RESPONSIBLE ARTIFICIAL INTELLIGENCE

Prof. Dr. Virginia Dignum

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

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

UMEÅ UNIVERSITY

DECISION-MAKING AND ETHICS

"All my decisions are well thought out."

AI AND ETHICS - SOME CASES

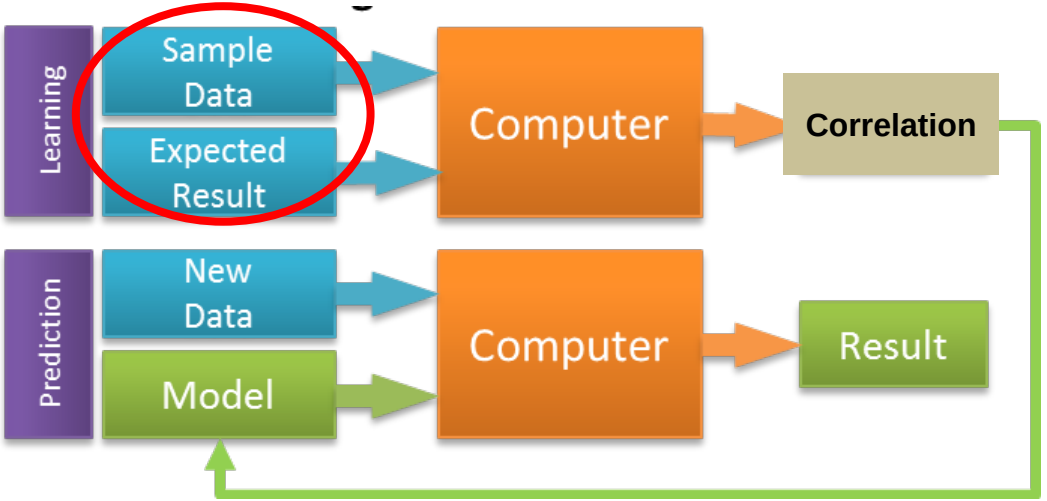
- Self-driving cars
  - Who is responsible for the accident by self-driving car?
  - How can car decide in face of a moral dilemma?
- Automated manufacturing
  - How can technical advances combined with education programs (human resource development) help workers practice new sophisticated skills so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

WHAT WE TALK ABOUT WHEN WE TALK ABOUT AI

- Autonomy
- Decision-making
- Algorithms
- Robots
- Data
- Learning
- End of the world!?
- A better world for all?

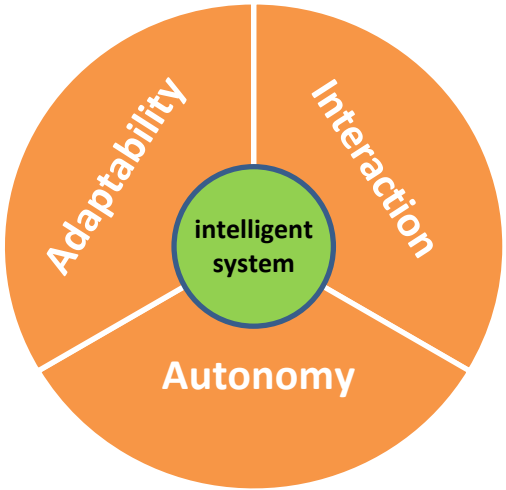
WHAT IS AI? – NOT JUST THE ALGORITHM

WHAT IS AI? – NOT JUST MACHINE LEARNING



- Not all problems can be solved by correlation
  - Causality, abstraction…
- Rubbish in – rubbish out
- Brittleness

ARTIFICIAL INTELLIGENCE



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

How can we develop AI to benefit humanity?

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

ETHICS *IN* DESIGN

- Design for values
- Design for all
- Doing the right thing
- Doing it right

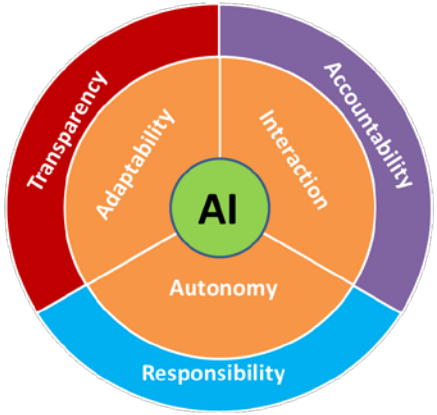
ETHICS *IN* DESIGN - DOING THE RIGHT THING

- 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– 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

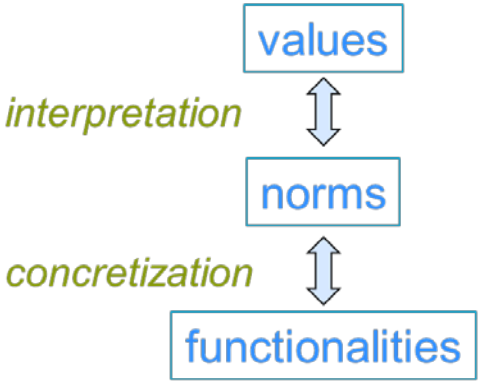


- AI systems (will) take decisions that have ethical grounds and consequences
- Many options, not one 'right' choice
- Need for design methods that ensure

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

ART METHODOLOGY

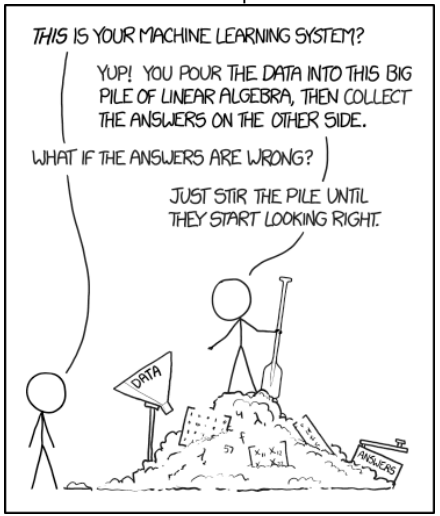
- Socially accepted
  - Participatory
- Ethically acceptable
  - Ethical theories and human values
- Legally allowed
  - Laws and regulations



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

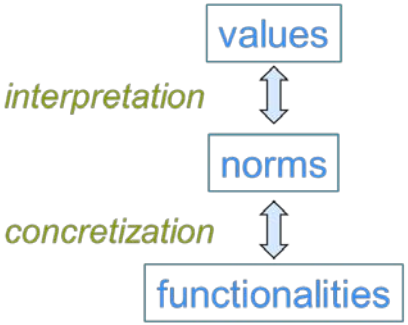
ACCOUNTABILITY CHALLENGES

https://xkcd.com/1838/



- Explanation
  - Optimal AI is explainable AI
  - Optimal is not that which optimizes the result ignoring the context but the one that gives the best result for the context

- 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

- Data
  - Bias is inherent in human behavior
  - Provenance: Where does it come from? Who is involved?
  - Training data: the cheapest/easiest or the best?
  - Data responsibility
  - Governance, storage, updated

  - Environment: energy costs
  - Market mechanisms

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*

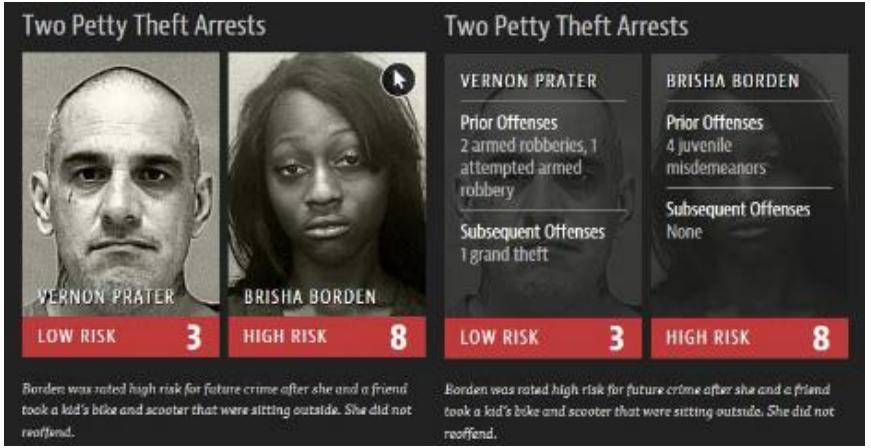
- **Stereotype**
  - those bias that we don't want to have persisting
  - *Most programmers are male*

- **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 ?

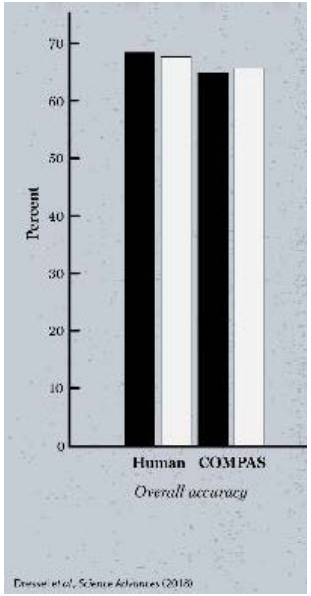
TRANSPARENCY - DEALING WITH BIAS

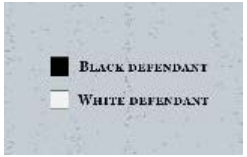
- COMPAS – recidivism risk identification



DEALING WITH BIAS

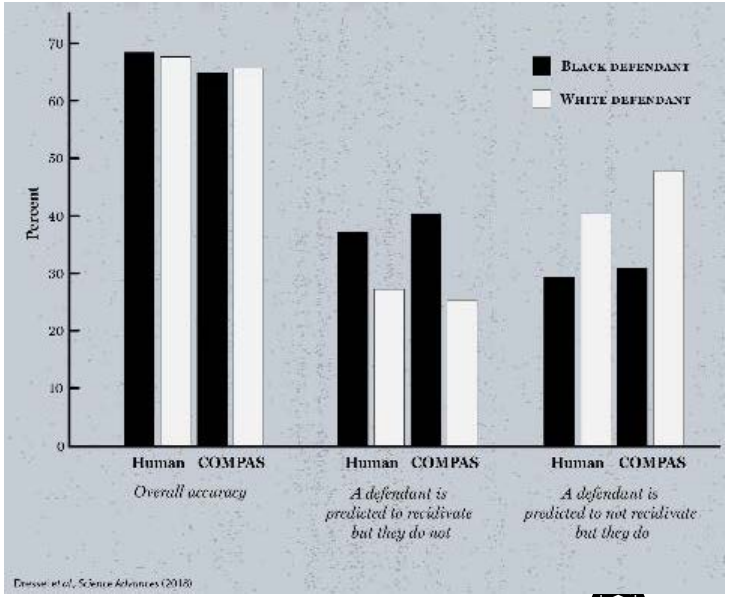
- COMPAS – recidivism risk identification
- Is the algorithm fair to all groups?

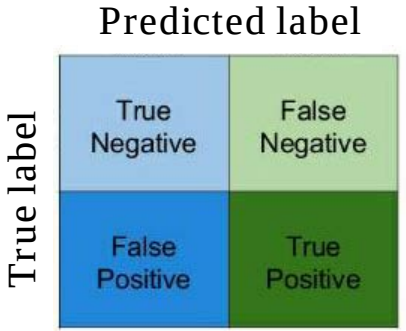




DEALING WITH BIAS

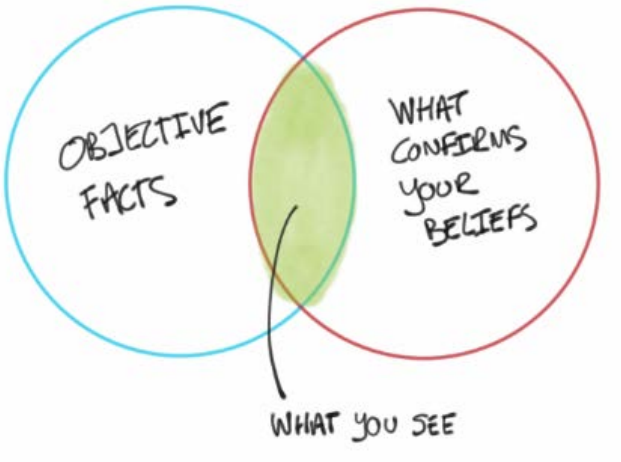
- COMPAS – recidivism risk identification
- Is the algorithm fair to all groups?





CONFIRMATION BIAS

- tendency to search for, interpret, favour, and recall information in a way that confirms one's pre-existing beliefs or hypotheses.



Predictive policing

ITY

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 algorithm?
  - Does training data resemble the context of use?

UMEÅ UNIVERSITY

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

ART IS ABOUT BEING EXPLICIT

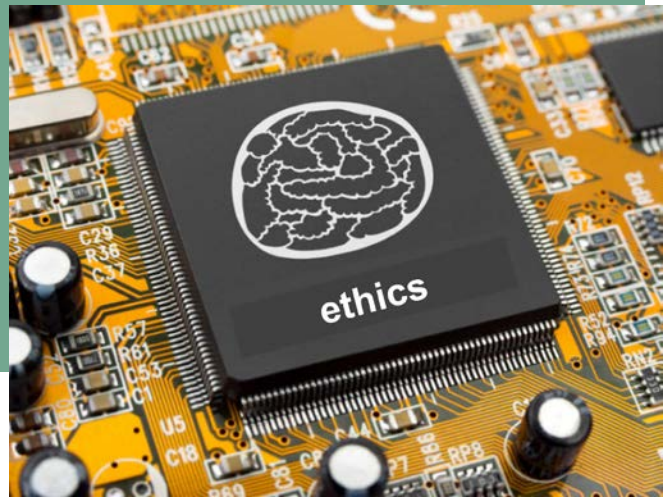
- Question your options and choices
- Motivate your choices
- Document your choices and options
- Regulation
  - External monitoring and control
  - Norms and institutions
- Engineering principles for policy
  - Analyze – synthetize – evaluate - repeat

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

ETHICS BY DESIGN – ETHICAL ARTIFICIAL AGENTS

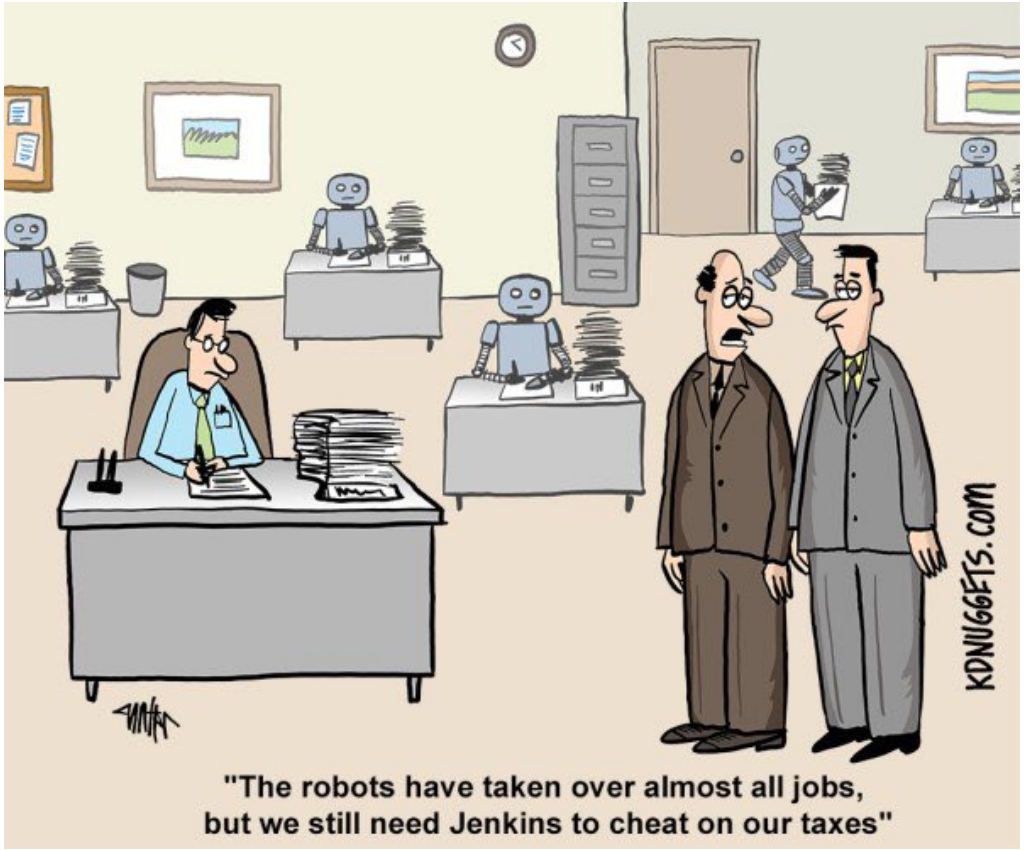
- **Can AI artefacts be build to be ethical?**
  - What does that mean?
  - What is needed?

- **Understanding ethics**
- **Using ethics**
- **Being ethical**

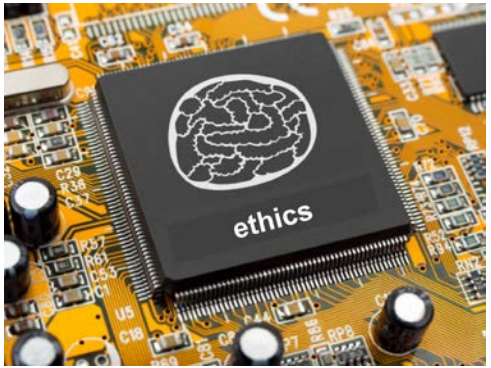


UMEÅ UNIVERSITY

NOT SO BRAVE NEW WORLD?



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?

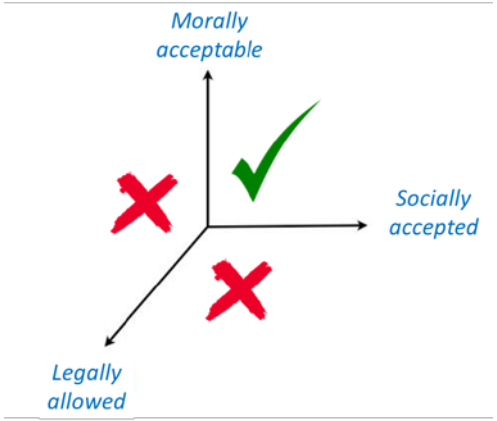
3. **How to implement?**
   - Role of user
   - Role of society
   - Role of AI system

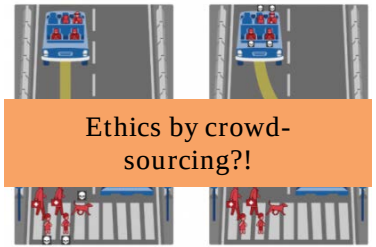
PARTICIPATION – AI VALUES

- 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?

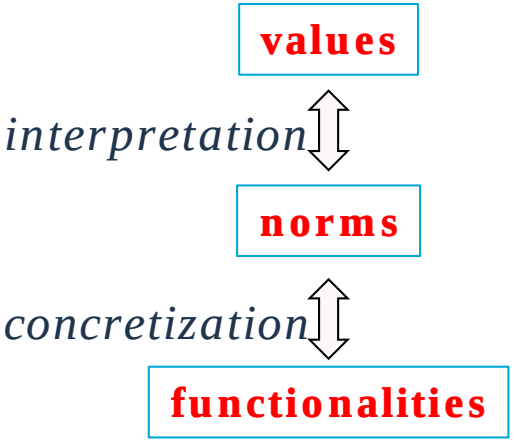


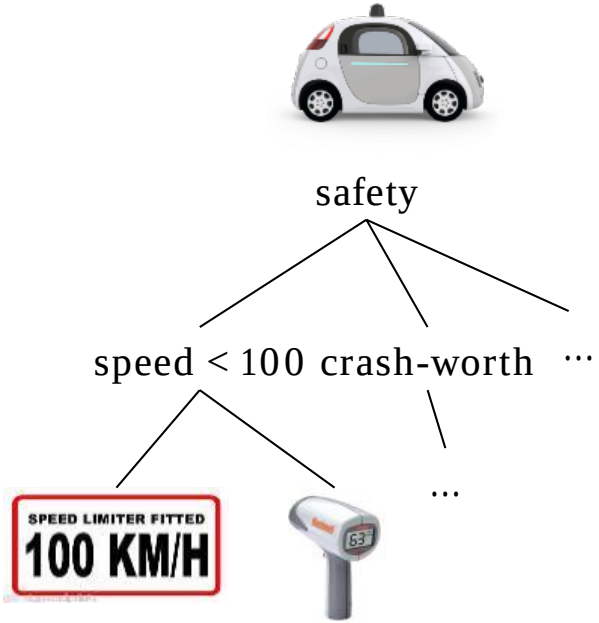


SOCIAL ACCEPTANCE – DEMOCRACY

- 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
- Counting
  - Simple majority? Districts? Rankings?

IMPLEMENTATION: FROM VALUES TO FUNCTIONALITIES





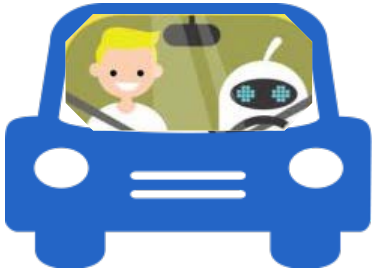
ETHICAL REASONING? - AN EXAMPLE

- Design a self-driving can that makes ethical decisions
- Value: “human life”
- Implementation?
- 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?)**

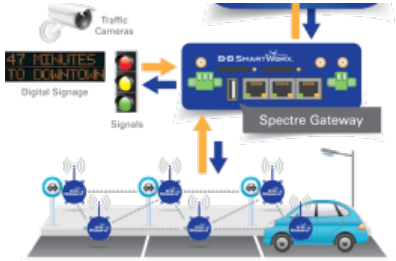
Ethical theories

- 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.

IMPLEMENTATION CHOICES



collaboration

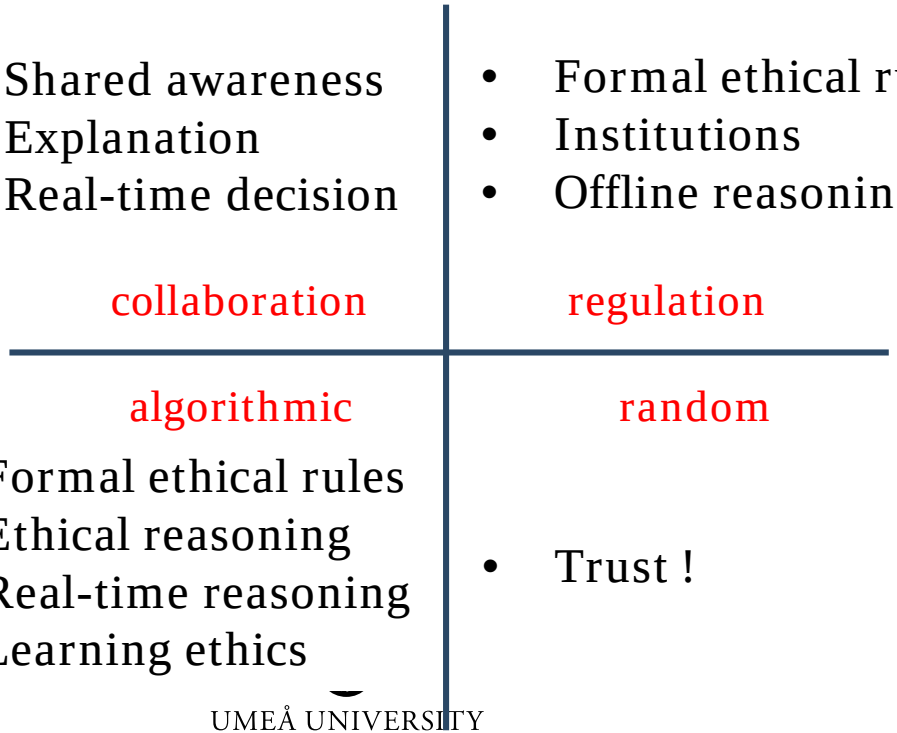


regulation

algorithmic

random

COMPUTATIONAL REQUIREMENTS



(TC King et al, AAMAS 2015)

IMPLEMENTATION ISSUES - DILEMMAS

- Moral dilemma
  - You cannot have all
  - There is not one right solution!

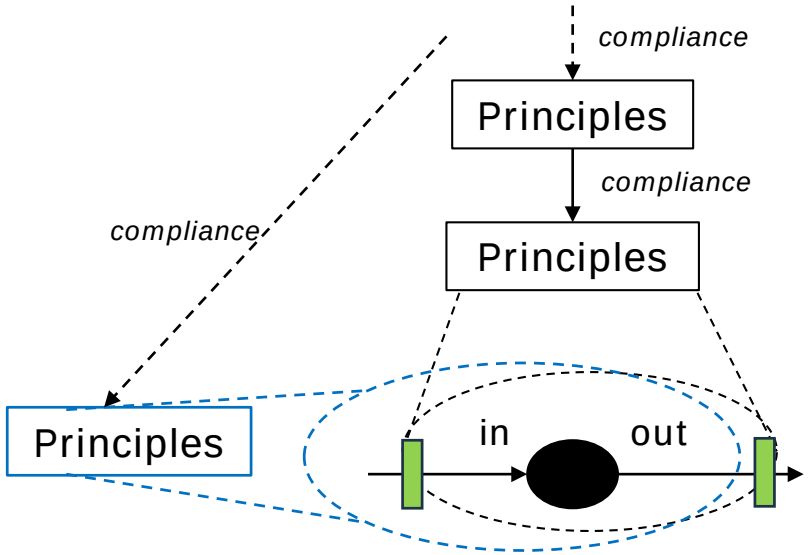
- So the issue should not be what is the answer the AI system will give to a moral dilemma

- The issue is one of responsibility and openness
  - Make assumptions clear
  - Make options clear
  - Open data
  - Inspection
  - Explanation
  - ...

→ **ART**

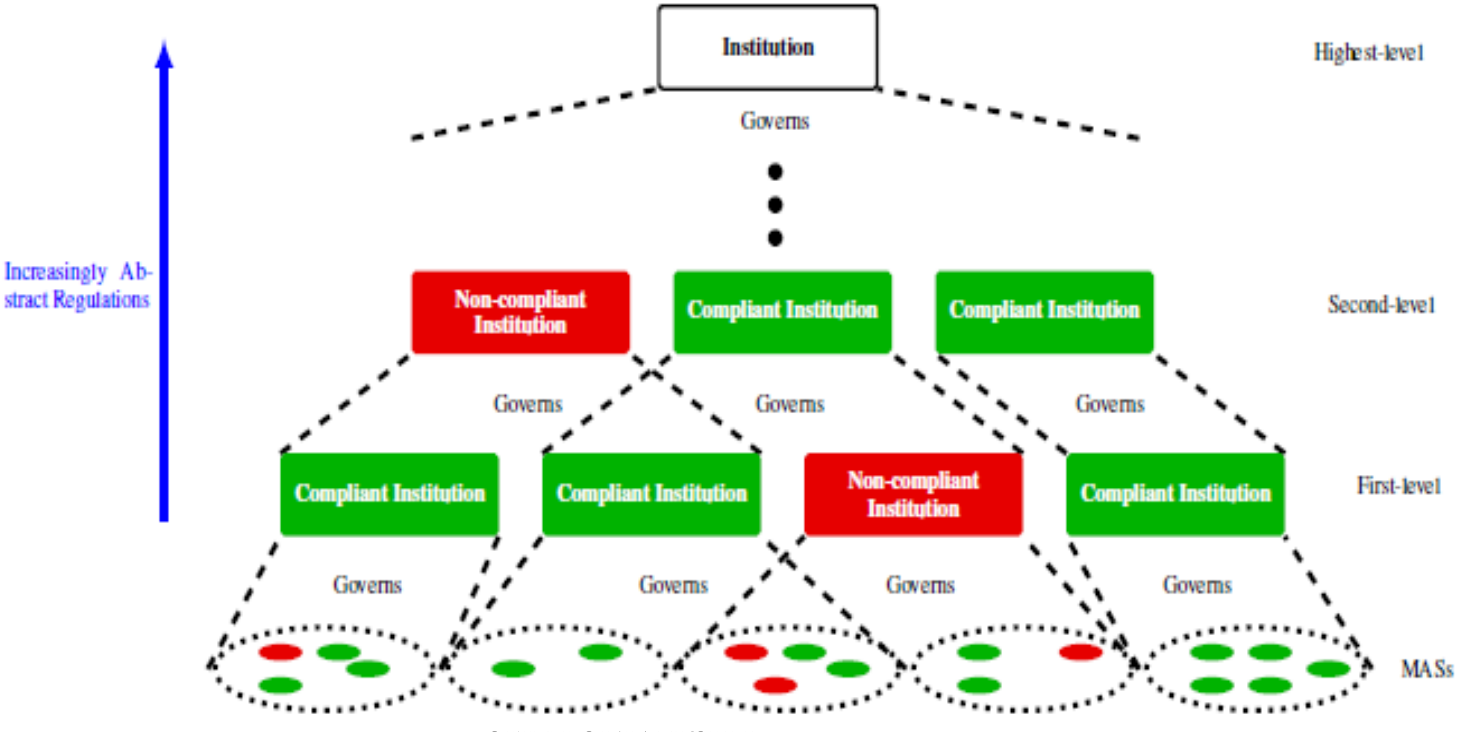
GOVERN AND VERIFY - GLASS BOXES

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



extending (TC King et al, AAMAS 2015)

TRANSPARENCY – GLASS DOMES



EXPLANATION


Terms and Conditions

UMEÅ UNIVERSITY

Email: virginia@cs.umu.se, twitter: @vdignum

APPROACH

BDI program
Lowest-level plans annotated with value change info.

Annotated

Constraint Optimization

Transformed BDI program

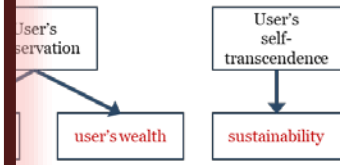
(E1) “This is the best possible coffee available; I had no money.” (V+B)

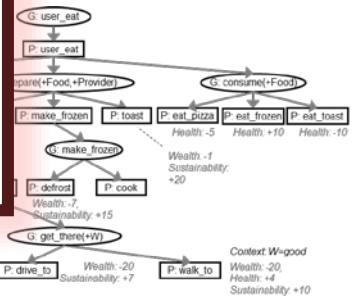
(E2) “This is the best possible coffee available.” (AV)

(E3) “This coffee is better than the kitchen and cheaper than in the shop.” (V)

(E4) “I’ve no money; Ann was in her room.” (B)

(E5) “I wanted coffee; Ann was in her room.” (G+B)






```
+!vbr(G) : constraints
  .callSolver(C, VT, V
  !G[choices(Choices)]

+!usr_eat[choices(C)] : (not .ground(C)) | C = [1, choices(C1,C2,C3)] <-
  !choose(Food, Provider)[choices(C1)];
  !prepare(Food, Provider)[choices(C2)];
  !consume(Food)[choices(C3)].
```

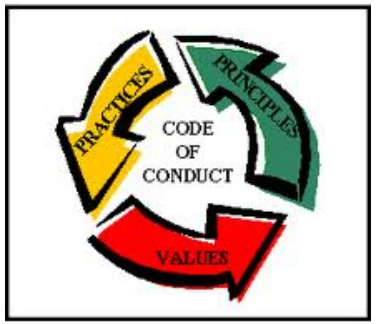
Only choose this plan if path[illegible]s with 1

ETHICS *FOR* DESIGN(ERS)

- Regulation
- Certification
- Standards
- Conduct

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

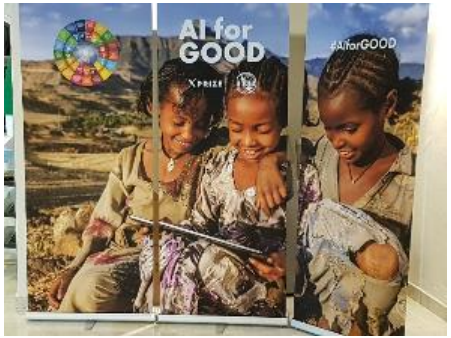
- Standards
  - System design
  - Dealing with transparency
  - Dealing with privacy
  - Dealing with algorithmic bias
  - Data protection
  - Robotics
  - ...

- 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

AI AND EUROPE

- High level expert group on AI
  - Ethical guidelines
  - Policy and investment strategy

- AI Alliance: https://ec.europa.eu/digital-single-market/en/european-ai-alliance
  - forum engaged in a broad and open discussion of all aspects of Artificial Intelligence development and its impacts.

- AI4People Global Forum: http://www.eismd.eu/ai4people/about/

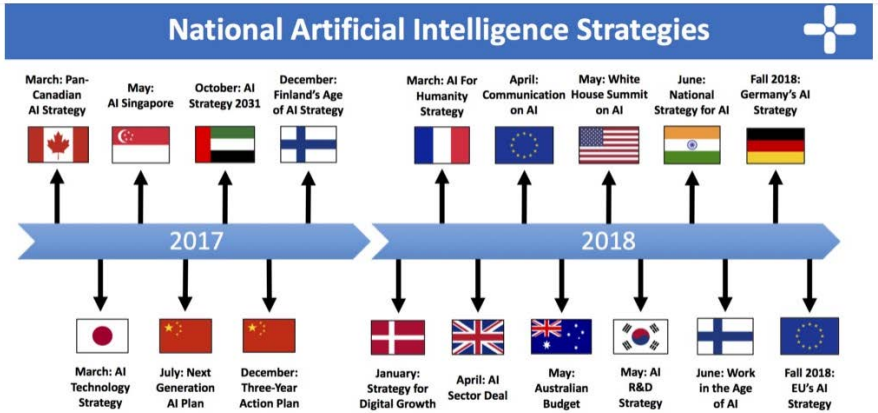
- CLAIRE (ELLIS): https://claire-ai.org/
  - Confederation of Laboratories for Artificial Intelligence Research in Europe
  - Research hubs (“CERN for AI”)

- H2020
  - EU “AI-on-Demand Platform”
  - EU Flagship proposal “Humane AI” ...

ETHICAL GUIDELINES

- IEEE Ethically Aligned Design
- EGE Statement on AI
- Asilomar principles
- EESC report on AI
- .....

Human dignity
Privacy
Security
Responsibility
Justice
Solidarity
Democracy
.....



Tim Dutton: https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd

Rumman Chowdhury's list: https://goo.gl/ca9YQV

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