Guy Hoffman - Fall 2003

A Reading of

Alison Gopnik and Henry M. Wellman's

Why the Child's Theory of Mind Really Is a Theory

Gopnik and Wellman argue primarily against the "Simulation Theory" (ST) in child development, which claims that children's understanding of other people's desires, beliefs and perceptions are simulations run on the platform of their own minds. ST does not presume, and actually does not need, a representation of another agent's mental state, but rather proposes a process of *perspective-taking*. In this process, the child, when trying to predict another person's actions based on real-world facts, runs the same facts through its own mind - only from the other person's perspective - and then examines the output of this simulation.

In the counter-argument proposed in this paper, called "Theory Theory" (TT), the child does actually have a representation of other people's minds, and has in fact a *theory* about how these other minds operate. The child's theory develops over the first few years of his life, from a simpler desire/perception theory to a full representational desire/belief/perception one. (In fact, the full theory could be viewed as a belief-only theory, since desires and perceptions can be subordinated to beliefs).

In addition, Gopnik and Wellman show that the developmental progress is similar to a evolution of a scientific theory, much along the lines of Kuhn - reinterpretation of evidence and auxiliary hypotheses on the way to a full paradigm shift.

The TT argument is well supported, mainly by findings that indicate the children under 3 do not bias towards ego-centric errors of belief interpretation, but rather towards desire-perception-centric mistakes. The understanding of beliefs and mental representations in others develops in parallel to one's own.

I find these arguments quite convincing, especially since the simulation theory seems unintuitive and overly complex. With today's methods, one would also expect to have found functional brain imaging support for the simulation theory, although I don't know whether such evidence exists. The types of mistakes that come up in the experiments look like solid support and in addition the belief-based "theory theory" does comply to what it subjectively feels like when trying to predict other people's actions.

The implications for artificial agents are complex: our natural tendency is to design behaviorist machines, drawing directly from perceptual (vision, speech-rec) data and mapping these to actions. Even when we want to model beliefs, it is likely that these are represented in similar symbols as the machine's own beliefs, and finally it seems that every sort of algorithmic interpretation of a human's mental state is basically a simulation using the same tools that the machine agent has. All is not bleak, though. A good lesson to be learned from this paper's analysis is the need to incorporate *beliefs* in addition to desires and perceptions in the human collaborator's artificial model. TT also urges us to make sure to model the human agent as an internal entity rather than analyzing him simply as a stream of perceptual input.