MIND-READING | Michaelmas 2018 | Maarten Steenhagen (ms2416@cam.ac.uk)
http://msteenhagen.github.io/teaching/2018mre/

Session 2: **Theory-theory**

Carruthers, Peter (2004), 'The case for theory-theory', in *The Nature of the Mind: An Introduction*. London: Routledge, 227-263.

Theory of Mind: Theory vs Simulation

How do we make judgements about the mental states of other human beings? This is now not a normative question about justification, but a descriptive question about mechanism or functioning. How should we answer the question? Imagine the following group discussion:

> **A:** "When we make judgements about the mental states of others, we rely on a theory about the role of mental states."
> **B:** "No, we don't rely on such a theory in these our judgements."
> **C:** "When we make judgements about the mental states of others, we rely on simulations of these mental states in ourselves and base our judgements on this."
> **D:** "No, we don't rely on such simulations in these our judgements."

There are different ways of making sense of the options. Here's one. Theory Theorists will side with **A**, while Simulation Theorists will side with **C**. But this leaves room for (i) those who side merely with **B** and **D** and who do not believe **A** and **C**, and (ii) those who believe **A** and **C** but disagree with **B** and **D**.

Our question is a descriptive one. One would think that settling the matter should be an empirical matter. Experimental findings bear on this, e.g. the false belief task. Alison Gopnik claims that young children operate like 'little scientists' and gradually construct a theory of mind. This is controversial.

What contribution can philosophical argument make to this discussion? The philosophical debate is about whether it is even *possible* for **A** or **C** to be right, or whether **A** or **C** *must* be right.

Versions of Theory Theory

There are several varieties of Theory Theory. First we see *externalists* and *internalists*:

> On the externalist reading of 'theory theory', our everyday talk about mental states implicitly constitutes a theory of mind: folk psychology (external). On the internalist reading of 'theory theory', our everyday capacity to predict and explain behavior is underpinned by an internally represented theory of mind: folk psychology (internal). Unfortunately, theory theorists are not always clear as one might hope about which sense of 'theory theory' they are endorsing. (Ravenscroft 1997, 2)

Externalists take the theory in question simply to be the folk psychology that we find in our culture (e.g. Lewis 1972). Others take the theory to be represented in our minds, without committing to an identity of this theory with folk psychology (e.g. Stich and Nichols 1995).

A further dimension of variation is the *acquisition* of the theory. How did we all get it?

> If our theory of mind isn't acquired by a process of theorizing, and it isn't acquired through simulation, then what other alternatives are there? One possibility is that it is taught. Perhaps theory of mind is learned much as chemistry is learned—not through theorizing, but through instruction by older children and adults. There is not a shred of evidence that any such instruction takes place, however; and quite a bit of evidence that it doesn't. The idea that theory of mind is taught to children is a complete non-starter, in fact. Which leaves biological maturation (innateness) as the only remaining alternative. (Carruthers 2004, 258)

Some believe that the theory in question is innate, others think it is acquired. If it's acquired then either you acquire it by yourself (by theorising, simulating,…), or others teach you the theory (or implant it in some other way!).

Argument for theory theory

The theory theorists needs to show that we cannot make judgements about the mental states of others without relying on a theory about the role of mental states.

One strategy focuses on the concepts used in these judgments. If applying the concepts used in making judgements about the mental states of others relies on a theory about the role of mental states, then those judgments themselves rely on such a theory.

David Lewis (1972) suggests that meaning of theoretical terms (T-terms) is determined by how they figure in the theory in which they are used alongside observational terms (O-terms):

> Suppose we have a new theory, *T*, introducing the new terms $t_1 \ldots t_n$. These are our T-terms. (Let them be names.) Every other term in our vocabulary, therefore, is an O-term. The theory *T* is presented in a sentence called the *postulate* of *T*. Assume this is a single sentence, perhaps a long conjunction. It says of the entities—states, magnitudes, species, or whatever—named by the T-terms that they occupy certain *causal roles*; that they stand in specified causal (and other) relations to entities named by O-terms, and to one another. (Lewis 1972: 253)

Mental state terms are not observational. And at least on a functionalist conception of the mind, mental states are defined by their causal roles. So the only way of making sense of mental concepts is by conceiving them as theoretical terms. If we follow Lewis, we must assume that understanding the meaning of a mental concept requires grasp of some theory about their role. So when we make judgements about the mental states of others, we rely on a theory about the role of mental states.

Arguments against theory theory

If we make judgments about the mental states of others through applying a theory, then we will locate the particular case among a range of possible ones. This means that our judgments is an inference to the best explanation: this case *x* most closely fits this theoretical possibility *P*. However, "our attitude to the mental states of others isn't tentative and indefinitely revisable in the way that our attitude towards a scientific theory usually is" (Carruthers 2004, 261). We can perhaps overcome this objection by adopting a nativist version of theory theory.

Jane Heal (1996) presents a different objection: as an explanation of our ability to make judgements about the mental states of others, the theory theory has to assume that we have a (tacit) grasp of a systematic and general theory of *relevance*.

> The theory-theorist is committed to the claim that we have—tacitly at least—solved an extremely important precursor problem to the famous Frame Problem in Artificial Intelligence, namely the problem of providing a general theory of relevance. And this claim is highly implausible.

The Frame Problem here is understood as a general epistemological problem for any kind of computation-based intelligence. As Fodor puts it: "How ... does the machine's program determine which beliefs the robot ought to re-evaluate given that it has embarked upon some or other course of action?" (1983, 114)

> As we stressed earlier, we can cope when circumstances are not normal and we understand very well that others can do so too. So if our imagined psychological theory is to account for our competence in these cases it must give systematically organised insight into the difference between our responses in usual and unusual cases, i.e. insight into a whole range of world view/question pairs and their possible upshots. It must specify the range of psychological factors which influence thoughts and decisions in response to given questions; it must lay out how they interact; it must say why some are important to outcomes in some settings and not in others; and it must be able to tell us how and why things would have been different, given this or that variation in the starting conditions. But given epistemological holism and our actual rationality, what all this amounts to is precisely a general and systematic theory of relevance. (Heal 1996, 83)

The worry for theory theory here is that in the case of human psychology, flexibility and unusual circumstances are the norm.