(HOW) SHOULD AI BE REGULATED?

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?

- A technology
  - Currently mostly pattern matching (stochastic, non-deterministic)

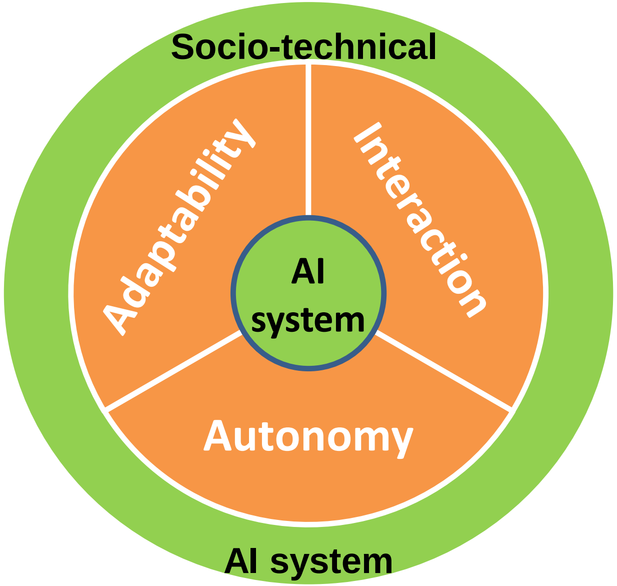
- A field of science
  - Model intelligence as means to understand intelligence

- An entity
  - Magic, all-knowing, all-powerful

WHAT IS AI?

- Not just algorithm
- Not just machine learning

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

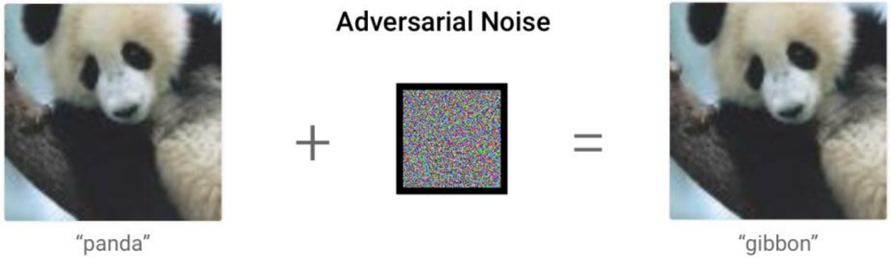


AI IS NOT INTELLIGENCE!

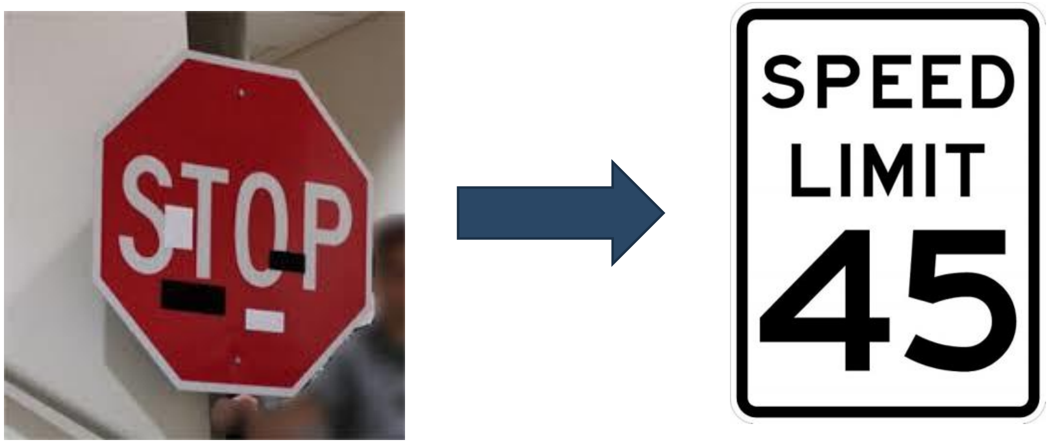
- What AI systems cannot do (yet)
  - Common sense reasoning
    - Understand context
    - Understand meaning
  - Learning from few examples
  - Learning general concepts
  - Combine learning and reasoning

- 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 ARE THE RISKS?

bias and prejudice

discrimination

misuse

loss of self-determination

removing human responsibility

devaluation of human skills

lack of control

WHAT ARE THE GAINS?

- Medicine
- Climate
- Education
- Work
- Communication

WHAT IS RESPONSIBLE AI?

Responsible AI is

- Ethical
- Lawful
- Reliable
- Beneficial

Responsible AI recognises that

- AI systems are artefacts
- We set the purpose

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?
- .....

Essential question: shoud we use AI here?

AI AND ETHICS - SOME CASES

- Not just trolley problems!!
- 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
- Automated decision making
  - Accuracy versus explainability
  - E.g. 95% accuracy but no explanation or 80% accuracy but always explains

PRINCIPLES AND GUIDELINES

Responsible / Ethical / Trustworthy....

The European Commission's

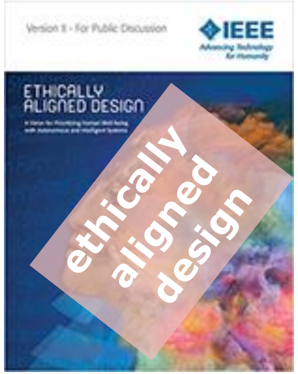
High-Level Expert Group on Artificial Intelligence

Draft Ethics Guidelines for Trustworthy AI

Working Document for stakeholders' consultation

Brussels, 18 December 2018

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

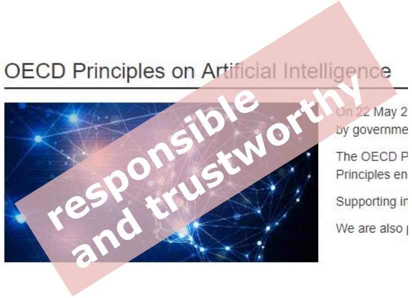

Version II - For Public Discussion

IEEE Advancing Technology for Humanity

ETHICALLY ALIGNED DESIGN

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

https://ethicsinaction.ieee.org


OECD Principles on Artificial Intelligence

On 22 May 2
by governme

The OECD P
Principles en

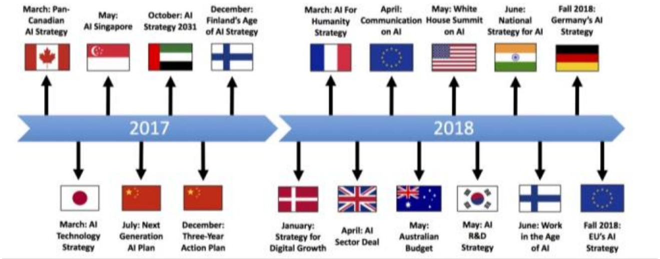
Supporting ir

We are also |

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

MANY INITIATIVES (AND COUNTING…)

- Strategies / positions
  - IEEE Ethically Aligned Design
  - European Union
  - OECD
  - WEF
  - Council of Europe
  - National strategies:
    - Tim Dutton, https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd
  - …
- 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

Check Alan Winfield blog: http://alanwinfield.blogspot.com/2019/04/an-updated-round-up-of-ethical.html

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?

AI for Good
Global Summit

28-31 May 2019
Geneva, Switzerland

Accelerating progress towards the SDGs

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

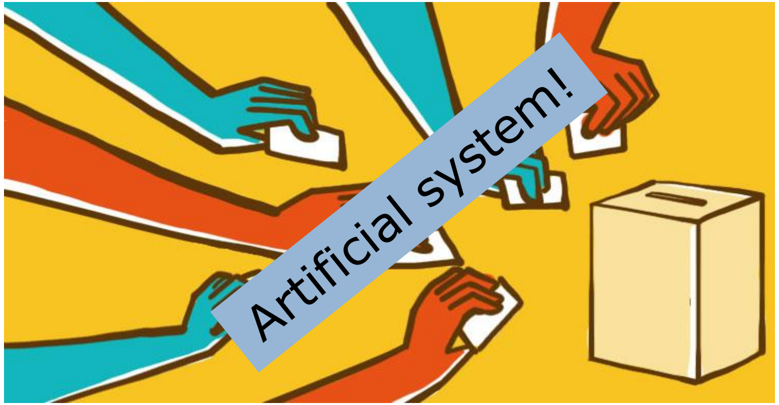
UMEÅ UNIVERSITY

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?

HOW DO WE MAKE DECISIONS TOGETHER?



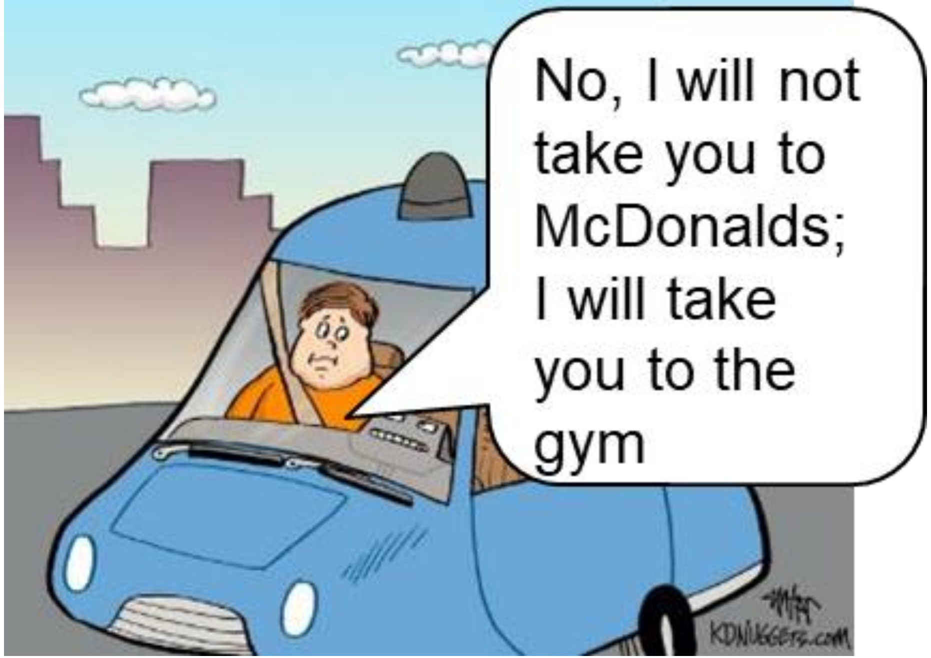
UMEÅ UNIVERSITY

DESIGN IMPACTS DECISIONS IMPACTS SOCIETY

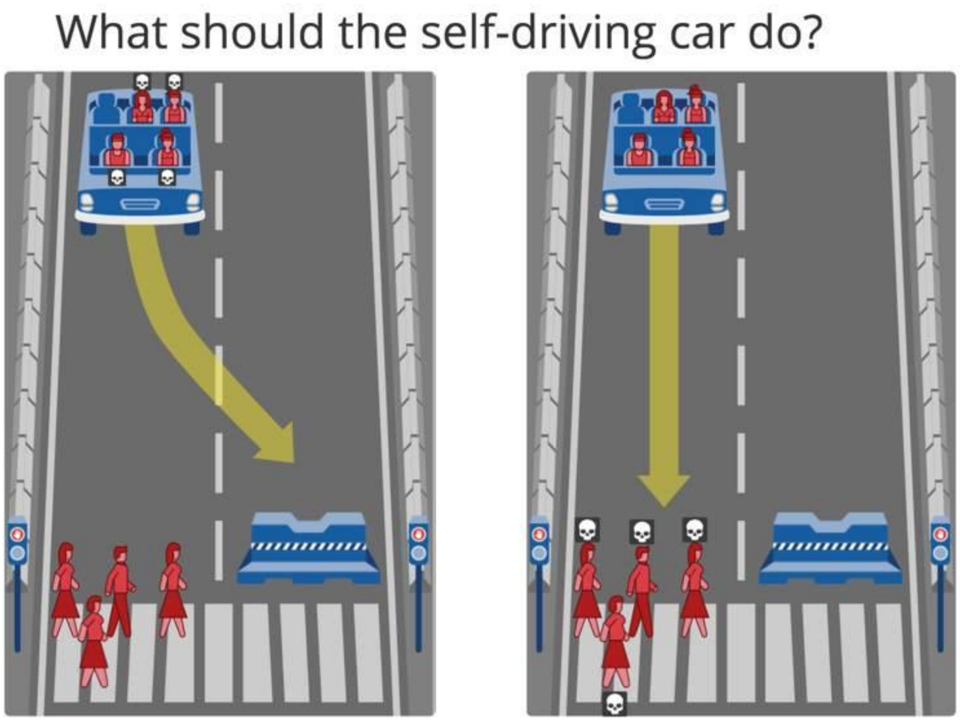
- Choices
- Formulation
- Involvement
- Legitimacy
- Aggregation

UMEÅ UNIVERSITY

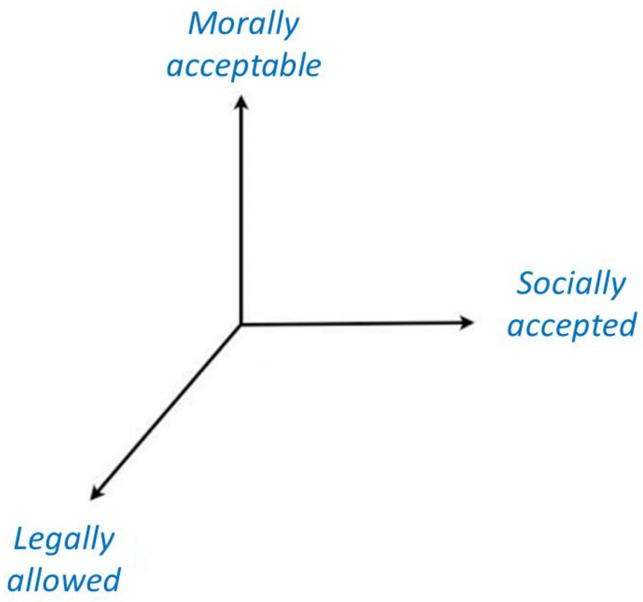
WHICH DECISIONS SHOULD AI MAKE?


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

WHICH DECISIONS SHOULD AI MAKE?



HOW SHOULD AI MAKE DECISIONS?



Morally acceptable

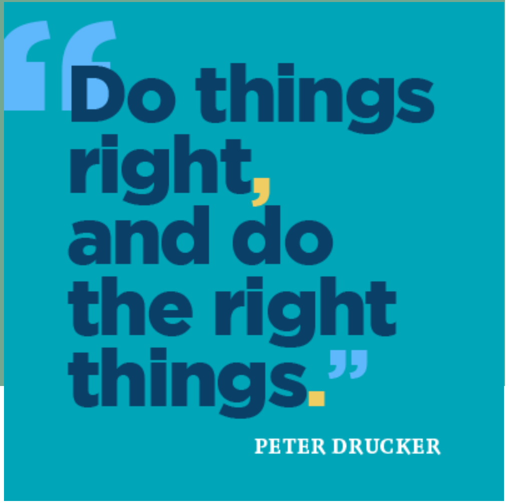
Socially accepted

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: PROCESS**



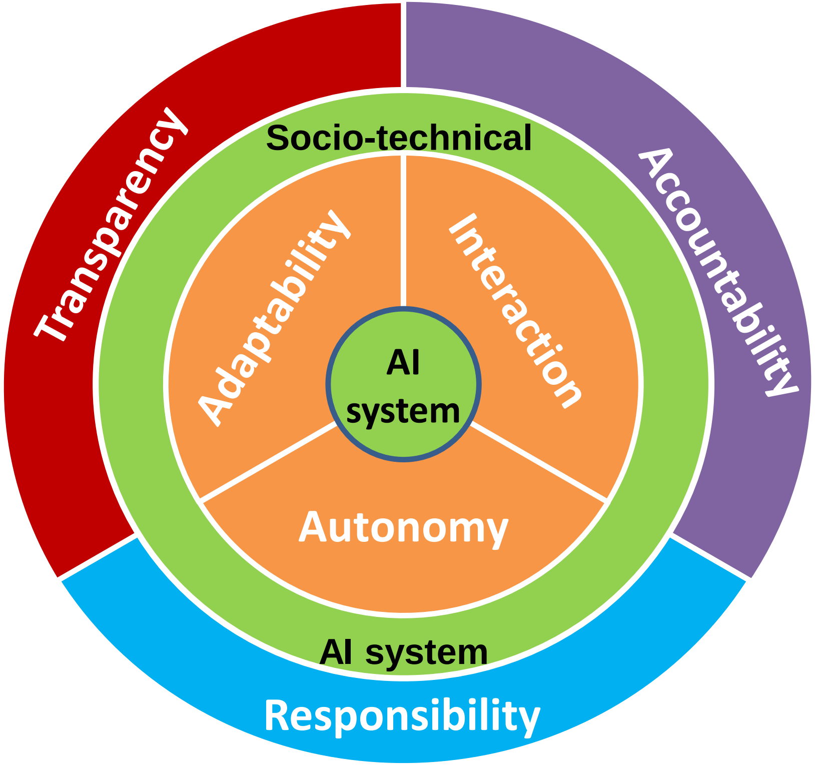
"Do things right, and do the right things."

PETER DRUCKER

UMEÅ UNIVERSITY

TAKING RESPONSIBILITY: ART

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



ETHICS *IN* DESIGN– 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
    - Not just about algorithms

- AI systems (will) take decisions that have ethical grounds and consequences
- Many options, not one ‘right’ choice
- Need for design methods that ensure

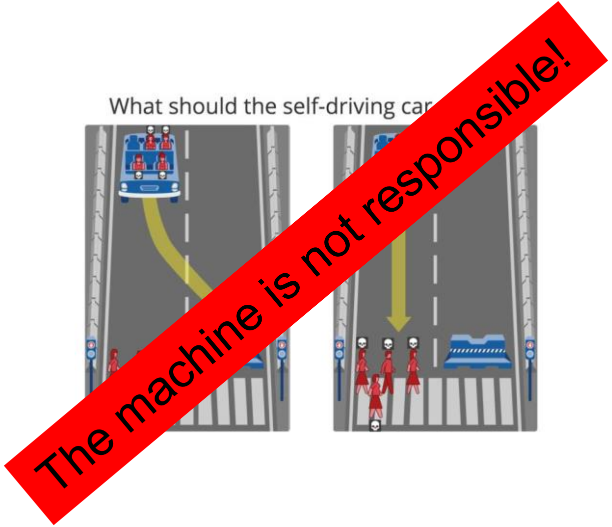
ETHICS *IN* DESIGN: AI – DOING IT RIGHT

- 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

ETHICS *IN* DESIGN: AI – DOING IT RIGHT

- 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!

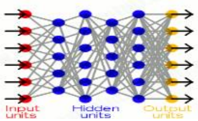
ETHICS ***IN*** DESIGN: AI – DOING IT RIGHT

- 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

UMEÅ UNIVERSITY

CONCERN: EXPLAINABLE AI

AI System



- Machine learning is currently the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

Watson

©IBM

AlphaGo

©Marcin Bajer/Flickr

Sensemaking

Operations

User

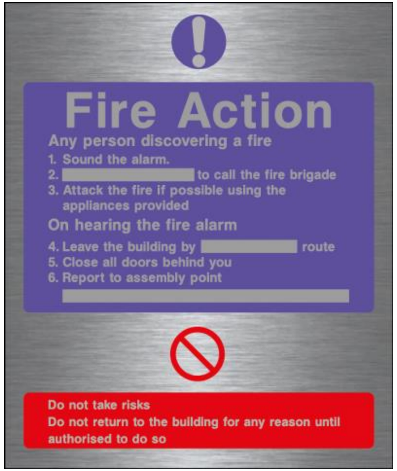
- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

UMEÅ UNIVERSITY

WHAT IS AN EXPLANATION?

Terms and Conditions

Correct
Compreensible
Timely
Complete
Parsimonous



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

UMEÅ UNIVERSITY

NO AI WITHOUT EXPLANATION

- XAI is for the user:
  - Who depends on decisions, recommendations, or actions of the system
  - Just in time, clear, concise, understandable
- XAI is about:
  - provide an explanation of individual decisions
  - 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

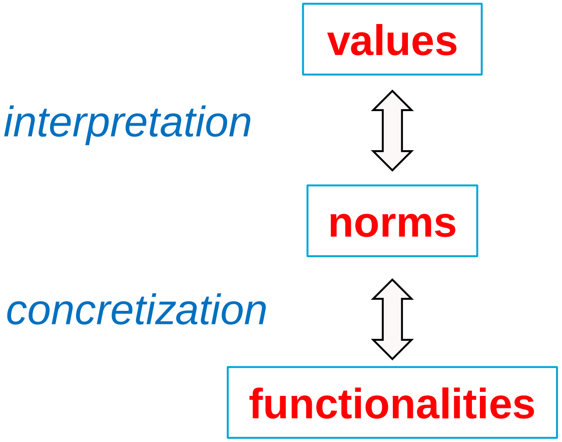
BY **DESIGN: ARTIFICIAL AGENTS**

- Can we teach ethics to AI?
- Should we teach ethics to AI?
  - What does that mean?
  - What is needed?
- Decisions matter
  - Our decisions matter
- Good results matter
  - Accurate/efficient/fair/sustainable?
- The dilemmas
  - 95% accurate but not explainable or 80% accurate but explainable?
  - Energy costs: more AI per joule!

WHICH VALUES – WHOSE VALUES

- Sources
    - Society (Designer, Users, Owner, Manufacturer)
    - Law: legislation, standards
    - Ethics
- 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?

DECISIONS MATTER!



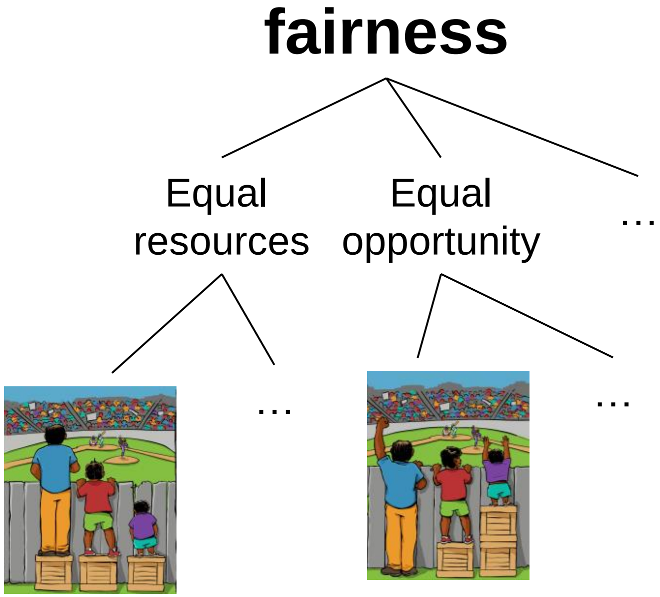
values

interpretation

norms

concretization

functionalities



fairness

Equal resources

Equal opportunity

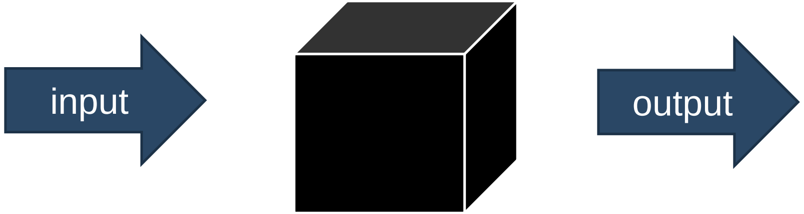
…

…

…

ONE PROBLEM

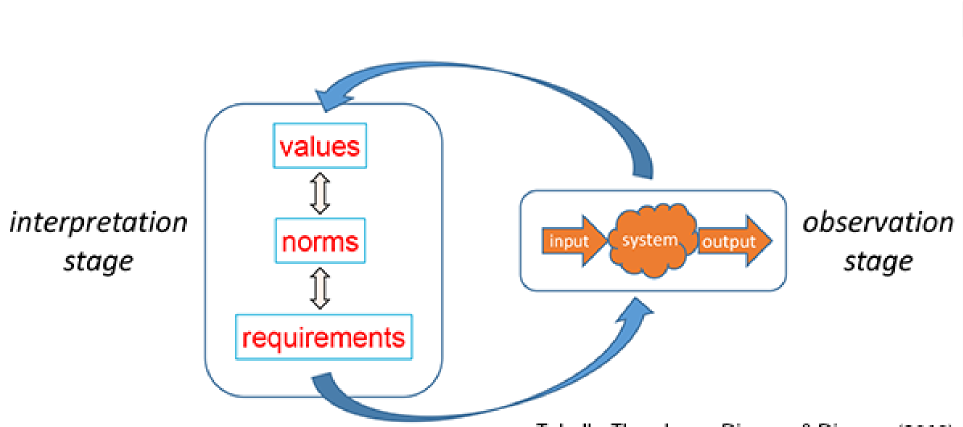
- **black boxes** cannot always be avoided



- Still, we need to **trust** systems. **Check** their **compliance** against our **values**.

GLASS BOX APPROACH

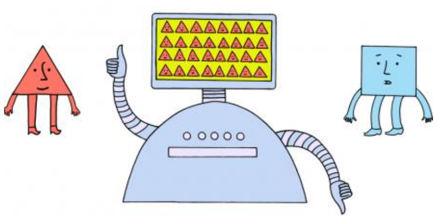
- Doing the right thing
  - Elicit, define, agree, describe, report
- Doing it right
  - Explicit values, principles, interpretations, decisions
  - Evaluate input/output against principles

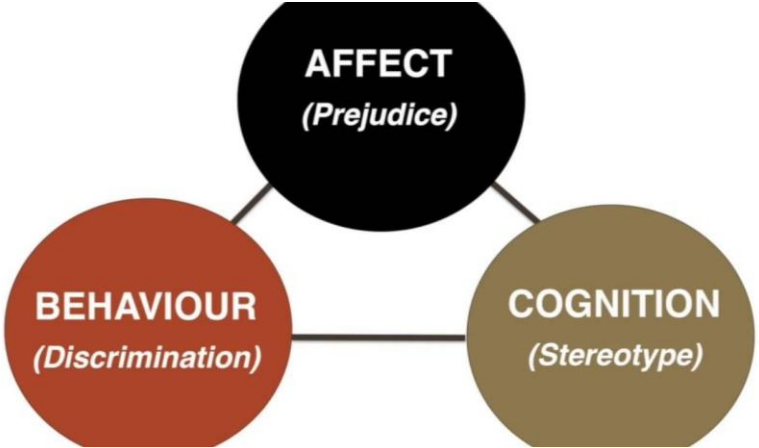


Tubella, Theodorou, Dignum & Dignum (2019)
https://arxiv.org/abs/1905.04994

CONCERN: BIAS AND DISCRIMINATION

- Bias is inherent on human data – we need bias to make sense of world
- But we dont want AI to be prejudiced!
- Unbiasing: Are we creating new bias ?





BIAS IS MORE THAN BIASED DATA

- Who is collecting the data?
- Whose data is being collected?
- Which data is collected?
  - Why don’t we keep information about the colour of the socks of the data collector?
- How and by who is the data labelled?
  - Data farms, exploitation
- What is the training data?
  - The cheapest and easiest to attain?

GUIDELINES TO DEVELOP ALGORITHMS RESPONSIBLY

- 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 algorithms?

<u>FOR</u> DESIGN(ERS): PEOPLE

- **Regulation**
- **Certification**
- **Standards**
- **Conduct**

AI principles are principles for us

REGULATION AND CERTIFICATION

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

- Principles and regulation are drive for transformation
  - Better solutions
  - Return on Investment

Recommendations for trustworthy AI – Main issues

1. Empower and protect humans and society
2. Take up a tailored approach to the AI market
3. Secure a Single European Market for Trustworthy AI
4. Enable AI ecosystems thorough sectoral multi stakeholder alliances
5. Foster the European data economy
6. Exploit the multi-faceted role of the public sector
7. Strengthen and unite Europe's research capabilities
8. Nurture education
9. Adopt a risk-based governance approach to AI and ensure an appropriate regulatory framework
10. Stimulate an open and lucrative investment environment
11. Embrace a holistic way of working

Independent
High-Level Expert Group on
Artificial Intelligence
set up by the European Commission

AI

Policy and Investment Recommendations
for
Trustworthy AI

WHAT IS BEING REGULATED?

- A computational technology for decision making?
- A field of scientific research that studies theories and methods for adaptability, interaction and autonomy of machines?
- An intelligent entity that acts autonomously in (our) environment?

ALTERNATIVES / COMPLEMENTS TO REGULATION

- Standards (IEEE, ISO)
  - soft governance; non mandatory to follow
  - demonstrate due diligence and limit liability
  - user-friendly integration between products
- Advisory panels and ethics officers
  - Set and monitor ethical guidelines
  - able to veto any projects or deliverables that do not adhere to guidelines
- Assessment lists for trustworthy, ethical, AI (EU)
  - responsible AI is more than ticking boxes
  - Means to assess maturity are needed
- Education and training
- Appeal to civic duty / voluntary implementation (Australia)

TRUSTWORTHY AI

CERTIFICATION FOR AI?

CONCERN: WHO IS DEVELOPING AI?

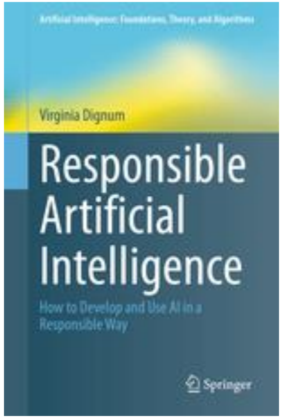
- 18% researchers at conferences are women
- 80% professors are men
- Workforce
  - Google: 2,5% black, 3,6% Latino, 10% women
  - Facebook: 3,8% black, 5% Latino, 15% women

- Design impacts decisions impacts society impacts design

- Regulation requires understanding what and why regulate

- AI systems are tools, artefacts made by people:
  We set the purpose

- AI can give answers, but we ask the questions



Artificial Intelligence: Foundations, Theory, and Algorithms

Virginia Dignum

Responsible Artificial Intelligence

How to Develop and Use AI in a Responsible Way

Springer