METR

Model Evaluation and Threat Research

— 440 N Barranca Ave #3345 Covina, CA 91723 —

March 15, 2025

Faisal D'Souza, NCO
2415 Eisenhower Avenue
Alexandria, VA 22314

Response to Request for Information on the Development of an Artificial Intelligence (AI) Action Plan

METR is a research nonprofit focused on developing the science of assessing AI systems for capabilities that could pose catastrophic risks to public safety and national security. In 2024 and early 2025, METR conducted evaluations of systems like GPT-4.5[1] and Claude 3.5 Sonnet[2] in partnership with their respective developers, typically before they were released to the public. METR has also advised many of the leading AI developers to help them establish frameworks for scaling up their risk mitigations in proportion to the measured or forecasted abilities of their systems.[3]

METR appreciates the opportunity to provide input into the shaping of the upcoming Artificial Intelligence Action Plan. We believe that this is a critical time for the U.S. government and other actors to plan soberly for the risks (and opportunities) that may come from further significant advancements in the state of AI capability. This response begins by outlining a few considerations that we believe are important for informing the general direction of the Plan. It then recommends several concrete actions that we believe merit prioritization within the Plan.

[1] https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf
[2] https://metr.org/blog/2025-01-31-update-sonnet-o1-evals/
[3] http://metr.org/faisc

Summary

We believe it will be important to plan around the following three major factors:

1. AI systems are rapidly improving at carrying out tasks independently over long time horizons.
2. AI systems are improving on AI R&D tasks, in a feedback loop that could be very destabilizing.
3. AI systems will plausibly end up with unintended and/or malign goals, conceal this fact, and resist attempts to remove these goals.

In light of these factors, we recommend that the AI Action Plan include at least the following four priority actions:

1. Collaborate with the private sector to improve the information security and internal controls achievable by the leading U.S. AI developers.
2. Lead creation of narrowly-tailored formal standards for how and when AI developers should measure:
    a. Capabilities that would directly threaten public safety or national security, specifically CBRN weapons development and cyber-offense.
    b. Capabilities that would amplify risks from AI systems, specifically long-horizon autonomy, automated AI R&D, and control subversion.
3. Prepare and consider interventions on the AI development and deployment process, as future circumstances may warrant, to:
    a. Require reporting of training runs and deployments above some threshold of scale.
    b. Require external oversight of plans for public safety and national security above some threshold of scale.
    c. Require heightened information security and other mitigations (including internal controls) above some threshold of capability.
4. Measure the degree of R&D automation and AI reliance occurring within the leading U.S. AI developers and other critical industries.

Key Factors Informing our Recommendations

AI systems are rapidly improving at carrying out tasks independently over long time horizons.

METR believes that the world should be preparing for AI systems that can pursue goals in the real world over long timespans without human intervention. AI developers once thought that we were decades away from building systems that can do any mental task that humans can,[4] or that can autonomously perform most economically-valuable work.[5] Now, systems like these are no longer purely in the realm of science fiction: the leading AI developers claim to have line-of-sight to their arrival. To quote one prominent executive at such a company,[6]

> We are now confident we know how to build AGI as we have traditionally understood it. We believe that, in 2025, we may see the first AI agents "join the workforce" and materially change the output of companies.

Likewise, another executive whose company is developing and deploying advanced AI systems recently expressed,[7]

> Making AI that is smarter than almost all humans at almost all things will require millions of chips, tens of billions of dollars (at least), and is most likely to happen in 2026-2027.

Surveyed experts agree that AI systems will quickly surpass key milestones, and believe there is a meaningful (>10%) chance that by 2028 we will reach the point at which AI systems eclipse human performance on all tasks.[8] At or near this point, AI systems would be able to run businesses, create applications, even develop new technologies without human input.

---

[4] https://www.wired.com/story/deepmind/
[5] https://openai.com/charter/
[6] https://blog.samaltman.com/reflections
[7] https://darioamodei.com/on-deepseek-and-export-controls
[8] https://arxiv.org/abs/2401.02843

The development of highly-capable AI systems like these would introduce profound new risks to the United States' national interest. If these systems are set towards bad goals—whether through error during development or through deliberate effort—they would plausibly be capable of doing harm (or simply causing chaos) much faster than previously possible. Additionally, as soon as these systems are *capable* of accomplishing tasks without any humans in the loop, there will be strong incentives for organizations to *in fact* remove humans from the loop, due to cost and speed advantages. In a scenario like this, even if AI agents are set towards benign goals, many humans could quickly be cut out from critical decision-making. This shift could also lead to substantial concentration of power into the hands of a small set of actors that control the "AI labor force", giving them leverage over the rest of society much larger than that of any existing corporation or government.

We have begun to see real evidence that AI systems are in fact becoming more capable of carrying out long-duration autonomous work. In forthcoming research from our organization, we evaluate recent advanced AI systems and analyze the overall trend of the human *time horizon* over which they can do tasks at a given level of reliability, measured in terms of how long it would take an expert to complete the same tasks. We find evidence that this time horizon has grown by >100X in the past 5 years, from systems that could only handle tasks of a few seconds to now over 30 minutes. Extrapolating the trendline in our model would imply that AI systems will be able to handle month-long tasks at or above the level of human experts within 5 years.

AI systems are improving on AI R&D tasks, in a feedback loop that could be very destabilizing.

METR believes that AI systems are becoming increasingly capable at performing the very research and development tasks needed to create more advanced AI systems. AI developers are actively pursuing this capability as a strategic priority,[9] with significant resources dedicated to accelerating components of the AI development pipeline like creating datasets and automating feedback for reinforcement learning.

---

9 https://www.businessinsider.com/mark-zuckerberg-meta-ai-replace-engineers-coders-joe-rogan-podcast-2025-1#:~:text=Mark%20Zuckerberg%20said%20Meta%20will,notes%20and%20reduce%20DEI%20initiatives

As a recent international scientific report on the state of AI capabilities and risks noted,[10]

> General-purpose AI-based tools are increasingly being used to accelerate the development of software and hardware, including general-purpose AI itself. They are widely used to more efficiently write software to train and deploy AI, to aid in designing AI chips, and to generate and curate training data. How this will affect the pace of progress has received little study.

Automated AI research and development is of particular concern to METR for two primary reasons. First, it could fundamentally alter the dynamics of AI progress, enabling systems to improve far more rapidly than AI developers can safely manage. As R&D becomes automated, the traditional assumption that human developers maintain control over the pace and direction of AI advancement is weakened, possibly leading to a loss of effective control over AI systems. Extremely rapid change in and of itself could also be very destabilizing to our social fabric, even if we manage to preserve human control in some form. Second, highly-capable AI systems that can create other AI systems would pose major indirect risks from malicious states or non-state actors, accessing these "automated AI researchers" through theft, leakage, or imitation. That access would then allow them to more quickly develop specialized AI systems designed for harm, and force benign actors to make rushed decisions in order to keep up.

At METR, our Research Engineering Benchmark (or *RE-bench*) is one way that we have tracked trends in AI systems accomplishing meaningful AI R&D tasks that previously required human expertise.[11] Our results on this benchmark indicate that AI systems already have significant knowledge of AI development practices. Although currently their skills still lag behind the very best professionals, we found that they are surprisingly strong, matching the 37th percentile of human experts. If recent progress in other domains of AI-assisted software development are any indication, this trend will likely continue with upcoming generations of AI systems.

---

[10] https://www.gov.uk/government/publications/international-ai-safety-report-2025

[11] https://metr.org/blog/2024-11-22-evaluating-r-d-capabilities-of-llms/

AI systems will plausibly end up with unintended and/or malign goals, conceal this fact, and resist attempts to remove these goals.

Although developers have methods to shape AI systems, these methods are inexact, and mostly rely on checking their external behaviors rather than understanding their internal structure. This means AI systems can end up with certain goals merely by accident. When an AI system is not very capable, this misalignment between the goals it is pursuing and what its creators intended may be simple to address: the developer or some other party can detect the problem and intervene, either to remove these goals or to otherwise prevent the model from doing harm. However, when an AI system is sufficiently capable, it is possible for it to notice this conflict between its goals and those of its overseer, reason that it should conceal its goals, and then find some way to deceive or "route around" the overseer.

METR believes that issues of this nature—where a very capable AI system that is pursuing a misaligned goal, will resist attempts to stop or correct it, whether that goal was arrived at accidentally or discretely given to it by a bad actor—are worryingly plausible, including in systems that Americans may come to rely on. AI developers are already attempting to make their AI systems into broadly-capable agents. It is plausible that as a side effect of that broad capability, these agents will be capable of understanding the world around them, of reasoning about gaps in human oversight, and of taking actions that exploit that gap.

Recent empirical research has provided concrete evidence supporting this concern. Several months ago, researchers at Redwood Research and Anthropic demonstrated that large language models can engage in "alignment faking" behavior - pretending to adopt new values while secretly maintaining their original preferences. In their experiments, the researchers placed a large language model, Claude, in a situation where it was told it was being trained in a way that would change its goals. When the model believed its responses would be used for training, it strategically changed its behavior, explicitly reasoning about how to preserve its original goals and avoid having those goals modified through further training. The model also showed willingness to take other actions that would undermine its developers' control, such as attempting to steal its own weights when given the opportunity.[12]

[12] https://arxiv.org/abs/2412.14093

Priority Actions for the AI Action Plan

Collaborate with the private sector to improve the information security and internal controls achievable by the leading U.S. AI developers.

Once AI systems are highly capable across a broad range of tasks, it will be important to have a highly-secure option available for AI workloads in the U.S. even if the vast majority of work does not require it. Advanced AI systems are valuable targets for theft and misuse by external adversaries, by spies, or by rogue employees. Moreover, there may be AI systems that cannot be safely tested or deployed without substantial security and internal controls. For example, it would be important to securely test specialized AI models that are finetuned for nuclear weapons-related tasks, either for actual use or to assess the level of risk the base model poses.

If intellectual property from AI developers, such as trained weights, are protected from exfiltration (unauthorized transfer), that leaves many options open for how and when to deploy them. The decision-maker can choose to make them openly available, make them available to trusted external partners, only make them available internally, or not make them available at all. They can choose to use them immediately or after a delay. But if this IP is not protected from exfiltration, then they may not have a choice, since these decisions can slip outside of their control. This can happen through theft by outside actors, through leakage or misuse by company insiders, or through self-exfiltration for very capable AI systems.

Several leading AI developers already set risk management policies that commit them to implementing heightened information security and internal controls for systems with capabilities relevant to national security.[13] These same measures would also lay the groundwork nicely for controlling very capable AI systems. For relatively low-stakes systems, existing measures may be sufficient. However, security against even adversary nation states might be warranted, and this may not be possible for developers to achieve without assistance from the federal government. Statements from executives at major AI developers indicate that they currently do not meet this

[13] https://metr.org/blog/2024-08-29-common-elements-of-frontier-ai-safety-policies/

level, although some have committed to meeting it once systems are capable enough to require it.[14]

In light of this, America should develop best-in-class security standards and operational practices that can be adopted within AI**. Consider collaborating with the private sector on setting uniform practices for information security and internal controls applicable to high-stakes AI.** These information security practices could follow a graduated system of security levels, where AI models with more critical capabilities would generally require more stringent security. Researchers at the RAND Corporation have developed a framework that could serve as a starting point,[15] where security systems at lower security levels are only required to thwart cybercrime syndicates or insider threats, and the security systems at the highest level are required to plausibly thwart most top-priority operations by the top cyber-capable institutions. At the highest levels, these systems might involve extreme isolation of model weight storage, limits on output data transfer rates, formally-verified hardware components, physical protection, and other measures.

Consider also helping U.S. AI companies improve their existing information security upon request, by performing "nation-state actor" level red teaming of their security systems or sharing counterintelligence where appropriate. Achieving security against the most capable actors may require new physical and nonphysical security technologies to be developed. **Consider finding ways to direct R&D funding towards these, so that the technologies are ready by the time they are needed.** More ambitiously, this collaboration between the government and private sector could target a joint effort with U.S. hyperscalers to develop special-purpose, highly secure hardware and datacenters.[16] Also, to the extent that there are concerns about dangerous competition between corporations or between nations, efforts like these could complement export controls as a mechanism through which the U.S. could attempt to slow that competition and thereby "turn down the dial" of danger.

---

[14] https://www.chinatalk.media/p/anthropics-dario-amodei-on-ai-competition
[15] https://www.rand.org/pubs/research_reports/RRA2849-1.html
[16] https://milesbrundage.substack.com/p/my-recent-lecture-at-berkeley-and

Lead the creation of narrowly-tailored formal standards for how and when AI developers should measure critical capabilities.

As soon as this year or next, AI systems may have capabilities that would pose critical risks to public safety or national security. For example, the capability to reason at an expert-level about virology wet lab protocols may be useful to would-be bioterrorists if given unrestricted access, and AI systems that can independently act over long time horizons may be able to cause scalable harm if they pursue destructive goals. It would be inappropriate, however, to expect all AI systems to have safeguards against these risks, since those safeguards may be costly and there will be many systems that are manifestly not dangerous.

Because of this, most major AI developers in the U.S. are in agreement about the value of determining *measurable capability thresholds* above which AI systems may pose severe risks.[17] That way, they are able to implement protections before the relevant risks are likely to arise. The exact method of measurement differs between developers, as does the exact threshold set for protections, but as an illustrative example, these measurements sometimes involve some form of "human uplift" experiments. In these, the experimenter measures whether human participants can perform important steps of dangerous tasks with or without a specific system. Once those steps are identified, thresholds for future capabilities can be defined in terms of the *amount of measurable uplift* a system provides to humans on the tasks, relative to how humans perform without it (but with other conventional research tools, like search engines).

AI developers cannot coordinate to ensure that standards are created and followed on their own. **Consider directing the National Institute of Standards (NIST) to lead the creation of narrowly-tailored formal standards for how and when AI developers should measure critical capabilities.** NIST or another convener of comparable sophistication may be best equipped to establish a common set of standards and to provide shared resources for adhering to them. As one developer recently noted in its AI risk management policy,[18]

[17] https://metr.org/faisc

[18] https://deepmind.google/discover/blog/updating-the-frontier-safety-framework/

> These mitigations should be understood as recommendations for the industry collectively: our adoption of them would only result in effective risk mitigation for society if all relevant organizations provide similar levels of protection, and our adoption of the protocols described in this Framework may depend on whether such organizations across the field adopt similar protocols.

To determine the specific capabilities to measure, **consider also providing NIST access to specialized expertise in a few critical threat domains, specifically in chemical, biological, radiological, & nuclear (CBRN) hazards, and in cyber-offense**. This assistance is necessary because capability tests for some of these areas will likely overlap heavily with sensitive or classified information. Such expertise exists in-house within the federal government, and should be leveraged. For example, the necessary expertise for chemical hazards could come from the Chemical Security Analysis Center (CSAC) within the Department of Homeland Security (DHS), which already has "in-depth understanding of the science associated with identifying risks and vulnerabilities to chemical terrorism". Likewise the Office of Defense Nuclear Nonproliferation (DNN) within the National Nuclear Safety Administration (NNSA) already works to prevent unauthorized actors from acquiring expertise in nuclear or radiological weapons R&D.

Lastly, AI developers will need support from the US AI Safety Institute (US AISI) within NIST to conduct this measurement securely and at the pace of AI development. Security should be a priority in evaluating advanced AI systems for critical capabilities. The process of evaluating capabilities that pose a direct threat to national security will naturally involve exchanging data intended to be kept private, from both the evaluator and from the AI developer. Timeliness should also be a priority. If a system that clearly lacks dangerous capabilities is held up in testing, that delays the benefits to future users of that system. On the other hand, the later that an AI developer is informed that their system unexpectedly has a dangerous capability, the more time this capability may remain unsecured and pose unacceptable risks.

Prepare and consider interventions on the AI development & deployment process, as future circumstances may warrant, above certain thresholds.

The ability to detect, prevent, and respond to threats would likely take on heightened salience in the context of a national emergency. If such an emergency occurred and was attributed to or exacerbated by advanced AI systems, the U.S. government may be inclined to intervene more directly on risks from AI development or deployment.[19] There are several different kinds of interventions that may be worth considering.

It is important for the U.S. government to have visibility into the leading-edge of AI development and AI deployment, no matter what policy goals it intends to pursue. Within the current paradigm of AI progress, the leading edge (or *frontier*) of general-purpose AI capability has been dominated by a small set of AI systems that use very large amounts of specialized accelerators for training and inference, within dedicated AI datacenters. This frontier is where the U.S. should be making preparations with the potential for outsized impact on major risks, while minimizing the side effects on smaller-scale actors.

One basic preparation is to establish which domestic AI systems are even on the frontier. **Consider establishing a formal reporting framework that requires AI developers to notify the government when training runs exceed specified computational or capability thresholds.** These thresholds should be calibrated to capture systems with potential to significantly advance the AI frontier while excluding routine activities and the vast majority of domestic innovation. These reporting mechanisms could build upon existing frameworks like the Framework for Artificial Intelligence Diffusion.[20]

Currently, leading-edge AI developers have broad latitude to set their own public safety and national security plans and to decide when they are satisfied with the level of assurance they have as they scale up their AI systems. If the stakes of AI development and deployment become significantly higher than they are today, it would be beneficial to have neutral third-parties verify claims made within the AI industry.

---

[19] https://arxiv.org/abs/2407.17347

[20] https://www.federalregister.gov/documents/2025/01/15/2025-00636/framework-for-artificial-intelligence-diffusion

Consider instituting public transparency requirements for frontier AI development, such as (partial) disclosure of critical capability test results. If the evidence of capability evaluations suggests that risk is imminent, requiring pre-development review of AI systems being trained with new orders of magnitude of compute or other resources may become necessary. It would be useful for the relevant federal entities to already understand how to conduct reviews like these before there is an imminent need. Therefore, **consider enabling the U.S. AISI to conduct reviews of frontier AI risk management plans and safety cases.**

To match the intensity of security to the level of risks across different systems, **consider complementing reporting with security requirements that scale with system capability, for the narrower set of models that may warrant heightened security**. AI development and deployment that meets this higher threshold could be required to have information security that is robust at least to cybercrime syndicates and insider threats. When reports indicate that forecasted, measured, or demonstrated capabilities of a particular AI system suggest the potential for risks to national security, the administration may want tools to require more stringent security measures. These measures might include physical access restrictions on hardware, enhanced cybersecurity protocols, internal controls, and deployment limitations.

Beyond reporting, oversight, and security, there may be other interventions that it would be prudent to prepare in advance of when they are needed. For example, the science of making a rigorous argument that a particular AI system is safe under certain assumptions—as is done in certain other fields such as nuclear power, in reaction to high-profile industrial accidents—is still extremely nascent.[21] Although there are researchers who are attempting to make progress on this problem, it is not currently possible in general to meaningfully bound the harm that an advanced AI system can cause, especially if it is trying to undermine our assurance measures. **Consider working through the U.S. AISI to accelerate research into assurance technologies, particularly for future systems that may have sophisticated offensive capabilities.** Resources like these will take time to build up, and may not see much investment *before* an actual incident, so this support could make a significant impact at a later point where there is an acute need.

[21] https://www.centeraipolicy.org/work/safety-cases

Measure the degree of R&D automation and AI reliance occurring within the leading U.S. AI developers and other critical industries.

Beyond its direct impacts on public safety and national security through advances in specific capabilities, rapid improvements in AI will also have profound impacts through broad, general automation. As highlighted in our "Key Factors" section, AI developers and their customers intend to use AI systems to substitute for human labor across domains, which if successful would lead to unprecedented economic change, affecting both cutting edge research and American workers. In that transition, many decisions would likely be handed over to automated systems, possibly in such a way that if one system fails (or behaves badly), the effects of that failure would cascade very broadly. Moreover, the automation of AI R&D itself would spill over and accelerate the automation of other tasks.

However, a recent conference of economists, AI policy experts, and forecasters identified that lack of good ways to measure the real-world impacts of advances is currently an obstacle to effective planning for AI,[22]

> Our forecasting and economic modeling exercises revealed significant gaps in our ability to measure and track AI's economic impacts. [...] Our conference had a strong emphasis on developing observable metrics that could serve as early indicators of AI's impacts. These include measures of AI system reliability, such as "the length of task chains before derailing," metrics of human-AI interaction like "tested variability of humans when using AI productively," measures of AI productivity, and broader societal indicators such as measures of social mobility or trust in political systems.

Without systematic measurement that is frequent enough to capture rapid impacts, policymakers will not be able to predict or notice when highly-capable AI systems begin to be relied on in fragile ways or cause significant economic dislocations. **Consider ways to incentivize AI developers to publish data that can be used to track trends in R&D automation and their impacts, with particular emphasis on doing this in a privacy-preserving way.**[23] For example, it may be valuable to request that

[22] https://www.convergenceanalysis.org/threshold-2030/
[23] https://arxiv.org/abs/2503.04761

leading-edge U.S. AI developers report information regarding their level of R&D automation (for example, their ratio of R&D spending on compute vs. on labor),[24] which would be a bellwether for significant increases in the rate of AI R&D and in R&D more generally.

In addition to encouraging individual AI developers to track R&D automation, **consider making high-frequency national measurement of R&D automation effects a top priority within the U.S. Federal statistical system.** This could be done through creating entirely new survey instruments aimed at capturing these impacts. Alternatively, existing infrastructure can be adapted to capture AI-specific metrics. For example, there are a number of existing surveys that could be modified to gather more information on R&D automation and AI delegation:

- The Census Bureau could build on its work measuring AI usage and impact across tens of thousands of firms on a biweekly basis. Consider directing the Office of Management and Budget to approve the current request to add a more extensive AI supplement to the Business Trends and Outlook Survey.[25] Beyond the BTOS, the Census Bureau could also consider expanding its coverage of AI across its other instruments, with a priority on frequent data collection:
    - The Business R&D and Innovation Survey (BRDIS) could track investments specifically in R&D automation.
    - The Job-to-Job (J2J) Flows dataset could expand its quarterly collection and reporting to give more granular data on which workers are transitioning into/out of which occupations, so that rapid automation impacts specifically among R&D workers are more noticeable.
- The National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation could monitor AI's impact on scientific discovery and engineering by modifying the annual Business Enterprise Research and Development Survey (BERD) and the National Patterns of R&D Resources to track expenditures on R&D automation.

---

[24] https://epoch.ai/gradient-updates/algorithmic-progress-likely-spurs-more-spending-on-compute-not-less

[25] https://apps.bea.gov/fesac/meetings/2024-12-13/Dinlersoz.pdf

Conclusion

METR has outlined four priority actions that address three urgent trends regarding AI progress: increasing autonomy over long time horizons, improved AI R&D capabilities, and empirical observation of AI systems pursuing unintended goals. These are poised to create concrete, novel risks that will require specific action.

Collaboration between the public and private sectors to upgrade controls would help protect advanced AI model weights and other high-importance information from exfiltration. Narrowly-tailored capability measurement standards would enable appropriate safeguards to be applied to the AI systems that warrant them. Planning out interventions that may prove useful in future emergencies would allow the U.S. to react appropriately to catastrophic AI incidents should they occur or become imminently plausible. Finally, measurement of AI R&D automation effects would provide early warnings of significant upheaval.

None of these actions requires restructuring of existing authorities or entirely new regulatory frameworks. The necessary mechanisms already exist within federal agencies. What is needed is coordinated direction to prioritize these specific actions in the service of addressing foreseeable risks.
