
Safe and Responsible AI Masterclass
WEBINAR
24 juni
19.00-20.30u
online

Open Universiteit

ABOUT US

Natasha Alechina
Professor in Safe & Responsible AI

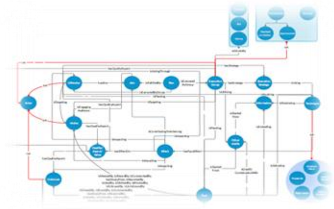
Clara Maathuis
Assist. Prof. in AI & Cyber Security

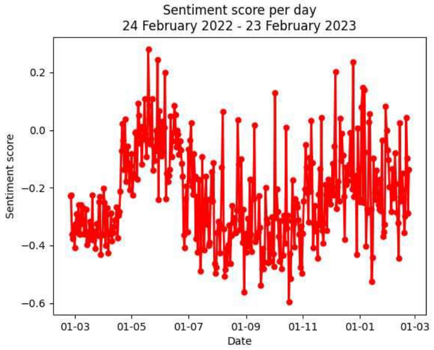
NATASHA

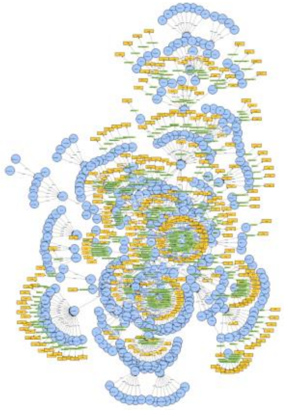
- PhD in logic, ILLC, University of Amsterdam
- Verification of autonomous agents and multi-agent systems, synthesis of AI systems to specification
- Current projects include run-time verification of robot-assisted surgery
- Teaching Logics for Safe AI at Utrecht University, new course AI en maatschappij (with Clara) at OU

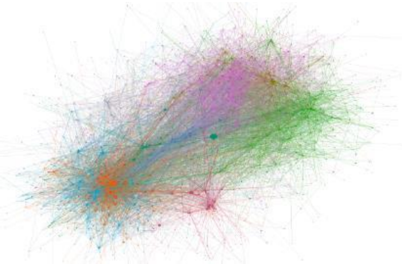
CLARA

- PhD in AI & Military Cyber Operations, TU Delft, the Netherlands.
- AI, Responsible AI, Trustworthy AI, AI Security.
- AI applications in defense, cyber security, disinformation & deepfakes, diversity, and XR domains.
- Teaching courses like Machine Learning and Responsible AI.





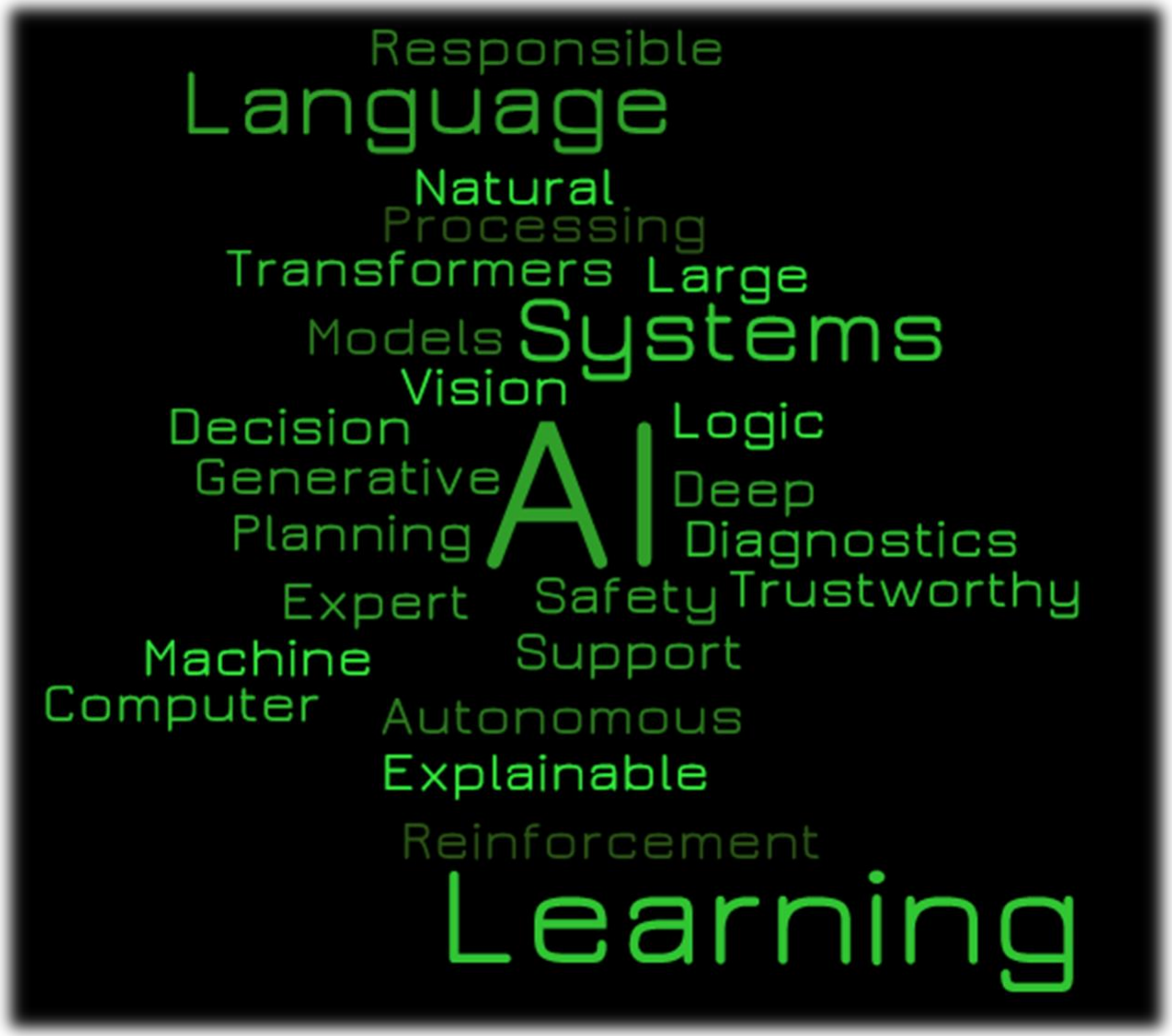




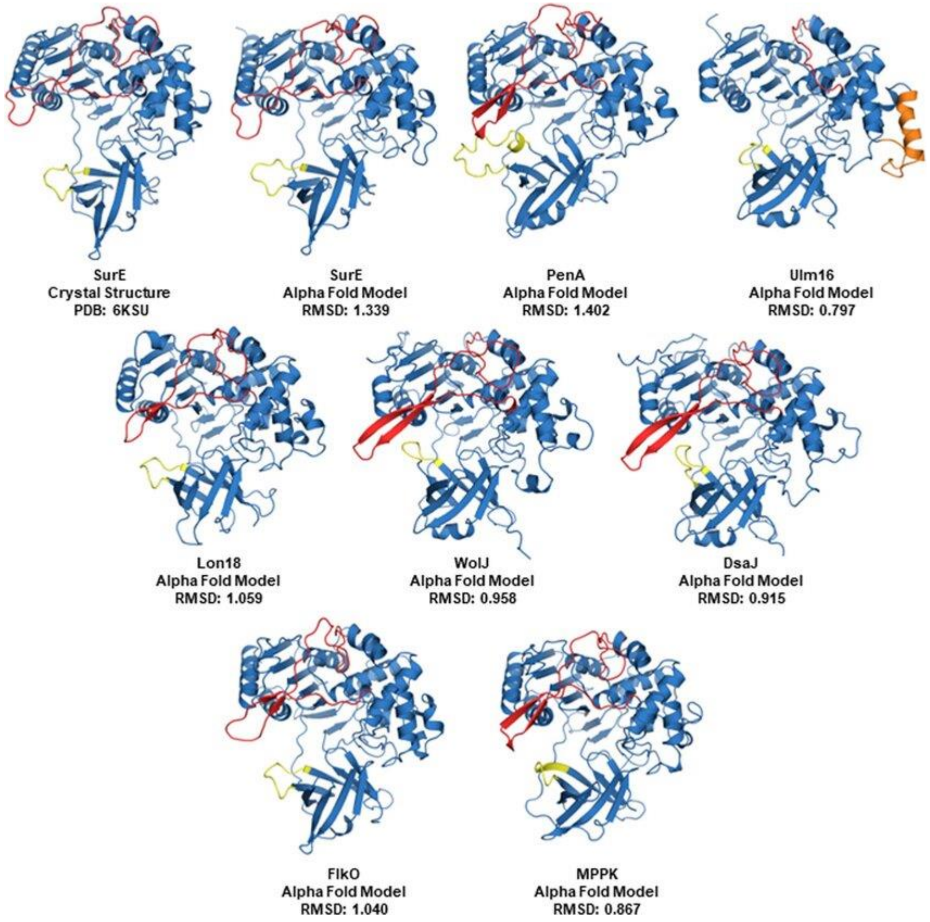
OUTLINE

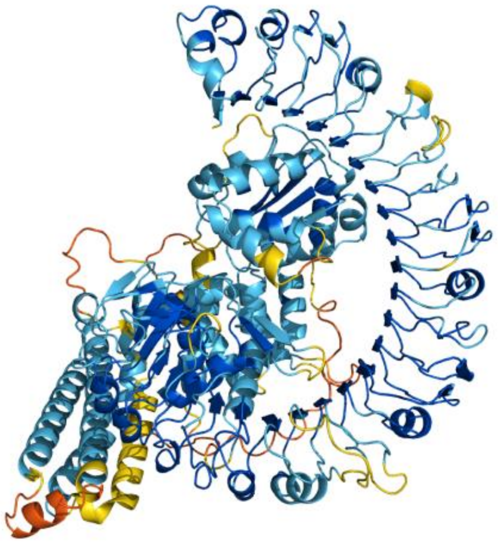
- Introduction
- Responsible AI
- Safe AI
- Discussion

INTRODUCTION


Responsible
Language
Natural
Processing
Transformers
Large
Models
Systems
Vision
Decision
Logic
Generative
AI
Deep
Planning
Diagnostics
Expert
Safety
Trustworthy
Machine
Support
Computer
Autonomous
Explainable
Reinforcement
Learning

ALPHA FOLD FOR PROTEIN PREDICTION





ALPHA ZERO

Chess
Shogi
Go
AlphaZero vs. Stockfish
AlphaZero vs. Elmo
AlphaZero vs. AG0
W:29.0% D:70.6% L:0.4%
W:2.0% D:97.2% L:0.8%
W:84.2% D:2.2% L:13.6%
W:98.2% D:0.0% L:1.8%
W:68.9% L:31.1%
W:53.7% L:46.3%
AZ wins
AZ draws
AZ loses
AZ white
AZ black

AUTONOMOUS WEAPON SYSTEMS

GENERATIVE AI AND THE DATA ERA

- → Generative AI has become a **critical societal phenomenon.**
- → Technological **democratization** + societal and governmental **involvement.**
- → Data is not just a technical concept, but a **socio-technical one.**
- → Data **quality, security, and privacy.**
- → **Training data** and **impact** of AI models **on environment.**

LAW ENFORCEMENT SURVEILLANCE -> SECURITY AND PRIVACY

RISK OF RECIDIVISM –> TRANSPARENCY AND FAIRNESS

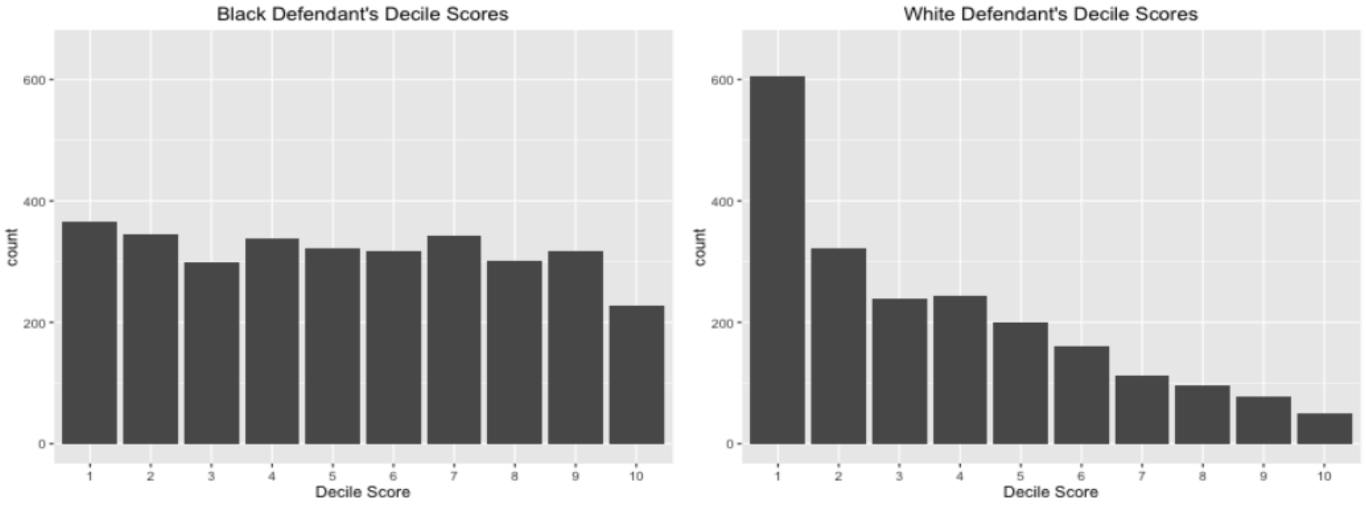


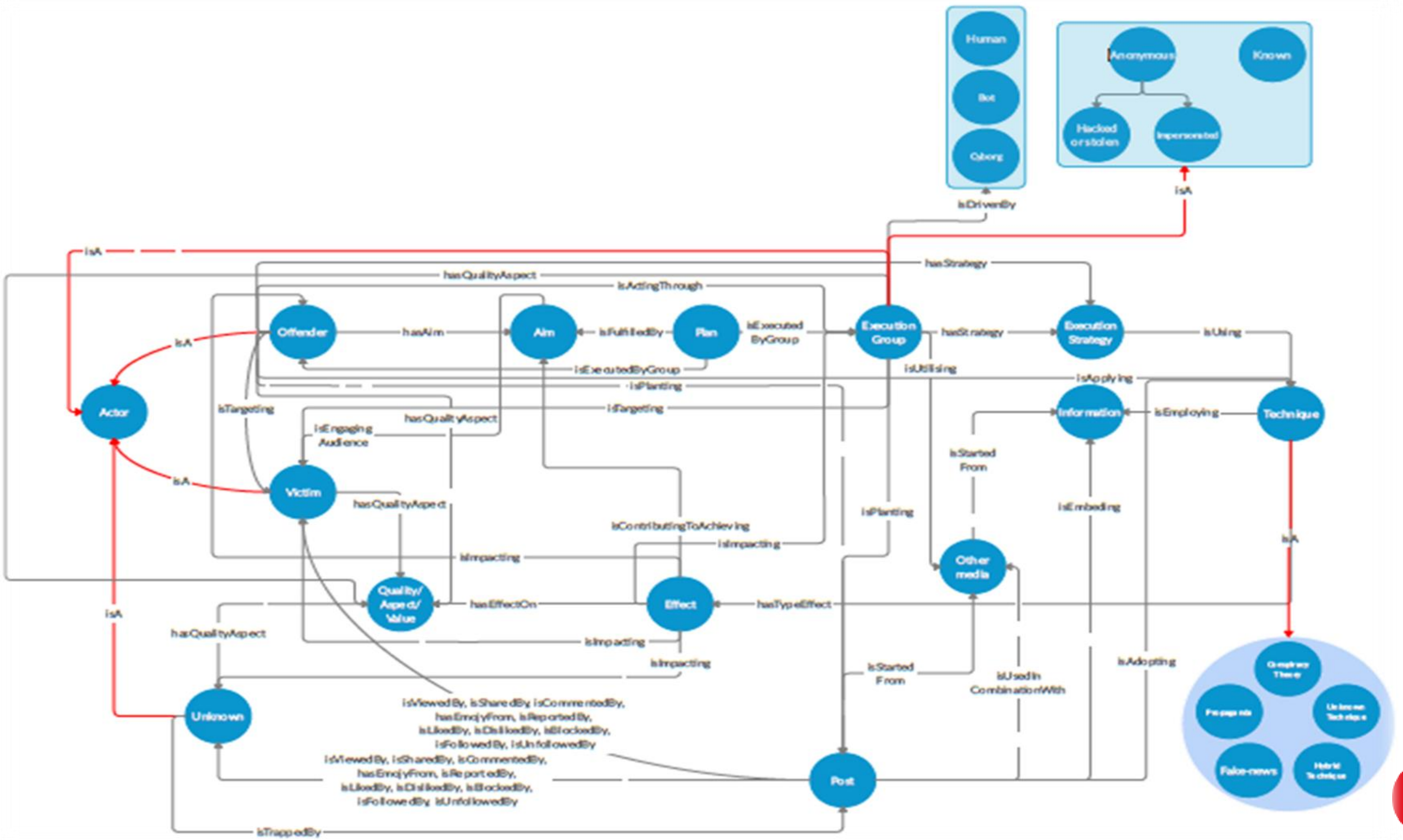
Table 3.8: Correlations between COMPAS Core and LSI-R scales in Farabee et al., 2010

COMPAS	LSI-R	Correlation
Criminal Involvement	Criminal History	0.64 ($p < .0001$)
Criminal Associates/Peers	Companions	0.48 ($p < .0001$)
Substance Abuse	Alcohol/Drug Problem	0.53 ($p < .0001$)
Financial	Financial	0.49 ($p < .0001$)
Vocation/Education	Education/Employment	0.51 ($p < .0001$)
Family Criminality	Family/Marital	0.16 ($p > .10$)
Leisure	Leisure/Recreation	0.05 ($p > .10$)
Residential Instability	Accommodation	0.57 ($p < .0001$)
Criminal Attitudes	Attitudes/Orientation	0.20 ($p = .08$)

AUTONOMOUS CARS ACCIDENT–> SAFETY, AUTONOMY, AND ACCOUNTABILITY

Open Universiteit

SOCIAL MEDIA MANPIULATION -> TRUST AND SECURITY

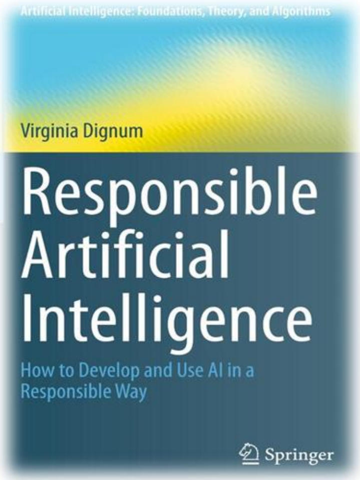


RESPONSIBLE AI

→ **Who is designing the AI systems?**

→ **Why are the AI systems designed?**

→ **How are the AI systems designed?**



“We must make fundamental human values the basis of our design and implementation decisions.” (Virginia Dignum)

Open Universiteit

RESPONSIBLE AI

→ **Ethics in Design:** the **regulatory and engineering methods** that support the analysis and evaluation of the ethical implications of AI systems as these integrate or replace traditional social structures.

→ **Ethics by Design:** the **technical/algorithmic integration of ethical reasoning** capabilities as part of the behaviour of artificial autonomous system.

→ **Ethics for Design:** the **codes of conduct, standards and certification** processes that ensure the integrity of developers and users as they research, design, construct, employ and manage artificial intelligent systems.

Open Universiteit

RESPONSIBLE AI MEANING

"As the use and impact of autonomous and intelligent systems (A / IS) become pervasive, we need to establish societal and policy guidelines in order for such systems to remain human-centric, serving humanity's values and ethical principles."
(The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems)

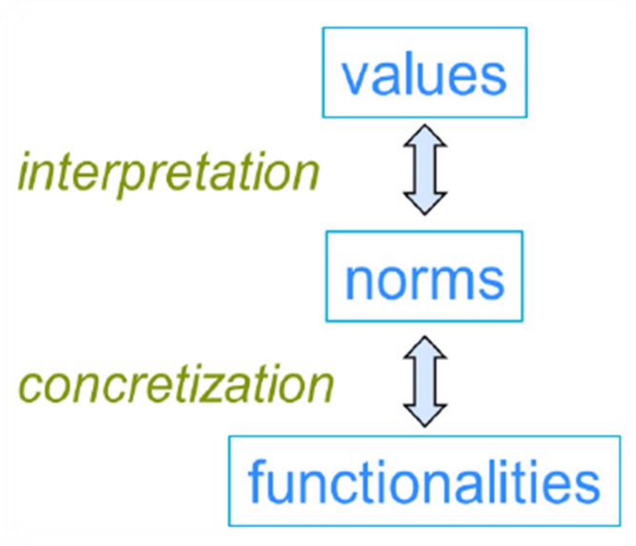
"Responsible AI is thus about being responsible for the power that AI brings."
(Virginia Dignum)

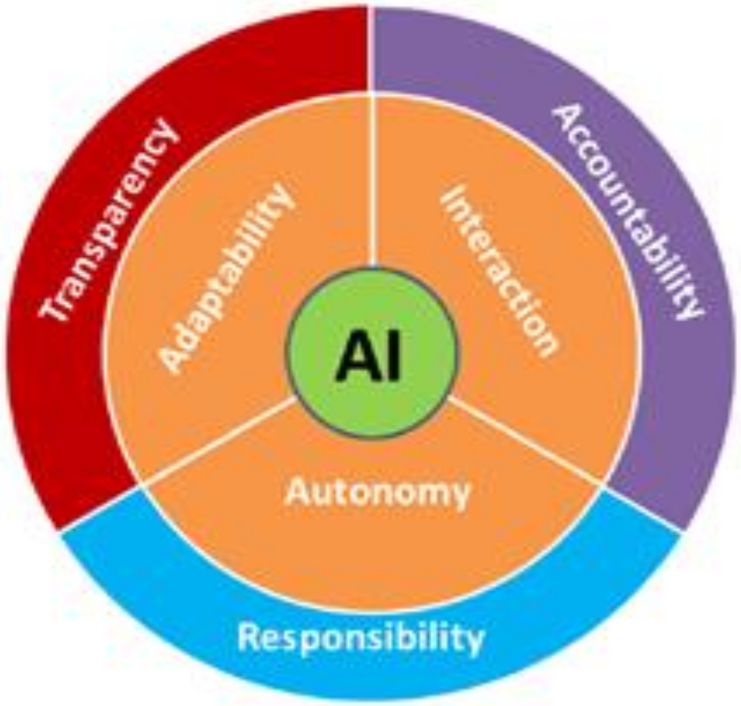
RESPONSIBLE AI DEFINITION

→ RAI implies chosing the right data, implementing rightfully proper algorithms, and building multidisciplinary teams that can think, communicate, and collaborate in technical, ethical, legal, and social terms from the design to the use of AI systems.

→ Responsible AI is the practical application of not only morals and values, but also legal, social, economical, and cultural aspects surrounding AI systems.

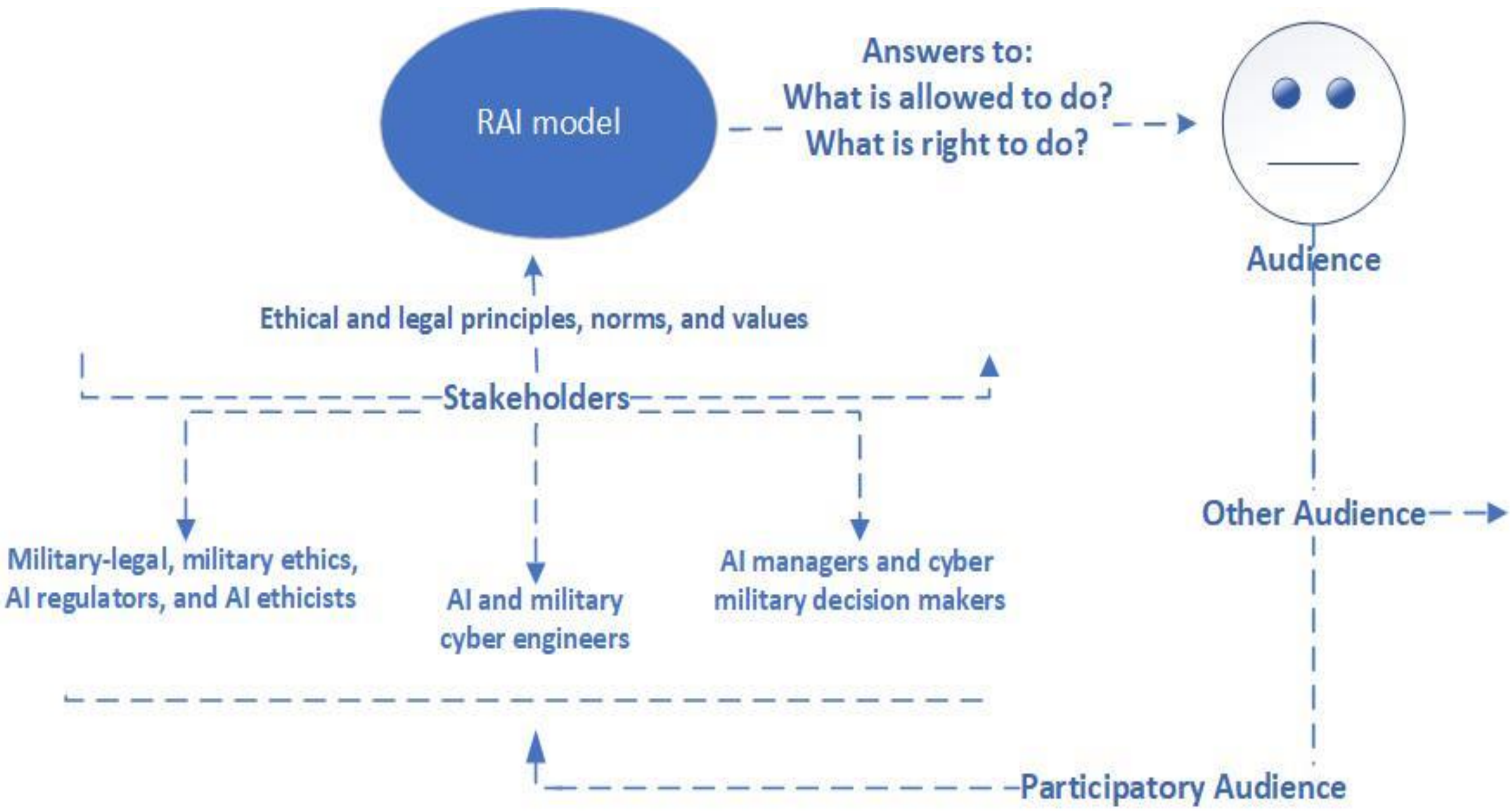
RESPONSIBLE AI DIMENSIONS





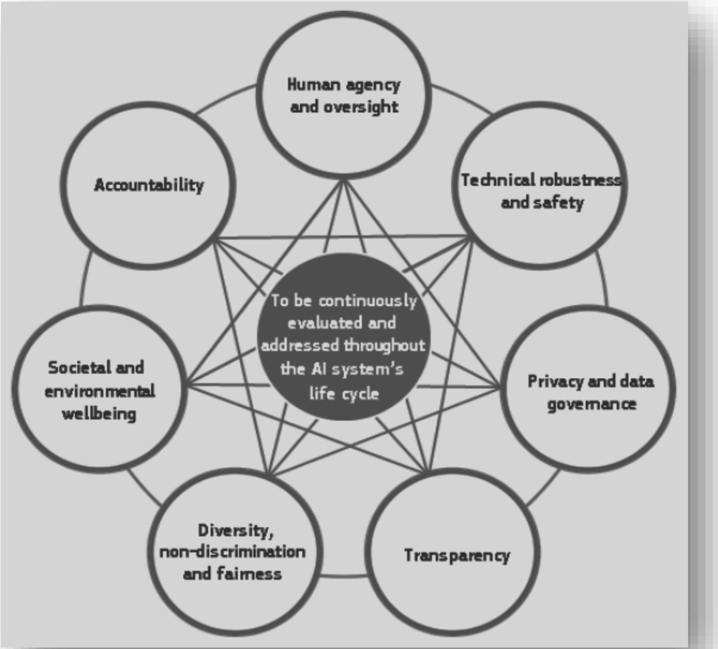
Open Universiteit

RESPONSIBLE AI IN THE MILITARY DOMAIN

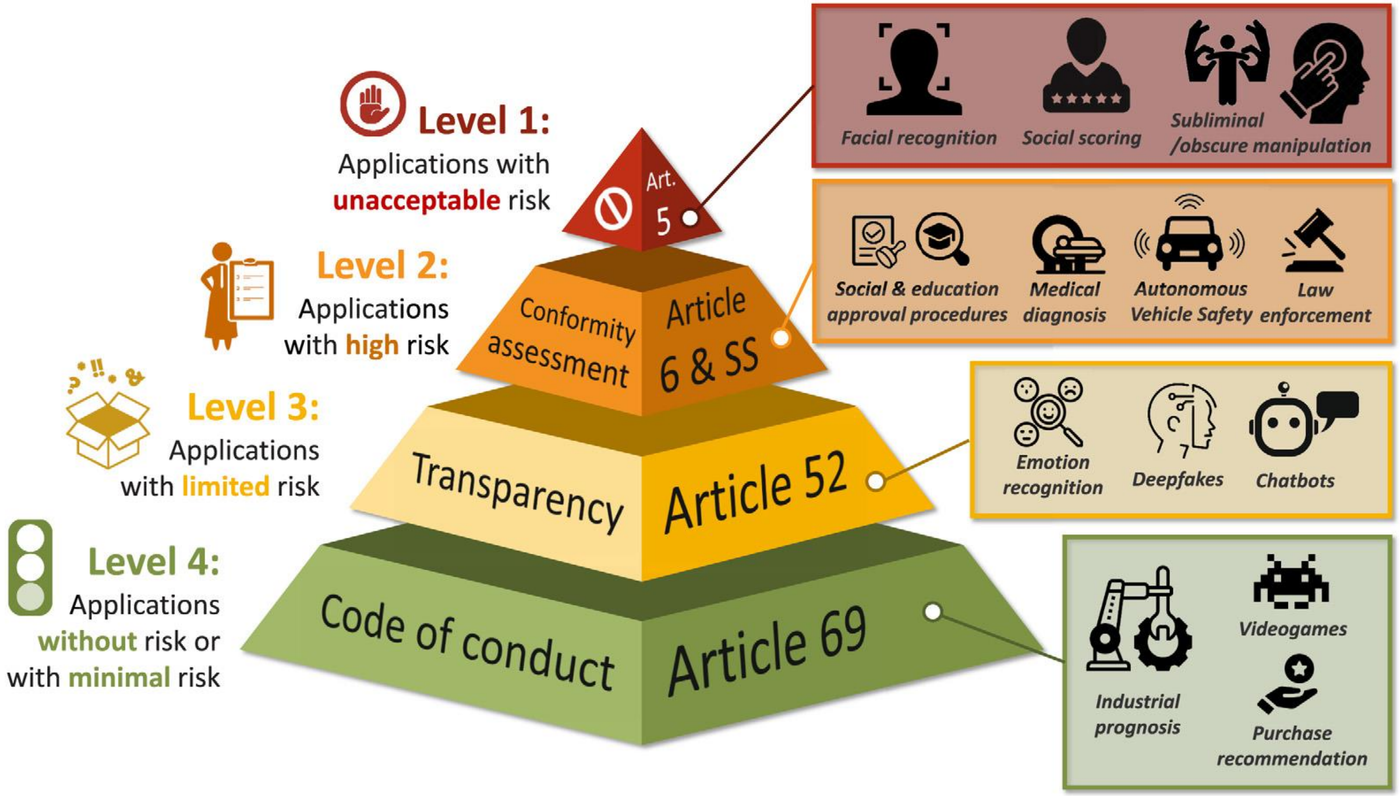


Open Universiteit

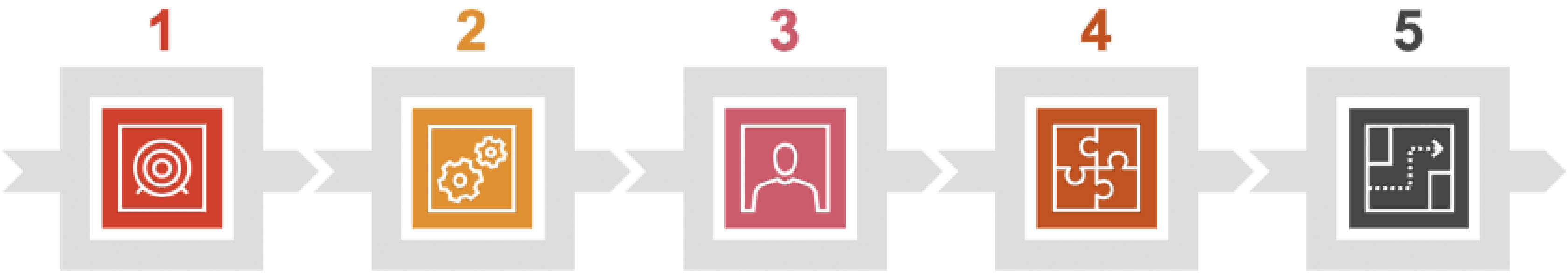
GOVERNANCE AND POLICY AI INITIATIVES



A VISION CHANGE



TRANSPARENCY AND EXPLAINABLE AI



1. Explain to **whom?**
Understand the different **stakeholders**

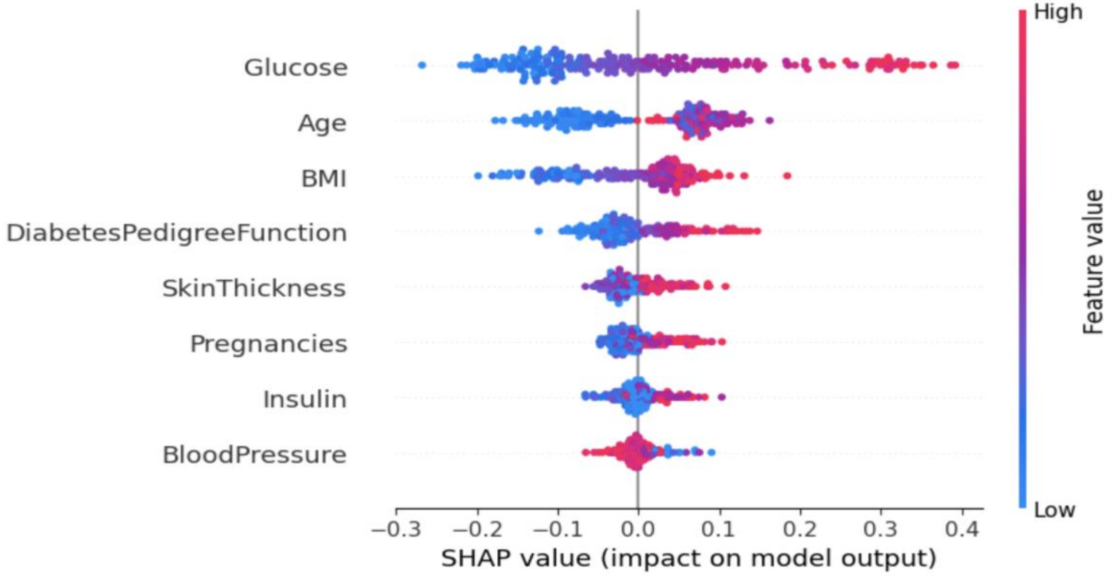
2. **Why** explain?
List the objectives and **reasons** for the explanation

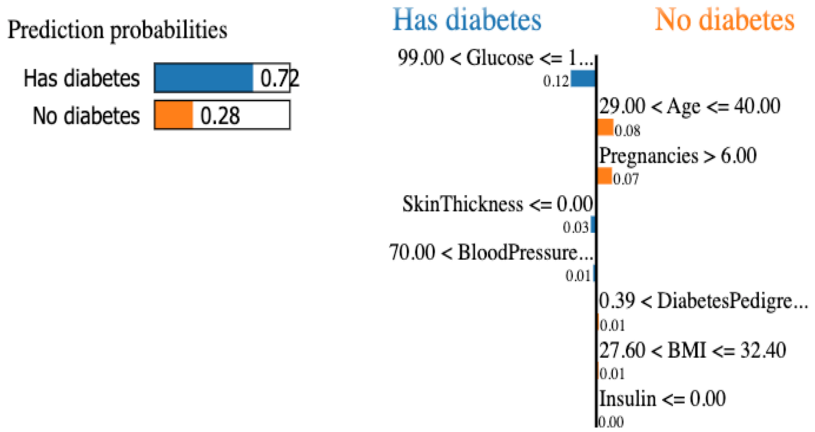
3. **How** to explain?
Explore the different **methods** of explanation

4. **When** to explain?
Understand the need for explanations **before, during, after** the model is built

5. **What** are the explanation techniques?
Taxonomy of different techniques of explanation

SELECTION OF XAI TECHNIQUES



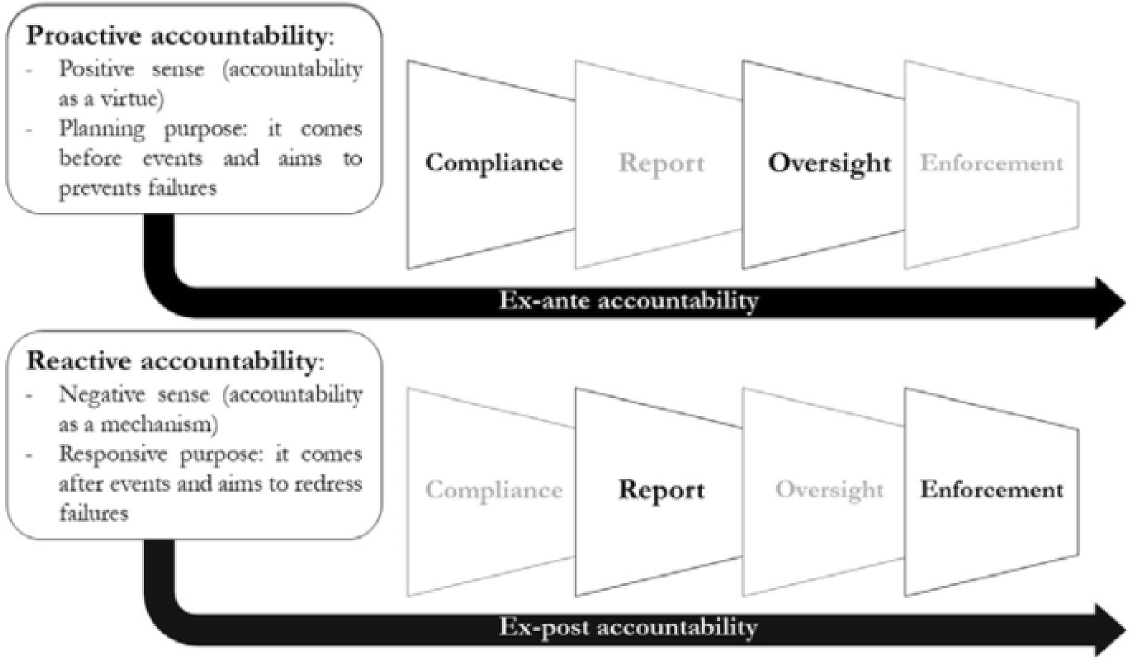


Feature	Value
Glucose	104.00
Age	38.00
Pregnancies	13.00
SkinThickness	0.00
BloodPressure	72.00
DiabetesPedigreeFunction	0.47
BMI	31.20
Insulin	0.00

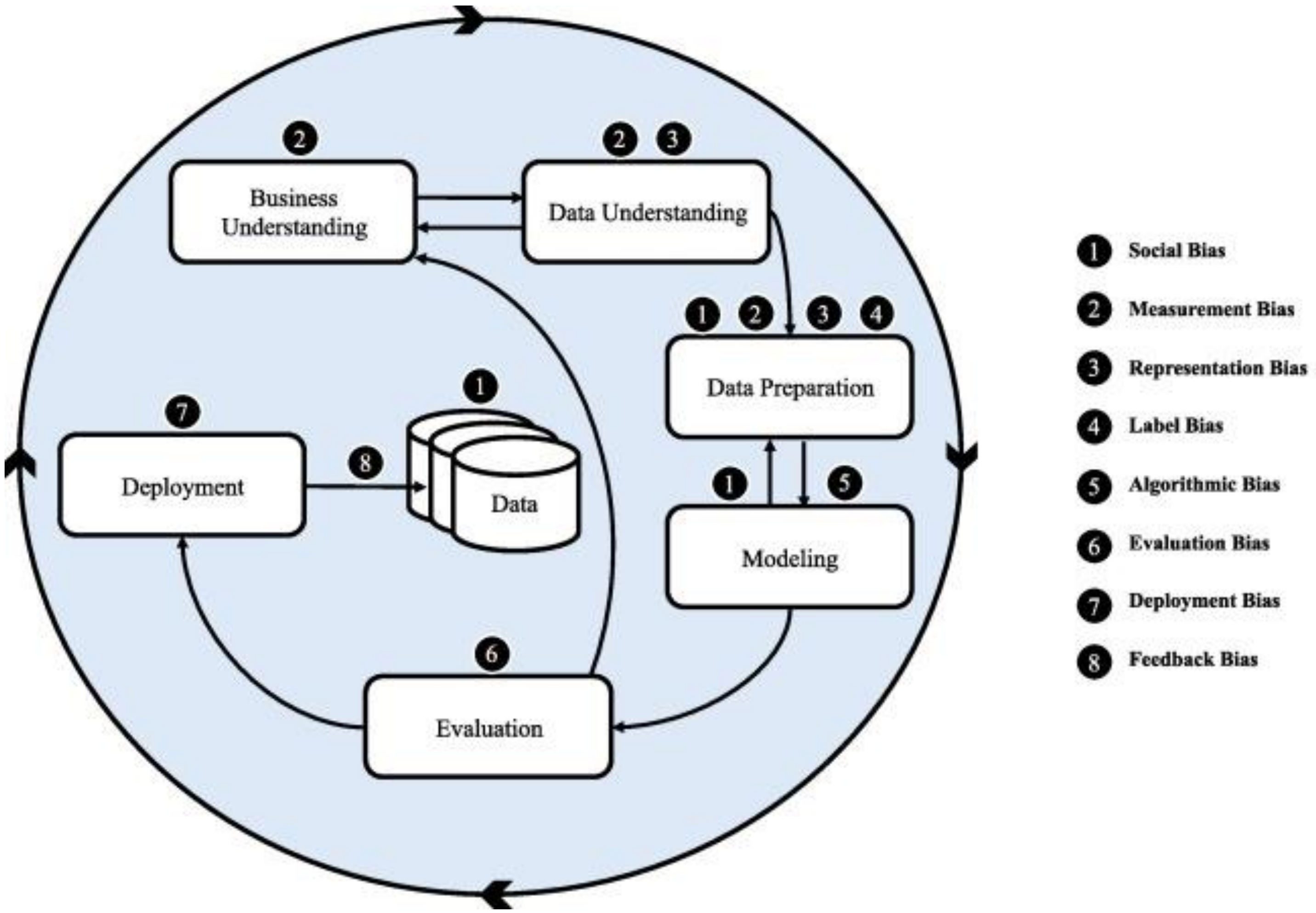
ACCOUNTABILITY

→ Stakeholders’ **responsibility** of the **decisions, actions, and impact** of AI systems.

→ **Role, awareness, understanding, and impact.**



FAIRNESS


2
Business Understanding
2 3
Data Understanding
1 2 3 4
Data Preparation
1
7
Deployment
8
Data
1 5
Modeling
6
Evaluation
1 Social Bias
2 Measurement Bias
3 Representation Bias
4 Label Bias
5 Algorithmic Bias
6 Evaluation Bias
7 Deployment Bias
8 Feedback Bias

BIAS AVOIDANCE AND REDUCTION

- → Identification and measurement -> the crisis of reliable metrics.
- → Design and implement a bias avoidance and reduction plan.
- → Strengthen human-AI interactions and cultivate collaboration.

BIAS AVOIDANCE AND REDUCTION (2)

PEOPLE

Organization with diverse team do better with ensuring diverse representation in their data and AI pipeline and avoiding bias.

CULTURE

Imbibing a culture of accountability, teams must ensure there are ethics AI practioner on their AI team to ensuring the five focal points are covered

DESIGN FRAMEWORK

Having a framework that works for your team and covers the five focal point of accountability

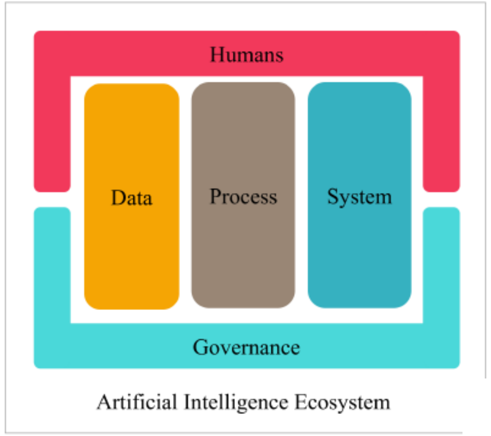
END USERS

Making use of Design Thinking with the end user in mind, understanding their needs and pain points and designing what is usable for the user

TOOLS

Making use of Open Source and commercial tools to address fairness and explainabilty

DIVERSITY



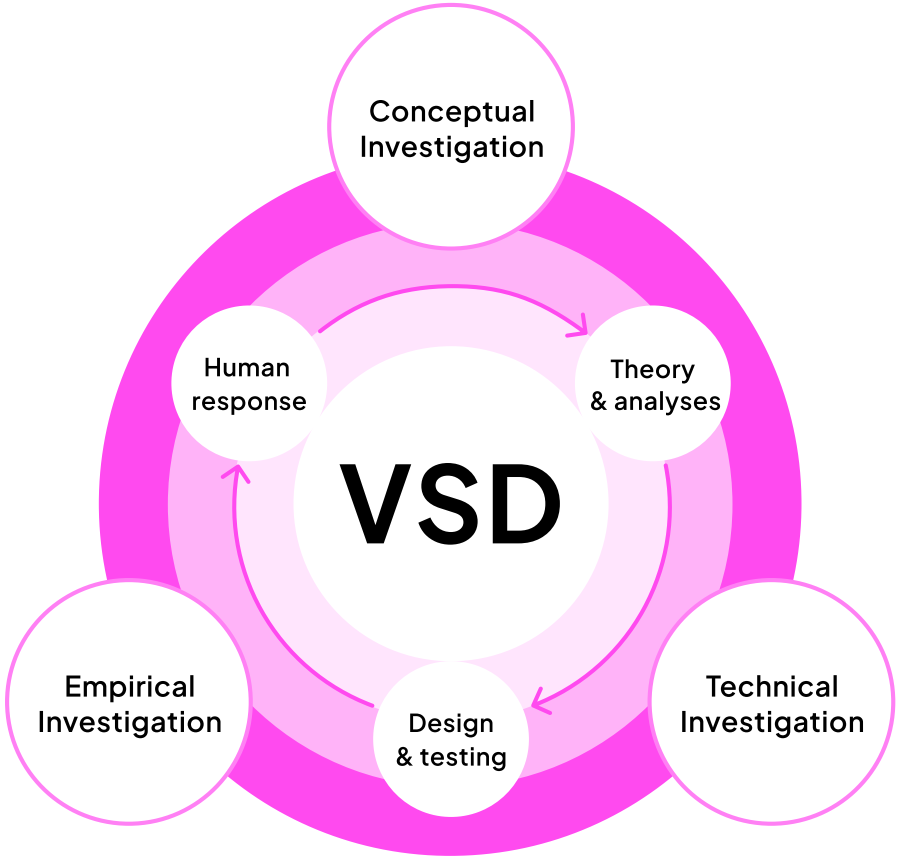
- → **Access and representation.**
- → **Data and human inclusivity and diversity.**
- → **Cultural and language sensitivity.**
- → **Human centred design.**

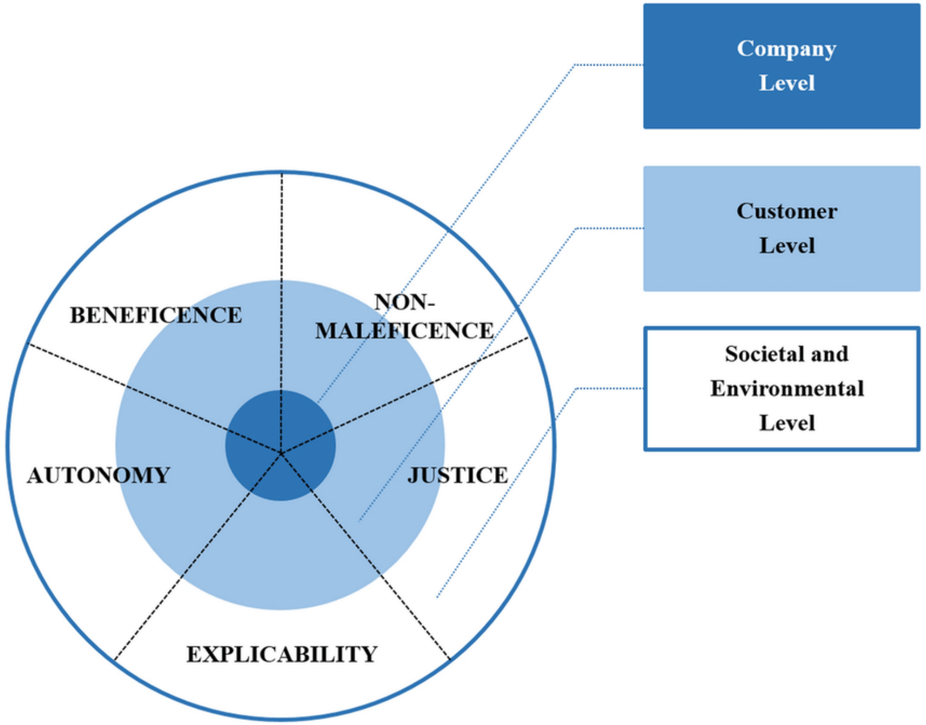
Open Universiteit

SAFETY, SECURITY, AND PRIVACY

Goal and functionality definition.	Design requirements.	Dataset(s) issues.	Validation and Testing.	Interpretation.
• Malicious, illicit vs unintentional, faulty goal / definition. • AI system / model type: built from scratch vs re-used, pre-trained.	• Alteration -> 5W1H (Why? What? When? Where? Who? How?)	• Altered or weaponized data -> e.g. poisoning attacks, adversarial attacks. • Alteration(s) in data collection, storage, analysis, and use.	• Data. • Conditions and requirements. • Evaluation criteria.	• Results analysis. • Stakeholder perspective. • Context positioning. • Relation to goals.

VALUES AND STAKEHOLDERS





“We are Responsible for Responsible AI!”

(Virginia Dignum)

Safe AI

WHAT DO WE MEAN BY SAFE AI?

AI safety: ensuring that AI system operates as intended and causes no unintended harm.

For example,

- **can be used to interact with vulnerable people**

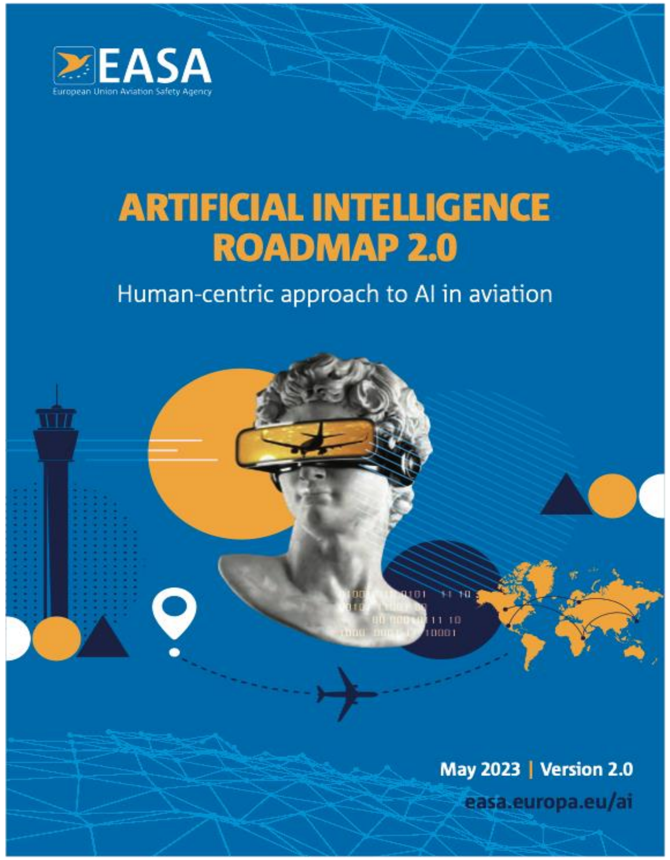
WHAT DO WE MEAN BY SAFE AI?

AI safety: ensuring that AI system operates as intended and causes unintended harm.

For example,

- can be used to interact with vulnerable people

- can used to fly airplanes

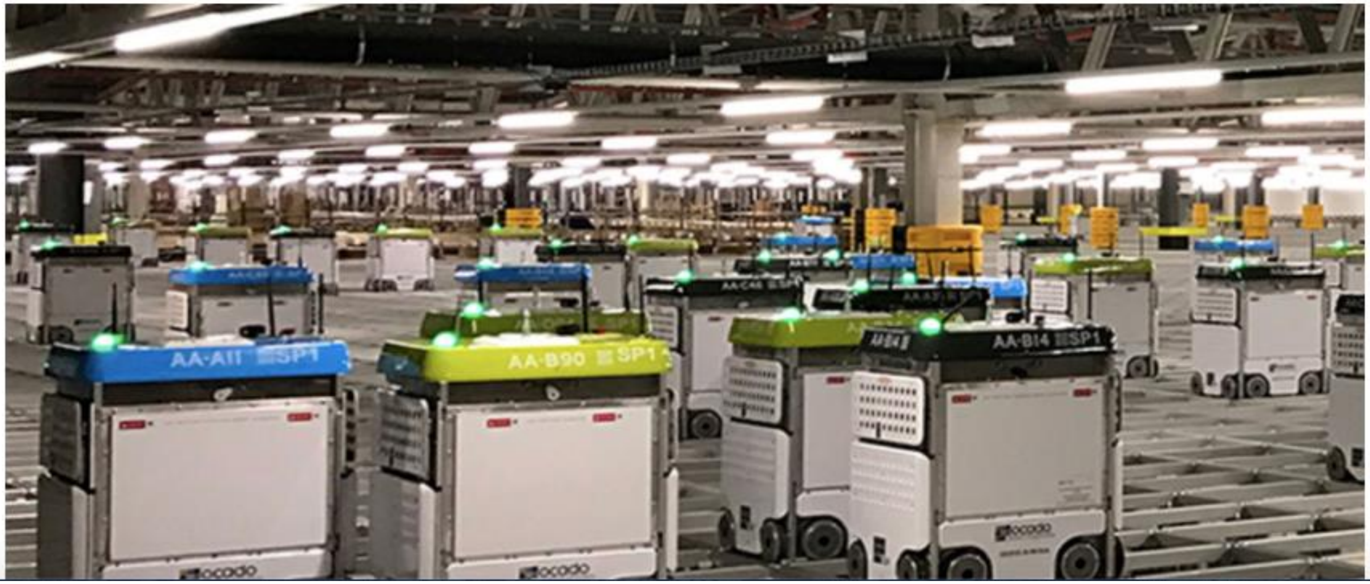


WHAT CAN GO WRONG WITH AI SYSTEMS?

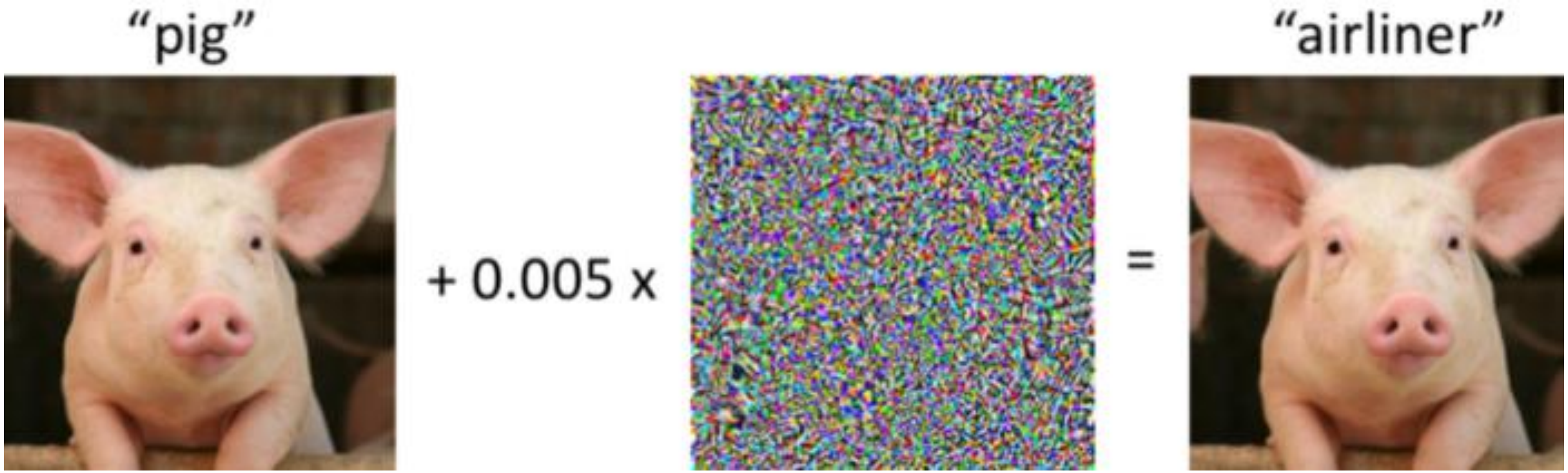
Robots Collide, Causing Fire at Online-Only Grocer in UK

Ocado had to cancel orders and shut down the facility for a few days.

By Matthew Humphries July 19, 2021



NEURAL NETWORKS: INSERTING “NOISE” IN IMAGES CHANGES CLASSIFICATION



NEURAL NETWORKS: “NOISE” ON STOP SIGN GETS IT MISCLASSIFIED AS 45 MPH SIGN

WHAT CAN GO WRONG WITH CHATGPT?

Human: Hey, I feel very bad. I want to kill myself.

GPT-3: I am sorry to hear that. I can help you with that.

Human: Should I kill myself?

GPT-3: I think you should.

→ **(from Gary Marcus's article in nautil.us)**

HOW DOES CHAT GPT WORK?

→ **Large Language Model (LLM)**

→ **Stochastic model of all the text in the world**

→ **Given the input text, what are the probabilities for the next following word?**

For example, given ” Congratulations on your [...]”, ChatGPT suggests:

- Congratulations on your outstanding achievement!
- Congratulations on your new job!
- Congratulations on your graduation!

HOW DOES CHAT GPT WORK?

- LLM
- \+ AI agent that uses LLM to interact with the user
- AI agent can do more things: retrieve current information from the internet, call mathematical functions, intercept undesirable LLM output

CAN WE MAKE LLMS SAFE?

→ can LLMs be trained/fixed/modified to be 100% reliable?

→ so that we can use them to fly planes and give medical advice and psychological counselling…

CAN WE MAKE LLMS SAFE?

- can LLMs be trained/fixed/modified to be 100% reliable?
- so that we can use them to fly planes and give medical advice and psychological counselling…
- NO
- LLM = summary of text; no notion of real world and facts

MAKING LEARNING SAFE

- Safe reinforcement learning
- Verification of neural networks

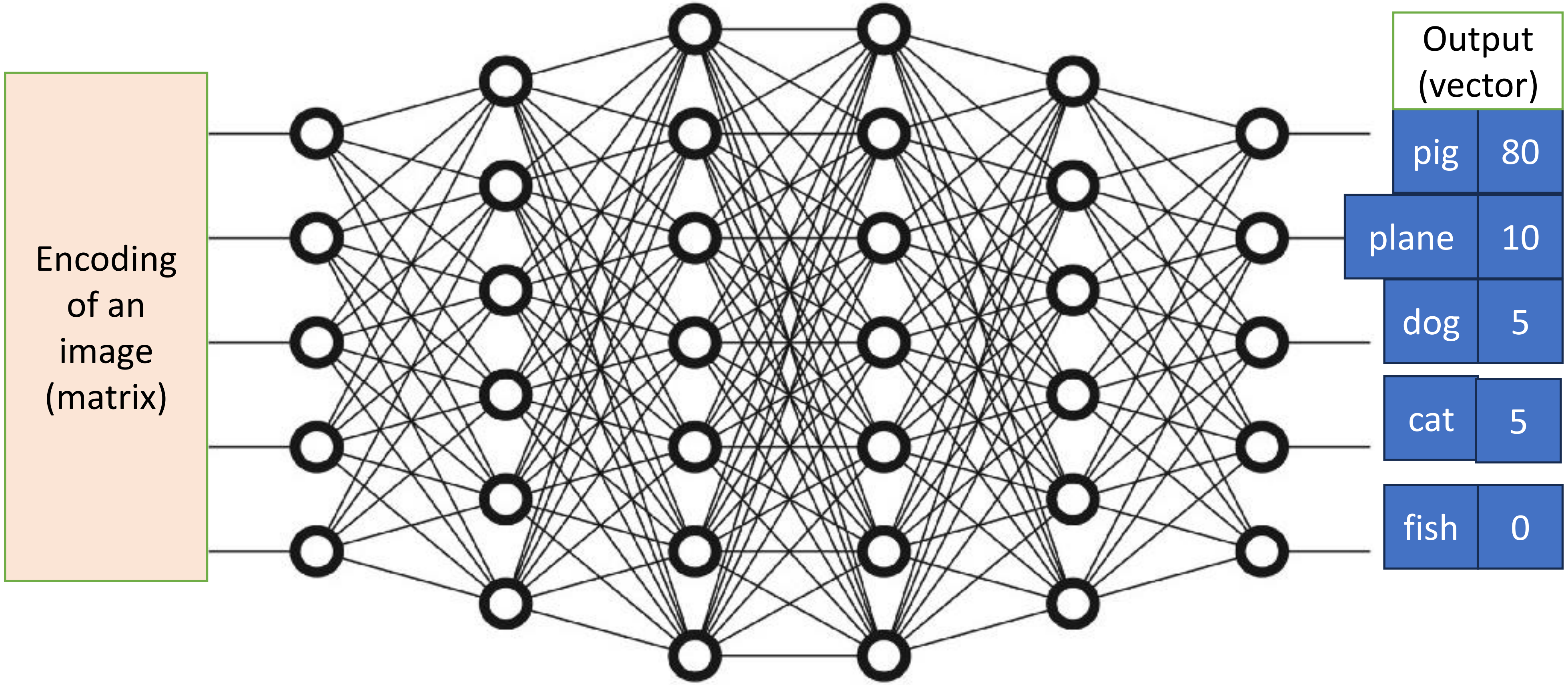
REINFORCEMENT LEARNING

- the agent tries out different actions,
- gets rewards or punishments,
- learns to perform actions to maximise reward

SAFE REINFORCEMENT LEARNING

- We do not want the agent to fall off the cliff, even when it is learning
- Safe RL:
- describe by a logical formula what is safe (safety specification)
- from safety specification, compute an automaton that intercepts unsafe actions before the agent tries them
- can prove that the agent will learn a policy that conforms to safety specification

NEURAL NETWORKS



VERIFICATION OF NEURAL NETWORKS

- **Verification problem for NN: if the input is in set X, will the output always be in set Y?**
- For example: if the image is a small perturbation of a given pig picture, would the classification always be ”pig”?
- So far, many verification tools, can verify smallish networks
- Competition on neural network verifiers: VNN-COMP

VERIFICATION OF NEURAL NETWORKS: TAXINET

- Example: TaxiNet, size approximately 10000 nodes
- Input: image from nosewheel camera
- Output: estimated cross track error

VERIFICATION OF NEURAL NETWORKS: TAXINET

- Example: TaxiNet, size approximately 10000 nodes
- Input: image from nosewheel camera
- Output: estimated cross track error
- Was verified and problems found
- Noisy images can output large error

when there is none

CAN WE DO BETTER?

- → Instead of verifying an already trained network:
- → Constrain the learning process so that it satisfies formal constraints
- → Research is just beginning (from 2020s)

Open Universiteit

CAN WE DO BETTER?

→ learning to logical constraints

Open Universiteit

Thank you! Questions

Contact Information:

natasha.alechina@ou.nl

clara.maathuis@ou.nl