A USAISI Workshop: Collaboration to Enable Safe and Trustworthy AI

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

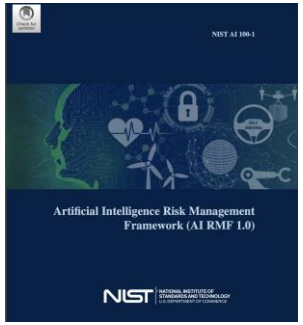
US AI Safety Institute Consortium

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

Elham Tabassi
Chief AI Advisor
elham.tabassi@nist.gov

Major achievements and announcements in 2023

Jan 2023
AI RMF 1.0



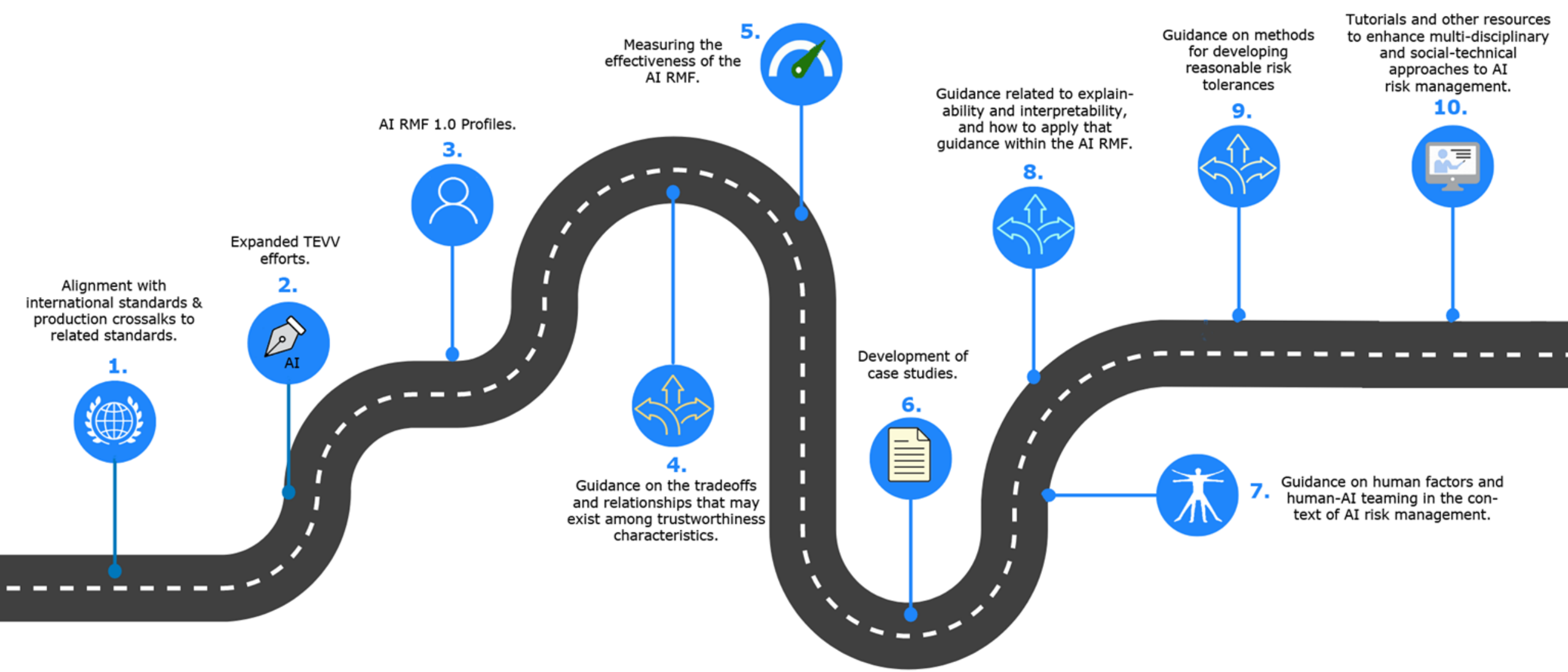
March 2023
Trustworthy AI Resource Center

June 2023
Generative AI Public Working Group

October 2023
EO 14110

November 2023
US AI Safety Institute Announced

Roadmap of future activities released in January 2023.



Building and maturing a measurement science for trustworthy and responsible AI.

Build and expand the science of AI evaluation

Create test environments

Develop guidelines and standards

Conduct evaluations

For more information, we encourage you to access NIST resources, or reach out directly!

www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute

airc.nist.gov

usaisi@nist.gov

US AI Safety Institute Consortium

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

Jake Taylor
Senior Advisor for Critical and Emerging Technologies
jacob.taylor@nist.gov

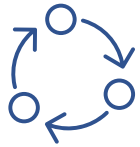
US AI Safety Institute: Three Pillars

Research

GOAL: Improve the science of AI safety

Example activities

- Fundamental research
- Technical building blocks
- Content authentication guidance and best practices



Implementation

GOAL: Guide implementation of scientific findings, tests, and risk management frameworks

Example activities

- Metrics and methodologies
- Development of testbeds
- Evaluations and red-teaming
- Use-case-specific risk management "profiles"

Consortium

GOAL: Drive work and oversee research collaboration with partners

Example activities

- Working groups
- Scientific collaborations
- Shared guidance
- Draft standards
- Aligned approaches

NIST Consortia: Public-Private Partnerships to Address Pre-competitive Challenges

NIST GENOME IN A BOTTLE (GIAB) CONSORTIUM

Provides authoritative characterization of benchmark human genomes

NIST GENOME EDITING CONSORTIUM

Develops measurement solutions and standards needed to increase confidence and reduce risk

NIST FLOW CYTOMETRY STANDARDS CONSORTIUM

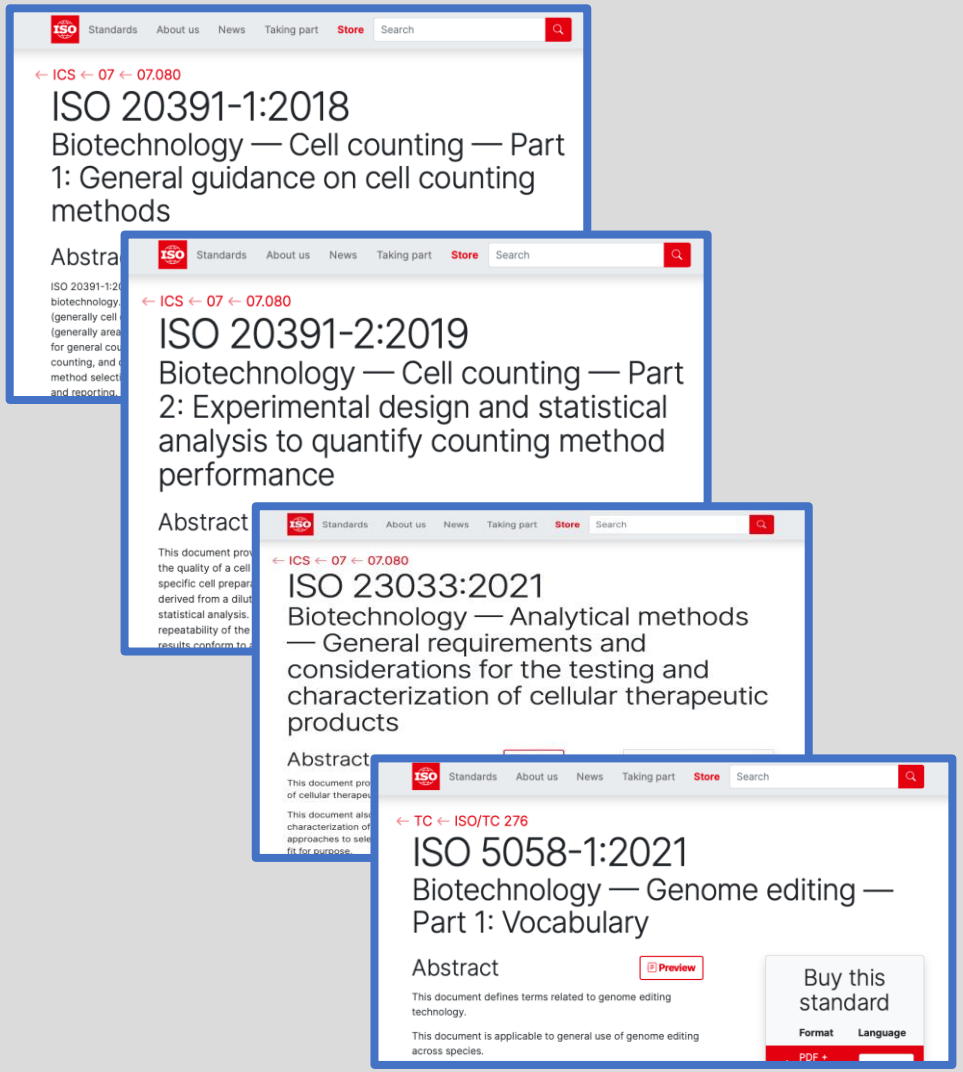
Accelerates the adoption of quantitative flow cytometry in biomanufacturing

NIST RAPID MICROBIAL TESTING METHODS CONSORTIUM

Addresses measurements and standards needed to increase confidence in the use of rapid testing

Global Documentary Standards Leadership: Example from Biotechnology

NIST-led Standards



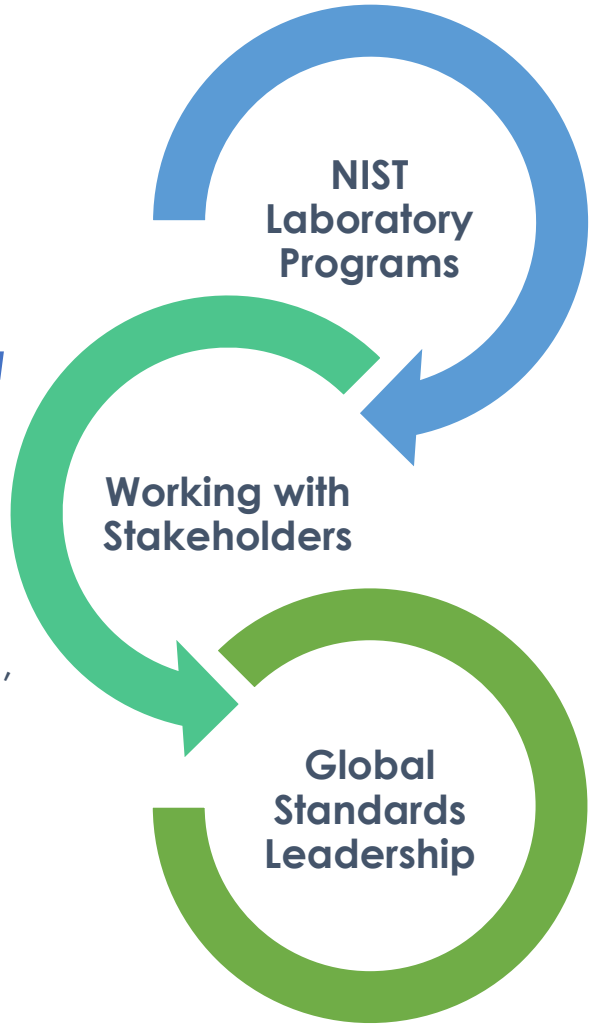
NIST Consortia

biofabusa

NIMBL

STANDARDS COORDINATING BODY
REGENERATIVE MEDICINE

Industry, USG, NGOs, etc.



- *Advanced metrology*
- *DBTL*
- *Standards*

ISO

ANSI
American National Standards Institute

NIST Reference Materials: Calibration, traceability, comparability

Why a consortium?

- Enable a shared space for complex conversations and research collaboration
  - Where possible, take an open and transparent approach

- Bring together actors from public and private sector and strategic international partners
  - Interested parties can work together in building and maturing a new measurement science for trustworthy AI

- Work together
  - to align capability evaluation and red-teaming guidance;
  - to enable AI testbeds and test environments;
  - to build the foundations and documentation for standards

What is the approach?

Send in your letter of interest!

https://www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute

NIST-led consortia built from a cooperative research and development agreement – everyone signs the same one

Illustrative example:
https://www.nccoe.nist.gov/publications/other/nccoe-consortium-crada-example

Initial activities guided through working groups – discussion after the break!

Driven from experiences with the Generative AI Public Working Group

What can we do together?

Guidance	Develop guidance and benchmarks for identifying and evaluating AI capabilities, with a focus on capabilities that could potentially cause harm
Security	Develop approaches to incorporate secure-development practices for generative AI, including special considerations for dual-use foundation models
Testing	Develop and ensure the availability of testing environments
Red-teaming	Develop guidance, methods, skills and practices for successful red-teaming and privacy-preserving machine learning
Tools	Develop guidance and tools for authenticating digital content
Workforce	Develop guidance and criteria for AI workforce skills, including risk identification and management, test, evaluation, validation, and verification (TEVV), and domain-specific expertise
Society	Explore the complexities at the intersection of society and technology, including the science of how humans make sense of and engage with AI in different contexts
Lifecycle	Develop guidance for understanding and managing the interdependencies between and among AI actors along the lifecycle

US AI Safety Institute Consortium

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

Kamie Roberts, NIST AI Executive Order Program Manager
(kathleen.roberts@nist.gov)
Reva Schwartz, Research Scientist, Principal Investigator for AI Bias,
(reva.schwartz@nist.gov)

Foundation for Success

Build the foundation for sustained and continuous efforts to create safe and trustworthy AI

- Advance research
- Facilitate consensus standards
- Create test environments
- Enable evaluations to assess the risk and impact of current and next generation AI on individuals and society

Working Groups: Role and Outputs

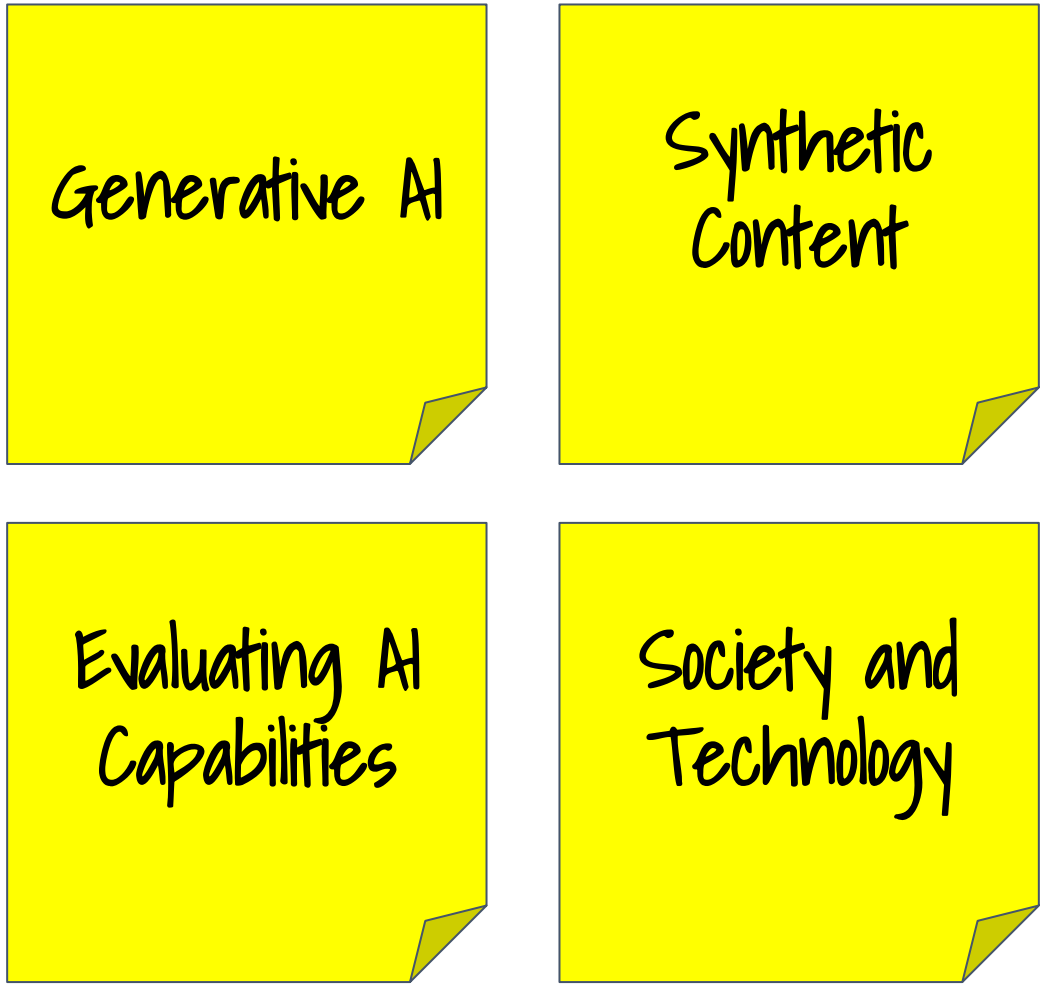
Support development of Safe and Trustworthy AI through

- Convening space for dialogue and information sharing
- Collaborative R&D
- Evaluations of test systems and prototypes

Outputs

- New guidelines, tools, methods, protocols and best practices
- Guidance and benchmarks for AI capabilities - particularly those that may cause harm
- Availability of testing environments

Proposed Working Groups



Proposed Working Groups

- Generative AI
  - Risks and capabilities
- Synthetic content
  - Authentication, detection, labeling
  - Deep fakes
- Evaluating AI capabilities
  - AI red-teaming and other testing methodologies
  - Pre-deployment testing and post-deployment monitoring
- Society and Technology
  - Standards
  - Operationalize the AI RMF