Lecture 3

Responsible AI & EU Legal Context

2026-04-27

LINKÖPING UNIVERSITY

Responsible AI & EU Legal Context

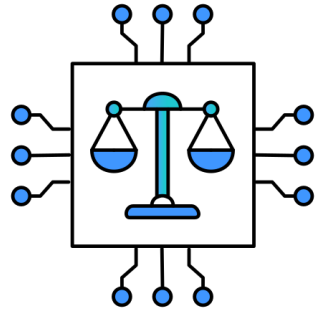
- Foundations of responsible AI
- Ethical dimensions of AI
- EU AI Act – structure & purpose
- Risk classification
- Compliance requirements
- Advanced challenges

Material is available at: **https://www.itn.liu.se/~siaih22/6fitn80.html**

Sources: EU AI Act - Regulation 2024/1689; AIMA 28

Meaning of responsible AI and why it matters in research & development

What is responsible AI?	Responsible AI refers to the design, development, deployment, and use of AI systems in ways that are ethical, transparent, safe, and aligned with societal values
Core objective:	Ensure that AI systems are trustworthy, accountable, and beneficial to individuals and society



Core principles of responsible AI

1. Fairness
2. Transparency
3. Accountability
4. Safety and reliability
5. Privacy and data protection

Why responsible AI matters in research & development

Scientific integrity	• Prevents misuse or misinterpretation of AI-generated results • Supports reproducible and transparent research practices
Risk mitigation	• Identifies potential harms such as bias, misinformation, or unsafe outputs
Legal and regulatory compliance	• Aligns research with emerging regulatory frameworks such as EU AI Act
Trust in scientific systems	• Ensures that AI-supported research remains credible to the scientific community and society

Responsible AI is not only an ethical concern but also a core component of rigorous scientific methodology in modern AI-assisted research

Sources: AIMA 28.3

Difference between ethical guidelines and legal requirements for AI

Two layers of governance in AI

- Ethics → normative guidance on what developers **should** do
- Law → enforceable rules defining what developers **must** do

Ethical guidelines

Normative recommendations
Voluntary adherence
Broad principles
Developed by communities

- Fairness and non-discrimination
- Transparency and explainability
- Human oversight
- Safety and societal benefit

Legal requirements

Binding regulations
Mandatory compliance
Specific obligations
Enforced by governments

- Risk classification of AI systems
- Mandatory documentation and transparency
- Data governance and quality standards
- Human oversight requirements

Why distinction matters in research

Researchers must:
- follow legal requirements to ensure compliance
- apply ethical principles to maintain scientific integrity and societal trust

Ethical AI defines what responsible researchers should strive for, while AI regulation defines the minimum standards that must be met

Societal impact of AI systems and importance of accountability

AI as socio-technical infrastructure

- AI systems increasingly influence economic, social, and political processes
- used in domains such as healthcare, finance, employment, education, and public services
- decisions once made by humans are now partially automated

AI systems can shape opportunities, rights, and life outcomes for individuals and communities

Positive societal impacts

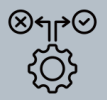
- scientific discovery and innovation
- improved healthcare diagnostics
- efficient public services
- automation of repetitive tasks
- new economic opportunities

Potential societal risks

- algorithmic bias and discrimination
- opacity in automated decision-making
- spread of misinformation
- loss of privacy and surveillance risks
- economic disruption and labor displacement

Why accountability is critical

- AI systems are developed and deployed responsibly
- human actors remain responsible for system outcomes
- harmful consequences can be identified, investigated, and corrected



Without accountability, AI systems risk becoming uncontrolled decision infrastructures

Elements of accountability in AI

- auditability - transparency of algorithms, data and design processes
- minimisation and reporting of negative impacts
- trade-offs: when implementing requirements for trustworthy AI, trade-offs should be addressed in rational and methodological manner within the state of the art.
- redress: when unjust adverse impact occurs, accessible mechanisms should be foreseen that ensure adequate redress.

As AI systems increasingly influence societal outcomes, accountability becomes essential to ensure that technological power remains aligned with democratic values and human responsibility

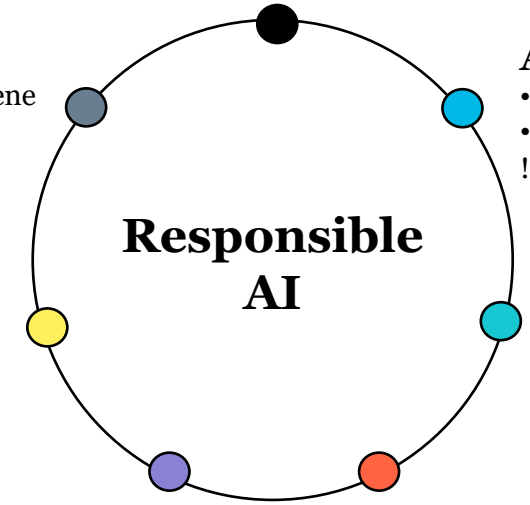
Source: Aima 28.3; European Commission (2019) Ethics Guidelines for Trustworthy AI

Core principles of responsible AI

Fairness
- avoid discrimination and bias
- ensure equitable outcomes across individuals and demographic groups

⚑ different definitions of fairness may conflict

Human agency and oversight
- humans must retain control and be able to intervene
- avoid over-automation in critical decisions



Accountability
- clear responsibility for AI system behavior
- organizations must monitor systems, document decisions, address harms

! AI systems do not bear responsibility — humans and institutions do

Alignment with societal values
AI should respect
- human rights
- democratic principles
- social well-being

Transparency
- make AI systems understandable, traceable, open to scrutiny
- includes:
  - explainability of decisions
  - documentation of models and dat

Safety and robustness
- systems must perform reliably under
  - normal conditions
  - unexpected inputs
- includes:
  - robustness to errors
  - resilience to adversarial attacks

Privacy and data governance
- protect personal data and sensitive information
- ensure:
  - lawful data use
  - data minimization
  - secure handling

Responsible AI is built on principles such as fairness, accountability, and transparency, ensuring that AI systems are not only technically effective but also ethically aligned and socially trustworthy.

Sources: AIMA 28.3

Bias in AI systems: sources and consequences

What is AI bias?

Systematic and unfair distortion in model outputs, which leads to different outcomes for individuals or groups **not justified by the task objective**

AI learns and amplifies patterns already present in society → Bias is therefore socially inherited through data

Main sources of bias

Data collection bias
unrepresentative samples, missing groups

Label bias
biased human annotations or historical decisions

Measurement bias
poor proxies and noisy variables

Algorithmic bias
model objectives and optimization choices

Deployment bias
mismatch between training and real-world use

Real-world consequences

Biased AI systems can produce unfair outcomes in:
- hiring and recruitment
- lending and insurance
- healthcare
- criminal justice
- facial recognition

Amplification effect

The model may not simply reproduce existing bias — it may strengthen and scale it

biased historical hiring decisions → biased hiring model → future exclusion

Research and legal relevance

Bias directly connects to:
- fairness principles
- anti-discrimination law
- EU AI Act high-risk requirements

Bias in AI systems can originate from data, modeling, and deployment processes, and its consequences extend from technical error to real-world discrimination and legal risk

Sources: AIMA 28.3; Cossette-Lefebvre & Maclure, AI's fairness problem: understanding wrongful discrimination in the context of automated decision-making, AI and Ethics (2023)

Fairness and discrimination risks

Why fairness matters

AI systems increasingly influence decisions in:
- hiring, lending, healthcare, education …

⚠ Unequal treatment or unequal outcomes across individuals and groups → This connects directly to anti-discrimination law and Responsible AI principles

Fairness VS **Discrimination**

normative and technical objective | legal and social harm

Common ML fairness criteria
Demographic parity equal positive rates across groups
Equalized odds equal error rates across groups
Individual fairness similar individuals treated similarly

Direct discrimination

Explicit use of protected attributes such as:
- gender
- age
- ethnicity
- disability

Example:
rejecting candidates because of gender

Indirect discrimination

Using variables that act as proxies for protected attributes
Examples:
- postal code
- education institution
- employment gap

often harder to detect

A model may appear "accurate" while still producing **unlawful discriminatory outcomes**
Real-world risk example: Recruitment model trained on historical hiring data may learn: "past hiring preference = future hiring rule"

Fairness risks in AI are not only technical optimization problems but also legal and societal discrimination risks, especially when automated systems affect people's rights and opportunities

Sources: AIMA 28.3

Transparency, explainability, and model complexity

Why this matters

Modern AI systems increasingly affect high-stakes decisions in:
- healthcare, finance, hiring, …

more powerful the model, the harder it often is to understand **why** it made a decision → ethical and legal risks

Transparency

visibility into the system design and process
- what data was used
- how the model was trained
- intended use and limitations

How the system works overall?

Explainability

ability to provide human-understandable reasons for a specific output
Example:
Why was this loan application rejected?

often case-level

Model complexity

refers to how difficult it is to interpret internal decision logic
Examples:
linear regression → low complexity
decision tree → interpretable
deep neural network / transformer → high complexity

black-box problem

Trade-off: **performance** vs **interpretability**

Explainability techniques
- feature importance / SHAP
- attention visualization
- counterfactual explanations

Under EU AI Act, high-risk systems must be sufficiently transparent to allow deployers to interpret outputs appropriately

Full mathematical transparency is often impossible → Responsible AI focuses on meaningful explanation for stakeholders, not total visibility into every parameter

As AI models become more complex, transparency and explainability become essential for trust, legal compliance, and meaningful human oversight—especially in high-risk applications

Sources: AIMA 28.3

Human oversight and human-in-the-loop

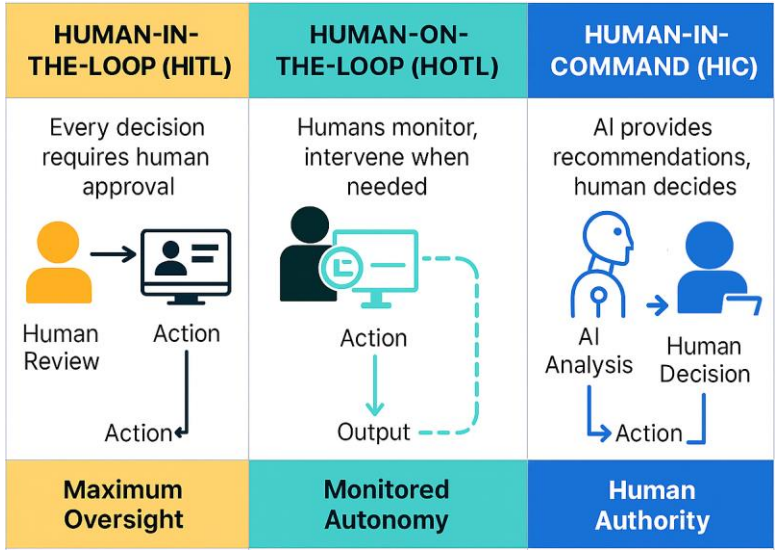
Humans must retain meaningful control over consequential decisions - central to both Responsible AI and the EU AI Act

What is Human-in-the-Loop (HITL)?
- A governance and system design principle where a human can review, intervene, correct, or override the AI output before final action
- Example: AI recommends candidate ranking → recruiter makes final decision

Three oversight models
- Human-in-the-Loop
- Human-on-the-Loop
- Human-in-Command

Why oversight is not just "sign-off"
- common failure is rubber-stamping: human formally approves outputs without genuine review
- oversight must be meaningful, not procedural

Research and development relevance
In research workflows, oversight applies to:
- AI-generated analyses
- literature synthesis
- coding assistants
- autonomous agents
- researchers remain responsible for: interpretation, validation, and final decisions



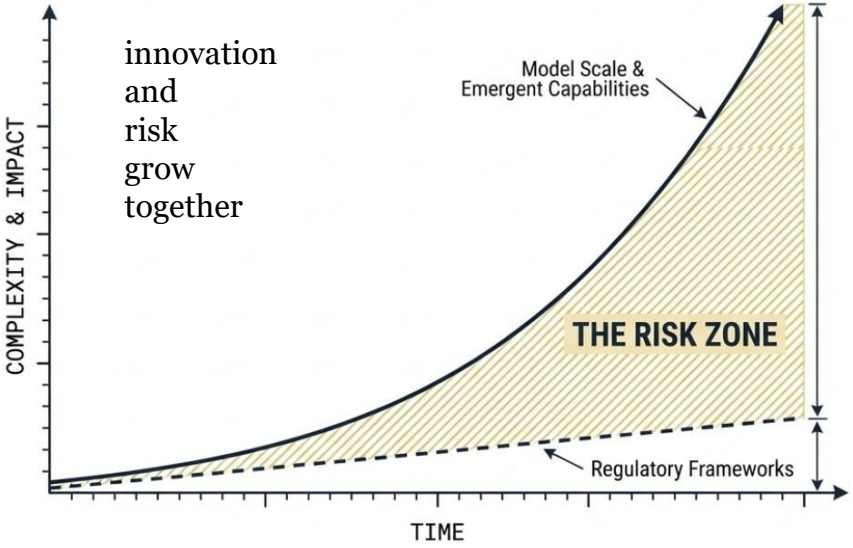
Human oversight ensures that AI augments rather than replaces human judgment, and under the EU AI Act it becomes a legal requirement for high-risk systems—not merely a design preference

From innovation to regulation: why AI needs governance

AI as a dual-use innovation

AI creates major opportunities in:
- healthcare
- scientific research
- industry automation
- public services

At the same time, the same systems can create:
- bias and discrimination
- privacy violations
- safety risks
- misinformation
- accountability gaps



Technology alone is not enough

Technical performance metrics, such as accuracy and F1 score, do not capture societal harm

Example:

A highly accurate hiring model may still be unfair or unlawful

⇒ This creates the need for governance beyond engineering

The governance gap

AI systems often affect:
- rights
- opportunities
- access to services
- public trust

Without governance, key questions remain unanswered:
- Who is responsible for harm?
- Who validates data quality?
- Who monitors drift after deployment?

⇒ This is known as the accountability gap

Governance frameworks emerged because AI decisions can have real-world consequences

From ethics to binding rules
- Early AI governance relied mainly on ethical guidelines, voluntary principles, corporate policies

Governance as lifecycle control: design → training → deployment → monitoring → audit

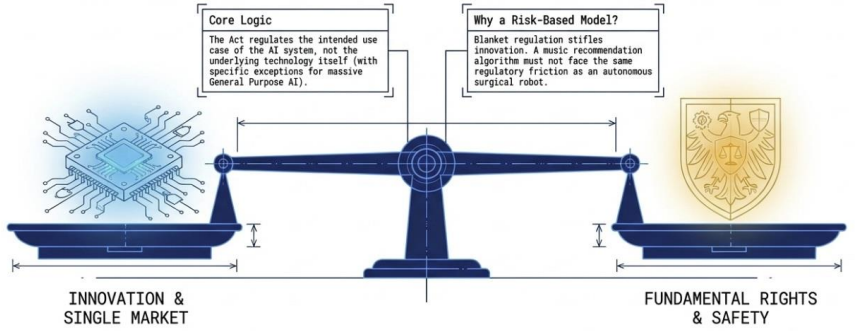
AI needs governance because technical excellence does not automatically ensure fairness, safety, accountability, or legal compliance in real-world use

Why EU chose a risk-based model

The stricter the potential harm, the stricter the legal obligations

Single rule for all AI would fail = regulation should be proportionate to the level of risk

Four levels:
1. minimal risk
2. limited risk
3. high risk
4. unacceptable risk



AI Act focuses not only on market safety but also on privacy, fairness, non-discrimination, human dignity, autonomy

Avoid two extremes:
- innovation without safeguards
- regulation that suppresses research and industry

For researchers, this means: the same model may be regulated differently depending on context of use
Example:
- same LLM in teaching → lower risk
- same LLM in clinical triage → higher risk

The EU chose a risk-based model to ensure that legal obligations are proportionate to the real-world harm an AI system may cause, balancing innovation with the protection of safety and fundamental rights

Structure and goals of the EU AI Act

EU Artificial Intelligence Act

EU AI Act
- Regulation (EU) 2024/1689
- the world's first comprehensive horizontal AI regulation
- applies to providers and deployers placing AI systems on the EU market

Key purpose: create harmonized rules for safe, trustworthy, and human-centric AI across the EU

Core goals of the Act
- protect health and safety
- protect fundamental rights
- support innovation and adoption
- harmonize the EU internal market

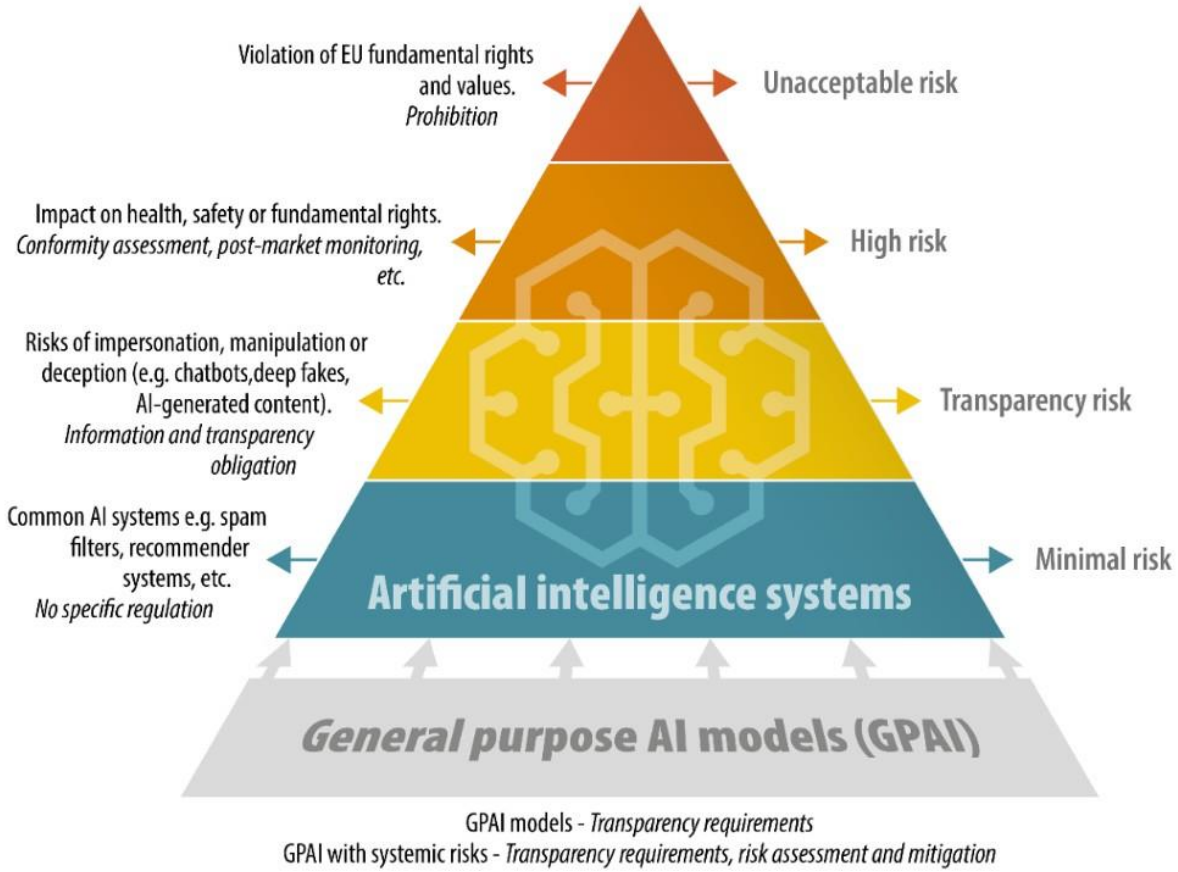
Structural logic of the Act

The Act is organized around a risk-based architecture
- prohibited practices
- high-risk systems
- transparency obligations
- general-purpose AI models
- governance and enforcement

A common misconception: the AI Act is not only restrictive
- regulatory sandboxes
- SME support
- startup innovation measures

The EU AI Act combines a risk-based legal structure with the dual goal of protecting fundamental rights and enabling trustworthy AI innovation across the European market

Risk Categories in the EU AI Act



The EU AI Act organizes AI systems into four risk categories—prohibited, high, limited, and minimal—so that legal obligations are proportionate to the potential harm caused by the system's real-world use

Prohibited and high-risk systems

EU Artificial Intelligence Act

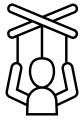
EU AI Act separates AI uses into two legally critical categories:
- prohibited systems → banned outright
- high-risk systems → permitted, but under strict obligations

not all harmful AI is treated the same way

Prohibited systems

These uses are considered incompatible with EU values and fundamental rights

Examples include:
- social scoring of individuals
- manipulative or deceptive systems causing harm
- exploitation of vulnerable persons or groups
- certain forms of real-time remote biometric identification
- some predictive policing and emotion recognition uses



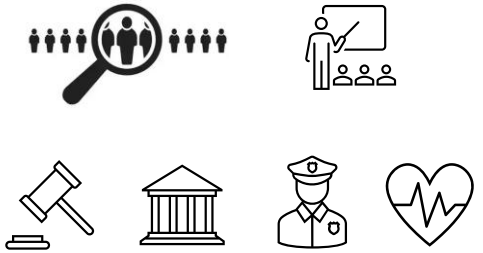
Why they are prohibited
- threaten human dignity, autonomy, privacy, equality, democratic rights

Example: AI that classifies citizens into “trust scores”

High-risk systems

These systems are allowed, but only with strict compliance requirements.

Typical sectors include:
- employment and recruitment
- education and admissions
- healthcare / medical devices
- critical infrastructure
- law enforcement
- migration and border control
- justice and public administration



The same model may be legal or prohibited depending on how and where it is used
- face recognition for phone unlock → generally allowed
- public real-time surveillance → heavily restricted / prohibited

The EU AI Act draws a sharp line between prohibited systems that pose unacceptable risks to rights and society, and high-risk systems that may be used only under strict governance, documentation, and human oversight requirements

Sources: https://artificialintelligenceact.eu/article/5/
https://artificialintelligenceact.eu/article/6/

Providers, deployers, and users

EU Artificial Intelligence Act

Provider
- develops AI system, or
- has it developed under its own name / trademark
- places it on the market
- puts it into service

Deployer
- organization that uses AI system in professional practice
  - hospital using clinical AI
  - HR department using CV screening software
  - university using plagiarism detection AI

User / End user
- under the Article 3, EU AI Act, the legal term is deployer, not "user"
- A user in everyday language often means:
  - employee interacting with the tool
  - researcher using ChatGPT
  - applicant affected by an AI decision

Lab or university may become a deployer when using AI systems in:
- student assessment
- participant screening
- medical research workflows
- automated decision support

'deployer' means a natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity
Article 3: Definitions

Under the EU AI Act, obligations depend on legal role: providers build and place AI systems on the market, deployers use them professionally, and affected users are the individuals subject to the system's outputs.

Sources: https://artificialintelligenceact.eu/article/3/

GDPR and interaction with existing law

EU Artificial Intelligence Act

EU AI Act does not replace existing law. Instead, it works alongside existing legal frameworks, especially:

- GDPR (data protection)
- consumer protection law
- anti-discrimination law
- product safety law
- sector-specific regulation (health, finance, education)

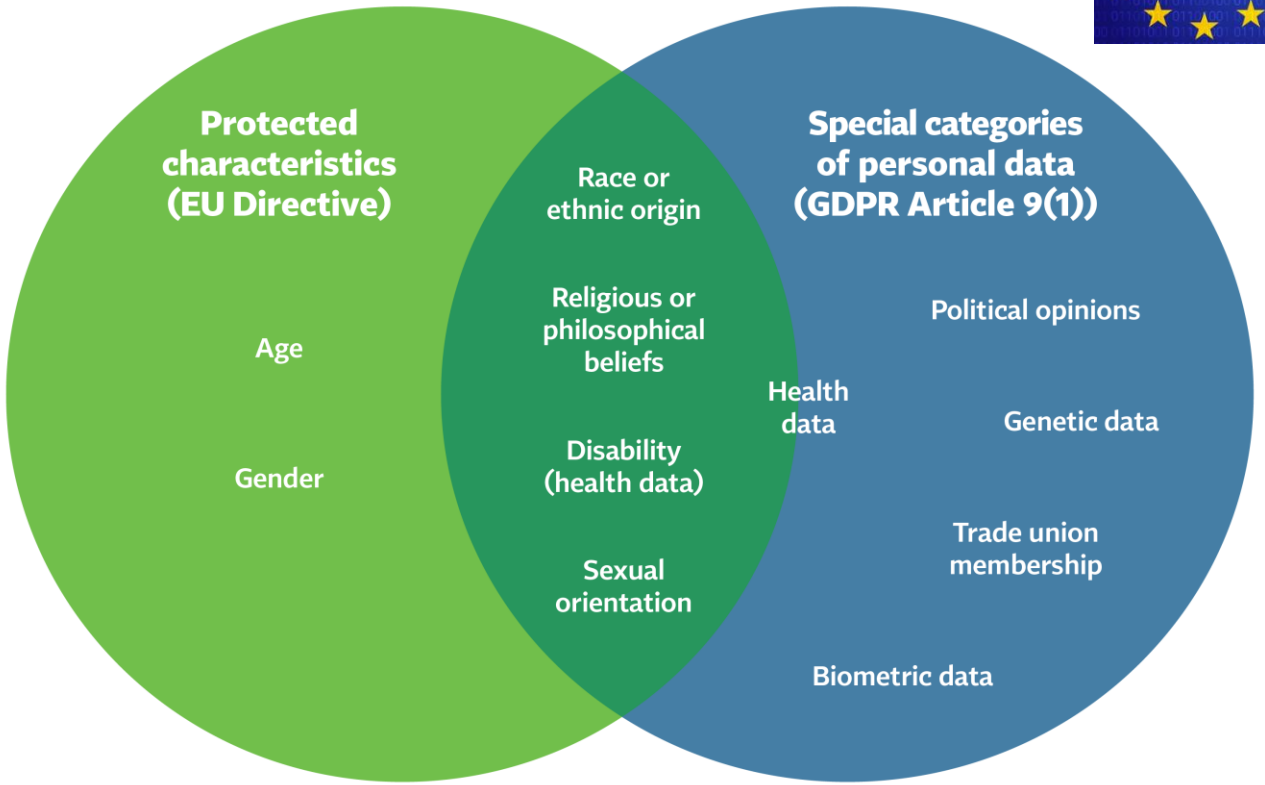
GDPR regulates personal data processing
AI Act regulates AI systems and their risks

Example:
AI recruitment system 
Under GDPR

- lawful basis for CV processing
- purpose limitation
- data minimization
- rights of access / correction
- automated decision safeguards

Under AI Act

- likely high-risk
- technical documentation
- human oversight
- bias monitoring
- post-deployment surveillance



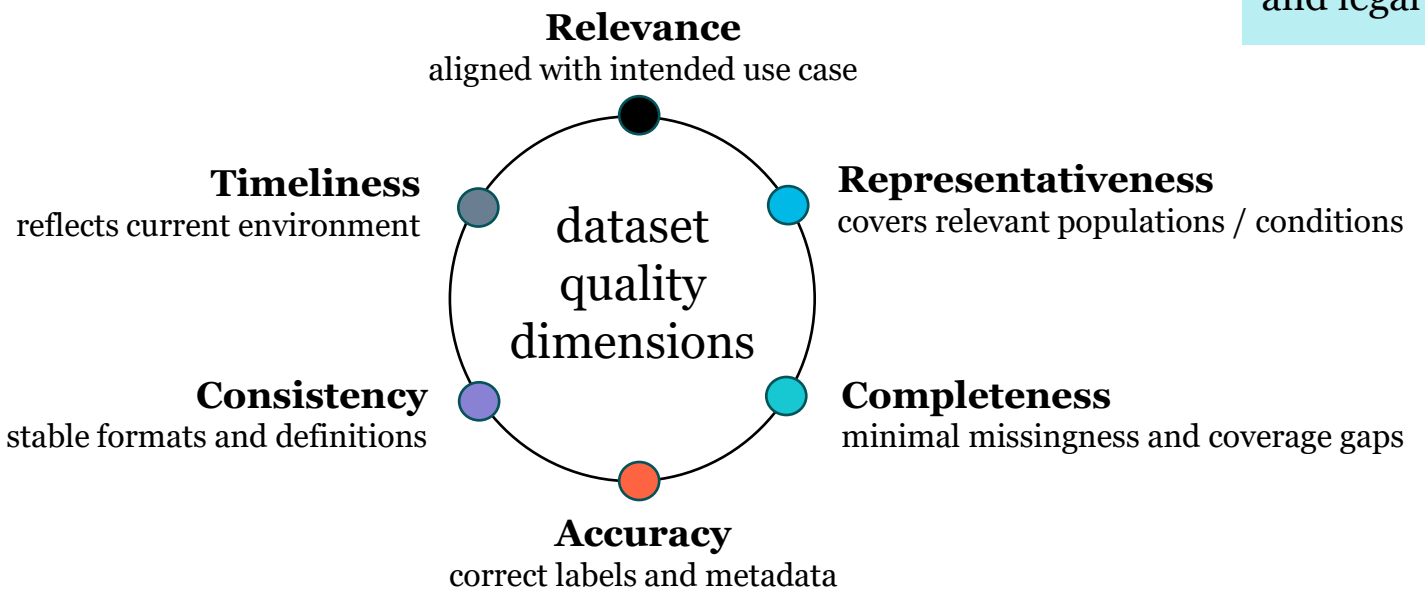
The EU AI Act and GDPR regulate different but overlapping dimensions of AI systems: GDPR focuses on lawful personal data processing, while the AI Act governs AI-specific risks, transparency, and accountability requirements

Data governance and dataset quality

In AI systems, model quality begins with data quality
- garbage in → garbage out
- under the EU AI Act, this is a formal legal requirement for high-risk systems

Data governance means the policies, processes, and controls used to manage data throughout its lifecycle
- who owns the data, where it came from, and how it changed

Responsible AI requires strong data governance: datasets must be representative, documented, traceable, and continuously audited for quality and bias to ensure both scientific validity and legal compliance



For high-risk AI systems, datasets must be:
- relevant
- sufficiently representative
- to the best extent possible, free of errors
- complete

Best practice for research / Recommended documentation tools:
- datasheets for datasets
- data cards
- version control for datasets
- experiment logs

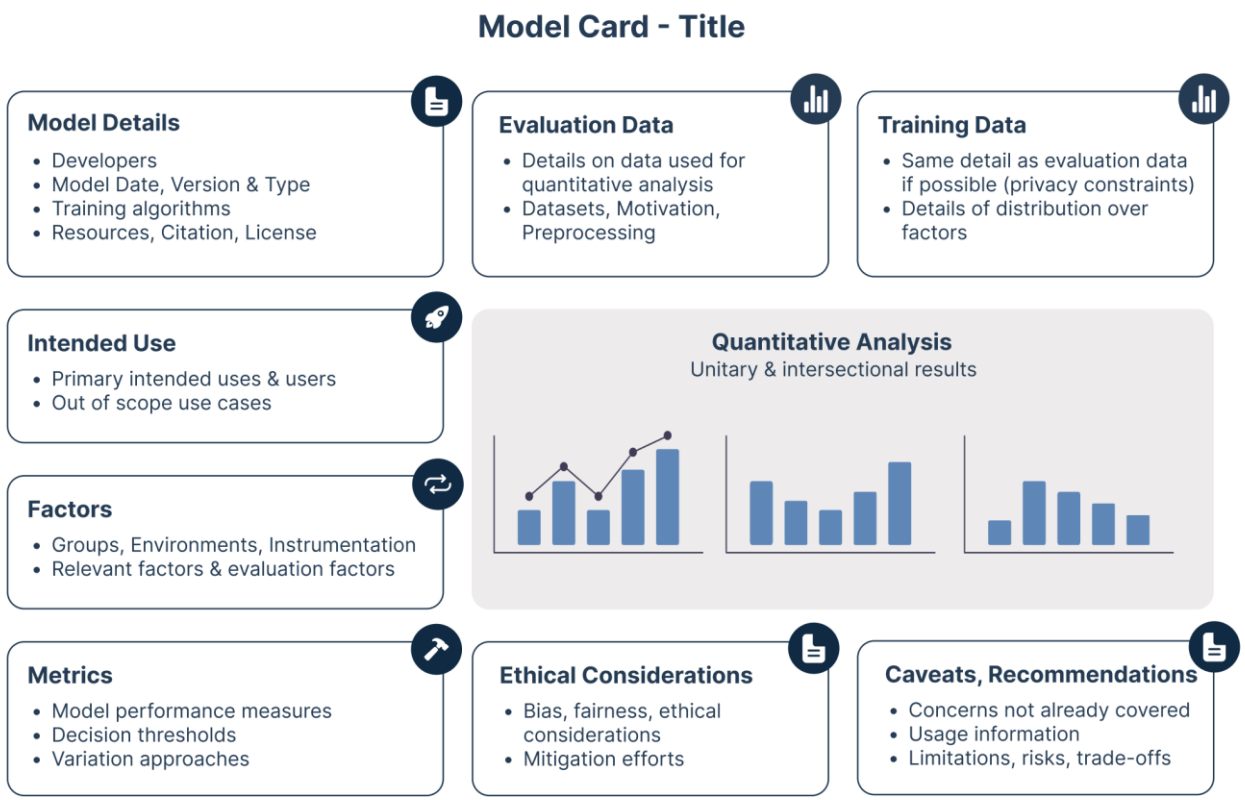
Documentation, model Cards, and auditing

If it is not documented, it cannot be audited, reproduced, or defended

A model card is a concise structured report describing an AI model
- ~ “nutrition label” for AI systems

Auditing means systematic review of model behavior and compliance
- Internal audit - conducted by provider / institution
- External audit - conducted by regulators, independent reviewers, or notified bodies

For PhD research and lab workflows, documentation supports:
- experiment reproducibility
- model comparison
- publication transparency
- ethics review submissions



Documentation, model cards, and auditing transform AI systems from opaque prototypes into transparent, reproducible, and legally accountable research and deployment artifacts

Monitoring and post-deployment surveillance

Deploying an AI system is not the end of responsibility
- AI systems change in real-world environments

The core risk:
- model drift
- data drift (input data distribution changes (patient population changes over time))
- concept drift (relationship between inputs and outputs changes (hiring criteria or market conditions evolve))

Monitoring must include both technical and ethical risks
- accuracy / error rates
- fairness metrics
- false positives / negatives
- subgroup performance
- latency and system failures
- security anomalies
- user complaints / override rates

For high-risk AI systems, providers must establish a post-market monitoring system

Lifecycle governance logic
Deploy → Monitor → Detect drift → Audit → Update → Revalidate

For research AI systems, this includes monitoring:
- reproducibility drift
- dataset shifts
- evaluation degradation
- pipeline failures in deployed experiments

Responsible AI requires continuous post-deployment surveillance to detect drift, bias, and emerging risks, ensuring that systems remain safe, fair, and legally compliant throughout their operational lifetime

Foundation models, GPAI, and autonomous agents

AI governance is moving beyond single-purpose models toward general-purpose and agentic systems

GPAI (General-Purpose AI)
Under the EU AI Act, the legal term is GPAI model
Definition: models capable of performing a wide variety of tasks and being integrated into many downstream systems

New EU AI Act obligations (Since Aug 2025)
Specific obligations for GPAI providers now apply across the EU
- technical documentation
- transparency about training data summaries
- copyright compliance
- evaluation and safety information
- disclosure of capabilities and limitations

The Act introduces a stricter category: GPAI models with systemic risk

Autonomous agents: goal → plan → act → observe → revise

Agents raise new governance risks. Compared with static models, agents introduce risks such as:
- reduced predictability
- tool misuse
- unsafe autonomous actions
- cascading errors
- unclear accountability

Example: autonomous literature review + drafting + emailing pipeline

Foundation models and GPAI systems are now treated as core digital infrastructure under the EU AI Act, while autonomous agents introduce a new frontier of governance challenges around autonomy, accountability, and systemic risk

Future of AI governance in Europe and beyond

Why this matters

AI governance is entering a **new phase: from regulation to implementation and global coordination**

The future challenge is no longer whether AI should be governed, but how governance keeps pace with rapid technological change

Relevant for:
- foundation models
- autonomous agents

Rule-making → Enforcement

Milestones:
- AI Act entered into force: 1 Aug 2024
- GPAI obligations active: 2 Aug 2025
- full application: 2 Aug 2026
- some high-risk product rules: 2027

Future focus on standards, enforcement capacity, AI Office, audits, and guidance

Global convergence vs Regulatory fragmentation

- rights-based
- risk-based
- legally binding

- innovation-oriented
- sectoral / agency-led
- less centralized

- state-centric
- control focus on security and content

New frontier: **agentic** and **autonomous** AI

- multi-agent systems
- tool-using agents
- autonomous decision chains
- self-improving workflows

Key challenge: accountability becomes distributed across multiple actors

From static compliance to adaptive governance
- traditional regulation assumes stable systems
- modern AI requires continuous lifecycle governance

The future of AI governance will depend on adaptive, globally coordinated frameworks that can regulate fast-evolving foundation models and autonomous agents while balancing innovation, rights, and societal trust

Lecture 4
Resources and Evaluation
2026-04-28 (Tuesday) K025 13.15