Social Behavior: Theory of Mind 105

Sara M. Schaafsma, Donald W. Pfaff, Robert P. Spunt, and
Ralph Adolphs

Contents

Abstract

The concept of how individuals perceive and comprehend the intentions of other individuals is presented and criticized. A road map toward the discovery of neural mechanisms for same is laid out in detail.

Keywords

Social perception · Empathy · Temporoparietal junction · Social brain

S. M. Schaafsma (✉) · D. W. Pfaff
Laboratory of Neurobiology and Behavior, The Rockefeller University, New York, NY, USA
e-mail: sschaafsma@rockefeller.edu; pfaff@mail.rockefeller.edu

R. P. Spunt · R. Adolphs
Division of Humanities and Social Science, California Institute of Technology, Pasadena, CA, USA
e-mail: bobspunt@gmail.com; radolphs@hss.caltech.edu

D. W. Pfaff et al. (eds.), *Neuroscience in the 21st Century*,
https://doi.org/10.1007/978-3-030-88832-9_165

Brief History

Neurobiological studies of social behaviors began with analyses of mechanisms of simple animal behaviors. For example, see ▶ Chap. 97, “Parental Behavior.” Then, following the lead of experimental psychologists, more complex social behaviors were considered, with an emphasis on social behaviors of humans. Human social behavior requires that we sometimes recognize the beliefs or intents of those with whom we are interacting. Following a literature in philosophy of mind and developmental and comparative psychology and a seminal 1978 article by Premack and Woodruff (1978), the ability to do so came to be called “Theory of Mind.” 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 inconsistently used between studies, 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. Here, adapted from our review in *Trends in Cognitive Science* (Schaafsma et al 2015), we consider use of the term in neuroscience, some of the brain mechanisms thought to support this broad social behavioral function, as well as a reformulation of the term ToM.

The Term “Theory of Mind” Has Several Interpretations

Since 1978, an ever-increasing number of studies have been published 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. A large number 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. Moreover, experimental approaches currently used often engage a large number of additional abilities, whose association with ToM is not always appropriate. Also, the term ToM is used interchangeably with mentalizing or mind reading, mind perception, and social intelligence (Baron-Cohen et al. 1999), to name only a few. This diversity of terms used is probably telling that different investigators have different concepts in mind. Confusion arises because many publications (i) implicitly treat ToM as a monolithic process, (ii) refer to a single brain network for ToM, and/or (iii) conflate varieties of ToM.

Humans have a competence to make sense of other people’s observed behavior, a competence shared with many other animals. How exactly we manage to do this is less clear and 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. The processes that enable us to think about other people's minds, in turn, are yet another matter. Distinctions among cognitive and emotional aspects of ToM are 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), with some schemes for relating them to one another. Humans likely use a mix of strategies that cuts across all these processes to figure out other people's minds.

Further, the different levels of description, together with the different terms used, make it difficult even for experts from different fields to understand exactly what is meant by ToM and how to study it using scientific methods. 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. Focusing just on the many papers that study ToM using neuroimaging yields no less heterogeneity.

Worst for the neuroscientist, the original definition of ToM does not permit easy downward translation to more basic processes such as those studied in behavioral neuroscience, leaving the interpretation of neuroimaging results opaque. We argue below 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 will be arrived at.

Thus, programmatic revision of ToM is the way forward. One might imagine going about this simply by constructing a kind of dictionary for the vocabulary of the scientific study of cognitive processes and attempting to relate these concepts to others that explain behavior at a lower level. Cognitive ontologies like this have seen some attention in recent years. For instance, there is the "Cognitive Atlas" project by Russ Poldrack (Poldrack et al. 2011), which aims to relate psychological concepts with one another and in particular aims to map concepts in terms of part-whole relationships. While, so far, no decomposition of ToM has resulted, the Cognitive Atlas (see www.cognitiveatlas.org) would seem an ideal platform in which to inform the project we sketch below, which proceeds in two main steps.

First, we propose that one needs to break ToM and its associated concepts apart into ones that describe more basic processes that also permit better identification with neural mechanisms. Second, one needs to reassemble different aspects of ToM from these more basic building blocks. The general approach bears considerable similarity to what Thomas Insel and the National Institute of Mental Health have recently advocated for the scientific study of psychiatric disorders Insel et al. (2010) by means of implementing the Research Domain Criteria Project (RDoC; http://www.nimh.nih.gov/research-priorities/rdoc/index.shtml). As with this RDoC approach to psychiatry, the original categories need to be revised and based on smaller dimensional constituents.

Theory of Mind Has More than One Form

A review of the neuroimaging data already suggests that there are likely to be a number of different varieties of ToM. A prominent distinction has been between a rapid, automatic form of ToM that may not require verbal competence, on the one hand, and a slower, deliberative form of ToM that is featured when we effortfully think about ToM, usually in language, and possibly culturally inherited, on the other hand (i.e., implicit versus explicit forms of ToM). Other distinctions, to which we have alluded already, are cognitive versus affective, a distinction closely related to one in work on empathy (Bernhardt and Singer 2012) and representation of one's own mental states (beliefs, intents, desires, etc.) *compared with* those of other people. All of these dual-process schemes have been arrived at more recently than the original one, which revolved around theorizing versus simulation.

Despite all these different flavors of ToM, one general observation is critical to note: these psychologically based ways of dichotomizing ToM are not generally intended to begin to disassemble ToM. The schemes offer psychological theories about ToM, but they all leave the original construct of ToM untouched.

Theory of Mind Must Be "Deconstructed" into Component Processes

While precursors to adult-level ToM abilities have been detailed in both nonhuman animals and in human infants, no systematic decomposition of the processes responsible for the ability in adult humans has been undertaken. Once agreed-upon tasks have been chosen, different components of ToM would need to be identified and separated in behavioral studies. Examples of such elements would include (but not be limited to) perceptual discrimination and categorization of the socially relevant stimuli, as well as of interoceptive signals elicited by those stimuli, semantic or conceptual knowledge, executive processes, and motivational processes.

It is still unclear exactly how to choose the best criteria for generating our list of more basic processes. One criterion should probably be that the basic processes are reasonably well understood already, and one might envision a hierarchical scheme, whereby ToM is first related to intermediate-level constructs, which may themselves still be further decomposed into more and more elemental processing components. The intermediate levels will then constitute combinations of component processes at lower levels (an illustrative example of a possible deconstruction and reconstruction scheme is depicted in Fig. 1).

A second criterion, at least for the most basic processes, is that these should have generally agreed-upon mappings to specific test instruments, such that propositions regarding them can be experimentally evaluated using behavioral tasks.

Finally, a related criterion is that they should ideally have a relatively clear relationship to neural networks, which will help to quantify their relationships to one another and in particular to lower-level processes grounded in our understanding of neural circuit function. An essential ingredient in this decomposition is attention to the construction of an array of behavioral tasks, which, at the top level of the hierarchy, should be representative of ToM as it appears in the real world. In addition

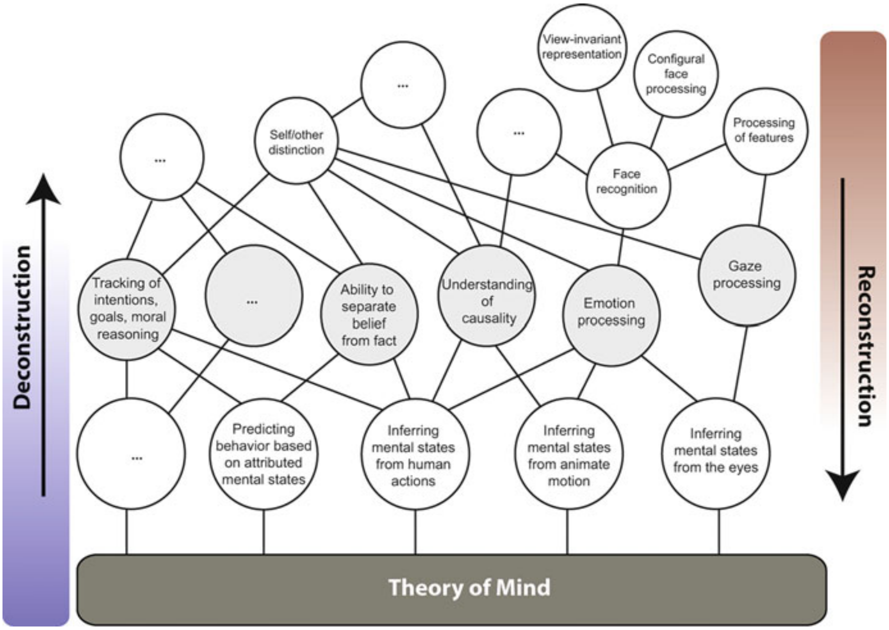
View-invariant representation
Configural face processing
Processing of features
...
Self/other distinction
...
...
Face recognition
Gaze processing
Tracking of intentions, goals, moral reasoning
...
Ability to separate belief from fact
Understanding of causality
Emotion processing
...
Predicting behavior based on attributed mental states
Inferring mental states from human actions
Inferring mental states from animate motion
Inferring mental states from the eyes
Theory of Mind
Deconstruction
Reconstruction

Fig. 1 An illustrative example of the reformulation of ToM by deconstruction into a comprehensive set of basic component processes on the one hand and a complementary reconstruction on the other hand, with the aim to construct a richer and scientifically tractable concept of ToM

to construct validity, such behavioral tasks must also offer convergent validity among their multiple measures, as well as discriminant validity, to distinguish ToM from processes not constitutive of ToM. Other psychometric features on our wish list for tasks are that they provide a range of performance, avoiding floor and ceiling effects and instead yielding a parametric measure that could reflect individual differences, that they show good test-retest reliability, and of course that they are practical to administer, ideally also within the environment of fMRI. Critically, we need to keep in mind that more basic tasks, taken in isolation, will be no better than higher-level tasks in selectively measuring ToM. Just as the higher-level constructs suffer from over-inclusion, so might the lower-level constructs suffer from both over- and under-inclusion. No basic task can capture all of ToM, and every basic task will involve cognitive processes in addition to those constitutive of ToM. ToM emerges from the basic tasks not because, at some point, we have captured a magic essence but because of the shared variance across all the basic tasks.

Following Deconstruction, Elementary Processes Can Be "Reconstructed" to Achieve Varieties of ToM

In the subsequent reconstruction stage, components of ToM would be identified by systematically recombining the most elementary, basic building blocks (Fig. 1). Mapping varieties of ToM within this space of more basic processes should allow us to relate the varieties to one another. Some should be more similar, in terms of their constituent basic processes, and some quite different. These similarity relations, in turn, would then need to be mapped onto the original concepts for varieties of ToM that we had to begin with. Once more complete, this exercise should yield two desirable outcomes. First, it will help us to separate valid instances of ToM from behaviors that should not qualify as ToM at all. Second – and this is the most subtle point – it should help us to revise our categories of ToM. Perhaps these revised categories will look similar to the ones we currently have or perhaps not – but in either case, they will be based on a more principled approach grounded in the similarity of relationships among simpler processes. Importantly, reconstruction will need to go hand in hand with conceptual refinement: neither neuroimaging results nor behavioral results will, in a single step, yield a new concept of ToM, but rather iterations among all these different levels of description will result in a more gradual revision.

While we stress the crucial role of well-designed behavioral tasks, an issue woefully ignored in much of the current literature on ToM, we also believe that the reconstruction of a new concept, or conceptual framework, for ToM can well be aided by the creative use of fMRI data. One such possibility is by using a tool such as Neurosynth, which can conduct an automated large-scale synthesis of the neuroimaging literature concerning ToM and produce core activation brain maps for ToM. By using simple ToM-related tasks that activate different parts of the neural network that fall within this Neurosynth-derived set of structures, it could be possible to help differentiate varieties of ToM.

For example, two of us recently compared two tasks that have been used in conjunction with fMRI to investigate the neural bases of two conceptually distinct uses of ToM (Spunt and Adolphs 2014). The first captures false-belief reasoning (the ability to attribute mental states to others and understand that those mental states may be different from one's own), while the second captures causal attributions about human behavior (Spunt and Lieberman 2012). Both tasks capture abilities that fall under the broad conceptual umbrella of ToM, and both evoke highly reliable activation in a circumscribed set of brain regions that have been labeled ToM regions in previous studies and meta-analyses. Yet, when directly compared in the same set of subjects, the two tasks evoke mostly nonoverlapping patterns of brain activation. Hence, although the cognitive abilities captured by these tasks may well have some shared components, their unshared components are more striking.

Neuroimaging results will of course also directly inform the decomposition of ToM and importantly will do so not merely through their similarity relationships with one another and to well-designed behavioral tasks but also because they will point to the computational processes. This, after all, is precisely the point of a tool such as Neurosynth: it is data driven with the aim to map activation patterns to processes. In very broad strokes, there are already plenty of examples from social neuroscience that suggest how this could work. For instance, medial prefrontal cortex, superior temporal sulcus, and temporal poles have been argued to implement, respectively, a decoupling between representations of the world and of other minds, processing of biological motion and agency, and semantic knowledge of social scripts. One can add additional components that could serve functions such as differentiating ToM about other individuals or about groups (a distinction found in multivoxel patterns of fMRI activation within shared regions) (Contreras et al. 2013) or that could add modulatory biases accounting for individual differences in dimensions such as egocentricity bias (Silani et al. 2013). It is clear that even a partially complete picture would look orders of magnitude more complex than the sparse sketch we show in Fig. 1, but it is also clear that eventually such a dense picture will be required to do justice to the complexity of the original construct of ToM.

Our suggested reconstruction could provide principled answers to a range of important questions. For instance, (i) which sets of basic processes are shared across the different varieties of ToM-related tasks currently studied? (ii) Are there collections of basic processes that can be seen as precursors to ToM abilities in infants and nonhuman animals? (iii) Are the constituent basic processes that come into play during ToM engaged in a particular temporal sequence? (McCleery et al. 2011) (iv) And which component processes (and the brain regions that implement them) are necessary for ToM, in the sense that ToM disabilities will arise if they are disrupted (either experimentally or through neurological or psychiatric illness)?

Outlook

One important question that arises is how do we know when a set of basic processes actually constitutes an instance of ToM (the concept cannot be synthesized simply from knowledge of the basic processes alone but requires some higher-level criteria to begin with)? Clearly, even in the face of massive revision, the project of reconstructing ToM as we have sketched it requires faith that there is indeed something distinctive about the core concept of ToM: our common way of understanding other people in terms of mental processes that cause their behavior (their desires, intentions, beliefs, and feelings).

Two candidates for criteria to provide this distinctiveness are (i) specific content and (ii) specific computational features. The *content* needs to be social and would be built into the tasks used for deconstructing ToM: they need to be about understanding desires, intentions, beliefs, and feelings. The *computational features* refer to the processes, whether inferred from careful assessment of behavior or from neuroimaging data. Some candidates for computational features, as used already in several recent neuroimaging studies (Dunne and Doherty 2013), include decoupling, recursion, and prediction, although all of these are, at present, too generic and descriptive to provide much mechanistic explanatory power. It may be that ToM recruits many processes (perception, attention, memory, motivation) through social-specific content but that its core processing (causal inference) makes particular computational, functional demands.

A recent neuroimaging example that, perhaps, comes closest to our idea has focused on the much-debated functions of the temporoparietal junction (TPJ). This region has been activated in studies engaging a large number of cognitive processes, although most of the focus has been on its role in signaling shifts in attention and in representing false beliefs. While there may be some anatomical segregation of these functions within the TPJ (Mar et al. 2012), another view has been to propose that its role in ToM emerges from the engagement of a number of other processes. That is, TPJ may serve as a sort of "nexus," extracting and synthesizing social context (from a large body of information) and guiding attention and decision-making (Carter et al. 2012).

Another issue will be how to characterize the most basic elements contributing to ToM in neuroanatomical terms. An important recent direction in all of cognitive and systems neuroscience has been to think of *networks* of brain regions, rather than individual regions. While such networks are often identified from fMRI resting-state data, they can be extracted from cognitive activation tasks as well. How they map onto basic behaviors is still very much a work in progress, although some initial schemes are starting to emerge. It seems clear that some such network-based inventory of neuroanatomical "basis functions" will need to replace the current region-based literature.

Some time ago, it might have been argued that the kind of decomposition and reconstruction we are envisioning might prove impossible, if it had not been done for any higher-level cognitive process. But it has been accomplished, at least in broad terms. The best example of this is for memory. The psychological concept of memory has been successfully fractionated into temporal stages (encoding, consolidation, retrieval), has been decomposed into types of memory (declarative, procedural, etc.), and has been identified with specific neural structures and systems (the hippocampus for declarative memory, the amygdala for Pavlovian fear conditioning, etc.) as well as cellular processes (long-term potentiation, spike-timing-dependent plasticity). Of course, our understanding of memory is by no means complete, and the above examples are much more complicated than their brief sketch would indicate. But, at least in broad strokes, we know a lot about the components of memory and how they generate a psychological instance of memory performance on a task. We anticipate doing something similar for ToM.

It may well be that decomposition of ToM should be expected to be more like a decomposition of fluid intelligence than a decomposition of memory: specific, basic neural mechanisms (like spike-timing-dependent plasticity) may not emerge, and a very distributed set of neural regions may be involved. Indeed, we may need to take into account factors outside the brain. In this effort, a possibly helpful tool emerges from the conceivable parallels between reading minds and reading print. That is, authors have argued that explicit ToM is a culturally inherited skill, analogous to reading print: there is no brain system "for" such a skill, but rather the skill emerges in a cultural context through recruitment of many available processes. It may prove fruitful to borrow theories and methods from research on other complex cognitive processes and behaviors, such as acquiring the ability to read print, and implement them in our efforts to decompose and subsequently reconstruct the mechanisms underlying ToM.

In summary, the project of mapping behavior to psychology to neurobiology in the case of Theory of Mind requires revising our concepts at multiple levels. All levels are valuable, since each captures regularities that are less economically described at other levels, and so none can be eliminated. Our core argument has been that ToM has problems with how it has been constrained. Specifically, it has been constrained too little in the diverse usages across our field, and it has been constrained too much by anchoring to a single concept (representing other minds) without easy translation downwards. Deconstructing ToM to a fully fleshed-out list of building blocks, and then reconstructing it, is, of course, a huge undertaking that will require a concerted effort across the scientific community. Our aim here has been to sketch what we hope could be a common vision to achieve that goal.

References

Baron-Cohen S et al (1999) Social intelligence in the normal and autistic brain: an fMRI study. Eur J Neurosci 11:1891–1898

Bernhardt BC, Singer T (2012) The neural basis of empathy. Annu Rev Neurosci 35:1–23

Carter RM et al (2012) A distinct role of the temporal-parietal junction in predicting socially guided decisions. Science 337:109–111

Contreras JM, Schirmer J, Banaji MR, Mitchell JP (2013) Common brain regions with distinct patterns of neural response during mentalizing about groups and individuals. J Cogn Neurosci 25:1406–1417

Dunne S, O’Doherty JP (2013) Insights from the application of computational neuroimaging to social neuroscience. Curr Opin Neurobiol 23:387–392

Insel T et al (2010) Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am J Psychiatry 167:748–751

Mars RB, Sallet J, Schüffelgen U, Jbabdi S, Toni I, Rushworth MFS (2012) Connectivity-based subdivisions of the human right “temporoparietal junction area”: evidence for different areas participating in different cortical networks. Cereb Cortex 22:1894–1903

McCleery JP et al (2011) The neural and cognitive time course of theory of mind. J Neurosci 31:12849–12854

Poldrack RA et al (2011) The cognitive atlas: towards a knowledge foundation for cognitive neuroscience. Front Neuroinform. https://doi.org/10.3389/fninf.2011.00017x

Premack D, Woodruff G (1978) Does the chimpanzee have a theory of mind? Behav Brain Sci 1:515–526

Schaafsma SM, Pfaff DW, Spunt R, Adolphs R (2015) Deconstructing and reconstructing theory of mind. Trends Cogn Neurosci 19:65–79

Silani G, Lamm C, Ruff CC, Singer T (2013) Right supramarginal gyrus is crucial to overcome emotional egocentricity bias in social judgments. J Neurosci 33:15466–15476

Spunt RP, Adolphs R (2014) Validating the why/how contrast for functional MRI studies of theory of mind. Neuroimage 99:301–311

Spunt R, Lieberman M (2012) Dissociating modality-specific and supramodal neural systems for action understanding. J Neurosci 32:3575–3583

Further Reading

Carrington S, Bailey A (2009) Are there theory of mind regions in the brain? A review of the neuroimaging literature. Hum Brain Mapp 30:2313–2335

Waytz A, Mitchell JP (2011) Two mechanisms for simulating other minds: dissociations between mirroring and self-projection. Curr Dir Psychol Sci 20:197–200