The Impact of Cultural Cognition on Explainable AI in Strategic Environments

Johns Hopkins SAIS – Anthropology for Strategists – Fall 2018

Leo Klenner

Outline

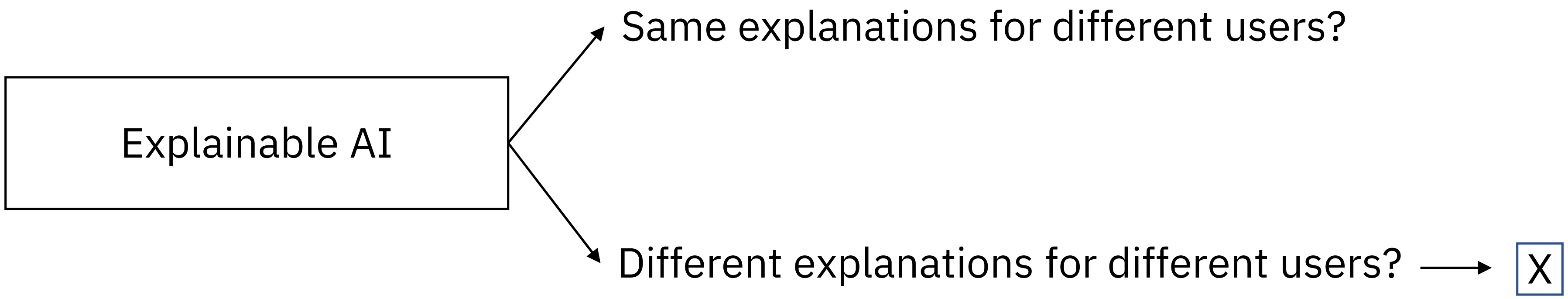
1. Background
2. Summary of the Paper
3. Structure of the Paper
    - > Explaination as Minimization v. Maximization
    - > Cognitive Anthropology – Culture as {Knowledge, Learning}
    - > Cognition in Combat
    - > Human-centered Explainable AI and Applications to Algorithms
4. Towards a Socio-Cultural Approach

Background

> 5-10 year strategic environment shaped by autonomous weapons systems (AWSs) with AI-based OODA (observe, orient, decide, act) capabilities

> Problem – Humans need to understand the decisionmaking of AI-based systems

> Solution – Enable systems to output explanations of logic that drives its decsions

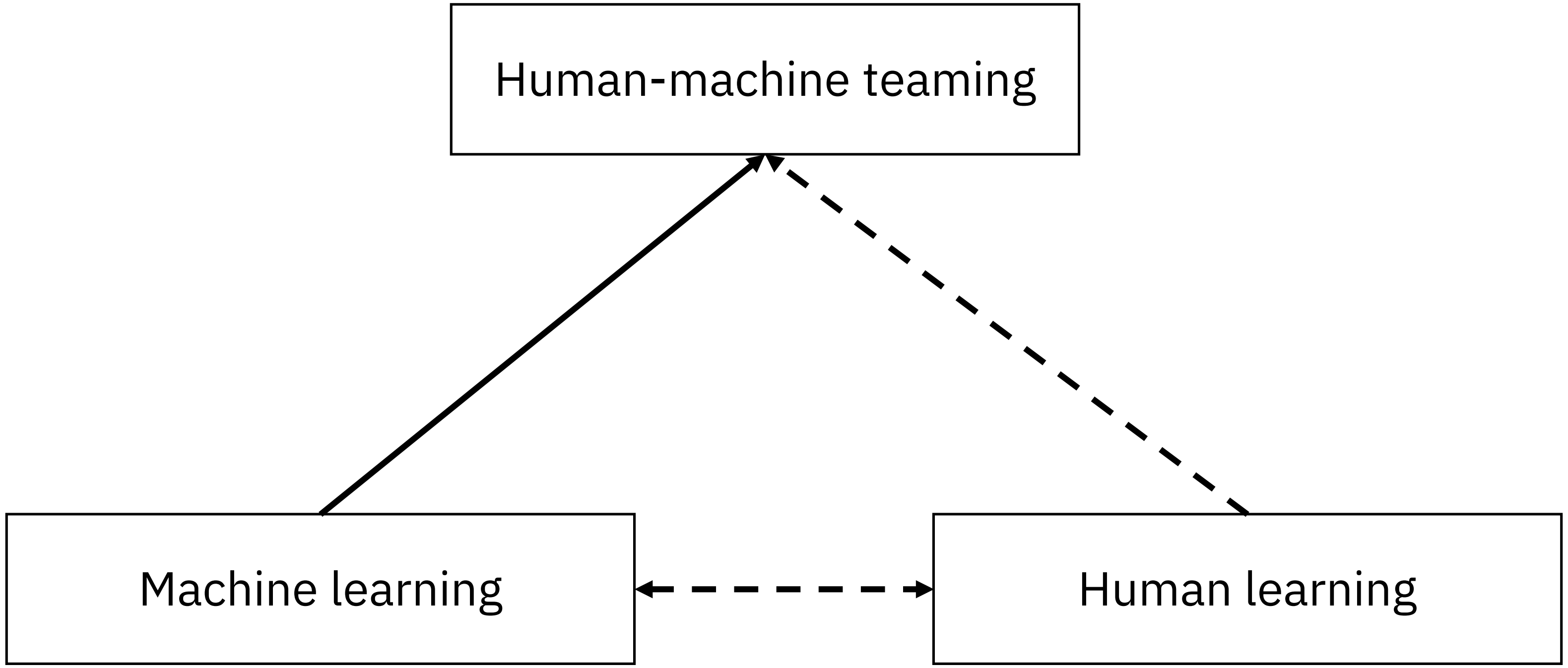


Explainable AI

Same explanations for different users?

Different explanations for different users? → X

Summary of the Paper

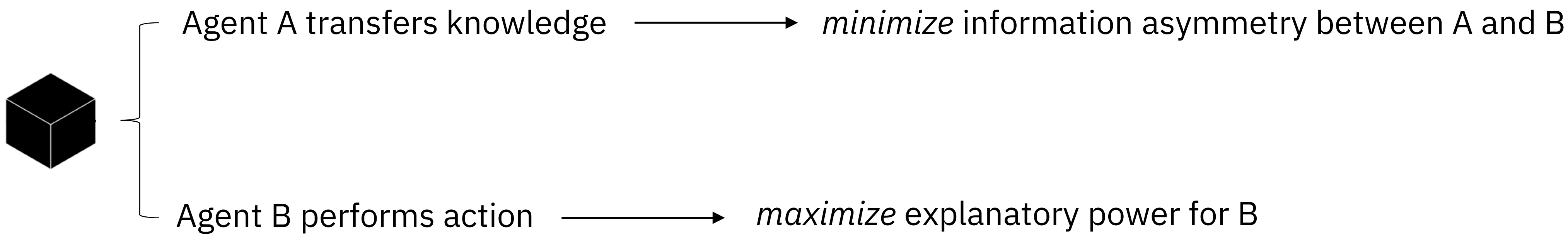


> For robust explainable AI, understand machine learning (different algorithms) *and* human learning (differnt groups -> cultures)

Structure of the Paper

Explanation as Minimization v. Maximization

- What do we do when we explain something to someone?
- Explanation as a blackbox – what constitutes an optimal explanation?



Agent A transfers knowledge → *minimize* information asymmetry between A and B

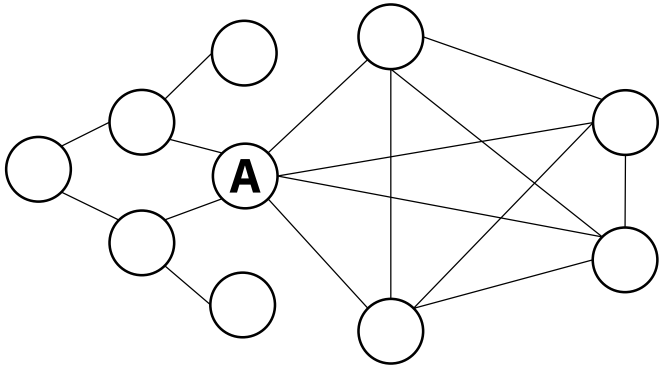
Agent B performs action → *maximize* explanatory power for B

- Tradeoffs for min v. max approach
  - Min: “Don‘t touch dead bodies because the Ebolavirus spreads through skin contact.”
  - Max: “Don‘t touch dead bodies because an evil ghost will enter your body and kill you.”

Cognitive Anthropology – Culture as {Knowledge, Learning}

> “Cognitive Anthropology is the study of the relation between human society and human thought. [It] studies how people in social groups conceive and think about the objects and events make up their world. Such a project […] inevitably leads to questions about the basic nature of […] cognitive processes.”
>
> Roy G. D‘Andrade. *The Developement of Cognitive Anthropology.* 1995.

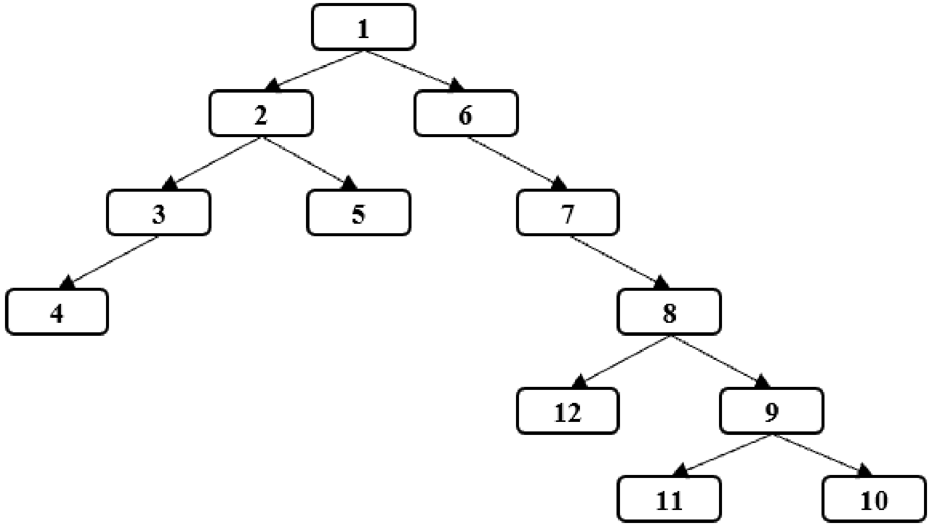
> Culture = information environment with distinct knowledge demand and supply



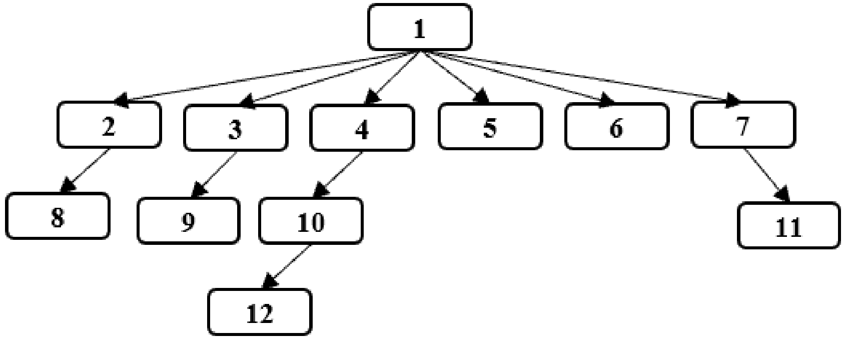
> **A** lives in two environments but each has different epistemic parameters

Cognition in Combat

- Optimal foraging for information in an environment changes based on culture
- Hence, different cultures have different optimal explanations



Foraging path of human accountant of AWS

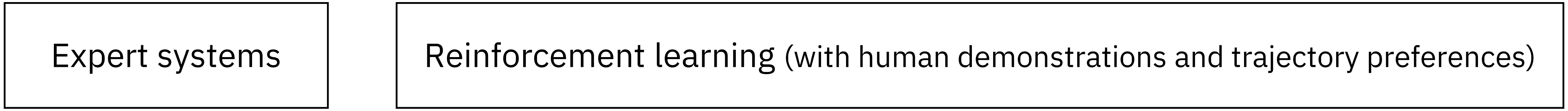


Foraging path of human team member of AWS

Human-Centered Explainable AI and App. to Algorithms

- Robust explainable AI =
  - optimized based on **sender** (AI – engineering) *and*
  - optimized based on **receiver** (humans – anthropology, cognitive science)
  - structured around on **joint human-machine learning**

- Joint-learning case study of two classes of algorithms

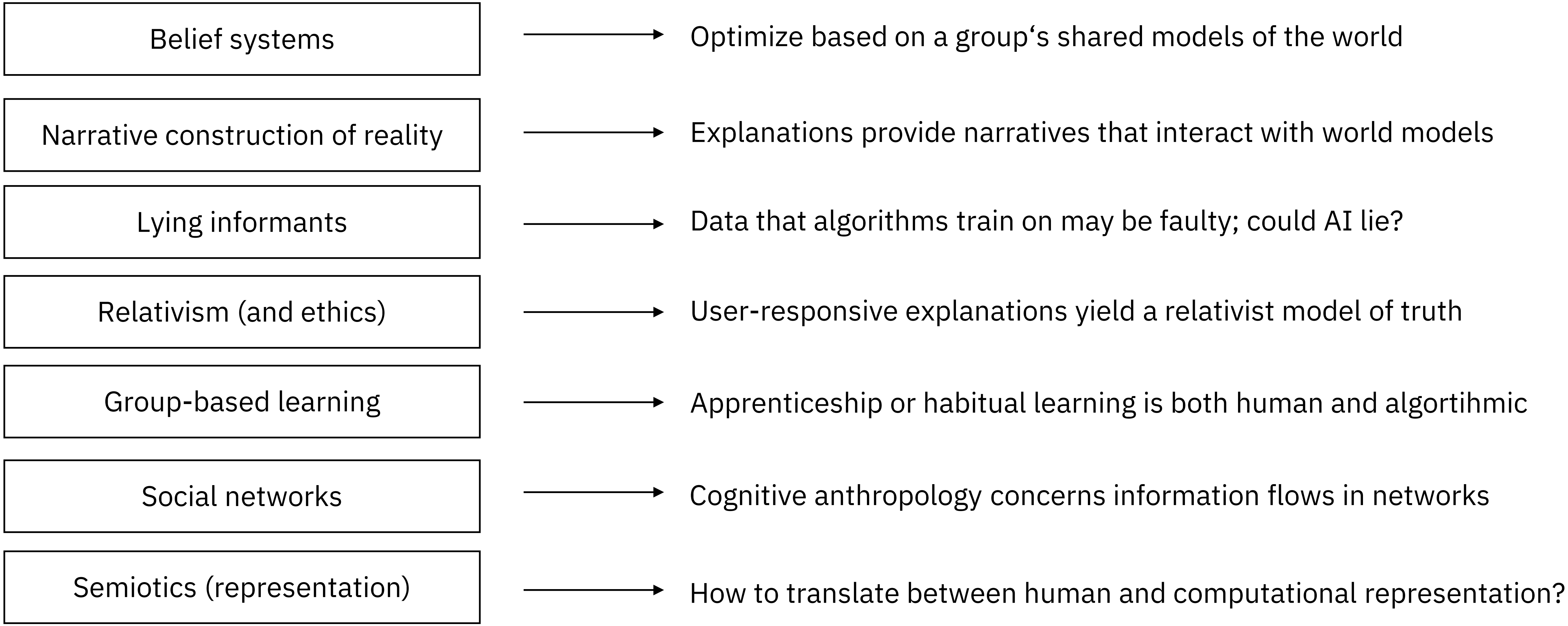

Expert systems

Reinforcement learning (with human demonstrations and trajectory preferences)

- **Conclusion**: Human-cent. explainable AI won‘t eliminate all ethical concerns, but can at least partially minimize them

Towards a Socio-Cultural Approach

Aspects of Anthropology in AI Research



Belief systems → Optimize based on a group‘s shared models of the world

Narrative construction of reality → Explanations provide narratives that interact with world models

Lying informants → Data that algorithms train on may be faulty; could AI lie?

Relativism (and ethics) → User-responsive explanations yield a relativist model of truth

Group-based learning → Apprenticeship or habitual learning is both human and algortihmic

Social networks → Cognitive anthropology concerns information flows in networks

Semiotics (representation) → How to translate between human and computational representation?