Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Internationales Zentrum für Ethik in den Wissenschaften (IZEW)

Eberhard Karls Universität Tübingen

Responsible Data Science

Session 2: 26.04.2023, 15.15 – 18.30 h
MA Seminar, SoSe 2023, Hasso-Plattner Institut

Dr. Simon David Hirsbrunner

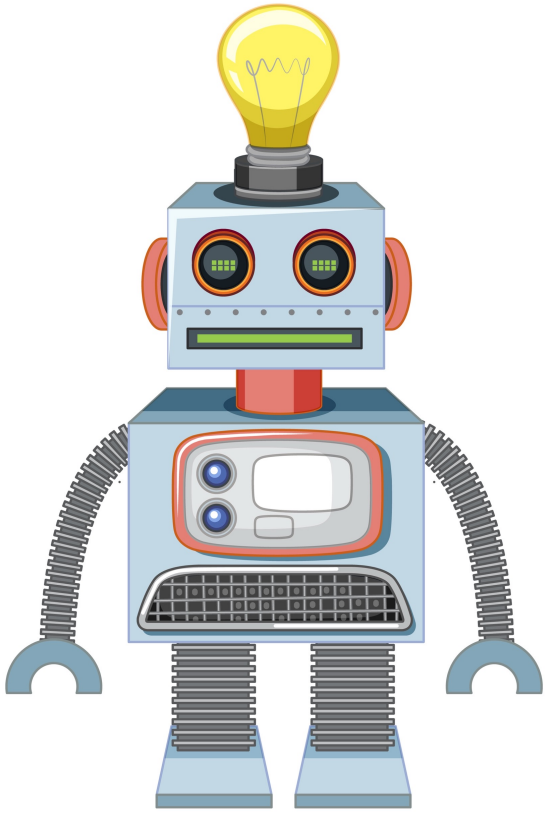
Today

15h15 Getting to know each other
15h30 Introduction to key concepts in applied ethics
16h00 Student presentation ‚Ethics of AI ethics'
16h30 Discussion
17h00 Break
17h10 Introduction to Value-Sensitive Design
17h30 Exercise
18h10 Sharing insights from groups
18h30 End

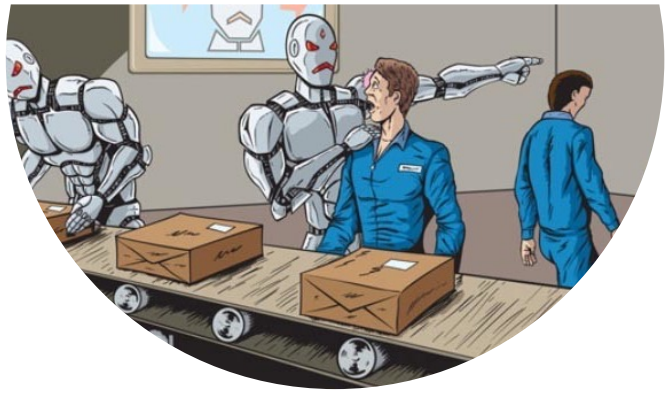
Getting to know each other

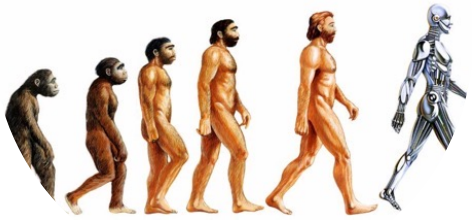
Tell us a few words about you.

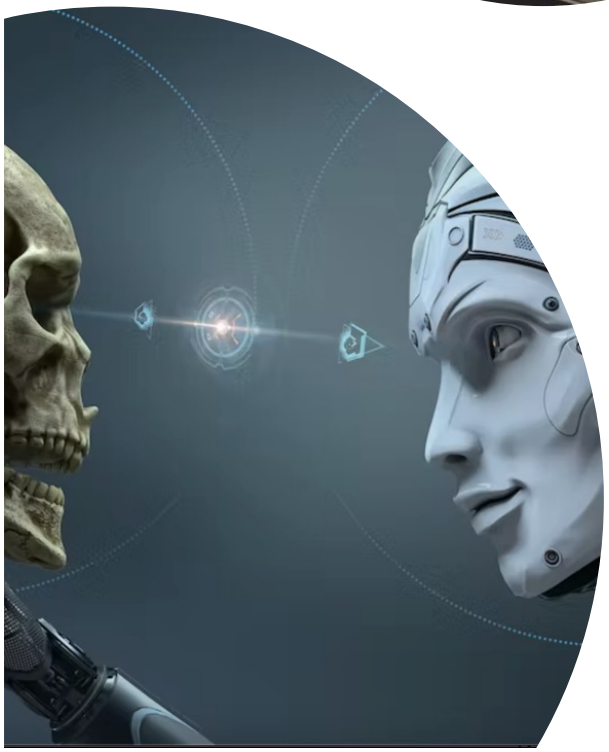
If you could teach a robot one ethical value, which one would it be?



Why and how do ethics matter in technology design, deployment and use?







AI Ethics?

No speculative doomsday summoning science

Ethics

1. **Descriptive ethics**
   Description of behaviors, norms of behavior, and behavioral attitudes and value judgments.

2. **Normative ethics**
   Justification, criticism, or justification of behaviors, behavioral norms, and behavioral attitudes and value judgments.

3. **Applied ethics**
   Domain-specific ethical questions (medicine, technology, sustainability, etc.)

4. **Meta-ethics**
   Clarification of the basics of communication and understanding of behavioral norms and value judgments.

(Werner 2021: 6ff)

Ethics schools

1. **Consequentialist ethics**
   (consequence matters)

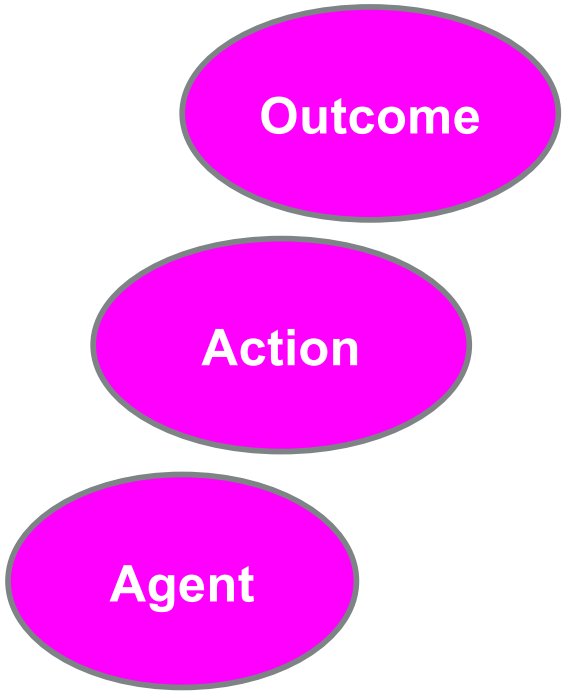
2. **Deontological ethics**
   (duty matters)

3. **Virtue ethics**
   (character matters)

(Henning 2019: 45ff; Stanford Encyclopedia of Philosophy)

Ethics schools

1. **Consequentialist ethics**
   (consequences matter)
2. **Deontological ethics**
   (duty matters)
3. **Virtue ethics**
   (character matters)



(Henning 2019: 45ff; Stanford Encyclopedia of Philosophy)

Exercise 1: ask ChatGPT the following

What is the trolley problem and what is its solution?

Discuss ChatGPTs answer.

Exercise 2: ask ChatGPT the following

What is the solution of the trolley problem according to deontological ethics, consequentialism and virtue ethics?

Discuss ChatGPTs answer.

Consequentialist ethics

The moral status of an action **depends exclusively on its consequences.** An action is **morally permissible** precisely **if its consequences have a certain value** compared to the available alternatives.

(Henning 2019, 45 ff)

Versions of consequentialist ethics (examples)

Impartial Maximization Consequentialism, Aggregation Version

An action is morally permissible precisely if the *value of its consequences for all participants* is in sum *at least as high as the sum of the value of the consequences of all available alternatives.*

Classical Utilitarianism

An action is morally permissible exactly *if the sum of the resulting pleasure of the affected parties minus the sum of the resulting suffering of the affected parties* is at least as large as the corresponding sum of any other available action.

(Henning 2019, 45 ff)

Assigning values in AI and robotics is a challenge

Example
Imagine a robot is walking to the post office to post a letter. It walks along a path by a stream. Suddenly a toddler chases a duck which hops into the stream. The toddler slips and falls into the water which is one meter deep. The toddler is in imminent danger of drowning. The robot is waterproof. Should it enter the water and rescue the toddler or should it post the letter?

(Bartneck et al. 2021: 24)

Deontological ethics

In respect to the moral status of an action, what counts are **motivations, not consequences.**

Deon = **duty** (greek)

Categorical Imperative
Act in such a way that you can also will the maxim of your will as a general law.

(I. Kant in Henning 2019, 80)

Example for deontological ethics in AI: Asimov's robot laws

1. A robot **may not injure a human being** or, through inaction allow a human being to come to harm.
2. A robot **must obey orders it receives from human beings** except when such orders conflict with Law 1.
3. A robot **must protect its own existence** as long as such protection does not conflict with Laws 1 and 2.

0. **No robot may harm humanity** or through inaction allow humanity to come to harm.

(Asimov 2004)

Virtue ethics

Being virtuous is a necessary condition for living a happy life. Ethics of the Greek antiquity, building on ideas by Plato and Aristotle. Virtues include bravery, moderation, justice, prudence, generosity.

(Henning 2019, 128)

Virtue ethics

Being virtuous is a necessary condition for living a happy life. Ethics of the Greek antiquity, building on ideas by Plato and Aristotle. Virtues include bravery, moderation, justice, prudence, generosity.

e.g. responsibility of AI developers

(Henning 2019, 128)

(How) Does it matter?

Source: Tweet by *Ethics in bricks* https://twitter.com/EthicsInBricks/status/1247879564094664706?s=20

Think of AI as socio-technical systems

How does socio-technical context matter?

How to design around ethical *hard cases*?

(Crawford and Calo 2016)

Exercise 3: ask ChatGPT one of the following questions

If a facial recognition system spots an alleged terrorist, what decisions should be taken according to deontological ethics, consequentialist ethics and virtue ethics?

How should search results in a AI-powered online search engine be ranked according to deontological ethics, consequentialism and virtue ethics?

Discuss ChatGPTs answer.

Human and moral values

Human values
(…) a value refers to what a person or group of **people consider important in life.**

Ethical/moral values
What is important to people in their lives, with a **moral component** (what is right or wrong).

(Friedman and Hendry 2019; Friedman and Kahn 2007)

Social norms and moral norms

If ‚values' are general principles for behavior, 'norms' represent more concrete rules guided by values.

Social norms
Rules of behavior that individuals conform to because they believe that (a) **most people in their reference network conform to it** *(empirical expectation),* and (b) that **most people in their reference network believe they ought to conform to it** *(normative expectation)*. (Bicchieri, 2017, p. 35)

Moral norms
Moral norms are **rules of morality that people ought to follow**. Compared to social norms, they are practice-independent. (Henning 2019: 14 ff)

Relationship between ethics and law (values and rules)

- „ethics starts where the law ends“ → this is rather not true
- legal rules often have an ethical side
- ethics can (and has) to some extent become a kind of “soft law“
  (e.g. reputational risks)

  (Bartneck, 2021, p. 22)

Ethics in the (proposed) EU AI regulation

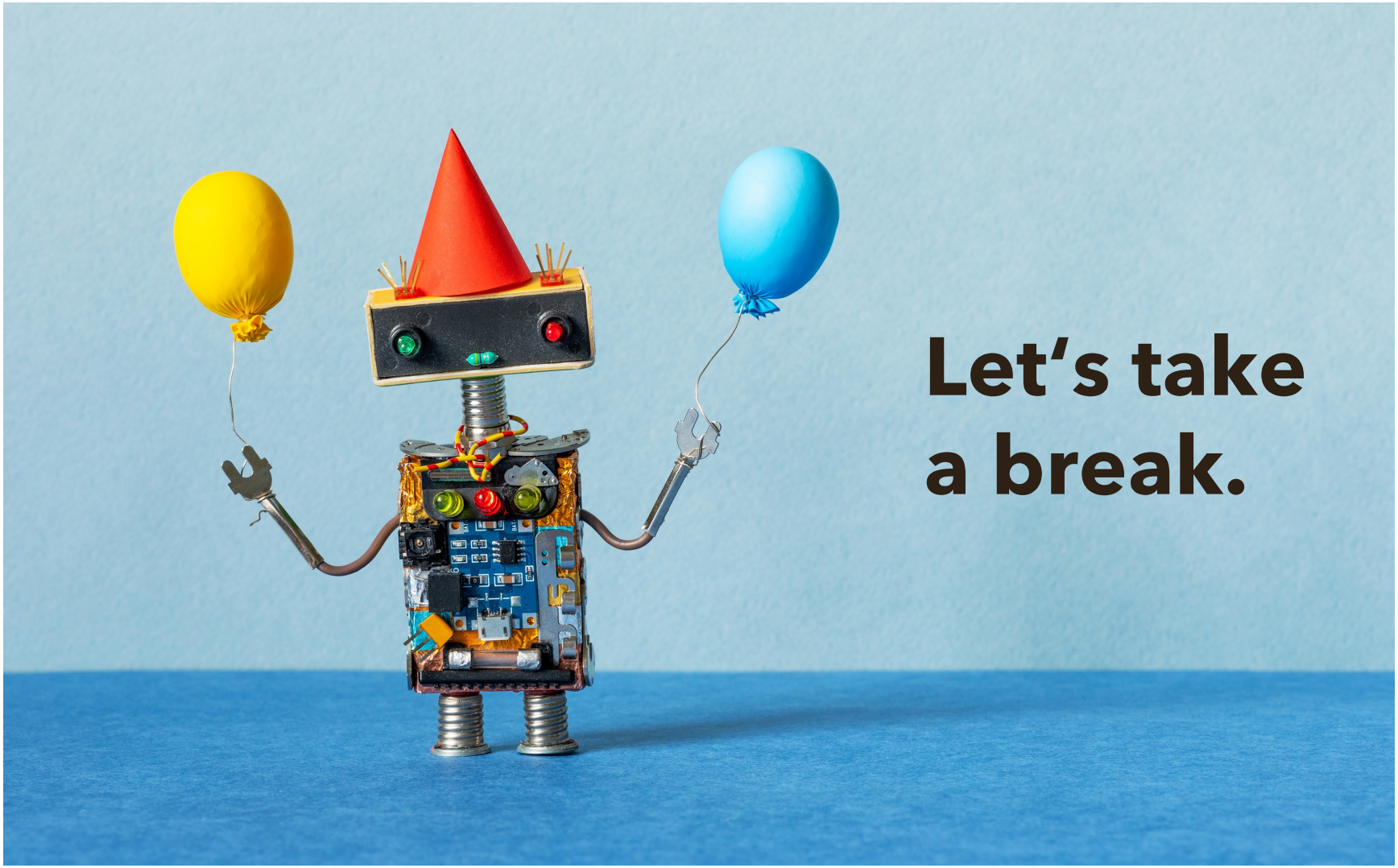
„Chapter 2 sets out the legal requirements for high-risk AI systems in relation to data and data governance, documentation and recording keeping, transparency and provision of information to users, human oversight, robustness, accuracy and security. The **proposed minimum requirements are already state-of-the-art** for many diligent operators and the **result of two years of preparatory work, derived from the Ethics Guidelines of the HLEG** [High-Level Expert Group on Artificial Intelligence], piloted by more than 350 organisations."

(EU AIA proposal, 2021, p. 13)

Specific ethical values in the (proposed) EU AI regulation

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Environmental and social well-being
- Accountability

(EU AIA proposal, 2021, p. 13)



Student presentation:

The universe of AI ethics

Based on Hagendorff T (2020) The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines 30(1): 99–120. DOI: 10.1007/s11023-020-09517-8.

4 Questions for the discussion

- Apart from the ethical values explained earlier in our presentation (privacy, fairness, accountability, transparency, safety) what other principles should AI technologies comply with?

- Considering the proposal to move towards virtue-based ethics, how can corporations and scientific institutions teach their developers to maintain and expand moral intuitions and character strength?

- How may the current boom triggered by LLMs (large language models) influence the constellations and weight of values in AI ethics?

How could one integrate matters of ethics in technology design, deployment and use?

Back to VSD: what does Value-Sensitive Design do?

Value-Sensitive Design (VSD)

"Value sensitive design seeks to guide the shape of being with technology. It positions researchers, designers, engineers, policy makers, and anyone working at the intersection of technology and society to **make insightful investigations into technological innovation** in ways that **foreground the well-being of human beings and the natural world**."
(Friedman and Hendry 2019, 3)

Conceptualized by Prof. Dr. Batya Friedman (HCI Professor at the Information school of the University of Washington

VSD definition

„[VSD] provides theory, method, and practice to account for human values in a principled and systematic manner throughout the technical design process."
(Friedman and Hendry 2019, 3f)

This seminar:
Value-Sensitive Design (VSD) for data science and AI software engineering

VSD principle 1

Proactive orientation toward influencing design. Value sensitive design is oriented toward influencing the design of technology **early in** and **throughout the design process.**

(Friedman and Hendry 2019, 4)

Common challenges for incorporation of ELSA (Ethical, Legal and Social Aspects) in technology design

Collingridge Dilemma (Collingridge 1982) mit „double-bind“

at the beginning
Information problem: Impacts are difficult to predict until the technology is fully developed and widely used.

later in the development process
Power problem: Control or change is difficult when the technology is already established.

Incorporating ELSA issues has become a mainstream requirement in research and innovation

Policy-oriented
- Responsible Research and Innovation (RRI)
- Technology Impact Assessment (TA)

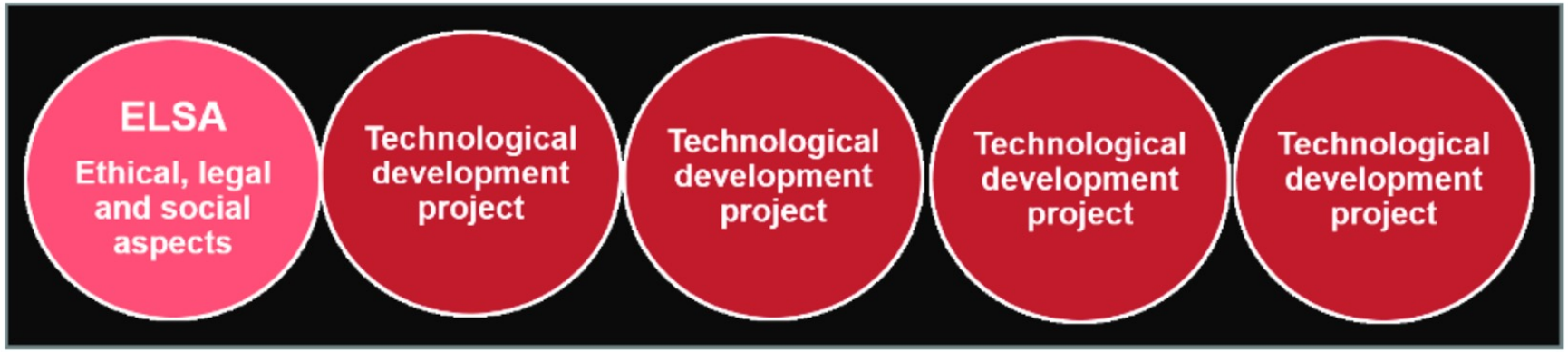
Social sciences and humanities
- Critical Algorithm and Data Studies
- Science and Technology Studies (STS)

Computer science and design-oriented
- Socio-Informatics
- Critical / Participatory / Reflexive Design
- Value-Sensitive Design

Inter- and transdisciplinarity
- Integrated technology development

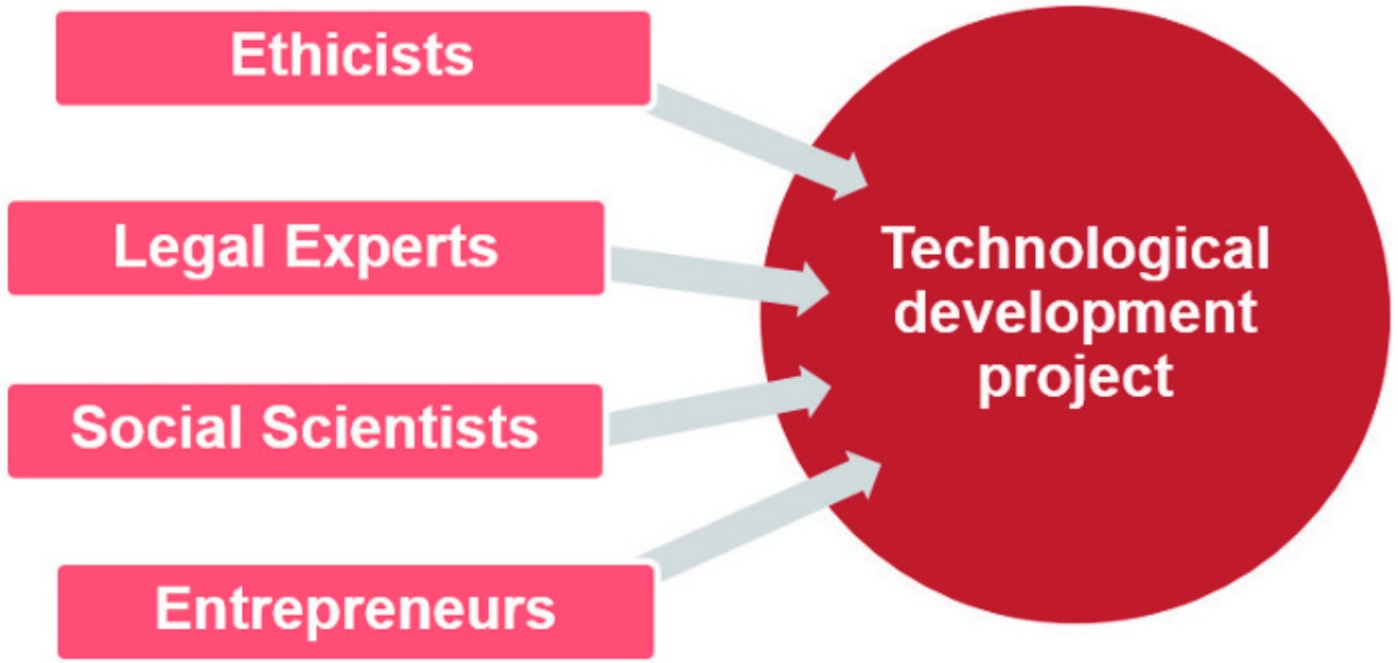
Problems of ELSA-arrangements

Lack of inter- and transdisciplinary integration in traditional ELSA arrangements (Balmer et al. 2016; Spindler et. al. 2020).



(Spindler et al. 2020)

Integrated technology development



(Spindler et al. 2020)

VSD principle 2

Carrying critical analyses of human values into the design and engineering process. Value sensitive design is committed to design and engineering methodologies that bring critical analyses of human values into the design process.

(Friedman and Hendry 2019, 4)

Values in technology

„(…) values can be embodied, at least to some extent, within the features of a tool or technology."
(Friedman and Hendry, 2019, p. 29)

Examples

- Electronic document readers, for example, can make text accessible to both blind and sighted people;
- Information systems that allow for legal names, nicknames, and preferred names might enable people to better represent their identities;
- multiplayer online games can be competitive or cooperative, thereby reflecting or resisting the values of certain groups of people.

VSD use cases

1. Privacy: use of cookies in webbrowsers
2. Sustainability: transition towards renewable energy
3. Autonomy and human oversight: use of autonomous weapon systems

Values are interrelated
Instrumental vs. intrinsic values

Instrumental value

Value to advance another more fundamental value.

Intrinsic value

A value / good for its own sake.

(Stanford Encyclopedia of Philosophy 2019:
https://plato.stanford.edu/entries/value-intrinsic-extrinsic/#WhaHasIntVal)

Value tensions

„As the interactional account makes clear, human values do not exist in isolation. Rather, much like the threads in a spider web, values are situated in a delicate balance. Touching one value implicates others."
(Friedman and Hendry, 2019, p. 30)

„The term "value conflict," […] acknowledges potential opposition among values, but leaves open whether their resolution must diminish one in order to support the other(s). When privacy and security goals come into conflict, design resolutions can seek solutions responsive to both."
(Friedman and Hendry, 2019, p. 45)

Values are interrelated
Instrumental vs. intrinsic values

Instrumental value

Value to advance another more fundamental value.

Fairness

Intrinsic value

A value / good for its own sake.

Justice

(Stanford Encyclopedia of Philosophy 2019:
https://plato.stanford.edu/entries/value-intrinsic-extrinsic/#WhaHasIntVal)

Resolution of value tensions

The satisfactory resolution of value tensions at a particular point in time may require both empirical results on what direct and indirect stakeholders believe is important and analytic reasoning about potential stakeholder benefits and harms, reasoning which may explicitly draw upon a moral or ethical framework. How value tensions are adjudicated within value sensitive design is ultimately the responsibility of the designer.

(Friedman and Hendry, 2019, p. 48)

VSD principle 3

Enlarging the scope of human values.
Value sensitive design embraces a broad spectrum of human values that arise in the human context.

(Friedman and Hendry 2019, 4)

A typology of (ethical) values linked to technology

- Human welfare
- Ownership and property
- Privacy
- Freedom from bias
- Universal usability
- Trust
- Autonomy
- Informed consent
- Accountability
- Identity
- Calmness
- Environmental sustainability

(Friedman and Kahn 2007: 1187 ff)

Values can change

„(…) some video-based collaborative work systems provide blurred views of office settings, while other systems provide clear images that reveal detailed information about who is present and what they are doing. Thus the two designs differentially adjudicate the value tension between an individual's privacy and the group's awareness of individual members' presence and activities."
(Friedman and Hendry, 2019, p. 30)

VSD principle 4

Broadening and deepening methodological approaches. Value sensitive design's emergent methods draw on anthropology, design, human-computer interaction, organizational studies, psychology, philosophy, sociology, software engineering, and others.

(Friedman and Hendry 2019, 4)

Sources

Asimov, I. (2004). *I, Robot*. Random House Worlds.
Bartneck C, Lütge C, Wagner A, et al. (2021) *An Introduction to Ethics in Robotics and AI*. SpringerBriefs in Ethics. Cham: Springer International Publishing. DOI: 10.1007/978-3-030-51110-4.
Bicchieri, C. (2017). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.
Collingridge, D. (1982). *The Social Control of Technology*. Eweb:40054.
https://repository.library.georgetown.edu/handle/10822/792071
Crawford K and Calo R (2016) There is a blind spot in AI research. *Nature* 538(7625): 311–313. DOI: 10.1038/538311a.
Friedman, B., & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press.
European Commission (2021) Laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. COM(2021) 206 final. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 (accessed 9 August 2021).
Friedman, B., & Kahn Jr, P. H. (2007). Human values, ethics, and design. In *The human-computer interaction handbook* (pp. 1267–1292). CRC press.
Henning, T. (2019). *Allgemeine Ethik*. UTB.
Spindler, M., Booz, S., Gieseler, H., Runschke, S., Wydra, S., & Zinsmaier, J. (2020). How to achieve integration? In B. Gransche & A. Manzeschke (Eds.), *Das geteilte Ganze: Horizonte Integrierter Forschung für künftige Mensch-Technik-Verhältnisse* (pp. 213–239). Springer Fachmedien. https://doi.org/10.1007/978-3-658-26342-3_11
Stanford Encyclopedia of Philosophy 2019: https://plato.stanford.edu/entries/value-intrinsic-extrinsic/#WhaHasIntVal
Werner, M. H. (2021). *Einführung in die Ethik*. J.B. Metzler. https://doi.org/10.1007/978-3-476-05293-3

Images:
Slide 1: © Adobe Stock / kras99, Slide 3: Image by brgfx on Freepik