AI Innovation Lab

Elham Tabassi, Chief AI Advisor

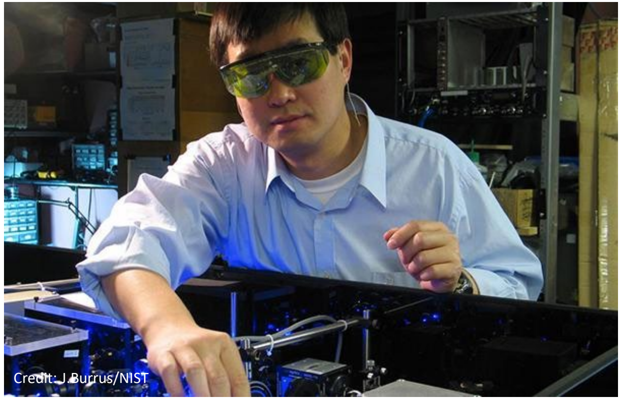
June 11, 2024

NATIONAL INSTITUTE OF
STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST’s Mission

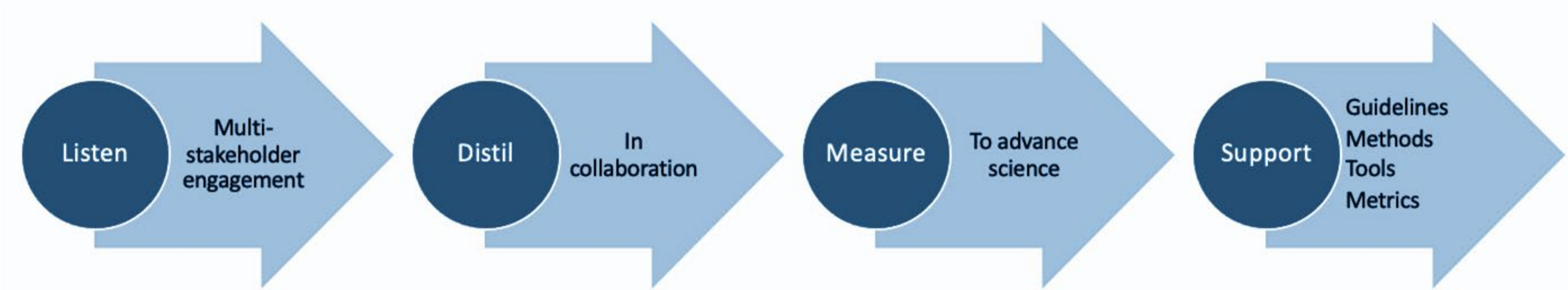
To promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life

©Robert Rathe


Credit: J.Burrus/NIST

©Nicholas McIntosh Photography

NIST helps industry develop valid, scientifically rigorous methods, metrics and standards.



Listen

Multi-stakeholder engagement

Distil

In collaboration

Measure

To advance science

Support

Guidelines
Methods
Tools
Metrics

NIST AI Risk Management Framework and Resources

AI Risk Management Framework Concept Paper

13 December 2021

NOTE TO REVIEWERS

This Artificial Intelligence Risk Management Framework (AI RMF) concept paper incorporates input from the Notice of Request for Information (RFI) released by the National Institute of Standards and Technology (NIST) on July 29, 2021[1], and discussions during the workshop, "Kicking off NIST AI Risk Management Framework," held October 19-21, 2021.[2]

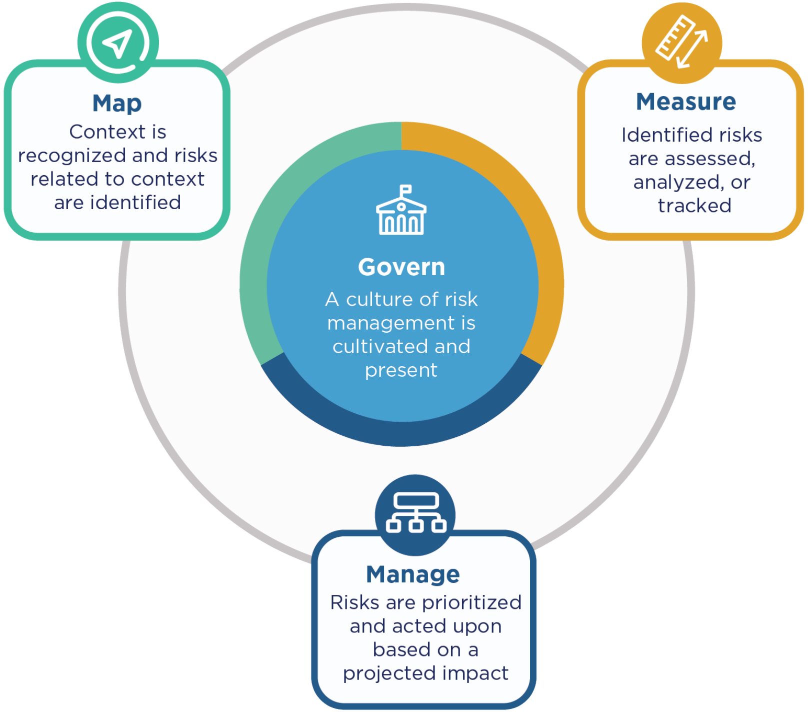
Feedback on this paper will inform further development of this approach and the first draft of the AI RMF

AI RMF PLAYBOOK

NIST AI 100-1

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

NIST AI RMF: A voluntary resource for organizations designing, developing, deploying, or using AI systems to manage AI risks and promote trustworthy and responsible AI



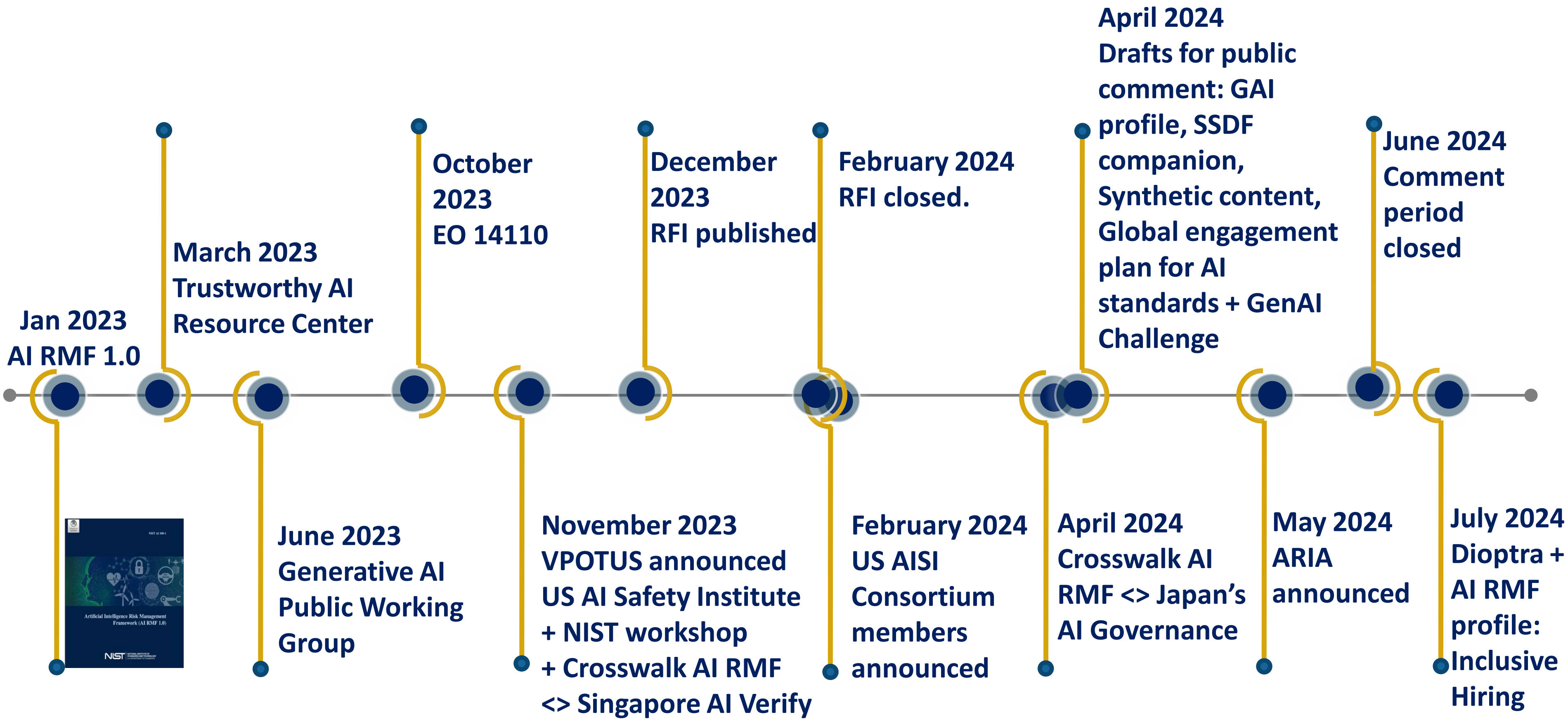
Map
Context is recognized and risks related to context are identified

Measure
Identified risks are assessed, analyzed, or tracked

Govern
A culture of risk management is cultivated and present

Manage
Risks are prioritized and acted upon based on a projected impact

Major Achievements and Announcements Since 2023



NIST AI 100-1
Artificial Intelligence Risk Management Framework (AI RMF 1.0)
NIST

Jan 2023
AI RMF 1.0

March 2023
Trustworthy AI
Resource Center

June 2023
Generative AI
Public Working
Group

October
2023
EO 14110

November 2023
VPOTUS announced
US AI Safety Institute
+ NIST workshop
+ Crosswalk AI RMF
<> Singapore AI Verify

December
2023
RFI published

February 2024
RFI closed.

February 2024
US AISI
Consortium
members
announced

April 2024
Crosswalk AI
RMF <> Japan's
AI Governance

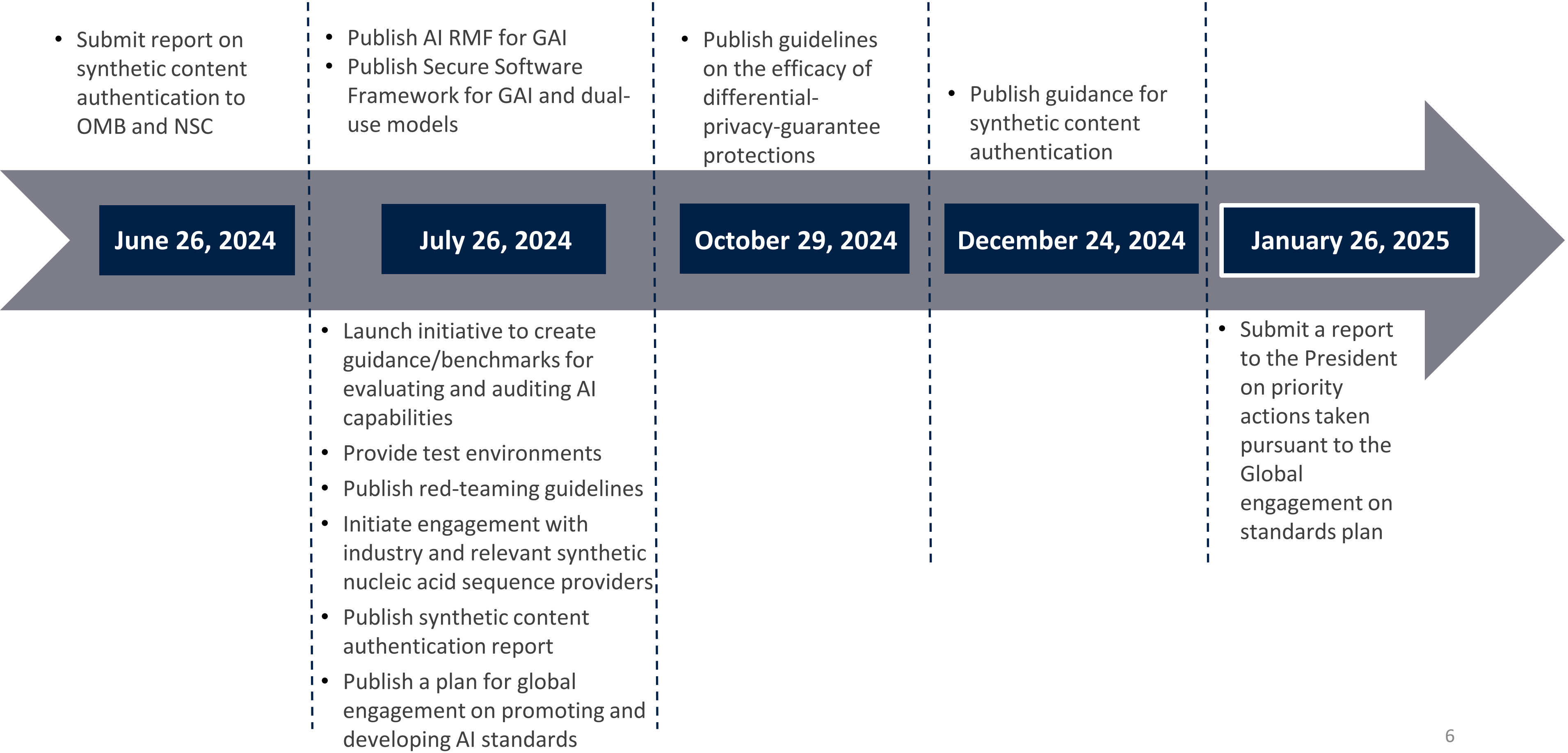
April 2024
Drafts for public
comment: GAI
profile, SSDF
companion,
Synthetic content,
Global engagement
plan for AI
standards + GenAI
Challenge

May 2024
ARIA
announced

June 2024
Comment
period
closed

July 2024
Dioptra +
AI RMF
profile:
Inclusive
Hiring

NIST Due Dates Under Executive Order 14100



June 26, 2024

- Submit report on synthetic content authentication to OMB and NSC

July 26, 2024

- Publish AI RMF for GAI
- Publish Secure Software Framework for GAI and dual-use models
- Launch initiative to create guidance/benchmarks for evaluating and auditing AI capabilities
- Provide test environments
- Publish red-teaming guidelines
- Initiate engagement with industry and relevant synthetic nucleic acid sequence providers
- Publish synthetic content authentication report
- Publish a plan for global engagement on promoting and developing AI standards

October 29, 2024

- Publish guidelines on the efficacy of differential-privacy-guarantee protections

December 24, 2024

- Publish guidance for synthetic content authentication

January 26, 2025

- Submit a report to the President on priority actions taken pursuant to the Global engagement on standards plan

Artificial Intelligence Safety Institute Consortium (AISIC)

AISIC brings more than 280 leading AI stakeholders together to develop science-based and empirically backed guidelines and standards for AI measurement and policy, laying the foundation for AI safety across the world.

AISIC working groups sustain, scale, and implement E.O. elements

- Risk Management for Generative AI
- Synthetic Content
- Capability Evaluations
- Red-Teaming
- Safety & Security

ARIA

Assessing Risks and Impacts of AI

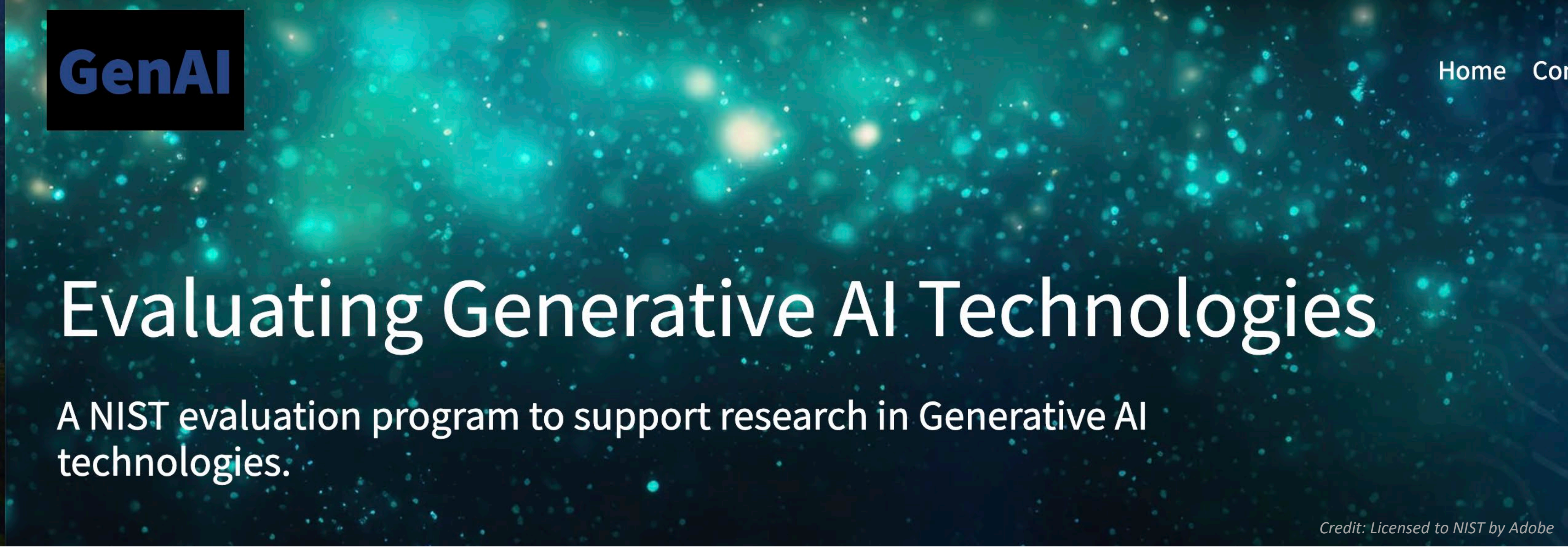
A compelling set of scenarios will aim to explore risks and related impacts across three levels of testing: model testing, red-teaming, and field testing.

Credit: Licensed to NIST by Adobe

Team: **Reva Schwartz (Lead)**

Jonathan Fiscus, Kristen K. Greene, Craig Greenberg, Afzal Godil, Kyra Yee, Razvan Amironesei, Theodore Jensen (Feds)

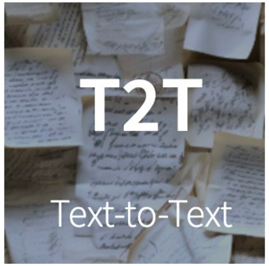
Rumman Chowdhury (associate), Gabriella Waters (PREP), Patrick Hall (associate), Shomik Jain (pathway),



Evaluating Generative AI Technologies

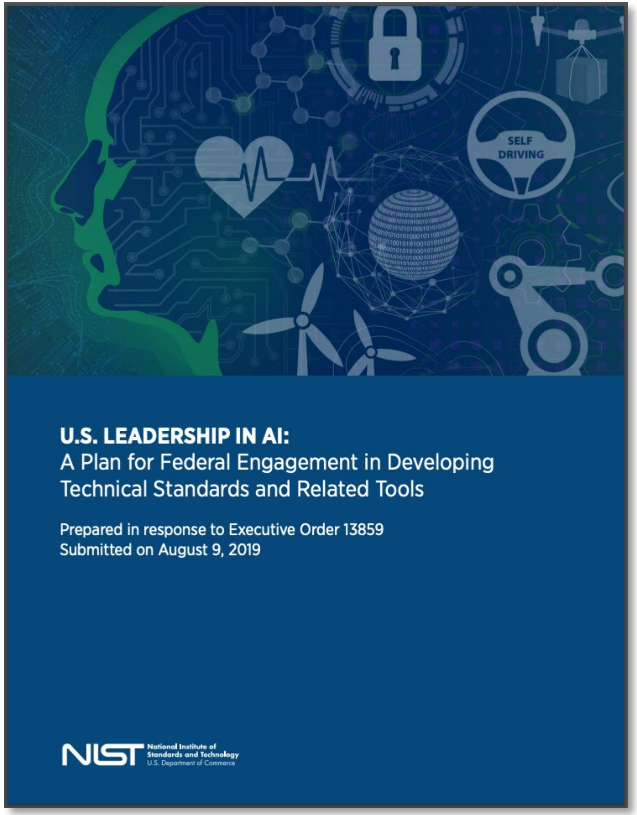
A NIST evaluation program to support research in Generative AI technologies.

Credit: Licensed to NIST by Adobe



Team: **Yooyoung Lee (Lead)**
George Awad, Kay Peterson, Peter Fontana (Feds)
Lowen DiPaula (pathway student),
Seungmin Seo (associate)

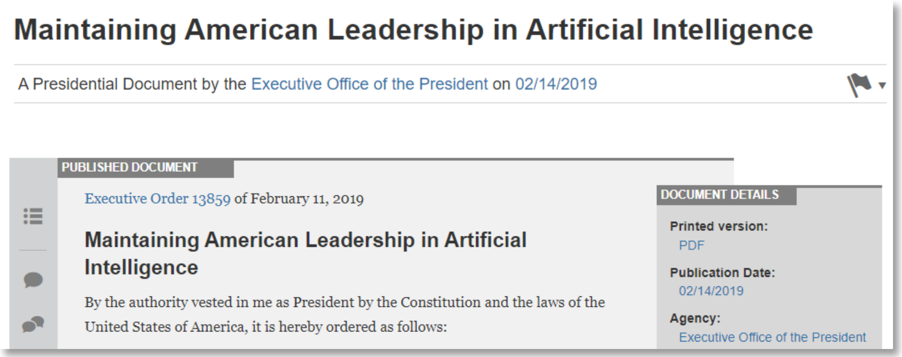
USG AI Standards Coordinator



U.S. LEADERSHIP IN AI:
A Plan for Federal Engagement in Developing Technical Standards and Related Tools

Prepared in response to Executive Order 13859
Submitted on August 9, 2019

NIST National Institute of Standards and Technology U.S. Department of Commerce



Maintaining American Leadership in Artificial Intelligence

A Presidential Document by the Executive Office of the President on 02/14/2019

PUBLISHED DOCUMENT

Executive Order 13859 of February 11, 2019

Maintaining American Leadership in Artificial Intelligence

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

DOCUMENT DETAILS

Printed version:
PDF

Publication Date:
02/14/2019

Agency:
Executive Office of the President

NIST works across government and with industry to identify critical standards development activities, strategies, and gaps

NIST is coordinating in part through the Interagency Committee on Standards Policy (ICSP) AI Standards Coordination Working Group

Global Engagements

Credit: OECD

ISO
JTC 1
IEC
INFORMATION TECHNOLOGY STANDARDS
SC 42 – Artificial Intelligence

Credit: ISO/IEC SC 42

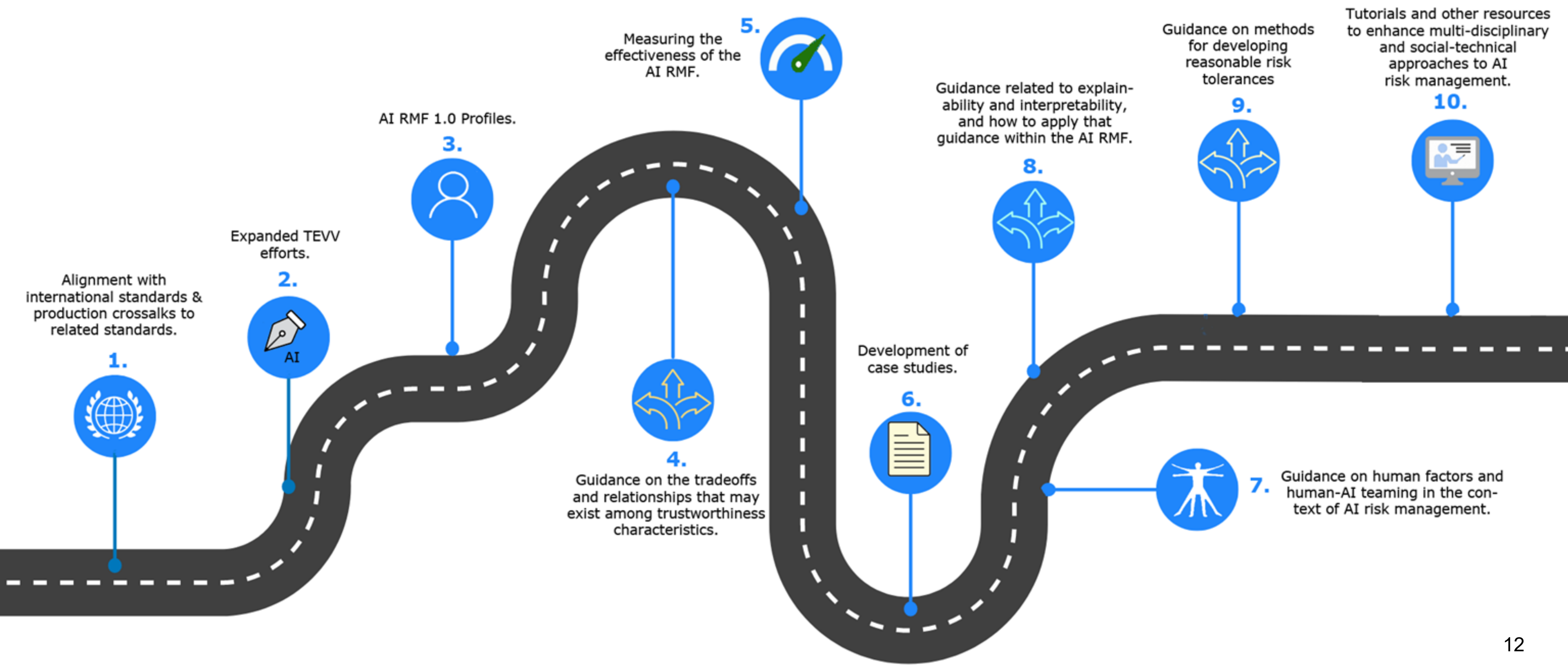
Credit: USembassy.gov

AI Risk Management Framework
Govern
Map
Measure
Manage

Internationalizing
AI RMF

Credit: NIST

A roadmap of future work was released along with the AI RMF in January.



1. Alignment with international standards & production crossalks to related standards.

2. Expanded TEVV efforts.

3. AI RMF 1.0 Profiles.

4. Guidance on the tradeoffs and relationships that may exist among trustworthiness characteristics.

5. Measuring the effectiveness of the AI RMF.

6. Development of case studies.

7. Guidance on human factors and human-AI teaming in the context of AI risk management.

8. Guidance related to explainability and interpretability, and how to apply that guidance within the AI RMF.

9. Guidance on methods for developing reasonable risk tolerances

10. Tutorials and other resources to enhance multi-disciplinary and social-technical approaches to AI risk management.

NATIONAL ARTIFICIAL INTELLIGENCE ADVISORY COMMITTEE (NAIAC)

The National Artificial Intelligence Advisory Committee (NAIAC) advises the President and the White House on the intersection of AI and innovation, competition, societal issues, the economy, law, international relations, and other critical areas.

Since first convening in May 2022

- 70+ experts interviewed across 26 public sessions
- 18 Recommendation reports/memos
- 6 Findings, including explainer documents and FAQs
- 2 Committee Statements
- 2 Annual Reports

The Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence operationalized many of NAIAC's recommendations.

WHAT'S AHEAD

NAIAC will continue to offer recommendations and insights as a committee, in public stakeholder panels and in the activity across five working groups:

AI Education & Awareness
International Collaboration
AI Futures – Preparedness, Opportunities, and Competitiveness
Safety, Trust, and Rights
AI in Work and the Workforce

Additionally, the NAIAC Subcommittee on AI and Law Enforcement has introduced three distinct working groups:

Performance, Evaluation and Bias; Processes; Identification and Surveillance Set

Click, Connect, Collaborate!

www.nist.gov/itl/ai-risk-management-framework

airc.nist.gov

ai-challenges.nist.gov

ai.gov/naiac

AIFramework@nist.gov
ai-inquiries@nist.gov

U.S. AI Safety Institute

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

Elizabeth Kelly, Director of the U.S. AI Safety Institute
June 11, 2024

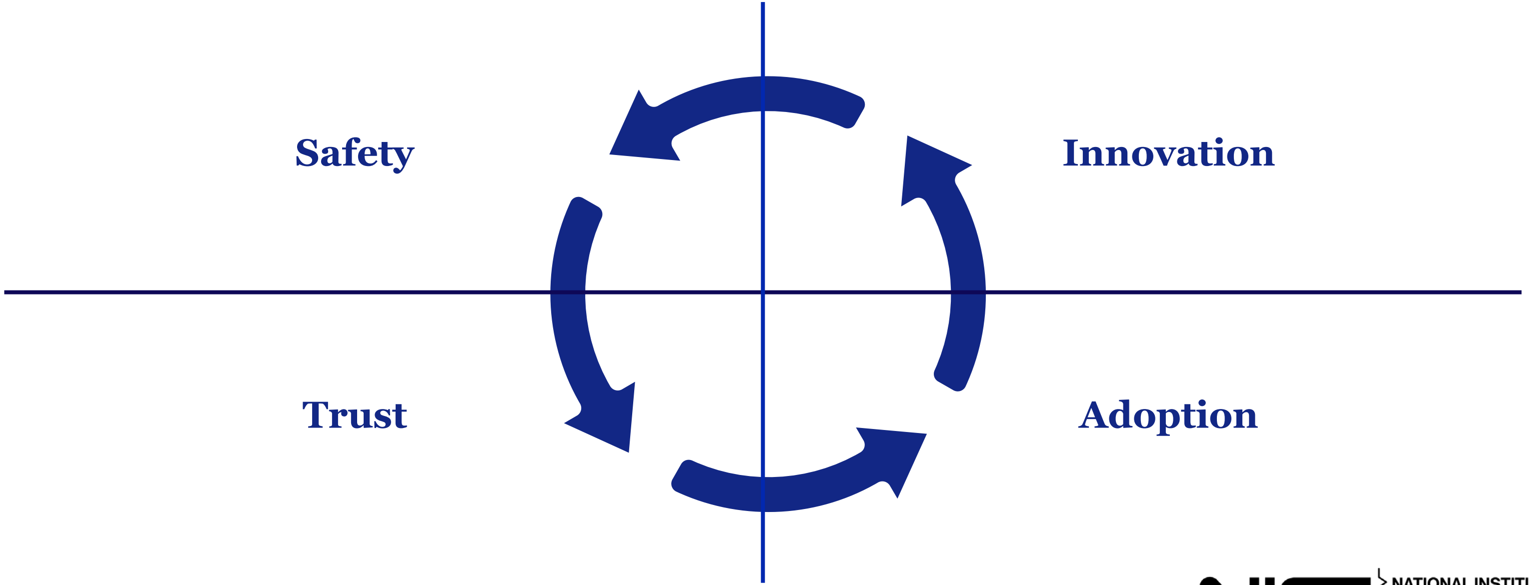
NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

OUR MISSION

To advance the science of AI safety, and at the same time advance the implementation and adoption of that science.

OUR BELIEF

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



Safety

Innovation

Trust

Adoption

OUR PLAN

Testing

- Establish a new government capability to directly test frontier models
- Initial focus on capabilities that could pose a threat to national security
- Share feedback with model developers

Guidance

- Catalyze a robust ecosystem of independent research and evaluation
- Develop guidance on testing and safety and security mitigations
- Create guidance on synthetic content detection

Research

- Conduct fundamental technical research on topics such as model interpretability
- Research and develop new and improved risk mitigations

OUR TEAM

Headquartered at NIST, the nation’s premier measurement institute with an impressive track record on AI, including the AI Risk Management Framework.

Top-notch talent, drawn from frontier labs, academia, and government. Our team connects Silicon Valley to the U.S. Capital, housed in San Francisco and Washington, DC offices.

Supported by the AISI Consortium, a group of over 280 organizations from academia, civil society, and industry at the frontier of AI safety.

INTERNATIONAL AI SAFETY NETWORK

Information and Research Exchange

Aligned Evaluations

Shared Guidance on Testing and Risk Mitigations

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Credit: Canva

We plan to host an international AI safety convening this fall in San Francisco.

More details forthcoming.

Credit: Canva