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

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?
- …..

AI AND ETHICS - SOME CASES

- Self-driving cars
  - Who is responsible for the accident by self-driving car?
  - (How) Can a 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 ABOUT OUR OWN ETHICS?

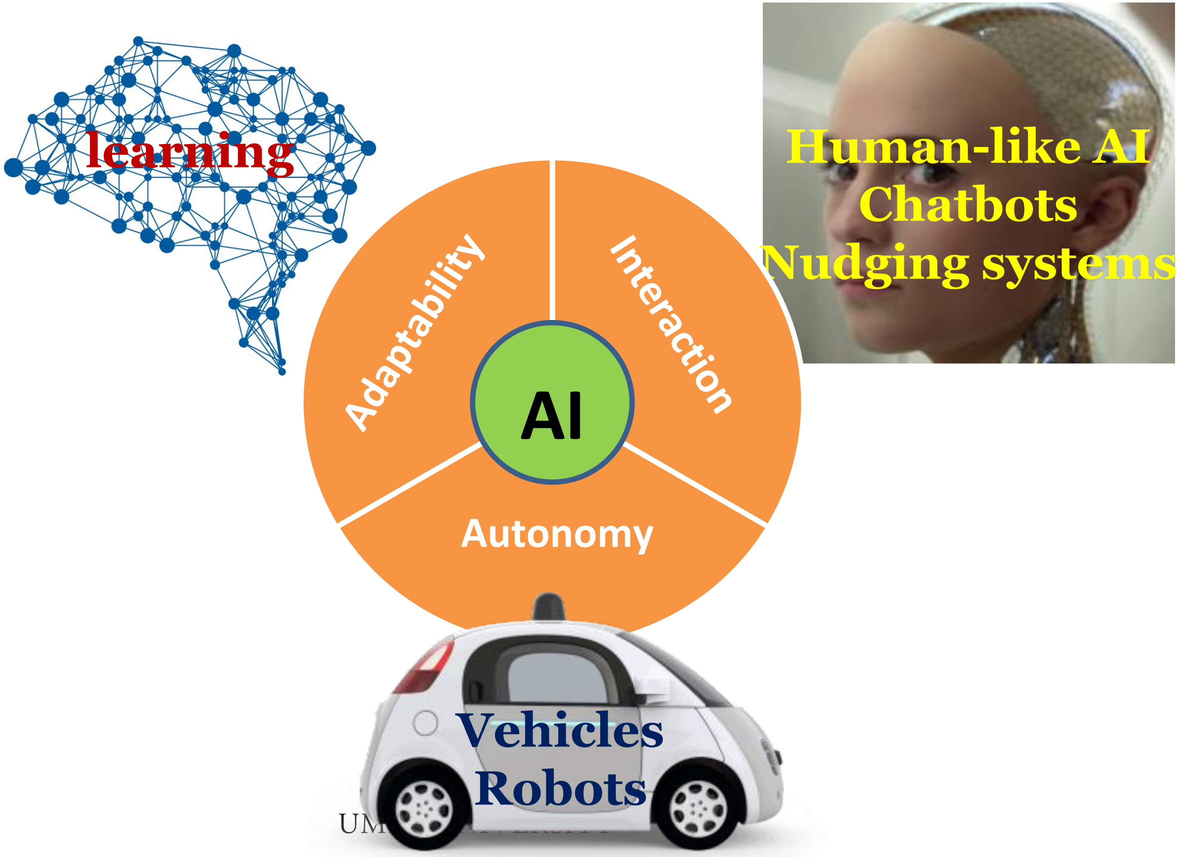
“All my decisions are well thought out.”

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

WHAT IS AI?

- Not just the algorithm
  - Algorithm is the recipe
  - Result is dependent on more
- Not just machine learning / deep learning
  - Current successes are in perception / pattern recognition
  - (Human) intelligence is more
- Not just data
  - Big data is big headache: governance, sustainability
  - Responsible AI demands more

ARTIFICIAL INTELLIGENCE



learning

Adaptability

Interaction

AI

Autonomy

Human-like AI
Chatbots
Nudging systems

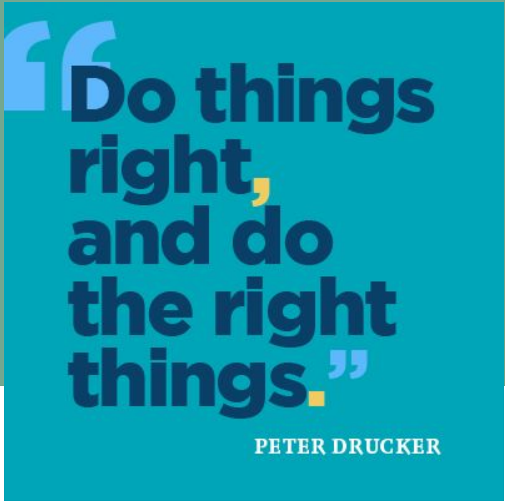
Vehicles
Robots

TAKING RESPONSIBILITY

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

ETHICS *<u>IN</u>* DESIGN

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



“Do things right, and do the right things.”

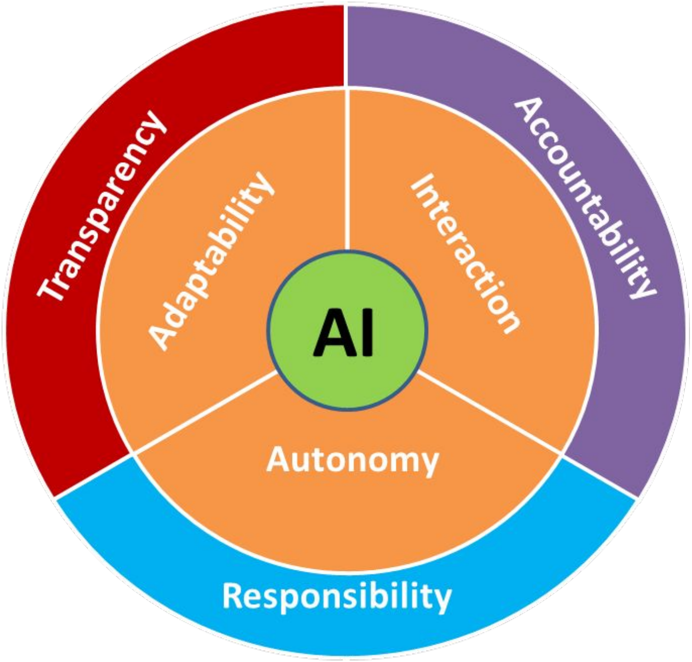
PETER DRUCKER

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

RESPONSIBLE ARTIFICIAL INTELLIGENCE



UMEÅ UNIVERSITY

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-ar
d-the-quest-for-transparency-in-artificial-intelligence-bcde64f5
9f5b

ETHICS ***IN*** DESIGN - DOING THE RIGHT THING

- 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

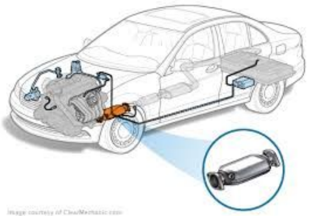
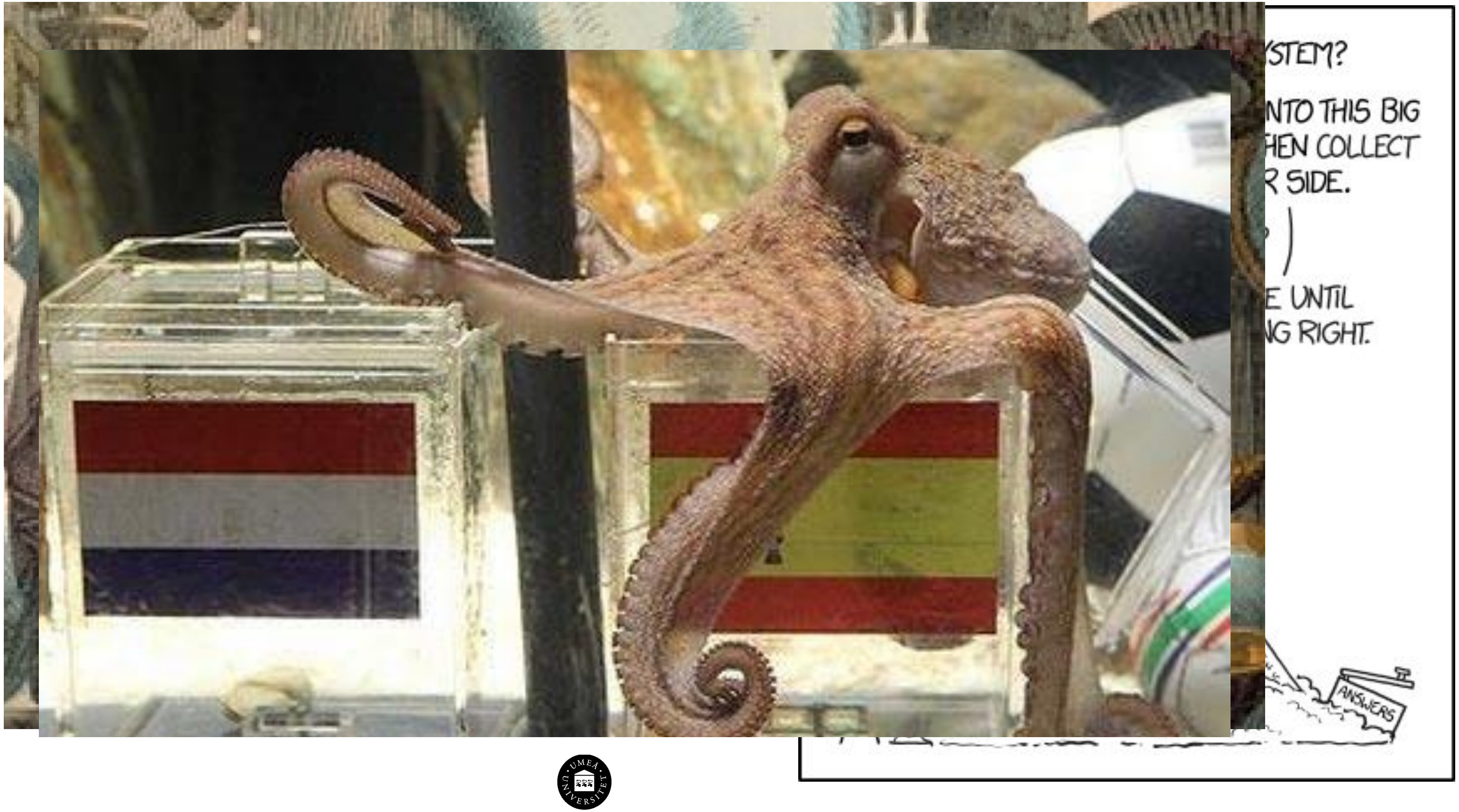


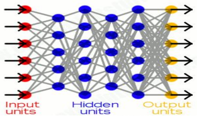
Image courtesy of ClearMechanic.com

DESIGN CHALLENGES



WHY 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

©Eric Keenan, U.S. Marine Corps

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?

WHAT IS AN EXPLANATION?

Terms and Conditions

[illegible]

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

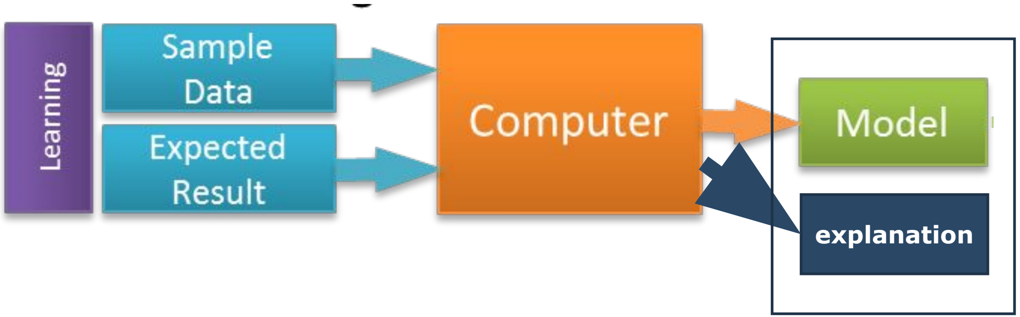
4. Leave the building by route
5. Close all doors behind you
6. Report to assembly point

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

UMEÅ UNIVERSITY

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

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

DESIGN FOR ALL

- Inclusion
- Diversity
- Dialogue

Optimal AI
=
AI for Good
=
AI for All
=
AI by All

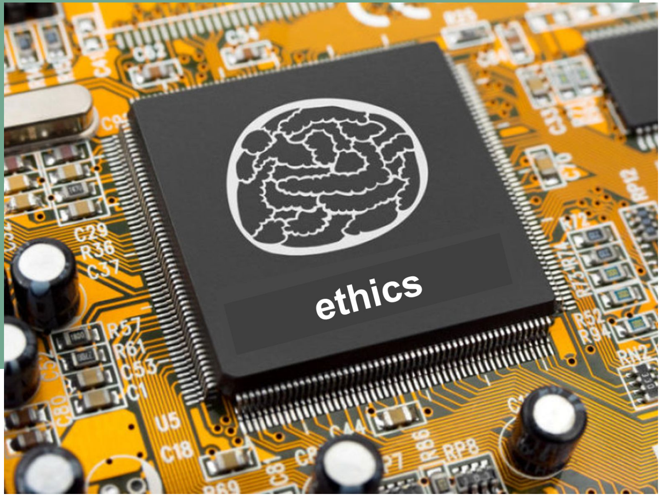
- Concerns
  - Safety
  - Replacement
  - Awareness
  - Privacy
  - Bias
  - Human dignity

Danger is not AI taking over the world,
but misuse and failures

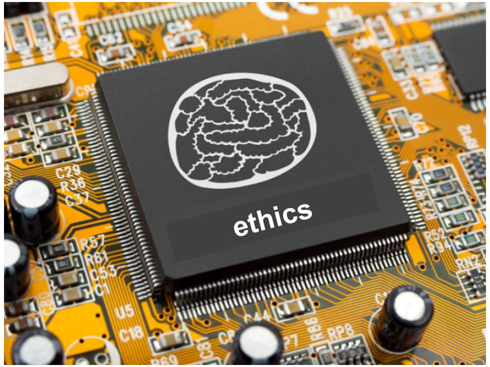
ETHICS BY DESIGN –
ETHICAL ARTIFICIAL AGENTS

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

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

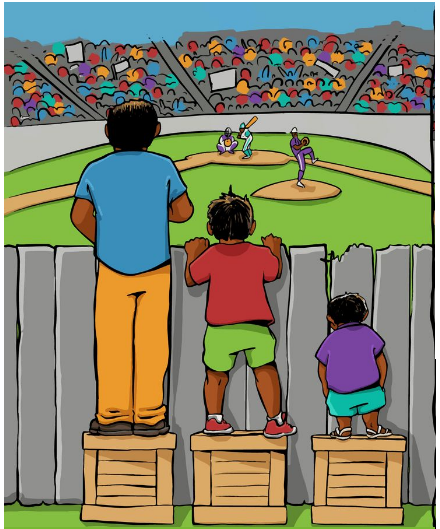


ETHICS BY DESIGN

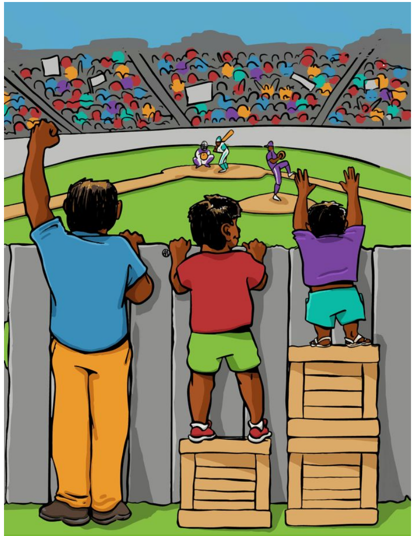


1. **Value alignment**
   - Identify *relevant* human values
   - Are there universal human values?
   - Who gets a say? Why these?
2. **How to behave?**
   - Ethical theories: How to behave according to these values?
   - How to prioritize those values?
3. **How to implement?**
   - Role of user
   - Role of society
   - Role of AI system

VALUES AND CONTEXT

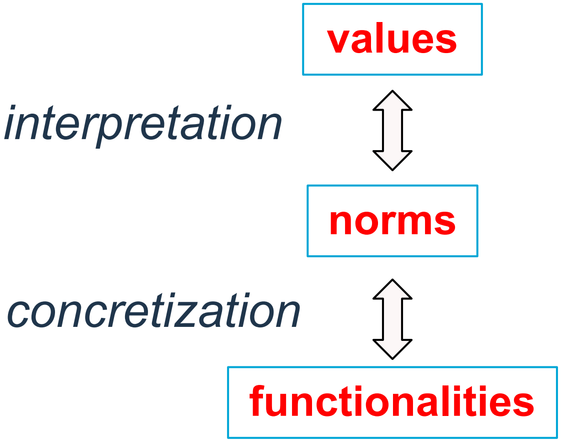


Fairness?



Fairness?

DECISIONS MATTER!



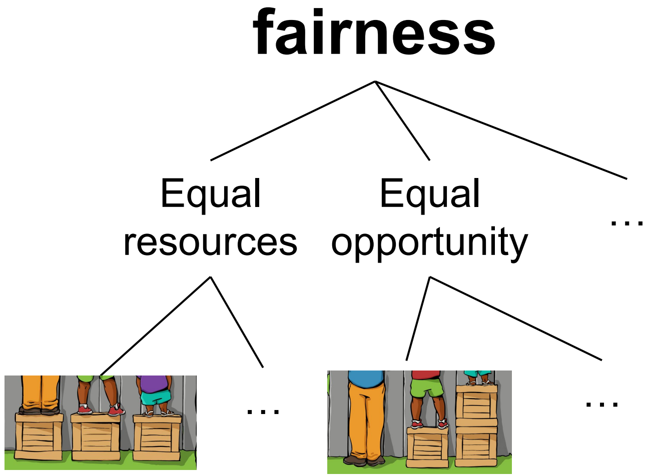
values

interpretation

norms

concretization

functionalities



fairness

Equal resources

Equal opportunity

…

…

…

ETHICAL REASONING?
- AN EXAMPLE

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.

- 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?)**

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

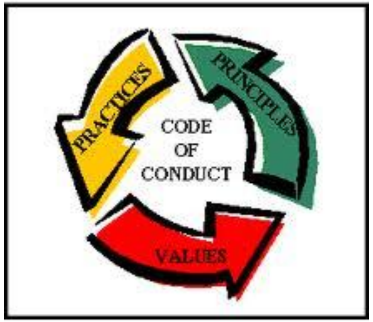
No, I will not take you to McDonalds; I will take you to the gym
KDNUGGETS.COM

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!

EU HIGH LEVEL EXPERT GROUP ON AI

- Ethical Guidelines
  - Guiding principles
    - Respecting Fundamental Rights, Principles and Values - Ethical Purpose
    - Critical concerns
  - Implementation
    - Realising trustworthy AI
    - Assessing Trustworthy AI

- Investment and policy strategy
  - Using AI to build an impact in Europe
    - Transforming Europe's Business landscape
    - Catalyzing Europe's Public Sector
    - Attaining World-Class Research Capabilities
    - Accomplishing Citizen's Benefits and Engagement
  - Leveraging Europe's enablers of AI
    - Attracting Funding and Investments in AI
    - Enabling AI with Data and Physical Infrastructure
    - Generating appropriate Skills and Education for AI
    - Ensuring an appropriate policy and regulatory framework

The European Commission's

High-Level Expert Group on Artificial Intelligence

AI

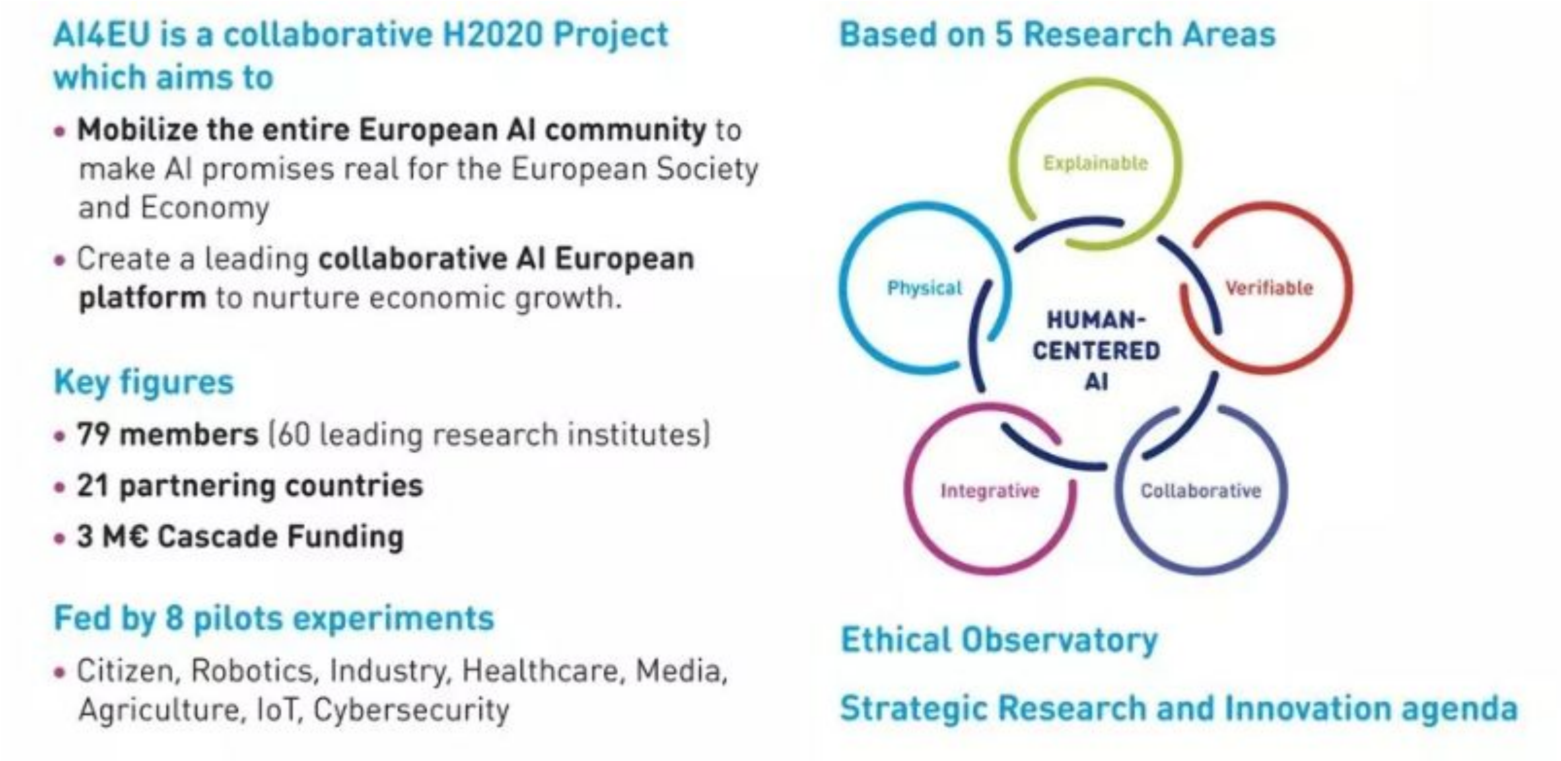
DRAFT

Ethics Guidelines for Trustworthy AI

Working Document for stakeholders' consultation

Brussels, 18 December 2018

AI4EU



AI4EU is a collaborative H2020 Project which aims to

- **Mobilize the entire European AI community** to make AI promises real for the European Society and Economy
- Create a leading **collaborative AI European platform** to nurture economic growth.

Key figures

- **79 members** (60 leading research institutes)
- **21 partnering countries**
- **3 M€ Cascade Funding**

Fed by 8 pilots experiments

- Citizen, Robotics, Industry, Healthcare, Media, Agriculture, IoT, Cybersecurity

Based on 5 Research Areas

Ethical Observatory

Strategic Research and Innovation agenda

IEEE

Global initiative for ethically aligned design of autonomous and intelligent systems

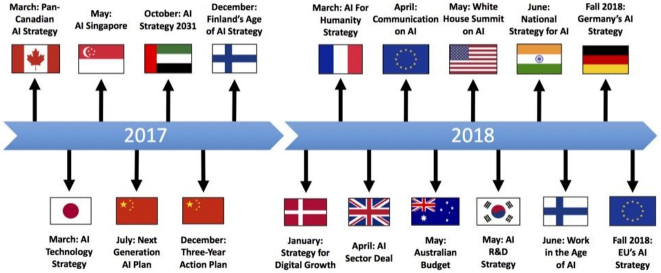
- since 2015
- identify and find broad consensus on pressing ethical and social issues and define recommendations regarding development and implementations of these technologies
- Standards
  - System design
  - Dealing with transparency
  - Dealing with privacy
  - Dealing with algorithmic bias
  - Data protection
  - Robotics
  - …
- Auditing
  - Certified agency

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/

MANY MORE (AND COUNTING…)

- Initiatives
  - CLAIRE (and ELLIS): https://claire-ai.org/
    - Confederation of Laboratories for Artificial Intelligence Research in Europe
  - AI4EU: on demand platform
  - ALLAI (NL)
- Strategies / positions
  - Council of Europe
  - OECD
  - National strategies: cf. 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

TAKE AWAY MESSAGE

- AI influences and is influenced by our social systems
- Design in never value-neutral
- Openness and explicitness are key!
  - Accountability, Responsibility, Transparency
- Optimal AI is explainable AI
- Optimal AI is AI for all
- AI systems are artefacts built by us for our own purposes
- We set the limits

RESPONSIBLE ARTIFICIAL INTELLIGENCE

WE ARE RESPONSIBLE

Email: virginia@cs.umu.se

Twitter: @vdignum

UMEÅ UNIVERSITY