

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)

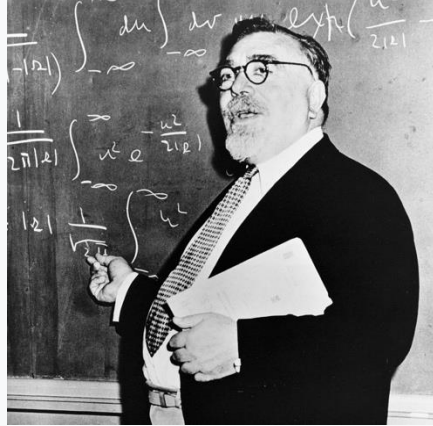


Ancient Greek

kybernan: 'To steer or pilot a ship.'

Ancient Latin

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



Mid-1900's Norbert Wiener

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.



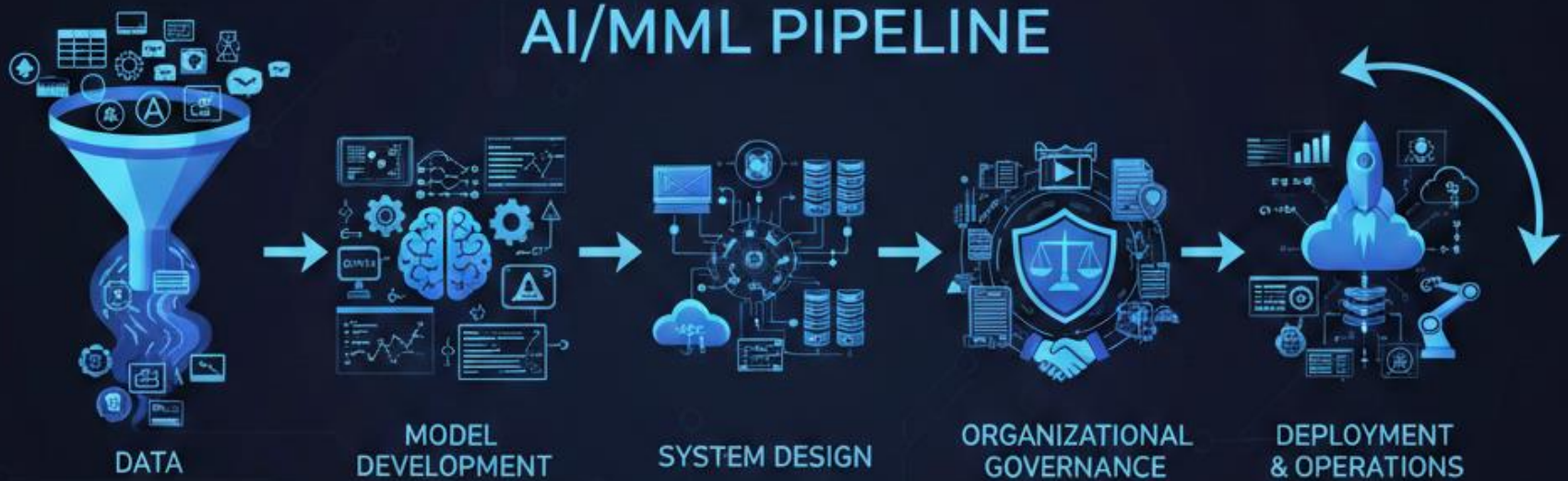
Not just what governments do; defined more broadly as *'the process through which some activity is organized, coordinated, steered, and managed.'*

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)



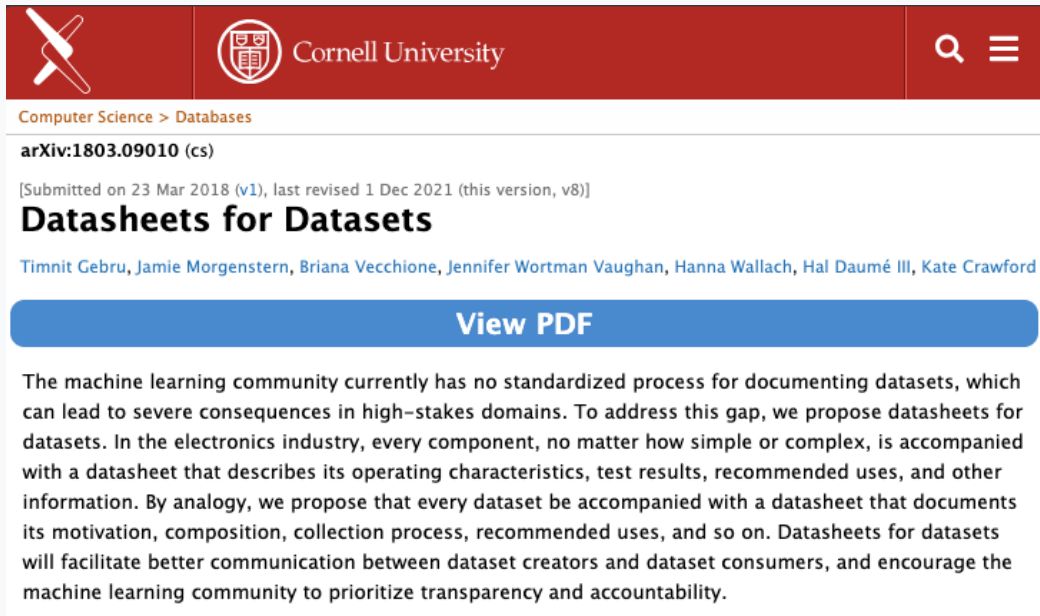
Data Governance

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

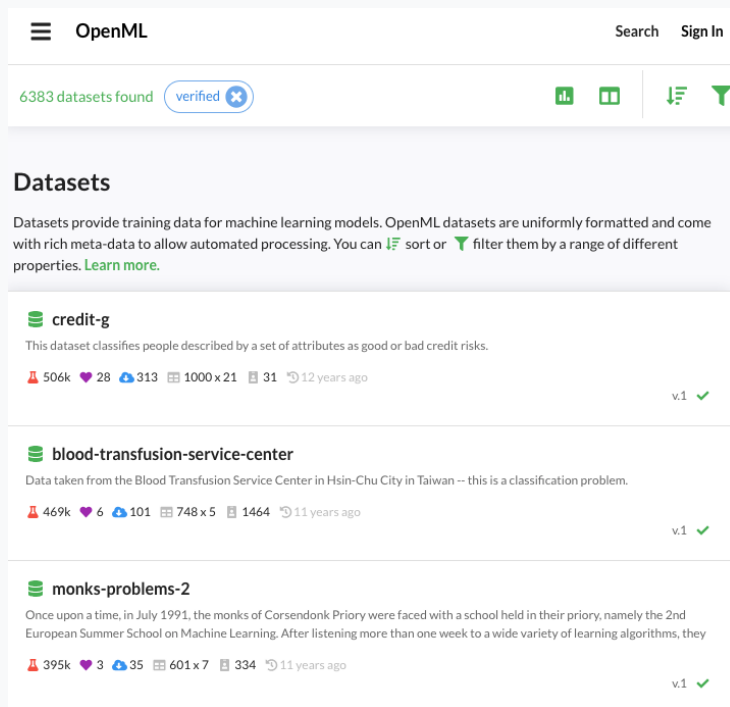
Influential on EU AI Act's data governance requirements (Article 10) and on NIST AI RMF's MAP function for data characterization



The screenshot shows the top portion of an arXiv preprint page. At the top left is a red navigation bar with a white logo of a crossed pencil and paper. To its right is the Cornell University logo and name. On the far right of the red bar are search and menu icons. Below the red bar, the breadcrumb "Computer Science > Databases" is visible. The preprint ID "arXiv:1803.09010 (cs)" is displayed, followed by submission and revision dates: "[Submitted on 23 Mar 2018 (v1), last revised 1 Dec 2021 (this version, v8)]". The title "Datasheets for Datasets" is prominently displayed in bold black text. Below the title, the authors are listed: "Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford". A blue button labeled "View PDF" is positioned below the authors. The main text of the abstract begins: "The machine learning community currently has no standardized process for documenting datasets, which can lead to severe consequences in high-stakes domains. To address this gap, we propose datasheets for datasets. In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet that describes its operating characteristics, test results, recommended uses, and other information. By analogy, we propose that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on. Datasheets for datasets will facilitate better communication between dataset creators and dataset consumers, and encourage the machine learning community to prioritize transparency and accountability."

<https://arxiv.org/abs/1803.09010>

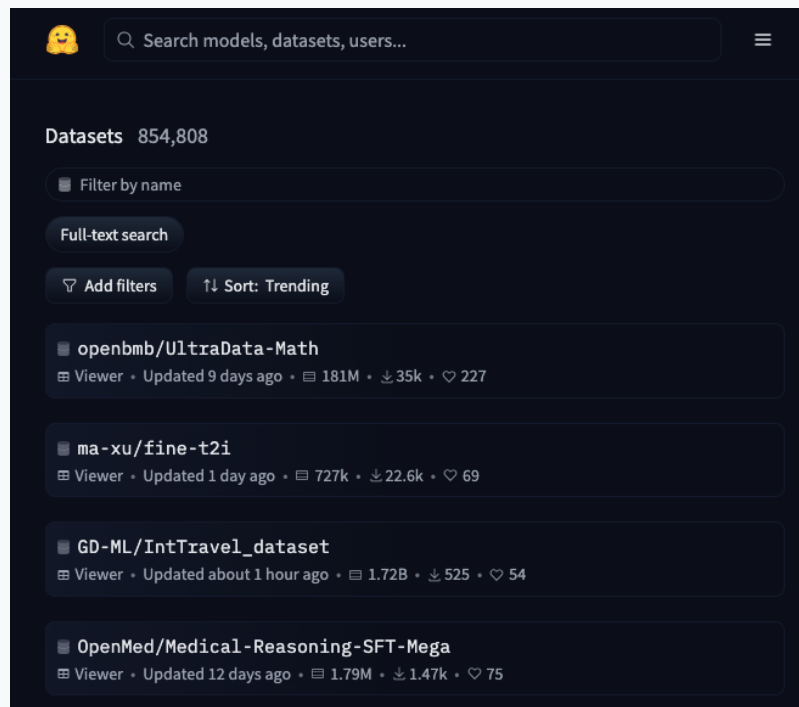
Datasheets for Datasets



The screenshot shows the OpenML website interface. At the top, there is a navigation bar with the OpenML logo, a search bar, and a 'Sign In' button. Below the navigation bar, it displays '6383 datasets found' with a 'verified' badge and several filter icons. The main content area is titled 'Datasets' and includes a brief description: 'Datasets provide training data for machine learning models. OpenML datasets are uniformly formatted and come with rich meta-data to allow automated processing. You can sort or filter them by a range of different properties. Learn more.' Below this, three dataset cards are visible:

- credit-g**: This dataset classifies people described by a set of attributes as good or bad credit risks. It has 506k likes, 28 hearts, 313 downloads, 1000 x 21 files, 31 versions, and was updated 12 years ago.
- blood-transfusion-service-center**: Data taken from the Blood Transfusion Service Center in Hsin-Chu City in Taiwan -- this is a classification problem. It has 469k likes, 6 hearts, 101 downloads, 748 x 5 files, 1464 versions, and was updated 11 years ago.
- monks-problems-2**: Once upon a time, in July 1991, the monks of Corsendonk Priory were faced with a school held in their priory, namely the 2nd European Summer School on Machine Learning. After listening more than one week to a wide variety of learning algorithms, they... It has 395k likes, 3 hearts, 35 downloads, 601 x 7 files, 334 versions, and was updated 11 years ago.

<https://www.openml.org>



The screenshot shows the Hugging Face Datasets page. At the top, there is a search bar with the text 'Search models, datasets, users...' and a menu icon. Below the search bar, it displays 'Datasets 854,808'. There are several filter and search options: 'Filter by name', 'Full-text search', 'Add filters', and 'Sort: Trending'. Below these options, a list of datasets is shown:

- openbmb/UltraData-Math**: Viewer • Updated 9 days ago • 181M • 35k • 227
- ma-xu/fine-t2i**: Viewer • Updated 1 day ago • 727k • 22.6k • 69
- GD-ML/IntTravel_dataset**: Viewer • Updated about 1 hour ago • 1.72B • 525 • 54
- OpenMed/Medical-Reasoning-SFT-Mega**: Viewer • Updated 12 days ago • 1.79M • 1.47k • 75

<https://huggingface.co/datasets>

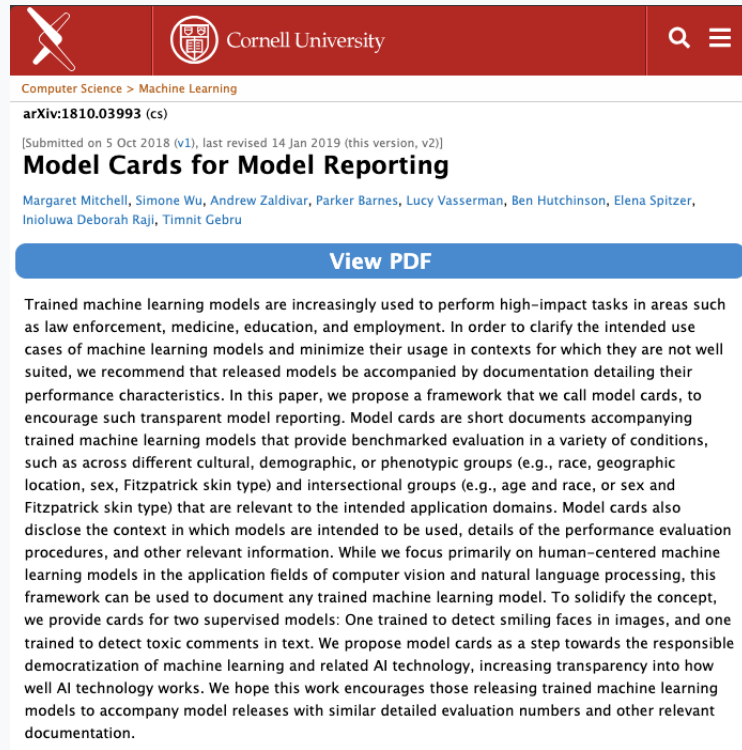
Model Development Governance

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



The screenshot shows the top portion of an arXiv paper page. At the top is a red navigation bar with the Cornell University logo and name on the right, and a search icon on the left. Below the bar, the breadcrumb 'Computer Science > Machine Learning' is visible. The paper ID 'arXiv:1810.03993 (cs)' is displayed, followed by the submission date '[Submitted on 5 Oct 2018 (v1), last revised 14 Jan 2019 (this version, v2)]'. The title 'Model Cards for Model Reporting' is prominently displayed in bold. Below the title, the authors' names are listed: 'Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru'. A blue button labeled 'View PDF' is positioned below the authors. The main text of the paper begins with the sentence: '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

AI

☰

Model System Cards

System cards document the capabilities, safety evaluations, and responsible deployment decisions for Claude models.

Model	
Date	
System card	
<hr/>	
Claude Opus 4.6	
February 2026	
Read system card	
<hr/>	
Claude Opus 4.5	
November 2025	
Read system card	
<hr/>	
Claude Haiku 4.5	
October 2025	
Read system card	

[Anthropic System Cards](#)

☰ Google DeepMind ▾

Model cards

Simple, structured overviews of how an advanced AI model was designed and evaluated.

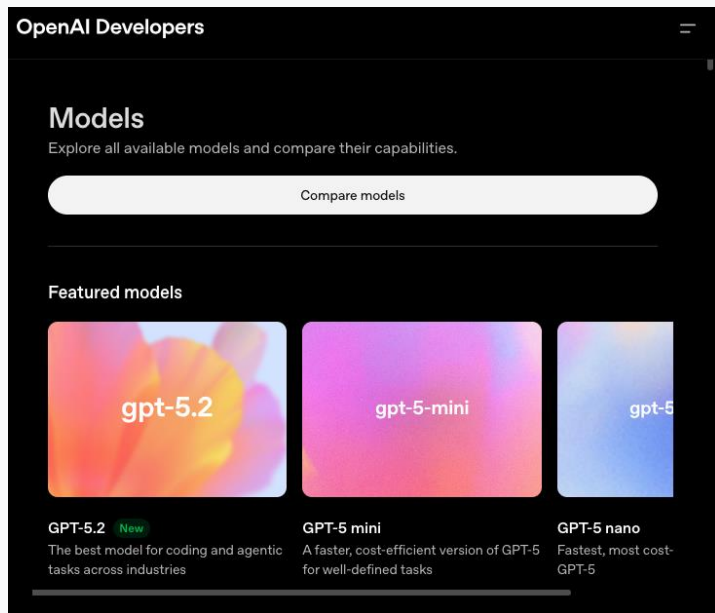
Gemini

Our most intelligent AI models

Gemini 3 Flash	Updated 17 December 2025	View model card
Gemini 3 Pro Image	Updated 20 November 2025	View model card
Gemini 3 Pro	Updated 18 November 2025	View model card
Gemini 2.5 Computer Use	Updated 7 October 2025	View model card

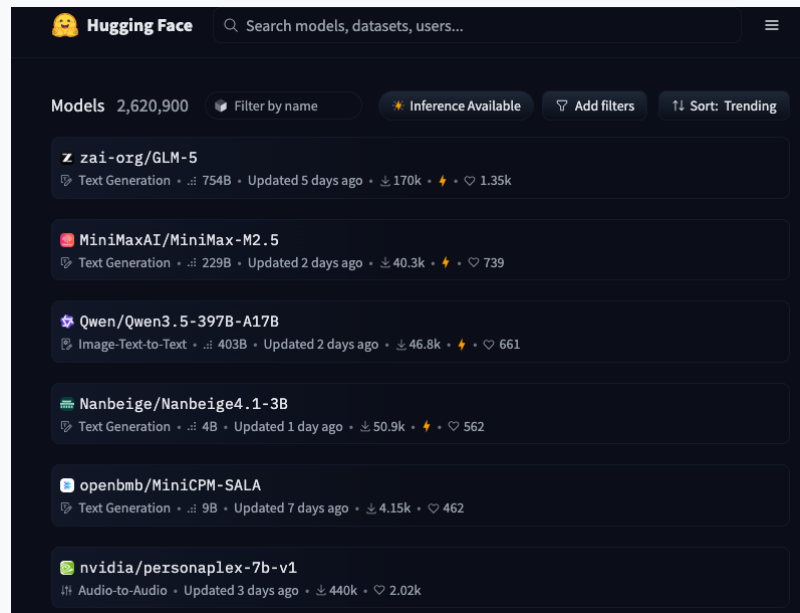
[Google Model Cards](#)

Model Card Examples



The screenshot shows the OpenAI Developers website's 'Models' section. At the top, it says 'OpenAI Developers' and 'Models'. Below this is a sub-header 'Explore all available models and compare their capabilities.' and a 'Compare models' button. The 'Featured models' section displays three cards: 'gpt-5.2' (described as 'The best model for coding and agentic tasks across industries'), 'gpt-5 mini' (described as 'A faster, cost-efficient version of GPT-5 for well-defined tasks'), and 'gpt-5 nano' (described as 'Fastest, most cost-efficient GPT-5').

[OpenAI Model Cards](#)



The screenshot shows the Hugging Face website's 'Models' section. At the top, it says 'Hugging Face' and 'Search models, datasets, users...'. Below this is a sub-header 'Models 2,620,900' and a 'Filter by name' button. There are also buttons for 'Inference Available', 'Add filters', and 'Sort: Trending'. The 'Featured models' section displays a list of models: 'zai-org/GLM-5' (Text Generation, 754B, Updated 5 days ago, 170k downloads, 1.35k likes), 'MiniMaxAI/MiniMax-M2.5' (Text Generation, 229B, Updated 2 days ago, 40.3k downloads, 739 likes), 'Qwen/Qwen3.5-397B-A17B' (Image-Text-to-Text, 403B, Updated 2 days ago, 46.8k downloads, 661 likes), 'Nanbeige/Nanbeige4.1-3B' (Text Generation, 4B, Updated 1 day ago, 50.9k downloads, 562 likes), 'openbmb/MiniCPM-SALA' (Text Generation, 9B, Updated 7 days ago, 4.15k downloads, 462 likes), and 'nvidia/personalex-7b-v1' (Audio-to-Audio, Updated 3 days ago, 440k downloads, 2.02k likes).

[Hugging Face Models](#)

System Design Governance

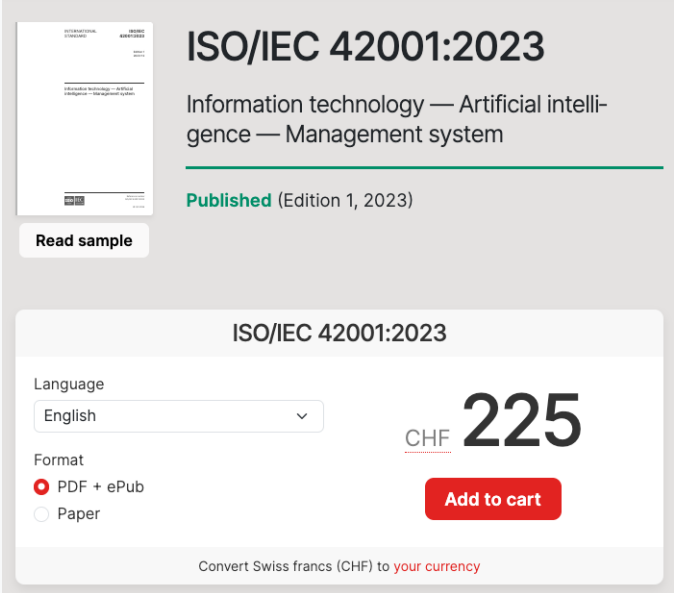
ISO/IEC 42001:2023

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



The screenshot shows the ISO/IEC 42001:2023 product page. At the top left is a thumbnail of the standard's cover. To its right, the title "ISO/IEC 42001:2023" is displayed in a large, bold font, followed by the subtitle "Information technology — Artificial intelligence — Management system". Below the subtitle, it says "Published (Edition 1, 2023)". A "Read sample" button is located below the thumbnail. The main content area features a white background with the title "ISO/IEC 42001:2023" at the top. Below this, there is a "Language" dropdown menu set to "English". To the right of the language menu, the price is shown as "CHF 225". Below the language menu, there are two radio button options for "Format": "PDF + ePub" (which is selected) and "Paper". A red "Add to cart" button is positioned to the right of the format options. At the bottom of the white area, there is a link to "Convert Swiss francs (CHF) to your currency".

<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



Active Standard

IEEE 7000-2021

IEEE Standard Model Process for Addressing Ethical Concerns during System Design

Access via the IEEE Get Program

Access via Subscription

This is a promotional card for the IEEE 7000-2021 standard. It features a dark background with white and blue text. At the top, it says 'Active Standard' with a blue underline. Below that is the standard number 'IEEE 7000-2021' and the full title 'IEEE Standard Model Process for Addressing Ethical Concerns during System Design' in bold. At the bottom, there are two blue buttons: 'Access via the IEEE Get Program' and 'Access via Subscription'.

<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/>

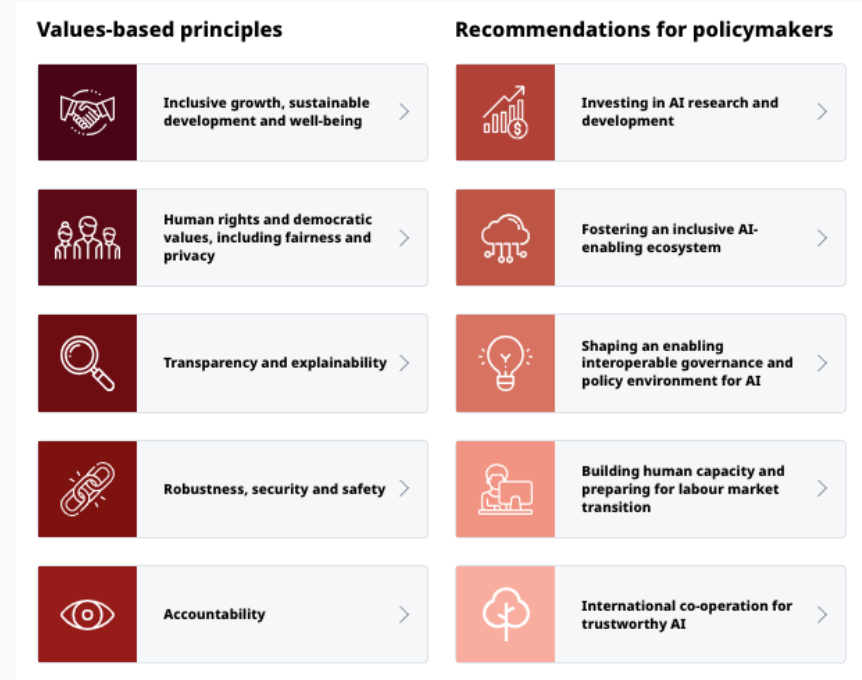
OECD AI Principles Overview

Organisation 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://owaspai bom.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!