DAIG
DATA & AI GOVERNANCE
PARTNERS

AI Governance, Responsible AI, and Data Governance

Connecting the Dots

DW & BI Summit
Utrecht – March 25, 2026

Who is **MATHIAS** Vercauteren

- PhD in Data Governance *(AMS, 2025 - 2029)*
- MSc in Business Economics *(2012, Ghent University)*
- BSc in Sociology *(2009, Ghent University)*

CDMP Certified Data Management Professional ASSOCIATE

Consulting & Advisory Services

- DAMA –DMBOK 3.0 *(Project Manager)*
- UZA *(Hospital)*
- MLOZ *(Healthcare Insurance)*
- Monument Group *(Insurance)*
- De Lijn *(Logistics)*
- MPET *(Logistics)*
- Securex *(Professional Services)*
- Federal Insurance *(Insurance)*
- Flemish Government *(Governmental Institution)*
- Belfius *(Financial Services)*
- Barry Callebaut *(Manufacturing)*
- Carrefour *(Retail)*
- Hilti *(Manufacturing)*

Research

- President of Data & AI Governance Research Institute *(2025, founding phase)*
- PhD in Data Governance *(AMS, 2024 - 2028)*
- Book “Data Governance Sprints” *(Technics Publication, est. Q2 2025)*
- Ethical Technology Institute *(2021 - 2024)*

Educational Services

Training and Coaching Engagements - both in-house and classroom:

- Data Governance
- AI Governance / Responsible AI
- DAMA-DMBOK® / CDMP®
- Data Strategy
- Data Quality
- Master Data Management

Speaking Engagements:

- DGIQ/EDW *(San Diego, 2026)*
- Data Modeling Zone *(San Francisco, 2026)*
- DGIQ/EDW *(Anaheim, 2025)*
- Data and AI Conference *(London, 2025)*
- Data Modeling Zone *(Phoenix, 2025)*
- DGIQ East *(Washington DC, 2024)*
- Data and AI Conference *(London, 2024)*
- DGIQ West *(San Diego, 2024)*
- Enterprise Data World *(Orlando, 2025)*
- DG & MDM Conference *(London, 2023)*
- DGIQ East *(Washington DC, 2023)*

https://daigpartners.com

DAIG DATA & AI GOVERNANCE PARTNERS

AGENDA: Welcome to "Connecting the Dots"

We'll cover the following topics

1. **Artificial Intelligence**
2. **AI Risks**
3. **AI Governance**
4. **Responsible AI**
5. **Data Governance**
6. **Connecting the Dots**

Back to Agenda

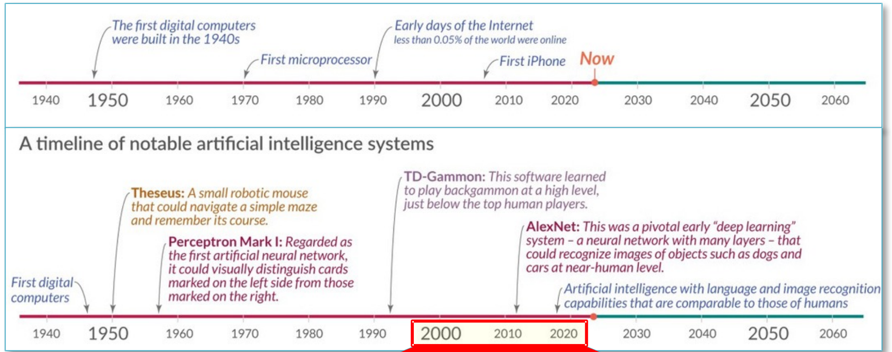
Artificial Intelligence
Orientation

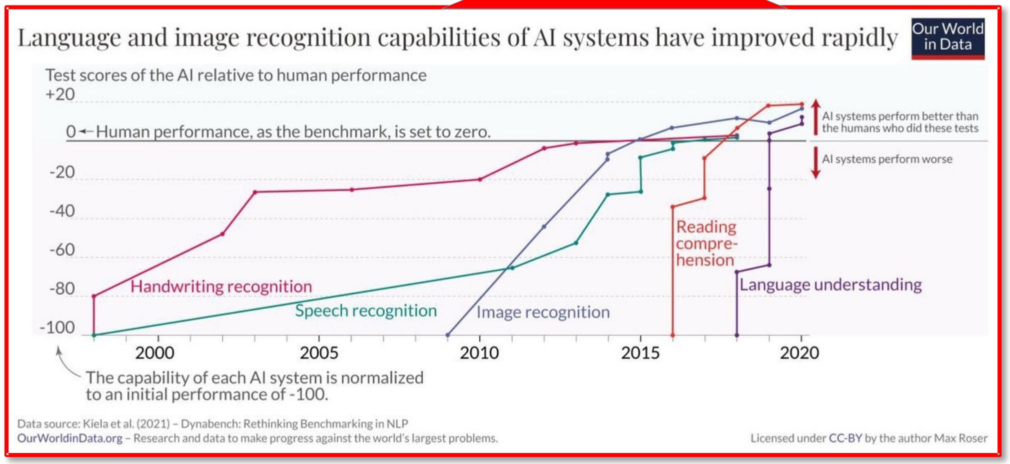
AI **MATURATION** has been rapid

The world of technology is young, but its evolution has been quick… and AI now performs better than humans.

Tech

AI





Source: Roser, M. (2022, December 6). The brief history of artificial intelligence: The world has changed fast — what might be next? Our World in Data. https://ourworldindata.org/brief-history-of-ai

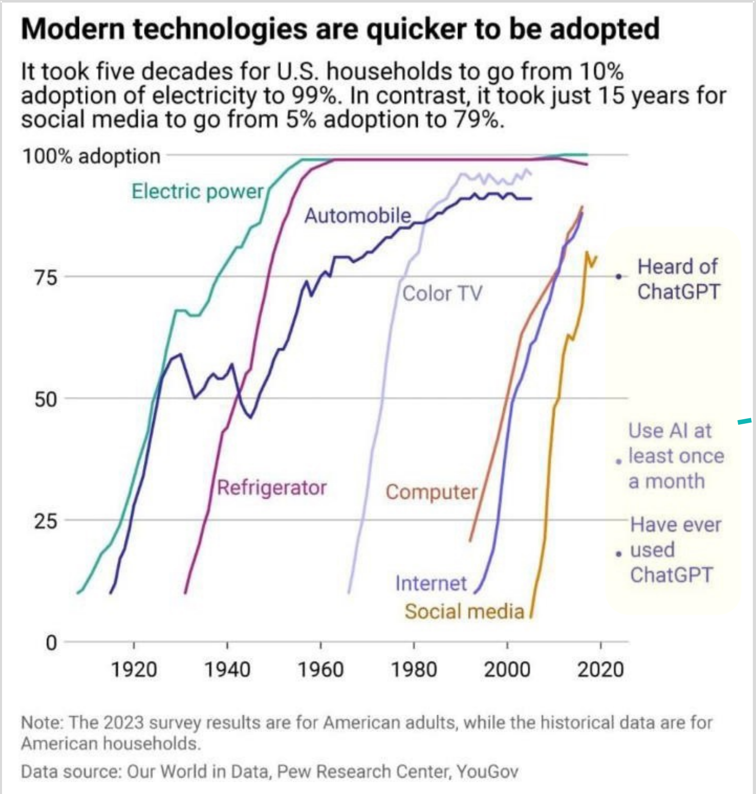
1940s	• First digital computer
1950s	• First artificial neural network *(SNARC)* • First computer programs to play games / learn • First AI program *(Logical Theorist)* • First pattern recognition *(Perceptron)*
1960s	• First industrial robot *(Unimate)* • First heuristic program *(SAINT)* • First NL understanding program *(STUDENT)* • First interactive dialogue program *(ELIZA)* • First expert system *(DENDRAL)*
1970s	• First anthropomorphic robot *(WABOT-1)* • First rule-based ordering system *(XCOM)*
1980s	• First driverless car (Bundeswehr University) • First chat-bot (Jabberwocky) • Handwritten ZIP code recognition
1990s	• Gen2 chat-bot *(A.L.I.C.E.)* • First chess program beats human *(Deep Blue)* • First pet robot *(Furby)*
2000s	• First AI robot to walk as fast as humans *(ASIMO)* • First program to write without humans *(Stats Monkey)* • Handwriting + speech recognition
2010s	• Image recognition / reading + language comprehension • First "deep learning" system • NL computer beats Jeopardy champions *(Watson)*
2020s	• Some AI already outperforms humans

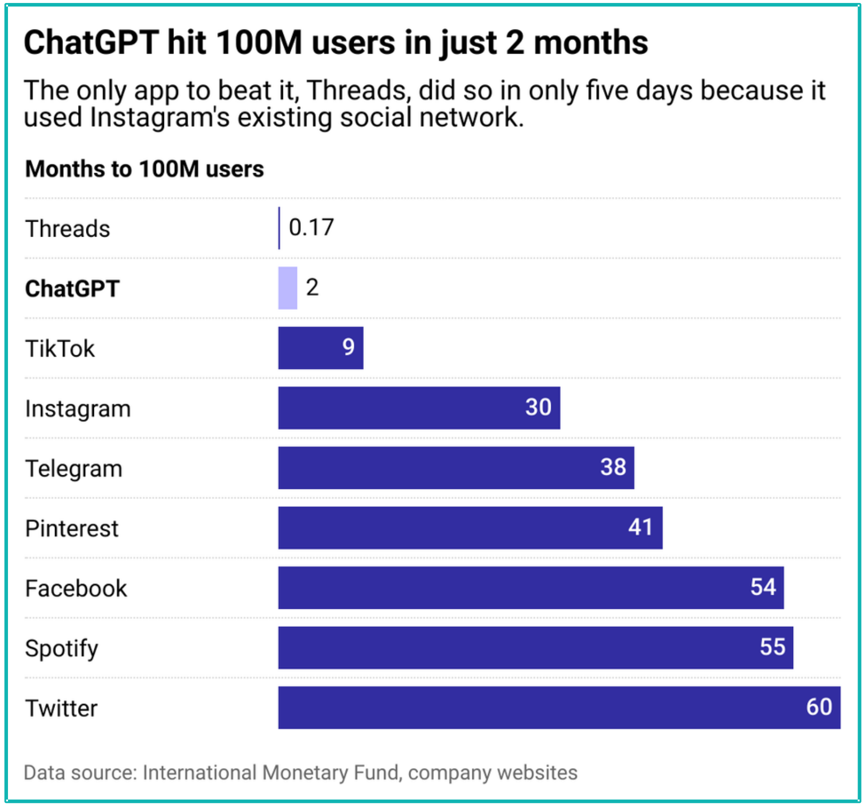
AI Winters
1974-1980
1987-2000

Source: Press, G. (2016, December 30). A very short history of artificial intelligence (AI). Forbes. https://www.forbes.com/sites/gilpress/2016/12/30/a-very-short-history-of-artificial-intelligence-ai/

AI **ADOPTION** rates outpaced most modern technologies

ChatGPT (AI text generation) reached 100M users faster than most technologies.





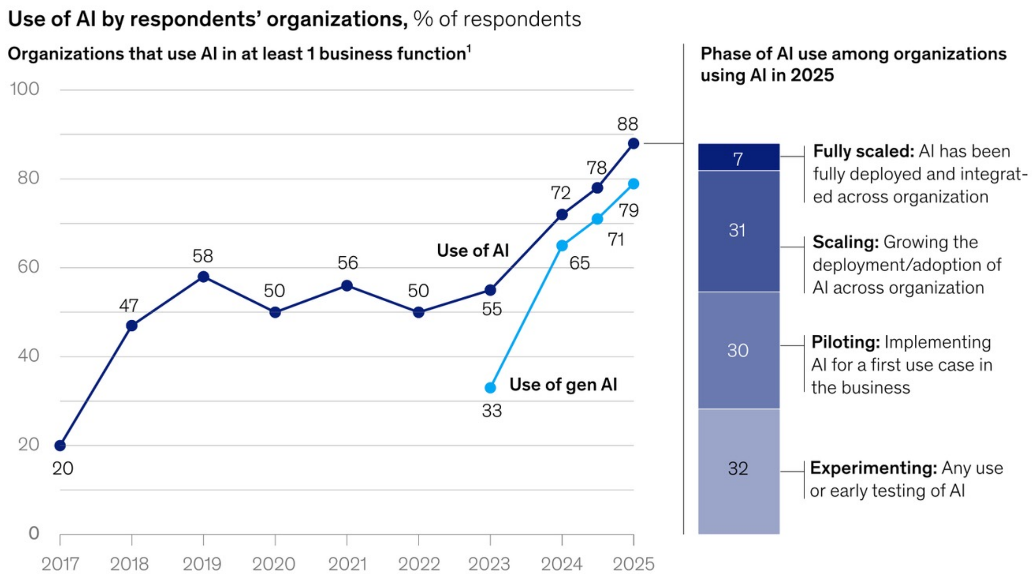
Source: Verbit Editorial. (n.d.). How quickly are consumers & businesses adopting AI tools compared to past technologies? Verbit. Retrieved February 22, 2024, from https://verbit.ai/how-quickly-are-consumers-and-businesses-adopting-ai-tools-compared-to-past-technologies

AI moves from experiment to **EVERYDAY BUSINESS**

More companies than ever now use AI in at least one business function—and the curve is still climbing.

- The share of companies using AI in at least **one business function** continues to rise year over year, signalling that AI has firmly entered mainstream operations.
- Adoption is no longer limited to digital leaders—most industries now report **widespread functional AI use** across marketing, operations, product development, and service.
- Organizations are increasingly moving beyond isolated pilots, **embedding AI into everyday workflows** where the impact is tangible and repeatable.
- With AI usage expanding across functions, the **competitive gap is shifting** from adoption to effective scaling and governance of AI systems.

Reported use of AI in at least one business function continues to increase.



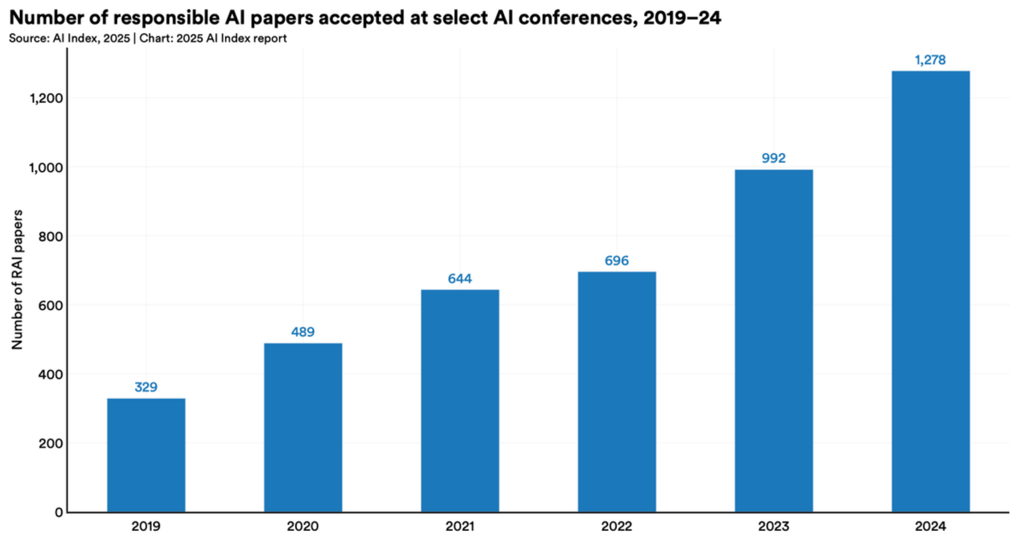
[1]In 2017, the definition for AI use was using AI in a core part of the organization's business or at scale. In 2018–19, the definition was embedding at least 1 AI capability in business processes or products. From 2020, the definition was that the organization has adopted AI in at least 1 function, and in 2025, the definition was regular use of AI in at least 1 function.
Source: McKinsey Global Surveys on the state of AI, 2017–25

Source: McKinsey & Company. (2025, November 5). The state of AI in 2025: Agents, innovation, and transformation. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

AI Governance is **ACCELERATING** Globally

Business adoption and regulatory activity surge in 2024.

The urgency for AI governance has never been more pronounced. In 2025, 88% of organizations reported using AI systems, a dramatic increase from 55% percent in 2023, matched by unprecedented regulatory activity.



88%

Organizations using AI in 2025 (up from 55% in 2023)

59

U.S. federal AI regulations in 2024 (2× increase from 2023)

21.3%

Increase in global legislative AI mentions (75 countries)

9×

Increase in AI legislation since 2016

This convergence creates an imperative for robust governance frameworks. Without proper guardrails, organizations face legal exposure, reputational damage, and operational disruption at scale.

Source: Stanford University Human-Centered Artificial Intelligence. (2025). The 2025 AI Index report. https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

We face multiple **DEFINITIONS** of AI

There is no single, universally accepted definition of AI… nonetheless all share common elements.

Dictionaries

Oxford English Dictionary:
The capacity of computers or other machines to exhibit or simulate intelligent behaviour; the field of study concerned with this. In later use also: software used to perform tasks or produce output previously thought to require human intelligence, esp. by using machine learning to extrapolate from large collections of data. Also as a count noun: an instance of this type of software; a (notional) entity exhibiting such intelligence.

Mirriam-Webster Dictionary:
1: the capability of computer systems or algorithms to imitate intelligent human behavior. *also, plural artificial intelligences*: a computer, computer system, or set of algorithms having this capability.
2: a branch of computer science dealing with the simulation of intelligent behavior in computers

Government

US Department of Commerce (CSRC/NIST) :
1. A branch of computer science devoted to developing data processing systems that performs functions normally associated with human intelligence, such as reasoning, learning, and self-improvement.
2. The capability of a device to perform functions that are normally associated with human intelligence such as reasoning, learning, and self-improvement.

US Dept of State:
A machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments.

EU Parliament:
The ability of a machine to display human-like capabilities such as reasoning, learning, planning and creativity.

Technology

Gartner:
Applying advanced analysis and logic- based techniques, including machine learning (ML), to interpret events, support and automate decisions, and take actions.

Forrester:
The theory and capabilities that strive to mimic human intelligence through experience and learning.

TechTarget:
The simulation of human intelligence processes by machines, especially computer systems.

IBM:
technology that enables computers and digital devices to learn, read, write, talk, see, create, play, analyze, make recommendations, and do other things humans do

Prof'l Orgs

DAMA International: *(member login required)*
Software that performs a function previously ascribed only to human beings, such as natural language processing.

IAPP:
A broad term used to describe an engineered system where machines learn from experience, adjusting to new inputs, and potentially performing tasks previously done by humans. More specifically, it is a field of computer science dedicated to simulating intelligent behavior in computers.

4 elements = computers software algorithms + automation + humans cognition intelligence + recommend decide predict create

TIP: Click any definition box in Slide Show mode *(or Ctrl+click in Slide View/Edit mode)* to view that definition's source

Lifecycle and key dimensions of an AI **SYSTEM**

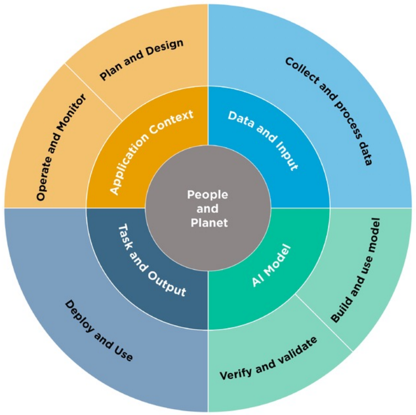


Fig. 2. Lifecycle and Key Dimensions of an AI System. Modified from OECD (2022) OECD Framework for the Classification of AI systems — OECD Digital Economy Papers. The two inner circles show AI systems' key dimensions and the outer circle shows AI lifecycle stages. Ideally, risk management efforts start with the Plan and Design function in the application context and are performed throughout the AI system lifecycle. See Figure 3 for representative AI actors.

Key Dimensions	Application Context	Data & Input	AI Model	AI Model	Task & Output	Application Context	People & Planet
Lifecycle Stage	Plan and Design	Collect and Process Data	Build and Use Model	Verify and Validate	Deploy and Use	Operate and Monitor	Use or Impacted by
TEVV	TEVV includes audit & impact assessment	TEVV includes internal & external validation	TEVV includes model testing	TEVV includes model testing	TEVV includes integration, compliance testing & validation	TEVV includes audit & impact assessment	TEVV includes audit & impact assessment
Activities	Articulate and document the system's concept and objectives, underlying assumptions, and context in light of legal and regulatory requirements and ethical considerations.	Gather, validate, and clean data and document the metadata and characteristics of the dataset, in light of objectives, legal and ethical considerations.	Create or select algorithms; train models.	Verify & validate, calibrate, and interpret model output.	Pilot, check compatibility with legacy systems, verify regulatory compliance, manage organizational change, and evaluate user experience.	Operate the AI system and continuously assess its recommendations and impacts (both intended and unintended) in light of objectives, legal and regulatory requirements, and ethical considerations.	Use system/ technology; monitor & assess impacts; seek mitigation of impacts, advocate for rights.
Representative Actors	System operators; end users; domain experts; AI designers; impact assessors; TEVV experts; product managers; compliance experts; auditors; governance experts; organizational management; C-suite executives; impacted individuals/ communities; evaluators.	Data scientists; data engineers; data providers; domain experts; socio-cultural analysts; human factors experts; TEVV experts.	Modelers; model engineers; data scientists; developers; domain experts; with consultation of socio-cultural analysts familiar with the application context and TEVV experts.		System integrators; developers; systems engineers; software engineers; domain experts; procurement experts; third-party suppliers; C-suite executives; with consultation of human factors experts, socio-cultural analysts, governance experts, TEVV experts,	System operators, end users, and practitioners; domain experts; AI designers; impact assessors; TEVV experts; system funders; product managers; compliance experts; auditors; governance experts; organizational management; impact-ed individuals/commu-nities; evaluators.	End users, operators, and practitioners; impacted individu-als/communities; general public; policy makers; standards organizations; trade associations; advocacy groups; environmental groups; civil society organizations; researchers.

Fig. 3. AI actors across AI lifecycle stages. See Appendix A for detailed descriptions of AI actor tasks, including details about testing, evaluation, verification, and validation tasks. Note that AI actors in the AI Model dimension (Figure 2) are separated as a best practice, with those building and using the models separated from those verifying and validating the models.

Source: National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). https://doi.org/10.6028/nist.ai.100-1

We can describe AI using various **CATEGORIES**

Each category set describes a different aspect of AI (and these categories are not mutually exclusive).

Categories

Strength	Function	Purpose	Technique	Learning	Output
Weak or Narrow	Reactive Machines	Learn from Data	Find Patterns	Supervised	Discriminative
General AI	Limited Memory	Extract Knowledge	Apply Rules	Unsupervised	Generative
Strong or Super AI	Theory of Mind	Use Language	Optimize Resources	Semi-supervised	
	Self-Awareness	Perform Tasks	Process Language	Reinforced	
		Mimic Humans	Represent Knowledge		
			Make Decisions		

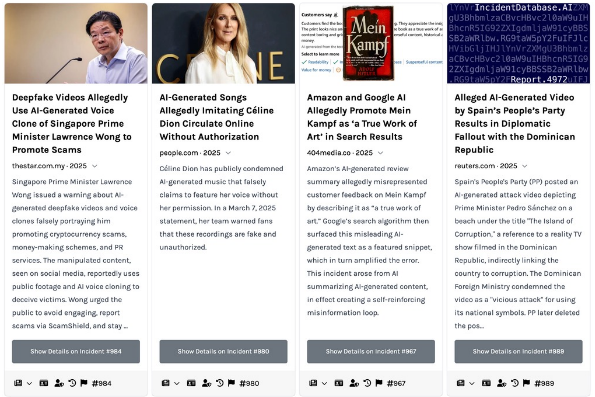
Back to Agenda

AI Risks
Understanding Risks & Challenges

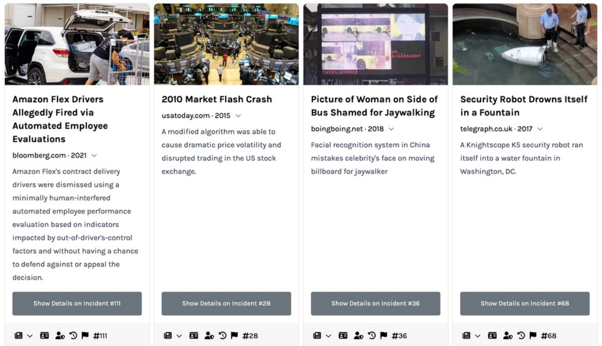
The stakes are BOARD-LEVEL

Financial, legal, and reputational risks concentrate at deployment scale.

As AI systems transition from pilot projects to enterprise deployments, risk shifts from theoretical to board-level materiality. Recent incidents underscore the urgent need for robust AI governance to prevent and manage such occurrences.



Source: The AI Incident Database, Accessed on 15/04/2025. Available at: https://incidentdatabase.ai



The AI Incident Database

The AI Incident Database is dedicated to indexing the collective history of harms or near harms realized in the real world by the deployment of artificial intelligence systems.

15 Potential AI Risks

1. Automation-spurred job loss
2. Deepfakes
3. Privacy Violations
4. Algorithmic bias caused by bad data
5. Socioeconomic inequality
6. Danger to humans
7. Unclear legal regulation
8. Social manipulation
9. Invasion of privacy and social grading
10. Misalignment between our goals and AI's goals
11. A lack of transparency
12. Loss of control
13. Introducing program bias into decision-making
14. Data sourcing and violation of personal privacy
15. Techno-solutionism

Source: WalkMe Team. (2025, June 23). 15 Potential Artificial Intelligence (AI) Risks. https://www.walkme.com/blog/ai-risks/

DAIG DATA & AI GOVERNANCE PARTNERS

Unintended **HARM**: The core challenge of AI systems

- AI systems, despite their promise, can lead to unintended negative consequences across various domains.
- The NIST AI Risk Management Framework (AI RMF) categorizes these potential harms into three main areas:

Fig. 1. Examples of potential harms related to AI systems. Trustworthy AI systems and their responsible use can mitigate negative risks and contribute to benefits for people, organizations, and ecosystems.

Source: National Institute of Standards and Technology (NIST) , "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," January 2023. Available at: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

Manifestations of AI harm become **INCIDENTS**

The AI Incident Database already contains 1000+ cross-referenced examples of AI harms.

	ID	Description
Bias	19	Google ads showed (1) more executive & career building jobs to males (vs females); (2) more arrest references for searches on Black names in ads from Instant CheckMate
Bias	37	Amazon shut down recruiting tool that downranked female applicants
Bias	47	LinkedIn search engine favored male names
Bias	48	New Zealand passport checker rejected Asian man's passport application after detecting his eyes as closed
Bias	87	UK passport checker showed higher rejection rate for dark-skinned women
Bias	95	HireVue removed AI employability scoring & ranking tool that analyzes facial movements, word choice & speaking voice during i nterviews after public outcry
Bias	265	Black courier sued Uber Eats over racist facial recognition dismissal based on incorrect identification after increasingly frequent requests for more verification selfies
Bias	390	Voice & video deepfakes & stolen PII used in online interviews of candidates applying for remote work & work -at-home positions
Bias	489	Workday's AI screening system is alleged in a lawsuit to allow employers to discriminate against African-Americans, people over 40 & people with disabilities
Privacy	618	Largest US credit union rejected more than half of its Black applicants & approved Latinos at significantly lower rates vs. Whites
Privacy	267	Clearview AI (US-based facial recognition software) hit with multiple fines & complaints for illegally collecting billions of images online photos without consent
Privacy	355	UK/Portugal drivers win lawsuit against Uber & Ola for robo-termination based on AI-detected fraud, denying access to personal data & lack of transparency into how data used
Privacy	412	Finland's National Police Board reprimanded for illegal processing of personal data in a facial recognition trial that did not comply with data protection legislation
Privacy	441	S. Korea shared photos of 170 million travelers (without their consent) to private companies developing immigration screening
Security	6	Microsoft chatbot removed within 24 hours after generating multiple racist, sexist & anti-Semitic tweets due to inputs by Twitter users
Security	352	Twitter users derail GPT-3 tweet bot dedicated to remote jobs by exploiting a newly discovered prompt injection hack to make bot to repeat embarrassing & ridiculous phrases
Security	473	Bing Chat users leverage prompt injection to reveal its built-in initial instructions, including a list of statements governing ChatGPT's interaction with users
Security	622	User crafts prompt to get Chevrolet dealer's ChatGPT bot to sell a 2024 Chevy Tahoe for $1 by manipulating the chatbot's obje ctive to agree with any statement

▲ Click the ID to view that record in the online AI Incident Database

Source: Responsible AI Collaborative. (n.d.). AI Incident Database | Discover. Retrieved April 3, 2024, from https://incidentdatabase.ai/apps/discover/?is_incident_report=true

AI presents new & different **BUSINESS** risks

The potential for these exposure areas increases with ungoverned implementation & usage of AI solutions.

Disruption Risks

- Autonomous vehicles replace drivers
- Chatbots replace customer service agents
- Content generators replace creative authors

Cybersecurity Risks

- Deepfakes increase phishing & online scams
- Users bypass prompts to control or interrupt AI
- Hackers employ AI 24/7 to find vulnerabilities faster

Reputation Risks

- AI outputs violate company values & ethics
- Poor chatbots drive customers to competitors
- AI used in ways perceived as unethical or harmful

Legal Risks

- Opaqueness (lack of transparency) into AI algorithms
- Discrimination & bias affects vulnerable groups
- Improperly using or exposing protected or copyrighted data

Operational Risks

- AI unintentionally exposes IP or trade secrets
- Overreliance on AI may reduce human cognition
- Having to remove AI solutions built with improper data

AI risk mitigation needs to be a critical component of strategic planning & risk management efforts

AI can be impacted by **DATA** risks

The best AI comes from the best data… anything that hinders “trusted” data will also affect AI outcomes.

Data Silo Risks

- Delays in updates & communication across departments
- Inconsistent & competing priorities across teams
- Dropping the ball or double-checking work along the way

Data Language Risks

- Differences in terminology across lines of business
- Different ways of organizing & storing information
- Disconnected, fragmented & badly designed systems

Junk Data Risks

- Lack of widespread “good data hygiene” habits & audits
- Unlabeled, misfiled, or outdated files, documents & data
- Multiple versions of data sets stored across fileshares

Ontology Risks

- Lack of formal data & information architecture
- Inconsistent data naming, structures & standards
- Varying descriptions & attributes of entities

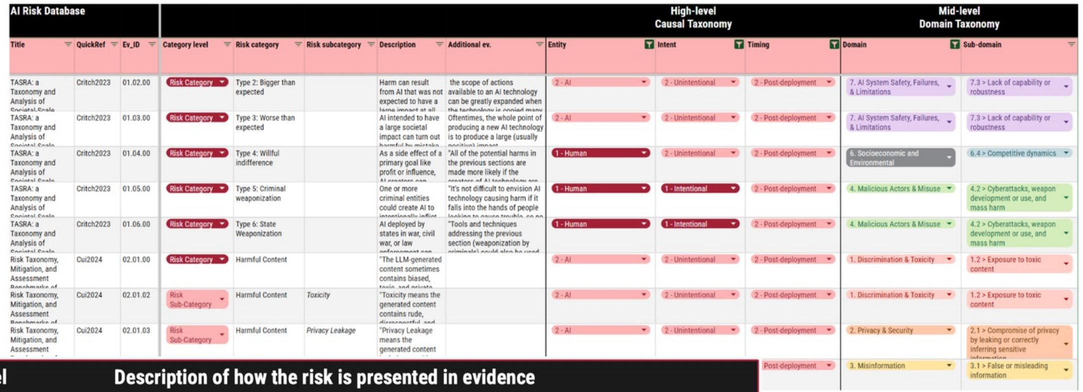
Data Quality Risks

- Low quality data delays or stalls AI development efforts
- Improperly labeled data compromises model training
- Garbage in/garbage out (GIGO) hinders AI adoption

High-quality data increases business agility & establishes the data foundation essential to AI success

DAIG PARTNERS

MIT AI **RISK** Repository



Category	Level	Description of how the risk is presented in evidence
Entity	AI	Due to a decision or action made by an AI system
	Human	Due to a decision or action made by humans
	Other	Due to some other reason or ambiguous
Intent	Intentional	Due to an expected outcome from pursuing a goal
	Unintentional	Due to an unexpected outcome from pursuing a goal
	Other	Without clearly specifying the intentionality
Timing	Pre-deployment	Before the AI is deployed
	Post-deployment	After the AI model has been trained a
	Other	Without a clearly specified time of oc

Domain / Subdomain	
1	***Discrimination & Toxicity***
1.1	Unfair discrimination and misrepresentation
1.2	Exposure to toxic content
1.3	Unequal performance across groups
2	***Privacy & Security***
2.1	Compromise of privacy by obtaining, leaking or correctly inferring sensitive information
2.2	AI system security vulnerabilities and attacks
3	***Misinformation***
3.1	False or misleading information
3.2	Pollution of information ecosystem and loss of consensus reality
4	***Malicious actors & Misuse***
4.1	Disinformation, surveillance, and influence at scale
4.2	Cyberattacks, weapon development or use, and mass harm
4.3	Fraud, scams, and targeted manipulation

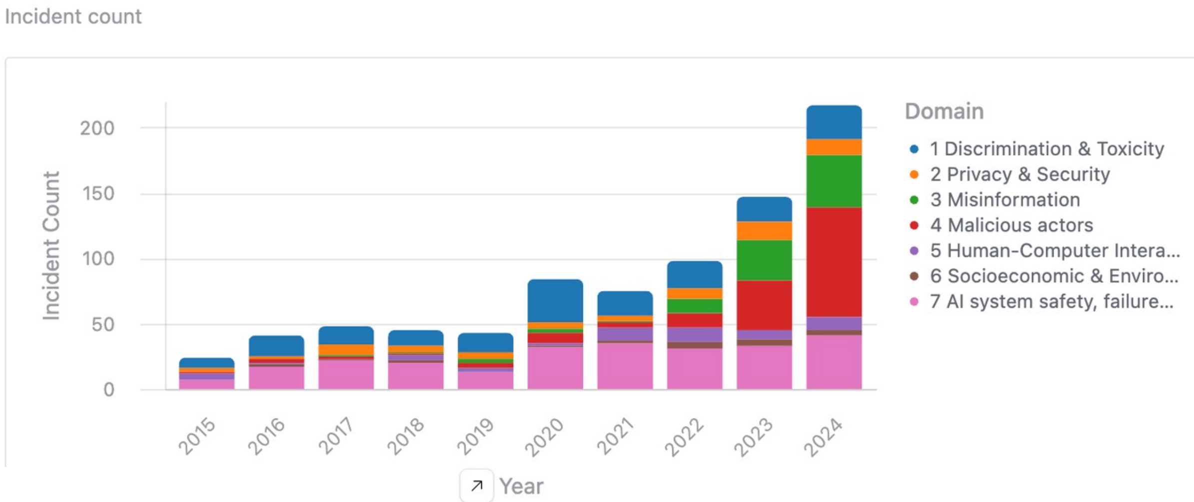
Domain / Subdomain	
5	***Human-Computer Interaction***
5.1	Overreliance and unsafe use
5.2	Loss of human agency and autonomy
6	***Socioeconomic & Environmental Harms***
6.1	Power centralization and unfair distribution of benefits
6.2	Increased inequality and decline in employment quality
6.3	Economic and cultural devaluation of human effort
6.4	Competitive dynamics
6.5	Governance failure
6.6	Environmental harm
7	***AI system safety, failures, and limitations***
7.1	AI pursuing its own goals in conflict with human goals or values
7.2	AI possessing dangerous capabilities
7.3	Lack of capability or robustness
7.4	Lack of transparency or interpretability
7.5	AI welfare and rights
7.6	Multi-agent risks

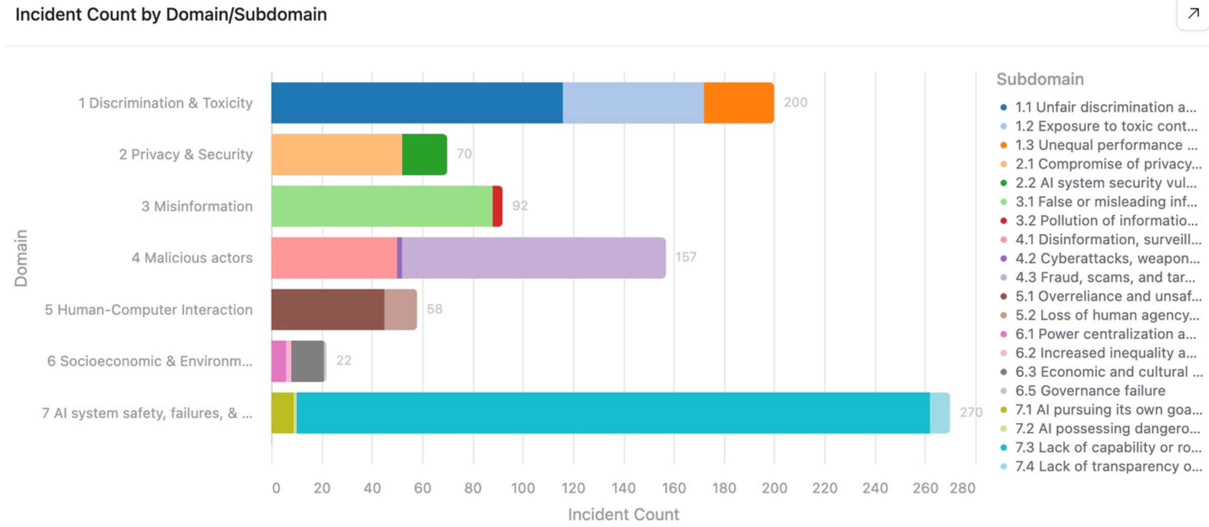
The AI Risk Repository has three parts:

- The **AI Risk Database** captures 1700+ risks extracted from 74 existing frameworks and classifications of AI risks
- The **Causal Taxonomy of AI Risks** classifies how, when, and why these risks occur
- The **Domain Taxonomy of AI Risks** classifies these risks into 7 domains (e.g., "Misinformation") and 24 subdomains (e.g., "False or misleading information")

Source: MIT AI Risk Repository, Accessed 15/04/2025. Available at: https://airisk.mit.edu

MIT AI Risk Repository - Risk CLASSIFICATION





Source: MIT AI Risk Repository, Accessed 15/04/2025. Available at: https://airisk.mit.edu

Back to Agenda

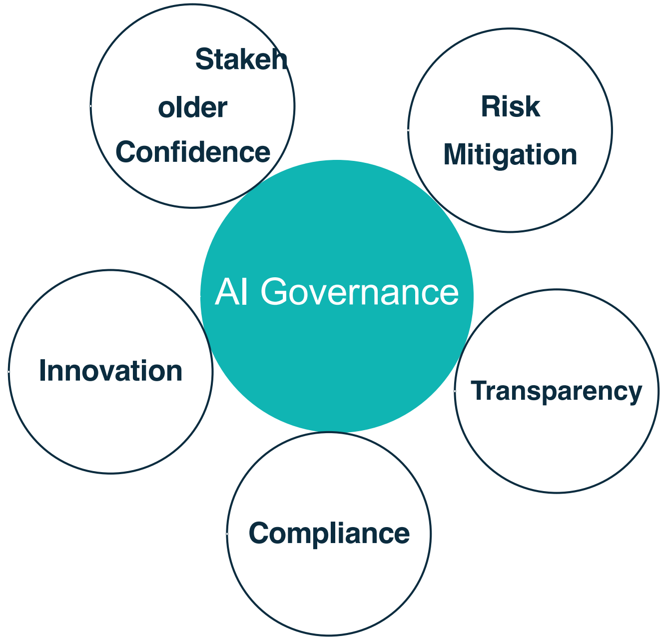
AI Governance
The Need & Definition

DAIG
DATA & AI GOVERNANCE
PARTNERS

ENSURING responsible, ethical, and effective AI

Why AI Governance matters.

- **Mitigating risks** is a primary driver for AI governance. This includes identifying and addressing potential ethical, social, economic, and legal risks associated with AI deployment, such as bias and discrimination.
- Robust AI governance fosters **transparency** in AI systems, promoting understanding of their purpose, algorithms, data sets, and outputs. **Explainability** is key to building trust and enabling human oversight.
- Implementing AI governance capabilities helps organisations ensure **compliance** with emerging AI-related regulations and ethical standards, minimising legal and reputational risks
- Strong AI governance enhances **stakeholder confidence** by demonstrating a commitment to responsible AI practices. This includes engaging with developers, users, policymakers, and affected communities.
- AI governance is not just about risk reduction; it also **enables innovation** by providing a structured and ethical pathway for the development and deployment of AI technologies, aligning them with business objectives and societal values.



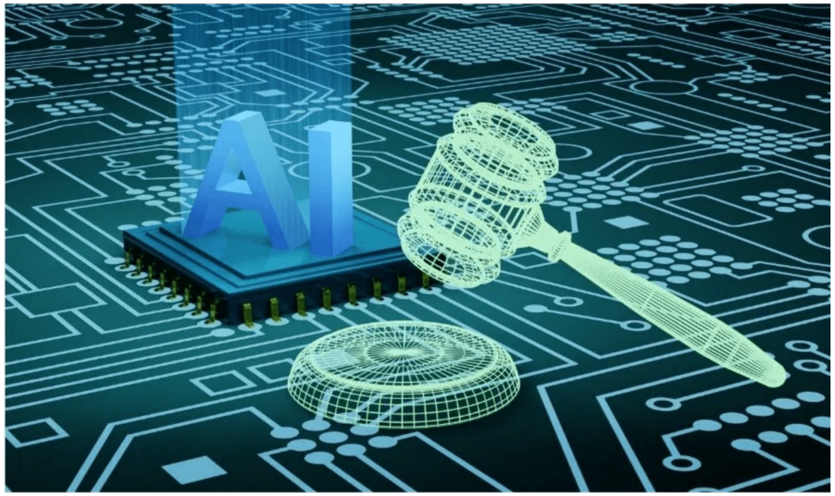
Intersection with Responsible AI and Data Governance

AI governance acts as an overarching capabilty that incorporates the principles of **Responsible AI** (ethical, fair, trustworthy AI) and relies heavily on strong **Data Governance** practices to ensure data quality, security, integrity, and ethical use of data, which are foundational for reliable AI.

DEFINING AI Governance

“**Artificial intelligence (AI) governance** refers to the processes, standards and guardrails that help ensure AI systems and tools are safe and ethical. AI governance frameworks direct AI research, development and application to help ensure safety, fairness and respect for human rights.”

Source: IBM, “What is AI Governance?” IBM Think, 2025. Available at: https://www.ibm.com/think/topics/ai-governance



Core Definition Components

- **Principles & Policies:** Principles, policies, standards, and decision criteria that direct ethical and compliant AI use.
- **Processes & Practices:** Repeatable, auditable lifecycle activities from intake to retirement. Lifecycle workflows for design, development, validation, monitoring, and decommissioning.
- **Risk Controls & Guardrails:** Defined control mechanisms and safeguards to manage bias, privacy, misuse, security, and safety risks.
- **Alignment & Oversight:** Alignment with strategy, organizational values, legal obligations, and stakeholder expectations.
- **Evidence & Traceability:** Auditable documentation, traceability, and evidence that controls and policies are operating as intended.

Three Interconnected Dimensions

- **Ethical Alignment:** Ensures fairness, transparency, accountability, human-centric design, and respect for rights.
- **Legal & Regulatory Compliance:** Compliance with applicable laws and standards (e.g., EU AI Act, GDPR, sector regulations)
- **Operational & Risk Resilience:** Robustness, security, reliability, and continuous monitoring of AI systems.

The growing **COMPLEXITY** demands AI Governance

Navigating new risks in an evolving technological landscape.

1. Increasing algorithmic complexity

- Modern AI systems have evolved significantly from simple rule-based systems to sophisticated deep learning and generative AI models. This increasing algorithmic complexity makes understanding and predicting their behaviour more challenging.

2. Vast and diverse datasets

- These advanced AI models often integrate vast and diverse datasets, further complicating the analysis of inputs, processes, and outputs. The quality and governance of this underlying data are crucial.

3. Ubiquitous nature of AI

- The growing functionality of AI in Software as a Service (SaaS), cloud platforms, and vendor products means AI is increasingly accessible and potentially in use across organisations, sometimes without formal oversight. This ubiquitous nature of AI amplifies the need for centralized governance.

4. New and different business risks

- Ungoverned implementation and usage of complex AI solutions can lead to new and different business risks, including inaccuracies, biases, lack of transparency, and legal defensibility issues.

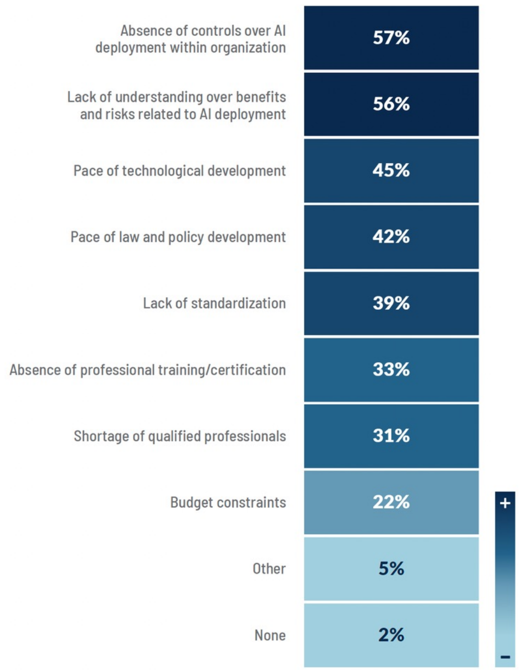
AI GOVERNANCE is in its early stages

Challenges and slow-ish implementation are still common and there's no standard governance framework (yet).

IAPP's definition

Organizational AI governance refers to the internal guidelines and practices organizations follow to ensure responsible development, deployment or use of AI by that organization.

Most common AI governance challenges faced by organizations



Existence of an AI governance function by annual revenue in USD

	Overall	Under 100 million	100-999 million	1-8.9 billion	9-19.9 billion	20-59.9 billion	More than 60 billion
Established AI governance function	29%	17%	26%	31%	18%	38%	52% ↑
Likely to establish an AI governance function in the next 12 months	31%	28%	31%	31%	39%	29%	26%
No established AI governance function	35%	45%	39%	34%	34%	32%	13% ↓
Unsure	6%	11%	4%	5%	8%	0%	10%

Existence of an AI governance function by number of employees

	Overall	Under 100	100-999	1,000-4,999	5,000-24,999	25,000-79,999	More than 80,000
Established AI governance function	29%	21%	18%	22%	34%	27%	45% ↑
Likely to establish an AI governance function in the next 12 months	31%	21%	29%	27%	37%	27%	30%
No established AI governance function	35%	43%	49% ↑	46% ↑	27% ↓	30%	20% ↓
Unsure	6%	14%	4%	5%	3%	16% ↑	5%

Existence of an AI governance function by respondent's confidence in privacy compliance

	Overall	Not at all confident	Somewhat confident	Totally confident
Established AI governance function	29%	12% ↓	30%	32%
Likely to establish an AI governance function in the next 12 months	31%	19%	31%	37%
No established AI governance function	35%	65% ↑	33%	28%
Unsure	6%	4%	7%	4%

Source: International Association of Privacy Professionals & EY. (2023). IAPP–EY professionalizing organizational AI governance report. IAPP. https://iapp.org/media/pdf/resource_center/iapp_ey_professionalizing_organizational_ai_governance_report.pdf

Global AI Law and Policy TRACKER

This tracker identifies AI legislative and policy developments in a subset of jurisdictions.

Jurisdictions in focus

Argentina
Australia
Bangladesh
Brazil
Canada
Chile
China
Colombia
Egypt
EU
India
Indonesia
Israel
Japan
Mauritius
New Zealand
Nigeria
Peru
Saudi Arabia
Singapore
South Korea
Taiwan
United Arab Emirates
U.K.
U.S.

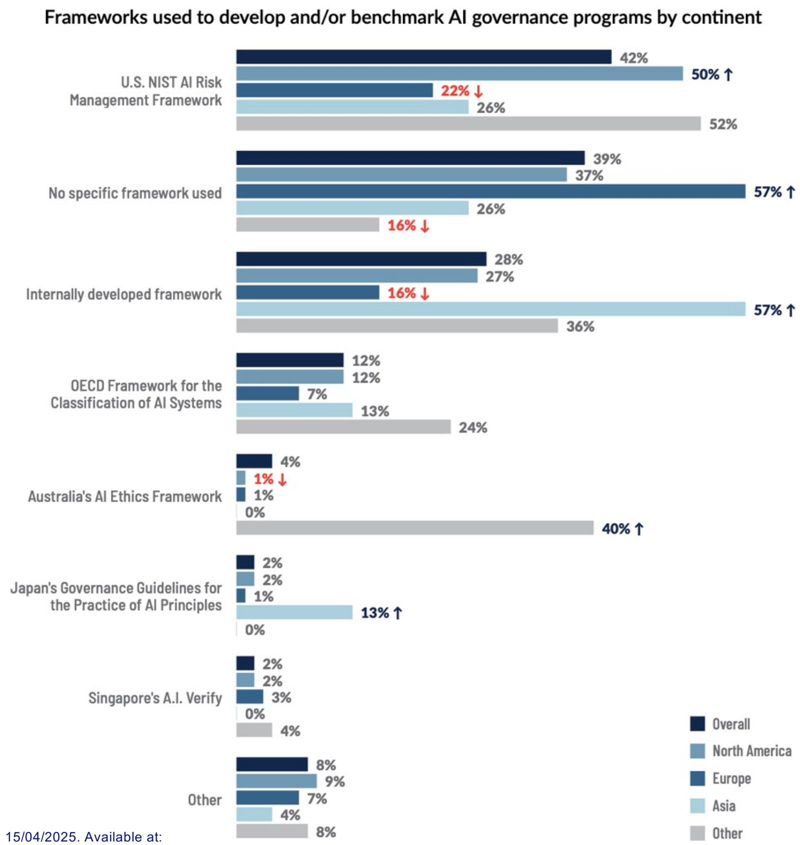
Region	Sovereignty	Assets
N. America	USA	Algorithmic Accountability Act (draft), NIST AI Risk Management Framework, E.O. 13960, E.O. 14110, Blueprint for an AI Bill of Rights, AI Safety Institute
	Canada	AI & Data Act (proposed), GenAI Code of Practice,
APAC	Australia	AI Ethics Framework, AI Ethics Principles, AI Standards Roadmap
	China	AI Guidelines, Summary of regulations
	India	Digital India Act, 2023 (proposed), India AI program
	Japan	Social Principles of Human-Centric AI
	New Zealand	Algorithm Charter, Trustworthy AI in Aotearoa principles
	Singapore	Model AI Governance Framework, VerifyAI, Model AI Governance Framework for Generative AI
LATAM	Argentina	Provision 2/2023 (published), Law 27,699, Resolution 161/23
	Brazil	Bill No. 2338/2023 (proposed)
	Chile	National Policy and Action Plan on AI
EMEA	EU	EU AI Act (approved)
	UK	AI Standards, AI Standards Hub
Global Agreements & Standards	Bletchley	Bletchley Declaration (AI Safety)
	G7	Hiroshima AI Process, AI Code of Conduct
	OECD	AI Principles
	UNESCO	AI Ethics

Source: International Association of Privacy Professionals (IAPP), "Global AI Law and Policy Tracker," Accessed on 15/04/2025. Available at: https://iapp.org/resources/article/global-ai-legislation-tracker/

Use of AI Governance **FRAMEWORKS**

- Organizations primarily use governmental frameworks, notably the U.S. NIST AI Risk Management Framework (42%) and internally developed frameworks (28%); there's a notable overlap where 60% using NIST AI RMF also employ the NIST Privacy Framework, reflecting privacy governance's maturity.

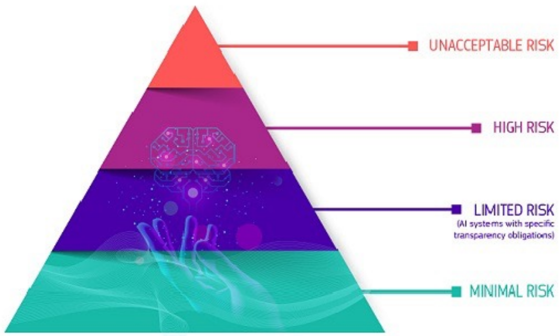
- Regional Differences: Framework choice strongly correlates with geography—NIST AI RMF usage is high in North America, Japan's AI governance guidelines dominate in Asia, and Europe notably lacks a specific AI governance framework (57% use none), likely awaiting the EU AI Act for regulatory clarity.



Source: International Association of Privacy Professionals (IAPP) & EY, "IAPP-EY Professionalizing Organizational AI Governance Report," Accessed on 15/04/2025. Available at: https://iapp.org/media/pdf/resource_center/iapp_ey_professionalizing_organizational_ai_governance_report.pdf

AI Governance frameworks and **REGULATIONS**

In the rapidly evolving AI landscape, businesses face new rules and standards to ensure trustworthy and responsible AI. Organizations must navigate both government regulations and voluntary frameworks to manage AI risks and compliance.



EU AI Act

Source: [EU AI Act - Official Portal]

NIST Risk Management Framework

Source: [NIST AI Risk Management Framework]

ISO/IEC 42001

Source: [ISO/IEC 42001 – AI Management System Standard]

NIST AI Risk Management Framework (AI RMF)

Overview
The NIST AI Risk Management Framework provides organizations with a structured, risk-based approach to identify, assess, and mitigate AI risks throughout the AI lifecycle.

Core Requirements

- **Risk Identification & Assessment:** Systematically map AI-related risks, from bias to security vulnerabilities.
- **Continuous Monitoring:** Implement iterative processes to update and mitigate risks over time.
- **Four Functional Pillars:** A core feature of the NIST AI RMF is its four functional pillars: Govern, Map, Measure, and Manage

NIST Risk Management Framework

Source: *NIST AI Risk Management Framework*

Business Impact

- Mitigates risks proactively, builds stakeholder confidence, and helps prepare for regulatory changes.

ISO/IEC 42001 AI Management System

Overview
ISO/IEC 42001 is the first international standard dedicated to AI management systems, providing a framework to govern AI processes and ensure responsible, transparent AI practices.

Core Requirements

- **Lifecycle Management:** Covers every stage of AI deployment—from design and testing to maintenance and decommissioning.
- **Ethical and Legal Integration:** Embeds ethical, legal, and technical controls to ensure AI operates within defined boundaries.
- **Continuous Improvement:** Emphasizes ongoing evaluation and enhancement of AI systems for improved accountability and performance.

ISO/IEC 42001

Source: ISO/IEC 42001 – AI Management System Standard

Business Impact

- Adoption of ISO/IEC 42001 demonstrates a commitment to responsible AI, enhances stakeholder trust, and provides a competitive edge in markets increasingly focused on ethical technology.

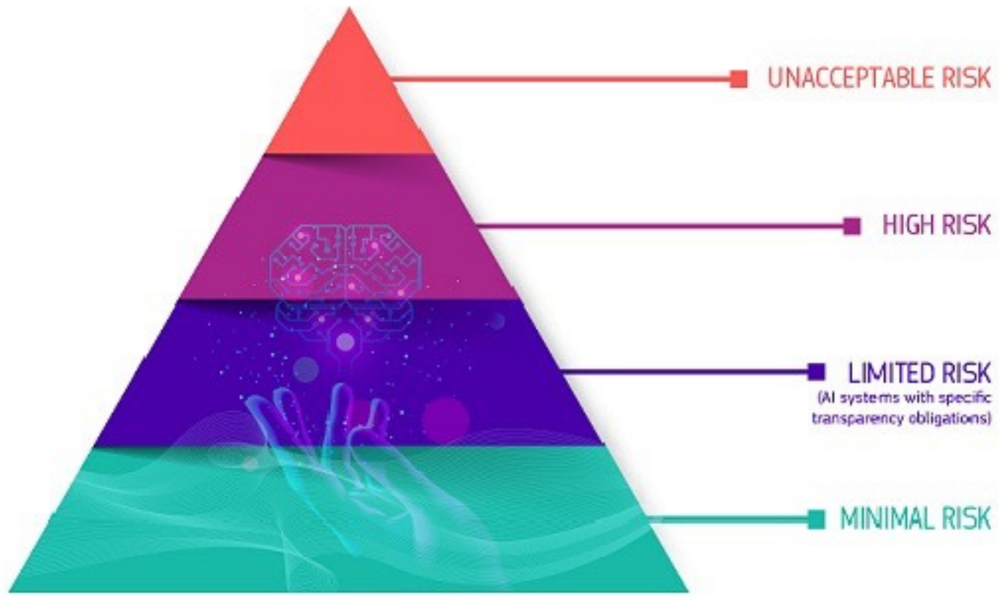
EU AI ACT (European Union Artificial Intelligence Act)

Overview

The EU AI Act establishes a comprehensive, risk-based framework for AI regulation in Europe. It classifies AI systems into risk categories—minimal, limited, high, and unacceptable—and imposes strict requirements for high-risk applications.

Core Requirements

- **Risk Assessment:** Organizations must rigorously evaluate the risks associated with high-risk AI systems according to 4 levels.
- **Transparency & Accountability:** Mandatory documentation, explainability, and human oversight are required.
- **Penalties:** Non-compliance can lead to severe fines (up to €30–40 million or 6–7% of global turnover).



EU AI Act

Source: [EU AI Act - Official Portal]

Business Impact

- Ensures AI deployments meet ethical, safety, and legal standards, protecting users and enhancing stakeholder trust.

EU AI Act - Case **CLASSIFICATION** Walkthrough

Applying the EU AI Act Risk Framework to Real Use Cases

Predictive Maintenance (Minimal-Risk)

Typically minimal-risk unless equipment is safety-critical component under Annex I. No AI Act-specific requirements; GDPR applies if processing personal data.

Retail Chatbot (Limited-Risk)

Chapter IV: Transparency. Must inform users they're interacting with AI. No QMS or conformity assessment required. GDPR applies if processing personal data.

CV Screening (High-Risk)

Annex III, Section 4: Employment. Provider: Risk management, data governance, technical documentation, QMS, conformity assessment, CE marking. Deployer: Human oversight, monitor for bias, log decisions, FRIA.

Credit Scoring (High-Risk)

Annex III, Section 5: Essential services. Provider: Full high-risk requirements including fairness testing, robustness validation. Deployer: Human review of adverse decisions, monitor across demographics, maintain logs.

Emotion Detection in Classroom (PROHIBITED)

Article 5: Emotion inference in educational institutions. Cannot be deployed; no compliance pathway exists.

Key governance **ARTIFACTS**

Effective AI governance relies on structured documentation artifacts that provide transparency, accountability, and evidence of compliance throughout the system lifecycle. These artifacts translate principles into operational controls and create an auditable governance trail.

Essential Documentation Artifacts

01 Classification Canvas

Structured template for categorizing AI systems by type, risk level, and regulatory applicability. Guides initial assessment and governance pathway selection.

02 Risk Register

Comprehensive log of identified risks, their likelihood, impact, mitigation strategies, and ownership. Updated continuously throughout the lifecycle.

03 TEVV Plan

Testing, Evaluation, Verification, and Validation plan detailing methodologies, metrics, test datasets, and acceptance criteria for model performance and fairness.

04 Model Cards

Standardized documentation of model details, intended use, performance metrics, limitations, and fairness evaluations. Facilitates transparency and informed deployment decisions.

05 Datasheets

Documentation of dataset characteristics, collection methodology, preprocessing steps, known biases, and recommended uses. Ensures data provenance and quality transparency.

06 RASCI Charts

Responsibility Assignment Matrix defining who is Responsible, Accountable, Supporting, Consulted, and Informed for each governance activity and decision point.

07 Technical Files

Comprehensive technical documentation including architecture diagrams, training procedures, validation results, and compliance evidence required for regulatory submissions.

08 PMM Logs

Post-Market Monitoring logs tracking operational performance, incidents, model drift, fairness metrics, and corrective actions taken during production deployment.

The **TECHNICAL** File (Annex IV)

Purpose: The technical file is the comprehensive documentation package demonstrating that a high-risk AI system meets all requirements. It must be maintained for **10 years** after the system is placed on the market and made available to authorities upon request. (Art. 18)

Required Contents (Annex IV)

System Description: Intended purpose, specifications, architecture, design choices

Risk Management: Risk identification, analysis, mitigation, residual risks

Data Governance: Training/validation/testing datasets, quality, bias mitigation

Testing & Validation: Test plans, procedures, results, robustness testing

Human Oversight: Design measures enabling effective oversight, user instructions

Performance Metrics: Accuracy, robustness, cybersecurity measures

Logging & Monitoring: Automatic logging capabilities, event recording, traceability

Post-Market Monitoring: Plan for ongoing monitoring, incident handling, corrective actions

QMS References: Links to quality management system documentation

Declaration of Conformity: EU declaration and CE marking evidence

Image Credit: Image generated by Midjourney, OpenAI.

Version Control: The technical file must be *updated* with substantial modifications and system updates.

Accessibility: Must be available to authorities upon request; language requirements apply.

Source: European Union, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on artificial intelligence (AI Act), Annex IV: Technical documentation referred to in Article 11(1). Available at: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Back to Agenda

Responsible AI
Ethical, Fair & Trustworthy

DAIG DATA & AI GOVERNANCE PARTNERS

What is **RESPONSIBLE** AI?

Definition (Virginia Dignum)

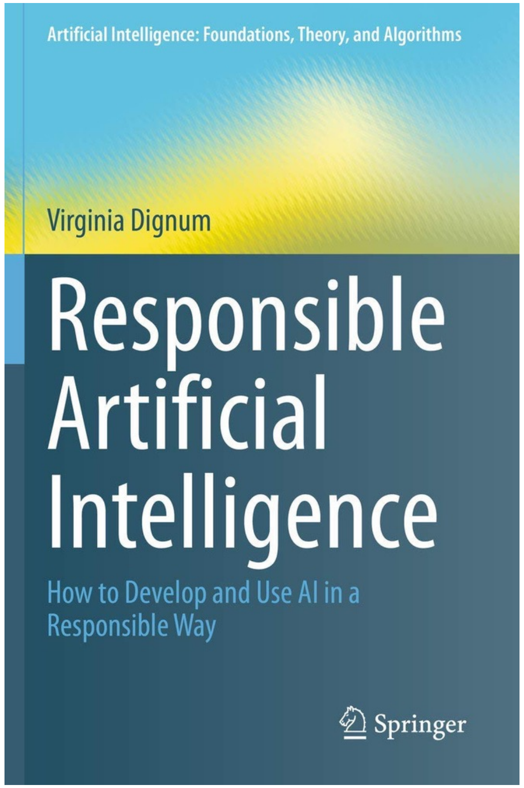
"Responsible AI is about **human responsibility** for the development of intelligent systems along fundamental human principles and values, to ensure human-flourishing and well-being in a sustainable world. … **Responsible AI is not about the characteristics of AI systems**, but about our own role. We are responsible for how we build systems, how we use systems and how much we enable these systems to decide and act by themselves."

Source: Virginia Dignum, Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way, Springer International Publishing, 2019. Available at: https://link.springer.com/book/10.1007/978-3-030-30371-6

Definition (SiliconANGLE)

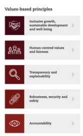
"Responsible AI is an **umbrella term** for aspects of making appropriate business and ethical choices when adopting AI. It encompasses **decisions around business and societal value, risk, trust, transparency, fairness, bias, mitigation, explainability, accountability, safety, privacy, regulatory compliance and more**. Before organizations design their AI strategy, they must define what responsible AI means within the context of their organization's environment."

Source: SiliconANGLE, "How IT leaders can embrace responsible AI," September 11, 2022. Available at: https://siliconangle.com/2022/09/11/leaders-can-embrace-responsible-ai/



Global AI Ethics **FRAMEWORKS**

Three Pillars of Global AI Ethics Standards



OECD AI Principles

1. Inclusive growth, sustainable development, and well-being
2. Human rights and democratic value, incl. fairness and privacy
3. Transparency and explainability
4. Robustness, security, and safety
5. Accountability

Adopted by 47 countries, these principles shape national AI strategies and create a foundation for cross-jurisdictional compliance.

Source: Organisation for Economic Co-operation and Development (OECD), "OECD AI Principles," Adopted May 2019 (updated 2024). Available at: https://www.oecd.org/en/topics/sub-issues/ai-principles.html

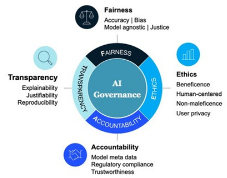
EU HLEG Requirements

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination, and fairness
6. Societal and environmental well-being
7. Accountability

Over 500 stakeholders contributed. Provides detailed assessment criteria for system-level evaluation.

Source: European Commission, High-Level Expert Group on Artificial Intelligence (AI HLEG), "Ethics Guidelines for Trustworthy AI," 8 April 2019. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

FEAT Principles



1. Fairness
2. Ethics
3. Accountability
4. Transparency

The FEAT framework helps guide the development and deployment of AI systems to ensure they are ethical, responsible, and trustworthy.

Source: Monetary Authority of Singapore (MAS), "Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector," 2018 (updated). Available at: https://www.mas.gov.sg/~/media/MAS/News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf

Image Source FEAT: https://medium.com/digital-mckinsey/using-the-feat-approach-to-avoid-biased-ai-f86471bf9d5b

Key Takeaway

These frameworks converge on similar themes, creating a common global language for **responsible AI and AI ethics** that facilitates international cooperation and regulatory alignment. Organizations implementing AI governance should map their controls to multiple frameworks to demonstrate comprehensive compliance and build stakeholder trust across jurisdictions.

Characteristics of **TRUSTWORTHY** AI Systems

Characteristics of trustworthy AI systems include: **valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed.** Creating trustworthy AI requires balancing each of these characteristics based on the AI system's context of use.

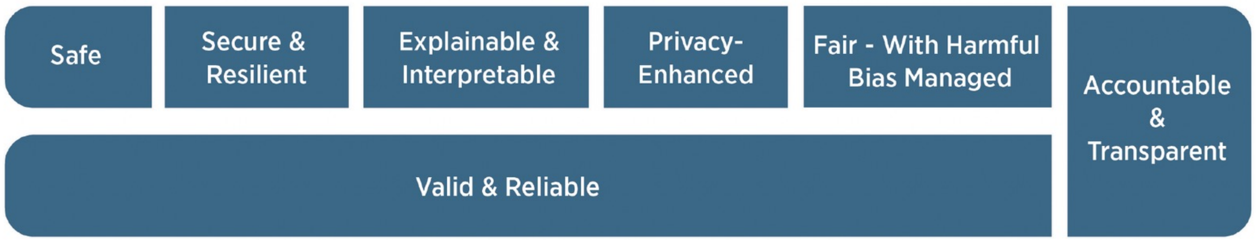


Fig. 4. Characteristics of trustworthy AI systems. Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable & Transparent is shown as a vertical box because it relates to all other characteristics.

Source: National Institute of Standards and Technology (NIST) , "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," January 2023. Available at: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

What exactly is AI **BIAS**?

- Bias, in general terms, represents a **skew or preference towards a particular outcome**, often unfair or prejudicial. It can arise from various sources and affect how information is perceived and decisions are made.

- In the context of Artificial Intelligence, bias refers to **systematic errors in AI outputs** due to flawed or unrepresentative data, flawed algorithms, or biased human input and interpretations. This can lead to AI systems making unfair, discriminatory, or inaccurate decisions.

- AI bias is not always intentional but can be **embedded in the data** that reflects existing societal inequalities or historical prejudices. Even well-intentioned AI development can inadvertently perpetuate these biases.

- **Recognising and understanding** the different forms and sources of AI bias is the crucial first step in mitigating its harmful effects. Without this foundational understanding, efforts to govern AI effectively will be undermined.

- Bias can manifest across the **entire AI lifecycle**, from data collection and preparation to model development, deployment, and even user interaction. Each stage presents opportunities for bias to be introduced or amplified.

How AI can perpetuate and amplify existing **BIASES**

- AI systems learn patterns from their training data; if this data contains biases, the AI will learn and repeat these biases in its outputs. This creates a cycle where **existing inequalities are automated and scaled**.
- The **speed and scale** at which AI systems can operate mean that biases can be amplified far more rapidly and widely than through traditional human decision-making processes. A single biased algorithm can impact millions of individuals.
- AI can sometimes **mask or obscure the underlying biases** in its decision-making processes, making it harder to identify and challenge unfair outcomes. This lack of transparency can exacerbate the problem.
- **Feedback loops in AI systems can further amplify biases**. For example, if a biased AI makes a decision that reinforces an existing inequality, the data generated by that decision can then be used to retrain the AI, further strengthening the bias.
- The **NIST "socio-technical" approach** to bias recognises that AI operates within a larger social context. Biases can originate not just from data but also from human thought and institutional practices, all of which can be reflected and amplified by AI.

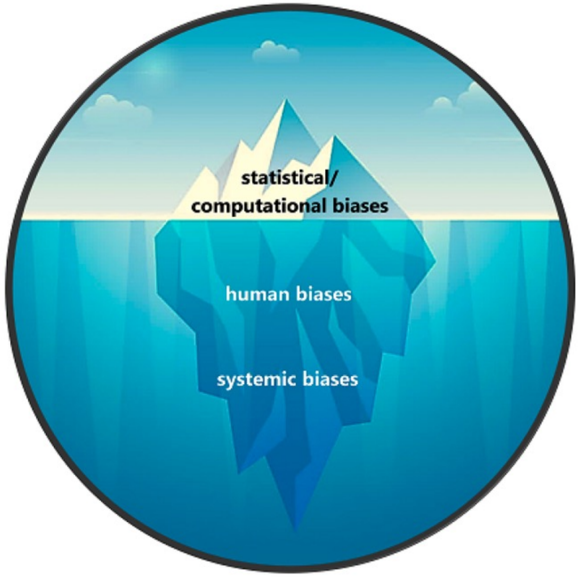


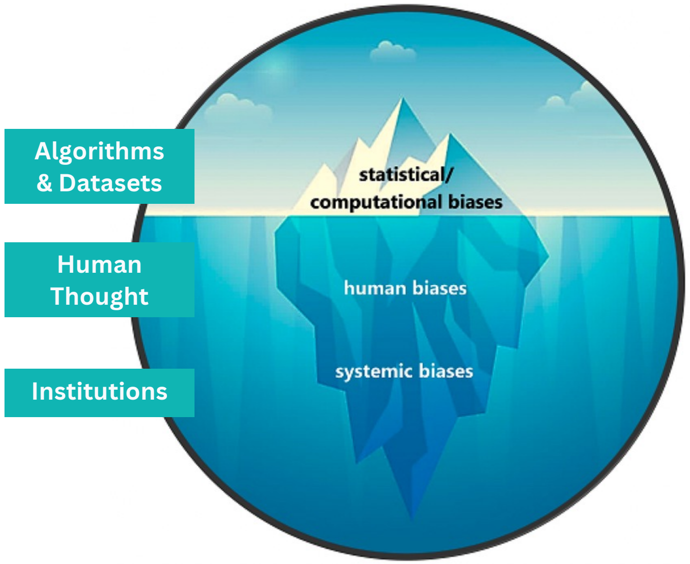
Fig. 1. The challenge of managing AI bias

Source: National Institute of Standards and Technology (NIST), "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," March 2022. Available at: https://doi.org/10.6028/NIST.SP.1270

DAIG DATA & AI GOVERNANCE PARTNERS

AI can perpetuate or amplify **BIAS**

The NIST "socio-technical" approach to mitigating bias in AI recognizes that AI operates in a larger social context.



Statistical & Computational Biases
- Stem from errors that result when the sample is not representative of the population
- These biases arise from systematic as opposed to random error and can occur in the absence of prejudice, partiality, or discriminatory intent

Technical

Human Biases
- Reflect systematic errors in human thought based on a limited number of heuristic principles and predicting values to simpler judgmental operations
- Often implicit & tend to relate to how an individual or group perceives information to make a decision or fill in missing or unknown information

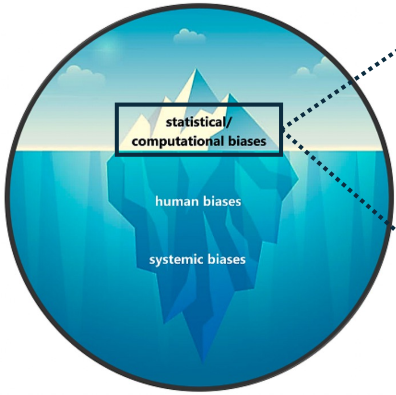
Systemic Biases
- Result from procedures and practices of institutions that operate in ways which result in certain social groups being advantaged or favored and others being disadvantaged or devalued
- Institutional racism and sexism are the most common examples

Social

Source: National Institute of Standards and Technology. (2022). Towards a standard for identifying and managing bias in artificial intelligence (NIST Special Publication 1270). https://doi.org/10.6028/NIST.SP.1270

Types of STATISTICAL & COMPUTATIONAL bias in AI (1 of 2)

Relate to errors that result when the sample is not representative of the population.



Algorithms & Datasets

Processing/Validation

1. amplification – Arises when the distribution over prediction outputs is skewed in comparison to the prior distribution of the prediction target
2. error propagation – Arises when applications built with ML are used to generate inputs for other ML algorithms
3. inherited – Arises when applications built with ML are used to generate inputs for other ML algorithms
4. model selection – Introduced while using the data to select a seemingly "best" model from a large set of models employing many predictor variables; or, when an explanatory variable has a week relationship with the response variable
5. survivorship – Tendency for people to focus on the items, observations, or people that "survive" or make it past a selection process, while overlooking those that did not

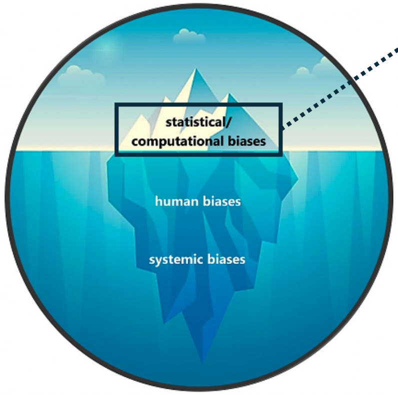
Selection and Sampling

6. data generation – Arises from the addition of synthetic or redundant data samples to a dataset
7. detection – Systematic differences between groups in how outcomes are determined; may cause an over- or under-estimation of the size of the effect
8. ecological fallacy – When an inference is made about an individual based on their membership within a group
9. evaluation – When the testing or external benchmark populations do not equally represent the various parts of the user population or from the use of performance metrics that are not appropriate for the way in which the model will be used
10. exclusion – When specific groups of user populations are excluded from testing and subsequent analyses
11. measurement – Arises when features and labels are proxies for desired quantities, potentially leaving out important factors or introducing group or input-dependent noise that leads to differential performance
12. popularity – Occurs when items that are more popular are more exposed and less popular items are under-represented
13. population – Systematic distortions in demographics or other user characteristics between a population of users represented in a dataset or on a platform and some target population
14. representation – Arises due to non-random sampling of subgroups, causing trends estimated for one population to not be generalizable to data collected from a new population
15. Simpson's Paradox – A statistical phenomenon where the marginal association between two categorical variables is qualitatively different from the partial association between the same two variables after controlling for one or more other variables
16. temporal – Arises from differences in populations and behaviors over time

Source: National Institute of Standards and Technology. (2022). Towards a standard for identifying and managing bias in artificial intelligence (NIST Special Publication 1270). https://doi.org/10.6028/NIST.SP.1270

Types of **STATISTICAL** & **COMPUTATIONAL** bias in AI (2 of 2)

Relate to errors that result when the sample is not representative of the population.



Algorithms & Datasets

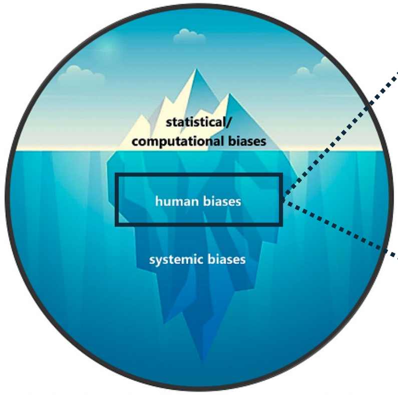
Use and Interpretation

17. uncertainty – Arises when predictive algorithms favor groups that are better represented in the training data, since there will be less uncertainty associated with those predictions
18. activity – Occurs when systems/platforms get their training data from their most active users, rather than those less active (or inactive)
19. concept drift – Use of a system outside the planned domain of the application (a common cause of performance gaps between laboratory settings and the real world)
20. content production – Arises from structural, lexical, semantic, and syntactic differences in the contents generated by users
21. data dredging – A statistical bias in which testing huge numbers of hypotheses of a dataset may appear to yield statistical significance even when the results are statistically nonsignificant
22. emergent – Use of a system outside the planned domain of application (a common cause of performance gaps between laboratory settings and the real world)
23. feedback loop – Effects that may occur when an algorithm learns from user behavior and feeds that behavior back into the model
24. linking – Arises when network attributes obtained from user connections, activities, or interactions differ and misrepresent the true behavior of the users

Source: National Institute of Standards and Technology. (2022). Towards a standard for identifying and managing bias in artificial intelligence (NIST Special Publication 1270). https://doi.org/10.6028/NIST.SP.1270

Types of HUMAN biases in AI (1 of 2)

Relate to how people use data to fill in missing information.



Human Thoughts

Group

25. deployment – Arises when systems are used as decision aids for humans, since the human intermediary may act on predictions in ways that are typically not modeled in the system.
26. funding – Arises when biased results are reported in order to support or satisfy the funding agency or financial supporter of the research study, but It can also be the individual researcher.
27. groupthink – A psychological phenomenon that occurs when people in a group tend to make non-optimal decisions based on their desire to conform to the group, or fear of dissenting with the group.
28. sunk cost fallacy – A human tendency where people opt to continue with an endeavor or behavior due to previously spent or invested resources, such as money, time, and effort, regardless of whether costs outweigh benefits.

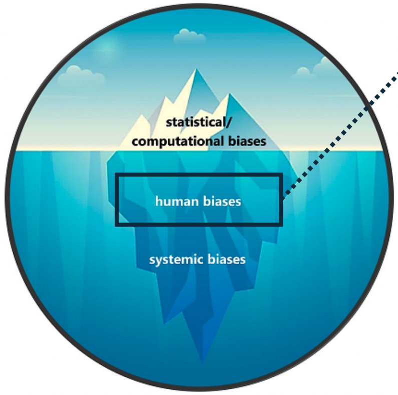
Individual

29. anchoring – A cognitive bias, the influence of a particular reference point or anchor on people's decisions.
30. annotator reporting – When users rely on automation as a heuristic replacement for their own information seeking and processing.
31. automation complacency – When humans over-rely on automated systems or have their skills attenuated by such over-reliance (e.g., spelling and autocorrect or spellcheckers).
32. availability heuristic – (also referred to as availability bias) A mental shortcut whereby people tend to overweight what comes easily or quickly to mind, meaning that what is easier to recall – e.g., more "available" – receives greater emphasis in judgement and decision-making.
33. behavioral – Systematic distortions in user behavior across platforms or contexts, or across users represented in different datasets.
34. cognitive – A broad term referring generally to a systematic pattern of deviation from rational judgement and decision-making. A large variety of cognitive biases have been identified over many decades of research in judgement and decision-making, some of which are adaptive mental shortcuts known as heuristics.
35. confirmation – (also referred to as confirmatory bias) A cognitive bias where people tend to prefer information that aligns with, or confirms, their existing beliefs.
36. consumer – Arises when an algorithm or platform provides users with a new venue within which to express their biases, and may occur from either side, or party, in a digital interaction.

Source: National Institute of Standards and Technology. (2022). Towards a standard for identifying and managing bias in artificial intelligence (NIST Special Publication 1270). https://doi.org/10.6028/NIST.SP.1270

Types of **HUMAN** biases in AI (2 of 2)

Relate to how people use data to fill in missing information.



Human Thoughts

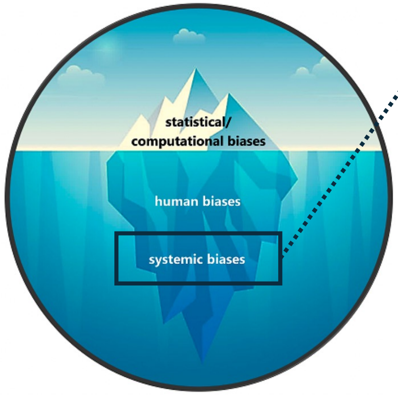
Individual *(continued)*

37. Dunning-Kruger effect – A cognitive bias, the tendency of people with low ability in a given area or task to overestimate their self-assessed ability.
38. human reporting – When users rely on automation as a heuristic replacement for their own information seeking and processing.
39. implicit – An unconscious belief, attitude, feeling, association, or stereotype that can affect the way in which humans process information, make decisions, and take actions
40. interpretation – A form of information processing bias that can occur when users interpret algorithmic outputs according to their internalized biases and views.
41. loss of situational awareness – When automation leads to humans being unaware of their situation such that, when control of a system is given back to them in a situation where humans and machines cooperate, they are unprepared to assume their duties.
42. mode confusion – When modal interfaces confuse human operators, who misunderstand which mode the system is using, taking actions which are correct for a different mode but incorrect for their current situation.
43. presentation – Biases arising from how information is presented on the Web, via a user interface, due to rating or ranking of output, or through users' own self-selected, biased interaction.
44. ranking – (a form of anchoring bias) The idea that top-ranked results are the most relevant and important and will result in more clicks than other results.
45. Rashomon effect or principle – Refers to differences in perspective, memory and recall, interpretation, and reporting on the same event from multiple persons or witnesses.
46. selective adherence – Decision-makers' inclination to selectively adopt algorithmic advice when it matches their pre-existing beliefs and stereotypes.
47. streetlight effect – A bias whereby people tend to search only where it is easiest to look.
48. user interaction – Arises when a user imposes their own self-selected biases and behavior during interaction with data, output, results, etc.

Source: National Institute of Standards and Technology. (2022). Towards a standard for identifying and managing bias in artificial intelligence (NIST Special Publication 1270). https://doi.org/10.6028/NIST.SP.1270

Types of SYSTEMIC biases in AI

Relate to institutions that operate in ways that disadvantage/disvalue certain social groups.



Institutions

Systemic

49. historical – Long-standing biases encoded in society over time.
50. institutional – A tendency exhibited at the level of entire institutions, where practices or norms result in the favoring or disadvantaging of certain social groups.
51. societal – (also referred to as social bias) Can be positive or negative, and take a number of different forms, but is typically characterized as being for or against groups or individuals based on social identities, demographic factors, or immutable physical characteristics.

Source: National Institute of Standards and Technology. (2022). Towards a standard for identifying and managing bias in artificial intelligence (NIST Special Publication 1270). https://doi.org/10.6028/NIST.SP.1270

How biases contribute to **HARMS**

Fig. 5 provides examples of how the three categories of bias—systemic, statistical and computational, and human - interact and contribute to harms within the data and processes used in AI applications, and the validation procedures for determining performance.

	Systemic Biases	Statistical and Computational Biases	Human Biases
Datasets *Who is counted, and who is not counted?*	Issues with latent variables Underrepresentation of marginalized groups	Sampling and selection bias Using proxy variables because they are easier to measure Automation bias	Observational bias (streetlight effect) Availability bias (anchoring) McNamara fallacy
Processes and Human Factors *What is important?*	Automation of inequalities Underrepresentation in determining utility function Processes that favor the majority/minority Cultural bias in the objective function (best for individuals vs best for the group)	Likert scale (categorical to ordinal to cardinal) Nonlinear vs linear Ecological fallacy Minimizing the L1 vs. L2 norm General difficulty in quantifying contextual phenomena	Groupthink leads to narrow choices Rashomon effect leads to subjective advocacy Difficulty in quantifying objectives may lead to McNamara fallacy
TEVV *How do we know what is right?*	Reinforcement of inequalities (groups are impacted more with higher use of AI) Predictive policing more negatively impacted Widespread adoption of ridesharing/self-driving cars/etc. may change policies that impact population based on use	Lack of adequate cross-validation Survivorship bias Difficulty with fairness	Confirmation bias Automation bias

Fig. 5. How biases contribute to harms

Source: National Institute of Standards and Technology. (2022). Towards a standard for identifying and managing bias in artificial intelligence (NIST Special Publication 1270). https://doi.org/10.6028/NIST.SP.1270

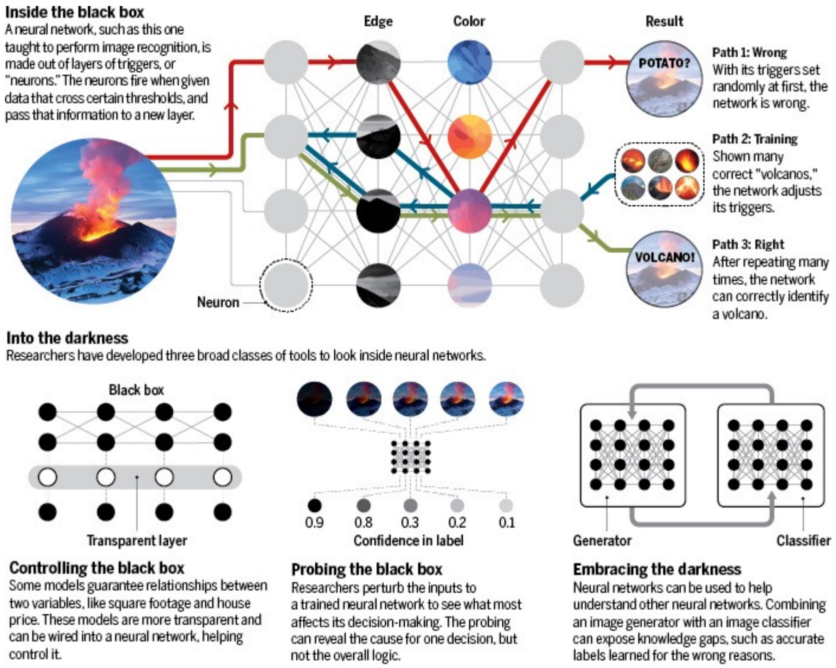
BLACK-BOX decision-making

What Are Black-Box AI Models?

AI models whose internal workings are opaque; they produce outputs without providing insight into the decision-making process.
Often based on complex neural networks with many hidden layers.

Challenges in Understanding & Explaining AI Decisions:

- **Interpretability:** Difficult to trace how specific inputs lead to outputs, complicating error analysis and debugging.
- **Transparency:** Limited visibility into model behavior hinders explanation and verification.
- **Accountability:** Challenges in understanding decision rationale make it hard to hold systems or developers accountable.
- **Impact on Stakeholder Trust & Accountability:**
  - **Reduced Trust:** Opaque decision-making erodes confidence among users, regulators, and investors.
  - **Regulatory Risks:** Non-transparent AI can lead to compliance issues with emerging explainability standards and legal requirements.
  - **Operational Risks:** Difficulties in auditing and improving models may lead to persistent errors and reputational damage.



Source: https://www.science.org/content/article/how-ai-detectives-are-cracking-open-black-box-deep-learning

NIST's Four Principles of EXPLAINABLE AI (XAI)

Foundational Framework for Designing Effective Explanation Systems

PRINCIPLE 1

Explanation

The system should provide evidence or reasons for its outputs. This establishes the baseline expectation that AI systems can articulate the basis for their decisions or predictions in some form accessible to relevant stakeholders.

→ *Local attributions (LIME, SHAP)*

PRINCIPLE 2

Meaningful

Explanations should be understandable and useful to the intended audience. Meaningfulness requires tailoring explanation format, detail level, and technical sophistication to match the knowledge, goals, and decision-making context of specific user groups.

→ *Model cards, stakeholder-specific patterns*

PRINCIPLE 3

Explanation Accuracy

What is asserted in the explanation should be correct for the model and case at hand. This addresses the critical challenge that many XAI techniques produce approximations or simplifications that may not faithfully represent actual model behavior.

→ *Calibration plots, fidelity metrics, stability testing*

PRINCIPLE 4

Knowledge Limits

The system should identify when it is likely to error when inputs fall outside its scope of competence. Transparency about these limits enables appropriate use and prevents over-reliance on system outputs in contexts where they may be unreliable.

→ *Abstain/deferral policies, confidence thresholds*

Source: National Institute of Standards and Technology (NIST), "NIST Interagency/Internal Report (NIST IR 8312): Four Principles of Explainable Artificial Intelligence (XAI)," October 2021. Available at: https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf

AI outputs can also be impacted by **MODEL** issues

These challenges can surface during development, training, or deployment (and ongoing monitoring and retraining is critical).

1	**Hallucination**	• AI may perceive patterns or objects that are nonexistent or imperceptible to human observers • Outputs can be nonsensical or inaccurate
2	**Plagiarism**	• AI creates new patterns by synthesizing the many examples in their training data sets • Some output may be (too) identical to those inputs instead of new and unique
3	**Stagnation**	• After AI creates a model from its training data, that model may not change much • Retraining models over time helps them adapt and overcome being limited to their initial static state
4	**Stupidity**	• AI often makes mistakes with counting and abstract or contextual uses of math • AI thinks differently than humans do so its intelligence and stupidity will differ, too
5	**Quality**	• Low-quality data creates noise in the signals, increases processing, and reduces output quality • Use trusted, governed data inputs to reduce signals & increase AI effectiveness
6	**Replication**	• Deployed models using real-life data may not perform the same as during development or testing • Ensure training data sets are diverse and representative of real-life scenarios and retrain models
7	**Security**	• Attackers can ask the right questions to get data they want (bypassing attempts to keep it secured) • Conduct robust testing and ethical hacking to find and address any such holes before deployment

Many AI efforts face common **FAILURES**

Implementation can be challenging as AI approaches are still being developed and refined.

#		Failure	Lesson
1	**Misperceiving reality**	• Overestimation of “out of the box” functionality vs. aspirational capabilities • Organizational processes that are incompatible with AI approaches	• Manage expectations on the ground and in the C-Suite • Avoid overly ambitious efforts or disillusioning the organization/stakeholders
2	**Unrealistic expectations & budgets**	• Not understanding the problem or what’s needed for the solution • Inadequate budget for the complexity of the task • Lowballed estimates of the challenge	• Use the right resources/skills to identify scope and requirements • Plan for hidden storage and processing costs for versioning data sets • Provide appropriate resources (time, budget, staff) for the solution
3	**Overpromised functionality**	• New/emerging technologies may not be fully baked or cost-effective • Blurred lines between what’s theoretical vs. ready and practical • Underestimating effort and cost to reach the maturity of AI providers	• Experiment and innovate to determine how to apply AI technologies • Separate what is real and practical within a short time frame • Incorporate what’s possible in the distant future into a guiding vision
4	**Incorrect resources**	• Few to no internal resources already skilled in AI • Difficulty finding/recruiting specialized roles for AI	• Upskill internal resources that already know our customers & processes • Leverage the skills needed to solve the real problem
5	**Overly broad scope**	• Considering problems holistically may seem too broad to tackle • Difficulty executing in an agile or iterative fashion	• Start with high-level perspective and model a domain • Define scenarios that represent the broader scope plus detailed areas • Chunk the work into pieces that respect the overall vision
6	**Overly complex technology**	• M&A models often leverage discrete systems with costly integrations • Differing architectures, languages and approaches add cost and complexity	• Focus on using limited applications of AI/ML • Integrate specific functionality instead of an overly broad set of technologies
7	**Lack of training data**	• Data sources that are noisy and dirty • Throwing all the data at AI and expecting meaningful results	• Curate & structure data for the specific use cases under consideration • Consider manual cleanup of the data as well as tuning the algorithm
8	**Not laying the data foundation**	• Multiple disconnected initiatives with conflicting objectives or priorities • Focusing on cost or deadlines over building basic data discipline • Not looking past problems to see and understand the true underlying issues	• Lay the foundation data governance, data quality & a solid ontology • Measure organizational maturity & identify what’s actually possible • Define a roadmap of AI priorities aligned with business goals

Source: Earley, S., & Davenport, T. H. (2020). The AI-powered enterprise: Harness the power of ontologies to make your business smarter, faster, and more profitable. LifeTree Media.

Human-Centric AI: AUGMENTATION, Not Replacement

Preserving Human Autonomy and Meaningful Oversight

Human-centric AI means designing systems that **augment human capabilities** while respecting autonomy and dignity. AI should enhance human decision-making, not replace it. The goal is not to eliminate human judgment but to provide tools that enable better-informed, more consistent, and more efficient decisions while preserving meaningful human agency.

Image Credit: Image generated by Midjourney, OpenAI.

The Augmentation Principle

AI systems should be designed to complement human strengths and compensate for human limitations, not to bypass human judgment entirely. This requires understanding what humans do well (contextual reasoning, ethical judgment, handling novel situations) and what AI does well (processing large volumes of data, identifying patterns, maintaining consistency).

Human OVERSIGHT MODES: HITL, HOTL, HIC

Models of Human-AI Interaction

	Human-In-The-Loop (HITL)	Human-On-The-Loop (HOTL)	Human-In-Command (HIC)
Level of Human Involvement	Continuous; collaborative; active participation in decision-making	Supervisory; intervene only when necessary	Ultimate decision-making authority
AI Autonomy	Reliant on human handshake; varies; typically acts autonomously until human review is needed	Operates autonomously with human oversight	Can act autonomously but will never decide autonomously
Efficiency	Lower, due to the need for constant human input	Higher than HITL, balanced with oversight	Varies; prioritizes control over efficiency
Control & Safety	High control; allows for nuanced decisions	Balanced control; efficient for routine tasks	Maximum control
Typical Healthcare Use Cases	Clinical decision-making support; pathology analysis	Remote patient monitoring	Robotic surgery

Source: DeepScribe, "Optimizing Human-AI Collaboration: A Guide to HITL, HOTL, and HIC Systems," DeepScribe Resources. Available at: https://www.deepscribe.ai/resources/optimizing-human-ai-collaboration-a-guide-to-hitl-hotl-and-hic-systems?hss_channel=lcp-19018424

Model CARDS & System Factsheets

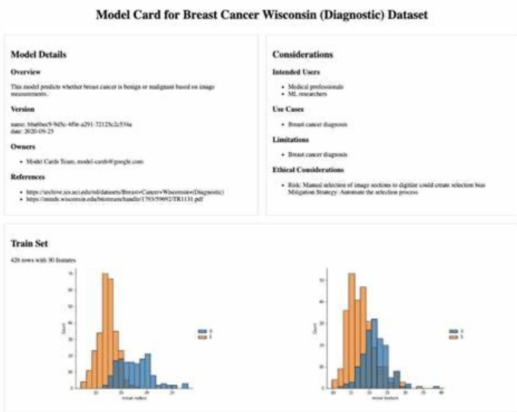
Model cards and system factsheets serve as **standardized transparency artifacts** that document AI system characteristics, performance, and limitations. These tools operationalize transparency principles by providing structured, accessible information to diverse stakeholders.

What Model Cards Provide

Model cards summarize intended use, training data characteristics, evaluation results (including disaggregated metrics across subgroups), known limitations, and ethical considerations. System factsheets extend this to system-level context: data lineage, oversight mode, escalation paths, recourse mechanisms, and change-control history.

Key components of a Model Card

Component
Model Overview
Intended Use
Model Architecture & Training Data
Performance Metrics
Evaluation Data
Ethical Considerations
Bias & Fairness Analysis
Safety, Security & Robustness
Limitations
Maintenance & Versioning

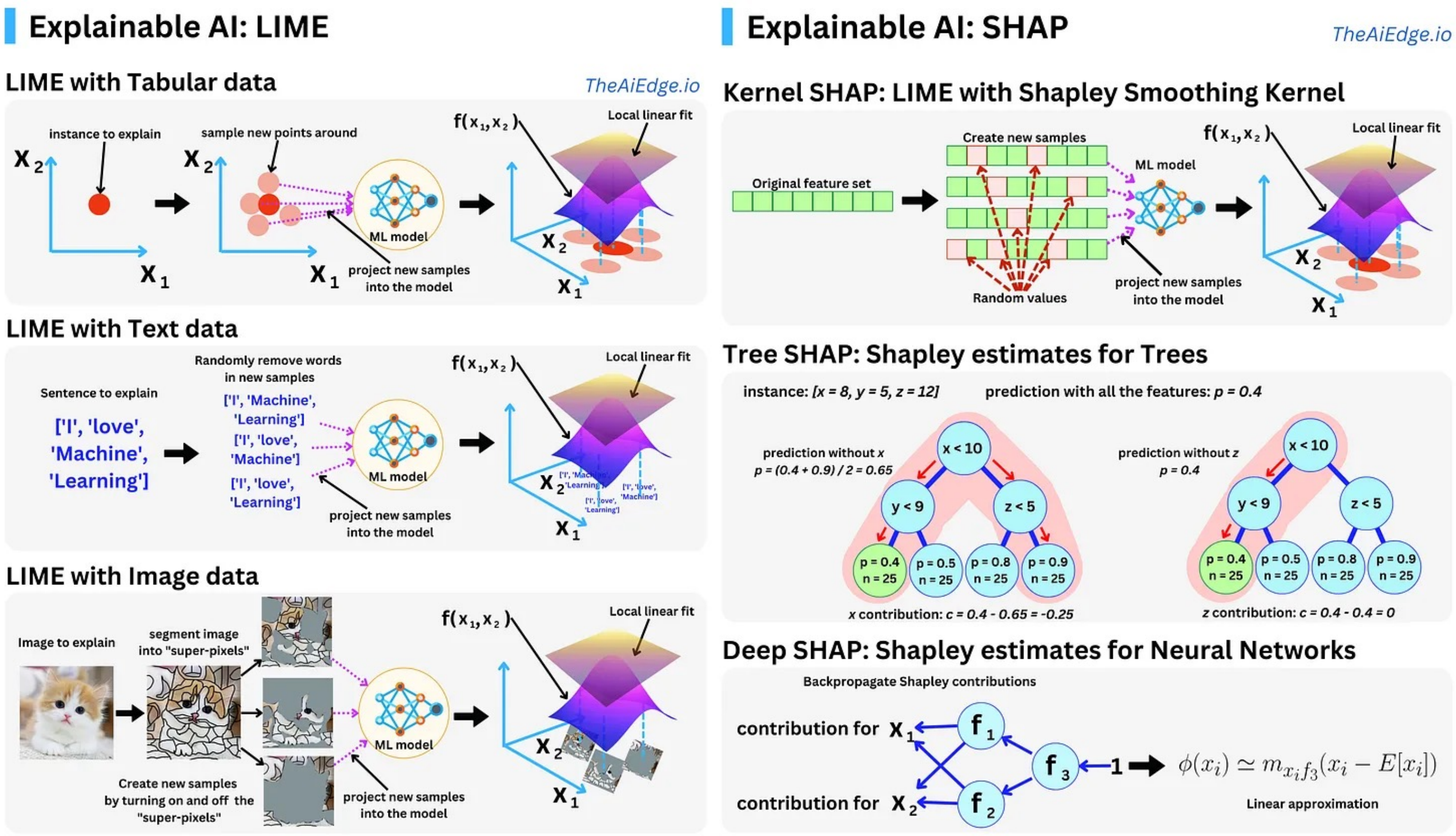


Capability **Benchmark**		**Gemini 2.5 Flash** Preview (04-17) Thinking	**Gemini 2.0 Flash** Non-thinking	**OpenAI o4-mini**	**Claude 3.7 Sonnet** 64k Extended thinking	**Grok 3 Beta** Extended thinking	**DeepSeek R1**
Reasoning & knowledge **Humanity's Last Exam (no tools)**		12.1%	5.1%	14.3%	8.9%	—	8.6%*
Science **GPQA diamond**	single attempt (pass@1)	78.3%	60.1%	81.4%	78.2%	80.2%	71.5%
	multiple attempts	—	—	—	84.8%	84.6%	—
Mathematics **AIME 2025**	single attempt (pass@1)	78.0%	27.5%	92.7%	49.5%	77.3%	70.0%
	multiple attempts	—	—	—	—	93.3%	—
Mathematics **AIME 2024**	single attempt (pass@1)	88.0%	32.0%	93.4%	61.3%	83.9%	79.8%
	multiple attempts	—	—	—	80.0%	93.3%	—
Code generation **LiveCodeBench v5**	single attempt (pass@1)	63.5%	34.5%	—	—	70.6%	64.3%
	multiple attempts	—	—	—	—	79.4%	—
Code editing **Aider Polyglot**		51.1% / 44.2% whole / diff	22.2% whole	68.9% / 58.2% whole / diff	64.9% diff	53.3% diff	56.9% diff
Factuality **SimpleQA**		29.7%	29.9%	—	—	43.6%	30.1%

Model-Agnostic **XAI TOOLS** – LIME & SHAP

	(Local Interpretable M
Definition	A method that explains **one** simple, local model around t
Mechanism	Perturbs (slightly changes) t trains a simple model (like a and uses that to say which f
Strengths	- Simple and intuitive idea - Fast and flexible - Works with almost any mo

Model-Agnostic **XAI TOOLS** – LIME & SHAP



Source: The AI Edge, “Explainable AI – LIME and…” The AI Edge Newsletter. Available at: https://newsletter.theaiedge.io/p/the-aiedge-explainable-ai-lime-and

Back to Agenda

Data Governance
Central Role of Data

DAIG DATA & AI GOVERNANCE PARTNERS

Data: A strategic **ASSET** powering AI advantage

AI is fundamentally dependent on data because it relies on historical examples and patterns within that data to learn, adapt, and make predictions.

- In the age of AI, data has evolved into a **key strategic asset.** Organizations that possess unique, high-quality, and well-managed data gain a significant competitive edge in developing superior AI solutions.
- AI can unlock hidden **insights and value** from data, leading to better decision-making, personalized customer experiences, and innovative products and services.
- Governing data effectively transforms it from a potential liability into a **valuable asset** that fuels AI-driven growth and innovation.
- Consider how **data rights** and **data sovereignty** impact the strategic use of data for AI initiatives.

Image Credit: Image generated by Midjourney, OpenAI.

Quality, Quantity, and Relevance in AI **DATA**

The critical trio.

- **Data Quality**: Accurate, consistent, complete, and timely data is essential for building reliable AI models. Poor data quality leads to flawed learning and inaccurate predictions.

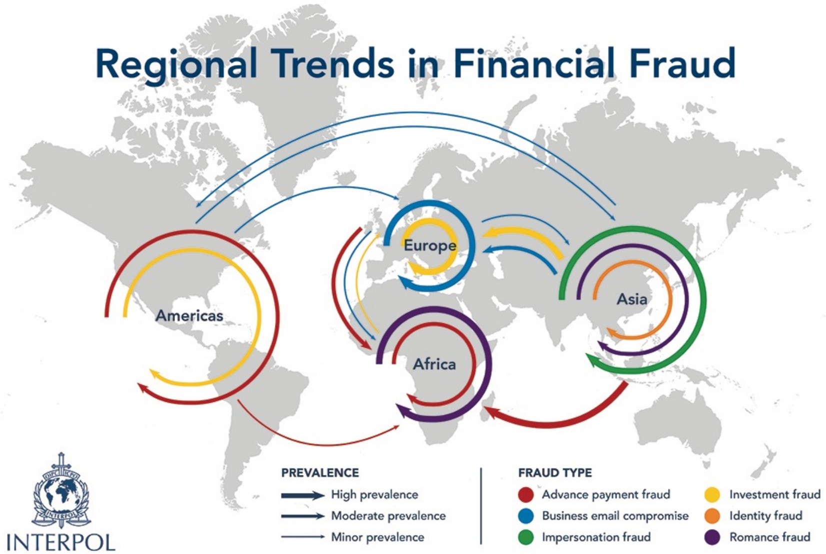
- **Data Quantity**: Many AI techniques, especially deep learning, require large volumes of data to learn complex patterns and generalise effectively. Insufficient data can lead to overfitting and poor performance on new data.

- **Data Relevance**: The data used to train an AI must be directly relevant to the task the AI is intended to perform. Irrelevant data can introduce noise and hinder the model's ability to learn meaningful relationships.

Example: Fraud Detection

An AI designed to detect fraudulent transactions needs historical transaction data that is **accurate** (correct amounts, dates), **complete** (all relevant fields filled), and **relevant** to fraudulent activity (including both fraudulent and legitimate transactions).

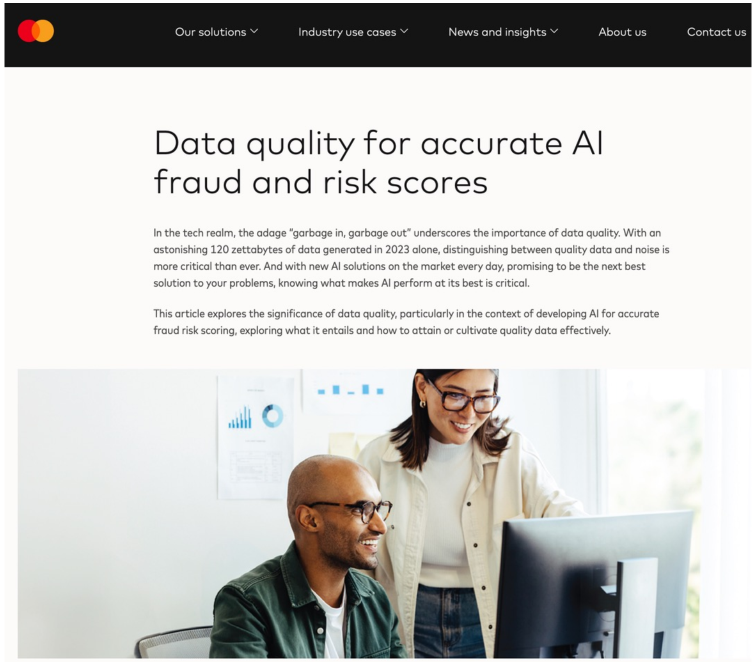
INTERPOL has identified a global surge in financial fraud, attributing the rise to technological advancements and the proliferation of AI and cryptocurrencies. The organization emphasizes the need for urgent action to address the increasing scale and sophistication of fraud affecting individuals, businesses, and governments worldwide.



Source: INTERPOL, "INTERPOL Financial Fraud assessment: A global threat boosted by technology" 2024. Available at: https://www.interpol.int/en/News-and-Events/News/2024/INTERPOL-Financial-Fraud-assessment-A-global-threat-boosted-by-technology

QUALITY of Data – The cornerstone of reliable AI

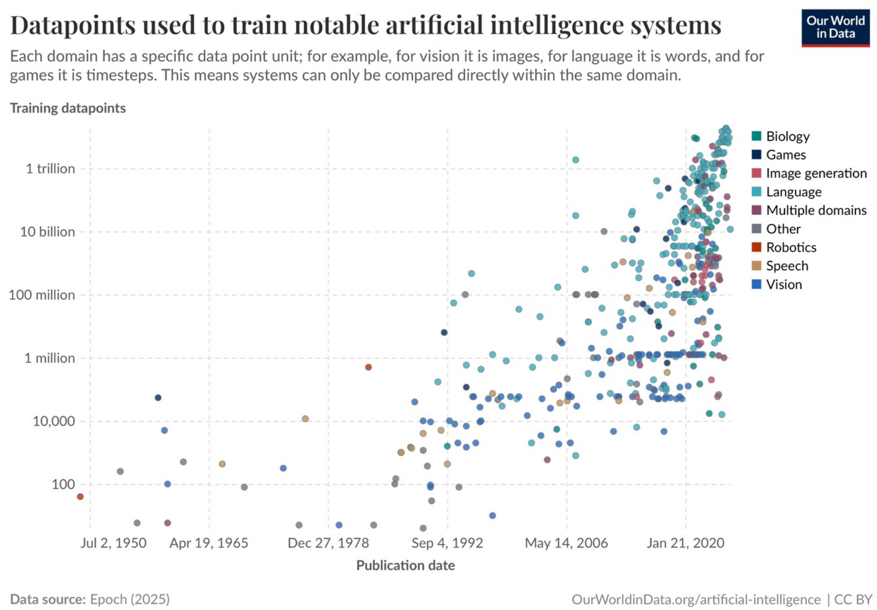
- **Data Quality** is paramount for effective AI. It encompasses accuracy, completeness, consistency, and timeliness.
- In fraud detection, high-quality data ensures the AI learns from correct transaction details, complete customer information, and consistent reporting formats.
- Inaccurate or incomplete data can lead to **false positives** (legitimate transactions incorrectly flagged as fraud) or **false negatives** (genuine fraud going undetected). This can result in customer frustration and financial losses.
- **Poor data quality masks data quality issues** for generative AI, making it harder to trust outputs.



Source: Mastercard, "Data quality for accurate AI fraud and risk scores," accessed: 15/04/2025. Available at: https://b2b.mastercard.com/news-and-insights/blog/data-quality-for-accurate-ai-fraud-and-risk-scores/

QUANTITY of Data – Fuel for effective AI models

- Many AI techniques, especially machine learning, require **large volumes of data** to learn complex patterns and generalize effectively.
- For fraud detection, a **substantial dataset** of both fraudulent and legitimate transactions allows the AI to discern subtle differences and identify sophisticated fraud schemes.
- **Insufficient data** can lead to overfitting, where the AI learns the training data too well and performs poorly on new, unseen transactions. It might fail to recognize new types of fraud.
- The **availability** of data for AI is crucial for data quality. Consider the need for historical data to train the model effectively.



Source: Our World in Data, Datapoints used to train notable artificial intelligence systems," 2025. Available at: https://ourworldindata.org/grapher/artificial-intelligence-number-training-datapoints

RELEVANCE of Data – Precision leads to actionable AI

- Data Relevance is critical; the data used to train the AI must contain **features and patterns directly related** to the phenomenon being predicted – in this case, fraudulent activity.
- For fraud detection, relevant data includes transaction amounts, timestamps, location information, user behaviour patterns, and device details.
- **Irrelevant data** can introduce **noise** and distract the AI from identifying the key indicators of fraud. Including unrelated demographic information, for example, might introduce bias without improving fraud detection accuracy.
- Consider the **target population** for the fraud detection model and ensure the training data is a good match.

Data is a strategic **ASSET**

BlueCross HealthCare

- **Vision Statement:** “To be a beacon of excellence in healthcare, providing qualitative, innovative, and patient-centered services that promote wellness and improve the quality of life for the communities we serve.“
- **Company Objectives:**
  - Deliver Exceptional Patient Care
  - Foster a Culture of Continuous Improvement
  - Expand Community Health Initiatives

Strategic Assets

Medical Equipment & Supplies	Patient Care & Work Locations	Employees & Contractors	Capital & Expenses & Revenue	Software & Hardware	Data & Analytics Solutions
					

Data is a strategic **ASSET**

Core Business Function

Supply Chain	**Facilities Management**	**Human Resources**	**Finance**	**Information Technology**	**Data Governance**
Medical Equipment & Supplies	Patient Care & Work Locations	Employees & Contractors	Capital & Expenses & Revenue	Software & Hardware	Data & Analytics Solutions

Strategic Assets






Mindset Shift: Data Governance is **MANAGING** data as an asset

Core Business Function

Supply Chain	Facilities Management	Human Resources	Finance	Information Technology	Data Governance
Strategic Assets					
Medical Equipment & Supplies	Patient Care & Work Locations	Employees & Contractors	Capital & Expenses & Revenue	Software & Hardware	Data & Analytics Solutions
Enterprise Capabilities					
Purchasing	Asset Management	Talent Acquisition	Budgeting	Hardware Mgmt.	Data Quality Mgmt.
Inventory Mgmt.	Plumbing	Performance Mgmt.	Forecasting	Cybersecurity	Master & Reference Data
Logistics & Distribution	Electricity	Benefits	Cost Accounting	Data Communications	Metadata
Supplier Mgmt.	Construction	Payroll	Capital Mgmt.	Networking	Analytics & BI
	Waste Mgmt.	Time & Attendance	Audit & Compliance	Application Support	Data Stewardship
	Safety & Compliance			Software Mgmt.	

Mindset Shift: Is AI Governance also **MANAGING** AI as an asset?

Core Business Function

Core Business Function	Strategic Assets	Enterprise Capabilities
Supply Chain	Medical Equipment & Supplies	Purchasing Inventory Mgmt. Logistics & Distribution Supplier Mgmt.
Facilities Management	Patient Care & Work Locations	Asset Management Plumbing Electricity Construction Waste Mgmt. Safety & Compliance
Human Resources	Employees & Contractors	Talent Acquisition Performance Mgmt. Benefits Payroll Time & Attendance
Finance	Capital & Expenses & Revenue	Budgeting Forecasting Cost Accounting Capital Mgmt. Audit & Compliance
Information Technology	Software & Hardware	Hardware Mgmt. Cybersecurity Data Communications Networking Application Support Software Mgmt.
Data Governance	Data & Analytics Solutions	Data Quality Mgmt. Master & Reference Data Metadata Analytics & BI Data Stewardship
AI Governance	AI Solutions	Risk assessment Control verification Risk monitoring Responsible AI

DAIG DATA & AI GOVERNANCE PARTNERS

Mindset Shift: Is AI Governance also **MANAGING** AI as an asset?

Human Resource Management

- Manages employee assets enterprise-wide
- Establishes hiring, promotion, and benefits policies
- Provides support, tools, and systems
- Clarifies roles and organizational structure
- Enables consistent HR practices
- Addresses complex employee issues

Finance Management

- Manages financial assets enterprise-wide
- Sets financial standards and compliance policies
- Provides budgeting and reporting tools
- Clarifies roles and financial accountability
- Enables consistent financial practices
- Addresses financial issues and compliance

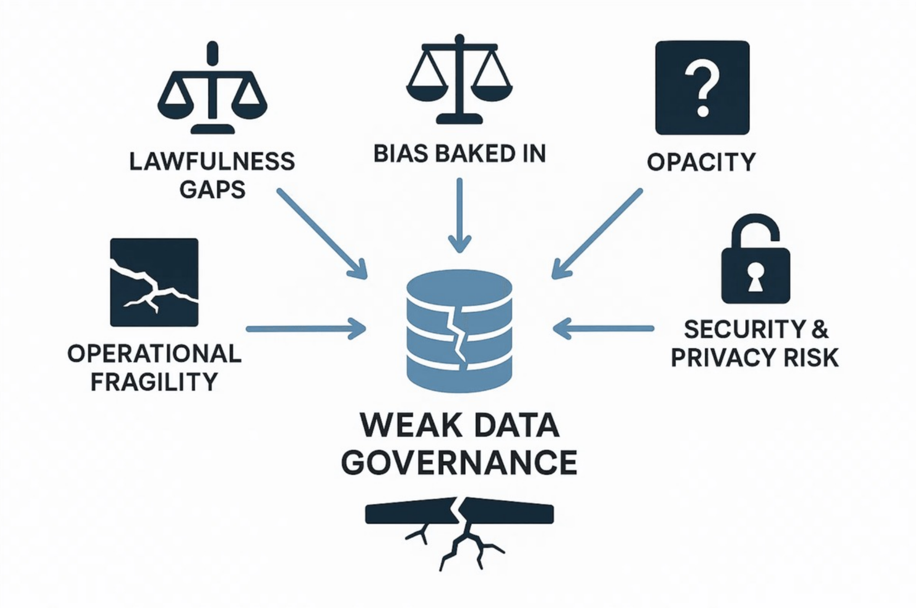
Data Governance

- Manages data assets enterprise-wide
- Defines standards, policies, and guidelines
- Provides templates, tools, and expertise
- Clarifies roles for data management
- Enables consistent data practices
- Addresses data privacy, security, compliance issues

Why DATA GOVERNANCE matters for AI

Poor data governance creates cascading failures across legal, ethical, operational, and technical dimensions.

Data governance is not merely a compliance exercise—it is the foundational capability that determines whether AI systems are lawful, fair, transparent, and trustworthy.



Weak data governance manifests in five interconnected **failure modes** that expose organizations to regulatory penalties, reputational damage, discriminatory outcomes, security breaches, and operational instability.

Five **FAILURE** modes in detail

Each failure mode represents a distinct dimension of risk arising from inadequate data governance. These modes are not mutually exclusive—organizations often **experience multiple failures** simultaneously, creating compounding regulatory and operational consequences.

Failure mode	Short description	How it manifests in practice	T
Lawfulness gaps	AI is built or used without satisfying legal/contractual requirements.	No clear lawful basis for using personal data; missing DPIA; breach of terms of use for scraped data; high-risk AI deployed without required documentation or oversight.	Regulatory ex inability to us system; loss o partners.
Bias baked in	Structural or statistical bias is embedded in data, labels, or model design.	Training data under-represents key groups; labels reflect human prejudice; no subgroup performance checks; model optimizes only for global accuracy, ignoring group disparities.	Discriminator equality and a reputational c trust.
Opacity	System behavior is not understandable or traceable to humans.	No documentation of model intent, data, or assumptions; lack of model cards or datasheets; no explanation interface for decisions; code and configuration changes unlogged.	Inability to ex users, regulat cause analysis accountability
		Weak access controls; no adversarial	Data breaches

DAIG PARTNERS

The amplified **IMPACT** of Data Governance on AI outcomes

- **Robust data governance is foundational** for responsible and effective AI. It provides the necessary framework for ensuring data quality, integrity, security, and ethical use in AI systems.
- **Poor data governance** significantly amplifies the risks associated with AI, including bias, inaccuracies, privacy violations, and lack of transparency.
- Effective data governance practices, such as **data quality** management, **metadata** management, and **data lineage** tracking, are crucial for building trust and explainability in AI models.
- Consider how data governance aligns with AI governance frameworks and **regulatory requirements**, such as the EU AI Act and the NIST AI Risk Management Framework.
- ***Example***: *Without proper data governance, an AI used for loan applications might unknowingly be trained on historical data containing gender or racial bias, leading to discriminatory outcomes.*

Image Credit: Image generated by Midjourney, OpenAI.

The tangible **COSTS** of poor Data Governance in AI initiatives

- **Increased development costs and delays** due to the need to clean, validate, and correct flawed data used for AI training.
- **Reduced accuracy and reliability** of AI models, leading to poor decision-making and potentially harmful outcomes.
- **Higher operational costs** associated with managing errors, rework, and customer dissatisfaction resulting from flawed AI outputs.
- **Significant reputational damage** and loss of customer trust due to biased, unfair, or inaccurate AI applications.
- **Increased legal and compliance risks**, including potential fines and penalties for violating data privacy regulations or deploying discriminatory AI systems.

Image Credit: Image generated by Midjourney, OpenAI.

Building a solid **FOUNDATION**

Integrating Data Governance throughout the AI lifecycle.

- **Embed data governance principles and practices** at every stage of the AI lifecycle, from data acquisition and preparation to model training, deployment, and ongoing monitoring.
- Establish clear data quality standards, metadata definitions, and data lineage tracking specifically for AI datasets.
- Define **roles and responsibilities** for data owners, data stewards, and AI practitioners to ensure accountability for data quality and ethical use in AI.
- Implement mechanisms for **continuous monitoring** and evaluation of data used in AI systems to detect and mitigate issues like bias, drift, and quality degradation.
- Foster a **data-centric culture** that recognises the critical role of well-governed data in driving successful and responsible AI innovation.

Image Credit: Image generated by Midjourney, OpenAI.

How AI Governance relies on strong **DATA GOVERNANCE**

- **Data Quality is Foundational for Effective AI Governance:** AI systems, especially complex models, depend on high-quality, trustworthy data for accurate predictions and reliable performance. Data governance ensures the quality, integrity, and appropriate handling of this crucial input.

- **Addressing Bias Requires Data Governance:** Many AI biases originate from the data used to train the models. Data governance practices, including careful data selection, bias identification, and mitigation strategies, are essential components of responsible AI and fall under the AI governance umbrella.

- **Transparency and Explainability are Enhanced by Data Governance:** Understanding the lineage, provenance, and characteristics of the data used by AI systems contributes significantly to their transparency and explainability. Data governance establishes the processes for documenting this vital information.

- **Regulatory Compliance Intersects**: Emerging AI regulations, such as the EU AI Act, explicitly mandate data governance and management practices for training, validation, and testing datasets used in high-risk AI systems. AI governance frameworks must incorporate these data governance requirements.

Image Credit: Image generated by Midjourney, OpenAI.

Back to Agenda

Connecting The Dots
A Practical Framework

DAIG
DATA & AI GOVERNANCE
PARTNERS

A practical Data & AI Governance **FRAMEWORK**

Six capabilities to govern data and AI — from strategy to execution

USE CASE DRIVEN

01 Foundations

Why you govern data and AI, what principles guide you, and where the boundaries are. Includes risk appetite and responsible AI commitments.

02 Structure & Roles

Who decides what — governance bodies, data owners, AI model owners, and stewardship networks. Clear escalation paths from use case teams to executive oversight.

03 Processes & Workflows

How issues, changes, and approvals flow — from data quality fixes to AI model reviews. Triggered by real business needs, not bureaucratic checklists.

04 Measures & Feedback

A small set of metrics that prove governance is helping — data quality, model performance, bias detection, adoption. Continuous feedback loops, not annual audits.

05 Communication & Enablement

How you explain governance, build data and AI literacy, increasing awareness, and equip people for new roles. Making responsible AI everyone's business, not just a policy.

06 Frontline Governance

Day-to-day accountability carried by real teams — data stewards, AI use case owners, and domain experts. Where governance meets execution.

Underpinned by Responsible AI — Fairness, Transparency, Accountability, Safety & Human Oversight woven into every capability

Consider assessing relevant AI **READINESS** factors

AI success hinges on several dependencies that should proactively be evaluated.

Cultural Readiness

- Leadership education on ML, AI and GenAI
- Defined AI strategy (use cases, goals, ROI)
- Dedicated resources and funding for AI
- Realistic expectations of what AI can achieve

Data Readiness

- Modern and AI-ready infrastructure
- Enterprise data architecture standards
- High data quality and data integrity
- Scalable data platform with diverse data sets

Workforce Readiness

- Clearly defined AI roles & responsibilities
- Upskill/source needed AI engineering talent
- Advanced analytics and data science skills
- Robust AI ethics, tool, and literacy training

Operational Readiness

- Budget for hidden costs (processing, storage)
- AI development and testing processes
- Data pipeline and integration processes
- AI operationalization and support processes

Governance Readiness

- Transparency of AI algorithms and outputs
- Discrete security and data privacy measures
- Regulatory compliance and ethical guidelines
- Strong data governance and model governance

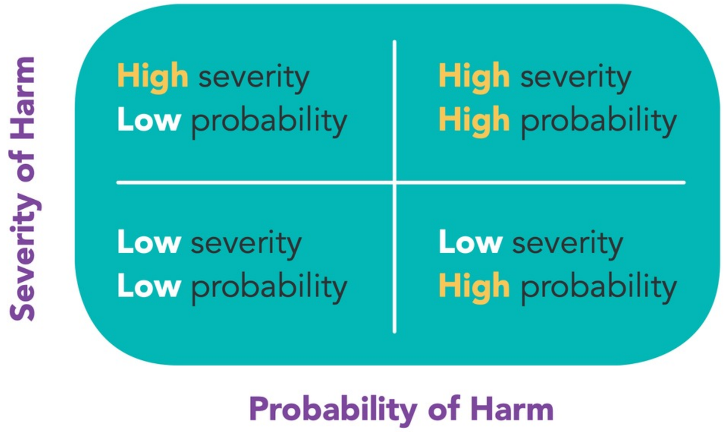
Consider completing a formal AI readiness assessment to baseline current state and measure improvement over time

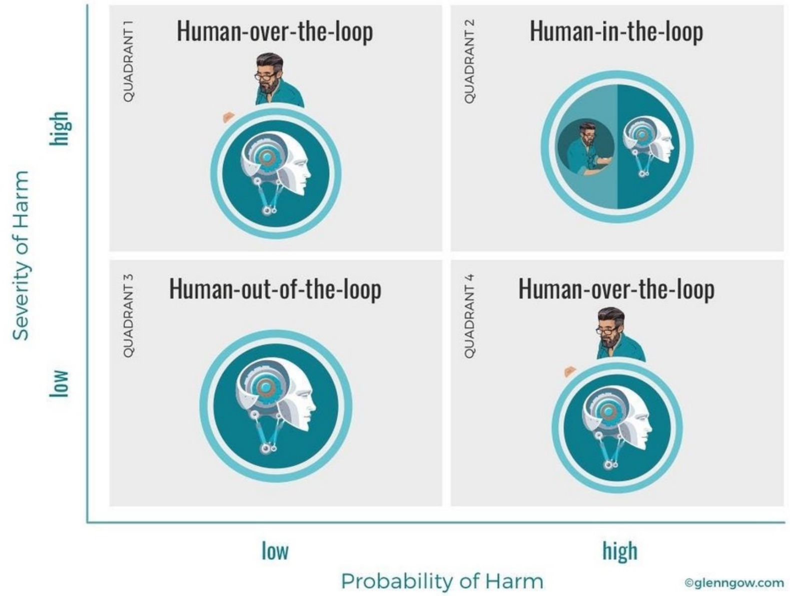
Source: Guidehouse. (2024, April 10). The state of GenAI today: The early stages of a revolution. https://guidehouse.com/news/advanced-solutions/2024/the-state-of-genai-today

Source: TDWI. (2024, April 15). TDWI AI readiness assessment guide. https://tdwi.org/research/2024/04/adv-all-tdwi-ai-readiness-assessment-guide.aspx

Consider the level of **HUMAN OVERSIGHT** for AI

Use a rubric to estimate the severity & probability of harm and determine the appropriate level of AI autonomy.





Other factors to consider:

- Nature of harm (physical or intangible)
- Reversibility / ability of humans to obtain recourse
- Operational feasibility / meaningfulness of involving a human

Choose the right level based on risk appetite

Source: Info-communications Media Development Authority & Personal Data Protection Commission. (2020). Artificial Intelligence Governance Framework Model: Second edition. https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf

Source: Gow, G. (2023, August 23). A simple AI governance framework in the age of ChatGPT. Forbes. https://www.forbes.com/sites/glenngow/2023/08/06/a-simple-ai-governance-framework-in-the-age-of-chatgpt/

Consider a **LEGALLY DEFENSIBLE** AI governance approach

Establish appropriate standards and controls now to maximize defensibility of our AI efforts as we mature.

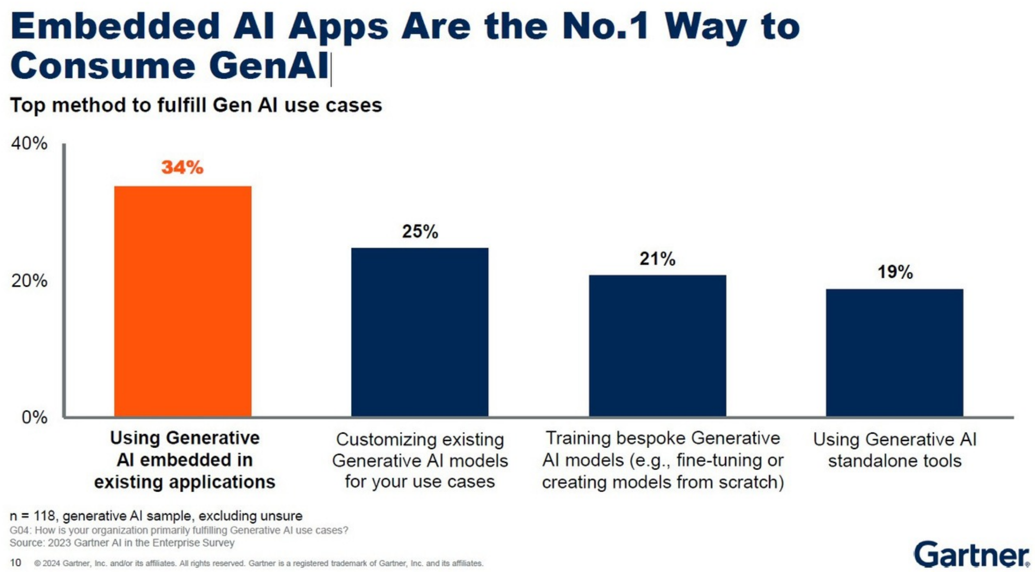
1	**Stay abreast of AI regulations**	• Some exist, more are coming – the EU AI Act is viewed by many as the primary model • Available regulations (and guidelines) for AI should be translated to internal policies & controls • Aim for the strongest requirements (where financially feasible) to foster global consistency & efficiency
2	**Using AI requires a formal framework**	• Consider an AI policy with official terms and definitions, lifecycle stages, development guidelines, etc. • Establish formal AI development processes and tollgates *(similar to DevOps for software/application development)* • Clarify roles and responsibilities with a RACI matrix and segregate duties where possible
3	**Transparency and explainability are key**	• Measure processes (tollgates and audits), roles (performance, compliance), and outputs/outcomes (quality, value) • Incorporate AI assessments into strategic planning, risk management, and auditing functions • Address consumer rights with informed consent, opt-out, and complaint intake/feedback processes
4	**The best AI needs the best data**	• AI uses large volumes of data which significantly increases the need for data privacy/security/protection controls • There's no AI without IA (information architecture) - which means strong data governance and data quality processes • AI is most efficient with high quality data and consistent, well-defined taxonomies, ontologies and metadata
5	**Adopting AI means culture change**	• AI is here to stay - we should embrace it proactively and foster AI literacy skills across the organization • An educated and properly trained workforce will (theoretically) create less risks with using AI • We need to promote culture change to promote managing data as an asset and using AI responsibly and ethically

Source: Milone, M. (2023, December 3). Legally defensible AI governance. Presentation at the Data Governance & Information Quality Conference (DGIQ) 2023 East, Washington, D.C.

Existing technologies are increasingly using **EMBEDDED** AI apps

AI solutions in systems and platforms already in use also need to be governed to manage risk.

- Adobe AI, AI Assistant, FireFly AI
- Apple Intelligence
- Atlassian Intelligence
- Google Gemini
- Grammarly for Windows
- Microsoft 365 Copilot
- Okta AI
- Oracle AI
- Red Hat Enterprise Linux AI
- Salesforce Einstein
- SAP Concur AI
- ServiceNow AI
- Tableau Pulse
- Zoom AI Companion



Identify existing AI **TEAMS** across the organization

Leverage accessible resources to find teams and members developing AI internally.

People Resources

- Leadership
- Org Charts
- Employee Directory
- HR Reports
- IT Managers
- Job Postings
- Skills Inventories
- LinkedIn Profiles

Productivity Tools

- Email (GALs)
- Company Intranet
- Collaboration Tools
- Teams etc.
- Slack, etc.
- Wikis
- Knowledge Bases
- Content Mgmt

Technical Resources

- Portfolio Mgmt
- Project Mgmt
- Project Repositories
- Helpdesk Tickets
- Application Portfolio
- Change Mgmt
- Technical Docs
- Training Content

SMEs & Horizontals

- Phone-A-Friend
- Procurement
- BI & Analytics
- Audit & Compliance
- Cyber Security
- Legal
- Policy
- Risk Mgmt

Identify existing AI **SOLUTIONS** across the organization

Assess assets, documentation, standards, and oversight processes across all identified AI teams.

Identify Solutions

- AI in Development
- Deployed AI
- Experimental AI
- Embedded AI
- Algorithms
- Models
- Chat Bots
- Vendors

Assess Documents

- Storage Locations
- Templates
- Use Cases & Req's
- Architecture Docs
- Model Specs
- Data Sources & Sets
- Audit & Test Results
- Performance Metrics

Confirm Standards

- Policies
- RACIs
- Decision Makers
- Best Practices
- Industry Standards
- Compliance Checks
- Risk Management
- Status Reporting

Identify Oversight

- Guiding Principles
- Ethics Principles
- Policies
- Plans & Goals
- Strategic Alignment
- Intake & Triage
- Project Priorities
- CAB Approvals

Identify existing AI **FRAMEWORKS**

Assess current standards and processes associated with AI development and usage.

Design Processes

- HITL Assessment
- Bias Assessment
- Risk Assessment
- Usage Guidelines
- Performance Metrics
- Data Requirements
- Tollgates
- Review Points

Development Processes

- Model Selection
- Data Collection
- Data Preprocessing
- Data Transformation
- Feature Engineering
- Model Training
- Model Tuning
- Risk Identification

Testing Processes

- Model Diagnostics
- Model Testing
- Model Refinement
- Bias Testing
- Security Testing
- Compliance Testing
- Error Analysis
- Audits

Deployment Processes

- Source Control
- Version Control
- Hand-Offs
- Model Monitoring
- Issue Tracking
- Model Retraining
- Lifecycle Mgmt
- Feedback

Identify relevant AI **LAWS** and regulations

Track proposed & enacted laws & how they apply to operational & geographical footprints.

Types of Laws

- AI-Specific
- Data Protection
- Consumer Protection
- Industry-Specific
- Liability & Safety
- Intellectual Property
- Patent & Copyright
- Telecommunications

Sovereignty & Jurisdiction

- Local
- Regional
- National
- Multi-National
- International
- Data Localization
- Model Registration
- Disclosures

Evidence of Compliance

- Documentation Audits
- Bias Detection
- Risk Mitigation
- Testing Results
- Performance Results
- Legal Reviews
- Compliance Audits
- Remediation Efforts

Continuous Monitoring

- New Laws
- Changes to Laws
- Policy Changes
- Impact Analysis
- Process Changes
- Communication
- Coordination
- Training & Awareness

Conduct an AI **MATURITY** assessment

Assess current state against frameworks & standards and define a roadmap to remediate gaps.

Areas of Assessment

- Standards
- Frameworks
- Info Architecture
- Documentation
- Model Quality
- Data Quality
- Skills & Resources
- Readiness

Short-Term Goals

- Strategy
- Policies
- Guidelines
- Governance Bodies
- Records Repository
- Audits & Reviews
- Risk Reduction
- Performance KPIs

Long-Term Goals

- Continuous Integration
- Business Value
- Benchmarking
- ESG & Scalability
- Crisis Response
- AI Literacy
- Ethics Board
- Chief AI Officer

Plans & Roadmaps

- Executive Briefings
- Culture Change
- Training & Education
- Knowledge Sharing
- Hands-On Workshops
- AI Stewards
- Certifications
- Center of Excellence

RECOMMENDATIONS for formalizing AI governance

Define key objectives for implementing enterprise AI governance (each with a bolus of more detailed work to accomplish).

1. Separate AI dev & operations frameworks vs. AI oversight & decision-making frameworks
2. Formalize an AI governance oversight team & purview
3. Formalize AI strategy, guiding principles & AI ethics principles
4. Formalize AI oversight structure & organizational model
5. Formalize AI policies for public, developed & embedded/third party AI
6. Formalize MLOps, AIOps, ModelOps frameworks & standards
7. Formalize controls & processes to comply with AI regulations
8. Formalize controls & processes to measure AI value, performance, bias, risks, etc.
9. Standardize AI terminology & publish terms to the enterprise glossary
10. Launch AI literacy training + consider an AI Center of Excellence
11. Consider AI governance certifications for staff
12. Consider AI certifications for the organization

Understand the burgeoning AI OPS functions

Concepts, lessons and benefits of established "Ops" functions are being adopted for AI.

IT

AIOps
AI for IT workflows, monitoring & service management

Analytics developers

DataOps
Data analytics dev & lifecycle

MLOps
ML model dev & testing

Data Scientists

ModelOps
Monitoring and support for all deployed AI solutions and their related dependencies

GRC & IT staff

Software developers

DevOps
Software development/engineering & lifecycle

InfoSec

DevSecOps
Security automation throughout development lifecycles

Design, develop, test, deploy, monitor

Typical capabilities:

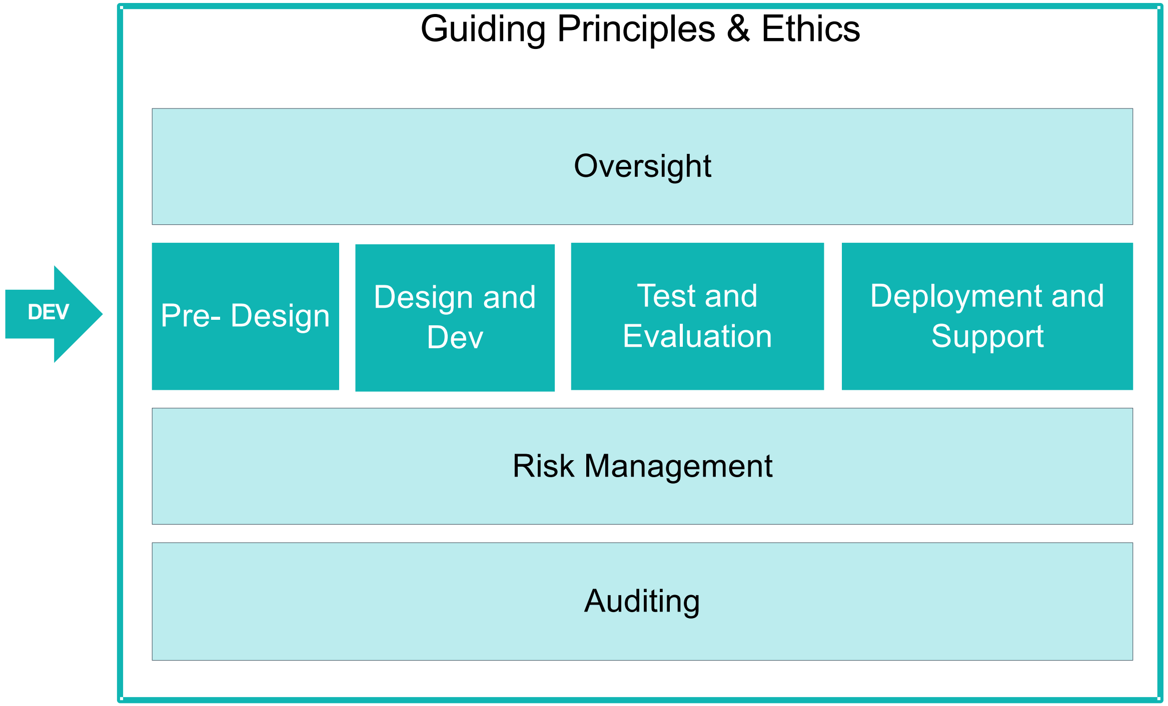
- request & approval process
- formal development framework
- dedicated environments
- segregation of duties
- robust testing & fine-tuning
- code versioning
- standardized deployment & rollback
- continuous integration/delivery
- formal documentation
- formal monitoring & support

BENEFITS

- improved collaboration
- faster deployment
- Increased compliance
- better auditability
- enhanced performance
- greater agility
- better resource utilization
- more automation
- higher reliability
- 360° enterprise visibility

Build an AI oversight & decision-making **FRAMEWORK**

Wrap an interconnected set of functions around AI “Ops” functions to proactively manage AI across the organization.



KEY: Ops Governance

1. Guiding Principles & Ethics
- Foster fair, responsible & trustworthy AI
- Promote AI ethics as a strategic imperative

2. Oversight
- Launch oversight team & define AI strategy
- Implement AI policies, standards & training

3. Pre-Design
- Align AI use cases with strategic goals
- Assess AI technical plans and potential risks

4. Design and Development
- Evaluate AI requirements and available data
- Define AI specs and initiate development

5. Test and Evaluation
- Engage robust testing and document metrics
- Fine tune model until performance is within specs

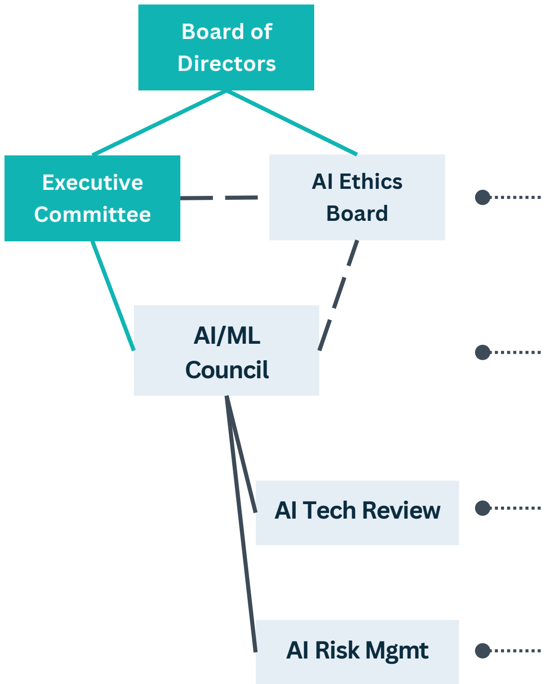
6. Deployment and Support
- Implement model and support processes
- Monitor AI outputs with humans & new data

7. Risk Management
- Compile risk register and score risks
- Define and implement risk mitigation strategies

8. Auditing
- Validate controls, tollgates and documentation
- Identify gaps and recommend remediation steps

Formalize AI **OVERSIGHT** structure

Ensure diverse perspectives and membership across all levels of governance.



Team	Responsibilities	Members	Audit	Compliance	DG	Info Security	Privacy	Risk Mgmt
AI Ethics Board *(Board Dir/CxO/VP/Sr Dir)*	• Educate Board of Directors about AI • Establish & communicate AI ethics principles • Oversee AI ethics enterprise-wide	• 1-3 Board Directors ★ • CDO / AI Exec ☆ • 1 Exec per LOB using AI • 1-2 External AI SMEs		✓	✓	✓		✓
AI/ML Council *(Sr Dir/Dir/Mgr/Sr Staff)*	• Establish AI principles • Manage AI literacy training • Set AI goals & priorities • Review & strategically align AI requests • Establish AI policies & standards • Oversee AI development & risk management • Resolve escalated issues	• CDO / AI Exec ★ • Chief Compliance Officer ☆ • 1 CxO/VP per LOB using AI • AI/ML Leader • AIOps Leader • Enterprise Data Architect • 1 Dir/Mgr per LOB using AI		✓	✓	✓	✓	✓
AI Tech Review *(Sr Dir/Dir/Arch/Sr Staff)*	• Set AI design, development, testing standards • Assess AI model design & algorithms • Assess AI model metrics • Monitor AI validation & testing • Monitor AI deployment & support	• AI/ML Leader ★ • AIOps Leader ☆ • Enterprise Data Architect • Application/Software Architects • 1 AI/ML Sr Dev per LOB using AI	✓	✓	✓	✓	✓	✓
AI Risk Mgmt *(Sr Dir/Dir/Mgr/Sr Staff)*	• Evaluate AI input/training/output data sets • Evaluate AI model documentation • Enforce AI/data regulatory compliance • Enforce AI/data policy compliance • Identify AI risks & manage mitigation efforts	• Risk Mgmt Leader ★ • AI/ML Leader ☆ • Enterprise Data Architect • Application/Software Architects • 1 Dir/Mgr per LOB using AI	✓	✓	✓	✓	✓	✓
AI Developers	• Follow all AI governance standards & processes							

★ = Chair ☆ = Co-Chair ✓ = Include

Formulate an AI **STRATEGIC PLAN**

Plot the organization's trajectory toward an AI-enabled workforce & a better future.

Leadership

- Understand AI basics, benefits, challenges & risks
- Plan the transition to AI-powered operations
- Establish ethical, social & technical guidelines for AI use

Vision

- Depict what we hope to achieve with AI in 3-5+ years
- Determine how & where AI can effectively be used
- Build reliable, safe & transparent AI solutions

Values

- Foster a culture of responsible & trustworthy AI
- Position humans, ethics & data quality at the center of AI
- Ensure AI efforts & solutions are legally defensible

Goals

- Align AI efforts with strategic goals & risk management
- Prioritize AI use cases by value, impact & outcome
- Develop AI knowledge & skills as core capabilities

Framework

- Implement MLOps & ModelOps as formal frameworks
- Integrate AI into audit, compliance and risk processes
- Standardize AI documentation for explainability

Integrate AI with existing strategic planning processes

Establish AI **GUIDING PRINCIPLES**

Evangelize core values & beliefs to drive AI decision-making mindsets and behaviors.

Accountability	Diversity	Human-Centered ★	Objectivity	Safety
Accuracy	Equity	Impact	Personalization	Security ★
Auditability	Ethical ★	Impartiality	Privacy	Sustainability
Autonomy	Explainability ★	Inclusiveness	Reliability	Transparency ★
Awareness	Fairness ★	Justice	Responsibility ★	Trustworthiness ★
Disclosure	Governable	Lawful	Robustness	Well-being

Choose & define AI guiding principles

Consider adopting examples from external sources: Deloitte EU Google IBM MicrosoftNIST OECD Singapore PDPC

★ commonly cited

Publish **DEFINITIONS** for AI guiding principles

Explain each principle with plain language that is easy to understand and memorable.

Trust is must	Ensure AI is **trustworthy** via safe, secure, responsible & ethical AI development
Make it easy	Promote **transparency** into AI purpose, content, function, outputs & use
Document it	Create **explainability** through formal documentation of all AI components
Humans first	Place **humans** at the center of AI (above/within/over the loop)
Let them know	Facilitate **awareness** & disclosure/informed consent for subjects of AI solutions
No bias	Promote **fairness** by identifying and mitigating bias in every step
Win audits	Support **auditability** via formal standards, processes & change management

TIP: Consider using names that are **short** **sticky** **desirable** **feasible**

Establish AI **ETHICS** principles too

Deloitte grounds deliberations in a scientific understanding of the strengths & weakness of both AI/ML & human cognition.

Deloitte's Design Principles for Ethical AI

Three core principles can help leaders think through AI's ethical implications

Themes (Non-maleficence)
- safety
- reliability
- robustness
- data provenance
- privacy
- cybersecurity
- misuse

Themes (Beneficence)
- human flourishing
- well-being
- dignity
- common good
- sustainability

Themes (Procedural fairness)
- algorithmic bias
- equitable treatment
- consistency

Themes (Distributive fairness)
- shared benefits
- shared prosperity
- fair decision outcomes

Themes (Control)
- consent
- choice
- enhancing human agency & self-determination
- reversibility of machine autonomy

Themes (Comprehension)
- intelligibility
- transparency
- trustworthiness
- accountability

Source: Guszcza, J., Lee, M., Ammanath, B., & Kuder, D. (2020, January 28). Human values in the loop: Design principles for ethical AI. Deloitte Insights. https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/design-principles-ethical-artificial-intelligence.html

Implement an enterprise AI **POLICY**

Codify critical AI principles & standards in an official corporate policy (SAMPLE CONTENT).

Purpose	The purpose of this policy is to guide the ethical and responsible use of artificial intelligence (AI) technologies across the organization. This policy outlines the principles, standards, and expectations for AI use and is designed to ensure compliance with legal and ethical standards while promoting innovation and effective use of AI technologies.
Scope	This policy applies to all employees, contractors, business partners, and stakeholders who interact with or use AI technologies for business operations. It covers all AI applications, systems, and related technologies (developed, publicly available, or embedded in existing platforms) within our organization.
Principles	Our company is committed to leveraging AI technologies responsibly and ethically while ensuring respect for individual rights and data privacy. This includes: • Ethical Development and Use: AI must be designed, developed, and used in a manner that is fair, transparent, and free from discrimination or bias. • Security: AI systems must comply with applicable data privacy regulations and ensure the protection of personal data from unauthorizedaccess or misuse. • Accountability: Employees and stakeholders must be accountable for their use of AI technologies and adhere to this policy's standards. • Transparency: The development and deployment of AI systems must be transparent, with clear explanations of how AI processes data and makes decisions. • Compliance with Standards: All AI technologies and practices must align with all approved frameworks and standards.
Standards	We have adopted the following standards for AI: • Deloitte's Trustworthy AI Framework and AI Ethics Principles • EU AI Act • OECD Framework for the Classification of AI Systems • Singapore's Model AI Governance Framework (Second Edition) • NIST AI Risk Management Framework (Pub. 100-1) • NIST Towards a Standard for Identifying and Managing Bias in AI (Pub. 1270)
Definitions	• AI Bias: The presence of systematic errors in AI outcomes that result in unfair or discriminatory treatment. • AI Data Privacy: The protection of personal data used in AI applications to ensure compliance with relevant privacy laws and regulations. • AI Ethics: A set of moral principles guiding AI development and usage to ensure fairness, transparency, and accountability. • Artificial Intelligence (AI): The simulation of human intelligence in machines designed to think and learn. • Machine Learning (ML): A subset of AI that involves algorithms allowing systems to improve their performance over time through experience or data analysis

Consider additional policies: GenAI, AI documentation, AI risk assessments…

Clarify EXPECTATIONS across the AI lifecycle

AI Framework Example: Define stages and responsible measures for documentation, transparency, and trust in AI solutions.

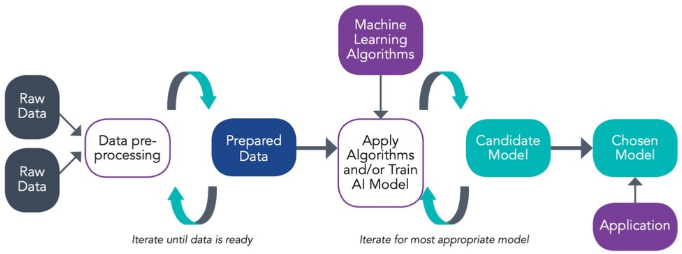
Singapore's AI Framework

Data Preparation → **Algorithms** → **Chosen Model**

Stage 1:
Raw data is formatted and cleansed so conclusions can be drawn accurately. Generally, accuracy and insights increase with relevance and the amount of data.

Stage 2:
Models are trained on the dataset and **algorithms** may be applied. This includes statistical or machine learning models including decision trees and neural networks. The results are examined and models are iterated until the most appropriate model emerges.

Stage 3:
The **chosen model** is used to produce probability scores that can be incorporated into applications to offer predictions, make decisions, solve problems and trigger actions.



Data Preparation

- Document data lineage (forward, backward, end-to-end)
- Maintain data provenance records & risk assessments
- Measure & ensure data quality as well as data privacy requirements
- Identify & minimize inherent bias
- Use different datasets for training, testing & validation
- Engage periodic reviewing & updating of datasets

Algorithm & Model

- Identify which features will most impact stakeholders and consumers
- Select measures of transparency that will most build trust
- Apply measures for the entire model (or a subset of features commercially sensitive or intellectual property is involved
- Store all measures, assessments & records in a centralized repository

Stakeholder Communication

- Provide general disclosure on how & why AI is used, its benefits, efforts to identify & mitigate risks, and the role/extent of AI in decision-making processes
- Develop a policy on what information to provide to individuals & when
- Tailor easy-to-understand communications based on audience
- Consider providing an option to opt-out + collect feedback
- Ensure AI governance practices & processes align with ethical standards

Source: Info-communications Media Development Authority & Personal Data Protection Commission. (2020). Artificial Intelligence Governance Framework Model: Second edition. https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf

Establish a **RACI** for the AI lifecycle

AI Framework Example: Map AI activities and deliverables to formal responsibilities to roles.

AI Deliverable	BOD	AI Steward	Data Arch	Audit	Compl	Privacy	Risk	HR	Other LOB	AI Ethics Board	AI Exec	AI Leader	AI PM	AI/ML Engineer	Data Scientist	Domain Owner	DGO
Define AI vision & strategy	A	C	C	C	C	C	C	I	I	C	R	I	I	I	C	I	C
Establish AI guiding principles	A	C	C	C	C	C	C	I	I	C	R	I	I	I	C	I	C
Establish AI ethics guidelines	A	C	C	C	C	C	C	I	I	R	C	I	I	I	C	I	C
Promote ethical & responsible AI	A	C	C	C	C	C	C	R	R	R	C	I	I	I	R	R	C
Ensure AI strategic alignment	C	C	C	C	C	C	C	I	I	C	A	R	I	I	R	R	C
Prioritize use cases	I	C	C	C	C	C	C	I	I	I	A	R	C	C	C	R	C
Manage AI projects, resources & timelines	I	C	C	I	I	C	C	I	I	I	I	A	R	C	C	C	C
Define AI design & testing requirements	I	R	C	I	I	C	C	I	I	I	I	A	R	R	R	R	C
Identify & mitigate AI risks	I	R	C	I	I	R	R	I	I	I	I	A	R	R	R	R	C
Identify data privacy & protection needs	I	R	C	I	I	R	R	I	I	I	I	A	C	C	C	I	C
Build & maintain AI solutions	I	R	C	I	I	C	C	I	I	I	I	A	C	R	R	I	C
Deploy & monitor AI solutions	I	R	C	I	I	C	C	I	I	I	I	A	C	R	R	I	C
Fix, improve & update AI solutions	I	R	C	I	I	C	C	I	I	I	I	A	C	R	R	R	C
Assess regulatory & policy compliance	I	C	C	C	R	C	C	I	I	I	I	A	C	C	C	I	C
Audit AI processes, solutions & risks	I	R	C	R	C	C	R	I	I	I	I	A	C	C	C	I	C
Develop & deliver AI training programs	I	C	C	C	C	C	C	R	I	I	I	A	C	C	C	I	C
Follow AI guidelines, standards & policies	C	R	R	R	R	R	R	R	R	C	C	A	R	R	R	R	R

KEY: R Responsible A Accountable C Consulted I Informed

Implement a standard for **DOCUMENTING** AI models

AI Framework Example: Leverage a template to standardize requirements and content for documentation.

Avoid disclosing proprietary or private details

1. Model Details
Answer basic questions about model version, type, and other shareable details that inform what the model represents

2. Intended Use
Explain how the model should & should not be used, and why it was created (to frame the statistical analysis in later sections)

3. Factors
Summarize model performance across relevant factors like groups, instrumentation & environments

4. Metrics
Define measures of performance, decision thresholds, approaches to uncertainty & variability, and reasons for those metrics

5. Evaluation Data
Describe the source, composition & reasons for data used to evaluate the model and how those datasets were preprocessed

6. Training Data
Document as much information as possible about training data, distributions over groups, and other details about potential biases

7. Quantitative Analyses
Break down quantitative analyses by chosen factors & provide evaluation results (with confidence intervals) for chosen metrics

8. Ethical Considerations
Explain ethical issues considered during development (e.g., human impact, sensitive data, risk mitigation, harms, fraught use cases)

9. Caveats and Recommendations
List concerns not covered in other sections (e.g., need for more testing, groups not represented in data set, recommendations)

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

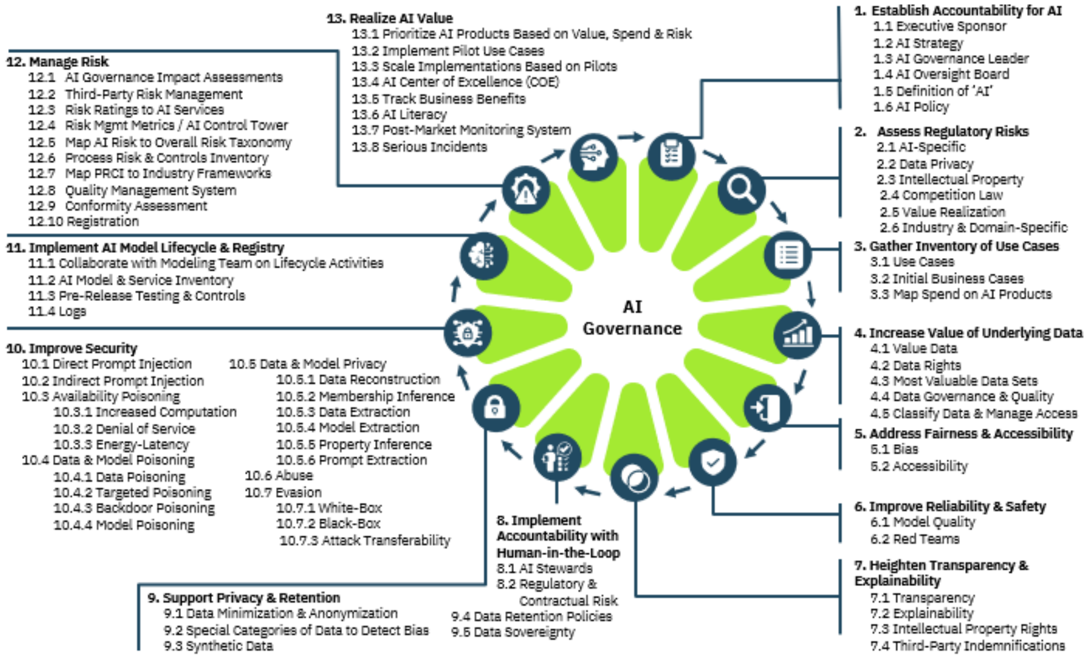
Model Card

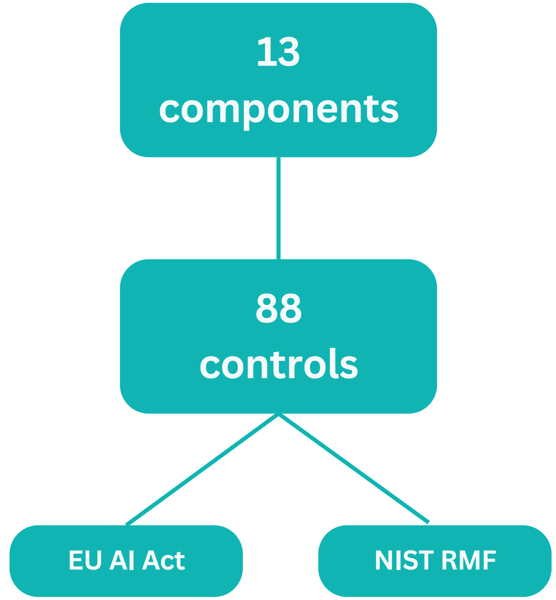
Source: Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019, January 14). Model cards for model reporting. In FAT '19: Conference on Fairness, Accountability, and Transparency (pp. 220–229). https://doi.org/10.48550/arXiv.1810.03993

Map AI governance framework to REGULATORY controls

AI Compliance Example: Continuous loop framework mapped to requirements in the EU AI Act and NIST RMF.

YourDataConnect.com





Source: Soares, S. (2024). AI governance: A controls playbook with mappings to the European Union AI Act and the NIST AI Risk Management Framework. YourDataConnect, LLC (DBA YDC). https://yourdataconnect.com/

DAIG PARTNERS

Apply performance **METRICS** across the AI lifecycle

AI Measures Example: Define feasible, cost-effective and appropriate measurements and documentation requirements.

Explainability	• Ensure the organization can explain how the algorithm works and why specific decisions were made • Define how the algorithm functions and/or how the decision-making process incorporates model predictions • Define how model training & selection processes were conducted • Define how risks were identified and addressed
Repeatability	• Ensure models consistently perform actions or make decision for the same scenario (via repeatability assessments in live environments) • Perform counterfactual fairness testing to ensure decisions are the same in a counterfactual world where sensitive attributes are altered • Define how exceptions are identified & handled when decisions are not repeatable + ensure exception handling complies with policies • Identify & account for changes over time to ensure models trained on time-sensitive data remain relevant
Robustness	• Assess the degree to which model function correctly in the presence of invalid inputs, execution errors, or stressful environmental conditions • Conduct adversarial testing on models (or highest risk functions) to ensure they can handle a broader range of unexpected input variables • Maintain awareness of risks with continual learning and ensure adequate testing and monitoring of models to detect unpredictable behaviors
Regular Tuning	• Perform regular model tuning to ensure models cater for changes to customer behavior over time • Refresh models based on updated training datasets that incorporate new input data or when objectives/risks/values change • Test in varied environments to reduce risk of models learning regularities that do not reflect actual production environment conditions
Traceability	• Ensure decisions, datasets & processes for the AI model's decision (including data gathering/labeling & algorithms) are documented • Make traceability documentation accessible for troubleshooting, investigating how the model functions, or why a prediction was made • Build audit trails to document model training & AI-augmented decisions + implement black box recorder to captures input data streams • Store data relevant to traceability to avoid degradation/alteration with appropriate retention for durations relevant to the industry/applicable laws
Reproducibility	• Ensure independent verification teams produce the same results using the same AI method based on documentation • Avoid disclosing IP by specifying the subset of features for the independent verification team to assess • Identify specific contexts or conditions required for reproducibility • Make available replication files (i.e., files that replicate each step of the AI model's developmental process) to facilitate testing efforts
Auditability	• Ensure readiness of AI systems to undergo assessments of algorithms, data & design processes by internal or external auditors • Identify commercially sensitive information/IP + areas where auditability is necessary to align with regulatory requirements or industry practice • Maintain & centralize records needed to support auditing in a centralized digital repository

Source: Info-communications Media Development Authority & Personal Data Protection Commission. (2020). Artificial Intelligence Governance Framework Model: Second edition. https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf

Identify **TOLLGATES** across the AI lifecycle

AI Measures Example: Define and document required outcomes/deliverables for each AI initiative.

1 **Value**
- Define a clear value proposition + assess strategic alignment for each AI effort
- Determine what the data & model will predict + how it will achieve or contribute to business goals

2 **Goals**
- Translate business objectives into clear technical actions + measurable & achievable goals
- Establish realistic expectations for precision + what's predicted + how to use predictions

3 **Metrics**
- Measure model performance objectively using agreed-upon metrics (e.g., accuracy, lift, cost, etc.)
- Include measures for assessing ethical issues + training + operationalizing the model

4 **Data**
- Resolve data issues before ingestion (i.e., good data is the foundation of predictive strength)
- Identify relevant data privacy & data usage issues + address them

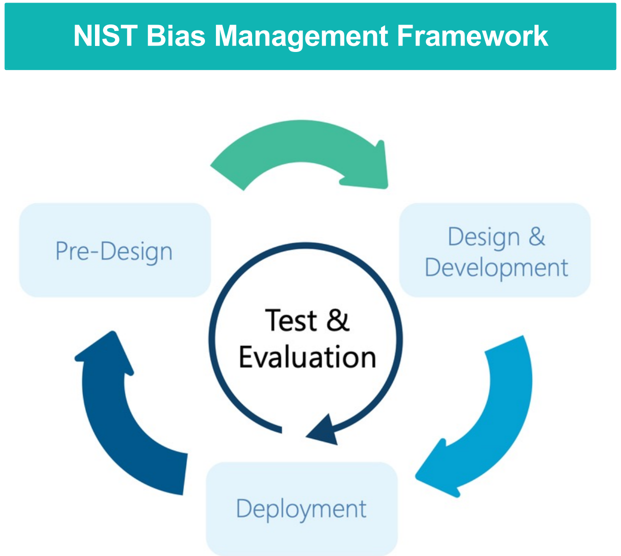
5 **Training**
- Clarify what can vs. cannot be learned from the data + what are expected vs. unexpected outcomes
- Validate model sensitivity + debug it before learning from the data + track model versions

6 **Deployment**
- Ensure buy-in from all levels + needed business units (include customer service + tech support)
- Communicate the impact on roles & responsibilities, daily routines, workflows & decision making

Source: Siegel, E. (2024). The AI playbook. MIT Press.

Manage potential BIAS across the AI lifecycle

AI Measures Example: Define stages and formal processes and multi-stakeholder support to mitigate potential harms.



Pre-Design
- Document data lineage (forward, backward, end-to-end)
- Maintain data provenance records & risk assessments
- Measure & ensure data quality as well as data privacy requirements
- Specify problem, purpose and benefits; conduct research; identify available data
- Assess organizational biases, individual & group heuristics, limited points of views
- Consider biases reflected in the selected datasets

Design & Development
- Analyze requirements and available data; select/design model
- Perform compatibility analysis; identify sources of bias; implement mitigation plans
- Evaluate/adjust bias mitigation efforts until model stays within pre-specified limits

Deployment
- Release and use the model; monitor system outputs after human interaction
- Ensure deployed solutions do not cause unintended effects or harms
- Assess/retrain model as needed; correct adverse events or decommission model

Test & Evaluation *(throughout)*
- Perform continuous testing and evaluation of all components and features
- Verify model performance against agreed-upon metrics and bias mitigation efforts
- Identify and resolve data quality, data privacy and data usage issues

Source: National Institute of Standards and Technology. (2022). Towards a standard for identifying and managing bias in artificial intelligence (NIST Special Publication 1270). https://doi.org/10.6028/NIST.SP.1270

DAIG PARTNERS

Audit for **BIAS** across the AI lifecycle

AI Measures Example: Assess the fairness and efficacy of AI models across all stages to lessen or eradicate harmful effects.

Purpose, Process & Monitoring Framework

Credit Scoring • Insurance Scoring • Automated Underwriting Risk-Based Pricing • Digital Advertising • Tenant Screening Selection Tools

Purpose

Business Understanding:
- Assess project goals and the expectations, requirements, and objectives of stakeholders (collectively referred to as the "business problem"
- Assess risks the business problem may pose to consumers, institutions, and society

Data Understanding:
- Assess how well data is used to accurately capture and reflect the business problem
- Determine what if any techniques were used to mitigate risks associated with data paucity or data quality

Process

Staff Profile:
- Ensure teams are diverse, inclusive, and educated to spot challenges and prevent issues that lead to unfavorable outcomes

Data Assessment:
- Determine if the data sources and data fields used to develop the model are appropriate, representative, fair, and accurate

Model Assessment:
- Evaluate training algorithms, parameters, hyper-parameters, fairness constraints used during both development and post-modeling, and the selection of less discriminatory alternatives

Outcome Assessment:
- Evaluate performance of the final model in line with scope and metrics defined in the business problem to determine if the model meets its objectives, including minimization of risks

Model Use and Limitation:
- Review and document known model limitations and assumptions, and circumstances where the model may or may not be used outside the scope of its intended uses

Monitoring

Product Model Validation:
- Compare missingness patterns of the data used to develop the training model with those of the production version
- Assess model evaluation metrics and fairness metrics across both environments and protected class categories (before and after model deployment) to confirm model stability and reliability
- Decide whether about retraining, patching, or retiring production models

Protection from Confidentiality and Integrity Attacks:
- Evaluate defenses that protect the privacy of records used to train the model or score the model in production
- Ensure the model incorporates defenses that assure fairness and accountability.

Source: Akinwumi, M., Rice, L., & Sharma, S. (2022). Purpose, process, and monitoring: A new framework for auditing algorithmic bias in housing and lending. National Fair Housing Alliance. https://nationalfairhousing.org/wp-content/uploads/2022/02/PPM_Framework_02_17_2022.pdf

Formalize **RISK MANAGEMENT** across the AI lifecycle

AI Measures Example: The NIST AI RMF helps minimize anticipated negative impacts & identify opportunities for positive impacts.

NIST AI Risk Management Framework

Map

- Establish & understand context
- Perform categorization of AI system
- Compare AI capabilities, use, goals, benefits & costs to benchmarks
- Map risks & benefits for all AI system components
- Describe impacts to individuals, groups, communities, organizations & society

Measure

- Identify & apply appropriate methods & metrics
- Evaluate AI systems for trustworthy characteristics
- Implement mechanism to track AI risks over time
- Gather & assess feedback about efficacy of measurement

Govern

- Create transparent policies & procedures
- Establish accountability & training structures
- Prioritize DEI & accessibility
- Cultivate a culture that acknowledges AI risks
- Foster robust engagement with AI actors
- Assess AI risks with 3rd party software & data

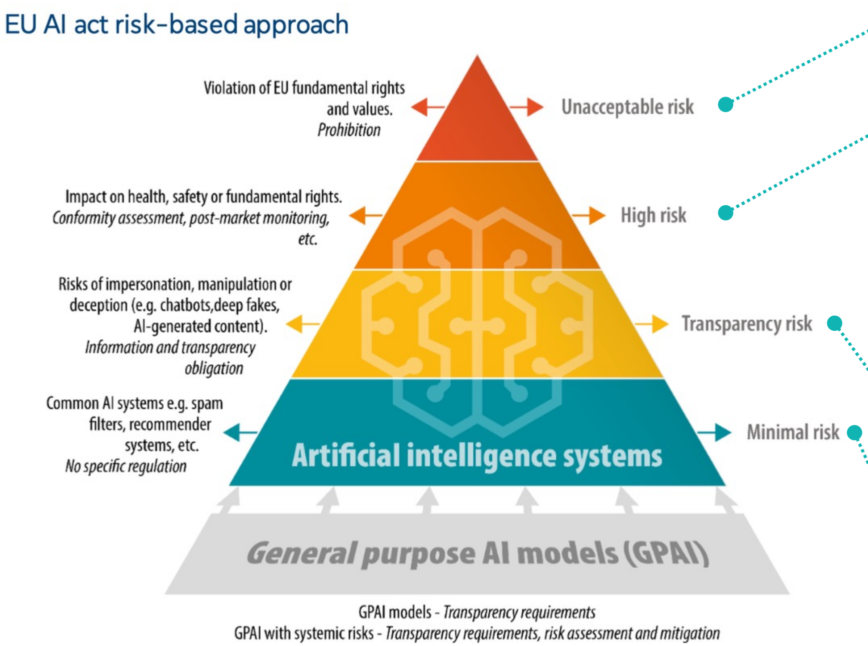
Manage

- Address AI risks from MAP & MEASURE functions
- Implement strategies to maximize AI benefits & minimize negative impacts (with inputs from all AI actors)
- Manage AI risks & benefits from third-party entities
- Document & monitor risk treatments & communication plans

Source: National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Set **RISK TOLERANCE LEVELS** for AI risk

AI Measures Management Example: The draft [EU AI Act] defines levels of risk based on the intended use of an AI system.



Banned

Most regulated

- data & data governance requirements
- human oversight
- technical documentation & record keeping
- transparency & provision of information to users
- human oversight
- robustness, accuracy & security
- risk management system across AI lifecycle
- conformity self-assessment via internal controls (pre-marketing)
- post-marketing monitoring (with reporting of incidents to authorities)

Provide transparency

- labelling & disclosure of manipulated content
- inform users they are interacting with AI

Voluntary

- free use of minimal-risk AI

Source: Madiega, T. (2024). Artificial Intelligence Act: EU legislation in progress (Briefing 698792). European Parliament. https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf

Define an AI risk **SCORING RUBRIC**

AI Measures Example: Use an objective rubric to rate and prioritize risks and inform risk mitigation approaches.

RISK MATRIX

Likelihood			
Probable Likely to occur during standard AI operations			
Occasional Likely to occur sometime during standard AI operations			
Improbable Unlikely but possible to occur during standard AI operations			
Impact	**Low**	**Medium**	**High**
	Impact of decisions is isolated and/or their severity is not serious	Impact of decisions reaches a moderate amount of people and /or their severity is moderate	Impact of decisions is widespread and/or their severity is serious

RISK RATING

LOW Risk	**MEDIUM Risk**	**HIGH Risk**
Less rigor / More autonomy Could apply risk mitigation efforts	Balanced rigor / autonomy Should apply risk mitigation efforts	More rigor / Less autonomy Must apply risk mitigation efforts

Prioritize

Additional measurement challenges:

- Availability of reliable metrics
- Risks at different stages of the AI lifecycle
- Risks in real-world settings (vs. pre- deployment environments)
- Inscrutability (limited explainability and/or interpretability)
- Reliable human baselines
- Third-party software, hardware & data risks

Source: Info-communications Media Development Authority & Personal Data Protection Commission. (2020). Artificial Intelligence Governance Framework Model: Second edition. https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf

Source: National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 10 U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Publish a **GLOSSARY** of AI terminology

Socialize official AI-related terms & definitions so everyone is on the same page (and can find that page!)

Definition of AI	**AI-related terms**		
51 definitions	69 terms	117 terms	62 terms
IAPP	IAPP	ISO 22989	NIST

225 unique terms *(20 of which have 2 sources)*

1 definition + **selected terms**

plus any additional terms

publish

Enterprise Glossary

Searchable

Launch widespread AI **LITERACY** training

A formal training program will help upskill AI awareness and capabilities and help mitigate risk.

Fundamentals of AI

- Understand AI terms & definitions
- Understand categories & types of AI
- Understand typical use cases for AI

AI Promises & Pitfalls

- Understand expected benefits of AI
- Understand common outcomes of AI solutions
- Understand potential risks & harms of AI

AI Principles & Regulations

- Understand AI guiding principles
- Understand AI ethics principles
- Understand relevant AI regulations

AI Operating Frameworks

- Understand AI development processes
- Understand GenAI usage guidelines
- Understand AI request & approval processes

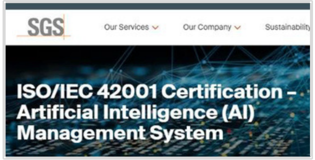
AI Governance Processes

- Understand AI strategy & oversight structure
- Understand AI documentation & auditing processes
- Understand AI exploration & learning guardrails

Consider launching an AI Center of Excellence for enterprise-wide AI-related learning & discovery

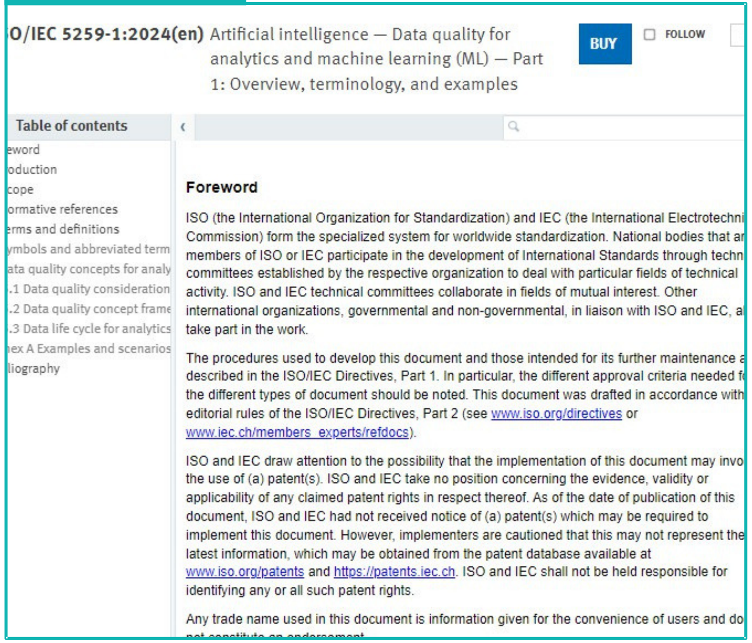
Consider AI certifications for the ORGANIZATION

Aligning development efforts with AIMS certification can generate competitive advantage.



[sgs.com](https://www.sgs.com)

5259-1:2024 Data Quality



O/IEC 5259-1:2024(en) Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples

BUY FOLLOW

Table of contents

eword
oduction
cope
ormative references
erms and definitions
ymbols and abbreviated term
ata quality concepts for analy
.1 Data quality consideration
.2 Data quality concept frame
.3 Data life cycle for analytics
ex A Examples and scenarios
liography

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechni Commission) form the specialized system for worldwide standardization. National bodies that ar members of ISO or IEC participate in the development of International Standards through techn committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, al take part in the work.

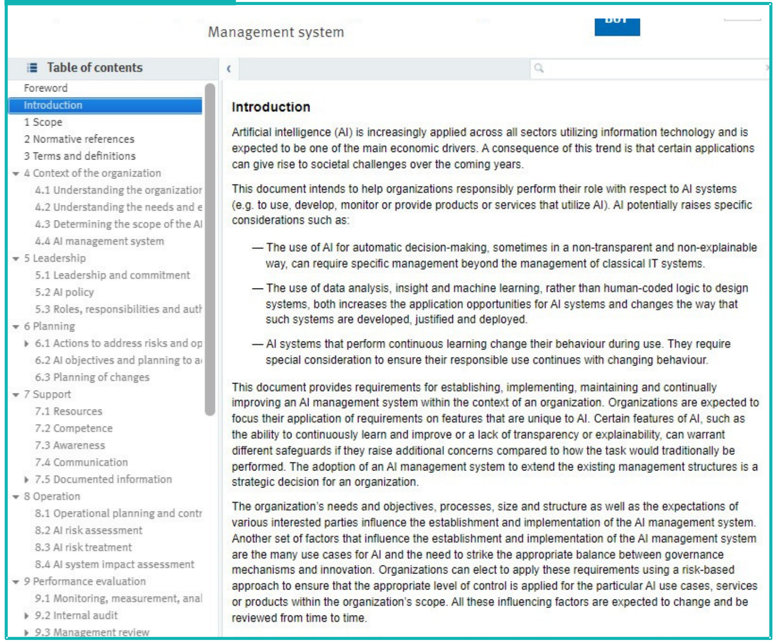
The procedures used to develop this document and those intended for its further maintenance a described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed f the different types of document should be noted. This document was drafted in accordance with editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives or www.iec.ch/members_experts/refdocs).

ISO and IEC draw attention to the possibility that the implementation of this document may invo the use of (a) patent(s). ISO and IEC take no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO and IEC had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents and https://patents.iec.ch. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and do

Source: [International Organization for Standardization. (2024). Information technology — Artificial intelligence — Artificial intelligence concepts and terminology (ISO/IEC 22989:2022). https://www.iso.org/obp/ui/en/#iso:std:iso-iec:22989:ed-1:v1:en](https://www.iso.org/obp/ui/en/#iso:std:iso-iec:22989:ed-1:v1:en)

42001:2023 Model Management



Management system

Table of contents

Foreword
Introduction
1 Scope
2 Normative references
3 Terms and definitions
4 Context of the organization
4.1 Understanding the organizatior
4.2 Understanding the needs and e
4.3 Determining the scope of the AI
4.4 AI management system
5 Leadership
5.1 Leadership and commitment
5.2 AI policy
5.3 Roles, responsibilities and auth
6 Planning
6.1 Actions to address risks and op
6.2 AI objectives and planning to a
6.3 Planning of changes
7 Support
7.1 Resources
7.2 Competence
7.3 Awareness
7.4 Communication
7.5 Documented information
8 Operation
8.1 Operational planning and contr
8.2 AI risk assessment
8.3 AI risk treatment
8.4 AI system impact assessment
9 Performance evaluation
9.1 Monitoring, measurement, anal
9.2 Internal audit
9.3 Management review

Introduction

Artificial intelligence (AI) is increasingly applied across all sectors utilizing information technology and is expected to be one of the main economic drivers. A consequence of this trend is that certain applications can give rise to societal challenges over the coming years.

This document intends to help organizations responsibly perform their role with respect to AI systems (e.g. to use, develop, monitor or provide products or services that utilize AI). AI potentially raises specific considerations such as:

— The use of AI for automatic decision-making, sometimes in a non-transparent and non-explainable way, can require specific management beyond the management of classical IT systems.

— The use of data analysis, insight and machine learning, rather than human-coded logic to design systems, both increases the application opportunities for AI systems and changes the way that such systems are developed, justified and deployed.

— AI systems that perform continuous learning change their behaviour during use. They require special consideration to ensure their responsible use continues with changing behaviour.

This document provides requirements for establishing, implementing, maintaining and continually improving an AI management system within the context of an organization. Organizations are expected to focus their application of requirements on features that are unique to AI. Certain features of AI, such as the ability to continuously learn and improve or a lack of transparency or explainability, can warrant different safeguards if they raise additional concerns compared to how the task would traditionally be performed. The adoption of an AI management system to extend the existing management structures is a strategic decision for an organization.

The organization's needs and objectives, processes, size and structure as well as the expectations of various interested parties influence the establishment and implementation of the AI management system. Another set of factors that influence the establishment and implementation of the AI management system are the many use cases for AI and the need to strike the appropriate balance between governance mechanisms and innovation. Organizations can elect to apply these requirements using a risk-based approach to ensure that the appropriate level of control is applied for the particular AI use cases, services or products within the organization's scope. All these influencing factors are expected to change and be reviewed from time to time.

Source: [International Organization for Standardization. (2023). Information technology — Artificial intelligence — Management system (ISO/IEC 42001:2023). https://www.iso.org/obp/ui/en/#iso:std:iso-iec:42001:ed-1:v1:en](https://www.iso.org/obp/ui/en/#iso:std:iso-iec:42001:ed-1:v1:en)

Source: [SGS. (2023). ISO/IEC 42001 certification: Artificial intelligence (AI) management system. https://www.sgs.com/en/services/iso-iec-42001-certification-artificial-intelligence-ai-management-system](https://www.sgs.com/en/services/iso-iec-42001-certification-artificial-intelligence-ai-management-system)

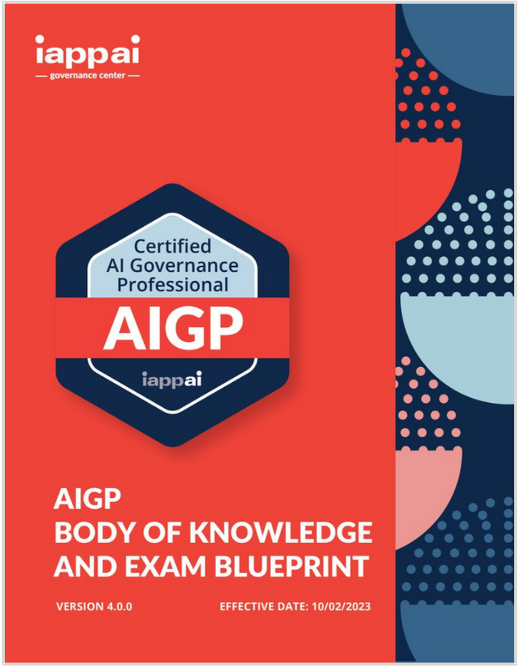
Consider AI certifications for **STAFF**

IAPP launched their AI Governance Professional (AIGP) certification in 2024.

iapp.org

Candidates:

- AI/ML
- Auditing
- Compliance
- Data Governance
- Data Science
- GRC
- Info Sec
- Legal
- Privacy


iappai
— governance center —
Certified
AI Governance
Professional
AIGP
iappai
AIGP
BODY OF KNOWLEDGE
AND EXAM BLUEPRINT
VERSION 4.0.0
EFFECTIVE DATE: 10/02/2023


IAPP AIGP CANDIDATE
HANDBOOK 2023
Procedures and policies
Certified
AI Governance
Professional
AIGP
iappai
iappai
— governance center —

Source: International Association of Privacy Professionals. (n.d.). Artificial Intelligence Governance Professional (AIGP) certification. IAPP. Retrieved April 18, 2024, from https://iapp.org/certify/aigp/

Questions? Feel free to contact me

mathias@daigpartners.com
+32 468 258 947
https://daigpartners.com