RESPONSIBLE ARTIFICIAL INTELLIGENCE

Virginia Dignum

Chair of Social and Ethical AI - Department of Computer Science

Email: virginia@cs.umu.se - Twitter: @vdignum

UMEÅ UNIVERSITY

RESPONSIBLE RESEARCH AND INNOVATION

'A transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view on the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products' (Von Schomberg 2011)

- Dynamic process
- Stakeholder involvement
- Mutual responsiveness
- Shared responsibility for outcomes and process requirements

- democratic values regarding participation and power
- social and moral values regarding the care for the future of our planet and its people
- individual and institutional values of open-mindedness or receptiveness to change

UMEÅ UNIVERSITY

http://ec.europa.eu/research/swafs/pdf/pub_rri/rri_indicators_final_version.pdf

RESPONSIBILITY IN THE AGE OF AI

- Ethics
  - Can AI systems behave ethically? Understand ethics?
- Law
  - Need laws that protect users and yet accelerate R&D and utilization of AI
- Economy
  - Maximize the benefit from AI while minimizing the income gap between people who can take advantage of AI and those who can't
- Society
  - How can we avoid excessive dependence on and exaggerated fear of AI?
- Education
  - What should we learn to cope with AI?
- R&D
  - What should researchers do to make AI secure, transparent, controllable, and ethical?

RESPONSIBLE AI: WHY CARE?

- AI systems act autonomously in our world
- Eventually, AI systems will make *better* decisions than humans

AI is designed, is an artefact

- We need to sure that the purpose put into the machine is the purpose which we really want

Norbert Wiener, 1960 (Stuart Russell)

King Midas, c540 BCE

RESPONSIBLE AI

- AI can potentially do a lot. Should it?
- Who should decide?
- Which values should be considered? Whose values?
- How do we deal with dilemmas?
- Who is responsible?
- Who is accountable?
- .....

Setting boundaries: technical, ethical, societal

SOME CASES

- Self-driving cars
  - Who is responsible for the accident by self-driving car?
  - How can car decide in face of a moral dilemma?
- Automated manufacturing
  - How can technical advances combined with education programs (human resource development) help workers practice new sophisticated skills so as not to lose their jobs?
- Chatbots
  - Mistaken identity (is it a person or a bot?)
  - Manipulation of emotions / nudging / behaviour change support

RRI IN THE AI AGE

- consideration of the social contexts and consequences of decisions in the AI design space;
- reflectiveness about one's emphasis on theoretical vs. applied work and choice of application domains;
- engagement with the public about what they desire from AI and what they need to know about it.

TAKING RESPONSIBILITY

- Ethics **in** Design
  - Ensuring that development processes take into account ethical and societal implications of AI as it integrates and replaces traditional systems and social structures
- Ethics **by** Design
  - Integration of ethical reasoning abilities as part of the behaviour of artificial autonomous systems
- Ethics **for** Design(ers)
  - Research integrity of researchers and manufacturers, and certification mechanisms

AI - DOING THE RIGHT THING

- Taking an ethical perspective
  - Ethics is the new green
  - Business differentiation
  - Certification to ensure public acceptance

- Regulation is drive for transformation
  - Better solutions
  - Return on Investment

ETHICS *IN* DESIGN: AI – DOING IT RIGHT

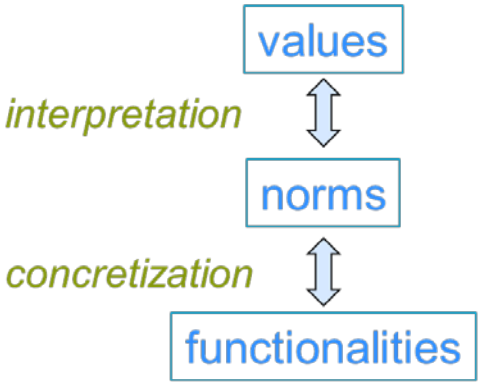
- Principles for Responsible AI = ART
  - **A**ccountability
    - Explanation and justification
    - Design for values
  - **R**esponsibility
    - Autonomy
    - Chain of responsible actors
    - Human-like AI
  - **T**ransparency
    - Data and processes
    - Algorithms

ACCOUNTABILITY CHALLENGES

- Optimal AI is explainable AI

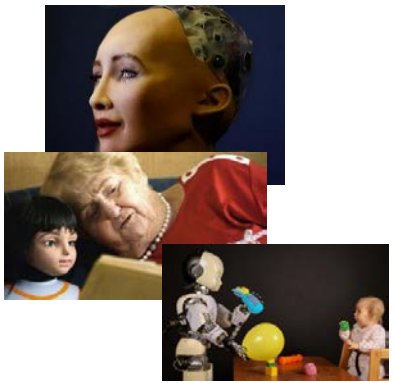
- Explanation
  - Human level understanding (“symbolic” vs “sub-symbolic”)
  - Possibly constructed ad posteriori
  - Social heuristics

- Design for values
  - include values of ethical importance in design
  - Explicit, systematic
  - Verifiable



RESPONSIBILITY CHALLENGES

- Chain of responsibility
  - researchers, developerers, manufacturers, users, owners, governments, ...
- Levels of autonomy
  - Operational autonomy: Actions / plans
  - Decisional autonomy: Goas/ motives
  - Attainable autonomy: dependent on context and task complexity
- Human-like AI
  - Mistaken identity / expectations
  - Vulnerable users: children / elderly



TRANSPARENCY CHALLENGES

- Manage expectations
  - Training wheels / L-plates
- Openness
  - Data, processes, stakeholders
- Algorithms and data
  - Bias is inherent in human behavior; acting on it not necessarily
  - Provenance: Where does it come from? Who is involved?
  - Training data: the cheapest/easiest or the best?
  - Governance, storage, history

ART IS ABOUT BEING EXPLICIT

- Question your options and choices
- Motivate your choices
- Document your choices and options
- Regulation
  - External monitoring and control
  - Norms and institutions
- Engineering principles for policy
  - Analyze – synthetize – evaluate - repeat

https://medium.com/@virginiadignum/on-bias-black-boxes-and-the-quest-for-transparency-in-artificial-intelligence-bcde64f59f5b

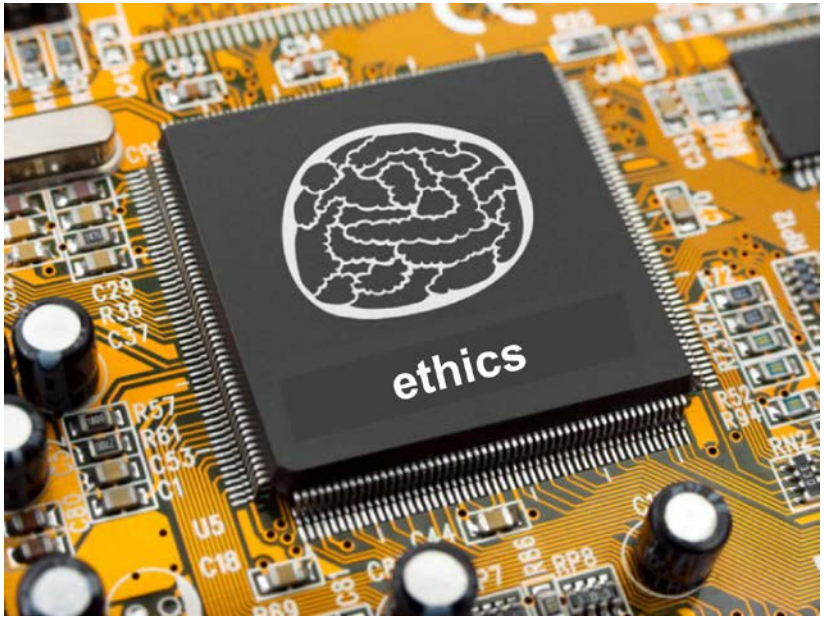
ART-FULL AI

- **Openness of data**
  - Which data was used to train the algorithm?
  - Which data does the algorithm use to make decisions?
  - How is this data governed (collection, storage, access....)
  - What are the characteristics of the data? How old is the data, where was it collected, by whom, how is it updated...
  - Is the data available for replication studies?
- **Openness of processes**
  - What are the assumptions?
  - What are the choices? And the reasons for choosing and the reasons not too choose
  - Who is making the design choices? And why are these groups involved and not others
  - How are the choices being determined? By majority, consensus, is veto possible...
  - What are the evaluation and validation methods used?
  - How is noise, incompleteness and inconsistency being dealt with?
- **Openness of stakeholders and stakes**
  - Who is involved in the process, what are their interests?
  - Who is paying and who is controlling?
  - Who are the users, and how are they involved (voluntary, paid, forced participation)

ETHICS *BY* DESIGN

- Can AI artefacts be build to be ethical?
- What does that mean?
- What is needed?

- Understanding ethics
- Using ethics
- Being ethical



ETHICS BY DESIGN

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?

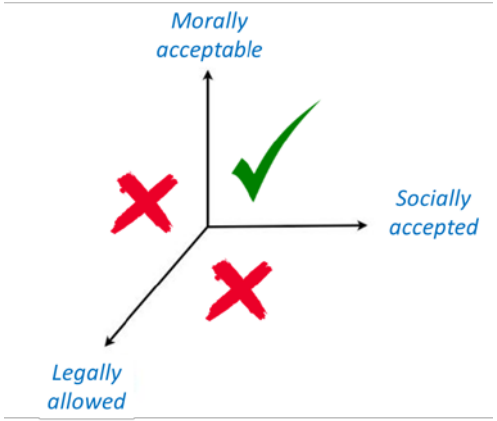
3. **How to implement?**
   - Role of user
   - Role of society
   - Role of AI system

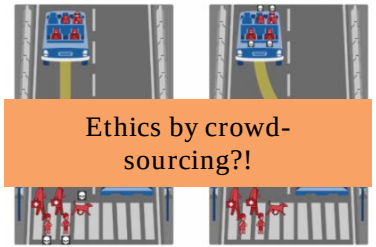
PARTICIPATION – AI VALUES

- Sources
  - Stakeholders: Designer, User, Owner, Manufacturer
  - Society: codes of ethics, codes & standards, law

- Identify what is
  - Socially accepted
  - Morally acceptable
  - Legally possible

- But
  - 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?





ETHICAL REASONING?

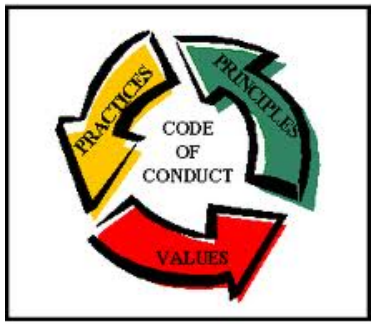
- Value: “human life”
- Reasoning:
- Utilitarian car
  - The best for most; results matter
  - **maximize lives**
- Kantian car
  - Do no harm
  - **do not take explicit action if that action causes harm**
- Aristotelian car
  - Pure motives; motives matter
  - **Harm the least; spare the least advantaged (pedestrians?)**

Ethical theories

- Many different theories, each emphasizing different points
  - Utilitarian, Kantian, Virtues….
- Highly abstract
- None provide ways to resolve conflicts
- Deontology and Virtue Ethics focus on the individual decision makers while Teleology considers on all affected parties.

ETHICS *FOR* DESIGN(ERS) – 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

- identify and find broad consensus on pressing ethical and social issues and define recommendations regarding development and implementations of these technologies

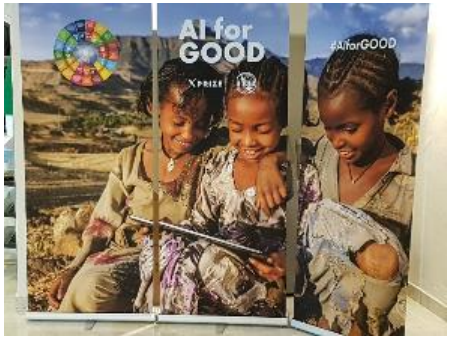
- Standards
  - System design
  - Dealing with transparency
  - Dealing with privacy
  - Dealing with algorithmic bias
  - Data protection
  - Robotics
  - ...

- Auditing
  - Certified agency

https://ethicsinaction.ieee.org/

AI ETHICS, AI FOR GOOD, AI FOR PEOPLE,...

- Harness the positive potential outcomes of AI in society, the economy
- Ensure inclusion, diversity, universal benefits
- Prioritize UN2020 Sustainable Development Goals
- The objective of the AI system is to maximize the realization of human values



https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

UMEÅ UNIVERSITY

http://www.ai4people.eu

TAKE AWAY MESSAGE

- AI influences and is influenced by our social systems
- Design in never value-neutral
- 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
  - Be explicit!
- AI systems are artefacts built by us for our own purposes
- We set the limits

RESPONSIBLE ARTIFICIAL INTELLIGENCE

WE ALL ARE RESPONSIBLE