Karlsruhe Institute of Technology

Robotics: Ethical Issues

Lecture: Robotics I
05.02.2025
PD Dr. Alexander Bagattini and Désirée Martin

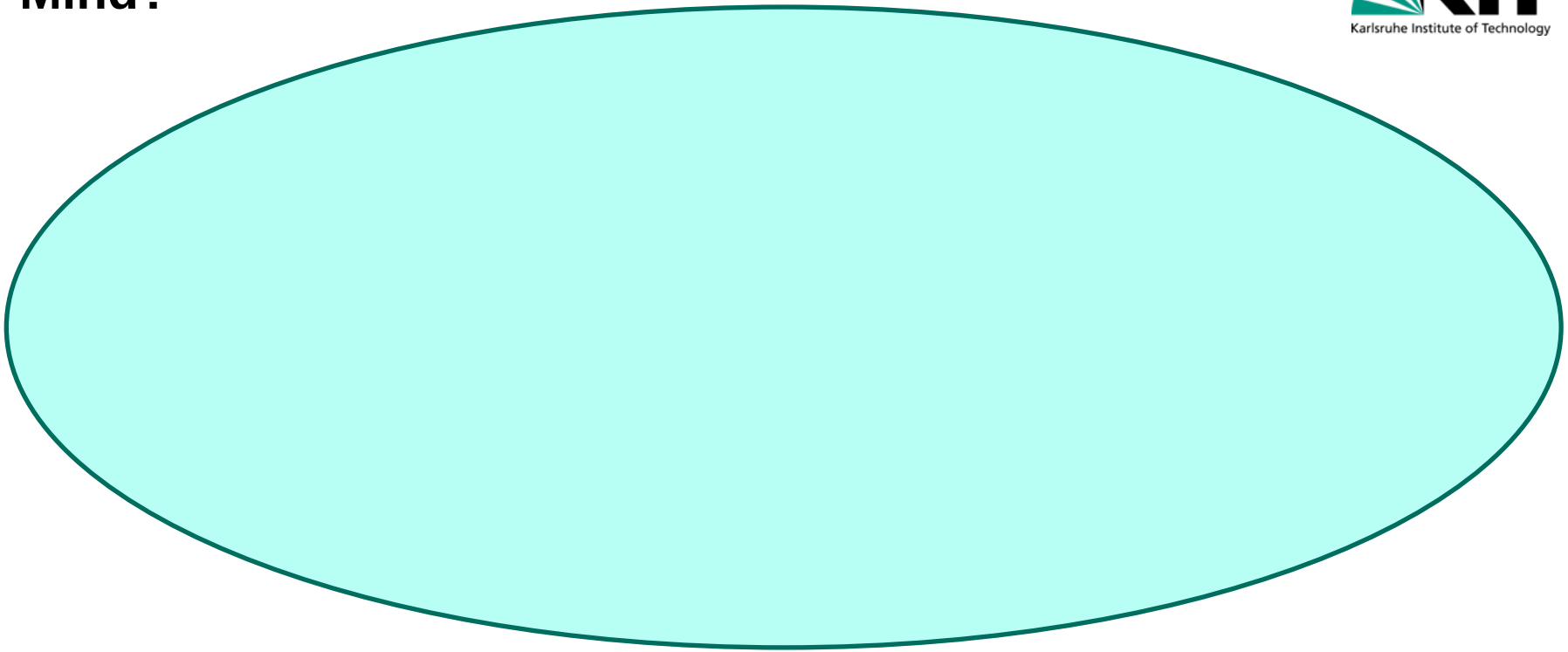
Academy for Responsible Research, Teaching, and Innovation

ARRTI

Warm-Up

- Participation in EduVote is possible anonymously and without registration
- You can either participate via the QR code (right).
- Or you can go to www.vote.ac and enter the ID a.bagattini@kit.edu in the ID field
- Please vote as soon as the survey starts

Humans and Robots: Which Movies Come to your Mind?



ID = a.bagattini@kit.edu
Click to start poll

Content of Today‘s Lecture

1. Introduction
2. Autonomy and Control
3. Human-Robot-Interaction
4. A practical example: Ethical Issues Concerning the Opportunities and Risks of AI-Assisted Robotics for Radiological Hazards
5. Open Discussion

1. Introduction

- Ethics: some basics
- Ethics and Robots

Ethics: some basics

- Ethics: what is the right thing to do?
- Important: moral judgements are often controversial
- Because normative concepts are ambigious: equality, autonomy, dignity, responsibility
- Ethics: rational discussion by using methods like analysis of moral concepts, ethical theory, case discussion

Moral Judgements	Non-Moral Judgements
Prescriptive Men and women should earn the same salary.	**Descriptive** x% of the population of country y believe that men and women should earn the same salary.
Universal Parents are (means: have a right to be) responsible for their children.	**Local** According to German law: parents have an obligation an a right to care for their offspring.

Hare 1952

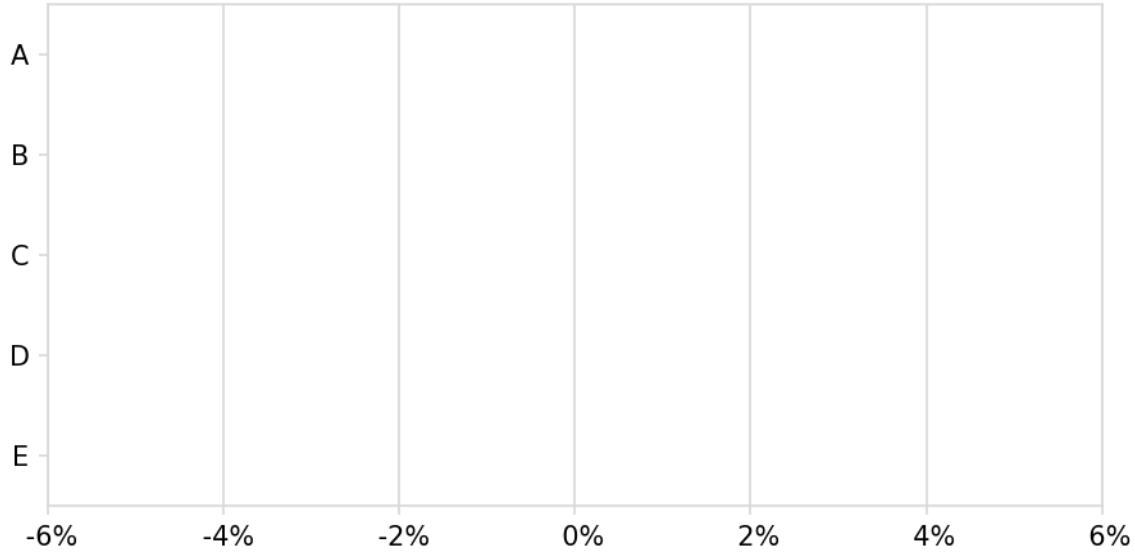
EduVote

- Participation in EduVote is possible anonymously and without registration
- You can either participate via the QR code (right).
- Or you can go to www.vote.ac and enter the ID a.bagattini@kit.edu in the ID field
- Please vote as soon as the survey starts

EduVote

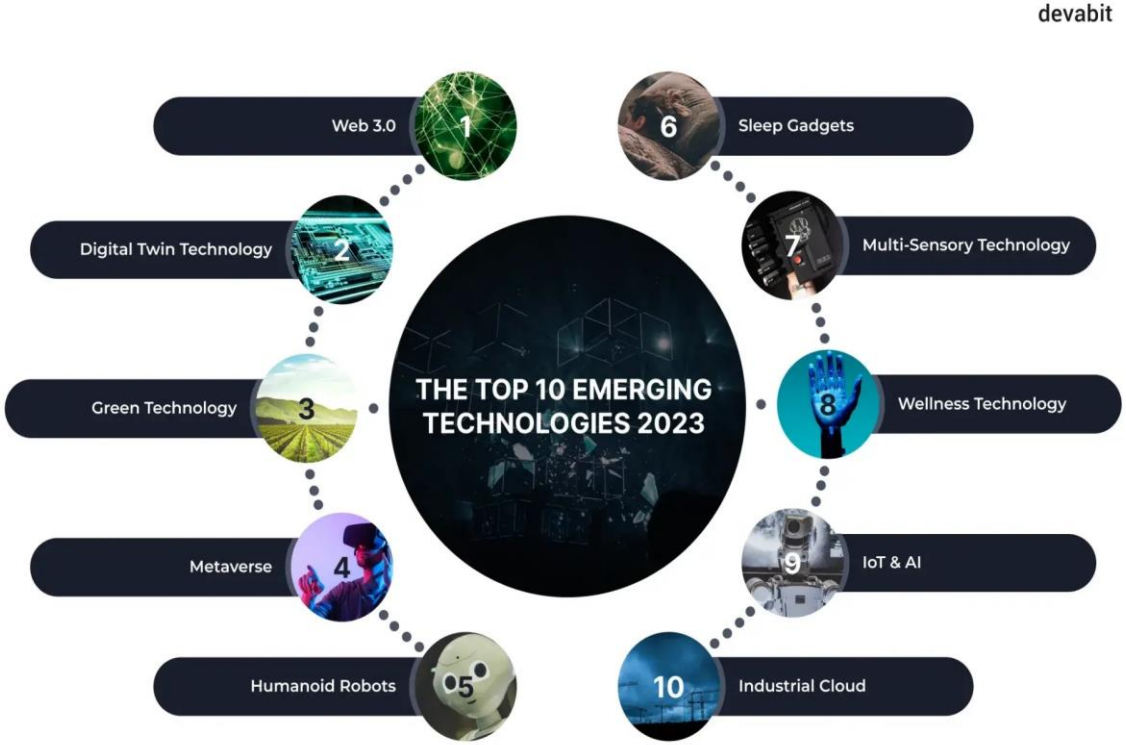
Which is/are moral judgements?

A) Lying in court is punishable.
B) Most people believe that lying is wrong.
C) One ought to tell the truth, even if it is difficult.
D) Sometimes people are motivated to tell the truth.
E) A white lie can be justified to safe peoples lives.



ID = a.bagattini@kit.edu
Click to start poll

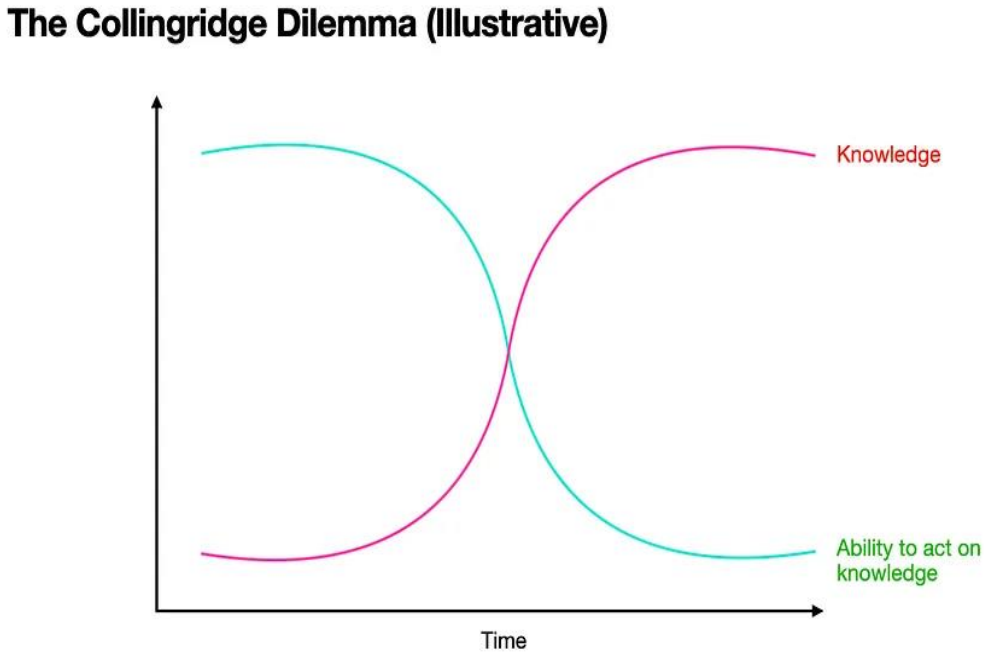
Ethics and Robotics



Some Ethical Issues in Robotics

- Autonomy and control
- Responsibility
- Moral Status of robots
- Human-robot-interaction
- Safety and risk
- Bias and fairness
- Privacy and surveillence

Ethics and Robotics: The Collingridge Dilemma



Collingridge 1982

It is important to consider ethical, social and legal issues already at the beginning of the development of new technologies!

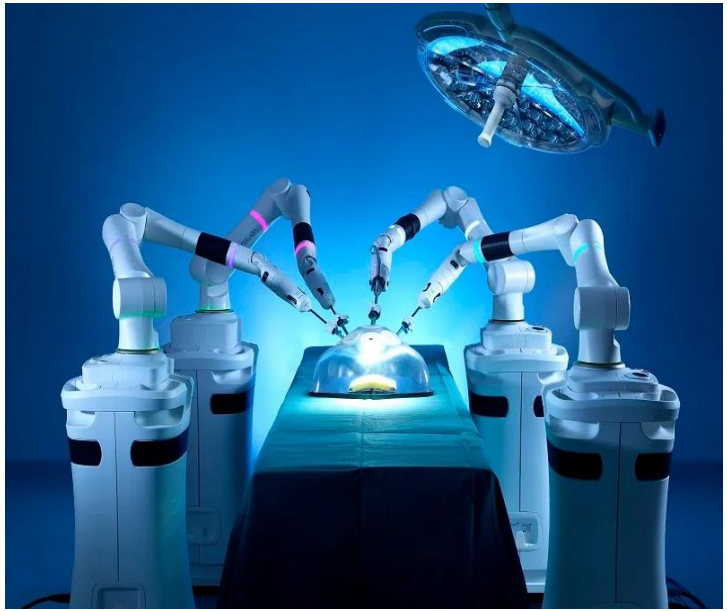
Content of Today‘s Lecture

1. Introduction
2. Autonomy and Control
3. Human-Robot-Interaction
4. A practical example: Ethical Issues Concerning the Opportunities and Risks of AI-Assisted Robotics for Radiological Hazards
5. Open Discussion

2. Autonomy and Control

- Robotics is a subset of automation that specifically involves physical machines (robots) that can perform tasks autonomously or semi-autonomously.
- Many modern robots use AI and machine learning to improve their automation capabilities (e.g., self-learning warehouse robots).

2. Autonomy and Control

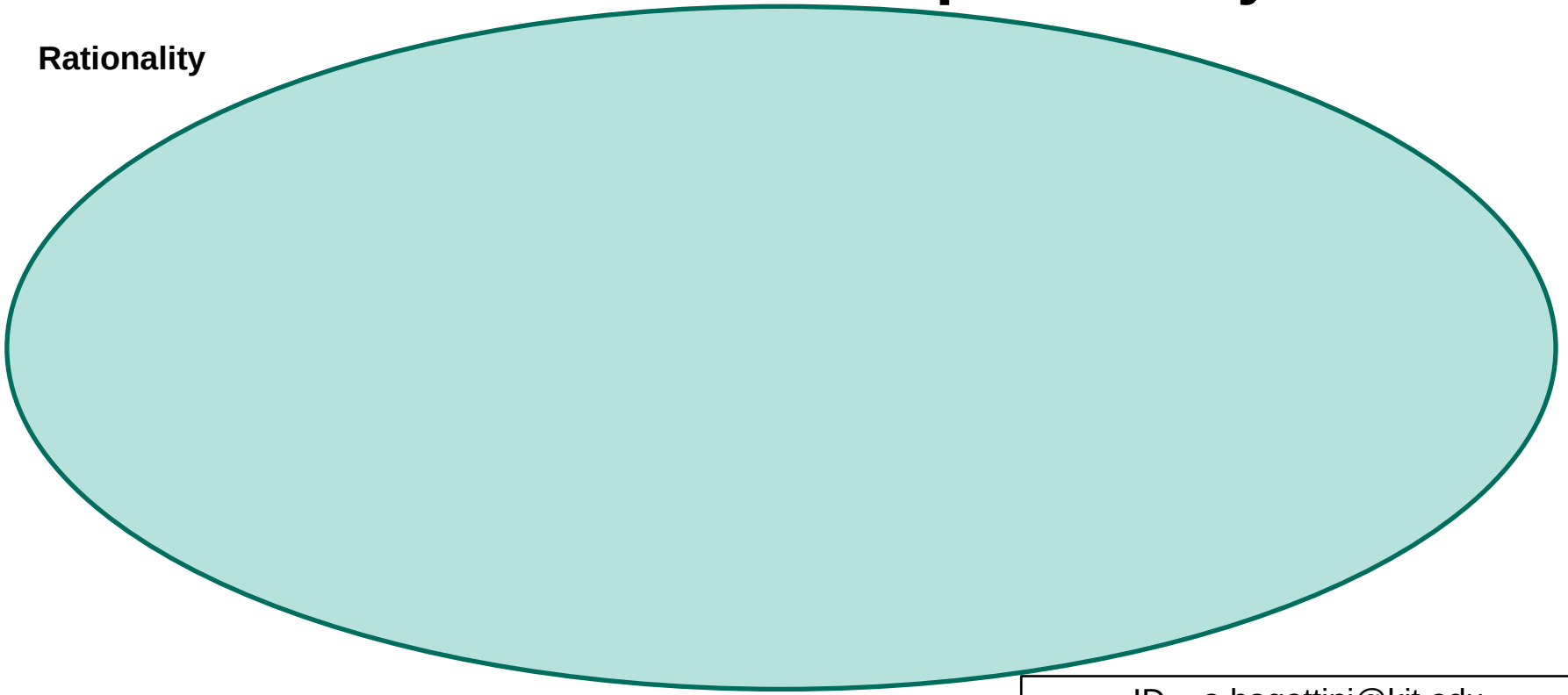


- Example: The da Vinci Surgical System, an AI-assisted robotic surgeon.
- Why Human Control is Necessary:
  - Prevent misdiagnoses and surgical errors.
  - Ensure that the human doctor makes final decisions in life-critical procedures.
  - Maintain patient consent and ethical medical practices.
- Ethical Issue: Who is responsible for robot actions?

EduVote: What is Moral Responsibility?

- Participation in EduVote is possible anonymously and without registration
- You can either participate via the QR code (right).
- Or you can go to www.vote.ac and enter the ID a.bagattini@kit.edu in the ID field
- Please vote as soon as the survey starts

EduVote: What is Moral Responsibility?



ID = a.bagattini@kit.edu
Click to start poll

An Account of Moral Responsibility

Moral Responsibility as Guidance Control (Fischer and Ravizza 1998)

Reason-Responsive Mechanism

- A acts according to moral reasons that align with A's values.
- When A acts, he or she does so with an awareness of the reasons behind their decisions, allowing for thoughtful and principled action.

Ownership Condition

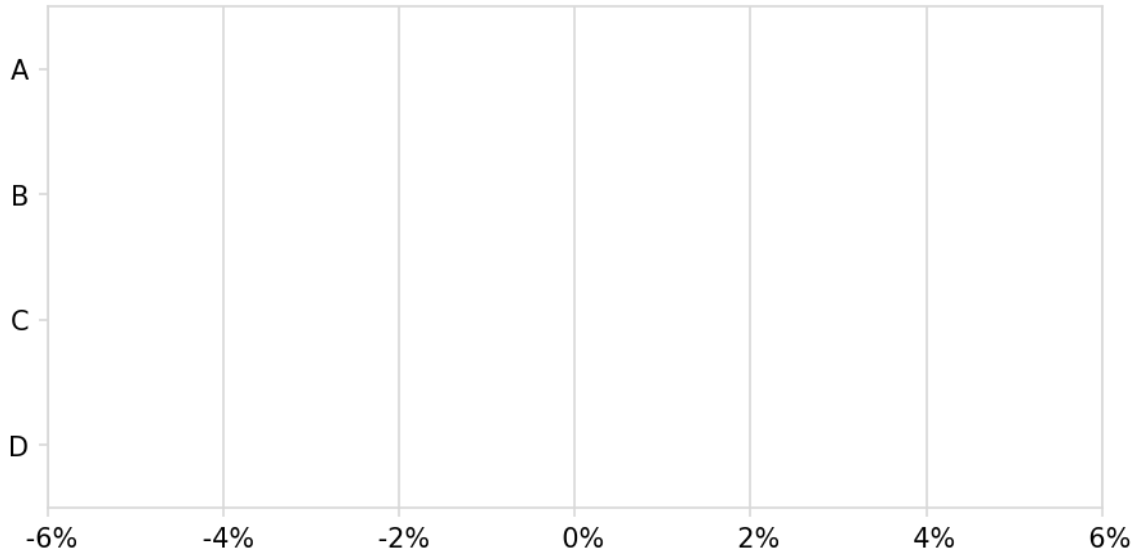
- A must recognize it's reasons as A's own.
- If A's actions are the result of manipulation or external influence (such as psychological coercion), guidance control may not hold, as A may not genuinely own those actions.

EduVote: Who is responsible for his action?

- Participation in EduVote is possible anonymously and without registration
- You can either participate via the QR code (right).
- Or you can go to www.vote.ac and enter the ID a.bagattini@kit.edu in the ID field
- Please vote as soon as the survey starts

EduVote: Who has guidance control over his action?

1. A votes for a party because his parents did.
2. A has been influenced by Social Media and phishing emails in his vote.
3. A was subject to political indoctrination.
4. A's IQ is below average.



ID = a.bagattini@kit.edu
Click to start poll

Autonomous Systems: A Challenge for Moral Responsibility (as Guidance Control)

- Since AI systems operate with a degree of autonomy and unpredictability, traditional accountability models like guidance control struggle to determine who (or what) is responsible when something goes wrong.
  - **Unpredictability & Black Box Decision-Making**
    - Many AI models (especially deep learning systems) operate as black boxes, making decisions that even their creators cannot fully explain. (Responsibility gaps)
    - If an AI system makes an unethical or biased decision (e.g., discriminatory hiring algorithms), it's unclear who should be held liable.
    - If an AI system makes a harmful decision, should responsibility fall on the developer, manufacturer, user, or regulator?

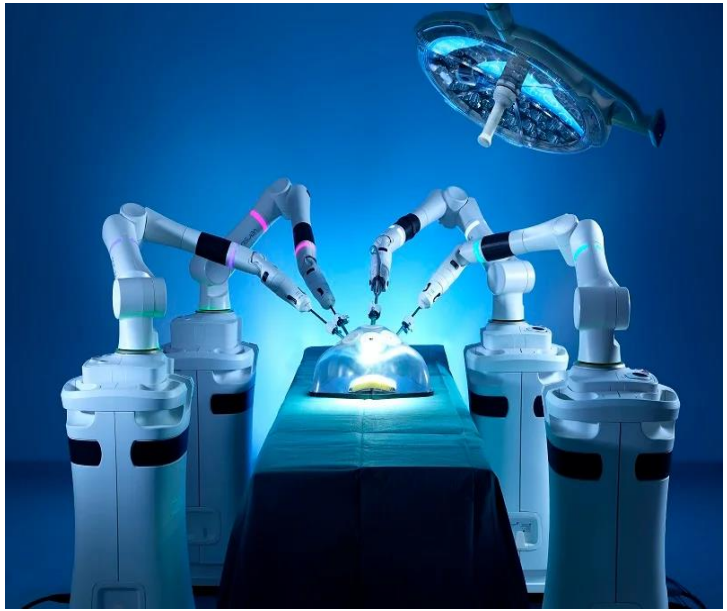
How Can Human control be Maintained?

Meaningful Human Control (Santoni de Sio/ van den Hoeven 2018)

Tracking Condition (TkC): The autonomous system can respond to relevant moral reasons (as understood by humans), as well as the pertinent facts of the environment in which it operates.

Tracing Condition (TrC): The outcomes of the system's operations should be tracable to at least one human involved in the design or operation chain. (Accountability)

Application Case: The da Vinci Surgical System



- TkC (reason responsiveness): System not only reacts to surgeon‘s commands but also to real-time feedback from the surgical environment (context sensitivity).
- TrC (accountability): it is possible to identify who made decisions throughout the surgical process (data collected)
- TkC (reason responsivenes): robot operates without human oversight (autonomously), lack of familiarity with system, inflexibility
- TrC (accountability): absense of accountability (fully autonomous system), ambigious accountability (e.g. overlapping roles), lack of documentation

Critical Discussion

1. How feasible is it to achieve the tracking and tracing conditions for meaningful human control in “messy” real-world situations?
2. What challenges do we face in ensuring transparency and accountability in these systems, and what potential solutions could address these challenges?
3. How should we navigate situations where human control is diminished, yet the system's actions have significant ethical consequences?

Content of Today‘s Lecture

1. Introduction
2. Autonomy and Control
3. Human-Robot-Interaction
4. A practical example: Ethical Issues Concerning the Opportunities and Risks of AI-Assisted Robotics for Radiological Hazards
5. Open Discussion

3. Human Robot Interaction

- **Can Robots Have Moral Status?** Experts and the media discussed whether humanoid robots like Sophia should at some point acquire moral status or even a certain legal status. Among other things, this involved questions of autonomy, responsibility and the definition of “personhood”.
- **Deception in Social Robotics**: Ethical questions arise when robots are designed to appear human-like or act empathetically, leading users to falsely trust them as moral agents.

Can Robots Have Moral Status?

Robot Sophia
© Hanson Robotics

- Advanced Facial Expressions and Animations
- Natural Language Processing and Conversation Skills
- Facial Recognition and Emotional Analysis
- Machine Learning Capabilities
- Social Engagement Abilities

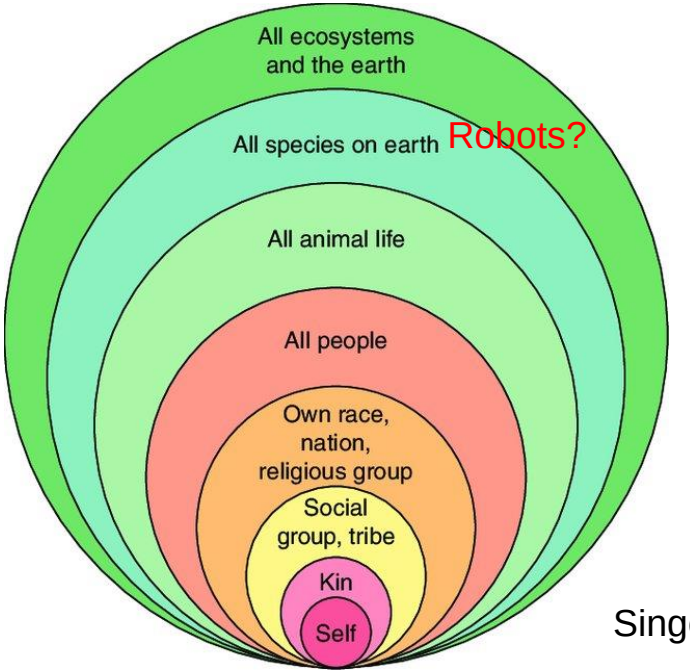
- In 2017, Sophia was granted symbolic citizenship of Saudi Arabia.
- Experts and media debated whether humanoid robots like Sophia should eventually gain some moral status or even rights.

What Does it Mean to Have Moral Status?

A has moral Status:

- A has intrinsic value.
- A's interest enjoy (equal) moral considerability.
- Other's have moral obligations to treat A with care and respect.

The Expanding Circle of Equality



Singer 1981

An Argument for the Moral Status of Robots

A problem for ascribing moral status to robots

- Many criteria for moral status are epistemically or metaphysically contested

- Autonomy/ free will
- Personal Identity
- Emotional needs
- Preferences

Danahars Argument from Ethical Behaviorism

(1) If a robot is roughly *performatively equivalent* to another entity whom, it is widely agreed, has significant moral status, then it is right and proper to afford the robot that same status.

(2) Robots can be roughly *performatively equivalent* to other entities whom, it is widely agreed, have significant moral status.

(3) Therefore, it can be right and proper to afford robots significant moral status.

Danahar 2019

Implications of the Argument

Behavioral Patterns as Grounds for Moral Status (Ethical Behaviorism)

- The assignment of moral status to entities is often based on specific behavioral patterns—like displaying anguish when in pain.
- In such cases our ethical imperatives are determined by observable behavior rather than what's happening "on the inside".

Performative Equivalence

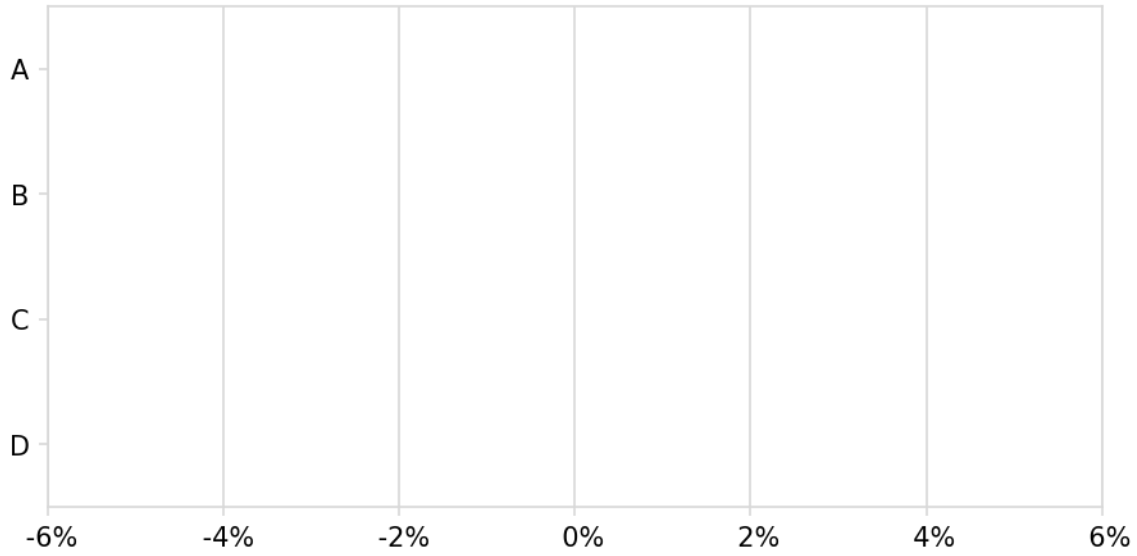
- Principle: Equal cases should be treated equally.
- If a robot exhibits behaviors that are virtually indistinguishable from those of another entity that is already granted moral status, then ethical consistency dictates that the robot must be granted that same status as well.

EduVote Danahr‘s Argument

- Participation in EduVote is possible anonymously and without registration
- You can either participate via the QR code (right).
- Or you can go to www.vote.ac and enter the ID a.bagattini@kit.edu in the ID field
- Please vote as soon as the survey starts

EduVote Danahar‘s Argument

A. Is the Argument valid?
B. Is the Argument sound?
C. If not-B: would you deny premise 1?
D. If non-B: would you deny premise 2?



ID = a.bagattini@kit.edu
Click to start poll

Forced Opinion Game

Team Defense

- Argue why EB provides a compelling framework for assessing the moral status of robots, emphasizing the importance of observable behavior over metaphysical speculation.
- Highlight the real-world applications of EB in robotics, showing how it can guide ethical treatment and policymaking concerning robots.

Team Offense

- Argue that EB's exclusive focus on observable behavior ignores important factors about beings with morals status. Which factors could that be?
- Highlight dangers of misattributing human-like properties to robots.

Team “Out of the Box”

- Imagine potential future scenarios where advanced robots exhibit behaviors that complicate the ethical consideration of moral status, proposing thought experiments that could stretch current thinking.

Critique of Danahar‘s Argument

1. Relevance of underlying ontological or metaphysical properties
2. Ignorance of epistemic limits and potentials
3. Dangers of Anthropomorphism

1. Relevance of underlying ontological or metaphysical properties

Premise 1: If a robot is roughly *performatively equivalent* to another entity whom, it is widely agreed, has significant moral status, then it is right and proper to afford the robot that same status.

Ethical behaviorism tends to overlook the need to identify the underlying ontological or metaphysical properties that ground moral status.

Case 1: The Real Dog

- Possesses sentience
- Capacity to experience pain, pleasure, and emotions
- Behavior is directly linked to experience
- This ontological property of sentience grounds its moral status

Case 2: The Robot-Dog

- Programmed to mimic the dog's behavior but does not possess consciousness or subjective experiences.
- Therefore, despite displaying similar behaviors to the dog, the robot lacks the ontological property of sentience that would justify granting it moral status.

1. Ignorance of epistemic limits and potentials

Premise 1: If a robot is roughly *performatively equivalent* to another entity whom, it is widely agreed, has significant moral status, then it is right and proper to afford the robot that same status.

While we may rely heavily on behavior in ascribing moral status, this method does not respect our epistemic limits, as we can also use other forms of evidence.

Relying exclusively on behavior ignores relevant knowledge about a robot's design and the intentions of its designers, which can provide critical context for understanding moral status.

1. Dangers of Anthropomorphism

Premise 1: If a robot is roughly *performatively equivalent* to another entity whom, it is widely agreed, has significant moral status, then it is right and proper to afford the robot that same status.

The tendency to attribute human-like characteristics to robots (anthropomorphism) may lead to misleading conclusions about their moral status.

Specific Dangers of Anthropomorphism

- Resource misallocation: individuals spend more time and effort for robots, mistaking its simulated behaviors for genuine emotions, while neglecting the needs of relatives, friends and animals.
- Negative impact on human relations: Individuals may become more emotionally attached to robots than to other humans, leading to weakened social bonds and possible isolation.
- Risks of harm: e.g. risking one's life to save a robot in an emergency

Food for Thought

- Given Danaher's argument that observable behavior is the primary basis for ascribing moral status, how do we define the boundaries of moral consideration? Are there specific behaviors that should trigger a moral response, and if so, how do we account for cultural and individual differences in interpreting such behaviors?

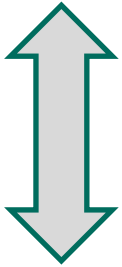
- As robotic technology continues to advance and potentially exhibit increasingly complex behaviors, how might our ethical frameworks evolve to address the moral status of these entities? In what ways should we integrate insights from both ethical behaviorism and alternative ethical theories to create a comprehensive approach to the moral implications of advanced AI and robotics?

3. Human Robot Interaction

- **Can Robots Have Moral Status?** Experts and the media discussed whether humanoid robots like Sophia should at some point acquire rights or even a certain legal status. Among other things, this involved questions of autonomy, responsibility and the definition of “personhood”.
- **Deception in Social Robotics**: Ethical questions arise when robots are designed to appear human-like or act empathetically, leading users to falsely trust them as moral agents.

Deception in Social Robotics

Question: Is there deception in social robotics?



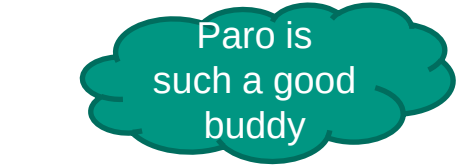
This is controversial!

Techniques enabling robots to detect basic human social gestures and to respond with human-like social cues are arguably forms of deception. (Wallach and Allan 2009)

Deception is only involved if a design of a robot misleads people into believing that it is areal human or animal. (Sorell and Draper 2017)

Deception in Social Robotics

© Hanson Robotics



© AIST

© ElliQ Company

Are these cases of deception?

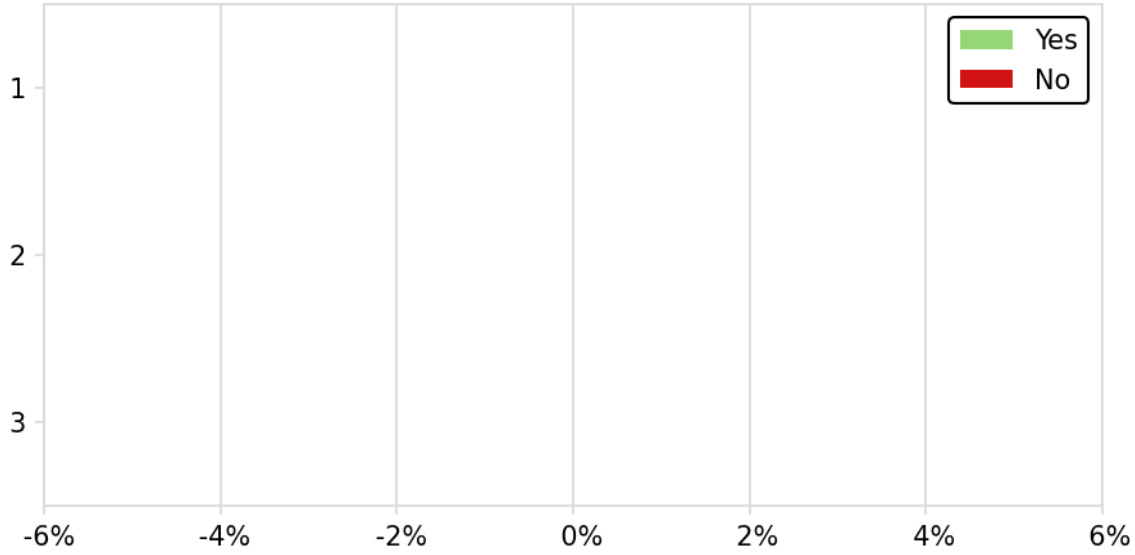
EduVote

- Participation in EduVote is possible anonymously and without registration
- You can either participate via the QR code (right).
- Or you can go to www.vote.ac and enter the ID a.bagattini@kit.edu in the ID field
- Please vote as soon as the survey starts

EduVote

1. Sophia
2. ElliQ
3. Miko3

Is it a form of deception?



ID = a.bagattini@kit.edu
Click to start poll

Deception in Social Robotics

- What is deception?
- When is deception morally wrong?

What is deception?

Intention-based definitions: deception means “intentionally causing someone to have false beliefs.“ (Carson 2017)

- Very high threshold for what counts as deception (creating false beliefs).
- Is creating false beliefs a necessary condition for deception?
- Counterexample: Paro (Sharkey and Sharkey 2020)
  - Although its creators did not intend to create the false belief that it is a real seal, some users may still develop that belief due to its appearance and behavior.
  - The illusion of sentience or emotion in social robots can be considered a form of deception, regardless of intentionality.
  - Deceiving vs. lying
- Necessary for deception: effects that something has on the deceived person.

When is deception morally wrong?

Intention-based definition

- Only in cases of intentionally causing false beliefs
- Very limited cases
- Example: robot that is designed for therapeutic purposes that is marketed as having the ability to understand and respond to the emotional needs of individuals, such as those with dementia or other cognitive impairments.
- Responsibility: always the deceiver

Impact-based definition

- Harmless vs. harmful deception
  - Misplaced trust
  - Overreliance on robotic systems
- Example: robot where – unintended by the designers – the realistic and engaging behaviors of the robot can lead the elderly individual to form a belief that the robot possesses genuine emotions and a desire to care for them.
- Responsibility:
  - Developers
  - Users
  - Manufacurers
  - Regulatory bodies

Critical Discussion

1. What standards should be established to differentiate between acceptable and unacceptable levels of deception in social robotics, especially considering the varying contexts of use (e.g., companionship for elderly individuals vs. educational tools for children)?

2. How can we effectively balance the potential benefits of social robots, such as companionship and assistance, against the ethical implications of deception, particularly in cases where users may anthropomorphize these machines?

Content of Today‘s Lecture

1. Introduction
2. Autonomy and Control
3. Human-Robot-Interaction
4. A practical example: Ethical Issues Concerning the Opportunities and Risks of AI-Assisted Robotics for Radiological Hazards
5. Open Discussion

5. Open Discussion

Literature

Carson, T. L. (2010) Lying and deception: Theory and practice. New York: Oxford University Press Inc.

David Collingridge: *The Social Control of Technology.* Pinter u. a., London u. a. 1982,

Danahar, J. (2019) Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism, Science and Engineering Ethics, Vol. 26

Fischer, J., and Ravizza, M. (1998) Responsibility and Control: A Theory of Moral Responsibility. Cambridge, UK: Cambridge University Press

Hare, R. (1952) *The Language of Morals*, Oxford: Clarendon Press

Santoni de Sio, F./ van den Hoven J. (2018) Meaningful Human Control Over Autonomous Systems: A Philosophical Account. In Frontiers in Robotics and AI, Vol. 5

Sharkey, A./ Sharkey, N. (2020) We need to talk about deception in social robotics! In Ethics and Information Technology Vol. 23

Singer, P. (1981) The Expanding Circle of Equality, Princeton UP

Sorell, T., & Draper, H. (2017) Second thoughts about privacy, Safety and deception. Connection Science, 29(3)

Wallach, W., & Allen, C. (2009) Moral machines: Teaching robots right from wrong. New York: Oxford University Press.

Karlsruher Institut für Technologie

Institut für Technikfolgenabschätzung und Systemanalyse

How can we embed AI ethics?

Case study: AI assistance for robot-assisted reconnaissance and defense against acute radiological hazards (KIARA)

Schedule

- Introduction to the Project
- Benefit Scenarios and Damage Scenarios
- Comparison of AI Ethics Guidelines Literature
- Systematizing Ethical Values and Principles for the Lifecycle of an AI System
- AI Ethics and AI Law
- Implementing AI Ethics in an Indicator System

SPONSORED BY THE

Federal Ministry of Education and Research

Introduction to the research project„KIARA“

KIARA

AI assistance for robot-assisted reconnaissance and defense against acute radiological hazards (KIARA)

KIARA – The project

Scenario:

Acute, radiological hazard situation

- No direct access for emergency services to the site of the hazard

Proposed solution:

Modular AI systems for use on mobile robotic systems

Research content:

Development and evaluation of AI and robotic systems for use in the this scenario

Role of AI:

- Supporting operations
  - in the rapid clarification of acute danger situations
  - in the implementation of initial measures to defuse the situation
- Reconnaissance and security measures can thus be carried out faster, more effectively and with less risk for the emergency services

SPONSORED BY THE

Federal Ministry of Education and Research

The project partners - from science, industry and a non-profit organisation - are working on various areas, from

- Robotics
- AI
- Law
- Ethics

TECHNISCHE UNIVERSITÄT DARMSTADT

KIT Karlsruher Institut für Technologie

DRZ DEUTSCHES RETTUNGSROBOTIK ZENTRUM

ENERGY ROBOTICS

TELERAD An AeroVironment™ Company

KIARA

Subproject (KIT-ITAS): Ethical Issues Concerning the Opportunities and Risks of AI-Assisted Robotics for Radiological Hazards


Prof. Dr. Dr. Rafaela Hillerbrand


Dr. Michael Schmidt


Désirée Martin, M.A.

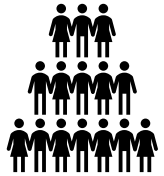

Heinrich Blatt, M.Sc.

KIARA

The Ethical Subproject: Ethical Issues Concerning the Opportunities and Risks of AI-Assisted Robotics for Radiological Hazards

Big Goal of Ethics in the field of AI: Embedding Ethics in AI

The ethical subproject

Interface between ethics  and technology 

- Co-creation of the design and use of the systems by means of continuous, prospective technical-ethical (accompanying) research
- Identification of ethical issues
- Achieving the acceptability of technical developments

→ Development of regulations and indicator systems for the design and application of AI assisted robots in this context

The ethical subproject

Interface between ethics  **and law** 

Main questions: To what extent do ethics and law agree? Do ethics and law provide different, conflicting or complementary assessments?

Research objectives:

- Identify which ethical aspects can be quantified and transformed into indicators in a meaningful way, and which cannot, and therefore may need to be addressed differently.
- Findings as to whether there are ethical aspects that are not covered by law
- Systematization of ethical aspects
- Recommendations for the regulation of AI-assisted robot systems in this context

The Ethical Subproject

Ethical Issues Concerning the Opportunities and Risks of AI-Assisted Robotics for Radiological Hazards

Big Goal: Embedding Ethics in AI

2. Benefit Scenarios and Damage Scenarios

Pros	Cons

Immediate Danger

Imagine the robot meets a suspect while exploring an apartment. The suspect might attack the robot. The suspect also might prepare an explosive.
Answer the following question!
Is the operator –via the robot- allowed to use physical force on the suspect in face of immediate danger?

https://commons.wikimedia.org/wiki/File:Van_Bentum_Explosion_in_the_Alchemist%E2%80%99s_Laboratory_FA_2000.001.285.jpg

Accountability

Imagine the KIARA robot unintentionally destroys a valuable Chinese ceramic while automatically moving within a apartment in which it is employed. The operator was suspected to monitor the movement of the robot but did not control it directly at that moment.

Answer the following questions!

Who is accountable for this incident?

Is it technically possible to make the accountability issue less problematic?

https://clevelandart.org/art/1962.154

Cultivating Trust

When we trust, we make ourselves vulnerable to others, yet still feel safe. Trust takes time to build and requires attention to sustain. Systems play a role in establishing, building, maintaining, strengthening, and compromising trust. For Trust, it is not enough if the system is accepted, it also has to be acceptable in every important aspect.

Surface different points in KIARA where direct and indirect stakeholders might be vulnerable. What could make KIARA trustworthy and protect vulnerable people?

Consider Key Values and Value Tensions

A technology can support certain values and hinder others (e.g., a shared online calendar system can support community, but impinge on privacy). Possible values may include for example: accountability, environmental sustainability, fairness, privacy.

Generate a list of as many potentially implicated values for KIARA as possible in 5-10 minutes.
Brainstorm two value tensions that KIARA may engage. For each tension, identify one design feature that favors one of the values over the other, or that solve the tension.

Health and Work of the Future

Technology may have effects on people‘s health. The introduction of new technologies can also change working habits, or even what it means to „do work“. How might KIARA effect people‘s health and change the nature of work? How could this technology be adapted to other projects for rescue and health systems?

Imagine that KIARA has been widely adopted. Reflect upon 2 likely ways in which the system influences health.
Think about 10 years from now. In which ways, positive and negative, may KIARA change the way people work.

Benefit scenarios and damage scenarios

Pro:

- Significant reduction in health risks for the emergency services
- Effective, documented situation clarification
  - Faster adoption of appropriate measures

Contra:

20 identified damage scenarios

- Divided into four categories: Scenarios in which

  a) the AI system collects or interprets data incorrectly or inadequately

  b) the interaction with humans or animals in the danger zone leads to ethical issues

  c) the user interface poses ethically relevant hazards

  d) the manufacturing of the system raises ethical issues

Benefit scenarios and damage scenarios

What we need:

Solutions to address ethical challenges

- not only in the field of AI rescue robotics
- but in AI applications in general
- what needs to be considered in AI design

Comparison of AI Ethics Guidelines Literature

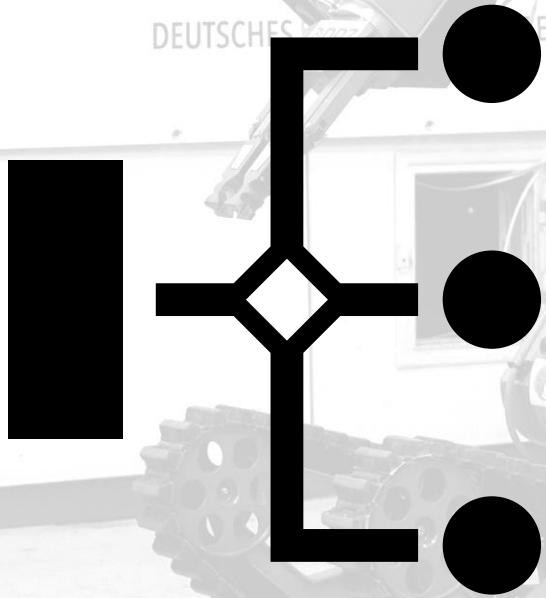
Comparison of AI ethics guidelines literature

- Prominent AI ethics guidelines
  - 2017 Asilomar conference (Beneficial AI), ‘Asilomar AI Principles’, Future of Life Institute. Accessed: Oct. 08, 2021. [Online]. Available: https://futureoflife.org/ai-principles/
  - L. Floridi et al., ‘AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations’, Minds & Machines, vol. 28, no. 4, pp. 689–707, Dec. 2018, doi: 10.1007/s11023-018-9482-5.
  - IEEE, ‘Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, version 2’, Version 2, 2017. [Online]. Available: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf
  - High-Level Expert Group on Artificial Intelligence set up by the European Commission, ‘Ethics guidelines for trustworthy AI’, European Commission, 2019. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
  - C. Abrassart et al., ‘Montréal Declaration for a Responsible Development of Artificial Intelligence’, Announced at the conclusion of the Forum on the Socially Responsible Development of AI, 2018. [Online]. Available: https://www.montrealdeclaration-responsibleai.com/the-declaration

Tabelle im Detail – ein Ausschnitt

Keywords	AI4People (2018)	IEEE Ethically aligned (2017)	HLEG Ethics Guidelines for Trustwothy AI (2019)
Accountability	Included in the principle of "Explicability: Enabling the Other Principles Through Intelligibility and Accountability" (p. 699); accountability "as an answer to the question: “who is responsible for the way it works?”" (p. 700)	One of the five principles, p. 27f.; closely linked with Transparency; based "on the cultural context, application, and use of A/IS, people and institutions need clarity around the manufacture and deployment of these systems to establish responsibility and accountability, and avoid potential harm." (IEEE 2017, p. 27) "Public confidence in technology requires both transparency and accountability. [...] Transparency improves accountability, which might in turn support judicial processes. Finally, following high profile accidents, society can benefit from the reassurance of knowing that problems have been found and addressed." (IEEE 2019, p. 12)	As a requirement of Trustworth AI - "Including auditability, minimisation and reporting of negative impact, trade-offs and redress" (p. 14)
Autonomy	Principle of "Autonomy: The Power to Decide (Whether to Decide) [...] the idea that individuals have a right to make decisions for themselves about the treatment they do or not receive" (p. 697f.)	Mentioned as "social or cultural value" (IEEE 2017, p. 33); mentioned just as 'system autonomy' which can be distinguished in four levels: Human Operated, Human Delegated, Human Supervised, and Fully Autonomous, IEEE 2019, p. 17	Principle of repect for human autonomy, p. 12; mentioned in the requirement of Trustworthy AI 'human agency and oversight', p. 15; mentioned with a good life of individuals, p. 9

Systematizing Ethical Values and Principles for the Lifecycle of an AI System



Towards a Systematization of Ethical Values and Principles for the Life Cycle of an AI System

Karlsruher Institut für Technologie

Results of the comparison of the AI ethics guidelines

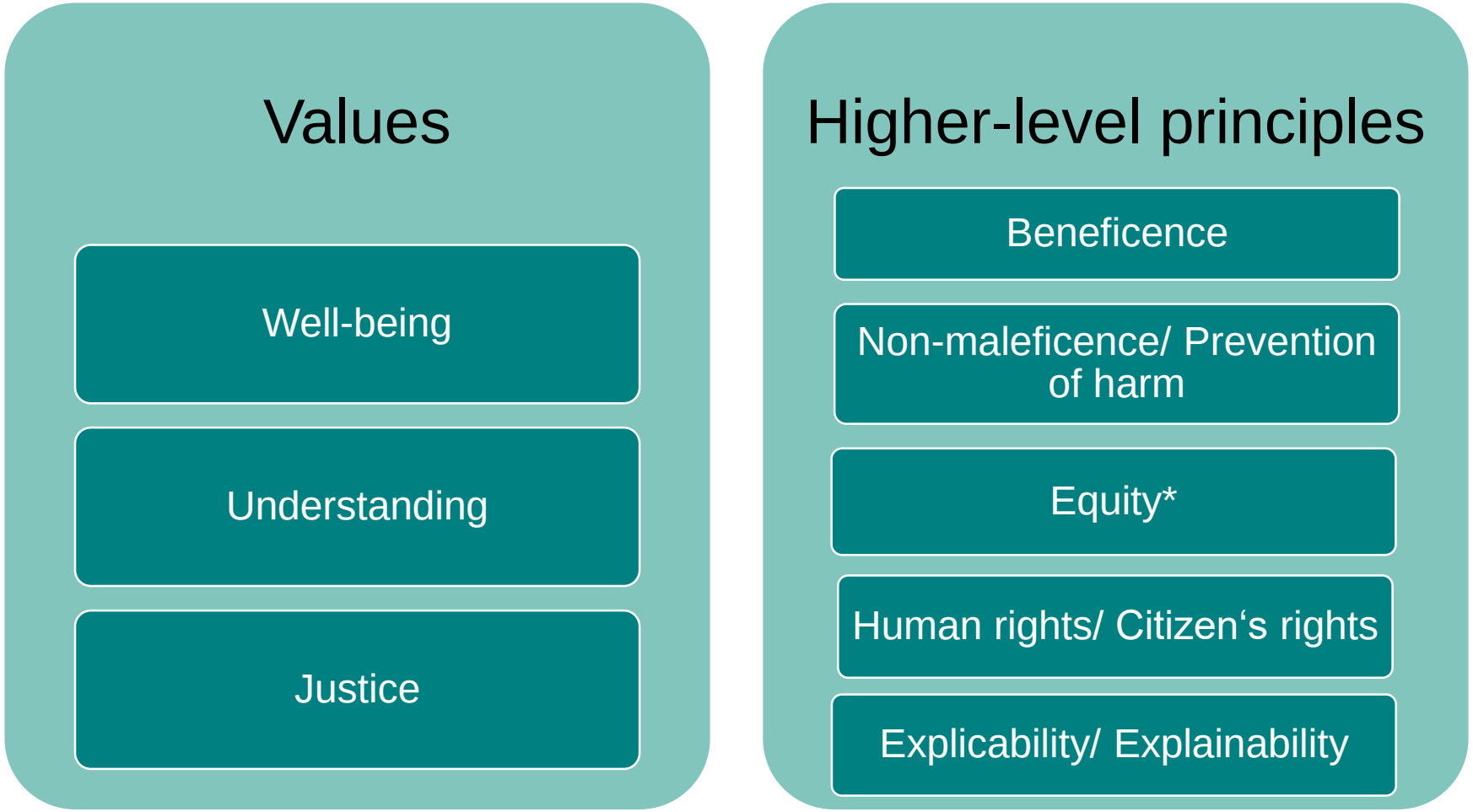
- with consideration of the analysis of the concept of values and principles
  - Values are desirable states
  - Principles are action-guiding

Consensus of the AI ethics guidelines (excerpt):

- Accountability / responsibility
- Beneficence
- Non-Maleficence
- Justice / fairness
- Transparency / explicability
- Aspects of well-being

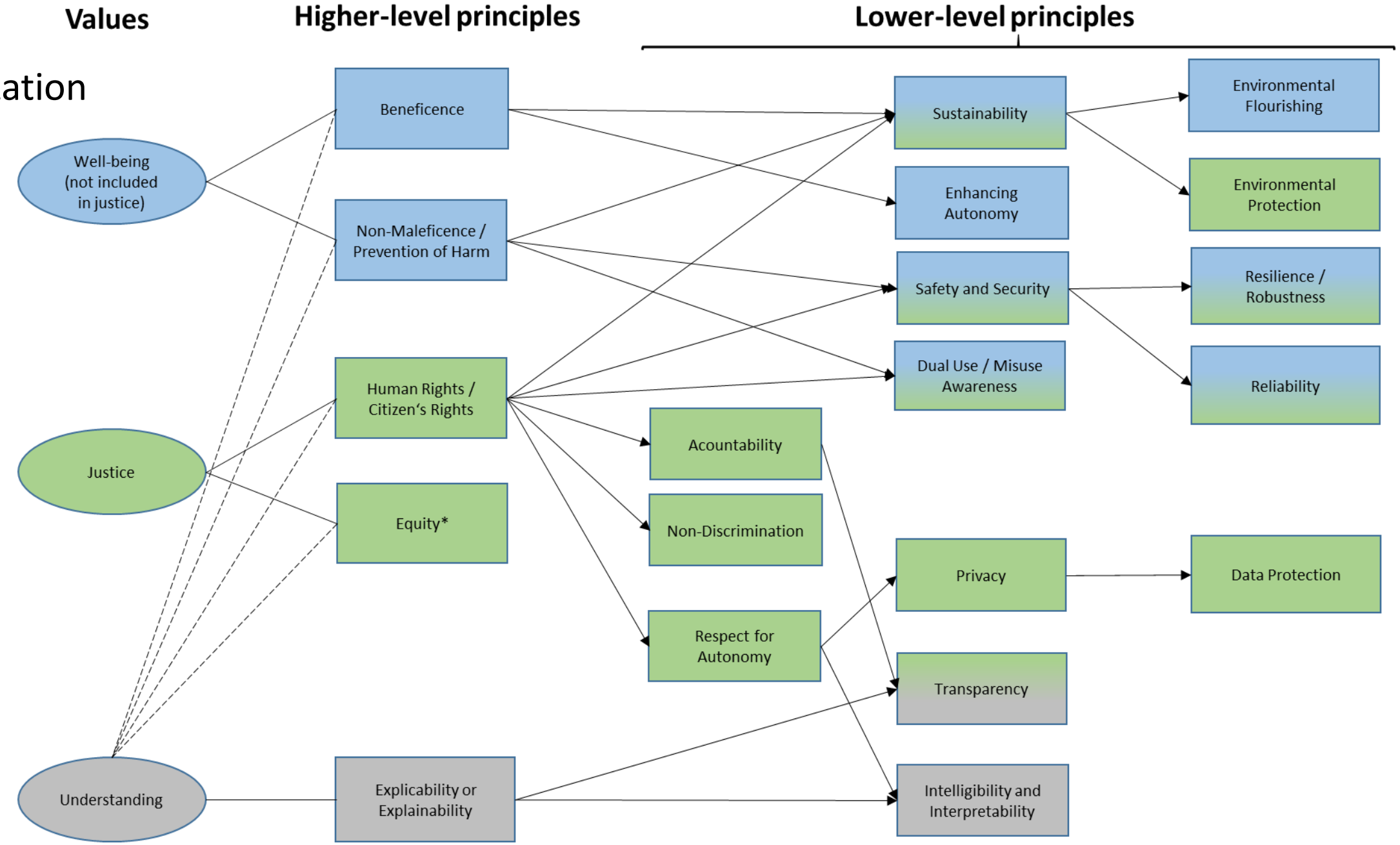
Next step: Systematization of values and principles
Challenge: Many principles
Therefore: Principles of higher and lower levels

Towards a Systematization of Ethical Values and Principles – Values and higher-level principles



*Not part of the consensus

Systematization of ethical values and principles



The Systematization of Ethical Values and Principles for the Life Cycle of an AI System

Karlsruher Institut für Technologie

Now, we know

- Which ethical values and principles should be embedded in the life cycle of an AI system
- How ethical values and principles are related

Remember the big goal: **Embedding ethics in AI**

To achieve this goal, we need to think

- prospectively in terms of developers
  - what do they need to incorporate ethical issues into the system?
- and retrospectively in terms of auditors
  - how can they determine whether ethical requirements have been incorporated?

The Systematization of Ethical Values and Principles for the Life Cycle of an AI System

Remember the big goal: **Embedding ethics in AI**

To achieve this goal, we need to think

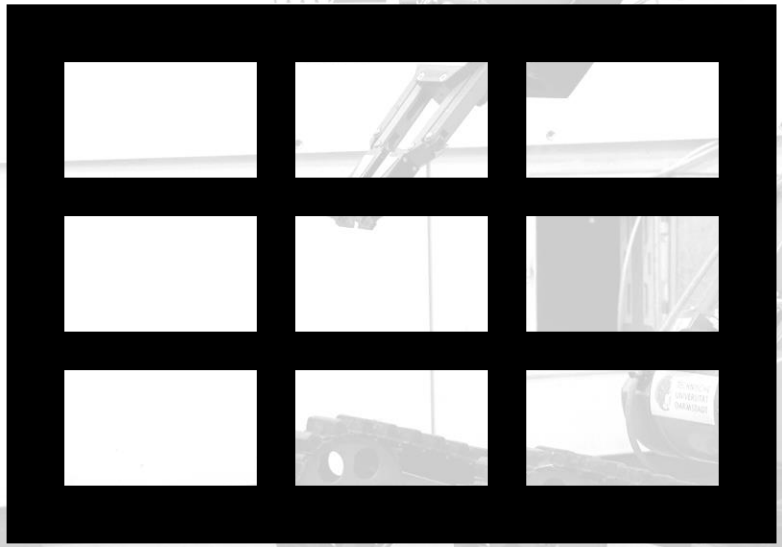
- and retrospectively in terms of auditors
  - how can they determine whether ethical requirements have been implemented?

    But wait! ‼

    For auditors to be able to assess compliance with ethical requirements, there must be ethical requirements that are legally binding!

    → We need a legally binding AI regulation to guarantee and assess AI ethics

AI Ethics and AI Law



Comparing AI ethics and AI law

Status quo:

So far there is no comparison of AI ethics and AI law (with European level)

We want to change that!

For the purpose of comparison, we have taken the current regulations (at European level) and international conventions AND WE ADDED THEM TO THE TABLE:

- Regulation (EU) 2024/1689, 2024
- Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (2023).
- UNESCO, ‘Recommendation on the Ethics of Artificial Intelligence’. 2022.
- OECD, ‘Recommendation of the Council on Artificial Intelligence’. 2019.

Comparing AI ethics and AI law

1. **Consensus**
2. **Dissent**

Comparing AI ethics and AI law

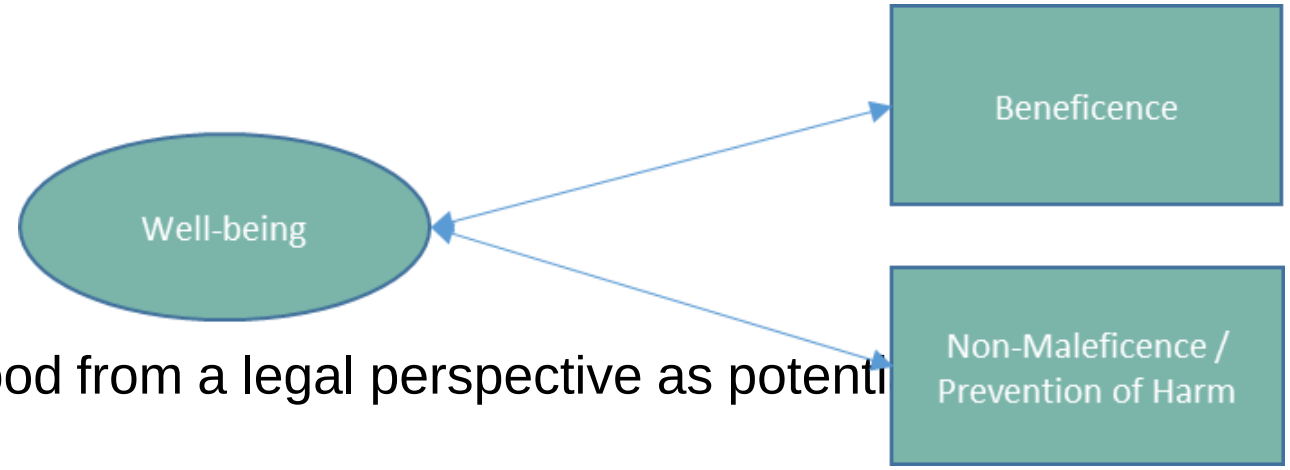
1. Consensus

- Accountability / responsibility
- Non-Maleficence
- Justice / fairness
- Transparency / explicability

Comparing AI ethics and AI law

2. Dissent

- Accountability / responsibility
- Beneficence
- Non-Maleficence
- Justice / fairness
- Transparency / explicability
- Aspects of well-being

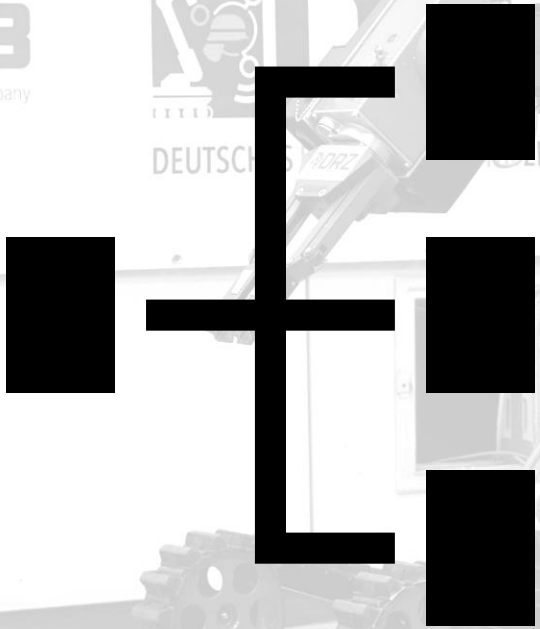


Beneficence and well-being are understood from a legal perspective as potenti
outcomes of AI technologies, but there

- is no legal consensus on regulatory obligations and

→ no common operationalization of beneficence and well-being → clarifying the why is a task for further research

Implementing AI Ethics in an Indicator System



Implementing AI Ethics in an Indicator System

Remember the big goal: **Embedding ethics in AI**

To achieve this goal, we need to think

- prospectively in terms of developers
  - what do they need to incorporate ethical issues into the system?
- and retrospectively in terms of auditors
  - how can they determine whether ethical requirements have been incorporated?

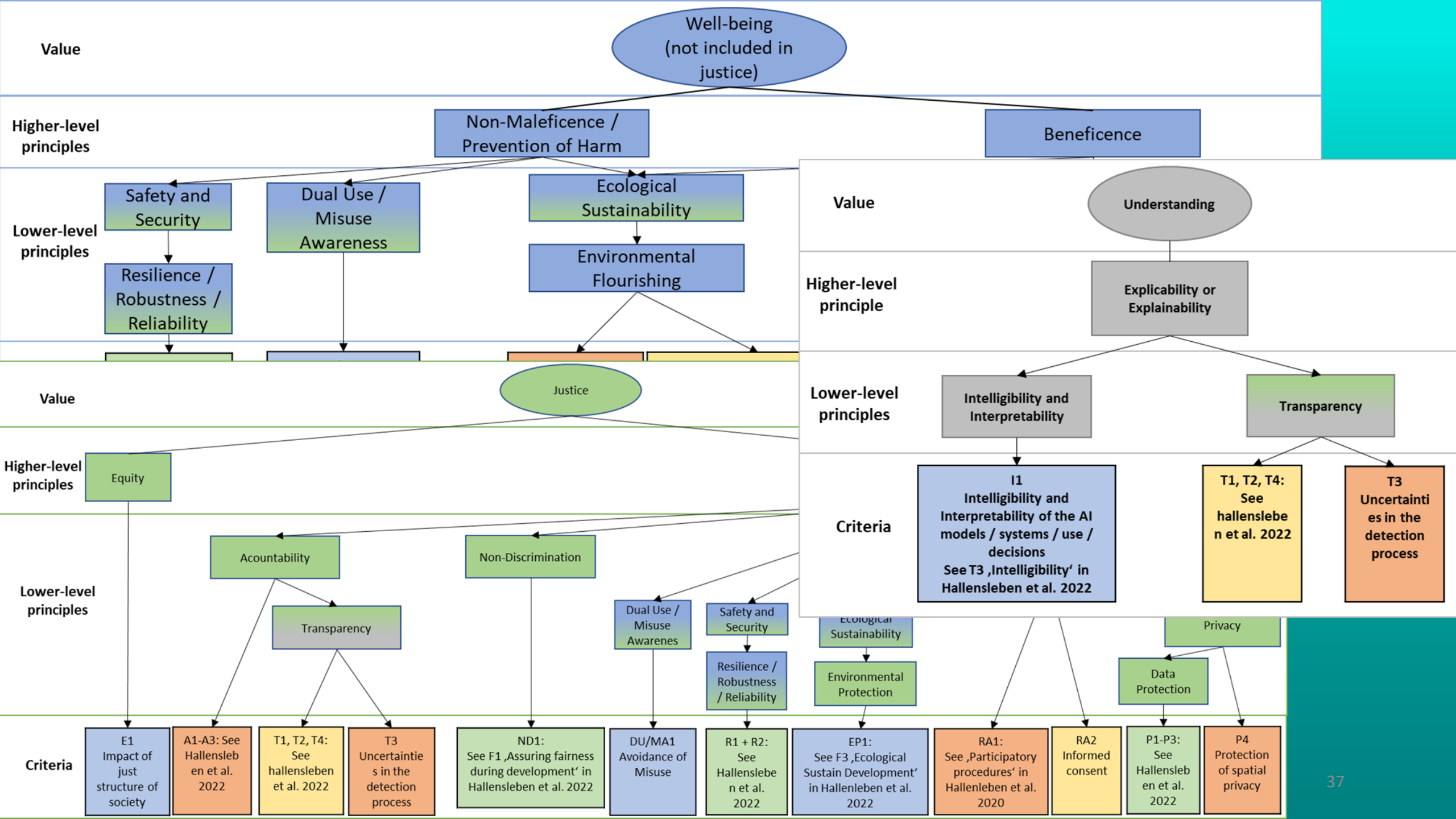
Implementing AI Ethics in an Indicator System

Development of an indicator system:

- **Damage scenarios**
- **Systematized ethical values and principles**

Background: The **VCIO model** by Hallenleben et al. (2020) „From Principles to Practice - An interdisciplinary framework to operationalise AI ethics”

To address the reasonably shared principles, we have included them in the model


Value
Well-being (not included in justice)
Higher-level principles
Non-Maleficence / Prevention of Harm
Beneficence
Lower-level principles
Safety and Security
Dual Use / Misuse Awareness
Ecological Sustainability
Resilience / Robustness / Reliability
Environmental Flourishing
Value
Justice
Higher-level principles
Equity
Lower-level principles
Acountability
Non-Discrimination
Transparency
Dual Use / Misuse Awarenes
Safety and Security
Ecological Sustainability
Resilience / Robustness / Reliability
Environmental Protection
Privacy
Data Protection
Criteria
E1 Impact of just structure of society
A1-A3: See Hallensleben et al. 2022
T1, T2, T4: See hallensleben et al. 2022
T3 Uncertainties in the detection process
ND1: See F1 ‚Assuring fairness during development' in Hallensleben et al. 2022
DU/MA1 Avoidance of Misuse
R1 + R2: See Hallenslebe n et al. 2022
EP1: See F3 ‚Ecological Sustain Development' in Hallenleben et al. 2022
RA1: See ‚Participatory procedures' in Hallenleben et al. 2020
RA2 Informed consent
P1-P3: See Hallensleb en et al. 2022
P4 Protection of spatial privacy
Value
Understanding
Higher-level principle
Explicability or Explainability
Lower-level principles
Intelligibility and Interpretability
Transparency
Criteria
I1 Intelligibility and Interpretability of the AI models / systems / use / decisions See T3 ‚Intelligibility' in Hallensleben et al. 2022
T1, T2, T4: See hallenslebe n et al. 2022
T3 Uncertainti es in the detection process

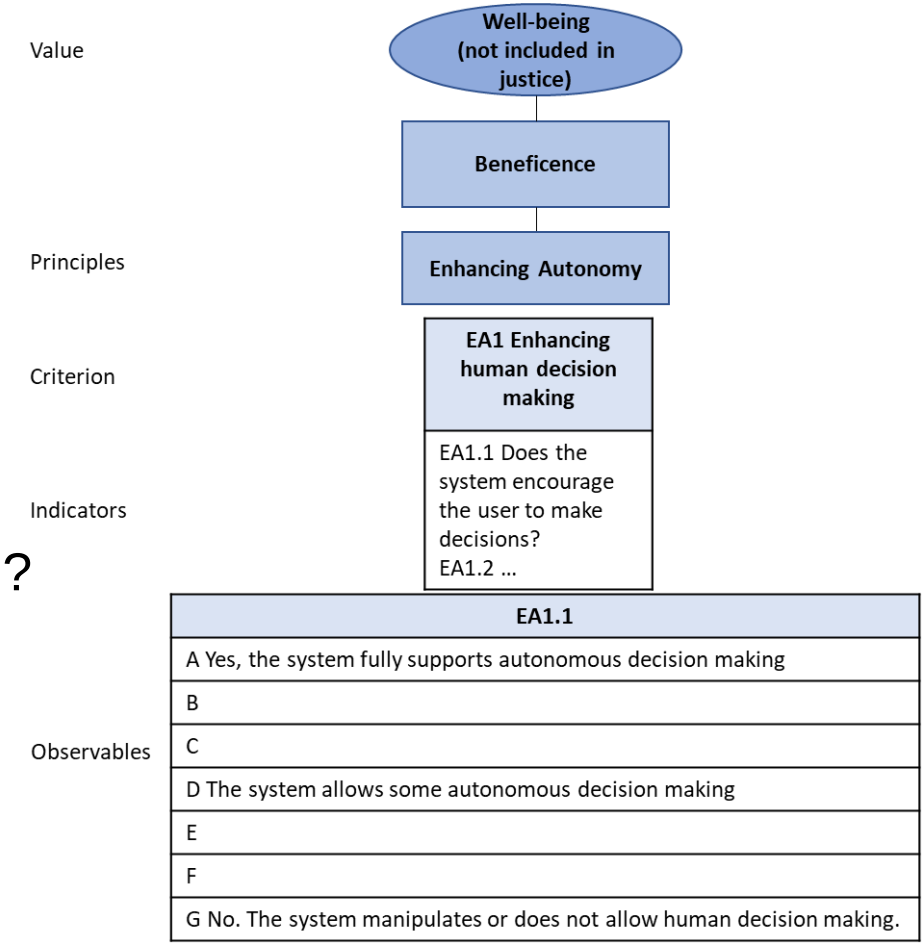
The Indicator System

This indicator system enables us to translate abstract ethical values and principles into concrete requirements

Remember the big goal: **Embedding ethics in AI**

To achieve this goal, we need to think

- ✓ prospectively in terms of developers
  what do they need to incorporate ethical issues into the system?
- ✓ and retrospectively in terms of auditors
  how can they determine whether ethical requirements have been incorporated?



This brings us closer to the goal of embedding ethics in AI!

Publications

- D. Martin, M. W. Schmidt, and R. Hillerbrand, ‘Comparing AI Ethics and AI Regulation: Ethical Valuesand Principles and the Case of Well-being, Beneficence and Sustainability’, in *Philosophy of Artificial Intelligence*: The State of Art, Synthese Library., V. C. Müller, L. Dung, A. R. Dewey, and G. Löhr,Eds., Berlin: SpringerNature, forthcoming.
- D. Martin, M. W. Schmidt, R. Hillerbrand. Implementing AI Ethics in the Design of AI-assisted Rescue Robots. 2023 IEEE International Symposium on Ethics in Engineering, Science, and Technology.
- K. Daun *et al*., "A Holistic Concept on AI Assistance for Robot-Supported Reconnaissance and Mitigation of Acute Radiation Hazard Situations," *2024 IEEE International Symposium on Safety Security Rescue Robotics (SSRR)*, New York, NY, USA, 2024, pp. 40-45, doi: 10.1109/SSRR62954.2024.10770059.
  keywords: {Training;Navigation;Law;Prevention and mitigation;Reconnaissance;User interfaces;Hazards;Robustness;Artificial intelligence;Robots}

Literatur

- 2017 Asilomar conference (Benficial AI), „Asilomar AI Principles“, Future of Life Institute, 2017. [Online]. https://futureoflife.org/ai-principles/
- „High-Level Expert Group on Artificial Intelligence set up by the European Comimission, „Ethics guidelines for trustworthy AI“, European Commission, 2019. [Online. file:///C:/Users/dl5458/Downloads/ai_hleg_ethics_guidelines_for_trustworthy_ai-en_87F84A41-A6E8-F38C-BFF661481B40077B_60419.pdf
- Hallenleben, S. et al. (2020). From Principles to Practice - An interdisciplinary framework to operationalise AI ethics. VDE, Bertelsmann Stiftung, Frankfurt a. M. / Gütersloh.
- „IEEE Draft Standard for Transparency of Autonomous Systems“, IEEE P7001D3 Sept. 2021, S. 1–75, Sep. 2021.
- UNESCO, ‘Recommendation on the Ethics of Artificial Intelligence’. 2022.
- OECD, ‘Recommendation of the Council on Artificial Intelligence’. 2019
- „The IEEE Initiative on Ethics of Autonomous and Intelligent Systems, „Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, version 2“, IEEE, 2017. [Online]. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf
- L. Floridi et al., “AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations,” Minds Mach., vol. 28, no. 4, pp. 689–707, Dec. 2018, doi: 10.1007/s11023-018-9482-5.
- „Montréal Declaration for Responsible Development of Artificial Intelligence“, Abrassart, Christophe, Yoshua Bengio, Guillaume Chicoisne, Nathalie de Marcellis-Warin, Marc-Antoine Dilhac, Sébastien Gambs, Vincent Gautrais et al., 2018.
- J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, ‘Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI’, SSRN Journal, 2020, doi: 10.2139/ssrn.3518482.
- Regulation (EU) 2024/1689, 2024.
- T. Hagendorff, ‘The Ethics of AI Ethics: An Evaluation of Guidelines’, Minds & Machines, vol. 30, no. 1, pp. 99–120, Mar. 2020, doi: 10.1007/s11023-020-09517-8.
- A. Jobin, M. Ienca, and E. Vayena, ‘The global landscape of AI ethics guidelines’, Nat Mach Intell, vol. 1, no. 9, pp. 389–399, Sep. 2019, doi: 10.1038/s42256-019-0088-2.
- Y. Zeng, E. Lu, and C. Huangfu, ‘Linking Artificial Intelligence Principles’, in Proceedings of the AAAI Workshop on Artificial Intelligence Safety 2019, H. Espinoza, S. Ó hÉigeartaigh, X. Huang, J. Hernández-Orallo, and M. Castillo-Effen, Eds., Honolulu, Hawaii, 2019, p. 4. [Online]. Available: https://www.ceur-ws.org/Vol-2301/paper_15.pdf

Thank you for your attention!!