Published in final edited form as:
Trends Cogn Sci. 2015 February ; 19(2): 65–72. doi:10.1016/j.tics.2014.11.007.

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

[2]California Institute Of Technology, 1200 East California Boulevard, Pasadena, CA 91125

Abstract

Usage of the term Theory of Mind (ToM) has exploded across fields ranging from developmental psychology to social neuroscience and psychiatry research. Yet 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 critically, 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 could be recovered.

What is Theory of Mind?

The term theory of mind (ToM), 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. 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, 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 phrase ToM here, it should be noted that this is merely for convenience in exposition, not an endorsement of current usage. Our aim is more of a general call to action than a specific

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

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

prescription, however; consequently we sketch a broad research programme rather than tackle its implementation.

BOX 1

Tasks typical for studying ToM in fMRI studies

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 > false photograph
1 study	false belief and subjective preference
7 studies	false belief > true belief
3 studies	false belief > 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 vs. self-judgments
3 studies	other judgments vs. 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 >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 > 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 > 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>how)
3 studies	only passively watch actions

Footnote: Table adapted from [29]

The Problem

Humans all 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 (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. Here, 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 going on 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 cuts 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] (see 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 just on the many papers that study ToM using neuroimaging yields no less heterogeneity (see Box 1 and 2, Figure 2). The problem is that these differences generally go unarticulated, and their basis is often not well grounded.

BOX 2

Is there a ToM network in the human brain?

Pioneering studies [67–69] suggested the feasibility of localizing ToM to particular brain regions, forming a reliably activated functional network in the human brain that is so specifically associated with the use of ToM that it is typically reified with phrases like “ToM Network” or “Mentalizing System”.

Is this reification warranted? There are two pieces of evidence that might suggest so. The first emerges from meta-analyses based on the spatial coordinates reported in a large number of studies, which find that there are some spatial foci that are consistently activated when people perform a ToM-related task [7, 70–73]. The second emerges from research on the neural bases of false-belief reasoning [19, 23, 74].

However, this case breaks down under closer scrutiny, and it does so for reasons that are evident in the pioneering neuroimaging studies of ToM referenced above. Specifically, there is massive heterogeneity in the neuroimaging methods used to investigate ToM, both in terms of the behavioral manipulations used to elicit it, and in terms of the imaging methods used to measure and analyze its neural correlates. This heterogeneity is so great that, until recently, meta-analyses have been underpowered to examine task-specific activations and have instead pooled data from all tasks.

Box 1 provides some examples from the inventory of different tasks and stimuli that have been used to study ToM. A recent meta-analysis [29] illustrates why this heterogeneity is highly problematic (Figure 2). The authors examined activation foci from 73 neuroimaging studies that spanned 6 distinct task groups that fall under the umbrella of ToM. When performing a meta-analysis on the pooled activation data, they observed an activation map that largely reproduced published meta-analyses. However, when performing the meta-analysis on the task groups separately, this map breaks down, revealing clearly distinct activation profiles (see also [7]). Lumping tasks together blurs out potentially meaningful distinctions at both the neural and cognitive levels of analysis and undermines the possibility that different brain networks subserve distinct aspects of ToM.

Although various counterarguments could be made, it is safe to say that the evidence for a ToM network is limited and contentious. Moving forward, our suggestion is not to abandon the effort to ground ToM in brain function. Rather, our suggestion is to abandon the notion that ToM is a single cognitive ability grounded in a single set of brain regions.

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 look like 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.

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].

On the other hand, 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 permits the reconstruction of a concept of ToM — albeit in a revised fashion, and now with the necessary links to lower levels of explanation in place. But it is important to keep in mind the distinction between our concepts associated with ToM, on the one hand, and the psychological processes constituting ToM, on the other hand. Indeed, it has been suggested that we should approach ToM as more of a conceptual framework, considering psychological processes as a separate issue [24–25]. This distinction could also permit domain-specific (i.e., specifically social) concepts for ToM, even though none of the constituent processes may be domain-specific (but perhaps rather generic processes simply operating on specific kinds of content).

In short, we believe that a 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 [26], 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, 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 Tom Insel and the National Institute of Mental Health have recently advocated for the scientific study of psychiatric disorders [27,28] 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.

A way forward

A review of the neuroimaging data (Box 2) already suggests that there are likely to be a number of different varieties of ToM [29]. 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 (implicit vs. explicit; [30–32]). Other distinctions, to which we have alluded already, are cognitive versus affective [33–34], a distinction closely related also to one in work on empathy [35–36], and representation of one's own mental states compared with those of other people [37]. All of these dual-process schemes are more recent than the original one, which revolved around theorizing versus simulation [3–6]. 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. An alternative view is analogous to what has been argued for social cognition more generally:

> "….a view of social cognition as a collection of mental processes, each specialized for making sense of others under specific circumstances. Just as humans make use of several different perceptual senses (sight, smell, taste, touch, hearing) to represent the physical world around them, humans use several different social-cognitive processes to construct a useful representation of the social world around them." [38]

Deconstruction

While precursors to adult-level ToM abilities have been detailed in both nonhuman animals and in human infants (e.g. [16,39]), 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 unclear at this stage 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 Figure 3).

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. Finally, a related criterion is that they should ideally have a relatively clear relationship to neural networks [40], 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 to construct validity, such behavioral tasks must also offer convergent validity amongst 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 do 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.

Reconstruction

In the subsequent reconstruction stage, components of ToM would be identified by systematically recombining the most elementary, basic building blocks (Figure 3). 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 (Table 2). 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, it should help us to revise our categories of ToM; perhaps these 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 relationships amongst 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 amongst 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 [41], which can conduct an automated large-scale synthesis of the neuroimaging literature concerning ToM and produce core activation brain maps (forward inference) for ToM. By using simple ToM-related tasks that activate different parts of the neural network that falls within this Neurosynth–derived set of structures (e.g., [42,43]), it could be possible to help differentiate varieties of ToM.

For example, we recently compared two tasks that have been used in conjunction with fMRI to investigate the neural bases of two conceptually distinct uses of ToM [44]. The first, already discussed above, captures false-belief reasoning about characters in short stories [23], while the second captures causal attributions about human behavior [45]. Both tasks capture abilities that fall under the umbrella of ToM, and both evoke highly reliable activation in a circumscribed set of brain regions that have been labeled ToM or mentalizing regions. Yet, when directly compared in the same set of subjects, the two tasks evoke mostly non-overlapping patterns of brain activation [44]. 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 [41]: 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 [46]. 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 [47]), or that could add modulatory biases accounting for individual differences in dimensions such as egocentricity bias [48]. It is clear that even a partially complete picture would look orders of magnitude more complex than the sparse sketch we show in Figure 3; 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? [49] (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 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 (as opposed to more eliminativist views) 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 andidates for criteria to provide this distinctiveness are specific content and 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 (e.g., [50–53]) include decoupling [54], recursion [55,56], and prediction [57–59], 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 temporal-parietal 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 [60, 61], 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 [62].

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 [63]. How they map onto basic behaviors is still very much a work in progress, although some initial schemes are starting to emerge (e.g., [64,65]). 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 (hippocampus for declarative memory, 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 solved, 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. (For more details, see the entry for memory in the Cognitive Atlas and click on some of the types of memory.) Why can't we do something similar for ToM?

It may well be that our decomposition of ToM is 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 [66]. 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 [32]. In that paper, the authors argue 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.

Acknowledgments

We would like to thank for helpful discussions and comments Jed Elison, Evelina Fedorenko, Ian Apperly, Nicholas Epley, Russell Poldrack, Bertram Malle, Frank van Overwalle, as well as Tania Singer and five of her colleagues. S.M.S. was supported by a postdoctoral fellowship from the Autism Science Foundation (13-1002). R.A. and B.S. were supported by a Conte Center grant from NIMH (P50MH094258). An initial discussion of some of the themes of this paper took place in a meeting organized by Don Pfaff and Terry Sejnowski at the Center for Research and Teaching in Anthropogeny (CARTA), funded by the Mathers Foundation.

References

1. Premack D, Woodruff G. Does the chimpanzee have a theory of mind? Behav Brain Sci. 1978; 1:515–526.
2. Bretherton I, Beeghly M. Talking about internal states: the acquisition of an explicit theory of mind. Developmental Psychology. 1982; 18:906–921.
3. Gopnik, A.; Wellman, HM. The theory theory. In: Hirschfeld, L.; Gelman, S., editors. Mapping the Mind. Cambridge University Press; 1994. p. 257-293.
4. Leslie AM, et al. Core mechanisms in ‘theory of mind’. Trends Cogn Sci. 2004; 8:528–533. [PubMed: 15556021]
5. Gallese V, Goldman A. Mirror neurons and the simulation theory of mind-reading. Trends Cogn Sci. 1998; 2:493–501. [PubMed: 21227300]
6. Gordon R. Folk Psychology as Simulation. Mind Lang. 1986; 1:158–171. reprinted: Davies M. and Stone, T. (1995) *In Folk Psychology: The Theory of Mind Debate*, pp 60–73, Blackwell Publishers.

7. Van Overwalle F, Baetens K. Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. Neuroimage. 2009; 48:564–584. [PubMed: 19524046]
8. Keysers C, Gazzola V. Integrating simulation and theory of mind: from self to social cognition. Trends Cogn Sci. 2007; 11:194–196. [PubMed: 17344090]
9. Mitchell JP. The false dichotomy between simulation and theory-theory: the argument's error. Trends Cogn Sci. 2005; 9:363–364. [PubMed: 16006173]
10. Apperly I. What is "theory of mind"? Concepts, cognitive processes and individual differences. Q J Exp Psychol A. 2012; 65:825–839.
11. Mahy CEV, et al. How and where: Theory-of-mind in the brain. Dev Cogn Neurosci. 2014; 9:68–81. [PubMed: 24552989]
12. Vogeley K, et al. Mind reading: Neural mechanisms of theory of mind and self-perspective. Neuroimage. 2001; 14:170–181. [PubMed: 11525326]
13. Gray K, et al. Distortions of mind perception in psychopathology. P Natl Acad Sci USA. 2011; 108:477–479.
14. Baron-Cohen S, et al. Social intelligence in the normal and autistic brain: An fMRI study. Eur J Neurosci. 1999; 11:1891–1898. [PubMed: 10336657]
15. Kaminski J, et al. Dogs steal in the dark. Anim Cogn. 2013; 16:385–394. [PubMed: 23179109]
16. Emery NJ, Clayton NS. Comparative social cognition. Ann Rev Psychol. 2008; 60:87–113. [PubMed: 18831684]
17. Call J, Tomasello M. Does the chimpanzee have a theory of mind? 30 years later. Trends Cogn Sci. 2008; 12:187–192. [PubMed: 18424224]
18. Penn DC, Povinelli DJ. On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. Phil Trans R Soc B. 2007; 362:731–744. [PubMed: 17264056]
19. Koster-Hale, J.; Saxe, R. Functional neuroimaging of theory of mind. In: Baron-Cohen, S.; Lombardo, M.; Tager-Flusberg, H., editors. Understanding Other Minds. 3. Oxford University Press; 2014. p. 132-163.
20. Heyes C. Submentalizing: I am not really reading your mind. Perspect Psychol Sci. 2014; 9:131–143.
21. Frith U, Happe F. Autism: Beyond "theory of mind". Cognition. 1994; 50:115–132. [PubMed: 8039356]
22. Bloom P, German TP. Two reasons to abandon the false belief task as a test of theory of mind. Cognition. 2000; 77:B25–B31. [PubMed: 10980256]
23. Saxe R, Powell LJ. It's the thought that counts: specific brain regions for one component of theory of mind. Psychol Sci. 2006; 17:692–699. [PubMed: 16913952]
24. Malle, BF. The fundamental tools, and possibly universals, of human social cognition. In: Sorrentino, RM.; Yamaguchi, S., editors. Handbook of Motivation and Cognition across Cultures. San Diego: Academic Press; 2008. p. 267-296.
25. Malle, BF. Folk theory of mind: Conceptual foundations of human social cognition. In: Hassin, R.; Uleman, JS.; Bargh, JA., editors. The new unconscious. New York: Oxford University Press; 2005. p. 225-255.
26. Poldrack RA, et al. The Cognitive Atlas: Towards a knowledge foundation for cognitive neuroscience. Front Neuroinform. 201110.3389/fninf.2011.00017
27. Kapur S, et al. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? Mol Psychiatr. 2012; 17:1174–1179.
28. Insel T, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am J Psychiat. 2010; 167:748–751. [PubMed: 20595427]
29. Schurz M, et al. Fractionating theory of mind: A meta-analysis of functional brain imaging studies. Neurosci Biobehav Rev. 2014; 42:9–34. [PubMed: 24486722]
30. Apperly IA, Butterfill S. Do humans have two systems to track beliefs and belief-like states? Psychol Rev. 2009; 116:953–970. [PubMed: 19839692]
31. van Overwalle F, Vandekerckhove M. Implicit and explicit social mentalizing: dual processes driven by a shared neural network. Front Hum Neurosci. 201310.3389/fnhum.2013.00560

32. Heyes C, Frith S. The cultural evolution of mind reading. Science. 2014; 344:1243091. [PubMed: 24948740]

33. Abu-Akel A, Shamay-Tsoory S. Neuroanatomical and neurochemical bases of theory of mind. Neuropsychologia. 2011; 49:2971–2984. [PubMed: 21803062]

34. Poletti M, et al. Cognitive and affective Theory of Mind in neurodegenerative diseases: neuropsychological, neuroanatomical and neurochemical levels. Neurosci Biobehav Rev. 2012; 36:2147–2164. [PubMed: 22819986]

35. Preston SD, de Waal FBM. Empathy: Its ultimate and proximate bases. Behav Brain Sci. 2002; 25:1–20. [PubMed: 12625087]

36. Bernhardt BC, Singer T. The neural basis of empathy. Annu Rev Neurosci. 2012; 35:1–23. [PubMed: 22715878]

37. van Veluw SJ, Chance SA. Differentiating between self and others: an ALE meta-analysis of fMRI studies of self-recognition and theory of mind. Brain Imaging Behav. 2014; 8:24–38. [PubMed: 24535033]

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

39. Baron-Cohen, S., editor. Mindblindness: an essay on autism and theory of mind. MIT Press; 1995.

40. Price CJ, Friston KJ. Functional ontologies for cognition: the systematic definition of structure and function. Cogn Neuropsychol. 2005; 22:262–275. [PubMed: 21038249]

41. Yarkoni T, et al. Large-scale automated synthesis of human functional neuroimaging data. Nat Methods. 2011; 8:665–670. [PubMed: 21706013]

42. Hartwright CE, Apperly IA, Hansen PC. Multiple roles for executive control in belief-desire reasoning: Distinct neural networks are recruited for self perspective inhibition and complexity of reasoning. Neuroimage. 2012; 61:921–930. [PubMed: 22440654]

43. Hartwright CE, Apperly IA, Hansen PC. Representation, control, or reasoning? Distinction functions for theory of mind within the medial prefrontal cortex. J Cogn Neurosci. 2014; 26:683–698. [PubMed: 24236763]

44. Spunt RP, Adolphs R. Validating the why/how contrast for functional MRI studies of theory of mind. Neuroimage. In Press.

45. Spunt R, Lieberman M. Dissociating modality-specific and supramodal neural systems for action understanding. J Neurosci. 2012; 32:3575–3583. [PubMed: 22399779]

46. Frith U, Frith CD. Development and neurophysiology of mentalizing. Phil Trans R Soc Lond B. 2003; 358:459–473. [PubMed: 12689373]

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

48. Silani G, Lamm C, Ruff CC, Singer T. Right supramarginal gyrus is crucial to overcome emotional egocentricity bias in social judgments. J Neurosci. 2013; 33:15466–15476. [PubMed: 24068815]

49. McCleery JP, et al. The neural and cognitive time course of theory of mind. J Neurosci. 2011; 31:12849–12854. [PubMed: 21900563]

50. Hampton AN, et al. Neural correlates of mentalizing-related computations during strategic interactions in humans. P Natl Aca Sci USA. 2008; 105:6741–6.

51. Behrens TEJ, et al. The Computation of Social Behavior. Science. 2009; 324:1160–1164. [PubMed: 19478175]

52. Boorman ED, et al. The behavioral and neural mechanisms underlying the tracking of expertise. Neuron. 2013 in press.

53. Dunne S, O'Doherty JP. Insights from the application of computational neuroimaging to social neuroscience. Curr Opin Neurobiol. 2013; 23:387–392. [PubMed: 23518140]

54. Gallagher HL, Frith CD. Functional imaging of 'theory of mind'. Trends Cogn Sci. 2003; 7:77–83. [PubMed: 12584026]

55. Yoshida W, et al. Game theory of mind. PLOS Comput Biol. 2008:4.10.1371/journal.pcbi.1000254

56. Frith CD. The role of metacognition in human social interactions. Phil Trans R Soc B. 2012; 367:2213–2223. [PubMed: 22734064]

57. Baker, CL. Doctoral Thesis. MIT Press; 2012. Bayesian Theory of Mind: modeling human reasoning about beliefs, desires, goals, and social relations.

58. Koster-Hale J, Saxe R. Theory of Mind: A neural prediction problem. Neuron. 2013; 79:836–848. [PubMed: 24012000]

59. Kilner JM, Friston KJ, Frith CD. The mirror system: a Bayesian perspective. Neuroreport. 2007; 18:619–623. [PubMed: 17413668]

60. Scholz J, et al. Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. PLOS ONE. 2009; 410.1371/journal.pone.0004869

61. Mars RB, Sallet J, Schüffelgen U, Jbabdi S, Toni I, Rushworth MFS. Connectivity-based subdivisions of the human right "temporoparietal junction area": evidence for different areas participating in different cortical networks. Cereb cortex. 2012; 22:1894–1903. [PubMed: 21955921]

62. Carter RM, et al. A distinct role of the temporal-parietal junction in predicting socially guided decisions. Science. 2012; 337:109–11. [PubMed: 22767930]

63. Smith SM, et al. Correspondence of the brain's functional architecture during activation and rest. P Natl Aca Sci USA. 2009; 106:13040–13045.

64. Seeley WW, et al. Dissociable intrinsic connectivity networks for salience processing and executive control. J Neurosci. 2007; 27:2349–2356. [PubMed: 17329432]

65. Laird AR, et al. Behavioral interpretations of intrinsic connectivity networks. J Cogn Neurosci. 2011; 23:4022–4037. [PubMed: 21671731]

66. Glaescher J, et al. The distributed neural system for general intelligence revealed by lesion mapping. PNAS. 2010; 107:4705–4709. [PubMed: 20176936]

67. Baron-Cohen S, et al. Recognition of mental state terms: a clinical study of autism, and a functional neuroimaging study of normal adults. Brit J Psychiat. 1994; 165:640–649.

68. Fletcher PC, et al. Other minds in the brain: a functional imaging study of 'theory of mind' in story comprehension. Cognition. 1995; 57 (2):109–128. [PubMed: 8556839]

69. Goel V, et al. Modeling other minds. NeuroReport. 1995; 6:1741–1746. [PubMed: 8541472]

70. Denny B, et al. A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. J Cogn Neurosci. 2012; 24:1742–1752. [PubMed: 22452556]

71. Mar RA. The Neural Bases of Social Cognition and Story Comprehension. Annu Rev Psychol. 2011; 62:103–134. [PubMed: 21126178]

72. Carrington S, Bailey A. Are there theory of mind regions in the brain? A review of the neuroimaging literature. Hum Brain Mapp. 2009; 30:2313–2335. [PubMed: 19034900]

73. Lieberman, MD. Social cognitive neuroscience. In: Fiske, ST.; Gilbert, DT.; Lindzey, G., editors. Handbook of Social Psychology. 5. McGraw-Hill; 2010. p. 143-193.

74. Saxe R, Kanwisher N. People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". Neuroimage. 2003; 19:1835–1842. [PubMed: 12948738]

Highlights

1. “ToM” has been intended to represent a crucial component of human cognitive capacity
2. ToM is not constrained adequately in its diverse usages and too much by anchoring to a single concept
3. We propose a two stage approach to refine “Theory of Mind”
4. Deconstructing ToM will disassemble its essences into simpler processes
5. By systematically and heuristically recombining basic building blocks, ToM can be reconstructed

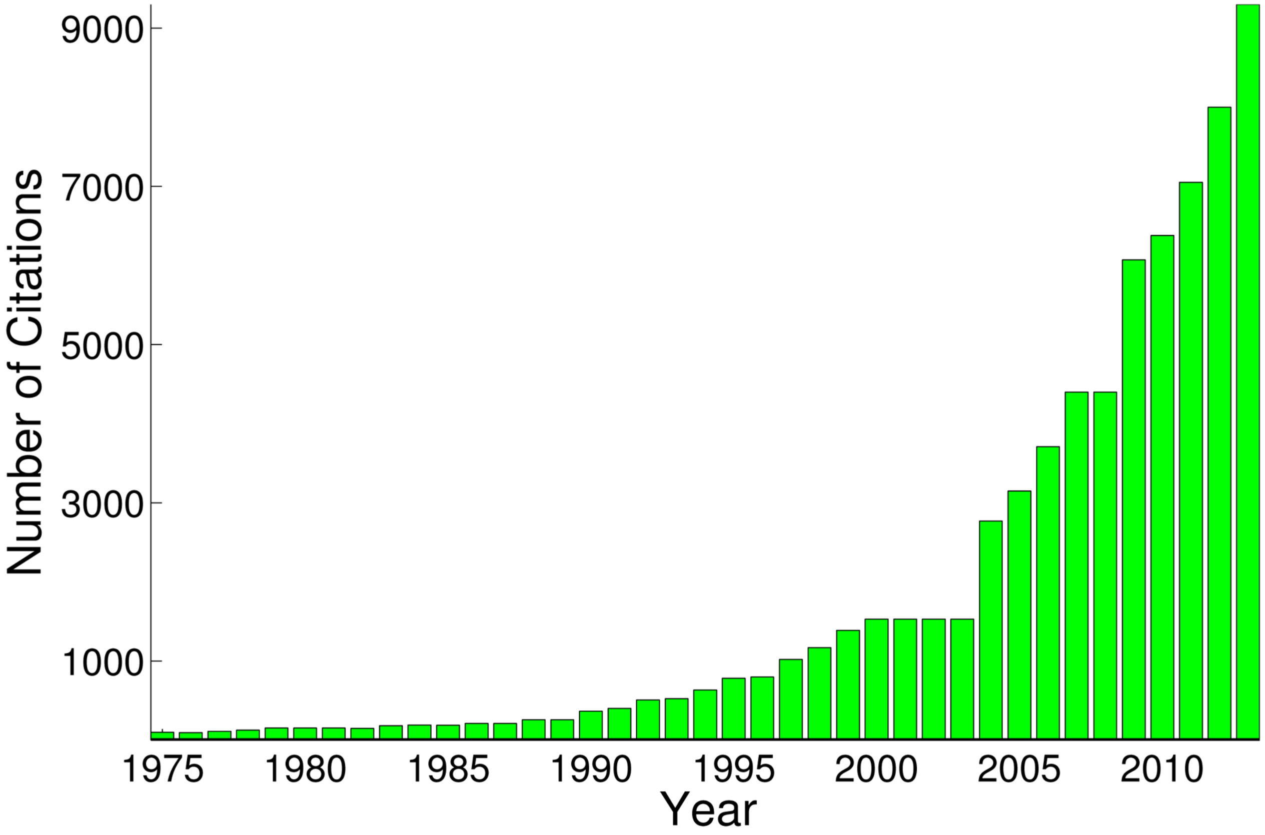

Figure 1.

Articles referencing Theory-of-Mind have increased markedly in recent years. Estimates are based on a per annum Google Scholar search (scholar.google.com) for articles that use the exact phrase “Theory of Mind”.

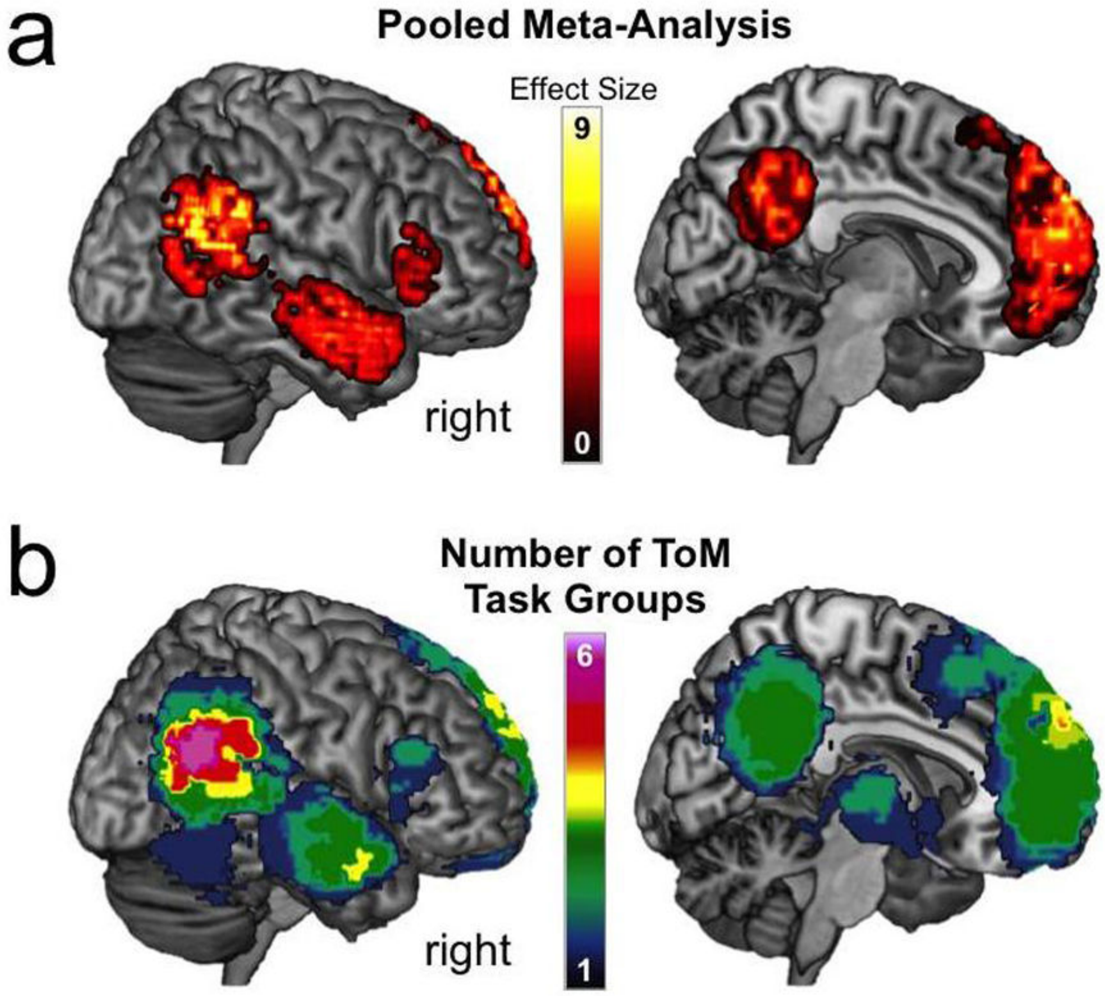

Figure 2.

Deconstructing the ToM Network. Data is adapted from meta-analytic results reported in [29], which included 73 neuroimaging studies that used one of the six distinct task groups listed in Box 1 and are rendered on a canonical brain in Montreal Neurological Institute (MNI) space. (a) Results of the meta-analysis when pooling data from all six distinct task groups. (b) Sum of results from six independent meta-analyses conducted for each task group. The color map indexes the number of the tasks that reliably produce activation at a given voxel.

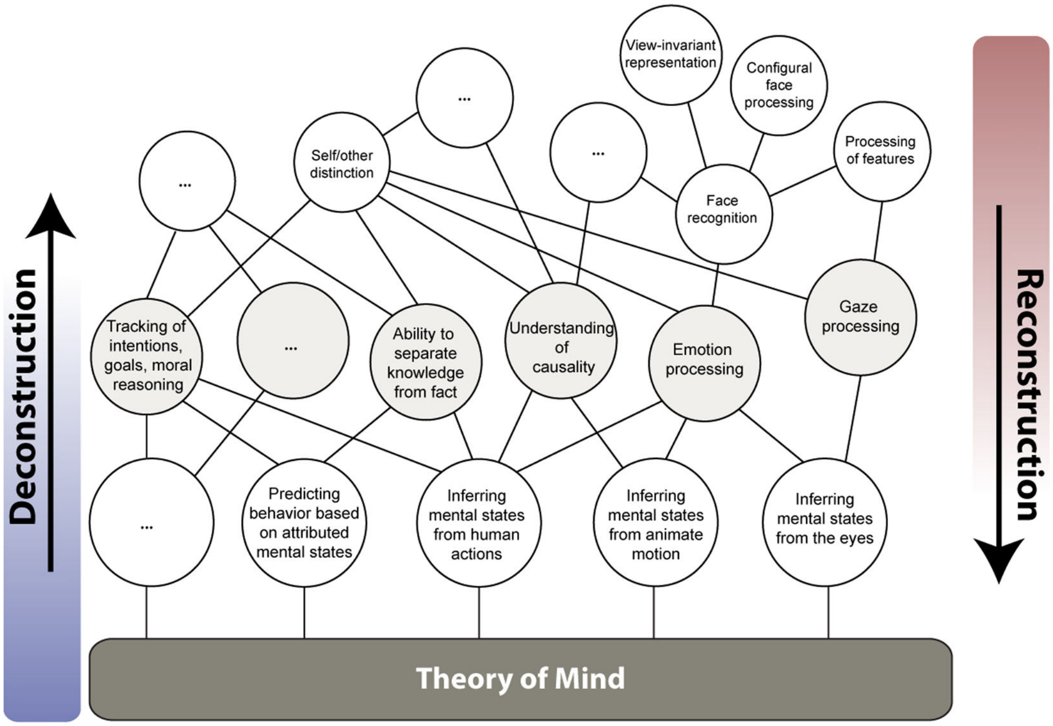
View-invariant representation
Configural face processing
...
Processing of features
Self/other distinction
...
...
Face recognition
Tracking of intentions, goals, moral reasoning
...
Ability to separate knowledge from fact
Understanding of causality
Emotion processing
Gaze processing
...
Predicting behavior 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

Figure 3.

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.

Table 1

Table 1 (within Box 1): Examples of different tasks and constructs for ToM

Test	**Reading the Mind in the Eyes Task**	**Sally and Anne Task**	**Heider & Simmel animation**
population	clinical populations; children; adults	children only; autism spectrum disorder	mostly adults
construct	emotion recognition	false-belief understanding	anthropomorphization
stimulus	photographs of eyes	vignettes/cartoons	videos of moving shapes
response	emotion identification	behavior prediction	subjective description