CS 435-L1 Lecture 6

Adnan Masood, PhD.

Generative AI Security

Recap Quiz

AI Regulations – I

Story time - Class Presentations on AI Horror Stories

AI Regulations - II

Assignment

Generative AI Security

Safety is the state of being protected from potential harm, while **security** is the measures taken to protect against threats.

Safety safeguards against unintentional harms or accidents (e.g., mechanical failures, human error), ensuring systems or environments don't accidentally cause damage.

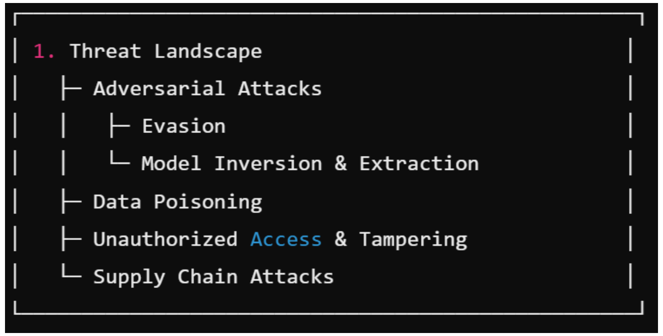
Security protects against deliberate threats or malicious acts (e.g., hacking, sabotage), ensuring systems or environments are shielded from intentional harm.

Safety
Freedom from accidental risks or harm.

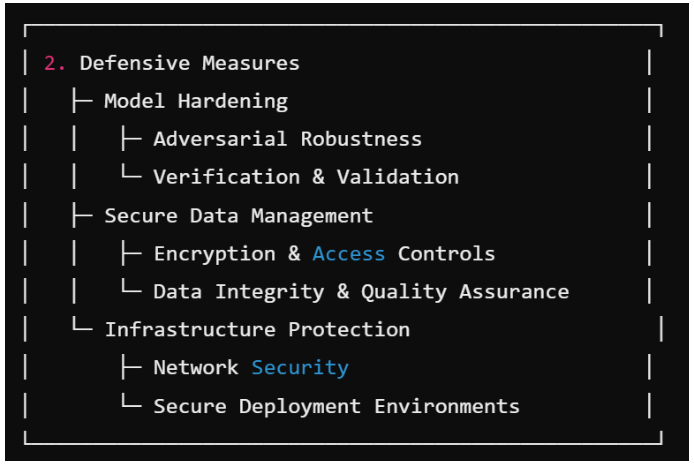
Engineering: Ensuring systems function without causing unintended damage.
Health & Environment: Minimizing hazards to well-being and the ecosystem.
Social: Creating conditions where people feel at ease and not under threat.

Security
Freedom from deliberate threats or attacks.

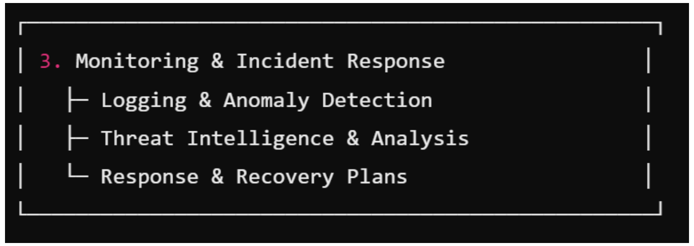
Cyber: Protecting data and systems from malicious intrusions.
National: Defending a country’s interests against espionage or aggression.
Personal: Safeguarding individuals against violence or theft.



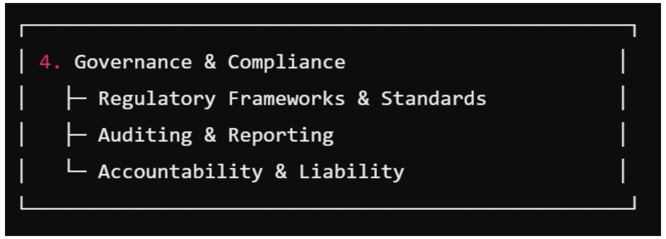
```
┌───────────────────────────────────────────┐
│ 1. Threat Landscape                        │
│   ├─ Adversarial Attacks                   │
│   │   ├─ Evasion                           │
│   │   └─ Model Inversion & Extraction      │
│   ├─ Data Poisoning                        │
│   ├─ Unauthorized Access & Tampering       │
│   └─ Supply Chain Attacks                  │
└───────────────────────────────────────────┘
```



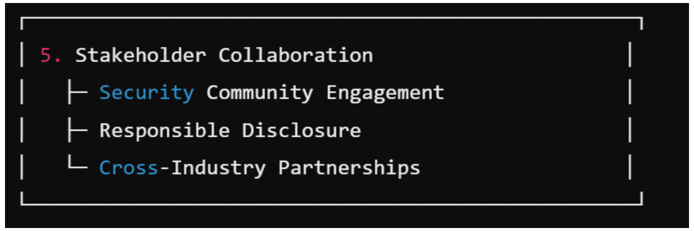
```
┌───────────────────────────────────────────────┐
│ 2. Defensive Measures                         │
│   ├─ Model Hardening                          │
│   │   ├─ Adversarial Robustness               │
│   │   └─ Verification & Validation            │
│   ├─ Secure Data Management                   │
│   │   ├─ Encryption & Access Controls         │
│   │   └─ Data Integrity & Quality Assurance   │
│   └─ Infrastructure Protection                  │
│       ├─ Network Security                     │
│       └─ Secure Deployment Environments       │
└───────────────────────────────────────────────┘
```



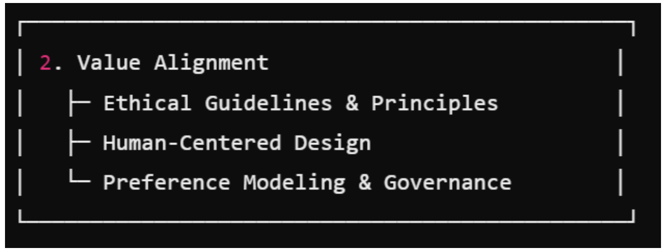
```
┌───────────────────────────────────────────┐
│ 3. Monitoring & Incident Response          │
│   ├─ Logging & Anomaly Detection           │
│   ├─ Threat Intelligence & Analysis        │
│   └─ Response & Recovery Plans             │
└───────────────────────────────────────────┘
```



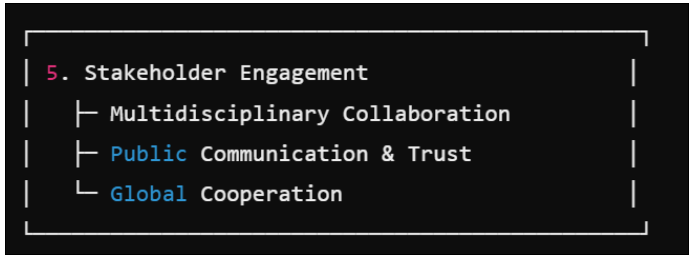
```
┌─────────────────────────────────────────┐
│ 4. Governance & Compliance              │
│   ├─ Regulatory Frameworks & Standards  │
│   ├─ Auditing & Reporting               │
│   └─ Accountability & Liability         │
└─────────────────────────────────────────┘
```



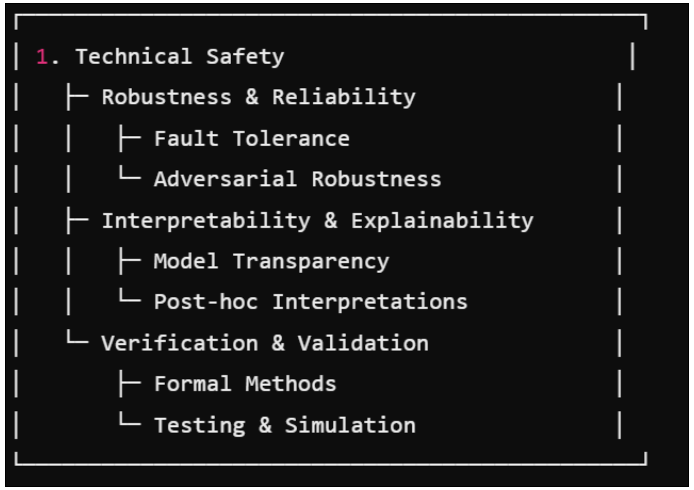
```
┌──────────────────────────────────────────┐
│ 5. Stakeholder Collaboration             │
│   ├─ Security Community Engagement       │
│   ├─ Responsible Disclosure              │
│   └─ Cross-Industry Partnerships         │
└──────────────────────────────────────────┘
```



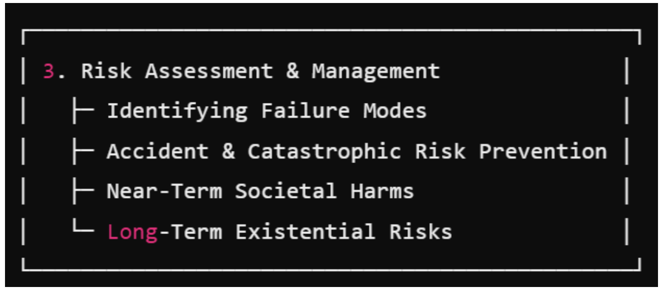
```
┌──────────────────────────────────────────────┐
│ 2. Value Alignment                           │
│   ├─ Ethical Guidelines & Principles         │
│   ├─ Human-Centered Design                   │
│   └─ Preference Modeling & Governance        │
└──────────────────────────────────────────────┘
```



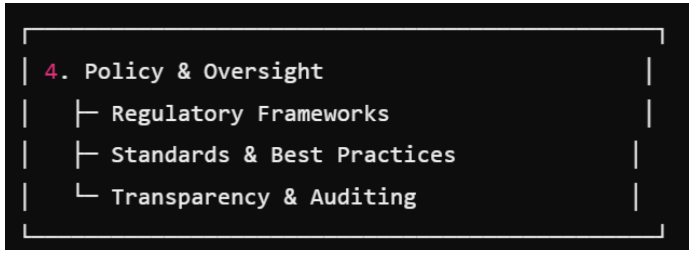
```
┌──────────────────────────────────────────────┐
│ 5. Stakeholder Engagement                    │
│   ├─ Multidisciplinary Collaboration         │
│   ├─ Public Communication & Trust            │
│   └─ Global Cooperation                      │
└──────────────────────────────────────────────┘
```



```
┌──────────────────────────────────────────────┐
│ 1. Technical Safety                          │
│   ├─ Robustness & Reliability                │
│   │   ├─ Fault Tolerance                     │
│   │   └─ Adversarial Robustness              │
│   ├─ Interpretability & Explainability       │
│   │   ├─ Model Transparency                  │
│   │   └─ Post-hoc Interpretations            │
│   └─ Verification & Validation               │
│       ├─ Formal Methods                      │
│       └─ Testing & Simulation                │
└──────────────────────────────────────────────┘
```

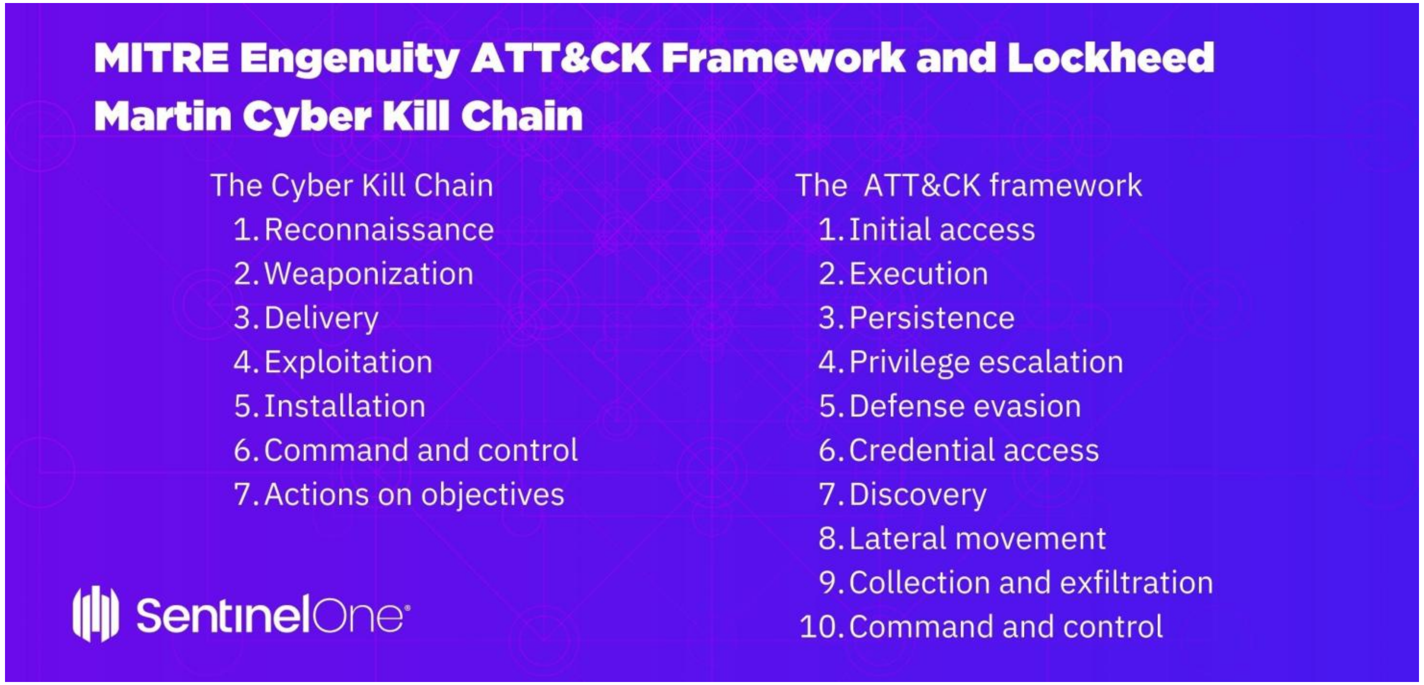


```
┌──────────────────────────────────────────────┐
│ 3. Risk Assessment & Management              │
│   ├─ Identifying Failure Modes               │
│   ├─ Accident & Catastrophic Risk Prevention │
│   ├─ Near-Term Societal Harms                │
│   └─ Long-Term Existential Risks             │
└──────────────────────────────────────────────┘
```



```
┌──────────────────────────────────────────────┐
│ 4. Policy & Oversight                        │
│   ├─ Regulatory Frameworks                   │
│   ├─ Standards & Best Practices              │
│   └─ Transparency & Auditing                 │
└──────────────────────────────────────────────┘
```

MITRE



MITRE Engenuity ATT&CK Framework and Lockheed Martin Cyber Kill Chain

The Cyber Kill Chain

1. Reconnaissance
2. Weaponization
3. Delivery
4. Exploitation
5. Installation
6. Command and control
7. Actions on objectives

The ATT&CK framework

1. Initial access
2. Execution
3. Persistence
4. Privilege escalation
5. Defense evasion
6. Credential access
7. Discovery
8. Lateral movement
9. Collection and exfiltration
10. Command and control

SentinelOne®

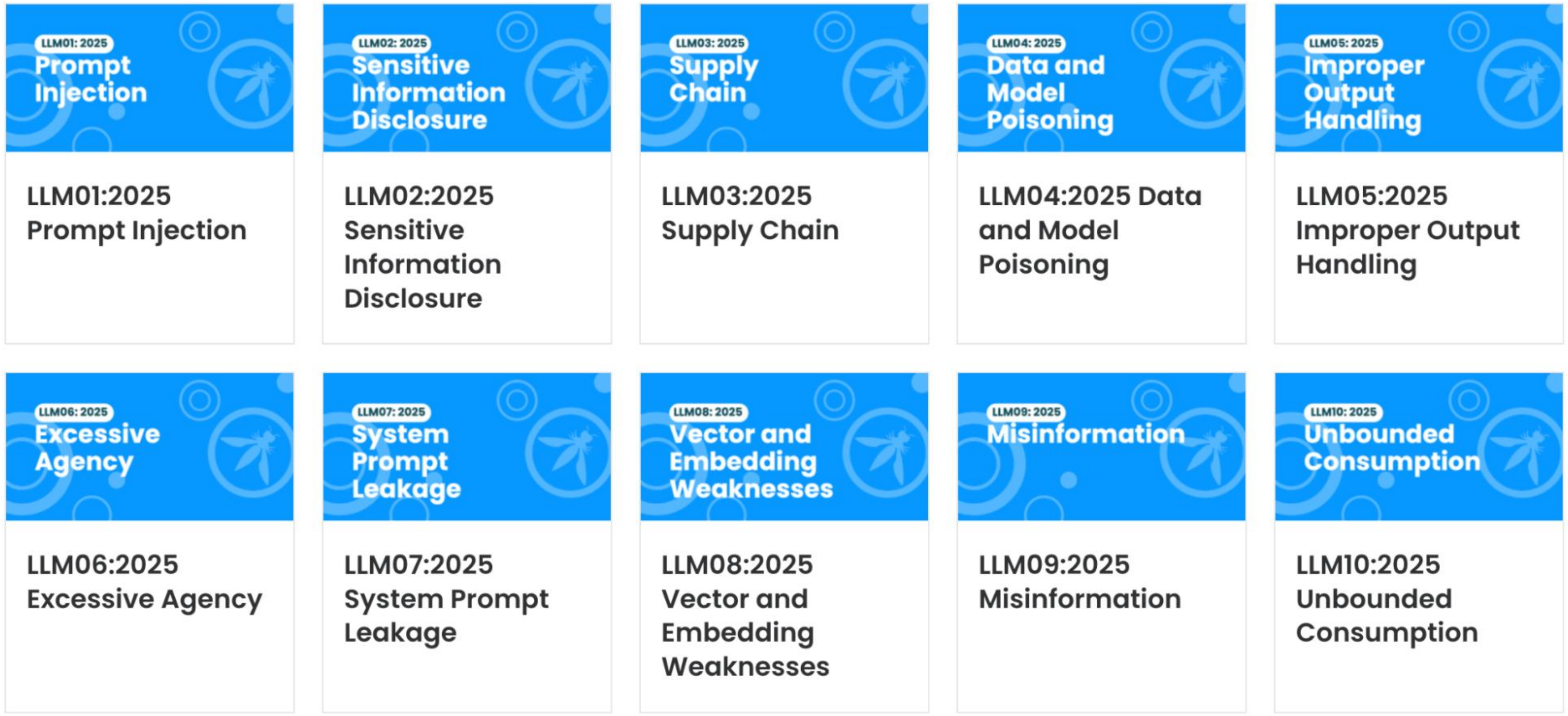
OWASP | TOP 10 LLM APPLICATIONS & GENERATIVE AI

OWASP Top 10 for LLM Applications 2025

Version 2025
November 18, 2024

OWASP PDF v4.2.0a 20241114-202703

https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/



LLM01:2025 Prompt Injection

LLM02:2025 Sensitive Information Disclosure

LLM03:2025 Supply Chain

LLM04:2025 Data and Model Poisoning

LLM05:2025 Improper Output Handling

LLM06:2025 Excessive Agency

LLM07:2025 System Prompt Leakage

LLM08:2025 Vector and Embedding Weaknesses

LLM09:2025 Misinformation

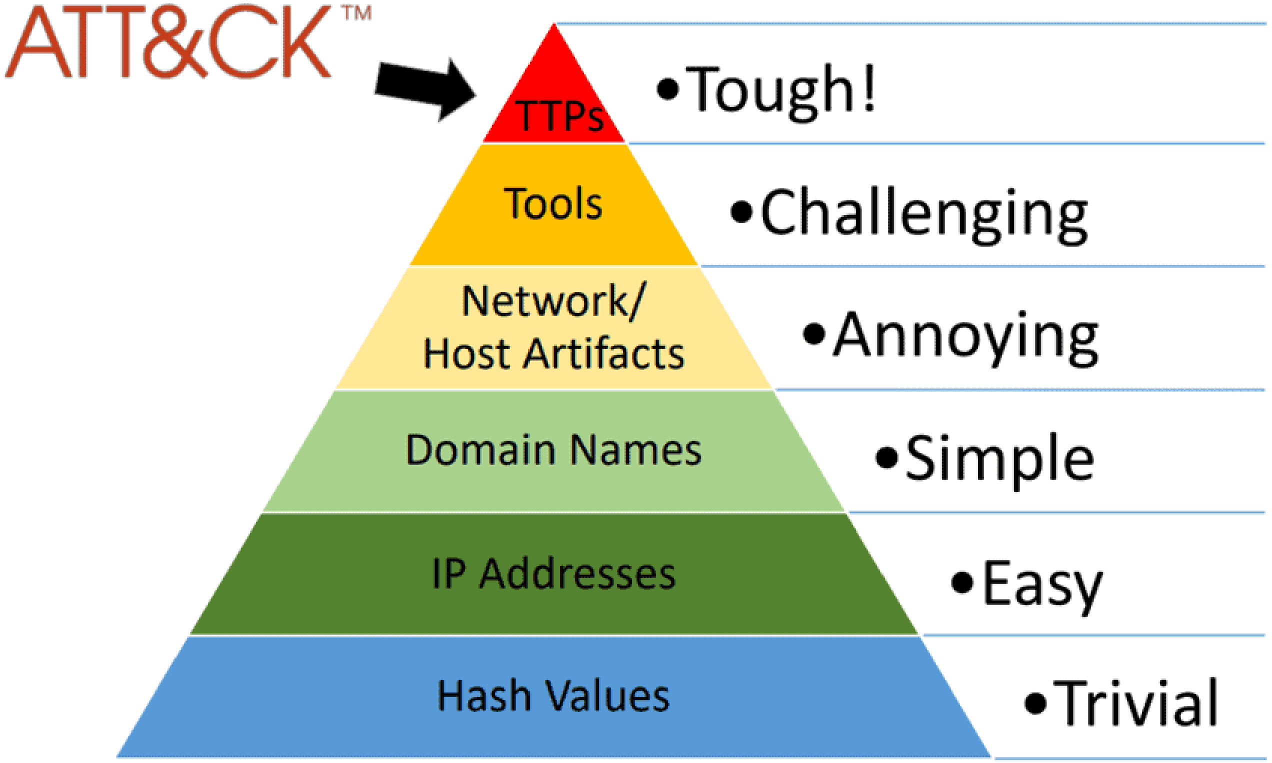
LLM10:2025 Unbounded Consumption

OWASP | TOP 10 LLM APPLICATIONS & GENERATIVE AI

OWASP Top 10 for LLM Applications 2025

Version 2025
November 18, 2024

OWASP PDF v4.2.0a 20241114-202703



ATT&CK™

TTPs •Tough!

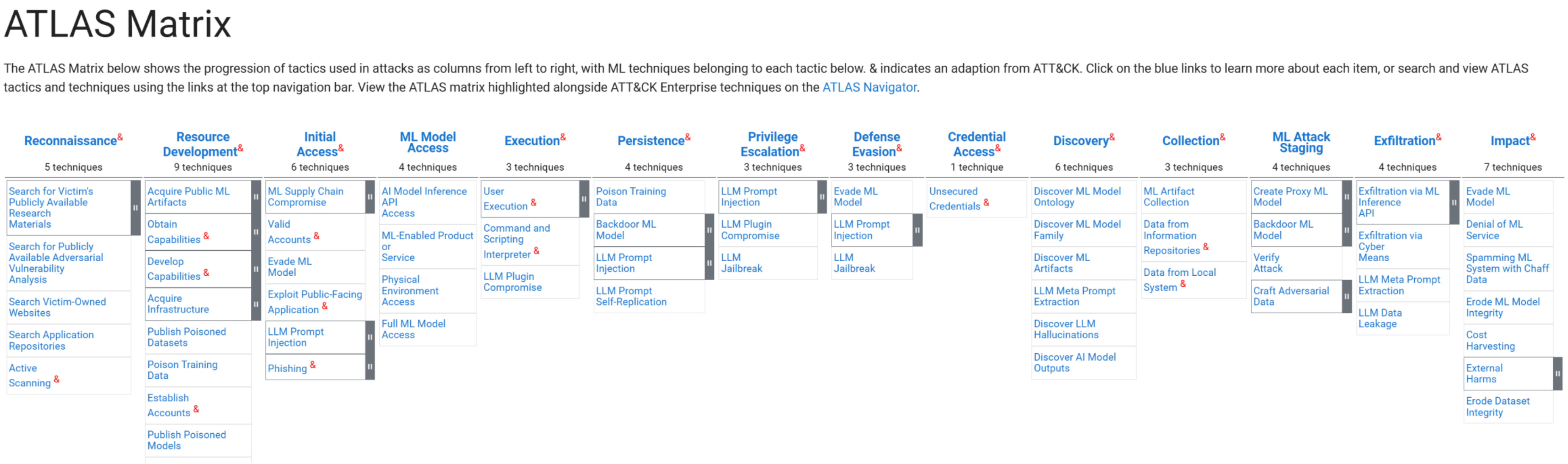
Tools •Challenging

Network/ Host Artifacts •Annoying

Domain Names •Simple

IP Addresses •Easy

Hash Values •Trivial



ATLAS Matrix

The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML techniques belonging to each tactic below. & indicates an adaption from ATT&CK. Click on the blue links to learn more about each item, or search and view ATLAS tactics and techniques using the links at the top navigation bar. View the ATLAS matrix highlighted alongside ATT&CK Enterprise techniques on the ATLAS Navigator.

Reconnaissance& 5 techniques	Resource Development& 9 techniques	Initial Access& 6 techniques	ML Model Access 4 techniques	Execution& 3 techniques	Persistence& 4 techniques	Privilege Escalation& 3 techniques	Defense Evasion& 3 techniques	Credential Access& 1 technique	Discovery& 6 techniques	Collection& 3 techniques	ML Attack Staging 4 techniques	Exfiltration& 4 techniques	Impact& 7 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	AI Model Inference API Access	User Execution &	Poison Training Data	LLM Prompt Injection	Evade ML Model	Unsecured Credentials &	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Command and Scripting Interpreter &	Backdoor ML Model	LLM Plugin Compromise	LLM Prompt Injection		Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Capabilities &	Evade ML Model	Physical Environment Access	LLM Plugin Compromise	LLM Prompt Injection	LLM Jailbreak	LLM Jailbreak		Discover ML Artifacts	Data from Local System &	Verify Attack	LLM Meta Prompt Extraction	Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model Access		LLM Prompt Self-Replication				LLM Meta Prompt Extraction		Craft Adversarial Data	LLM Data Leakage	Erode ML Model Integrity
Active Scanning &	Publish Poisoned Datasets	LLM Prompt Injection							Discover LLM Hallucinations				Cost Harvesting
	Poison Training Data	Phishing &							Discover AI Model Outputs				External Harms
	Establish Accounts &												Erode Dataset Integrity
	Publish Poisoned Models												

Recap Quiz

Policy

A policy is a set of guiding principles or rules designed to influence decisions and actions within an organization or system.

A university enacts a policy on academic integrity, requiring students to follow strict guidelines to prevent plagiarism and ensure honest scholarly work.

A tech company implements a policy that requires rigorous bias testing for all AI models, mandating regular evaluations and updates to address any discriminatory outcomes.

Regulation

Regulation is the creation and enforcement of rules by a governing authority to guide or restrict the behavior of individuals and organizations.

Environmental protection laws limit industrial emissions, penalizing companies that exceed legally mandated thresholds to preserve air and water quality.

The European Union’s proposed AI Act sets standards for high-risk AI systems, specifying transparency, data governance, and human oversight requirements to ensure safe and responsible AI deployment.

Compliance

Compliance is the act of adhering to applicable laws, regulations, standards, and internal policies to ensure that organizational activities meet required legal and ethical standards.

A financial institution implements strict anti-money laundering policies, conducting thorough customer due diligence and transaction monitoring to meet regulatory obligations and prevent illegal activity.

An AI startup ensures it follows data protection regulations (e.g., GDPR) when collecting, storing, and analyzing user data—obtaining proper consent and safeguarding information to avoid privacy violations.

Governance

Governance is the framework of rules, structures, and processes that guide decision-making, ensure accountability, and balance the interests of stakeholders.

A municipal government sets up a procurement committee that follows strict procedures for public tenders, monitors budget spending, and holds officials accountable—ensuring fair and efficient use of taxpayer money.

A tech company establishes an AI ethics board to set guidelines on data use, oversee model development, and enforce policies to prevent bias—ensuring AI systems are transparently managed and ethically deployed.

General Oversight & Control

- Supervision
- Management
- Oversight
- Stewardship
- Accountability
- Administration
- Control

Legal & Regulatory

- Legislation
- Statutes
- Mandates
- Directives
- Ordinances
- Bylaws

Compliance & Risk

Enforcement

Adherence

Conformance

Standardization

Audit

Risk Management

Due Diligence

Assurance

Corporate & Organizational Governance

- Board Oversight
- Ethics
- Best Practices
- Internal Controls
- Decision Rights
- Enterprise Risk Management

Policy & Frameworks

- Guidelines
- Principles
- Protocols
- Procedures
- Standards
- Frameworks

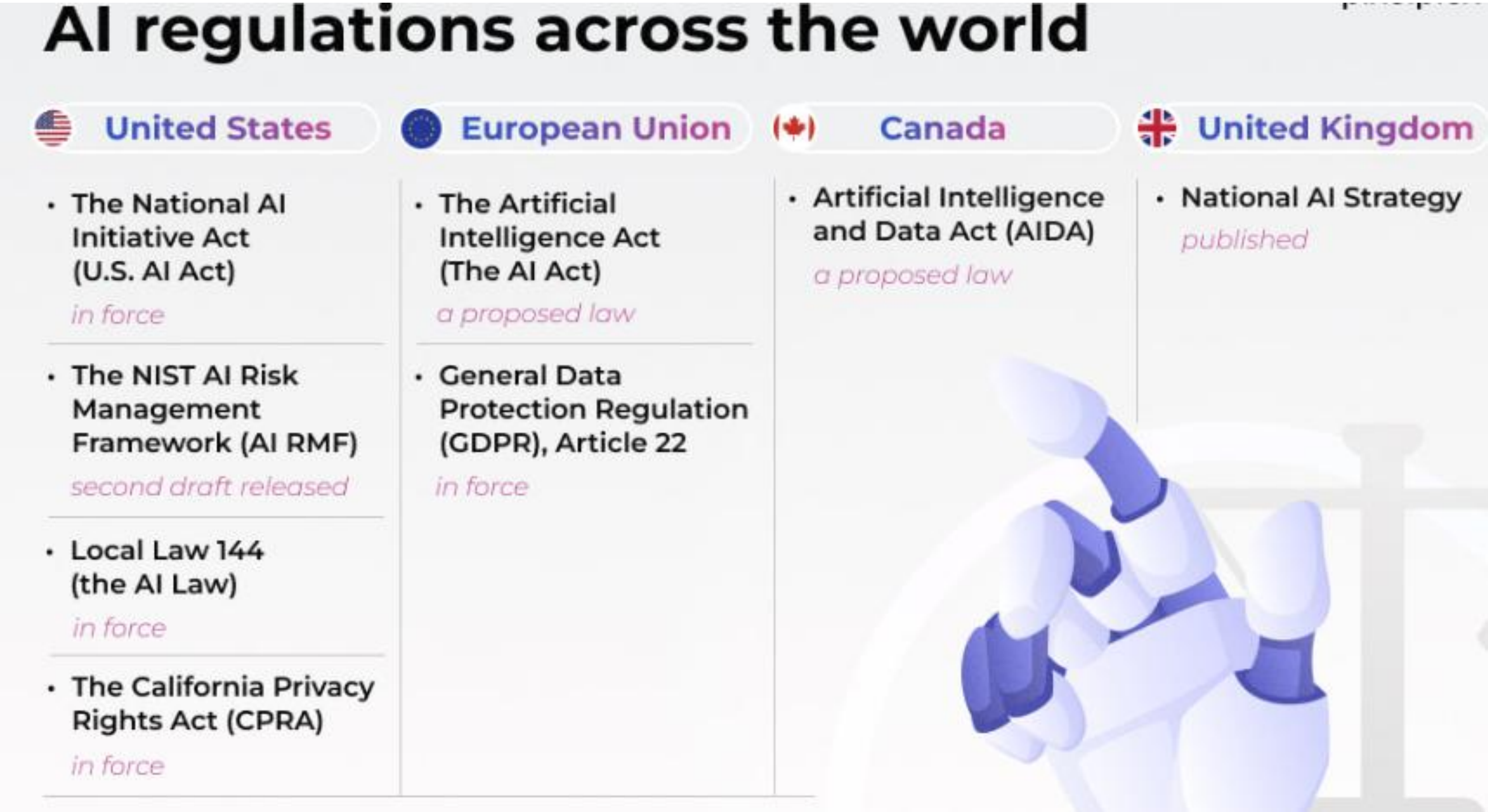
In-depth Analysis of the Current State of AI Regulations Globally

Covering compliance frameworks, ethical implications, and enforcement mechanisms.

Agenda:

- National AI regulations in the U.S. (federal and state-level policies) AI
- regulations in Europe (EU AI Act)
- AI regulatory frameworks in Asia, Latin America, and Africa Role of
- academic organizations (IEEE) in AI governance
- Key non-profit and industry organizations contributing to AI policy
- Impact of AI regulations on businesses, especially in finance and healthcare
- Overview of Pakistan’s digital services and AI regulations

Current State of AI Governance, Compliance, and Regulation Globally



AI regulations across the world

United States

- The National AI Initiative Act (U.S. AI Act)
  in force
- The NIST AI Risk Management Framework (AI RMF)
  second draft released
- Local Law 144 (the AI Law)
  in force
- The California Privacy Rights Act (CPRA)
  in force

European Union

- The Artificial Intelligence Act (The AI Act)
  a proposed law
- General Data Protection Regulation (GDPR), Article 22
  in force

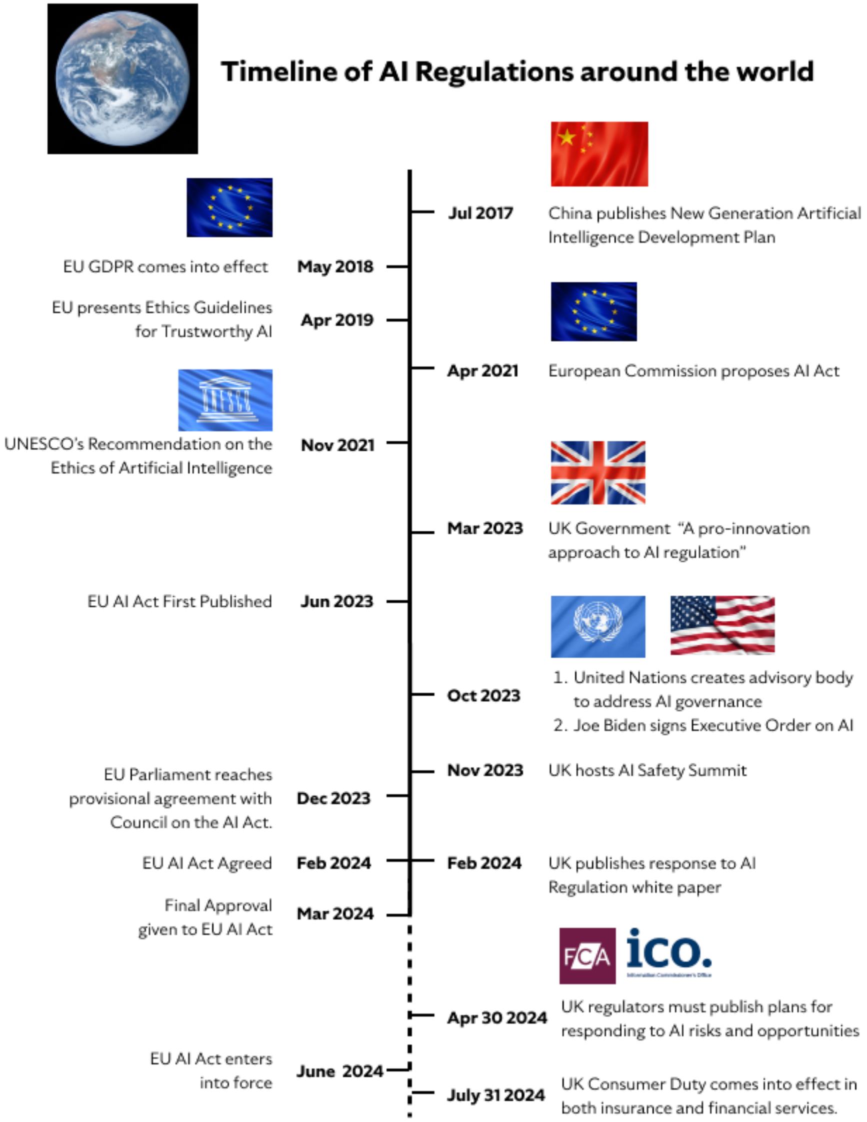
Canada

- Artificial Intelligence and Data Act (AIDA)
  a proposed law

United Kingdom

- National AI Strategy
  published

Timeline of AI Regulations around the world



Jul 2017 China publishes New Generation Artificial Intelligence Development Plan

May 2018 EU GDPR comes into effect

Apr 2019 EU presents Ethics Guidelines for Trustworthy AI

Apr 2021 European Commission proposes AI Act

Nov 2021 UNESCO's Recommendation on the Ethics of Artificial Intelligence

Mar 2023 UK Government "A pro-innovation approach to AI regulation"

Jun 2023 EU AI Act First Published

Oct 2023

1. United Nations creates advisory body to address AI governance
2. Joe Biden signs Executive Order on AI

Nov 2023 UK hosts AI Safety Summit

Dec 2023 EU Parliament reaches provisional agreement with Council on the AI Act.

Feb 2024 EU AI Act Agreed

Feb 2024 UK publishes response to AI Regulation white paper

Mar 2024 Final Approval given to EU AI Act

Apr 30 2024 UK regulators must publish plans for responding to AI risks and opportunities

June 2024 EU AI Act enters into force

July 31 2024 UK Consumer Duty comes into effect in both insurance and financial services.

Global AI Regulations Overview

International Principles: There is no single global AI law, but several international bodies have set **principles and frameworks** to guide AI governance. The **OECD AI Principles (2019)**, endorsed by 47 countries, outline values for *"innovative and trustworthy AI"* that respects human rights (AI Principles Overview - OECD.AI). Key OECD principles include fairness, transparency, robustness, safety, and accountability (AI Principles Overview - OECD.AI). Similarly, UNESCO's **Recommendation on the Ethics of AI (2021)** established ten guiding principles – such as transparency, non-discrimination, accountability, and human oversight – to align AI with human rights and societal well-being. These global guidelines are voluntary but influential, encouraging nations to adopt **ethical AI practices**.

Global Cooperation: Multilateral efforts are emerging to coordinate AI policy.

The **G20** and **G7** have issued statements on trustworthy AI, and the **Global Partnership on AI (GPAI)** was launched by governments to facilitate collaboration on responsible AI development (Global Partnership on Artificial Intelligence - OECD). The **United Nations** has also begun examining AI's risks; in 2023 the UN Secretary-General proposed creating an international AI watchdog, reflecting a growing consensus that some global oversight is needed. Enforcement of AI ethics globally still relies on national laws, but these international principles lay the groundwork for interoperable regulations. In practice, many countries are using the OECD and UNESCO guidelines as a basis to craft national AI policies.

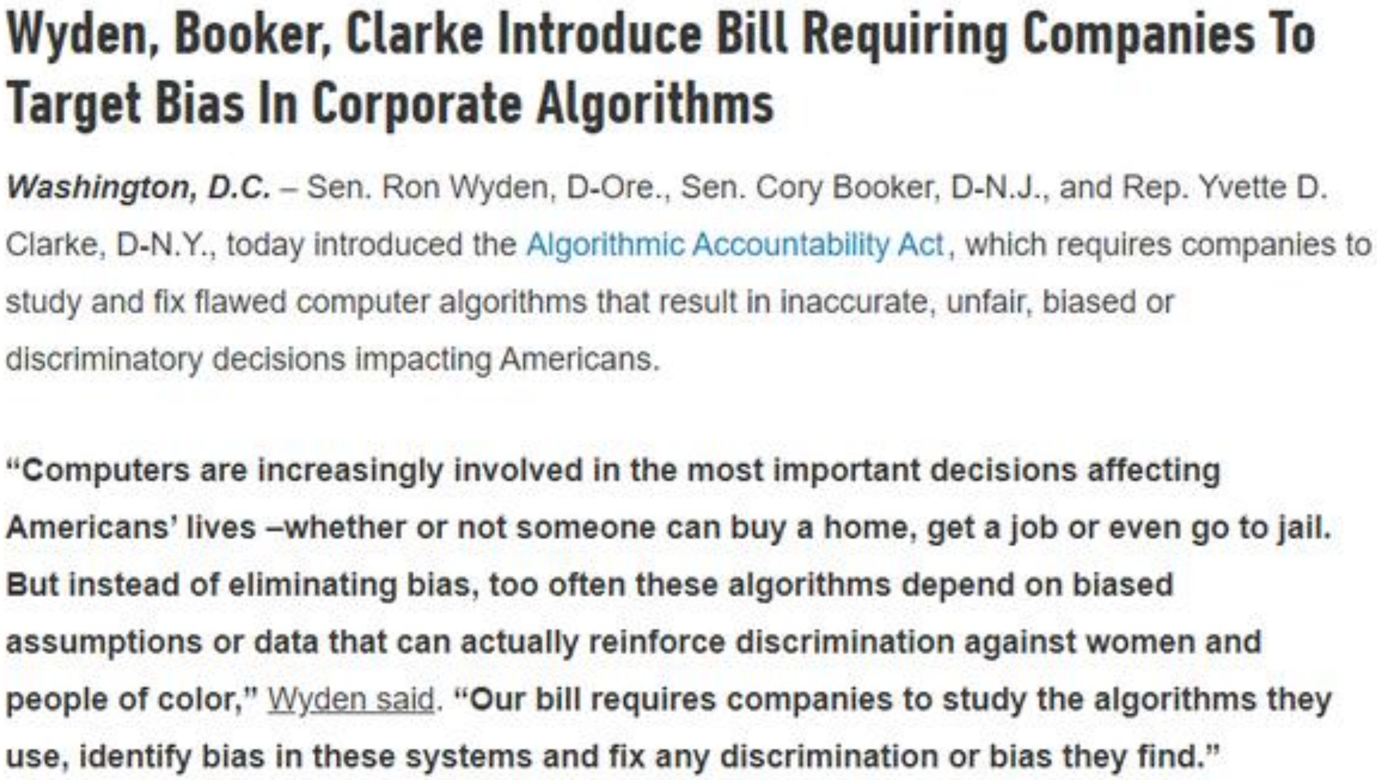
https://oecd.ai/en/wonk/evolving-with-innovation-the-2024-oecd-ai-principles-update

Global AI Regulations Overview

United States (National-Level Regulations)

Federal AI Frameworks: The U.S. has not enacted a comprehensive AI law, but it has developed **frameworks and policies** to guide AI governance. In early 2023, NIST released its **AI Risk Management Framework (RMF) 1.0**, a voluntary guidance for organizations to manage AI risks (AI Risk Management Framework | NIST). It outlines functions like *Map, Measure, Manage, and Govern* to help identify and mitigate risks such as bias, lack of transparency, and security issues. While not mandatory, the NIST framework has become a de facto standard for AI risk governance in industry.

Federal Policies and Executive Actions: The White House has issued policies emphasizing AI ethics. In October 2022, the OSTP published a **Blueprint for an AI Bill of Rights**. In October 2023, President Biden signed a **landmark Executive Order on "Safe, Secure, and Trustworthy AI"**, directing actions to manage AI's safety and security risks, protect privacy, advance equity, and promote innovation (Fact Sheet: Key AI Accomplishments...). This Executive Order invokes the Defense Production Act to require companies developing powerful AI models to share safety test results with the government and address AI security risks. It also mandates NIST to set new AI safety standards and directs federal agencies to ensure their AI use is transparent, unbiased, and accountable. Earlier, under the prior administration, the **2019 American AI Initiative** and a **2020 EO on trustworthy AI in government** laid groundwork for balancing innovation with precautions.



Wyden, Booker, Clarke Introduce Bill Requiring Companies To Target Bias In Corporate Algorithms

Washington, D.C. – Sen. Ron Wyden, D-Ore., Sen. Cory Booker, D-N.J., and Rep. Yvette D. Clarke, D-N.Y., today introduced the Algorithmic Accountability Act, which requires companies to study and fix flawed computer algorithms that result in inaccurate, unfair, biased or discriminatory decisions impacting Americans.

"Computers are increasingly involved in the most important decisions affecting Americans' lives –whether or not someone can buy a home, get a job or even go to jail. But instead of eliminating bias, too often these algorithms depend on biased assumptions or data that can actually reinforce discrimination against women and people of color," Wyden said. **"Our bill requires companies to study the algorithms they use, identify bias in these systems and fix any discrimination or bias they find."**

- **Proposed Legislation:** U.S. lawmakers have introduced bills to directly regulate AI, though none have passed yet. The **Algorithmic Accountability Act** would require companies to conduct impact assessments for automated decision systems in finance, healthcare, housing, andemployment.

- Another bill, the **National AI Commission Act (2023)**, proposes creating a national commission to study AI regulation. Federal regulators assert existing laws already apply to AI (for example, the FTC's stance on deceptive AI practices). In practice, U.S. AI governance at the national level is a patchwork of *voluntary frameworks (NIST), executive directives, sector-specific regulations, and enforcement of existing laws.*

United States (State-Level Regulations)

State AI Laws and Initiatives: In the absence of a broad federal AI law, U.S. states are introducing their own regulations. **California** is developing rules for automated decision systems. Under the **CPRA (2020)**, the state's privacy agency must issue regulations on *"automated decision-making technology"*, giving consumers rights to opt-out of profiling and receive *meaningful information about the logic* behind AI decisions. A proposed **AB 331** sought to require developers to perform impact assessments and notify individuals when AI is used in consequential decisions. California's legislature has introduced multiple AI bills ranging from AI transparency labels to creating an Office of AI.

New York (City): NYC's **Local Law 144 (2021)** requires employers using Automated Employment Decision Tools (AEDTs) for hiring or promotions to conduct an **annual bias audit** of those AI tools and **notify** candidates about AI usage. Since July 2023, any AI used in NYC hiring must undergo independent bias audits and public disclosure of results. Illinois similarly passed the **Artificial Intelligence Video Interview Act (2019)**, mandating notice, consent, and limited data sharing for AI-based video interviews.

Focus Areas:

Bias mitigation and fairness (NYC's bias audit law, Illinois' interview act)

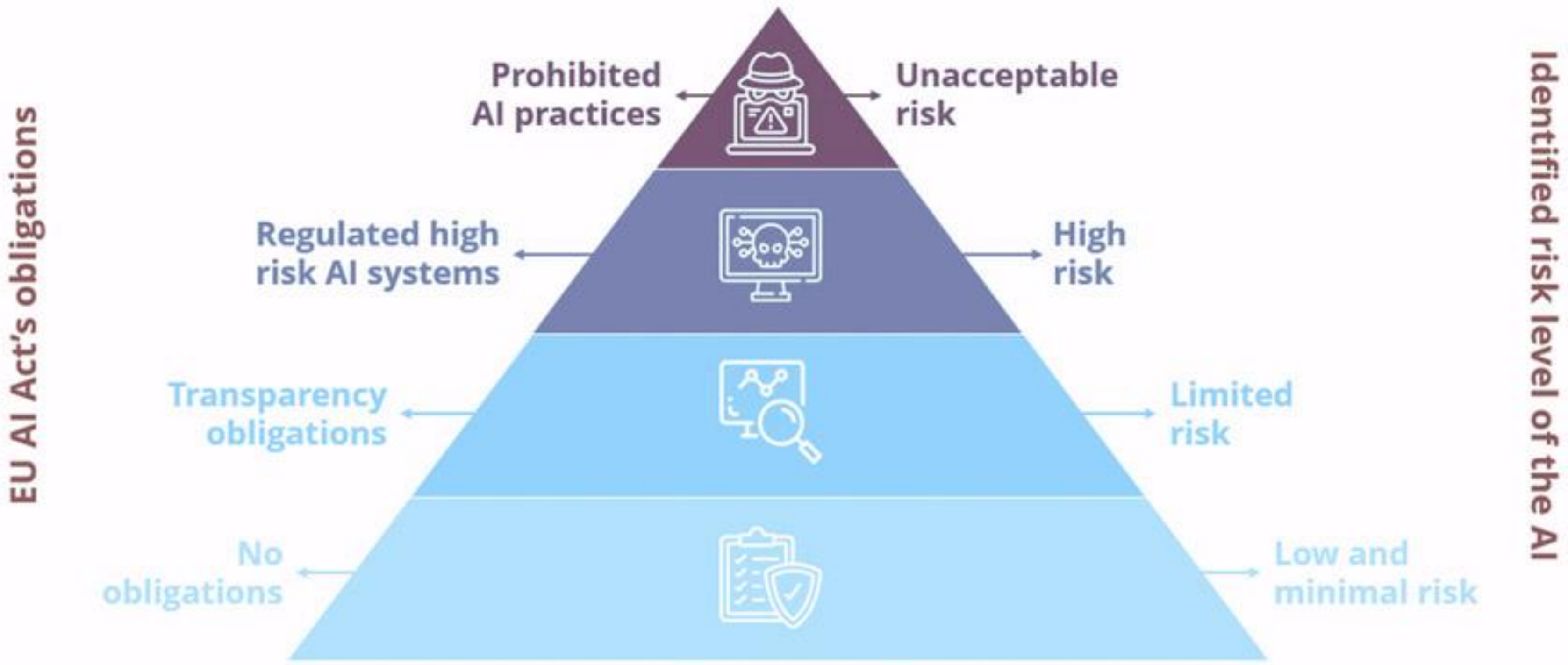
Privacy (states with data protection laws granting rights related to automated profiling)

Accountability and transparency (algorithmic impact assessments, public disclosure when AI is used in high-stakes decisions)

These state initiatives illustrate how sub-national governments are stepping in to regulate AI's risks in specific domains.

Europe’s AI Regulations

The EU AI Act employs a risk-based approach to regulate AI systems based on their level of risk



EU AI Act's obligations

Identified risk level of the AI

Prohibited AI practices ← → Unacceptable risk

Regulated high risk AI systems ← → High risk

Transparency obligations ← → Limited risk

No obligations ← → Low and minimal risk

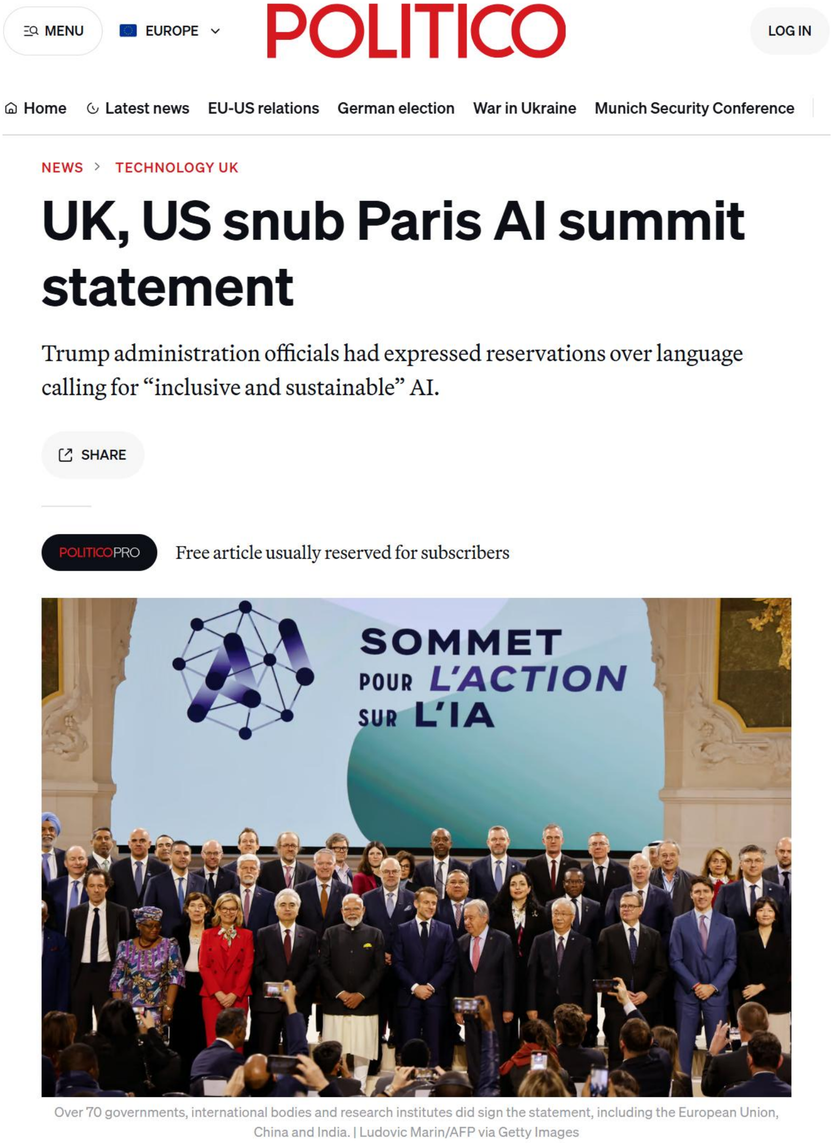
EU AI Act: The European Union is enacting the **world’s first comprehensive AI law** – the *EU Artificial Intelligence Act*. A provisional agreement was reached in late 2023, aiming for implementation by 2025–2026. The Act takes a *risk-based approach*:

Unacceptable Risk: AI uses threatening safety/fundamental rights are banned (e.g. social scoring, manipulative AI).

High Risk: AI in critical areas (infrastructure, healthcare, education, employment, law enforcement, credit scoring) is allowed but heavily regulated. Providers must perform risk assessments, ensure data quality, maintain technical documentation, transparency, human oversight, etc.

Limited Risk: Transparency obligations (e.g. chatbots must disclose they’re AI, generative AI must label deepfakes).

Minimal/No Risk: Mostly unregulated (e.g. spam filters).

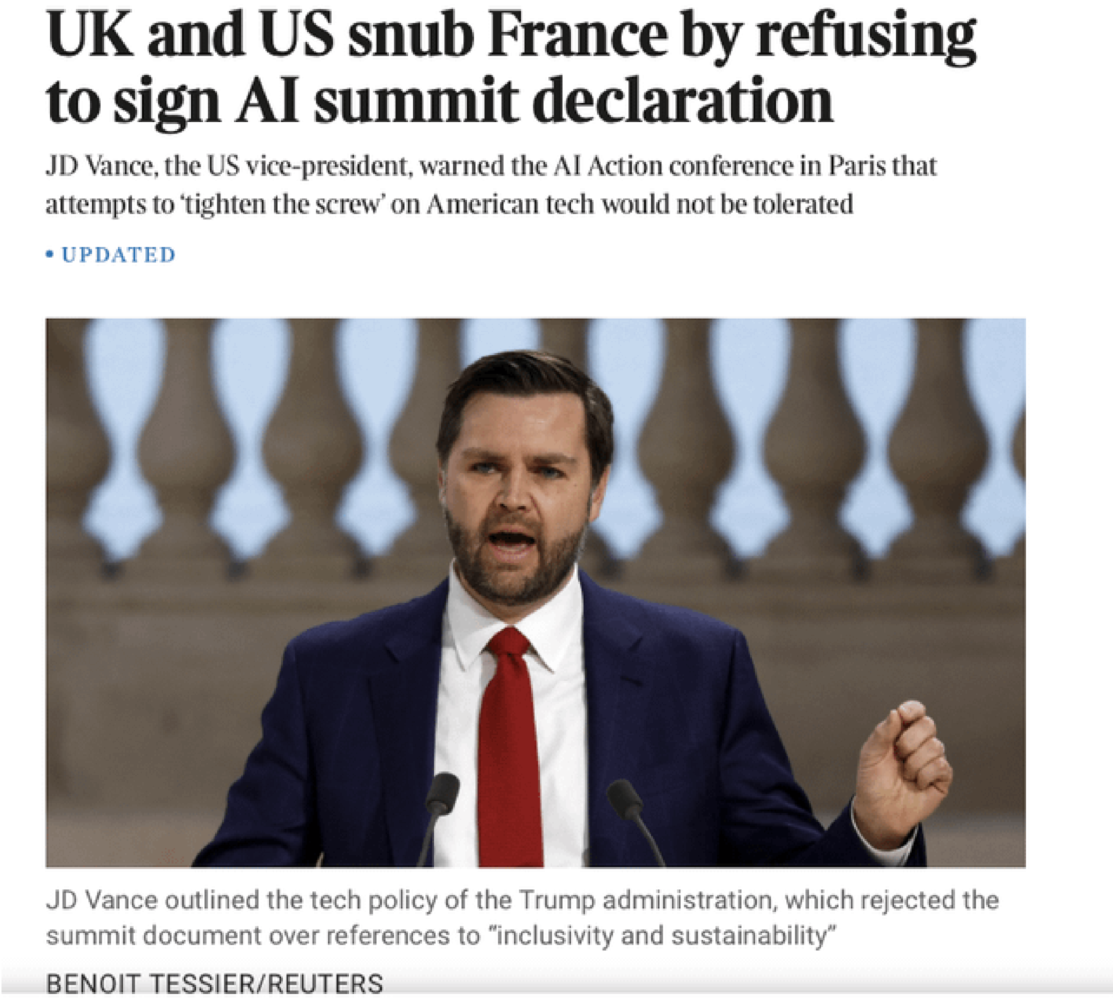


UK, US snub Paris AI summit statement

Trump administration officials had expressed reservations over language calling for "inclusive and sustainable" AI.

Free article usually reserved for subscribers

Over 70 governments, international bodies and research institutes did sign the statement, including the European Union, China and India. | Ludovic Marin/AFP via Getty Images



UK and US snub France by refusing to sign AI summit declaration

JD Vance, the US vice-president, warned the AI Action conference in Paris that attempts to 'tighten the screw' on American tech would not be tolerated

• UPDATED

JD Vance outlined the tech policy of the Trump administration, which rejected the summit document over references to "inclusivity and sustainability"

BENOIT TESSIER/REUTERS

PRINCIPLED ARTIFICIAL INTELLIGENCE

A Map of Ethical and Rights-Based Approaches to Principles for AI

Authors: Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, Madhulika Srikumar

Designers: Arushi Singh (arushisingh.net) and Melissa Axelrod (melissaaxelrod.com)

HOW TO READ:

Date, Location
Document Title
Actor

COVERAGE OF THEMES:

Higher ← → Lower

● ● ● • • ◦

Not referenced

◆ References International Human Rights

✸ Explicitly Adopts Human Rights Framework

The size of each dot represents the percentage of principles in that theme contained in the document. Since the number of principles per theme varies, it's informative to compare dot sizes within a theme but not between themes.

The principles within each theme are:

Privacy:
Privacy
Control over Use of Data
Consent
Privacy by Design
Recommendation for Data Protection Laws
Ability to Restrict Processing
Right to Rectification
Right to Erasure

Accountability:
Accountability
Recommendation for New Regulations
Impact Assessment
Evaluation and Auditing Requirement
Verifiability and Replicability
Liability and Legal Responsibility
Ability to Appeal
Environmental Responsibility
Creation of a Monitoring Body
Remedy for Automated Decision

Safety and Security:
Security
Safety and Reliability
Predictability

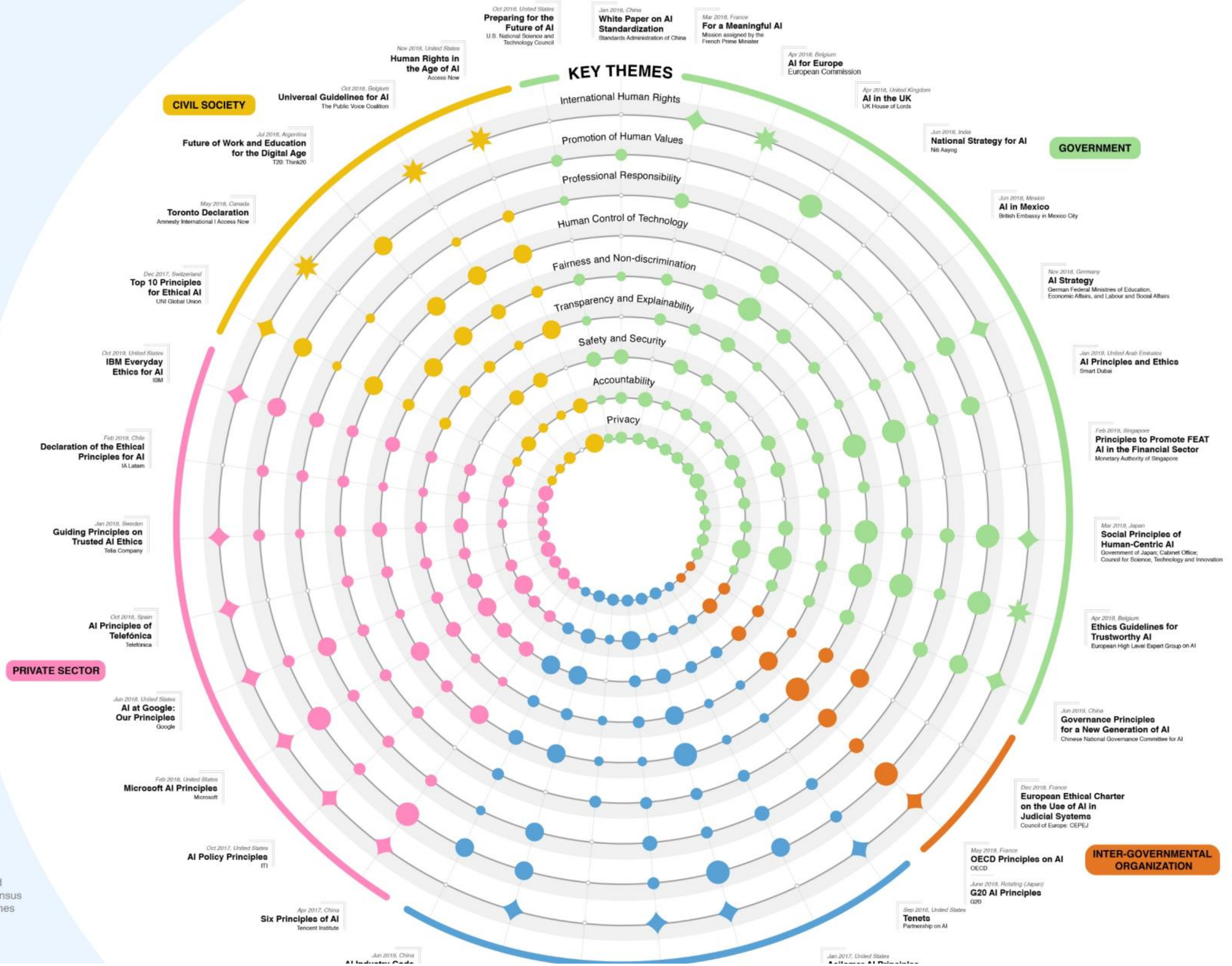
Transparency and Explainability:
Explainability
Transparency
Open Source Data and Algorithms
Notification when Interacting with an AI
Notification when AI Makes a Decision about an Individual
Regular Reporting Requirement
Right to Information
Open Procurement (for Government)

Fairness and Non-discrimination:
Non-discrimination and the Prevention of Bias
Fairness
Inclusiveness in Design
Inclusiveness in Impact
Representative and High Quality Data
Equality

Human Control of Technology:
Human Control of Technology
Human Review of Automated Decision
Ability to Opt out of Automated Decision

Professional Responsibility:
Multistakeholder Collaboration
Responsible Design
Consideration of Long Term Effects
Accuracy
Scientific Integrity

Further information on findings and methodology is available in Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches (Berkman Klein, 2020) available at ***cyber.harvard.edu****.*



There will be **EU-wide enforcement**, with national regulators supervising compliance and a European AI Board coordinating consistency. Fines can reach *€30 million or 6% of global turnover*. The Act also covers **foundation models and generative AI**.

AI Governance under GDPR: Even before the AI Act, the EU's **GDPR** impacts automated decision-making. *Article 22* of GDPR grants individuals the right *"not to be subject to a decision based solely on automated processing"* producing significant effects, unless safeguards apply. This effectively requires a **human-in-the-loop** or a mechanism to contest such decisions. GDPR also demands transparency and fairness regarding personal data used in AI.

National Initiatives: EU Member States have their own AI strategies, generally aligning with EU-wide rules. France's CNIL, Spain's new AI supervisory agency, and Germany's AI Ethics Commission exemplify national efforts. Sector-specific regulations (medical devices, finance) also govern AI. The EU's *Ethics Guidelines for Trustworthy AI (2019)* shaped many of these measures, emphasizing human agency, technical robustness, privacy, transparency, diversity, and accountability.

AI Regulations in Other Regions

Asia (China, Japan, India)

China: China focuses on **state control, content governance, and ethical norms**. In 2022, the Cyberspace Administration issued *Algorithms Regulation*, requiring algorithm providers to *“observe social morality and ethics”*, register major algorithms, avoid misinformation, and prevent discrimination. In 2023, **Deep Synthesis Regulation** mandates labeling AI-generated media (deepfakes) and restricts deceptive uses. China’s **Interim Measures for Generative AI Services** impose content restrictions aligned with “core socialist values,” require measures to prevent discriminatory outputs, label AI-generated content, and register with authorities if influencing public opinion. The *Personal Information Protection Law (2021)* imposes GDPR-like privacy rules on AI data handling.

Japan: Japan employs *“agile governance”* and a soft-law approach. *AI Governance Guidelines for Business (2023)* offer voluntary best practices, building on the *Social Principles of Human-Centric AI (2019)*. Existing laws (privacy, consumer protection) and sectoral rules apply to AI. Japan avoids heavy-handed regulation to foster innovation, though it debates legal frameworks for high-risk AI. Japan advocates internationally (e.g. G7 Hiroshima AI Process) for shared AI governance.

India: India has **no AI-specific law** yet. Policy documents like *NITI Aayog’s National Strategy for AI (2018)* and *Principles for Responsible AI (2021)* guide *“AI for social good”*. The government is pro-innovation, relying on existing frameworks (IT laws, sectoral regulation). A new **Digital Personal Data Protection Act (2023)** will affect AI by regulating data processing. India is forming committees to explore AI regulation; more concrete measures may emerge as adoption grows.

Latin America (Brazil, others)

Brazil: Brazil leads AI governance in Latin America. It has a *National AI Strategy (2021)* and is close to passing a comprehensive **AI law** (Bill No. 2338/2023) with a risk-based approach similar to the EU. The proposed law bans harmful AI practices and requires transparency, oversight, and nondiscrimination for high-risk AI. Brazil's **LGPD** (data protection law) also constrains AI. Approval of the AI bill would make Brazil one of the first non-European nations with a dedicated AI Act.

Other LATAM: Mexico, Chile, Argentina, Colombia all have AI strategies focusing on ethics, innovation, and privacy. Sector-specific usage (fintech, facial recognition) has prompted calls for regulation. Regulatory sandboxes are popular (Brazil, Mexico) to test AI under oversight. Latin America's AI governance largely relies on *soft law and existing data protection rules*, with momentum building for formal AI regulations (led by Brazil).

Africa (South Africa, Nigeria)

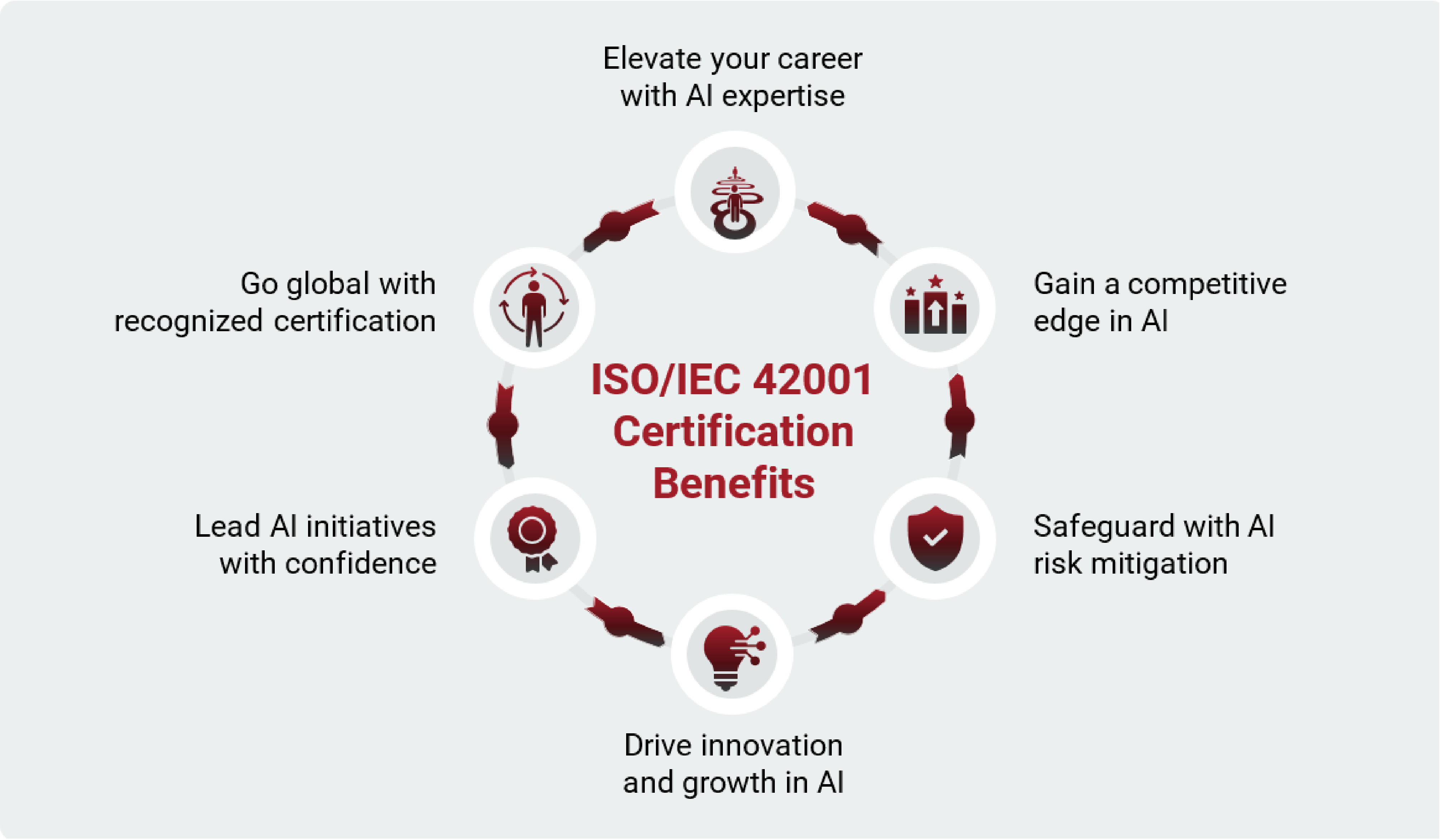
South Africa: No AI-specific law yet. A *National AI Policy Framework (2021)* lays out an ethical and responsible AI vision. AI is regulated indirectly via existing laws (e.g., Protection of Personal Information Act). Sector regulators in finance/telecom oversee AI under broader mandates. South Africa launched an **AI Institute** and consults on a future regulatory approach. It also participates in the African Union's draft AI Strategy.

Nigeria: Formulating a **National AI Policy** to guide adoption and address risks. Current regulation relies on general frameworks (Nigeria Data Protection Regulation, financial rules). Government uses AI in security/surveillance, prompting civil society concerns about transparency.

A new *Data Protection Act (2023)* sets stricter rules for personal data in AI. Nigeria's approach, similar to many African nations, emphasizes *innovation balanced with ethical safeguards*.

Academic & Industry Standards

- **Role of Standards Organizations:**
  - IEEE: The Global Initiative on Ethical AI produced *Ethically Aligned Design* (2019) and the *P7000 series* of standards (e.g. IEEE 7000-2021) for addressing ethical concerns in system design. IEEE’s Ethics Certification Program for AI Systems (ECPAIS) certifies AI products against ethical criteria.
  - ISO/IEC: The joint committee *ISO/IEC JTC 1/SC 42* develops AI standards on terminology, architecture, governance, bias mitigation, and risk management. *ISO/IEC 42001* is a draft management system standard for trustworthy AI.
- These standards shape best practices, bridging academia, industry, and regulators. Many laws reference or adopt these international standards.
- Ethical Frameworks: Academic groups have published principles (Asilomar AI Principles,
- Montreal Declaration, ACM code updates). These inform IEEE/ISO standards and help define “*trustworthy AI.*”
- Industry Self-Regulation: Tech companies have *AI Ethics Charters* (Google, Microsoft, IBM), often guided by standard-setting bodies. Collaborative platforms like the *Partnership on AI* and NIST workshops develop open frameworks (e.g. Model Cards for transparency). Standards are crucial in operationalizing high-level ethical principles into tangible requirements.



ISO/IEC 42001 Certification Benefits

Elevate your career with AI expertise

Gain a competitive edge in AI

Safeguard with AI risk mitigation

Drive innovation and growth in AI

Lead AI initiatives with confidence

Go global with recognized certification

Non-Profit & Policy Organizations

Several nonprofits and coalitions shape AI policy and promote responsible AI worldwide:

Partnership on AI (PAI): A global multi-stakeholder nonprofit with 100+ partners from industry, academia, civil society. Works on fairness, transparency, labor impacts, etc. Issues best practices, research, and policy advice.

OpenAI: Develops frontier models (e.g. GPT-4). Advocates for AI regulation, has proposed safety standards, independent audits for powerful AI systems, etc. Participates in voluntary commitments for watermarking AI content and other safeguards.

Future of Life Institute (FLI): Focuses on existential risks. In March 2023, coordinated an Open Letter calling for a pause on training AI more powerful than GPT-4 to allow time for safety research. Engages policymakers and promotes licensing/liability for advanced AI.

AI Now Institute: Produces critical studies on social implications (bias, facial recognition, etc.). Calls for bans on certain harmful uses (e.g. government facial recognition). Influential in policy discussions, advocating *algorithmic accountability* frameworks.

Other Key Orgs:

- *World Economic Forum* (AI Governance initiative)
- *OECD AI Policy Observatory*
- *Berkman Klein Center, Alan Turing Institute, Ada Lovelace Institute*
- *Center for AI and Digital Policy (CAIDP)*

These groups research impacts, raise awareness, and advocate for civil liberties. Their recommendations help shape the emerging regulatory landscape worldwide.

Business Impacts of AI Regulations

Compliance Requirements: Companies see new AI regulations increasing costs (data privacy, security, fairness). Many are *reviewing and updating data practices* and investing in **AI transparency and fairness** measures. They're creating *AI governance programs*, with risk/ethics committees, and *algorithmic impact assessments*.

Accountability and Documentation: Regulations (EU AI Act, U.S. bills) require detailed **documentation of AI systems** (training data, design, risk mitigation). Firms must provide explanations for decisions (e.g. credit denials under fair lending laws). *Legal accountability* rests on companies, not the "black box" algorithm.

Impact on Innovation vs. Risk Management: Clear rules can boost public trust in AI but add compliance costs. Many large companies adopt *AI Ethics Principles* and internal review boards, partly to demonstrate good faith. Responsible AI reduces scandal risk and fosters sustainable use.

Data Protection and Governance: Privacy laws (e.g. GDPR) heavily influence AI. Consent, data minimization, user rights to opt out or correct data. Some laws require *data audits* and bias testing in high-risk AI systems.

Sectoral Adjustments:

- **Finance:** Banks apply Model Risk Management for AI (stress-testing, validation).
- **Healthcare:** AI-based medical devices require regulatory approval (FDA, MDR in EU).

Competitive Advantage: Compliance can become a differentiator; trustworthy AI attracts consumers and avoids reputational damage. This drives growth in *AI compliance tools and consultancies*. Responsible AI is increasingly seen as *good business*.

Sectoral Impact of AI Regulations

Healthcare

Medical Device Regulation: AI diagnostic tools are often classified as *medical devices*, requiring FDA (U.S.) or CE (EU) approval. Regulators are adapting frameworks for continuously learning algorithms. FDA’s AI/ML-based SaMD action plan requires **transparency** and **bias mitigation** strategies. The EU AI Act designates AI in medical contexts as *high risk*.

Patient Safety and Efficacy: AI must undergo clinical validation. Regulators demand thorough **testing** and **oversight** to ensure reliability across demographics. Human oversight is typically mandated to keep clinicians responsible. WHO guidelines emphasize ethical principles (safety, transparency, accountability).

Data Protection (Health Privacy): Healthcare AI handles sensitive data, so HIPAA (U.S.), GDPR (EU) apply. Hospitals must sign BAAs with AI vendors, de-identify data, or obtain consent. *Data Protection Impact Assessments* may be required under GDPR.

Liability and Malpractice: Responsibility typically remains with healthcare providers. If AI misdiagnoses, providers or manufacturers could face liability. Product liability laws might evolve for defective AI. EU’s draft AI Liability Directive eases burden of proof for harm claims.

Finance

Fair Lending & Anti-Discrimination: Laws like ECOA (U.S.) bar discrimination in credit decisions. Regulators demand *explainable AI* for credit approval/denial. Black-box models don’t shield lenders from liability. AI must pass *bias audits*.

Transparency in Automated Decisions: Many jurisdictions require reasons for financial decisions. Under GDPR, individuals can request *meaningful information* about the logic of automated credit decisions. Some proposals require **impact assessments** for AI used in finance.

Algorithmic Trading & Market Stability: Regulators (SEC, EU's MiFID II) demand safeguards against disorderly markets. Firms must have kill-switches. Robo-advisors must register and act in clients' best interest. AI-driven trading is subject to ongoing scrutiny.

Fraud Detection & AML: AI is popular for fraud monitoring. Regulators encourage it but insist on *accuracy*, fairness, and data-protection compliance. Model Risk Management guidelines (Fed SR 11-7) apply to AI.

Model Risk Management: Banks must inventory, validate, and monitor *all* models, including AI/ML. Regulators expect rigorous governance structures. Overreliance on untested AI can lead to enforcement actions.

Consumer Protection and Accountability: Agencies (CFPB, FTC, DOJ, etc.) vow to enforce anti-discrimination in AI. In the EU, the AI Act classifies credit scoring as high-risk, requiring strict compliance. Global financial stability bodies (FSB, Basel) watch for systemic risks.

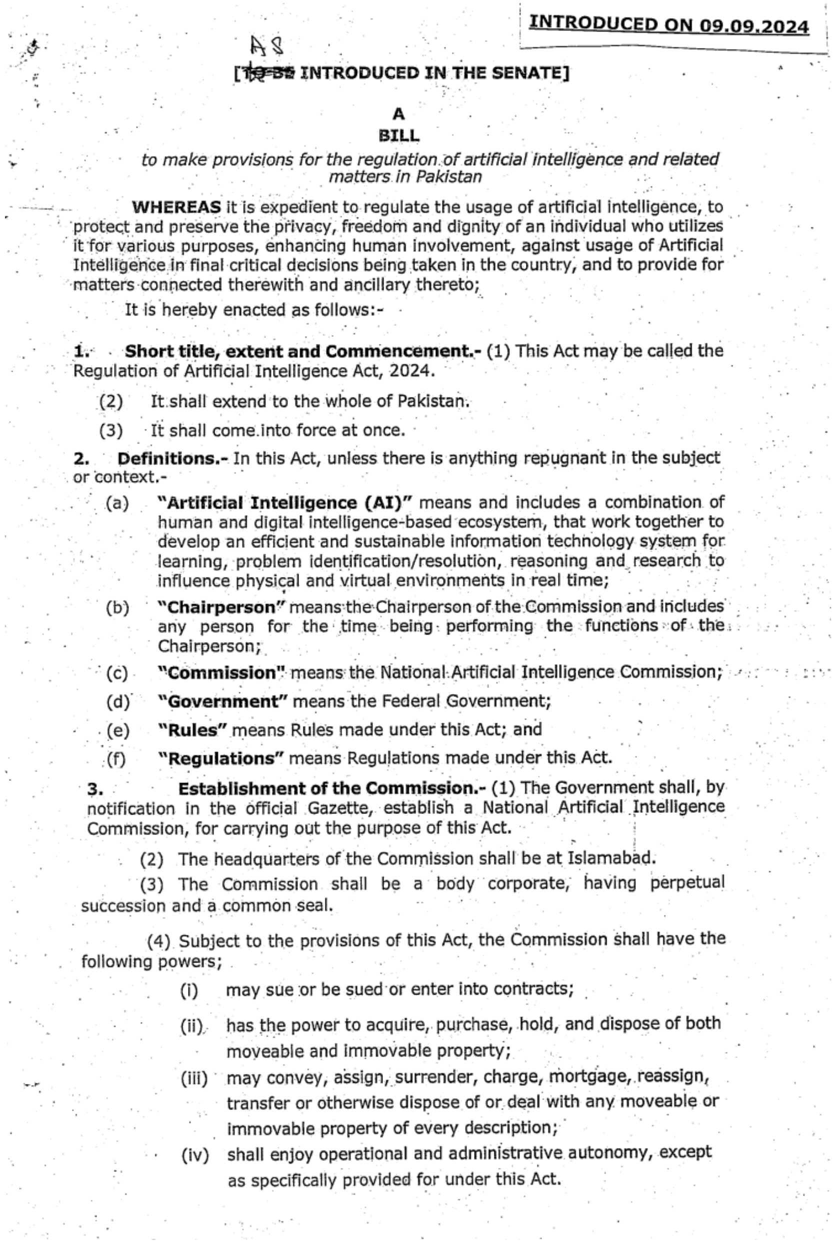
AI & Digital Regulations in Pakistan

National AI Policy: Pakistan is formulating a *Draft National AI Policy* (May 2023), part of “Digital Pakistan.” It aims to *“create a conducive ecosystem for the responsible adoption of AI”*, emphasizing **ethical and responsible use**, R&D, and addressing downsides like job displacement. The final policy, expected by 2025, will likely establish a **National AI Coordination** body, pilot projects, and guidelines for government AI.

Digital Governance and Laws: A **Personal Data Protection Bill (PDPB) 2023** is near enactment, similar to GDPR. It will regulate collection and processing of personal data, requiring consent, user rights (access, correction, deletion), data localization for sensitive data, and possibly *Data Protection Impact Assessments* for high-risk AI. Pakistan’s cybersecurity and electronic transactions laws (PECA 2016) also apply to malicious AI uses. Further updates might come to constrain government surveillance AI.

Sectoral AI Use and Regulation:

- **Banking:** The State Bank oversees AI in credit scoring/fraud detection.
- **Healthcare:** Basic regulation so far; AI diagnostics might need device-like approvals.
- **Government:** AI for smart cities, policing. Calls for *transparency and human rights safeguards*.
- **Education:** Draft AI policy suggests integrating AI tools and curricula.



INTRODUCED ON 09.09.2024

[~~To Be~~ AS INTRODUCED IN THE SENATE]

A

BILL

to make provisions for the regulation of artificial intelligence and related matters in Pakistan

WHEREAS it is expedient to regulate the usage of artificial intelligence, to protect and preserve the privacy, freedom and dignity of an individual who utilizes it for various purposes, enhancing human involvement, against usage of Artificial Intelligence in final critical decisions being taken in the country, and to provide for matters connected therewith and ancillary thereto;

It is hereby enacted as follows:-

1. **Short title, extent and Commencement.-** (1) This Act may be called the Regulation of Artificial Intelligence Act, 2024.

(2) It shall extend to the whole of Pakistan.

(3) It shall come into force at once.

2. **Definitions.-** In this Act, unless there is anything repugnant in the subject or context.-

(a) **"Artificial Intelligence (AI)"** means and includes a combination of human and digital intelligence-based ecosystem, that work together to develop an efficient and sustainable information technology system for learning, problem identification/resolution, reasoning and research to influence physical and virtual environments in real time;

(b) **"Chairperson"** means the Chairperson of the Commission and includes any person for the time being performing the functions of the Chairperson;

(c) **"Commission"** means the National Artificial Intelligence Commission;

(d) **"Government"** means the Federal Government;

(e) **"Rules"** means Rules made under this Act; and

(f) **"Regulations"** means Regulations made under this Act.

3. **Establishment of the Commission.-** (1) The Government shall, by notification in the official Gazette, establish a National Artificial Intelligence Commission, for carrying out the purpose of this Act.

(2) The headquarters of the Commission shall be at Islamabad.

(3) The Commission shall be a body corporate, having perpetual succession and a common seal.

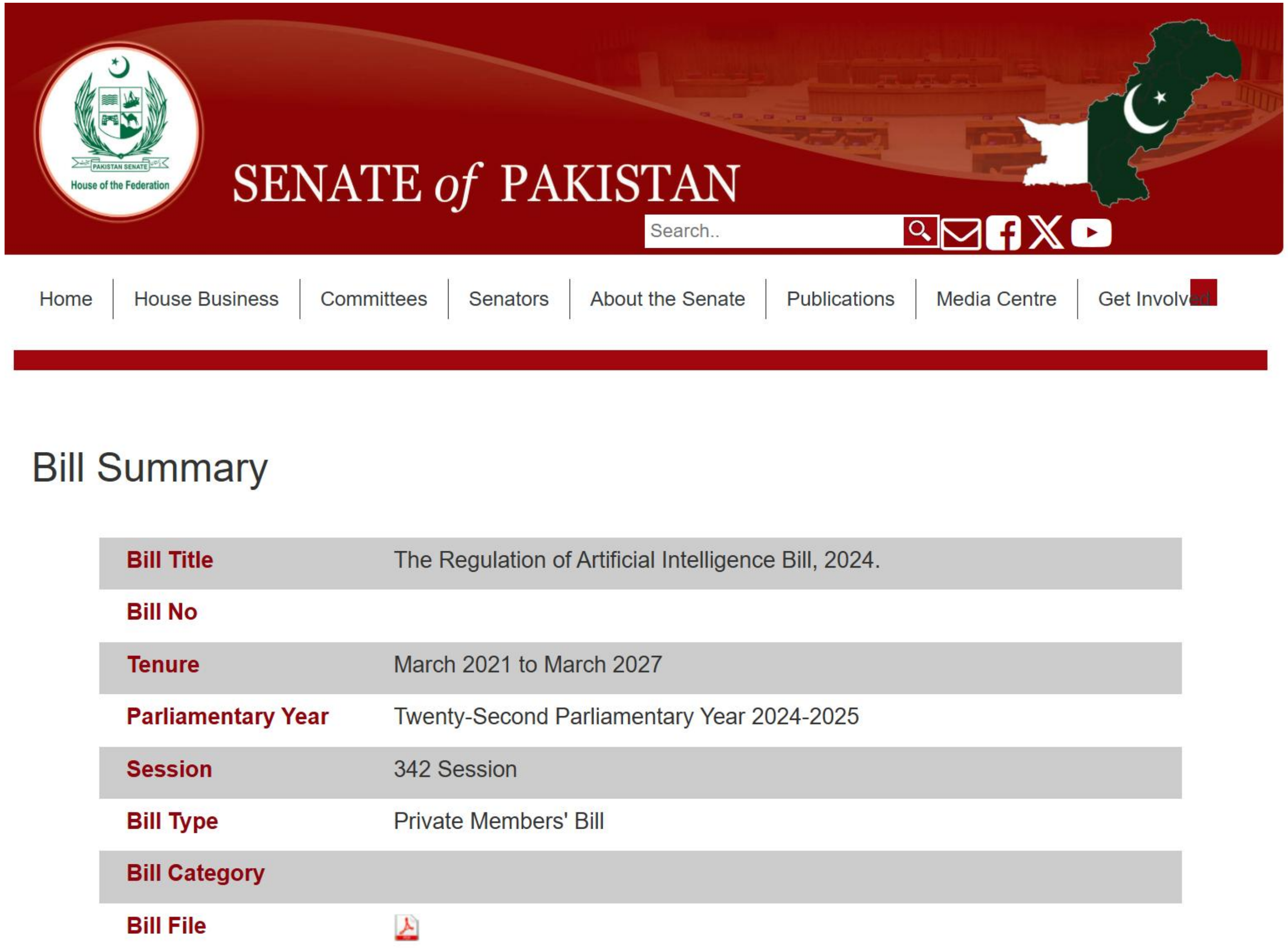
(4) Subject to the provisions of this Act, the Commission shall have the following powers;

(i) may sue or be sued or enter into contracts;

(ii) has the power to acquire, purchase, hold, and dispose of both moveable and immovable property;

(iii) may convey, assign, surrender, charge, mortgage, reassign, transfer or otherwise dispose of or deal with any moveable or immovable property of every description;

(iv) shall enjoy operational and administrative autonomy, except as specifically provided for under this Act.



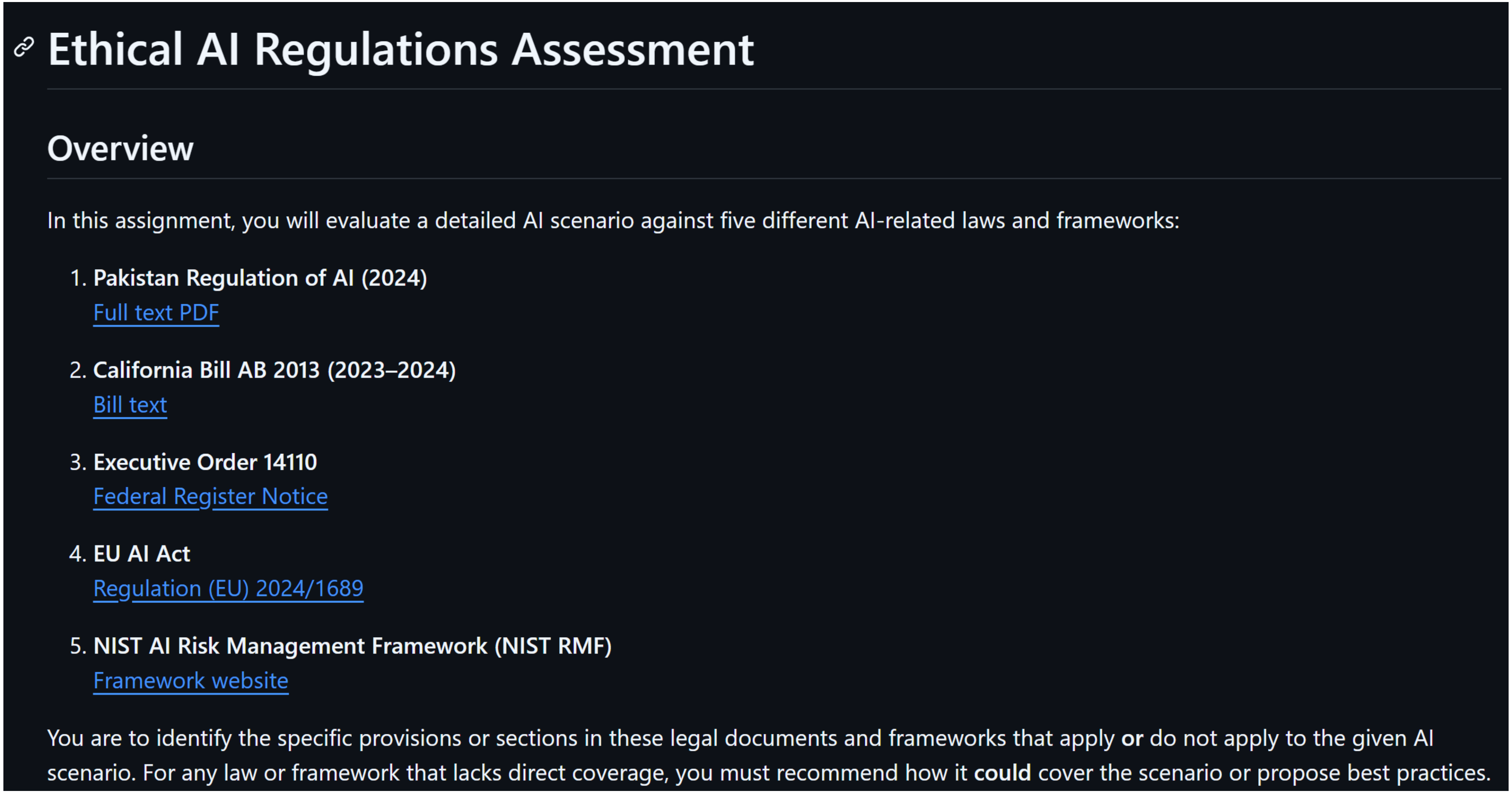
Bill Summary

Bill Title	The Regulation of Artificial Intelligence Bill, 2024.
Bill No	
Tenure	March 2021 to March 2027
Parliamentary Year	Twenty-Second Parliamentary Year 2024-2025
Session	342 Session
Bill Type	Private Members' Bill
Bill Category	
Bill File	📄

International Alignment: Pakistan references **OECD AI Observatory** and UNESCO’s AI Ethics Recommendation. Harmonization with trading partners (e.g. EU) is considered. Implementation challenges remain—capacity-building, establishing a Data Protection Authority, etc. The *Draft National AI Policy 2023* suggests enthusiasm for **AI-driven growth** paired with **ethical guardrails** to build public trust.

Future Trends in AI	Associated Risks	Potential Solutions
Autonomous Decision-Making	Unreliable decisions in high-stakes areas; reduced human oversight	Implement rigorous testing and validation protocols; establish frameworks for human-AI collaboration
Advances Machine: Learning Techniques	"Black box" models leading to transparency issues; difficulty in bias correction	Develop explainable AI models; invest in research for more interpretable algorithms
Increased AI Integration in Daily Life	Privacy concerns; over-reliance affecting human skills	Enforce strict data privacy laws; promote digital literacy and human skill development
AI in Cybersecurity	Sophisticated cyber threats; potential errors in AI-based security systems	Create AI defense systems with adaptive learning capabilities; establish multi-layered security protocols
AI in Job Markets and Employment	Job displacement; skills gap in the workforce	Foster education and training programs for new skills; implement policies supporting workforce transition
Ethical and Governance Challenges	Misuse in surveillance, military; balancing innovation with ethics	Establish comprehensive AI governance frameworks; promote ethical AI development practices
AI and Global Inequalities	Wider gap between AI-advanced and less advanced regions	Promote global cooperation in AI development; invest in AI education and infrastructure in underdeveloped regions
Human-AI Interaction and Societal Impact	Dependency on AI; erosion of human interactions	Encourage the development of AI that complements human abilities; create awareness about the healthy use of AI

Assignment



Ethical AI Regulations Assessment

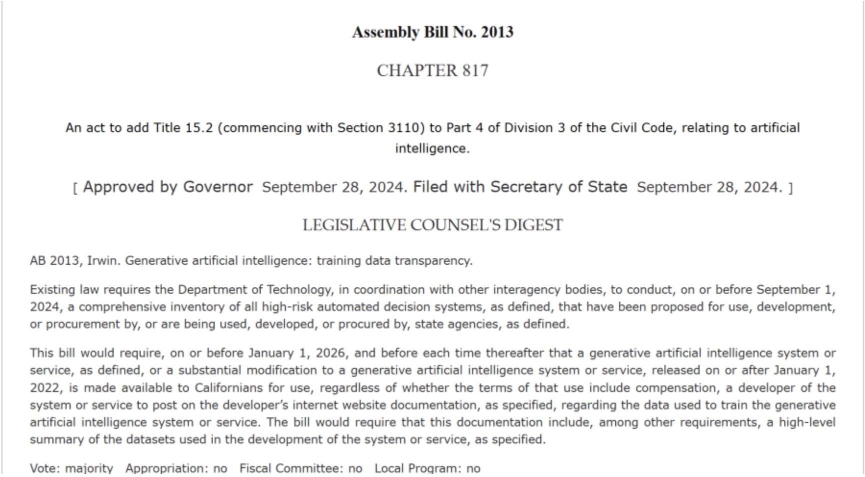
Overview

In this assignment, you will evaluate a detailed AI scenario against five different AI-related laws and frameworks:

1. **Pakistan Regulation of AI (2024)**
   Full text PDF
2. **California Bill AB 2013 (2023–2024)**
   Bill text
3. **Executive Order 14110**
   Federal Register Notice
4. **EU AI Act**
   Regulation (EU) 2024/1689
5. **NIST AI Risk Management Framework (NIST RMF)**
   Framework website

You are to identify the specific provisions or sections in these legal documents and frameworks that apply **or** do not apply to the given AI scenario. For any law or framework that lacks direct coverage, you must recommend how it **could** cover the scenario or propose best practices.

https://github.com/adnanmasood/cs-435-spring-2025/blob/main/regulations-assessment.md



Assembly Bill No. 2013

CHAPTER 817

An act to add Title 15.2 (commencing with Section 3110) to Part 4 of Division 3 of the Civil Code, relating to artificial intelligence.

[Approved by Governor September 28, 2024. Filed with Secretary of State September 28, 2024.]

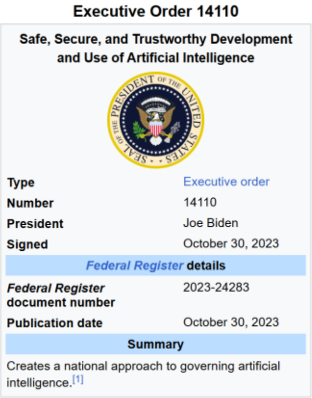
LEGISLATIVE COUNSEL'S DIGEST

AB 2013, Irwin. Generative artificial intelligence: training data transparency.

Existing law requires the Department of Technology, in coordination with other interagency bodies, to conduct, on or before September 1, 2024, a comprehensive inventory of all high-risk automated decision systems, as defined, that have been proposed for use, development, or procurement by, or are being used, developed, or procured by, state agencies, as defined.

This bill would require, on or before January 1, 2026, and before each time thereafter that a generative artificial intelligence system or service, as defined, or a substantial modification to a generative artificial intelligence system or service, released on or after January 1, 2022, is made available to Californians for use, regardless of whether the terms of that use include compensation, a developer of the system or service to post on the developer's internet website documentation, as specified, regarding the data used to train the generative artificial intelligence system or service. The bill would require that this documentation include, among other requirements, a high-level summary of the datasets used in the development of the system or service, as specified.

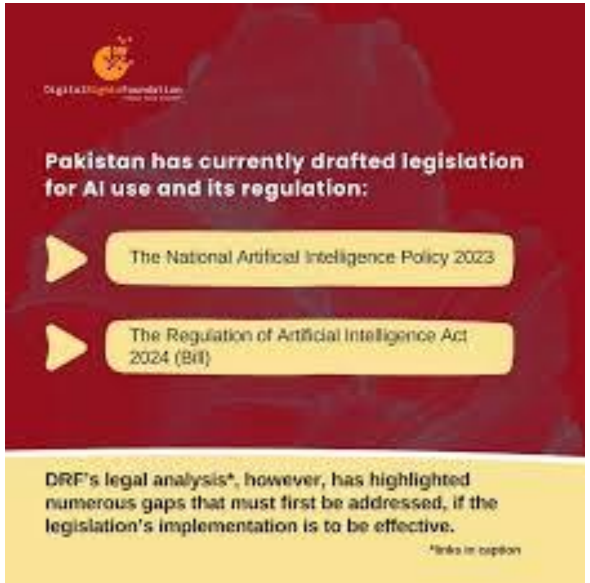
Vote: majority Appropriation: no Fiscal Committee: no Local Program: no



Executive Order 14110

Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

Type	Executive order
Number	14110
President	Joe Biden
Signed	October 30, 2023
Federal Register details	
Federal Register document number	2023-24283
Publication date	October 30, 2023
Summary	
Creates a national approach to governing artificial intelligence.[1]	



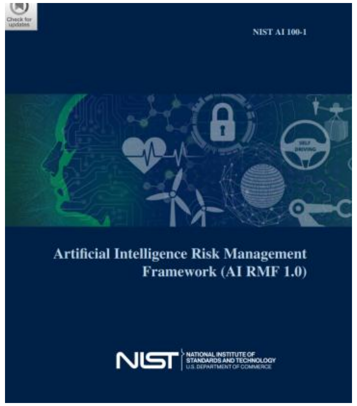
DigitalRightsFoundation

Pakistan has currently drafted legislation for AI use and its regulation:

- The National Artificial Intelligence Policy 2023
- The Regulation of Artificial Intelligence Act 2024 (Bill)

DRF's legal analysis*, however, has highlighted numerous gaps that must first be addressed, if the legislation's implementation is to be effective.

*links in caption



Check for updates

NIST AI 100-1

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

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

https://github.com/adnanmasood/cs-435-spring-2025/blob/main/coding-with-ai-assignment.md

https://github.com/adnanmasood/cs-435-spring-2025/blob/main/coding-with-ai-assignment-problems.md