Hybrid Intelligence

AI systems that collaborate with people, instead of replacing them

Frank van Harmelen

Vrije Universiteit Amsterdam

(opinions are my own)

VU VRIJE UNIVERSITEIT AMSTERDAM
TU Delft
UvA
university of groningen
Universiteit Leiden
Universiteit Utrecht

Hybrid Intelligence

Augmenting human intellect

Frank van Harmelen

Vrije Universiteit Amsterdam

VU VRIJE UNIVERSITEIT AMSTERDAM TUDelft UvA university of groningen Universiteit Leiden Universiteit Utrecht

Intro & Motivation

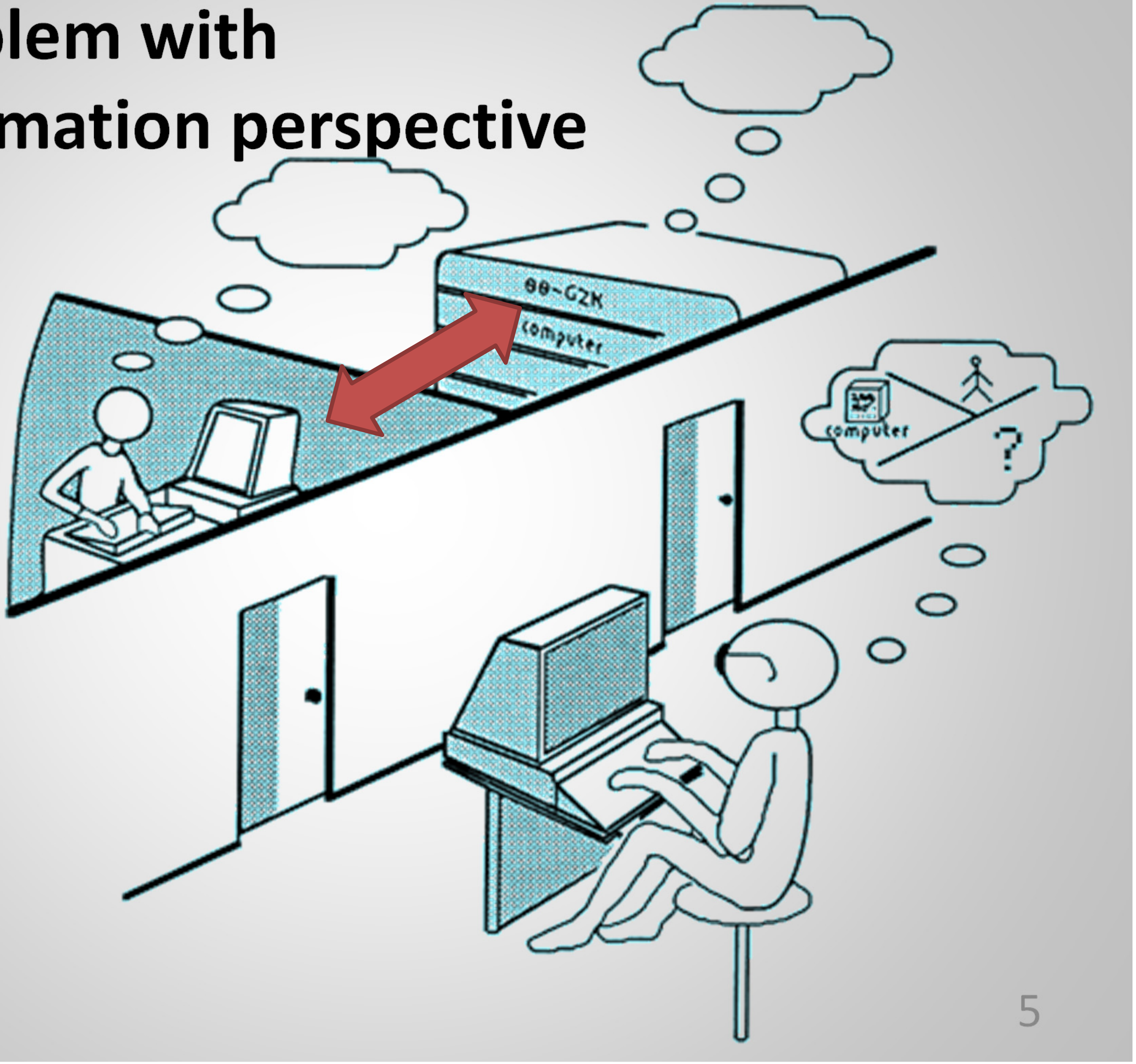
The automation perspective on AI

“My guess for when we will have **full autonomy** [in cars] is approximately three years” (Elon Musk, 2015)

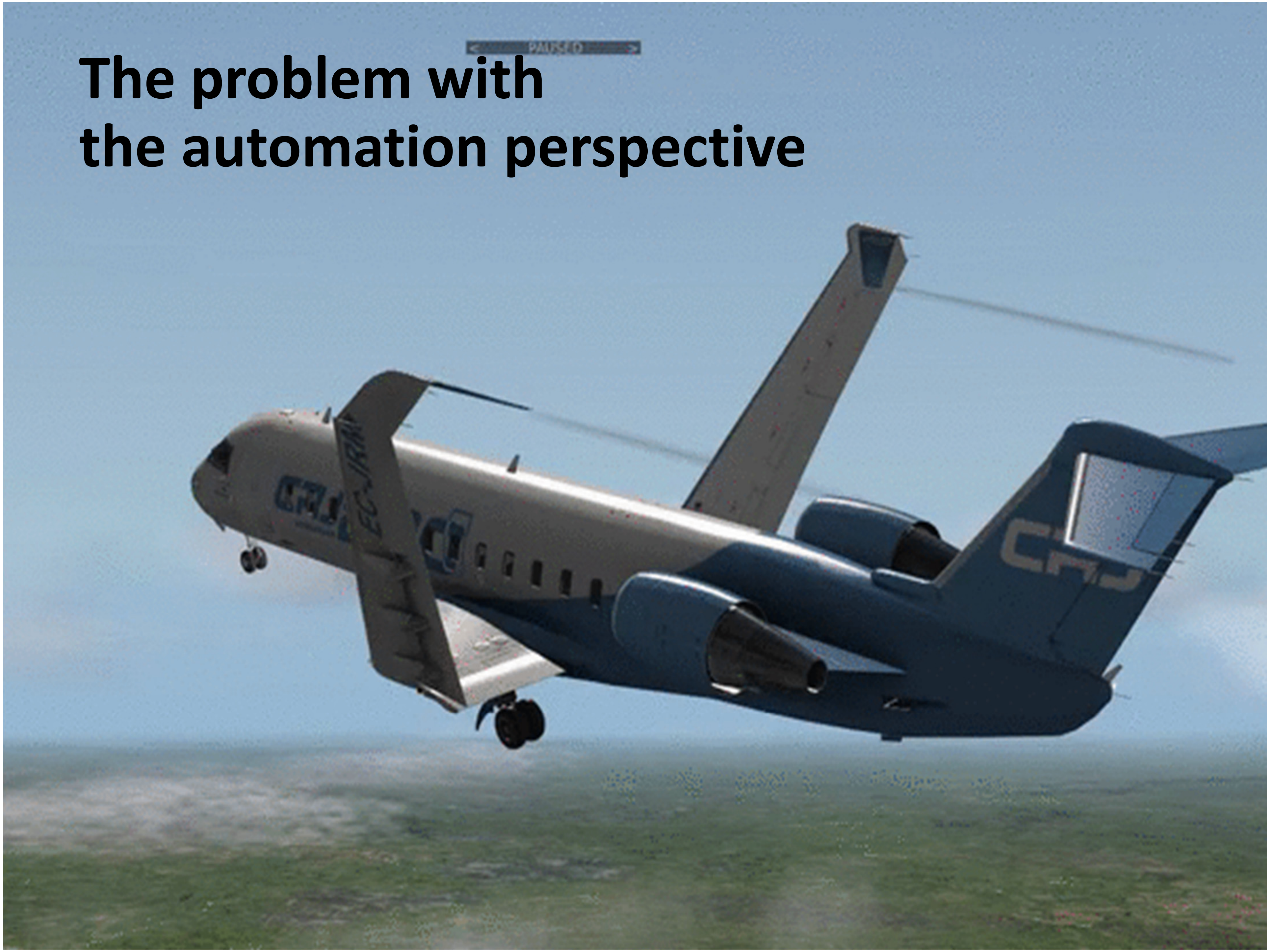
“[a] highly-trained and specialised radiologist may now be in greater danger of **being replaced by a machine** than his own executive assistant” (Andrew Ng, The Economist, 2016)

“People should stop training radiologists now. It’s just completely obvious that within 5 years, deep learning is **going to do better than radiologists**”
(Geoffrey Hinton, The New Yorker, 2017)

The problem with the automation perspective

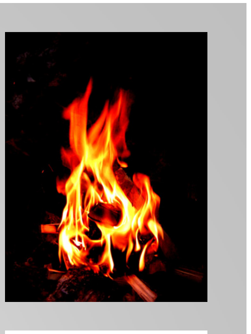


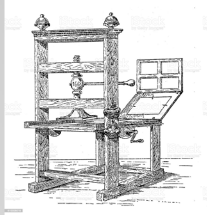
The problem with the automation perspective



Replacing humans?...☹

An alternative perspective on AI





Consider AI as:

fire, the wheel, the printing press,

the computer, the Internet

Enabling humans to scale up their capabilities.

Hybrid intelligence (HI):

- the combination of human and machine intelligence,
- augmenting human intellect and capabilities instead of replacing them
- achieving goals that were unreachable by either humans or machines alone.

Humans need AI

global pandemics,
resource scarcity,
environmental conservation,
climate change,
eroding democratic institutions

© The Scream 1895/Edvard Munch

Solutions are hampered by human cognitive biases:

Handling of probabilities
Short termism
Functional fixedness
In-group favoritism

Entrenchment
Confirmation bias
Stereotypes
....

We could use some help in cooperative problem solving…

AI needs humans

- AI performs well on very narrow tasks, poor generalisation outside the training data
  - face recognition trained on Caucasian faces,
  - MRI images trained on scanner from a single vendor

- AI is unaware of
  - norms and values
  - the reason for the computation
  - the context of the computation

So….

“It is better to view AI systems not as “thinking machines” but as cognitive prostheses that can help humans think and act better” (Deloitte, 2018)

Challenge of Hybrid Intelligence

How to build adaptive intelligent systems that

- augment rather than replace human intelligence,
- leverage our strengths,
- compensate for our weaknesses
- taking into account ethical, legal, and societal considerations.

A research agenda in four parts

COLLABORATIVE

ADAPTIVE

RESPONSIBLE

EXPLAINABLE

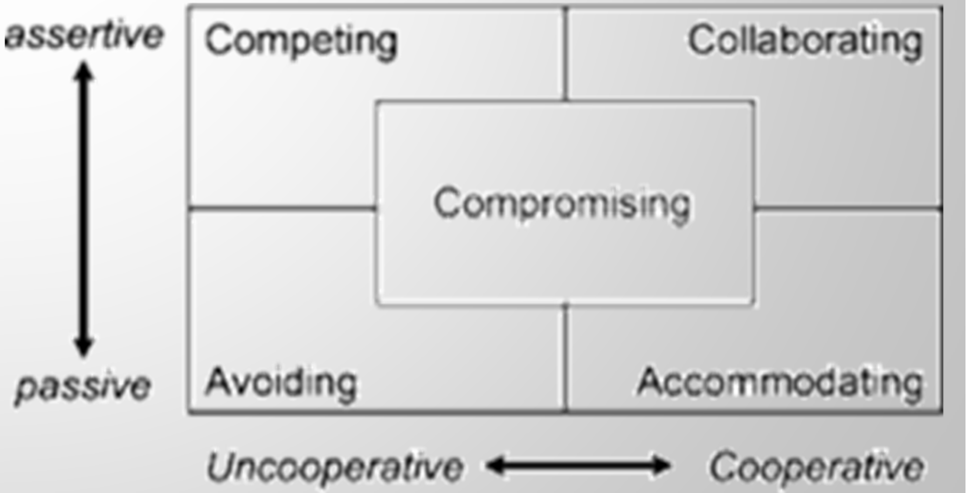
A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence

IEEE Computer, 53(8), 2020

Collaborative HI

State of the art in Collaborative AI

- **Negotiation**
- Planning
- Behaviour change support
- **Centaur Chess**

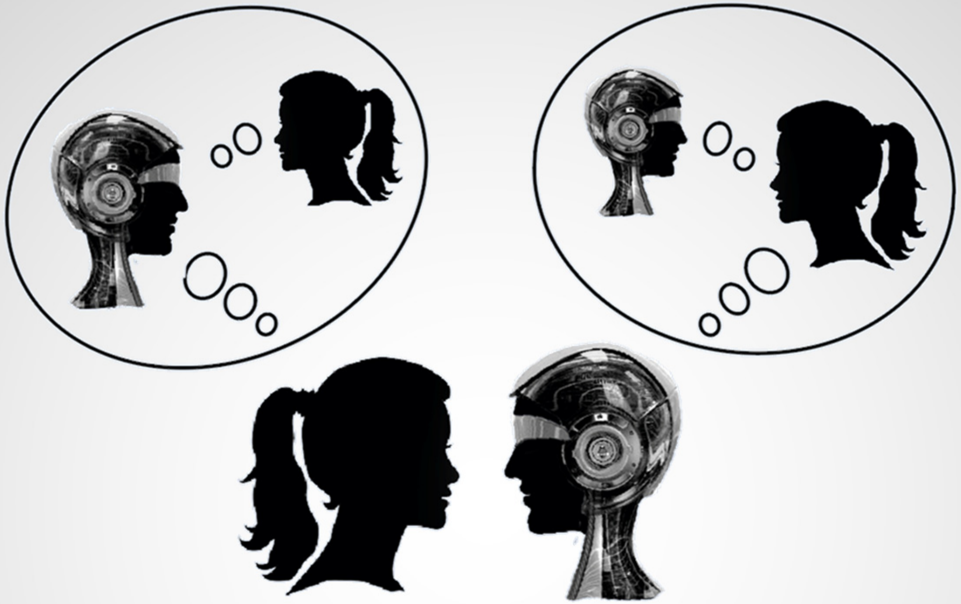


Challenges in Collaborative AI

- **perceive** social behavior by collaborators (language, vision)
- **communicate** with their collaborators (language, other modalities)
- a computational **understanding of human actors**
- an **understanding of joint actions** in teams, and
- **social norms** such as reciprocity, which are crucial in such teamwork.

Beyond traditional “human-in-the-loop”:
HI aims for reciprocity

Example: Theory of Mind



- 2nd order ToM is beneficial in competitive, cooperative, and mixed-motive situations
- software agents with deeper ToM levels give better support to humans on negotiation outcomes. (de Weerd et al, AI Journal 2013)

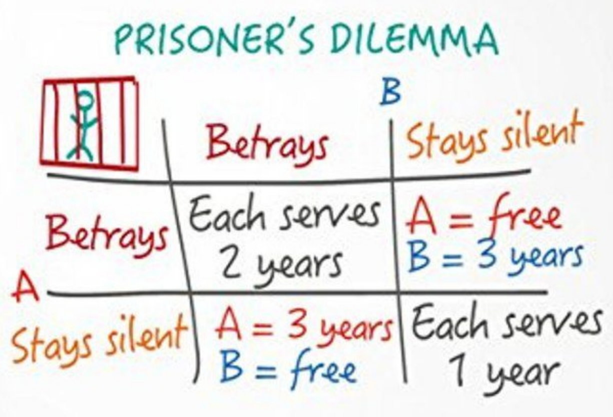
Max Planck Institute for Evolutionary Anthropology

Example: multi-agent systems

Human cooperation is based on kinship, direct reciprocity, indirect reciprocity (Romano & Balliet, Psychol. Science 2017).

- **Game theory**: maths of direct & indirect reciprocity
- **Epistemic logic**: maths of mutual knowledge and belief

omniscience: $P \to \Box P$
introspection: $\Box P \to \Box\Box P$
transparency: $\Box_i P \to \Box_j \Box_i P$



PRISONER'S DILEMMA

		B	
		Betrays	Stays silent

A	Betrays	Each serves 2 years	A = free
			B = 3 years

	Stays silent	A = 3 years
		B = free
			Each serves 1 year

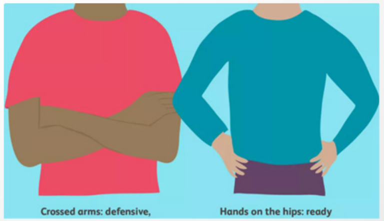
Example: multi-modal interaction

Interaction beyond language:

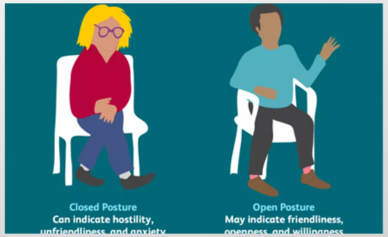
- Facial expression

Happiness
Confusion
Sadness
Surprise
Anger
Excitement

- Gesture



- Posture


Closed Posture
Can indicate hostility,
unfriendliness, and anxiety
Open Posture
May indicate friendliness,
openness, and willingness

Research questions for Collaborative HI

- **Computational models** for negotiation, agreements, planning, and delegation in hybrid teams
- A **computational Theory of Mind** for collaboration between humans and artificial agents
- How can **multimodal** messages, expressions and gestures be understood and generated for the purpose of collaboration?

Adaptive HI

Challenges for Adaptive HI

AI systems need to

- Adapt to change in environment
- Adapt to change in team
- Balance with desire for safety and reliability

State of the art for Adaptive HI

- Transfer learning
- Multi-task learning
- Auto-ML and meta-learning

Example: reinforcement-learning agent

- safety constraints encoded
  - in the reward/loss functions (*preferably don’t do this*)
  - as symbolic constraints (*never do this*)
  - as restriction on the exploration process (*don’t try this*)

Research questions for Adaptive AI

- **Constrained ML:** How can learning systems change during training, but still respect the societal, legal, ethical, safety, and resource constraints?
- **Transfer learning:** How can learning systems accommodate changes (in user preferences, environments, tasks, available resources) without having to completely relearn each time something changes?
- **Neurosymbolic ML:** How can the adaptivity of machine learning techniques be integrated with the precision and interpretability of symbolic knowledge representation and reasoning?

Responsible HI

Challenges for Responsible HI

- AI increasingly makes key decisions
  - for individuals
    (job selection, financial decisions, medical screening)
  - for society
    (spam filtering, fake news & hate speech detection)
- The reasons for these decisions are often unknown, and hence cannot be disputed
- Urgency of this os increasingly acknowledged (IEEE, UNESCO, EU, gov's in France, UK, others)
- Need to ground explanations in values, norms, motives, commitments, goals

Example: Ethical reasoning *about* HI systems

Ethics accounted for during the *design* process

Methods to

- identify stakeholders,
- identify values and goals,
- identify conflicts,
- align values and goals

“Design for values”
(Robert Moses’
racist bridge)

Example: Ethical reasoning *by* HI systems

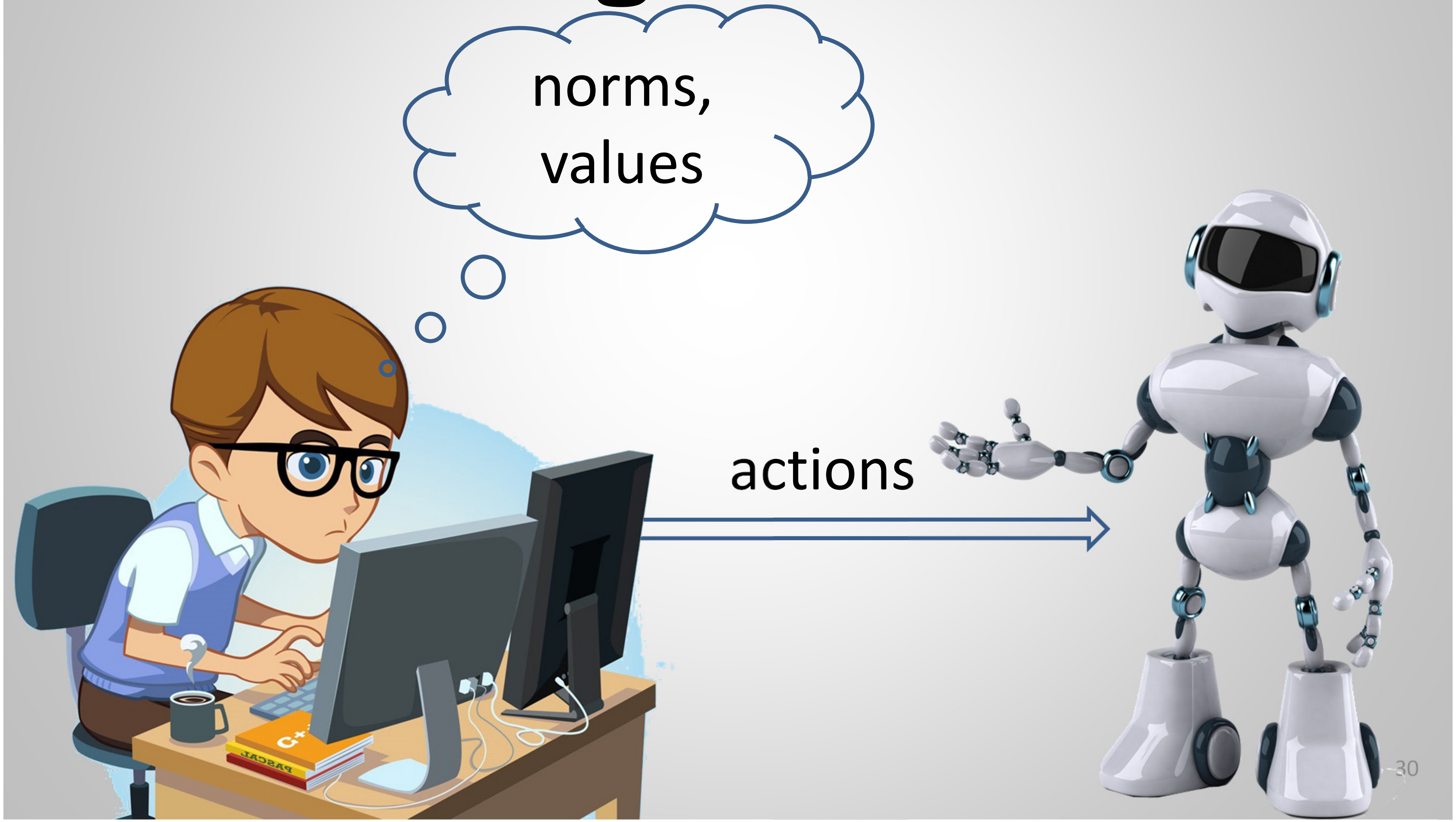
Ethics accounted for during the *computation* process

- encode/model moral reasoning,
  ethical decision making done by the system
  (presumes some encodable moral theory)

- allow humans to express their norms and values to the system at runtime,
  ethical decision making emerges
  from the human-machine interaction
  (still presumes some encodable moral theory)

Responsible:

ethics in design

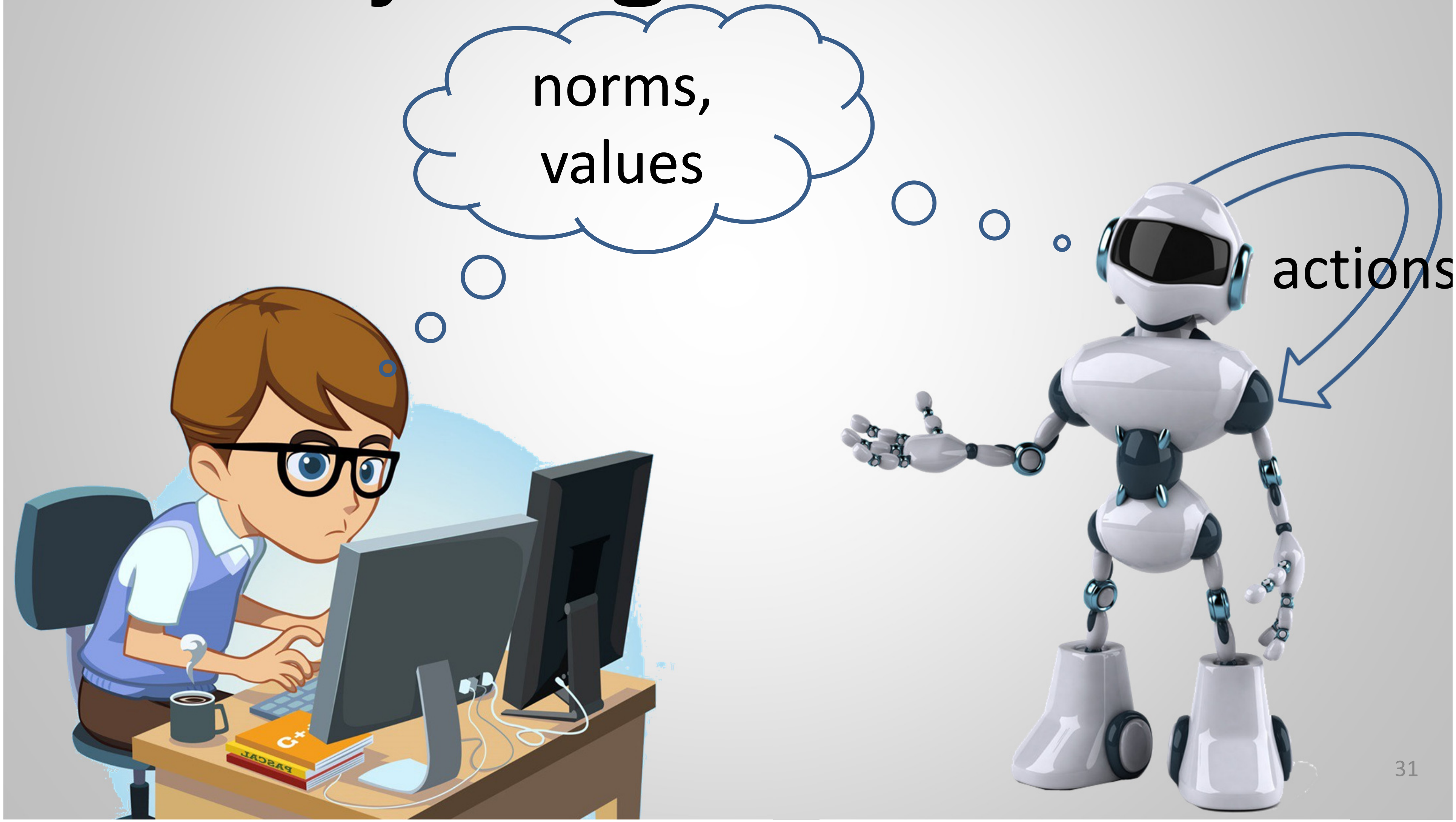


norms, values

actions

Responsible:

ethics by design

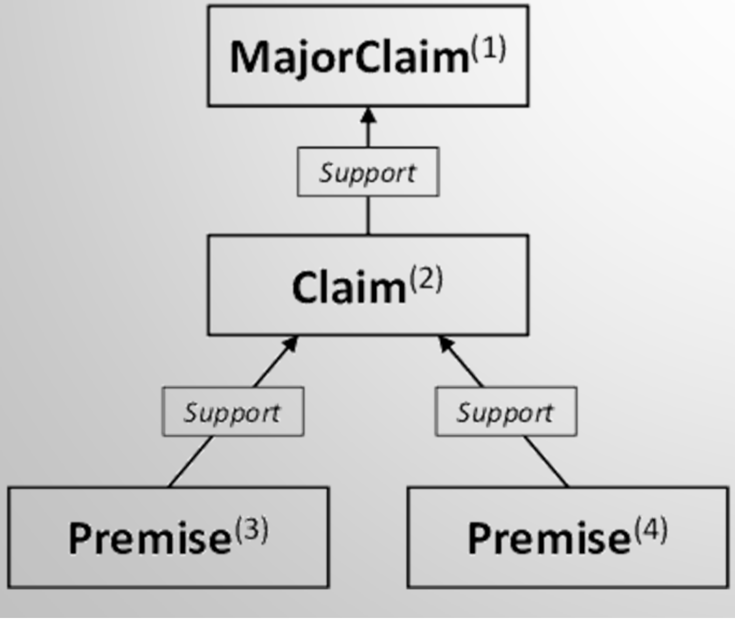


norms, values

actions

Example: argumentation theory

- the argumentation structure is encoded in the system, and argumentation is performed by the system (presumes an encodable theory of argumentation)
- the arguments themselves are provided by humans, either interactively or by text-mining



MajorClaim(1)

Support

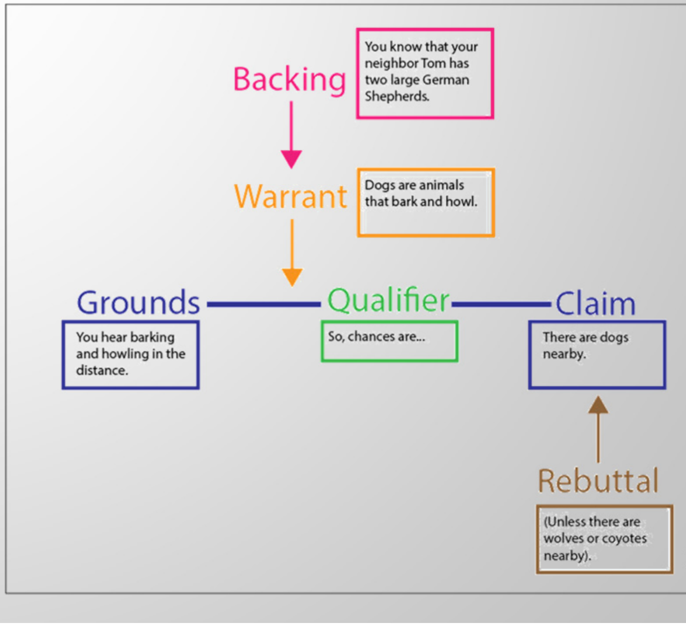
Claim(2)

Support

Support

Premise(3)

Premise(4)



Backing

You know that your neighbor Tom has two large German Shepherds.

Warrant

Dogs are animals that bark and howl.

Grounds

You hear barking and howling in the distance.

Qualifier

So, chances are...

Claim

There are dogs nearby.

Rebuttal

(Unless there are wolves or coyotes nearby).

Research questions in Responsible HI

Ethics in design

- How to include ELS considerations in the development process?
- How to verify the agent's architecture and behavior w.r.t. ELS requirements?

Ethics by design

- What new computational techniques are required for ELS by design
- What are the ELS concerns around the development of systems that can reason about ELS consequences of their decisions and actions?

Explainable HI

Challenges in Explainable HI

- Explanations are crucial for building trust, essential in collaboration

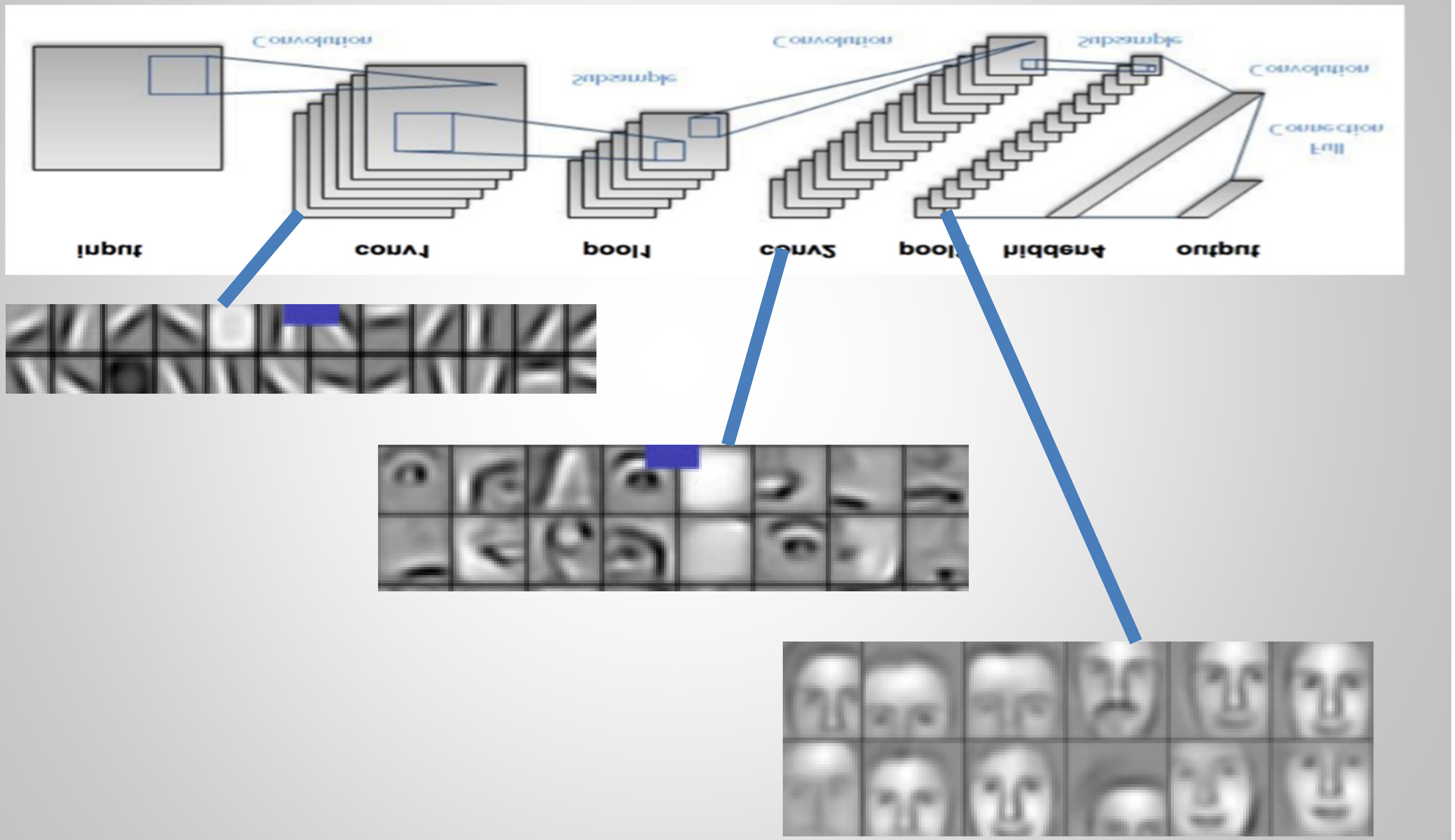
- **Faithful explanations**:
  explain the mechanics of the machine model, possibly at some higher level of abstraction
- **Rational reconstructions**:
  give a justification for the decision,
  without it being necessarily faithful to how it was derived.

Challenges in Explainable HI

- **Contrastive explanations**:
  explain not why an event happened but
  explain why it happened instead of something else

- **Social explanations**:
  an explanation serves a social purpose
  (convince someone , transfer knowledge)
  so must be related to the receiver's beliefs
    (or: to the explainer's beliefs about the receiver's belief;
     or to the explainer's beliefs about the receiver’s
    believes about the explainer's beliefs)

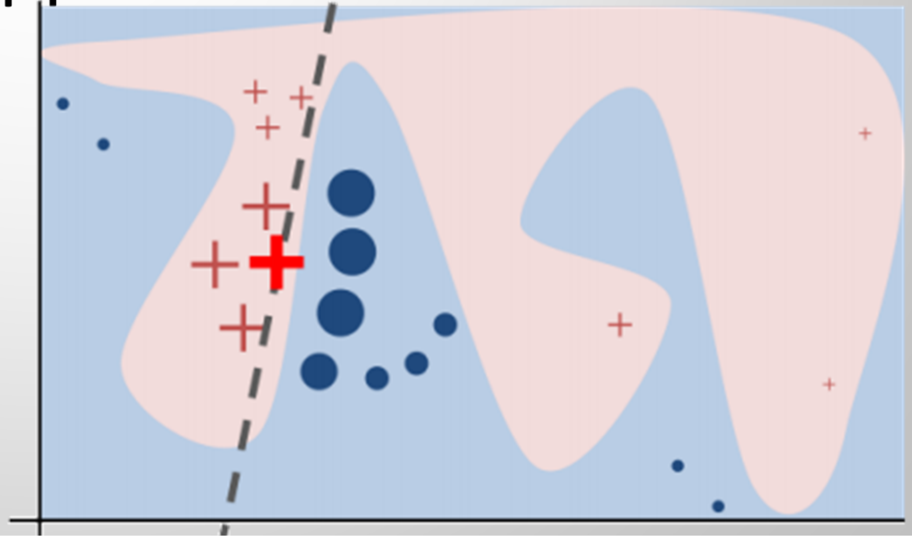
Example: faithful explanations



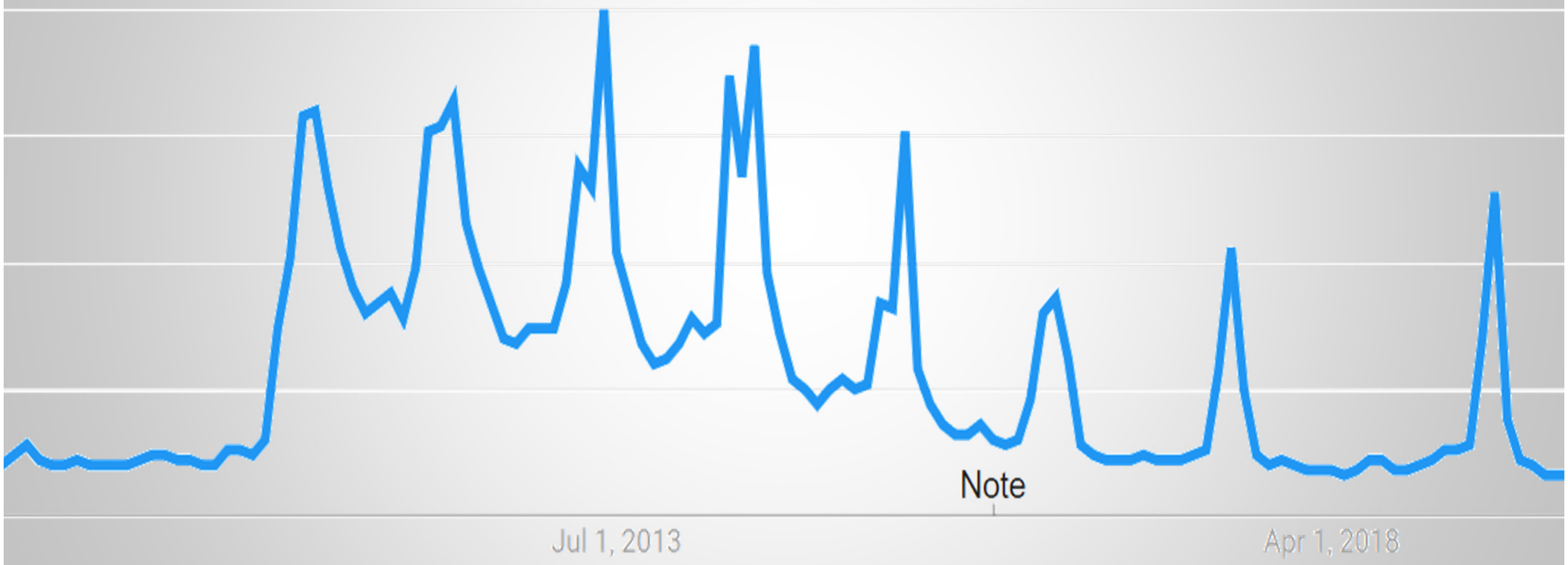
Example: faithful explanations

Other examples:

- Find the most influential training example
- Use the gradient of the output probability to find the most important features
- Give a locally linear approximation of the classification surface

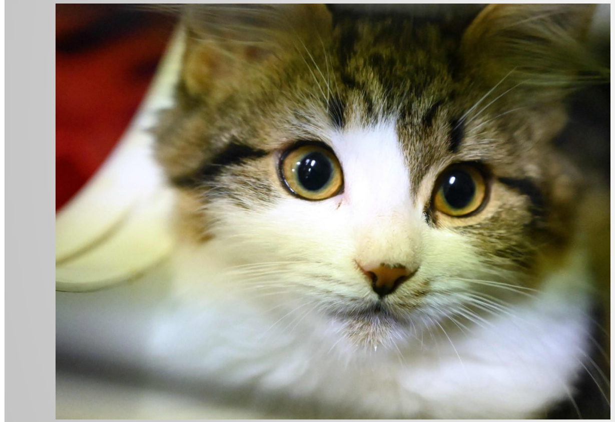


Example: rational reconstruction

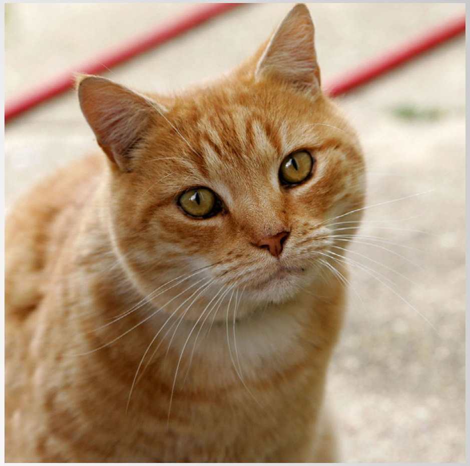


Google Trends for “Song of Ice and Fire”

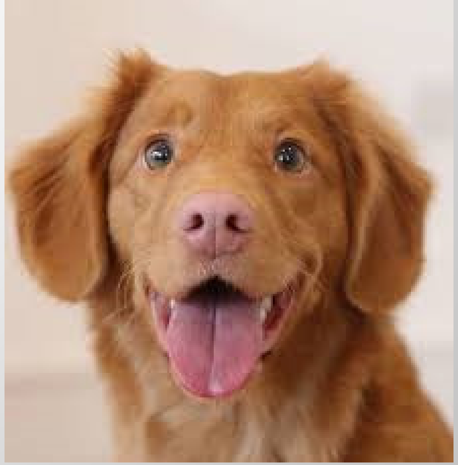
Example: contrastive explanation



is similar to



is different from



Example: contrastive explanation

1. Because I dropped it.

2. Because I dropped it,
    and it has mass,
    and the earth has mass,
    and Newton's gravitational law,
    and air resistance lower than momentum of cup,
    and ....

Research questions for Explainable HI

- What are the **different types of explanations** that make the decision-making process more transparent and understandable?
- How can explanations be **communicated** to users such that they improve the user’s trust
- How can explanations be **personalized** to align with the users’ needs and capabilities
- What are **shared representations** as the basis for explanations, covering both the external world and the internal problem-solving process?
- How to **evaluate** quality and strength of explanations?

Potential HI application scenario's

- **Education**: teacher-system collaboration to give extra attention to children to slow-learners or to fast-learners
- **Health-care**: nurse-system collaboration for patient observation and question-answering)
- **Health-care**: care pathway management between patients, GPs, nurses, specialists, family
- **Public health**: personalised coaching during a pandemic, reconciling personal goals with public goals
- **Science**: collaboration in all parts of the scientific cycle:

HYBRID INTELLIGENCE

- Aimee van Wynsberghe
- Annette ten Teije
- Antske Fokkens
- Bart Verheij
- Birna van Riemsdijk
- Catholijn Jonker
- Christof Monz
- Dan Balliet
- Davide Grossi
- Eliseo Ferrante
- Florian Kunneman
- Frank Dignum
- Frank van Harmelen
- Frans Oliehoek
- Guszti Eiben
- Hayley Hung
- Henry Prakken
- Herke van Hoof
- Holger Hoos
- Jakub Tomczak
- Koen Hindriks
- Maarten de Rijke
- Mark Neerincx
- Max Welling
- Myrthe Tielman
- Piek Vossen
- Rineke Verbrugge
- Roel Dobbe
- Silja Rennooij
- Stefan Schlobach
- Victor de Boer
- Virginia Dignum

VU VRIJE UNIVERSITEIT AMSTERDAM TUDelft UvA university of groningen Universiteit Leiden Universiteit Utrecht