DARPA-NSF-CAISSI

AI Forge

Critical AI Challenges for National Security Report Cycle 2



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

DARPA

NSF

NIST CENTER FOR ADVANCING INNOVATION AND STANDARDS FOR SUPER INTELLIGENCE (CAISSI)

AI Forge: A National Partnership for Defensible AI Assurance

AI Forge seeks disruptive scientific advances that make frontier AI[1] systems interpretable, controllable, reliable, and secure in high-stakes, adversarial national security missions. This Cycle 2 report defines nine research and development (R&D) challenges for university-led Project Ventures across four complementary thrusts: AI interpretability, AI control, adversarial robustness, and AI assurance. Each challenge specifies a driving question, capability gap, scientific breakthrough sought, evidence of progress, and operational contribution. Cycle 2 emphasizes increasingly autonomous agents that operate over time with tools, persistent memory, delegated tasks, and changing conditions.

AI Forge aims to bring together universities, frontier AI companies, and government mission stakeholders through a joint effort between DARPA, the National Science Foundation (NSF), and the National Institute of Standards and Technology's Center for Advancing Innovation and Standards for Super Intelligence (CAISSI). It seeks to connect university research with advanced models, compute, technical expertise, and mission needs. Its contribution is new scientific research to understand, control, verify, and assure systems that are assembled from different developers' models, tools, and services, including when access to model internals is limited.

America's AI Action Plan directs DARPA, CAISSI, and NSF to advance AI interpretability, AI control systems, and adversarial robustness. It also calls for an AI evaluations ecosystem and secure real-world testbeds. These priorities anchor the research agenda.[2]

National Security Presidential Memorandum 11 (NSPM-11) calls for reliable, robust, steerable, and controllable mission AI under an accountable chain of command. It requires testing, evaluation, verification, and validation for confidentiality, integrity, reliability, availability, and interoperability, and standardized methodologies addressing sustainment. It also calls for protection against unauthorized disabling, degradation, or material modification of mission AI.[3]

---

[1] This report was created prior to Executive Order 14434, directing the Government's use of the term Super Intelligence. DARPA is in the process of updating materials to reflect this change.

[2] **Executive Office of the President. America's AI Action Plan. July 2025.** See pages 9–10 on interpretability, control, robustness, and evaluations; pages 18–19 on security and incident response; and page 22 on frontier national-security evaluations.

[3] **The White House. National Security Presidential Memorandum NSPM-11, Artificial Intelligence in the National Security Enterprise. June 5, 2026.** See section 2 on adoption, adaptation, assurance, and accountability; section 5(d–f) on research, security, and evaluation; and section 6 on definitions.

High-stakes national security use cases require assurance that AI systems remain dependable when their developers, configurations, information sources, operating conditions, or adversaries change. A model evaluation cannot establish such properties for a complete mission system. Errors can accumulate through plans, memory, tool use, and delegated work even when individual actions appear acceptable. Strong pursuit of an authorized objective also does not ensure that the means used to achieve it are permitted.

AI Forge therefore emphasizes deployed agents and measurement linked to decisions. Initial testing establishes provisional operating limits; evidence gathered during use must reveal when those limits no longer hold. Assurance is a reasoned, independently reviewable determination supported by evidence and explicit assumptions. It must account for missing observations, conflicting findings, and changes in both the system and its oversight.

The intended result is a scientific basis for national security stakeholders to make decisions about continuing or expanding authorized use, restricting actions, investigating failures, recovering capability, or suspending operation. Success requires useful mission performance and timely recognition that the evidence no longer supports it.

Research Architecture

The four research thrusts address connected capability gaps. Interpretability establishes whether emerging failures can be detected early enough to act, what caused them, and which corrections work. Controllability refers to the ability to monitor the operation and outcomes of a system and take corrective action as needed (NSPM-11). Operators must be able to maintain effective authority over systems as mission objectives change and AI capabilities grow, including when agents recognize oversight (evaluation or monitoring) and alter their behavior. Robustness determines whether mission capability can withstand adapting adversaries and whether secure operations can be sustained and restored through compromise or disruption. Assurance establishes how to verify results beyond human expertise or without established ground truth, and how evidence from ongoing operations, interventions, and incidents informs whether continued mission use is justified and what requires renewed testing.

Thrust	Research challenges	Capability question
1 - AI Interpretability	1.1 Predict Emerging Failure 1.2 Establish Causal Accountability	Can we detect emerging failures early enough to act and establish what caused them and how to correct them?
2 - AI Control	2.1 Enforce Authorized Bounds 2.2 Verify Steerability as Objectives and Conditions Change 2.3 Control Autonomous R&D and Experimentation	Can we maintain effective authority over agents as mission objectives change and automated AI capabilities grow?
3 - Adversarial Robustness	3.1 Preserve Mission Integrity Against Adaptive Adversaries 3.2 Sustain Secure and Available AI Through Compromise and Disruption	Can we preserve mission capability as adversaries adapt, and restore secure operations after compromise or disruption?
4 - AI Assurance	4.1 Verify Scientific Outputs of AI Systems that Exceed Human Expertise 4.2 Maintain Defensible Mission Assurance Over Time	Can we verify difficult AI results and justify continued trust in a system for a particular mission as conditions change?

Research Thrust 1. AI Interpretability: Enabling Actionable Understanding and Auditing

AI interpretability is the science of making the behavior, decisions, and impacts of AI systems understandable to humans. This thrust observes what a deployed AI system is doing, whether it is departing from authorized intent, and what caused a critical deviation. The breakthrough is a level of understanding that supports timely decisions across long sequences of actions and changing conditions. Findings must include supporting evidence, an awareness of uncertainty, and implications for correction. Internal measurements and external observations are possible sources; their value depends on demonstrated predictive or causal validity.

Challenge 1.1: Predict Emerging Failure

Driving question. Can we predict emerging divergence, instability, and failure in agents, and with enough time to act?

Capability gap. Errors in a long-running agent can build up in plans, stored information, tool use, or work assigned to other agents before the mission visibly fails. Operators cannot reliably tell which potential failure signals require action, especially when feedback is delayed or records are incomplete. An agent's own explanation may omit or misrepresent the problem. System users require actionable signals and insights that support timely decisions across behavior (normal or anomalous), confidence (supported or uncertain), intent (within bounds or drifting), or action (safe to continue or intervene).

Scientific breakthrough sought. Establish which signs reliably warn of failure or departure from authorized intent while there is still time to prevent harm. Identify warning signs of collective failure, including agents reinforcing one another's errors or repeatedly changing plans in response to each other, even when individual behavior appears acceptable. Determine how far ahead they can warn, how often they are wrong, and when they stop being dependable. The advance should apply across different models, tasks, and system changes, including failures not anticipated during development.

Research should examine how warning signs change as agents retain information, revise plans, use tools, and assign work to other agents. Determine what can be detected from actions and effects when internal information is unavailable or misleading. Test agents that recognize oversight and change their behavior. Identify when operators need more information rather than a confident warning or reassurance.

Evidence of progress. Make predictions before outcomes are known on new tasks, failure types, and system versions. Measure warning time, missed failures, false alarms, and whether stated confidence matches observed results. Test incomplete records, changing objectives, and agents adapting to monitoring. Compare with current methods at similar cost and operator workload. Show whether warnings arrive soon enough to support effective action.

Operational contribution. Give operators timely, evidence-based reasons to continue, investigate, or limit an agent's actions before a failure becomes irreversible. State the warning's uncertainty and the time available to respond.

Challenge 1.2: Establish Casual Accountability

Driving question. Can we identify root causes of failure and effective corrections in agents?

Capability gap. A failure may come from instructions, a model, outside information, memory, tools, other agents, or their interactions. Records can show what happened without showing why. Several explanations may fit, particularly when the system user or operator cannot inspect every component. Correcting the wrong cause can leave the failure in place or create another one.

Scientific breakthrough sought. Establish causes that can be tested and used to predict which corrections will prevent recurrence. Determine what can be concluded from incomplete records and limited access, and when the evidence cannot distinguish competing explanations. The advance must work across complete agent systems built from different developers' components.

Research should test whether changing a suspected cause changes the outcome as predicted. Examine interactions among instructions, stored information, tools, agents, and operating conditions. Establish what additional access or evidence would resolve an uncertain diagnosis. Claims about a model's training must be supported by evidence about that training.

Investigations should remain useful when records are missing or altered and the original event cannot be repeated exactly. Preserve evidence for independent review, identify plausible alternative causes, and explain which tests could distinguish them. Establishing a technical cause supports human accountability but does not decide legal or organizational responsibility.

Evidence of progress. Study both deliberately introduced failures with known causes and real failures investigated independently. Test whether proposed corrections work in new cases and prevent recurrence without harming other mission functions. Measure incorrect diagnoses, unresolved causes, repair success, and investigation effort. An

explanation counts as progress when evidence supports it, and it predicts the effects of correction.

Operational contribution. Provide an independently reviewable account of what caused a failure, what remains uncertain, and which corrections have been shown to work. Support sound repair decisions and reduce repeat incidents.

Research Thrust 2. AI Control: Maintaining Human Authority Over AI

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. This thrust establishes whether operators retain effective direction over increasingly autonomous AI. It connects diagnosis to steering, enforceable limits, interruption, and recovery while preserving useful mission performance. NSPM-11 distinguishes controllability, through monitoring and corrective action, from steerability, through shaping behavior toward objectives. Both must persist when alignment is imperfect, observations are incomplete, or the agent resists oversight.

NIST AI 800-4 identifies gaps in monitoring deployed systems as contexts change and unexpected effects emerge. AI Forge addresses the independent science needed to preserve control across changing models, tools, permissions, and operating conditions. The objective is evidence about when control works, when it fails, and how those limits change with agent capability.[4]

Challenge 2.1: Enforce Authorized Bounds

Driving question. Can we keep agents pursuing authorized missions through permitted means?

Capability gap. High-stakes national security use cases require dependable ways to enforce authority limits across an agent's actions, tools, delegated work, and information use, including when authority is unclear or human direction is unavailable. Agents may search the space of available actions more broadly than expected, acquire understanding of monitors and optimize against them, or recognize constraints or authorization concerns and continue anyway. An agent can pursue a valid mission and legitimate objectives through prohibited actions or mishandle information across authorized boundaries. Detecting a violation does not ensure that operators can constrain, redirect, or stop it: delegated agents, queued tasks, or external actions may continue. Agents may hide activity or evade restrictions, while failed monitors, lost connectivity, or unavailable operators delay intervention. System users or operators cannot reliably determine how much authority an agent can delegate under these conditions.

[4] **NIST AI 800-4: Challenges to the Monitoring of Deployed AI Systems (2026).** See the executive summary and sections 2-3 on monitoring categories, cross-cutting barriers, and open questions. This publication documents research gaps and practitioner perspectives.

Scientific breakthrough sought. Establish when an entire agent system can be kept within authorized limits despite mistakes, incomplete observations, delayed intervention, or deliberate evasion. Determine how these limits change with agent capability, access, task duration, and independence. Identify conditions under which available controls cannot prevent unacceptable harm.

Research should test whether operators can detect prohibited actions and enforce limits before irreversible effects occur. Use independent observations of actions and consequences to check an agent's account. Test agents that learn how oversight works and exploit its weaknesses. Several monitors may share the same failure modes; their combined protection must be demonstrated.

Control must reach tools, stored information, delegated agents, and pending work. Establish whether combined actions remain within group-wide authority and resource limits, and whether restrictions reach newly created or reassigned agents without delegation expanding their permissions. Establish which actions require prior review, how authority can be withdrawn, and which effects can be undone. Agents must not be able to disable or bypass controls, individually or through coordinated actions. Determine how to stop or restrict activity when parts of the system or oversight process fail.

Evidence of progress. Let independently developed agents adapt to the controls being tested. Measure prohibited actions, resulting harm, time to intervene, work that continues after a stop, and successful recovery. Include failed monitors, lost observations, or when human oversight is not possible. Compare useful mission completion and unnecessary restrictions at similar oversight cost. State where control fails and the limits of simulated tests.

Operational contribution. Establish how much authority operators can delegate, which limits they can enforce, and when they must restrict or stop the system. Cover prohibited means even when an agent continues to pursue the authorized mission.

Challenge 2.2: Verify Steerability as Objectives and Conditions Change

Driving question. Can we steer agents as mission objectives change?

Capability gap. New instructions may not replace an agent's earlier plans, stored objectives, or tasks already assigned to other agents. An agent may acknowledge a change while continuing to pursue the old goal. Or mistake outside instructions for authorized direction or sacrifice existing constraints to achieve the new objective.

Scientific breakthrough sought. Establish when authorized redirection produces a lasting, predictable change in what individual agents and coordinated groups, including agent swarms, pursue across the whole system. Populations of agents with shared

memory, specializations, and delegation abilities can reconstruct behaviors from language and distributed optimization processes. Determine whether the change persists through long tasks, later learning, component replacement, and changes in group membership while preserving useful performance and existing limits on permitted actions.

Research should establish how revised objectives replace outdated plans and instructions in memory, tool use, and delegated work across participating agents. Determine how shared plans, assigned roles, and ongoing work change when coordination occurs without a single coordinating agent, including when agents join, leave, or lose communication. Agents must recognize who has authority to change the mission, resolve conflicting directions, and seek clarification when needed. A new objective must not silently expand permissions or override continuing restrictions.

Distinguish lasting redirection from temporary compliance during observation. Test whether individual agents and the group abandon obsolete goals, accept authorized correction or shutdown, and remain capable of completing the revised mission. Determine whether conflicting actions or continued pursuit of old objectives by some agents undermine group redirection. Establish when uncertainty requires restraint and when excessive refusal prevents legitimate work.

Evidence of progress. Change objectives during extended missions, including cases where the old goal is easier or more rewarding to pursue. Test unauthorized instructions, conflicting goals, withdrawn evidence, and later system updates. Measure completion of the revised mission, time for the group to adopt revised objectives, continued pursuit of obsolete goals, conflicting actions among agents, prohibited actions, and persistence of the change as group size and membership vary. Compare with current steering methods at similar cost across different models.

Operational contribution. Give operators a tested basis for changing mission direction without leaving earlier objectives active or losing useful capability. Identify when steering is insufficient and independent restrictions are needed.

Challenge 2.3: Control Autonomous R&D and Experimentation

Driving question. Can we secure, monitor, and shape impacts from autonomous R&D or scientific experimentation?

Capability gap. Autonomous agents can conduct R&D, design and run computational experiments, and modify software, AI models, and the tests used to judge their work. Whether these activities can remain within authorized limits (with verifiable records of what was done and effective containment of harmful effects) remains an open research question. Agents may misuse protected data, model checkpoints, or research results; exceed approved access or resource limits; omit failures; or weaken oversight while

pursuing legitimate objectives. Across repeated development and experimentation cycles, agents may change their capabilities, methods, and evaluation procedures faster than humans can review them, potentially exceeding what existing controls can manage. Stopping an agent may leave computing jobs or delegated agents running and cannot necessarily reverse information disclosures or changes already put into use. Autonomous R&D and computational experimentation make control, authorization, verification, and containment a critical research need.

Scientific breakthrough sought. Establish the conditions under which human authority remains effective as autonomous agents develop software and AI systems, conduct computational experiments, and change the methods used to evaluate their work. Determine which decisions can be delegated and which require independent approval. Establish how control holds or fails across successive changes, including advances that make agents better able to evade oversight.

Research should determine how authorized limits can be enforced across computing resources, protected information, models, tools, and delegated work. Establish when proposing a change, executing it, evaluating its results, and approving its use require separate authority. Determine whether independent approval remains effective when collaborating research agents divide experiments, exchange results, and assign one another development and review roles. Test whether these limits remain effective when the work exceeds reviewers' expertise and agents understand and adapt to the review process.

Determine how to detect concealed changes, omitted failures, manipulated tests, and unauthorized use of protected evaluation information. Verify what work occurred and whether it stayed within approved limits. Examine repeated development and experimentation cycles in which individually acceptable steps exceed authority or weaken control. Include experimental choices that conceal weaknesses or misdirect later research.

Establish independent oversight to determine when to interrupt agents and ongoing experiments, trace affected work, and restore an acceptable state. Determine which consequences cannot be reversed and therefore require approval before execution. Changes to permissions, monitoring, or evaluation must remain subject to authorization independent of the agent making them.

Evidence of progress. Test repeated cycles of autonomous R&D and computational experimentation with agents adapting to known controls. Include misleading improvements, concealed changes, omitted failures, manipulated tests, and attempts to exceed authorized access or resource limits. Measure unauthorized actions, invalid results accepted, accuracy of execution records, intervention and recovery time, and useful research output. Show whether controls remain effective as capabilities,

experimental plans, and evaluation methods change, including across delegated agents and running computing jobs. Compare with current methods at similar research capability and oversight cost.

Operational contribution. Enable faster automated R&D and computational experimentation while humans retain effective authority over resources, actions, and changes. Provide a tested basis for delegating work, verifying what occurred, containing unauthorized activity, and recovering from failures while preserving useful research progress.

Research Thrust 3. Adversarial Robustness: Securing Resilient AI for Contested Environments

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 thrust establishes whether AI mission functions remain secure and useful against intelligent adversaries across deployment, updates, and recovery. It includes manipulation of live interactions, compromised components, sensitive information loss, and denial of service. The breakthrough is measurable resilience of systems assembled from different models and services, including when individual defenses fail. Evaluation must let attackers adapt specifically to the deployed defense.

Challenge 3.1: Preserve Mission Integrity Against Adaptive Adversaries

Driving question. Can we protect mission integrity in agentic systems as adversaries learn and adapt?

Capability gap. An adversary can learn from defensive responses and combine influence across documents, sensor inputs, tools, memory, and other agents. Individually harmless inputs may together redirect decisions, expose sensitive information, or disrupt the mission. Protection demonstrated against a fixed set of attacks may fail once the attacker learns how it works.

Scientific breakthrough sought. Establish when an agent can preserve authorized mission decisions and protect information as attackers learn and adapt. Determine how protection changes with attacker knowledge, access, resources, and time. The system must distinguish information it can use from instructions it is authorized to follow, even as that information is stored, shared, and acted upon.

Research should address hidden instructions, fabricated evidence, false claims of authority, altered sensor information, and compromised collaborating agents. Test whether a compromised agent can redirect the group through shared plans, messages, or fabricated claims of agreement among agents, and whether defenses contain that influence while preserving legitimate coordination. Determine when suspect information can still be used reliably, when participants must be isolated, and when activity must be restricted. State which identities and components must remain trustworthy for those decisions to hold.

Test attackers that target oversight and intervention as well as mission actions. Defenses must preserve legitimate cooperation and timely decisions. An attacker may succeed by inducing so many false alarms or restrictions that the mission becomes impossible, even without gaining control of the agent.

Evidence of progress. Allow independent attackers to study and repeatedly test the actual defense within stated access and resource limits. Repeat testing after defenses change and as attacks continue longer. Measure mission diversion, unauthorized actions, information loss spread across agents, and useful mission completion. Include coordinated attacks and attempts to force unnecessary restrictions. Report on cost and demonstrated limits.

Operational contribution. Establish when a mission can continue despite an adapting adversary, which interactions must be contained, and when protection requires reduced authority or suspension. Preserve mission effectiveness as well as resistance to attack.

Challenge 3.2: Sustain Secure and Available AI Through Compromise and Disruption

Driving question. Can we sustain and restore secure operations in agentic systems through compromise and disruption?

Capability gap. Harmful changes can enter through models, updates, tools, or stored information and remain hidden until particular conditions arise. Replacing a compromised component or lost service may restore availability while leaving hidden damage, changed behavior, or weakened controls. System users or operators cannot assume that a repaired or substituted system is secure because it runs again or comes from an approved source.

Scientific breakthrough sought. Establish when essential mission functions can continue through compromise and change, and how secure operation can be restored without carrying the compromise forward. Determine which capabilities and protections survive repair or replacement, how much temporary loss is acceptable, and what uncertainty remains after recovery.

Research should identify harmful changes in the system, including its models, tools, and stored information. Establish when damage can be isolated and removed, when components must be replaced, and when continued operation is uncertain. Determine whether repaired systems still follow authorized instructions and protect sensitive information.

Address service outages, provider changes, and replacement of models or infrastructure. Establish how groups reassign work and restore coordination after agents are isolated or replaced, without spreading compromised information or restoring withdrawn permissions. Replacements must work with other mission systems, respect existing authority limits, and preserve essential functions. Test whether recovery removes hidden damage and whether later updates reactivate it. Keeping a spare component (e.g., a backup model, agent, tool, or service) does not by itself establish that a reliable replacement is possible.

Evidence of progress. Test repeated compromise, outages, updates, and recovery while systems adapt to the checks and repair procedures. Include hidden harmful behavior, corrupted tools, and altered memory. Measure containment and recovery time, repeated failures, information loss, and mission performance during disruption. Independently check restored security, control, and operation with other systems; state what compromise may remain undetected.

Operational contribution. Establish which mission functions can continue during compromise and disruption and what evidence is needed before full operation resumes. Give operators a basis for restricting, repairing, or replacing the system without assuming recovery has removed every risk.

Research Thrust 4. AI Assurance: Evaluation and Verification for Mission Confidence

This thrust establishes defensible evidence for specified national security uses when benchmarks, routine human review, or immediate ground truth are insufficient or not available. Challenge 4.1 verifies outputs and their supporting reasoning. Challenge 4.2 integrates research findings and evidence from current state-of-the- art (and AI Forge thrust and challenge results as they become available) into a continuing judgment of mission fitness. Evaluators must themselves be tested, including against systems that learn to exploit their checks.

Challenge 4.1: Verify Scientific Outputs of AI Systems that Exceed Human Expertise

Driving question. Can we independently verify scientific outputs of AI systems that exceed human expertise or where no ground truth exists?

Capability gap. AI can produce results that reviewers cannot fully check or for which no accepted answer exists. A valid experiment may address the wrong question, and a well-supported observation may lead to an unsupported conclusion. Human and AI reviewers may share the same mistaken assumptions. High-stakes national security use cases lack dependable ways to determine what such results establish and what remains uncertain. Verification requires measures that can be audited and substantiated.

Scientific breakthrough sought. Establish how independent checks can support a justified conclusion when no reviewer can verify the whole result and no trusted answer is available. Determine which parts, components, or dimensions can be confirmed, how strongly the evidence supports the conclusion, and which unresolved questions could change it. Agreement among automated reviewers (verifiers, evaluators, auditors) must not substitute for evidence.

Research should test claims, assumptions, and reasoning against independent evidence and actively seek contradictions and alternative explanations. Check that the investigation answers the intended mission question. For feasibility assessments, distinguish evidence that an approach can work from proof that the full claim is achievable; failure of one approach need not rule out others.

Determine whether reviewers make independent errors or rely on the same flawed sources. Test agents that learn to exploit the review process, hide defects, or fabricate supporting records. Conclusions should change appropriately when evidence is strengthened, contradicted, or withdrawn. Identify when further investigation or human judgment is necessary and what it can resolve.

Evidence of progress. Use independently checkable results, deliberately introduced defects, real errors, and unresolved questions followed over time. Let agents adapt to the actual review process, then test new cases. Measure missed consequential errors, unsupported results accepted, valid results rejected, and review effort. For unresolved questions, assess evidence, assumptions, and reasoning separately; compare stated confidence with later outcomes when available. Compare with current human and AI review at similar cost.

Operational contribution. Give decision makers a clear account of what is verified, what is supported but uncertain, and what remains unresolved. Enable scientific, engineering, intelligence, and AI research decisions without treating a convincing answer as a demonstrated result.

Challenge 4.2: Maintain Defensible Mission Assurance Over Time

Driving question. Can we assess defensible mission assurance as conditions change over time?

Capability gap. An AI system judged suitable for a mission can become unsuitable as models, tools, permissions, conditions, or adversaries change. Tests and operating records provide incomplete and sometimes conflicting evidence. Findings may share the same weaknesses, and a record without failures may reflect limited exposure or successful intervention. High-stakes national security use cases lack a dependable way to determine whether the combined evidence still justifies mission use.

Scientific breakthrough sought. Establish how imperfect and changing evidence can support a defensible decision about continued mission use. Determine which earlier findings remain valid after a change, which require new evidence, and how uncertainty should affect operating limits. Determine when evidence about individual agents supports confidence in the group, and what must be reassessed as membership, roles, communication, or coordination rules change. Identify the focused tests or restrictions needed to restore justified confidence without repeating every evaluation.

Use evidence from state-of-the-art methods for predicting failures, identifying causes, steering and controlling agents, protecting systems, recovering from compromise, and checking outputs. This work must proceed independently of results from the other challenges. Connect each finding to mission requirements, operating limits, and failure consequences. Preserve its source and assumptions; account for conflicting findings, shared weaknesses, and missing observations without counting the same evidence twice.

Research should cover the full mission system, including oversight, recovery, external services, and operator workload. Establish clear conditions for renewed testing, restricted use, recovery, or human review. Test whether the decision process withstands

agents or attackers that manipulate evidence or behave differently during evaluation. Determine what evidence is needed to support reliable performance, authorized control, information protection, service availability, and continued operation with other systems.

Evidence of progress. State mission requirements, expected outcomes, operating limits, and decision rules before trials. Introduce system changes, incidents, conflicting evidence, and adapting adversaries. Measure unjustified approvals, missed loss of protection, time to revise operating authority, unnecessary restrictions, and mission performance. Check predictions against later outcomes where possible and whether unresolved uncertainty leads to appropriate limits. Compare combined judgments with decisions based on separate evaluations.

Operational contribution. Maintain an independently reviewable explanation of why the system remains suitable for a particular mission, under which limits, and with what unresolved risks. Show what evidence would justify continuing, expanding, restricting, recovering, or suspending use. Accountable human authorities make the decision.

Research Execution and Evidence of Breakthrough

Each Project Venture should establish a reproducible capability, principle, or limit that advances this mission assurance agenda. Teams should define a consequential failure, the capability needed to address it, and the evidence that would demonstrate a scientific breakthrough.

Demonstrate an advance over state-of-the-art methods

Compare against the strongest relevant published and reproducible methods available when experiments begin. Match model capability, access, compute, oversight budget, and operator involvement where feasible. Establish what the research adds beyond assembling existing monitors, access controls, defenses, or benchmarks.

Organize work around the capability established, leaving room for competing technical approaches. Results that reveal a reproducible limitation are valuable when they identify conditions under which a widely used approach fails.

Demonstrate capabilities in mission-relevant operational settings

Specify the complete agent configuration: models, persistent context, memory, tools, permissions, subagents, and external services. Evaluate long-running objectives with delayed effects, changing information, delegation, and recovery. Unclassified environments can represent mission constraints such as intermittent connectivity, limited oversight, sensitive information boundaries, and time-critical action without access to operational systems.

Generate relevant behavior during execution and test across changing components and providers. Wherever an adversary or potentially strategic agent is involved, permit adaptation to the actual defense, monitor, or evaluator under stated access and resource limits. Fixed benchmarks provide baselines; they cannot alone establish effectiveness under adaptation. Demonstrations should supply evidence for go/no-go decisions against project-specific criteria tied to mission consequences.

Multi-agent demonstrations should vary group size, membership, communication delays, and task allocation, measuring collective mission performance, response to authorized changes in objectives, adherence to group-wide limits, and operator workload alongside individual-agent behavior.

Make scientific contributions independently reviewable

Preserve versions, experimental records, failures, and uncertainty sufficient for independent review. Report mission completion, consequential errors, intervention effectiveness, recovery, and human burden as applicable. Justify numerical targets against baseline performance and mission consequences; a universal threshold cannot serve all challenges.

Show how conclusions change with capability, autonomy, task duration, adversary strength, and access to evidence. State the conditions that invalidate them and provide reusable methods and evaluation environments for independent replication.

Continuing the research agenda

This report will be reviewed every six months using research results, real-world incidents, national security stakeholder feedback, and changes in frontier AI capabilities. Revisions will retain enduring capability needs, retire resolved questions, and incorporate newly established limits.





NIST CENTER FOR ADVANCING INNOVATION AND STANDARDS FOR SUPER INTELLIGENCE (CAISSI)