Susanne Hindennach
Contact: susanne.hindennach@vis.uni-stuttgart.de

Redirecting Theory of Mind for Explainable AI

Introduction

- Theory of Mind (ToM) is the ability to attribute a mind and beliefs, desires, intents
- Terms like *intelligence*, *learning*, and *thinking* attribute a mind to AI systems
- ToM directed at AI systems disregards the involved stakeholders' minds
- Redirecting Theory of Mind on the stakeholders puts the focus on stakeholders' thoughts and decisions when creating the AI systems

Attribution of a Mind in XAI research

- Assessment of current direction of ToM in XAI
- Text analysis of terms used in scientific work that introduces XAI methods using an established ToM coding scheme [1]
- Examples from the LIME paper [2] are: "Words that algorithm 1 considers important" and "that predictions are made for quite arbitrary reasons"

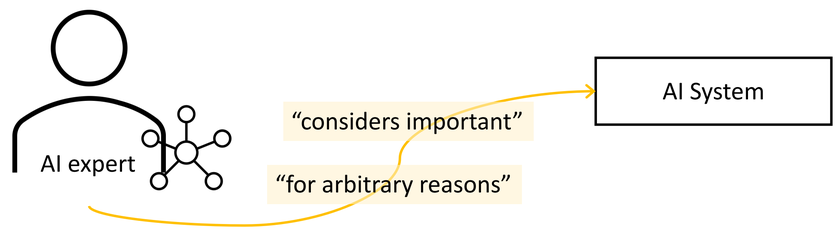

Figure 1. Current direction of ToM in XAI.

Explaining AI Systems by Inferring Stakeholders' Beliefs

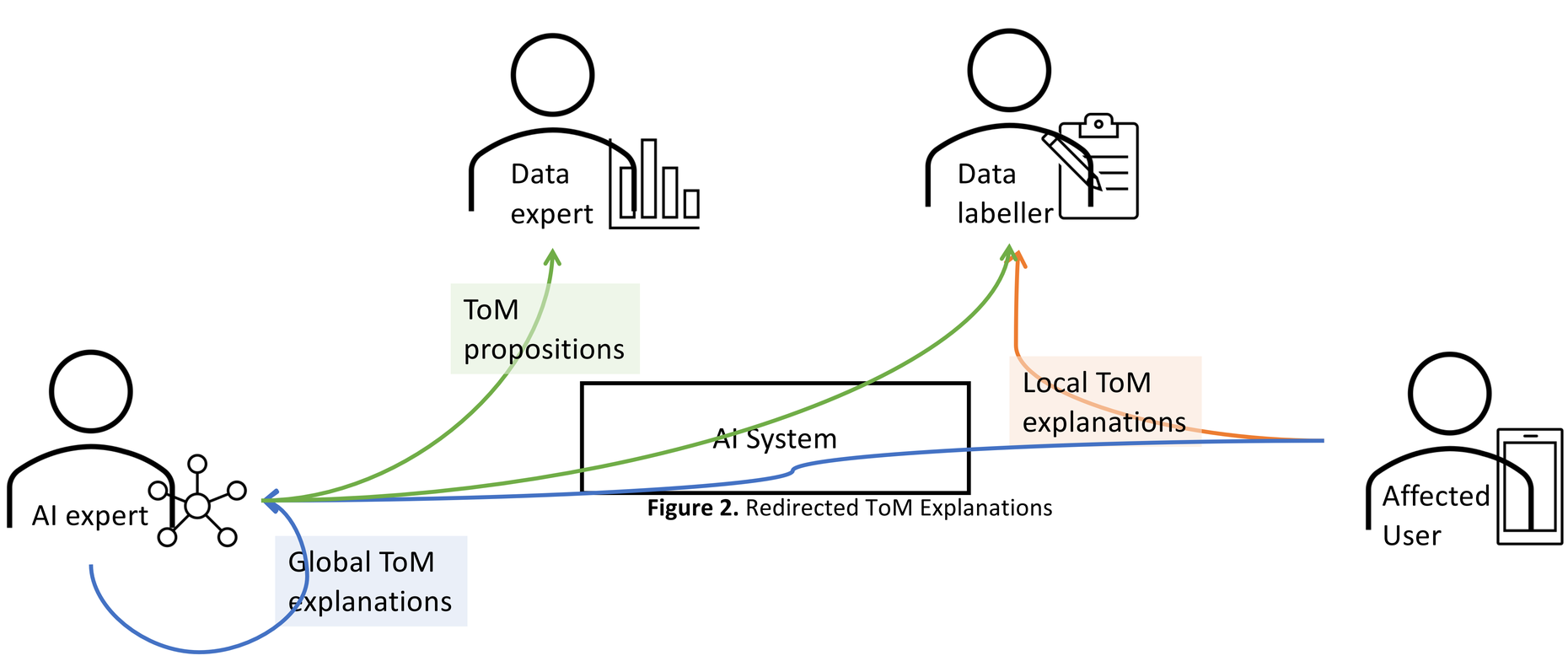

Figure 2. Redirected ToM Explanations

- **Local ToM explanations** are natural language rationales based on [3] that explain single predictions by referring to the data labellers' beliefs. They require the collection of explanations from the data labellers, which are used to train an encoder-decoder network.
- **ToM propositions** are beliefs held during AI system creation that centre the thought processes of data experts and data labellers. They are used guide the creation of AI systems with shared proposition tables.
- **Global ToM explanations** give an overview of the entire model by inferring the beliefs held by AI experts. They are selected from the ToM propositions using Bayesian model selection based on the AI expert's actions.

Contributions

- Assessment of mental state attribution in explainable AI
- Novel XAI methods that leverage computational models of ToM for explainable AI
- Shift in power by enabling affected users to question the (unconscious) assumptions and beliefs that were held during AI system creation

References

[1] Bertram Malle. 2014. F . EX A Coding Scheme for Folk Explanations of Behavior (Issue February).
[2] Ribeiro, M. T., Singh, S., & Guestrin, C. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. international conference on knowledge discovery and data mining
[3] Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., & Riedl, M. O. 2019. Automated rationale generation: A technique for explainable AI and its effects on human perceptions. International Conference on Intelligent User Interfaces