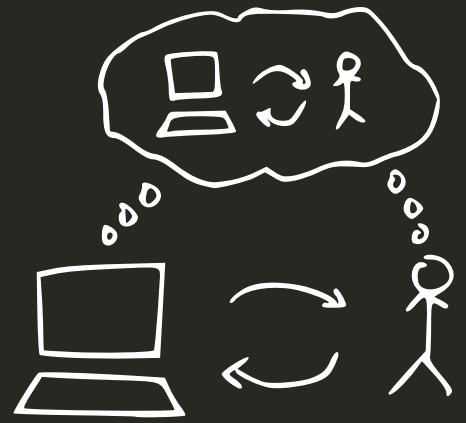


USER AND MACHINE THEORY OF MIND

AI Day 2018

Tomi Peltola
Aalto University

December 12, 2018

GOAL

Make human–AI collaboration as efficient as human–human.

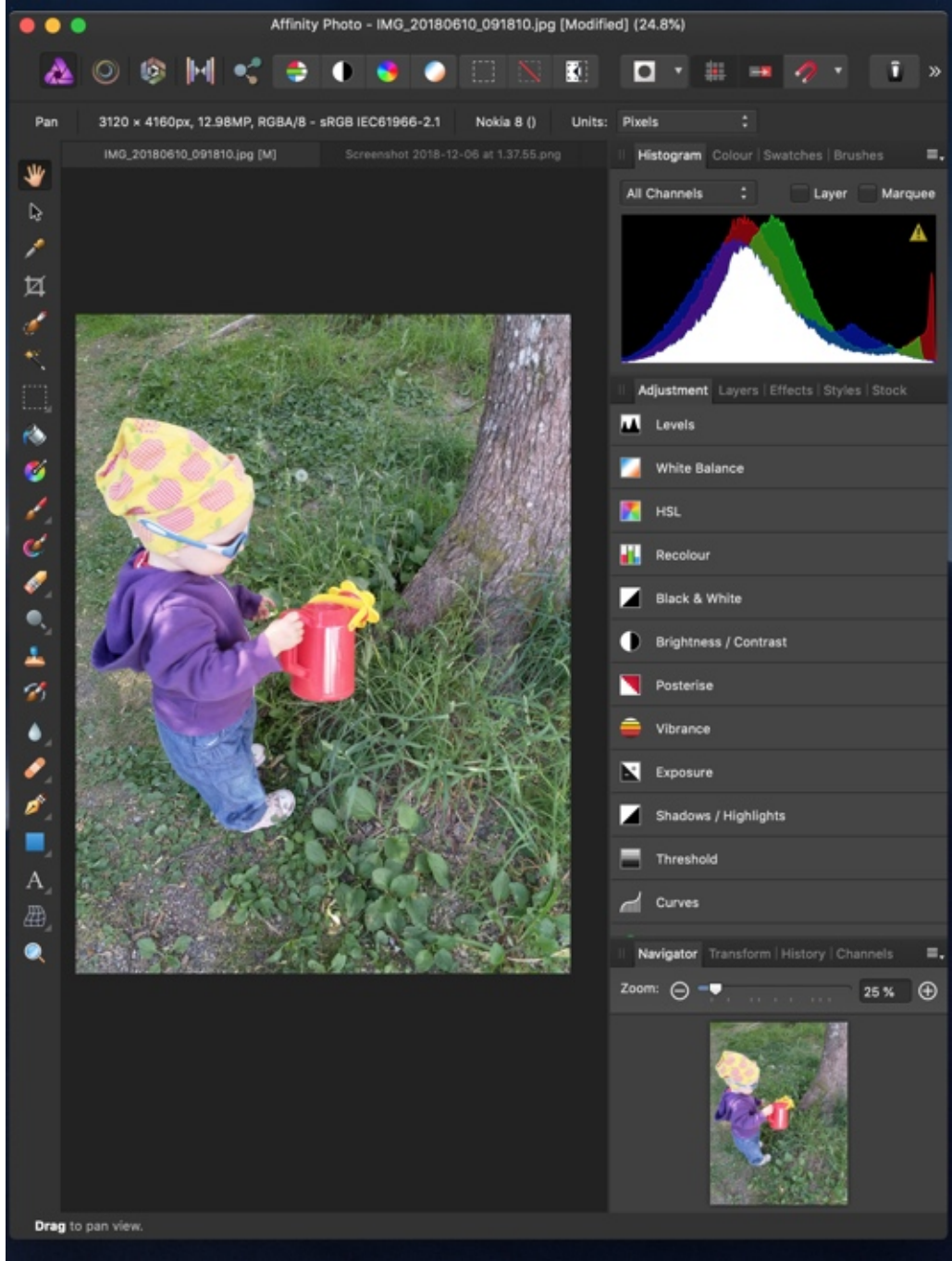
Humans form **mental models and representations** of the world.

We **make sense** of the world and **plan and act** based on them.

Humans form **mental models and representations** of the world.

We **make sense** of the world and **plan and act** based on them.

This extends to **the digital world**.



WE ALSO REASON ABOUT OTHER HUMANS

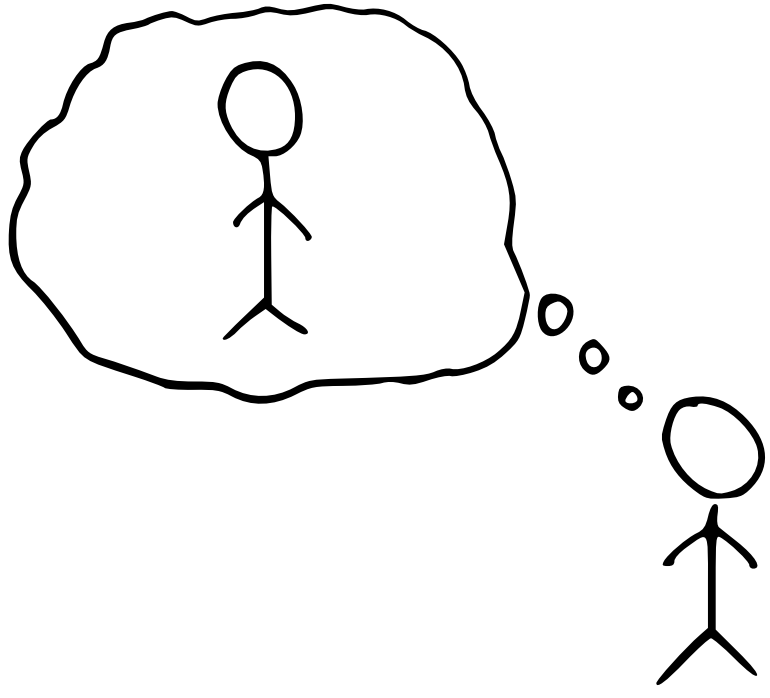
beliefs, knowledge, intent, purpose, goals, desires, emotions, thinking, ...

We also reason about other humans

beliefs, intents, purpose, desires, emotions, knowledge, goals, thinking, ...

THEORY OF MIND

"An individual has a theory of mind if he imputes mental states to himself and others. A system of inferences of this kind is properly viewed as a theory because such states are not directly observable, and the system can be used to make predictions about the behavior of others." (Premack and Woodruff, 1978)

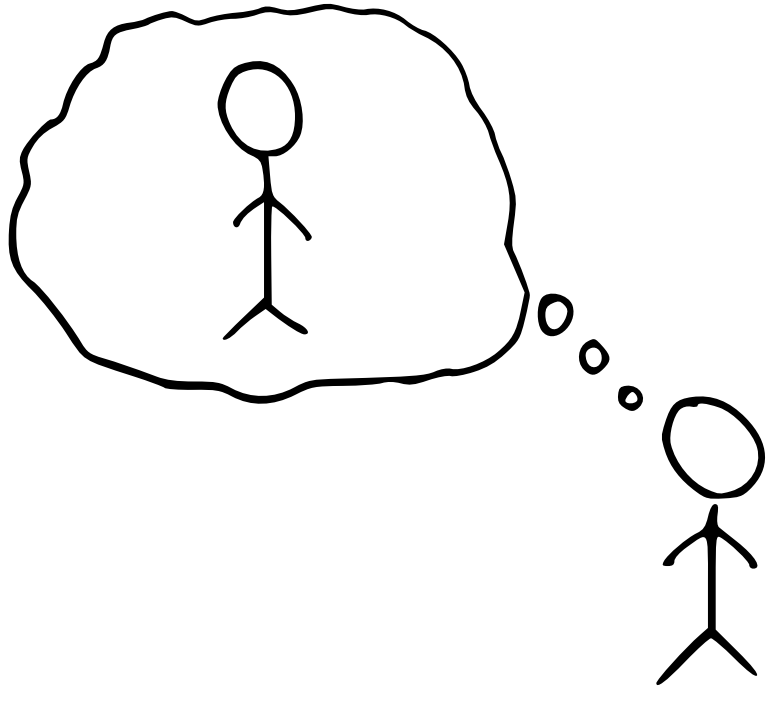


We also reason about other humans

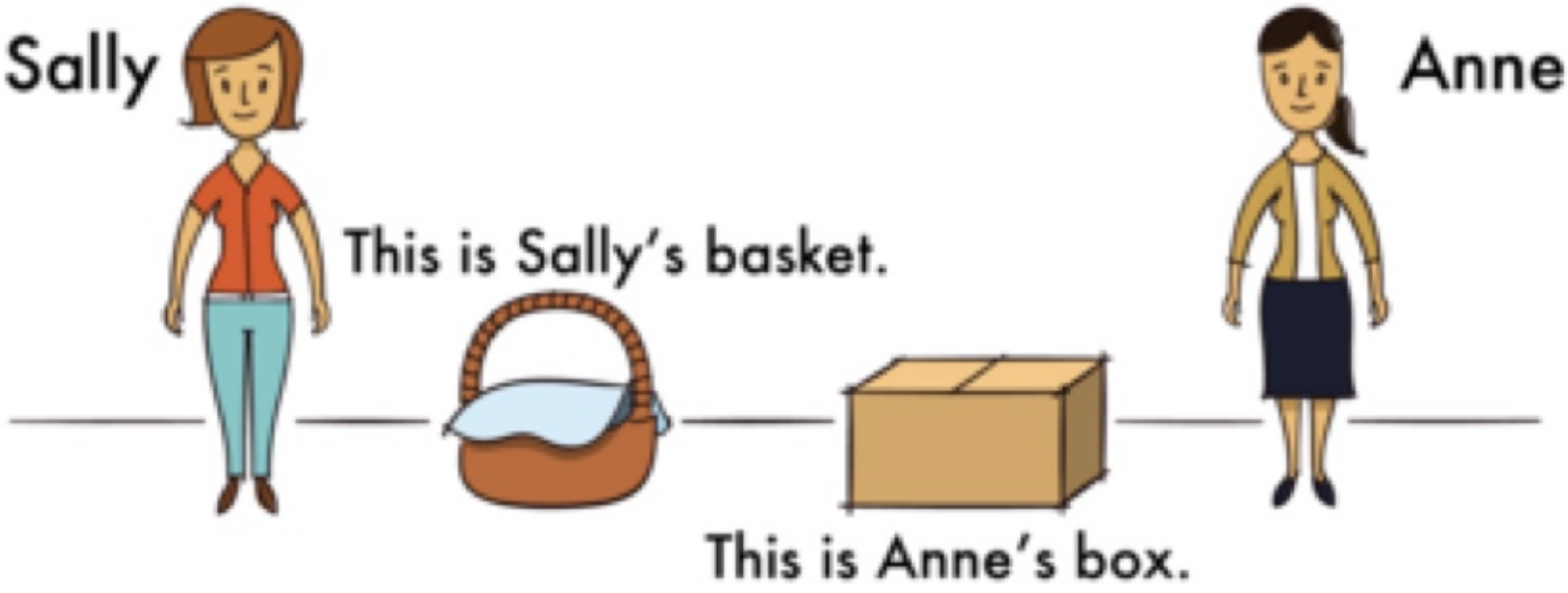
beliefs, intents, purpose, desires, emotions, knowledge, goals, thinking, ...

THEORY OF MIND

"An individual has a theory of mind if he imputes mental states to himself and others. A system of inferences of this kind is properly viewed as a theory because such states are not directly observable, and the system can be used to make predictions about the behavior of others." (Premack and Woodruff, 1978)

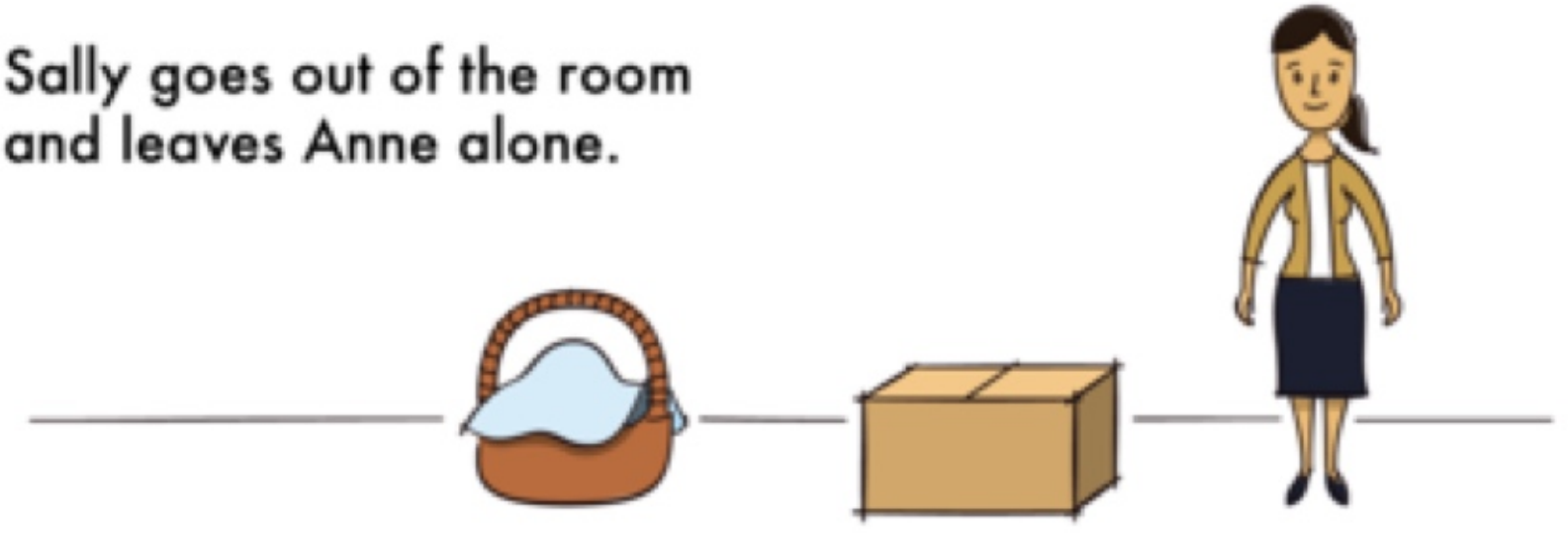


Social cognition, mindreading, mentalizing, folk psychology



[figure: https://nobaproject.com/modules/theory-of-mind]

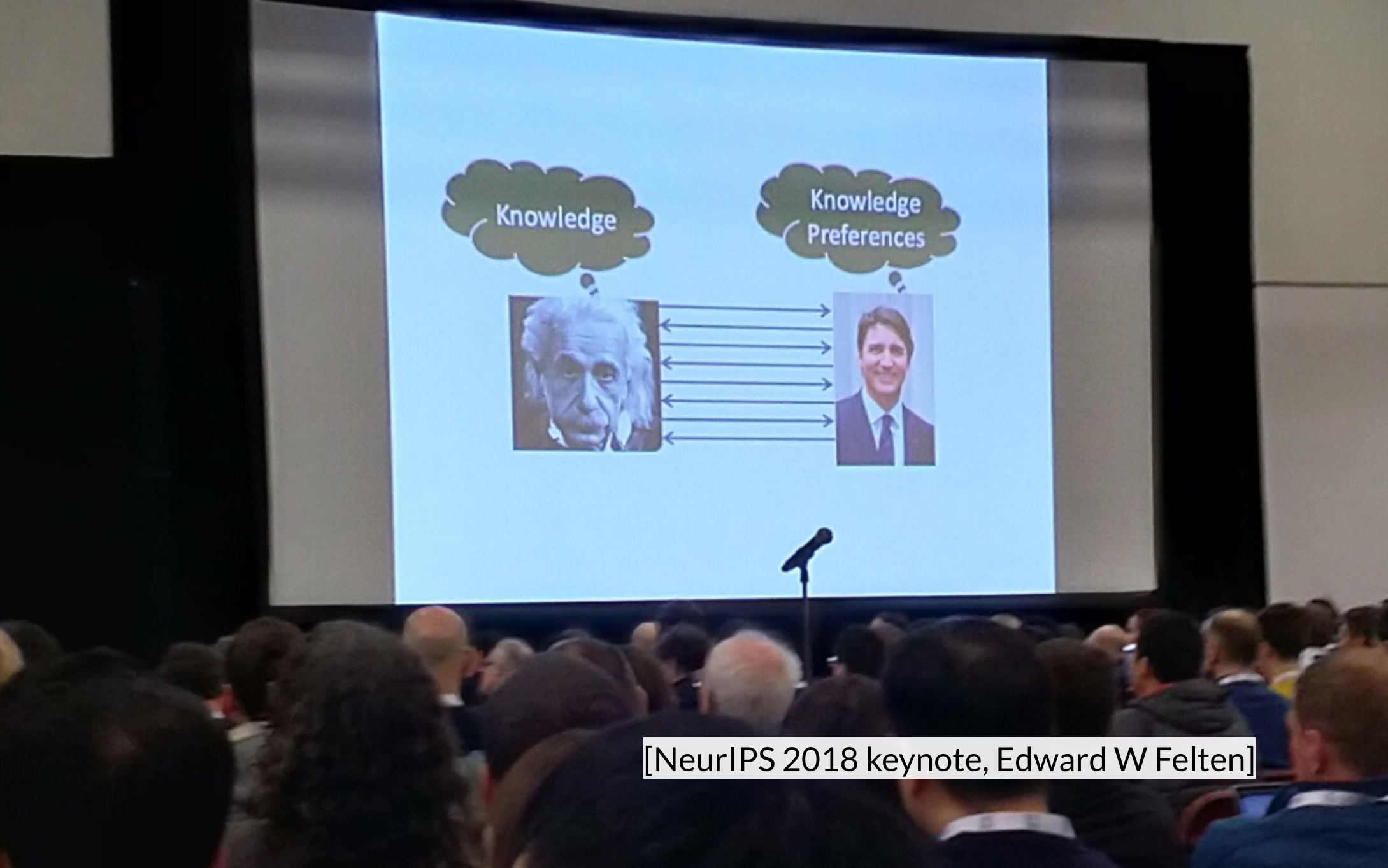
[figure: https://nobaproject.com/modules/theory-of-mind]



[figure: https://nobaproject.com/modules/theory-of-mind]

[figure: https://nobaproject.com/modules/theory-of-mind]

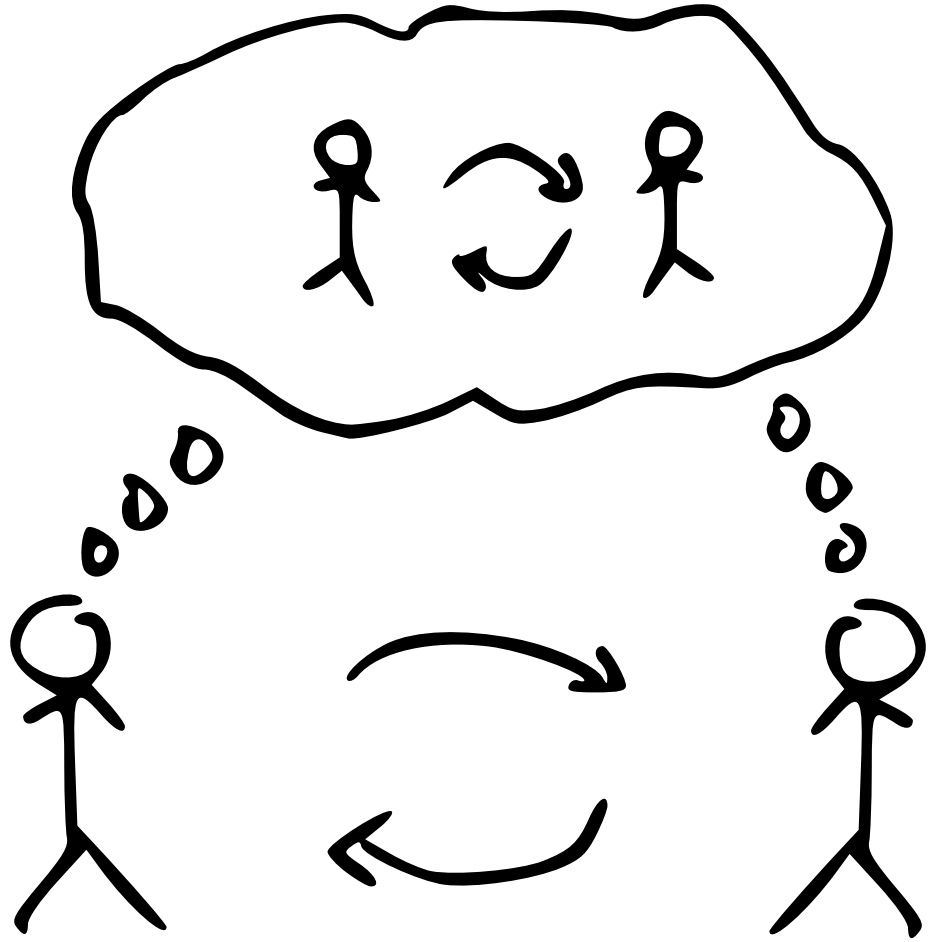
[figure: https://nobaproject.com/modules/theory-of-mind]



[NeurIPS 2018 keynote, Edward W Felten]

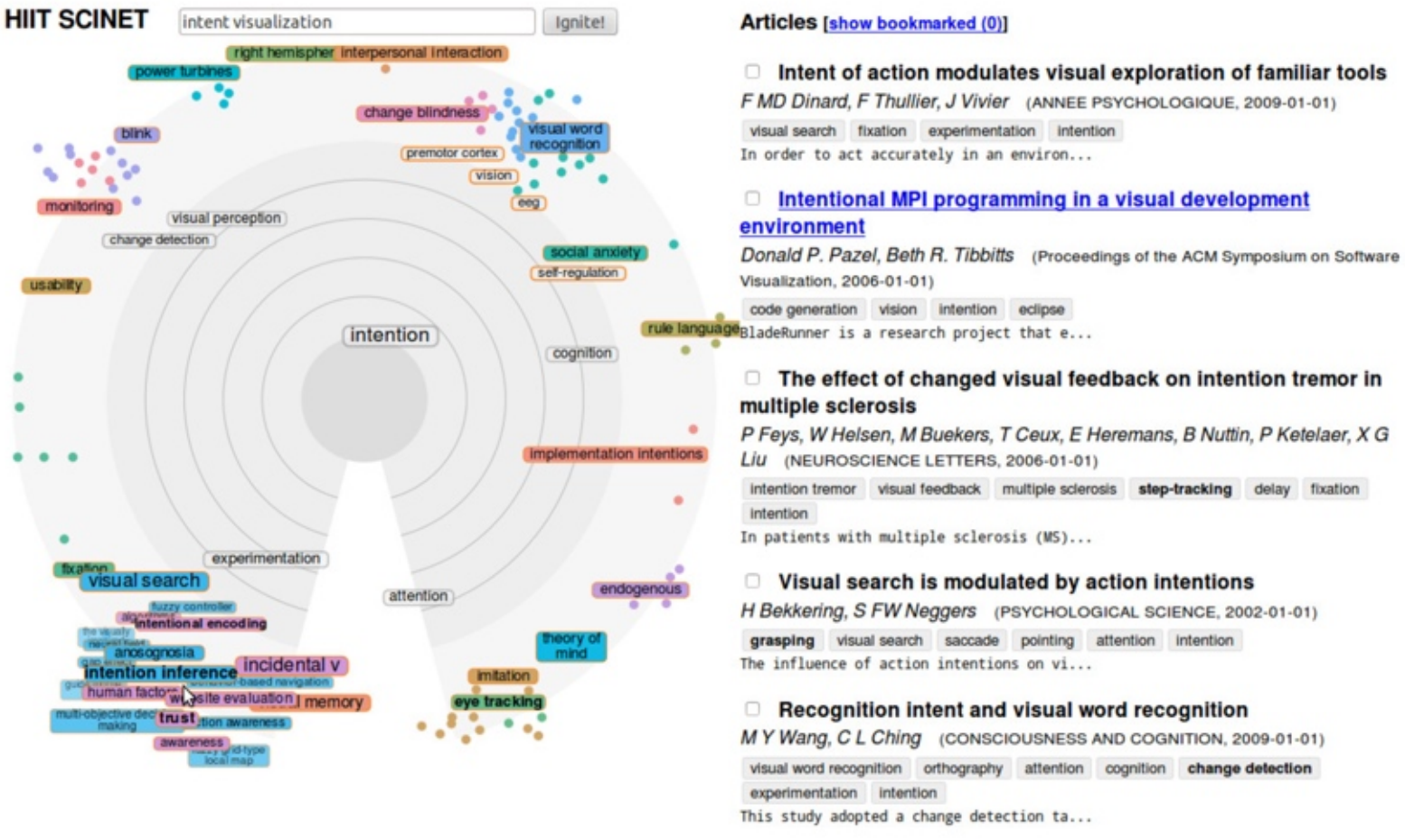
THEORY OF MIND

is essential for efficient human–human collaboration.



Increasingly, we interact with machine learning based adaptive systems.

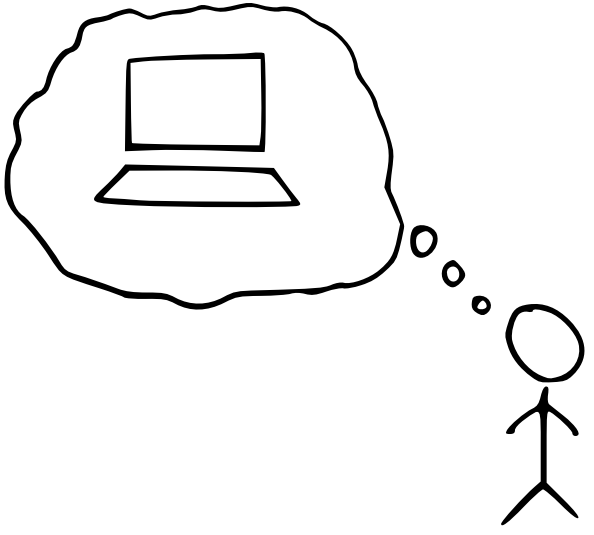
amazon Google NETFLIX



THEORY OF AI'S MIND

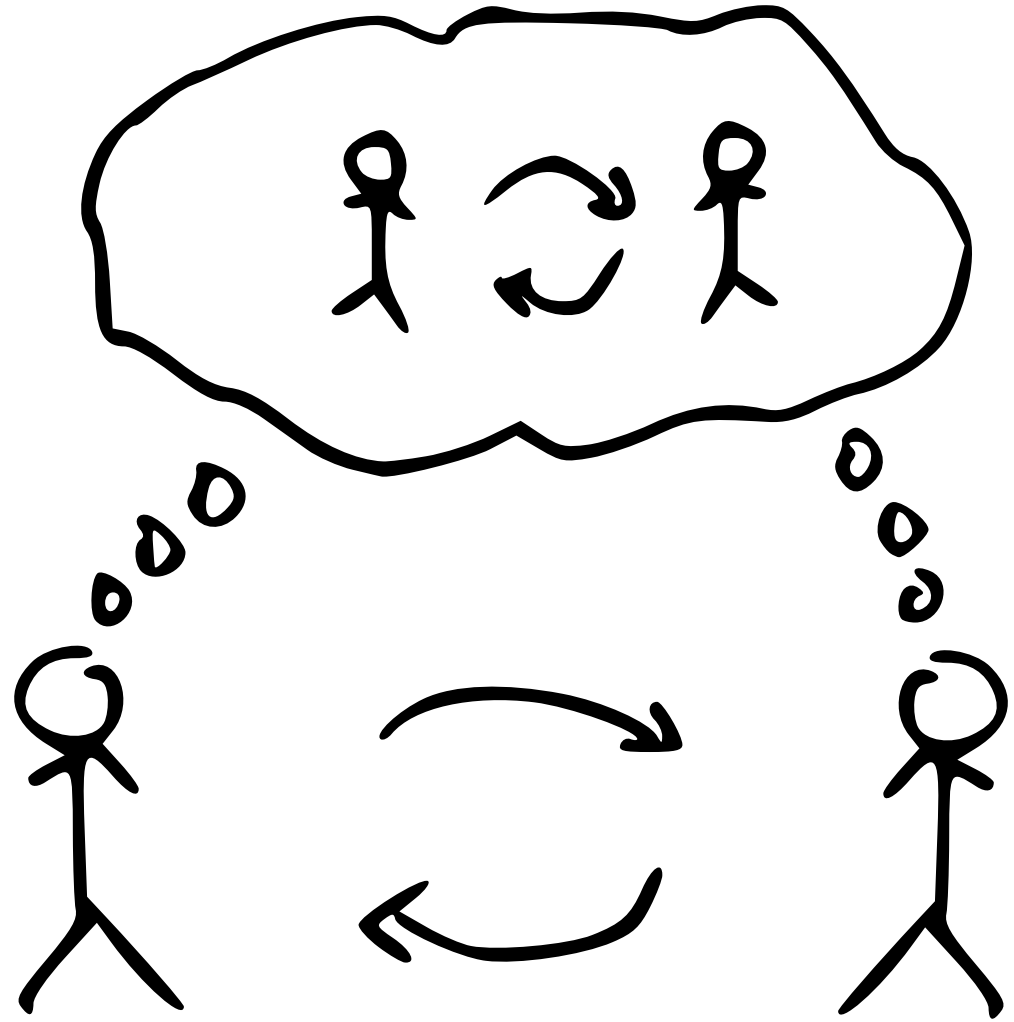
Humans create mental models of the adaptive systems and can predict their behaviour (Chandrasekaran et al., arXiv, 2017).

- Understandability/predictability is essential.

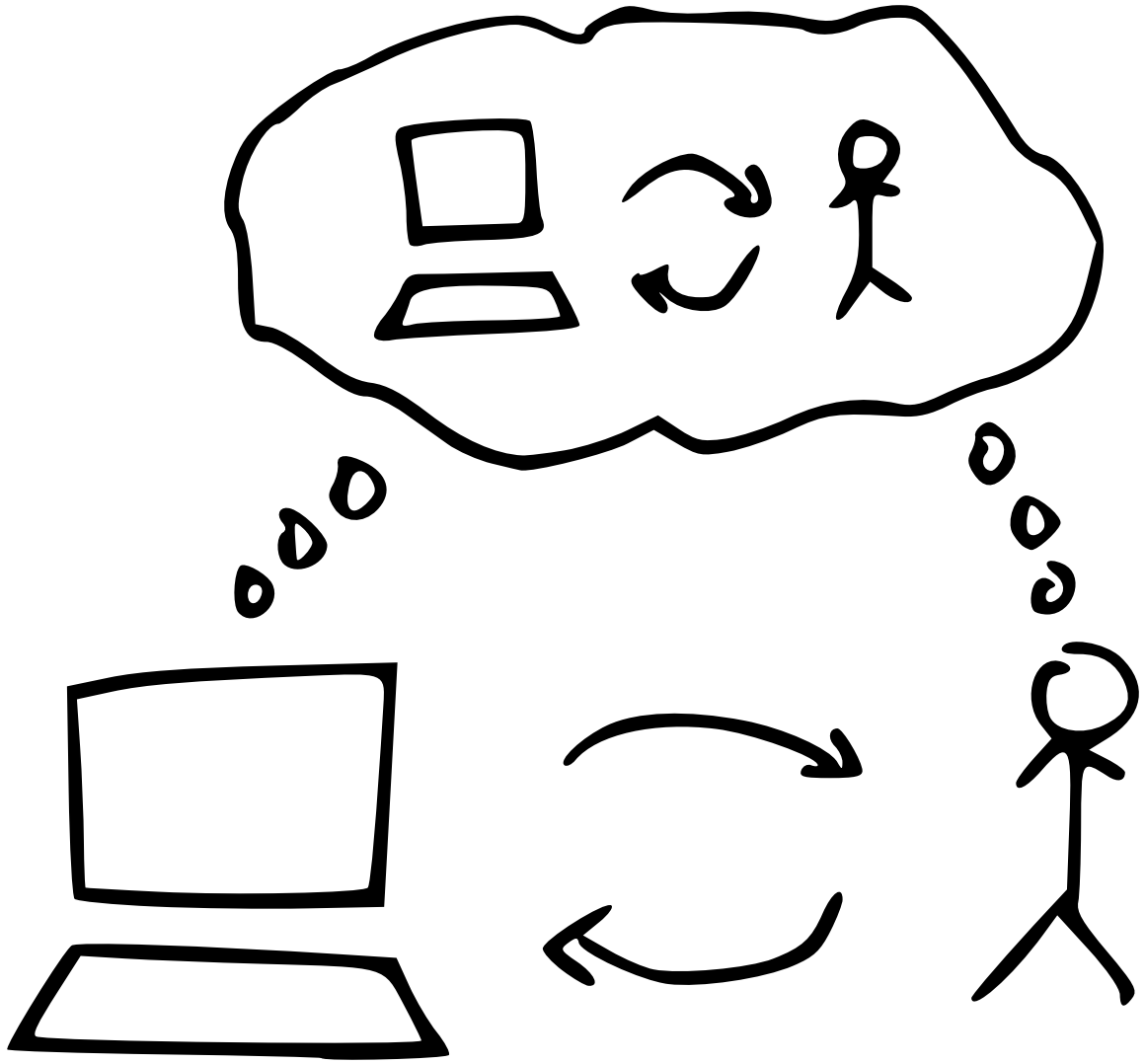


Goal: **Make human–AI collaboration as efficient as human–human**.

Goal: **Make human–AI collaboration as efficient as human–human.**

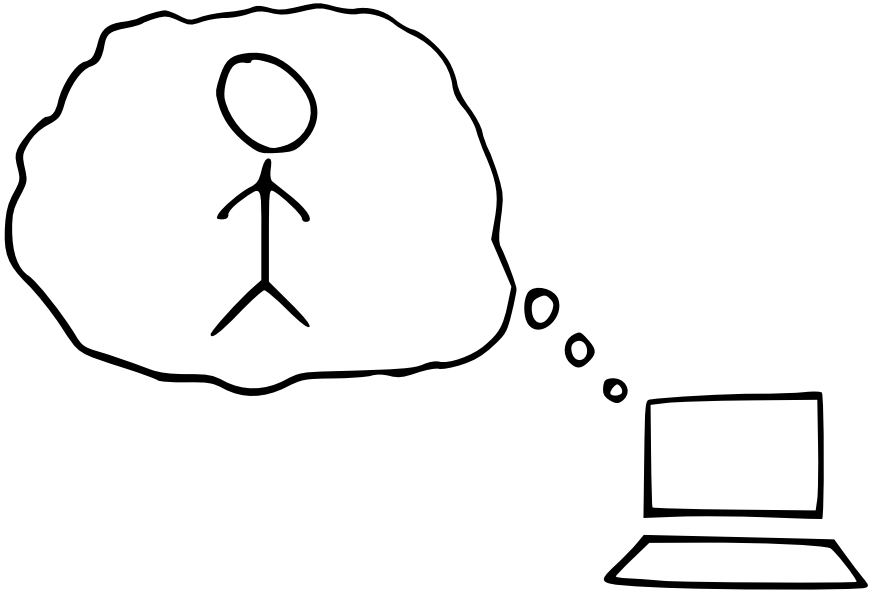


Goal: **Make human–AI collaboration as efficient as human–human.**

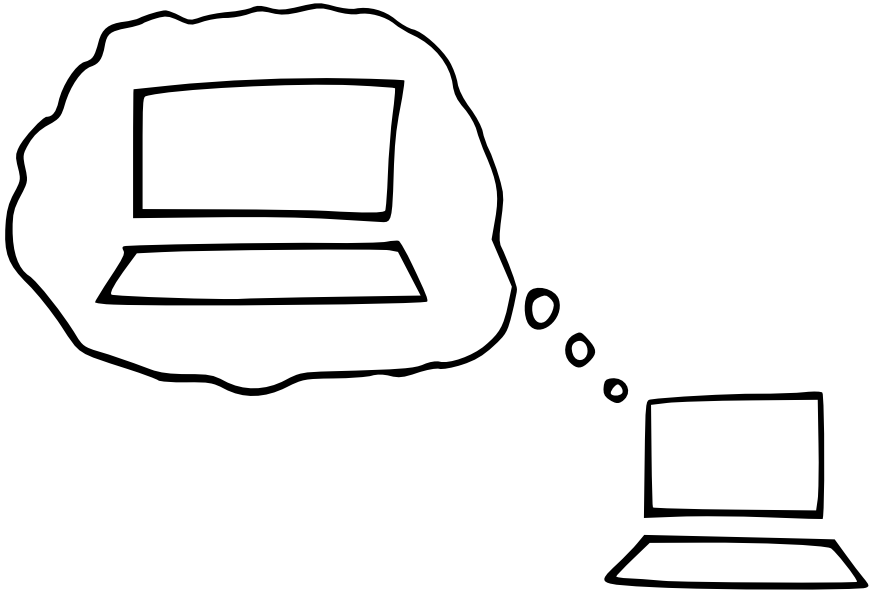


MACHINE'S POINT OF VIEW

AI's theory of human mind

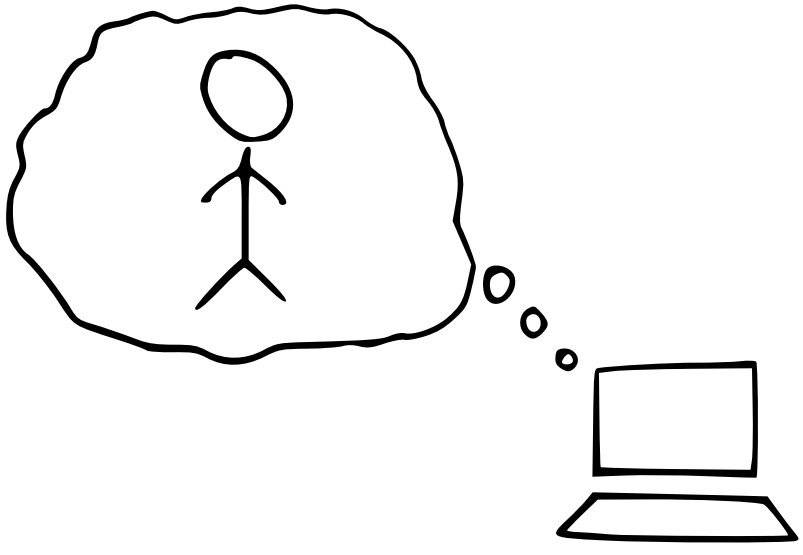


AI's theory of AI's mind



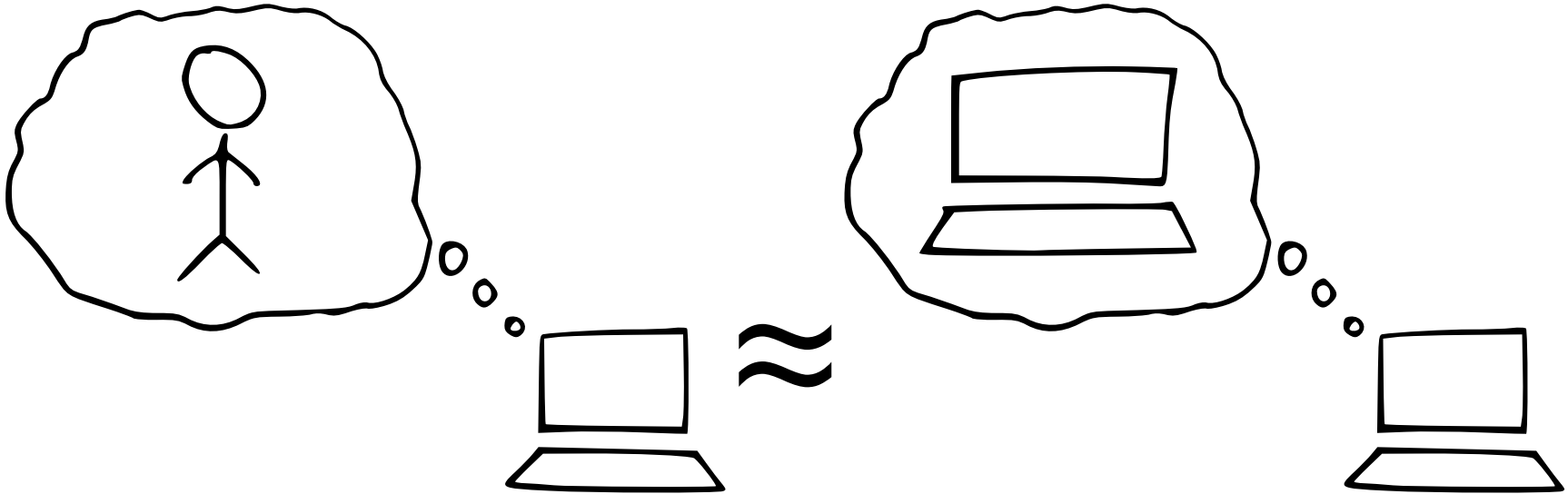
HOW DO MACHINES MODEL HUMANS?

How to infer reasons/goals/intent from observed behaviour?



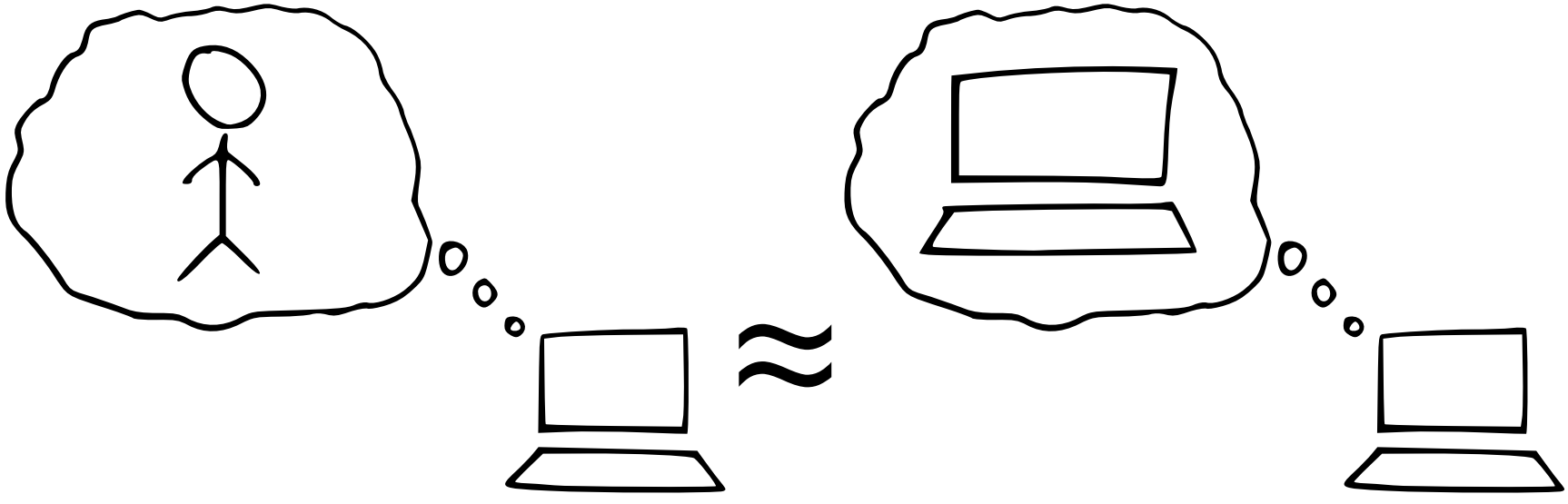
HOW DO MACHINES MODEL HUMANS?

How to infer reasons/goals/intent from observed behaviour?



HOW DO MACHINES MODEL HUMANS?

How to infer reasons/goals/intent from observed behaviour?



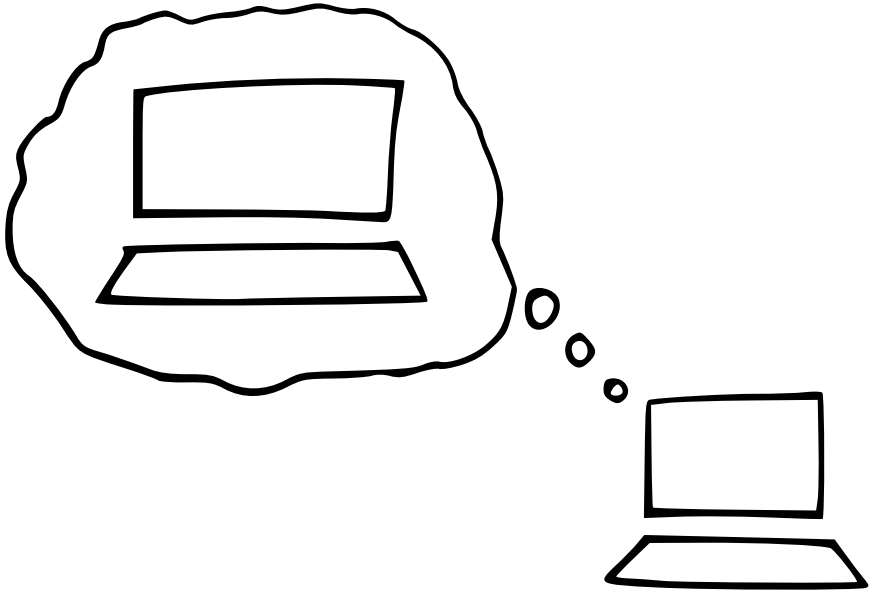
- Computational/bounded rationality,
- and other computational models in cognitive science.

HOW DO MACHINES MODEL MACHINES?

(among artificial autonomous agents)

How to infer reasons/goals/intent from observed behaviour?

Ill-posed inverse problem.



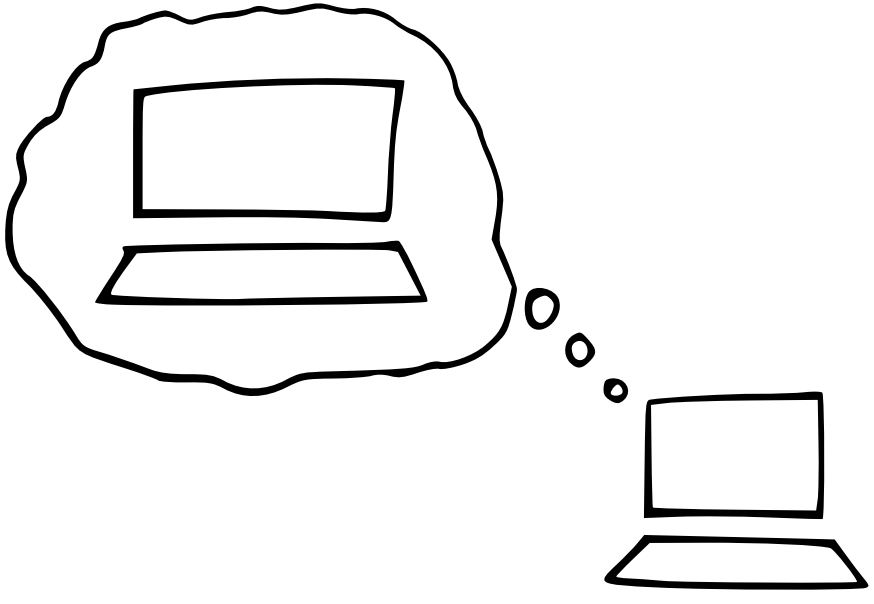
HOW DO MACHINES MODEL MACHINES?

(among artificial autonomous agents)

How to infer reasons/goals/intent from observed behaviour?

Ill-posed inverse problem.

- Deep learning based "model-free"/"black box" inversion.
- "Model based" inverse reinforcement learning.
- Multi-agent modelling.



MODELLING USER'S THEORY OF AI'S MIND IN INTERACTIVE INTELLIGENT SYSTEMS

Joint work with Mustafa Mert Çelikok, Pedram Daee, Samuel Kaski.

- Most current statistical models view **users as passive data sources**.
- Intelligent systems should acknowledge **the active strategic behaviour of humans**.

MODELLING USER'S THEORY OF AI'S MIND IN INTERACTIVE INTELLIGENT SYSTEMS

Joint work with Mustafa Mert Çelikok, Pedram Daee, Samuel Kaski.

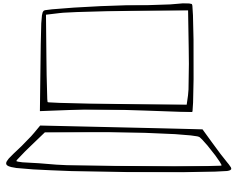
- Most current statistical models view **users as passive data sources**.
- Intelligent systems should acknowledge **the active strategic behaviour of humans**.

We propose **a user model that explicitly accounts for the user's theory of the AI's mind.**

- A nested probabilistic model of user's interest/intent based on sequential interaction,
- explicitly acknowledging the user as an active agent that has a model of the system.
- Currently for multi-armed bandit based systems.

User model that explicitly accounts for the user's theory of the AI's mind

1. User knows that the system has beliefs and/or state and can anticipate how these change with her actions.



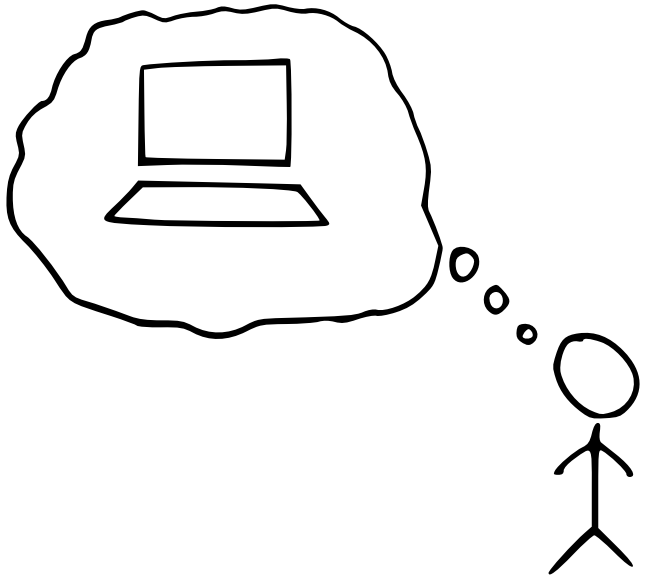
"Naive" multi-armed bandit

User model that explicitly accounts for the user's theory of the AI's mind

1. User knows that the system has beliefs and/or state and can anticipate how these change with her actions.

2. User plans her actions, based on the model of the system, to achieve good future states.

Markov decision process, with "naive" bandit providing transition dynamics



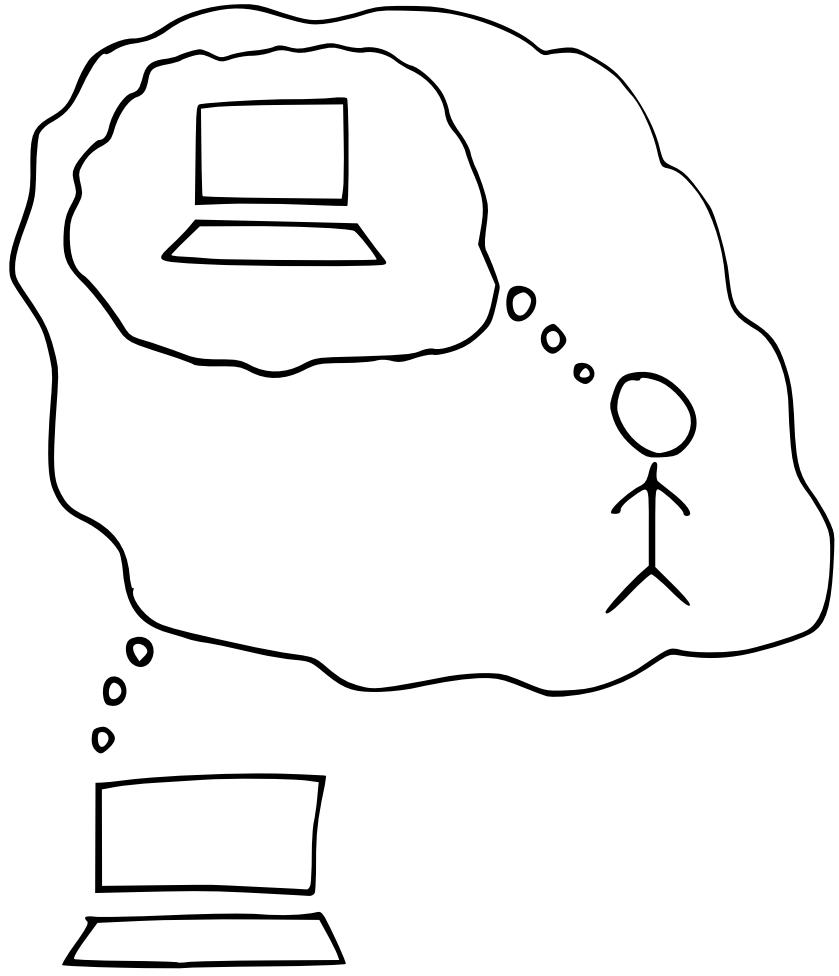
User model that explicitly accounts for the user's theory of the AI's mind

1. User knows that the system has beliefs and/or state and can anticipate how these change with her actions.

2. User plans her actions, based on the model of the system, to achieve good future states.

3. System interprets the observed user's actions based on the user model and infers the user's intent/interests/goals.

"Sophisticated" bandit, with observation model defined via the state-action value function of the MDP



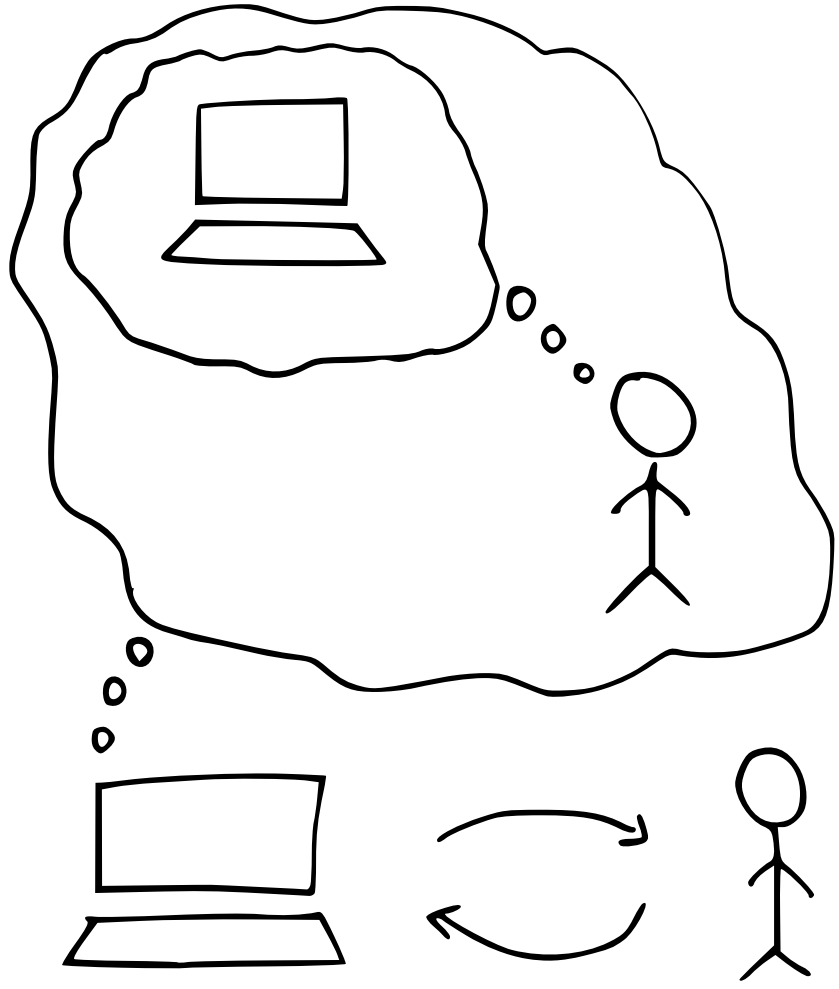
User model that explicitly accounts for the user's theory of the AI's mind

1. User knows that the system has beliefs and/or state and can anticipate how these change with her actions.

2. User plans her actions, based on the model of the system, to achieve good future states.

3. System interprets the observed user's actions based on the user model and infers the user's intent/interests/goals.

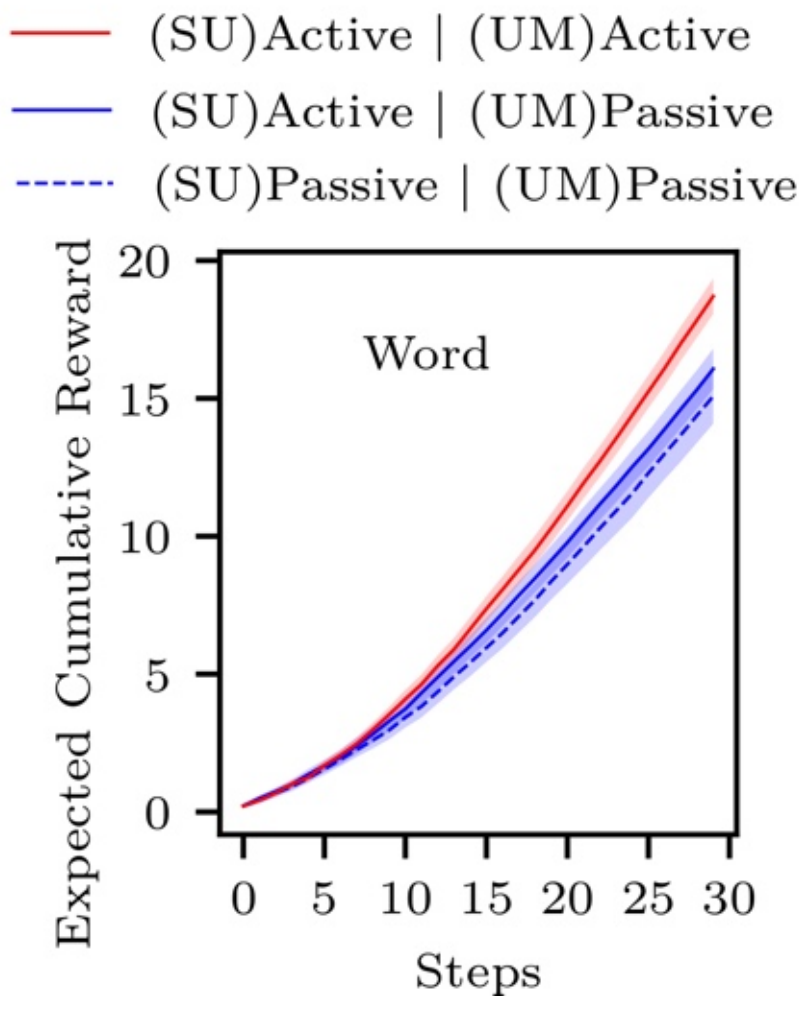
"Sophisticated" bandit, with observation model defined via the state-action value function of the MDP



SIMULATION EXPERIMENT

Task: search by sequential interaction where system suggest an item and user provides binary feedback for it.

- Active user can steer a predictable system towards the target item faster.
- If system models the user as active, performance increases further.

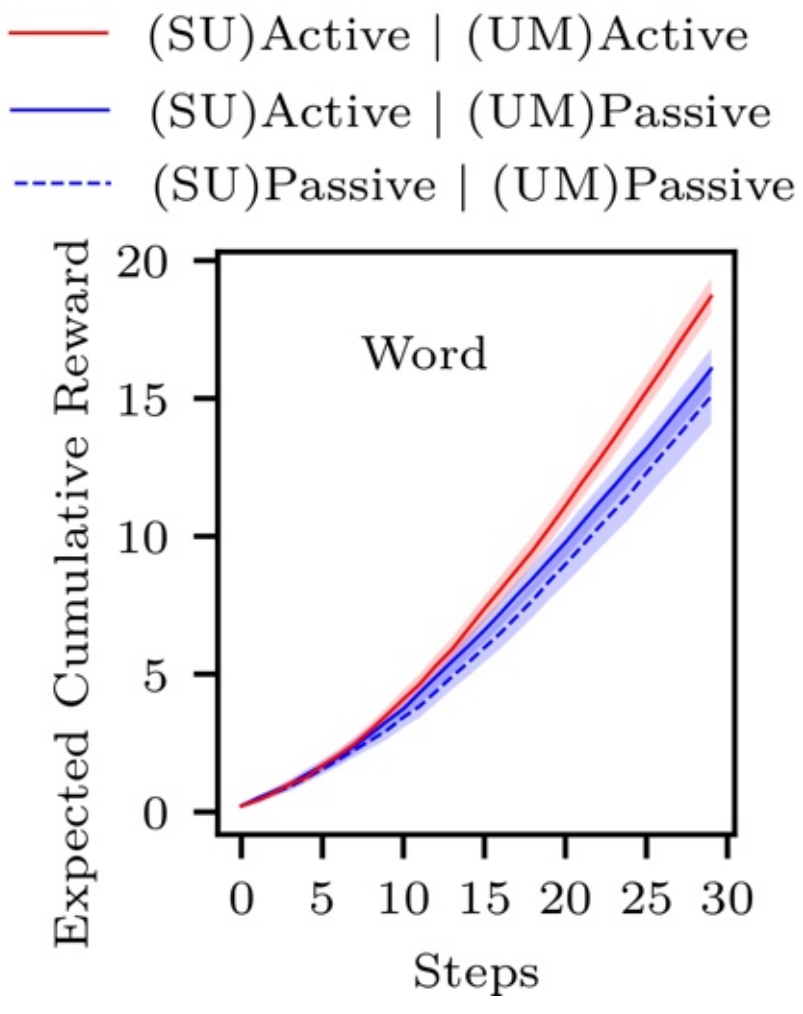


SIMULATION EXPERIMENT

Task: search by sequential interaction where system suggest an item and user provides binary feedback for it.

- Active user can steer a predictable system towards the target item faster.
- If system models the user as active, performance increases further.

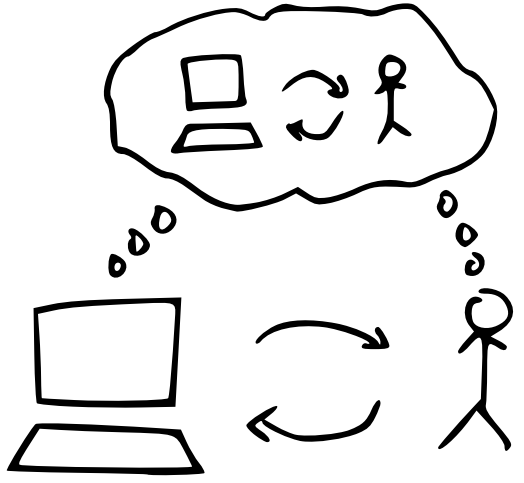
Come see our poster for details.



SUMMARY

Modelling theory of mind provides a path towards better human–AI collaboration.

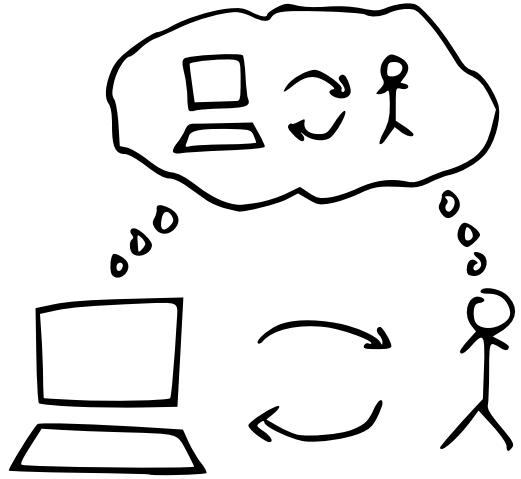
- Being understandable is easier if one knows what the other can understand (shared grounding).
- In interactive systems, understandability and predictability are important not only for user experience but also for the statistical models.
- Current limitations: simple settings, narrow view of theory of mind.



SUMMARY

Modelling theory of mind provides a path towards better human–AI collaboration.

- Being understandable is easier if one knows what the other can understand (shared grounding).
- In interactive systems, understandability and predictability are important not only for user experience but also for the statistical models.
- Current limitations: simple settings, narrow view of theory of mind.



Thanks!

tomi.peltola@aalto.fi / http://www.tmpl.fi / https://research.cs.aalto.fi/pml/

Pre-print of *Modelling User's Theory of AI's Mind in Interactive Intelligent Systems*: https://arxiv.org/abs/1809.02869