What actually was done?

Humans have the ability to observe other humans and can make assumptions about their mental state, such as desires, beliefs, and intentions. On the other hand, when using a machine learning algorithm to solve complex tasks, the resulting solution is mostly difficult to grasp for humans and is labelled as a black-box. Therefore, Rabinowitz et al. attempt to build a Theory of Mind neural network (ToMnet) to learn "a strong prior model for agents' future behaviour, and, using only a small number of behavioural observations, can bootstrap to richer predictions about agents' characteristics and mental states" (Rabinowitz). In other words, they design a neural network to observe, track, and therefore try to understand the black-box of another machine learning structure with ultimate the goal to be able to predict future behavior of a novel agent that they have never met before.

What was the experiment (with numbers and results)?

Rabinowitz et al. created "a number of different species of random agents" (Rabinowitz) that move around in a virtual room and collect colored boxes. Mainly there were three different species: "One couldn't see the surrounding room, one couldn't remember its recent steps, and one could both see and remember" (Hutson). The goal for the ToMnet after observing many agents was to correctly identify an agent's species which would allow for a prediction of future behavior. Therefore, the agents had to reach a subgoal on a gridworlds environment (3D visual environments) and then consume a preferred object that was different form agent to agent. 120 (3 species x 4 preferred objects x 10 initial random seeds) agents were trained and used in the experiment. Not only was the ToMnet able to grasp the different types of personalities but could also discover unexpected substructures.
In a further experiment, they considered the fact that humans sometimes depict reality different than it actually is and build their behavior on top of this false belief. The goal for this additional exeperiment was to identify such false beliefs by the observed agent. Indeed, the ToMnet was able to identify unusual behavior patters of the agent that did not match their actual belief while the agent himself did not notice it.

What ideas from ToM were used?

The general theory of mind which predicts the common behavior of all agents in the training set and the agent-specific theory of mind which takes a single agent's character as a distinct one from others.

Rabinowitz et al. also tried to replicate the ability of young human beings to depict other people's beliefs and whether it diverges from reality.

What is the solution architecture?

The solutions architecture is neural network composed of three modules, i.e. the character net, the mental state net, and the prediction net. While the character net focuses on past trajectories, the mental net parses the current trajectory up to time t-1. The prediction net then leverages these two modules to predict an agent's behavior.