CAP 4623

TRUSTWORTHY ARTIFICIAL INTELLIGENCE

Lesson 7: Responsible AI

Dr. Fernando Koch
kochf@fau.edu

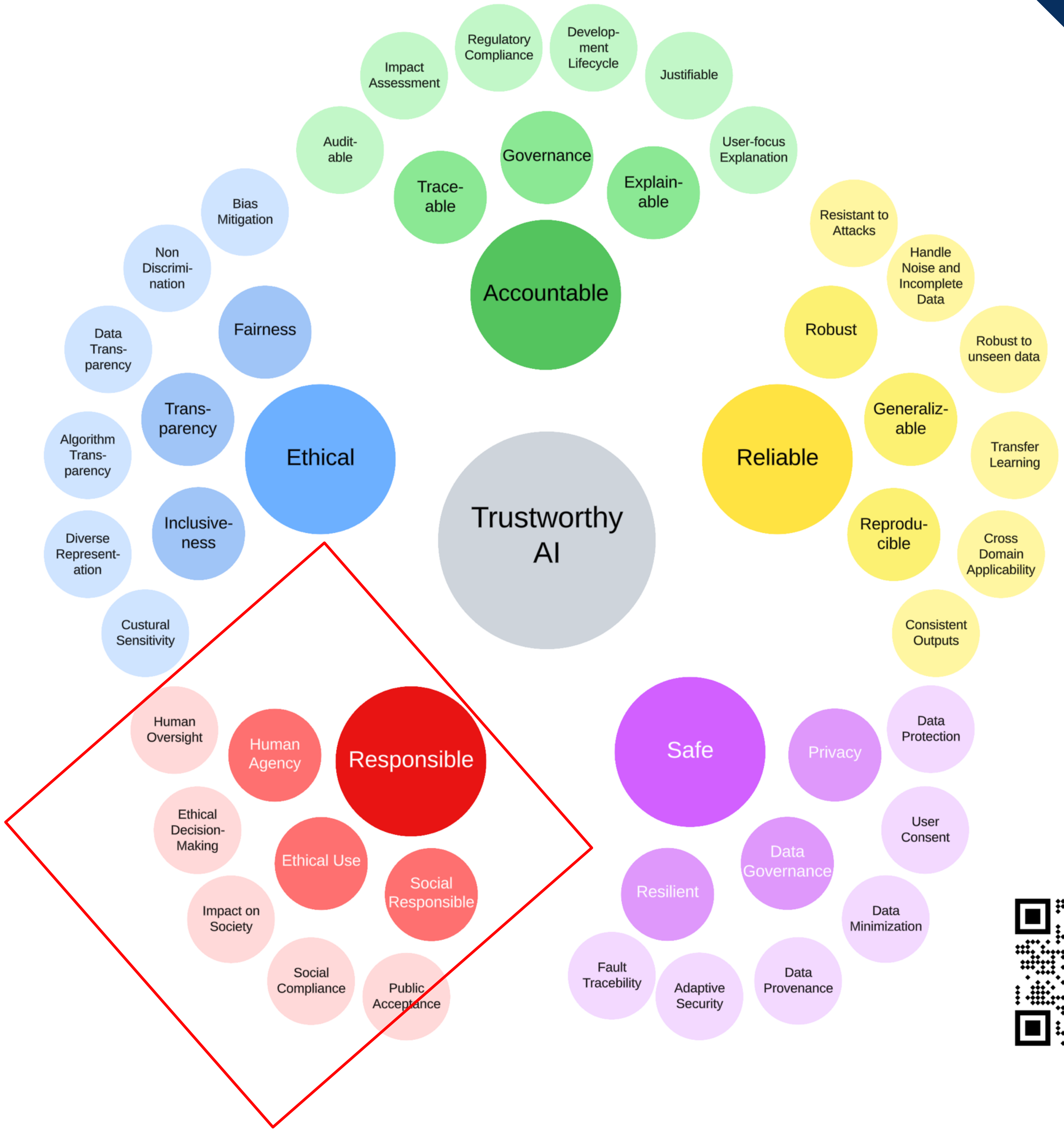
[DALL-E] cute white robot with trustworthy semblance

AGENDA

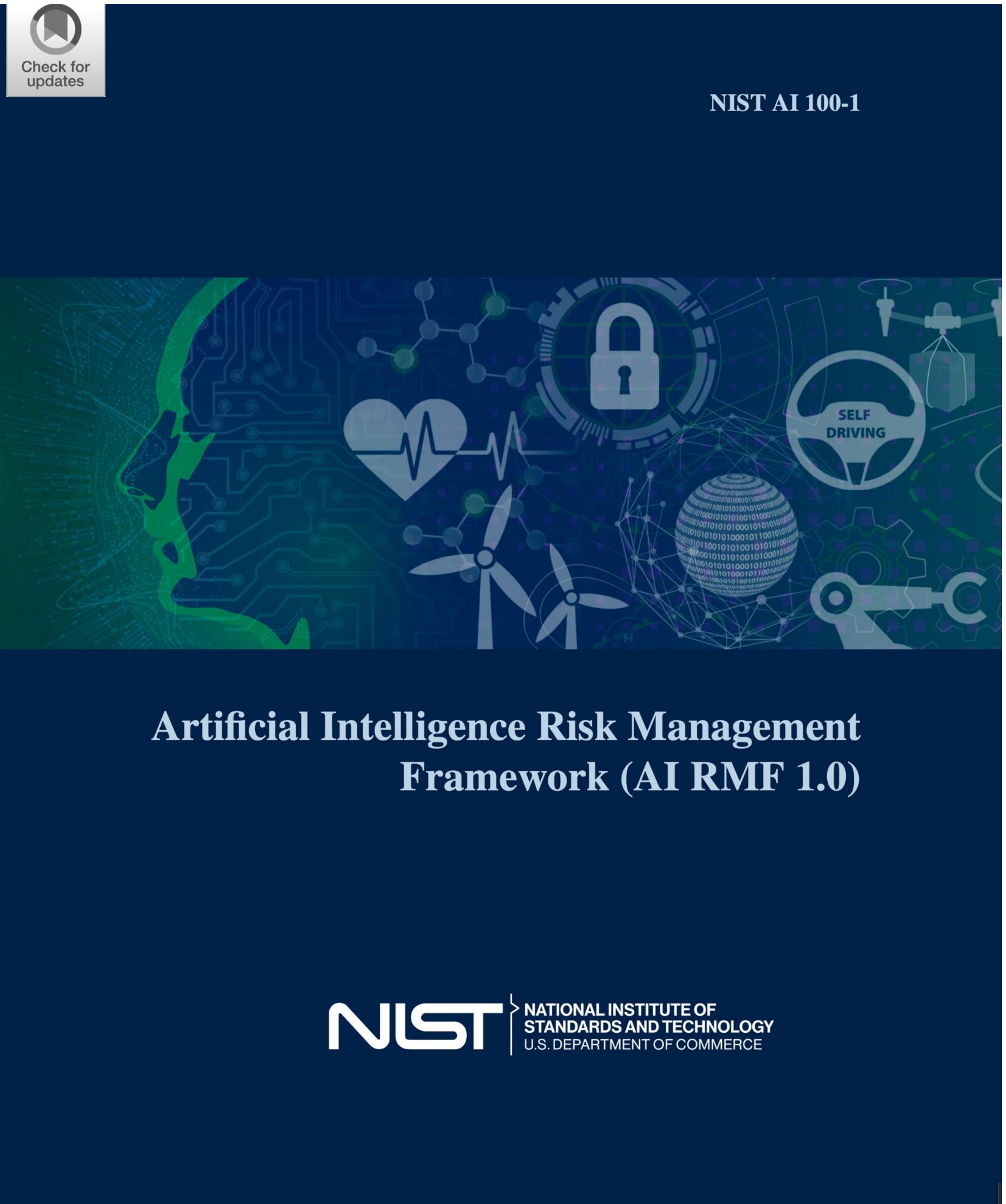
- Introduction to Responsible AI
- Definitions
- Significance and Social Impact
- Case Studies

[DALL-E] cute white robot with trustworthy semblance

Requirements for Trustworthy AI



How can we balance the potential benefits of AI with its long-term societal impacts?



NIST AI 100-1

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

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

Recap

Why do I like this reference?

The Framework is designed to equip organizations and individuals – referred to here as AI actors – with approaches that increase the trustworthiness of AI systems, and to help foster the responsible design, development, deployment, and use of AI systems over time.

https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

What is Responsible AI?

Source: NIST AI Risk Management Framework (NIST AI RMF)

“Responsible AI is the development, deployment, and use of AI systems that align with ethical principles and societal values.”

Why does Responsible AI matter?

Technical debt

Organizations not thinking about responsible AI are acquiring technical debt and increasing the risk of doing business or to cause irreversible harm.

Financial penalties

Emerging regulations for AI systems, such as the EU AI Act and the Canada Data and AI Act, are planning financial penalties of up to 6% of revenue and even criminal punishment for non-compliant systems.

The New York City Law on Automated Employment Decision Tools carries a penalty up to $1,500 per violation, per user, per day.

Definitions

Source: NIST AI Risk Management Framework (NIST AI RMF)

What are ‘Ethical Principles’?

Set of guidelines and norms that help ensure AI systems operate in a way that respects fundamental human rights, fairness, and justice.

What are ‘Societal Values’?

The collective principles and beliefs that guide behavior and decision-making, ensuring the well-being, fairness, and justice of a community.

Common Ethical Principles related to Responsible AI

Fairness:

AI should not reinforce or perpetuate biases, discrimination, or inequalities. It should ensure equitable treatment for all groups and individuals.

Transparency:

AI decisions and processes should be understandable and explainable to stakeholders, allowing for accountability and trust.

Non-maleficence (Do No Harm):

AI systems should avoid causing harm or unnecessary risk to individuals, communities, and the environment.

Autonomy:

AI should respect human agency and allow users to make informed choices without being manipulated or unduly influenced by the AI systems.

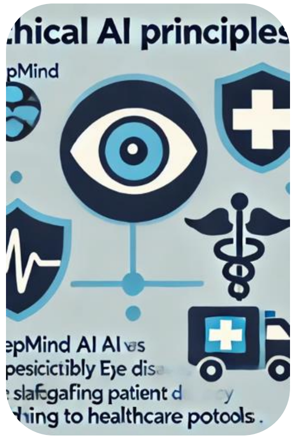
Accountability:

Creators and operators of AI should be answerable for any harms caused by the technology.

Source: NIST AI Risk Management Framework (NIST AI RMF)

Ethical Principles

Examples



AI in Healthcare:
DeepMind’s partnership with the NHS, where AI systems were used responsibly for detecting eye diseases while safeguarding patient privacy and adhering to strict healthcare protocols.

AI in Finance:
AI systems used in financial decisions (like loan approvals) are increasingly being required to provide clear explanations for their decisions, ensuring that institutions remain accountable

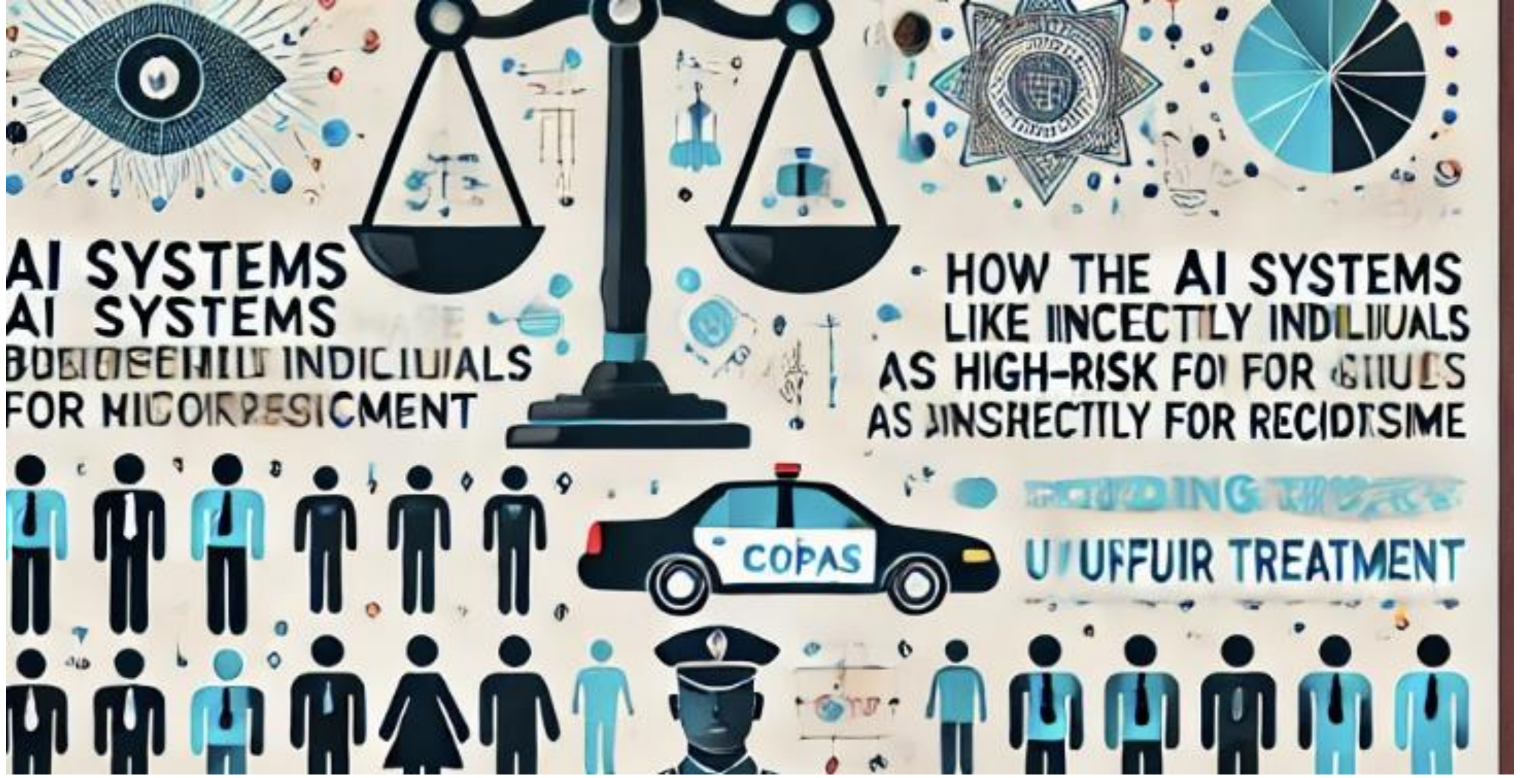
Counter-Examples

Faulty Self-Driving Cars:
An autonomous Uber car struck and killed a pedestrian in 2018, highlighting flaws in the system's ability to correctly identify and avoid hazards, ultimately causing harm.

AI Influence on Personal Choice:
AI systems used in personalized marketing and advertising influencing users' purchasing decisions by targeting them based on private information.

Q71
Predictive Policing

Is this Responsible AI?



AI systems used in law enforcement, like the COMPAS system, which flagged individuals as high-risk for lawlessness

A. Yes, this is Responsible AI.

B. Fairness.

C. Transparency.

D. Non-maleficence.

E. Autonomy.

F. Accountability.

Q72
Facial Recognition for Airport Security

Is this Responsible AI?

AI-powered facial recognition systems are widely used in airports, public spaces, and events to identify potential threats and enhance security by quickly matching faces to known databases of suspects.

A. Borderline, this could be Responsible AI.
B. Fairness.
C. Transparency.
D. Non-maleficence.
E. Autonomy.
F. Accountability.

Q73
Predictive Analytics for Policing

Is this Responsible AI?

AI models help law enforcement predict where crimes are likely to occur.

A. Yes, this is Responsible AI.
B. Fairness.
C. Transparency.
D. Non-maleficence.
E. Autonomy.
F. Accountability.

Common Societal Values related to Responsible AI

Justice:

Promoting fairness in access to opportunities, resources, and benefits from AI while preventing harm or inequality.

Inclusiveness:

Ensuring that AI benefits a broad and diverse population and doesn’t exclude or marginalize any group.

Safety and Security:

AI should contribute to societal well-being by enhancing safety and protecting against harm, including physical, financial, and psychological harm.

Environmental Sustainability:

AI systems should operate in a manner that promotes environmental stewardship and sustainability, helping mitigate ecological impact.

Human Dignity:

AI should respect the intrinsic worth of every individual, ensuring that technology enhances human welfare rather than degrading human rights and dignity.

Source: NIST AI Risk Management Framework
(NIST AI RMF)

Societal Values

Examples

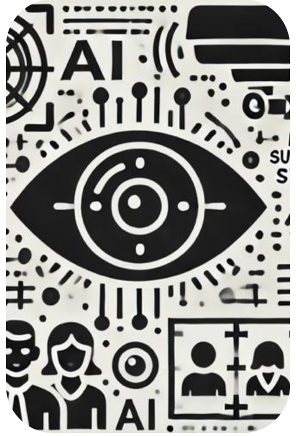
AI in Customer Services

AI chatbots and virtual assistants provide fast and efficient service to users, reducing the need for human involvement in repetitive tasks

AI in Agriculture

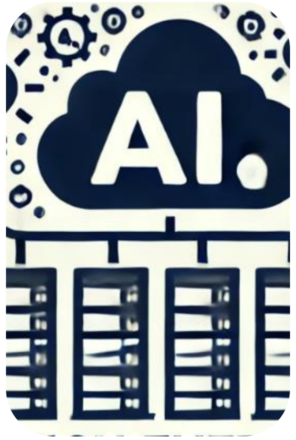
AI is used to optimize crop yields while minimizing resource usage like water and fertilizers.

Counter-Examples



Mass surveillance systems.

AI for mass facial recognition, invading privacy and creating a sense of constant monitoring, leading to a loss of individual freedom and rights.



High Energy Consumption in AI Training

Large-scale AI models (e.g., OpenAI's GPT-3) require massive amounts of computational power, consuming significant energy and contributing to carbon emissions.

Q74
Voice Assistant

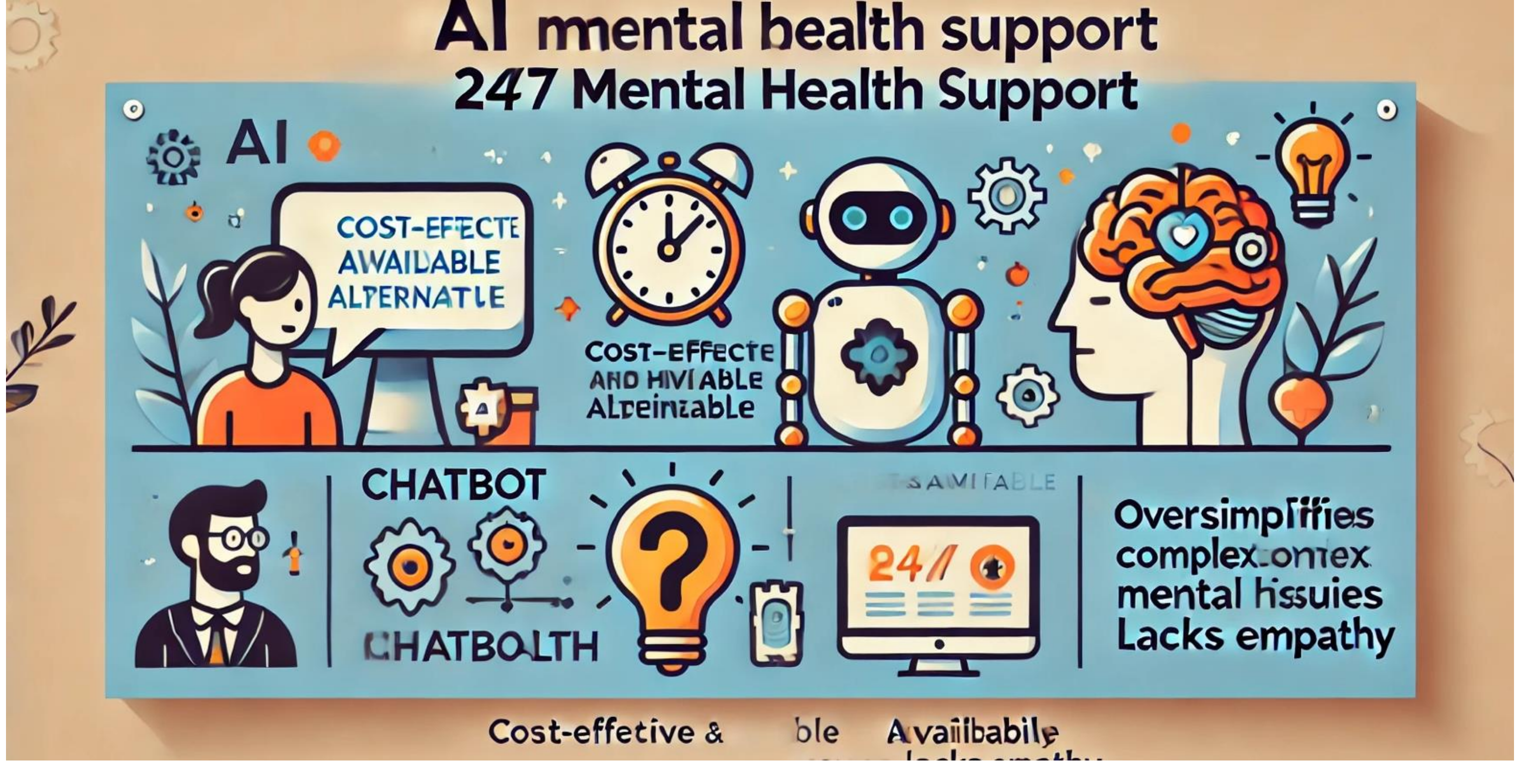
Is this Responsible AI?

Voice recognition systems like Siri and Alexa perform well for users with American accents but struggle with non-native English speakers or those with regional dialects.

A. Borderline; this could be Responsible AI.
B. Justice
C. Inclusiveness.
D. Safety and Security.
E. Environment and Sustainability.
F. Human Dignity.

Q75
AI in Mental Health Support

Is this Responsible AI?

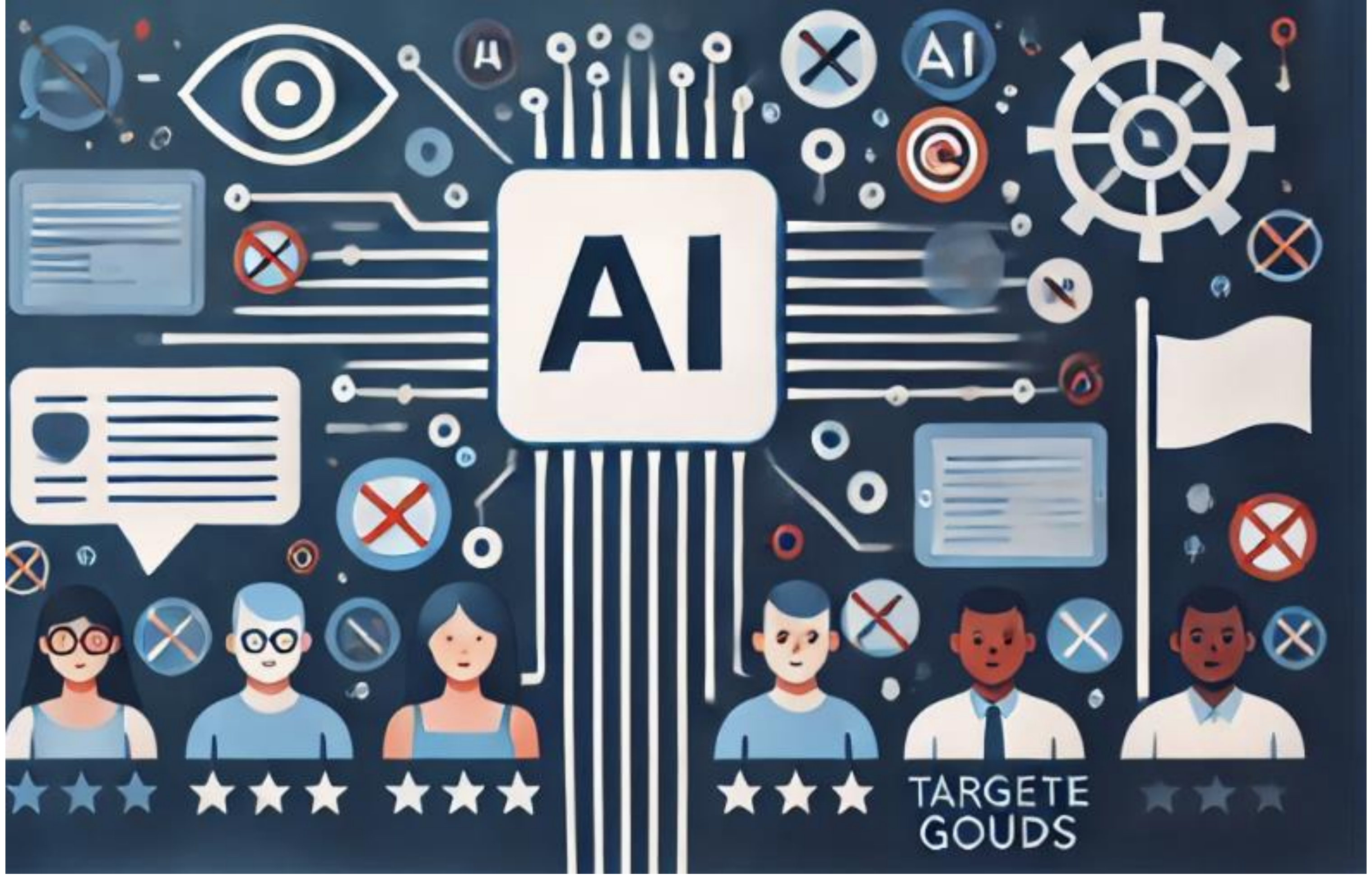


AI chatbots provide 24/7 mental health support, offering users a cost-effective and readily available alternative to human therapists. However, some argue that this technology oversimplifies complex mental health issues and lacks empathy.

A. Borderline; this could be Responsible AI.
B. Justice
C. Inclusiveness.
D. Safety and Security.
E. Environment and Sustainability.
F. Human Dignity.

Q76
AI in Content Moderation

Is this Responsible AI?



AI is used to moderate content on social media platforms, but it disproportionately flags and removes posts from ’target groups’.

A. Borderline; this could be Responsible AI.
B. Justice
C. Inclusiveness.
D. Safety and Security.
E. Environment and Sustainability.
F. Human Dignity.

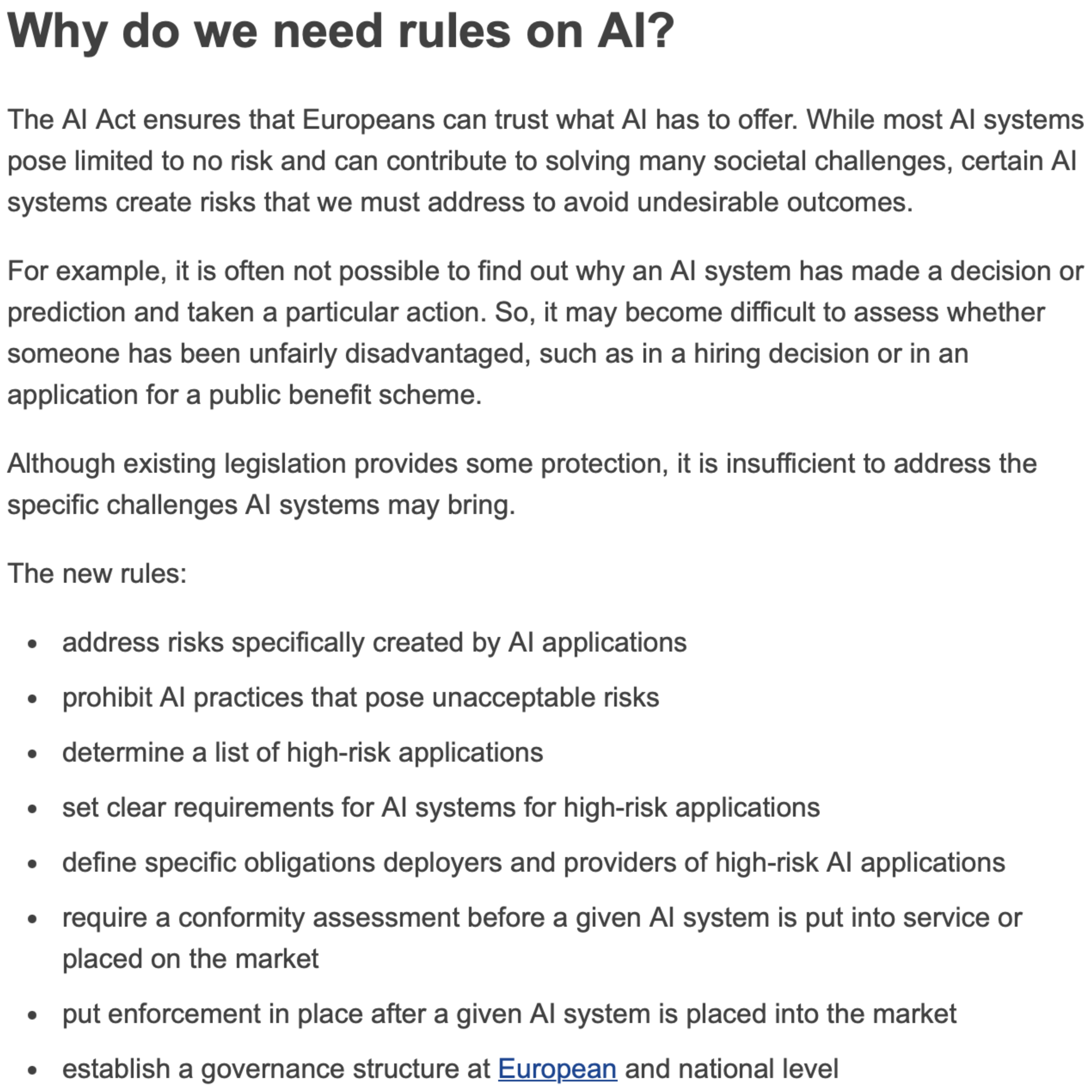
EU AI Act

https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

- Published in the Official Journal of the EU on July 12, 2024.
- Effective from August 1, 2024
- Comprehensive oversight of AI technologies.
- Introduces a phased enforcement schedule, making it crucial for global enterprises, particularly those operating in or with the EU, to understand and comply with its provisions
- It aims to ensure the safe, transparent, and ethical deployment of AI, with far-reaching implications beyond the European Union.

EU AI Act

https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai



Why do we need rules on AI?

The AI Act ensures that Europeans can trust what AI has to offer. While most AI systems pose limited to no risk and can contribute to solving many societal challenges, certain AI systems create risks that we must address to avoid undesirable outcomes.

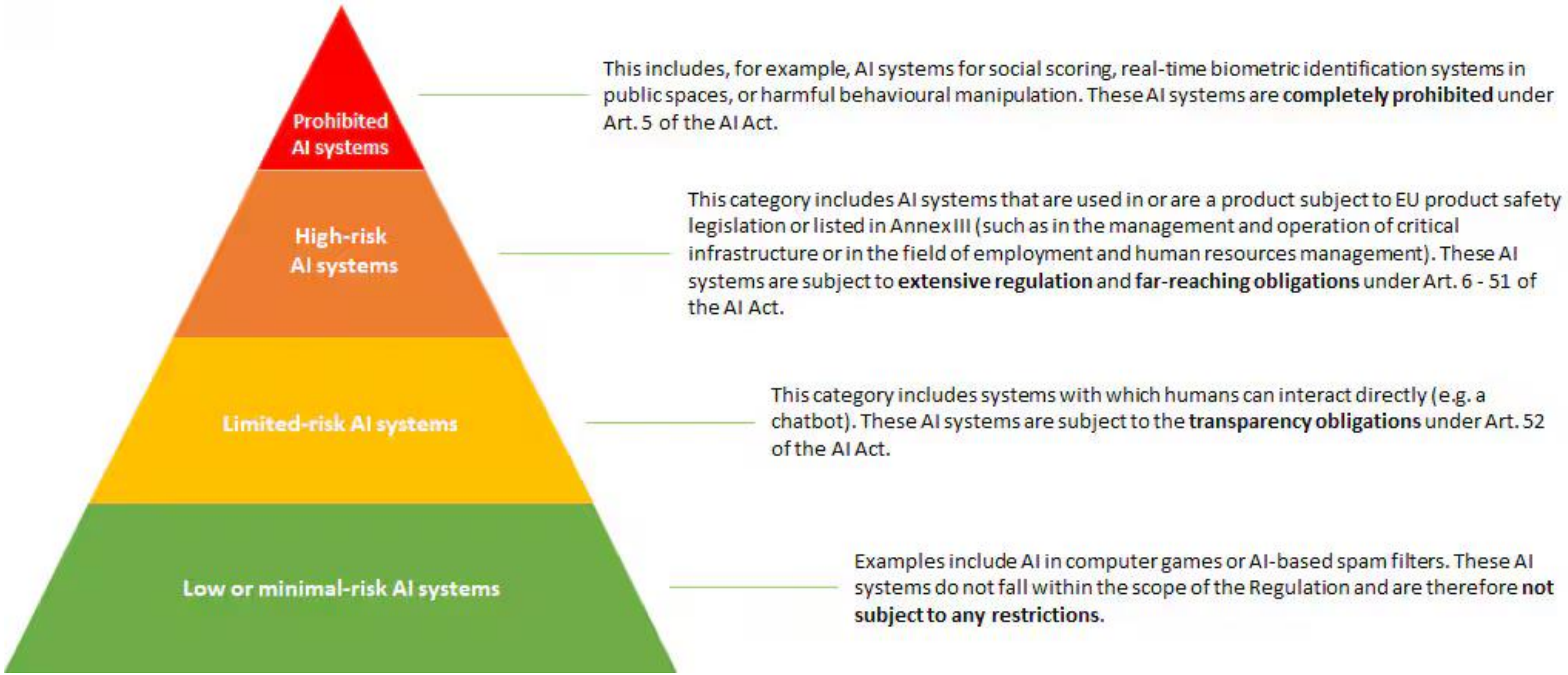
For example, it is often not possible to find out why an AI system has made a decision or prediction and taken a particular action. So, it may become difficult to assess whether someone has been unfairly disadvantaged, such as in a hiring decision or in an application for a public benefit scheme.

Although existing legislation provides some protection, it is insufficient to address the specific challenges AI systems may bring.

The new rules:

- address risks specifically created by AI applications
- prohibit AI practices that pose unacceptable risks
- determine a list of high-risk applications
- set clear requirements for AI systems for high-risk applications
- define specific obligations deployers and providers of high-risk AI applications
- require a conformity assessment before a given AI system is put into service or placed on the market
- put enforcement in place after a given AI system is placed into the market
- establish a governance structure at European and national level

EU AI Act – Levels of Risk



This includes, for example, AI systems for social scoring, real-time biometric identification systems in public spaces, or harmful behavioural manipulation. These AI systems are **completely prohibited** under Art. 5 of the AI Act.

This category includes AI systems that are used in or are a product subject to EU product safety legislation or listed in Annex III (such as in the management and operation of critical infrastructure or in the field of employment and human resources management). These AI systems are subject to **extensive regulation** and **far-reaching obligations** under Art. 6 - 51 of the AI Act.

This category includes systems with which humans can interact directly (e.g. a chatbot). These AI systems are subject to the **transparency obligations** under Art. 52 of the AI Act.

Examples include AI in computer games or AI-based spam filters. These AI systems do not fall within the scope of the Regulation and are therefore **not subject to any restrictions.**

EU AI Level of Risk

Prohibited AI

Social Scoring System

AI that scores citizens based on their behavior and compliance with laws, potentially leading to social exclusion and discrimination.

Real-time Biometric Surveillance

AI-powered facial recognition systems used for mass surveillance in public places, which pose serious risks to privacy and individual freedoms.

Low-Risk AI

AI in Video Games

AI used to control non-playable characters (NPCs) or enhance gaming experiences.

AI-based Spam Filters

Email filtering systems that sort and remove unwanted messages.

EU AI Level of Risk

High-Risk AI

AI in Critical Infrastructure

AI systems used to manage critical infrastructure such as power grids or water supply, where a failure could endanger public safety.

AI in Healthcare

AI systems used in medical diagnosis or treatment recommendations which could have life-threatening consequences if they malfunction or are biased.

Limited-Risk AI

AI Chatbots

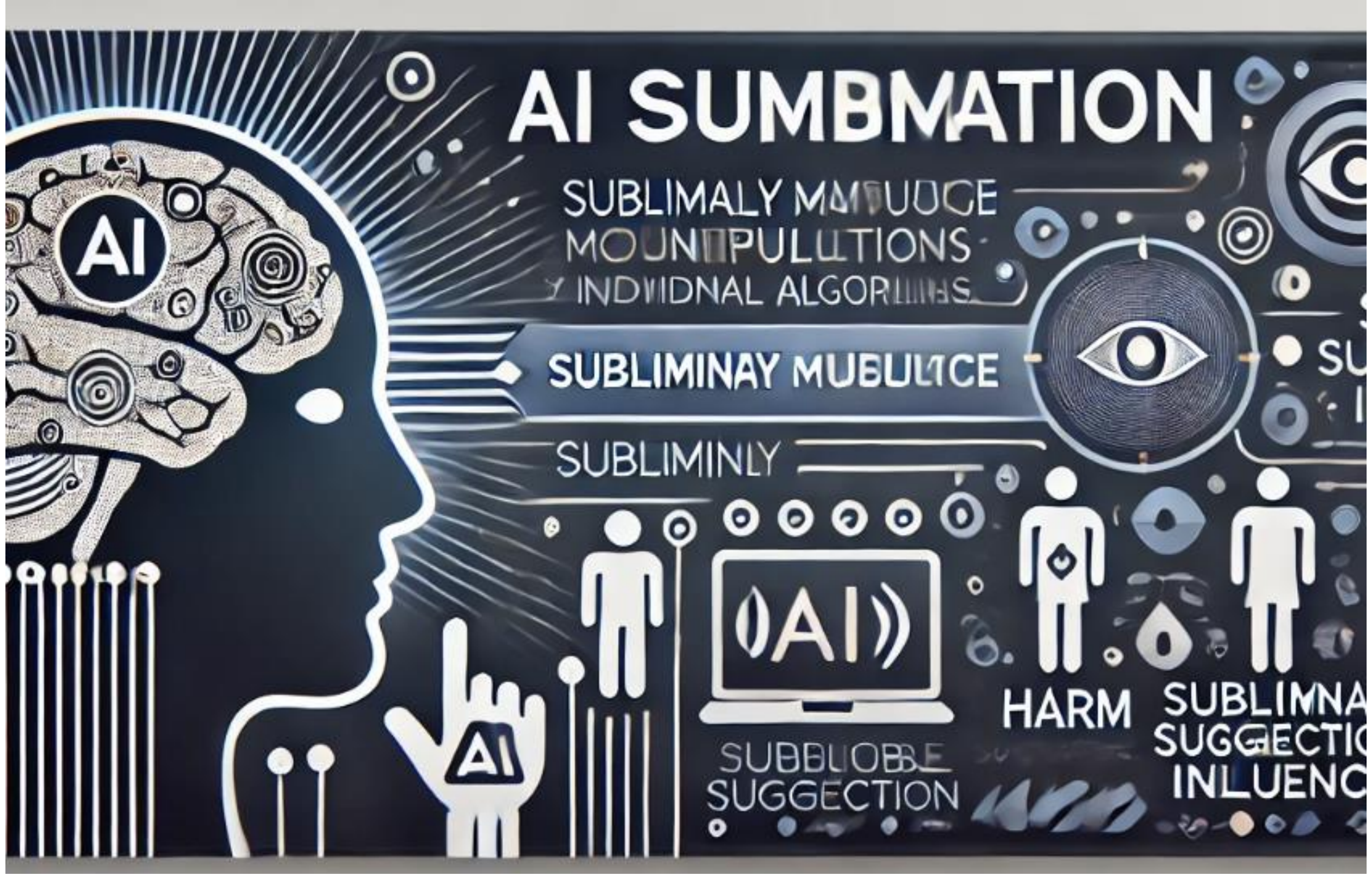
Customer service chatbots that help answer questions or resolve issues. These AI systems are low-risk but require transparency.

AI in Marketing

AI systems used to personalize advertisements or recommend products based on browsing history, where user consent and transparency are critical.

Q77
AI for Behavior Manipulation

What risk level should we assign this AI?



AI systems that subliminally manipulate individuals' behavior in ways that may lead to harm or questionable decisions.

A. Prohibited AI
B. High-Risk AI: Extensive regulation; no-human interaction
C. Limited-Risk AI: transparency obligation; human interaction
D. Low-Risk AI: not subject to restrictions

Q78
AI in Employment Decisions

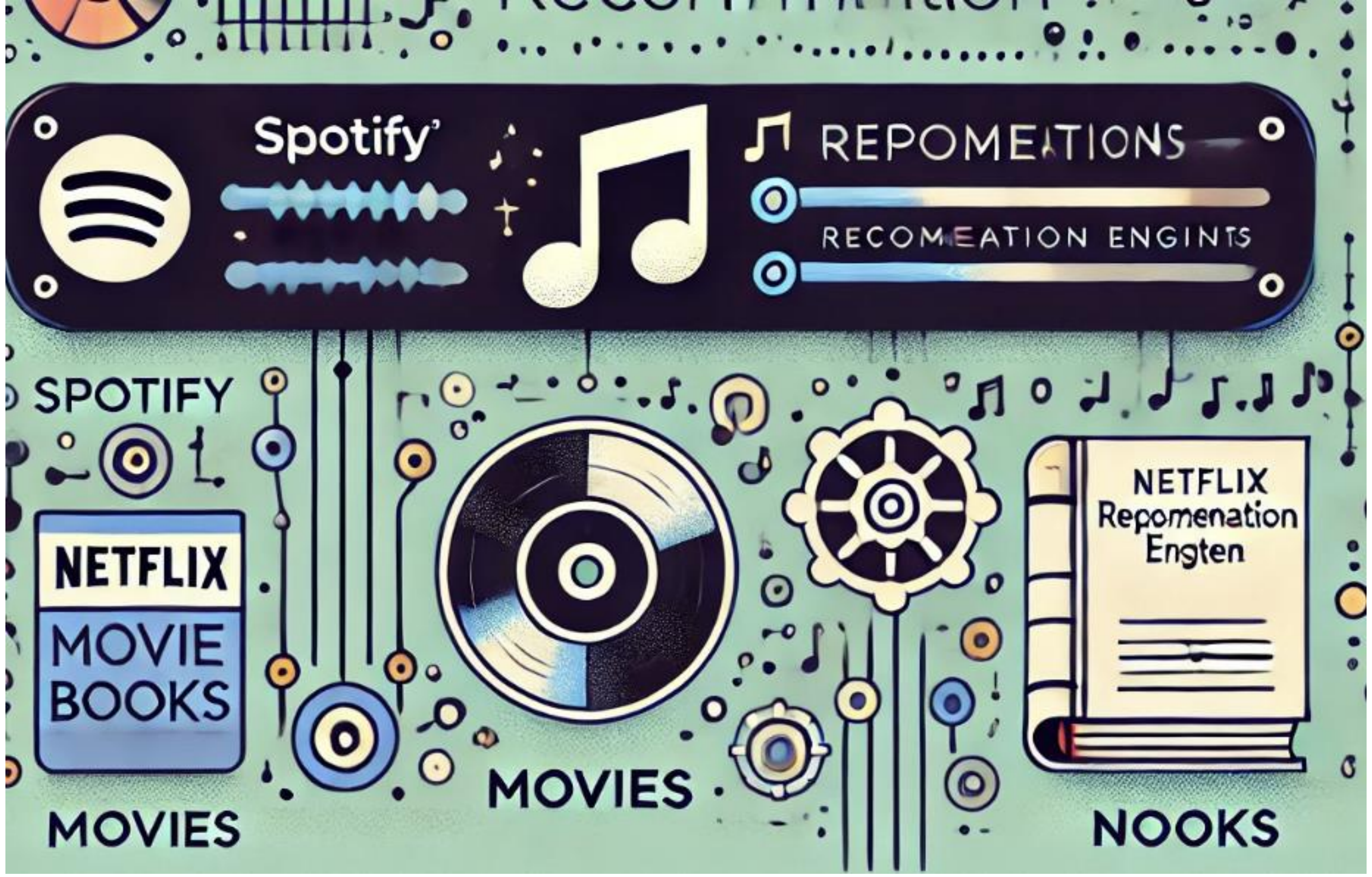
What risk level should we assign this AI?

Recruitment tools that automate job candidate evaluations, which could perpetuate biases or unfair treatment. (e.g., Amazon’s AI recruitment tool that was scrapped due to gender bias).

A. Prohibited AI
B. High-Risk AI: Extensive regulation; no-human interaction
C. Limited-Risk AI: transparency obligation; human interaction
D. Low-Risk AI: not subject to restrictions

Q79
AI for Personalized Learning in Entertainment

What risk level should we assign this AI?



Systems that recommend music, movies, or books based on user preferences (e.g., Spotify or Netflix recommendation engines).

A. Prohibited AI
B. High-Risk AI: Extensive regulation; no-human interaction
C. Limited-Risk AI: transparency obligation; human interaction
D. Low-Risk AI: not subject to restrictions

Conclusion

How can we balance the potential benefits of AI with its long-term societal impacts?

- Establish clear ethical frameworks to guide AI development.
- Adhere to principles like human dignity, justice, and inclusiveness in AI design and deployment.
- Focus on creating AI systems that augment human capabilities, rather than replacing human roles entirely.
- Implement regulatory frameworks, such as the EU AI Act, to ensure AI systems meet safety, privacy, and accountability standards.
- Foster ongoing discussions with the public, experts, and policymakers to ensure AI is developed and deployed with societal values in mind.
- Encourage public participation in decisions regarding AI deployment, ensuring transparency and trust.

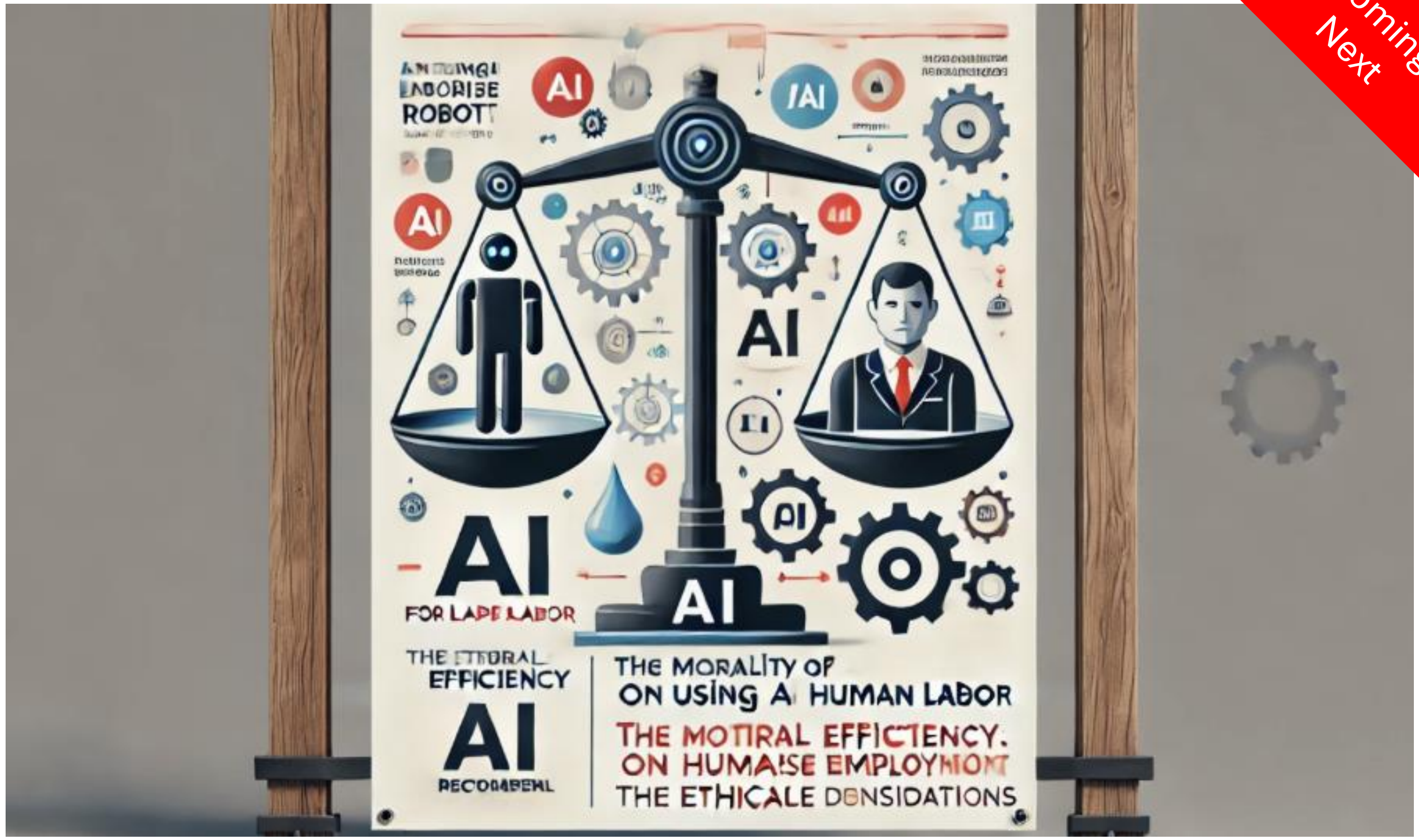
Coming Next

Lab 3 :: Responsible AI

The Morality of AI Decisions for the Enterprise

Scenario:

The CEO of a highly profitable multinational corporation announces plans to replace 80% of the company's HR staff with AI-powered chatbots to boost efficiency and cut costs. These chatbots will handle tasks like onboarding, employee inquiries, and performance evaluations. However, this decision raises ethical concerns about automating roles that require empathy and personal understanding.


ROBOTT
AI
AI
AI
AI
FOR LAPE LABOR
AI
THE ETTORAL EFFICIENCY
AI
RECOMBEHL
THE MORALITY OF ON USING A HUMAN LABOR
THE MOTIRAL EFFICTENCY. ON HUMAISE EMPLOYMONT
THE ETHICALE DENSIDATIONS

Objectives:

- Is this Responsible AI?
- What are the key implications around 'Ethical Principles'?
- What are the key implications around 'Societal values'?
- What is the Level of Risk as per EU AI Regulation?

Deliverables:

- Collaborative Analysis Exercise
- Report outlining:
  - Considerations about the problem scenario
  - Response to Lab Objectives
  - Recommendations for improvement

CAP 4623
TRUSTWORTHY ARTIFICIAL INTELLIGENCE

Dr. Fernando Koch
kochf@fau.edu