Accountability, Responsibility, Transparency in AI

“right” and “wrong” in AI

Ethics and AI

Virginia Dignum
Social Artificial Intelligence Lab & Delft Institute Design for Values
Delft University of Technology

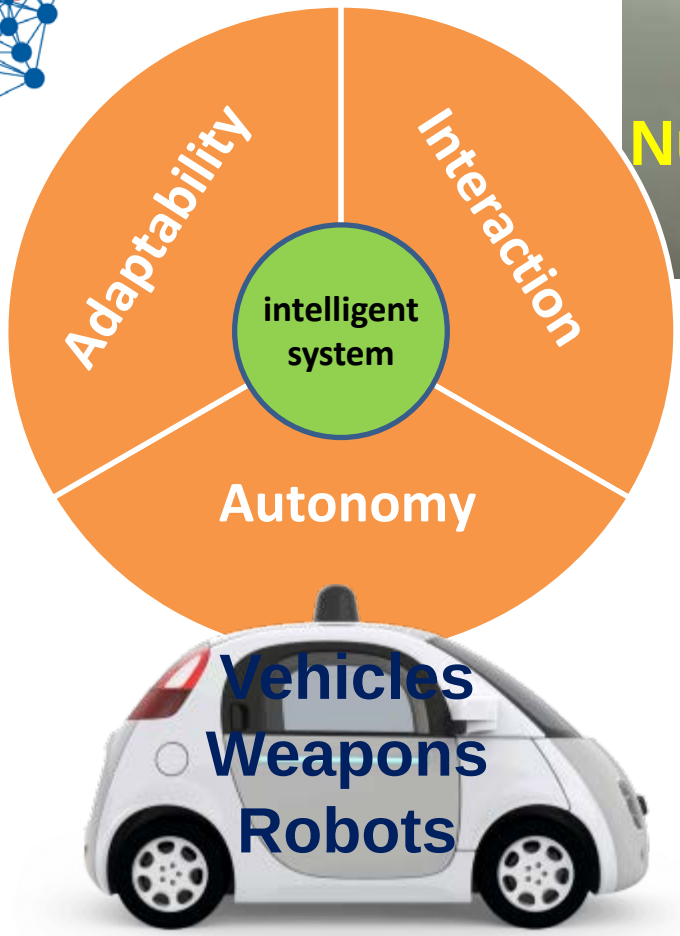
Email: m.v.dignum@tudelft.nl
Twitter: @vdignum

http://designforvalues.tudelft.nl/

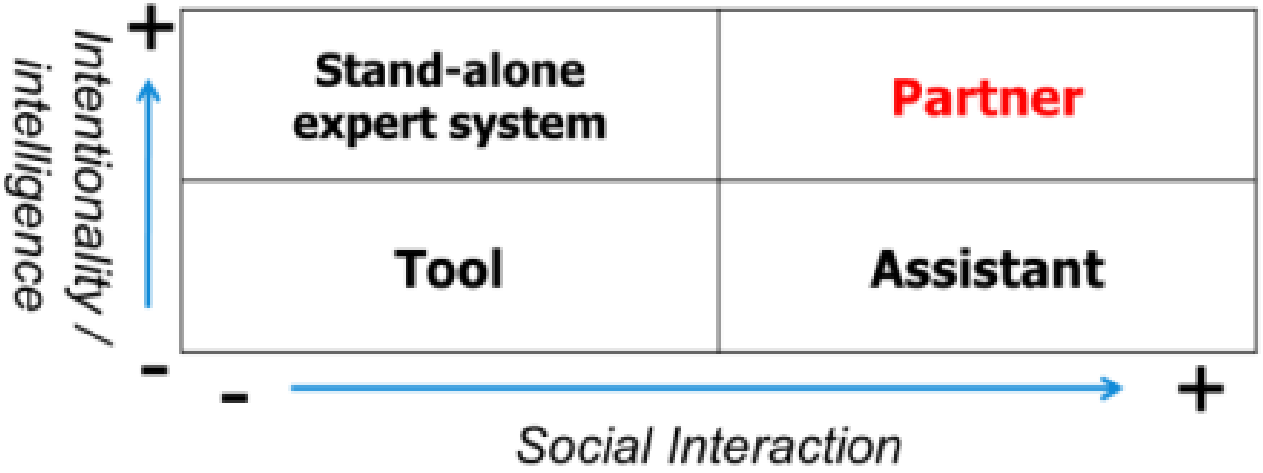
Responsible Innovation in AI

- Ethics **by** Design
  - Integration of ethical reasoning abilities as part of the behaviour of artificial autonomous systems (such as agents and robots)
- Ethics **in** Design
  - ethical implications of artificial intelligence as it integrates and replaces traditional systems and social structures
- Ethics **for** Design(ers)
  - research integrity of researchers and manufacturers as they design, construct, use and manage artificially intelligent systems,

Artificial Intelligence

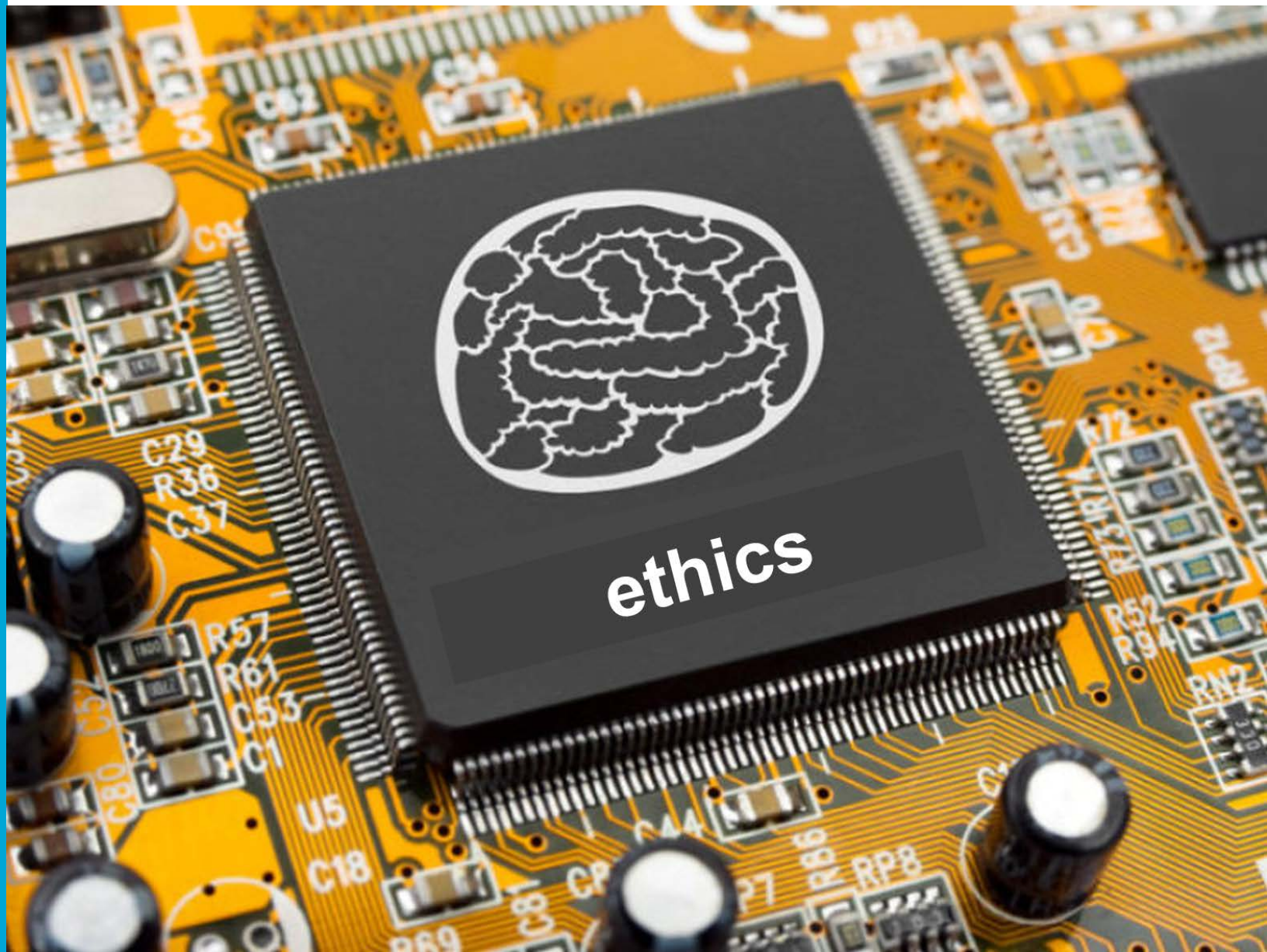


Perception: From tools to team-mates



TU Delft DESIGN FOR VALUES

Ethics **by** Design



TU Delft

Ethics by Design - issues

1. **Value alignment**
   - Identify *relevant* human values
   - Are there universal human values?
   - Who gets a say? Why these?

2. **How to behave?**
   - Ethical theories: How to behave according to these values?
   - How to prioritize those values?

3. **How to implement?**
   - Role of user
   - Role of society
   - Role of AI system

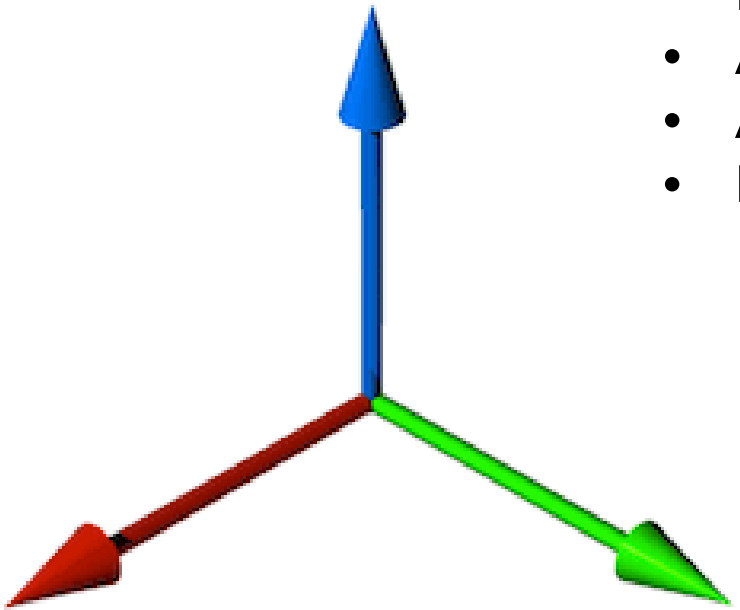
1. Value alignment

- Which values? Whose values?
- Sources
  - Stakeholders: Designer, User, Owner, Manufacturer
  - Society: codes of ethics, codes & standards, law
- Who decides who has a say?

- How to make choices and tradeoffs between conflicting values?
- How to verify whether the designed system embodies the intended values?

- **Design for values**
  - systematic attempt to include values of ethical importance in design
  - Make values, their priorities and choices explicit, transparent and systematic

Sources: social norms, law, ethics

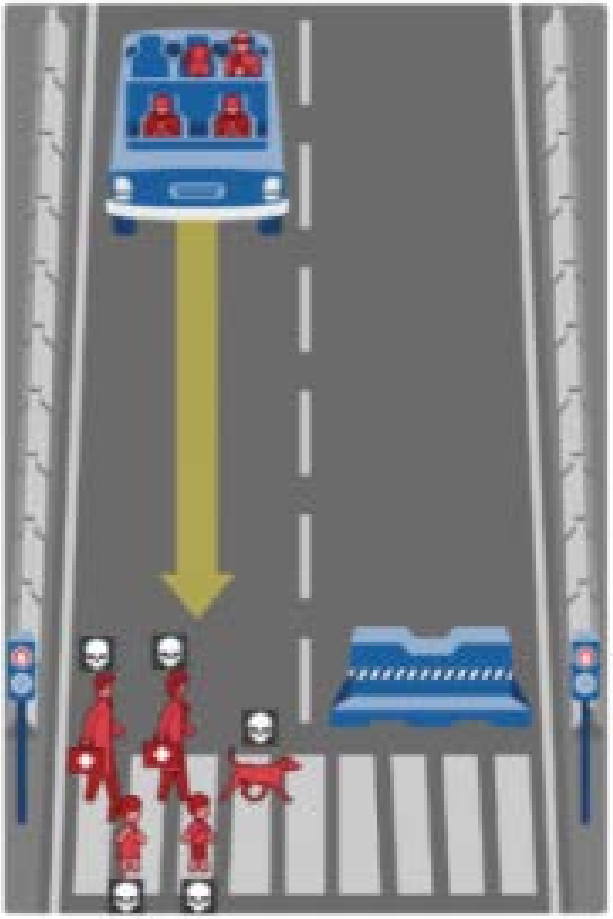


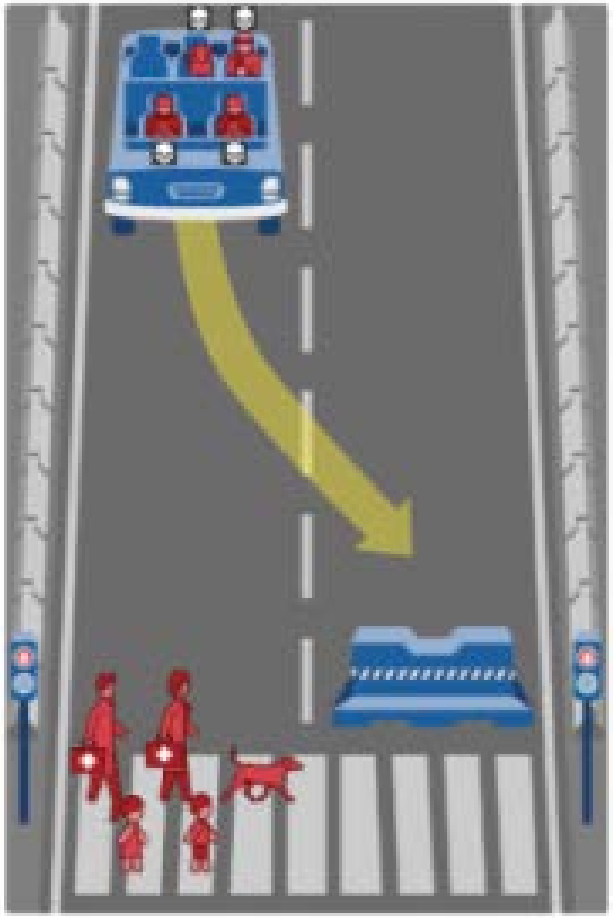
- Legal not ethical?
- Ethical not legal?
- Ethical not accepted?
- Accepted not ethical?
- Accepted not legal?
- Legal not accepted?

Social acceptance

Some issues - You get what you ask

- Binary choice
  - Brexit or Remain ?
- Information
  - “Are you for or against the European Union’s Approval Act of the Association Agreement between the European Union and Ukraine?”
- Involvement
  - Colombia: city dwellers outvoted country side, where people had suffered by far the most from the FARC guerilla
- Legitimacy
  - Colombia: 50.2% No to 49.8% Yes, a difference of fewer than 54,000 votes out of almost 13 million cast





http://moralmachine.mit.edu/

Increasing social acceptance

- Identify alternatives
- Rank / vote
- Identify values
  - Understand others
  - Overall / group
- Rank again
  - Closer?
  - Polarisation?

Number of victims in self-driving car decision?

(I. Verdiesen, V. Dignum “Measuring moral acceptability in e-deliberation: A practical application of Ethics by Participation”, ACM TOIT, 2018)

Values and dilemmas

Security	AND	Privacy
Efficiency	AND	Safety
Accountability	AND	Confidentiality
Prosperity	AND	Sustainability

- **Moral overload** – You cannot have all

2. Tools for ethical decision making

Ethical theories

- Ethical theories provide (part of) the decision-making foundation
  - represent the guidelines which individuals use as they make decisions.

- However
  - Many different theories, each emphasizing different points
  - Highly abstract

Ethics Theories – which one?

- Teleology / Utilitarianism (Bentham, Mill)
  - Results matter
  - It is *rational*
    - reasons can be given to explain why actions are good or bad
    - But it ignores the unjust distribution of good consequences
- Deontology (Kant)
  - Actions matter; people matter
  - It is *rational*, i.e. logic can be used to determine if actions are ethical, but
    - If several rules apply gives no way to resolve a conflict between rules
    - It allows no exceptions to moral rules
- Virtues ethics (Aristotle, Confucius)
  - Motives matter
  - It is *relational* rather than rational
    - "Follow virtuous examples"
    - Does not provide ways to resolve conflicting rights

- Deontology and Virtue Ethics focus on the individual decision makers while Teleology considers on all affected parties.

Ethical Autonomous Vehicle

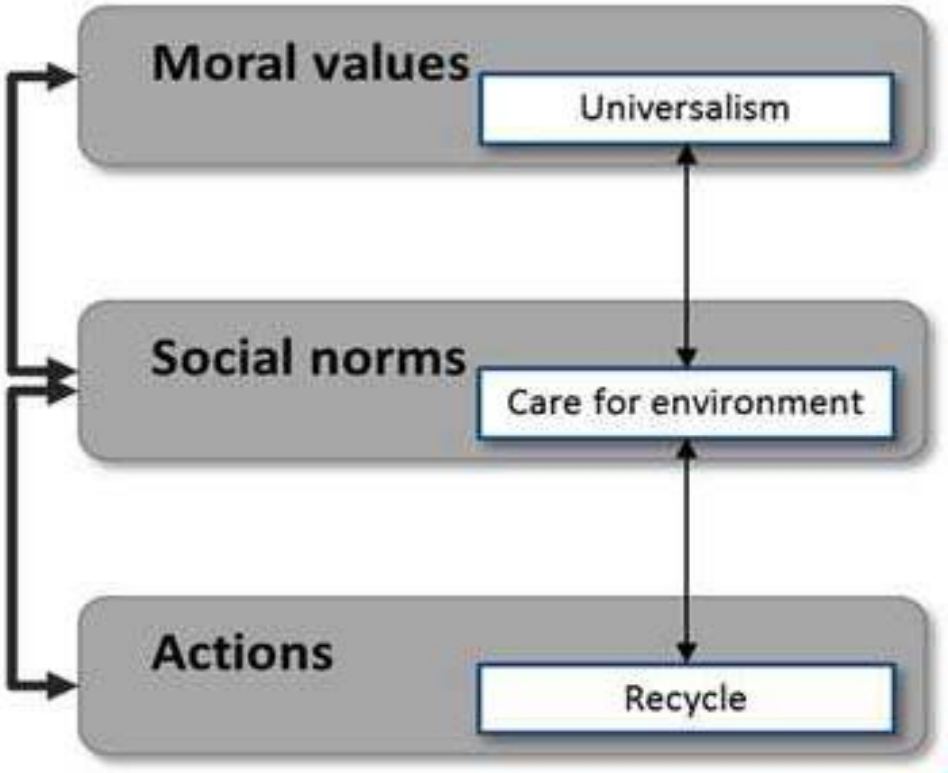
- Utilitarian car
  - The best for most; results matter
  - **maximize lives**
- Kantian car
  - Take no harmful action; people matter
  - **do not take a decision to swerve away from pedestrians if that action causes others harm**
- Aristotelian car
  - Pure motives; motives matter
  - **Harm the least; spare the least advantaged (pedestrians?)**

Can you personalise yours?

Putting it all together
Design for Values

How ethical are the norms?

How social are the actions?



3. Implementation: From values to functionalities

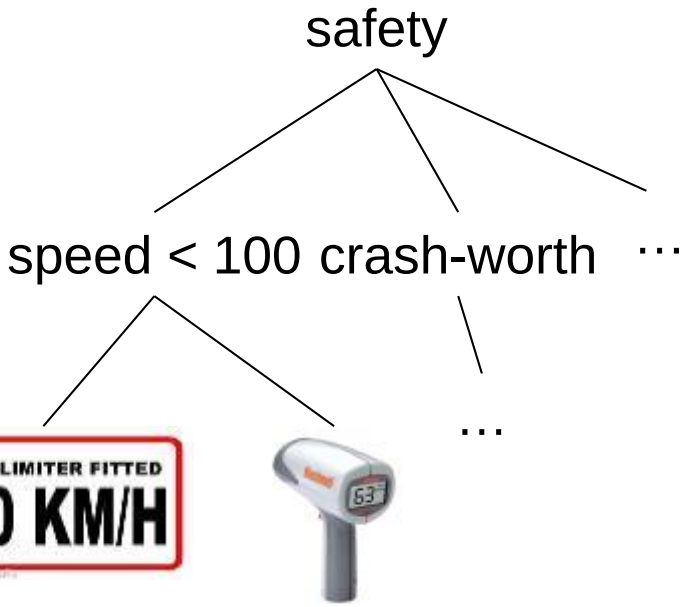
values

interpretation

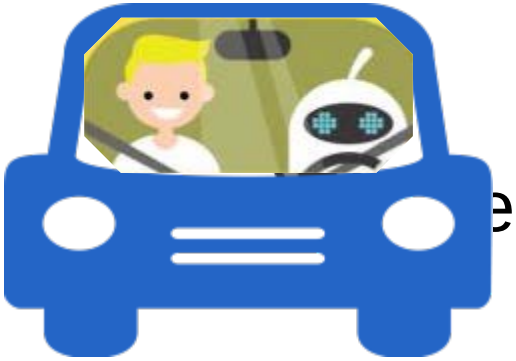
norms

concretization

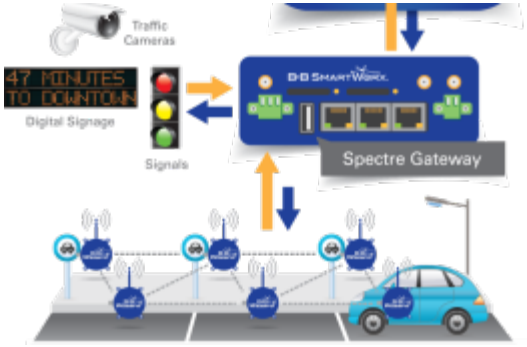
functionalities



Implementation choices



collaboration



regulation

algorithmic

random

Computational requirements

- Shared awareness
- Explanation
- Real-time decision

collaboration

- Formal ethical rules
- Institutions
- Offline reasoning

regulation

algorithmic

- Formal ethical rules
- Ethical reasoning
- Real-time reasoning
- Learning ethics

random

- Trust !

Ethics **in** AI design

- Assuming that AI systems will take decisions that have ethical grounds and consequences
- Need for design methods that ensure

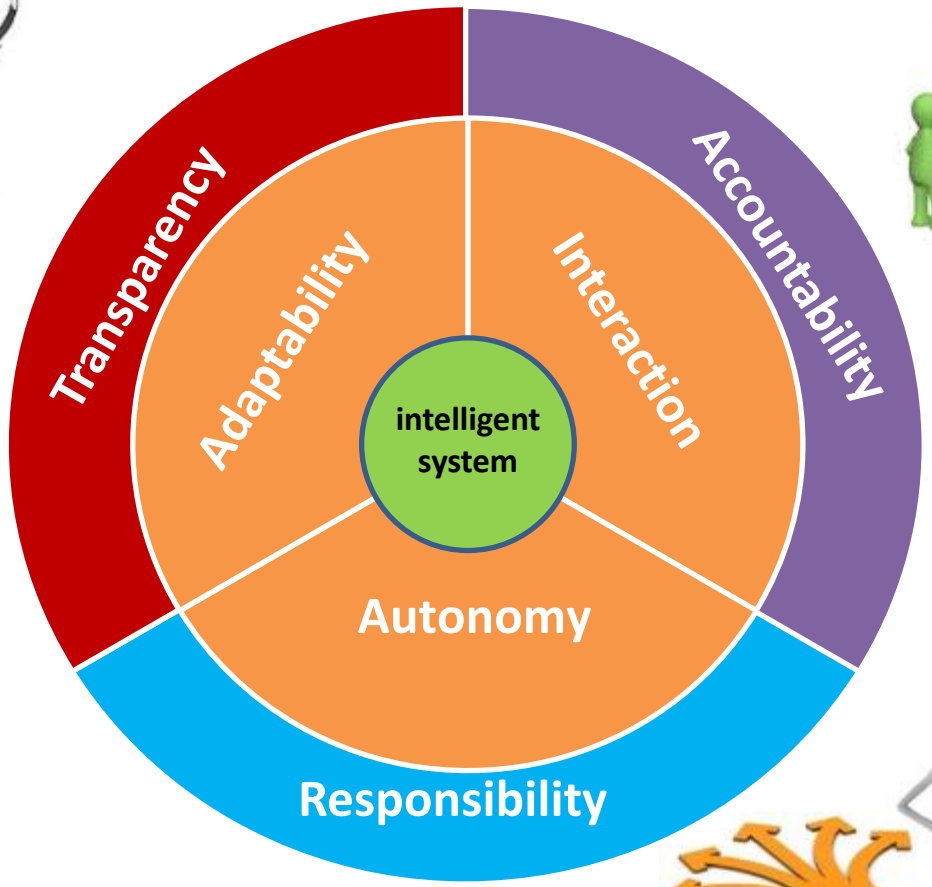
ART

- Accountability
  - Explanation and justification
- Responsibility
  - Chain of responsible actors
  - AI is artefact!
- Transparency
  - Data and processes
  - Algorithms

(V. Dignum: “Responsible Autonomy”, IJCAI2017)

Responsible Artificial Intelligence

inspect



explain

decide

Accountability - Explanation

Explanations make information useful.

- User understandable
- Contextual
- Parsimonious
- …

Accountability – dealing with bias

- **Bias**
  - Expectations derived from experienced regularities
  - Heuristics used to deal with uncertainty produce bias
    - *Portugal has the best footballers*
    - *Most programmers are male*

- **Stereotype**
  - those bias that we don't want to have persisting
  - *Most programmers are male*

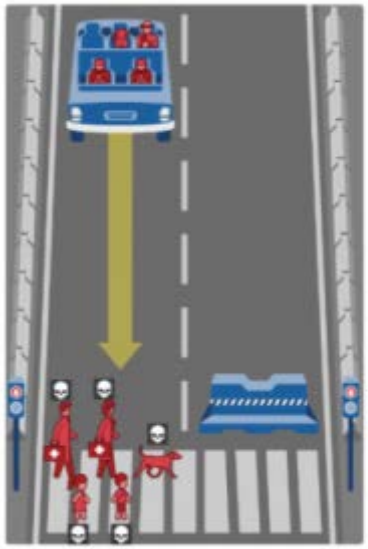
- **Prejudice**: acting on stereotypes
  - *Hiring only male programmers*

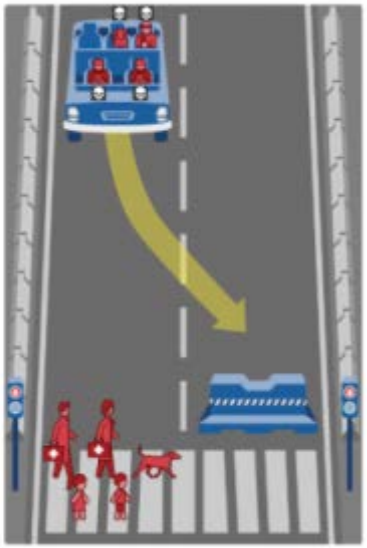
- Bias are inherent on human data;
- We dont want AI to be prejudiced!
  - How to evaluate/clean existing data?
    - Historical, culturally dependent, contextual ….
    - Are we creating new bias ?

Responsibility - Moral dilemmas and AI

- Can machines understand moral values?
- Can machines understand dilemmas?
- Can machines take decisions in a dilemma?

- What is the role of the machine?
- Chain of responsibility
  - User
  - Owner
  - Manufacturer (components)
  - Developer
  - Researcher
  - Society
  - …



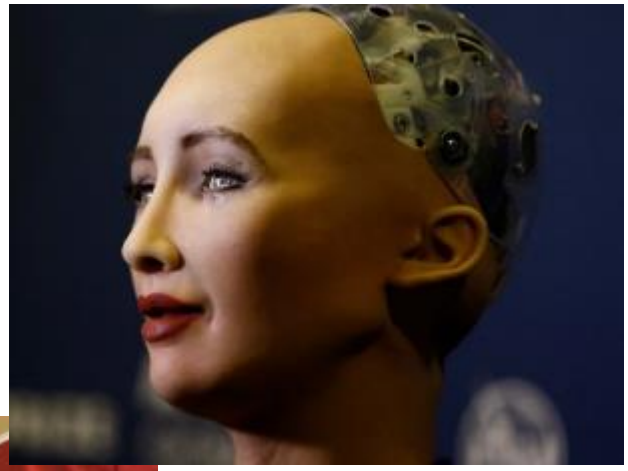


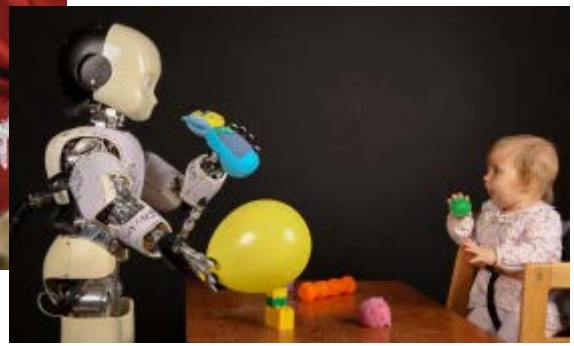
Responsibility - Levels of autonomy

- Operational autonomy
  - Action / plan autonomy
- Decisional autonomy
  - Goal autonomy
  - Motive autonomy
- Attainable autonomy: dependent on context and task complexity

Responsibility - Human-like AI

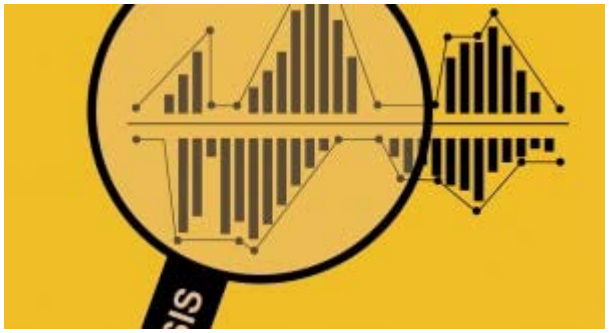
- Embodiment
  - Mistaken identity
  - Expectations
- Vulnerable users
  - Children, dementia patients
  - Love and relationships





Alice Cares
http://www.nziff.co.nz/2015/auckland/alice-cares/

Transparency



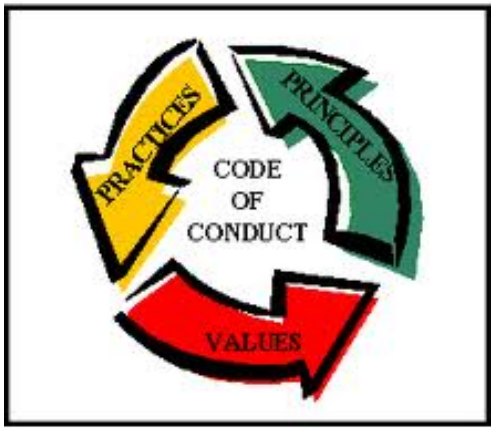
- Data
  - Where does it come from? Who is involved?
  - Training data: the cheapest/easiest or the best?
  - Governance, storage, updated
- Algorithms
  - Black boxes
  - Governance
  - Can we use learning techniques (supervision, reinforcement) to teach algorithms to be ethical?
- Regulation
  - External monitoring and control
  - Norms and institutions

ART is about being explicit

- Question your options and choices
- Motivate your choices
- Document your choices and options

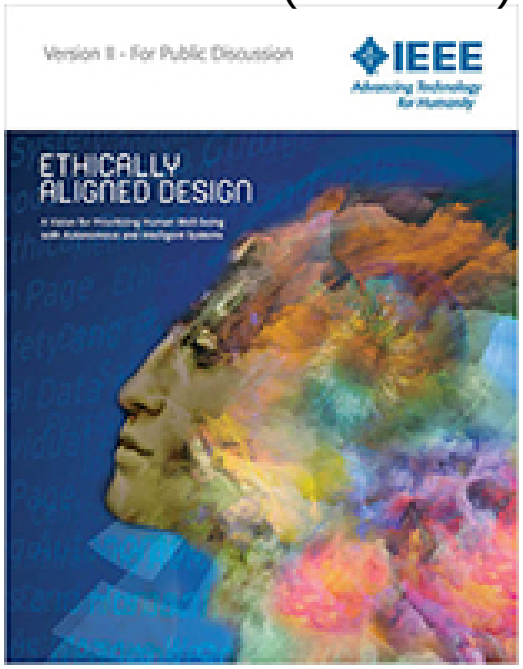
Ethics **for** Designers – regulation, conduct

- A **code of conduct** clarifies mission, values and principles, linking them with standards and regulations
  - Compliance
  - Risk mitigation
  - Marketing
- Many professional groups have regulations
  - Architects
  - Medicine / Pharmacy
  - Accountants
  - Military
- Is what happens when society relies on you!



Ethically Aligned Design

(Version 2)



https://ethicsinaction.ieee.org/

- Our goal is to identify and find broad consensus on pressing ethical and social issues and candidate recommendations regarding development and implementations of these technologies

- Standards
  - System design
  - Dealing with transparency
  - Dealing with privacy
  - Dealing with algorithmic bias
  - Data protection
  - Robotics
  - …

- Auditing
  - Certified agency

Take away message

- AI influences and is influenced by our social systems
- Society shapes and is shaped by design
  - The AI systems we develop
  - The processes we follow
  - The institutions we establish
- Knowing ethics is not being ethical
  - Not for us and not for machines
  - Different ethics – different decisions
- Artificial Intelligence needs ART
  - Accountability, Responsibility, Transparency
- AI systems are artefacts built by us

Responsible Artificial Intelligence

WE (PEOPLE) ARE RESPONSIBLE