RESPONSIBLE ARTIFICIAL INTELLIGENCE - A GLASS BOX APPROACH

Prof. Dr. Virginia Dignum

Chair of Social and Ethical Artificial Intelligence - Department of Computer Science

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

UMEÅ UNIVERSITY

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

TAKING RESPONSIBILITY

- Responsiblity / 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
- Responsibility /Ethics **by** Design
  - Integration of ethical abilities as part of the behaviour of artificial autonomous systems
- Responsibility /Ethics **for** Design(ers)
  - Research integrity of researchers and manufacturers, and certification mechanisms

TAKING RESPONSIBILITY

- Responsiblity / 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
- R
- Responsibility /Ethics **for** Design(ers)
  - Research integrity of researchers and manufacturers, and certification mechanisms

Can we guarantee that behaviour is ethical?

cial

ETHICS BY DESIGN

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

- Which values?
- Whose values?
- Which ethical rules?
- **Which interpretation?**

VALUES IN CONTEXT

DECISIONS MATTER!

values

interpretation

norms

concretization

functionalities

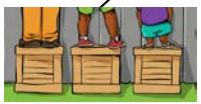
Design for Values

fairness

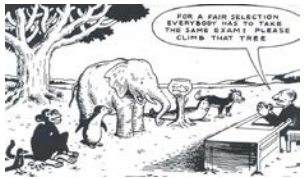
Equal resources

Equal opportunity

…



…



DECISIONS MATTER!

safety

Limit speed

Ensure crash-worthiness

…

…

values

interpretation

norms

concretization

functionalities

Design for Values

GUIDELINES – BE OPEN AND EXPLICIT

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

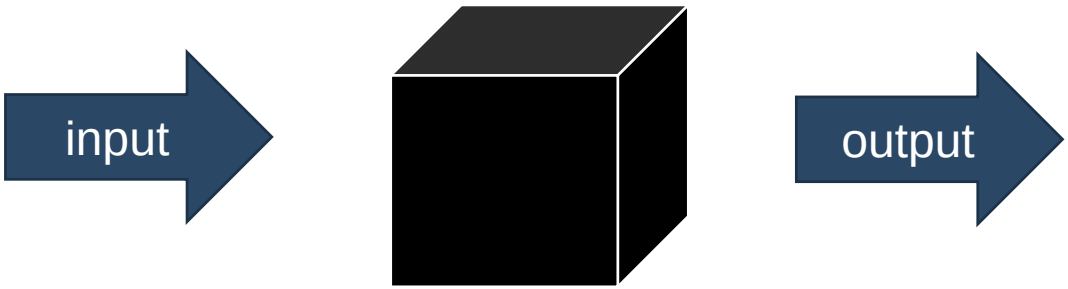
ASK YOURSELF

- Who will be affected?
- What are the decision criteria we are optimising for?
- How are these criteria justified?
- Are these justifications acceptable in the context we are designing for?
- How are we training our algorithm?
  - Does training data resemble the context of use?

UMEÅ UNIVERSITY

IEEE standard Algorithmic bias
https://standards.ieee.org/project/7003.html

ALGORITHMS - THE BLACK BOX?



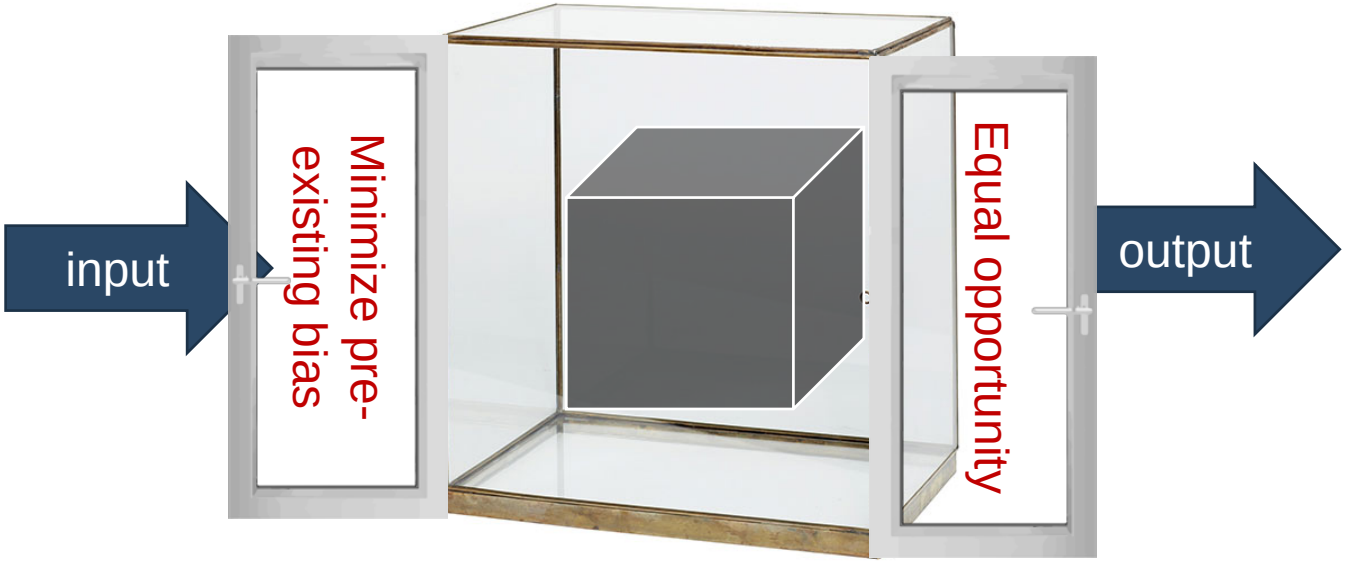
GOVERNANCE - THE GLASS BOX

input

output

GOVERNANCE - THE GLASS BOX

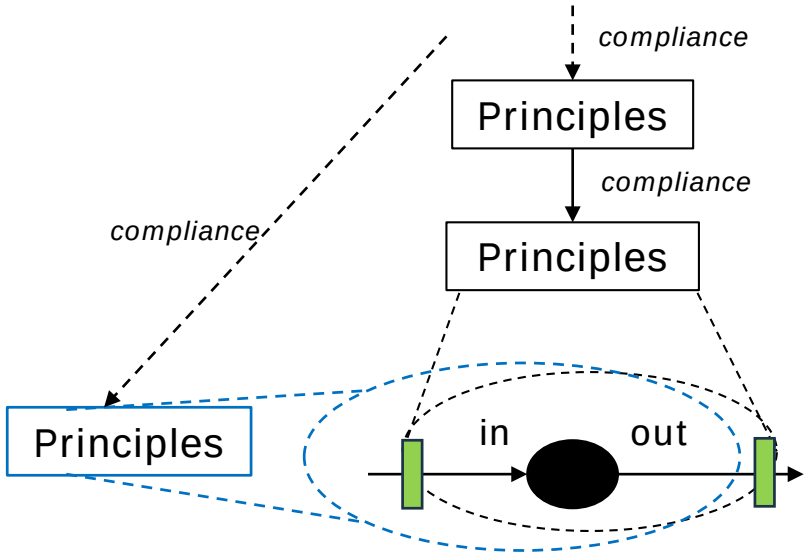
Fairness



UMEÅ UNIVERSITY

GOVERN AND VERIFY - GLASS BOXES

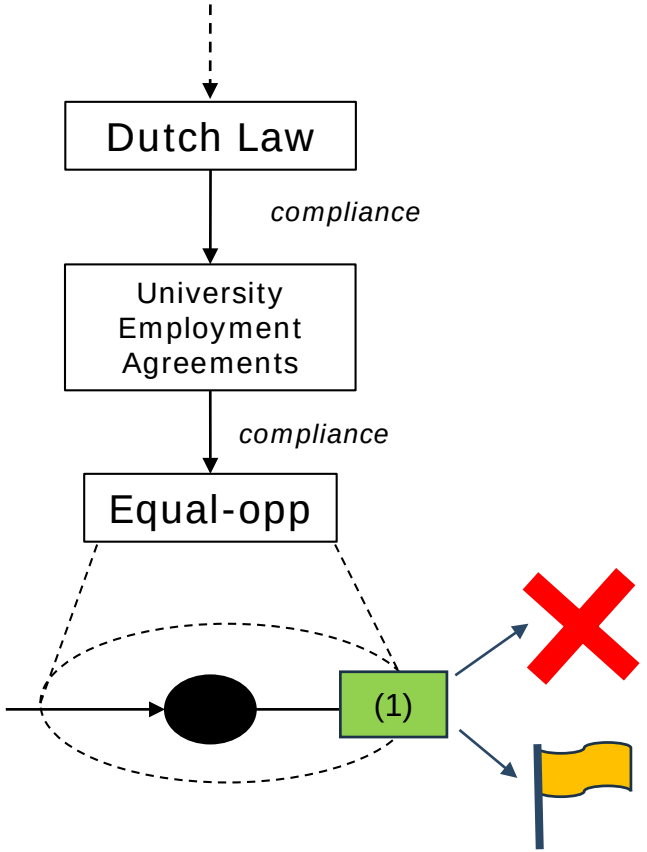
- Verify limits to action and decision
- Define the ethical borders
  - Formal
  - Monitoring input – output
- Governance
  - Monitor
  - “block” undesirable



EXAMPLE - FAIRNESS

- Value: Fairness
- Norm: Equal opportunity
- Implementation:
  - Output evaluation

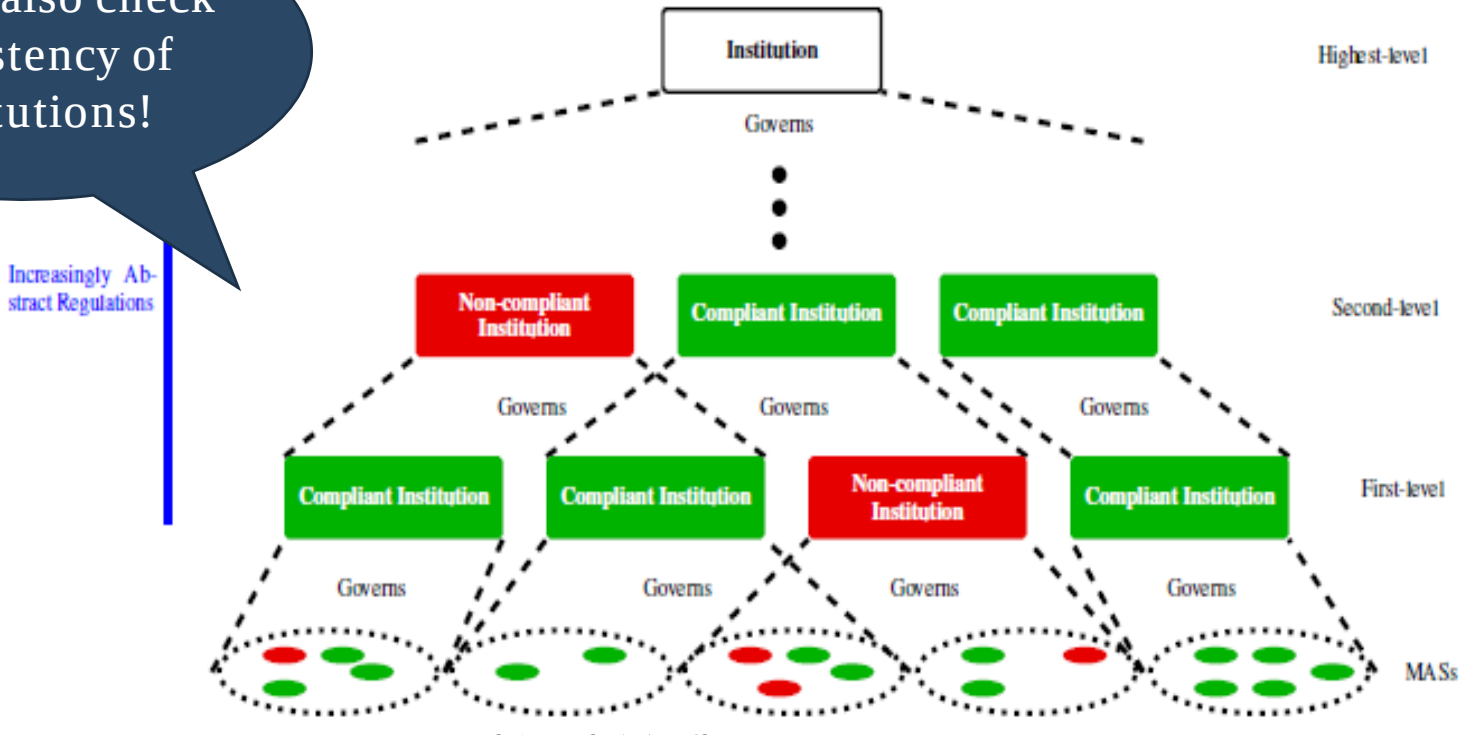
    (1) P(job | female) = P(job | male)
- Governance
  - Cut-off
  - Flag-out



extending (TC King et al, AAMAS 2015)

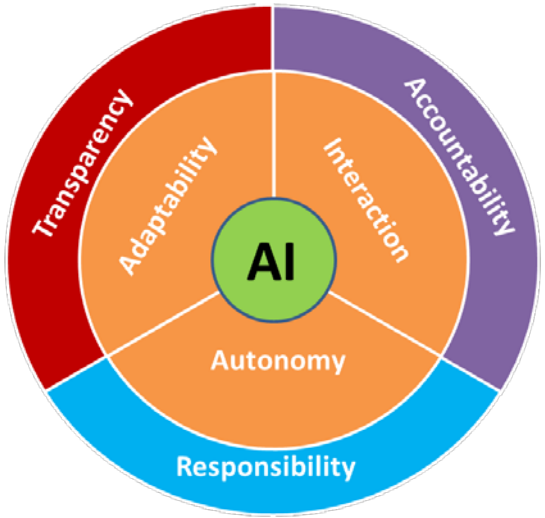
GOVERNANCE TRANSPARENCY

We can also check consistency of institutions!



RESPONSIBLE DESIGN – ART OF AI

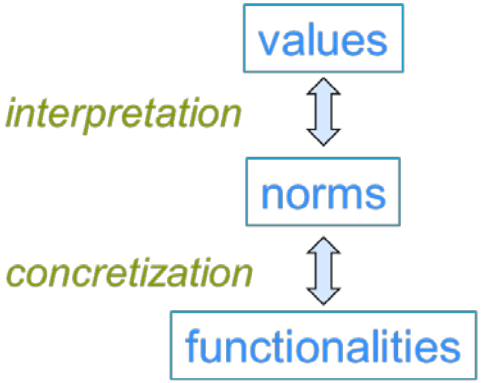
- Principles for Responsible AI = ART
  - **A**ccountability
    - Explanation and justification
    - Design for values
  - **R**esponsibility
    - Autonomy
    - Chain of responsible actors
  - **T**ransparency
    - Data and processes
    - Algorithms



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

ART METHODOLOGY

- Socially accepted
  - Participatory
- Ethically acceptable
  - Ethical theories and human values
- Legally allowed
  - Laws and regulations



- Engineering principles
  - Cycle: Analyse – synthetize – evaluate –repeat
  - Report: Identify, Motivate, Document

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
- Openness and explicitness are key!
  - Accountability, Responsibility, Transparency
- AI systems are artefacts built by us for our own purposes
- We set the limits

Center for Responsible AI @Umeå

A research institute dedicated to develop AI systems that meet their social responsibility:

- Understand social implications
- Develop theories, models and tools for oversight, accountability and verification
- Methods to design, measure and audit social implications

http://people.cs.umu.se/virginia

We are hiring!!

UMEÅ UNIVERSITY