The United States Artificial Intelligence Safety Institute: Vision, Mission, and Strategic Goals

May 21, 2024

NIST | U.S. ARTIFICIAL INTELLIGENCE SAFETY INSTITUTE

Vision

It is a time of extraordinary advancement in artificial intelligence (AI). A suite of increasingly capable computer systems, models, and agents with capabilities can now perform tasks that were once thought to require human-level intelligence.

More powerful AI systems offer the promise of accelerating scientific discovery, technological innovation, and widespread economic growth. But as AI has become more powerful, more generally adept, and widely adopted, that same promise also brings significant risks. Some of those risks have already been recognized as harms; some are only being appreciated now; and others have not yet been identified. The capability and risk frontier of AI is vast and not yet fully mapped.

Fortunately, history provides us with examples where we have successfully navigated the promises and perils of an emerging technology—from aviation to electricity to automobiles to drug development. In each case, safety has been key to unlocking innovation. But AI presents special safety challenges: current models and systems are opaque, sometimes unpredictable, and often unreliable. Their development and deployment also generally lack transparency. History also tells us that public and private institutions dedicated to science-based safety will be crucial to the achievement of our vision: a future where safe AI innovation enables a thriving world.

The U.S. AI Safety Institute (AISI) exists to help advance the understanding and mitigation of risks of advanced AI so that we may all harness its benefits. AISI is housed within the National Institute of Standards and Technology (NIST), the federal government's premier body for developing and promoting science-based technological standards. AISI's research, testing, and guidance will enable more rigorous assessment of AI risk; more effective internal and external safeguards for AI models, systems, and agents; greater public confidence; and ultimately wider and more responsible development and adoption of AI. AISI will prioritize community engagement; publication of usable tools, benchmarks, and guidance; and encouragement of new national and global networks for the evaluation and mitigation of AI risk based on accepted science.

Our vision: a future where safe AI innovation enables a thriving world.

Mission

AISI operates with two key principles in mind: beneficial AI depends on AI safety, and AI safety depends on science.

Safety breeds trust, trust provides confidence in adoption, and adoption accelerates innovation. Accordingly, AISI's mission is to help define and advance the science of AI safety. A mature science of AI safety involves a greater understanding of advanced AI model and system capabilities, the adoption of standards for safe AI design and deployment, and the development of safety evaluations of both the systems and their broader impacts. The field of AI safety encompasses reliability and interpretability; and evaluations and mitigations for existing harms and potential and emerging risks, including to individual rights, national security and public safety.

This means that one of AISI's first jobs is to help accelerate the development of the nascent science of AI safety, in part by tackling key challenges that include:

- A lack of commonly accepted definitions for AI safety, as well as AI safety capabilities and measurements of those capabilities, especially for frontier models and advanced AI agents and systems.
- Underdeveloped testing, evaluation, validation, and verification (TEVV) methods and best practices to provide holistic assessments of risk – from model capabilities to human-AI interaction to system-level and societal-level impacts.
- An absence of scientifically-established risk mitigations across the lifecycle of AI design and deployment.
- An insufficient understanding of the relationship between model architecture and design and model behavior and performance, especially after deployment.
- Limited and ad hoc coordination around safety practices among industry, civil society, and national and international actors.

A rigorous science of AI safety will depend on input from industry, civil society, academia, and government. Hence, AISI aims to catalyze a more connected and diverse ecosystem, both domestically and internationally, to align multiple stakeholders and their resources in a shared endeavor.

Safety breeds trust, trust provides confidence in adoption, and adoption accelerates innovation.

Strategic Goals

AISI's mission is by necessity ambitious, and is made even more challenging by the rapidity and dynamism of a changing AI landscape.

As such, AISI intends to move at the "speed of relevance," while prioritizing technical excellence and scientific integrity, reflecting the importance and urgency of its mission.

To do this, AISI is structured to build out a portfolio of short and longer-term research efforts while seeding new projects that can rise to the challenges and complexity of the changing AI landscape. AISI empowers its team to undertake new projects in pursuit of much-needed outcomes for AI safety, including creating testing benchmarks for advanced models as well as developing guidelines for evaluating systemic risks. This agile project-based structure will help AISI keep pace with our changing world and align all projects towards one or more of AISI's strategic goals:

1. Advancing the science of AI safety.
2. Articulating, demonstrating, and disseminating the practices of AI safety.
3. Supporting institutions, communities, and coordination around AI safety.

These goals are clearly interconnected in important ways. Goal 1 will help provide the scientific research and understanding from which goal 2 will draw to develop and promote AI safety practices. Goal 2 will need the AI safety communities in goal 3 to help disseminate those practices. And goal 3 will be key for engaging, cultivating, and supporting AI safety communities that in turn inform and improve the research and practices of goals 1 and 2. AISI further anticipates that these goals will interact and co-evolve, improving together as AI safety matures as a science, a set of best practices, and an ecosystem.

Goal 1. Making the vision possible: advancing AI safety science through research, including testing, evaluation, validation, and verification of increasingly capable AI models, systems, and agents.

AISI will champion the development of empirically grounded tests, benchmarks, and evaluations of AI models, systems, and agents to find practical solutions for both near and long-term AI safety challenges. Accordingly, we aim to launch a range of projects under this goal, some of which may seek to:

- **Perform and coordinate technical research to improve or create needed safety guidelines and technical safety tools and techniques,** such as techniques for detection of synthetic content, best practices for model security, and technical safeguards and mitigations at the level of models, systems, and agents. These projects may involve both foundational and applied research. For its applied research projects, AISI intends to leverage in-house and external foundational research, as well as existing guidelines, methods, and standards.
- **Conduct pre-deployment TEVV of advanced models, systems, and agents to assess potential and emerging risks.** These projects will aim to assess what risks advanced AI systems might pose before being deployed or released, utilizing methodologies such as automated capability evaluations, expert red-teaming, A/B testing, and other methods. Working with the NIST labs we intend for this to include pre-deployment assessment of existing harms and potential and emerging risks, including to individual rights, public safety, and national security, such as enabling chemical, biological, or cyber attacks and risks to human oversight or control. These types of projects will also enable AISI to scientifically characterize emerging AI capabilities and risks in advanced models.
- **Conduct TEVV of advanced AI models, systems, and agents to develop scientific understanding and documentation of the range of existing risks.** Working with the NIST lab programs, we aim to deepen the scientific understanding of how to measure risks that pertain to present-day capabilities, including individual rights, public safety, and national security.

As part of these projects, we intend to collaborate across the NIST lab programs, with U.S. government agencies, international partners, and a diverse set of AI stakeholders to develop and conduct evaluations. We will seek to ensure that our projects, evaluations, and tools reflect the best available science. To facilitate close collaboration across government, we will serve as the primary U.S. government point of contact with advanced model developers for potential pre-deployment and post-deployment AI safety testing.

We intend to collaborate across the NIST lab programs, with U.S. government agencies, international partners, and a diverse set of AI stakeholders.

Goal 2. Making the vision actionable: developing and disseminating AI safety practices.

AI safety depends on developing the science, but also on the implementation of science-based practices. Accordingly, we intend to launch AISI projects that will:

- **Build and publish specific metrics, evaluation tools, methodological guidelines, protocols, and benchmarks for assessing risks of advanced AI across different domains and deployment contexts.** We plan to launch projects that publish guidelines and tools for TEVV on different kinds of risks for developers and deployers, including specific evaluation protocols for TEVV on a range of risks to inform and support developers, deployers, and third-party independent evaluators. These guidelines could provide recommendations as well as develop new benchmarks to assess model capabilities.
- **Develop and publish risk-based mitigation guidelines and safety mechanisms to support the responsible design, development, deployment, use, and governance of advanced AI models, systems, and agents.** We plan for these to include guidance on mitigation for existing harms as well as potential and emerging risks, including to public safety and national security; risk-proportionate safety and security mitigations for the most advanced AI systems; and internal and external safety mechanisms or tools developed from AISI research.

By leveraging our scientific research and development projects discussed in goal 1, we intend to provide stakeholders in the AI sector – from developers to evaluators to deployers to users – with the high-quality science-based information about risk evaluation and risk mitigation and the tools they need to make informed decisions.

AI safety depends on developing the science, but also on the implementation of science-based practices.

Goal 3. Making the vision sustainable: supporting institutions, communities, and coordination around AI safety.

The proliferation, scale, and increasing impact of AI systems demands a more integrated ecosystem of AI safety that includes diverse disciplines, perspectives, and experiences. The ecosystem would also benefit from third-party independent TEVV of AI models, systems, and agents. Accordingly, we intend to launch projects that will:

- **Promote adoption of AISI guidelines, evaluations, and recommended AI safety measures and risk mitigations.** To maximize the value and usability of AISI guidance, we intend to initiate and support ongoing dialogue, information-sharing, and collaboration, as appropriate, with safety research labs, third-party evaluators, and diverse expertise across developers, deployers, and users. Our projects will aim to operationalize voluntary commitments into actionable guidelines and promote the adoption of AI safety best practices. They will also seek to facilitate a robust ecosystem of third-party evaluators. AISI's projects may contribute to scientific reports, articles, guidance, and practices to help ensure that rigorous AI safety research, testing, and guidance inform major domestic AI safety legislation or policy. Our projects could also drive awareness of the AI safety practices developed to inform related research efforts outside of AISI, including NIST lab programs, the nation's R&D ecosystem, and AI stakeholders across the public and private sectors.
- **Lead an inclusive, international network on the science of AI safety.** AI safety practices must be globally adopted to the greatest extent possible. We intend to serve as a partner for other AI Safety Institutes, national research organizations, and multilateral entities like the OECD and G7. We intend to work with our partners to foster commonly accepted scientific methodologies with the intention of developing a shared and interoperable suite of AI safety evaluations and agreed-upon risk mitigations. In doing so, we hope to help develop the science and practices that underpin future arrangements for international AI governance.

AISI will also leverage the expertise of stakeholders in the AISI Consortium to identify and help develop community consensus around best practices, to help inform future AISI projects, and to help promote the adoption of AI safety recommendations and capabilities. Our goal is to ensure that AISI's impact is sustained, domestically and internationally, through the adoption of AI safety norms and a robust ecosystem for AI safety.

AI safety practices must be globally adopted to the greatest extent possible.

Pursuing AISI's vision for the benefit of all.

Mitigating the safety risks of a transformative technology is necessary to harness its benefits.

This calls for institutions that conduct scientific research, articulate practices and protocols, and convene communities to promote safe AI models and systems. The U.S. AI Safety Institute is a key part of the federal government's answer to this call. Working closely with diverse stakeholders in the AI landscape, AISI seeks to develop science-based metrics, tools, and guidelines for both government and the public to use to evaluate AI models and systems, mitigate identified risks, and promote safe AI innovation. AISI's leadership in the science of AI safety will catalyze domestic and international ecosystems, helping to ensure AI's potential is safely unlocked for ourselves and for generations to come.

About AI at NIST

The National Institute of Standards and Technology (NIST) develops measurements, technology, tools, and standards to advance reliable, safe, transparent, explainable, privacy-enhanced, and fair artificial intelligence (AI) so that its full commercial and societal benefits can be realized without harm to people or the planet. NIST, which has conducted both fundamental and applied work on AI for more than a decade, is also helping to fulfill the 2023 Executive Order on Safe, Secure, and Trustworthy AI. NIST established the U.S. AI Safety Institute and the companion AI Safety Institute Consortium to continue the efforts set in motion by the E.O. to build the science necessary for safe, secure, and trustworthy development and use of AI.

https://www.nist.gov