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
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
  o Who is responsible for the accident by self-driving car?
  o (How) Can a car decide in face of a moral dilemma?

• Automated manufacturing
  o 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
  o Mistaken identity (is it a person or a bot?)
  o 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.”
WHAT IS AI?

• Not just the algorithm
  o Algorithm is the recipe
  o Result is dependent on more

• Not just machine learning / deep learning
  o Current successes are in perception / pattern recognition
  o (Human) intelligence is more

• Not just data
  o Big data is big headache: governance, sustainability
  o Responsible AI demands more
ARTIFICIAL INTELLIGENCE

- Human-like AI
- Chatbots
- Nudging systems
- Learning
- Adaptability
- Interaction
- Autonomy
- Vehicles
- Robots
TAKING RESPONSIBILITY

• **in** Design
  o 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
  o Integration of ethical reasoning abilities as part of the **behaviour** of artificial autonomous systems

• **for** Design(ers)
  o Research integrity of **researchers** and manufacturers, and certification mechanisms
ETHICS IN DESIGN

• Doing the right thing
• Doing it right
• Design for values
• Design for all
ETHICS IN DESIGN—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
    ▪ 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

- Autonomy
- Interaction
- Adaptability
- Accountability
- Transparency
- Responsibility
ART IS ABOUT BEING EXPLICIT

• Question your options and choices

• Motivate your choices

• Document your choices and options

• Regulation
  o External monitoring and control
  o Norms and institutions

• Engineering principles for policy
  o Analyze – synthetize – evaluate - repeat

ETHICS IN DESIGN - DOING THE RIGHT THING

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

• Principles and regulation are drive for transformation
  o Better solutions
  o Return on Investment
DESIGN CHALLENGES

• 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
WHY EXPLAINABLE AI

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

Watson

AlphaGo

Sensemaking

Operations

• 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?

Correct
Compreensible
Timely
Complete
Parsimonous
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
Inclusion • Diversity • Dialogue

DESIGN FOR ALL

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
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

- Design a self-driving car 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.
RESPONSIBILITY CHALLENGES

• Chain of responsibility
  o researchers, developerers, manufacturers, users, owners, governments, ...

• Levels of autonomy
  o Operational autonomy: Actions / plans
  o Decisional autonomy: Goals / motives
  o Attainable autonomy: dependent on context and task complexity

No, I will not take you to McDonalds; I will take you to the gym
ETHICS FOR DESIGN(ERS)

- Regulation
- Certification
- Standards
- 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
  o Guiding principles
    ▪ Respecting Fundamental Rights, Principles and Values - Ethical Purpose
    ▪ Critical concerns
  o Implementation
    ▪ Realising trustworthy AI
    ▪ Assessing Trustworthy AI

• Investment and policy strategy
  o 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
  o 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
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

- Explainable
- Verifiable
- Integrative
- Collaborative
- HUMAN-CENTERED AI

Ethical Observatory

Strategic Research and Innovation agenda
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
  o System design
  o Dealing with transparency
  o Dealing with privacy
  o Dealing with algorithmic bias
  o Data protection
  o Robotics
  o ...
• Auditing
  o Certified agency

https://ethicsinaction.ieee.org/
MANY MORE (AND COUNTING...)

• Initiatives
  o CLAIRE (and ELLIS): https://claire-ai.org/
    ▪ Confederation of Laboratories for Artificial Intelligence Research in Europe
  o AI4EU: on demand platform
  o ALLAI (NL)

• Strategies / positions
  o Council of Europe
  o OECD
    o ...

• Declarations
  o Asilomar
  o Montreal
  o ...

UMEÅ UNIVERSITY
TAKE AWAY MESSAGE

• AI influences and is influenced by our social systems
• Design in never value-neutral
• Openness and explicitness are key!
  o 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