FACILITATE-AI

GUIDELINES FOR FACILITATING THE LEARNING OF ARTIFICIAL INTELLIGENCE BY SCHOOL STUDENTS OF GRADES 7-12

FACILITATE – AI: Guidelines for facilitating the learning of Artificial Intelligence (AI) by School Students of Grades 7-12

Project Number: 2021-1-CY01-KA220-SCH-000032567

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

FACILITATE-AI

Co-funded by the European Union

C1 Training course: **Ethics About AI**

- **Result 1, A1/T1**
  Module Number and Area/ Topic: Ethics About AI
  Module owners: UCY, CY
- Coordinator: Antonis Kakas

Project Number: 2021-1-CY01-KA220-SCH-000032567

FACILITATE-AI

Co-funded by the European Union

AI Ethics
A high-level Overview

- **AI Ethics is very Important**
  - **Dangers vs Opportunities**
  - **Levels of Dangers**
- **AI Ethics is Hard**
  - **How do we build Ethical Machines?**
  - **Teach or Grow?**
- **AI Ethics in EU**
  - **Guidelines, Regulations and Laws!**
    - **EU AI act**
  - **But**

Project Number: 2021-1-CY01-KA220-SCH-000032567

FACILITATE-AI

Co-funded by the European Union

EU AI Ethics

European-approach-artificial-intelligence

European Artificial Intelligence Act: the right framework for trustworthy AI and innovation?

Shaping Europe's digital future

Home Policies Activities News Library Funding Calendar Consultations

Home > Policies > A European approach to artificial intelligence

A European approach to artificial intelligence

The EU's approach to artificial intelligence centers on excellence and trust, aiming to boost research and industrial capacity while ensuring safety and fundamental rights.

The way we approach Artificial Intelligence (AI) will define the world we live in the future. To help building a resilient Europe for the Digital Decade, people and businesses should be able to enjoy the benefits of AI while feeling safe and protected.

The European AI Strategy aims at making the EU a world-class hub for AI and ensuring that AI is

Project Number: 2021-1-CY01-KA220-SCH-000032567

FACILITATE-AI

Co-funded by the European Union

AI Ethics
Teacher Training Videos

FROM EU ERASMUS+ Trustworthy AI

- https://www.trustworthyaiproject.eu/teaching-trustworthy-ai-oers/

Project Number: 2021-1-CY01-KA220-SCH-000032567

Responsible ai 1-0-1: from values to requirements

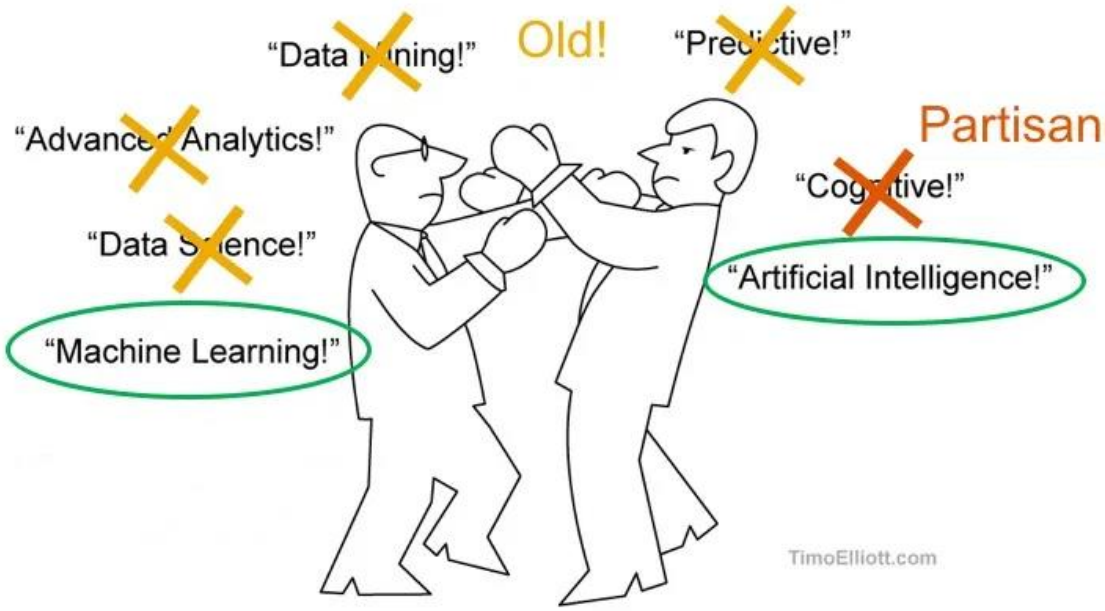
Dr. Andreas Theodorou

andreas.theodorou@umu.se

@recklesscoding

UMEÅ UNIVERSITY

What is AI?


"Data Mining!"
Old!
"Predictive!"
"Advanced Analytics!"
Partisan(?)
"Cognitive!"
"Data Science!"
"Artificial Intelligence!"
"Machine Learning!"
TimoElliott.com

What is AI?

- **Simulation of natural intelligence;** the field of AI includes the study of theories and methods for adaptability, interaction and autonomy of machines (virtual or embedded)

- A (computational) technology that is able to infer patterns and possibly draw conclusions from **data.**

- **An (autonomous) entity** (e.g. when one refers to **'an' AI**); this is the most usual reference in media and science fiction.

“AS SOON AS IT WORKS, NO ONE CALLS IT AI ANY MORE.”

John McCarthy

This Lack of agreement leads to…

- A constant **re-writing of similar high-level policy statements:**
  - Opens loopholes to be exploited.
  - We will see later more about this!
- **Increases misconceptions** (true AI, superintelligence, etc):
  - AI is not a magic abstraction.
  - Computation is a **physical process**; it requires *energy*, *space*, and *time*.

Theodorou, A. and Dignum V. (2020), *Towards ethical and socio-legal governance in AI.* Nature Machine Intelligence, 2(1).

Narratives in the news

"Alexa, Can I Trust You?"

Hyunji Chung, Michaela Iorga, and Jeffrey Voas, NIST
Sangjin Lee, Korea University

BBC Sign in | News | Sport | Weather | Shop | Reel | Travel | More | Search

NEWS

Home | Video | World | UK | Business | Tech | Science | Stories | Entertainment & Arts | Health | World News TV | More

Technology

Amazon scrapped 'sexist AI' tool

10 October 2018

Top Stories

US attorney general held in contempt

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By James Vincent

Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum

COMPROP RESEARCH NOTE 2016.1

Philip N. Howard
Oxford University
philip.howard@oii.ox.ac.uk

Bence Kollanyi
Corvinus University
kollanyi@gmail.com

DISINFORMATION AND SOCIAL BOT OPERATIONS IN THE RUN UP TO THE 2017 FRENCH PRESIDENTIAL ELECTION

EMILIO FERRARA
UNIVERSITY OF SOUTHERN CALIFORNIA, INFORMATION SCIENCES INSTITUTE

ABSTRACT

Recent accounts from researchers, journalists, as well as federal investigators, reached a unanimous conclusion: social media are systematically exploited to manipulate and alter public opinion. Some

TIME

SPOTLIGHT STORY INSIDE THAILAND'S HAZY LEGALIZATION OF WEED | SIGN IN | SUBSCRIBE

BUSINESS • COVID-19

Millions of Americans Have Lost Jobs in the Pandemic—And Robots and AI Are Replacing Them Faster Than Ever

Andreas Theodorou | t: @recklesscoding

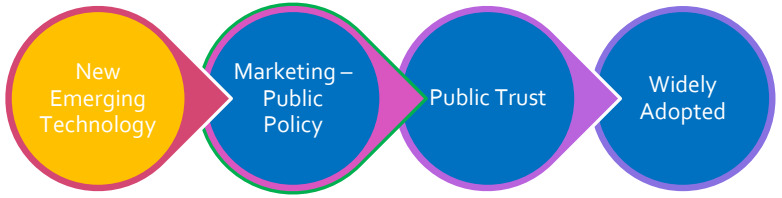
2017 Eurobarometer

- **61%** of respondents have a **positive view** of robots
- **84%** of respondents agree that **robots can do jobs** that are too **hard**/**dangerous** for people
- **68%** agree that robots are a **good** thing **for society** because they help people
- **88%** of respondents consider robotics a technology that **requires careful management**
- **72%** of respondents think robots **steal people's jobs**

We need to build trust for our systems

- To **perform as we expect them to.**
- The implications from their development and deployment fall within:
  - **E**thical
  - **L**egal
  - **S**ocial
  - **E**conomic
  - **C**ultural

  (**ESLEC**) specifications and values we want to protect and **communicating that** to the public.



Slide provided by DG Connect

European Union Background on AI

EU STRATEGY ON ARTIFICIAL INTELLIGENCE

published in April 2018

- Boost AI uptake
- Tackle socio-economic changes
- Ensure adequate ethical & legal framework

In this context: appointment of Independent High-Level Expert Group on Artificial Intelligence (AI HLEG) in June 2018

European Commission

Slide provided by DG Connect

Ethics Guidelines for AI – Requirements

Human agency and oversight

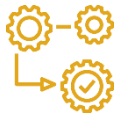
Technical Robustness and safety

Privacy and data governance

Transparency

Diversity, non-discrimination and fairness

Societal & environmental well-being



Accountability

To be continuously implemented & evaluated throughout AI system's life cycle

European Commission

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	• infringe **human rights**? • Traditional metrics of prosperity do not take into account the full 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**?

Which Values to implement?
How?

FACILITATE-AI

EXAMPLE ASSESSMENT

Co-funded by the European Union

- Did you build in mechanisms for notice and control over personal data depending on the use case (such as valid consent and possibility to revoke, when applicable)?
- Did you assess what level and definition of accuracy would be required in the context of the AI system and use case?
  - Did you assess how accuracy is measured and assured?
  - Did you put in place measures to ensure that the data used is comprehensive and up to date?
  - Did you put in place measures in place to assess whether there is a need for additional data, for example to improve accuracy or to eliminate bias?
- Did you assess:
  - to what extent the decisions and hence the outcome made by the AI system can be understood?
  - to what degree the system's decision influences the organisation's decision-making processes?
  - why this particular system was deployed in this specific area?
  - what the system's business model is (for example, how does it create value for the organisation)?
- In case the AI system interacts directly with humans:
  - Did you assess whether the AI system encourages humans to develop attachment and empathy towards the system?
  - Did you ensure that the AI system clearly signals that its social interaction is simulated and that it

Laborious
Abstract
Not-actionable
Generic

Responsible ai & Human Control

WHAT IS AI ETHICS?

- Umbrella term for many things:
  - moral agency;
  - existential crisis/superintelligence;
  - **adherence (or not) to human moral values;**
  - **trustworthiness.**

- It is not telling people if they are 'morally right' or morally wrong.'

RESPONSIBLE AI

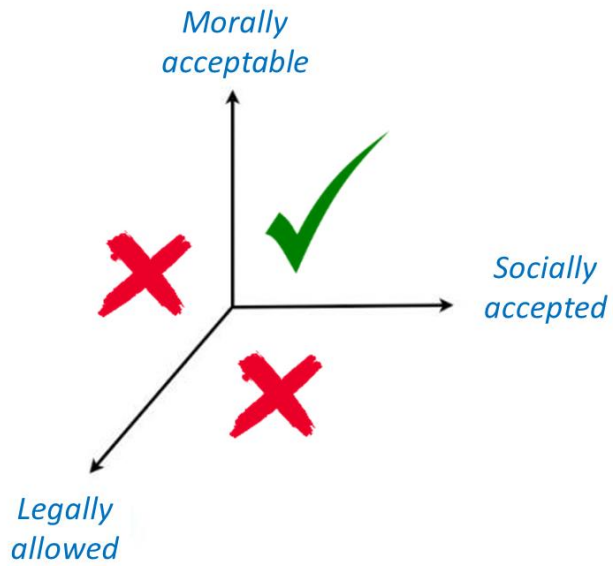
- AI research includes the study of theories and methods for adaptability, interaction and autonomy of machines.

- Responsible AI is the understanding of how we can **develop socially-beneficial systems** and **ensure human control**.

Responsible AI = ethical + legal + robust + **verifiable AI**

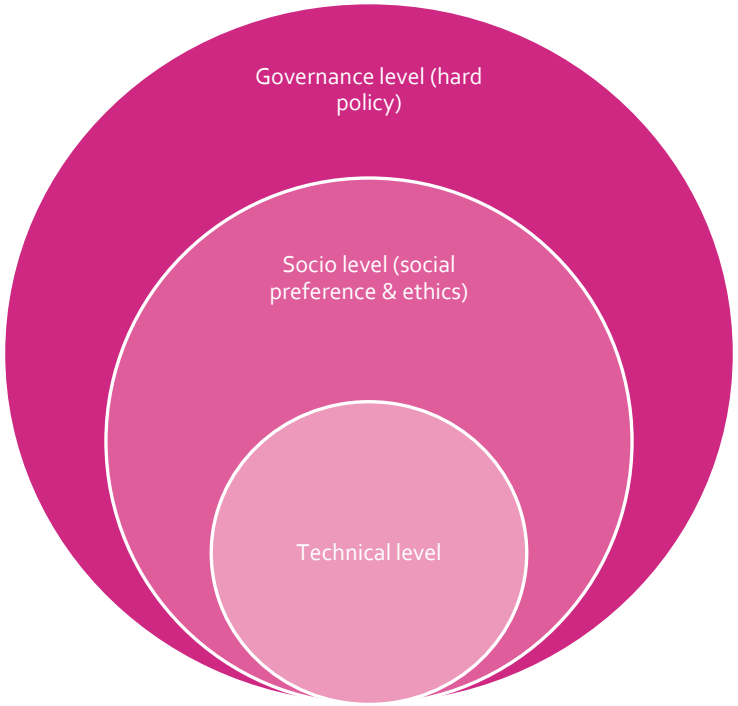
Trustworthy AI (HLEG)

ETHICS != LAW != SOCIAL NORMS



Theodorou, A. and Dignum V. (2020), *Towards ethical and socio-legal governance in AI.* Nature Machine Intelligence, 2(1).

Socio-technical systems



AI involves

- ***In*** Design: **Development** is **influenced** by ESLEC issues.
- ***By*** Design: **Integration** of ethical abilities as part of the **behaviour** of artificial intelligent systems.
- ***For*** Design: Codes of conduct, standards, and certification processes that ensure the **integrity** of **key stakeholders**.

Dignum, V (2018). *Ethics in Artificial Intelligence: Introduction to the special issue.* Ethics and Information Technology, 20(1):1–3, 3 2018.

Values integration

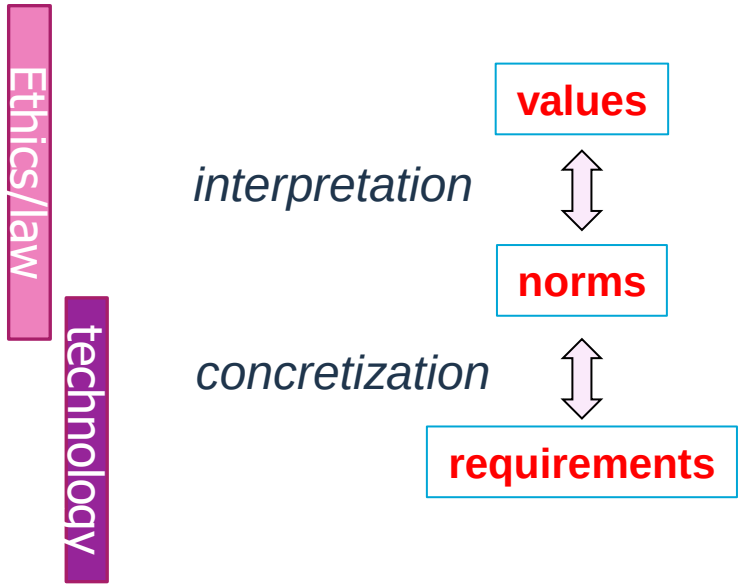
WORKING AROUND THE PROBLEM

- Stop thinking of AI as 'magic data' instead of software.
- We do not necessary have to agree what *each value* means.
- We need to make their interpretations **explicit and transparent**.

Theodorou, A. and Dignum V. (2020), *Towards ethical and socio-legal governance in AI*. Nature Machine Intelligence, 2(1).

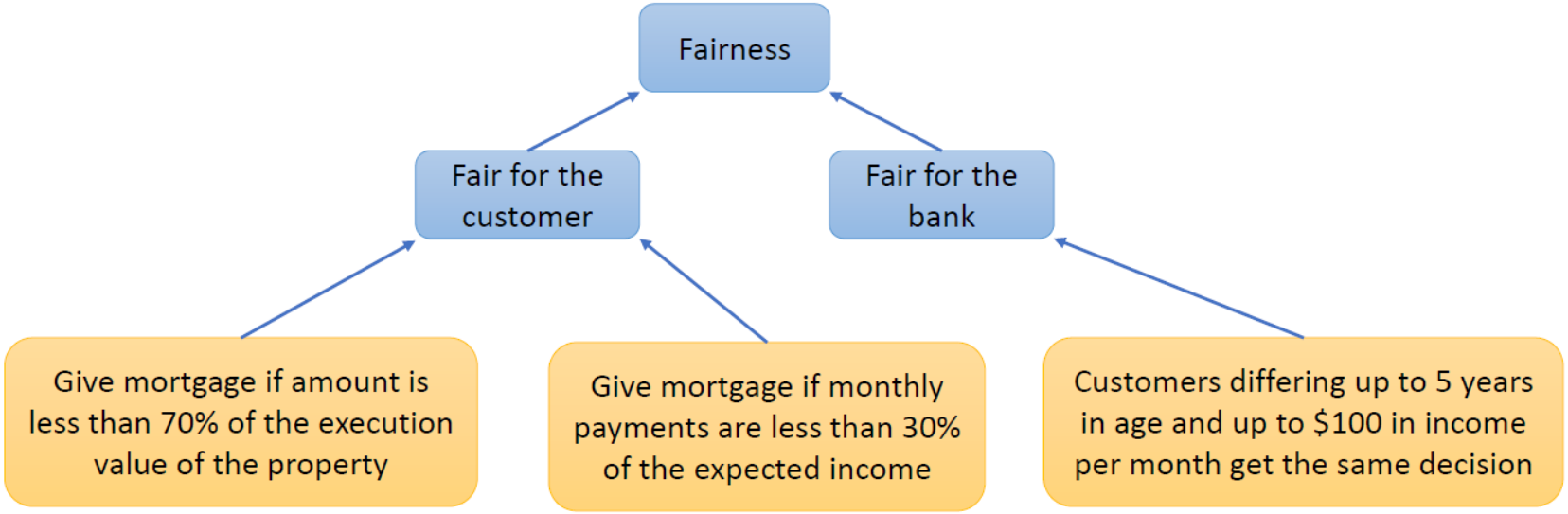
Interpretating values

- Structured and explicit process of translating translate **abstract values** into **concrete norms** and **requirements**.
- We aim to not only **describe** the norms themselves, but also **the exact connection** between abstract and concrete concepts **in each context**.
- Fulfilling the norm will be considered as adhering to the value.



Aler Tubella A., **Theodorou A.**, Dignum F., Dignum V. (2019). *Governance by Glass-box: implementing transparent moral bounds for AI behaviour*. IJCAI

Interpreting values



Aler Tubella A., **Theodorou A.**, Dignum F., Dignum V. (2019). *Governance by Glass-box: implementing transparent moral bounds for AI behaviour*. IJCAI

Accountability & responsibility

- Accountability is ***backwards thinking*****,** provides an account of events after they have occurred.
- Responsibility is ***forwards thinking***, i.e., acting to deter incidents and violations of our ethical and legal values from occurring.
- To be held accountable, you need to be held responsible…

accountability

- Umbrella term, covering amongst others:
  1. legal accountability ;
  2. professional accountability;
  3. political accountability;
  4. administrative accountability; and
  5. social accountability.
- Used not only to indicate punishment but also acceptance of responsibility.

Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework. Eur. L. J.

accountability

- Incidents will occur —and sometimes reoccur.
- It is essential for **maintaining the public's trust** in the technology.
- It is a **mean of control**: that compels adherence to specified regulations and practices to demonstrate *due diligence*.

Transparency in the IEEE 7001 Standard

- "To consider an autonomous system transparent to inspection, the stakeholder should have the ability to request meaningful explanations of the **system's status either at a specific moment or over a specific period** or of the general principles by which decisions are made (as appropriate to the stakeholder) (Theodorou et. al., 2017)"

Winfield AFT, Booth S, Dennis LA, Egawa T, Hastie H, Jacobs N, Muttram RI, Olszewska JI, Rajabiyazdi F, **Theodorou A,** Underwood MA, Wortham RH and Watson E (2021) *IEEE P7001: A Proposed Standard on Transparency*. Front. Robot. AI

Theodorou A., Wortham R.H., and Bryson J. (2017). *Designing transparency for real time inspection of autonomous robots*. Connection Science, Vol. 29, Issue 3

Transparency as Contestability

- Our right to contest decisions made for us is not only protected by the Regulation (EU), 2016 GDPR

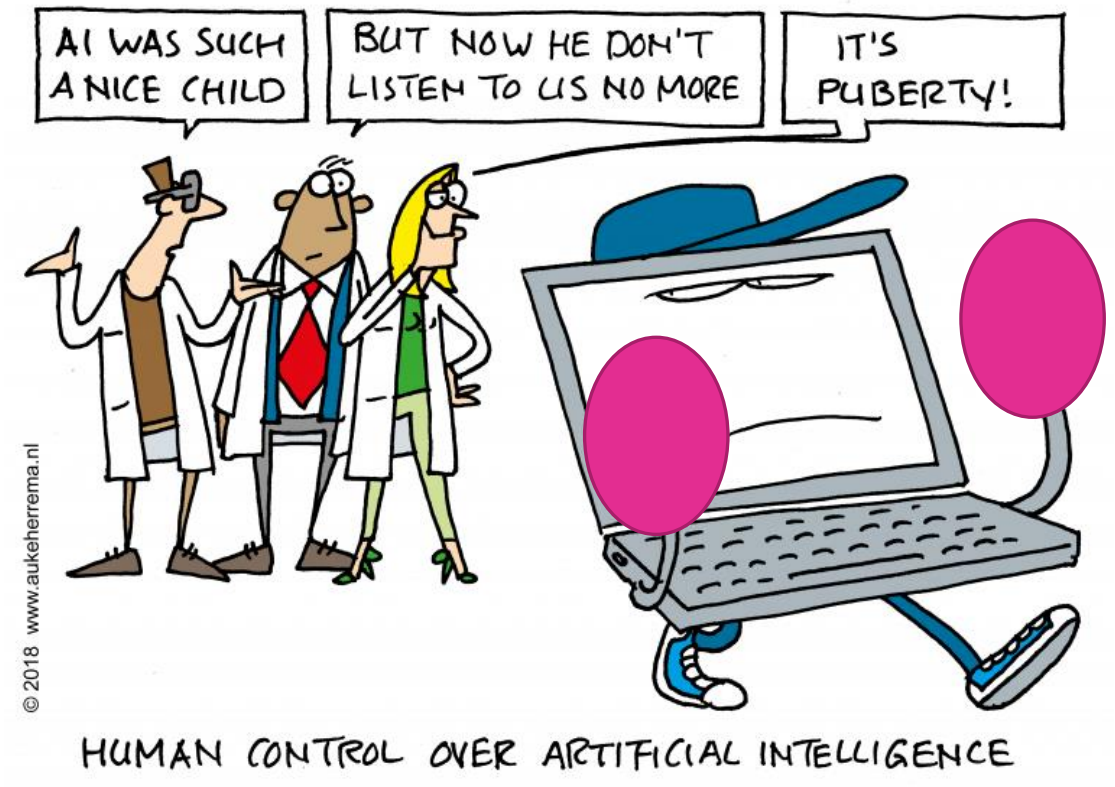
- Requires looking beyond why a decision was made: a decision needs to be both ***correct*** and ***permissible***.

Aler Tubella, A., **Theodorou, A.**, Dignum, V., and Michael, L. (2020). *Contestable Black Boxes.* International Joint Conference on Rules and Reasoning.

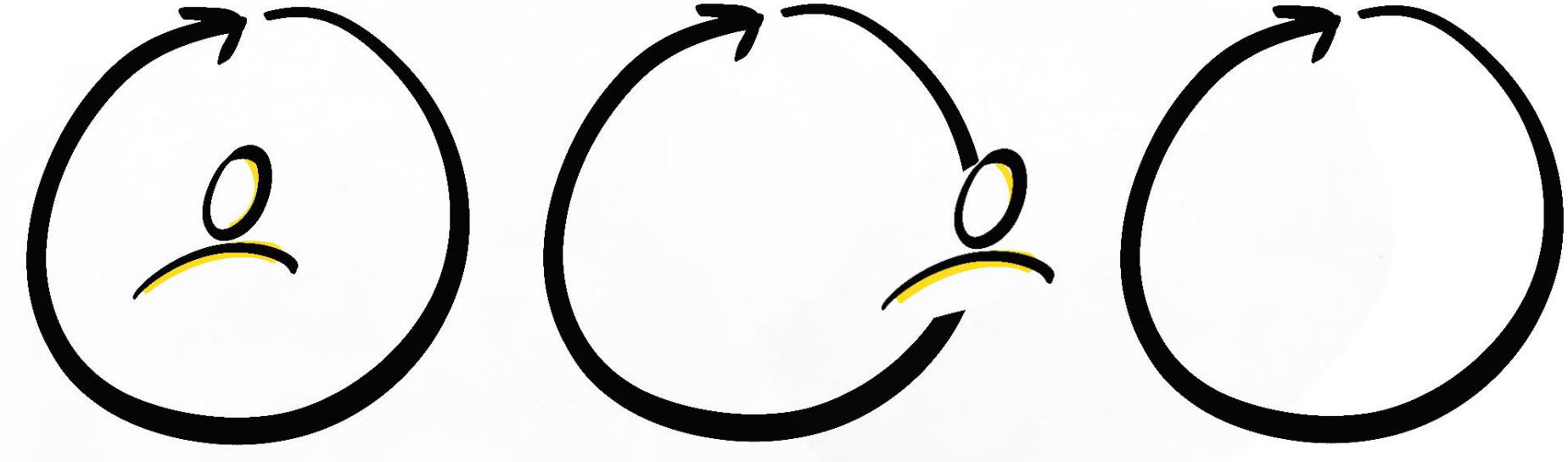
Human control

UMEÅ UNIVERSITY

Human control



Current frameworks



Human-in-the-loop
(HITL)

Human-on-the-loop
(HOTL)

Human-out-of-the-loop
(HOOTL)

Drawings made by Methnani L. (UMU)

Meaningful Human control

- Human presence is not sufficient.

- We need to be able to:
  - track relevant reasons behind a decision;
  - verify their compliance to any policy; and
  - trace back to an individual along the chain who is aware and accepting of their responsibility.

Santoni de Sio, F., and Van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account

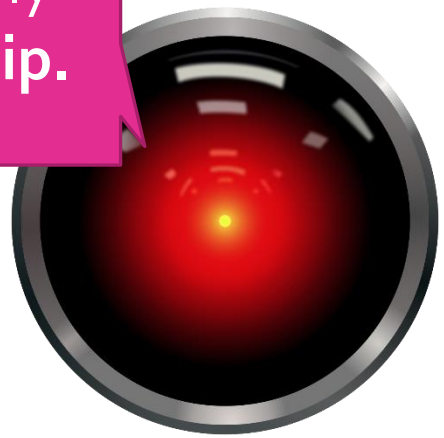
Meaningful Human control

- **The world is not static**: contexts that change will demand changing levels of responsiveness.

- Requires us to consider wider issues around Responsible AI.

Variable autonomy

I will fly the **ship.**

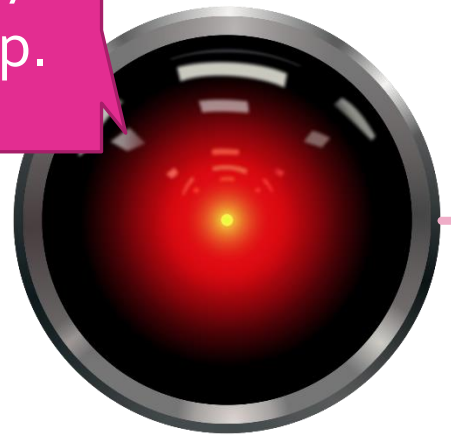


I don't care.

Methnani L., Aler Tubella A., Dignum V and **Theodorou A.** (2021). *Let Me Take Over: Variable Autonomy for Meaningful Human Control*. Front. Artif. Intell.

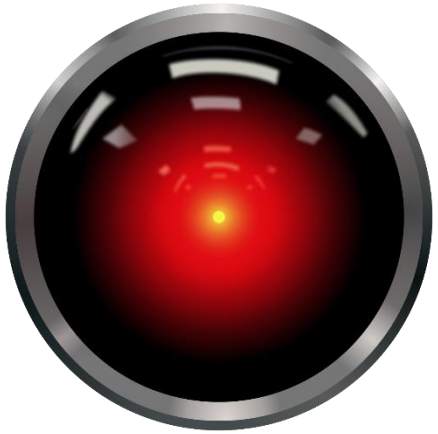
Variable autonomy

I will fly the ship.



Takes action

Variable autonomy



Variable autonomy

Takes action

You never trained in this environment. I will take over now.

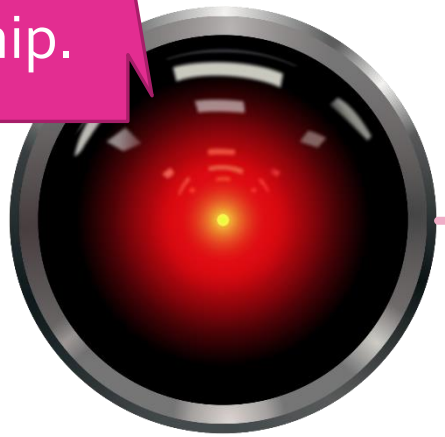
Variable autonomy

Takes action

I am bored and taking control.

Variable autonomy

Asteroid ahead, I will stop the ship.

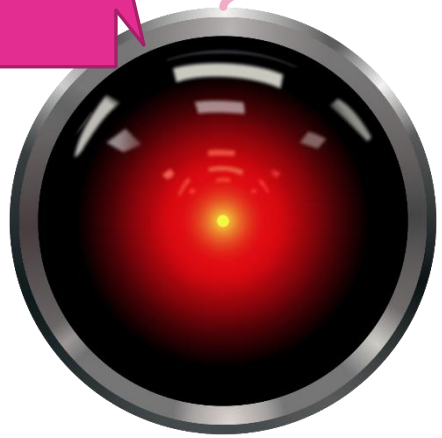


Takes action

Variable autonomy

Informs of an action

I will take care of ship navigation.



OK. I will take care of our life-support system

Variable autonomy

- **Maximises performance**: shifts autonomy both ways.
- **Maximises human control:** human presence adjusted as the environment changes.
- Ensures **ART**.

Methnani L., Aler Tubella A., Dignum V and **Theodorou A.** (2021). *Let Me Take Over: Variable Autonomy for Meaningful Human Control*. Front. Artif. Intell.

Mental Models

- With poor transparency, **robots mislead us due to our lack of theory of mind for them.**
- Misunderstanding leads to anxiety, mistrust, fear, and misuse/disuse. Introduces user self-doubt.
- **Transparency helps us build accurate mental models.**

Technical solutions

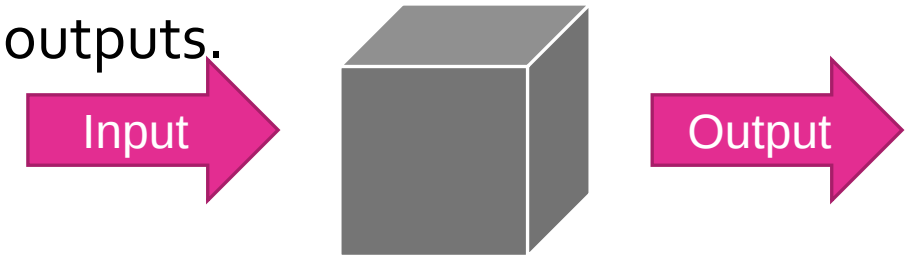
Keeping the black box

- **Sometimes black boxes are inevitable.**
- Some of the best-performing methods for pattern recognition, e.g. deep learning, are black boxes right now.
- Yet, we still need to audit our systems.

testing black boxes

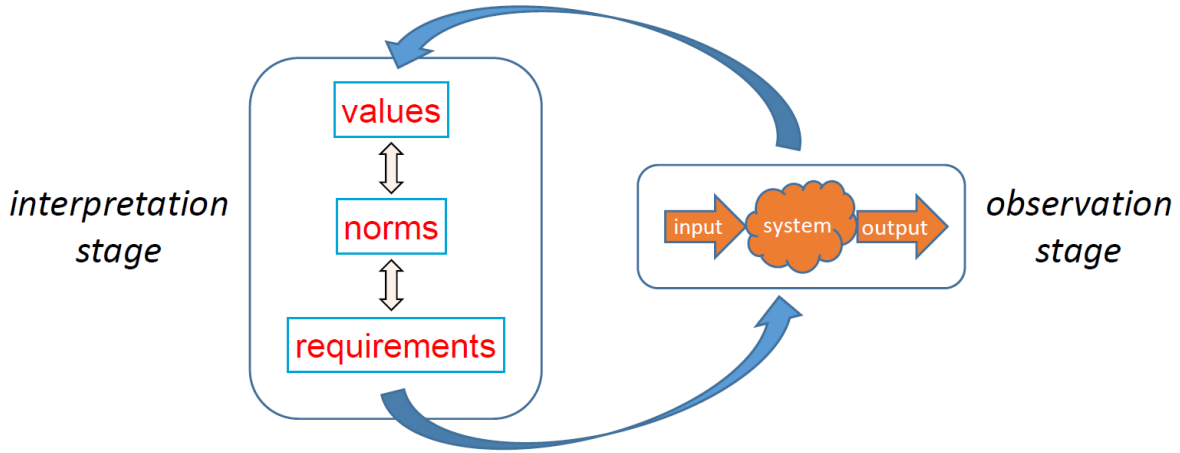
Black-box testing for Responsible ai

- **AI systems are software.**
- White-box testing: the focus on testing is in the internal structures an application
- Black-box testing: ignores the 'inside' of a software, the focus on the inputs and outputs.



Black-box testing for Responsible ai

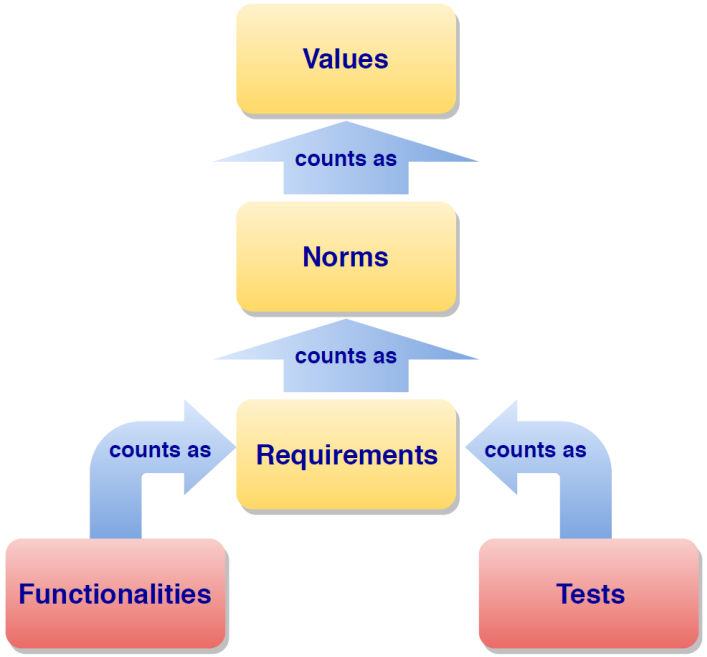
- Testing a system adherence to ethical values.



Aler Tubella A., **Theodorou A.**, Dignum F., Dignum V. (2019). *Governance by Glass-box: implementing transparent moral bounds for AI behaviour*. International Joint Conference on Artificial Intelligence (IJCAI).

Formalising the interpretation stage

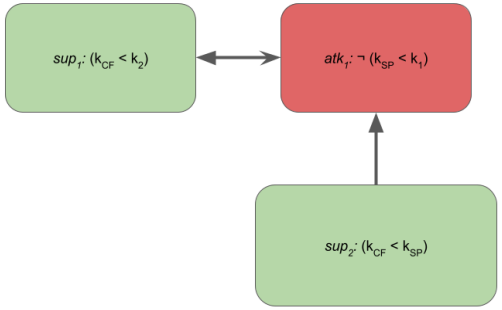
- We encode statements of the form: "***A* counts-as *B* in context *C***".
- It allows for verification in reasonable time.

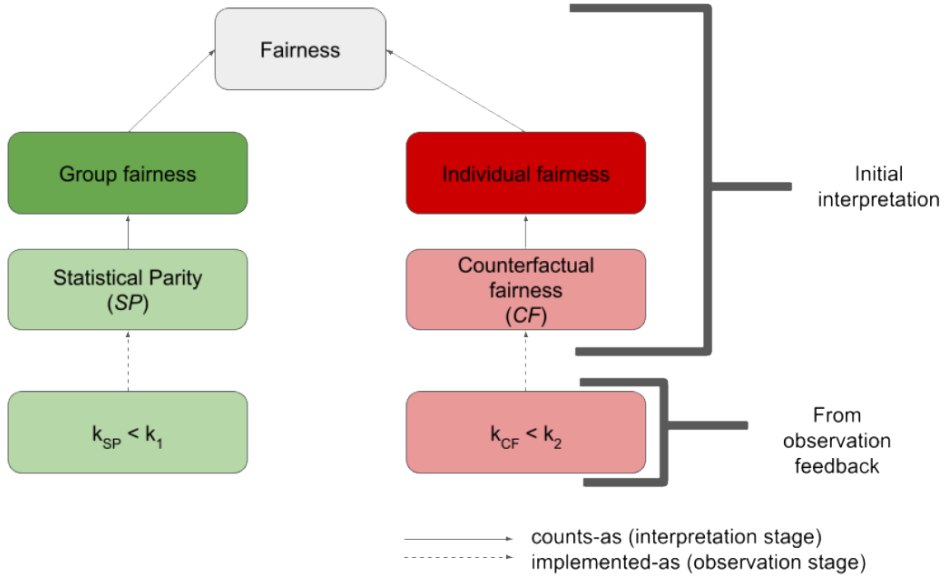


Aler Tubella A., Dignum V. (2019). *The Glass Box Approach: Verifying Contextual Adherence to Values*. Workshop in AI Safety 2019

Formalising the OBSERVATION stage

- Tests can be established metrics in binary format.
-



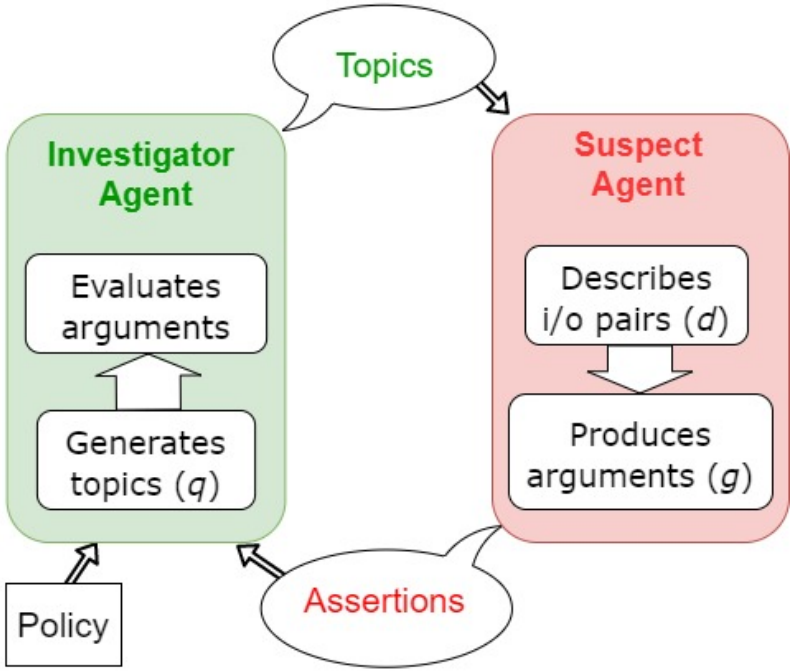


But how consistently do we adhere to a policy?

Why not just ask the system itself?

Explainability and Dialogue

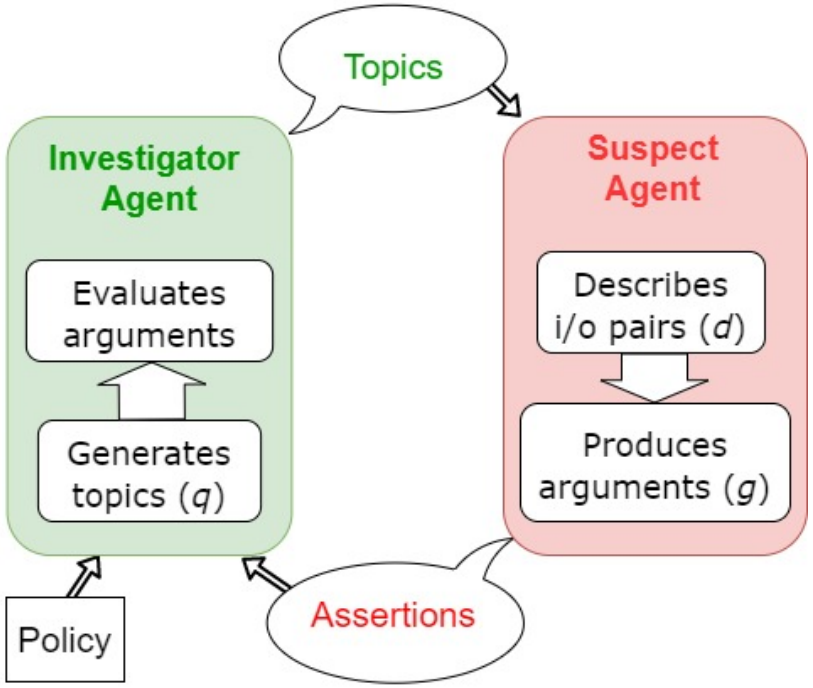
Transparency through Information-Seeking Dialogues



Aler Tubella A., **Theodorou A.**, Nieves J.C. (2021). *Interrogating the Black Box: Transparency through Information-Seeking Dialogues*. AAMAS 2021

Agents-modelling-other-agents

- Inspired by agents-modelling-agents literature.
- Agent communicate through **information-seeking dialogues**.
- Aggregation of information with **argumentation semantics**.



Aler Tubella A., **Theodorou A.**, Nieves J.C. (2021). *Interrogating the Black Box: Transparency through Information-Seeking Dialogues*. AAMAS 2021

The Awkward architecture

Socially-Aware Agents

for

Ethics By Design

The social behavior dilemma

- Fulfilment of societal goals requires **altruistic** or other **pro-social behaviour**.
- Fulfilment of individual goals may require **freeriding, selfish,** and **anti-social** behaviours.
- There is a need to alternate between different behaviours given the situation.

EU AI Ethics

european-approach-artificial-intelligence

European Artificial Intelligence Act: the right framework for trustworthy AI and innovation?

Shaping Europe's digital future

Home Policies Activities News Library Funding Calendar Consultations

Home > Policies > A European approach to artificial intelligence

A European approach to artificial intelligence

The EU's approach to artificial intelligence centers on excellence and trust, aiming to boost research and industrial capacity while ensuring safety and fundamental rights.

The way we approach Artificial Intelligence (AI) will define the world we live in the future. To help building a resilient Europe for the Digital Decade, people and businesses should be able to enjoy the benefits of AI while feeling safe and protected.

The European AI Strategy aims at making the EU a world-class hub for AI and ensuring that AI is

Many Thanks to

RAI Group @ UmU

Prof. Dr. Virginia Dignum
Professor of Responsible AI
@vdignum
virginia@cs.umu.se

Dr. Andrea Aler Tubella
Senior Research Engineer
andrea.aler@umu.se

Mattias Brännström
Research Engineer
mattias.brannstrom@umu.se

Dr. Juan Carlos Nieves
Associate Professor
@jcnieves
juan.carlos.nieves@umu.se

Leila Methnani
PhD Student
leila.methnani@cs.umu.se

Other members:
A. Brännström, *PhD student*
Dr. P. Eriksson, *postdoc*
A. Horned, *PhD student*
Dr. A. Johnsson, *postdoc*
K. Mendez, *PhD student*
Dr. L. Vanhée, *Associate Professor*

Other collaborators

Dr. A. Antoniades – Builder.AI (UK)
Prof. J.J. Bryson – Hertie Schol (DE)
Prof. A. Colman – Princeton University (USA)
Dr. A. Cortes – Barcelona Supercomputer Centre (ES)
Dr. M. De Vos – Uni. Of Bath (UK)
Prof. F. Dignum — UmU (SE)
Prof. A. Kakas – Uni. Of Cyprus (CY)
Dr. L. Michael – Open Uni. Cyprus (CY)
C. Muller – ALLAI (NL)
A. Rotisidis – Uni. Of Bath
Dr. L. Sartori - Univ. of Bologna (IT)
Dr. T. Scantamburlo -Univ. Of Venice (IT)
B. Wilder – Harvard University (USA)
H. Wilson – Uni. Of Bath (UK)
Prof. A. Winfield – Bristol Robotics Lab/UWE (UK)
Dr. R. H. Wortham – Uni. Of Bath (UK)

Current Funding Sources:

Knut and Alice Wallenberg Foundation
WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM
WASP—HS
Horizon2020 European Union Funding for Research & Innovation
HUMANE AI NET
Grant agreement 761758
VINNOVA
ITEA 4
explAIn
Erasmus+
Trustworthy AI
Project 2020-1-ES01-KA203-082983

FACILITATE-AI

Co-funded by the European Union

FACILITATE – AI Partners

Plovdiv University "Paisii Hilendarski"

DOUKAS SCHOOL 1917

IASA

ITCSPACLE "E. MORANTE " LIMBIATE

University of Cyprus

P.PORTO

Prof. Ivan Apostolov ENGLISH LANGUAGE SCHOOL

Project Number: 2021-1-CY01-KA220-SCH-000032567