CYB-4203/6203

Secure and Trustworthy AI

Presentation 8: Governance, Documentation, and Auditability Across the AI/ML Pipeline

Wednesday, February 18, 2026

Topics: 4.3, 4.4

Today's Agenda

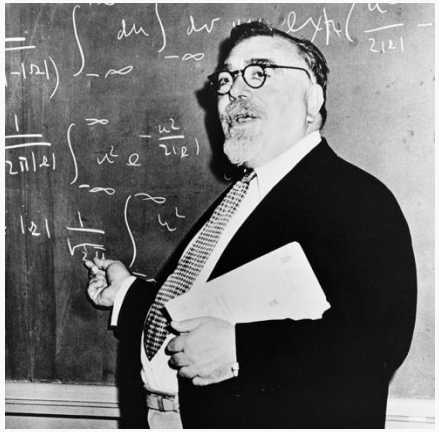
- 4.3: Organizational governance frameworks: NIST, ISO, MITRE, industry best practices
- 4.4: Documentation and auditability: model cards, datasheets, transparency reporting

4.3: Organizational AI Governance
4.4: Documentation & Auditability Across the AI/ML Pipeline

What is Governance? (an etymological-evolutionary excursion)

Ancient Greek
kybernan: 'To steer or pilot a ship.'

Ancient Latin
gubernare: 'To direct, rule, or govern.'



Mid-1900's Norbert Weiner
cybernetics: 'The scientific study of control and communication in the animal and the machine.'

- Control & communication
- Animal & machine
- Feedback mechanisms
- Homeostasis

Late 20th/early 21st Century Arnold
- 1993: "I'm a *cybernetic* organism. Living tissue over a metal endoskeleton."
- 2003: Elected *Governator* of California

Coincidence?

What is Governance?

In general: Rules and processes that coordinate behavior.

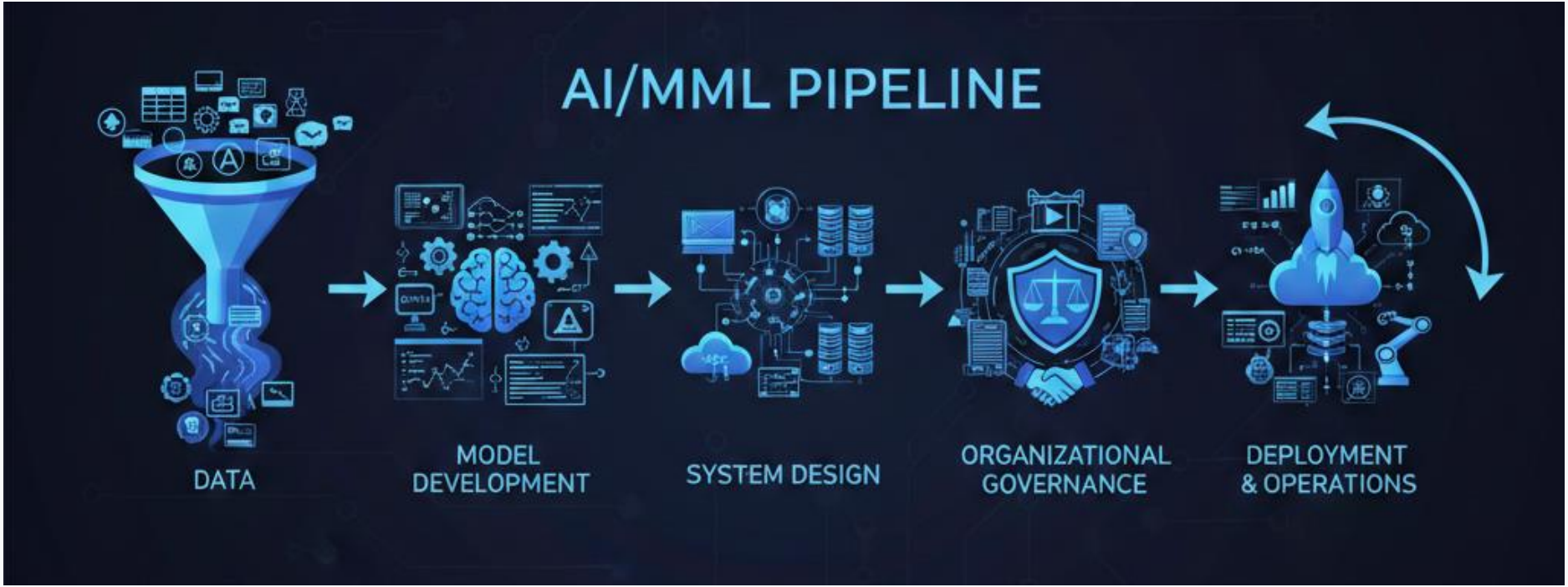
Not just what governments do; defined more broadly as '*the process through which some activity is organized, coordinated, steered, and managed*.'

Includes the norms, policies, and institutions that influence stakeholders' actions to achieve socially desirable outcomes.

Examples

- **Healthcare:** Patient care norms, professional ethical standards, licensing organizations…
- **Financial services**: Fiduciary duty norms, Basel III capital requirements, FDIC and SEC
- **Aviation**: Crew safety culture, FAA airworthiness certifications, mandatory maintenance schedules, NTSB for independent accident investigation

The AI/ML Pipeline (Nano Banana typo intentionally retained)



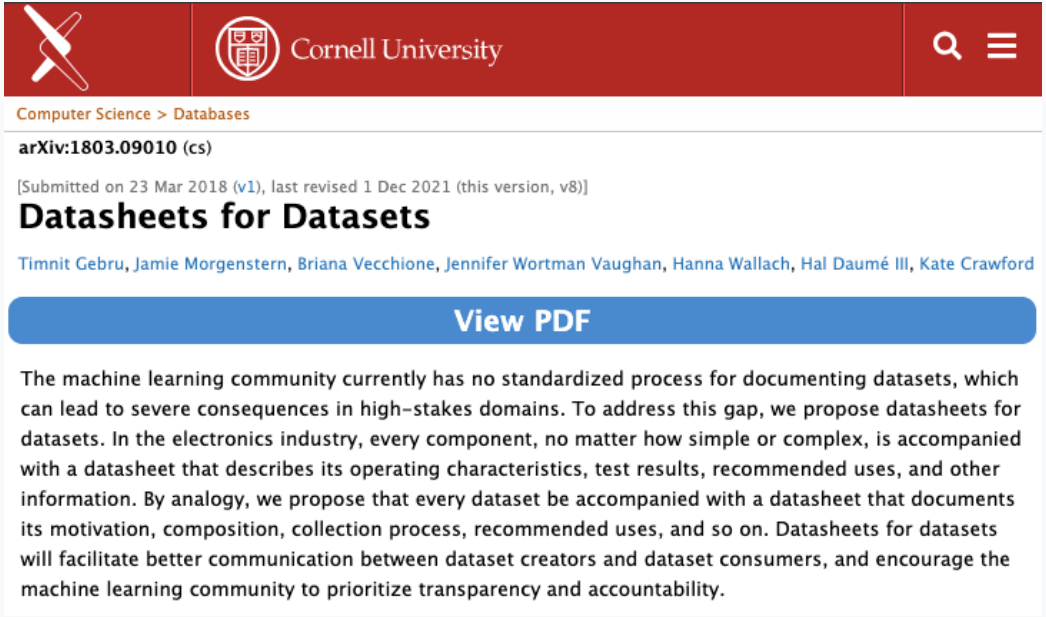
Data Governance

Datasheets for Datasets

Analogous to electronics datasheets: documents a dataset's motivation, composition, collection process, preprocessing, recommended uses, and distribution

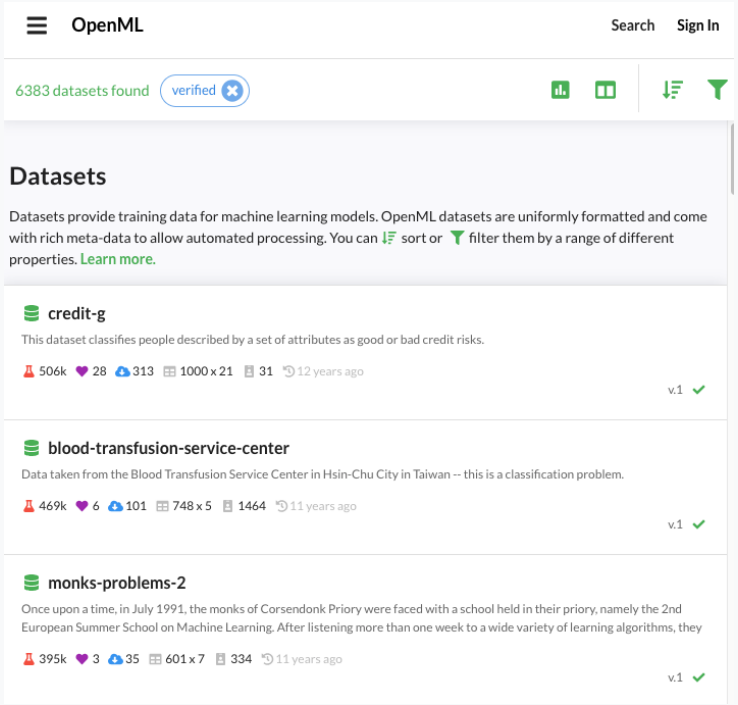
Targets the **data collection/curation stage** of the pipeline — the earliest point where bias and provenance issues can be caught and documented

Influential on EU AI Act's data governance requirements (Article 10) and on NIST AI RMF's MAP function for data characterization

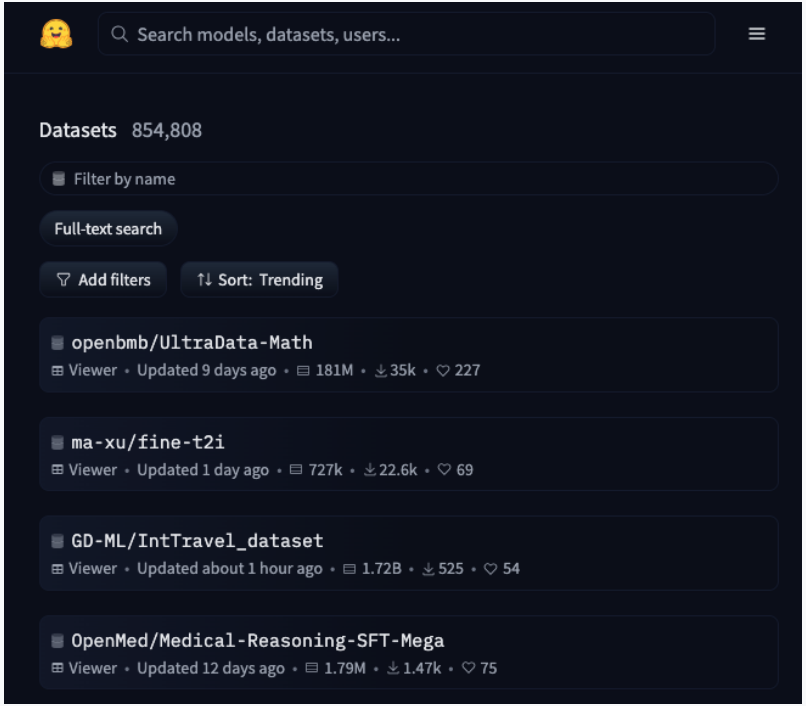


https://arxiv.org/abs/1803.09010

Datasheets for Datasets



https://www.openml.org



https://huggingface.co/datasets

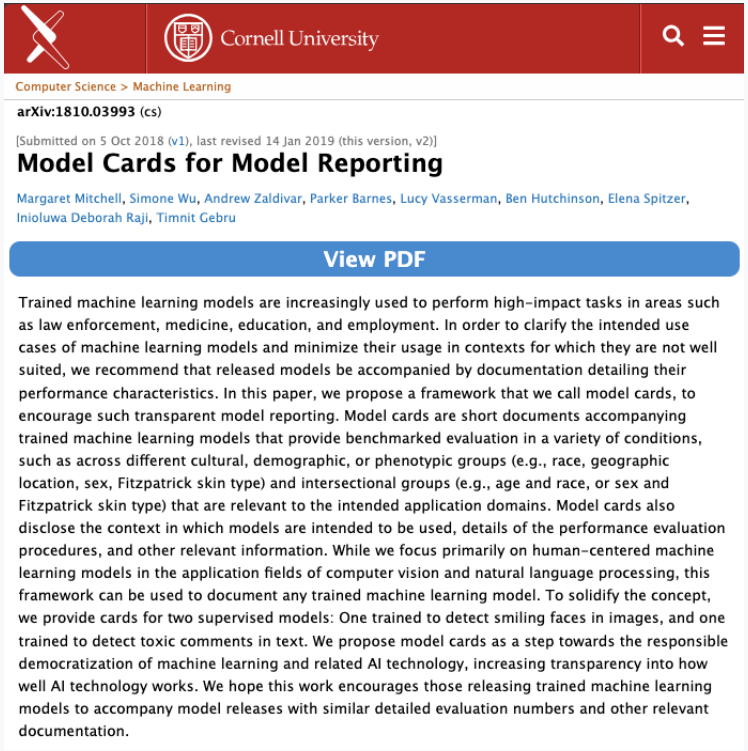
Model Development Governance

Model Cards for Model Reporting

Proposes short, standardized documents accompanying trained ML models that disclose intended use, performance benchmarks across demographic groups, and limitations

Directly addresses the **model developer → deployer handoff** — the doc artifact that lets downstream users make informed decisions about fitness for purpose

Now widely adopted: Hugging Face has Model Cards built into every model repo; Google, Meta, and OpenAI publish variants



Cornell University

Computer Science > Machine Learning

arXiv:1810.03993 (cs)

[Submitted on 5 Oct 2018 (v1), last revised 14 Jan 2019 (this version, v2)]

Model Cards for Model Reporting

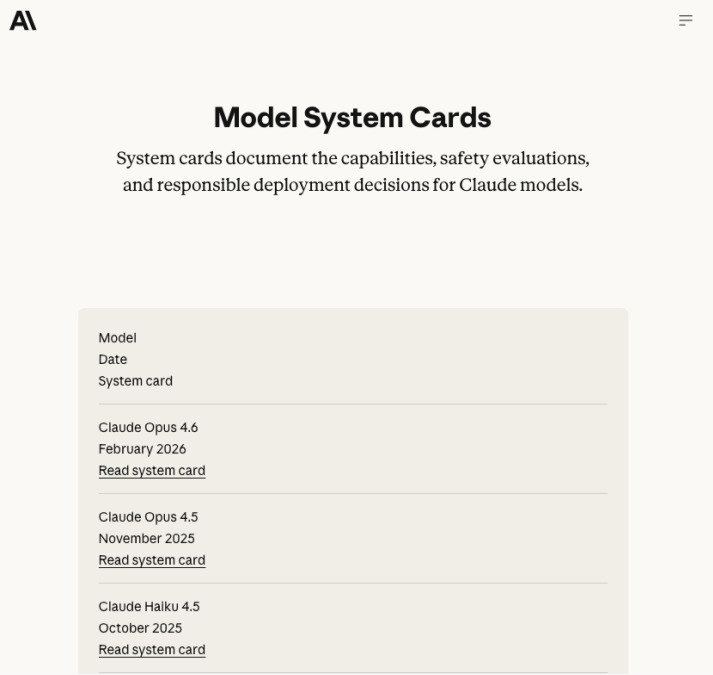
Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru

View PDF

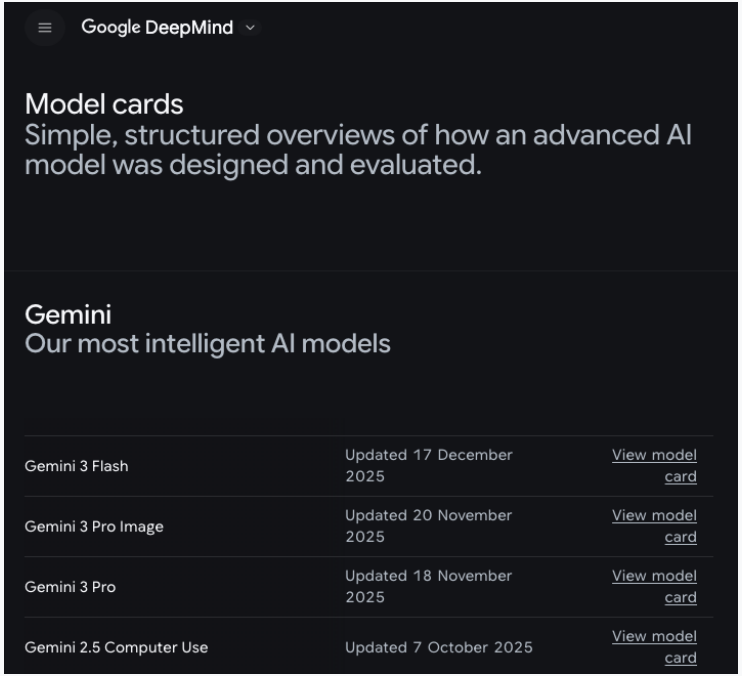
Trained machine learning models are increasingly used to perform high-impact tasks in areas such as law enforcement, medicine, education, and employment. In order to clarify the intended use cases of machine learning models and minimize their usage in contexts for which they are not well suited, we recommend that released models be accompanied by documentation detailing their performance characteristics. In this paper, we propose a framework that we call model cards, to encourage such transparent model reporting. Model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex, Fitzpatrick skin type) and intersectional groups (e.g., age and race, or sex and Fitzpatrick skin type) that are relevant to the intended application domains. Model cards also disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information. While we focus primarily on human-centered machine learning models in the application fields of computer vision and natural language processing, this framework can be used to document any trained machine learning model. To solidify the concept, we provide cards for two supervised models: One trained to detect smiling faces in images, and one trained to detect toxic comments in text. We propose model cards as a step towards the responsible democratization of machine learning and related AI technology, increasing transparency into how well AI technology works. We hope this work encourages those releasing trained machine learning models to accompany model releases with similar detailed evaluation numbers and other relevant documentation.

https://arxiv.org/abs/1810.03993

Model Card Examples

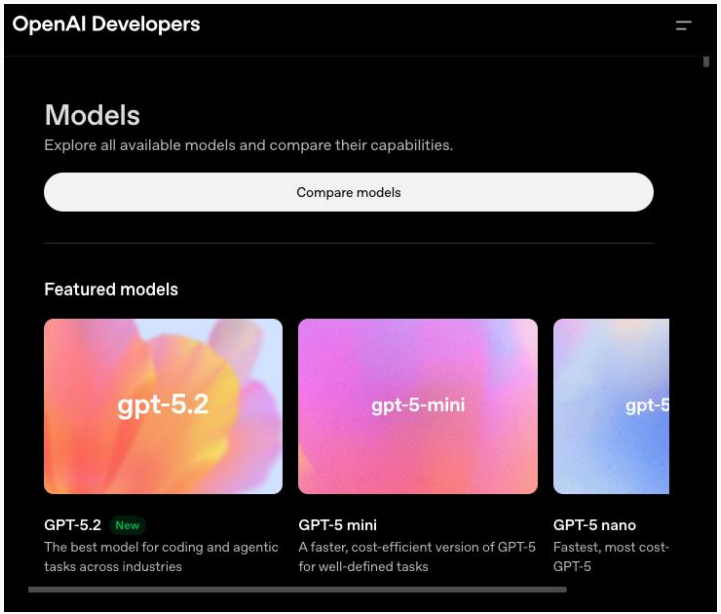


Anthropic System Cards

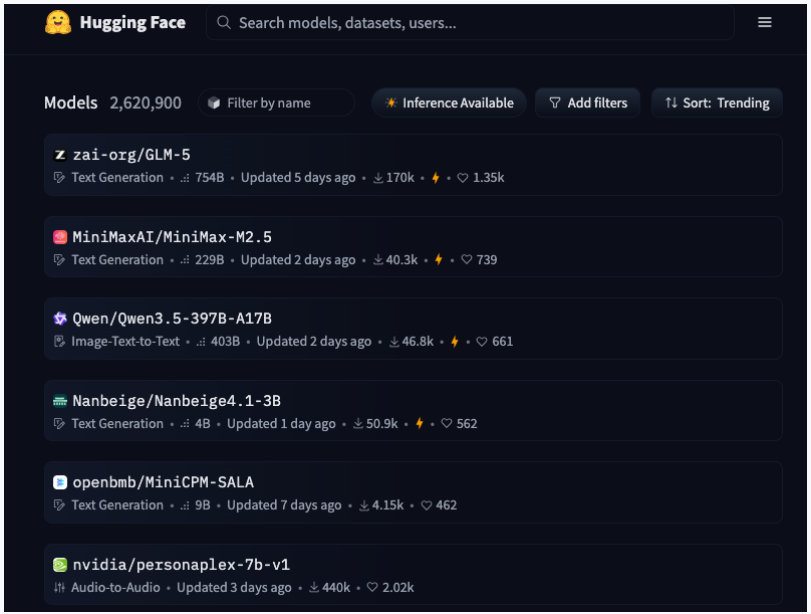


Google Model Cards

Model Card Examples



[OpenAI Model Cards]



[Hugging Face Models]

System Design Governance

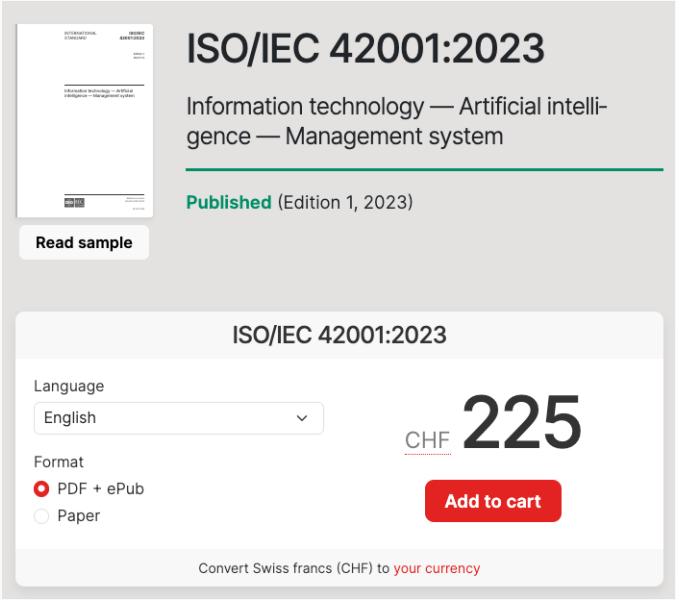
ISO/IEC 42001:2023

International Organization for Standardization / International Electrotechnical Commission

First international **certifiable** standard for AI management systems — organizations can get audited and certified against it, which makes it a *compliance differentiator*

Covers AI policy, risk assessment, data governance, and continuous improvement

Referenced by the EU AI Act as a way to demonstrate conformity



https://www.iso.org/standard/42001

IEEE 7000-2021 – Ethical System Design Process

Establishes a **model process for addressing ethical concerns during system design** — not just AI-specific but highly relevant to AI systems

Provides traceability from ethical values through concept of operations → value propositions → ethical requirements → risk-based design decisions

Unique angle: a *process standard* focused on stakeholder elicitation and value prioritization, not just a checklist — good for orgs building AI products from scratch

https://standards.ieee.org/ieee/7000/6781/

Organizational Governance

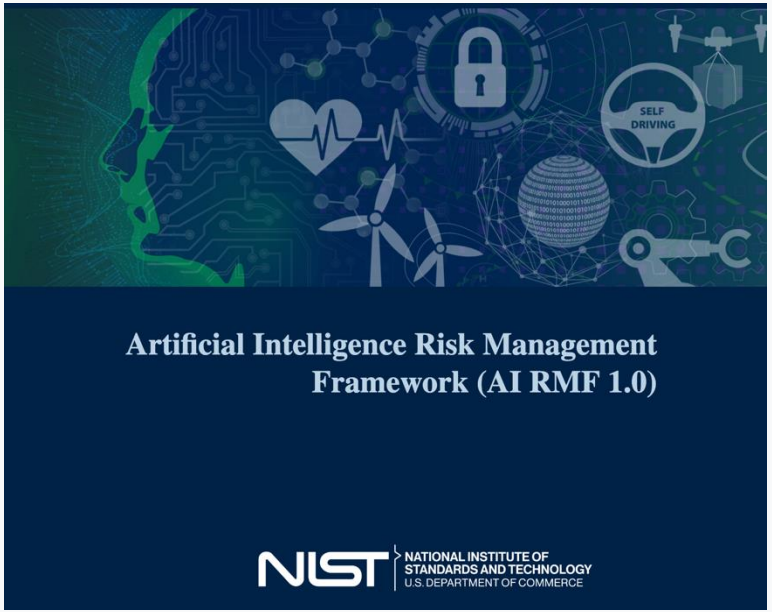
NIST AI Risk Management Framework (AI RMF 1.0)

‘The de facto US framework’

Remains the de facto US framework even after EO 14110 was rescinded in Jan 2025; widely adopted by industry as a baseline

Structured around four core functions — **Govern, Map, Measure, Manage** — providing a voluntary, flexible lifecycle approach to identifying and mitigating AI risks

Comes with a companion **Playbook** with suggested actions per sub-category, plus crosswalk documents mapping to other frameworks (ISO, OECD, EU AI Act)



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

NIST AI Risk Management Framework (AI RMF 1.0)

Four Core Functions

Govern - Policies, processes, procedures and practices across the organization related to the mapping, measuring and managing of AI risks are in place, transparent, and implemented effectively.

Map - Context is established and understood.

Measure - Appropriate methods and metrics are identified and applied.

Manage - AI risks based on assessments and other analytical output from the Map and Measure functions are prioritized, responded to, and managed.

https://airc.nist.gov/airmf-resources/playbook/

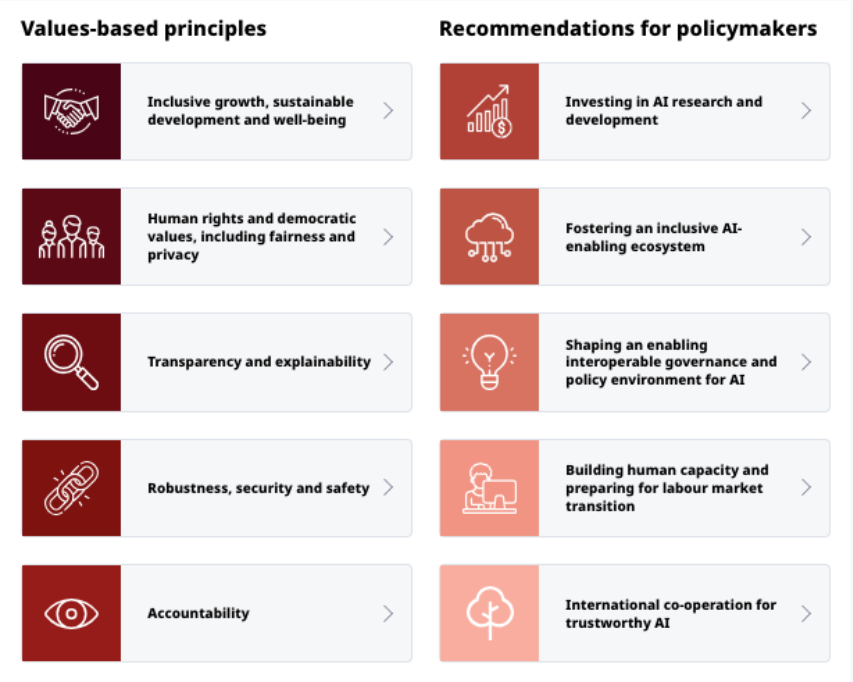
OECD AI Principles Overview

Oranisation for Economic Co-operation and Development

Five principles (inclusive growth, human-centered values, transparency, robustness/security, accountability)

Adopted by 46 countries — the closest thing to a global baseline. Directly influenced the G7 Hiroshima AI Process, the EU AI Act's risk categories, and NIST's AI RMF design

The OECD AI Policy Observatory provides a live tracker of national AI policies, useful for comparing governance approaches across jurisdictions



https://oecd.ai/en/ai-principles
https://oecd.ai/en/ - AI Policy Observatory (Global Tracker)

Deployment & Operational Governance

AI Bill of Materials (AI BOM) / ML Supply Chain Documentation

Emerging concept extending Software Bill of Materials (SBOM) to AI

Supply chain security; tracking model provenance, training data sources, fine-tuning history, and dependency chains

Addresses the **deployment and operations stage** — critical for incident response, vulnerability management, and regulatory audit trails

https://owaspaibom.org/

https://csrc.nist.gov/presentations/2024/securing-ai-ecosystems-the-critical-role-of-aibom

Assigned Reading

Hendrycks Ch. 8: Governance

Let’s talk about Assignment 5

Give me your ideas!