CellPress

Deconstructing and reconstructing theory of mind

Sara M. Schaafsma[1], Donald W. Pfaff[1], Robert P. Spunt[2], and Ralph Adolphs[2]

[1] The Rockefeller University, 1230 York Avenue, New York, NY 10065, USA
[2] California Institute Of Technology, 1200 East California Boulevard, Pasadena, CA 91125, USA

Usage of the term 'theory of mind' (ToM) has exploded across fields ranging from developmental psychology to social neuroscience and psychiatry research. However, its meaning is often vague and inconsistent, its biological bases are a subject of debate, and the methods used to study it are highly heterogeneous. Most crucially, its original definition does not permit easy downward translation to more basic processes such as those studied by behavioral neuroscience, leaving the interpretation of neuroimaging results opaque. We argue for a reformulation of ToM through a systematic two-stage approach, beginning with a deconstruction of the construct into a comprehensive set of basic component processes, followed by a complementary reconstruction from which a scientifically tractable concept of ToM can be recovered.

What is theory of mind?

The term, together with an approach for measuring it through the ability to attribute false beliefs, was first introduced in a highly influential article in 1978 [1]. Since then, an ever-increasing number of studies have been published (Figure 1) probing the emergence of ToM in typical human development, debating its possible presence in nonhuman animals, and diagnosing its breakdown in diseases such as autism spectrum disorders. Many of these studies have employed neuroimaging methods to identify the neural correlates of ToM, and their results have fostered the view that ToM relies on a specific set of brain regions now commonly known as the ToM network. The original usage of the term ToM (to infer the representational mental state of another individual, such as a belief or intention) already encompasses a diversity of processes, and the experimental approaches currently used often engage a large number of additional abilities whose association with ToM is not always appropriate (Box 1). Confusion arises because many publications (i) implicitly treat ToM as a monolithic process, (ii) refer to a single brain network for ToM, or (iii) conflate varieties of ToM. While we will continue to use the term ToM here, it should be noted that this is merely for convenience in exposition, not an endorsement of current usage. Our aim is more a general call to action than a specific prescription, however; consequently we sketch a broad research program rather than tackle its implementation.

The problem

Humans all have a competence to make sense of the observed behavior of others, a competence shared with many other animals. How exactly we manage to do this is less clear, and is probably less similar to how other animals do it. For one thing, we can think and talk about it; the concepts we employ when we do so are part of our folk psychology (indeed, it may be that the concepts develop in service of our need to talk about them [2]). The processes that enable us to think about other people's minds, in turn, are yet another matter. Debate has focused on whether these psychological processes are analogous to those involved in constructing a scientific theory (the theory-theory of ToM [3], closely related to cognitive ToM and often invoking a module for ToM [4]), or whether they involve more intuitive ways of simulating what is taking place in the other person (the simulation-theory of ToM, closely related to empathy and emotional ToM [5,6]). This distinction among processes is thought to be reflected in distinct brain networks that can be revealed in functional neuroimaging studies (the ToM network versus the mirror neuron system, respectively) [7], with some schemes for relating them to one another (e.g., [8]). In some instances, additional components of ToM are added, including executive control processes, and several other dual-process ways of carving up the conceptual landscape are often invoked (see further below). Humans likely use a mix of strategies that cut across all these processes to figure out other people's minds [9,10].

The different levels of description, together with the different terms used, make it difficult even for experts from different fields to navigate both what is meant by ToM and how to study it using scientific methods [11] (Box 1); to the uninitiated, the topic becomes bewildering. Even a preliminary survey of recent papers illustrates the problem that the field faces: some usages of ToM pertain to early cognitive development, whereas others pertain to adult social cognition; some refer to understanding of the self, whereas others refer to the perception of others; some refer to logical inferences, whereas others refer to emotional or empathic reactions. The term ToM is used interchangeably with mentalizing or mindreading [12], mind perception [13], and social intelligence [14], to name only a few. This diversity of terms used is probably telling: different investigators have different concepts in mind. Focusing only on

Corresponding author: Adolphs, R. (radolphs@hss.caltech.edu).

1364-6613/
 http://dx.doi.org/10.1016/j.tics.2014.11.007

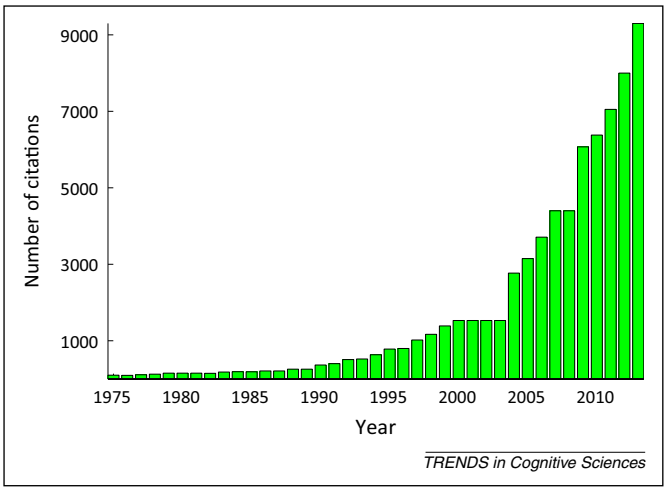

Figure 1. Articles referencing theory of mind have increased markedly in recent years. Estimates are based on a per annum Google Scholar search (http://scholar.google.com) for articles that use the exact phrase 'theory of mind'.

the many papers that study ToM using neuroimaging yields no less heterogeneity (Boxes 1 and 2, Figure 2). The problem is that these differences generally go unarticulated, and their basis is often not well grounded.

Difficulties in clarifying our concept of ToM have been there all along. Comparative studies in species ranging from dogs (e.g., [15]), to corvids (e.g., [16]), to, most famously, great apes [1,17,18] have all left ongoing debates in their wake about the status of the psychological processes those species use. They all exhibit behaviors that certainly suggest that they are using ToM, but it has been elusive to triangulate the actual processes involved. The discrepancies among views on the status of ToM are extreme. Some tasks, especially in developmental and comparative psychology, have taken great pains to isolate highly specific competences (Table 1). And some localizers used in neuroimaging studies result in highly reproducible patterns of brain activation. It is perhaps unsurprising, then, that some claim,

> *Unlike many aspects of higher-level cognition, which tend to produce small and highly variable patterns of [brain] responses across individuals and tasks, ToM tasks generally elicit activity in an astonishingly robust and reliable group of brain regions* [19].

By contrast, it has been suggested that ToM could be deconstructed into other processes, with no domain-specificity at the core of human ToM ability at all:

> *....dedicated mentalizing processes may not be necessary....the same jobs can be done just as effectively by domain-general processes, such as those involved in automatic attentional orienting and spatial coding of stimuli and responses* [20].

There is already a body of literature criticizing how ToM is used and investigated [21,22]. Proposed solutions have ranged from banning the term altogether to reserving it for a very specific task [23]. We have no intention of eliminating the term ToM, but it does need radical revision. We believe that our current concept of ToM hinges on the essence of a mental representation of minds, but that a scientific concept of ToM needs to disassemble that essence into a collection of simpler processes. Furthermore, we think this could actually work in a way that

Box 1. Tasks typical for studying ToM in fMRI studies[a]

False belief attribution
Tests the ability to attribute mental states (beliefs, intents, desires, etc.) to others and understand that those mental states may be different from one's own.

15 studies	False belief versus false photograph
1 study	False belief and subjective preference
7 studies	False belief versus true belief
3 studies	False belief versus physical reality
10 studies	Story-based format for false belief, with various comparison tasks

Trait judgments
Tests the ability to judge whether a specific trait is descriptive of a particular person.

12 studies	Read written descriptions of a person that convey a trait
3 studies	Read trait descriptions accompanied by a photo of the face (all with a variety of control tasks)
4 studies	Other judgments versus self-judgments
3 studies	Other judgments versus diverse mental state judgments
3 studies	Self-judgments
1 study	Trait judgments about animations

Strategic games with another person (or computer)

9 studies	Compete or cooperate; contrast human versus computer
2 studies	Play with another human, but no computer contrast
3 studies	Only low-level control conditions
3 studies	No contrast, only model-based fMRI

Social animations

14 studies	Shapes moving intentionally versus shapes moving physically/randomly
3 studies	Cartoons, high-level stories
3 studies	Causal, but not social relationships conveyed

Reading the mind in the eyes task
Tests the ability to recognize mental states based on just the area of and around the eyes.

10 studies	Mental state judgments versus physical judgments on photos of eyes
2 studies	Basic emotion judgments only

Rational actions
Tests the ability to infer mental states.

10 studies	Attributing intentions from nonverbal material (why versus how)
3 studies	Only passively watch actions

[a] Adapted from [29].