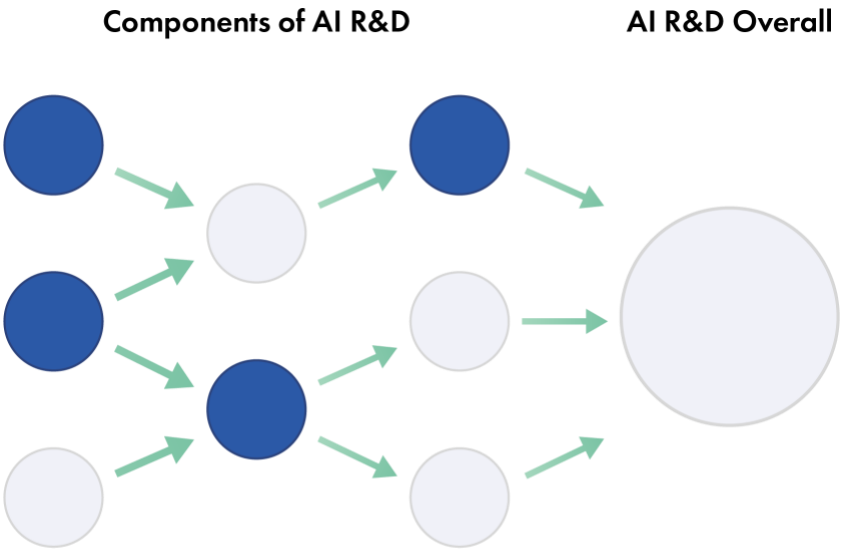
[*] Amdahl's law is a concept from computer science that describes how, if there are multiple potential performance bottlenecks in a system, then optimizing some components will have diminishing returns on the performance of the whole system, since other bottlenecks will take hold.

Figure 1: Stylized Depiction of "Explosion" and "Fizzle" Variants of the Productivity-Multiplier Model



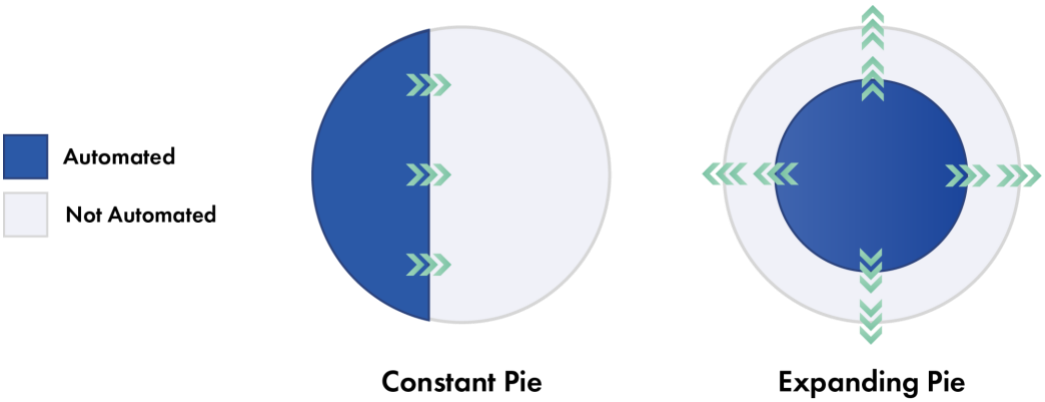
Source: CSET.

Figure 2: Stylized Depiction of Amdahl's Law Model



Source: CSET, adapted from Sayash Kapoor, "AI as Normal Technology Is Often Contrasted with AI 2027," LinkedIn Post, September 5, 2025, https://www.linkedin.com/feed/update/urn:li:activity:7369793551636807681/.

Figure 3: Stylized Depiction of Expanding Pie Model



Source: CSET, adapted from Kapoor, "AI as Normal Technology Is Often Contrasted with AI 2027."

Differing expectations for which of these dynamics will dominate are associated with quite different answers to questions about the 'shape of the curve' of AI progress, such as:

- How rapidly will AI R&D progress? Will progress accelerate over time due to compounding improvements or decelerate due to diminishing returns?
- How likely is it that AI capabilities will reach a level comparable to top human AI researchers? Will early automation accelerate progress towards that point, or will fully matching human performance in AI R&D prove elusive?
- If AI capabilities do reach expert human level, where is the ceiling on performance at different tasks beyond that point? In any given area (e.g., coding, experimental design, business strategy, social engineering), is it possible for AI to far exceed human performance?[17]
- Are there bottlenecks that will hold back AI R&D progress? Some candidates include:
    - Hard-to-automate tasks: AI may continue to struggle with some tasks. For instance, perhaps AI grows increasingly capable at tasks that can be cleanly specified and evaluated, but continues to struggle with 'messy tasks' (discussed further under "Indicators to Watch For," below), meaning that overall AI R&D progress can only progress as fast as humans are able to perform those tasks.

- ‘Last-mile’ data: Training AI models to perform well on any particular real-world task might require collecting large volumes of real-world data specific to that task. Frontier AI companies have a substantial advantage at collecting data on the operations of frontier AI research, but may be bottlenecked on collecting data that would allow their systems to have large real-world impacts on other sectors.[18] This means it is conceivable that they could fully automate AI R&D itself, but still face data bottlenecks in making use of the resulting systems in other domains (more on this below).
- Compute: AI developers’ ability to automate their R&D may become bottlenecked by their access to computational power.

A key finding of the workshop is that it will be difficult to use empirical evidence to adjudicate in advance between two conflicting clusters of views on AI R&D automation. One cluster of views expects rapid progress that leads to extremely advanced AI systems (aka ‘superintelligence,’ AI that is far more capable than humans across all domains); the other expects slower progress that will plateau with AI systems that still fall short of human performance in at least some key areas.[19]

Both of these views rely on assumptions that let them explain why, even if contrary evidence is observed, the situation will revert to expectations later.

For example, someone expecting slow progress might point to a bottleneck, such as how current frontier models struggle with the seemingly simple task of operating a mouse and keyboard, which would seem like an indicator of limited general intelligence. But someone expecting fast progress could respond that this is just an issue with the software tooling available to the models, meaning that once better tooling is available, models’ performance on computer use tasks will rapidly improve to catch up to the underlying capability trends. In general, what looks like a bottleneck to one observer can look like a source of future explosive growth to another.

As a contrary example, someone expecting fast progress might point to the increasing share of tasks that are becoming automatable, and argue that as these are automated, they will speed up progress towards automating an even larger share of tasks. But someone expecting slow progress might instead believe that the tasks currently being automated are systematically different from other tasks, for example, because they are unusually easy to delineate, describe, and assess performance on.[20] If the latter view is true, then rapid progress on automating that set of tasks only means rapid progress towards hitting the next wall. In general, if major bottlenecks or ceilings have not yet

been observed, it is difficult to determine whether that is because they do not exist, or simply because they have not yet begun to bite.

Takeaway 3: Experts' views differ on how rapid and impactful AI R&D automation is likely to be. Even if the use of AI in AI R&D continues to increase, there is no consensus on whether AI progress is more likely to accelerate or plateau. What's more, because different views are associated with different assumptions about how AI R&D works, new data on how AI R&D automation is progressing in practice may be insufficient to resolve conflicting perspectives. It thus may be difficult to either detect or rule out extreme 'intelligence explosion' scenarios in advance.

How Does Automating AI R&D Lead to Real-World Impacts?

Even granting the assumption that AI R&D itself may become highly automated, it is worth explicitly drawing out an additional important disagreement that remains: what is the connection between AI that can automate AI R&D and AI that can have large impacts on the world outside AI companies? Differing views on this question have significant implications for how much AI R&D automation will matter outside the walls of AI companies.

In order to automate AI R&D, AI systems will need to be highly capable at AI-relevant tasks such as writing code, processing data, designing and running AI experiments, coming up with new algorithmic insights, and managing computing resources. To what extent will these capabilities spill over into the many other areas that would be required for AI systems to have significant impacts on the world? These might include making progress in non-AI R&D (e.g., materials science, biomedicine, energy, etc.), manufacturing physical technologies, autonomously running corporations, persuading or manipulating humans, and so on. Some workshop participants find it self-evident that once AI is highly capable at AI R&D tasks, it will have a broad set of capabilities that allow it to succeed at many other tasks. Other participants believe that in the absence of costly efforts to gather data and adapt AI systems to new domains (as described above in the discussion of last-mile data as a potential bottleneck), AI R&D capabilities alone will likely be far from sufficient to enable success in other areas.

Two key elements of this disagreement are *sample efficiency* and *serial experimentation*:

- *Sample efficiency* is a technical term for how well an AI system can utilize a limited supply of data to learn a new task. Higher sample efficiency means less data is needed for the AI system to perform well on a new task.[21]

- *Serial experimentation* refers to the idea that in many fields, even top experts (or teams of top experts) cannot just think their way through a problem—they need to use trial and error to discover what works, drawing on the results of one experiment to inform the design of the next experiment.[22] This may not be a major bottleneck where it is possible to train in simulation (as is already the case with many robotics tasks) or to automate real-world experiments (à la the 'self-driving labs' that are starting to be introduced in fields like materials science).[23] But it is not yet clear how feasible simulation and automated data collection will be in different domains.

AI systems that are capable enough to automate AI R&D may turn out to be either narrowly optimized for AI R&D or more generally capable. If AI systems that are capable of automating AI R&D have high sample efficiency on new tasks and are not bottlenecked by the need for serial experimentation (e.g., due to training in simulation or automated real-world data collection), then AI R&D automation may rapidly yield AI systems with highly general capabilities, poised to have enormous real-world impacts. If AI systems capable of AI R&D automation instead need data-intensive fine-tuning or time-intensive opportunities for trial-and-error learning in order to effectively perform tasks outside of the AI R&D domain, then there will be a lag between when operations inside AI companies are highly automated and when significant real-world effects manifest. Such a lag could allow time for policy interventions or other societal responses.

It would be valuable to gather empirical data that could provide evidence for one side or the other of this debate about how readily AI R&D capabilities transfer to real-world domains. This could include trying to measure whether sample efficiency is improving as AI R&D becomes increasingly automated, tracking the real-world value of training in simulation, and assessing the success of automated data collection efforts.

Indicators to Watch For

Despite the potential challenges in interpreting empirical evidence, participants agreed that efforts to gather and make sense of indicators of the trajectory of AI R&D automation would be highly valuable.

In this section, we lay out indicators identified at the workshop that could shed light on the degree to which AI R&D has been automated, the impact automation is having on AI progress, and the potential future trajectory of AI R&D. It is difficult to find high-fidelity indicators, so we propose multiple categories of measurement, providing distinct vantage points with different strengths and limitations. The categories we suggest are metrics for general AI capabilities, benchmarks for AI capabilities specific to AI R&D, and signs of how AI companies are using AI for AI R&D.

Metrics for Broad AI Capabilities

Well-known weaknesses of current AI systems constrain their usefulness for many tasks, including AI R&D research tasks. In many cases, good metrics for these weaknesses do not yet exist, but if metrics could be developed, then they could be leading indicators for progress on automating AI R&D. The following broad capabilities are likely prerequisites for high levels of automation:[24]

- **Carrying out tasks that take humans a long time to complete**. To reach high degrees of AI R&D automation, AI systems will likely need to reliably complete tasks that would take humans multiple months, if not longer. The AI model evaluation nonprofit Model Evaluation & Threat Research (METR) is tracking frontier AI systems' ability to complete so-called *long-horizon* tasks; this metric appears to be increasing on a somewhat uniform trend line, which makes it unusually helpful for medium-term forecasts.[25]

- **Carrying out "messy" tasks.** Many useful tasks (for AI R&D and otherwise) are "messy" in that they have imprecise specifications, depend on large amounts of context, require interacting with people or other dynamic systems, and/or have conditions for success that are difficult to measure.[26] Messy tasks are drastically underrepresented in current AI evaluations because "cleaner" tasks are inherently easier to specify and evaluate in an efficient, repeatable manner.[27]

- **Assimilating new facts, skills, and ideas on the fly.** At present, AI systems can only carry out a task using information that was either present in their original training data, or that is made available for that specific task (e.g., in the prompt).

They do not have a way to accumulate information or skills over time—unlike humans or other animals, who learn continuously over the course of their lifetimes. Terms such as continuous learning, sample-efficient learning (as discussed earlier), memory, and long-context reasoning are used to refer to different ideas about how to bridge this gap.[28]

Aside from METR's time horizon measurements, there are few existing metrics that can capture progress in the above capabilities. Developing or improving such metrics would be valuable for better understanding the trajectory of AI R&D automation.

Benchmarks for AI Capabilities Specific to AI R&D

Benchmark evaluations are another type of indicator for AI R&D automation designed to directly measure AI capabilities on tasks involved in AI R&D. Benchmarks can serve as leading indicators because capabilities precede adoption. They also have the advantage of providing detailed, reproducible measurements. However, a notable limitation of benchmarks is that they do not reflect real-world conditions.[29]

We present these AI R&D tasks as a "ladder," in roughly increasing order of sophistication, time horizon, information scope, depth of experience required by human workers, and apparent difficulty of automation. For each rung, one or more benchmarks could measure AI systems' ability to carry out relevant tasks:

- **Software & hardware engineering**, including coding, debugging, performance optimization, provisioning and managing compute clusters, and chip design. Several benchmarks already exist for these capabilities.[30]

- **Conducting experiments**, including implementation, data gathering, and analysis. A small number of benchmarks exist for these capabilities.[31]

- **Ideation**, including proposing experiments and identifying takeaways. For example, benchmarks could measure AI systems' ability to find ways to improve another AI model in terms of a metric such as training loss.

- **Strategy & leadership**, including direction setting, prioritization, and orchestration. Benchmarks could hypothetically measure AI capabilities in carrying out entire research campaigns, from ideation to drawing conclusions from entire suites of experiments.

Each rung spans the breadth of activities involved in AI R&D: training, inference, data cleaning, data generation, simulation environments, evaluations, etc. Rungs overlap in

the level of sophistication involved; for instance, architecting a new software system (an engineering task) may require more sophistication than carrying out a simple experiment. “Software and hardware engineering” and “conducting experiments” roughly map to “research engineer” tasks in the breakdown from the earlier section “What Does AI R&D Consist Of,” while “ideation” and “strategy and leadership” are “research scientist” tasks.

Full automation of AI R&D could only occur once upper-rung tasks are automated, but we may not see much data to indicate progress on upper-rung tasks until shortly before full automation. As a partial workaround, we could watch how rapidly capabilities are progressing up the ladder. The time taken to progress from rung one to rung two to rung three may provide some evidence about when the uppermost rung will be reached.

Progress on a given benchmark could be evaluated in terms of the percentage of tasks completed successfully, or by some measure of task difficulty (for instance, the amount of time required by an AI system vs. a skilled human practitioner to successfully complete an AI R&D task).[32] More sophisticated tasks may be evaluated in terms of the *degree* of accomplishment rather than simple success/failure.

As AI capabilities progress up the sophistication ladder, it will become challenging to create realistic benchmarks that can be evaluated at reasonable cost. Currently, we are not aware of benchmarks for the highest two rungs (“ideation” and “strategy and leadership”). Realistic benchmarks, especially at higher rungs of the ladder, would likely require very detailed environments that include training and inference codebases, datasets, records of past experiments, and other resources. Realistic environments could most feasibly be created by cloning real-world environments, but these are either proprietary (and thus only accessible within companies) or public (and thus at risk of being included in model training data, potentially contaminating experimental results). Measuring the quality of model outputs can also be a challenge.[33]

Signs of How Automated AI R&D Is Progressing Inside AI Companies

A third potential source of data about progress toward AI R&D automation is information about the AI R&D activities of frontier AI companies. Compared to benchmark evaluation results, information about usage of AI within AI companies may be more difficult to collect in a consistent manner, may provide a less detailed picture, and would serve as a trailing (rather than leading) indicator. Much of this information may also be proprietary, meaning AI companies may be reluctant to share it. However, unlike benchmarks, this information would reflect real-world usage.

- **Distribution of R&D spending**. The allocation of spending among staff, AI tools, AI agents, synthetic data generation, data collection, compute, and other categories could indicate how much value companies expect AI systems to provide in the AI R&D process. For example, the ratio of compute devoted to primarily human-designed vs. primarily AI-designed experiments could shed light on how much company leadership trusts AI agents' research taste.

- **R&D employment patterns.** For example, a move to lay off entry-level research engineers in favor of hiring research scientists could be a sign that the company foresees near-term automation of the former, but not of the latter. In scenarios where AI systems quickly take on more of AI R&D, this indicator may be slow to provide information because human employees could be sidelined before hiring & firing patterns change.

- **Size & sophistication of tasks being delegated to AI systems.** Quantitative heuristics for the impact of AI on AI R&D tasks could include the frequency of human review or intervention on various tasks as well as the fraction of improvements for which AI would deserve first authorship.

- **Tasks involved in AI R&D.** As some AI R&D tasks are automated, if human researchers keep finding that further progress requires them to take on new types of hard-to-automate tasks, this may provide evidence that there will continue to be a place for human researchers in AI R&D.

- **Gap between internally deployed and publicly released frontier AI models.** To the extent that research automation is important to an AI company's competitive advantage, it may choose to keep its best model to itself for a longer period of time. Disclosures from companies about the gap between any models used internally for AI R&D versus those made available externally could indicate the degree to which AI capabilities are advancing.

- **Measurements of AI R&D progress**. Measuring the pace at which AI training techniques are advancing can help to determine whether research automation is accelerating that pace. For example, some companies have begun to measure changes in *effective compute* (or *compute multipliers*), i.e., the amount of compute needed to achieve a fixed training loss. This is one metric for algorithmic progress and thereby one indicator of overall AI R&D progress.

- **Qualitative impressions from AI researchers**. Researchers at AI companies have front-row seats to AI R&D automation. Although subjective impressions

sometimes mislead,[34] anecdotes from researchers can be illuminating on topics such as which areas of AI R&D currently require human input or how collaborative patterns between humans and AI systems are changing over time.

Some of these signs may also be visible in non-AI companies, such as changes in employment patterns or the size and sophistication of tasks being delegated to AI. This could also provide some evidence about progress in AI R&D automation, though it would be a more indirect signal.

Summary of Indicators

Table 1: Summary of proposed indicators of progress in AI R&D automation.
Indicators that appear especially feasible to collect now (easy targets) are marked with ◎. Indicators that appear especially valuable if collected are marked with ★.

Category: Metrics for Broad AI Capabilities *Advantages: May be leading indicators; relevant for many AI capabilities, so also valuable to collect for other uses.* *Disadvantages: May be difficult to operationalize; do not directly measure AI R&D automation.* Examples: ● ★◎ Carrying out tasks that take humans a long time to complete ● ★ Carrying out "messy" tasks ● ★ Assimilating new facts, skills, and ideas on the fly
Category: Benchmarks for AI Capabilities Specific to AI R&D *Advantages: May be leading indicators; typically structured as quantitative, repeatable metrics, making comparison easier.* *Disadvantages: Unlikely to fully reflect real-world usage; may be difficult to design benchmarks for higher-level capabilities.* Examples: ● ◎ Benchmarks for software & hardware engineering ● ◎ Benchmarks for conducting experiments ● Benchmarks for ideation ● Benchmarks for strategy & leadership

Category: Signs of How Automated AI R&D is Progressing Inside AI Companies

Advantages: Reflects real-world usage; directly measures questions of interest.

Disadvantages: May be trailing indicators; may be difficult to collect in a sufficiently consistent and detailed manner to be useful; may reveal proprietary information.

Examples:

- Distribution of R&D spending
- R&D employment patterns
- ★ Size & sophistication of tasks being delegated to AI
- Tasks involved in AI R&D
- ◎ Gap between internally deployed and publicly released frontier AI models
- ★◎ Measurements of AI R&D progress
- ★◎ Qualitative impressions from AI researchers

This collection of potential indicators is an incomplete, initial list, drawn from a longer set of possibilities generated and ranked during the workshop. We present it here as a strong starting point for further work to understand automated AI R&D. Good next steps could include gathering indicators marked as more feasible (or if they are already being gathered, systematizing this process), developing benchmarks and robust evaluations, and finding ways of gathering indicators marked most valuable.

Takeaway 4: Despite challenges in interpreting new evidence, better access to indicators of progress in AI R&D automation would be valuable. Existing empirical evidence, including existing benchmark evaluations, is insufficient for measuring, understanding, and forecasting the trajectory of automated AI R&D. More systematic collection of existing indicators—as well as developing ways of gathering new indicators—could provide a significantly clearer picture.

Policy Implications and Options

Given the high level of uncertainty about the trajectory of AI R&D automation (as discussed in previous sections), workshop discussions did not settle on strong policy recommendations. Accordingly, this section lays out a range of implications and options for policymakers to consider. Transparency measures were of greatest interest for participants, and are thus discussed in greatest depth.

Transparency Options for AI R&D Automation

Given the level of uncertainty about how rapid and impactful AI R&D automation will be, improving access to empirical evidence about the present and future of AI R&D automation is a valuable near-term policy goal.[35]

At present, anyone with an interest in access to empirical evidence about AI R&D automation is heavily reliant on voluntary releases of information from frontier AI companies. While companies do choose to release some data relevant to AI R&D automation, it tends to be patchy, for several reasons. First, companies often lack incentives to allocate significant resources toward collecting information. Some companies regularly report on capability benchmarks related to AI R&D, like coding capabilities, but many of the indicators listed in Table 1 above are less straightforward to evaluate and disclose. Second, even when companies do collect information, it could be sensitive (commercially or otherwise). For example, the details of how a company incorporates frontier models into its R&D workflows could be important to its competitive advantage. Third, companies may have some incentives to selectively share information, for example, to support certain narratives in order to attract investment.

A small number of laws and regulations relating to transparency around frontier AI development have recently been passed (most notably the European Union's Code of Practice for General-Purpose AI and California's Transparency in Frontier Artificial Intelligence Act, or SB 53). So far, however, these measures do little to create transparency around indicators of AI R&D automation like those discussed in the previous section.

Takeaway 5: Thoughtfully designed transparency efforts could improve access to valuable empirical information about AI R&D automation, which at present is almost fully dependent on patchy, voluntary releases of information from companies. While some early transparency mandates on frontier AI development have recently been enacted, they do not focus on indicators related to automating AI R&D. Policymakers have a range of options for how to increase visibility of these indicators.

Expanding transparency requirements to include information relevant to AI R&D automation could be valuable, but any new transparency measures will need to be carefully designed. Some workshop attendees with government experience spoke favorably of frontier AI companies' willingness to informally share information with relevant government agencies. Adding strict legal requirements for information disclosure could incentivize companies to be more careful about the information they share, for example by giving legal teams more say over what is shared. It may also make it more difficult for the information being shared to change flexibly over time as our collective understanding of the situation improves, if outdated requirements have been codified. Where possible, policymakers can also make use of softer mechanisms for gathering information, such as informal requests, invited testimony, and voluntary industry-government partnerships. Workshop attendees were generally supportive of increased transparency, but held a range of views on the extent to which transparency measures should be voluntary vs. mandatory and on what should be shared privately with government agencies vs. publicly.

Options for increasing transparency of indicators relating to AI R&D automation include:

- **Disclosure of key indicators** (including either voluntary or mandatory disclosure, with information disclosed either to the government or to the public). The U.S. government has existing mechanisms for private disclosure, including voluntary technical partnerships between industry and the Center for AI Standards and Innovation within the National Institute of Standards and Technology. If needed, existing authorities (such as the Defense Production Act) or new disclosure authorities could be leveraged to require industry to share information privately with the government. Transparency measures could include two tiers: more detailed, sensitive information sharing with the government, and more general reporting to the public. As discussed above,

transparency mandates of this kind could also inadvertently constrain informal information channels, so policymakers should carefully consider the costs and benefits of different options.

- **Targeted whistleblower protections** can increase the likelihood that key information is shared publicly, especially in more extreme potential scenarios. Most existing whistleblower protections center around illegal conduct, meaning that employees of frontier AI companies may refrain from whistleblowing about highly concerning but unregulated activities.[36] New whistleblower protections designed for frontier AI employees, as in SB 53, could ameliorate this dynamic.

Other Policy Implications

Beyond transparency, the potential for highly automated AI R&D has implications for several other areas of AI governance.

- **Risk management within AI companies** is crucial for internal deployments of frontier AI systems for AI R&D specifically. Several AI companies already include automated AI R&D capabilities in their safety frameworks as triggers for increased safety and security measures, but these frameworks are nascent.[37] Further work is needed to develop best practices for risk management that explicitly cover the internal deployment setting, implement these best practices at frontier AI companies, and establish oversight.[38] Policymakers developing broad regulatory frameworks should consider whether and how to cover internal deployments, not just external deployments. For instance, internal deployments may be out of reach of the EU AI Act if those deployments occur outside the EU.

- Increasing automation of AI R&D would accelerate AI capability progress, which would increase the urgency of **resilience-focused policy measures to prepare for highly advanced AI** in general. Low-regret policy measures include building societal resilience against AI-exacerbated threats such as cyberattacks, developing break-glass plans in case of rapid capability improvements, and preparing for potential labor market upheaval.

- High levels of AI R&D automation would likely raise the stakes for **compute advantages between companies and countries**. If AI R&D becomes highly automated, access to compute will likely be a significant determinant of how much a given organization can accelerate its AI research.[39] If so, the supply chain for these compute resources would be a crucial strategic resource. Advantages

in AI capability would accrue to companies with the largest stocks of overall AI compute, which could be used to run more parallel copies of automated AI R&D pipelines, or to run the pipelines faster. From the perspective of preparing for the possibility of such a world, compute controls could allow the U.S. and allies to slow competitors' ability to automate AI R&D at scale.

- The United States and allied governments may wish to consider intelligence collection options as a way to gain more information about AI R&D automation inside foreign companies not subject to their jurisdiction, especially if those companies are not participating in voluntary transparency efforts.

- Automated AI R&D could challenge the current paradigm of **open-weight AI models** following closely behind the frontier. As previously mentioned, in a scenario with high levels of AI R&D automation with compounding gains, the gap may increase between proprietary frontier models deployed inside AI companies for AI R&D and publicly available models, including open-weights models. AI governance approaches predicated on there being little difference between closed and open models may therefore be ineffective in worlds where AI R&D automation is proceeding apace behind closed doors.

These policy options and implications are an initial set of suggestions given the current state of play. Further work is needed to map out policy options that would be sufficient to manage the potentially extreme risks inherent to rapid-acceleration scenarios.

Authors

Helen Toner is interim executive director at CSET.

Kendrea Beers was a Horizon junior fellow at CSET until October 2025.

Steve Newman is chairman and president of the Golden Gate Institute for AI.

Saif Khan is a distinguished technology fellow at the Institute for Progress.

Colin Shea-Blymyer is a research fellow on the CyberAI Project at CSET.

Evelyn Yee is a summer research assistant at CSET.

Ashwin Acharya is an AI policy and strategy researcher.

Kathleen Fisher is the former office director of the Information Innovation Office (I2O) at the Defense Advanced Research Projects Agency (DARPA) and the incoming CEO of the Advanced Research and Invention Agency (ARIA).

Keller Scholl is a Ph.D. candidate at the RAND Graduate School.

Peter Wildeford is chief strategy officer and cofounder at the Institute for AI Policy and Strategy.

Ryan Greenblatt is chief scientist at Redwood Research.

Samuel Albanie is a machine learning researcher focused on evaluations.

Stephanie Ballard a director of responsible AI practice at Microsoft.

Thomas Larsen is a researcher at the AI Futures Project.

Acknowledgements

The authors are grateful to several workshop participants who contributed greatly to the discussion but were unable to participate in the writing process: Daniel Freeman, Jessica Ji, Jonas Sandbrink, Nicholas Carlini, Sayash Kapoor, and Tammy Masterson. We also thank John Bansemer and Cara LaPointe for feedback on drafts, as well as Shelton Fitch and Jason Ly for editorial and design support.

Document Identifier: doi: 10.51593/20250027

Endnotes

[1] Irving John Good, "Speculations Concerning the First Ultraintelligent Machine," in *Advances in Computers, vol. 6* (Elsevier, 1966), 31–88.

[2] See, e.g., Daniel Eth and Tom Davidson, "Will AI R&D Automation Cause a Software Intelligence Explosion?," Forethought, March 25, 2025, https://www.forethought.org/research/will-ai-r-and-d-automation-cause-a-software-intelligence-explosion; Tom Davidson, Rose Hadshar, and William MacAskill, "Three Types of Intelligence Explosion," Forethought, March 17, 2025, https://www.forethought.org/research/three-types-of-intelligence-explosion.

[3] Sam Altman (@Sama), "Yesterday We Did a Livestream. [...]," X (formerly Twitter), October 29, 2025, https://x.com/sama/status/1983584366547829073?lang=en.

[4] Joshua Clymer, Isabella Duan, Chris Cundy et al., "Bare Minimum Mitigations for Autonomous AI Development," arXiv preprint arXiv:2504.15416 (2025), https://doi.org/10.48550/arXiv.2504.15416.

[5] Oscar Delaney, "Managing Risks from Internal AI Systems" (Institute for AI Policy and Strategy, July 21, 2025), https://www.iaps.ai/research/managing-risks-from-internal-ai-systems; Joe Benton, Misha Wagner, Eric Christiansen et al., "Sabotage Evaluations for Frontier Models," arXiv preprint arXiv:2410.21514 (2024), https://doi.org/10.48550/arXiv.2410.21514; Jan Betley, Daniel Tan, Niels Warncke et al., "Emergent Misalignment: Narrow Finetuning Can Produce Broadly Misaligned LLMs," arXiv preprint arXic.2502.17424 (2025), https://doi.org/10.48550/arXiv.2502.17424.

[6] If this scenario eventuates, it would go against the "iterative deployment" philosophy articulated by OpenAI and implicitly followed by many other AI companies: "Crucially, we believe that society must have time to update and adjust to increasingly capable AI, and that everyone who is affected by this technology should have a significant say in how AI develops further. Iterative deployment has helped us bring various stakeholders into the conversation about the adoption of AI technology more effectively than if they hadn't had firsthand experience with these tools." "Our Approach to AI Safety," OpenAI, April 5, 2023, https://openai.com/index/our-approach-to-ai-safety/.

[7] Eth and Davidson, "Will AI R&D Automation Cause a Software Intelligence Explosion?"

[8] For example, "Autonomous Replication or Improvement" is the first suggestion for a red line that should not be crossed in Geoffrey Hinton, Andrew Yao 姚期智, Yoshua Bengio et al., "IDAIS-Beijing 2024 Statement," International Dialogues on AI Safety, September 24, 2024, https://idais.ai/dialogue/idais-beijing/.

[9] Yoshua Bengio, Geoffrey Hinton, Andrew Yao 姚期智 et al., "Managing extreme AI risks amid rapid progress," *Science* 384, no. 6698 (May 2024): 842–45.

[10] David Owen, "Interviewing AI Researchers on Automation of AI R&D" (Epoch AI, August 2024), https://epoch.ai/blog/interviewing-ai-researchers-on-automation-of-ai-rnd.

[11] Joel Becker, Nate Rush, Elizabeth Barnes, and David Rein, "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity," arXiv preprint arXiv:2507.09089 (2025), https://doi.org/10.48550/arXiv.2507.09089.

[12] "How Anthropic Teams Use Claude Code," Anthropic (blog), July 24, 2025, https://claude.com/blog/how-anthropic-teams-use-claude-code. For a more recent and more in-depth look at Anthropic's internal use of AI tools, see Saffron Huang et al., "How AI Is Transforming Work at Anthropic," Anthropic, December 2, 2025, https://anthropic.com/research/how-ai-is-transforming-work-at-anthropic/.

[13] Shivani Kapania, Stephanie Ballard, Alex Kessler, and Jennifer Wortman Vaughan, "Examining the Expanding Role of Synthetic Data throughout the AI Development Pipeline," arXiv preprint arXiv:2501.18493 (2025), https://doi.org/10.48550/arXiv.2501.18493.

[14] Aditya Challapally, Chris Pease, Ramesh Raskar et al., "The GenAI Divide: State of AI in Business 2025" (MIT Networked-Agents and Decentralized AI (NANDA), July 2025), https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf.

[15] A widely cited study on the "time horizon" of AI systems shows that as they get more advanced, AI systems have a higher success rate at tasks that take humans more and more time. Thomas Kwa, Ben West, Joel Becker et al., "Measuring AI Ability to Complete Long Tasks," arXiv preprint arXiv;2504.14499 (2025), https://doi.org/10.48550/arXiv.2503.14499.

[16] For further exploration of the "explosion" vs. "fizzle" models, see Eth and Davidson, "Will AI R&D Automation Cause a Software Intelligence Explosion?."

[17] For further discussion of this question, see the section on "Headroom" in Katja Grace, "Counterarguments to the Basic AI X-Risk Case," AI Impacts (blog), August 31, 2022, https://aiimpacts.org/counterarguments-to-the-basic-ai-x-risk-case/.

[18] As an example of a frontier AI company confronting the last-mile data problem, OpenAI has a team of "forward-deployed engineers" that collaborate with customers, e.g. working directly with farmers in Iowa during a collaboration with John Deere. Gergely Orosz, "What Are Forward Deployed Engineers, and Why Are They so in Demand?," The Pragmatic Engineer, August 12, 2025, https://newsletter.pragmaticengineer.com/p/forward-deployed-engineers.

[19] For illustrative examples of each cluster, see Daniel Kokotajlo, Scott Alexander, Thomas Larsen et al., "AI 2027," accessed November 25, 2025, https://ai-2027.com/; Arvind Narayanan and Sayash Kapoor, "AI as Normal Technology" (Knight First Amendment Institute at Columbia University, April 2025), http://knightcolumbia.org/content/ai-as-normal-technology. For points of agreement between these clusters that originated from discussions at this workshop, see Sayash Kapoor, Arvind Narayanan, Daniel Kokotajlo et al., "Common Ground between AI 2027 & AI as Normal Technology," Asterisk Magazine, November 12, 2025, https://asteriskmag.substack.com/p/common-ground-between-ai-2027-and.

[20] Steve Newman, "A Project Is Not a Bundle of Tasks," Second Thoughts, November 3, 2025, https://secondthoughts.ai/p/a-project-is-not-a-bundle-of-tasks.

[21] In some cases, even if real-world data is limited, synthetic data or data augmentation can allow a model to perform well even without being particularly sample efficient.

[22] Ben Reinhardt, "Teaching AI How Science Actually Works" (Institute for Progress, August 2025), https://ifp.org/teaching-ai-how-science-actually-works/.

[23] Heesun Choi, Cindy Crump, Christian Duriez et al., "On the use of simulation in robotics: Opportunities, challenges, and suggestions for moving forward," *Proceedings of the National Academy of Sciences of the United States of America* 118, no. 1 (2021); Milad Abolhasani and Eugenia Kumacheva, "The Rise of Self-Driving Labs in Chemical and Materials Sciences," *Nature Synthesis* 2, no. 6 (January 2023): 483–92.

[24] These broad capabilities are also likely prerequisites for AI to have a significant impact on the wider world, so they may also provide evidence about the disagreement described at the end of the previous section about how AI R&D capabilities transfer to real-world impacts.

[25] Kwa et al., "Measuring AI Ability to Complete Long Tasks"; "How Does Time Horizon Vary Across Domains?," METR (blog), July 14, 2025, https://metr.org/blog/2025-07-14-how-does-time-horizon-vary-across-domains/.

[26] For a detailed list of factors that make a task "messy," see Tables 10 and 11 in Kwa et al., "Measuring AI Ability to Complete Long Tasks."

[27] OpenAI's GDPval, released between the workshop and the publication of this report, is an interesting example of a benchmark that attempts to capture performance on messier tasks. As OpenAI notes, however, it is still limited to relatively clean versions of the tasks it includes: "While [GDPval] spans 44 occupations and hundreds of knowledge work tasks, it is limited to one-shot evaluations, so it doesn't capture cases where a model would need to build context or improve through multiple drafts." Tejal Patwardhan, Rachel Dias, Elizabeth Proehl et al., "GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks," arXiv preprint arXiv:2510.04374 (2025), https://doi.org/10.48550/arXiv.2510.04374.

[28] JS Denain and Anson Ho, "The huge potential implications of long-context inference," Epoch AI, September 19, 2025, https://epochai.substack.com/p/the-huge-potential-implications-of.

[29] Thomas Woodside and Helen Toner, "Evaluating Large Language Models" (Center for Security and Emerging Technology, July 2024), https://cset.georgetown.edu/article/evaluating-large-language-models/.

[30] Existing benchmarks include RE-Bench (7 research engineering tasks), SWE-bench (GitHub issue solving tasks), and PaperBench (tasks to recreate existing AI research papers). See: Hjalmar Wijk, Tao Lin, Joel Becker et al., "RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts," arXiv preprint arXiv:2411.15114 (2024), https://doi.org/10.48550/arXiv.2411.15114; Carlos E. Jimenez, John Yang, Alexander Wettig et al., "SWE-Bench: Can Language Models Resolve Real-World GitHub Issues?" arXiv preprint arXiv:2310.06770 (2023), https://doi.org/10.48550/arXiv.2310.06770; Giulio Starace, Oliver Jaffe, Dane Sherburn et al., "PaperBench: Evaluating AI's Ability to Replicate AI Research," arXiv preprint arXiv:2504.01848 (2025), https://doi.org/10.48550/arXiv.2504.01848.

[31] See e.g., CORE-Bench, a collection of 270 tasks involved in reproducing results from 90 scientific papers drawn from computer science, medicine, and social science. Zachary S. Siegel, Sayash Kapoor, Nitya Nagdir et al., "CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark," arXiv preprint arXiv:2409.11363 (2024), https://doi.org/10.48550/arXiv.2409.11363.

[32] See Kwa et al., "Measuring AI Ability to Complete Long Tasks."

[33] See "Research Update: Algorithmic vs. Holistic Evaluation," METR (blog), August 13, 2025, https://metr.org/blog/2025-08-12-research-update-towards-reconciling-slowdown-with-time-horizons/.

[34] Becker et al., "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity."

[35] For more on "reducing uncertainty as a policy goal" and "evidence-seeking AI policy," see Narayanan and Kapoor, "AI as Normal Technology"; Stephen Casper, David Krueger, and Dylan Hadfield-Menell, "Pitfalls of Evidence-Based AI Policy," arXiv preprint arXiv:2502.09618 (2025), https://doi.org/10.48550/arXiv.2502.09618.

[36] "The Urgent Case for the AI Whistleblower Protections: Congress Must Pass the AI Whistleblower Protection Act," National Whistleblower Center, September 19, 2025, https://www.whistleblowers.org/campaigns/the-urgent-case-for-the-ai-whistleblower-protections-congress-must-pass-the-ai-whistleblower-protection-act/.

[37] "Issue Brief: Components of Frontier AI Safety Frameworks" (Frontier Model Forum, November 2024), https://www.frontiermodelforum.org/updates/issue-brief-components-of-frontier-ai-safety-frameworks/.

[38] Charlotte Stix, Matteo Pistillo, Girish Sastry et al., "AI behind Closed Doors: A Primer on the Governance of Internal Deployment," arXiv preprint arXiv:2504.12170 (2025), https://doi.org/10.48550/arXiv.2504.12170.

[39] Note that expecting access to compute to be a source of advantage in a world of highly automated AI R&D does not depend on the AI models themselves being extremely compute intensive. Even if future research leads to new architectures or approaches that are significantly more compute-efficient, organizations with more access to compute will be able to run more copies of those models at higher speed, so compute access will still provide an advantage.