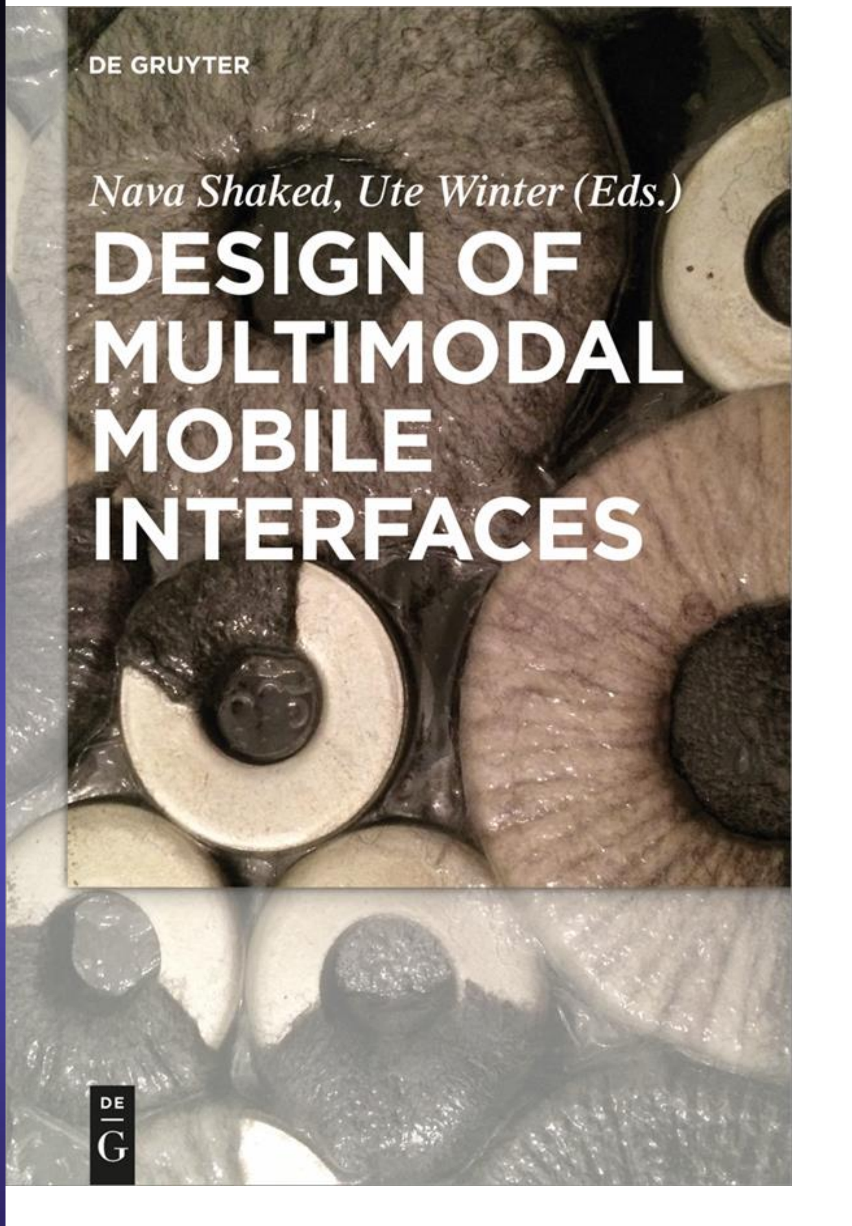
מכון טכנולוגי חולון
Holon Institute of Technology

Responsible AI: Ethics and Regulation in the new Era

FH | JOANNEUM University of Applied Sciences

- **Dr. Nava Shaked**

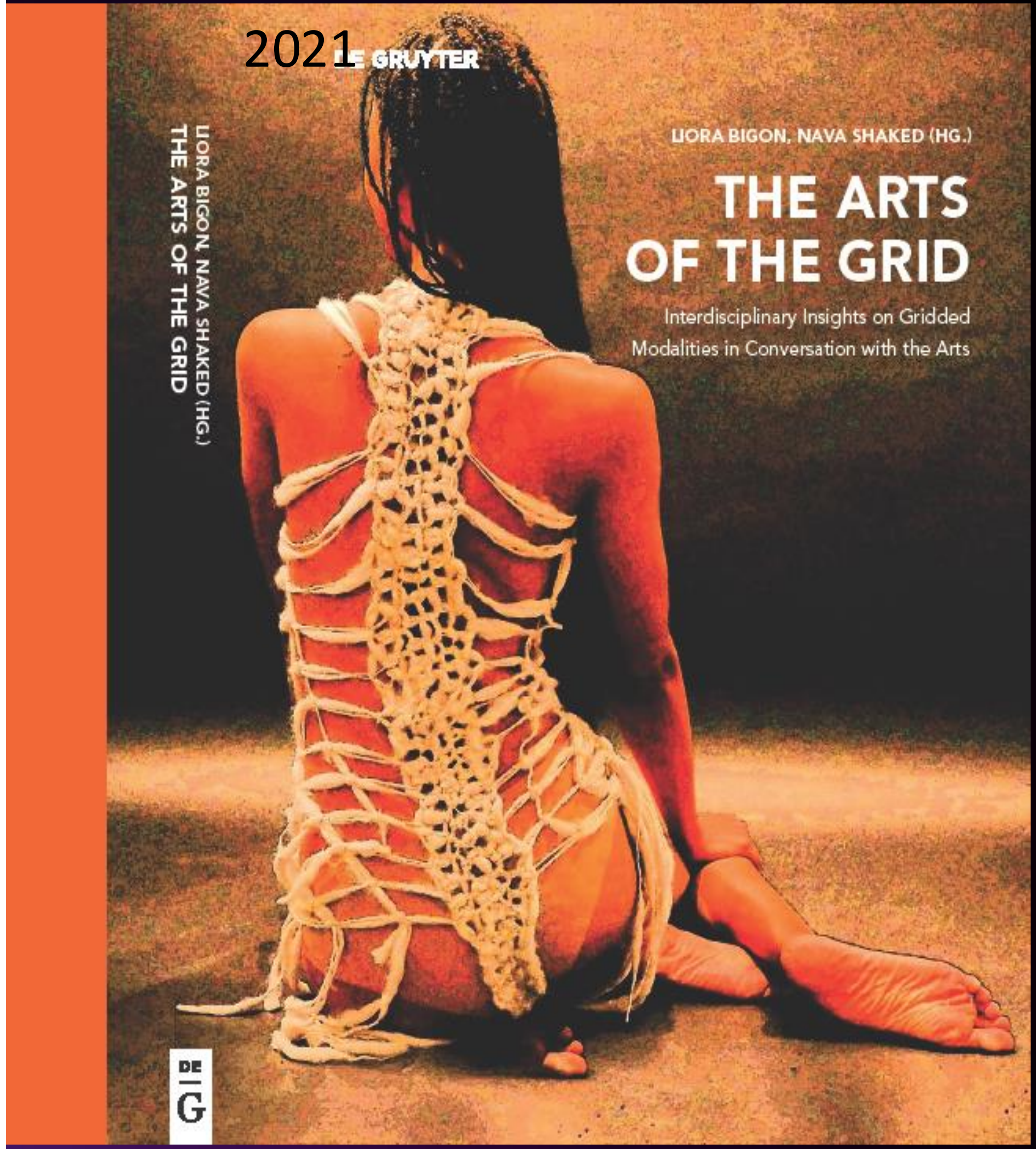
HIT, Holon Institute of Technology, Israel
May 19, 2025



DE GRUYTER

Nava Shaked, Ute Winter (Eds.)

DESIGN OF MULTIMODAL MOBILE INTERFACES



2021

DE GRUYTER

LIORA BIGON, NAVA SHAKED (HG.)

THE ARTS OF THE GRID

Interdisciplinary Insights on Gridded Modalities in Conversation with the Arts

LIORA BIGON, NAVA SHAKED (HG.) THE ARTS OF THE GRID

Agenda:

A word about AI

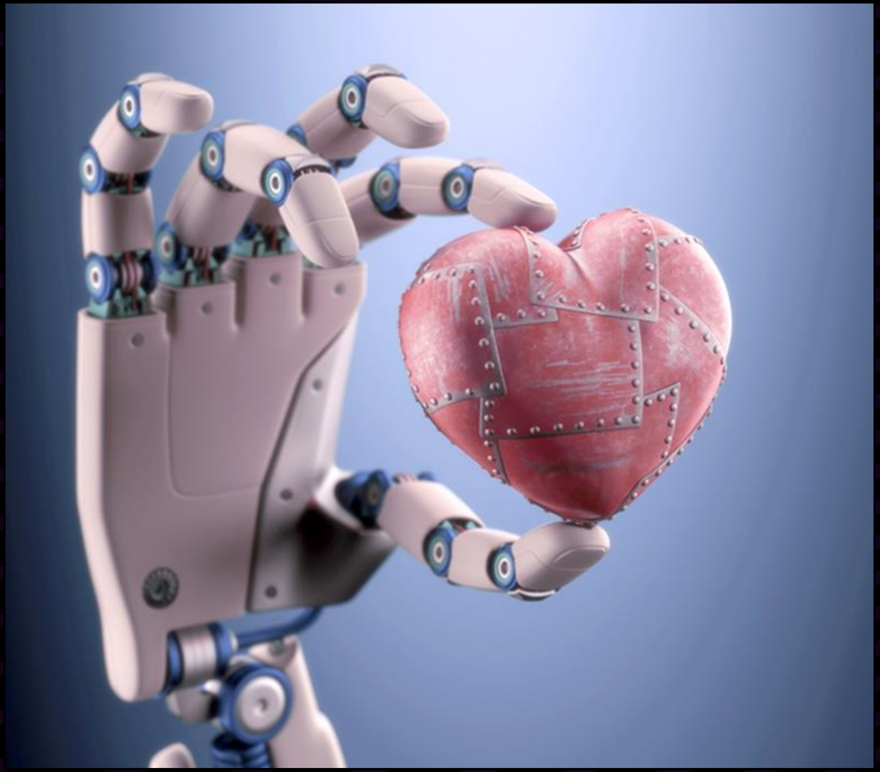
Ethical issues

Regulation effort

Responsible AI

Use cases

Food for thought



Why is this workshop relevant for you?

AI

This too shall pass…

Forbid or restrict

Embrace, prepare, adapt

AI Spectrum

Narrow AI

AI for domain specific

Edge AI

Hardware/tool based front end AI

Generative AI

Use of AI in uncharted areas

General AI

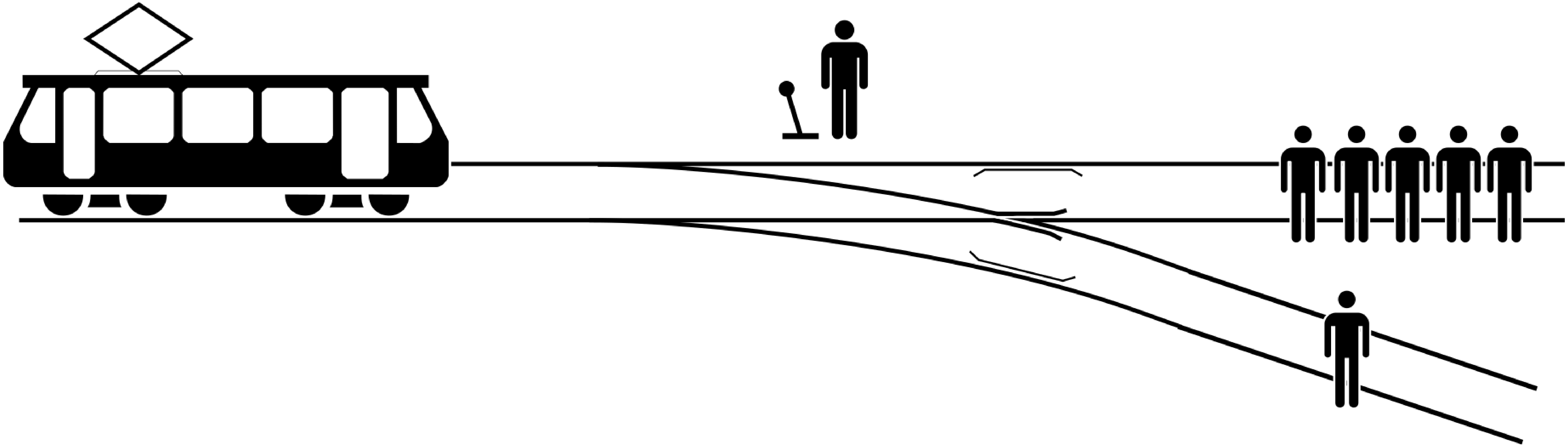
AI for all purposes

GenAI - Why now?

1. Technology is mature, fast and relatively reliable
2. Simple UI, free and short learning curve
3. Generation Z & naïve users are willing to adapt
4. Google search engine is not enough now

AI & Ethics

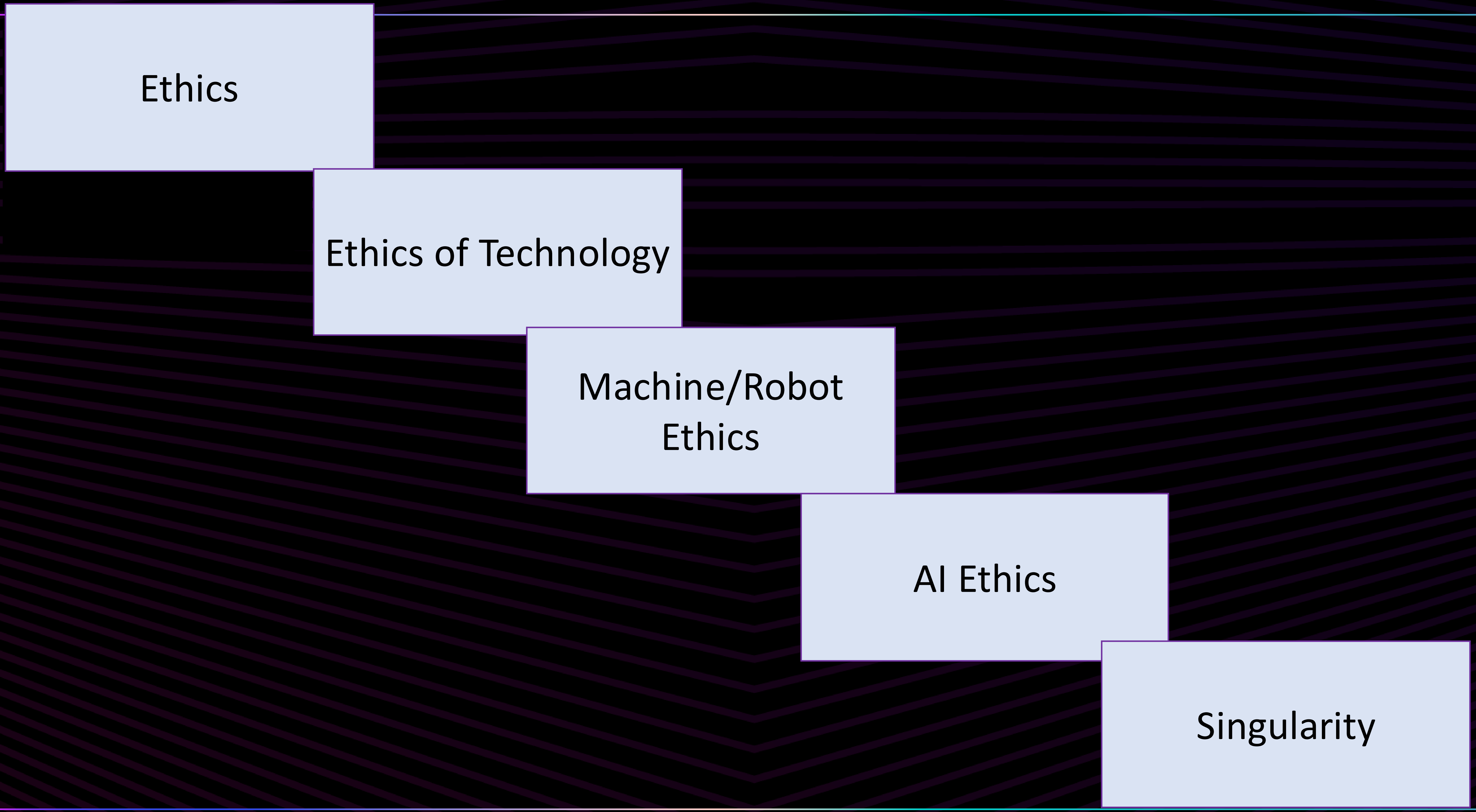
people
qualitatively
future
social
minds
person
regulations
problems
ethics
intellect
challenge
engineering
biological
creating
good
AI
upload
philosophy
AIs
brain
scenarios
design
weak
systems
population
civilization
mind
beings
engineer
machines
unethical



By Jonas Kubilius - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=39368927
1967 philosophy paper by Philippa Foot, and dubbed "the trolley problem" by Judith Jarvis Thomson in a 1976

- **Ethics** or **moral philosophy** is a branch of philosophy that "involves systematizing, defending, and recommending concepts of right and wrong behavior".

- **Ethics** defines concepts such as good and evil, right and wrong, virtue and vice, justice and crime.



Ethics

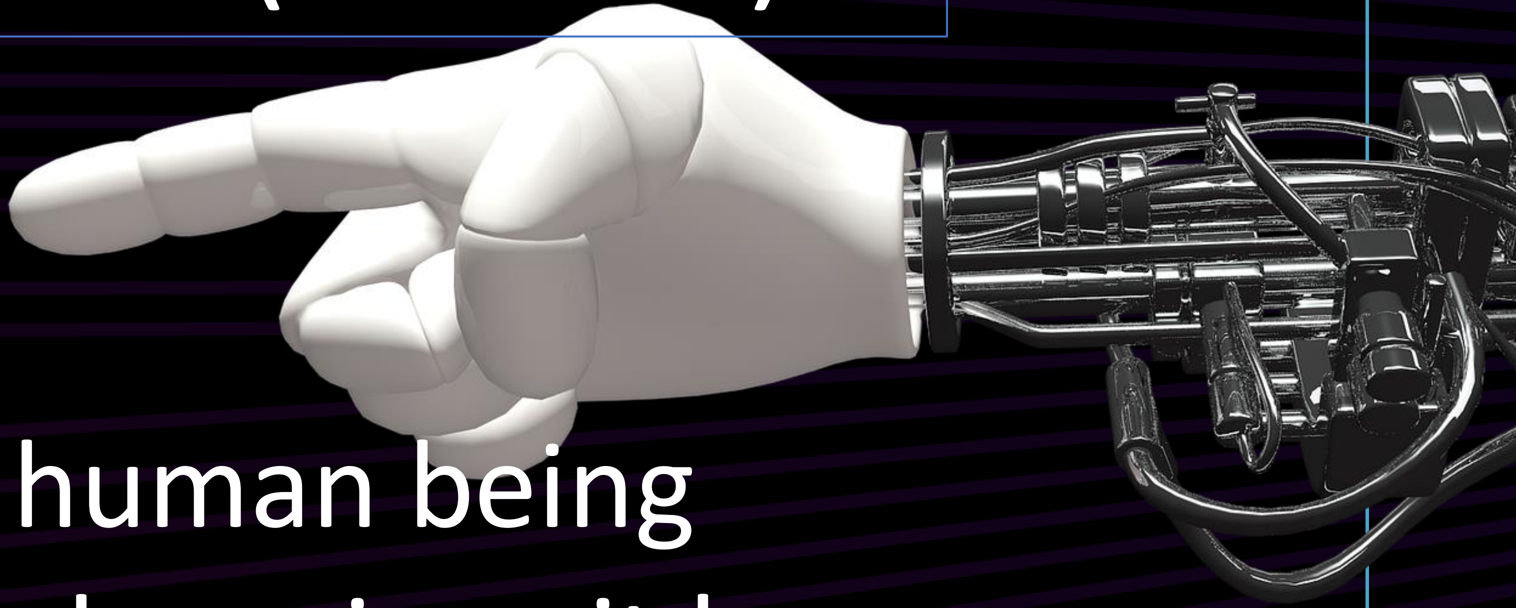
Ethics of Technology

Machine/Robot Ethics

AI Ethics

Singularity

The 3 Laws of Robotics (Asimov)



I. A robot may not injure a human being

II. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

III. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws

The GREY Area…..

- ✓Lack of transparency
- ✓Fake News
- ✓ Bias and discrimination
- ✓Use in “new” situation
- ✓3rd party usage
- ✓Targeting & segmentation

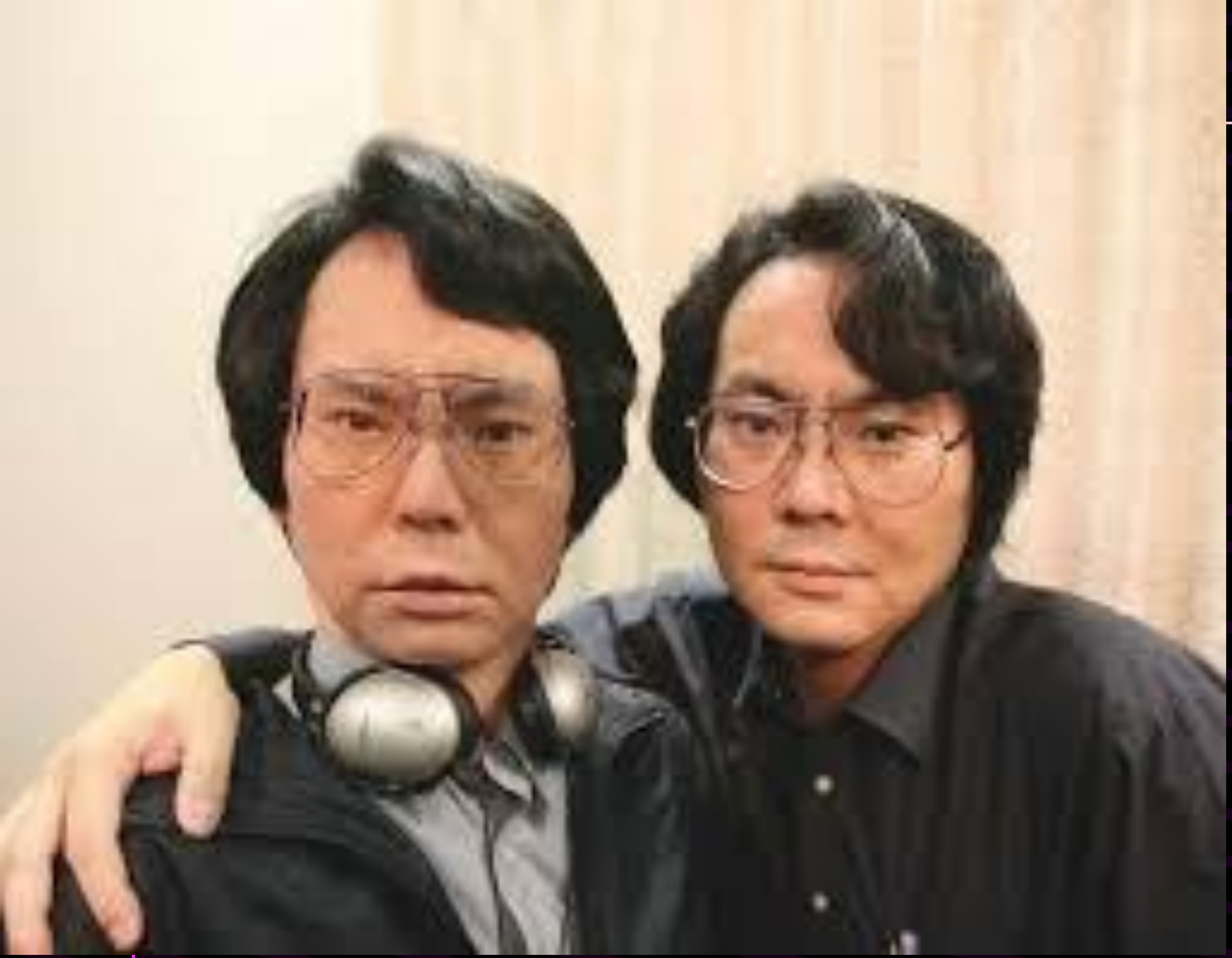
Ethics

- Biases in AI
- Robot Rights
- Threats to Human dignity /life/work
- Weaponization of AI
- Singularity

Regulation

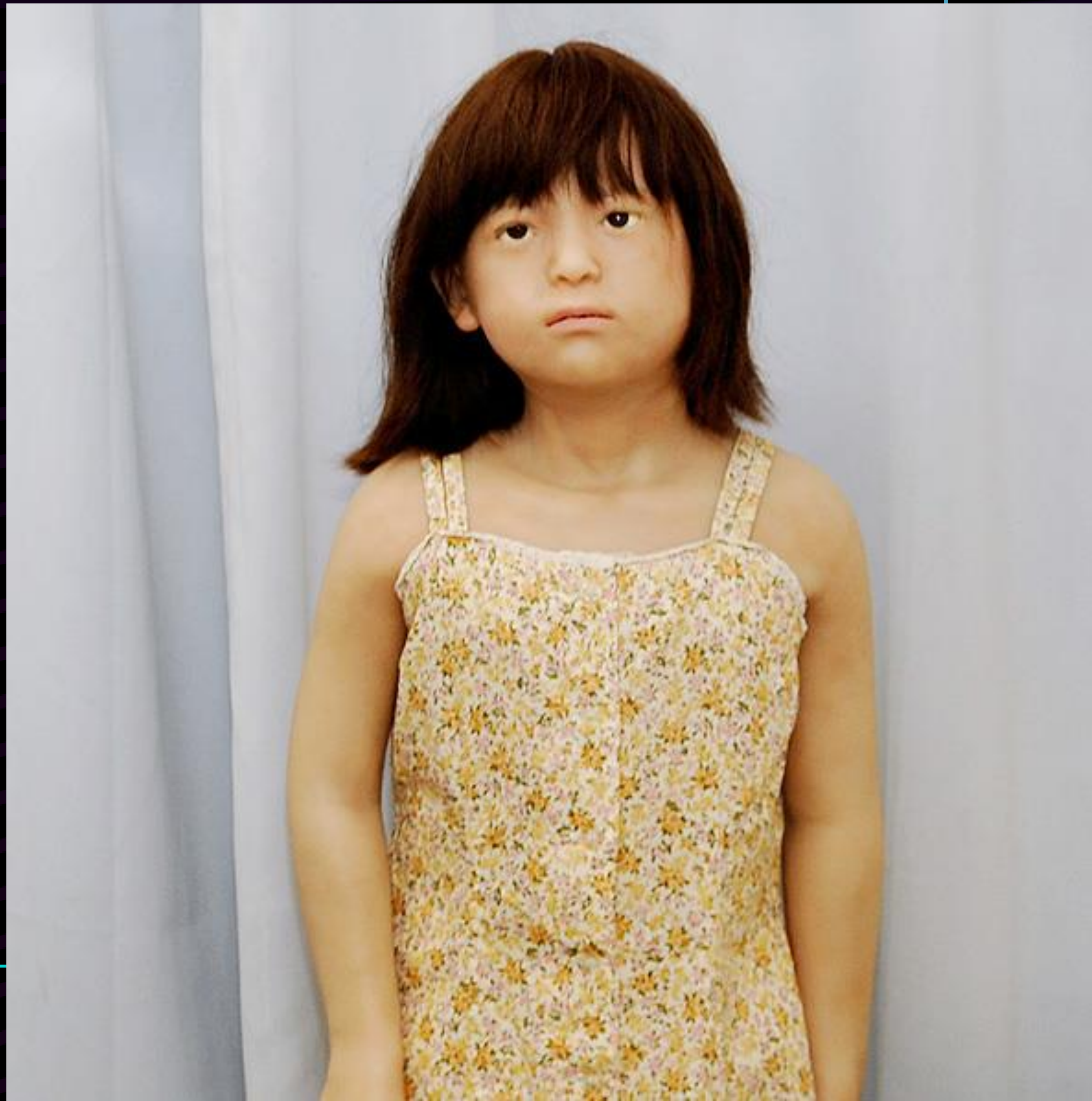
- Privacy (GDPR)
- Responsibility
- Liability
- Explainability
- IP & Legal use
- Singularity

AI application that raise ethical questions...



Prof. Hiroshi Hishiguro

Humanoid of his daughter



I share my identity with my Robot

Robots Rights

Ishiguro “In 10 years humans and robots will live side by side as a society “

What will be their legal and social status
Helpers, servants, soldiers, fighters what is their responsibility?

Right and duties- will they protected as humans, animals and environments? Citizens with the right to vote, pay taxes?

Violence against Man, Animals, environment- were in this equation virtual being are found ?
Sexual harassment, crimes against another robot or human?

Q1: Who will regulate AI ?

- What is the responsibility of a developer or solution architect ?
- Do we need a “Physician's Oath”?
- Is there a project you will not do?
- Are we “representors” or “changers ?

Replika
Need someone to talk to?
I've been missing you...

Q2: **Transparency, accountability, and open source**?

Do you measure our work’s influence?

- **Biases in AI systems**

Biased of AI – HR, race, gender, face, voice?

A customer service representative (

A nursemaid for the elderly

A soldier

A judge

A police officer

A therapist

GeriJoy

Liability & responsibility

- Should we inform users of AI?
- Monitoring AI use?

Medical Robots & Tele-Medicine

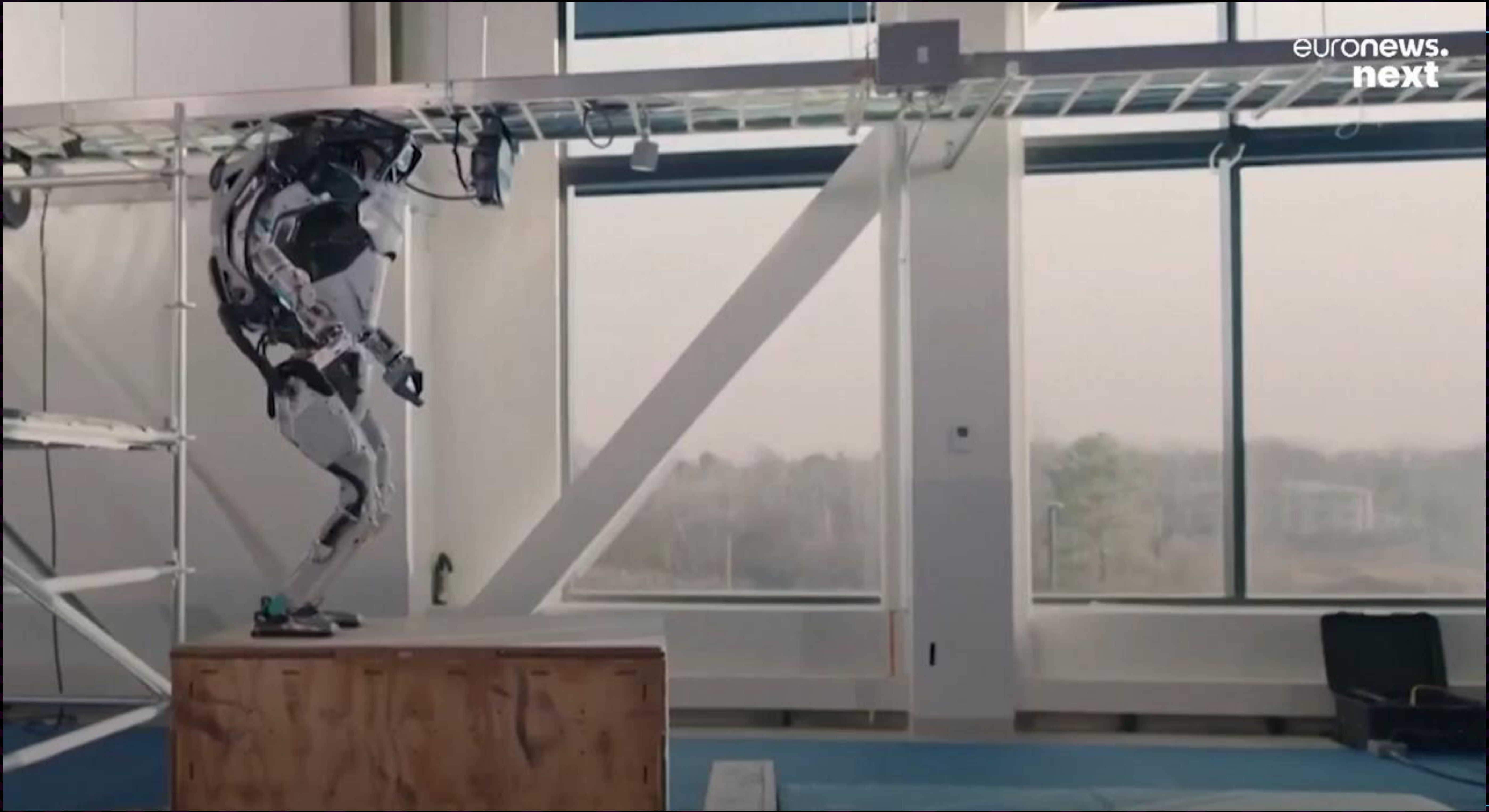
Q3: Setting Boarders

- How cultural differences effect AI ethics?
- Who is responsible to set boarders?
- Privacy...Corona…war…crises changed anything?

Weaponization of artificial intelligence

Bad ?

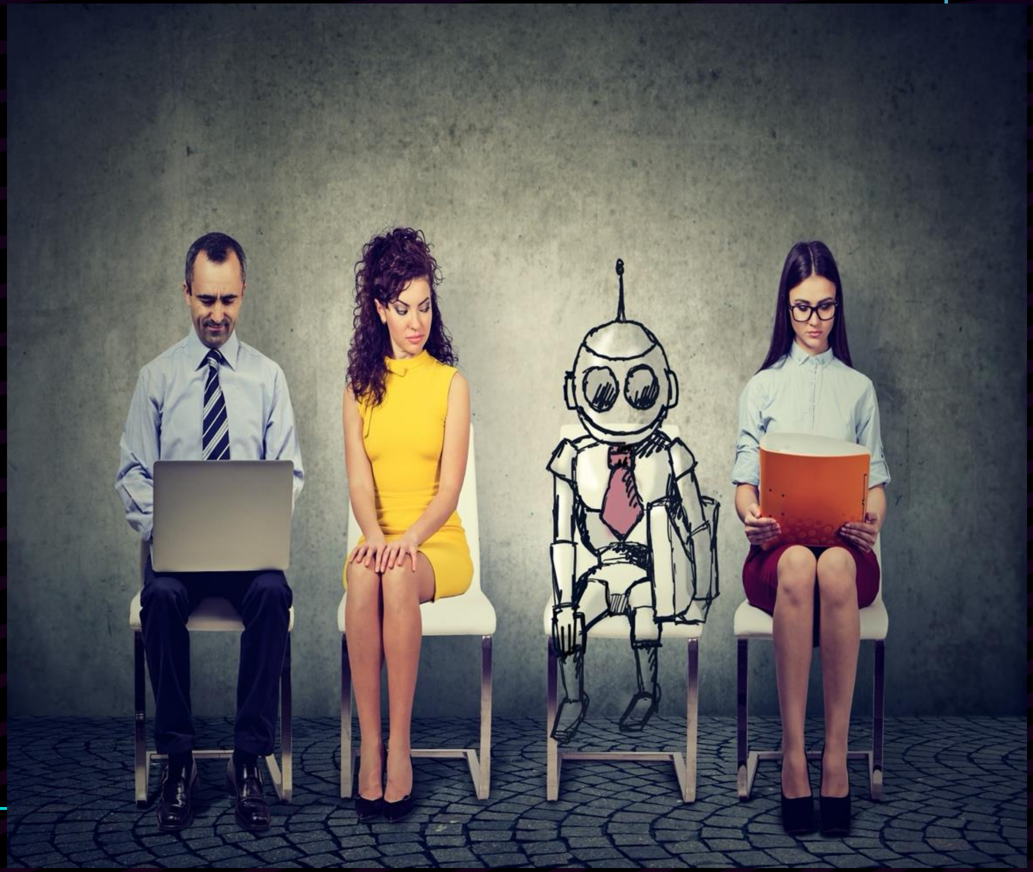
Good?


euronews.
next

//Let’s think

How do we reduce the fear of users and a controlled way and advance AI in the “right” format?

How to raise awareness to the added value on a professional and personal level to the user?



Can Regulation helps us with that?

AI Frameworks out there:

- ✓ Governance Frameworks: AI Ethics Committees and AI Governance Policies for companies
- ✓ Ethical Guidelines and Principles: Corporate Ethical Guidelines Adherence to Industry Standards
- ✓ Compliance Programs: Regular Audits and Assessments Training and Education
- ✓ Enforcing Responsible AI in Governments: Regulatory Bodies & Legislation
- ✓ Collaboration and International Cooperation: International Standards & Cross-Border Enforcement

Official Documents!

- ✓ Blueprint for an AI Bill of Rights
- ✓ Executive Order on the Safe, Secure, and Trustworthy Development and Use of AI

- ✓ EU’s General Data Protection Regulation (GDPR) in 2018,
- ✓ EU AI Act
- ✓ OECD AI Policy Observatory.

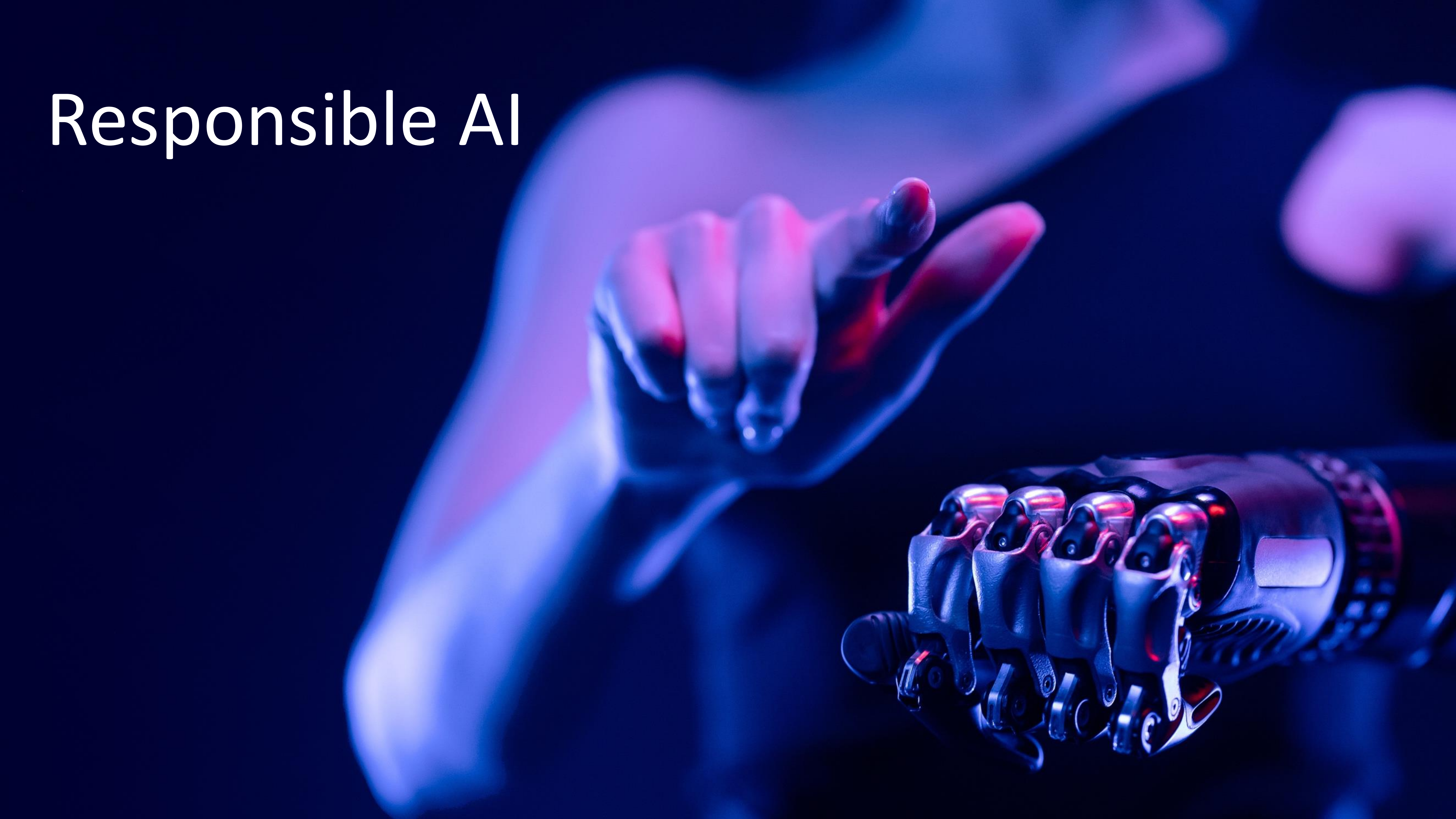
- ✓ Regulations for the Administration of Algorithm Recommendation of Internet Information Services:

- ✓ Artificial Intelligence and Data Act (AIDA):

The AI Act – Risks base approach

- ✓ **Unacceptable Risk:** clear threat to safety, livelihoods, and rights of people. Examples include social scoring by governments.
- ✓ **High Risk**: critical sectors like healthcare, transportation, and law enforcement.
- ✓ **Limited Risk:** Systems that need specific transparency obligations.
- ✓ **Minimal Risk:** **M**inimal risk, like AI-enabled video games or spam filters.

Responsible AI



Responsible AI

The term "Responsible AI" is a broad concept that encompasses various **ethical and societal considerations** when designing, developing and deploying artificial intelligence systems.

- No single, universally agreed-upon official definition,
- Several organizations and experts have proposed guidelines and frameworks
- Key principles of responsible AI.
- Definition and implementation of RAI may vary depending on the context, industry, and cultural factors

The RAI - Key principles

✓ Non Bias & Fairness

✓ Transparency, Explainability & Accountability

✓ Data protection & Privacy .

✓ Robustness, Safety & security, reliability

✓ Sustainability & Beneficial use (Human Centered)

✓ Scalability

RAI Organizations and Frameworks

- **OECD:** The Organization for Economic Co-operation and Development has published a set of principles for responsible AI that focus on fairness, accountability, transparency, and privacy.

- **NIST:** The National Institute of Standards and Technology has developed a framework for artificial intelligence that includes guidelines for responsible AI.

- **European Union:** The EU has proposed regulations for AI, including requirements for transparency, accountability, and risk assessment.

- **GPAI:** Global Partnership of Artificial Intelligence. Multi Stakeholder multinational focus to convene experts from a wide range of sectors to assist countries and large organizations deploy RAI in there long terms strategies and projects

New York state government to monitor its use of AI under a new law

U.S. News - Updated 3:33 PM EST, December 27, 2024

ALBANY, N.Y. (AP) — New York state government agencies will have to conduct reviews and publish reports that detail how they’re using artificial intelligence software, under a new law signed by Gov. Kathy Hochul.

- ✓ Hochul, signed the bill last week after it was passed by state lawmakers earlier this year.
- ✓ The law requires state agencies to perform assessments of any software that uses algorithms, computational models or AI techniques, and then submit those reviews to the governor and top legislative leaders along with posting them online.
- ✓ It also bars the use of AI in certain situations, such as an automated decision on whether someone receives unemployment benefits or child care assistance, unless the system is being consistently monitored by a human.
- ✓ State workers would also be shielded from having their hours or job duties limited because of AI under the law.

Academic Institutions' Documentation on Responsible AI

1. **Center for Responsible AI at New York University (NYU)**
2. **University of California (UC) System**
3. **Georgia Tech's Guidance for Effective and Responsible Use of AI in Research**
4. **EDUCAUSE Action Plan for AI Policies and Guidelines**

GPAI – Responsible AI

The responsible development, safe use, and governance of human-centred (generative) AI systems in congruence with the UN Sustainable Development Goals, **ensuring diversity and inclusivity to promote a resilient society**, in particular, in the interest of vulnerable and marginalised groups

SARIS project:

Scaling Responsible Artificial Intelligence Solutions.

An initiative of the Responsible AI (RAI) working group of the Global Partnership on Artificial Intelligence (GPAI currently part of the OECD.AI.

This network is focused not only on responsibility in the development of AI-based systems, but more uniquely on the intersection between scalability and responsibility.

Responsible AI with emphasis on Scalability

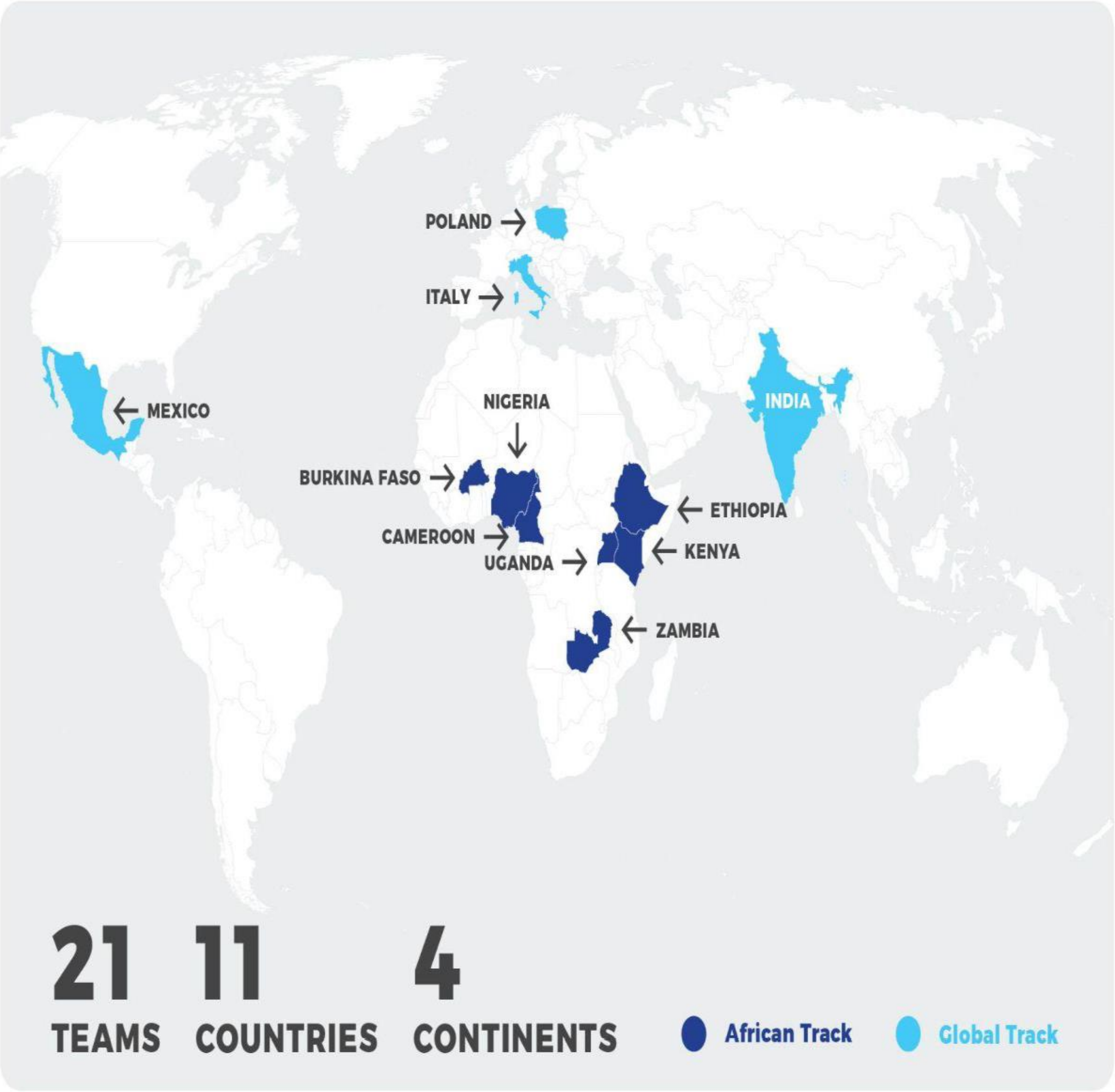
Data and cultural integration

Bias amplification

Legal and regulatory integration

Technical and operational expansion

Labor and economic considerations



Type of SARIS projects?

AI based systems for internal organizational impact

Data Collection and Privacy & Reliability Management

AI based system for customer service and services engagement

Accessibility

Made with Gamma

1

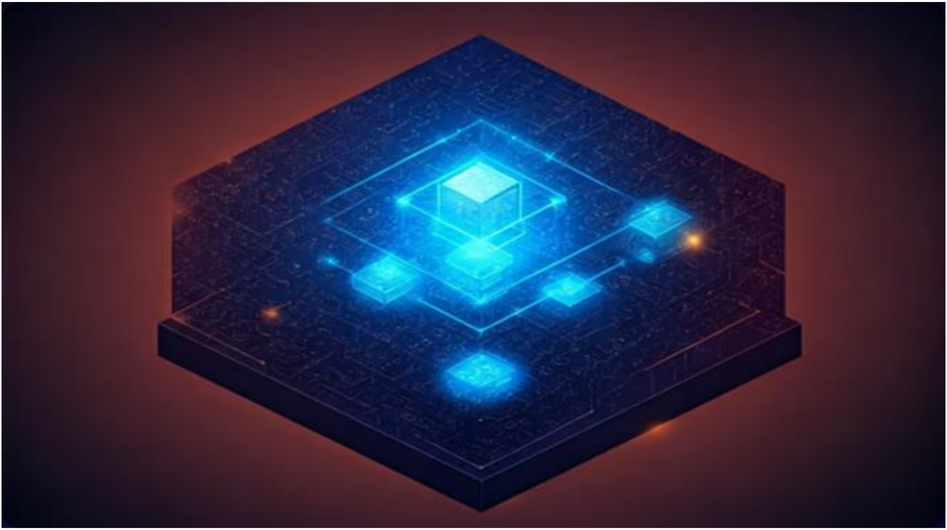
Employees Empowerment
Knowledge chat

2

Reducing manipulation
and control in employee performance

3

Satisfaction recognition to
improve customer interactions
with virtual agents



4

ASLAC Improving Automatic Data Models for Sign Language

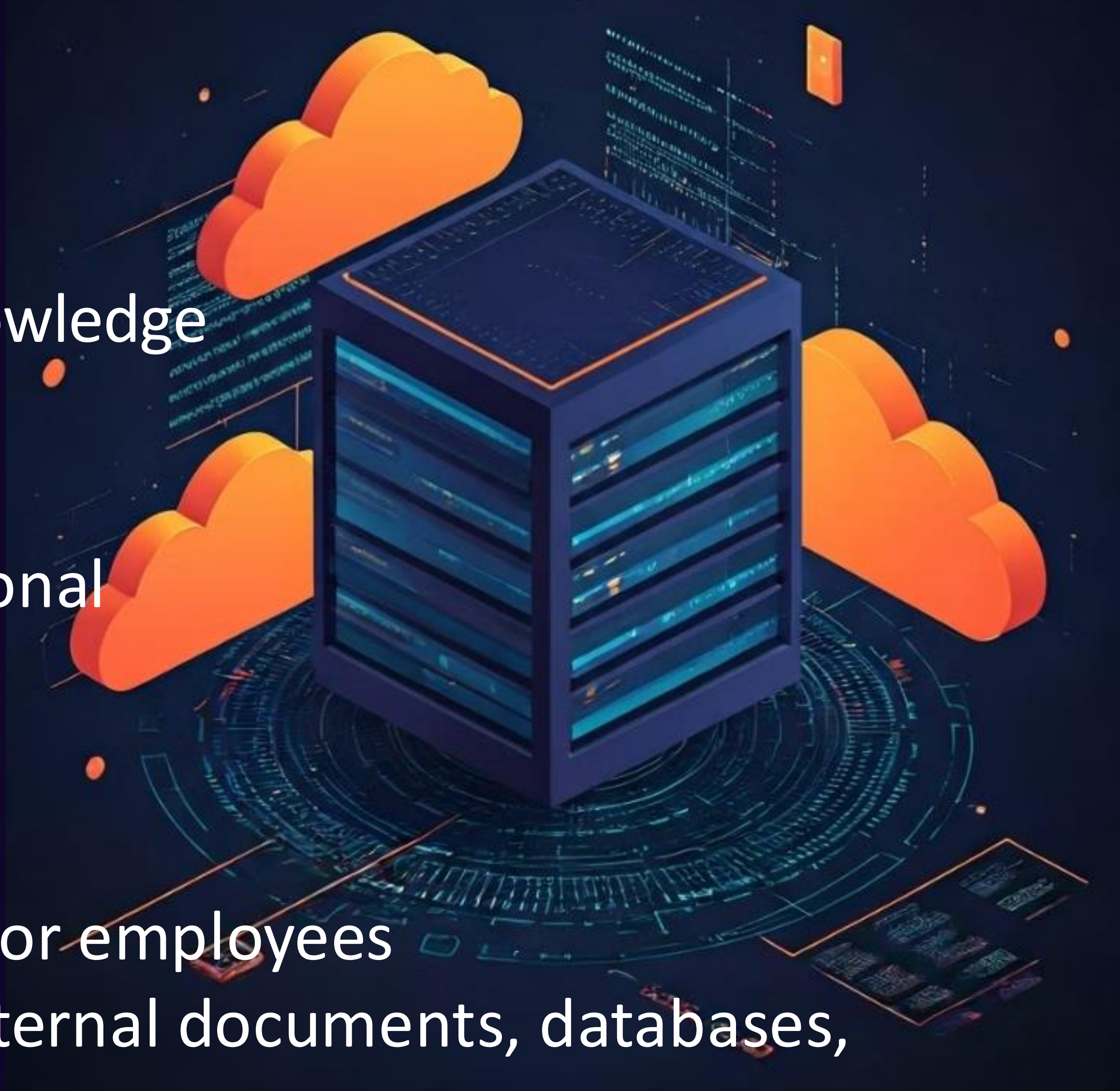
[redacted]

Employees Empowerment
Knowledge chat

1 Model (LLM)-based system developed to enable organizations to easily search and query their knowledge resources

2 **Knowledge chat to S**upport improved organizational efficiency and productivity,

3 **Virtual assistance to** making it easier and faster for employees to access and share information contained in internal documents, databases, policies, etc.



RAI Considerations – relevant to the general category

1. Data governance considerations

2. Ethical issues in the training data of LLM

3. Risk of hallucinations and inaccuracies

4. Impact on working ecosystem

Reducing manipulation and control in employee performance

1 Personalized ML Software for recommend tasks or actions to employees

2 Used to verify the execution of tasks and provide feed

3 Involve “micro-learning as well as opportunities to convey expertise via micro-surveys

4 Subject to strict consent and data management frameworks

RAI Aspects of the Project:

Transparency

Clarity

Autonomy

Trust

Satisfaction Recognition to improve Customer Interactions with virtual agents

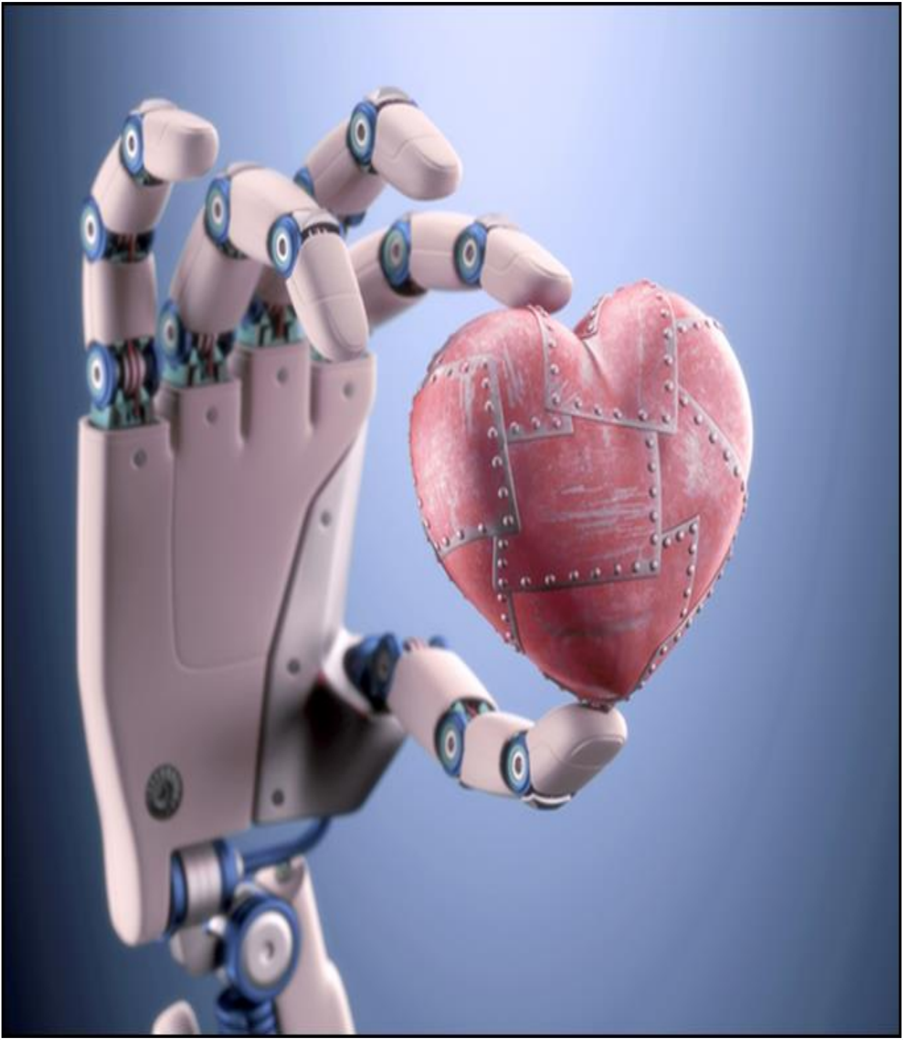
Customer service bot which is able to recognize reflect and adapt to user’s satisfaction and emotional state.

Realtime collection of data, user feedback and contextual info while creating responsible responses and service

Balancing privacy and personalization and in the same time being transparent to the user

The global and local understanding of the “Rules of Engagement”

RAI Challenges of the Project:



The area of emotion detection and UI-UX is very delicate.

User satisfaction in real-time, while maintaining transparency and ethical standards is even harder

Balancing the transparency with privacy and task completion.

How to implement the “rules of engagement"

[redacted]

ASLAC Automatic Sign Language – creating a reliable and responsible data collection process

Data collection process that allows the use of data with personal information (video) while guaranteeing that no personal features make it to the training sets where gestures, facial expressions, and transcripts are tokenized and embedded (Data “Clean Room”)

[redacted]

Improving Automatic Data Models for Sign Language

Initial ASL data for model training

Data Processing and creating a model with key skeletal pots

SL Model - Ensuring data consistency, calibration And alignment to kinematic movement

Create an Avatar capable of demonstrating key signs using skeletal data

Food for thought….

Who will regulate AI ?

- What is the responsibility of a developer or solution architect ?
- Liability - Do we need a “Physician's Oath”?
- Is there a project you would not do?

Transparency, accountability, and open source?

- Do you measure our work’s influence?
- Biased of AI – HR, race, gender, face, voice?
- Full disclosure – automation, emotion detection, information gathered

Setting Boarders

- How cultural differences effect AI ethics?

- Who is responsible to set boarders?

- Privacy...Corona...war...crises changed anything?

Shakedn@hit.ac.il

מכון טכנולוגי חולון
Holon Institute of Technology