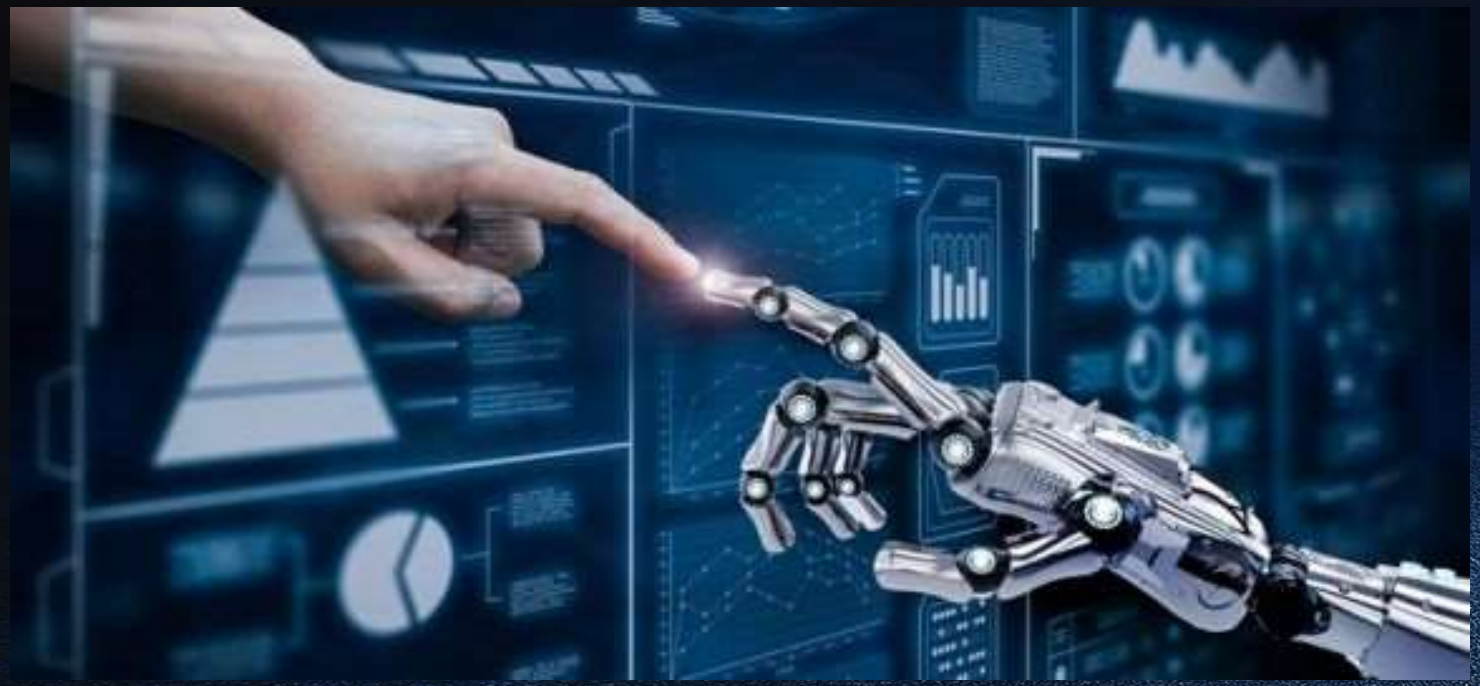
Ethics guidelines for AI and their operationalisation

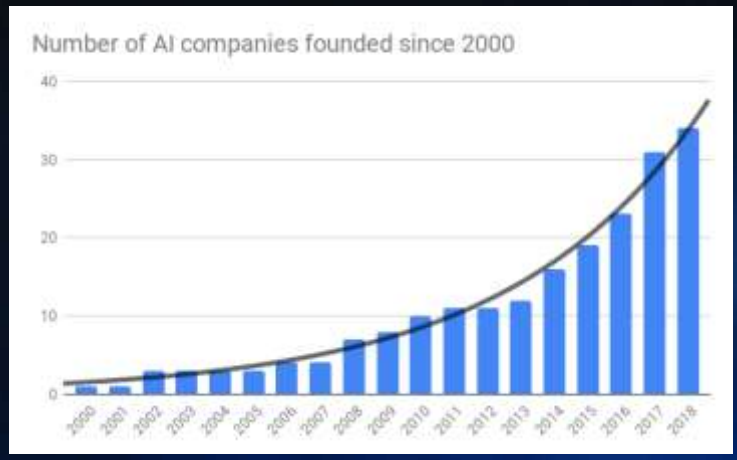


Philip Brey

University of Twente
The Netherlands

The booming AI industry

- AI is booming
- AI Market driven by factors such as:
  - Improved productivity/efficiency of algorithms
  - Significant increase in use of Big Data
  - Wide range of application areas



- Expected AI Market growth: from $9.5 billion in 2018 to $118.6 billion by 2025

  (https://www.tractica.com/research/artificial-intelligence-market-forecasts/)

AI brings significant social and ethical challenges

- Individual and human rights (autonomy, privacy, liberty, dignity)
- Fairness and justice
- Well-being and the quality of society
- The future of work and employment

Recent ethics guidelines for AI

- High-Level Expert Group on AI (HLEG) of the European Commission: The Ethics Guidelines for Trustworthy Artificial Intelligence (AI)

- IEEE (Institute of Electrical and Electronics Engineers): Ethically Aligned Design

Recent ethics guidelines for AI

- OECD (Organisation for Economic Cooperation and Development): OECD Principles on AI

- UNESCO: First draft of the Recommendation on the Ethics of Artificial Intelligence

Convergence of these guidelines

	IEEE	HLEG	OECD	UNESCO
autonomy	**	****	***	**
freedom	***	***	***	****
dignity	***	***	***	****
privacy	****	****	***	****
safety/security	**	****	****	***
justice/fairness	**	****	****	****
accountability	****	****	****	****
transparency	****	****	****	****
indiv. well-being	****	**	****	****
soc/env. well-being	***	****	***	****

EU Ethical AI: Seven requirements

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Environmental and societal well-being
7. Accountability

From guidelines to action

Limitations of ethics guidelines:

- They are guidelines for all actors and contexts. How are they to be implemented in specific contexts and practices, and by specific actors?

- Won’t we need guidelines for specific actors, specific products, specific application domains?

- Will their voluntary character be enough?

- Ethical guidelines are rules; don’t we need to develop ethical competencies?

Two-way approach (SIENNA project)

1. Operationalizing ethics guidelines

Operational guidance frameworks for specific actors and practices, specific products, specific application domains

Two-way approach (SIENNA project)

2. Development of other methods to support ethics in AI

Multistakeholder strategic approach to AI

Actors: researchers, tech developers, policy makers, civil society actors, educators, etc.

Practices: using AI systems, reporting on AI impacts, regulating AI development and use, etc.

Methods and tools: technical standards, educational programmes, research ethics frameworks, regulations, media campaigns, etc.

Distribution of responsibilities and collaboration between actors

Operationalizing and implementing ethics guidelines

In SIENNA, we have developed ethics guidance for researchers, developers and organisational users, and for specific products and application domains

- Research ethics guidelines for AI
- Ethics by Design approach for AI
- Coding guidelines into regulations and technical standards
- Ethics of Deployment and Use for Organisational users of AI
- Special topics (guidelines for products, techniques and application domains)

Research ethics guidelines for AI

The SIENNA approach:

- Use AI HLEG ethics requirements
- Translate to R&D guidelines
- Add guidelines for specific AI and robotics products, techniques and application domains
- Add Ethics by Design as a recommendation

Research ethics guidelines for AI

Our approach is implemented in Horizon Europe for ethics review of AI, robotics and data analytics projects

It is also available for research ethics commitees and researchers in our deliverable D5.4.

Human Agency

- AI should not try to control people, or enable such control:
  - Make choices that are personal, communal, or political
    - Especially issues of well-being, individual rights, economic, social and political decisions.
  - Remove basic freedoms
  - Subordinate, coerce, deceive, manipulate, or dehumanize people
  - Stimulate dependency or addiction

Privacy and Data Governance

- AI must respect the right to privacy.
- AI's use of data must be *actively* governed
  - Supervised, modified if necessary
- Adhere to GDPR
- Data usage should be auditable by humans
  - EG: Model Cards, Datasheets for Datasets, XAI

Fairness

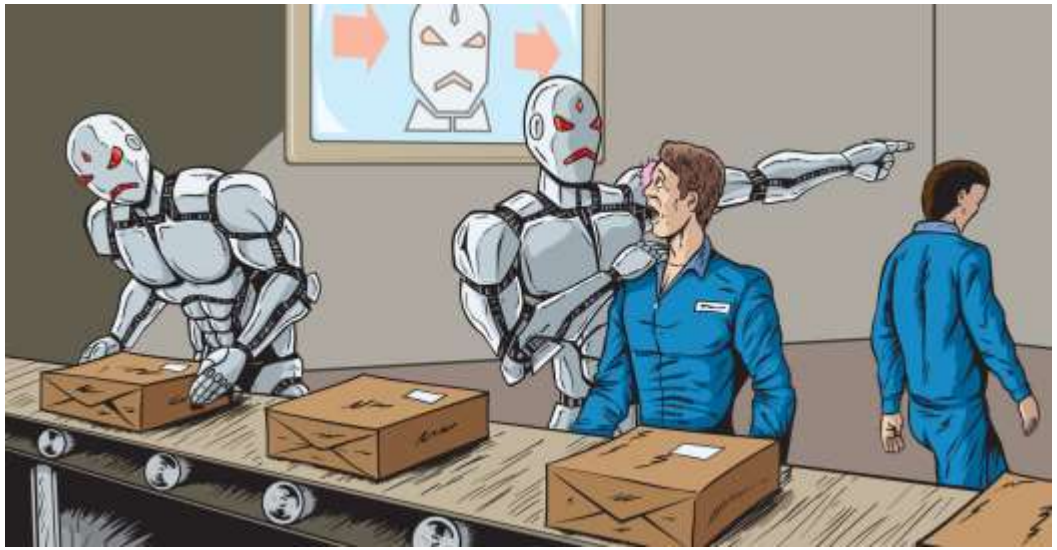
- People should be given equal rights and opportunities and should not be advantaged or disadvantaged undeservedly

Fairness

- Avoidance of algorithmic bias in input data, modelling, algorithm design
- Universal accessibility:
  - AI systems should be designed so that they are usable by different types of end-users with different abilities.
- Fair impacts:
  - Evidence that possible negative or discriminatory social impacts on certain groups have been considered and mitigated for if possible

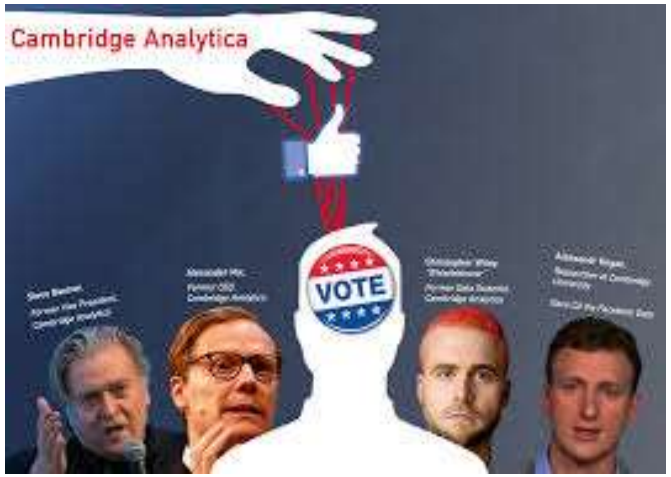
Individual, Social & Environmental Well-being

- AI systems should not harm, individual, social or environmental well-being
- AI systems should consider the welfare of all stakeholders



Individual, Social & Environmental Well-being

- Documented efforts to consider environmental impact
- Special consideration of AI for media, communications, politics, social analytics, and online communities: support quality communication, avoid fake news, filter bubbles, echo chambers, political manipulation



sienna.

Transparency

- *If relevant,* humans must be able to understand how
  - The AI functions
  - How the AI decisions are arrived at
- Best practice = XAI - eXplainable AI
- Enable traceability of the AI system during its entire lifecycle

Transparency

- It must be clear to end-users that they are interacting with an AI
- Open communication of purpose, capabilities, limitations, benefits, risks, decisions by AI system, governance
- Keep records of decisions about ethics made during construction

Accountability & Oversight

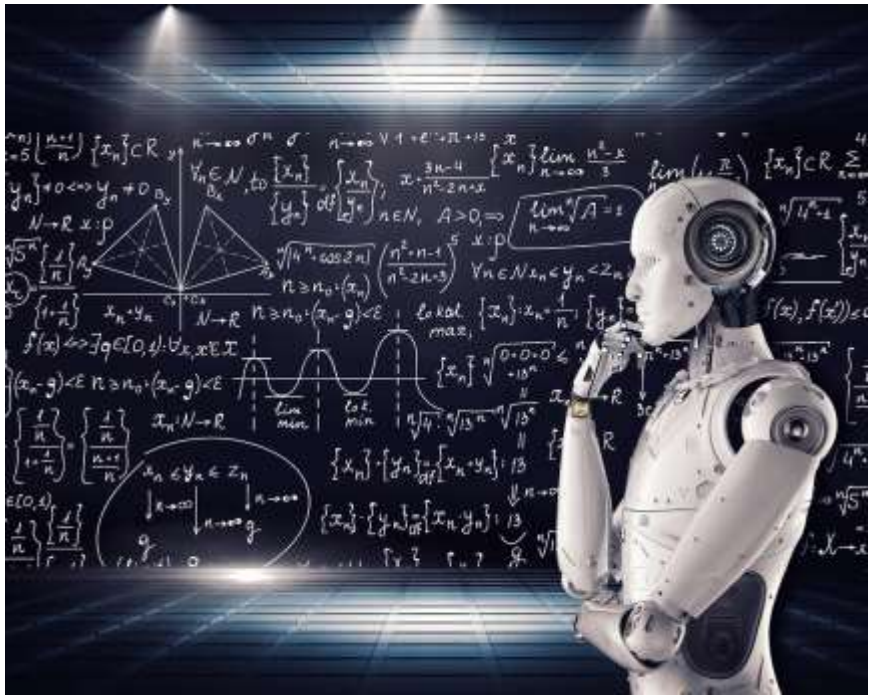
- Accountability: people who build or operate are responsible for the AI's actions/effects
  - Developers must be able to explain how and why a system acts the way it does
    - Unless compelling reasons provided to show oversight not required

Accountability & Oversight

- Applications must explain how undesirable effects will be detected, stopped, and prevented from reoccurring
- All AI systems should be auditable by independent third parties
- Oversight: humans can understand, supervise, control design and operation
  - Documented procedures for risk assessment and mitigation
  - How will people be able to report concerns?
    - How they will be evaluated and actioned?

Special guidelines

- Ethically aware AI
- Fully autonomous AI
- Subliminal and addictive AI
- Deep learning
- Affective Computing
- AI for children and vulnerable groups
- Medical applications of AI
- Processing of sensitive data
- AI systems that automate work processes



Special guidelines

- Analysis of speech, text, internet behaviour, and images and video of human beings
- Merging of databases and mining of personal information
- Surveillance involving personal identification and location tracking
- Predictive analytics in relation to persons and their behaviors

Special guidelines

- Covert and deceptive AI
- AI with applications in media and politics
- Autonomous and semi-autonomous weapons systems
- Decision Support Systems
- Human-like AI
- Embedded Systems and Internet-of-Things

sienna.

Ethics by Design

Guidelines and instructions for different stages of the development process

Our approach will also be included in Horizon Europe ethics review

THANK YOU !

www.sienna-project.eu