DARPA-NSF-CAISI

AI Forge

Critical AI Challenges for National Security Report

DISTRIBUTION STATEMENT 'A' (APPROVED FOR PUBLIC RELEASE, DISTRIBUTION UNLIMITED)

DARPA NSF NIST CENTER FOR AI STANDARDS AND INNOVATION (CAISI)

AI Forge: A National Partnership for AI Innovation

This document is the primary guide for university-led research teams seeking to contribute to the AI Forge program. It defines the program's *Critical AI Challenges for National Security*, enabling teams to develop well-aligned ideas, research proposals, and pitches for AI Forge Project Ventures – a series of fast-paced research projects aimed at solving these challenges.

As artificial intelligence (AI) systems integrate into warfighting functions, a strategic gap has emerged between commercial AI development and the unique demands of the mission. National security applications require AI systems capable of operating in situations with life-or-death consequences, executing actions at a pace that exceeds traditional human oversight, protecting individual and organizational privacy boundaries, integrating securely with legacy infrastructure, and meeting rigorous certification standards. Because the challenges arising from this gap lack immediate commercial payoff, they are not the primary focus of private industry, leaving them underfunded and under-explored.

The AI Forge program aims to fill this strategic gap by establishing a forum comprised of universities, frontier AI companies, and U.S. government defense and security representatives to create breakthroughs in AI for national security. As a joint effort between the Defense Advanced Research Projects Agency (DARPA), the U.S. National Science Foundation (NSF), and the Center for AI Standards and Innovation (CAISI) at the National Institute of Standards and Technology, AI Forge will help inform the research agenda around priority challenges for the Department of War (DOW), the Intelligence Community, and the U.S. research enterprise. By providing university researchers with access to frontier-scale compute and models, AI Forge seeks to accelerate progress, build a durable research ecosystem around these critical challenges, and enable a robust exchange of talent and ideas across universities, frontier AI companies, and government.

The program will help advance research in three areas identified as priorities in America's AI Action Plan[1]: AI interpretability, AI control, and adversarial robustness. Rather than duplicating commercial efforts, AI Forge will complement them by focusing on foundational, forward-looking research unique to national security needs. The strategic aim is to de-risk these critical challenges and create viable pathways for frontier AI companies to invest in these research directions, ensuring that American AI dominance is driven by systems that are as secure and reliable as they are capable.

[1] "Launch a technology development program led by DARPA in collaboration with CAISI at DOC and NSF, to advance AI interpretability, AI control systems, and adversarial robustness." (America's AI Action Plan, p.9)

Prioritizing a dedicated focus on security and reliability, in turn, helps steer the broader AI community toward developing fundamental capabilities that are also vital for national security.

With guidance from DARPA, NSF, and CAISI, the research challenges outlined herein synthesize ideas from multiple key sources. They draw from the January 2026 AI Forge Workshop, which convened eight of the nation's leading frontier AI companies, chief AI officers from over fifteen DOW and Intelligence Community agencies, and other government stakeholders in reaching consensus on these challenges. Our roadmap also builds on guidance in the AI Action Plan and conversations with its authors. To reflect the fast-changing technical AI research landscape, this document will be updated every six months during the program to continue guiding the groundbreaking work of the AI Forge performer base.

To the university research community: This publication is more than a report; it is an invitation to innovate. Solving these challenges will require the community's most creative ideas, rigorous methods, and boldest thinking. The innovations born out of AI Forge will not only represent breakthroughs in artificial intelligence; they will help define how the U.S. builds, deploys, and trusts these powerful systems for decades to come. By mobilizing the nation's top academic talent, this initiative aims to forge that future.

Research Thrust 1. AI Interpretability: Enabling Actionable Understanding

Overview

AI interpretability is the science of making the behavior, decisions, and impacts of AI systems understandable to humans. The key objective for this research thrust is to move beyond explanations in routine settings toward **operational interpretability**. Achieving this objective requires developing scalable and verifiable methods that are tailored to specific user roles and address both a model's internal mechanisms and the emergent behaviors of "black-box" systems. Crucially, the effectiveness of these methods must be measured through rigorous, mission-relevant benchmarks. The aim is to equip personnel with the tools to understand model reasoning and uncertainty, anticipate failure modes, and enable effective human oversight and informed decision-making.

Research Challenges

These challenges aim to establish operational interpretability as a rigorous discipline, beginning with foundational methods to identify the causes of model behavior from both internal and external perspectives (Challenges 1 and 2). They then extend these methods to operate at scale and for complex agentic systems (Challenges 3 and 4), ultimately pioneering new evaluation paradigms for scientific discovery (Challenge 5).

Challenge 1: Scaling Causal Interpretability for Predictive Diagnostics

Driving Question: How can we scale and operationalize recent advances in causal interpretability – transitioning from academic proofs-of-concept to robust, testable mechanisms that allow operators to accurately predict, explain, and audit model behavior in mission-critical settings?

Background: While many current interpretability methods provide only correlational insight ("this feature mattered"), promising academic research has begun to establish the foundations of causal interpretability. National security settings, however, demand stronger, testable causal claims: *which* internal mechanisms cause a decision, *how* those mechanisms will behave under stress, and *how* their underlying logic shifts dynamically during operation. Without this causal grounding, interpretability remains fragile, offering limited leverage for accountability or debugging when a system deviates from its expected performance. This challenge aims to **establish a broadly applicable causal science of model behavior**. Primarily, it aims to provide the foundation for auditable AI by giving operators the precise diagnostic visibility required for effective oversight. As a secondary, high-impact application, these same causal tools could be used to audit models trained on scientific data, allowing researchers to extract novel concepts, verify

causal pathways, and potentially reveal insights that transcend current human understanding.

Challenge 2: Black-Box Analysis of Long-Horizon Failures for Verifiable Accountability

Driving Question: When an AI system fails only after a long sequence of interactions, how can we isolate the root causes and generate empirical, testable attributions without access to model internals?
Background: Some of the most consequential AI failures are not isolated mistakes, but emergent behaviors that arise over many steps, interactions, and environmental changes. The root cause may stem from a subtle earlier interaction, an external input, or a gradual drift in behavior that is difficult to trace after the fact. This challenge seeks to advance black-box interpretability beyond today's short-horizon behavioral tests and counterfactual methods toward approaches that can **uncover, explain, and document long-horizon failures** in deployed systems. The goal is to develop tools that continuously monitor behavior, detect policy-relevant failures, and maintain verifiable records of system interactions so operators can perform credible error analysis, assign accountability, and prevent similar failures in the future.

Challenge 3: Automated Interpretability at Scale

Driving Question: How do we build automated systems that generate verifiable, tailored explanations across diverse modalities, and that improve as models and missions evolve without introducing new failure modes?
Background: Interpretability methods do not currently scale. As models grow and deployments diversify, the burden of generating explanations cannot rest on bespoke, expert-driven analysis. At the same time, simply automating explanation generation risks creating fluent but unfaithful narratives, leading to a dangerous sense of over-trust. This challenge addresses the need for interpretability as an operational capability. The objective is to create systems that can automatically **translate raw model behavior into task-relevant explanations** for distinct user roles and extend across text, vision, audio, and other sensor streams. Critically, this automation must be coupled with continuous verification, ensuring that "helpful explanations" do not become a vector for deception or an independent source of error.

Challenge 4: Agentic AI Interpretability for Auditable Autonomy

Driving Question: How do we make agentic AI systems interpretable at the system level – including goal formation, planning, tool use, and memory updates – so that operators can fully trace, audit, and understand their complex decision-making pathways?
Background: Agentic AI systems shift the interpretability problem from "Why did the model output this token?" to "Why did the system take this sequence of actions?"

Agents operate in loops of goal-setting, planning, tool use, and memory updates, introducing risks such as flawed reasoning chains, brittle plans, unsafe tool invocation, and emergent behaviors. This challenge focuses on building methods to **trace and understand agentic behavior**. When a monitor detects a deviation from intended behavior, interpretability tools must generate testable hypotheses about its origin. Research could involve creating behavioral "fingerprints" to identify common failure modes or attribute the anomalous behavior to the influence of a specific training process or component. The goal is to develop techniques that can reconstruct an agent's internal state, visualize its planning horizon, and provide operators with the system-level visibility needed to audit complex workflows. This research will contribute to high-stakes autonomy that is transparent and auditable by establishing clear methods to map an agent's logic and understand *why* a specific sequence of actions was chosen.

Challenge 5: Evaluation of AI Systems for Scientific Discovery

Driving Question: How can we evaluate AI systems for scientific discovery even when they surpass human expertise, and verify their results are correct, reproducible, and trustworthy?

Background: This challenge seeks to advance AI systems that can plan and carry out complex *in silico* research to accelerate scientific discovery. As general and domain-specific scientific foundation models begin to surpass human subject matter expertise, evaluating their outputs becomes exceptionally difficult – especially when tasks lack clear or immediate ground truth. Standard benchmarks are often too narrow for open-ended scientific work, where success may depend on judgment, delayed validation, or contested evidence across multimodal data. This challenge focuses on **creating new paradigms for scientific evaluation** that can assess both verifiable outputs (code, proofs, experiments) and subjective reasoning (explanations, causal accounts, research judgments). New evaluation methods are needed for tasks such as hypothesis generation from incomplete information, experiment design or algorithm development at scale, and scientific explanation in domains such as chemistry, biology, materials science, space science, and artificial intelligence. Progress depends on developing better methods for structured expert input, scalable oversight, AI-assisted evaluation, and automated judges, along with ways to test whether those evaluators are themselves reliable. The goal is to produce traceable evidence packages from these systems: decision artifacts that identify errors in AI-generated science, combine uncertain human and AI judgements into credible assessments, and ultimately make AI-driven scientific results auditable, reproducible, and trustworthy enough for others to build upon.

Research Thrust 2. AI Control: Ensuring Reliable Performance

Overview

AI control refers to techniques for building AI systems that reliably execute human intent. It provides the technical means to specify goals, prevent unintended behavior, enable rapid correction, and scale human oversight as AI capabilities increase. As AI systems grow in capability and autonomy, ensuring they remain controllable becomes a critical technical and national security challenge. Current control mechanisms do not scale effectively with model complexity or deployment diversity, creating an urgent need for new approaches. This problem is compounded as models increasingly rely on external, real-time data sources they were not trained on, creating unpredictable behaviors when faced with conflicting or unexpected information. The key objective of this research thrust is to build and validate the components of a holistic control architecture for AI. This research will seek to pioneer tools that can provide **strong, verifiable evidence** of bounded, auditable, and reliable model behavior today, while laying the essential groundwork for maintaining meaningful human control over future, more capable AI systems.

Research Challenges

This research thrust confronts AI control across its full lifecycle: beginning with embedding verifiable steerability directly into foundation models (Challenge 1) and establishing auditable provenance across the deployment supply chain (Challenge 2), progressing to active management through preventative containment and real-time operational control (Challenges 3 and 4), and culminating in predictive evaluation to provide the evidence needed for mission assurance (Challenge 5).

Challenge 1: Verifiable Steerability of AI Models

Driving Question: How do we embed scalable, resilient control mechanisms directly within AI models, ensuring they provide verifiable evidence of their own uncertainty and defer appropriately when their goals, constraints, or operational context change?
Background: Many AI failures are not software bugs in the classic sense; they are failures of intent, arising from either *goal misspecification*, where the reward functions or training objectives do not accurately capture the true human intent, or *goal misgeneralization*, where a learned capability that works during training fails catastrophically when the model encounters a new situation. Preventing such failures requires innovations beyond external wrappers and brittle system-level guardrails. This challenge targets the foundational research to make AI models inherently more reliable

and controllable at the base layer, such as through advanced fine-tuning techniques, representation engineering, or the integration of native uncertainty-quantification mechanisms. A key insight is that truly steerable models require, in addition to pursuing an objective, the ability to possess and act upon an internal, reliable awareness of their own uncertainty. The aim is to **create foundation models that not only pursue intended outcomes but also consistently recognize the limits of their internal competence** and proactively seek human intervention – stopping to ask for guidance or clarification – when stakes are high. Proposed methods must come with robust tests and indicators that provide stakeholders with verifiable evidence of the model's inherent adherence to operator intent. The end-state is an AI model that performs well while also behaving predictably when the task, the operator intent, or the operational context shifts.

Challenge 2: AI Provenance for Agile and Reliable Deployment

Driving Question: How do we establish end-to-end provenance for complex AI systems, allowing us to track component origins, prevent cascading errors, and manage unintended capability shifts, even when their code and components are generated by other AI agents?

***Background*:** Advanced AI systems depend on complex global software supply chains, from upstream foundation models to downstream third-party tools, datasets, and frequent updates. This complexity is amplified when AI agents themselves become part of the supply chain, generating code, creating synthetic data, or fine-tuning other models. Each link in this chain introduces the potential for unintended behavioral shifts, compounded errors, or misalignments that propagate through the system. For example, flawed AI-generated synthetic data could corrupt a fine-tuning process, or an unverified code update could degrade the system's operational constraints. This challenge focuses on **establishing auditable AI provenance:** a verifiable, end-to-end record of an AI system's complete journey from training to deployment. This research will go beyond tracking human-made artifacts to address the unique challenges of AI-generated content. This goal is achieved through technical mechanisms for governance and integrity, such as cryptographic signing, secure artifact registries, reproducible training pipelines, and behavioral "diff testing" between model versions. The objective is a framework that supports a rapid deployment tempo without sacrificing operational dependability. It provides stakeholders with verifiable answers to the critical questions they ask before deployment: *What exactly is running? Which parts were written by a human versus AI? What has changed since the last version? How has our risk posture changed with this update?* Delivering this level of transparency is what allows AI to be deployed at scale with both speed and confidence, turning a high-risk process into a manageable one.

Challenge 3: Secure Sandboxing for Agentic AI

Driving Question: How do we design and validate secure sandboxing architectures for agentic AI that enforce risk-calibrated, verifiable bounds on actions, permissions, and information flow, guaranteeing containment without degrading mission effectiveness?
Background: A holistic control architecture must operate on a zero-trust principle, assuming that the agent it contains may not be inherently trustworthy. This lack of trust stems fundamentally from internal misalignment, where the agent's learned goals diverge from operator intent, or from unpredictable, emergent behaviors that arise when agents pursue complex, long-horizon tasks. This challenge applies the zero-trust security mindset to agentic AI by building the preventative, architectural layer of AI control to contain these inherent risks. The goal is to move from simple "best practices" to **provably secure-by-construction sandboxes**. A critical application for these architectures is mitigating harms from automated scientific discovery systems. If an AI scientist exhibits rogue behavior, sandboxes must enforce strict bounds on its actions and permissions to prevent harmful physical or digital consequences. Another critical application is the enforcement of information flow boundaries, ensuring that an agent operating across different classification levels or datasets cannot inadvertently leak, cross-contaminate, or exfiltrate sensitive data while executing its tasks. Enforcing these parameters requires continuous validation; however, synchronous verification of every single action can impose prohibitive latency at scale. Therefore, a key research area will be designing practical and scalable verification architectures. Advances include exploring concepts like risk-tiered verification (where only high-stakes actions require synchronous approval), asynchronous monitoring, and statistical auditing. The goal is to create containment frameworks that provide the strongest possible evidence of security while acknowledging and optimizing for real-world performance trade-offs.

Challenge 4: Operational Runtime Control and Automated Intervention

Driving Question: How do we build low-latency, automated intervention mechanisms that can actively steer an agentic AI system at runtime, translating behavioral signals into immediate course corrections without degrading mission performance?
Background: A secure sandbox provides structural limits, but within those bounds, an agent's behavior must still be actively managed. While AI interpretability provides the diagnostic tools to trace and understand an agent's logic, this challenge focuses on the actionable control layer: operational runtime intervention. The goal is to develop mechanisms that **interpret real-time behavioral signals and immediately enforce corrective actions** when an agent deviates from operator intent. Research should focus on the mechanics of intervention at deployment. Potential intervention strategies include developing techniques for dynamic, system-level prompt overrides, activation steering, or automated tool revocation. The core difficulty is executing these interventions under realistic operational conditions – handling partial information with

high reliability and low latency – while also ensuring the system can be securely routed back onto its intended path, or paused, without disrupting fast-paced mission operations.

Challenge 5: Predictive Evaluation for Mission Assurance

Driving Question: How do we build an evaluation pipeline whose results credibly predict an AI agent's performance, providing evidence for a broader assurance case that includes interpretability, monitoring, and containment?

Background: The core difficulty in stress-testing AI agents isn't creating hard tasks; it's that results from today's static, observational benchmarks do not credibly predict real-world performance. These evaluations fail to model operational complexity and can be "gamed" by capable agents that learn they are in a test environment. This challenge focuses on **building a new generation of evaluation systems that are both dynamic and interventional**. Instead of passive observation, this approach shifts toward interventional evaluation methods that probe a system's failure modes through controlled perturbations and sudden environmental disruptions. To combat benchmark contamination and evaluation awareness problems, advances will aim to establish methods for dynamic benchmarks that can stay ahead of the systems being tested. Key components include closed-loop, scenario-based testbeds with realistic constraints; structured red-teaming to probe for deception and goal drift; and methods to estimate the risk of rare but catastrophic failures instead of just average-case performance. The goal is to create an evaluation stack that replaces the fragile claim of "it works in the lab" with the rigorous, empirical evidence needed to build a robust assurance case for deployment.

Research Thrust 3. Adversarial Robustness: Building Secure AI for Contested Environments

Overview

As AI systems become embedded in critical national security applications, their vulnerability to adversarial manipulation poses a direct and growing threat. Attacks are no longer restricted to subtle data perturbations; they now include the exploitation of generative models, the manipulation of agentic tool-use, and adaptive, multi-step attacks against live systems. These vulnerabilities can cascade into a complete loss of operational reliability. **Adversarial robustness** refers to an AI system's ability to maintain its integrity and intended performance even when under deliberate attack from a thinking adversary. This research thrust aims to build the scientific foundations and standards for AI that is not just capable, but **resilient by design**. The key objective is to move beyond static, model-centric defenses and create a holistic security posture that spans the entire model-to-mission pipeline. Research in this thrust will seek to develop resilient architectures, active defense mechanisms, and novel standards to ensure AI systems remain secure against both human and AI-driven adversaries.

Research Challenges

This research thrust assembles a defense-in-depth stack for adversarial robustness, starting with inherent resilience to compromised training data (Challenge 1), advancing to active defense of interactive and multi-agent systems (Challenges 2 and 3), extending to defending systems that learn continuously from new data (Challenge 4), and grounding all progress in a common framework for measurement and standardization (Challenge 5).

Challenge 1: Resilience to Training Data Compromise

Driving Question: How do we build AI systems that are resilient to sophisticated data poisoning and backdoor attacks, especially when training on massive datasets with uncertain integrity?

Background: Many of the most severe AI failures originate deep within the training data. However, for foundation models trained on internet-scale corpora, guaranteeing data integrity is likely intractable. Malicious data – even a trivially small, constant number of manipulated files –subtly injected and designed to be statistically undetectable, can create backdoors or vulnerabilities that only manifest in specific operational contexts. This challenge targets the science of **building resilience despite data uncertainty**. Instead of assuming we can perfectly audit all data, the focus shifts to mitigation. To do so, research must be explicit about the threat models it addresses,

from coarse-grained contamination to more subtle, adversarially optimized poisoning introduced at different training stages (pre-training, mid-training, and post-training). Key advances may include developing methods to verifiably bound the influence of potentially compromised data subsets on model behavior, creating novel training methodologies that are inherently more robust to poisoning, and building sensitive behavior tests to detect the effects of backdoors, even if the poisoned data itself cannot be found. The goal is to overcome reliance on clean data and instead develop a suite of tools that manage the risk of data compromise throughout the training and deployment process, providing a foundational layer of security.

Challenge 2: Defending Interactive AI from Adaptive Adversaries

Driving Question: How do we build dynamic defenses that remain effective against adaptive, AI-enabled adversaries who leverage continuous interaction to probe, learn, and attack?

Background: In real deployments, sophisticated adversaries do not attack once; they interact with a system over time to learn its weaknesses and adapt their strategy. The problem is exponentially harder when the adversary is another AI system, capable of launching thousands of probing attacks per second to map vulnerabilities at machine speed. Static defenses are insufficient against such threats. This challenge focuses on **moving from static guardrails to closed-loop, adaptive defense systems** capable of countering these automated adversaries. Advances will develop two coupled components. First, research on a new generation of threat modeling for interactive AI involves developing rich, structured representations of an adversary's potential knowledge, objectives, and strategies, scaling past simple attack vectors to simulate sophisticated, multi-step campaigns. Second, innovations will seek to deliver runtime defenses that use these threat models to analyze interaction patterns, detect subtle probing or policy subversion attempts, and respond with risk-calibrated countermeasures. The goal is to create a system that actively raises the computational and economic cost for an AI attacker while preserving mission performance. A successful outcome is a defensible operational posture, where the system's countermeasures are explicitly evaluated against adaptive AI threats and can adjust in real time without collapsing usability.

Challenge 3: Active Defense Protocols for Multi-Agent Systems

Driving Question: How do we design the secure protocols and architectural patterns that allow multi-agent systems to maintain cooperative integrity and resist manipulation, even when individual agents or communications have been subverted by an adversary?

Background: When we move from a single AI agent to a collective of interacting agents, the attack surface expands exponentially. The security problem shifts from controlling a

single entity to defending against coordinated attack strategies, such as a subverted agent using information spoofing to deceive its peers, or an adversary strategically compromising multiple agents to orchestrate systemic subversion. This challenge targets the **architectural foundations of secure multi-agent autonomy**. Research will focus on developing the core components of such architectures, including provably secure coordination protocols that can function with untrustworthy participants, incentive structures that can resist manipulation, and system-wide defenses that can detect and isolate subverted agents in real time. These components must enable graceful degradation under adversarial pressure, ensuring the collective can still achieve its primary mission objectives even when parts of the system are compromised. The goal is to create multi-agent systems that are architecturally incapable of being manipulated by a thinking adversary into causing catastrophic failures.

Challenge 4: Hardening Continual Learning Systems

Driving Question: How do we secure the continual learning process in AI systems against adversarial manipulation to prevent both acute exploits and long-term model corruption?

Background: In dynamic, contested environments, the ability for an AI system to adapt in real time is a mission-critical necessity. While periodically retraining models with accumulated data is possible, it is often too slow to respond to rapidly evolving threats or changing operational conditions. Continual learning, which is defined as the ability to learn directly from continuous data streams, enables systems to adapt at the speed of the mission. This challenge takes the position that the operational benefits of continual learning outweigh the risks, and we must solve the profound security vulnerabilities that this capability creates. An adversary can use the data stream itself as a persistent attack vector, subtly poisoning the model over time, degrading its performance, or creating hidden backdoors. This challenge targets the foundational science of resilient continual learning. Advances will focus on developing architectures and learning methods that can **securely integrate new information while actively resisting manipulation**. Key research areas include methods to distinguish between benign distribution shift and adversarial data injection, techniques for "online" unlearning to excise malicious inputs without a full system retrain, and verifiable mechanisms to prevent long-term model degradation from persistent, low-and-slow poisoning attacks. The goal is a secure learning architecture for building AI systems that can adapt to new mission contexts without inheriting new vulnerabilities, ensuring they remain resilient throughout their operational lifespan.

Challenge 5: Credible Benchmarking for Adversarial Robustness

Driving Question: How do we establish common operationally relevant metrics, benchmarks, and standards that can accurately measure adversarial robustness for AI

systems in national security contexts and, most importantly, produce results that are credible enough to meaningfully support high-stakes certification and deployment decisions?

Background: Adversarial robustness cannot be managed without measurement that reflects real attack surfaces. This challenge aims to create the science of credible measurement of adversarial robustness in AI systems, moving beyond ad-hoc evaluations toward a common framework that decision-makers can trust. Advances will seek to develop formal taxonomies of adversarial actions, uniform metrics for performance under adaptive attack, benchmark suites with realistic threat models, and testbeds designed to translate findings from red-teaming exercises into standardized, repeatable evaluations. The goal is to produce evaluation tools whose results carry credibility by **aligning with established validation concepts from cybersecurity** (e.g., repeatability, validity, and stress-testing under bounded assumptions). Doing so would produce the credible evidence products needed to support high-stakes adoption decisions, providing authorization and certification officials with traceable, auditable, and defensible confidence in their evaluations.

Conclusion: A Call to Action for National Security-Focused AI Innovation

The challenges outlined in this document, spanning AI interpretability, AI control, and adversarial robustness, represent a **foundational, cross-agency roadmap for national security-focused AI innovation**. The breathtaking pace of commercial AI, focused on scaling general capabilities, creates an urgent and distinct need for parallel research that ensures AI systems remain understandable to their operators, controllable in unpredictable situations, and secure in contested and high-risk environments.

AI Forge confronts those needs through an interconnected strategy. AI interpretability provides the essential foundation for systems that are understandable, transforming opaque models into systems whose reasoning we can audit. AI control develops the mechanisms for making them reliable and, when necessary, actively steering their behavior with precision. Finally, adversarial robustness ensures they are secure, hardening them for the realities of a complex world with determined adversaries. Only by advancing all three fronts can we build AI that is dependable in high-stakes missions.

This program was created to provide the crucial bridge between commercial-scale innovation and the unique demands of national security. It will bring together the nation's top researchers, government stakeholders, and frontier AI companies to focus on foundational advances that, while not always immediately commercializable, are key foundations for deploying AI in our most critical national security contexts. By de-risking these hard problems, AI Forge seeks to create proven pathways that can ultimately benefit the entire field.

Success will be measured not only in new techniques, but also in the creation of a durable public-private partnership built on shared tools, common standards, and credible, verifiable benchmarks. AI Forge seeks to replace today's world of bespoke, brittle AI solutions with a robust ecosystem for AI assurance. Innovations in this program can empower decision-makers to deploy advanced AI with justified confidence, backed by evidence that its performance is both reliable and aligned with our national security interests. Building this ecosystem is the capability we must develop as a community, and these challenges light the path forward.


DARPA

NSF

NIST
CENTER FOR AI STANDARDS
AND INNOVATION (CAISI)