Lecture 1

Some definitions

Technology – application of knowledge for practical goals, immaterial (eg.alphabet, AI) and material(eg.hammer, robots)

Robot - embodied machines that have sensors and actuators, that possess some degree of artificial intelligence and functional autonomy, which enable the robot to perform certain tasks that humans would perform" (Nyholm, 2018, p. 9)

AI (general or strong) –machine ability to perform a wide range of tasks in different domains in ways that seemingly exhibit a humanlike or even more impressive intelligence

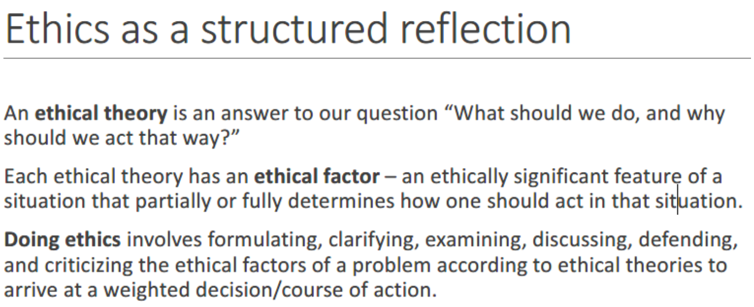
Different perspectives on technology

1. Instrumentalist- technologyis neutral, a means to a humanly-determined end:"Guns don't kill people, people kill people"

2. Deterministic –technology follows a predefined trajectory of its own, people have no influence on it: "Robots will take over our jobs!"

3. Interactionist –technology is created by people, embedded in interaction with people and in turn co-creates society

What is ethics

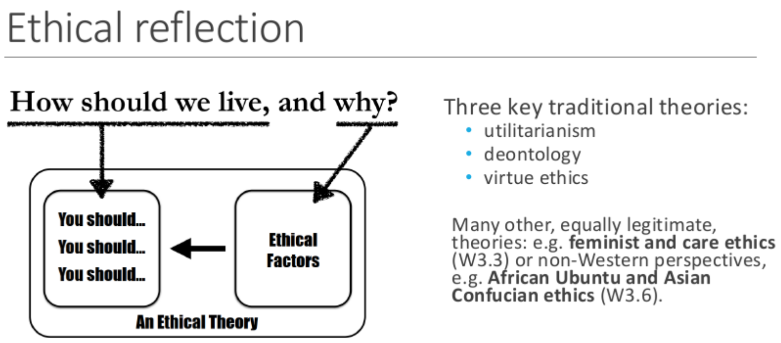


Ethics as a structured reflection

An **ethical theory** is an answer to our question "What should we do, and why should we act that way?"

Each ethical theory has an **ethical factor** – an ethically significant feature of a situation that partially or fully determines how one should act in that situation.

Doing ethics involves formulating, clarifying, examining, discussing, defending, and criticizing the ethical factors of a problem according to ethical theories to arrive at a weighted decision/course of action.

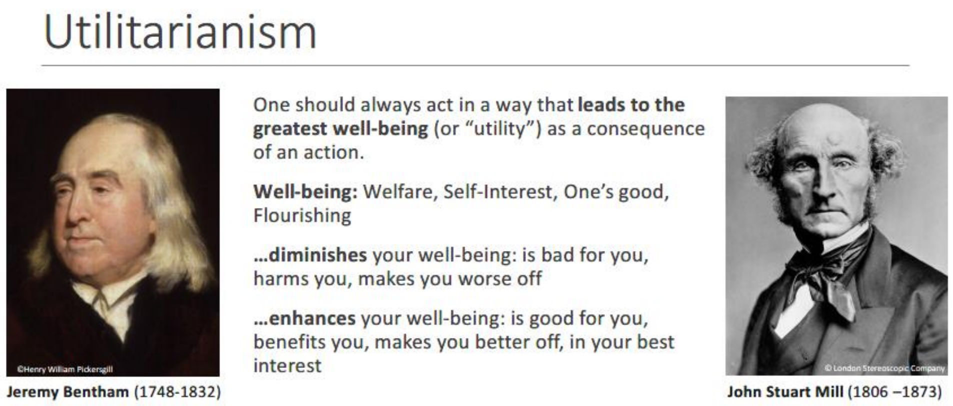


Ethical reflection

Three key traditional theories:

- utilitarianism
- deontology
- virtue ethics

Many other, equally legitimate, theories: e.g. **feminist and care ethics** (W3.3) or non-Western perspectives, e.g. **African Ubuntu and Asian Confucian ethics** (W3.6).



Utilitarianism

One should always act in a way that **leads to the greatest well-being** (or "utility") as a consequence of an action.

Well-being: Welfare, Self-Interest, One's good, Flourishing

…diminishes your well-being: is bad for you, harms you, makes you worse off

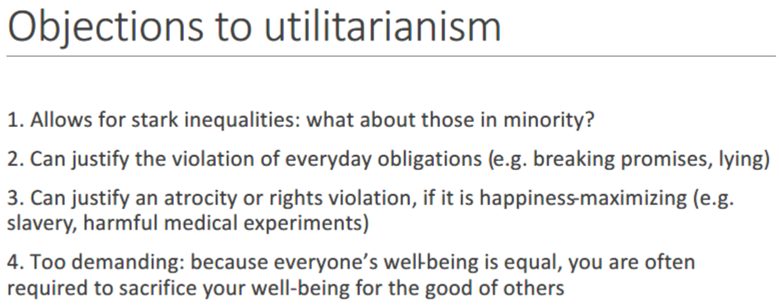
…enhances your well-being: is good for you, benefits you, makes you better off, in your best interest

©Henry William Pickersgill

Jeremy Bentham (1748-1832)

© London Stereoscopic Company

John Stuart Mill (1806 –1873)



Objections to utilitarianism

1. Allows for stark inequalities: what about those in minority?

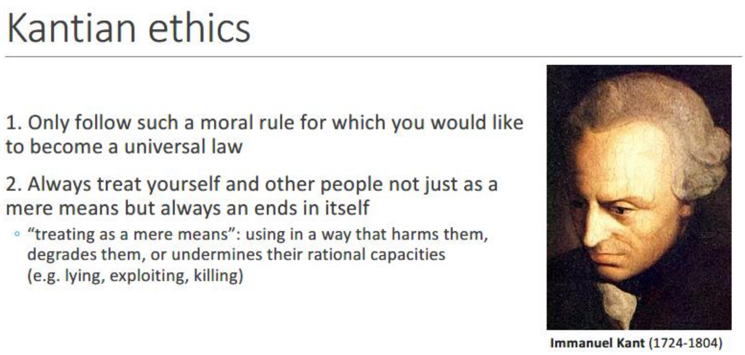
2. Can justify the violation of everyday obligations (e.g. breaking promises, lying)

3. Can justify an atrocity or rights violation, if it is happiness-maximizing (e.g. slavery, harmful medical experiments)

4. Too demanding: because everyone's well-being is equal, you are often required to sacrifice your well-being for the good of others

Deontology

Not concerned with the consequences but with **doing the right thing** under a series of rules, e.g. rights and duties.



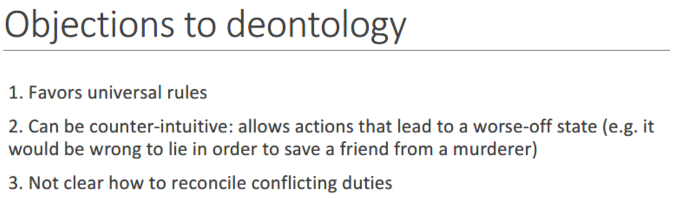
Kantian ethics

1. Only follow such a moral rule for which you would like to become a universal law

2. Always treat yourself and other people not just as a mere means but always an ends in itself

- "treating as a mere means": using in a way that harms them, degrades them, or undermines their rational capacities (e.g. lying, exploiting, killing)

Immanuel Kant (1724-1804)

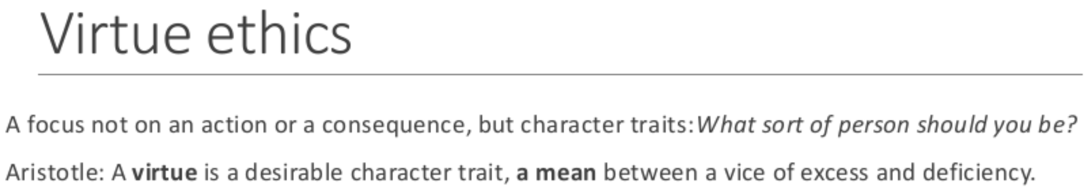


Objections to deontology

1. Favors universal rules

2. Can be counter-intuitive: allows actions that lead to a worse-off state (e.g. it would be wrong to lie in order to save a friend from a murderer)

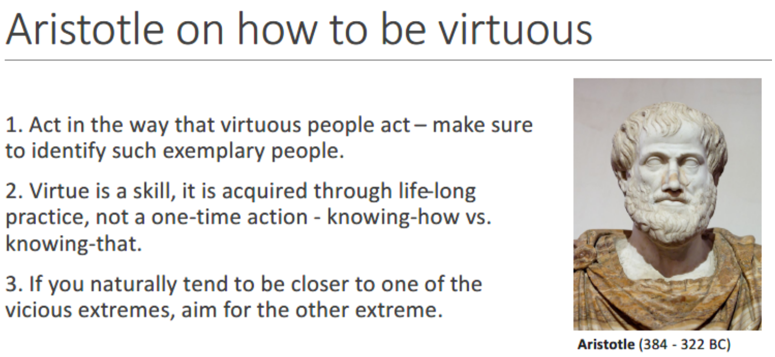
3. Not clear how to reconcile conflicting duties



Virtue ethics

A focus not on an action or a consequence, but character traits: *What sort of person should you be?*

Aristotle: A **virtue** is a desirable character trait, **a mean** between a vice of excess and deficiency.



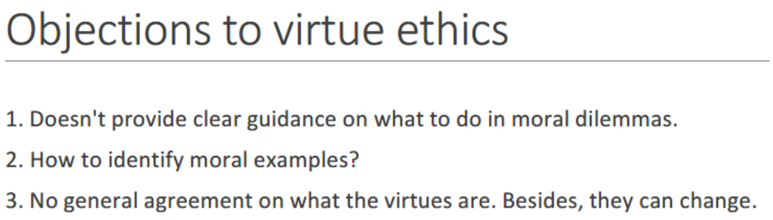
Aristotle on how to be virtuous

1. Act in the way that virtuous people act – make sure to identify such exemplary people.

2. Virtue is a skill, it is acquired through life-long practice, not a one-time action - knowing-how vs. knowing-that.

3. If you naturally tend to be closer to one of the vicious extremes, aim for the other extreme.

Aristotle (384 - 322 BC)



Objections to virtue ethics

1. Doesn't provide clear guidance on what to do in moral dilemmas.

2. How to identify moral examples?

3. No general agreement on what the virtues are. Besides, they can change.

Obligatory reading

3.1.1

What is a robot

What is human being

Are human being unfit for the future?

We should either adopt AI to fit us or adapt human to fit robot.

Four aspects of mind: “mind-reading” “dual processing” “tribal tendencies” “laziness”

3.1.2

The definition of “agent” “autonomy” “intelligence”

Overestimate of AI risks.

Underestimation of AI risks. Automatic weapon and facial recognition. insurance

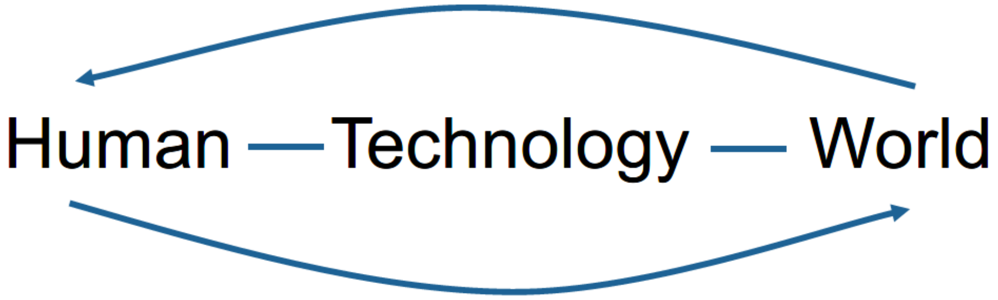
Make machine moral. Model ethical reasoning. Ethical attitude and values. Intrinsic limitaions

Predictive science. CDS pose a threat on scientific understanding. Crisis of Causal knowledge

Sociotechnical system. We should design AI to help us to act more ethically.

Lecture 2

Technological mediation approach



Subject (Human) and Object (World) are not separate from each other, but **co-shape each other** through mediation (Technology)

How can technology influence morality?

In at least three ways:

- By co-shaping moral perceptions
- By co-shaping moral actions
- By co-shaping – and changing - values

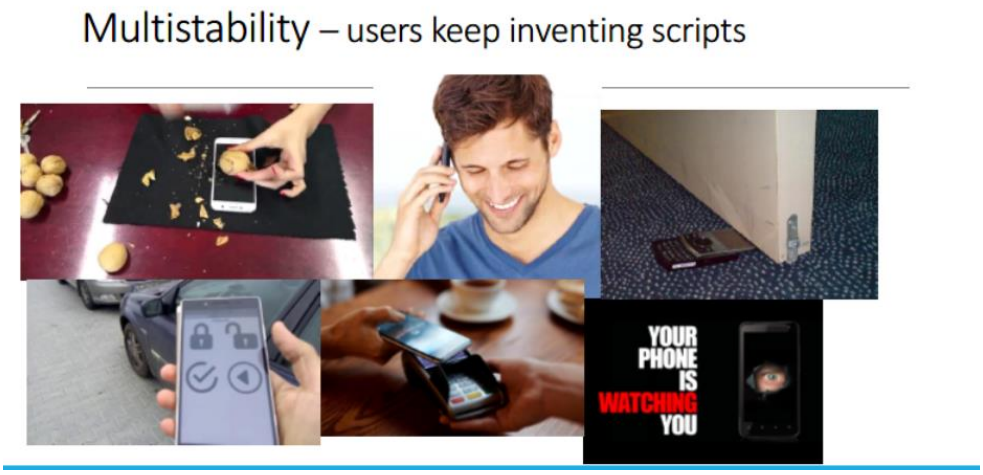
Non-neutrality of technology:

Co-shaping perceptions and actions

Amplification/reduction of certain aspects of reality

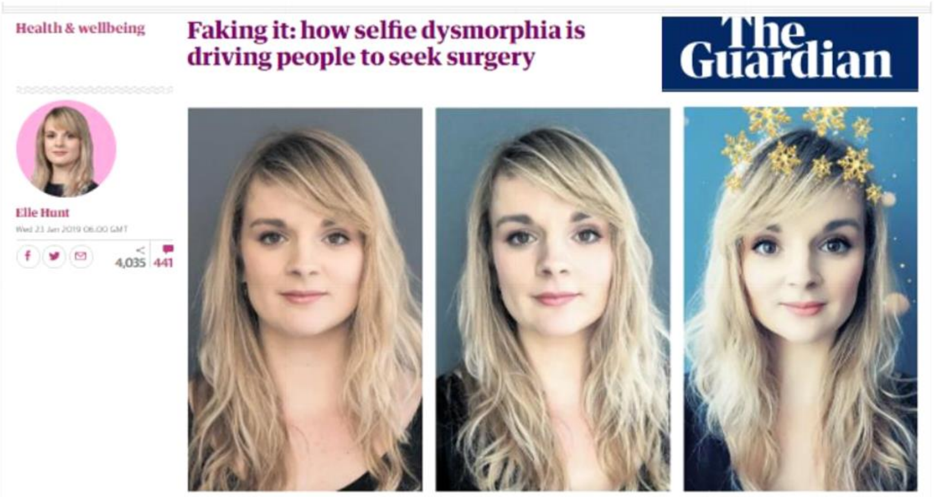
Invitation/inhibition - suggesting certain actions/making alternatives less appealing

@OlyaKudina



Multistability – users keep inventing scripts

Moral perceptions



Health & wellbeing

Faking it: how selfie dysmorphia is driving people to seek surgery

The Guardian

Elle Hunt

Wed 23 Jan 2019 06.00 GMT

4,035 441

Moral actions

Amazon use AI to select candidates and pre-score they before interview.

Moral value

Co-shape the definition of privacy.

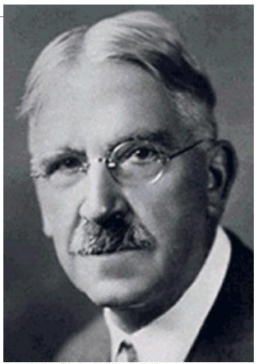
What is value

Practice-based approach to values

Inspired by pragmatism of John Dewey (1859-1952), **values** as the evaluative devices that originate in and guide our practices.

Values as:

1. Lived realities
2. Interactive with their context
3. Dynamic



Retrieved from https://ffrf.org/news/day/dayitems/item/14605-john-dewey

Values as ends-in-view and evaluative devices

Values - evaluative devices that carry over from earlier experiences and are (to some extent) shared in society. Present values offer a generalized response to earlier encountered problematic situations.

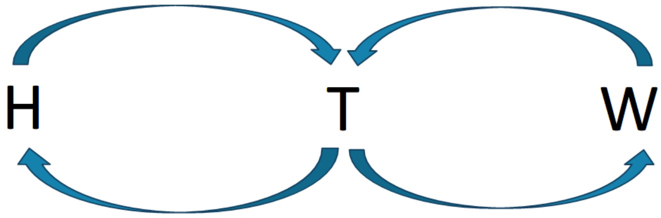
Values – ends-in-view: are always provisional and means to new future ends (not final ends).

Functions of values

- Help to discover what is (morally) salient in a situation
- Help to normatively evaluate situations
- Provide clues/guidance for action: existing values act as hypotheses to be verified

Sociotechnical system

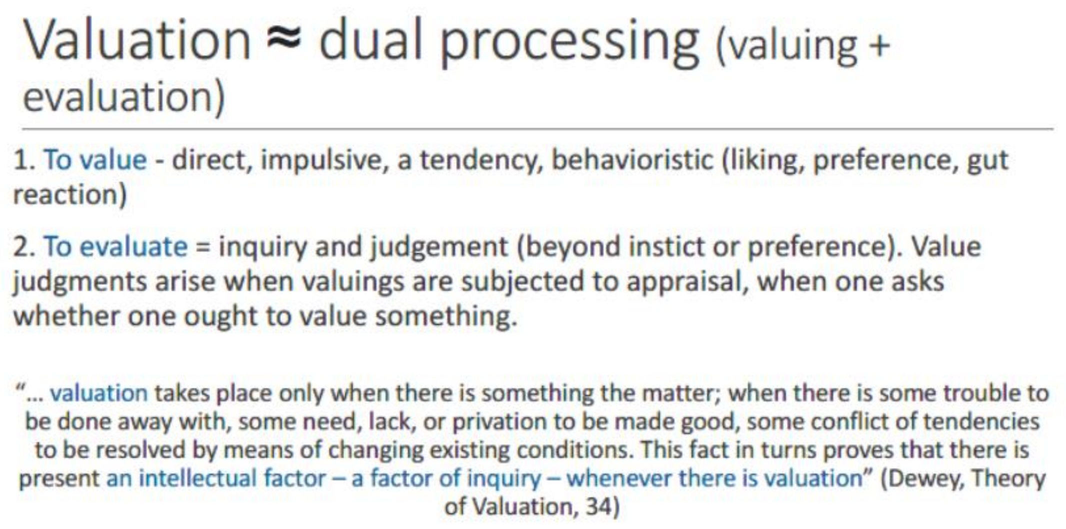
Values and robotic/AI sociotechnical systems



H – social actors, individual and group level, intelligent agents

T – technology, traditional and artificial agents

W – world, sociocultural, institutional norms, traditions, rules, expectations, priorities, etc.



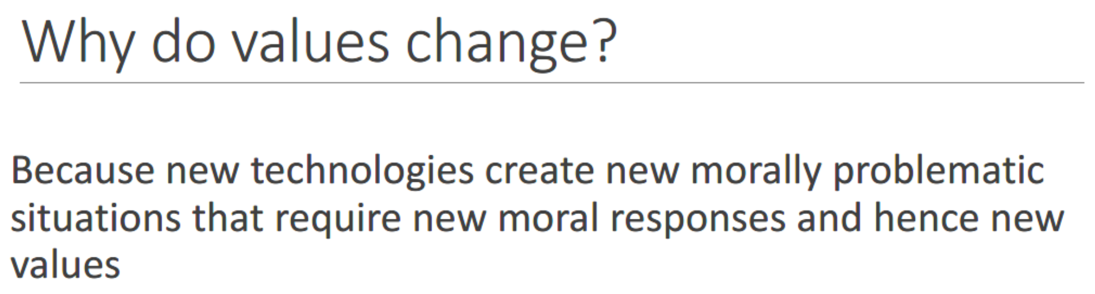
Valuation ≈ dual processing (valuing + evaluation)

1. To value - direct, impulsive, a tendency, behavioristic (liking, preference, gut reaction)

2. To evaluate = inquiry and judgement (beyond instict or preference). Value judgments arise when valuings are subjected to appraisal, when one asks whether one ought to value something.

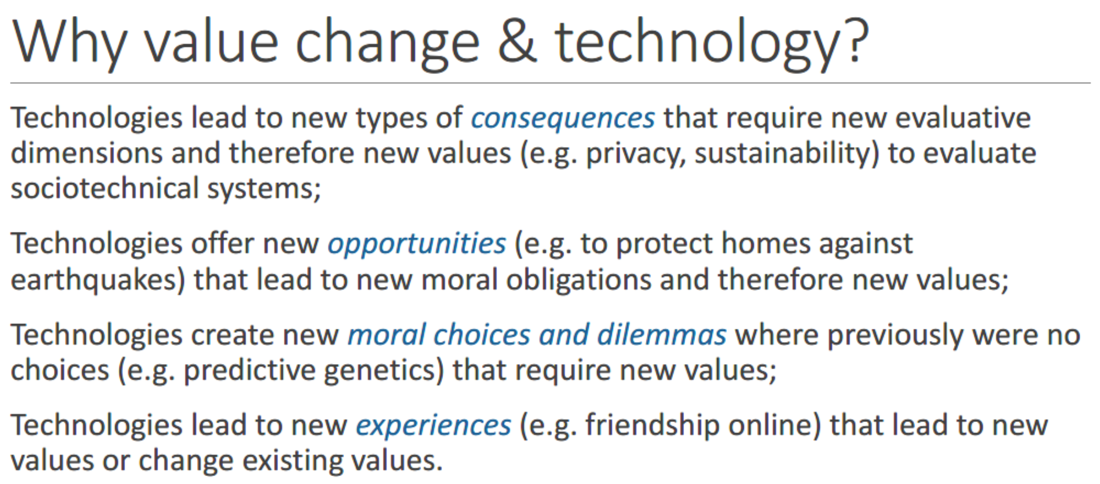
"… valuation takes place only when there is something the matter; when there is some trouble to be done away with, some need, lack, or privation to be made good, some conflict of tendencies to be resolved by means of changing existing conditions. This fact in turns proves that there is present an intellectual factor – a factor of inquiry – whenever there is valuation" (Dewey, Theory of Valuation, 34)

Why do value change



Why do values change?

Because new technologies create new morally problematic situations that require new moral responses and hence new values



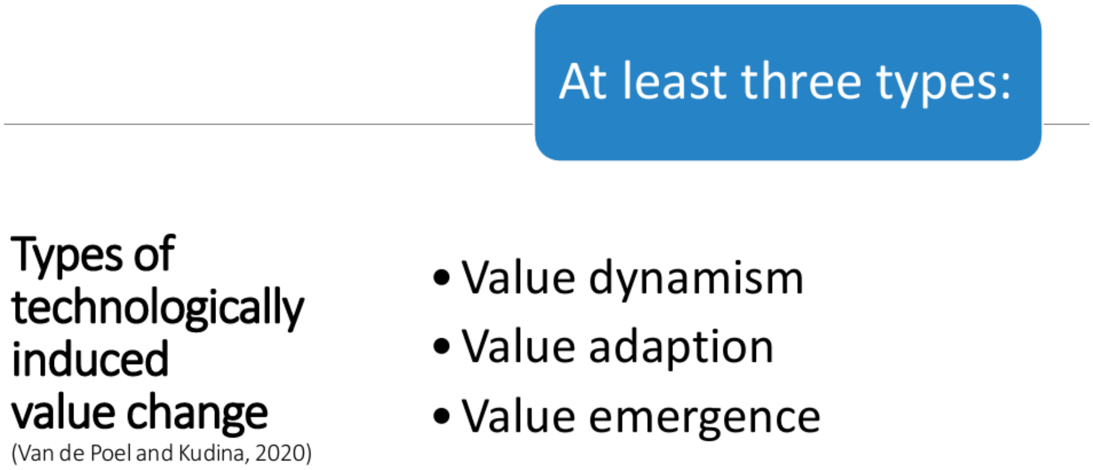
Why value change & technology?

Technologies lead to new types of *consequences* that require new evaluative dimensions and therefore new values (e.g. privacy, sustainability) to evaluate sociotechnical systems;

Technologies offer new *opportunities* (e.g. to protect homes against earthquakes) that lead to new moral obligations and therefore new values;

Technologies create new *moral choices and dilemmas* where previously were no choices (e.g. predictive genetics) that require new values;

Technologies lead to new *experiences* (e.g. friendship online) that lead to new values or change existing values.



At least three types:

Types of technologically induced value change

(Van de Poel and Kudina, 2020)

- Value dynamism
- Value adaption
- Value emergence

Value dynamism

The reaction and opinion on the same robot is different in Japan and Europe.
The definition of value privacy changes in different situation.

Value adaption

ICT technology tries to remember things. Nowadays, people try to be forgotten on the internet.

Value emergency

reading

3.2.1

Technological mediation.

3.2.2

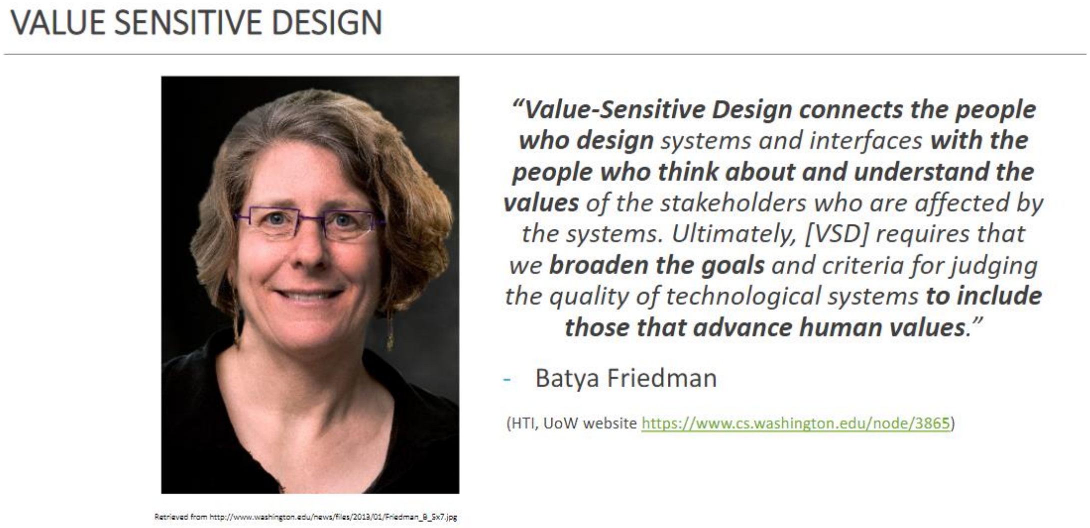
Value change taxonomy(emergency of value; value change in how value is conceptualized).
Technical features that allow better dealing with value change(adaptation, robust, flexibility).

3.2.3

Autonomy-safety-paradox.

Lecture 3

What is VSD



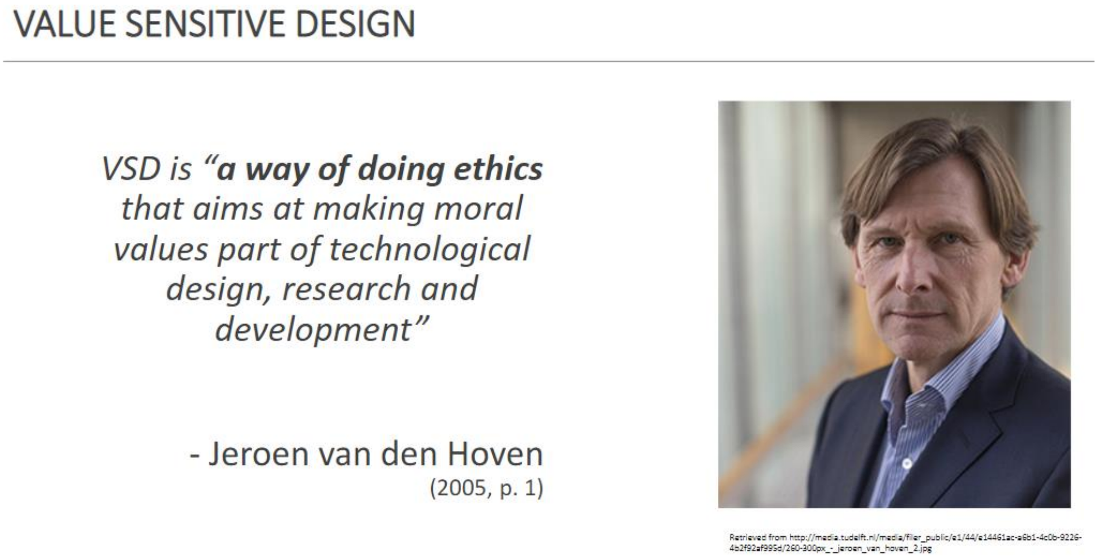
VALUE SENSITIVE DESIGN

*“**Value-Sensitive Design connects the people who design** systems and interfaces **with the people who think about and understand the values** of the stakeholders who are affected by the systems. Ultimately, [VSD] requires that we **broaden the goals** and criteria for judging the quality of technological systems **to include those that advance human values**.”*

- Batya Friedman

(HTI, UoW website https://www.cs.washington.edu/node/3865)

Retrieved from http://www.washington.edu/news/files/2013/01/Friedman_B_5x7.jpg



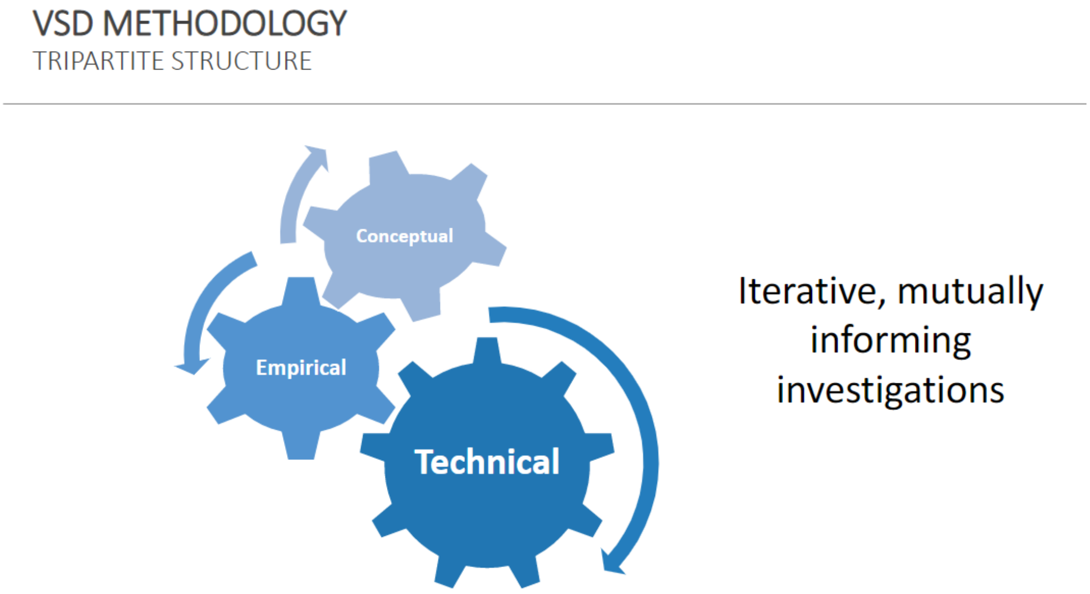
VALUE SENSITIVE DESIGN

*VSD is “**a way of doing ethics** that aims at making moral values part of technological design, research and development”*

- Jeroen van den Hoven
(2005, p. 1)

Retrieved from http://media.tudelft.nl/media/filer_public/e1/44/e14461ac-a6b1-4c0b-9226-4b2f92af995d/260-300px_-_jeroen_van_hoven_2.jpg

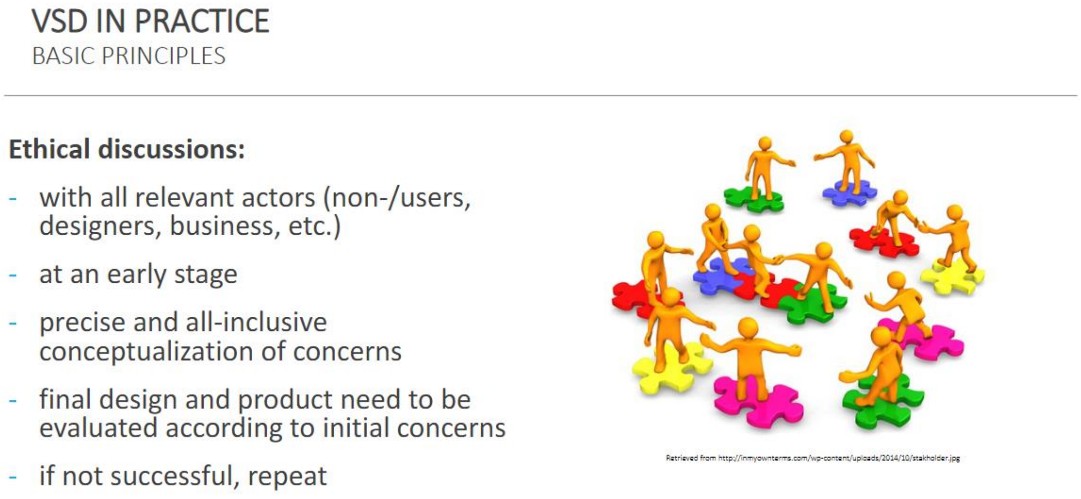
The process of VSD



VSD METHODOLOGY

TRIPARTITE STRUCTURE

Iterative, mutually informing investigations



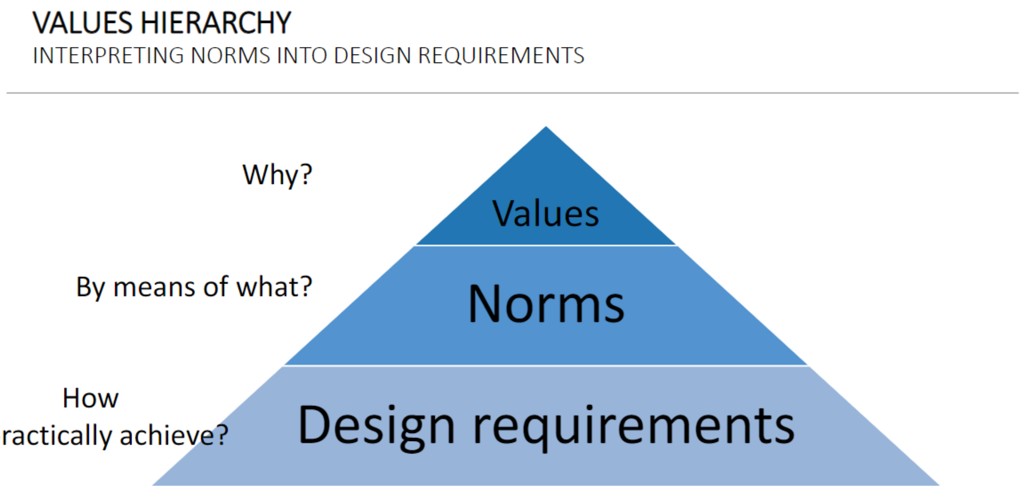
VSD IN PRACTICE

BASIC PRINCIPLES

Ethical discussions:

- with all relevant actors (non-/users, designers, business, etc.)
- at an early stage
- precise and all-inclusive conceptualization of concerns
- final design and product need to be evaluated according to initial concerns
- if not successful, repeat

Retrieved from http://inmyownterms.com/wp-content/uploads/2014/10/stakeholder.jpg



VALUES HIERARCHY

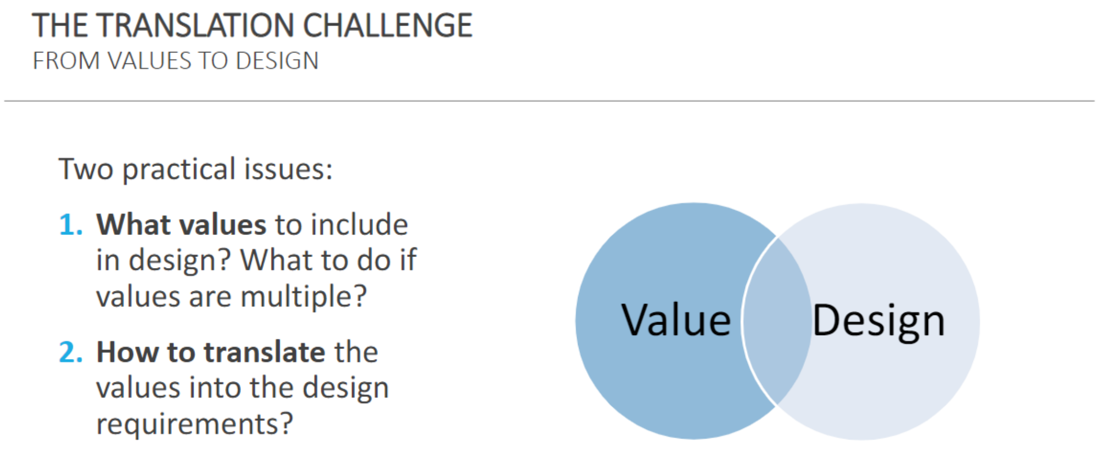
INTERPRETING NORMS INTO DESIGN REQUIREMENTS

Why? — Values

By means of what? — Norms

How to practically achieve? — Design requirements

The challenge of VSD (the problem we need to notice)

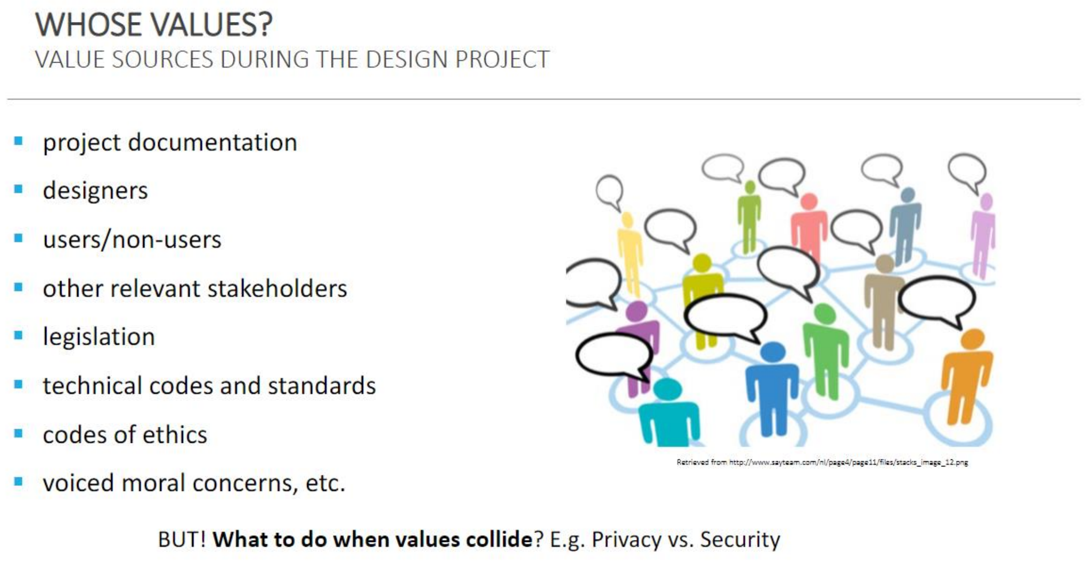


THE TRANSLATION CHALLENGE

FROM VALUES TO DESIGN

Two practical issues:

1. **What values** to include in design? What to do if values are multiple?
2. **How to translate** the values into the design requirements?



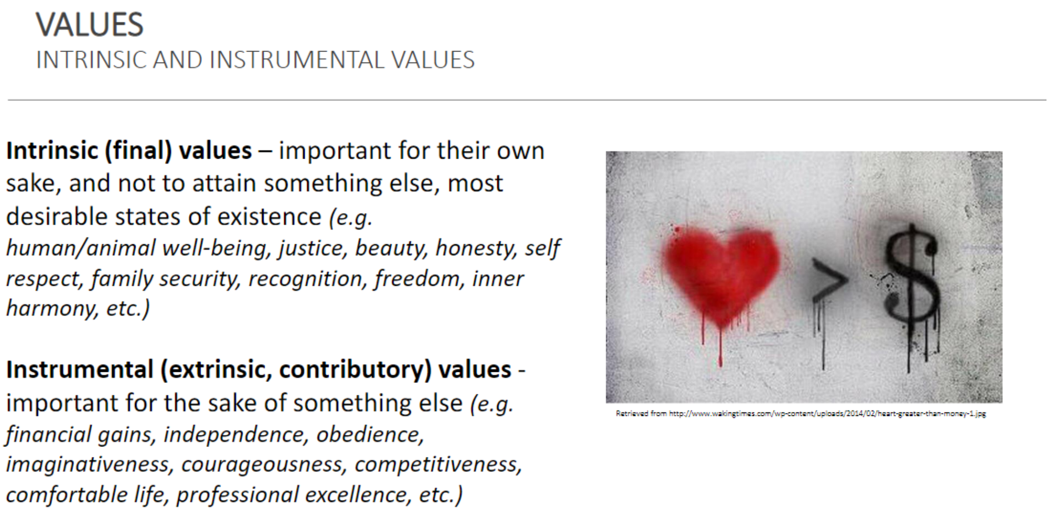
WHOSE VALUES?

VALUE SOURCES DURING THE DESIGN PROJECT

- project documentation
- designers
- users/non-users
- other relevant stakeholders
- legislation
- technical codes and standards
- codes of ethics
- voiced moral concerns, etc.

Retrieved from http://www.sayteam.com/nl/page4/page11/files/stacks_image_12.png

BUT! **What to do when values collide**? E.g. Privacy vs. Security



VALUES

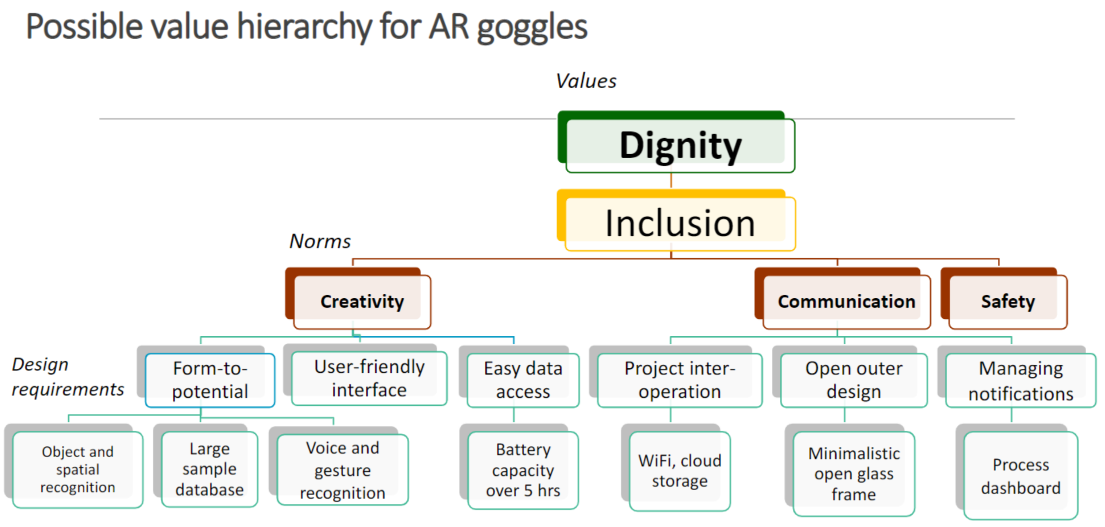
INTRINSIC AND INSTRUMENTAL VALUES

Intrinsic (final) values – important for their own sake, and not to attain something else, most desirable states of existence *(e.g. human/animal well-being, justice, beauty, honesty, self respect, family security, recognition, freedom, inner harmony, etc.)*

Instrumental (extrinsic, contributory) values - important for the sake of something else *(e.g. financial gains, independence, obedience, imaginativeness, courageousness, competitiveness, comfortable life, professional excellence, etc.)*

Retrieved from http://www.wakingtimes.com/wp-content/uploads/2014/02/heart-greater-than-money-1.jpg

Care, safety, autonomy, efficiency, satisfaction, responsibility, attentiveness, competence, responsiveness. Truth. Trustiness. Serviceability, survivability, surveillance, usability, awareness inclusivity, loneliness, disconnect, empathy, justice, social order, proximity, transparency.



Possible value hierarchy for AR goggles

Values

Dignity

Inclusion

Norms

Creativity

Communication

Safety

Design requirements

Form-to-potential

User-friendly interface

Easy data access

Project inter-operation

Open outer design

Managing notifications

Object and spatial recognition

Large sample database

Voice and gesture recognition

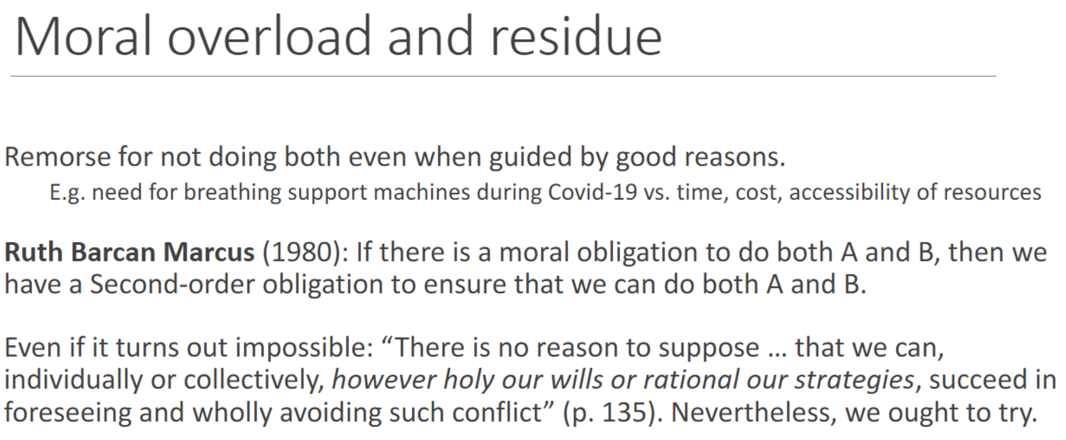
Battery capacity over 5 hrs

WiFi, cloud storage

Minimalistic open glass frame

Process dashboard

Moral overload



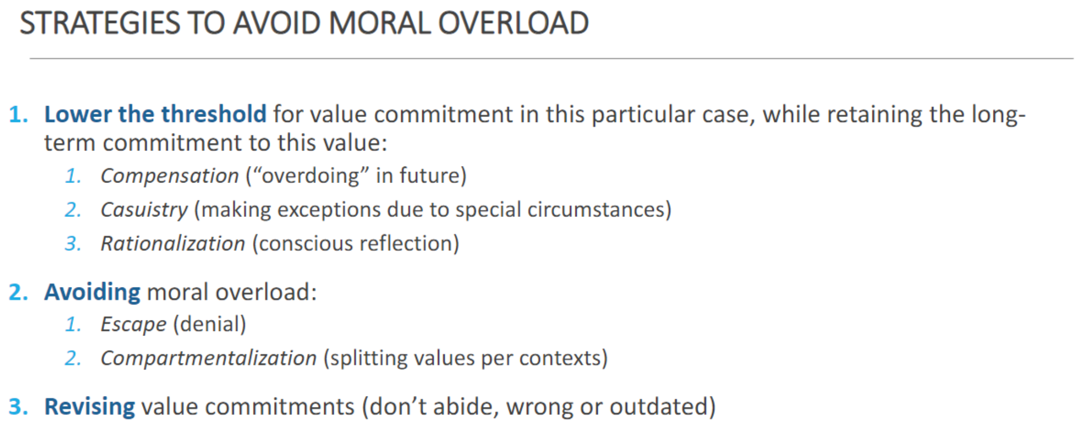
Moral overload and residue

Remorse for not doing both even when guided by good reasons.

E.g. need for breathing support machines during Covid-19 vs. time, cost, accessibility of resources

Ruth Barcan Marcus (1980): If there is a moral obligation to do both A and B, then we have a Second-order obligation to ensure that we can do both A and B.

Even if it turns out impossible: “There is no reason to suppose … that we can, individually or collectively, *however holy our wills or rational our strategies*, succeed in foreseeing and wholly avoiding such conflict” (p. 135). Nevertheless, we ought to try.



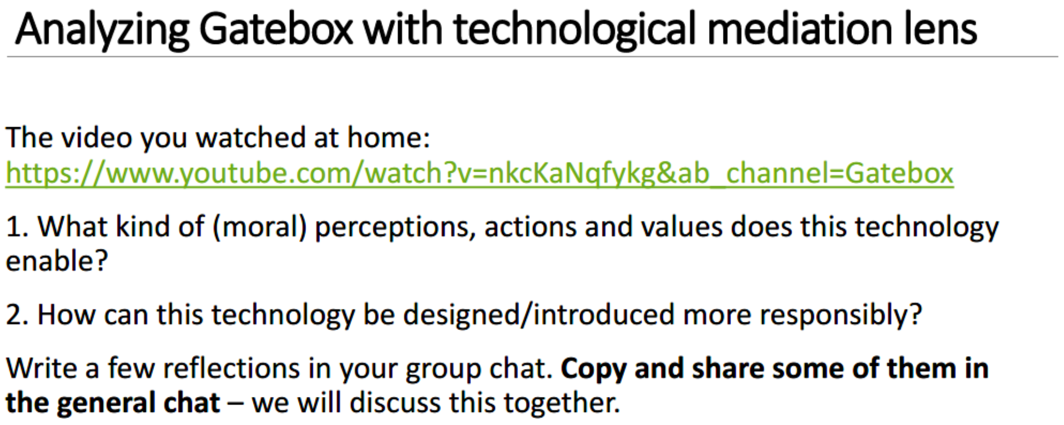
STRATEGIES TO AVOID MORAL OVERLOAD

1. **Lower the threshold** for value commitment in this particular case, while retaining the long-term commitment to this value:
   1. *Compensation* (“overdoing” in future)
   2. *Casuistry* (making exceptions due to special circumstances)
   3. *Rationalization* (conscious reflection)
2. **Avoiding** moral overload:
   1. *Escape* (denial)
   2. *Compartmentalization* (splitting values per contexts)
3. **Revising** value commitments (don’t abide, wrong or outdated)

Reading

3.3.1 embodied values, sociotechnical systems, Institutions.
3.3.2 VSD in care robot. Framework for ethical evaluation of care robots. P14. Compare two different robots
3.3.3 what is VSD. Conceptual investigation. Empirical investigation, technical investigation. Some examples of VSD

Lecture 4



Analyzing Gatebox with technological mediation lens

The video you watched at home:
https://www.youtube.com/watch?v=nkcKaNqfykg&ab_channel=Gatebox

1. What kind of (moral) perceptions, actions and values does this technology enable?

2. How can this technology be designed/introduced more responsibly?

Write a few reflections in your group chat. **Copy and share some of them in the general chat** – we will discuss this together.

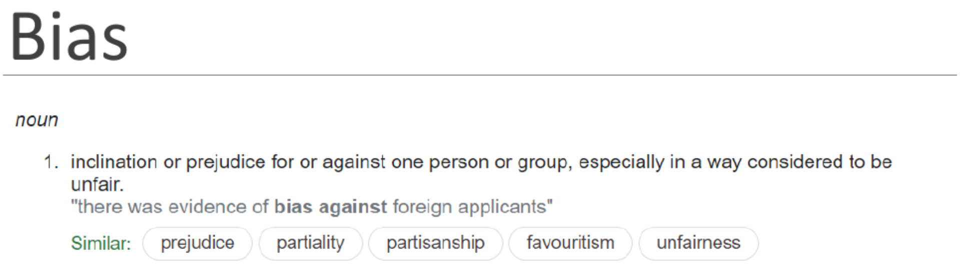
Reading

3.4.1 ethics of our relationship with seemingly sentient.

3.4.2 the human right of robots. Should robots be slaves?

3.4.3 meaningful human control

Lecture 5



Bias

noun

1. inclination or prejudice for or against one person or group, especially in a way considered to be unfair.
   "there was evidence of **bias against** foreign applicants"

   Similar: prejudice partiality partisanship favouritism unfairness

Benefits

shortcuts to constructing meaning and making sense of information that comes to our attention, determine what's important to remember, acting quickly, preventing information overload, cognitive "muscle memory"

Drawbacks

Prevents from having an overview, can conjure illusion, quick decisions can be seriously flawed, our memory reinforces errors and perpetuate injustice (self-perpetuating ideas that go without check)

Understanding bias

Technical definition: *bias* – deficiencies and imbalances in human perception

Algorithms are—in contrast with human beings—harbours of objectivity.

Garbage in, garbage out idea – AI merely reproduces human imperfections.

Goal: eradicate bias from AI and robotic application.

Understanding bias

Philosophical definition: *bias* – ontological condition (a condition of what it means to be human), seeing something AS something

Bias = pre-judgment, prior decisions, choices, awareness

We cannot avoid bias!

But… does not mean we shouldn't be aware of it, it's our responsibility to discern bias and mitigate it when it leads to harmful effects

Bias in AI and Robotics: Long before data collection

Framing the problem. What needs to be achieved (e.g. a credit card company wants to predict a customer's creditworthiness, but "creditworthiness" is a fuzzy idea, how should it be defined? Within the context)

Collecting the data. Collected data can be (1) unrepresentative of reality or (2) reflect existing prejudices. E.g. feeding more photos of light-skinned faces than dark-skinned faces. The resulting face recognition system would inevitably be worse at recognizing darker-skinned faces. Or when Amazon discovered that its internal recruiting tool was dismissing female candidates.

Preparing the data. Selecting which attributes you want the algorithm to consider. E.g. in modeling creditworthiness, an "attribute" could be the customer's age, income, or number of paid-off loans. In the case of Amazon's recruiting tool, an "attribute" could be the candidate's gender, education level, or years of experience.

1. Co-shaping moral perceptions:

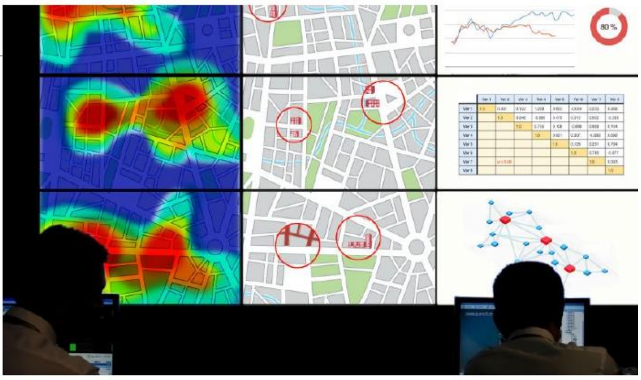
Safety in the city

© Stratumseind TU

: Smart Cities Program, Living labs: https://www.tue.nl/en/our-university/departments/built-
:onment/research/smart-cities-program/collaboration/living-labs/
, M. (2019). Surveillance and privacy in smart cities and living labs: Conceptualising privacy for public
e.: https://research.tilburguniversity.edu/en/publications/surveillance-and-privacy-in-smart-cities-and-
g-labs-conceptu

2. Co-shaping moral actions:

Fair treatment in the city



© Council of Europ

Amnesty International. (2020). *We sense trouble. Automated discrimination and mass surveillance in predictive policing in the Netherlands*. Available at: https://www.amnesty.org/download/Documents/EUR3529712020ENGLISH.PDF

Change the pattern of buying a house or sending kids to which school.

Sociotechnical systems lens

https://medium.com/pcmag-access/healthcare-algorithms-are-biased-and-the-results-can-be-deadly-da11801fed5e

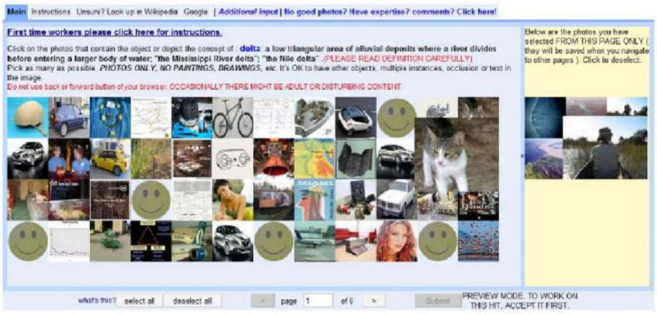
1. How are the practice and experience of healthcare changed by AI? In giving the answer, analyze the case from a perspective of the sociotechnical systems.

Recap: this means navigating across 3 levels of analysis: (1) social, (2) technological and (3) world/institutional (e.g. culture, norms, traditions, political regimes, governance structures).

2. How can this technology be designed/introduced more responsibly?

Write a few reflections in your group chat. **Copy and share some of them in the general chat** – we will discuss this together.

IMAGENET Basic User Interface


Main
Instructions
Unsure? Look up in Wikipedia
Google
[Additional input] No good photos? Have expertise? comments? Click here!
what's this?
select all
deselect all
<
page
1
of 6
>
Submit

First time workers please click here for instructions.

Click on the photos that contain the object or depict the concept of : **delta**: **a low triangular area of alluvial deposits where a river divides before entering a larger body of water; "the Mississippi River delta"; "the Nile delta"** .(PLEASE READ DEFINITION CAREFULLY)

Pick as many as possible. *PHOTOS ONLY, NO PAINTINGS, DRAWINGS*, etc. It's OK to have other objects, multiple instances, occlusion or text in the image.

Do not use back or forward button of your browser. OCCASIONALLY THERE MIGHT BE ADULT OR DISTURBING CONTENT.

Below are the photos you have selected FROM THIS PAGE ONLY (they will be saved when you navigate to other pages). Click to deselect.

PREVIEW MODE. TO WORK ON THIS HIT, ACCEPT IT FIRST.

No easy technological fix to complex social problems

The problems of fairness, justice, equality cannot be fixed with algorithms

But we can proactively help to identify bias and mitigate it

https://excavating.ai/

Summary

We cannot eradicate bias fully, it's a feature of human nature ➔ proactive responsibility, e.g. in team composition, design and use

Transparency, like bias, is a family value involving trust, informed choice, independence, dignity, fairness and other values

Transparency requires a (sociotechnical) system analysis to understand the needs and appropriate strategies: labelling, watermarks and other public disclosure

SO

Ethics ≠ assessment (external) BUT accompaniment (from within process, relation, practice)

≠ not a one-time check (static) BUT dynamic (continuous, iterative, learning)

Reading

3.5.1 bias on women or other group. google sexual harassment but no punishment. amazon automated hiring tool. Back neighborhood criminal. Automatic gender recognition. GANs to create fake pornographic.

3.5.2 transparency, restaurant inspection can improve transparency and food safety. How to improve transparency in data, model, human involvement. But improving transparency may introduce violation on privacy and manipulation.

Lecture 6

3.6.1 the threat of algocracy(hiddenness, opacity).

Should we eradicate the threat?

How can we accommodate

3.6.2 the decision-making is context based.

Different decisions are made according to different ethical theory.

Autonomy

Solution to global engagement with ethics of AI. Maximize engagement of people from all over the world. Global participation. Diverse views.