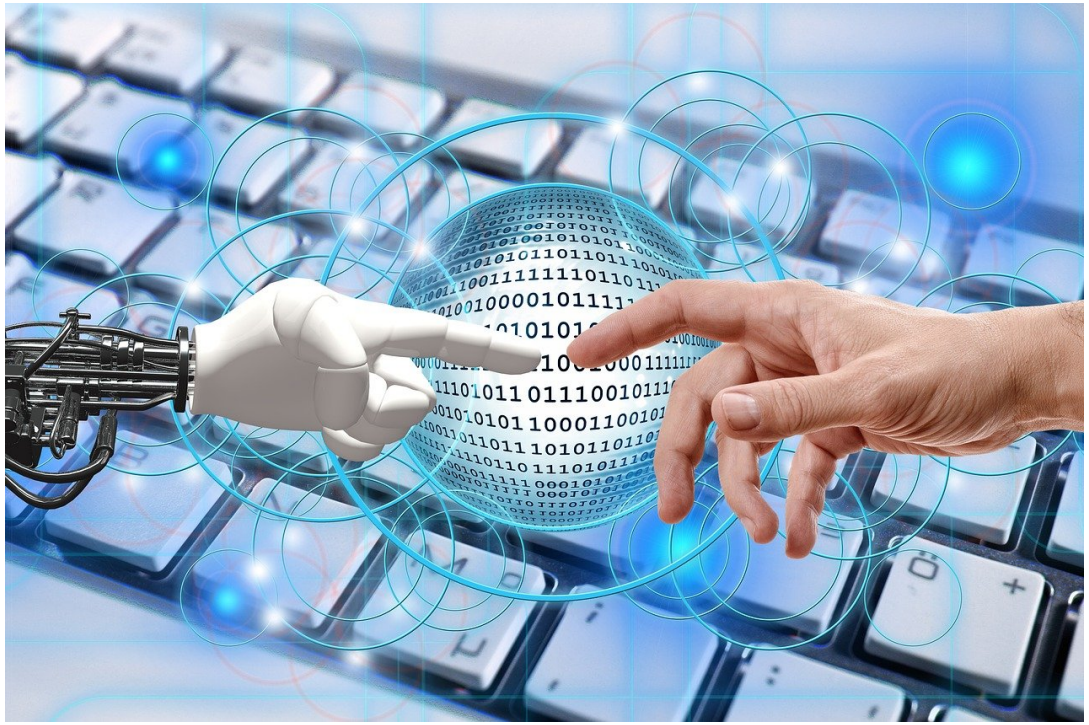
Ethical Aspects in Human-Robot Interaction



Prof. Dr. Teena Hassan
teena.hassan@h-brs.de

Department of Computer Science
Hochschule Bonn-Rhein-Sieg
Sankt Augustin

20 June 2024

A²S

Learning Goals

- At the end of today's lecture, you will be able to:
  1. Explain why ethics is relevant in Human-Robot Interaction (HRI).
  2. Elucidate the concept of trust and explain the relation between ethics and trust.
  3. Identify the constructs involved in the acceptance of technology in general and assistive social robots in specific.
  4. Explain the principles and some of the important aspects of GDPR, and its application in HRI design and experiments.

What is Ethics?

- According to Immanuel Kant, ethics tries to answer the question “What should I do?“
- Ethics deals with the principles, judgements, and norms that help to determine the answer to the above question.
- Morality is ethics in action.

(Bartneck et al., 2021)

Brainstorming

- Why is ethics relevant in human-robot interaction?

- Enter your answers here: https://www.menti.com/alk9mqip65st

Descriptive versus Prescriptive Ethics

- Descriptive ethics:
  - Studies what attitudes, beliefs, and intuition people have regarding moral principles by conducting experiments.
  - Derives philosophical insights by observing how people make moral decisions.

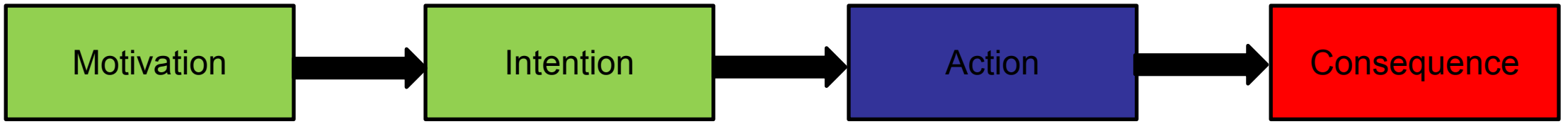
(Bartneck et al., 2021)

Descriptive versus Prescriptive Ethics

- Prescriptive ethics:
  - Applies the insights gained from descriptive ethics to **evaluate** if an action is right or wrong / good or evil.
  - Also known as normative ethics, because it defines the norms that determine morality of actions.

(Bartneck et al., 2021)

Deontological versus Consequentialist Ethics

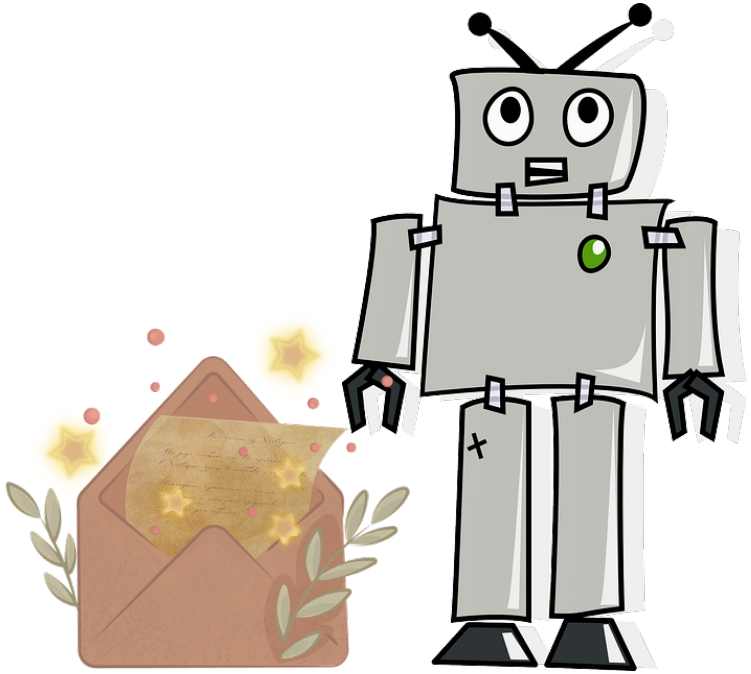


- **Deontological**: Normative ethics based on duty
  - Determines whether a course of action is ethically correct based on „characteristics that affect the action itself“, e.g. the **intention** or **motivation** behind the action.

- **Consequentialist**: Normative ethics based on outcomes
  - Determines whether a course of action is ethically correct based on the predictable **consequences** of that action.

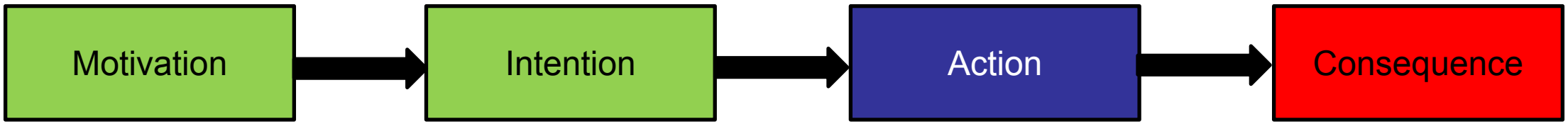
(Bartneck et al., 2021)

Designing Ethical Machines – An Example



??

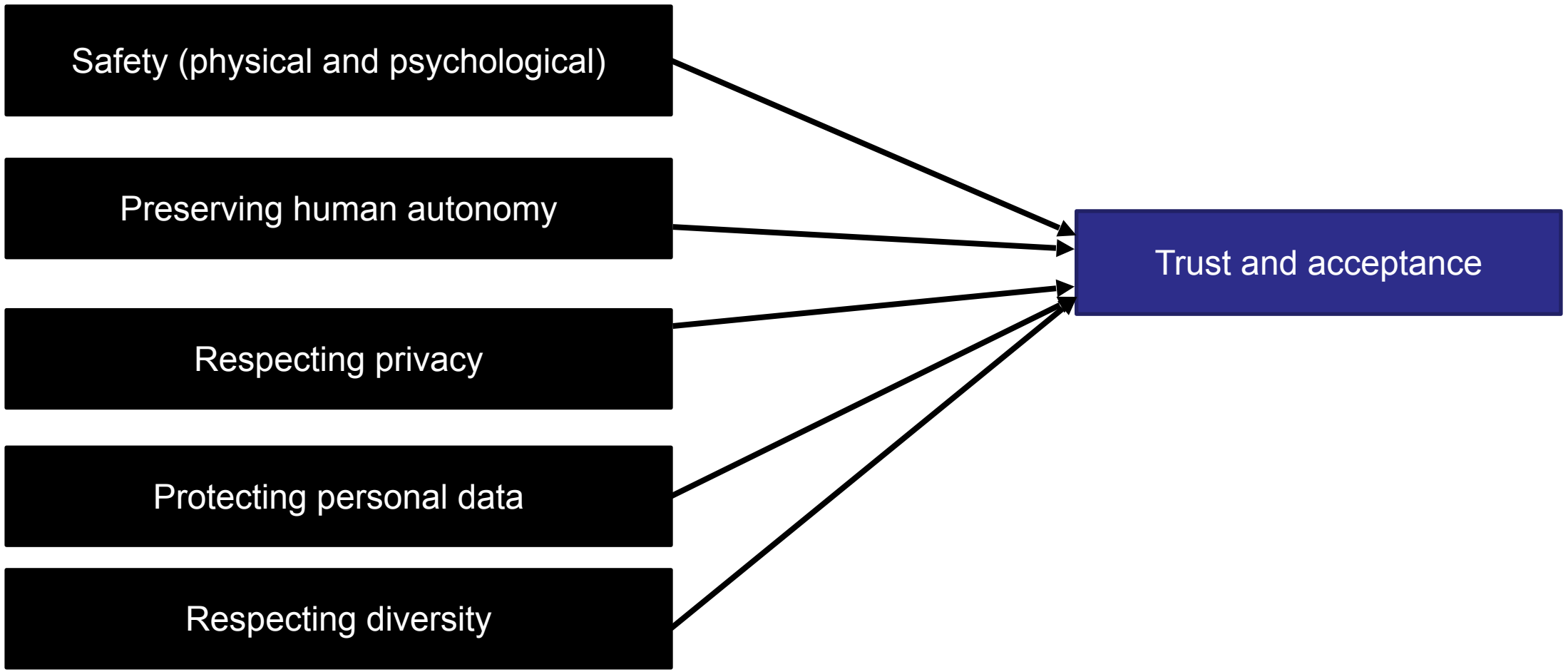
(Bartneck et al., 2021)



Ethics should inform the motivation, the intention and the utility of an action.
We need both deontological and consequentialist ethics.

A²S

Ethics and Trust in HRI



Part 1: Trust and Acceptance

What is Meant by Trust?

trusts

Trustor

Agent who is trusting another agent

Vulnerability

Trustee

Agent who is trusted by another agent

Uncertainty
Lack of control

Definition of Trust

- Lee and See (2004) defined trust as “the **attitude** that an agent will help achieve an individual’s goals in a situation characterized by **uncertainty** and **vulnerability**.”

- “Trust is the willingness of a party to be **vulnerable** to the action of another party based on the **expectation** that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party” (Mayer et al., 1995).

- The trustor is vulnerable and has something to lose (loss of life, property, reputation, money, mental/physical health, etc.).

- Trust comes into picture when there is uncertainty and lack of control associated with the actions or outcomes of the trustee.
  - No uncertainty means there is no vulnerability. Then there’s no need for trust.

Trust in AI and Robots

Can trust?

Is trustworthy for?

Trustor:

Adopts / Uses / Consumes AI

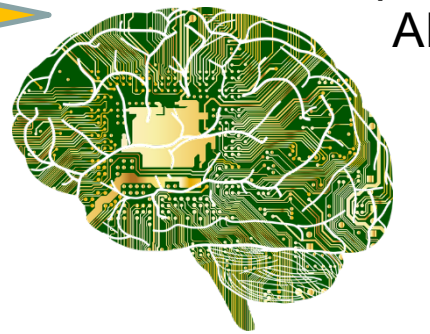
Individuals, groups of individuals, businesses, schools, governments, judiciary…

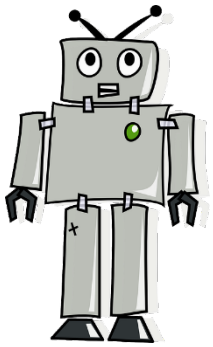
Acceptance

Trustee:

Offers / Deploys / Produces AI

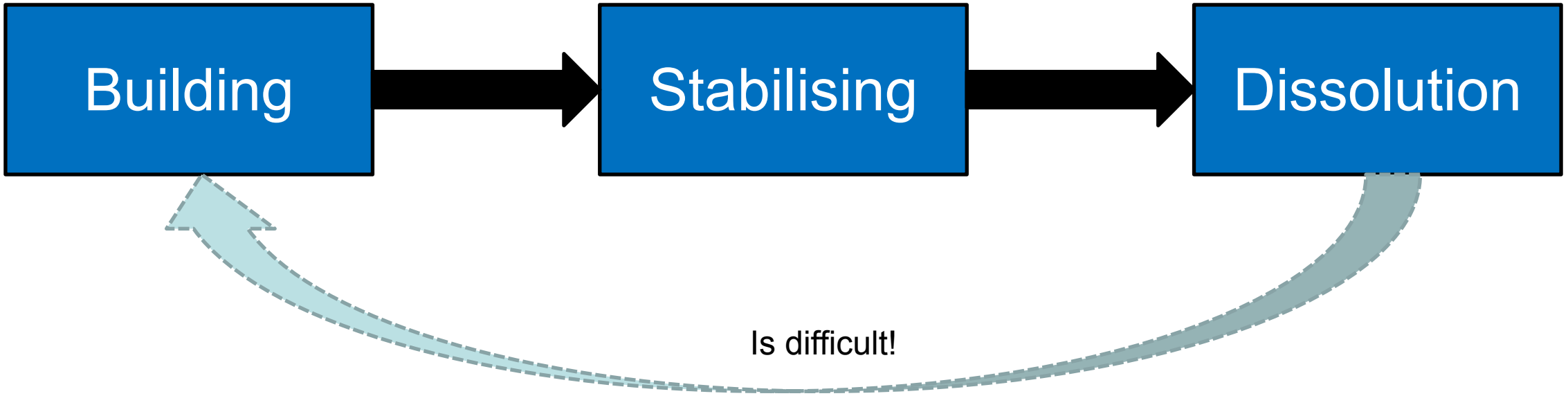
AI technologies, AI creators, AI-based service providers, AI users





Phases of Trust

- Trust develops over time.
- Three phases: building, stabilising, dissolution (T. Kautonen and Karjaluoto, 2008)



Can you think of an example showing the evolution of trust?

Trust in Automation

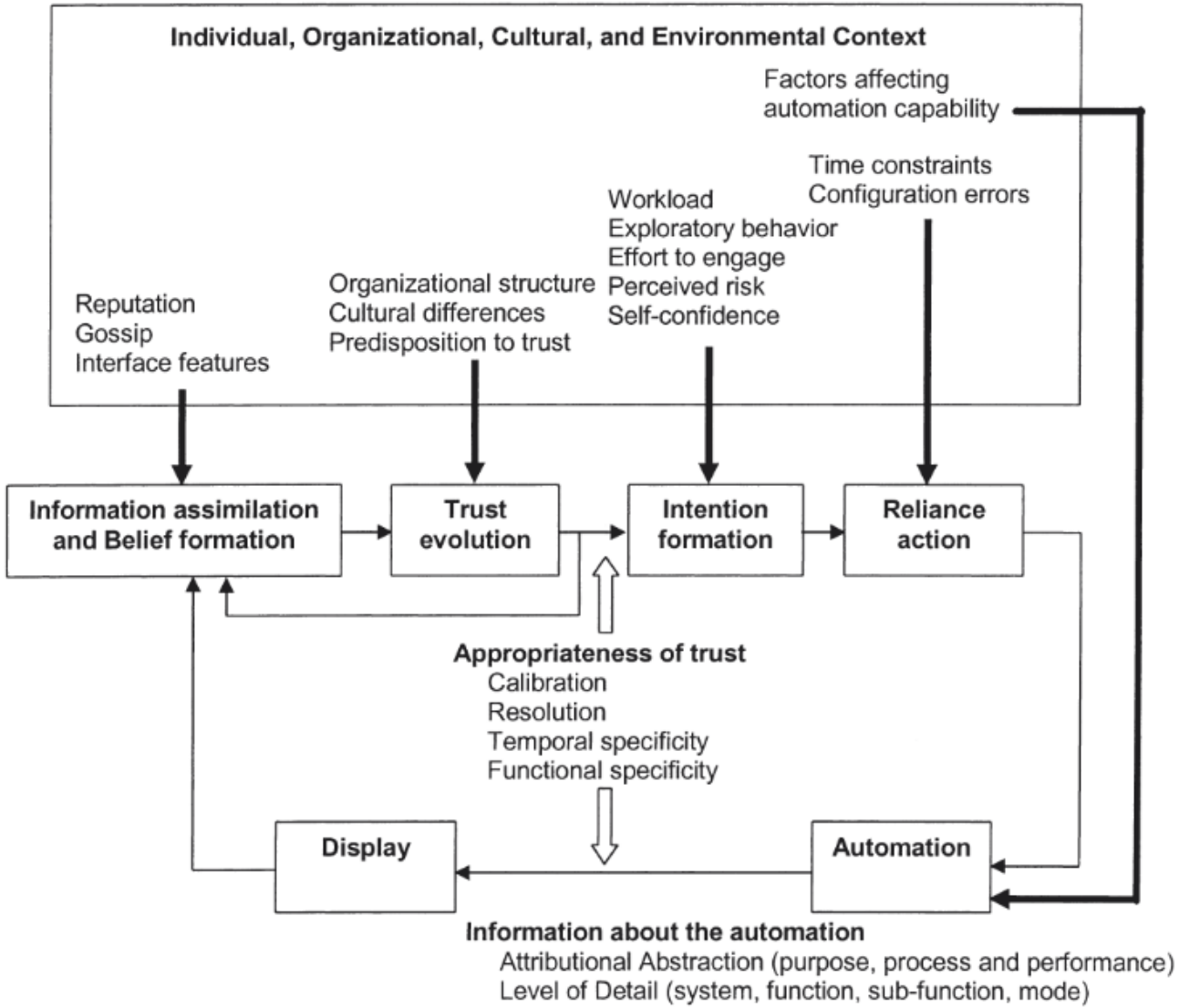
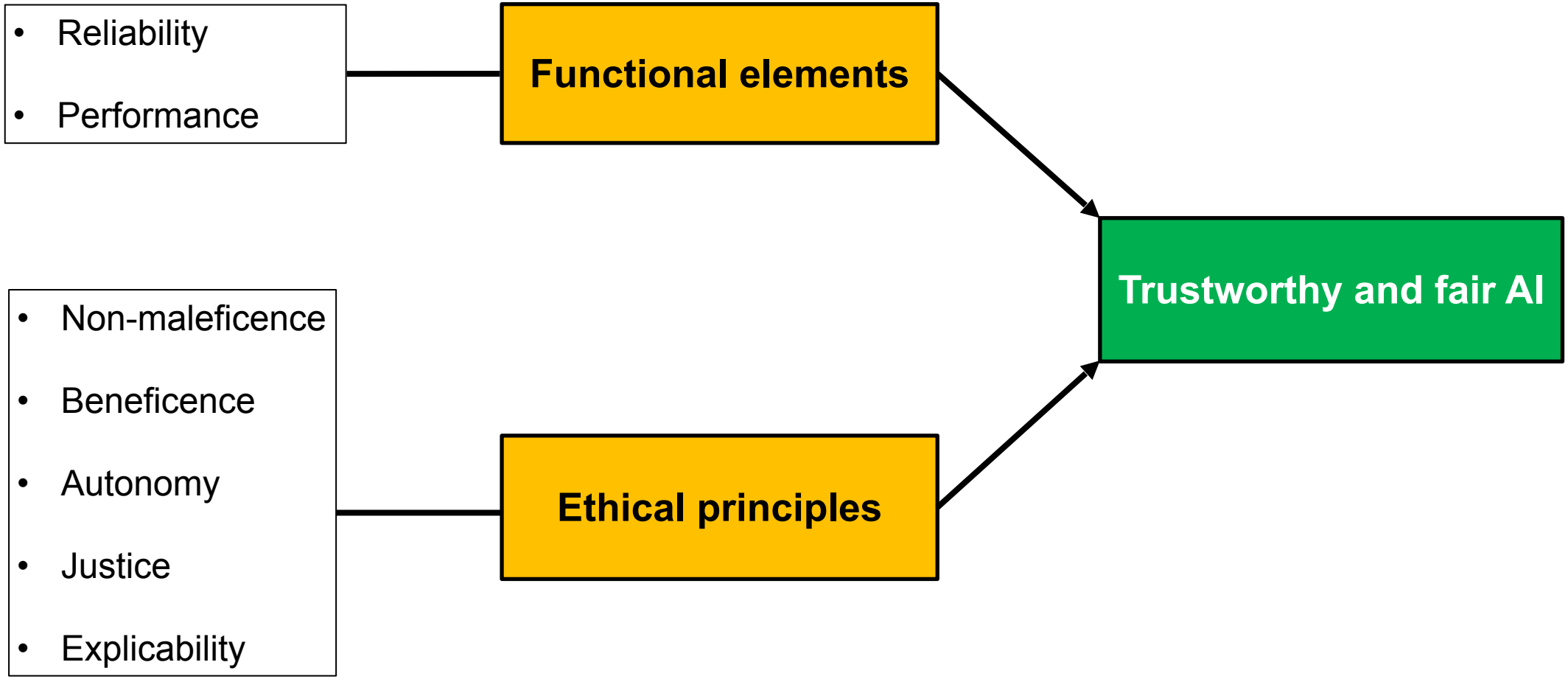


Figure 4. A conceptual model of the dynamic process that governs trust and its effect on reliance.

Source: Figure 4 in (Lee and See, 2004)

Elements of Trust in Machines

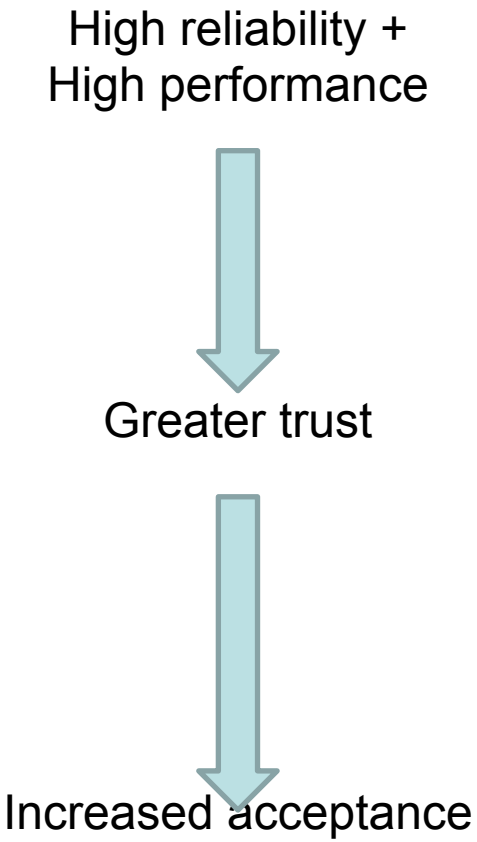


(Bartneck et al., 2021)

Functional Elements Influencing Trust

1. Performance
   - How well does the machine perform?
   - E.g. Accuracy, false alarm rate, time and memory complexity, energy efficiency, etc.

2. Reliability
   - How long does the machine run without failure?
   - How often does software/hardware failures occur?
     - Software failure (e.g. „segmentation fault“; crashing app after update,)
     - Hardware failure (e.g. sensors not working)
   - Under which conditions is the performance guaranteed?



(Bartneck et al., 2021)

Ethical Principles for Trustworthy and Fair AI

AI4People 2018 (Floridi et al., 2018):

1. Beneficence: „Promoting Well-Being, Preserving Dignity, and Sustaining the Planet”
2. Non-maleficence: Privacy, Security and “Capability Caution”
3. Autonomy: The Power to Decide (Whether to Decide)
4. Justice: Promoting Prosperity and Preserving Solidarity.
5. Explicability: Enabling the Other Principles Through Intelligibility and Accountability

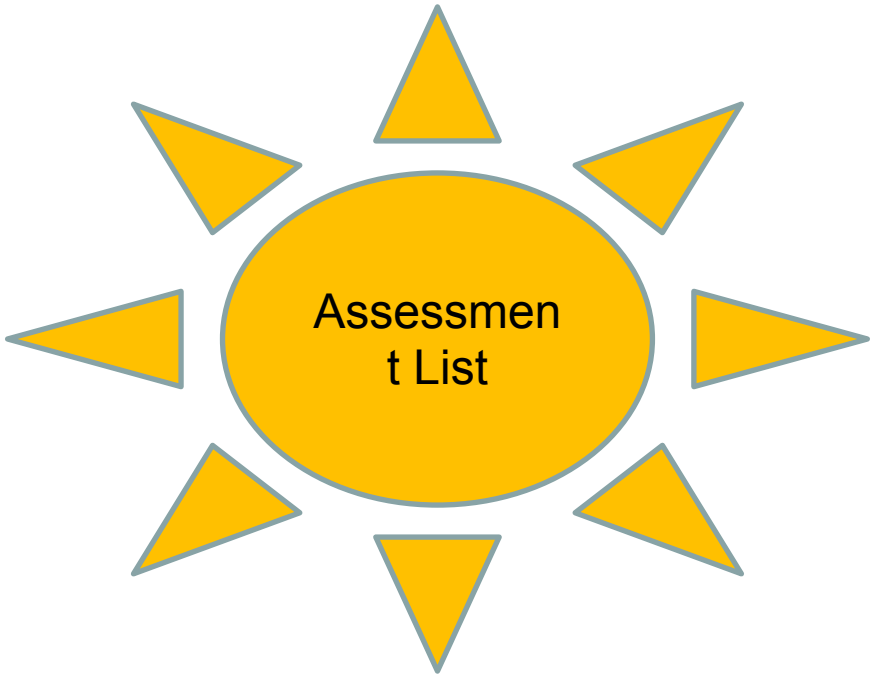
Ethics Guidelines for Trustworthy AI

- Presented by European Commission's High-Level Expert Group on AI on 8th April 2019.
- This states that "trustworthy AI should be:
  1. lawful - respecting all applicable laws and regulations
  2. ethical - respecting ethical principles and values
  3. robust - both from a technical perspective while taking into account its social environment"

  (Source: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, Accessed: 09.04.2023)

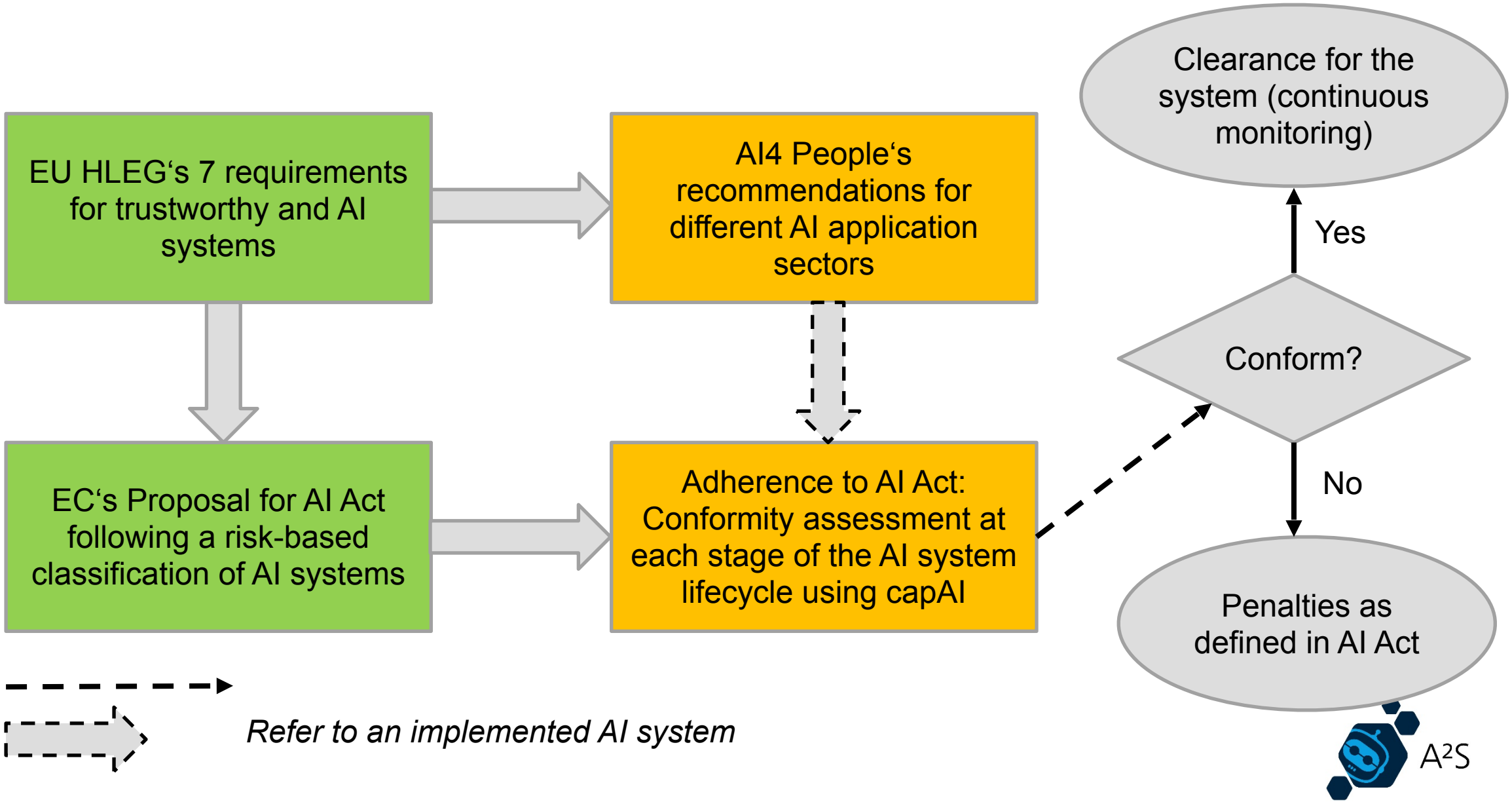
Realisation of Trustworthy AI

- **7 requirements** laid forth by European Commission‘s high-level expert group on AI, which should be addressed through technical and non-technical methods:
  1. Human agency and oversight
  2. Technical robustness and safety
  3. Privacy and data governance
  4. Transparency
  5. Diversity, non-discrimination and fairness
  6. Societal and environmental well-being
  7. Accountability



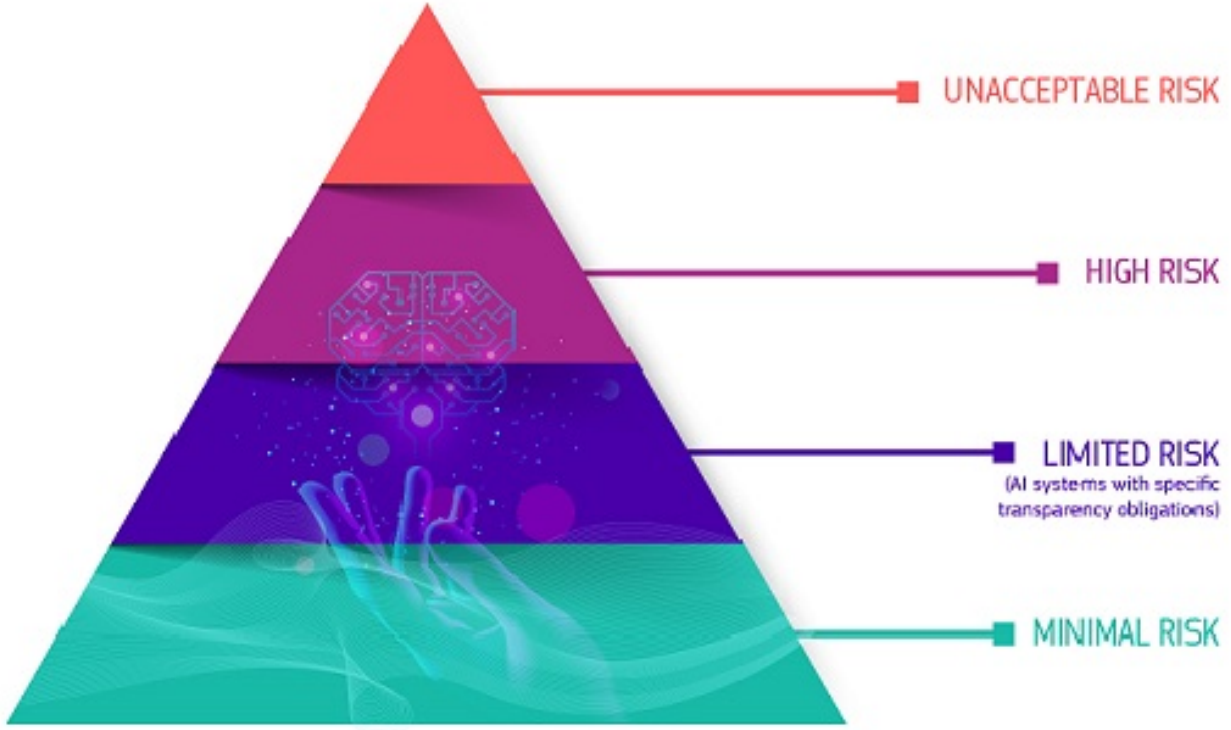
- Read more here:
  - https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
- Application to different application sectors:
  - https://2020.ai4people.eu/wp-content/pdf/AI4People7AIGlobalFrameworks.pdf

Trustworthy AI and Regulatory Framework in the EU



European Commission‘s Proposal for an AI Act

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



Source: https://ec.europa.eu/information_society/newsroom/image/document/2021-17/pyramid_7F5843E5-9386-8052-931F5C4E98C6E5F2_75757.jpg

Further Developments…

- capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act
  - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091

- European Commission‘s Liability Rules for AI
  - https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en

How Can We Measure User Acceptance?

- By conducting hypothesis-driven experimental research.
- Technology Acceptance Model (TAM) (Review: (Marangunić and Granić 2015)):

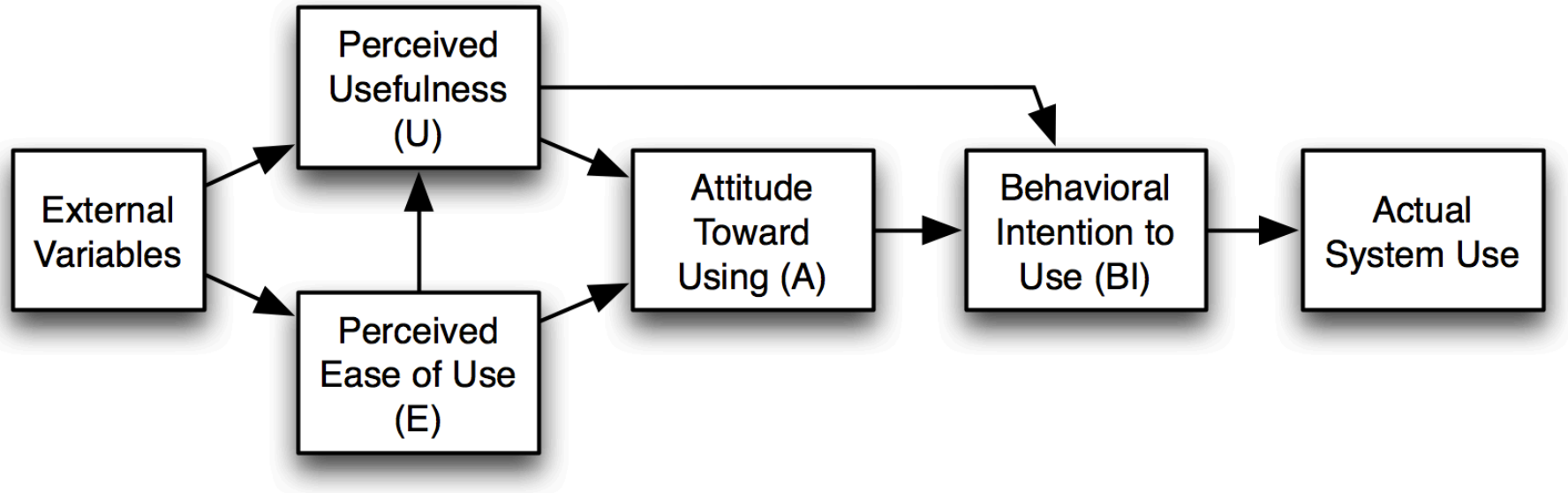


Image source: https://es.wikipedia.org/wiki/Modelo_de_aceptaci%C3%B3n_de_tecnolog%C3%ADa#/media/Archivo:Technology_Acceptance_Model.png
Licence: CC BY 3.0

Acceptance of Social Robots: (Heerink et al. 2009)

TABLE 1 – MODEL OVERVIEW

Code	Construct	Definition
ANX	Anxiety	Evoking anxious or emotional reactions when using the system.
ATT	Attitude	Positive or negative feelings about the appliance of the technology.
FC	Facilitating conditions	Objective factors in the environment that facilitate using the system.
ITU	Intention to use	The outspoken intention to use the system over a longer period in time.
PAD	Perceived adaptability	The perceived ability of the system to be adaptive to the changing needs of the user.
PENJ	Perceived enjoyment	Feelings of joy or pleasure associated by the user with the use of the system.
PEOU	Perceived ease of use	The degree to which the user believes that using the system would be free of effort
PS	Perceived sociability	The perceived ability of the system to perform sociable behavior.
PU	Perceived usefulness	The degree to which a person believes that using the system would enhance his or her daily activities
SI	Social influence	The user's perception of how people who are important to him think about him using the system
SP	Social presence	The experience of sensing a social entity when interacting with the system.
Trust	Trust	The belief that the system performs with personal integrity and reliability.
Use	Use/Usage	The actual use of the system over a longer period in time

These constructs can be measured using subjective questionnaires or objective task-relevant metrics.

Part 2: Personal Data, Privacy, GDPR

What is Personal Data?

- https://youtu.be/-_hLUi4AlNU

- Data that can help in direct or indirect identification of a person

- Direct: Name, address, phone no.
- Indirect: IP address

- Sensitive data: health data, biometric data, genetic data, sexual orientation, religion, ethnicity, political opinion.
  - In principle, processing of sensitive data is prohibited.
  - Explicit consent is mandatory.

https://www.gdprhandbook.eu/

Personal Data Processing

- *“Processing” is defined very broadly under the GDPR:*
  - *Any operation or set of operations;*
  - *Which is performed on personal data or on sets of personal data;*
  - *Whether or not by automated means.*

- *Recording – Collecting – Organising – Structuring – Adapting or altering – Storing – Retrieving – Consulting – Disclosing by transmission – Using – Disseminating or otherwise making available –Aligning or combining – Erasing or destructing – Restricting*

https://www.gdprhandbook.eu/

Who is Processing Personal Data?

- Questions to identify whether you are a data controller: https://www.gdprhandbook.eu/data-controller-processor
- Controller determines the purpose for which data is being collected.
- Processor performs processing of data on behalf of the controller and as per the controller‘s instructions.
- Both are responsible for GDPR compliance and have to take Technical and Organisational measures (TOMs) to enforce GDPR compliance.
- Data Controller has more responsibility than Data Processor.

What About Producers?

- “when developing, designing, selecting and using (. . .) products that are based on the processing of personal data or process personal data to fulfil their task, **producers** of the products (. . .) should be encouraged to take into account the right to data protection (. . .) with due regard to the state of the art, to make sure that controllers and processors are able to fulfil their data protection obligations” (Recital 78 S. 4. GDPR).

- Therefore, designers and producers should also take TOMs for data protection.

(Horstmann et al. 2020)

GDPR – Seven Principles

Lawfulness
Fairness
Transparency

Purpose limitation

Data minimisation

Accuracy

https://www.gdprhandbook.eu/

Accountability

Integrity and Confidentiality

Storage limitation

GDPR – Data Subject and Their Rights

- Data subject is the person whose personal data is being collected and processed.
- What rights do they have?
  - Right to be informed (inform)
    - For which purpose is the data corrected; which data has been collected; who will receive this data; for how long will it be stored, etc.
  - Right to rectification (correct)
  - Right to erasure (delete, forget)
  - Right to restrict processing (stop partially or fully)
  - Right of access (get a copy of their data)
  - Right to appeal, object

(Horstmann et al. 2020)
https://www.gdprhandbook.eu

GDPR Principles <->Data Subject Rights

Principles binding Controllers	Rights of Data Subjects
Lawfulness, fairness	Rights to appeal, object, erasure
Transparency	Right of access, right to be informed
Purpose limitation, data minimisation, storage limitation	Right to restrict processing, right to erasure
Accuracy	Right to rectification

(Horstmann et al. 2020)

Personal Data Protection and Privacy

- Privacy and Data Protection are not the same.
- When personal data is protected, it contributes to privacy protection.
- However, privacy protection is more than just personal data protection.

H-BRS Privacy Statement

- https://www.h-brs.de/en/data-privacy-statement

- Controller?
- Data Protection Officer?
- Purpose?
- Duration?
- Types of personal data collected?

User Studies

- Human-robot interaction experiments
- We need to recruit participants.

- Prepare an information and consent form.

- Inform about: purpose, duration, type of personal data, data protection officer.

- Anonymisation (remove identifiers)
  - Usually when you publish results.
- Pseudonymisation (replace identifiers)
  - Usually when you store results, especially to handle request for erasure, rectification, etc.

https://www.gdprhandbook.eu/

Recommendations and Resources

- The GUIDE Project
  - Fraunhofer IOSB and Bielefeld University of Applied Sciences and Arts
  - Funded by German Federal Ministry of Education and Research (BMBF)
  - Outcome: A set of guidelines on processing personal data, checklists, sample consent forms…
    - https://www.iosb.fraunhofer.de/content/dam/iosb/iosbtest/documents/kompetenzen/bildauswertung/IAD/projekte-und-produkte/GUIDELine.pdf

- Checklist from DFG for handling research data:
  - https://www.dfg.de/resource/blob/174736/92691e48e89bf4ac88c8eb91b8f783b0/forschungsdaten-checkliste-en-data.pdf

Conclusion

- In this lecture, you learned to:
  1. Explain why ethics is relevant in Human-Robot Interaction (HRI).
  2. Elucidate the concept of trust and explain the relation between ethics and trust.
  3. Identify the constructs involved in the acceptance of technology in general and assistive social robots in specific.
  4. Explain the principles and some of the important aspects of GDPR, and its application in HRI design and experiments.

References

- Bartneck C, Lütge C, Wagner A, Welsh S. An introduction to ethics in robotics and AI. Springer Nature; 2021.
- Floridi, L., Cowls, J., Beltrametti, M. *et al.* AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds & Machines* **28**, 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5
- Heerink, M., Kröse, B.J., Evers, V., & Wielinga, B.J. (2009). Measuring acceptance of an assistive social robot: a suggested toolkit. RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication, 528-533.
- B. Horstmann, et al., "Towards Designing Privacy-Compliant Social Robots for Use in Private Households: A Use Case Based Identification of Privacy Implications and Potential Technical Measures for Mitigation", *Proceedings of the 29th IEEE International Conference on Robot and Human Interactive Communication (Ro-Man 2020)*, 2020, pp.869–876.
- T. Kautonen and Karjaluoto, Eds., Trust and New Technologies: Marketing and Management on the Internet and Mobile Media. Edward Elgar, 2008.
- Lee, John D., and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46 (1): 50–80. https://doi.org/10.1518/hfes.46.1.50_30392. PMID: 15151155.
- Marangunić, N., Granić, A. Technology acceptance model: a literature review from 1986 to 2013. *Univ Access Inf Soc* **14**, 81–95 (2015). https://doi.org/10.1007/s10209-014-0348-1
- R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," The Academy of Management Review, vol. 20, no. 3, pp. 709–734, 1995.