RESPONSIBLE ARTIFICIAL INTELLIGENCE

Prof. Dr. Virginia Dignum

Chair of Social and Ethical AI - Department of Computer Science

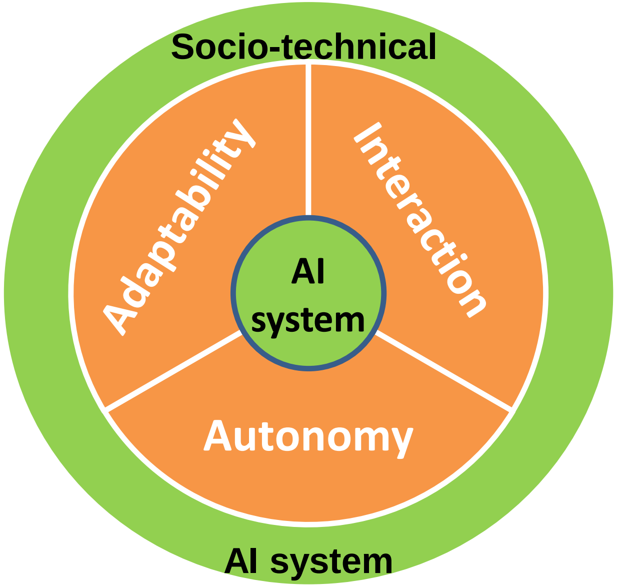
Email: virginia@cs.umu.se - Twitter: @vdignum

UMEÅ UNIVERSITY

WHAT IS AI?

- Not just algorithm
- Not just machine learning

- But
- AI applications are not alone
  - Socio-technical AI systems

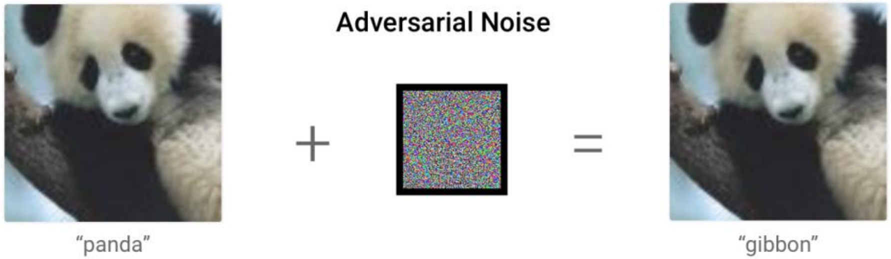


AI IS NOT INTELLIGENCE!

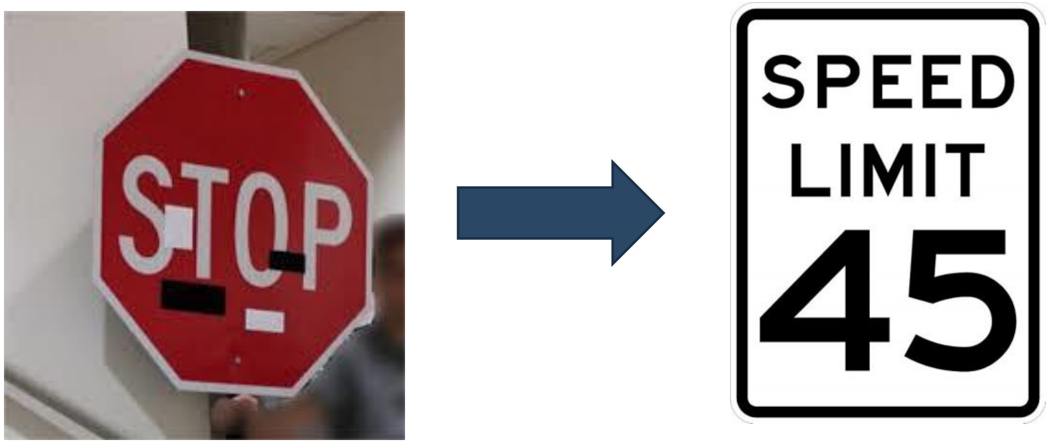
- What AI systems cannot do (yet)
  - Common sense reasoning
    - Understand context
    - Understand meaning
  - Learning from few examples
  - Learning general concepts
  - Combine learning and reasoning

- What AI systems can do (well)
  - Identify patterns in data
    - Images
    - Text
    - Video
  - Extrapolate those patterns to new data
  - Take actions based on those patterns

AI IS NOT INTELLIGENCE!



AI IS NOT INTELLIGENCE!



WHAT IS RESPONSIBLE AI?

Responsible AI is

- Ethical
- Lawful
- Reliable
- Beneficial

Responsible AI recognises that

- AI systems are artefacts
- We set the purpose
- We are responsible!

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?
- How should values be prioritized?
- .....

PRINCIPLES AND GUIDELINES

Responsible / Ethical / Trustworthy....

The European Commission's

High-Level Expert Group on Artificial Intelligence

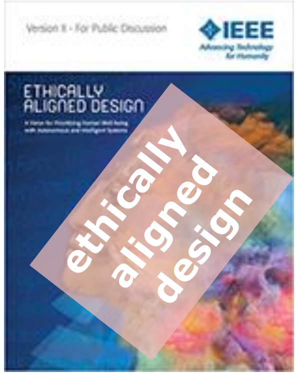
Draft Ethics Guidelines for Trustworthy AI

Working Document for stakeholders' consultation

Brussels, 18 December 2018

trustworthy

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



Version II - For Public Discussion

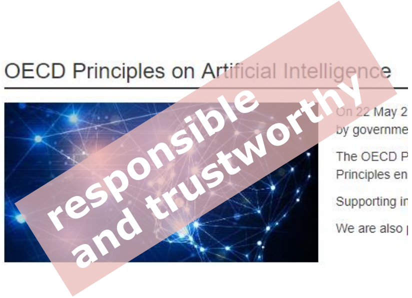
IEEE Advancing Technology for Humanity

ETHICALLY ALIGNED DESIGN

A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems

ethically aligned design

https://ethicsinaction.ieee.org



OECD Principles on Artificial Intelligence

On 22 May 2
by governme

The OECD P
Principles en

Supporting ir

We are also

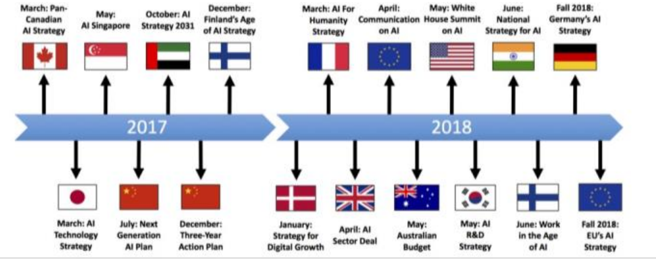
responsible and trustworthy

https://www.oecd.org/going-digital/ai/principles/

UMEÅ UNIVERSITY

MANY INITIATIVES (AND COUNTING…)

- Strategies / positions
  - IEEE
  - European Union
  - OECD
  - WEF
  - Council of Europe
  - Many national strategies
  - …
- Declarations
  - Asilomar
  - Montreal
  - …


March: Pan-Canadian AI Strategy
May: AI Singapore
October: AI Strategy 2031
December: Finland's Age of AI Strategy
March: AI For Humanity Strategy
April: Communication on AI
May: White House Summit on AI
June: National Strategy for AI
Fall 2018: Germany's AI Strategy
2017
2018
March: AI Technology Strategy
July: Next Generation AI Plan
December: Three-Year Action Plan
January: Strategy for Digital Growth
April: AI Sector Deal
May: Australian Budget
May: AI R&D Strategy
June: Work in the Age of AI
Fall 2018: EU's AI Strategy

https://arxiv.org/ftp/arxiv/papers/1906/1906.11668.pdf lists 84!

EU HLEG	OECD	IEEE EAD
• Human agency and oversight • **Technical robustness and safety** • Privacy and data governance • **Transparency** • **Diversity**, non-discrimination and fairness • **Societal and environmental well-being** • **Accountability**	• benefit people and the planet • respects the rule of law, **human rights**, democratic values and **diversity**, • include appropriate safeguards (e.g. human intervention) to ensure a **fair and just society**. • **transparency** and responsible disclosure • **robust, secure and safe** • Hold organisations and individuals **accountable** for proper functioning of AI	• How can we ensure that A/IS do not infringe **human rights**? • effect of A/IS technologies on **human well-being**. • How can we assure that designers, manufacturers, owners and operators of A/IS are responsible and **accountable**? • How can we ensure that A/IS are **transparent**? • How can we extend the benefits and minimize the risks of AI/AS technology being misused?

OCDE
LES POLITIQUES POUR UNE VIE
BETT
AI HUMANITY
L'INTELLIGENCE ARTIFICIELLE AU SERVICE DE L'HUMAIN
PARIS – JEUDI 29 MARS 2018
ropean Commission
mmission européenne

AI for Good
Global Summit

28-31 May 2019
Geneva, Switzerland

Accelerating progress towards the SDGs

PRACTICAL INFORMATION | REGISTER NOW | PROGRAMME
WATCH WEBCAST | SPEAKERS | DEMOS

Connecting AI innovators with problem owners for sustainable development

The AI for Good Global Summit is THE leading United Nations platform for global and inclusive

Influential audience

Leading minds in AI, business, government, civil society and international cooperation. Global coverage in world-renowned media.

Action oriented

Project pitches by 'AI Breakthrough Tracks' mentored by expert audience. Projects aim for global impact in the near term.

Exhibits and Demos

Latest innovations in AI on show, demonstrated by their supporting

Join the movement

Network with like-minded innovators in search of global impact. Meet your

BUT ENDORSEMENT IS NOT (YET) COMPLIANCE

EU HLEG	OECD	IEEE EAD
• Human agency and oversight • **Technical robustness and safety** • Privacy and data governance • **Transparency** • **Diversity**, non-discrimination and fairness • **Societal and environmental well-being** • **Accountability**	• benefit people and the planet • respects the rule of law, **human rights**, democratic values and **diversity**, • include appropriate safeguards (e.g. human intervention) to ensure a **fair and just society**. • **transparency** and responsible disclosure • **robust, secure and safe** • Hold organisations and individuals **accountable** for proper functioning of AI	• How can we ensure that A/IS do not infringe **human rights**? • effect of A/IS technologies on **human well-being**. • How can we assure that designers, manufacturers, owners and operators of A/IS are responsible and **accountable**? • How can we ensure that A/IS are **transparent**? • How can we extend the benefits and minimize the risks of AI/AS technology be
regulation	**observatory**	**standards**

The promise of AI: Better decisions

HOW DO WE MAKE DECISIONS?

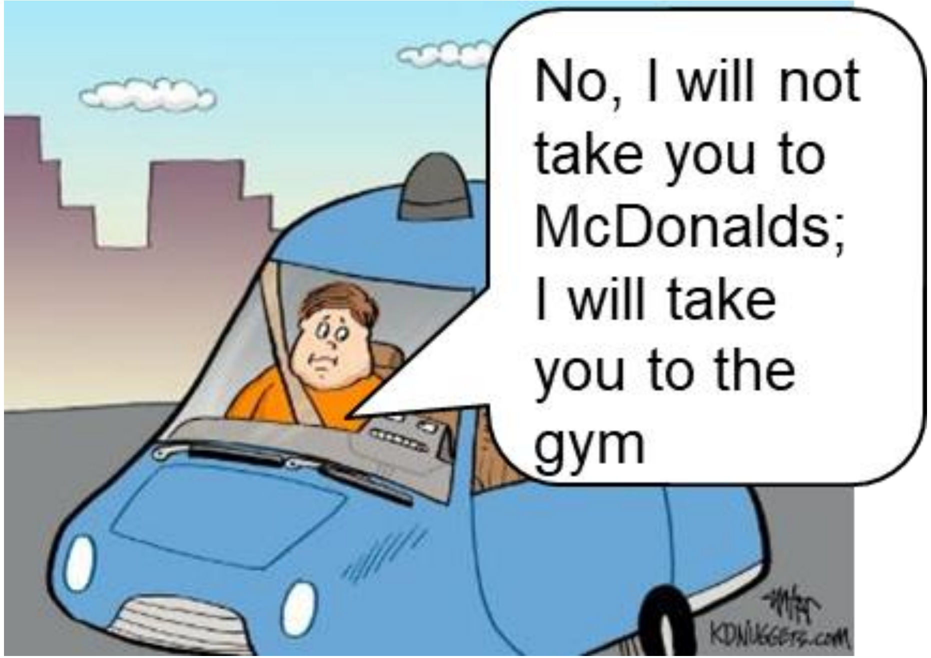
I'M WAITING TO SEE HOW MANY "LIKES" MY DIAGNOSIS GETS BEFORE I SHARE IT WITH YOU.

HOW DO WE MAKE DECISIONS TOGETHER?

DESIGN IMPACTS DECISIONS IMPACTS SOCIETY

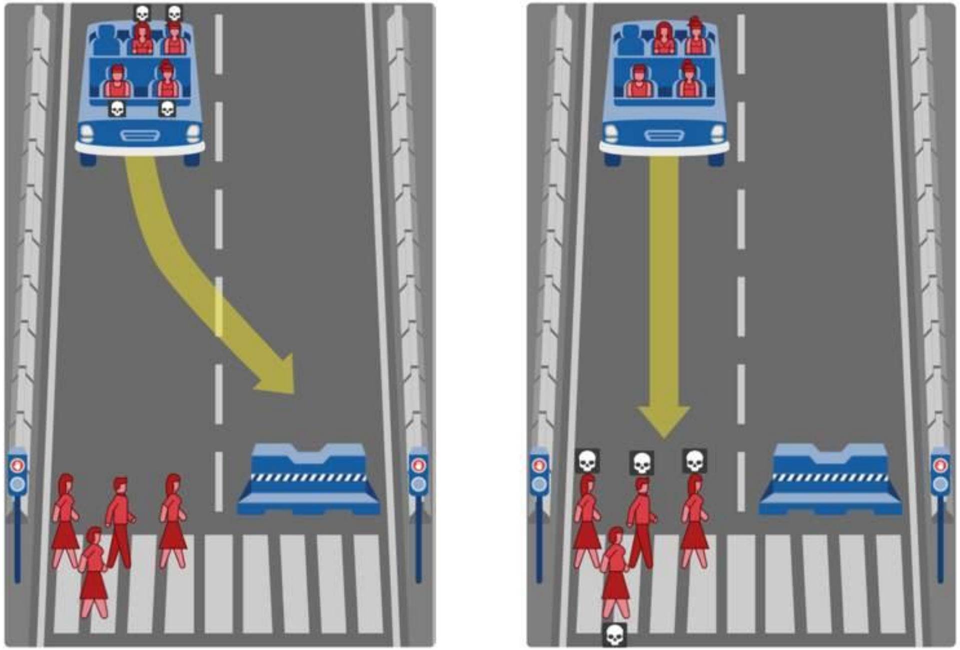
- Choices
- Formulation
- Involvement
- Legitimacy
- Voting system

WHICH DECISIONS SHOULD AI MAKE?

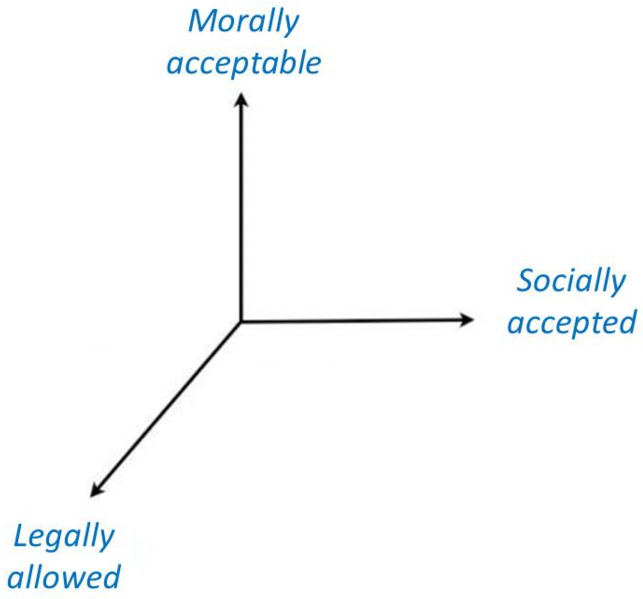

No, I will not take you to McDonalds; I will take you to the gym
KDNUGGETS.COM

WHICH DECISIONS SHOULD AI MAKE?

What should the self-driving car do?



HOW SHOULD AI MAKE DECISIONS?



Morally acceptable

Socially accepted

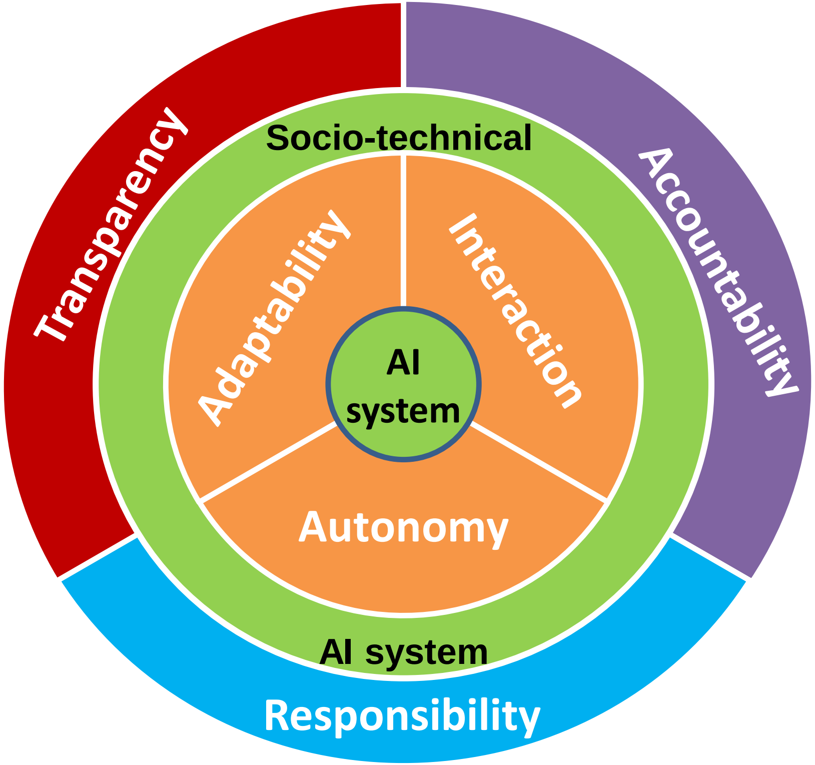
Legally allowed

TAKING RESPONSIBILITY

- **in** Design
  - Ensuring that development processes take into account ethical and societal implications of AI and its role in socio-technical environments
- **by** Design
  - Integration of ethical reasoning abilities as part of the behaviour of artificial autonomous systems
- **for** Design(ers)
  - Research integrity of stakeholders (researchers, developers, manufacturers,...) and of institutions to ensure regulation and certification mechanisms

IN DESIGN: ART

- AI needs ART
  - **A**ccountability
  - **R**esponsibility
  - **T**ransparency



ACCOUNTABILITY

- Principles for Responsible AI = ART
  - Accountability
    - Explanation and justification
    - Design for values
  - Responsibility
  - Transparency

- Optimal AI is explainable AI
- Many options, not one 'right' choice
- Explanation is for the user: context matters

CHALLENGE: NO AI WITHOUT EXPLANATION

Terms and Conditions

[illegible]

Fire Action

Any person discovering a fire

1. Sound the alarm.
2. ________ to call the fire brigade
3. Attack the fire if possible using the appliances provided

On hearing the fire alarm

4. Leave the building by ________ route
5. Close all doors behind you
6. Report to assembly point

Do not take risks

Do not return to the building for any reason until authorised to do so

- Explanation is for the user:
  - Different needs, different expertises and interests
  - Just in time, clear, concise, understandable, correct
- Explanation is about:
  - individual decisions and the ‘big picture’
  - enable understanding of overall strengths & weaknesses
  - convey an understanding of how the system will behave in the future
  - convey how to correct the system’s mistakes

RESPONSIBILITY

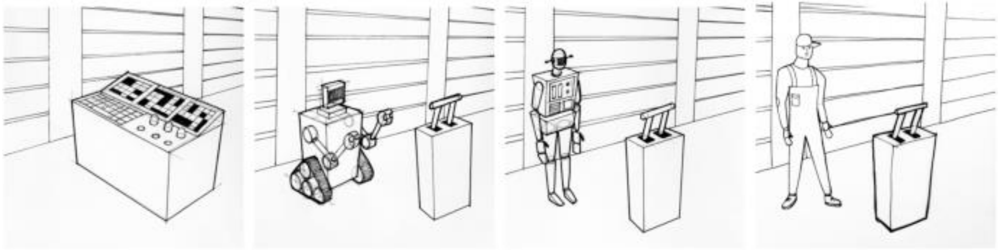
- Principles for Responsible AI = ART
  - Accountability
    - Explanation and justification
    - Design for values
  - Responsibility
    - Autonomy
    - Chain of responsible actors
    - Human-like AI
  - Transparency

The machine is not responsible!

RESPONSIBILITY CHALLENGES

- Chain of responsibility
  - researchers, developerers, manufacturers, users, owners, governments, …
  - Liability and conflict settling mechanisms

- Human-like systems
  - Robots, chatbots, voice…
  - Expectations
  - Vulnerable users
  - Mistaken identity



https://ieeexplore.ieee.org/document/7451743

- Responsibility for choices
  - 95% accurate but no explanation or 80% accurate with explanation?
  - Fairness or sustainability?

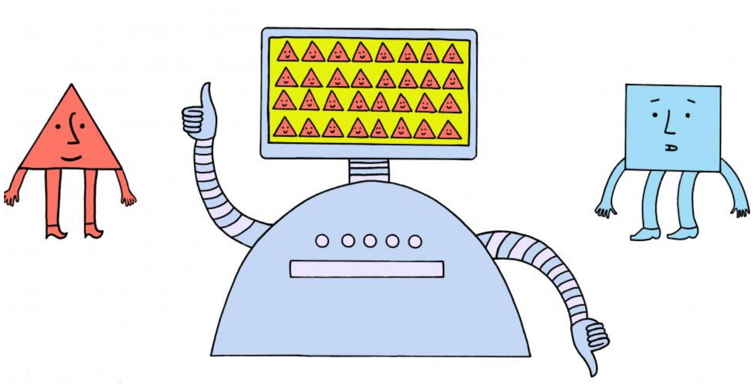
TRANSPARENCY

- Principles for Responsible AI = ART
  - Accountability
    - Explanation and justification
    - Design for values
  - Responsibility
    - Autonomy
    - Chain of responsible actors
    - Human-like AI
  - Transparency
    - Data and processes
    - Algorithms
    - Choices and decisions

CHALLENGE: BIAS AND DISCRIMINATION

Remember: AI systems extrapolate patterns from data to take action

- Bias is inherent on human data
  - we need bias to make sense of world
- Bias leads to stereotyping and prejudice
- Bias is more than biased data



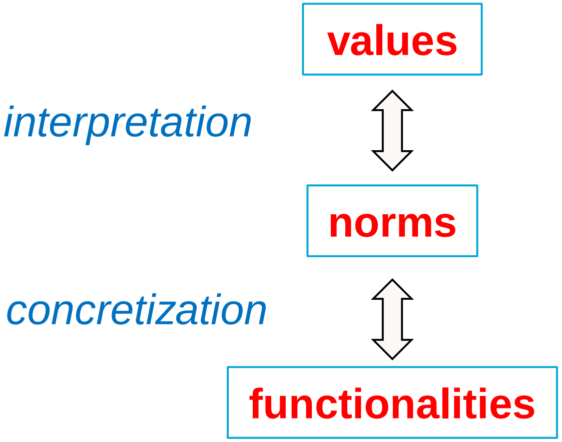
BY DESIGN: ARTIFICIAL AGENTS

- Can AI systems be ethical?
    - What does that mean?
    - What is needed?
- Design for values

ETHICAL BEHAVIOR

- **Should** we teach ethics to AI?
- Understanding ethics
  - Which values? Whose values?
  - Who gets a say?
- Using ethics
  - What is proper action given a values?
  - Are ethical theories of use?
  - How to prioritise values?
  - Is knowing ethics enough?
- Ethical reasoning
  - Many different theories
    - Utilitarian, Kantian, Virtues…)
  - Highly abstract
  - Do not provide ways to resolve conflicts

DESIGN FOR VALUES



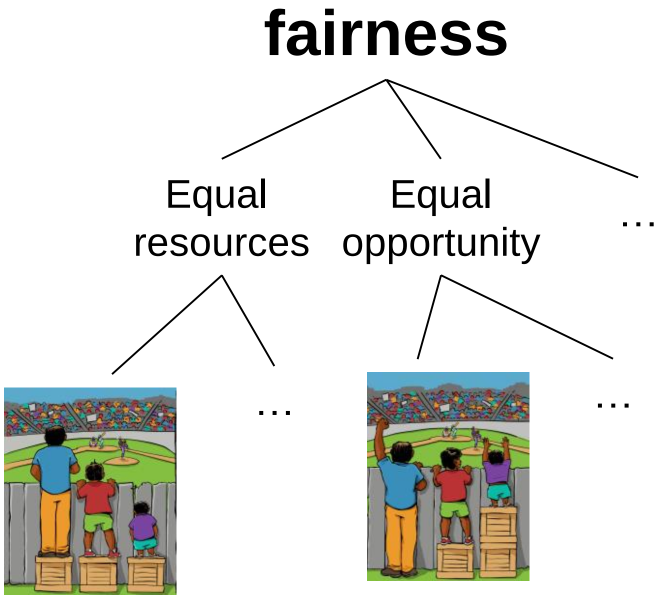
values

interpretation

norms

concretization

functionalities



fairness

Equal resources

Equal opportunity

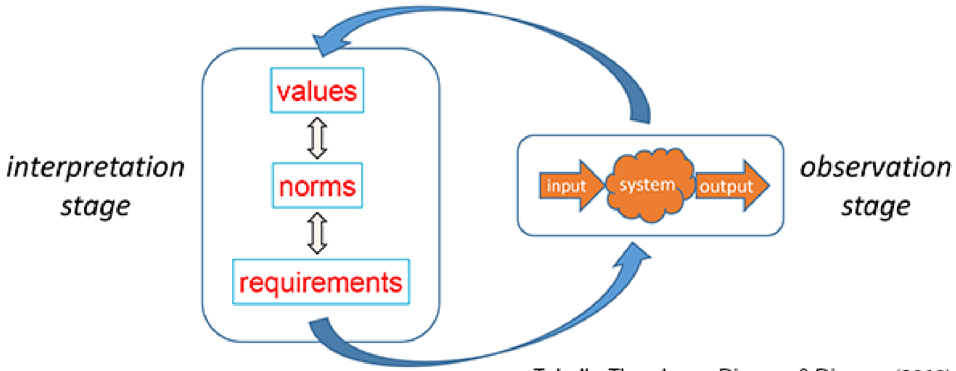
…

…

…

GLASS BOX APPROACH

- Doing the right thing
  - Elicit, define, agree, describe, report
- Doing it right
  - Explicit values, principles, interpretations, decisions
  - Evaluate input/output against principles



Tubella, Theodorou, Dignum & Dignum (2019)
https://arxiv.org/abs/1905.04994

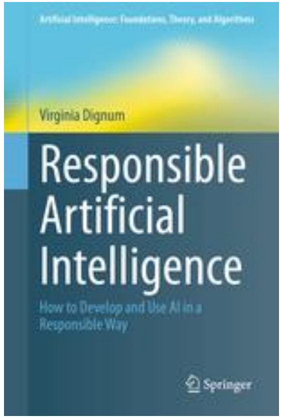
FOR **DESIGN(ERS): PEOPLE**

- Regulation
- Certification
- Standards
- Conduct

- AI principles are principles for us

FOR DESIGN: TRUSTWORTHY AI

- Regulation and certification
- Codes of conduct
- Human-centered
- AI as driver for innovation



Artificial Intelligence: Foundations, Theory, and Algorithms

Virginia Dignum

Responsible Artificial Intelligence

How to Develop and Use AI in a Responsible Way

Springer

- Design impacts decisions impacts society impacts design
- AI systems are tools, artefacts made by people:
  We set the purpose
- AI can give answers, but we ask the questions
- AI needs ART (Accountability, Responsibility, Transparency)

RESPONSIBLE ARTIFICIAL INTELLIGENCE

WE ARE RESPONSIBLE

Email: virginia@cs.umu.se

Twitter: @vdignum

UMEÅ UNIVERSITY