Design for values

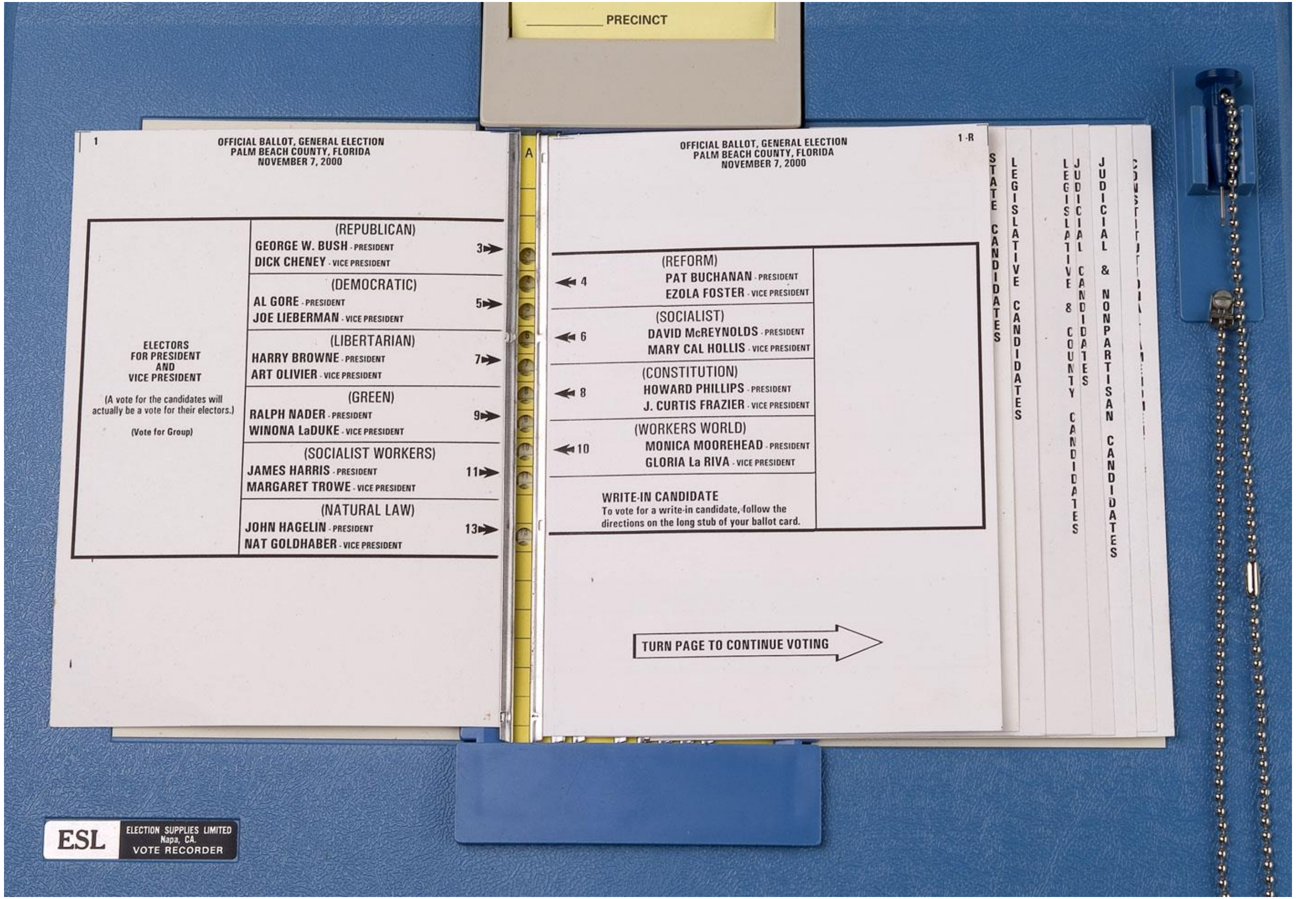
Ibo van de Poel

TU Delft

Overview

- Why design for values?
- How to identify values?
- How to translate values into design requirements?
- Dealing with conflicting values

Why design for values?



1

OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000

ELECTORS FOR PRESIDENT AND VICE PRESIDENT

(A vote for the candidates will actually be a vote for their electors.)

(Vote for Group)

(REPUBLICAN)
GEORGE W. BUSH - PRESIDENT
DICK CHENEY - VICE PRESIDENT 3➡

(DEMOCRATIC)
AL GORE - PRESIDENT
JOE LIEBERMAN - VICE PRESIDENT 5➡

(LIBERTARIAN)
HARRY BROWNE - PRESIDENT
ART OLIVIER - VICE PRESIDENT 7➡

(GREEN)
RALPH NADER - PRESIDENT
WINONA LaDUKE - VICE PRESIDENT 9➡

(SOCIALIST WORKERS)
JAMES HARRIS - PRESIDENT
MARGARET TROWE - VICE PRESIDENT 11➡

(NATURAL LAW)
JOHN HAGELIN - PRESIDENT
NAT GOLDHABER - VICE PRESIDENT 13➡

1-R

OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000

⬅ 4 (REFORM)
PAT BUCHANAN - PRESIDENT
EZOLA FOSTER - VICE PRESIDENT

⬅ 6 (SOCIALIST)
DAVID McREYNOLDS - PRESIDENT
MARY CAL HOLLIS - VICE PRESIDENT

⬅ 8 (CONSTITUTION)
HOWARD PHILLIPS - PRESIDENT
J. CURTIS FRAZIER - VICE PRESIDENT

⬅ 10 (WORKERS WORLD)
MONICA MOOREHEAD - PRESIDENT
GLORIA La RIVA - VICE PRESIDENT

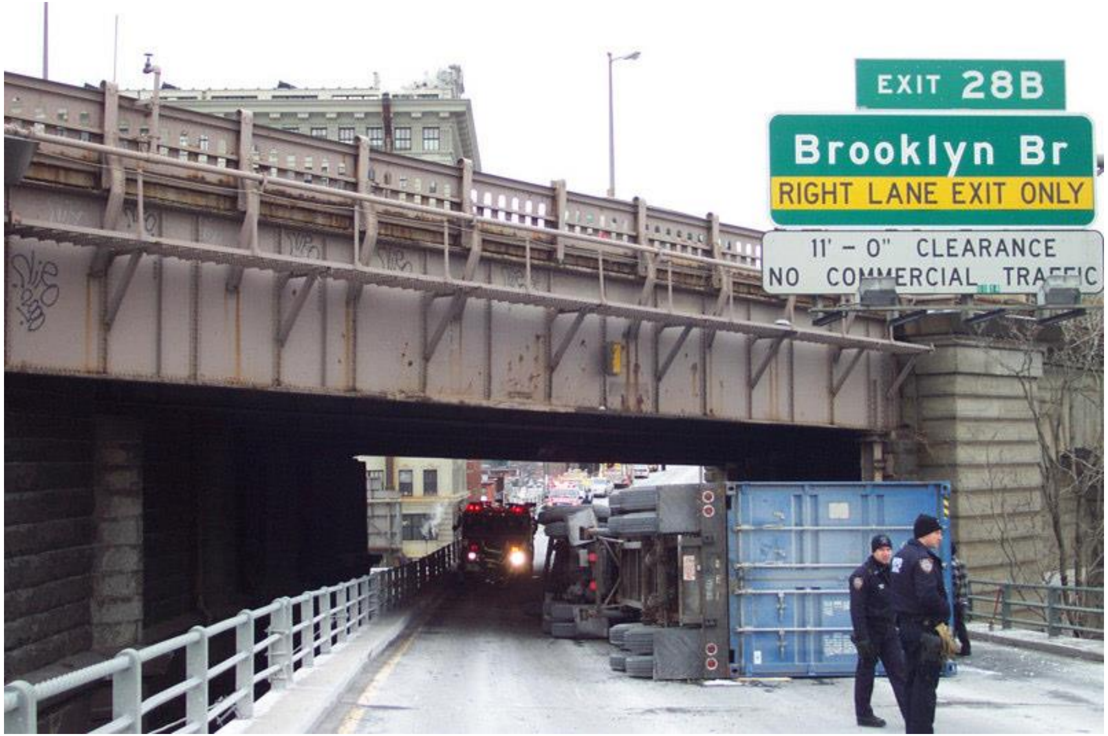
WRITE-IN CANDIDATE
To vote for a write-in candidate, follow the directions on the long stub of your ballot card.

TURN PAGE TO CONTINUE VOTING ➡

Voting computers the Netherlands

- Judge forbids current models in 2007
- Voting secrecy not guaranteed
- No possibility for independent control on counting

Low overpasses Long Island



Low overpasses Long Island

- Designed by urban planner Robert Moses (1888-1981)
- Deliberate low as to avoid buses to go to the beaches
- Black people usually travelled by bus
- “Racist overpasses”

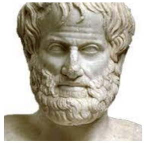
Design for Values

Systematic attempt to include values of moral importance in design

Three types of investigations (Friedman et al.):

Empirical

- Stakeholders and their values



Conceptual

- Conceptualizations of relevant values
- Trade-offs

Technical/engineering

- Embodiments of value
- Value issues raised by technology

Identifying values

Identifying relevant values

How to identify what values a system should be designed for?

Requires both:

- Open-ended (empirical) inquiry what values might be *relevant*
- Normative choice/criterion: what values *should* we design for?

Five possible (but individually unsatisfactory) answers

- Designers’ values
- Stakeholders’ values
- Based on a moral theory
- European HL Expert Group on AI (or a similar authoritative source)
- Elicitation through an AI system

Designers’ values

Important to be aware and reflexive about, as they will affect design choices anyway.

But too limited a basis in designing for values:

- Bias
- Blind spots

As designs will affect more parties than just the designers, focusing only on designer’s value is morally unjust

Stakeholders’ values

- Focus in approaches like VSD
- Very important, but not enough

- Not all stakeholders’ values are normatively important, or even normatively significant
- Need to distinguish normative values from needs, preferences, desires etc.

- Stakeholders may disagree about normative values
- Some normative values may not be discerned by stakeholders

Moral theory

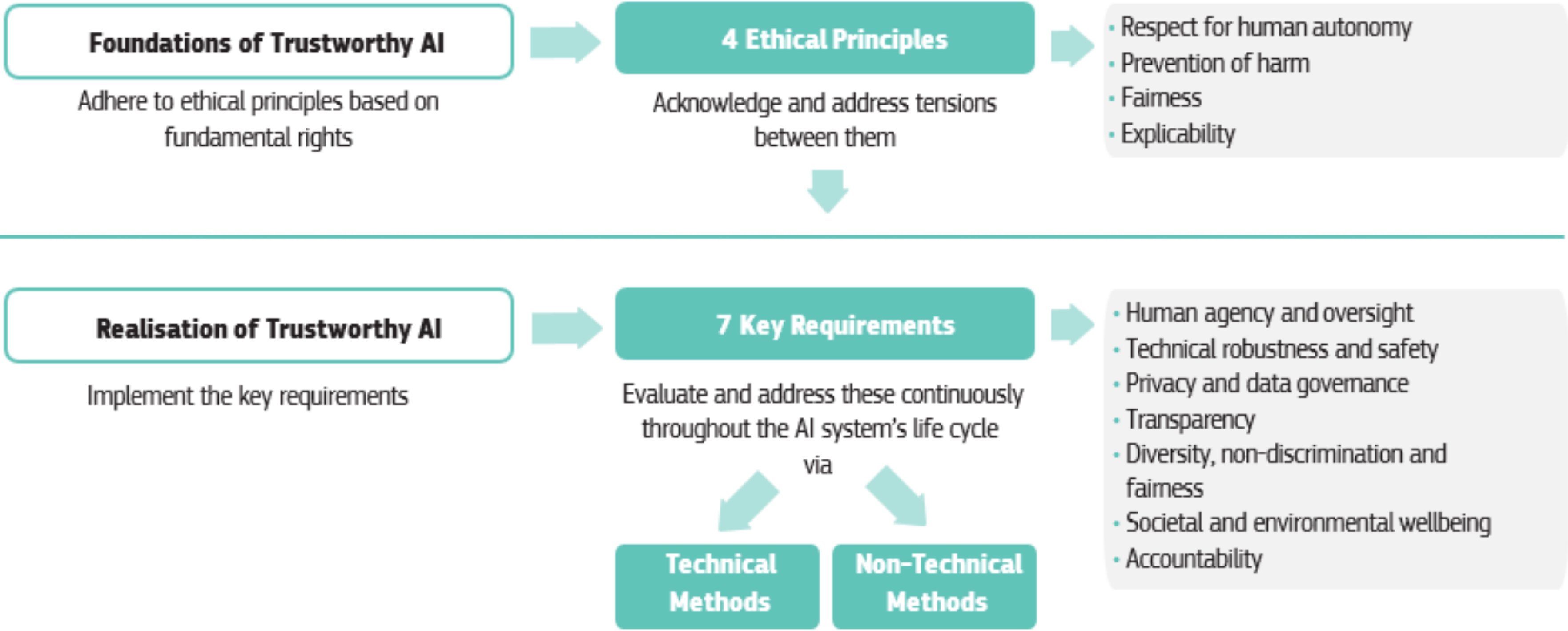
Advocated by some in VSD

But: there are conflicting substantive moral theories and there is no agreement on what is the right one

Does therefore not solve issues of moral disagreement and normative diversity

Still moral theories may be relevant for better understanding certain relevant values

EU High-level Expert group on AI



Foundations of Trustworthy AI

Adhere to ethical principles based on fundamental rights

4 Ethical Principles

Acknowledge and address tensions between them

- Respect for human autonomy
- Prevention of harm
- Fairness
- Explicability

Realisation of Trustworthy AI

Implement the key requirements

7 Key Requirements

Evaluate and address these continuously throughout the AI system's life cycle via

Technical Methods

Non-Technical Methods

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental wellbeing
- Accountability

Some authorative source

Naturalistic fallacy

Still: might reflect legitimate societal consensus

But they may be too general and abstract

- For concrete applications and context: other moral values will be important as well
- If these are ignored, we might end up with normatively undesirable designs
- Also: need to understand these values in context
  - E.g., justice and no harm require contextual understanding
- Need to combine it with bottom-up elicitation of values

Value elicitation through an AI system

- Artificial agents may elicit values from humans
- But how do they distinguish what values are normatively relevant?
  - Substantive moral theory: but which one?
  - At least some meta-ethical assumptions needed
- Artificial agents may learn new values but they may also unlearn values
  - Introduces risks for value alignment (later more)

How to proceed?

Pragmatically, combining the 5 partial answers might bring us some way

More fundamentally:

- Don’t expect substantive normative agreement about values!
- But we might come to agree what count as proper arguments and what not

What philosophers (like me) might contribute:

- Suggest a general (meta-ethical) account of what values of moral importance are

Values: a proposed definition

Values are:

Properties of entities that correspond to reasons for a positive response or pro-attitude (towards that entity)

Litmus test

V is a (positive) value for an entity E in context C

Iff E having the property V in context C corresponds to a pro-tanto reason for a pro-attitude towards E in context C

For example:

If the car were safe would there be a pro-tanto reason for positively responding to it? yes

If the care were dangerous would there be a pro-tanto reason for positively responding to it? no

If the care were red would there be a pro-tanto reason for positively responding to it? Probably not

Translating values into design requirements

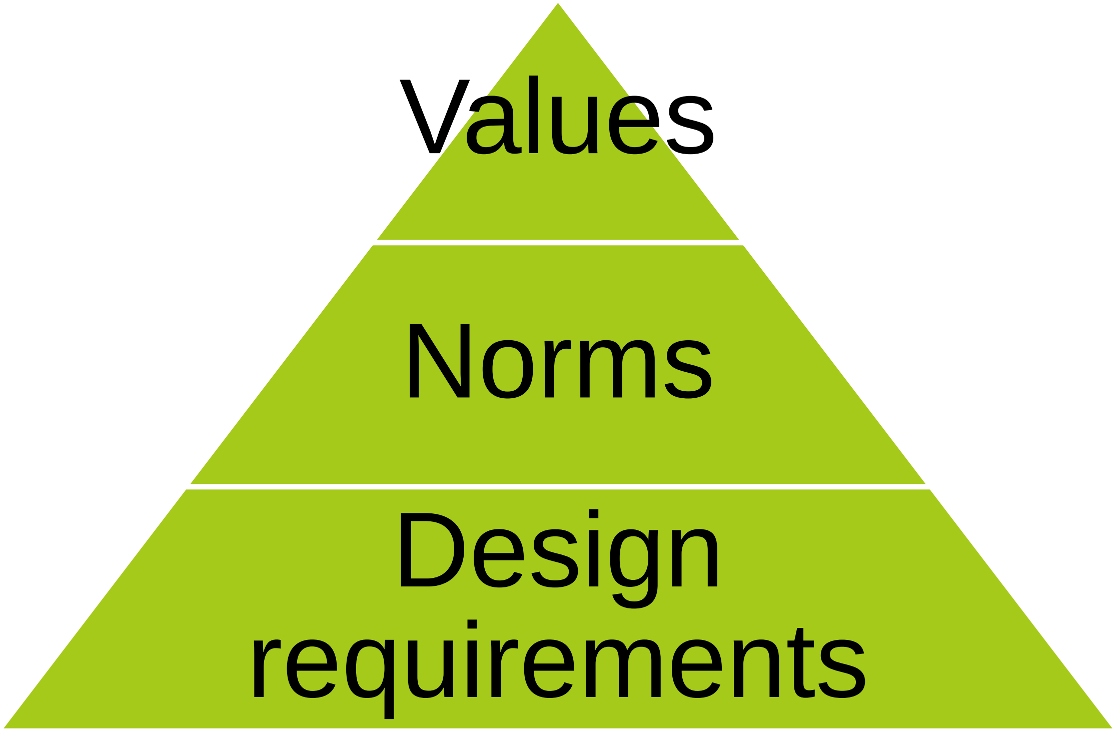
Conceptualization

the providing of a definition, analysis or description of a value that clarifies its meaning and its applicability *in general*

Specification

makes a value more specific so that it can guide action and decisions *in a specific context*

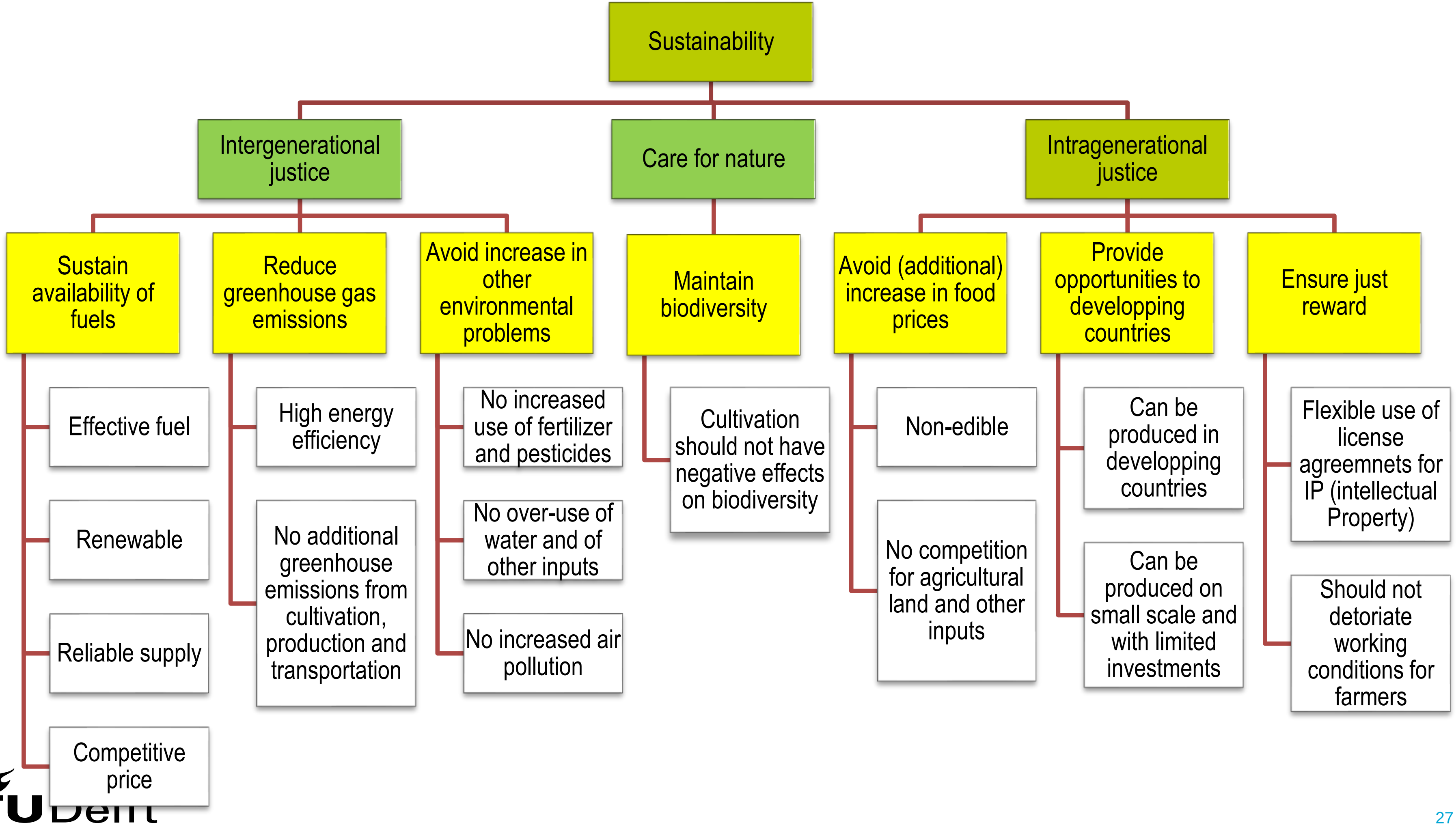
Values hierarchy



Values

Norms

Design requirements



- Sustainability
  - Intergenerational justice
    - Sustain availability of fuels
      - Effective fuel
      - Renewable
      - Reliable supply
      - Competitive price
    - Reduce greenhouse gas emissions
      - High energy efficiency
      - No additional greenhouse emissions from cultivation, production and transportation
    - Avoid increase in other environmental problems
      - No increased use of fertilizer and pesticides
      - No over-use of water and of other inputs
      - No increased air pollution
  - Care for nature
    - Maintain biodiversity
      - Cultivation should not have negative effects on biodiversity
  - Intragenerational justice
    - Avoid (additional) increase in food prices
      - Non-edible
      - No competition for agricultural land and other inputs
    - Provide opportunities to developping countries
      - Can be produced in developping countries
      - Can be produced on small scale and with limited investments
    - Ensure just reward
      - Flexible use of license agreemnets for IP (intellectual Property)
      - Should not detoriate working conditions for farmers

Adequacy

Does meeting the design requirements count as an instance of respecting/addressing the relevant value(s)?

Different from computer science approaches to e.g., fairness

- Fairness metrics
- But:
  - No clear connection to philosophical conceptions of fairness/justice
  - Ignore important moral dimensions of fairness/justice
    - E.g., procedural justice, recognition justice

Fairness in Design for Values

Instead

- Consider fairness in context: Start with a broad identification of possible fairness concerns
- Identify and apply relevant philosophical conceptions and technical metrics for fairness: you might need to develop new ones for your design project
- Move back and forth between general conceptions and specific metrics

Dealing with value conflict in design

TUDelft

Artificial Intelligence Makes Bad Medicine Even Worse

A new study out from Google seems to show the promise of AI-assisted health care. Actually, it shows the threat.

AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind

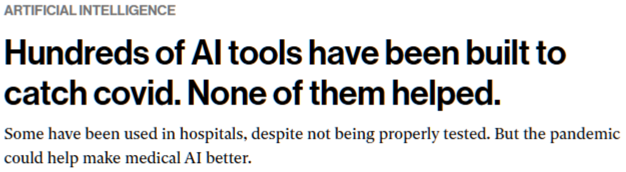
Machine learning has the potential to save thousands of people from skin cancer each year—while putting others at greater risk.

By Angela Lashbrook

Artificial intelligence Oct 25

A biased medical algorithm favored white people for health-care programs

A study has highlighted the risks inherent in using historical data to train machine-learning algorithms to make predictions.



ARTIFICIAL INTELLIGENCE

Hundreds of AI tools have been built to catch covid. None of them helped.

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

Example: NarxCare

- “analytics tool and care management platform that purports to instantly and automatically identify a patient’s risk of misusing opioids”
- Kathryn was refused opioids on basis of high risk-score
- Probably based on medicine for her pets!

(https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/)

Value conflict

- More explainable AI may help to avoid such undesirable situations

- But may well intrude on people’s privacy

- How to navigate that value conflict?

Four types of approaches

Calculative approaches

- “Only the best”

Satisficing

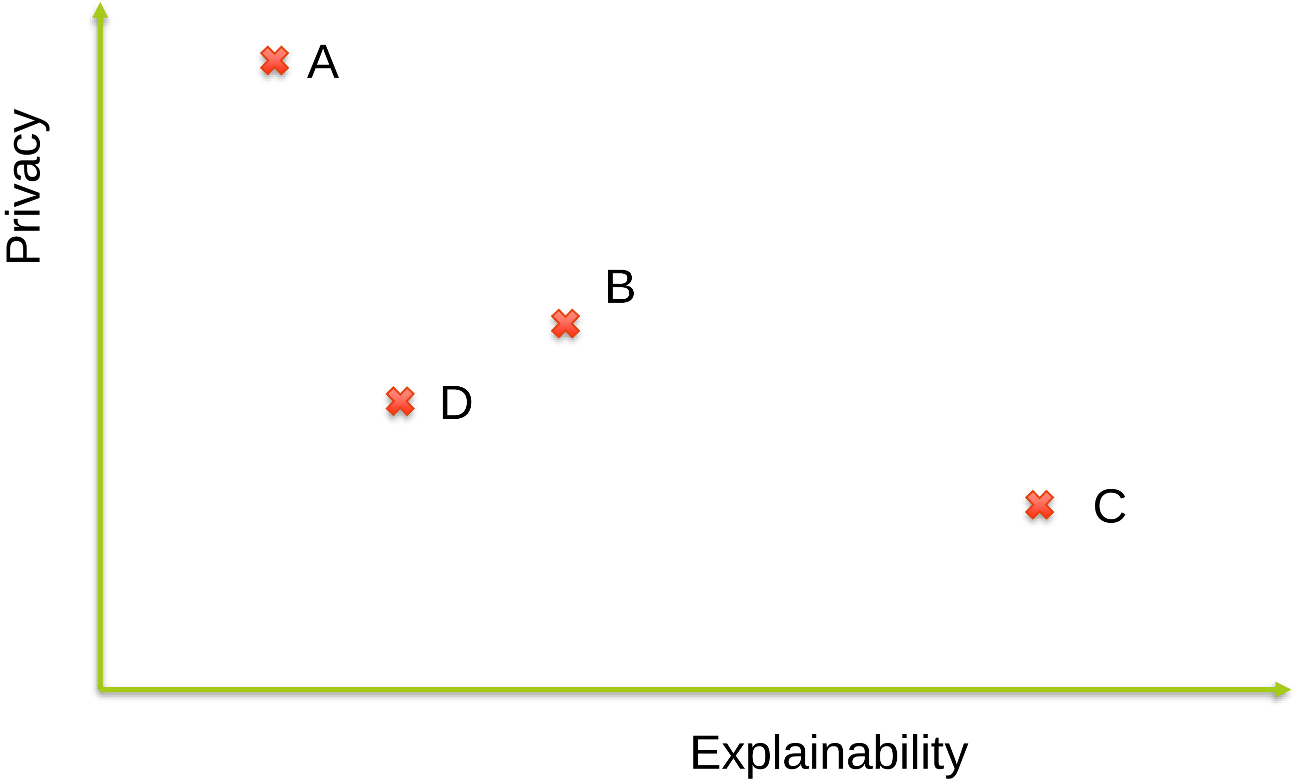
- “Good enough”

Respecification

- “Think again”

Innovation

- “Try harder”


Privacy
A
B
D
C
Explainability

Calculative approaches: only the best

Multi Criteria Analysis

Option	Privacy weight	Privacy score	Explainability weight	Explainability score	Total
A	3	5	2	1	17
B	3	3	2	3	15
C	3	2	2	5	16

Problems

- Calculative is not necessarily more objective

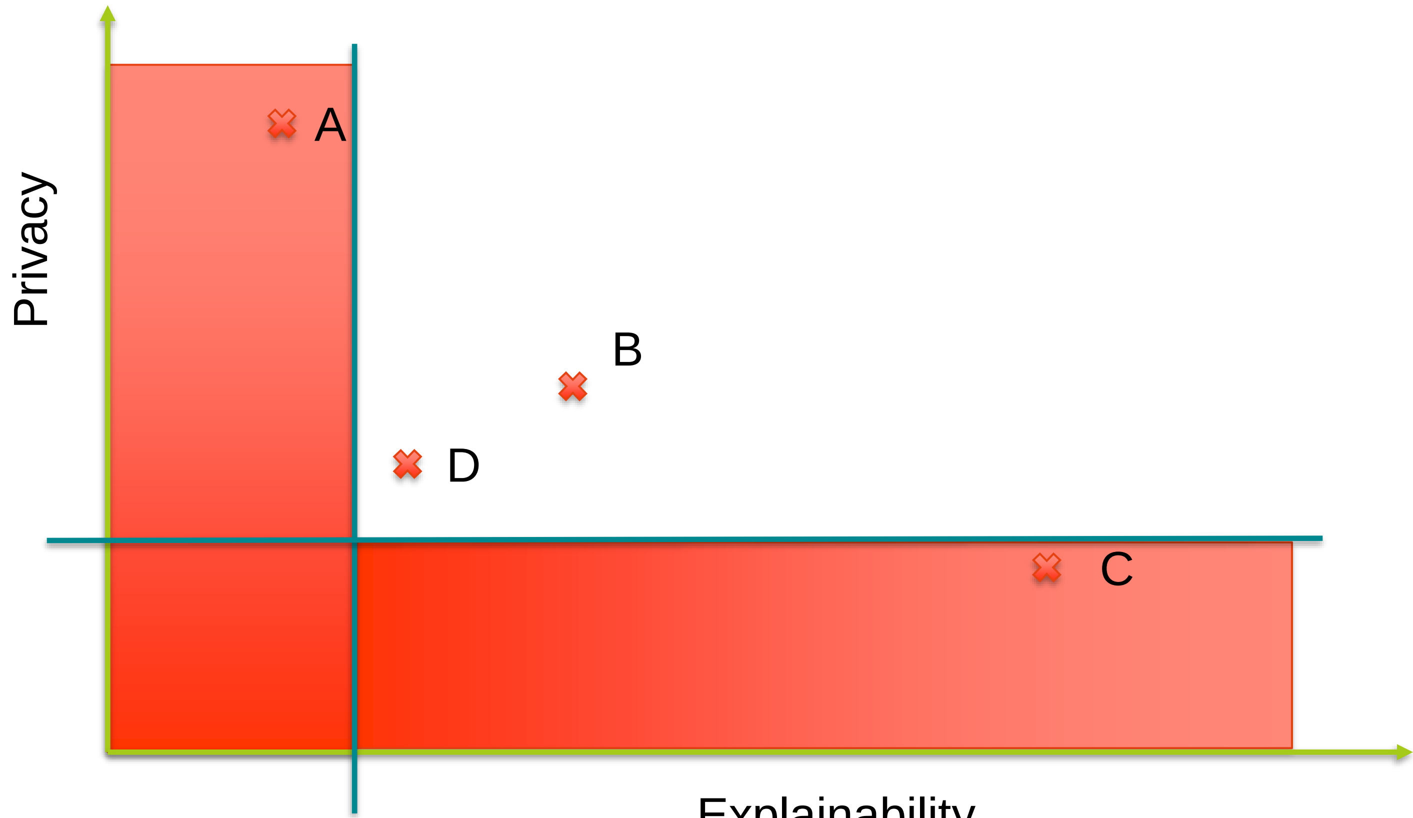
- Outcome depends on the measurement scale (and/or other arbitrary choices)

- Value incommensurability

Satisficing: good enough

- Look for alternative that is ‘good enough’

- Set threshold values for each of the relevant values


Privacy
A
B
D
C
Explainability

Problems

- How to set the thresholds?
  - Maybe legal requirements?

- You may end up with no or multiple options

- Is ‘good enough’ good enough or should we do better?

TU Delft

Respecification: think again

- Usually, values are not conflicting, but their specification is conflicting

- So, value conflict may perhaps be solved by respecifying values

Privacy

- For example, privacy can be specified as:
  - Secrecy: do not collect any personal information
  - Informed consent: personal data can be collected with the consent of data subject

- The first specification may conflict with explainability, but not the second

Explainability can mean a lot of different things:

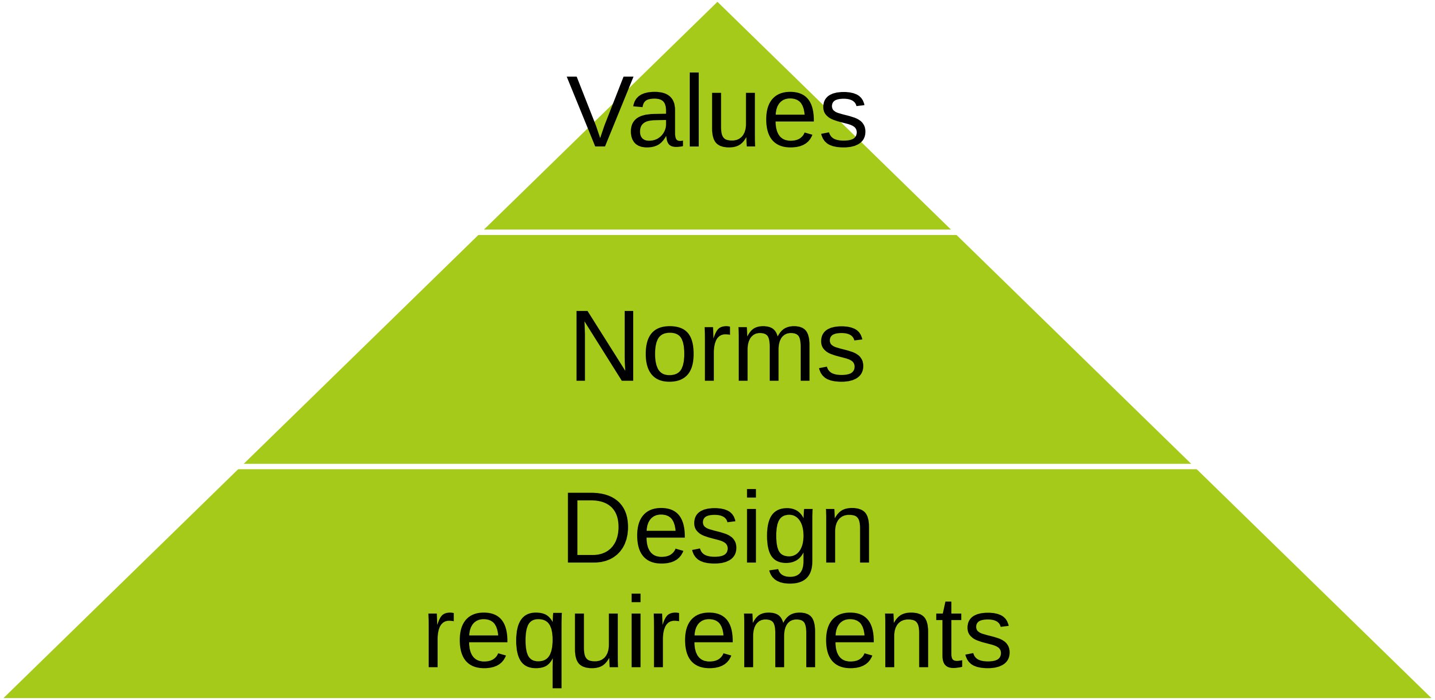
- Explaining (understanding) why the algorithm produces a certain outcome
- Explaining (understanding) how the algorithm learned something (adapted itself)
- Finding/understanding causal relations in the underlying data (rather than correlations that can be spurious)
- Providing reasons/arguments for an advice or decision
- Moreover: explainable to whom?

Overarching values

- Privacy and explainability may be based on similar values. E.g.,:
  - No harm
  - Human moral autonomy

- These overarching values may provide an argumentative framework to solve tensions and to respecify values

Values hierarchy

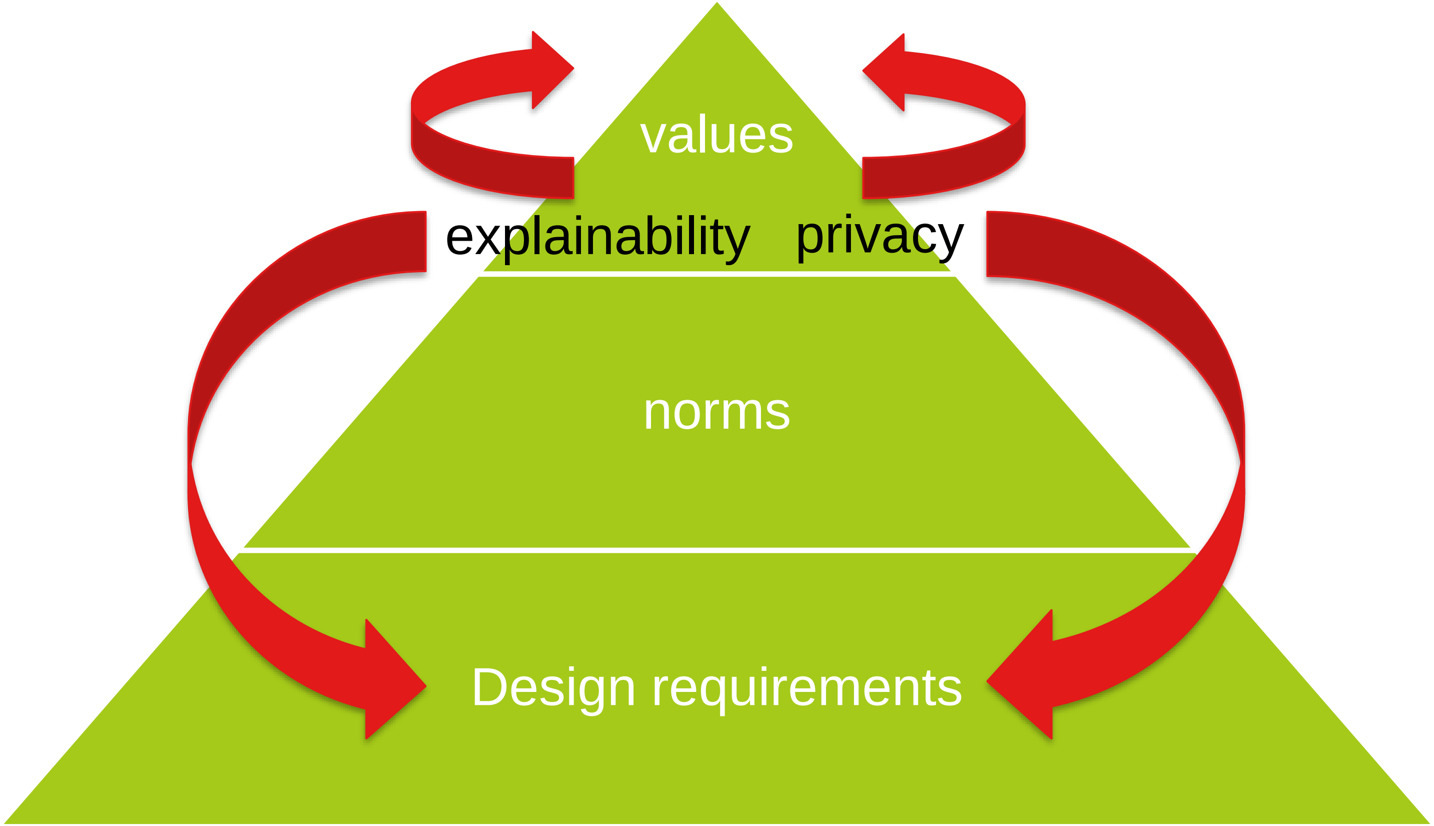


Values

Norms

Design requirements

Values hierarchy



values

explainability privacy

norms

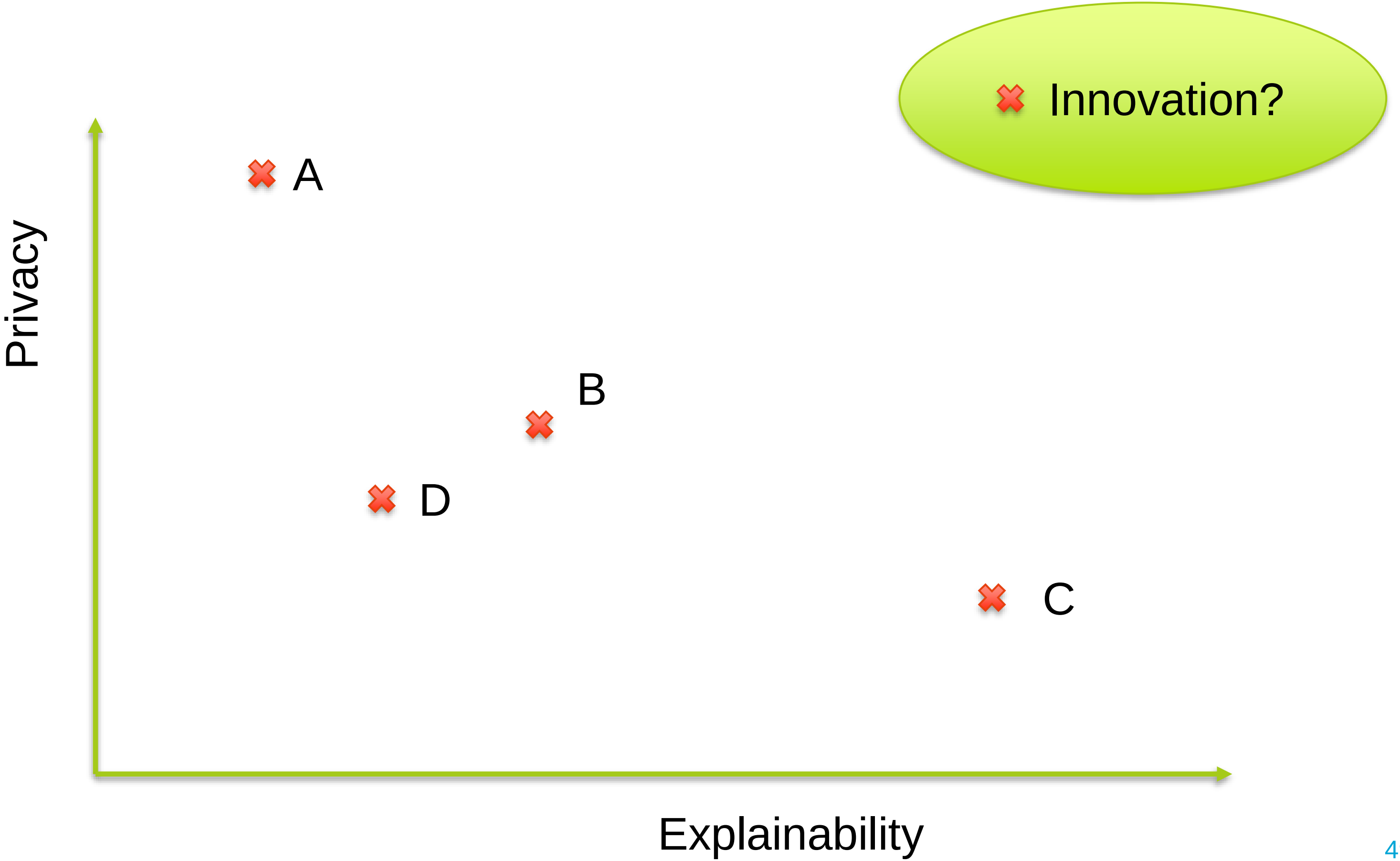
Design requirements

Problems

- May lead to unacceptable relaxation of values and norms

- Not all values conflicts can be solved in this way

Innovation: try harder



Value dams and flows

Value dams

Features that are strongly opposed by some stakeholders

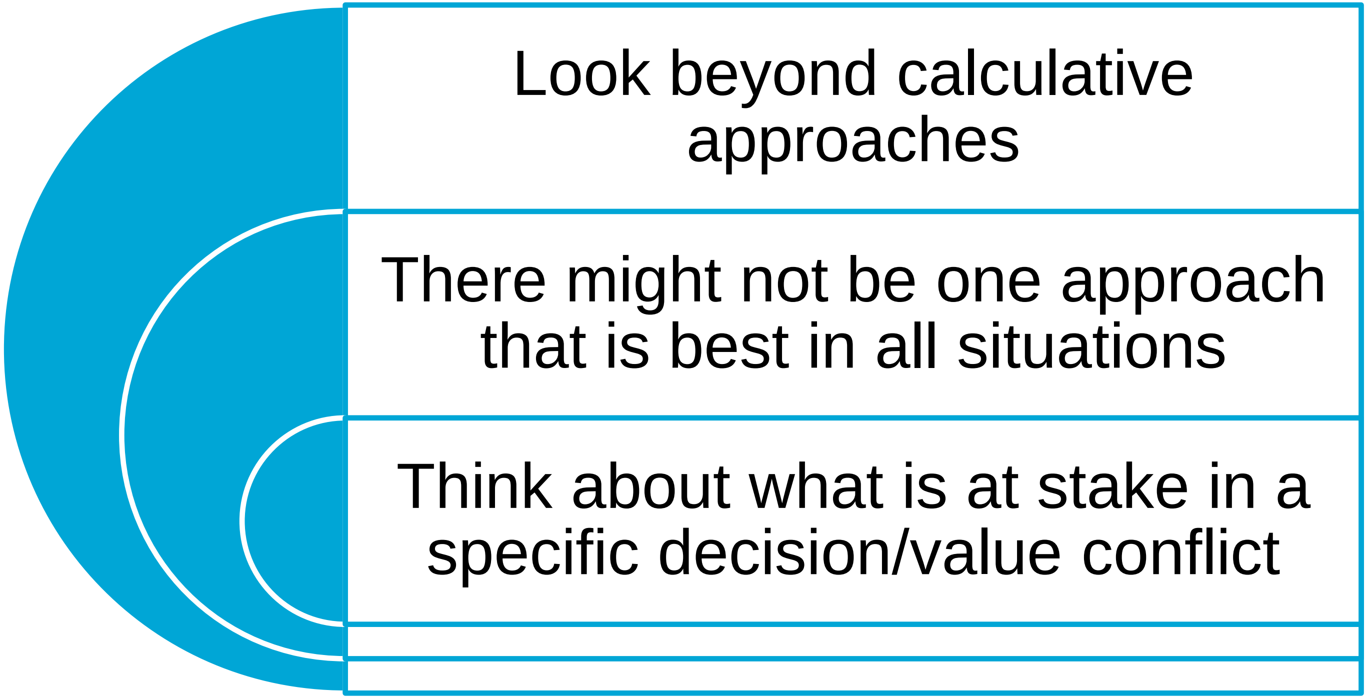
Value flows

Features that a large number of stakeholders support

Problems

- May lead to a technological fix
- There may not be a solution that solves the value conflict

Take aways



Look beyond calculative approaches

There might not be one approach that is best in all situations

Think about what is at stake in a specific decision/value conflict