

Tutorial: AI Ethical challenges and considerations



Jim Torresen

Research group Robotics and Intelligent Systems (ROBIN)

Department of Informatics and RITMO

University of Oslo, Norway

Email: jimtoer@ifi.uio.no

*2024 IEEE International Joint Conference on Neural Networks
(WCCI IJCNN-2024)*



UiO : **University of Oslo**



MorphCast for Zoom

Emotion AI for Zoom: AI decoding facial reactions in real-time in Zoom conference

Enhance your Zoom video conferences by adding the **MorphCast Attention Tracking Analytics** from the Zoom Marketplace. Unlock the advanced features for **tracking attention, engagement, and emotions**.

Use it for free or at market's best value!

MorphCast for Zoom

MorphCast for Zoom enables you to:

- **Detect emotions:** Understand your audience's emotional reactions in real-time.
- **Measure attention:** Check the participants' level of attention.
- **Assess engagement:** Quantify how engaged your audience is.
- **And much more:** Key features evolve weekly.



TECHNOLOGY > AI AND AUTOMATION | May 14, 2024

OpenAI launches GPT-4o, flaunting ability of model to detect user emotions

GPT-4o, claims OpenAI, is faster and more adept at handling text, audio and video – and even detecting user emotions, a major source of controversy in AI research.

Tutorial Agenda – Ethical perspectives and technical challenges and opportunities with care robots

1. Introduction and motivation
2. Ethical challenges and perspectives
3. How to address ethical considerations in research including examples
4. Ethical considerations in own research at University of Oslo
5. Future opportunities in ethics-related research

Why do we not want AI and robots?

- Privacy: Collects a lot of **data** that **can be published or misused?**
- Safety and security: Can **harm us physically** (and mentally)?
- Society: **Colder society?**
- Future jobs: **Taking away our jobs?**



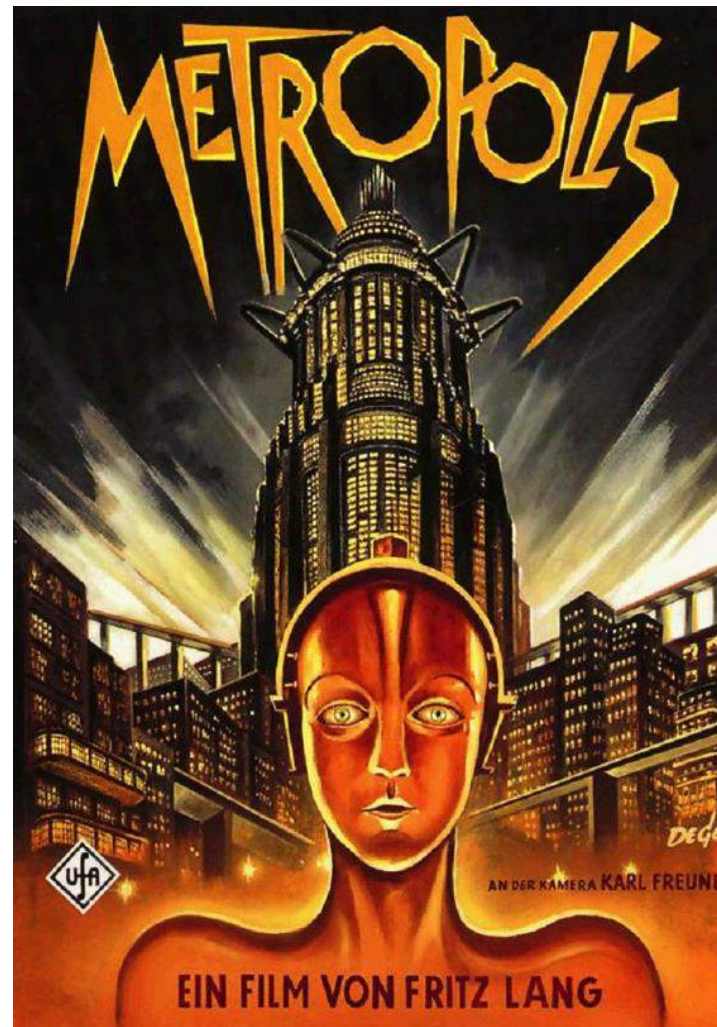
Why AI and robots?

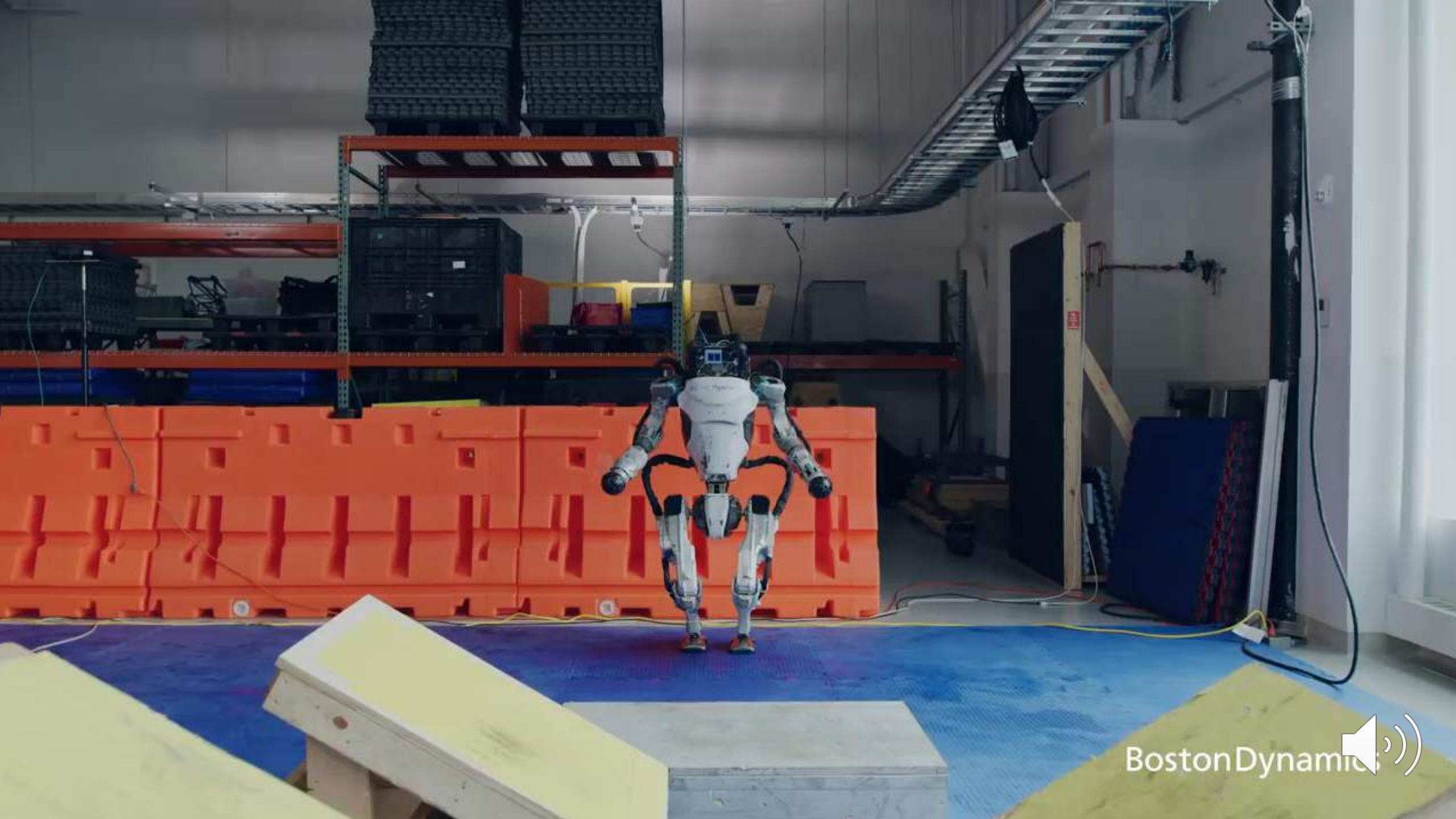


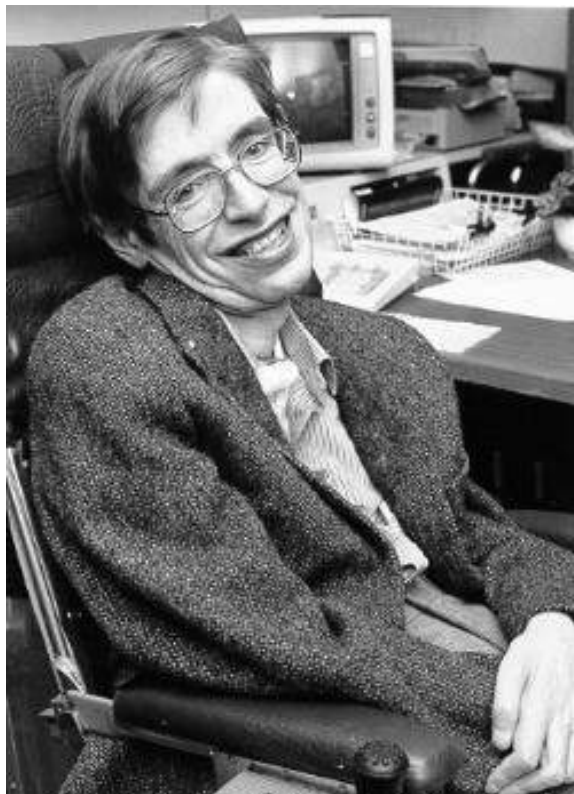
- We can do (mental and physical) **work with less effort**
- Reduced **manual routine work**
- Improve dignity by making us **more independent** of the help of others
- Giving us **better health and longer life**



Is Terminator Coming Soon?



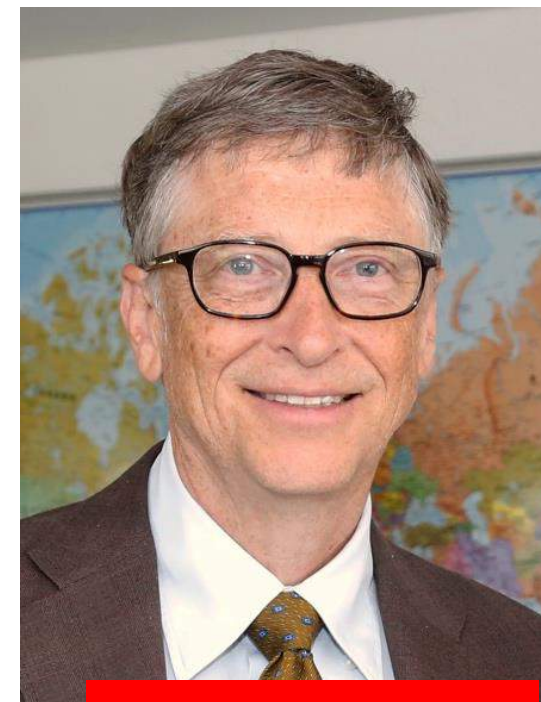




“Humans, limited by slow biological evolution, couldn’t compete and would be superseded by A.I.”



AI is our “biggest existential threat”



“I am in the camp that is concerned about super intelligence.”



Recent AI Concern Initiatives

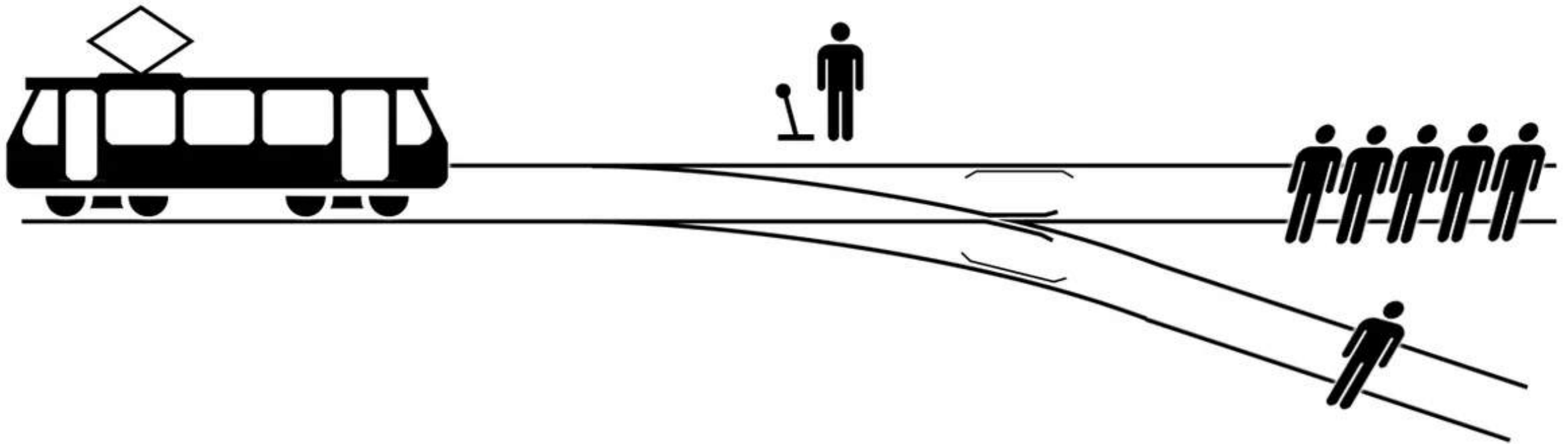
- **Future of Life Institute Open Letter:** "*call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4*" (published March 22, 2023).
 - Currently 33,707 signatures (including AI pioneers Yoshua Bengio and Stuart Russell, Apple co-founder Steve Wozniak, ++)
- AI 'godfather' **Geoffrey Hinton** (75) warns of dangers as he **quits Google**
 - in a statement to the New York Times, he is saying he is now regretted his work.



<https://futureoflife.org/open-letter/pause-giant-ai-experiments/>

<https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html>

Automatic programming – Moral dilemmas



Originates in a 1967 essay by the British philosopher Philippa Foot about the distinction between doing harm and allowing harm

Ethical Risks of Developing AI Systems

- **Jobs:** People may become unemployed because of automation.?
- **Jobs:** We get too much free time.?
- **Technology:** Loosing human skills?
- **Technology:** Artificial intelligence can be used for destructive and unwanted tasks.?
- **Technology:** Successful KI can lead to the extinction of mankind?

Levels of driving automation

Increased machine autonomy

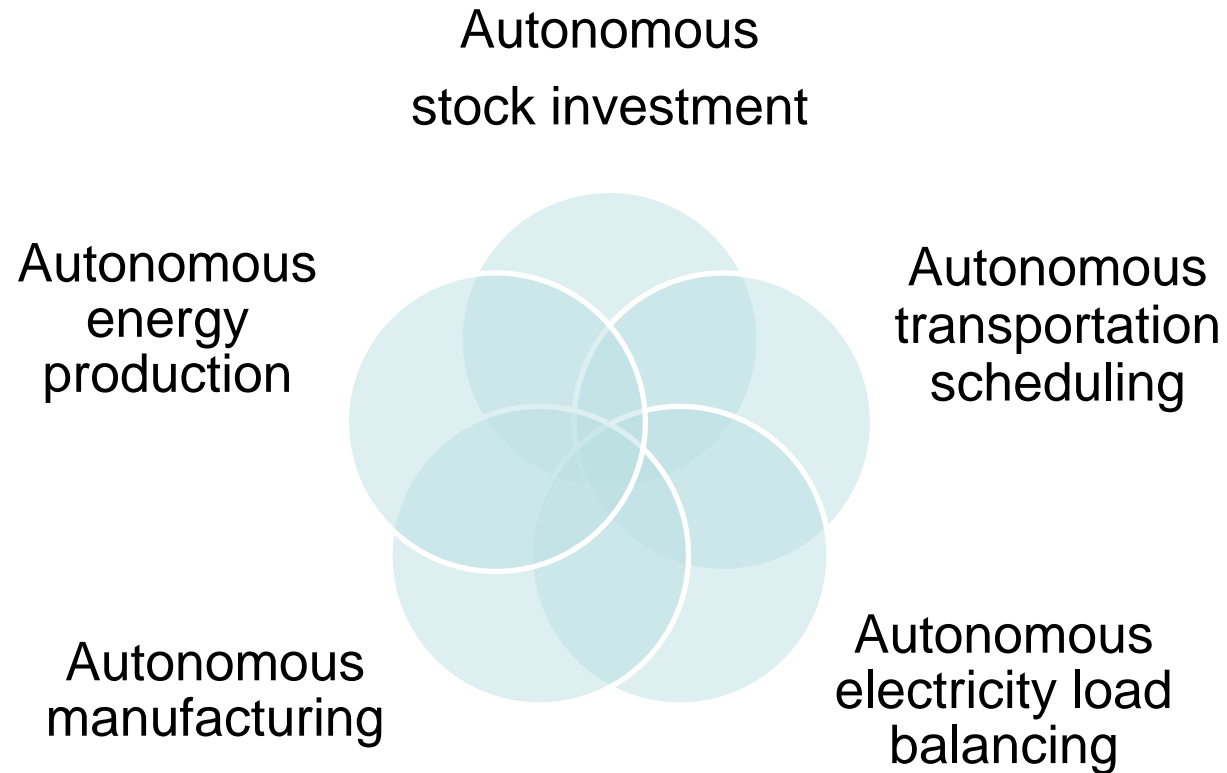
- Level 0: Automated system issues warnings
- Level 1 ("hands on"): Driver and automated system shares control over the vehicle.
- Level 2 ("hands off"): The automated system takes full control of the vehicle (accelerating, braking, and steering).
- Level 3–4 ("eyes off"): The driver can safely turn their attention away from the driving tasks
- Level 5 ("steering wheel optional"): No human intervention is required (robot taxi)



Image by rawpixel.com



Future Scenario with Autonomous Interacting AI Systems





Ethical Countermeasures

- **Designers, procurers and users need to be aware of possible ethical challenges that should be considered**
 - e.g. avoiding misuse and allowing for human inspection of the functionality
- The **systems themselves** be able to **do ethical decision making** to reduce the risk of unwanted behavior
 - Decide when a human is to be contacted or the machine should stop



Machine Ethics: Moor's four categories of ethical agency

- 1) *Ethical Impact Agents*: Any machine that can be evaluated for its ethical consequences.
- 2) *Implicit Ethical Agents*: Machines that are designed to avoid unethical outcomes.
- 3) *Explicit Ethical Agents*: Machines that can reason about ethics.
- 4) *Full Ethical Agents*: Machines that can make explicit moral judgments and justify them.

A. F. Winfield, K. Michael, J. Pitt and V. Evers, "Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems [Scanning the Issue]," in *Proceedings of the IEEE*, vol. 107, no. 3, pp. 509-517, March 2019, doi: 10.1109/JPROC.2019.2900622.

J. H. Moor, "The nature, importance, and [33] difficulty of machine ethics," *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 18–21, Jul./Aug. 2006.

ELSI – E (ethical), L (legal), and S (social) Implications

- Named by **James Watson**, at the press conference in 1988 announcing his appointment as **director of the Human Genome Project (HGP)**.
- A term initiated by the **ELSI Research Program** to foster basic and applied research on the **ethical, legal and social implications** of genetic and genomic research for individuals, families and communities.
- Still mainly consider **ethical, legal and social implications** (ELSI) or **aspects** (ELSA) of emerging sciences, notably genomics and nanotechnology

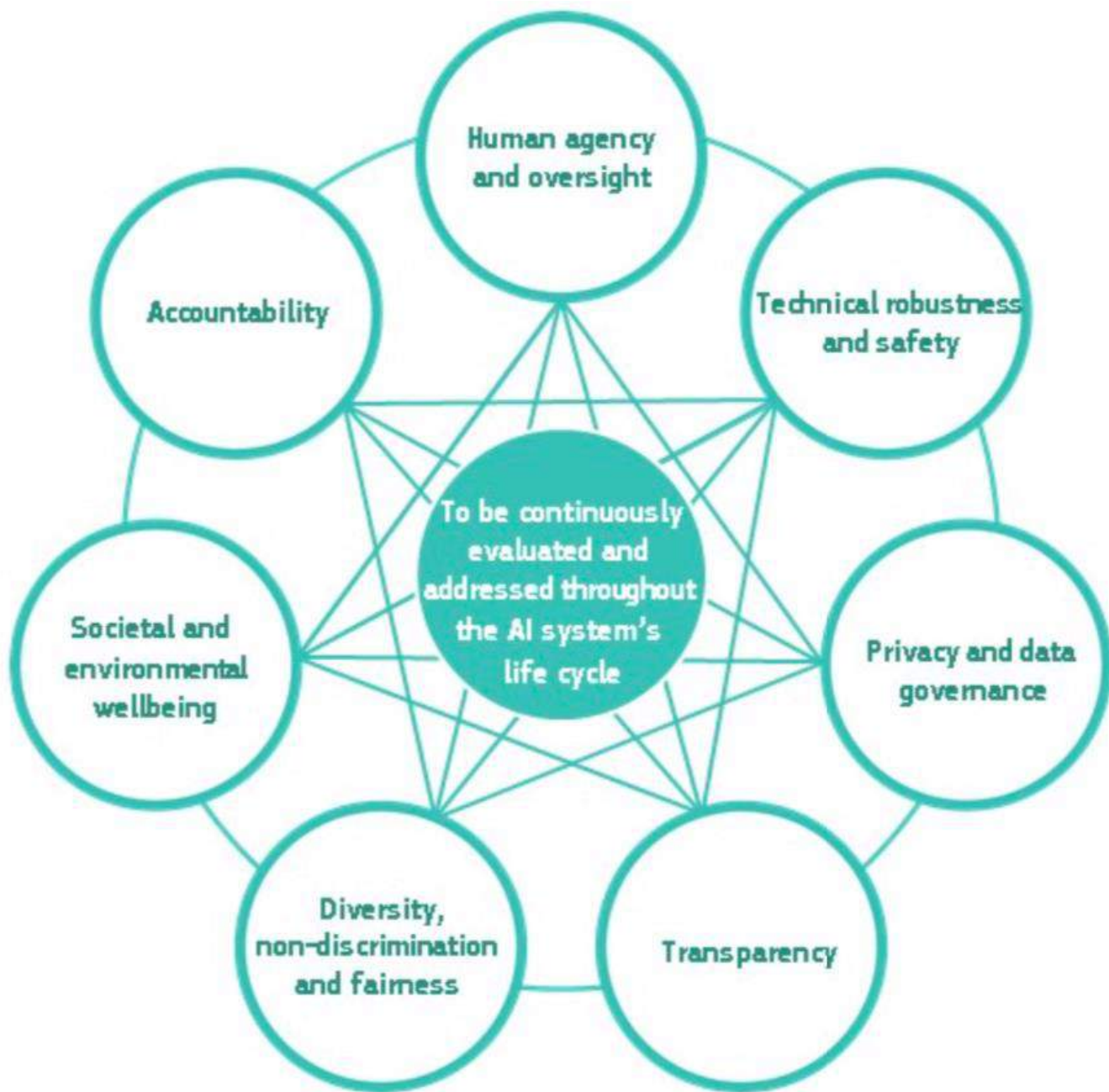
UNESCO – United Nations Educational, Scientific and Cultural Organization

- UNESCO produced the first-ever global standard on AI ethics – the ‘Recommendation on the Ethics of Artificial Intelligence’ in November 2021. This framework was adopted by all 193 Member States.
- Four core values which lay the foundations for AI systems that work for the good



International Work on Ethics and AI

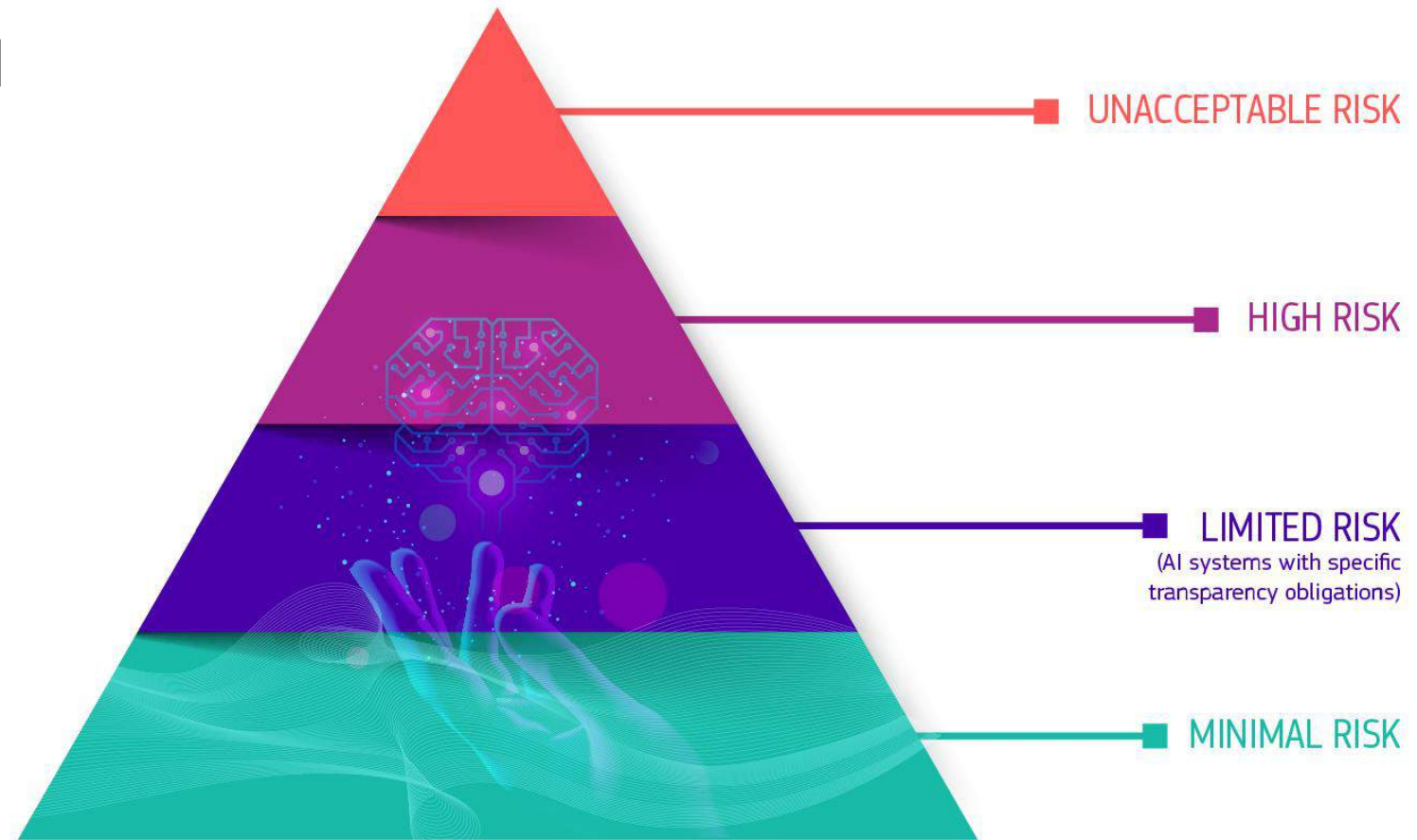
- **IEEE Standards Association**
 - Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems
- **ISO/IEC JTC 1/SC 42 Artificial intelligence**
 - Standardization in the area of Artificial Intelligence
- **EU High-Level Expert Group on Artificial Intelligence**
 - 52 experts on Artificial Intelligence, comprising representatives from academia, civil society, as well as industry.
 - support the implementation of the European strategy on Artificial Intelligence.
 - recommendations on future-related policy development and on ethical, legal and societal issues related to AI, including socio-economic challenges.



<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

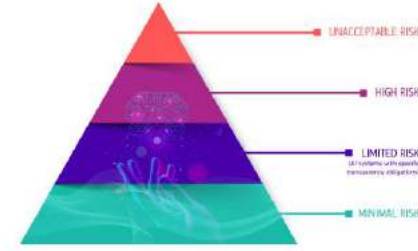
EU's risk-based approach to AI

- EU has adopted harmonised rules regarding AI applications:
- **EU AI Act:**
 - Prohibition of unacceptable AI practices
 - Regulation of high-risk AI systems
 - Adopted 2024



Source:

https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence_en



EU AI Act Proposal – Prohibited AI practices

AI systems with an unacceptable level of risk to people's safety will be strictly prohibited, including

- systems that deploy subliminal or **purposefully manipulative techniques**
- **exploit people's vulnerabilities or are**
- **used for detrimental treatment based on social scoring** (classifying people based on their social behaviour, socio-economic status, personal characteristics); or
- **emotion recognition** systems in workplace and educational institutions.

How to address ethical considerations in research: NENT AI Research Ethics Considerations Report

NENT – The National Committee for Research Ethics
in Science and Technology (**Norway**)

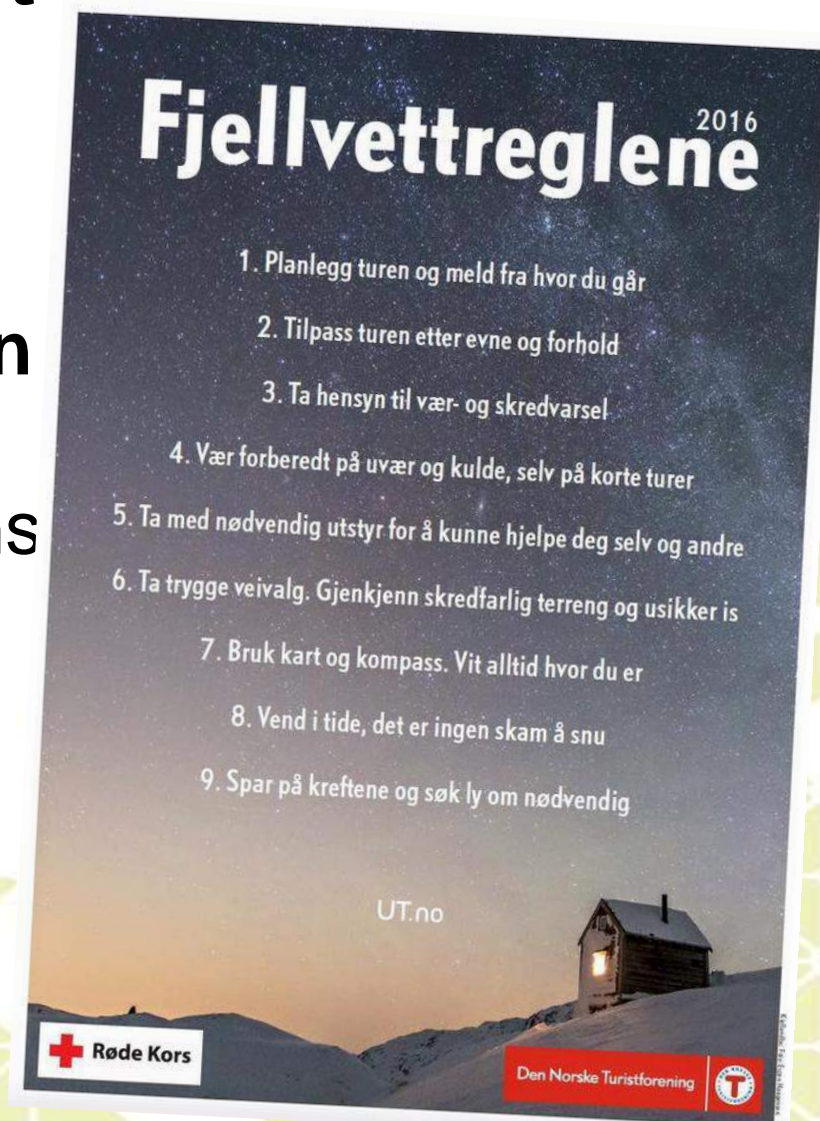
- **Input** from relevant **academic/research institutions** involved in artificial intelligence research
- Review of **international and national reports** and guidelines
- Given by NENT in 2019 (Norwegian). An English translation published in October 2020.



Look at the report? Google for "statements AI NENT"

Characteristics of AI – NENT AI report considerations

1. AI mimics, **replaces and extends human intelligent action** and human decision-making and assessment (four considerations)
2. AI has **numerous applications** (two considerations)
3. AI uses and generates **big data** (three considerations)



Assign responsibility

The more adaptive, autonomous and complex an AI system is, the harder it will be to control it, and the responsibility for its actions becomes more difficult to assign.



Image by rawpixel.com

Inspectability

Identifying the sources of the data used by the systems as well as how the systems make decisions is important.



Image by rawpixel.com

Recognize Uncertainty

The uncertainty and unpredictability associated with AI systems have many dimensions related to their training and use.



Image by rawpixel.com

Safeguard the privacy and consideration of individuals

There are contradictory desires for data maximization for machine learning on the one hand and data minimization for privacy on the other.



Image by rawpixel.com

Ethical considerations in own research at Univ of Oslo

University of Oslo, Norway

Robotics and Intelligent Systems group



Robotics and Intelligent Systems (ROBIN)



Jim Tørresen
Professor, Group leader



Mats Høvin
Assoc. Prof.



Kyrre Glette
Professor



Kai Olav Ellefsen
Assoc. Prof.



Yngve Hafting
Ass. Prof.



Vegard D Søyseth
Principal Engineer



Adrian Bergflødt
Assistant Engineer

**Postdoktor /
forsker:**
Benedikte Wallace
(RITMO)



Diana Saplacan
Lindblom (VIROS)



Adjunct positions (20%):

Alexander Wold (assoc.prof.)
Ole Jakob Elle (Prof.)
Roar Skogstrøm (lecturer)
Ståle Skogstad (assoc.prof.)
Tønnes Nygaard (lecturer)

PhD students

(ROBIN main superv.):

Adel Baselizadeh
Bjørn Thor Jonsson (RITMO)
Ege du Bruin
Emma H Stensby
Ivar-Kristian Waarum (NGI)
Katrine Nergård
Marieke van Otterdijk
Mojtaba Karbasi (RITMO)
Mateusz Wasiluk (BioAI)
Pedro Lucas (RITMO)
Shin Watanabe
Tom Frode Hansen (NGI)

Students: Bachelor ~200; Master: ~60
Robotics and Intelligent Systems program

Students hired on hourly basis

Visiting researchers

Robotics and Intelligent Systems (ROBIN) research group



Artificial Intelligence in industrial applications



Interactive music

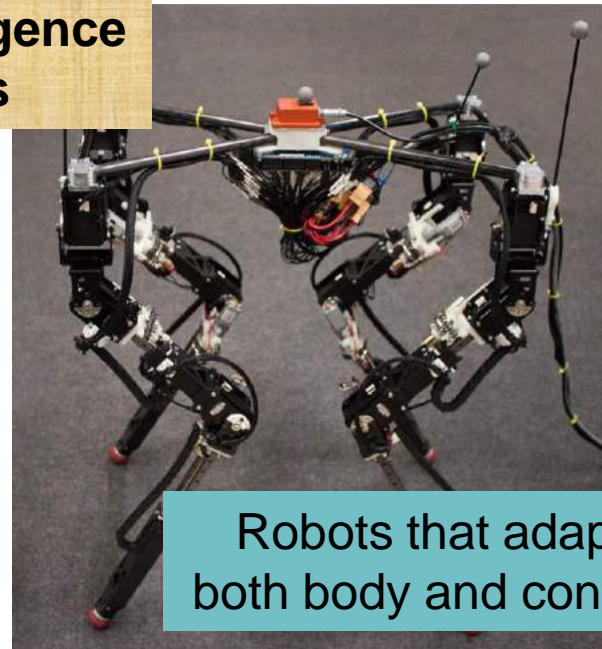
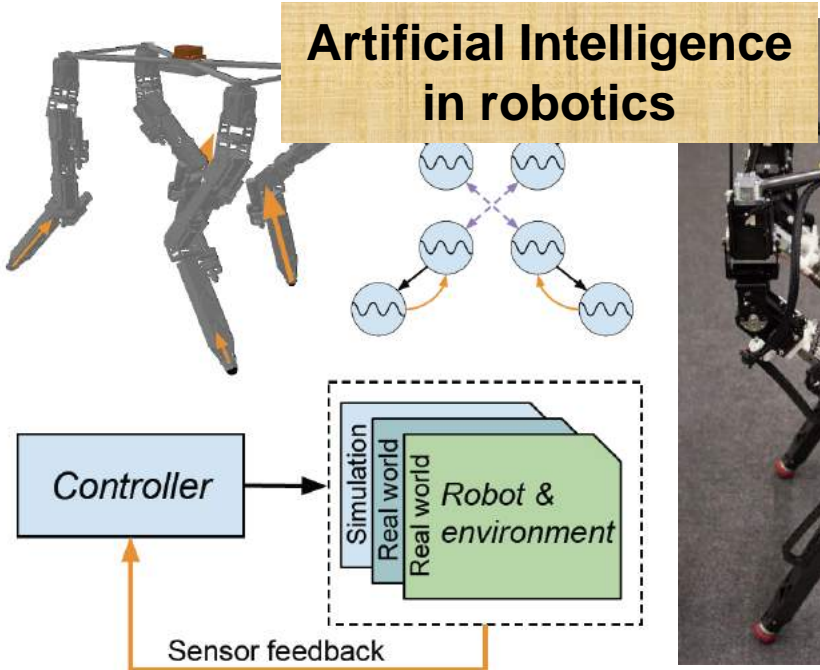


Adaptive and autonomous mental health treatment

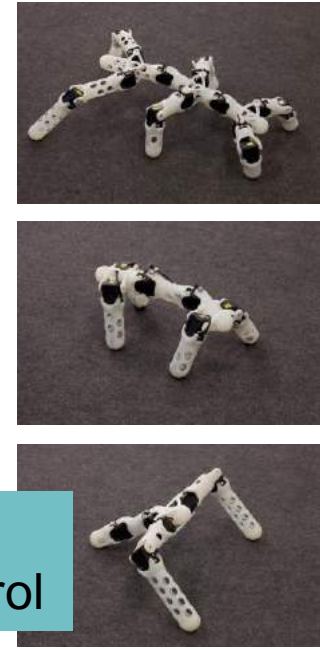
Artificial Intelligence in smartphones



Robots that look after and assist older people living at home



Robots that adapt both body and control



ROBIN Research Projects and Centre Funded by the Research Council of Norway

- **INtroducing personalized TReatment Of Mental health problems using Adaptive Technology** (INTROMAT, 2016-2021, LightHouse project)
- **Multi-sensor Elderly Care Systems/Robots** (MECS, 2015–2021, IKTPLUS)
- **Vulnerability in the Robot Society** (VIROS, 2019-2023, IKTPLUS)
- **Predictive and Intuitive Robot Companion** (PIRC, 2020-2027, IKTPLUS)
- **Centre of Excellence for Interdisciplinary Studies in Rhythm, Time and Motion** (RITMO, 2017-2027, CoE)



INTROMAT: INtroducing personalized TReatment Of Mental health problems using Adaptive Technology (2016-2021)

Research Council of Norway grant 259293



Goal: Increase access to **mental health** services for common mental health problems by developing **smartphone technology** which can **guide patients**.

<http://intromat.no>

Project Manager:

Haukeland Univ. Hospital, Bergen

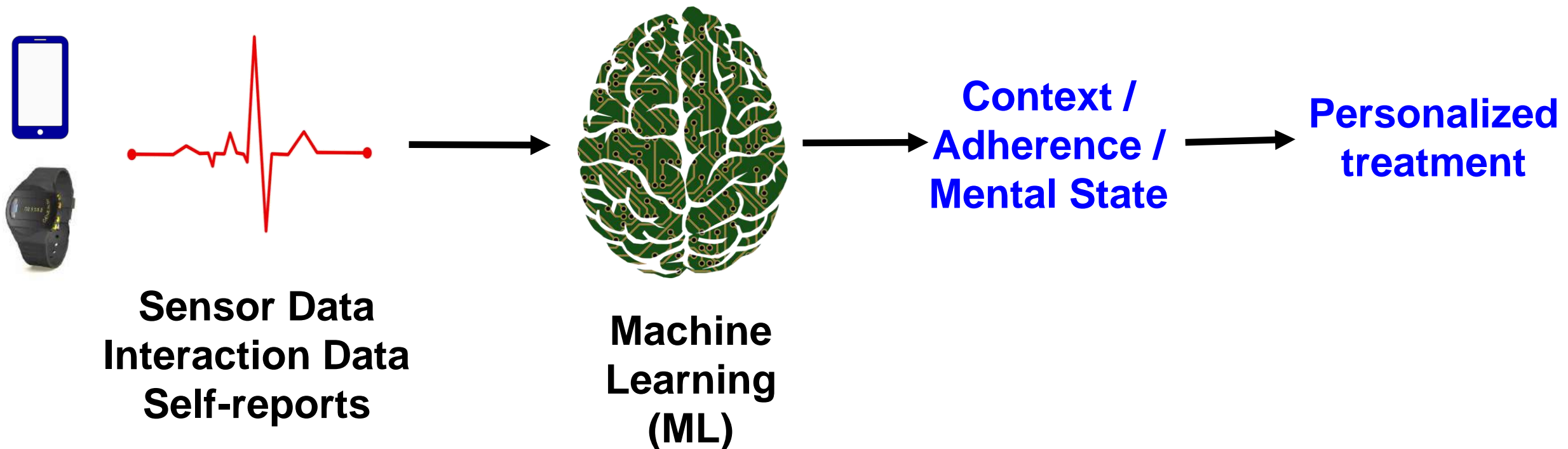
INTROMAT

Funding: IKTPLUS Lighthouse,
Research Council of Norway



The Research Council
of Norway

From data to Adapted and Personalized Treatment



Mental Health Research Challenges

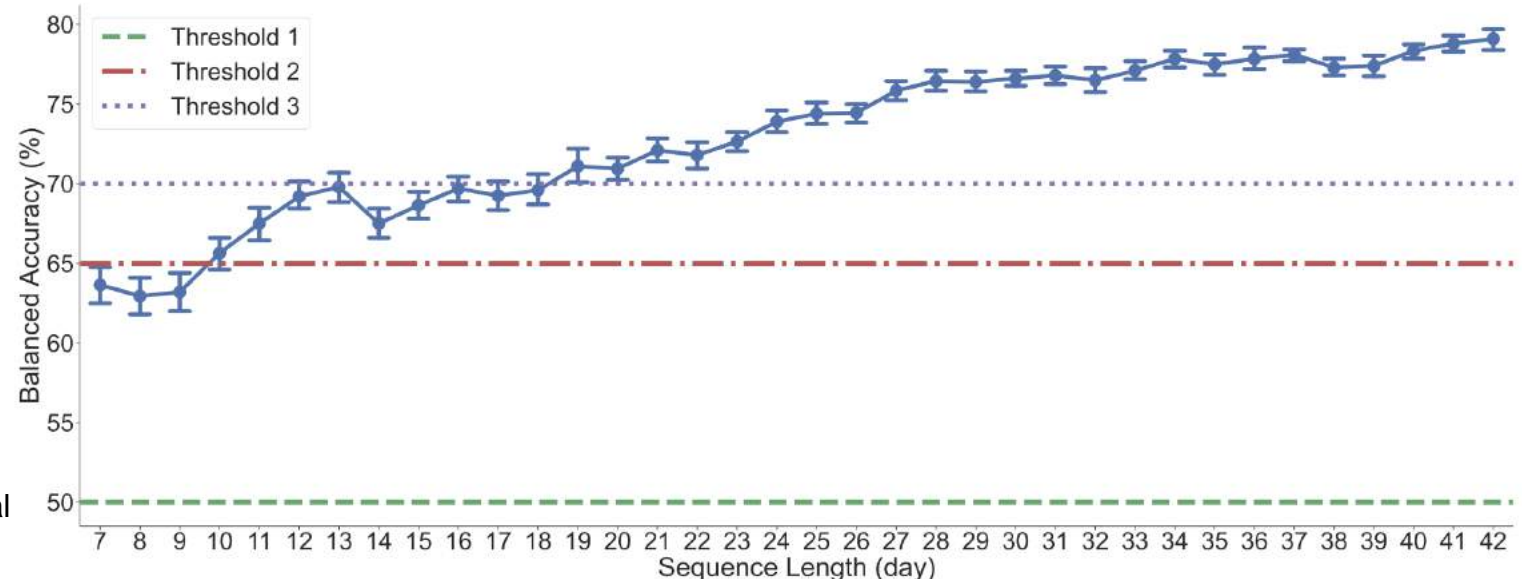
- **Access to data**
 - **Limited access** to relevant data from outside the project due to data protection rules. This also makes comparison of results difficult.
 - **Long approval process** before collecting own data.
- **Labeled data**
 - To make effective classification systems, the **quality of labeling** sensor data is crucial.
 - Frequent **self-reporting** of patients is time consuming and can make the patient less interested in participating in the study.

Adherence Forecasting for Internet-delivered Psychological Treatments (misc disorders)

- A dataset containing 342 patients undergoing guided internet-delivered cognitive behavioural therapy (G-ICBT) treatment.
- The proposed Self-Attention Network (deep learning) achieved over **70% average balanced accuracy**, when only **1/3 of the treatment duration had elapsed**.



Ulysse Côté-Allard, Minh H. Pham, Alexandra K. Schultz, Tine Nordgreen, Jim Torresen, *Adherence Forecasting for Guided Internet-Delivered Cognitive Behavioral Therapy: A Minimally Data-Sensitive Approach*, in journal in IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 6, pp. 2771-2781, June 2023, doi: 10.1109/JBHI.2022.3204737



ROBIN Research Projects and Centre Funded by the Research Council of Norway

- **INTroducing personalized TRreatment Of Mental health problems using Adaptive Technology** (INTROMAT, 2016-2021, LightHouse project)
- **Multi-sensor Elderly Care Systems/Robots** (MECS, 2015–2021, IKTPLUS)
- **Vulnerability in the Robot Society** (VIROS, 2019-2023, IKTPLUS)
- **Predictive and Intuitive Robot Companion** (PIRC, 2020-2025, IKTPLUS)
- **Centre of Excellence for Interdisciplinary Studies in Rhythm, Time and Motion** (RITMO, 2017-2027, CoE)



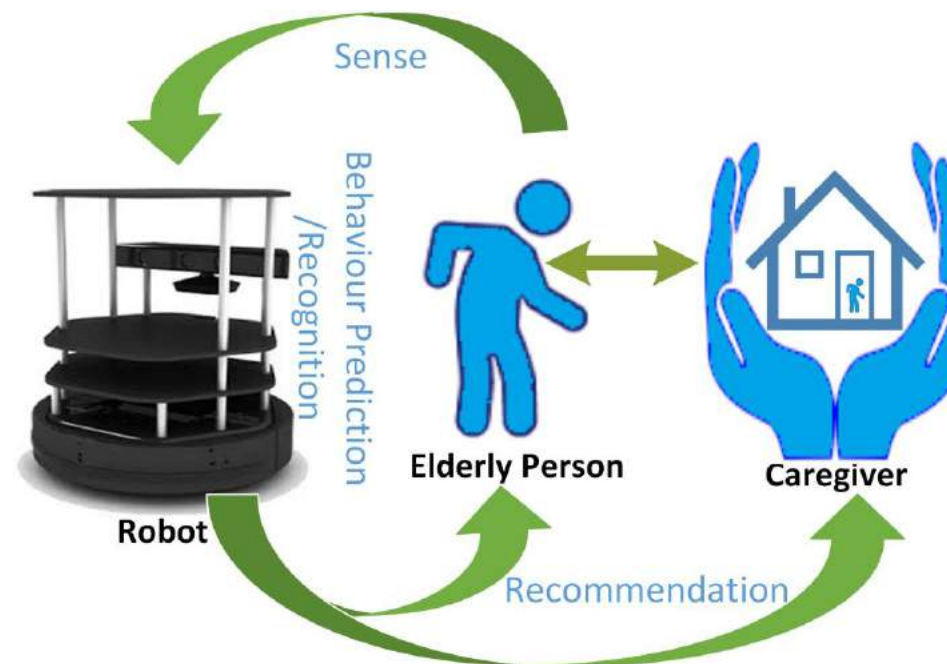


MECS: Multi-sensor Elderly Care Systems

Research Council of Norway grant 247697 (2015–2021)

<https://www.mn.uio.no/ifi/english/research/projects/mecs>

Goal: Create and evaluate multimodal mobile human supportive systems that are able to **sense, learn and predict future events.**

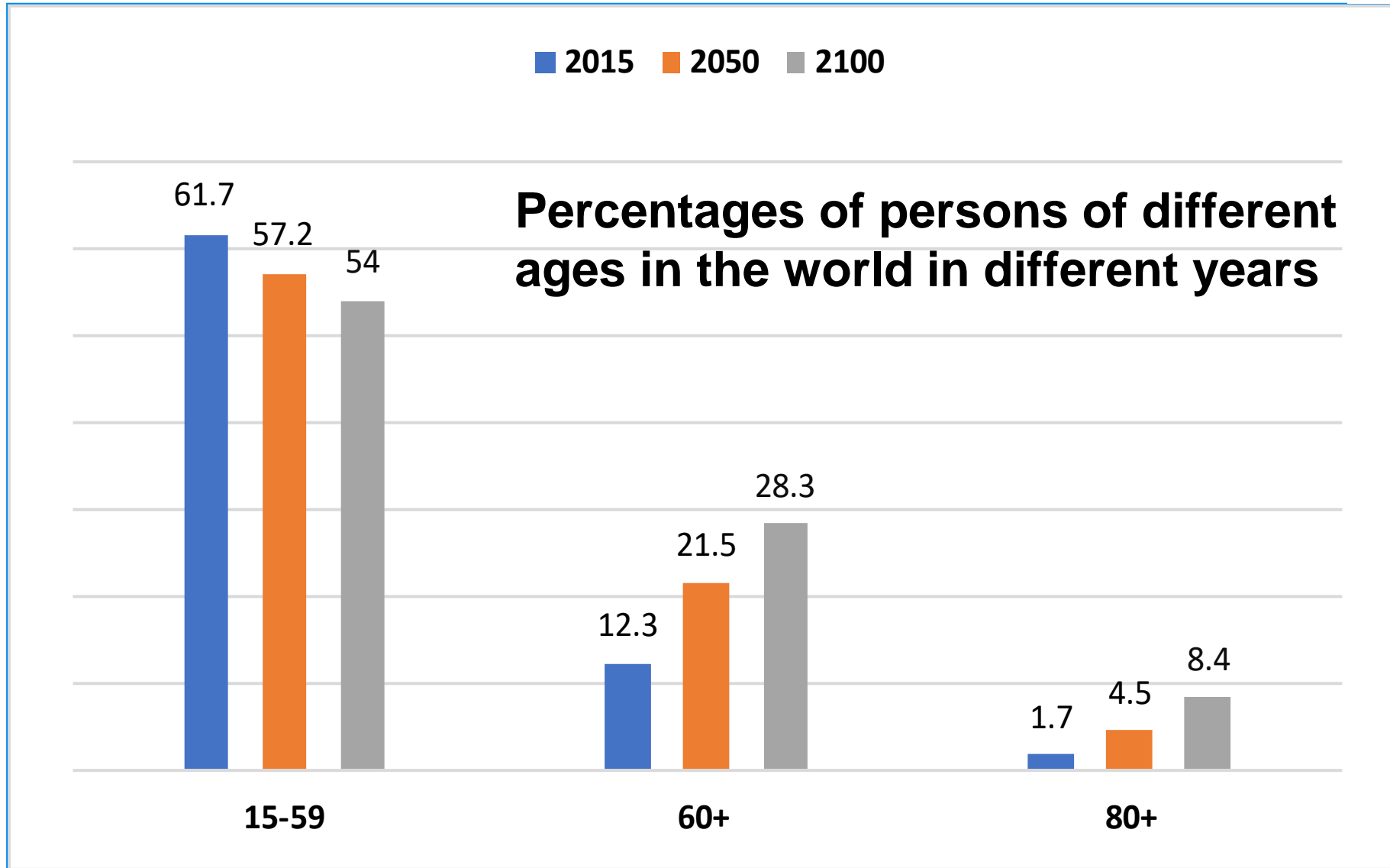


Funding: FRINATEK
Research Council of Norway

Web page: Google for "MECS IFI"



The Research Council
of Norway



United Nations (2015) World population ageing. United Nations, New York.

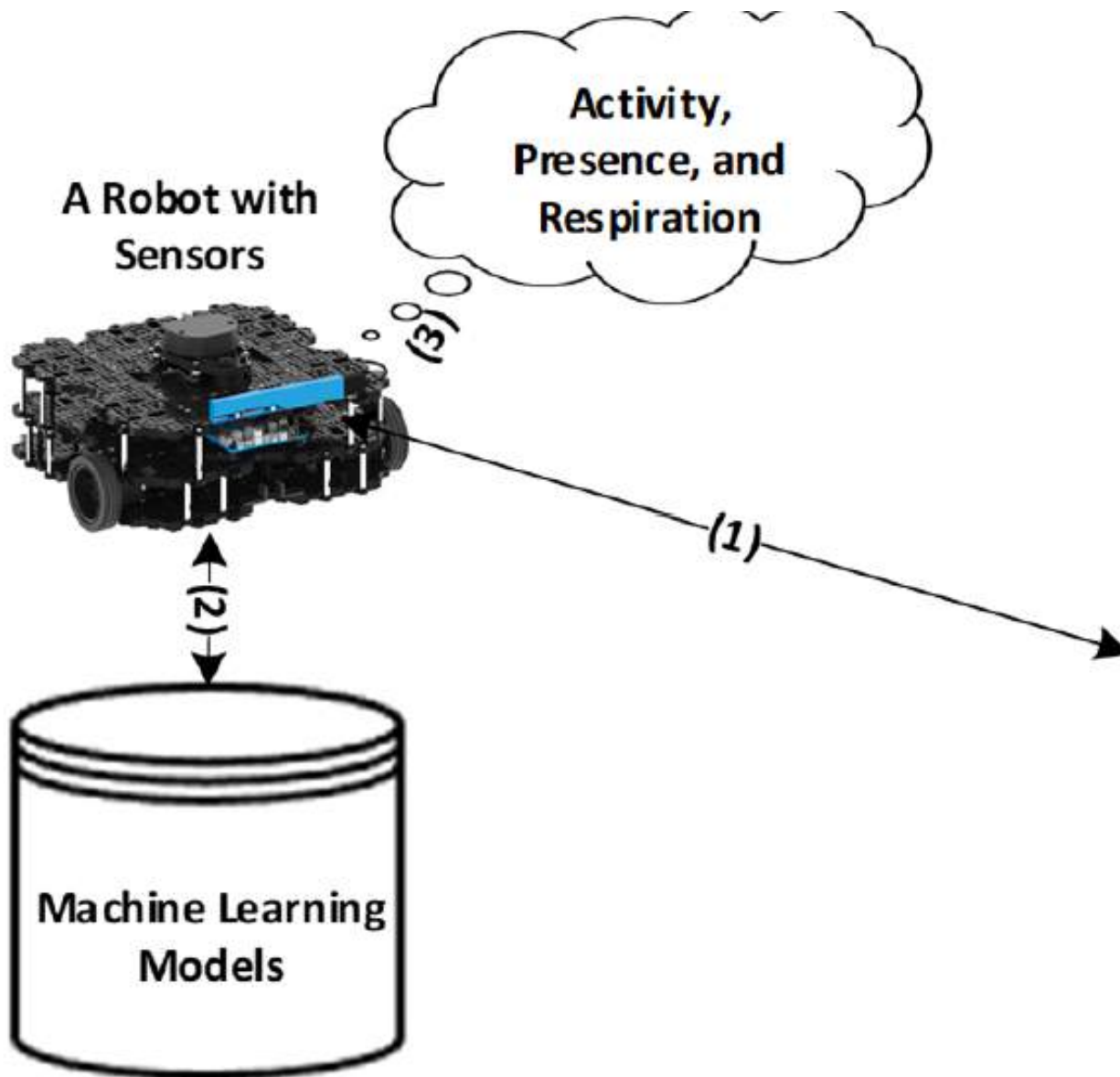
Robots and Older People

**Would we like to be surrounded
by robots rather than humans?**

Robots and Older People

Would we like – with some help from robots – to be independent with regards to our key needs like personal care, eating and transportation?

Myself: Yes, but my preference would be impacted by the ease of use and performance and to some extent the look of the robot.



MECS Research

Diana Saplacan
Rebekka Soma
Trenton Schulz



Apply sensors that provides non/less-intrusive sensing



Farzan M. Noori
Md. Zia Uddin



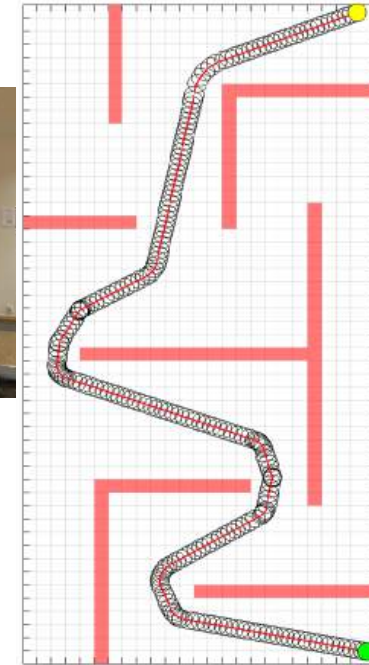
User needs and preferences

+ Master students

Robot sensing

Robot control

Weria Khaksar



Navigation without a map

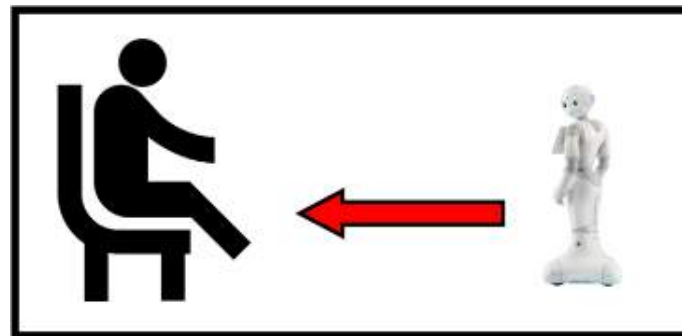


User Centered Design – Participatory Design

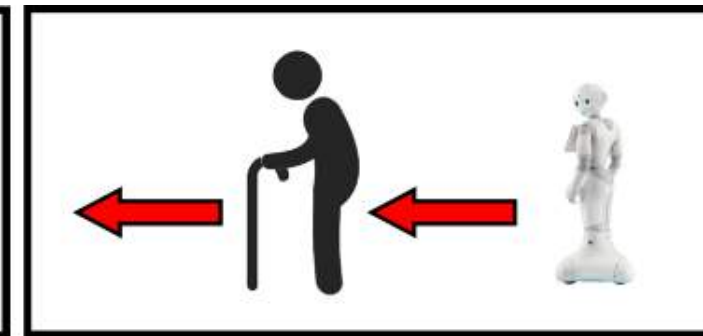
- involve real users in **actual use contexts** (home of older people)
- focus on behavior and **satisfying the needs** and desires of the users
- achieve improvements through **iterative testing and improvement**
- Oslo municipality elderly care facility: **Kampen Omsorg +**
- **Vitalis home** for elderly in Eindhoven in the Netherlands



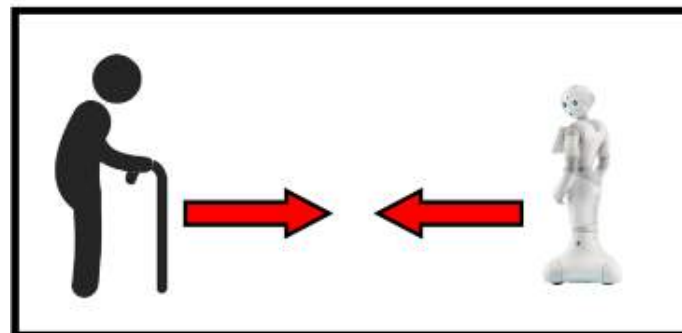
User studies on robot and user encounters



(a)

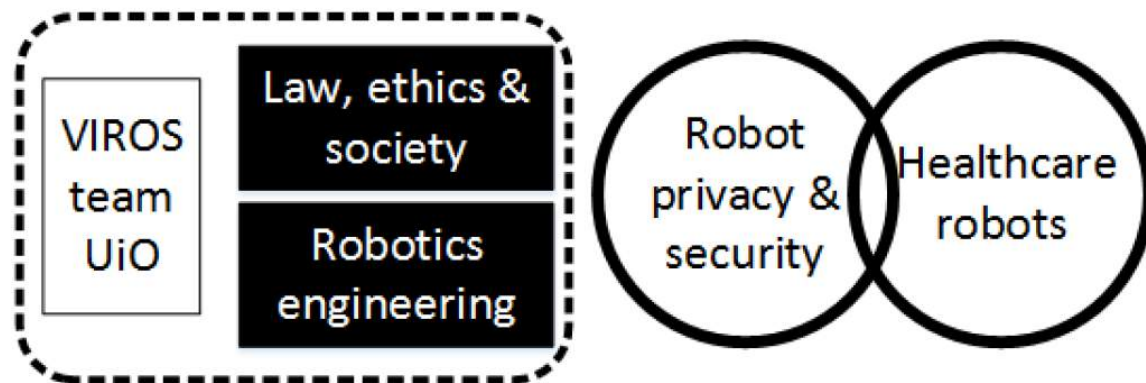


(b)



VIROS: Vulnerability in the Robot Society (2019-2023)

Research Council of Norway grant 288285



Dep. of Private Law (below)
Dep. of Informatics (left)



Goal:

Develop technology and proposals for regulatory measures to reduce vulnerabilities regarding robotics.

Focus on privacy, security and safety, particularly in healthcare contexts.

Technology partner: Robotics and Intelligent Systems (**ROBIN**) group

Diana Saplacan
Researcher



Adel Baselizadeh
PhD student



Funding: *IKTPLUS*, Research Council of Norway



The Research Council of Norway

<https://www.jus.uio.no/ifp/english/research/projects/nrccl/viros/index.html>



Ethical Concerns:

1. Privacy

2. Security

3. Safety



=



ROBOT

Sense

Think

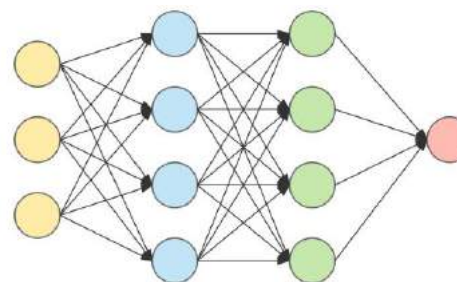
Act



Sensors

Artificial
Intelligence

Motors +
Mechanics

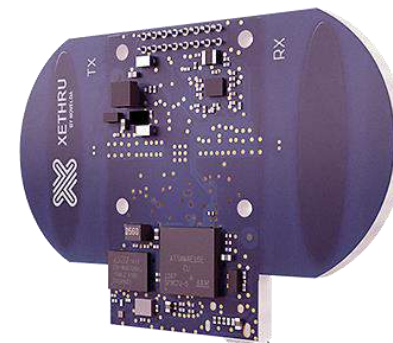




Ethical Concerns: 1. Privacy

- Challenge 1: Balance the **privacy of the elderly** against the **needs for data** collection for having an efficiently functioning elderly care systems.
- Challenge 2: Protection of sensitive data to **avoid unwanted distribution and misuse** of such data.
- Mitigation:
 - **Sensor type:** Use sensors collecting less privacy related information (person identifiable vs bio-signals/medical diagnostic)
 - **Sensor data processing:** Process data locally rather than sending sensor data over Internet

The MECS project has explored a number of novel non/less-intrusive sensors

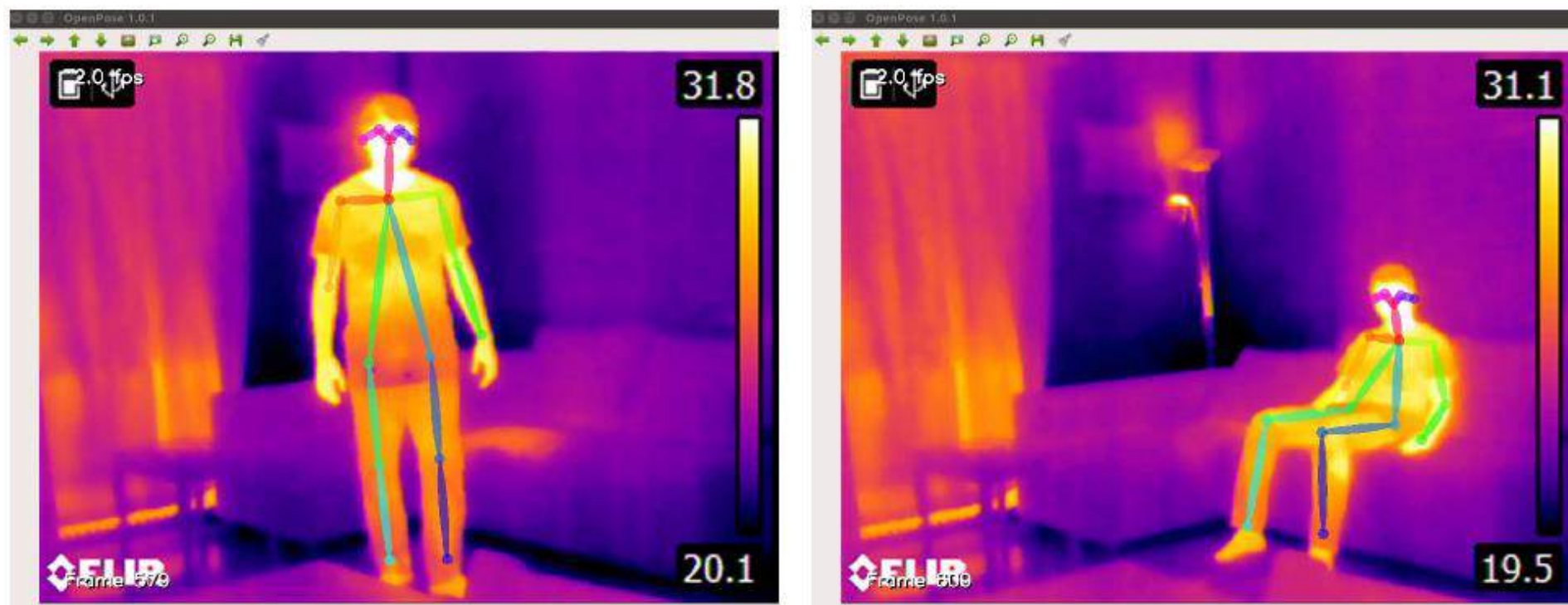


Novelda XeThru

FLIR
Thermal
Camera

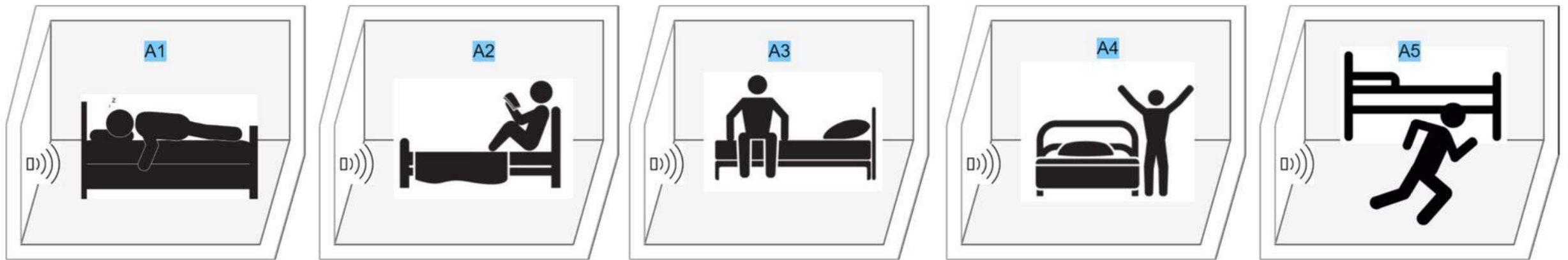


Non/less-Intrusive Sensing (thermal camera)



Md. Zia Uddin and Jim Torresen. A Deep Learning-Based Human Activity Recognition in Darkness, The 9th IEEE Colour and Visual Computing Symposium 2018 (CVCS 2018), Sept. 19-20, 2018

Ultra-Wideband (UWB) Radar-Based Activity Recognition



- **LSTM-based activity recognition** approach performed better than conventional approaches, with an **accuracy of 99.6%**.
- We applied 5-fold cross-validation to test our approach.

F. M. Noori, M. Z. Uddin and J. Torresen, "Ultra-Wideband Radar-Based Activity Recognition Using Deep Learning," in *IEEE Access*, vol. 9, pp. 138132-138143, 2021, doi: 10.1109/ACCESS.2021.3117667.



Ethical Concerns: 2. Security

- Concern 1: **Sensing** – possible theft and unwanted distribution of sensor data from a robot.
- Concern 2: **Control** – risk of misbehaviour of the robot in similar ways as computers can be attacked with malware.
- **Mitigation 1:** Regular security measures with **passwords and authentication**
- **Mitigation 2:** Add an **external user assessment** module that can consider the current context (ref. ethical reasoning engine)

Mutual trust cycles between a human and a machine

Do we now spend more effort protecting computing systems against unintentional attacks compared to recovering from software and hardware malfunction?

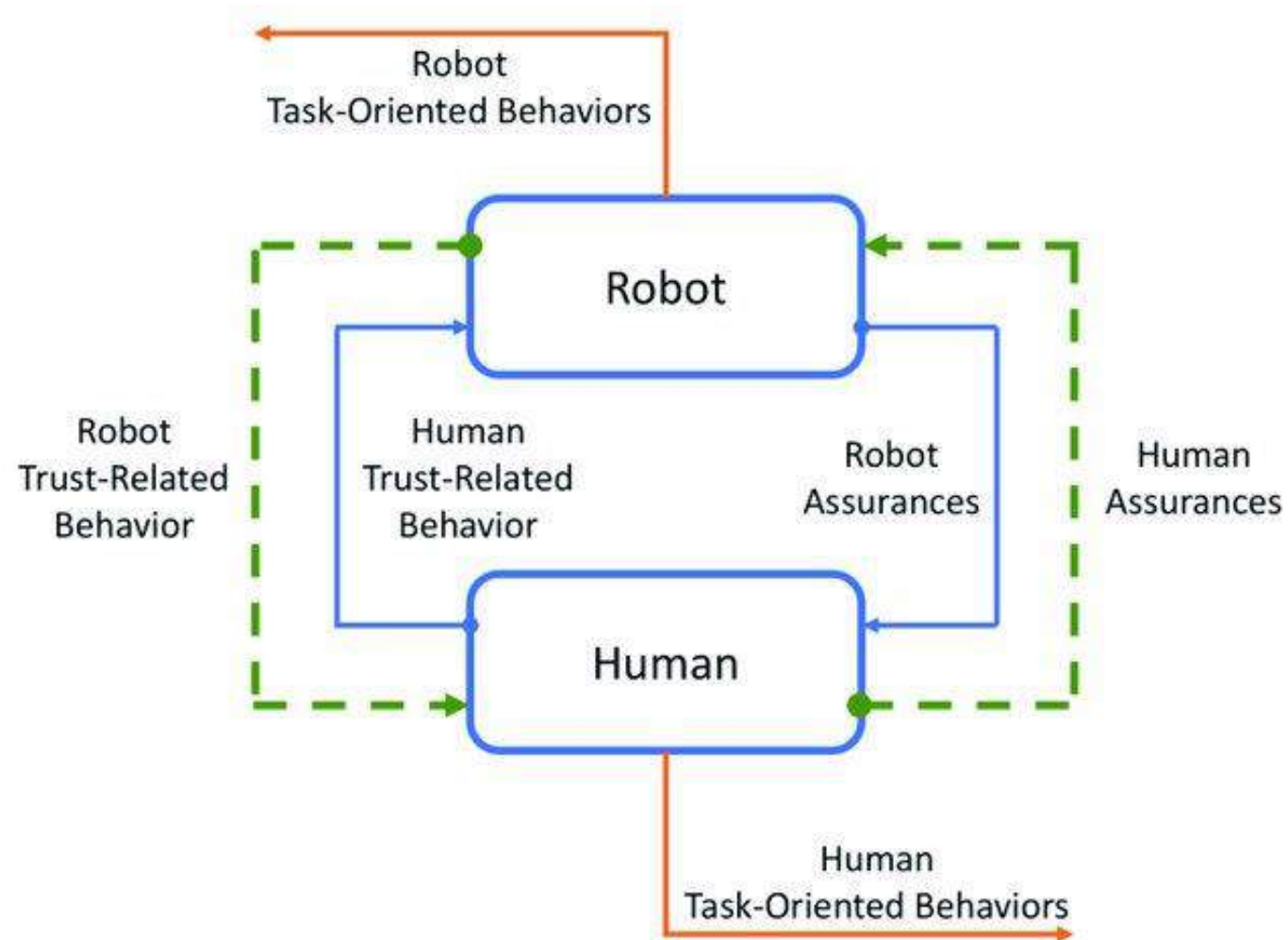


Figure available via license: CC BY 4.0



RITMO Centre of Excellence for Interdisciplinary Studies in Rhythm, Time and Motion grant 262762 (2017-2027)

- The center will study the **perceptual, cognitive** and **acting mechanisms** underlying our ability to experience rhythm and act rhythmically.
- Interdisciplinary **collaboration** between **musicology, psychology, computer science** and **robotics**.
- Machine learning and robotics to be applied

<https://www.uio.no/ritmo/english>



The Research Council
of Norway



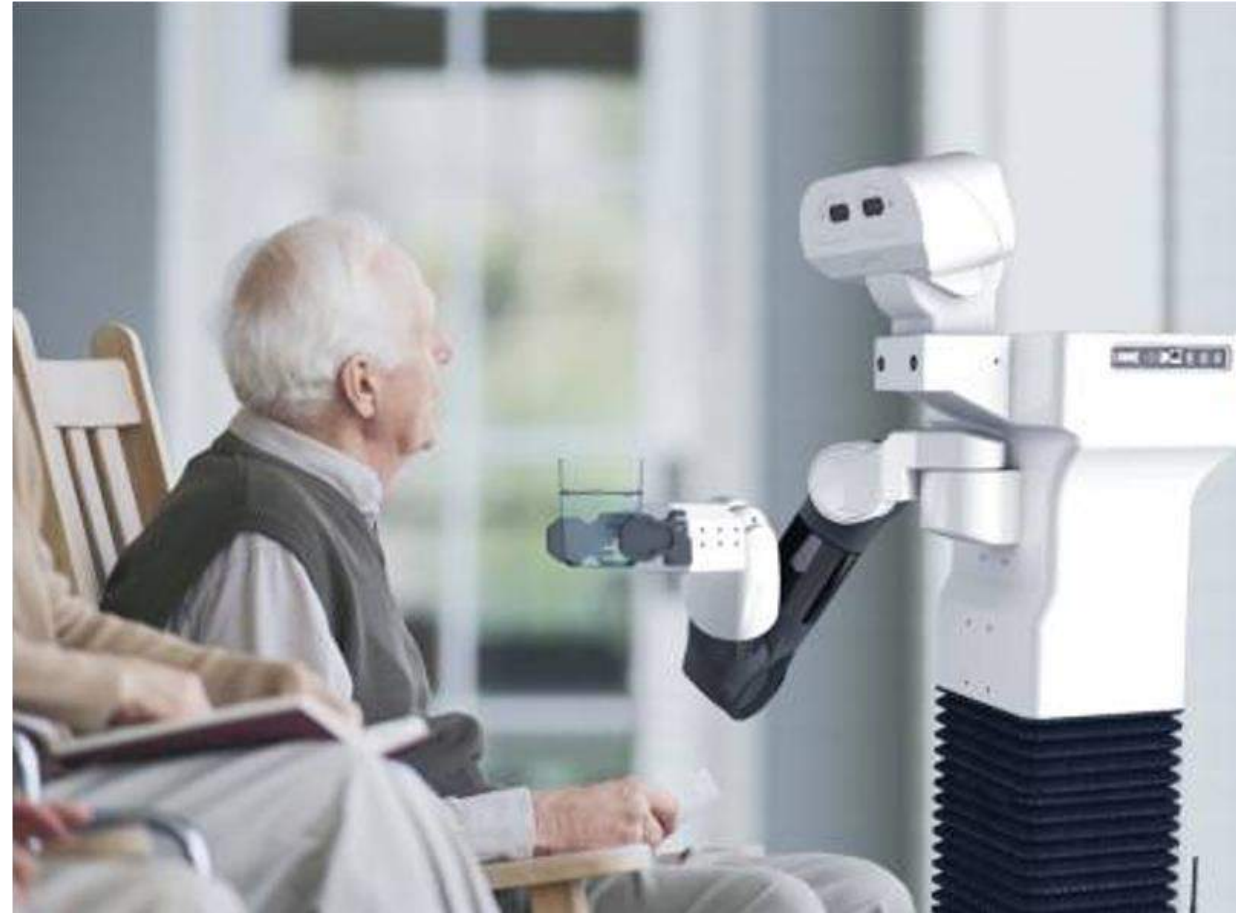
Norwegian
Centre of
Excellence



Human – Robot Interaction

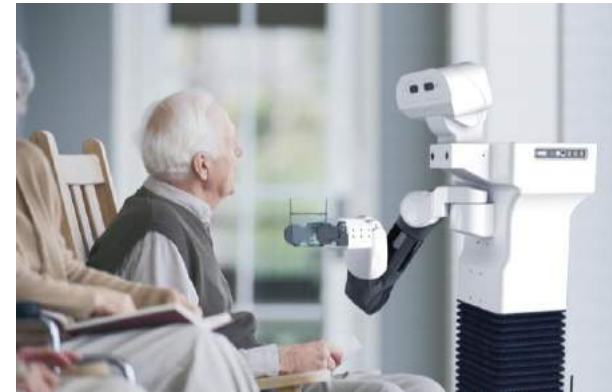
Slow Versus Safe Robot

- sloppy vs too slow
- must have general capabilities



Predictive and Intuitive Robot Companion (PIRC) (2020-2025)

Research Council of Norway grant 312333

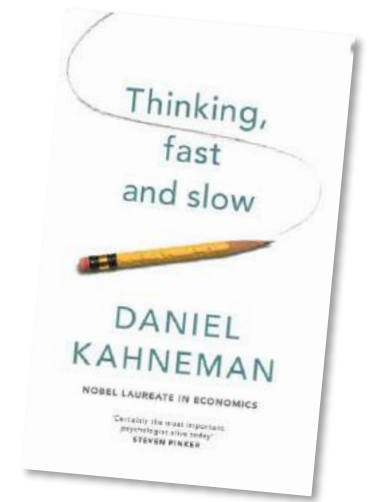


TIAGo
mobile robot
assistant



Goal: Build **models** that **forecast** future events and **respond dynamically by psychology-inspired computing:**

- Apply recent models of **human prediction** to perception-action loops of future intelligent robot companions.
- Include mechanisms for **adaptive response time** from quick and intuitive to slower and well-reasoned
- **Applications:** Physical rehabilitation and home care robot support for older people.



Funding: IKTPLUSS, Research
Council of Norway



The Research Council
of Norway

Future work and opportunities in ethics related research

- **privacy**
 - work with sensors collecting less privacy related data
 - work on algorithms for local /edge computing (rather than cloud computing)
- **security**
 - external user assessment module that can consider the current context
- **safety**
 - user-aware systems
 - explainable AI/transparent systems to be able to correct for unwanted or harming behavior
- **potential lack of contact with other humans**
 - **human dignity** impacted by our **independence**
 - politicians and society decide on the development as much as technology providers
 - researchers can contribute to a **balanced public discussion**

ACKNOWLEDGEMENTS

This tutorial is partly supported by the **Research Council of Norway** through following projects/centre:

- Vulnerability in the Robot Society (**VIROS**) (2019-2024) grant 288285
- Predictive and Intuitive Robot Companion (**PIRC**) (2020-2025) – grant 312333
- Collaboration on Intelligent Machines Norway (**COINMAC**) – grant 309869
- **RITMO** Centre of Excellence for Interdisciplinary Studies in Rhythm, Time and Motion – grant 262762 (2017-2027)



Collaboration on Intelligent Machines Norway (COINMAC)

↔ US, Brazil and Japan

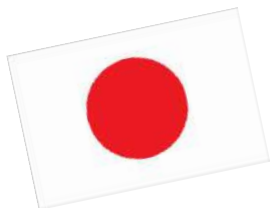
Project manager: Jim Torresen, Research Council of Norway grant 309869



- Short term mobility stay for meetings
- Sharing and development of curriculum and teaching material for courses



- Long term mobility stay (students)
- Intensive course/student workshop
- Workshop/conference org. activities



- New collaborations/guest lectures (can be in Canada, South-Korea, ++)



The Research Council
of Norway

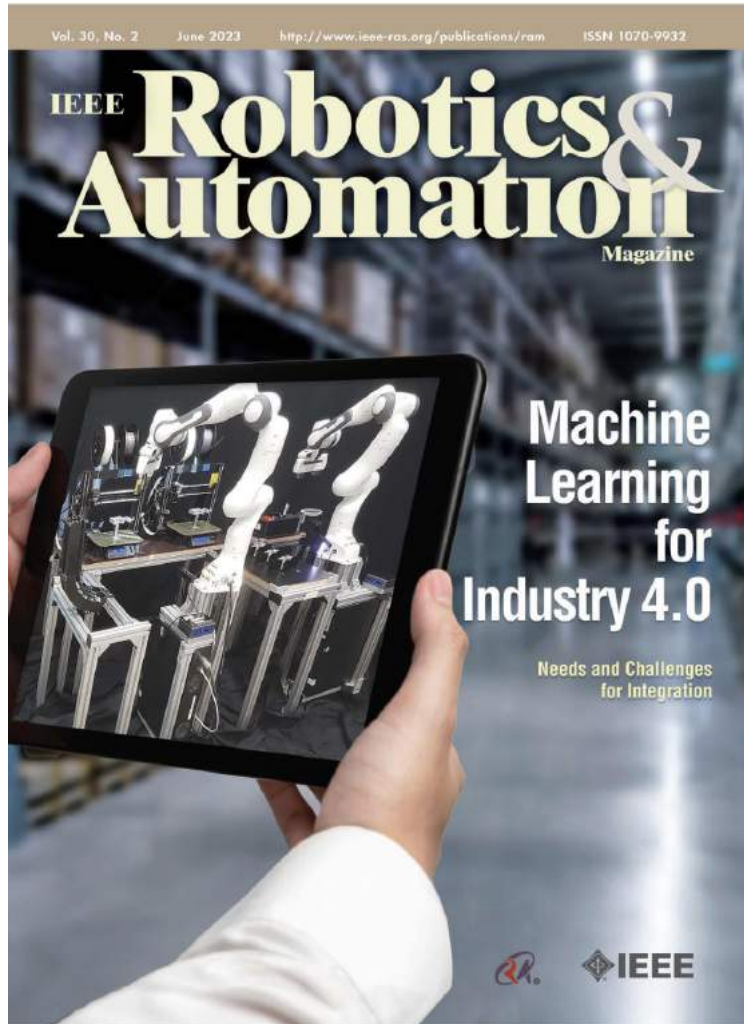
Funding: *INTPART 2017-2025*

Diku

Norwegian Agency for
International Cooperation
and Quality Enhancer
in Higher Education



Future work



**Special issue on Robot Ethics –
Ethical, Legal and User Perspectives
in the Development and Application
of Robotics and Automation**

Paper deadline: late 2024



We should focus as least as much on improved quality of life as reducing the cost by the technology being developed

Questions or Comments?

Make contact: jimtoer@ifi.uio.no

www.jimtoer.no