MITRE

The AI Mission Testing Gap

Securing the use of frontier AI models for economic and national security advantage

Charles Clancy, Ph.D., Douglas Robbins, Mikel Rodriguez, Ph.D., and Katie Enos

The AI Mission Testing Gap

Securing the use of frontier AI models for economic and national security advantage

CORE ARGUMENT

Maintaining leadership in the frontier AI race requires the United States to put its most advanced AI capabilities to work in the scientific, technological, and security missions where they create strategic advantage. **Doing so requires evaluation that reduces uncertainty and provides the evidence needed to make informed investment, adoption, and policy decisions about how those capabilities can be used effectively.**

From AI Leadership to Strategic Advantage

The United States and China are in a technoeconomic race to establish global leadership in the development, deployment, and strategic use of artificial intelligence.

The two countries are pursuing that leadership through increasingly distinct approaches: The U.S. ecosystem is driven by closed, proprietary models, while China is close behind with open-weight models. Benchmarks generally put the United States consistently months ahead of China, but the capability gap continues to narrow with Chinese national-level initiatives such as "Eastern Data, Western Computing" and "Digital China." These aim to follow the same success as "Made in China 2025," which used comprehensive mobilization of state resources, private enterprise, and national priorities to meet or exceed the ambitious goal to make China a global manufacturing powerhouse.

In the United States, the government and private sector are pursuing efforts to preserve and widen the U.S. lead in AI capabilities, while constraining the pace of China's progress. These efforts include restricting China's access to advanced graphics processing units (GPUs), strengthening cybersecurity and counterintelligence protections, and limiting China's ability to distill capabilities from leading U.S. models. Similarly, both the public and private sector are working to accelerate the pace of U.S. capability, from large R&D investments to national initiatives on energy and data centers.

U.S. leadership in AI will provide a strategic advantage **only if the United States can translate its technological lead into meaningful economic and national security outcomes**. Maintaining this lead requires deploying AI at scale across strategically important scientific, technological, and security

domains, particularly those at the center of technoeconomic and military competition. The latest frontier models can provide authorized users with capabilities beyond those available in widely released versions, creating the potential for meaningful mission advantage in high-consequence environments. These users should include authorized U.S. researchers and government agencies that are chartered to maintain our U.S. leadership. For example, biotechnology researchers cannot fully harness AI to accelerate scientific discovery if model safeguards unnecessarily restrict legitimate research involving biosecurity-relevant topics. Similarly, cybersecurity experts cannot fully understand the threat new models pose to our critical infrastructure without access and appropriate evaluation ranges.

Realizing this AI advantage requires a testing paradigm that can generate the evidence needed to determine where these models improve mission performance and establish the conditions for their effective use. Evaluation must account for a model's capabilities and limitations, as well as access controls, user authorization, deployment environments, monitoring mechanisms, and other safeguards. The resulting evidence should inform who receives access, where and how models can provide mission value, what safeguards and operating limits are appropriate, and how performance should be evaluated over time. This requires a framework spanning access control, mission evaluation, behavioral alignment, and system and lifecycle testing. Further, to determine the U.S. comparative advantage, testing should generate insight into the U.S. lead in science, technology, and national security missions fueled by U.S. industry compared to the latest open-weight models, particularly those developed by international competitors.

Access Control

Frontier models require careful access control. Adversaries who gain access can use these systems directly or distill their capabilities into competing models. Overly restrictive access policies, however, can prevent innovations and scientific and technological breakthroughs that maintain our strategic advantage. Access should therefore use Know Your Customer (KYC)-style controls, including organizational and individual vetting, cybersecurity requirements, identity management, authentication, and authorization. Existing personnel and cybersecurity controls, including the National Industrial Security Program Operating Manual (NISPOM), National Institute of Standards and Technology Special Publication 800-171 (NIST SP 800-171), NIST Cybersecurity Framework, and National Security Presidential Memorandum 33 (NSPM-33), may provide building blocks for extending trusted access beyond traditional government users and contractors.

Controls should:

- **Create a trusted-access framework:** Access should allow qualified U.S. users to employ advanced models while providing reasonable assurance that those capabilities will not be transferred to unauthorized actors.
- **Be commensurate with use:** The sensitivity of the model, the risk of the intended use, and the security posture of the organization receiving access should inform controls.

Capability and Mission Evaluation

Frontier models should be evaluated against the best publicly available alternatives to determine whether their advanced capabilities produce meaningful improvements in federal agency mission performance.

After all, a frontier model is only strategically valuable if its additional capabilities catalyze better scientific, technological, economic, health, or security outcomes. Evaluation should compare the latest models with publicly available closed and open-weight models in realistic mission environments and against outcomes that matter to the agencies using them. For especially consequential capabilities, evaluations may also need to use controlled or sequestered test materials that are not publicly available, reducing the risk that models have encountered the test during training or have been optimized specifically to perform well against it.

Where evaluation demonstrates a meaningful mission advantage, that finding should inform access, deployment, and protection decisions. Where less sensitive or more widely available models perform comparably, agencies may have little reason to assume the additional cost and security requirements associated with frontier systems.

Evaluation can also accelerate adoption. By providing evidence about where advanced models deliver meaningful mission advantage, where their limitations lie, and the conditions under which they can be used effectively, testing gives agencies and model developers greater confidence to expand access and deployment. Clear operating limits can enable faster decisions about where and how advanced capabilities can be put to work.

Evaluation should:

- **Measure mission impact:** Determine whether the model produces meaningful improvements in mission performance and identify the economic, scientific, or security value of those improvements.
- **Identify comparative advantage:** Establish where specialized frontier models materially outperform publicly available alternatives and where they do not.
- **Define effective operating conditions:** Characterize model limitations and determine the tasks, users, environments, and workflows in which the model provides the greatest value.
- **Improve future capabilities:** Provide structured feedback to model developers and users that can inform training, evaluation, deployment, and subsequent model development.

Matching Safeguards to Mission and User

The use of advanced AI for legitimate but potentially sensitive missions requires a different approach to determining appropriate model behavior.

Protections designed for publicly available AI systems may be unnecessarily restrictive for authorized users conducting legitimate work in sensitive fields. At the same time, providing greater access to model capabilities should not mean removing safeguards altogether. Models should operate consistently with the laws, policies, security requirements, and professional standards that govern the institutions using

them. Effective evaluation of frontier models allows potential users to understand where, when, and how a model will be highly valuable in streamlining tasks or accelerating workflows.

The appropriate safeguards will therefore vary by mission and user—as well as by the environment in which the model is employed. AI models operate within larger ecosystems, engaging humans, novel datasets, memory limitations, power limitations, and unique security controls. A biotechnology researcher, cybersecurity analyst, intelligence professional, and general public user may require access to different capabilities and may warrant different safeguards based on their missions and operating environments. Rather than applying the same restrictions to every user, specialized models can be configured for specific communities and operating environments, allowing greater capability where it is warranted while retaining appropriate protections.

Greater specialization can make safeguards more relevant to a particular mission and help limit inappropriate use, but maintaining many different versions of a model creates additional technical, security, and administrative complexity. Testing should provide the evidence needed to accelerate decisions about how much specialization is warranted, which safeguards are appropriate for particular users and missions, and whether those safeguards enable legitimate use while effectively constraining unauthorized activity.

Mission matching should:

- **Determine whether safeguards are appropriate:** Assess whether they prevent inappropriate uses without unnecessarily interfering with legitimate work.
- **Evaluate misuse resistance and monitoring:** Assess whether users can circumvent established restrictions and whether systems can detect and respond to attempted misuse.
- **Assess compliance with requirements:** Determine whether models follow applicable security, handling, classification, and other domain-specific requirements.
- **Refine safeguards over time:** Provide feedback to model developers and mission owners to optimize operating policies.

System and Lifecycle Testing

Testing a frontier model before deployment is necessary but is only a snapshot in time. Models increasingly operate as part of larger systems, interacting with other AI models, software tools, data, and human users over extended periods that continuously evolve.

These interactions can create behaviors and risks that do not appear in isolated testing. Testing must also, therefore, continue throughout deployment and operation, using realistic environments and changing conditions to identify shifts in performance or behavior and determine when safeguards or operating limits need to be adjusted.

These environments can also be used to test the broader systems designed to oversee AI use. Organizations can evaluate monitoring and misuse-detection tools, conduct exercises to test responses to AI-related incidents, and assess whether technical and organizational safeguards work together effectively. As new AI monitoring and threat-mitigation technologies become available, realistic test environments can provide

a common basis for determining whether those tools are effective before they are relied upon in high-consequence missions.

Pre-deployment testing is particularly important for agentic AI systems, where vulnerabilities may arise through interactions with tools, data, memory, credentials, and other software components. Testing should examine how models and agents behave under realistic adversarial conditions, including attempts to manipulate inputs, misuse authorized tools, bypass controls, or exploit weaknesses across system layers. Pre-deployment is where measuring the ability and propensity of models to exploit vulnerabilities in sandboxes and testing harnesses would occur.

System testing should:

- **Go beyond measures of model knowledge and capability:** It should include controlled benchmarks, realistic emulation, and representative environments with the security controls and operational friction the system will encounter in use.
- **Be treated as an ongoing process:** Especially in high-consequence applications, a one-time evaluation or approval will not be sufficient, and authorization should instead require test before use, control during use, and continuous verification that the system continues to perform within established parameters.

Test before use	Use with control	Verify in operation
Compare capabilities, test safeguards, define authorized users, and establish operating limits.	Integrate models into mission workflows, train users, enforce access controls, and monitor use.	Measure mission impact, identify changes in performance or behavior, exercise incident response, and adjust controls as needed.

Building a Testing Ecosystem for Specialized Frontier AI

Many frontier models will be deployed in high-consequence scientific, technological, national security, and critical infrastructure environments, where both the potential benefits and the consequences of failure or misuse are significant.

Effective testing must therefore determine not only what a model is capable of, but whether it can be used effectively and responsibly for its intended mission, by authorized users, and under the conditions in which it will operate. Doing so requires an uncommon combination of AI technical expertise, deep knowledge of the mission and operating environment, cybersecurity and risk-management capabilities, systems engineering, and sophisticated test infrastructure.

As advanced AI becomes embedded in high-consequence missions, independent testing and evaluation become part of the national infrastructure required to expedite employment of advanced AI at scale. Trusted evaluation capabilities provide a repeatable means of comparing systems, informing deployment decisions, identifying weaknesses, and verifying that safeguards continue to work as capabilities evolve. Testing accelerates safer AI model adoption because it reduces uncertainty and risk.

Ultimately, federal agencies will need to determine whether and under what conditions frontier AI capabilities should be used to advance U.S. government missions or made available to authorized users supporting broader national objectives. These decisions include whether a specialized model provides sufficient advantage over available alternatives, who should be authorized to access it, what safeguards and operating limits are required, and what level of risk is acceptable. Testing should provide the evidence needed to support those decisions by assessing the additional capability gained, operational limitations, safeguard effectiveness, and changes in performance as models are integrated into larger systems and workflows.

An effective testing ecosystem should:

- Depend less on any single benchmark or evaluation tool: Ecosystems should have the ability to integrate and adapt the best available methods, build mission-specific tests, protect sensitive test materials, instrument realistic environments, and validate results.
- Depend on trusted arrangements: Ecosystems should provide evaluators sufficient and timely access to advanced models, system information, and sensitive capabilities to conduct meaningful testing while protecting developers' proprietary information.
- Use common evaluation frameworks and reusable testing methods: Ecosystems should be able to create consistency across government and reduce duplicative investment. They should have a shared vocabulary for capabilities, threats, and failure modes that allows agencies to tailor evaluations to their missions while still comparing results, sharing lessons, and incorporating findings from other domains.

Realizing the U.S. AI Advantage

U.S. leadership in frontier AI will translate into strategic advantage when advanced capabilities can be effectively deployed in the scientific, technological, economic, and security domains that matter most.

New models can extend those capabilities to authorized users working in high-consequence environments, provided testing establishes the evidence, safeguards, and operating parameters necessary for their use.

A robust testing regime can accelerate the transition from frontier capability to real-world impact. It can identify where advanced models provide meaningful advantage, support decisions about access and deployment, establish appropriate operating limits, and verify performance as models are integrated into consequential missions. What is needed is a boundary around evaluation. The evaluator provides the evidence. The evaluator does not make the ultimate policy choices or risk-acceptance decisions, certify a model as universally safe, or replace the accountability of mission owners, users, developers, or government officials.

If the global AI capability gap continues to narrow, the United States' ability to effectively deploy its most advanced AI in consequential missions may become the critical source of strategic advantage.

Effective testing *should*	**Effective testing *should not***
Develop tests and evaluation infrastructure tailored to specific applications and operating environments.	Set government-wide AI policy or regulate frontier model developers.
Evaluate performance, limitations, safeguards, and risks associated with use.	Certify a model as universally safe.
Protect proprietary, classified, and other sensitive information.	Require public disclosure of sensitive model details.
Provide evidence to support access, deployment, and operating decisions.	Determine the level of risk an agency or other responsible organization must accept.
Support reusable evaluation methods and continued verification.	Replace mission-owner, user, developer, or government accountability.
Be conducted by conflict-free parties, with access to cleared test environments, mission understanding, and ability to steward frameworks.	Be left up to developers or mission owners who don't have the infrastructure the testing function requires.

About MITRE

Through its public-private partnerships and federally funded research and development centers (FFRDCs), MITRE works across government and in partnership with industry to tackle mission-driven challenges to the safety, stability, and well-being of our nation.

MITRE's perspective on frontier AI testing is informed by its work at the intersection of artificial intelligence, systems engineering, cybersecurity, and government missions. As the operator of multiple FFRDCs, MITRE works across national security, science and technology, health, transportation, critical infrastructure, and other federal missions in which emerging technologies must be evaluated in the context of complex operational systems, mission objectives, and the broader public interest.

That perspective is particularly relevant to the testing gap described in this paper. MITRE's AI work spans model and system evaluation, AI assurance and red teaming, autonomy, cybersecurity, human-machine teaming, and the integration of AI into mission environments. Its multidisciplinary AI workforce combines more than 600 AI researchers and engineers with domain experts and systems engineers, allowing questions about model capability to be considered alongside questions about mission performance, operational constraints, security, and system behavior.

MITRE also has experience developing approaches for evaluating technologies as they evolve. Its work includes secure environments for AI experimentation and evaluation, adversarial testing of AI systems, and the development and stewardship of frameworks such as MITRE ATT&CK® and MITRE ATLAS™, which organize changing threats and techniques into shared structures that can support analysis and testing. ATT&CK, originally developed at MITRE to characterize real-world adversary behavior, is used across government and industry worldwide. Its evolution illustrates how a common technical framework can provide a shared foundation for understanding threats, evaluating defenses, and adapting as technologies and adversaries change.

These experiences inform the central perspective of this paper: Evaluating frontier AI for consequential government missions requires looking beyond model performance in isolation. It requires understanding how AI capabilities interact with users, data, safeguards, software, infrastructure, adversaries, and mission requirements over time. The framework presented here reflects MITRE's experience approaching emerging technology through that broader mission and systems lens. Finally, MITRE's independence, objective technical role, and lack of a commercial interest in any particular AI model or platform position it to serve as a trusted contributor to an ecosystem that requires credible evidence across government, industry, and mission communities.

Authors

Charles Clancy, Ph.D., MITRE's Senior Vice President of Technology and Engineering and Chief Technology Officer, where he leads advanced research and technology strategy for the company. A computer scientist and engineer, he previously held leadership roles at Virginia Tech and the National Security Agency and is recognized for his contributions to wireless communications, security, and spectrum innovation.

Douglas Robbins, Vice President of MITRE's Center for AI, Cyber, and Digital Technologies, where he leads research and prototyping to advance emerging technologies for government sponsors. His work spans autonomous cyber operations, frontier AI model evaluation, AI-enabled missions, and acquisition transformation. He also helps connect MITRE with New England's business, government, and innovation communities.

Mikel Rodriguez, Ph.D., a MITRE Fellow and internationally recognized leader in artificial intelligence with more than two decades of experience developing AI solutions for complex challenges. Previously at Google DeepMind, he co-founded a team focused on the secure deployment of frontier AI systems and contributed to Project Gemini. He previously led MITRE's AI and Autonomy Innovation Center.

Katie Enos, Director of External Affairs at MITRE and a seasoned government affairs professional with nearly two decades of experience on Capitol Hill, including more than ten years as chief of staff to a member of the U.S. House of Representatives.

Acknowledgements

The authors would like to thank Wendy Ginsberg, Government Relations, Principal; Jessica Rajkowski, Artificial Intelligence, Director; and Chris Glasz, Principal AI Engineer, Model Threat Evaluation Group Leader for their thoughtful input and review of this document.

MITRE