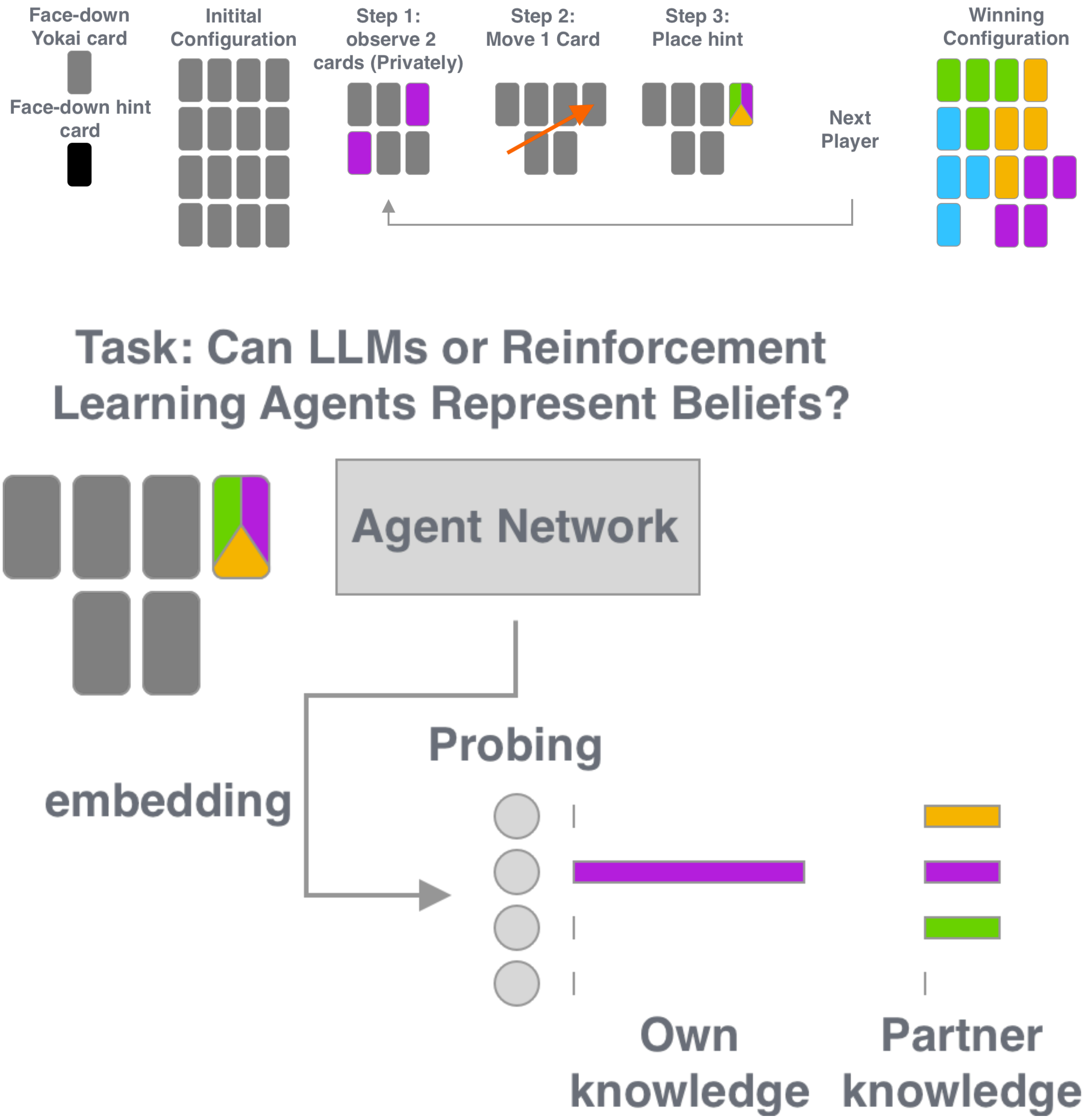
Probing for Theory of Mind in the Multi-Agent Cooperation Game Yokai



Description: The Yokai [1] board game has recently gained an interest as a simple test for the Theory of mind [4,3] capabilities of artificial agents. Yokai is a simple game that combines partial observability with a cooperative sorting task. The rules are sketched out in the accompanying figure but in essence: (0) 16 cards with one of 4 colors each are organized in a 4x4 grid face-down, (1) one player looks at two cards in private, (2) they then move one card, (3) they place a (1-, 2- or 3-color) hint, (4) the next player takes a turn. The game is won if all cards of the same color form a cluster. Any form of communication is forbidden, except for using hint cards. The game is hypothesized to require Theory of Mind reasoning as players need to reason about the moves of their partners while both have incomplete knowledge, i.e. reasoning about their moves "they moved the card to a card we both know is blue, therefore they must know that this card is also blue" and also about their believes "they moved a card I know is blue next to a card I know is green. They must mistakenly belief that they have the same color!".

Theory of Mind reasoning in games was explored in previous research [2] but Yokai has the unique property of only featuring very few elements that need to be tracked. This allows us to easily interface the game with text-based [6] and RL-based agents and makes probing work easier. To be specific, in this thesis we want to use tools from mechanistic interpretability (i.e. probing and others [5,7,8,9]) to explore whether artificial agents (LLMs and RL-based agents) can distinguish between their own and others knowledge and store a concept of belief in their latent representation in games versus texts [9]. This has previously been somewhat shown in text-based benchmarks [9]. Since LLMs are trained on huge amounts of text, we can never be sure whether part of these tests were in the training dataset but we can know that LLMs were likely not tasked with playing games during training. We are thus interested in studying them on Yokai using mechanistic interpretability tools and to evaluate the content of their embeddings for evidence of own- and other-beliefs.

To this end the following tasks are proposed:

- Built a text-interface for a Yokai Environment
- Explore an LLMs ability to play Yokai
- Hand design inputs for assessing LLM's Machine Theory of Mind capabilities
- Explore the representations via Linear Probing (see [7,8,9])
- Compare to results on text benchmarks [9]
- Train RL Policies on the Yokai environment and compare findings

Supervisor: Constantin Ruhdorfer

Distribution: 20% literature review, 50% implementation, 30% analysis/p>

Requirements: Good knowledge of deep learning and reinforcement learning, strong programming skills in Python and PyTorch and/or Jax, self management skills. An interest in interpretability and in understanding LLMs deeply is helpful.

Literature: [1] Fernandez, J., Longin, D., Lorini, E., & Maris, F. (2023). A logical modeling of the Yōkai board game. _AI Communications. The European Journal on Artificial Intelligence_, 1–34. doi:10.3233/aic-230050

[2] Bard, N., Foerster, J. N., Chandar, S., Burch, N., Lanctot, M., Song, H. F., ... & Bowling, M. (2020). The hanabi challenge: A new frontier for ai research. _Artificial Intelligence_, _280_, 103216.

[3] Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018, July). Machine theory of mind. In _International conference on machine learning_ (pp. 4218-4227). PMLR.

[4] Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind?. _Behavioral and brain sciences_, _1_(4), 515-526.

[5] Kosinski, M. (2023). _Evaluating Large Language Models in Theory of Mind tasks_. doi:10.48550/ARXIV.2302.02083

[6] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., … Wen, J.-R. (2023). A survey on large language model based autonomous agents. Retrieved from http://arxiv.org/abs/2308.11432

[7] K. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small, Nov. 2022. URL http://arxiv.org/abs/2211.00593.

[8] S. Marks and M. Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2023.

[9] Zhu, W., Zhang, Z., & Wang, Y. (2024). Language models represent beliefs of self and others. Retrieved from http://arxiv.org/abs/2402.18496.

Contact Us

pui-office@vis.uni-stuttgart.de

University of Stuttgart

Institute for Visualisation and Interactive Systems
Human-Computer Interaction
and Cognitive Systems

Pfaffenwaldring 5a
70569 Stuttgart
Germany