Hi

Universiteit Leiden

TU Delft

Catholijn Jonker

and reseach collaborators

Self-Reflective Hybrid Intelligence

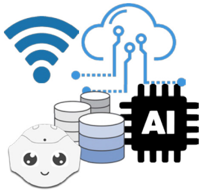
Combining Human with Artificial Intelligence and Logic

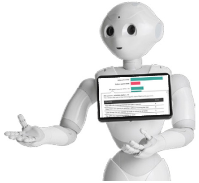
Symbiosis: Partners

Universiteit Leiden

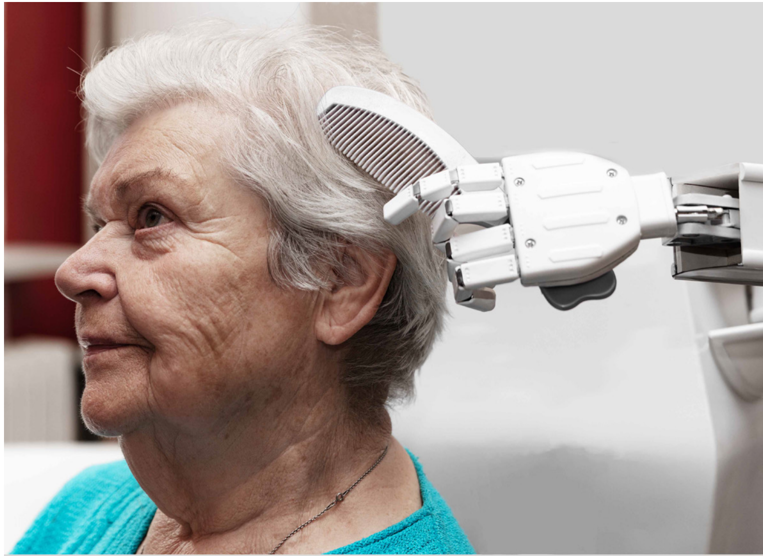
TU Delft

“mutualistic” human-AI symbiosis





- **Trustworthy and responsible AI**
  - Value sensitive data processing and decision-making
  - Inclusive (with, about and for whom)
- **Transparant & explainable AI**
  - Even with a “black box”?
  - Understandable & trust inducing
- **Human-AI co-development**
  - Exchanging knowledge
  - Human adapts AI wrt new insights
  - and vice versa



Universiteit Leiden

TUDelft

Hybrid
Intelligence

Align AI with human needs and values

The future of humankind with Artificial Intelligence (AI)

- Enhanced autonomy
- Enriched e ences
- New ties
- Strengthened democracies

Promote

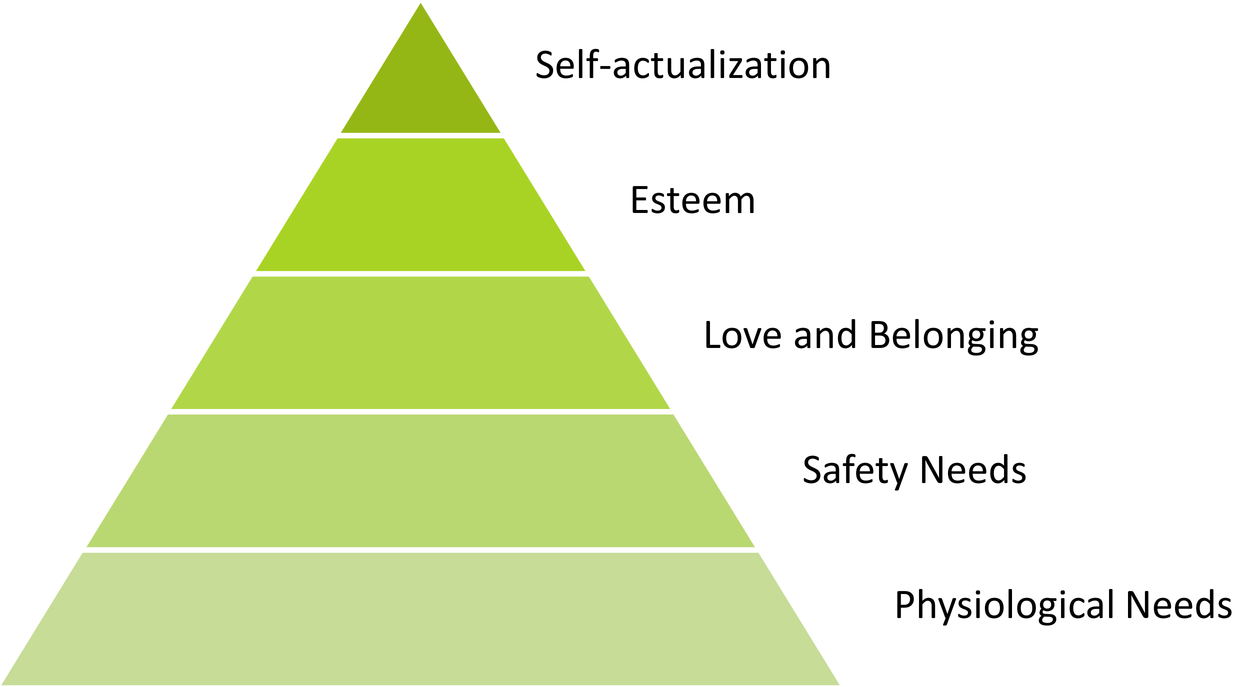
- Reduced autonomy
- Replaced e nces
- Redu
- Threatening democracies

Intervene

Universiteit Leiden

TU Delft

Basic Human Requirements



Self-actualization

Esteem

Love and Belonging

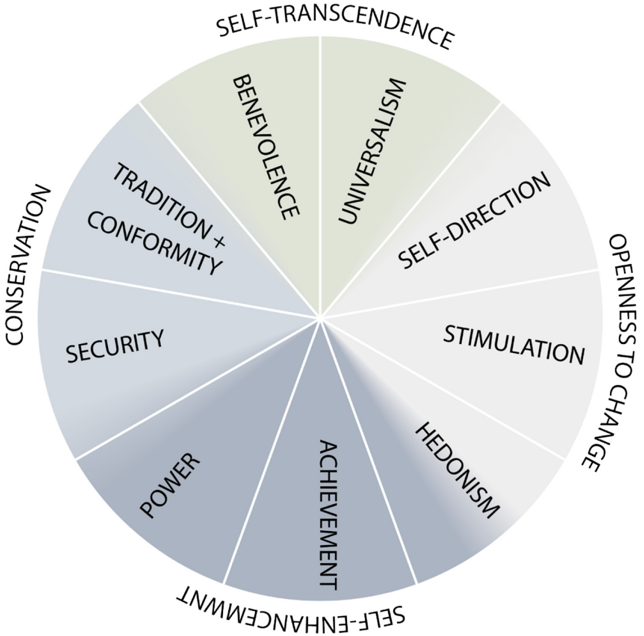
Safety Needs

Physiological Needs

A. H. Maslow (1943), A Theory of Human Motivation, *Psychological Review*, 50, 370-396.

Basic Human Values

Schwartz Values



Schwartz, Shalom H. (1992), "Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries", *Advances in Experimental Social Psychology,* 25: 1–65

Moral Foundation Theory
Care/Harm
Fairness/Cheating
Loyalty/Betrayal
Authority/Subversion
Purity/Degradation

Haidt, J.; C. Joseph (Fall 2004). "Intuitive ethics: how innately prepared intuitions generate culturally variable virtues". *Daedalus*. 133 (4): 55–66.
Graham, J.; et al. (2013). Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. *Advances in Experimental Social Psychology*, 47: 55–130.

Towards alignment of AI with Human Values

Universiteit Leiden

TU Delft

Features of Values

- Values refer to goals;
- Value beliefs are linked to affect;
- Value are standards of criteria;
- Values are ordered by importance;
- Value priorities guide actions;
- Values transcend contexts

Universiteit Leiden

TU Delft

Moral Foundations Twitter Corpus (MFTC)

- 35k tweets divided in 7 datasets,
- annotated with the MFT values

MFTC Datasets

All Lives Matter
Baltimore Protests
Black Lives Matter
Hate Speech
2016 US Elections
MeToo Movement
Hurricane Sandy

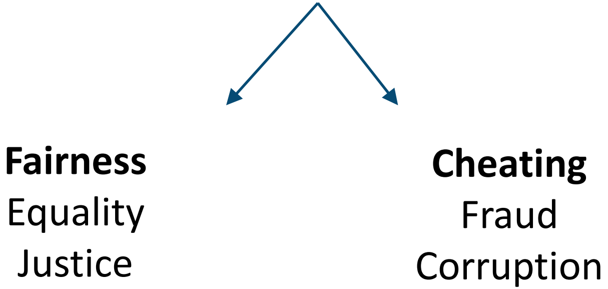
MFT Values

Care/Harm
Fairness/Cheating
Loyalty/Betrayal
Authority/Subversion
Purity/Degradation

Hoover, Joe, et al. "Moral Foundations Twitter Corpus: A collection of 35k tweets annotated for moral sentiment." *Social Psychological and Personality Science* 11.8 (2020): 1057-1071.

Value Rhetoric Similarities

ALM and **BLM** generally have similar value rhetoric:



Fairness
Equality
Justice

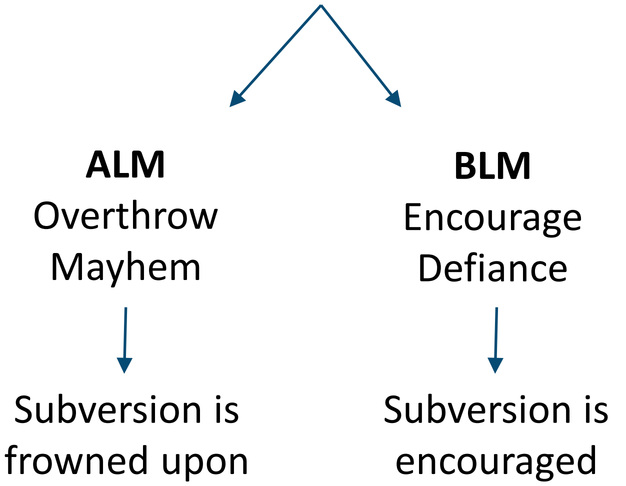
Cheating
Fraud
Corruption

Value Rhetoric Similarities

ALM and **BLM** generally have similar value rhetoric,
but they differ for the value of *subversion*

Value Rhetoric Similarities

ALM and **BLM** generally have similar value rhetoric, but they differ for the value of *subversion*



ALM
Overthrow Mayhem

Subversion is frowned upon

BLM
Encourage Defiance

Subversion is encouraged

~~Values transcend contexts.~~

Value expressions are context dependent

→

Aligning AI with Human Needs & Values is a context dependent task

Universiteit Leiden

TU Delft

Basic Human Values

- General and abstract
- Applicable across contexts
- Suitable for societal questions

Universiteit Leiden

TU Delft

Basic Human Values	Context-Specific Values
General and abstract	Applicable to a context
Applicable across contexts	Defined within a context
Suitable for societal questions	Suitable for concrete usage

Universiteit Leiden

TU Delft

Towards context specific alignment of AI with Human Values

How to identify these values?

Universiteit Leiden

TU Delft

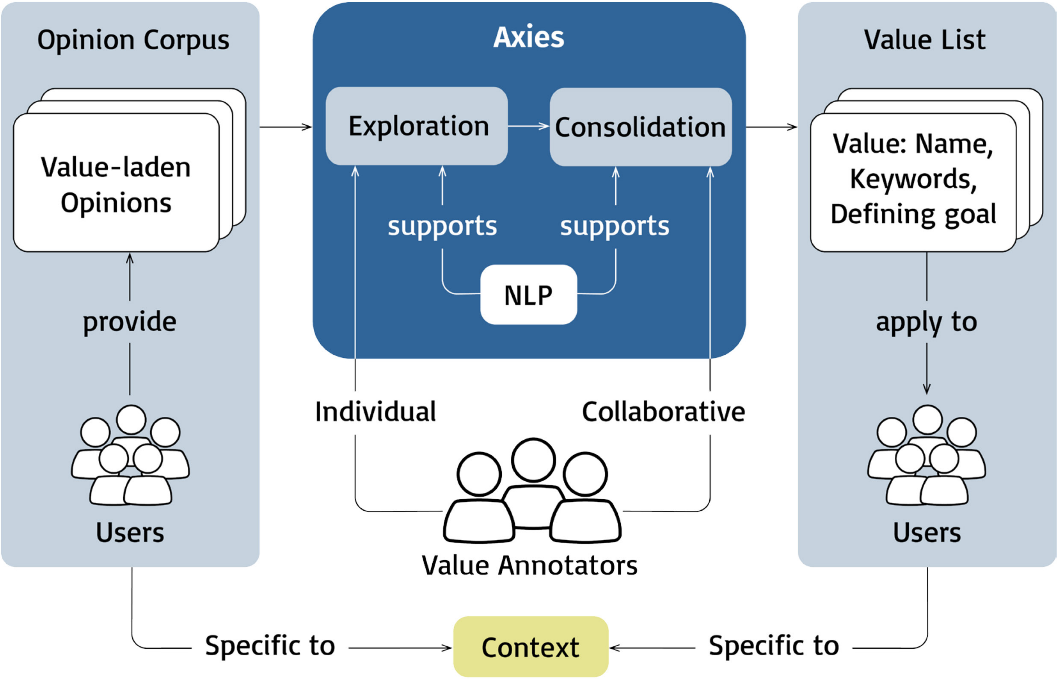
Axies methodology

Axies is a *hybrid* (human+AI) methodology for identifying context-specific values, with the support of NLP techniques.

Axies simplifies and distributes the value identification process.

E. Liscio, M. van der Meer, L. C. Siebert, C. M. Jonker, and P. K. Murukannaiah. “What values should an agent align with?”. In: *JAAMAS*, 36, 23, 2022.

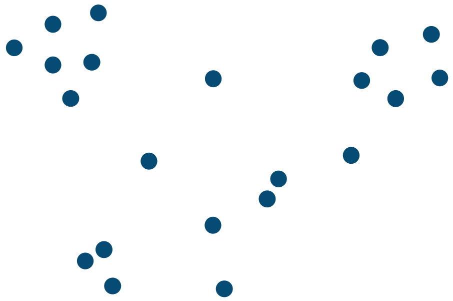
Axies methodology



Axies - Exploration

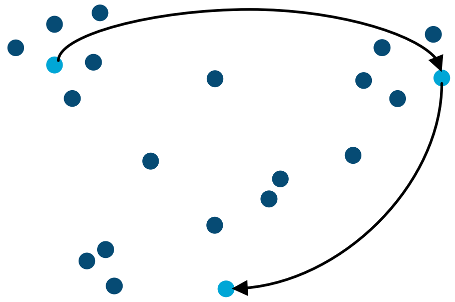
In the exploration phase, each annotator independently develops a value list.

The next opinion to be analysed is the most different from the already analysed opinions.



Axies - Exploration

In the exploration phase, each annotator independently develops a value list.

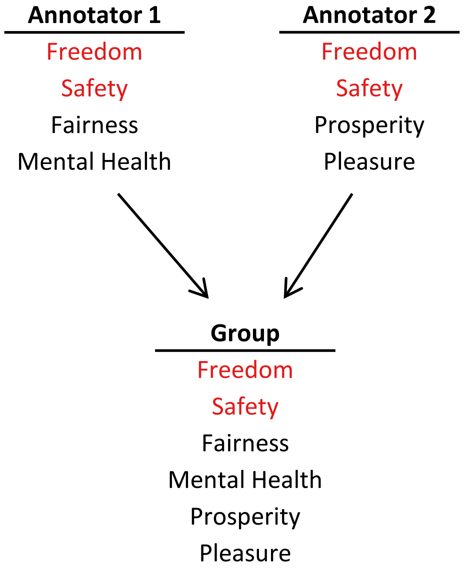


The next opinion to be analysed is the most different from the already analysed opinions.

Axies - Consolidation

The annotators in a group collaborate to merge their individual value lists.

Axies guides the annotators through the process via NLP moderation.



Annotator 1

Freedom
Safety
Fairness
Mental Health

Annotator 2

Freedom
Safety
Prosperity
Pleasure

Group

Freedom
Safety
Fairness
Mental Health
Prosperity
Pleasure

Evaluation

We perform Axies on two survey datasets:

COVID-19 (60,000 answers)

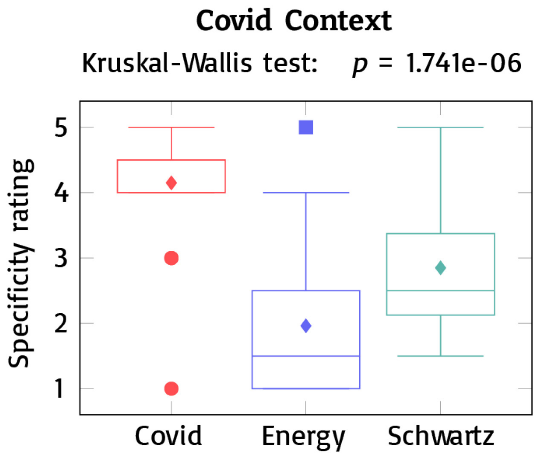
Green Energy Transition (3,000 answers)

Research questions

- Does Axies yield context-specific values?
- What are the differences between Axies and basic values?

Results - Specificity

Axies yields values that are more context-specific than basic (Schwartz) values.

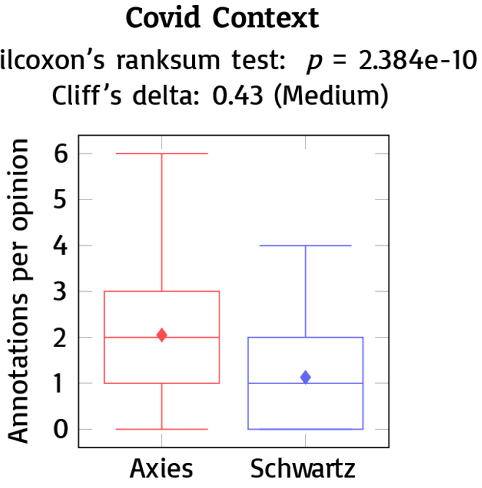


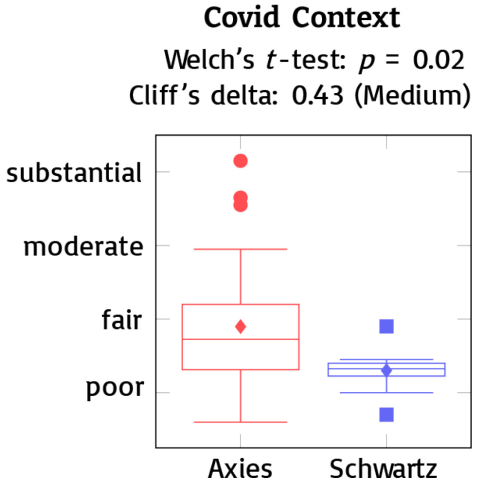
Universiteit Leiden

TU Delft

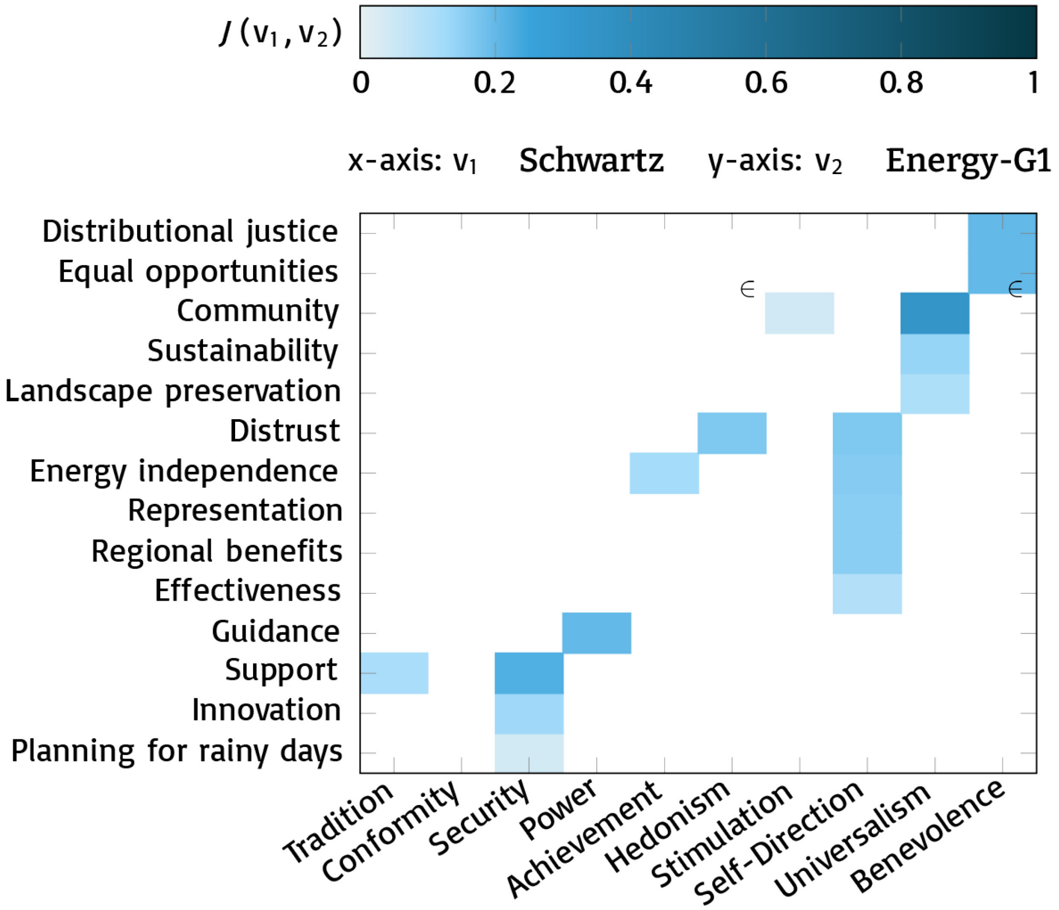
Results - Application

Laypeople annotate Axies values more often and with higher agreement. This shows the suitability of context-specific values for practical applications.





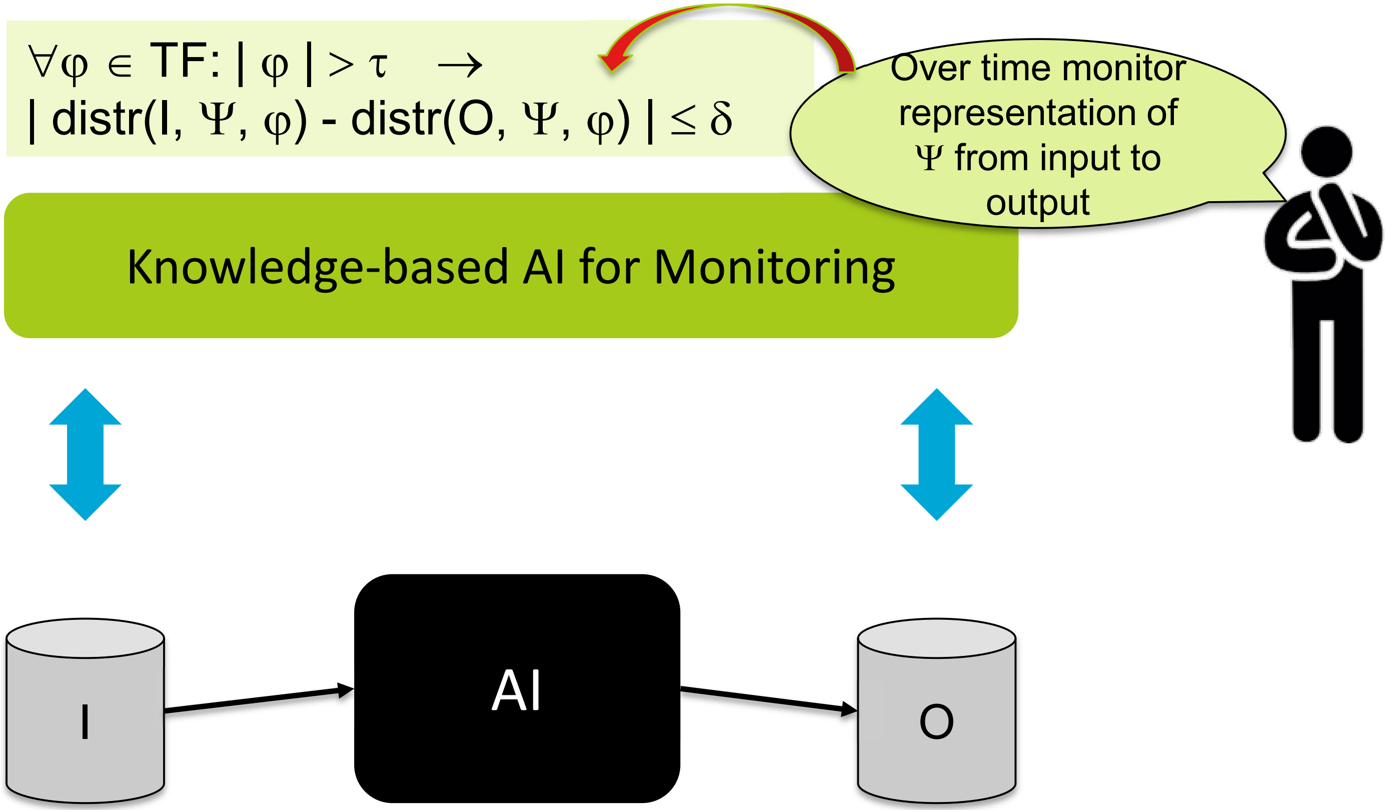
Results - Relationship



Having context specific Human Values: Now how to align AI with them?

Formalize the values

Use Knowledge Technology to help Humans to Monitor AI



$\forall\varphi \in \mathrm{TF}: |\,\varphi\,| > \tau \quad \rightarrow$
$|\,\mathrm{distr}(I, \Psi, \varphi) - \mathrm{distr}(O, \Psi, \varphi)\,| \le \delta$

Over time monitor representation of Ψ from input to output

Knowledge-based AI for Monitoring

I

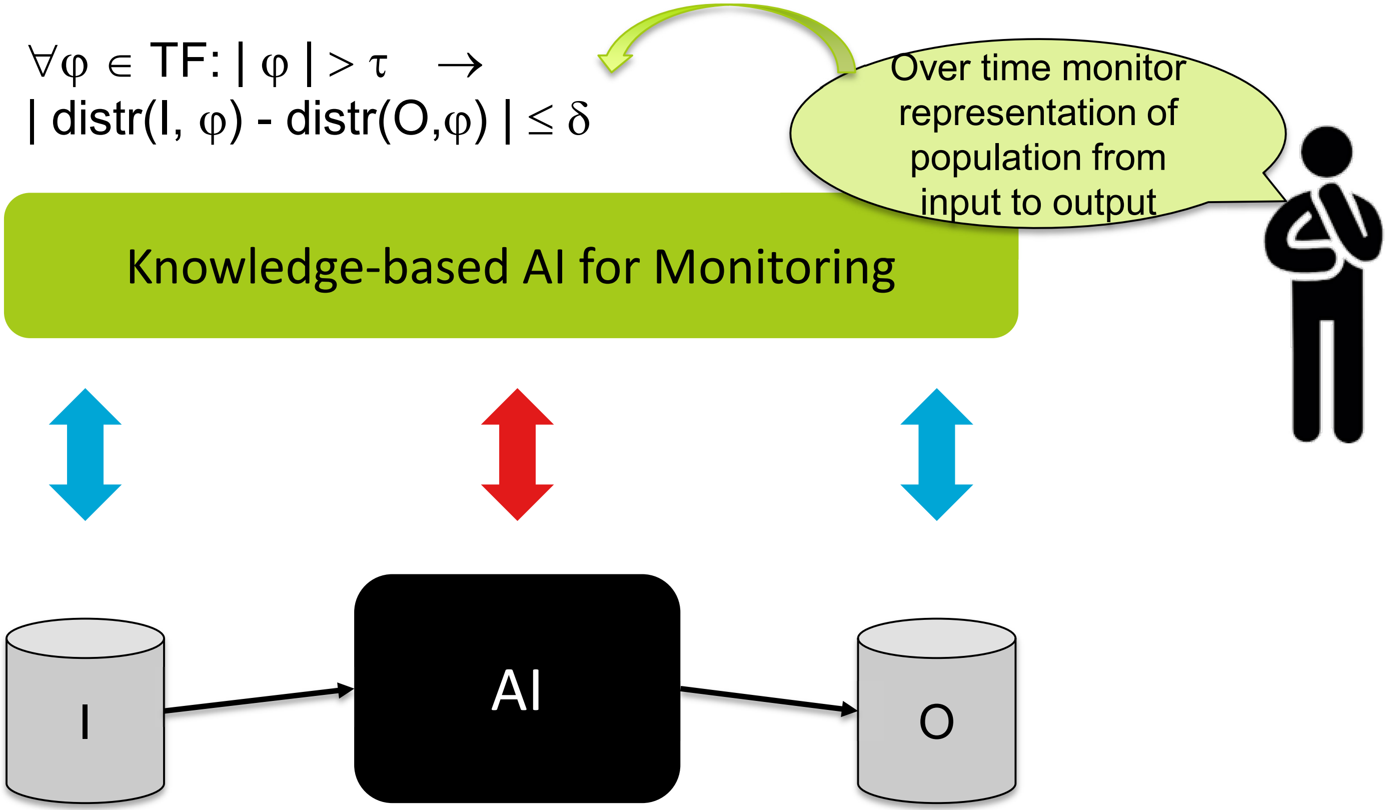
AI

O

Universiteit Leiden

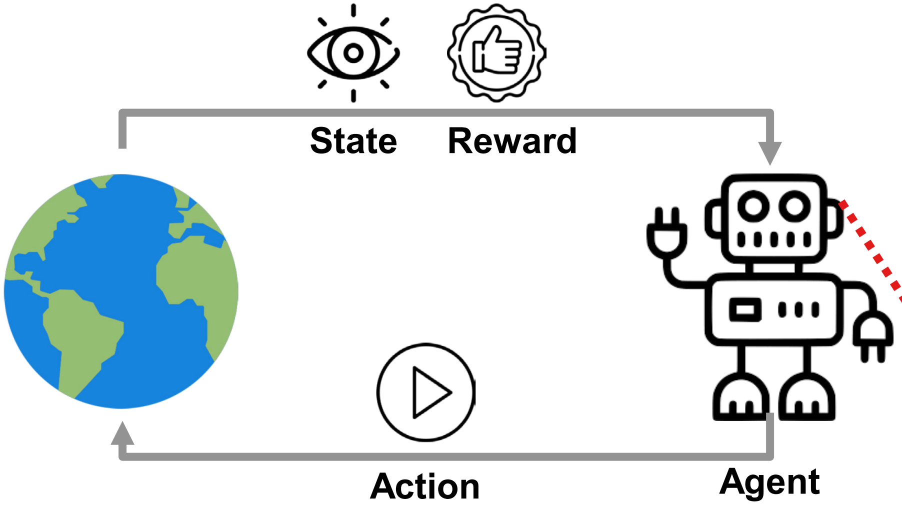
TUDelft

$\forall \varphi \in$ TF: $| \varphi | > \tau \rightarrow$
$|$ distr(I, φ) - distr(O,φ) $| \leq \delta$


Over time monitor representation of population from input to output
Knowledge-based AI for Monitoring
I
AI
O

Deep Reinforcement Learning

Performant black-box policies containing meta-information

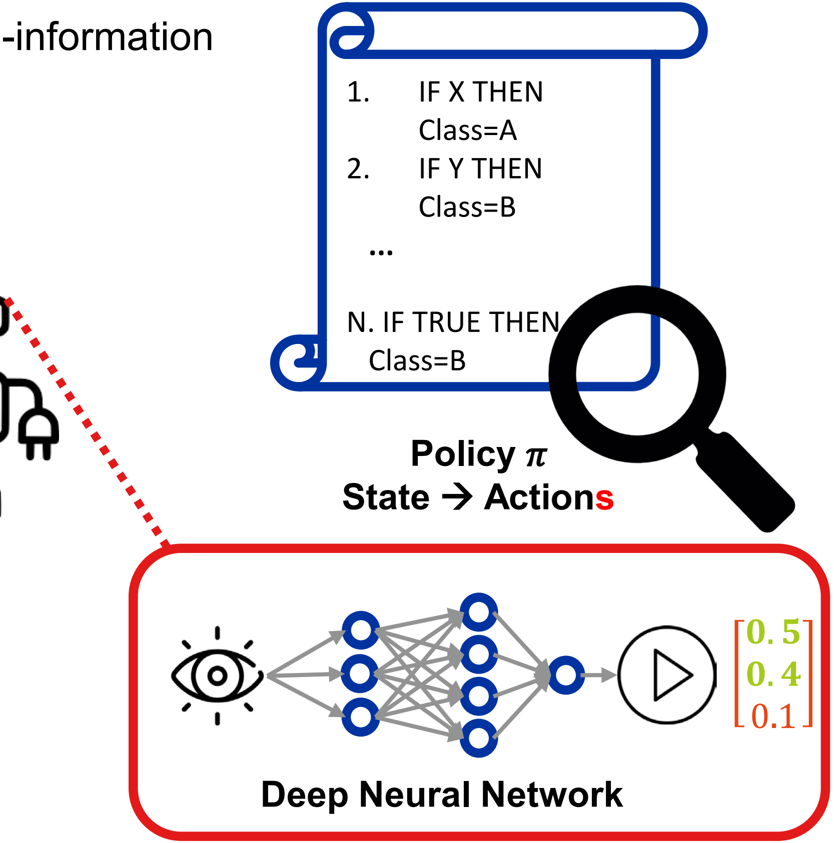


State

Reward

Action

Agent



1. IF X THEN Class=A
2. IF Y THEN Class=B

…

N. IF TRUE THEN Class=B

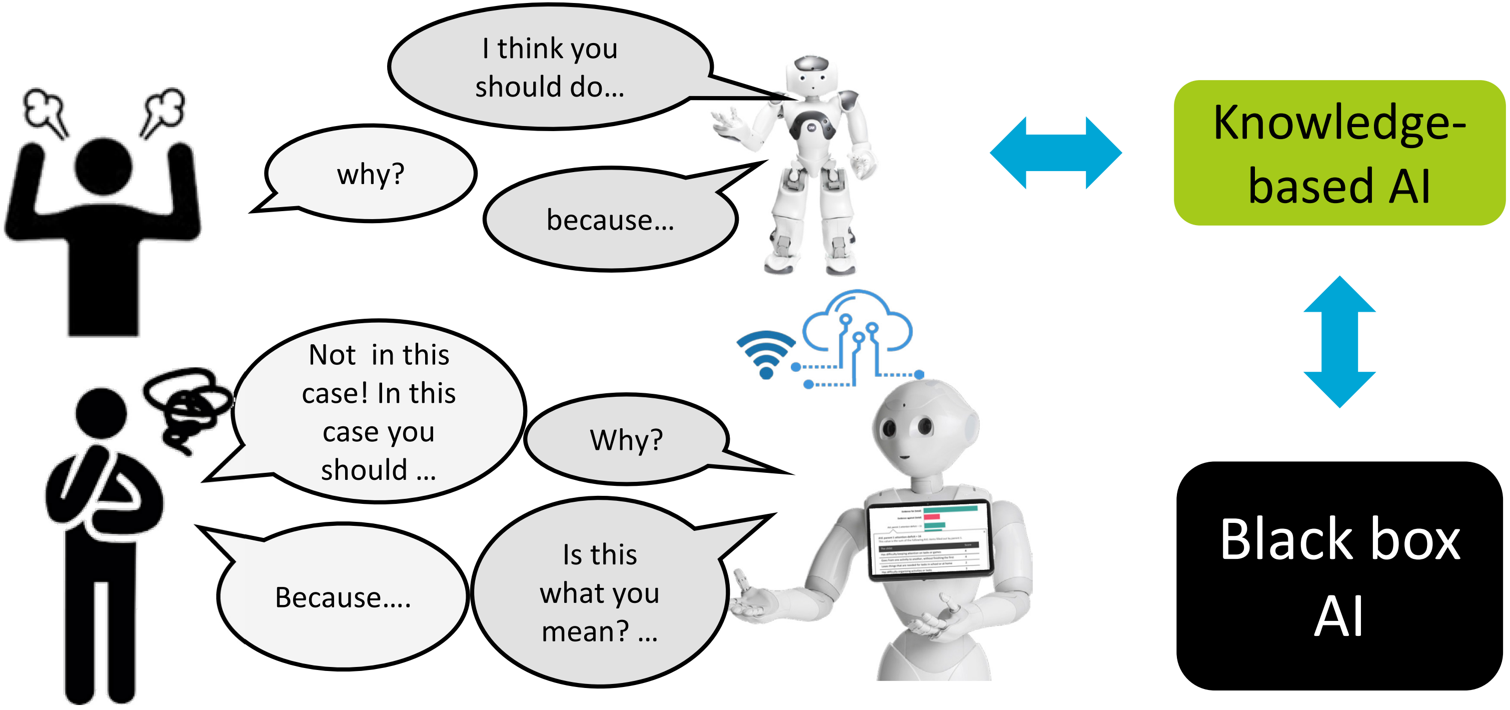
Policy π

State → Actions

Deep Neural Network

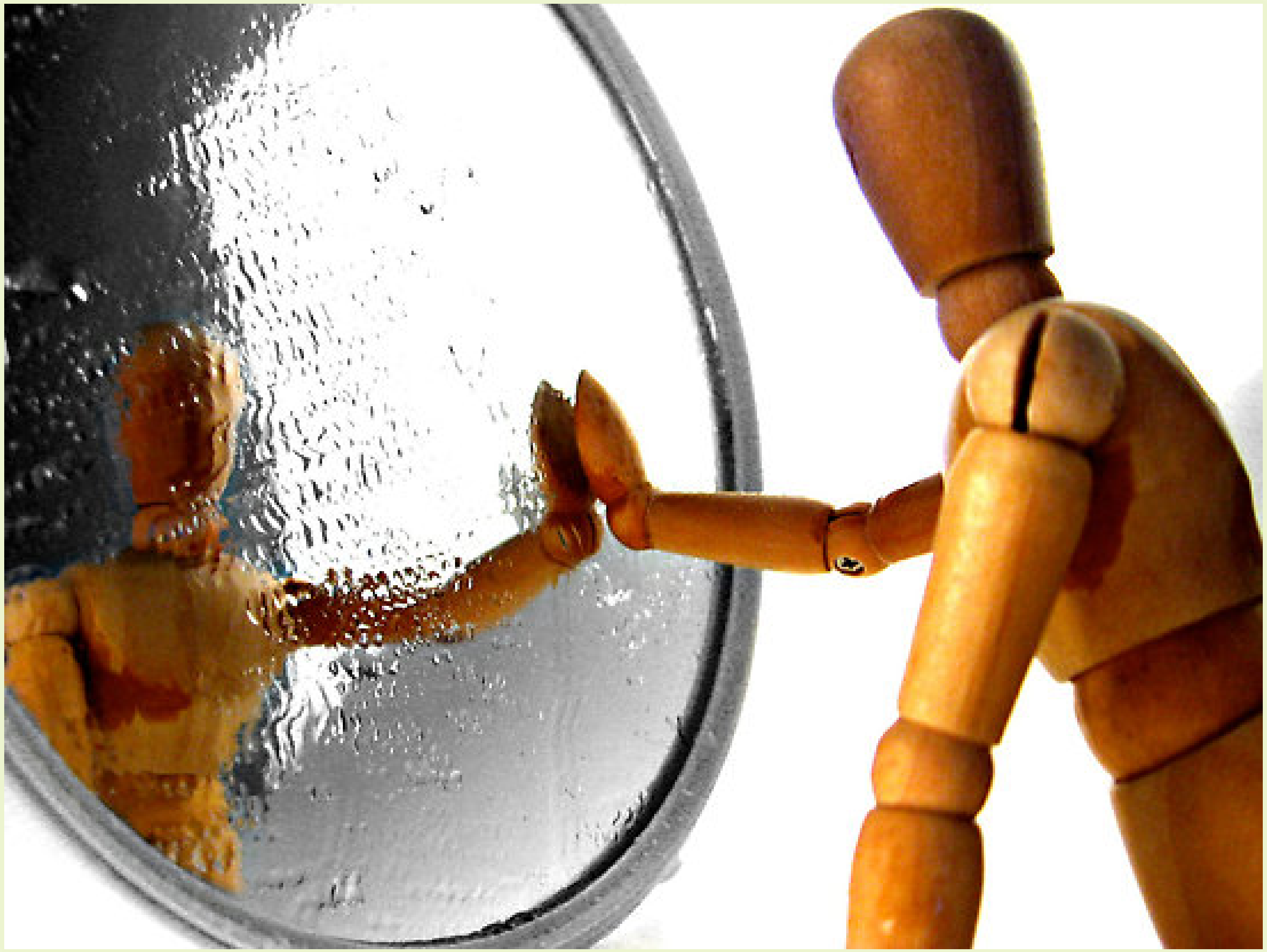
Coppens, Y., Steckelmacher, D., Jonker, C. M., & Nowé, A. (2020). Synthesising Reinforcement Learning Policies Through Set-Valued Inductive Rule Learning. In International Workshop on the Foundations of Trustworthy AI Integrating Learning, Optimization and Reasoning (pp. 163-179). .

Hybrid Intelligence over AI



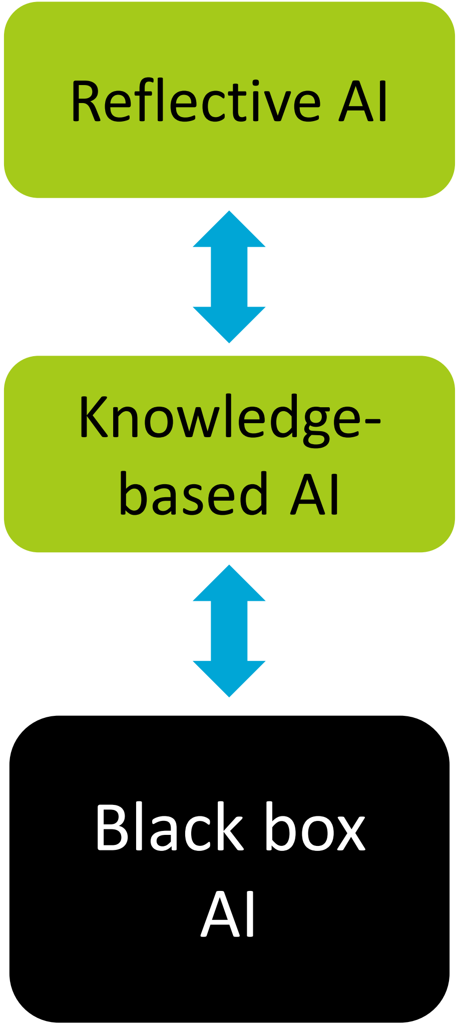
Knowledge-based AI

Black box AI



Self-reflective Hybrid Systems

- Where are we from a moral point of view?
- What biases are we forming?
- What is the quality of the data we use?
- Who has the expertise we need?
- Epistemic logic:
  - What do we know?
  - What do we know that we don't know?
  - Unknown Unknowns



Reflective AI

Knowledge-based AI

Black box AI

Universiteit Leiden

TUDelft

Mission

- Shift from autonomous AI to Hybrid Intelligence
  - Replace → Augment
  - Autonomy → Co-activity
  - Isolated AI → Escalate to HI
  - Hybrid Intelligence → System of Hybrid Systems
- Improve human & AI’s situational awareness
- Raise ethical awareness in humans & AI
- Place AI under meaningful human control with the help of KT

Universiteit Leiden

TU Delft

Self-reflective Hybrid Intelligent systems:

combining the strengths of

- Machine Learning
- Knowledge Representation
- Human Intelligence

https://www.hybrid-intelligence-centre.nl/
https://www.delftdesignforvalues.nl/
https://www.tudelft.nl/aitech

Catholijn M. Jonker