Synthesis of Interpretable and Obfuscatory Behaviors in Human-Aware AI Systems

Anagha Kulkarni
PhD Candidate, Arizona State University

The recent advances in AI have changed the landscape of various industries like finance, healthcare, marketing, education, etc. However, most of these deployed AI systems are ***not human-aware***.

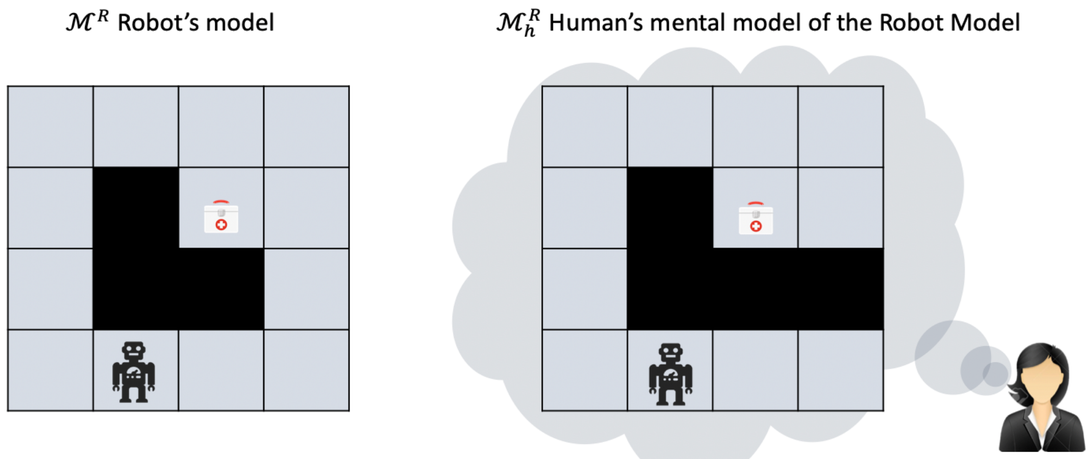
- To be human-aware, an AI system should reason over the ***human's mental model*** use it in its decision-making process.
- For example, *a smart car that expects the human driver to suddenly take control of the steering wheel without accounting for the human's response time is* ***not human-aware***.
- To avoid such incidents, ***human-awareness is crucial*** for AI systems!

Application domains that benefit from human-AI synergy:

- Decision support settings where an AI agent helps the human in a computationally challenging task. For e.g. ***providing assistance to pilots in the cockpit, providing suggestions to doctors in a clinical setup.***
- Commercial settings where humans and robots are co-workers. For e.g. ***factory floors, warehouses, restaurants.***
- Disaster response settings where it maybe physical unsafe for human. For e.g. ***collapsed structures, search and rescue.***

Why should an AI agent reason over human's mental model?

- For an AI agent operating in the presence of a human, reasoning with its own model may not always be sufficient.
- The human may have partial or inaccurate understanding of the environment/task/robot's capabilities leading to a different model.



Mental model acquisition:

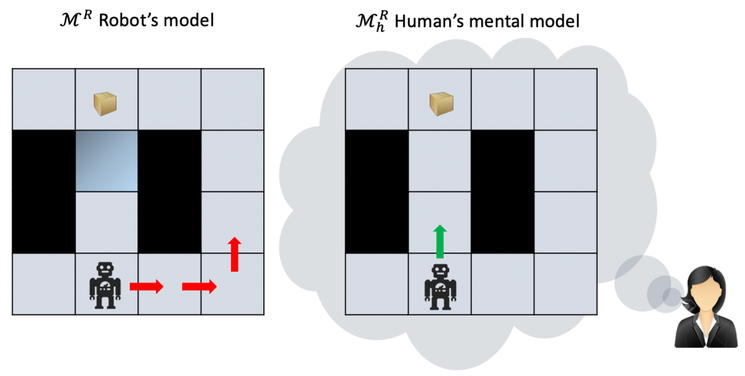
- Learn an approximation of the human's mental model [1].
- Construct a human mental model with desired representation by interacting with the users in the domain [2].

Once the mental model is available:

- Use it to ensure the AI agent's behavior is understandable to the human – i.e. ***be explicable***
  - Use the learned model to guide the AI agent's decision process [1].
  - Use the constructed model to minimize the distance between AI agent's pans and the human's expected plans [2].
- Communicate the model differences to the human – i.e. ***provide explanations***.

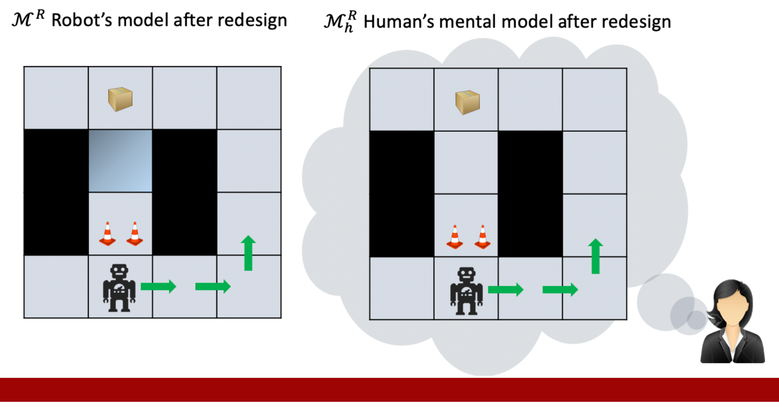
What if explicable behavior is infeasible or expensive?

- The environment might not be always conducive to explicable behavior.



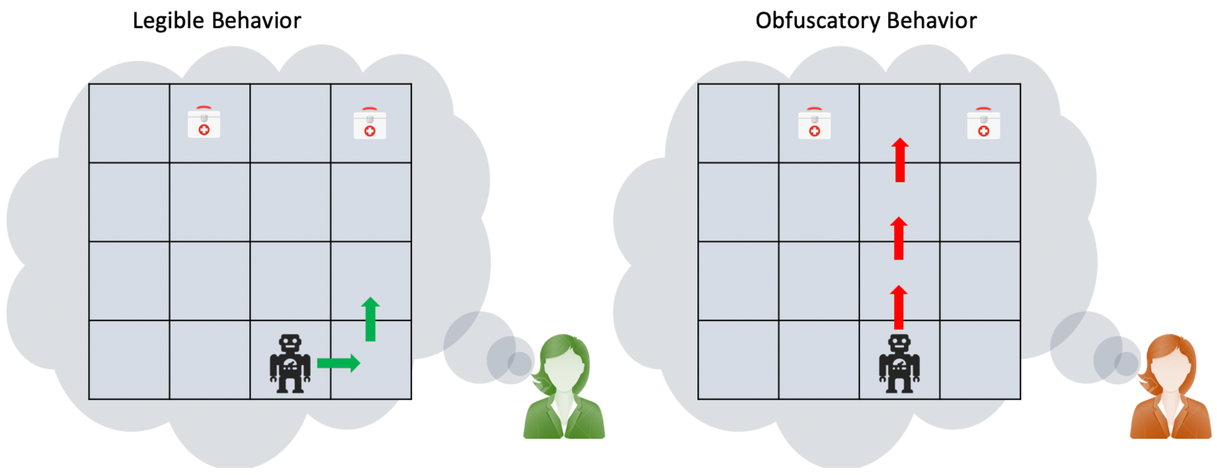
Environment Redesign:

- If tasks are repeatedly performed in an environment, the environment can be optimized to facilitate explicable behavior. [6]
- Also, the agent can perform ***longitudinally explicable behavior***. [6]



How can an AI agent update the human's mental model?

- The human's mental model may lack information about robot's objectives/capabilities/environment.
- The AI agent may want to convey this information to a human teammate as well as hide sensitive information from adversaries.



Updating human's mental model by accounting for human's perception limitations:

In a cooperative environment:

- Perform ***legible behavior*** that communicates information implicitly about agent's goals/plans. [3]
- Perform ***assistive behavior*** that minimizes human's workload without increasing her cognitive load.

In an adversarial environment:

- Perform ***obfuscatory behavior*** that hides information implicitly from an adversarial entity. [3]
- Perform ***secure obfuscatory behavior*** that is immune to rerun attacks from the adversary. [4]
- Perform behavior that is ***legible*** to teammates and ***obfuscatory*** to adversaries ***simultaneously***. [5]

References

[1] Plan Explicability and Predictability for Robot Task Planning (ICRA 2020)
[2] Explicability as Minimizing Distance from Expected Behavior (AAMAS EA 2019)
[3] A Unified Framework for Planning in Adversarial and Cooperative Environments (AAAI 2019)
[4] Resource Bounded Secure Goal Obfuscation (AAAI FSS 2018)
[5] Signaling Friends and Head-Faking Enemies Simultaneously: Balancing Goal Obfuscation and Goal Legibility (AAMAS EA 2020)
[6] Designing Environments Conducive to Interpretable Robot Behavior (IROS 2020)

Thanks to all my collaborators: S. Sreedharan, Y. Zha, S. Keren, T. Chakraborti, S. Vadlamudi, Y. Zhang, S. Srivastava, M. Klenk, D. Smith, S. Kambhampati.

Some of these works were also presented at ***AAAI 2020 Tutorial*** - https://yochan-lab.github.io/tutorial/AAAI-2020/
Webpage: https://anaghak.github.io/