Developing a Theory of Mind: Insights from FMRI Studies of Children

by

Hilary L. Richardson

B.S. Brain, Behavior, and Cognitive Science
University of Michigan, 2010

SUBMITTED TO THE DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN NEUROSCIENCE
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2018

Signature of Author: Signature redacted
Department of Brain and Cognitive Sciences
April 25, 2018

Certified by: Signature redacted
Rebecca R. Saxe
Professor of Cognitive Neuroscience
Thesis Supervisor

Accepted by: Signature redacted
Matthew A. Wilson
Sherman Fairchild Professor of Neuroscience and Picower Scholar
Director of Graduate Education for Brain and Cognitive Sciences

Developing a Theory of Mind: Insights from FMRI Studies of Children

by

Hilary L. Richardson

Submitted to the Department of Brain and Cognitive Sciences
On April 12, 2018 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Neuroscience

Abstract
Social cognitive abilities undergo drastic changes throughout childhood. Theory of mind (ToM), the ability to reason about the mental states of others, is a core social cognitive ability that is crucial for navigating the social world. A majority of prior fMRI research on ToM has characterized the functional response in brain regions that are preferentially recruited to reason about the minds of others in adults. By contrast, a majority of prior developmental research on ToM has used behavioral methods to describe milestones in theory of mind acquisition in early childhood. The experiments described in this thesis draw heavily from these two approaches, in order to link them: what is the relationship between the development of functionally selective responses in ToM brain regions, and developmental changes in ToM reasoning in childhood? Chapter 1 describes two longitudinal fMRI experiments that test for developmental change and stable individual differences in neural and behavioral measures of ToM, and for predictive relationships between the two measures. Chapter 2 describes a large, cross-sectional study that measures the development of the cortical dissociation between brain regions that process minds (the ToM network) and those that process bodies (the Pain Matrix). Chapter 2 additionally provides insight into the neural correlates of passing the false-belief task – the best known developmental milestone in ToM reasoning. Chapter 3 uses a publicly available dataset in order to provide confirmatory evidence for the results described in Chapter 2, and clarifies the relationship between stimulus-driven functional responses, and inter-region correlations within and between ToM and pain brain regions. Chapter 4 characterizes ToM development, neurally and behaviorally, in children who have experienced delayed access to sign language. Finally, Chapter 5 provides a discussion of challenges and strategies in developmental cognitive neuroscience research. This interdisciplinary thesis has three broad goals: 1) to characterize kinds of neural change that support and/or predict behavioral improvements in theory of mind, 2) to gain novel insight into the nature of specific behavioral milestones in social reasoning, and 3) to better understand the impact of experience (e.g., linguistic input) on ToM development, behaviorally and neurally.

Thesis Supervisor: Rebecca Saxe
Title: Professor of Cognitive Neuroscience

Acknowledgements
It feels impossible to express my gratitude to my advisor, Rebecca Saxe. This thesis would not have been possible without Rebecca's encouragement, guidance, generosity, and support. I am so thankful to Rebecca for being there through it all, and for continuing to inspire me.

I would also like to thank my amazing committee members: Nancy Kanwisher, Laura Schulz, and Jason Yeatman, for being so incredibly encouraging, and for providing key insights and new perspectives along the way. The encouragement and insights from Rebecca and my committee members kept me excited, and substantially improved and shaped this thesis and my research program.

I am thankful to have had the support of the members of The Saxe Lab, who have become a second family to me, and have provided guidance, advice, technical support, and friendship. Thanks in particular to my collaborators on the projects included in this thesis: Hyowon Gweon & Lyneé Alves (Chapter 1), Grace Lisandrelli & Alexa Riobueno-Naylor (Chapter 2), and Jorie Koster-Hale, Naomi Caselli, Rachel Magid, Rachel Benedict, Halie Olson, & Jennie Pyers (Chapter 4). Thanks to lab managers past and present (with extra thanks to Nick Dufour and Todd Thompson), and Mika Asaba and Natalia Velez who were amazing undergraduate researchers and contributed to projects not included in this thesis. Many thanks to Marina Bedny for her mentorship; working with Marina shaped my research interests, and made me better able to articulate them. I would also like to thank Steven Shannon, Atsushi Takahashi, and Sheeba Anteraper for imaging support and infinite patience.

I would not have found my way to the Saxe lab without the guidance and support from Lindsay Bowman and Henry Wellman. Lindsay and Henry introduced me to the phrase "theory of mind," and to questions about theory of mind development that continue to fascinate me. I am so grateful for your continued support.

Many many thanks to my fellow graduate students and friends who have made this journey more fun. In particular, thanks to Dorit Kliemann for being an incredibly supportive office mate and friend, and to Julia Leonard for our many conversations, zumba and pottery breaks, and hugs.

Finally, I am so grateful for the support of my family and Neal. Thank you for continually cheering me on.

Table of Contents

Notes:

1. Content in the Introduction, Chapter 5, and General Discussion and Conclusion appeared as:
   Richardson, H. & Saxe, R. (2016). Using MRI to Study Developmental Change in Theory of Mind. In Social Cognition: Development Across the Lifespan.

2. A version of Chapter 2 appeared as:
   Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. Nature Communications, 9(1), 1027.

Introduction

By the time children enter first grade, they have learned a lot about human minds – their own, and other people's. A five-year-old can tell whether someone is happy or sad, wanted raisins or carrots, and knows or doesn't know where the missing cookie is. Of course, children and even adults still have a lot to learn about people's thoughts and emotions; for example, the difference between someone taking a cookie intentionally, accidentally, or negligently, or the difference between feeling happy versus acting happy.

This long and impressive developmental progression is sometimes called "acquiring a Theory of Mind." A full account of Theory of Mind (ToM) development would require us to describe (1) the mature structure of the theory, (2) infants' initial conceptual repertoire, and (3) how maturation and different learning mechanisms capitalize on children's experience to bridge this gap[1]. To build such an account, psychologists must use many different empirical approaches; while initial studies of ToM typically asked three to six year old children to explain and predict others' actions[2,3], more recent studies of ToM use reaction times, eye tracking measures of anticipation and surprise, live-action measures of intervention, and more[4-9]. In this thesis, I argue that noninvasive neuroimaging measurements of children's developing brains offer a promising additional approach, providing a complementary and, in some cases, unique window on key aspects of ToM development, while simultaneously addressing basic questions about the development of functionally specialized cortical regions.

Remarkably, the past fifteen years of cognitive neuroscience research on ToM in adults has found that a particular network of brain regions, including bilateral temporoparietal junction (R/LTPJ), precuneus (PC), and medial prefrontal cortices (D/M/VMPFC), is reliably and robustly recruited for theory of mind reasoning tasks[10-12]; for reviews, see[13-15]). Responses in the right TPJ are particularly selective for mental state content in adults: this region responds to mental state stimuli (e.g. beliefs, desires), but not to other forms of meta-representation, like false-signs or photographs[10,16], other internal states, like bodily feelings[17-19], or other descriptions of people, like personality traits[19,20]. Thus, preferential, and in some cases highly selective, responses to mental states is one feature of the mature structure of brain regions recruited for theory of mind reasoning. A second feature is an apparent division of labor between "cognitive" and "affective" processes among ToM brain regions[21-24]. While the medial prefrontal cortex contains representations of motivational states and preferences[25-27], bilateral TPJ contains representations of the epistemic history of beliefs[25,28]. Relatedly, within each ToM brain region, mental state information is organized along abstract dimensions[29]. For example, emotion representations are organized according to attribution information, rather than simply valence or arousal[30-32], and belief representations are organized according to epistemic features like justification and source modality of the evidence[25,28].

Much less is known about the "starting state" of ToM brain regions, and the kinds of neural changes that occur during childhood to support ToM development. Though there are relatively few fMRI studies on theory of mind reasoning in children, those that exist, in addition to studies using other methods such as EEG and fNIRS, converge to suggest that children recruit these same regions to reason about mental states by age six years (fMRI[33-35]; fNIRS[36,37]; EEG/ERP[38-41]). By providing clear evidence that ToM brain regions are generally in the same location in

children as in adults, these studies enable subsequent studies to test specific hypotheses about the cognitive processes carried out by these regions[42,43].

Using Neuroimaging Evidence to Gain Insight into Debates about ToM Development

Neuroimaging evidence can provide key insight into the nature of the most known milestone in ToM development: the behavioral shift from failure to success on the false-belief ("Sally – Anne") task. In the classic version, young children see and hear a story about a character (Sally) who has a false belief about the location of her ball ("in the basket"), which is actually in the box, and are asked to predict where Sally will look for the ball[2]. The "false-belief task" is considered the gold standard measure of ToM because to pass the task, children must recognize that another person's mind contains a distinct representation of the world, which may be incomplete or even inaccurate, and that people tend to act based on their personal mental representation, regardless of how well it matches reality. Children reliably pass various versions of explicit false-belief tasks around age four years[44,3].

The major controversy around the false-belief task is about why three-year-olds fail and four-year-olds pass this task[3]. Initially, researchers proposed that false-belief task performance indexes conceptual development: that three-year-old children didn't have a full concept of belief as a person's (mis)representation of the world[44-47]. However, many others noted that classic false-belief tasks are long and complex, requiring children to hold many ideas in mind simultaneously and to choose among competing response options[48,49], so task performance may be a conservative measure of conceptual competence[50]. Indeed, an alternative view emerged, proposing that the concept of belief is fully developed (much) earlier[51,52], and that explicit false-belief task performance only reflected changes in children's ability to meet the task demands, related to development in language and executive functions[53].

Evidence from neuroimaging studies can contribute to this debate because, in both adults and children, distinct brain regions are associated with these different cognitive processes[10,12,54,55]. If initial success on explicit false-belief tasks is driven by conceptual change in theory of mind knowledge, we would expect to see differences in the response of ToM brain regions between children who pass and fail this task. While fMRI studies have contributed evidence about the role of executive function and theory of mind brain regions during false-belief tasks by studying adults and older children, they have yet to contribute to the debate about the role of executive functions in *initially* passing false-belief tasks, due to the inherent challenges of obtaining sufficient, high quality fMRI data from young children. However, EEG evidence has provided initial support for the hypothesis that four-year-olds' emerging success on the false-belief task reflects maturation of brain regions involved in ToM. Sabbagh et al. (2009) measured resting state EEG in four-year-old children who also completed independent behavioral assessments of executive function, language, and theory of mind reasoning, including questions about false-beliefs[41]. Resting state EEG can be used to measure the alpha rhythm (8-13 Hz in adults, 6-9 Hz in children); changes in alpha coherence reflect synchronization of neural firing within and across neural populations, which increases with maturation[56-58]. The maturation of each "region" of cortex was inferred using source localization techniques (sLORETA[59]). Controlling for differences in executive function and language, success on the theory of mind battery was uniquely predicted by maturation of two regions, both in the ToM network: right temporoparietal junction and dorsomedial prefrontal cortex. These results are intriguing, but there is significant

need for future research. First, it will be important to replicate the results of Sabbagh et al. (2009), and to rigorously assess both the source localization technique, and the inference from alpha coherence measured at the scalp to regional cortical maturation. Second, these results leave important open questions. How does synchrony in the resting alpha rhythm of theory of mind regions relate to the function of these regions during task performance, or to the concepts these regions represent? Finally, it is important to situate neural development that supports false-belief task performance in four year olds within the broader context of development in these brain regions, which could begin before and continue after false-belief task reasoning. By doing so, neuroimaging studies could provide insight into whether successful false-belief reasoning involves a conceptual leap in theory of mind development, or is one step in a long, gradual developmental progression of theory of mind achievements.

A second debate about theory of mind development that could be informed by neuroimaging studies concerns the role of developmental experience on ToM development. For example, language abilities in childhood are clearly related to theory of mind reasoning, but the precise role of language in theory of mind development continues to be debated. While some studies suggest that language plays a causal role in theory of mind development[60], others suggest that language may simply be one mode for expressing understanding[51,52]. In typical development, language and theory of mind develop simultaneously, making it difficult to tease apart these two hypotheses. Progress has been made by studying theory of mind reasoning in Deaf children, who have varying ages of exposure to language, corresponding to when they were first exposed to sign language. Deaf children born into non-signing (often hearing) families tend to receive exposure to sign language after an initial delay, whereas Deaf children born to signing families receive exposure to sign language at birth. Interestingly, Deaf children who experience delayed access to sign language have corresponding delays in theory of mind development[61], whereas Deaf children born to signing families do not[62]. Here again neuroimaging studies can contribute key insight because, in children and adults, distinct cortical regions are involved in language[63,64] and theory of mind processing. If behavioral delays in theory of mind reasoning in children with delayed access to language are domain-specific, there should be corresponding delays in the development of theory of mind brain regions.

These two debates are in essence about whether ToM development is domain-specific: do humans have specially designed cognitive and neural mechanisms for representing others' minds that undergo development in childhood? In both cases, domain-specific ToM development should correspond to developmental change or differences specifically in ToM brain regions. The experiments in this thesis use fMRI measurements of functional responses in children who pass and fail false-belief tasks, and in children who experience delayed access to language, in order to directly inform these two debates. In doing so, the experiments in this thesis additionally provide insight into the *kinds* of neural changes that support ongoing theory of mind development in childhood. Even if we assume that developmental change in ToM corresponds to functional changes in the brain regions associated with ToM, many different neural signatures are possible, with correspondingly different implications for cognitive theories. For example, increasingly sophisticated ToM could reflect an increased amount of selective cortex dedicated to theory of mind, faster, less noisy communication between ToM regions, and/or new representational capacities within theory of mind brain regions. I discuss these hypotheses below,

drawing heavily from research in other cognitive domains, and discussing implications for research in theory of mind.

What Kinds of Neural Changes Support Theory of Mind Development?

Development and Refinement of Functionally Selective Responses

As children's theory of mind develops, cortical regions recruited selectively for theory of mind processes may become larger and more selective. Increased cortical real estate for a given cognitive task might reflect an increased number of distinct concepts stored, or the increased use or application of these concepts. The selectivity of a brain region quantifies its response to a certain stimulus category, relative to other (control) categories. In adults, measuring the selectivity of the response in ToM brain regions has refined theories about the functions of these regions[19,65]. Increasing selectivity in childhood could reflect fine-tuning of specialized brain regions to the distinctive conceptual distinctions and computational demands of a particular cognitive domain.

Developmental cognitive neuroscience studies in multiple cognitive domains have provided evidence for the idea that larger, more selective cortical responses develop in childhood and support cognitive change. For example, the magnitude of selective responses in cortical regions specialized for faces (fusiform face area, FFA) and places (parahippocampal place area, PPA) is larger in adults, relative to children (despite similar whole-brain volume), and is correlated with behavioral recognition memory of faces and scenes[66,67,68]. These cortical regions respond less to non-preferred categories with age, and the reduced response to non-preferred categories corresponds to improvements on category-relevant behavioral recognition tasks[69]. Similarly, the volume of symbol selective cortex (visual word form area, VWFA)[70] increases rapidly in children upon learning to read[71], and development of this region involves reduced responses to non-preferred visual categories[72]. A recent longitudinal study measured the emergence of the VWFA in individual children over time, and found that increases in selectivity occur via encroachment of symbol-selective cortex into peripheral, relatively uncommitted cortex[71], rather than into nearby cortex that is particularly selective for non-preferred categories. Low responses to non-preferred categories in the periphery remained stable in individuals over time. It will be important to conduct similar longitudinal studies in other cognitive domains, in order to determine the relative roles of competition for uncommitted cortex versus encroachment into nearby cortex via reduced responses to non-preferred categories, in the development of other functionally specialized regions[73].

What kinds of *structural* changes in the brain are involved in the development of increasingly functionally selective responses? One hypothesis is that increasingly selective functional responses are achieved via synaptic or dendritic pruning: synapses and dendrites that support transmission of non-preferred inputs may weaken over time, leaving those that support transmission of the preferred stimuli (i.e., "use it or lose it"[74-76]). Gomez et al. (2017) used quantitative MRI (qMRI), a neuroimaging technique that directly measures physical tissue properties in vivo, in order to determine whether the development of face-selective responses corresponded to (1) pruning of synapses/dendritic arbors, (2) microproliferation of synapses/dendritic arbors, (3) myelination of relevant axons, or (4) strengthening of potentiation – connections at the molecular level (e.g., via receptor exchange and upregulation)[67]. In contrast to pruning, they found evidence for continued microstructural proliferation in the FFA, and a

significant correlation between proliferation and response selectivity, as well as with behavioral performance on a face recognition memory task. Their evidence was additionally consistent with some role for myelination in the development of selective responses, but they used model predictions of white matter development to argue that it is unlikely that myelination alone drives functional specialization (see the *Improved Communication Between Brain Regions*, below). This evidence is intriguing, and suggests that functional selectivity may develop as a result of the increasing spatial extent over which preferred information is stored and processed, and/or over which non-preferred responses can be inhibited[77-79].

To what extent is the development of functionally selective responses the result of maturational versus experiential factors? While these two factors are conflated in typical development (older children have more life experience), training studies have suggested that the amount of experience may matter for the refinement of functionally selective responses. For example, neural responses in IT show improved differentiation between stimuli across trained category boundaries[80-82]). Thus, developing expertise through exposure to relevant stimuli appears to fine-tune functionally selective responses.

Is extensive experience also necessary for the *initial* development of functionally selective responses? This question is much harder to address. However, a recent study provided evidence for face- and scene-selective responses in four-month-old infants[83], suggesting a more constrained timeline of development for, and a smaller role of experience on, functionally selective responses than previously hypothesized[68] (though also see[84]). This timeline is consistent with that suggested by longitudinal evidence in macaques[73]. These studies do not rule out the hypothesis that some initial relevant experience is necessary to trigger the development of selective responses, and are not mutually exclusive with the hypothesis that some aspects of neural maturation are necessary precursors for the brain to capitalize on experiential input. However, they suggest that initial functionally selective responses are present quite early, and that more extensive developmental experience may be necessary to maintain and refine the boundary between preferred and non-preferred stimuli. While face selective regions in infants responded preferentially to faces over scenes, there was not a preferential response between faces and objects[83]. Many prior studies have suggested that early developmental experience is important for the maintenance and refinement of functionally selective responses. Multiple brain regions appear to have early sensitive or critical periods during which the brain is particularly responsive to and shaped by experiential input. For example, blind adults who regain vision do not appear to subsequently develop typical face-selective responses[85], and a lack of visual input in the first few months of life results in enduring deficits in face perception[86,87]. Studies of adults who transition from illiteracy to literacy have found that reduced responses to non-preferred stimuli in the VWFA are not as robust as those that have been measured in children learning to read[72]. There is significant need for research characterizing early functional responses in ToM brain regions (see General Discussion and Conclusion), as well as the relevant experiential inputs for maintaining and refining selective responses for ToM processes in development.

Is theory of mind development in childhood supported by increasingly selective responses within ToM brain regions? Approaching this question involves addressing two key challenges. The first challenge is to determine the relevant categorical boundaries that would enable measuring selective responses for ToM processes. For example, contrasting faces and scenes evokes face-

and place-selective responses in infants, but contrasting faces and objects does not[83]: sensitivity to the categorical boundary between faces and objects apparently unfolds later in development, and the development of this categorical boundary is relevant for face recognition tasks[69]. What is the equivalent of the faces vs. scenes contrast in the domain of ToM, and what categorical divisions are relevant for cognitive improvements in ToM? Similar to visual categories (faces, objects, symbols), categorical boundaries relevant for ToM processes may be best described in terms of their functional use. The same cortical region does not process faces, objects, and symbols selectively, presumably because conducting cognitive processes over each of these stimulus categories requires sensitivity to different stimulus features (e.g., face recognition and memory involves attending to the distance between eyes, whereas object recognition and memory involves attending to shape and texture). Among adults, ToM brain regions are sensitive to the difference between mentalistic and non-mentalistic stimuli[10], and preferentially respond to minds over bodily states[17-19,88,89]), appearances[19], and enduring personality traits[19,20]. The experiments in this these will characterize the extent to which neural responses in ToM brain regions are sensitive to these across-category boundaries, and the extent to which these boundaries are refined throughout childhood.

Two previous studies have relied on these categorical divisions to provide initial evidence that increases in response selectivity in childhood support ToM reasoning. Saxe et al. (2009) asked children ages six to ten years old to listen stories that contained mental state information (Mental condition), descriptions of social interactions without explicit mental states (Social condition), or physical descriptions of the world (Physical condition) while lying in an fMRI scanner[34]. While children recruited the same ToM regions as adults in response to Mental stories, relative to the Physical control stories, the responses in the right TPJ of the youngest children were also high for the more general Social stories. The response to the Social condition decreased significantly with age, whereas the response to mental state content remained high (and the response to physical control stories remained low). Gweon et al. (2012) replicated this finding in five- to ten-year-old children, and additionally found evidence for a correlation between selectivity in the RTPJ and theory of mind reasoning, assessed behaviorally (see Chapter 1 for a more detailed review of these studies)[35]. Together, these studies suggest that (1) ToM brain regions are recruited for thinking about mental states by age six, (2) the RTPJ becomes more selective for processing mental state content throughout late childhood, and (3) increased specialization is related to ongoing developmental change in theory of mind reasoning. These results are promising, but future work is necessary to determine whether these categorical boundaries are the most relevant for measuring developmental change in ToM. As the size of the selective response increases, the range of stimuli evoking the response should decrease. Future studies that measure developmental change in the refinement of the preferred category and relate these measures to behavioral ToM tasks may provide insight into the conceptual changes involved in the development of ToM.

A second challenge for measuring response selectivity in children (especially children under age 5 years) is that these measures typically depend on collecting high quality responses to many conditions, in order to allow for a stable measure of the relative response. Young children often do not tolerate such long experiments. However, an alternative approach may facilitate acquiring relatively sensitive measures of neural responses across a range of conditions. Inter-subject correlation analyses (ISC[90]) measure reliable and meaningful differences in the timing and

selectivity of responses in functionally specific brain regions[91]. A key benefit of this method is that it is applied to functional data collected while participants view "naturalistic" stimuli, e.g. movies. In developmental contexts, ISC measures can serve as an index of neural maturity by comparing the extent to which timecourses of neural activity are correlated across children, and by comparing each participant's response timecourse to that of adult populations. This "neural maturation" measure can also be related to behavioral abilities. For example, Cantlon & Li (2013) measured activity in the intra-parietal sulcus (IPS; a region implicated in processing number) and the left inferior frontal gyrus (LIFG; a region implicated in processing language) while four- to ten-year-old children watched Sesame Street[92]. Neural maturity (e.g., similarity to the adult response timecourse) in IPS was significantly correlated with behavioral assessments of math, and neural maturity in LIFG was significantly correlated with behavioral assessments of verbal abilities. Thus, these methods are promising for measuring response selectivity, and for relating neural and cognitive change, in young children.

Improved Communication Between Brain Regions

The two studies reviewed above suggest that theory of mind development may be supported by increases in the functional selectivity of ToM brain regions. Improvements in theory of mind might additionally be reflected in faster communication *between* ToM brain regions. Diffusion tensor imaging (DTI) offers a way to directly measure the strength of white matter connections between different brain regions. In other cognitive domains, strengthening of specific white matter tracts, as measured via DTI, corresponds to improvements in particular cognitive skills. For example, development of the arcuate fasciculus is predictive of reading level and improvement[93], and related to performance on a phonological awareness task[94].

Theory of mind reasoning (e.g. understanding why Sally told Anne the marble was in the cupboard, when really it is in the box) requires integrating multiple different inputs (e.g. Does Sally know where the marble is? Is Sally a nice person? Are Sally and Anne friends? Does Sally want Anne to have the marble?). Different regions within the theory of mind network appear to preferentially encode different aspects of a person and his or her mental state. For example, while the RTPJ responds selectively to the content of another person's beliefs[19], and encodes information about the epistemic history of beliefs[25,28], the MPFC responds selectively to thinking about the preferences of others[95], and encodes information about motivational states[25,30-32]. Thus, faster and less noisy communication between brain regions could render theory of mind judgments more accurate and less costly, and could even enable inputs from some regions to "tutor" selective responses in other regions[96]. The strength of white matter tracts may very well be directly related to increasingly functionally selective responses in ToM regions. While some studies suggest that anatomical connections predict and constrain the location[97], spatial layout[98], and function[99] of functionally selective responses, other studies suggest that anatomical connectivity does not preclude[83], constrain[100], or drive[67] the development of functionally selective responses, across a number of cortical regions. Thus, the causal direction of white matter tract development and the development of functionally selective responses could plausibly go either way: enhanced communication between regions could *enable* each region to become even more functionally selective, or increasingly functionally selective regions could *require* faster communication between regions in the ToM network. Interestingly, one study has found a positive correlation between the strength of white matter connections between ToM brain regions and children's performance on false-belief tasks, independent of age[101]. More work

is necessary to relate development of white matter tracts to the development of functionally selective responses, and to clarify the contributions of each for developmental improvements in ToM.

While not a measure of the physical connections between brain regions, inter-region correlation analyses measure the extent to which a set of brain regions have functionally similar response profiles. High inter-region correlations between brain regions indicate that the responses of these brain regions are driven by similar content, and negative inter-region correlations indicate that the responses of brain regions are driven by distinct content[90]. Thus, measuring inter-region correlations during relevant functional tasks may be a promising way to measure network properties and functionally selective responses simultaneously.

Development of New and Refined Representations Within Brain Regions

While response selectivity measures typically focus on distinctions *across* category boundaries (e.g., RTPJ responds to beliefs but not photographs), children's theory of mind development may also predict new conceptual distinctions represented *within* a given category (e.g., beliefs that are justified versus those that are not). These emerging distinctions may be measurable in the responses of ToM brain regions. Representational Similarity Analyses (RSA) and Multi-Voxel Pattern Analyses (MVPA) exploit spatial patterns of activation across voxels to decode many perceptual and conceptual stimulus features[102], and to relate neural response patterns to models of similarity constructed from cognitive theories or behavioral tasks[103]. Some of these representations appear sensitive to cognitive change in adults. For example, when adults learn to discriminate between visual forms based on a particular distinction (e.g. spiral angle), that distinction becomes more linearly decodable in relevant cortical areas[104]. Children's construction of novel conceptual distinctions, even within highly abstract intuitive theories, may correspond to changes in neural representational spaces, and these changes could be measured using MVPA or RSA.

Recent studies with adults have provided evidence that abstract organizational features of beliefs, such as their source modality (e.g. whether someone believes something because of something they heard vs. saw) and justification (e.g. whether someone has strong vs. weak evidence for their belief), are reflected in neural response patterns of theory of mind brain regions[25,28]. In ongoing work not included in this thesis, I test the hypothesis that neural responses patterns can additionally be used to capture the reorganization and refinement of conceptual knowledge that occurs throughout development.

Brief Overview of Thesis

Characterizing the neural changes that support theory of mind development in childhood is one approach to gain novel insight into the developmental processes that bridge the gap between early and adult-like theory of mind capacities. Chapters 1 – 3 describe fMRI experiments conducted with typically developing children with this goal in mind. Specifically, **Chapter 1** includes two longitudinal experiments conducted to (1) determine whether later developments in ToM reasoning reflects changes in social reasoning, as opposed to domain-general skills, and (2) test for predictive relationships between ToM development and response selectivity. Experiment 1 was conducted with children ages 5 – 13 years old; Experiment 2 used identical methods and a younger population (5 – 7 years old), in order to focus on earlier developmental change.

Chapter 2 describes a large, cross-sectional fMRI study that includes even younger children (n=122, ages 3 – 12 years). As described above, a substantial number of behavioral studies involve characterizing young children and infants' performance on false-belief tasks: a task that has served as a gold standard for assessing whether a participant (primate, child, infant) represents the internal states of others' minds[3,4,105]. The participant sample described in Chapter 2 is unique in that it includes a substantial number of children who fail explicit false-belief tasks. We employed a short, engaging naturalistic movie-viewing stimulus to obtain high quality data from young children, and applied timecourse analyses to measure response selectivity and magnitude in ToM brain regions. In doing so, this study provides a glimpse into the neural correlates of passing false-belief tasks, in addition to the neural changes that occur throughout childhood. **Chapter 3** describes confirmatory results based on identical analyses in a large, publicly available dataset of 5 – 12 year old children (n=186) and adolescent/young adults (13 – 20 years old, n=55).

A second approach is to study the impact of specific environmental and experiential factors on ToM development. What are the relevant experiential inputs for developing increasingly selective functional responses in ToM brain regions? **Chapter 4** describes an experiment that directly investigates the impact of early language experience on theory of mind development, behaviorally and neurally. By measuring neural responses in addition to ToM behavior, this study provides unique insight into the nature of the apparent behavioral theory of mind deficits in children who experience delayed access to language, and additionally provides rare insight into the role of language in the development of functionally selective brain regions for theory of mind.

Noninvasive neuroimaging with young children is a relatively new technique, and faces many methodological and theoretical challenges. Both throughout the thesis, and in a concentrated chapter (**Chapter 5**), I address these limitations. The thesis ends with a **General Discussion and Conclusion**, which discusses the results of these experiments in a broader context, and suggests directions for future research.

References

1. Carey, S. *The origin of concepts*. (Oxford University Press, 2009).
2. Wimmer, H. & Perner, J. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *COGNITION* **13,** 103–128 (1983).
3. Wellman, H. M., Cross, D. & Watson, J. Meta-analysis of theory-of-mind development: the truth about false belief. *Child Dev* **72,** 655–684 (2001).
4. Onishi, K. H. & Baillargeon, R. Do 15-month-old infants understand false beliefs? *Science* **308,** 255–258 (2005).
5. Southgate, V., Senju, A. & Csibra, G. Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science* **18,** 587–592 (2007).
6. Knudsen, B. & Liszkowski, U. 18-Month-Olds Predict Specific Action Mistakes Through Attribution of False Belief, Not Ignorance, and Intervene Accordingly. *Infancy* **17,** 672–691 (2012).
7. Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J. & Bodley Scott, S. E. Seeing it their way: evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance* **36,** 1255 (2010).
8. Apperly, I. A., Back, E., Samson, D. & France, L. The cost of thinking about false beliefs: Evidence from adults' performance on a non-inferential theory of mind task. *COGNITION* **106,** 1093–1108 (2008).
9. Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y. & Plumb, I. The 'Reading the Mind in the Eyes' Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry* **42,** 241–251 (2001).
10. Saxe, R. & Kanwisher, N. People thinking about thinking peopleThe role of the temporo-parietal junction in 'theory of mind'. *NeuroImage* **19,** 1835–1842 (2003).
11. Gallagher, H. L. *et al.* Reading the mind in cartoons and stories: an fMRI study of 'theory of mind'in verbal and nonverbal tasks. *Neuropsychologia* **38,** 11–21 (2000).
12. Saxe, R. & WEXLER, A. Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia* **43,** 1391–1399 (2005).
13. Carrington, S. J. & Bailey, A. J. Are there theory of mind regions in the brain? A review of the neuroimaging literature. *Hum. Brain Mapp.* **30,** 2313–2335 (2009).
14. Dubois, J. & Adolphs, R. How the brain represents other minds. *Proceedings of the National Academy of Sciences* **113,** 19–21 (2016).
15. Kliemann, D. & Adolphs, R. The social neuroscience of mentalizing: challenges and recommendations. *Current opinion in psychology* (2018).
16. Dodell-Feder, D., Koster-Hale, J., Bedny, M. & Saxe, R. fMRI item analysis in a theory of mind task. *NeuroImage* **55,** 705–712 (2011).
17. Bruneau, E. G., Pluta, A. & Saxe, R. Distinct roles of the 'shared pain'and "theory of mind"networks in processing others' emotional suffering. *Neuropsychologia* **50,** 219–231 (2012).
18. Jacoby, N., Bruneau, E., Koster-Hale, J. & Saxe, R. Localizing Pain Matrix and Theory of Mind networks with both verbal and non-verbal stimuli. *NeuroImage* **126,** 39–48 (2016).
19. Saxe, R. & Powell, L. J. It's the thought that counts specific brain regions for one

component of theory of mind. *Psychological Science* **17,** 692–699 (2006).
20. Mitchell, J. P., Banaji, M. R. & Macrae, C. N. General and specific contributions of the medial prefrontal cortex to knowledge about mental states. *NeuroImage* **28,** 757–762 (2005).
21. Schlaffke, L. *et al.* Shared and nonshared neural networks of cognitive and affective theory-of-mind: A neuroimaging study using cartoon picture stories. *Hum. Brain Mapp.* **36,** 29–39 (2014).
22. Shamay-Tsoory, S. G. & Aharon-Peretz, J. Dissociable prefrontal networks for cognitive and affective theory of mind: A lesion study. *Neuropsychologia* **45,** 3054–3067 (2007).
23. Shamay-Tsoory, S. G., Tibi-Elhanany, Y. & Aharon-Peretz, J. The ventromedial prefrontal cortex is involved in understanding affective but not cognitive theory of mind stories. *Social Neuroscience* **1,** 149–166 (2006).
24. Sebastian, C. L. *et al.* Neural processing associated with cognitive and affective Theory of Mind in adolescents and adults. *Social Cognitive and Affective Neuroscience* **7,** 53–63 (2012).
25. Koster-Hale, J. *et al.* Mentalizing regions represent distributed, continuous, and abstract dimensions of others' beliefs. *NeuroImage* **161,** 9–18 (2017).
26. Winecoff, A. *et al.* Ventromedial Prefrontal Cortex Encodes Emotional Value. *Journal of Neuroscience* **33,** 11032–11039 (2013).
27. Leopold, A. *et al.* Damage to the left ventromedial prefrontal cortex impacts affective theory of mind. *Social Cognitive and Affective Neuroscience* **7,** 871–880 (2011).
28. Koster-Hale, J., Bedny, M. & Saxe, R. Thinking about seeing: Perceptual sources of knowledge are encoded in the theory of mind brain regions of sighted and blind adults. *COGNITION* **133,** 65–78 (2014).
29. Tamir, D. I., Thornton, M. A., Contreras, J. M. & Mitchell, J. P. Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences* **113,** 194–199 (2016).
30. Peelen, M. V., Atkinson, A. P. & Vuilleumier, P. Supramodal Representations of Perceived Emotions in the Human Brain. *Journal of Neuroscience* **30,** 10127–10134 (2010).
31. Skerry, A. E. & Saxe, R. A common neural code for perceived and inferred emotion. *J. Neurosci.* **34,** 15997–16008 (2014).
32. Skerry, A. E. & Saxe, R. Neural Representations of Emotion Are Organized around Abstract Event Features. *Current Biology* **25,** 1945–1954 (2015).
33. Kobayashi, C., Glover, G. H. & Temple, E. Children's and adults' neural bases of verbal and nonverbal "theory of mind". *Neuropsychologia* **45,** 1522–1532 (2007).
34. Saxe, R. R., Whitfield-Gabrieli, S., Scholz, J. & Pelphrey, K. A. Brain regions for perceiving and reasoning about other people in school-aged children. *Child Dev* **80,** 1197–1209 (2009).
35. Gweon, H., Dodell-Feder, D., Bedny, M. & Saxe, R. Theory of Mind Performance in Children Correlates With Functional Specialization of a Brain Region for Thinking About Thoughts. *Child Dev* **83,** 1853–1868 (2012).
36. Bowman, L. C., Kovelman, I., Hu, X. & Wellman, H. M. Children's belief-and desire-reasoning in the temporoparietal junction: evidence for specialization from functional near-infrared spectroscopy. *Front. Hum. Neurosci.* **9,** 560 (2015).
37. Hyde, D. C., Simon, C. E., Ting, F. & Nikolaeva, J. Functional organization of the

temporal-parietal junction for theory of mind in preverbal infants: A near-infrared spectroscopy study. *Journal of Neuroscience* 0264–17 (2018).
38. Liu, D., Sabbagh, M. A., Gehring, W. J. & Wellman, H. M. Decoupling beliefs from reality in the brain: an ERP study of theory of mind. *NeuroReport* **15,** 991–995 (2004).
39. Liu, D., Sabbagh, M. A., Gehring, W. J. & Wellman, H. M. Neural correlates of children's theory of mind development. *Child Dev* **80,** 318–326 (2009).
40. Bowman, L. C., Liu, D., Meltzoff, A. N. & Wellman, H. M. Neural correlates of belief- and desire-reasoning in 7- and 8-year-old children: an event-related potential study. *Dev Sci* **15,** 618–632 (2012).
41. Sabbagh, M. A., Bowman, L. C., Evraire, L. E. & Ito, J. M. B. Neurodevelopmental correlates of theory of mind in preschool children. *Child Dev* **80,** 1147–1162 (2009).
42. Saxe, R., Brett, M. & Kanwisher, N. Divide and conquer: a defense of functional localizers. *NeuroImage* **30,** 1088–1096 (2006).
43. Kanwisher, N. The quest for the FFA and where it led. *Journal of Neuroscience* **37,** 1056–1061 (2017).
44. Perner, J., Leekam, S. R. & Wimmer, H. Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology* **5,** 125–137 (1987).
45. Wimmer, H. & Weichbold, V. Children's theory of mind: Fodor's heuristics examined. *COGNITION* **53,** 45–57 (1994).
46. Flavell, J. H. The development of children's knowledge about the mind: From cognitive connections to mental representations. *Developing theories of mind* 244–267 (1988).
47. Gopnik, A. Theories and illusions. *Behav Brain Sci* **16,** 90–100 (1993).
48. Zaitchik, D. When representations conflict with reality: The preschooler's problem with false beliefs and "false" photographs. *COGNITION* **35,** 41–68 (1990).
49. Riggs, K. J., Peterson, D. M., Robinson, E. J. & Mitchell, P. Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? *Cognitive Development* **13,** 73–90 (1998).
50. Bloom, P. & German, T. P. Two reasons to abandon the false belief task as a test of theory of mind. *COGNITION* **77,** B25–B31 (2000).
51. Baillargeon, R., Scott, R. M. & He, Z. False-belief understanding in infants. *Trends in Cognitive Sciences* **14,** 110–118 (2010).
52. Scott, R. M. & Baillargeon, R. Early False-Belief Understanding. *Trends in Cognitive Sciences* (2017).
53. German, T. P., Leslie, A. M., Mitchell, P. & Riggs, K. Attending to and learning about mental states. *Children's reasoning and the mind* 229–252 (2000).
54. Duncan, J. & Owen, A. M. Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences* **23,** 475–483 (2000).
55. Fedorenko, E., Duncan, J. & Kanwisher, N. Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences* **110,** 16616–16621 (2013).
56. Klimesch, W. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain research reviews* **29,** 169–195 (1999).
57. Thatcher, R. W. Cyclic cortical reorganization during early childhood. *Brain and Cognition* **20,** 24–50 (1992).
58. Nunez, P. L. & Cutillo, B. A. *Neocortical dynamics and human EEG rhythms.* (Oxford

University Press, USA, 1995).
59. Pascual-Marqui, R. D. Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. *Methods Find Exp Clin Pharmacol* **24,** 5–12 (2002).
60. Pyers, J. E. & Senghas, A. Language Promotes False-Belief Understanding: Evidence From Learners of a New Sign Language. *Psychological Science* **20,** 805–812 (2009).
61. Peterson, C. C. & Siegal, M. Representing inner worlds: Theory of mind in autistic, deaf, and normal hearing children. *Psychological Science* **10,** 126–129 (1999).
62. de Villiers, P. A. The role of language in theory-of-mind development: what deaf children tell us. in (Oxford University Press, 2005).
63. Fedorenko, E., Hsieh, P. J., Nieto-Castanon, A., Whitfield-Gabrieli, S. & Kanwisher, N. New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *Journal of Neurophysiology* **104,** 1177–1194 (2010).
64. Fedorenko, E., Behr, M. K. & Kanwisher, N. Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences* **108,** 16428–16433 (2011).
65. Saxe, R. & Young, L. Theory of Mind: How brains think about thoughts. *The Oxford Handbook of Cognitive Neuroscience* **2,** 204–213 (2013).
66. Golarai, G. *et al.* Differential development of high-level visual cortex correlates with category-specific recognition memory. *Nature Publishing Group* **10,** 512 (2007).
67. Gomez, J. *et al.* Microstructural proliferation in human cortex is coupled with the development of face processing. *Science* **355,** 68–71 (2017).
68. Grill-Spector, K., Golarai, G. & Gabrieli, J. Developmental neuroimaging of the human ventral visual cortex. *Trends in Cognitive Sciences* **12,** 152–162 (2008).
69. Cantlon, J. F., Pinel, P., Dehaene, S. & Pelphrey, K. A. Cortical Representations of Symbols, Objects, and Faces Are Pruned Back during Early Childhood. *Cerebral Cortex* **21,** 191–199 (2010).
70. McCandliss, B. D., Cohen, L. & Dehaene, S. The visual word form area: expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences* **7,** 293–299 (2003).
71. Dehaene-Lambertz, G., Monzalvo, K. & Dehaene, S. The emergence of the visual word form: Longitudinal evolution of category-specific ventral visual areas during reading acquisition. *Plos Biol* **16,** e2004103 (2018).
72. Dehaene, S. *et al.* How learning to read changes the cortical networks for vision and language. *Science* **330,** 1359–1364 (2010).
73. Livingstone, M. S. *et al.* Development of the macaque face-patch system. *Nature Communications* **8,** 14897 (2017).
74. Changeux, J.-P. & Danchin, A. Selective stabilisation of developing synapses as a mechanism for the specification of neuronal networks. *Nature* **264,** 705 (1976).
75. Changeux, J.-P. & Dehaene, S. Neuronal models of cognitive functions. *COGNITION* **33,** 63–109 (1989).
76. Bourgeois, J.-P. & Rakic, P. Changes of synaptic density in the primary visual cortex of the macaque monkey from fetal to adult stage. *Journal of Neuroscience* **13,** 2801–2820 (1993).
77. Allison, T., Puce, A. & McCarthy, G. Category-sensitive excitatory and inhibitory processes in human extrastriate cortex. *Journal of Neurophysiology* **88,** 2864–2868 (2002).
78. Pelphrey, K. A., Mack, P. B., Song, A., Güzeldere, G. & McCarthy, G. Faces evoke

spatially differentiated patterns of BOLD activation and deactivation. *NeuroReport* **14,** 955–959 (2003).
79. Purves, D., White, L. E. & Riddle, D. R. Is neural development Darwinian? *Trends in Neurosciences* **19,** 460–464 (1996).
80. Sigala, N. & Logothetis, N. K. Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature* **415,** 318–320 (2002).
81. Freedman, D. J. & Assad, J. A. Experience-dependent representation of visual categories in parietal cortex. *Nature* **443,** 85 (2006).
82. Srihasam, K., Vincent, J. L. & Livingstone, M. S. Novel domain formation reveals proto-architecture in inferotemporal cortex. *Nature Publishing Group* **17,** 1776 (2014).
83. Deen, B. *et al.* Organization of high-level visual cortex in human infants. *Nature Communications* **8,** 13995 (2017).
84. McKone, E., Crookes, K., Jeffery, L. & Dilks, D. D. A critical review of the development of face recognition: Experience is less important than previously believed. *Cognitive Neuropsychology* **29,** 174–212 (2012).
85. Fine, I. *et al.* Long-term deprivation affects visual perception and cortex. *Nature Publishing Group* **6,** 915 (2003).
86. Grand, R. L., Mondloch, C. J., Maurer, D. & Brent, H. P. Impairment in holistic face processing following early visual deprivation. *Psychological Science* **15,** 762–768 (2004).
87. Le Grand, R., Mondloch, C. J., Maurer, D. & Brent, H. P. Expert face processing requires visual input to the right hemisphere during infancy. *Nature Publishing Group* **6,** 1108 (2003).
88. Lombardo, M. V. *et al.* Shared neural circuits for mentalizing about the self and others. *Journal of Cognitive Neuroscience* **22,** 1623–1635 (2010).
89. Spunt, R. P., Kemmerer, D. & Adolphs, R. The neural basis of conceptualizing the same action at different levels of abstraction. *Social Cognitive and Affective Neuroscience* nsv084 (2015).
90. Hasson, U. Intersubject Synchronization of Cortical Activity During Natural Vision. *Science* **303,** 1634–1640 (2004).
91. Hasson, U., Malach, R. & Heeger, D. J. Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences* **14,** 40–48 (2010).
92. Cantlon, J. F. & Li, R. Neural activity during natural viewing of Sesame Street statistically predicts test scores in early childhood. *Plos Biol* **11,** e1001462 (2013).
93. Yeatman, J. D. *et al.* Anatomical properties of the arcuate fasciculus predict phonological and reading skills in children. *Journal of Cognitive Neuroscience* **23,** 3304–3317 (2011).
94. Saygin, Z. M. *et al.* Tracking the roots of reading ability: white matter volume and integrity correlate with phonological awareness in prereading and early-reading kindergarten children. *Journal of Neuroscience* **33,** 13251–13258 (2013).
95. Jenkins, A. C. & Mitchell, J. P. Medial prefrontal cortex subserves diverse forms of self-reflection. *Social Neuroscience* **6,** 211–218 (2011).
96. Johnson, M. H. Functional brain development in humans. *Nat Rev Neurosci* **2,** 475–483 (2001).
97. Saygin, Z. M. *et al.* Connectivity precedes function in the development of the visual word form area. *Nat Neurosci* **19,** 1250–1255 (2016).

98. Osher, D. E. *et al.* Structural connectivity fingerprints predict cortical selectivity for multiple visual categories across cortex. *Cereb. Cortex* **26,** 1668–1683 (2015).
99. Melchner, Von, L., Pallas, S. L. & Sur, M. Visual behaviour mediated by retinal projections directed to the auditory pathway. *Nature* **404,** 871 (2000).
100. Bedny, M., Pascual-Leone, A., Dodell-Feder, D., Fedorenko, E. & Saxe, R. Language processing in the occipital cortex of congenitally blind adults. *Proceedings of the National Academy of Sciences* **108,** 4429–4434 (2011).
101. Wiesmann, C. G., Schreiber, J., Singer, T., Steinbeis, N. & Friederici, A. D. White matter maturation is associated with the emergence of Theory of Mind in early childhood. *Nature Communications* **8,** 14692 (2017).
102. Kamitani, Y. & Tong, F. Decoding the visual and subjective contents of the human brain. *Nat Neurosci* **8,** 679–685 (2005).
103. Kriegeskorte, N. & Kievit, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences* **17,** 401–412 (2013).
104. Zhang, J., Meeson, A., Welchman, A. E. & Kourtzi, Z. Learning alters the tuning of functional magnetic resonance imaging patterns for visual forms. *J. Neurosci.* **30,** 14127–14133 (2010).
105. Premack, D. & Woodruff, G. Does the chimpanzee have a theory of mind? *Behav Brain Sci* **1,** 515–526 (1978).

Chapter 1: Longitudinal Studies of Behavioral and Neural Theory of Mind Development

Children's understanding of others' beliefs and desires (or Theory of Mind, ToM) undergoes dramatic change throughout childhood. A plausible neural correlate of ToM development is increased selectivity of right temporoparietal junction (RTPJ) and dorso-medial prefrontal cortex (DMPFC). Two longitudinal fMRI studies were conducted with children ages five to twelve years old (Study 1; n=31) and five to seven years old (Study 2; n=27). ToM reasoning improved with age and showed stable individual differences across children. However, selectivity in RTPJ and DMPFC increased with age in Study 2 only, and was not correlated with ToM ability. Measures of a region's overall selectivity may not be stable or sensitive enough to capture conceptual change in ToM.

Note: This manuscript is currently under review as:
Richardson, H., Gweon, H., Alves, L., Saxe, R. (under review). Longitudinal Studies of Behavioral and Neural Theory of Mind Development.

Introduction

Theory of mind – the commonsense "folk psychology" that guides our reasoning about the minds of others – is a core component of social cognition. While research on theory of mind (ToM) typically focuses on development during infancy and preschool years, our ability to reason about the minds of others undergoes continual change throughout childhood. This "late" developmental change from early childhood to adolescence supports sophisticated inferences, decisions, and evaluations in complex social contexts[1-4]. FMRI studies have suggested that adults recruit a specific set of brain regions when engaging in such social processes (for reviews, see[5-8], and that these same brain regions are recruited for theory of mind reasoning in five to twelve year old children[9-12]. However, most behavioral research and almost all neuroimaging studies of ToM in childhood is cross-sectional. Here, we conduct two longitudinal studies in order to test if later developing aspects of ToM reasoning are conceptually related to earlier ToM abilities, and to characterize the neural signatures that support, reflect, and predict the development of theory of mind in childhood, within individual children.

A mature "theory of mind" employs many different concepts for reasoning about the minds of others, and behavioral research on ToM development largely focuses on describing the age of acquisition of specific ToM concepts. For example, hundreds of behavioral studies have documented children's transition to explicitly (verbally) reasoning about others' false beliefs, which occurs around four years of age[13]. More recently, a growing number of studies have begun to describe the conceptual repertoire for and precise limitations of earlier social cognitive abilities in toddlers and infants[14,15,16]. Similarly, research with older children describes aspects of ToM reasoning that continue to improve dramatically after age five years, like understanding non-literal speech[1,2,17], making judgments about moral blame-worthiness based on intention[3], and recognizing that the way someone feels may not match the emotion they express[4].

What is the relationship between these cognitive changes in childhood and development in ToM brain regions? Across multiple cognitive domains, increased selectivity of responses in functionally selective brain regions is a neural signature of development in childhood. For example, responses to non-preferred categories in face-selective and symbol-selective cortex decrease with age[18-21], and among children, increased selectivity is correlated with improvement on relevant behavioral tasks[18,19,21]. Similarly, there is evidence that the response in the right superior temporal sulcus becomes more selective for biological motion between the ages of seven and ten years of age[22], but this result has yet to be replicated.

Two previous studies suggest that the response in brain regions reliably recruited for ToM reasoning tasks becomes more selective for mental states between ages five to twelve years. In the first study, children were asked to listen to stories that contained information about mental states (Mental), social facts about people, like descriptions of kinship or appearance (Social), or physical descriptions of the world (Physical) while undergoing fMRI[11]. Children recruited the same ToM regions as adults (bilateral temporoparietal junction (TPJ), precuneus, and medial prefrontal cortex) when listening to the Mental stories, as compared to the Physical control stories. However, in the youngest children, the right TPJ also had high responses to the Social stories. While the response to the Mental condition remained high, the response to the Social condition decreased significantly with age (cross-sectionally). These data suggest that responses in ToM brain regions, and specifically RTPJ, become more selective for processing mental state

content throughout late childhood. In the second study, Gweon and colleagues used the same experimental conditions to replicate the finding that selectivity increases with age, and additionally found evidence that the selectivity of the response in the RTPJ was correlated with performance on a behavioral ToM task administered outside of the scanner, controlling for age[12]. Together, these two studies provide evidence for ongoing behavioral and neural change in ToM after age five, and suggest that behavioral ToM development is related to changes in selectivity within the RTPJ.

Although prior work provides suggestive evidence for a relationship between selectivity and developmental change in ToM in five to twelve year old children, it has relied on cross-sectional observations. Correlational relationships between neural and cognitive change cannot provide information about predictive relationships: does early selectivity support and drive behavioral change in ToM, or vice versa? Studying predictive relationships is critical for obtaining a basic understanding of the relationship between brain development and cognitive change. For example, cross-lagged correlations can reveal the causal direction of an association: e.g. whether earlier brain maturity predicts later gains in conceptual sophistication, or vice versa. This kind of information can be particularly informative for designing cognitive or neural interventions. To date, most longitudinal studies using MRI have characterized gross anatomical changes across development, rather than changes in functional responses[23-25]. Longitudinal MRI studies of functional responses are a promising approach for studying stable individual differences and developmental change at the neural level, but have not yet been used to study theory of mind development in childhood (see[26-28] for relevant longitudinal fMRI studies in adolescence).

Longitudinal behavioral studies of ToM have revealed that there are stable individual differences in theory of mind abilities, and that measures of ToM at different ages tap the same core social cognitive ability. For example, the ability to reason about diverse desires, diverse beliefs, and knowledge access at age four is predictive of reasoning about false beliefs and hidden emotions a year later[29-33]. However, the oldest children included in these studies tend to be five or six years old. The first goal of the current study was to extend longitudinal studies of behavioral ToM development to include older children, in order to test whether later milestones reflect domain-specific conceptual change in theory of mind. To accomplish this goal, we designed a novel longitudinal behavioral measure of ToM that includes questions that require reasoning about ToM concepts that are generally mastered by five years of age (e.g., similar and diverse desires, knowledge access, true and false beliefs, ambiguous referents[13,34,35], as well as concepts that are generally mastered later in childhood (e.g., non-literal speech, moral blameworthiness, and hidden emotions[1-4,17]. If this behavioral measure successfully captures developmental change as well as stable individual differences in theory of mind, longitudinally, it could then be used for studying predictive relationships between behavioral change in ToM and developmental changes in the brain. The second goal of this study was to use longitudinal behavioral and neural data to describe the relationship between cognitive and neural changes in ToM.

We conducted two longitudinal studies of behavioral and neural theory of mind development in children ages five to twelve (Study 1) and five to seven (Study 2), with these two key goals: (1) to develop a novel behavioral task that measures longitudinal behavioral change in theory of mind in relatively older children, and (2) to relate behavioral change in ToM to reliable markers of neural development in ToM brain regions. All children completed two visits, two years (Study 1) or one year (Study 2) apart, each of which included an fMRI scan and a behavioral ToM

battery. First, we tested for developmental change and stable individual differences in behavioral and neural measures of theory of mind. Finding behavioral and neural measures of ToM that reflect change with age and stable individual differences is essentially a prerequisite for relating these measures to one another, and discovering predictive relationships between neural and behavioral change. Second, we tested if behavioral theory of mind ability 1) predicts later measures of selectivity of the response in ToM brain regions), 2) is predicted by an earlier measure of selectivity in ToM brain regions, and 3) improves with increases in selectivity across the two visits. All planned analyses were pre-registered via the Open Science Framework (OSF; https://osf.io/jh68b/).

Results

Despite some methodological differences between Study 1 and Study 2 (in particular, the ages of the children studied), the results of the two studies are largely the same. Results are described for the two studies together, and differences between the studies are noted.

1. Behavioral Results

During Visit 1, children completed the behavioral ToM battery used in Gweon et al. (2012). In order to measure ToM during the second visit, we designed a novel ToM behavioral battery that included analogous questions to those used in V1, as well as more challenging ToM questions (see Methods). The more challenging questions were included in order to more sensitively capture individual differences among older children. In order to measure developmental change over time, we used questions that were matched (denoted by a subscript M) in difficulty across visits; matched items include all V1 questions and a subset of V2 questions (ToM Development = $V2_M - V1$). We tested for developmental change with age, cross-sectionally (between children) and longitudinally (within children), as well as for stable individual differences in ToM across visits. See Table 1 for full regression statistics.

1.1 Developmental Change in Behavioral Theory of Mind

In both studies, theory of mind behavior improved with age cross-sectionally (effect of age: ps < $5x10^{-7}$; regression did not include subject identifier as a random effect), as well as within individual children (longitudinally; effect of age: ps=0; regression included subject identifier as a random effect). Across both studies, all but two children showed improvement in ToM, one of whom performed at ceiling at both visits. In **Study 1**, in a regression that simultaneously tested for effects of within- and between-subject differences in age on the matched ToM score, both variables had significant positive effects on ToM, but the within-subject variable was a stronger predictor of ToM (effect of between-subject age: p=.0001, effect of within-subject age: p=0). In **Study 2**, only within-subject change in age significantly predicted ToM behavior (effect of between-subject age: p=.17, effect of within-subject age: p=0). See Table 1 for regression statistics, Figure 1, Figure 2, and Supplementary Figures 1 and 2.

1.2 Stable Individual Differences in Behavioral Theory of Mind

In both studies, theory of mind behavior at V1 was a significant positive predictor of theory of mind behavior at V2, controlling for average age (ps<.01, Table 1 and Figure 1). In order to control for other relevant behavioral variables, we first tested if verbal IQ, nonverbal IQ, or response inhibition independently predicted theory of mind, and subsequently controlled for variables that independently predicted ToM. In **Study 1**, standardized verbal IQ was the only

additional behavioral measure to independently predict (complete) theory of mind score at V2, when including age as a covariate (Verbal IQ (PPVT): b=-.40, t=-2.4, p=.02, NS effect of age: b=.27, t=1.7, p=.11; other behavioral measures: Nonverbal IQ (KBIT): p=.57; Response Inhibition (Flanker): p=.70). The relationship between theory of mind at Visit 1 and Visit 2 remained significant in a regression that included verbal IQ and average age as covariates (p=.0005, see Table 1). In **Study 2**, none of the other behavioral measures (besides ToM score) independently predicted theory of mind score at V2, when including age as a covariate (Verbal IQ (PPVT): p=.89; Nonverbal IQ (KBIT): p=.60; Response inhibition (DCCS): p=.67).

Table 1.

1.1 Developmental Change in Behavioral Theory of Mind	**Study**	**Predictor**	**Beta**	**T-value**	**p-value**
Cross-sectionally: lm(ToM ~ Age)	**Study 1**	Age	0.72	8.1	**3.5x10-11**
	Study 2	Age	0.64	5.8	**4.3x10-7**
Longitudinally: lme(ToM ~ Age + 1|SubID)	**Study 1**	Age	0.72	8.1	**0**
	Study 2	Age	0.68	7.8	**0**
Simultaneous test of within- and between- subject age differences: lme(ToM ~ Age_{Av}+ $Age_{w/i\text{-}sub}$ + 1|SubID)	**Study 1**	Age_{Av}	0.37	4.4	**0.0001**
		$Age_{w/i\text{-}sub}$	0.68	8.2	**0**
	Study 2	Age_{Av}	0.18	1.4	0.17
		$Age_{w/i\text{-}sub}$	0.63	7.7	**0**
1.2 Stable Individual Differences in Behavioral Theory of Mind	**Study**	**Predictor**	**Beta**	**T-value**	**p-value**
Controlling for age: lme(ToM_{V2C} ~ ToM_{V1} + Age_{Av} + 1|SubID)	**Study 1**	ToM_{V1}	0.58	3.0	**0.005**
		Age_{Av}	-0.05	-0.3	0.80
	Study 2	ToM_{V1}	0.49	2.8	**0.01**
		Age_{Av}	0.31	1.8	0.09
Controlling for variables that independently predict V2 ToM: lme(ToM_{V2C} ~ ToM_{V1} + Age_{Av} + VIQ + 1|SubID)	**Study 1**	ToM_{V1}	0.64	3.9	**0.0005**
		VIQ	-0.46	-3.4	**0.002**
		Age_{Av}	-0.09	-0.6	0.57

Table 1. Full Regression Statistics for Behavioral Results. Linear regression equations are shown in the left column; statistical results are shown in the right column. Section headers correspond to those used in the results section (sections 1.1 and 1.2 in Results of main text). Abbreviations: ToM: proportion correct on ToM behavioral task; matched score is used unless otherwise specified (ToM_{V2C}: Visit 2 "complete" score, which uses all items instead of only items that were matched across visits); Age: chronological age per participant per visit; Age_{Av}: average age per participant, across the two visits (between-subject age differences); $Age_{w/i\text{-}sub}$: difference between participant's average age and their age at each visit (within-subject change in age); V1: Visit 1; V2: Visit 2; VIQ: standardized verbal IQ, as measured by PPVT; 1|SubID: random effect of subject. P-values of significant results (p<.05) are in bold.

2. fMRI Results

2.1 Developmental Change in Response Selectivity

Following Gweon et al. (2012), we tested for significant increases in selectivity with age, and for a significant positive correlation between selectivity and ToM behavior. In **Study 1** (ages 5-12 years), selectivity did not increase with age, cross-sectionally or longitudinally (with subject identifier included as a random effect; ps=.13). In a regression simultaneously testing for effects of within- and between-subject differences in age, neither variable had a significant effect on selectivity (ps>.3). Similarly, within- and between-subject differences in ToM behavior did not have significant effects on selectivity in Study 1 (ps>.3; Figure 2). Descriptively, selectivity was similarly high across visits in RTPJ, and decreased between visits in DMPFC (M(SE) selectivity in RTPJ: V1: 62.3(5.5), V2: 62.3(4.8); DMPFC: V1: 56.6(5.8), V2: 48.3(5.9)). In sum, in Study 1, where participants' age ranged from 5-12 years, we did not find a significant relationship between age or ToM score and response selectivity in ToM brain regions.

In **Study 2** (ages 5-7 years), selectivity increased significantly with age, cross-sectionally and longitudinally (ps=.01). Descriptively, selectivity increased across visits in both regions (M(SE) selectivity in RTPJ: V1: 55.8(6.4), V2: 58.9(6.8); DMPFC: V1: 46.9(5.6), V2: 71.3(6.7)). In a regression simultaneously testing for effects of within- and between-subject differences in age, only within-subject change in age had a marginal positive effect on selectivity (p=.07). As in Study 1, there was not a significant relationship between selectivity and within- or between-subject differences in ToM behavior (ps>.35). See Table 2 for full regression statistics, Figure 1, Figure 2, and Supplementary Figures 2 and 3.

Figure 1

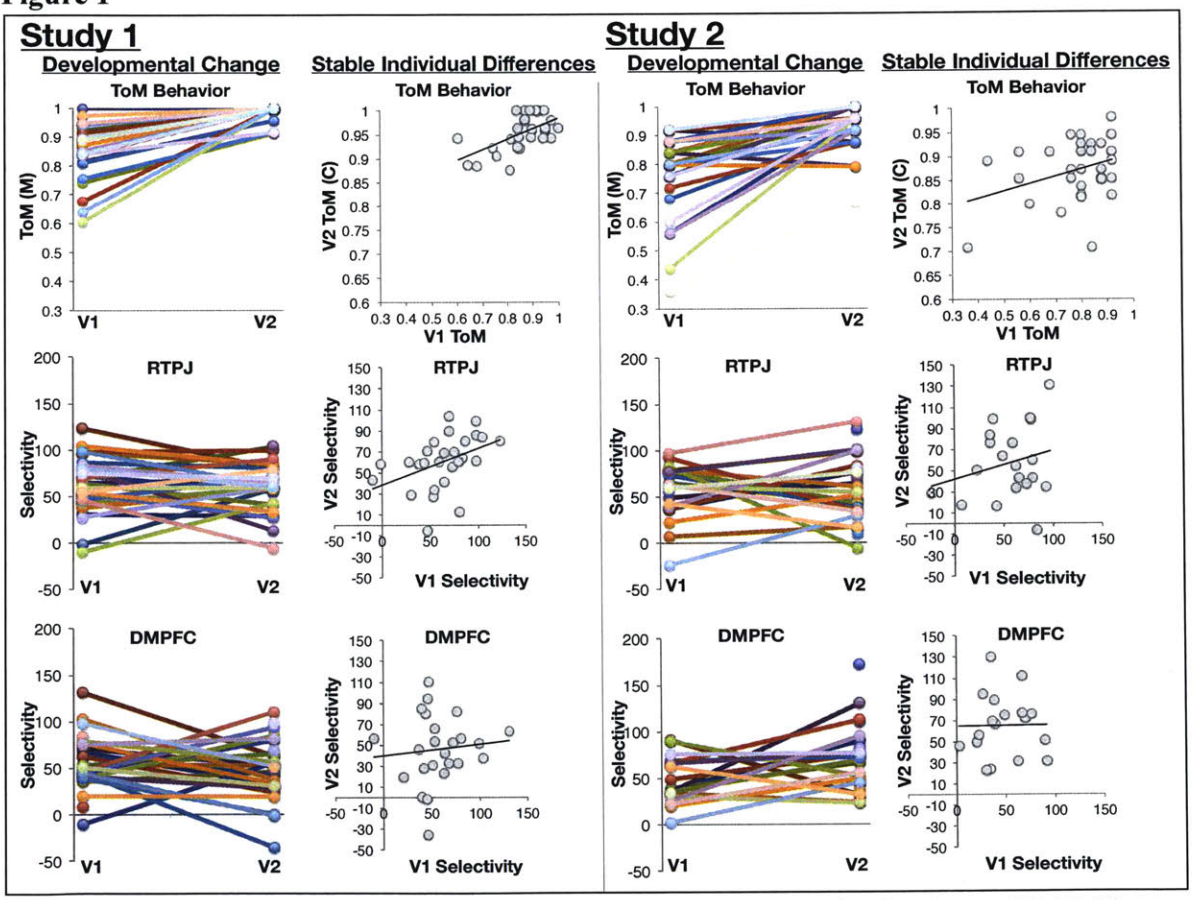


Figure 1. Developmental change and stable individual differences in behavioral and neural ToM. First two columns show data from Study 1; second two columns show data from Study 2. For each study, the left column shows developmental change within individual participants between the two visits in (top row) theory of mind behavioral performance (matched score), (middle row) selectivity of RTPJ, and (bottom row) selectivity of DMPFC. The right column shows stable individual differences in ToM behavior (top row) between the two visits, and a lack of stable individual differences in selectivity of RTPJ (middle row) and DMPFC (bottom row). For the stable individual difference plots of ToM behavior, the Visit 2 theory of mind measure is the "complete" score; e.g., using all items and not only those that are matched across visits.

Figure 2

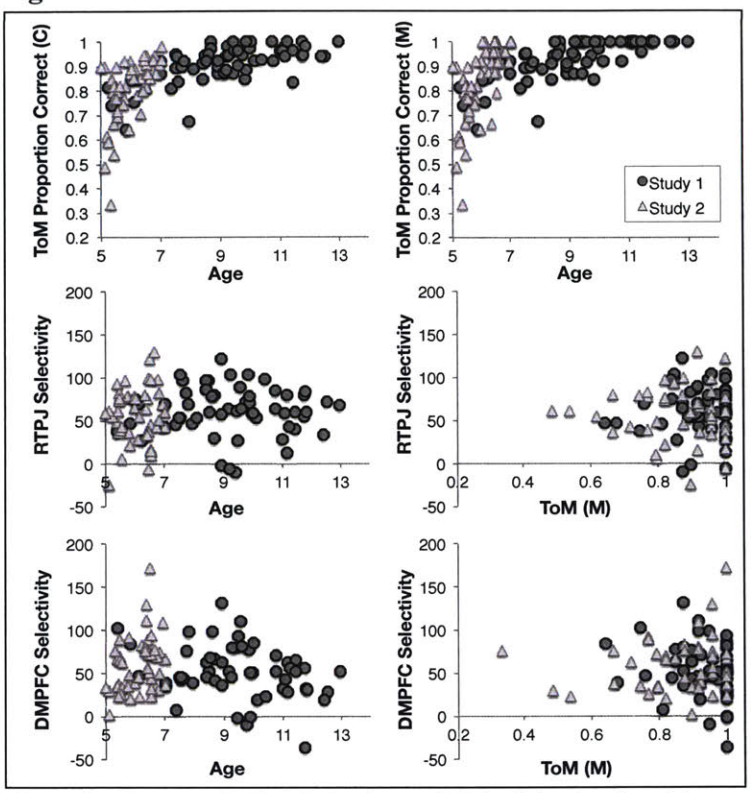


Figure 2. Theory of Mind behavior and Selectivity by Age. The top row shows theory of mind score (left: complete score; right: matched score) (y-axis) by age (x-axis). The bottom two rows show selectivity (y-axis) by (left) age (x-axis) and (right) theory of mind score (x-axis; matched score) for RTPJ (middle row) and DMPFC (bottom row). All scatterplots show data from both visits (e.g., two data points per participant per region of interest). Data from Study 1 is shown in dark grey circles; data from Study 2 is shown in light grey triangles.

2.2 Stable Neural Individual Differences

We tested for stable individual differences in selectivity: does a participant with a more selective response (relative to other children) at Visit 1 also have a more selective response at Visit 2? In **Study 1**, V1 selectivity marginally positively predicted V2 selectivity (p=.10, see Table 2). Subsequent exploratory analyses of each ROI separately found that this marginal effect was driven by a stable individual difference in selectivity in RTPJ (see Figure 1, and the "Stable Individual Differences in RTPJ & Relationship to ToM behavior" section, below). In **Study 2**, V1 selectivity did not significantly predict V2 selectivity (across both ROIs: p=.66; in either ROI individually: ps>.4, Table 2).

2.3 Predictive Relationships

Finally, we tested if behavioral ToM ability 1) predicted later measures of selectivity in ToM brain regions, or 2) was predicted by earlier measures of selectivity in ToM brain regions. We also tested if the amount of change in ToM between the two visits was correlated with the amount of change in selectivity. In both studies, ToM behavior at V1 did not predict selectivity at V2 (ps>.3, see Supplementary Table 3), and selectivity at V1 in RTPJ and DMPFC (tested separately) did not predict ToM behavior at V2 (ps>.14). In **Study 1**, theory of mind development ($V2_{Matched}$ – V1 ToM score) was not related to change in selectivity between visits (p=.33). In **Study 2**, ToM development had a marginal negative effect on change in selectivity (p=.07), such that children who showed more developmental increases in selectivity underwent (marginally) less improvement in ToM across visits. In this regression, there was a significant negative effect of ROI, such that there was more developmental change in selectivity in DMPFC than in RTPJ (p=.048). See Supplementary Table 3 and Supplementary Figure 4.

See Supplementary Information for results from planned supplementary analyses of responses in group ROIs (e.g., regions of interest that are not tailored to functional responses of individual

participants, but can be examined in all participants; Supplementary Table 5 and Supplementary Table 6) and for results of planned exploratory analyses (Supplementary Table 7).

Table 2

2.1 Developmental Change in Response Selectivity	**Study**	**Predictor**	**Beta**	**T-value**	**p-value**
Cross-sectionally: lm(Sel ~ Age + ROI + Motion)	**Study 1**	Age	-0.14	-1.5	0.13
		ROI	0.34	1.8	0.07
		Motion	-0.07	-0.7	0.49
	Study 2	Age	0.27	2.6	**0.01**
		ROI	-0.09	-0.5	0.65
		Motion	0.22	2.0	**0.047**
	Combined	Age	-0.06	-0.8	0.40
		ROI	0.13	0.9	0.37
		Motion	0.02	0.3	0.80
Longitudinally: lme(Sel ~ Age + ROI + Motion + 1|SubID)	**Study 1**	Age	-0.14	-1.5	0.13
		ROI	0.34	1.9	0.06
		Motion	-0.07	-0.7	0.50
	Study 2	Age	0.26	2.6	**0.01**
		ROI	-0.10	-0.6	0.58
		Motion	0.20	1.8	0.08
	Combined	Age	-0.07	-0.9	0.36
		ROI	0.13	1.0	0.33
		Motion	0.01	0.1	0.92
Simultaneous test of within- and between-subject age differences: lme(Sel ~ Age_{Av}+ $Age_{w/i\text{-}sub}$ + ROI + Motion + 1|SubID)	**Study 1**	Age_{Av}	-0.10	-1.0	0.31
		$Age_{w/i\text{-}sub}$	-0.09	-1.0	0.32
		ROI	0.33	1.8	0.07
		Motion	-0.06	-0.6	0.57
	Study 2	Age_{Av}	0.08	0.6	0.56
		$Age_{w/i\text{-}sub}$	0.17	1.9	0.07
		ROI	-0.11	-0.6	0.54
		Motion	0.11	0.8	0.44
	Combined	Age_{Av}	-0.08	-0.8	0.40
		$Age_{w/i\text{-}sub}$	-0.02	-0.4	0.72
		ROI	0.13	1.0	0.33
		Motion	0.02	0.2	0.86
Simultaneous test of within- and between-subject ToM differences: lme(Sel ~ ToM_{Av}+ $ToM_{w/i\text{-}sub}$ + ROI + Motion+ 1|SubID)	**Study 1**	ToM_{Av}	-0.05	-0.5	0.60
		$ToM_{w/i\text{-}sub}$	-0.09	1.0	0.34
		ROI	0.34	1.9	0.07
		Motion	-0.06	0.6	0.59
	Study 2	ToM_{Av}	-0.07	-0.5	0.65
		$ToM_{w/i\text{-}sub}$	0.09	0.9	0.36
		ROI	-0.11	-0.6	0.55
		Motion	0.10	0.6	0.54
	Combined	ToM_{Av}	-0.08	-0.9	0.37
		$ToM_{w/i\text{-}sub}$	0.02	0.2	0.80
		ROI	0.15	1.1	0.27
		Motion	0.003	0.03	0.98
2.2 Stable Neural Individual Differences	**Study**	**Predictor**	**Beta**	**T-value**	**p-value**
lme(Sel_{V2} ~ Sel_{V1} + Age_{Av} + ROI + 1|SubID)	**Study 1**	Sel_{V1}	0.24	1.7	0.10
		Age_{Av}	-0.14	-1.0	0.31
		ROI	0.42	1.5	0.14
		Motion	-0.11	-0.8	0.42
	Study 2	Sel_{V1}	0.09	0.5	0.66
		Age_{Av}	0.02	0.1	0.91
		ROI	-0.35	-1.2	0.24
		Motion	0.15	0.8	0.46
	Combined	Sel_{V1}	0.21	1.9	0.06
		Age_{Av}	-0.17	-1.4	0.18
		ROI	0.10	0.5	0.65
		Motion	-0.04	-0.3	0.76

Table 2. Full Regression Statistics for fMRI Results. Linear regression equations are shown in the left column; statistical results are shown in the right column. Section headers correspond to those used in the results section (sections 2.1, 2.2, and section 4 in Results of main text). Abbreviations: Sel: Selectivity index: (Mental-Social)/(Mental-Physical)*100; ROI: Region of interest (RTPJ or DMPFC); Motion: Number of artifact timepoints; Age: chronological age per participant per visit; Age_{Av}: average age per participant, across the two visits (between-subject age differences); $Age_{w/i\text{-}sub}$: difference between participant's average age and their age at each visit (within-subject change in age); ToM_{Av}: average (matched) ToM score per participant, across the two visits (between-subject differences in ToM); $ToM_{w/i\text{-}sub}$: difference between participant's average ToM and the ToM score at each visit (within-subject change in ToM); V1: Visit 1; V2: Visit 2; 1|SubID: random effect of subject. P-values of significant results ($p<.05$) are in bold.

3. Unplanned Analyses

Unplanned analyses were not included in the pre-registered analysis plan (https://osf.io/jh68b/), but were exploratory analyses conducted based on the results of planned analyses. See Supplementary Table 4 for full regression results from these analyses.

3.1 Stable Individual Differences in RTPJ & Relationship to ToM behavior

Motivated by the visualization of stable individual differences in each ROI (see Figure 1), we conducted unplanned analyses to test for stable neural individual differences in each ROI, separately. This analysis found evidence for a stable neural individual difference in Study 1: Visit 1 selectivity predicted Visit 2 selectivity in RTPJ ($p=.04$, see Supplementary Table 4). This result was not significant in Study 2 ($p=.42$). Still, we tested if selectivity in RTPJ was significantly correlated with, predictive of, or predicted by individual differences in ToM behavior. There was no evidence for a predictive relationship between selectivity in RTPJ and ToM behavior, in either study (all $ps>.05$; see Supplementary Table 4).

3.2 Predictive Relationships between Neural and Behavioral ToM Measures

Given the relative lack of evidence for developmental change in selectivity, and lack of evidence for a relationship between selectivity and ToM behavior, we tested if early ToM score predicted the amount of change in selectivity between visits, and if early selectivity predicted the amount of improvement in ToM between visits. Selectivity at V1 (rather than V1 ToM score) predicted the amount of developmental change in selectivity between visits, such that children who had the least selective responses at Visit 1 showed more developmental change in selectivity (effect of V1 selectivity: $ps<.005$; effect of V1 ToM: $ps>.4$, see Supplementary Table 4). Similarly, ToM score at V1 (rather than V1 selectivity) predicted the amount of improvement in ToM between visits, such that children with the lowest ToM scores Visit 1 underwent more developmental change (effect of V1 ToM: $ps<5x10^{-5}$; effect of V1 selectivity: $ps>.2$, see Supplementary Table 4 and Supplementary Figure 5).

4. Unplanned Analysis of the Combined Dataset

The results presented here ran counter to our predictions that response selectivity would show developmental increases throughout late childhood, and that selectivity would be correlated with ToM behavior (https://osf.io/jh68b/), which were based on previously published results[11,12]. Response selectivity showed developmental increases in Study 2 (ages 5-7 years), but not Study 1 (ages 5-12 years), and selectivity was not related to ToM behavior in either study. While null results are hard to interpret, confidence can be gained by replication (as already presented here), and by ensuring sufficient power to detect relevant effects. In an effort towards the latter, we

repeated the analyses above on the combined dataset (n=58 children, studied longitudinally at two timepoints). Full regression results from the combined dataset are included in Table 2.

4.1 Developmental Change in Response Selectivity
In the combined dataset, there was no evidence for an increase in selectivity with age, cross-sectionally or longitudinally (ps>.3). In a regression simultaneously testing for effects of within- and between-subject differences in age, neither variable had a significant effect on selectivity (ps>.3). There was no evidence for a significant relationship between selectivity and within- or between-subject differences in ToM behavior (ps>.3). See Table 2 and Supplementary Figure 2.

4.2 Stable Neural Individual Differences
In the combined sample, V1 selectivity marginally positively predicted V2 selectivity (p=.06, see Table 2). In unplanned analyses of RTPJ alone, V1 selectivity was a significant positive predictor of V2 selectivity (p=.03, see Supplementary Table 4 and Supplementary Figure 2).

4.3 Predictive Relationships between Behavioral and Neural ToM
In the combined sample, early ToM ability did not predict later selectivity (p=.25), and early selectivity did not predict later ToM ability in RTPJ (p=.54) or in DMPFC (p=.12). Finally, theory of mind development ($V2_{Matched}$ – V1 ToM score) had a marginal negative effect on change in selectivity (V2 – V1 selectivity) (p=.052). See Table 2.

See Supplementary Information for unplanned analyses of group ROIs in the combined dataset (Supplementary Table 5 and Supplementary Table 6), and for planned exploratory analyses (Supplementary Table 7).

Discussion
The current project aimed to measure developmental change and stable individual differences in theory of mind, behaviorally and neurally, within individual children. Specific hypotheses and planned tests were pre-registered prior to data analysis (https://osf.io/jh68b/). Across two longitudinal studies, we found strong evidence for developmental change and stable individual differences in behavioral theory of mind abilities. We found evidence for developmental change in response selectivity of ToM brain regions between ages five and seven, but not after age seven years. We also did not replicate previous observations of a relationship between selectivity and behavioral theory of mind, in either study.

The first goal of the current study was to develop a behavioral measure to capture stable individual differences in theory of mind, longitudinally, in older children. Across two studies, we provide evidence that earlier theory of mind behavioral performance predicts later theory of mind performance, controlling for variables that independently predicted theory of mind score. These data are consistent with previous work suggesting that theory of mind development continues throughout childhood[1-4], and suggest that even later improvements in ToM (e.g., reasoning about moral blameworthiness and non-literal speech) reflect continued development in social reasoning, rather than maturation of systems like language or executive functions alone. This behavioral task can be used to reliably measure theory of mind reasoning in typically developing children as old as ten years of age. This task is publicly available for download (https://osf.io/g5zpv/).

The second goal of this study was to relate behavioral change in ToM to reliable markers of neural development in theory of mind brain regions. Contrary to previously published results, our data suggest that responses in ToM brain regions are fairly selective for processing mental states (as opposed to general social information) in middle childhood. We found evidence for increases in response selectivity in a sample of five to seven year old children (Study 2), but not in a sample that extends this age range to twelve years of age (Study 1). Within individual participants in Study 1, response selectivity did not increase between visits in a majority of children: just as many children had similar or *less* selective responses at Visit 2. Measurements of change within individual children likely reflect real developmental change in addition to regression to the mean; e.g., a closer estimate of average (unchanging) selectivity upon repeated measurement[51]. Consistent with this idea, unplanned analyses in both studies found that children who had more selective responses at Visit 1 showed less developmental change in selectivity between visits. Interestingly, in unplanned analyses that examined selectivity in each region of interest separately, we found evidence for stable individual differences in selectivity of RTPJ in Study 1 and in the combined dataset, but still no evidence for a relationship between RTPJ selectivity and behavioral theory of mind. Given that response selectivity did not show developmental increases in many children, it may be unsurprising that we did not detect significant relationships (cross-sectionally or longitudinally) between this neural measure and behavioral theory of mind, which improved in almost all children.

Of course, we chose to measure response selectivity based on previous studies providing cross-sectional evidence for change with age and theory of mind behavior. How can our results be reconciled with these previous studies? It is unlikely that the results described here are due to insufficient power, given the reasonably large sample sizes, longitudinal design, similar pattern of results in analyses of group regions of interest (which include data from more participants), and similar pattern of results in the analysis of the combined datasets. One possible explanation is that the relatively small number of participants in the previous cross-sectional experiments, in addition to the wide age range studied, placed particular importance on the few young children who participated (Saxe et al. (2009): individually defined RTPJ ROIs defined in 11 of 13 6-11 year old children studied, including 1 child under age 7 years; Gweon et al. (2012): individual defined RTPJ ROIs defined in 17 of 20 5-12 year olds studied, including 3 children under age 7 years). It is possible that, by chance, these studies measured responses in 5-6 year old children who had relatively less selective responses, compared to their age-matched peers (e.g., as studied here, in Study 2). The correlation between selectivity and ToM behavior in these studies may result from underpowered neural evidence from the youngest children, coupled with (appropriately) low performance on the theory of mind task in this age range.

Still, multiple studies conducted using different methodologies converge to suggest that theory of mind behavior is indeed related to development in theory of mind brain regions, especially in young children. First, Sabbagh et al. found that success on explicit false belief tasks in four year old children, controlling for executive function and language differences, was predicted by maturation of the right temporal parietal junction (RTPJ) and dorsomedial prefrontal cortex (DMPFC)[49]. "Maturation" in this study reflects alpha coherence: changes in alpha coherence (as measured by resting state EEG) reflect synchronization of neural firing within and across neural populations, which increases with maturation. The location in cortex of these maturational

changes was inferred using source localization techniques (sLORETA)[52]. Second, using diffusion tensor imaging (DTI), Wiesmann et al. (2017) found that ToM brain regions become increasingly physically connected during childhood, and that the strength of white matter connections around ToM brain regions was significantly positively correlated with children's performance on false-belief tasks, independent of age[53]. Finally, using fMRI to measure functional responses during a short, animated movie, a recent study found that theory of mind brain regions (RTPJ and DMPFC, but also left TPJ, precuneus, and middle and ventro-medial prefrontal cortex) are more functionally correlated with one another in children who pass false-belief tasks, relative to children who fail, controlling for age[54] (see Chapter 2). Together, these three studies suggest that maturational changes in ToM brain regions are specifically related to theory of mind development in children ages three to five years old.

Here, we provide longitudinal fMRI evidence that ToM brain regions continue to develop *after* age five: we find increases in selectivity between ages five to seven years of age (Study 2, and consistent with[11,12,54]). However, we do not find clear evidence for a relationship between continued neural development and theory of mind behavioral change. Future work is necessary to determine the kinds of neural changes that support theory of mind development in later childhood. Response selectivity may simply be too coarse to reflect individual differences in theory of mind behavior, even in samples in which this selectivity shows developmental change with age (e.g. Study 2). A previous study using a large sample of neurotypical adults (n=462) and adults diagnosed with autism (n=31) provides evidence that the magnitude of selective responses does not always reflect real-world differences in social cognitive abilities[50]. Neural measures that capture finer-grained distinctions in representations (e.g., multi-voxel pattern analyses[55], representational similarity analyses[56]) may be promising approaches for studying developmental change in representational content of theory of mind brain regions[57,58]. In a similar vein, while the behavioral ToM task used here captures developmental change with age as well as stable individual differences over time, it is also a coarse tool. In being designed to capture variability across a wide-age range, this task includes a wide variety of ToM concepts. It is therefore less well designed for capturing developmental change concerning specific conceptual milestones, like understanding the distinction between beliefs based on strong vs. weak evidence, or the distinction between causing harm accidentally vs. intentionally (though for some conceptual milestones (e.g. false beliefs), composite measures can be derived[54]).

The current studies shed light on three important components of theory of mind development. First, we provide evidence that later behavioral theory of mind development, even after age 6 years, is uniquely predicted by earlier ToM abilities. Second, our evidence suggests that ToM brain regions respond selectively to mental state content earlier in childhood than previously suggested by cross-sectional studies using smaller sample sizes. Finally, the current results call into question the use of response selectivity as a sensitive measure of individual differences in theory of mind. These results are an important contribution towards understanding the relationship between behavioral and neural theory of mind development, and are highly informative for future studies in this domain.

Methods

The methods used in Study 1 and 2 were nearly identical. Thus, methods are described for both studies simultaneously, and differences between the studies are noted within each section.

1. Participants

Study 1 includes data from 31 typically developing children, ages 5-11 at Visit 1 (V1) (M(SD)=8.1(1.5)), and 7-13 at Visit 2 (V2) (M(SD)=10.4(1.4); 8 females, 1 left-handed). An additional 3 participants completed the study but were excluded for excessive motion during the scan (see Methods). Participants were initially recruited for a cross-sectional study involving a single visit (manuscript in progress). Participants were re-contacted and were recruited for the second visit if they contributed more than two runs of functional MRI data during the first visit, were younger than 13 years old at the time of V2, and if their V2 date could plausibly be scheduled less than four years after the V1 date (n=53). 19 children met these criteria but did not participate in the second visit because they had moved away (n=3), were not interested in participating (n=6), had braces (n=3), or were difficult to contact or schedule (n=7). Data collection occurred between August 2009 and August 2013.

Study 2 includes data from 27 typically developing children, ages 5-6 at V1 (M(SD)=5.5(.26)), and 6-7 at V2 (M(SD)=6.5(.26); 14 females, 4 left-handed, 1 ambidextrous). An additional 16 children were recruited and excluded for not completing the first visit (n=3), not completing the second visit (attrition; n=2), language delays (n=2), or excessive motion during the scan (n=9). Data collection occurred between March 2015 and March 2017. Study 2 participants are on average younger than participants in Study 1, and the range in ages within a single visit was smaller (1 year vs. 6 year range). The duration of time between visits also differs between Study 1 (2 years) and Study 2 (1 year).

In both studies, children were recruited from the local community (Boston, MA, USA), were native speakers of English, had no known neurological or cognitive disabilities, had normal or corrected-to-normal vision, and received an Amazon gift card for participation in addition to small prizes throughout each visit. Participants signed an assent form and parents of participants signed a consent form approved by the Committee on the Use of Humans as Experimental Subjects (COUHES) at MIT. See Supplementary Table 1 for additional information about participants.

2. Data Collection

Prior to each visit, participants received materials preparing them for the study and introducing them to the experimenters, as well as practice earbuds.

2.1 Behavioral Battery

In both studies, children completed a custom-made theory of mind behavioral battery at each visit. The V1 story booklet assessed participants' ability to make predictions and provide explanations about the beliefs, desires, actions, and emotions of various characters. The ToM concepts included in this booklet were largely drawn from work describing the successive ToM achievements in early childhood[34], with the addition of questions involving reasoning about moral blameworthiness. This task was used in a previous cross-sectional study[12]. For longitudinal measurement of ToM reasoning, we developed a second booklet for use at V2. The V2 story booklet asked questions analogous to the V1 story booklet as well as novel questions designed to be more challenging, including questions about second-order false beliefs, the use of sarcasm, and more difficult moral judgments. Analogous questions across the two booklets were similar in type of question as well as syntax of the story and questions, but different in semantic

content (e.g. V1 questions were about helping children find their books, and V2 questions were about helping children find snacks). By including analogous questions as well as novel, challenging questions in the V2 booklet, we could directly quantify the improvement in ToM performance (by comparing V1 performance to performance on the V1 analogues at V2) as well as obtain an overall performance score for each child at each visit (by calculating proportion of questions correct at each visit). ToM booklet stimuli are available via the Open Science Framework (https://osf.io/g5zpv/).

The ToM behavioral battery was coded off-line (e.g., by watching a video recording); the summary score of this measure is calculated as the proportion of questions answered correctly. For V2, participants received two scores: a score that reflected the proportion of all questions answered correctly ("complete score") and a score that reflected the proportion of all V1-analogous questions answered correctly ("matched score"). Improvement on the ToM task across visits was calculated as the V2 "matched score" – V1 score ($V2_M - V1$).

Participants in Study 1 additionally completed measures of verbal (PPVT[36]) and nonverbal (KBIT-II[37]) IQ at V1, and a custom-made computerized flanker task[38] at V2, in order to measure response inhibition (one component of executive functions that is correlated with ToM abilities in childhood[39]). Participants in Study 2 completed the same nonverbal IQ task and an age-appropriate measure of response inhibition (computerized Dimension Change Card Sort task[40] at V1, and the same verbal IQ task at V2). For both studies, age-standardized scores for the IQ tasks were scored and calculated based on the task instructions. For the executive function tasks, we used the difference in accuracy between congruent – incongruent trials for the flanker task, and DCCS summary score[40] as measures of response inhibition.

2.2 FMRI Experiment

During both fMRI scans, children listened to English stories involving characters and their mental states (Mental condition), characters and their physical appearance or social relationships (Social condition), or descriptions of physical objects and events in the world (Physical condition). Each story was read by one of three female speakers in child-directed prosody. V1 stimuli have been used in a previous study[12]. In order to match the stimuli across visits while minimizing practice effects, we varied the names of characters, verbs, and nouns across the two visits. Despite different content, the stories presented during V2 were syntactically identical and semantically very similar to the V1 stories. The stories were matched across condition and across visit for number of words (V1: 52.6, V2: 52.5) and Flesch Reading Ease Level (V1: 85.8, V2: 85.7), and had the same number of sentences (average: 4.7) and same length (20s) (linear regression on number of words; Flesch Reading Ease with Visit (1 or 2; as a factor) and condition (Mental, Social, and Physical) as fixed effects and stimulus (item) number as random effect: number of words: Visit: b=-.17, $t(23)$=-.61, p=.55, Condition: bs<.38, t(21)<.19, ps>.85; Flesch Reading Ease: Visit: b=-.06, $t(23)$=-.05, p=.96, Condition: bs<5.03, t(21)<1.6, ps>.12).

After each story or music clip (20s), children were asked, "Does this come next?" (1.5s). They then heard a clip containing the story or song ending or the ending of an unrelated story or song (3s), followed by an 6.5s pause during which they responded to the prompt by pushing one of two buttons ("Yes" or "No"). This was followed by an encouragement clip: "Way to go!" for correct responses, or "Let's try another!" for incorrect responses (5s). Half of the presented

stories were followed by the correct ending ("Yes" response). In Study 1, incorrect responses were drawn randomly from all other English story conditions (V1) or within condition (V2). In Study 2, half of the incorrect endings came from each of the other two conditions. The story ending was not included in subsequent analyses. All story stimuli are publicly available for download via OSF (https://osf.io/jh68b/).

Stimuli were presented in Matlab 2010a running on an Apple MacBook Pro. Participants heard 24 stories (8 per condition) across four 6.6 (Study 1) or 4.2 minute (Study 2) runs during each visit. In Study 1, participants also heard 8 clips of instrumental music, 8 stories read in a foreign language (V1 only), and 8 stories involving embedded mental states (V2 only); these conditions were excluded from Study 2 and the present analyses. Each run included ten (Study 1) or six (Study 2) 36-second blocks (2 per condition), as well as 12 seconds of rest at the beginning, halfway point, and end. The order of conditions in each run was palindromic (e.g., [rest] A B C D E [rest] E D C B A [rest]) and counterbalanced across runs. In Study 1, stories were counterbalanced across runs and participants. In Study 2, stimulus order was fixed across participants and visits, in order to ensure that differences across participants and visits are not driven by stimulus-order effects. A colorful swirl image was presented visually during the stories, as well as during the rest period. During the prompt, story ending, and response portion of the experiment, an image of a check (left) and an "X" (right) was displayed to encourage participants to answer the question, and remind them which buttons corresponded to "yes" and "no" answers. Children were introduced to the task and completed five practice trials prior to the scan.

Behavioral performance on the task was measured via accuracy (proportion of questions answered correctly) and reaction time (average speed of answering correctly), on trials from included functional runs only (trials from runs that were excluded due to excessive motion were not analyzed). Children performed well on this task during both visits, indicating that they attended to the stimuli (**Study 1**: Accuracy *M*(*SE*): Mental: V1: .85(.03), V2: .93(.02); Social: V1: .89(.03), V2:.87(.03); Physical: V1: .89(.03), V2: .91(.02); **Study 2**: Mental: V1: .75(.05), V2: .84(.03); Social: V1: .69(.06), V2:.83(.02); Physical: V1: .70(.04), V2: .83(.03)). There were no differences in accuracy across visit or condition in **Study 1** (NS effect of Visit: b=.15, t=1.6, p=.11; NS effect of Physical condition: b=.05, t=.47, p=.64; NS effect of Social condition: b=-.06, t=-.51, p=.61 (both relative to Mental condition)). In **Study 2**, children responded more accurately during V2, but there were no differences in accuracy across conditions (effect of Visit: b=.58, t=6.3, p=0; NS effect of Physical condition: b=-.13, t=-1.19, p=.23; NS effect of Social condition: b=-.13, t=-1.16, p=.25 (both relative to Mental condition)). In both studies, children responded faster in all conditions during their second visit (**Study 1**: effect of Visit: b=-1.01, t=-16.4, p=0; NS effect of Physical condition: b=-.04, t=-.50, p=.62; NS effect of Social condition: b=-.10, t=-1.32, p=.19; **Study 2**: effect of Visit: b=.44, t=4.93, p=0, effect of Physical condition: b=.30, t=2.8, p=.006, NS effect of Social condition: b=.02, t=.22, p=.83). As reported, in **Study 2**, there was additionally a significant effect such that children responded more slowly during the Physical condition (see previous regression results).

2.3 FMRI Data Acquisition

Prior to each fMRI scan, children watched a movie of their choice in a mock scanner while practicing lying still on their back and listening to a recording of scanner sounds for 10-15

minutes. If participants moved during the mock scan, their movie paused for three seconds, reminding and training them to stay still.

During the scan, participants were monitored by an experimenter in the control room as well as a second experimenter who stood in the MRI room near the participant's feet. If the participant moved noticeably during the scan, this experimenter would place her hand on the child's leg, as a reminder to stay still.

Whole-brain structural and functional MRI data were acquired on a 3-Tesla Siemens Tim Trio scanner located at the Athinoula A. Martinos Imaging Center at MIT, using one of two custom 32-channel phased-array head coils made for younger (Study 1: n=7, all during V1) or older (Study 1: n=34, 20 from V1) children [41] or the standard Siemens 32-channel head coil (Study 1: n=21, 4 from V1; both visits for all Study 2 participants). T1-weighted structural images were collected in 176 interleaved sagittal slices with 1mm isotropic voxels (GRAPPA parallel imaging, acceleration factor of 3; adult coil: FOV: 256mm; pediatric coils: FOV: 192mm). Functional data were collected with a gradient-echo EPI sequence sensitive to Blood Oxygen Level Dependent (BOLD) contrast in 3mm isotropic voxels in 32 interleaved near-axial slices aligned with the anterior/posterior commissure, and covering the whole brain (EPI factor: 64; TR: 2s, TE: 30ms, flip angle: 90°). Prospective acquisition correction was used to adjust the positions of the gradients based on the participant's head motion one TR back[42]. In both studies, functional data were acquired across four runs (Study 1: 198 volumes per run; Study 2: 126 volumes per run). Four dummy scans were collected at the beginning of each run to allow for steady-state magnetization.

2.4 FMRI Data Analysis

In order to constrain analysis decisions based on our hypotheses prior to analyzing data, all analysis decisions (including preprocessing, region of interest selection and definition, motion exclusion and treatment procedures, calculation of selectivity indices) and planned analyses were published via OSF (https://osf.io/jh68b/)[43,44]. Unplanned analyses are specifically marked as such in the results section.

FMRI data were analyzed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm) and custom software written in Matlab. Functional images were registered to the first image of the first run; that image was registered to each child's anatomical scan from the corresponding visit, and each child's anatomical scan was normalized to a common brain space (Montreal Neurological Institute (MNI) template). All data were smoothed using a Gaussian filter (5mm kernel).

Motion artifact timepoints were identified using the ART toolbox (https://www.nitrc.org/projects/artifact_detect/)[45] as timepoints for which there was 1) more than 2mm of motion in any direction relative to the previous timepoint or 2) a fluctuation in global signal that exceeded a threshold of three standard deviations from the mean global signal. Runs were excluded from analyses if one-third or more of the timepoints collected were identified as motion artifact timepoints, and participants were excluded from all analyses if they had fewer than two runs of usable data (Study 1: n=3; Study 2: n=9). The total number of included timepoints, which is highly correlated with mean translation (pre- and post- artifact removal; Study 1: rs>.51, ps<.0001; Study 2: rs>.54, ps<.0001) did not differ across visits (M(SD) Study

1: visit 1: 91.7(52.2), visit 2: 74.9(44.8), paired t-test: t(30)=1.5, p=.14; Study 2: visit 1: 48.2(26.5), visit 2: 39.5(30.1), paired t-test: t(26)=1.03, p=.31). Number of artifact timepoints (henceforth, "Motion") was not significantly correlated with age or ToM behavior (across both visits) in either study (Study 1: Motion-Age: $r_p(60)$=-.12, p=.34; Motion-ToM: $r_k(60)$=-.02, p=.86; Study 2: Motion -Age: $r_p(52)$=-.12, p=.38; Motion-ToM: $r_k(49)$=-.07, p=.61). In unplanned analyses, we did not find stable individual differences in motion across visits ($Motion_{V1}$- $Motion_{V2}$: Study 1: $r_p(29)$=.20, p=.28; Study 2: $r_p(25)$=-.19, p=.33). See Supplementary Table 1 for amount of motion per participant. The total number of artifact timepoints was included as a covariate in linear regression models in all ROI analyses. Data were high-pass filtered with a cutoff of 128 seconds, in order to remove low-frequency noise, after interpolating over artifact timepoints[46, 47].

We used a general linear model to analyze BOLD activity of each participant as a function of condition. Data were modeled in SPM8 using a standard hemodynamic response function (HRF). Boxcar regressors for each condition and the response period were convolved with the standard HRF, and nuisance covariates were included for run effects, motion artifact timepoints, and signal of no interest (five PCA-based regressors generated from signal extracted from eroded individual white matter masks, e.g. CompCor regressors[48]). SPM's global image scaling was applied to functional images.

Based on previous neuroimaging studies on ToM in adults and children, we conducted Region of Interest (ROI) analyses on two ROIs: the right temporoparietal junction (RTPJ) and dorsal middle prefrontal cortex (DMPFC). Development of these two regions has previously been related to behavioral theory of mind abilities in childhood[12,49]. Individual ROIs were defined as contiguous (minimum k=10) suprathreshold (p<.001) voxels within a 9mm radius sphere of the peak voxel to the Mental > Physical contrast, within previously defined search spaces for each region. Region search spaces were defined based on a random effects analysis using a False Belief > False Photograph contrast in a separate group of 462 typically developing adults[50]. We extracted the mean beta value per condition from these two regions, and calculated selectivity as (Mental – Social) / (Mental – Physical)*100. This calculation has been used in a previous study[12]. Because the difference between Mental and Physical conditions is used to identify ROIs, this measure focuses on the *relative difference* between Mental and Social conditions. In supplementary analyses we additionally measure selectivity in group ROIs, which were 10mm spheres drawn around the peak coordinates of the random effects analysis in a large-scale study of adults[50], excluding voxels that overlapped with language group ROIs not used in the current project. We used these group ROIs for easy comparison of results to other projects (see analysis plan: https://osf.io/jh68b/). Unlike individual ROIs, the voxels analyzed in group ROIs did not necessarily respond more to the Mental condition compared to the Physical condition (voxels in group ROIs are not selected based on their functional response profile; see Supplementary Information for more discussion of group ROIs). Thus, we calculated selectivity from extracted beta values as the (Mental – Social)* 100. See Supplementary Table 2 for additional information about individual and group regions of interest.

Based on previous analyses, we expected the selectivity measure to be between -50 and 200 in individual ROIs. As stated in our analysis plan (https://osf.io/jh68b/), we planned to exclude participants whose selectivity values fell outside of this range. However, selectivity values for all

participants fell within our expected range, so zero participants were excluded from individual ROI analyses based on this criterion.

3. Linear Mixed Effect Regressions

The longitudinal design employed here provided the opportunity to obtain sensitive measurements of development within individual participants. We used the nlme package and lme function in R (https://www.r-project.org/) to conduct linear mixed effect regressions in order to test for developmental change in ToM, behaviorally and neurally. Effects of age on ToM behavioral performance and selectivity were assessed via three variables: (1) "age," which is the chronological age of each participant, per visit; (2) "between-subject age difference," which is the average age of each participant, across visits (e.g., to test for effects of differences in age across participants), and (3) "within-subject age difference," which is the difference between a participant's age at a given visit, and their average age across both visits (e.g., to test for effects of an individual's change in age). We used these variables in order to study effects of age cross-sectionally (across participants) as well as longitudinally (within-participant). We used a similar approach to test for effects of ToM. All regressions on response selectivity included data from both ROIs (RTPJ and DMPFC), and tested for a significant effect of ROI. These regressions also included the number of artifact timepoints as a between-subject predictor ("Motion"). Regressions that included multiple data points per individual (e.g., for two visits, or two ROIs) included a subject identifier as a random effect in order to account for non-independence. The longitudinal design additionally enabled testing for stable individual differences in and predictive relationships between behavioral and neural measures of ToM. We used linear regressions to test if behavioral ToM ability 1) predicted later measures of selectivity in ToM brain regions, 2) was predicted by earlier measures of selectivity in ToM brain regions, and 3) improved with increases in selectivity across the two visits. Regression equations are displayed with statistical results in Table 1, Table 2, and in Supplementary Tables 3-7.

4. Pilot Experiment

To ensure that any neural differences between the visits were not introduced by differences in fMRI task stimuli, we collected pilot fMRI data from ten children while they listened to the stimuli from V1 and V2 in interleaved runs, during a single visit. Two children were dropped from analyses due to failure to complete the scan (n=1) and excessive motion (n=1), for a final pilot sample of eight children (M(SD) age: 10.5(1.3) years; 4 females; 1 LH, 1 Ambidextrous). Participant recruitment, fMRI data acquisition, and fMRI data analysis procedures were identical to the procedures described above. All pilot participants were scanned with the larger pediatric head coil. Participant motion did not differ across stimulus sets (M(SD) number of artifact timepoints: V1 stimuli: 56.3 (29.4); V2 stimuli: 61.4 (35.8); NS effect of stimuli: b=-.16, t=-1.3, p=.2). There were no differences in neural measures of interest across the two stimulus sets (**Selectivity:** individual ROIs: NS effect of stimuli: b=-.18, t=-.35, p=.7, NS effect of ROI: b=.67, t=1.3, p=.23, NS effect of motion: b=-.12, t=-.44, p=.68; group ROIs: NS effect of stimuli: b=-.12, t=-.44, p=.66, NS effect of ROI: b=.08, t=.27, p=.79, NS effect of motion: b=.48, t=2.17, p=.07; **Mental – Physical beta difference**: individual ROIs: NS effect of stimuli: b=-.70, t=1-.99, p=.09, NS effect of ROI: b=-.67, t=-1.9, p=.10, effect of motion: b=.54, t=3.0, p=.03; group ROIs: NS effect of stimuli: b=-.26, t=-.82, p=.42, NS effect of ROI: b=-.10, t=-.34, p=.74, effect of motion: b=.52, t=3.29, p=.02).

Acknowledgements
We thank the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT; Lindsey Powell, Todd Thompson, Julia Leonard, and Dorit Kliemann for feedback on the analysis plan; Grace Lisandrelli, Caitlin Malloy, Ellen Olson-Brown, and Julianne Herts for recruiting and scheduling participants, and Mika Asaba, Alexa Riobueno-Naylor, Hannah Pelton for help with data collection and behavioral data analysis, David Dodell-Feder for help with data collection, and Nick Dufour for help with fMRI analysis scripts. We also gratefully acknowledge support of this project by a NSF Graduate Research Fellowship (#1122374 to HR), a Whitaker Health Sciences Fund Fellowship (to HR), an NSF CAREER award (#095518 to RS), and support from the David and Lucile Packard Foundation (#2008-333024 to RS) and the Ellison Medical Foundation.

References

1. Peterson, C. C., Wellman, H. M. & Slaughter, V. The mind behind the message: Advancing theory-of-mind scales for typically developing children, and those with deafness, autism, or Asperger syndrome. *Child Dev* **83,** 469–485 (2012).
2. Baird, J. A. & Astington, J. W. The role of mental state understanding in the development of moral cognition and moral action. *New Directions for Child and Adolescent Development* **2004,** 37–49 (2004).
3. Cushman, F., Sheketoff, R., Wharton, S. & Carey, S. The development of intent-based moral judgment. *COGNITION* **127,** 6–21 (2013).
4. Harris, P. L. *Children and emotion: The development of psychological understanding.* (Basil Blackwell, 1989).
5. Adolphs, R. The Social Brain: Neural Basis of Social Knowledge. *Annu. Rev. Psychol.* **60,** 693–716 (2009).
6. Carrington, S. J. & Bailey, A. J. Are there theory of mind regions in the brain? A review of the neuroimaging literature. *Hum. Brain Mapp.* **30,** 2313–2335 (2009).
7. Frith, C. D. & Frith, U. Mechanisms of Social Cognition. *Annu. Rev. Psychol.* **63,** 287–313 (2012).
8. Bowman, L. C. & Wellman, H. M. Neuroscience contributions to childhood theory-of-mind development. *Contemporary perspectives on research in theories of mind in early childhood education* 195–224 (2014).
9. Kobayashi, C., Glover, G. H. & Temple, E. Children's and adults' neural bases of verbal and nonverbal "theory of mind". *Neuropsychologia* **45,** 1522–1532 (2007).
10. Moriguchi, Y., Ohnishi, T., Mori, T., Matsuda, H. & Komaki, G. Changes of brain activity in the neural substrates for theory of mind during childhood and adolescence. *Psychiatry Clin. Neurosci.* **61,** 355–363 (2007).
11. Saxe, R. R., Whitfield-Gabrieli, S., Scholz, J. & Pelphrey, K. A. Brain regions for perceiving and reasoning about other people in school-aged children. *Child Dev* **80,** 1197–1209 (2009).
12. Gweon, H., Dodell-Feder, D., Bedny, M. & Saxe, R. Theory of Mind Performance in Children Correlates With Functional Specialization of a Brain Region for Thinking About Thoughts. *Child Dev* **83,** 1853–1868 (2012).
13. Wellman, H. M., Cross, D. & Watson, J. Meta-analysis of theory-of-mind development: the truth about false belief. *Child Dev* **72,** 655–684 (2001).
14. Knudsen, B. & Liszkowski, U. Eighteen-and 24-month-old infants correct others in anticipation of action mistakes. *Dev Sci* **15,** 113–122 (2012).
15. Southgate, V., Senju, A. & Csibra, G. Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science* **18,** 587–592 (2007).
16. Powell, L. J., Hobbs, K., Bardis, A., Carey, S. & Saxe, R. Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development* (2017).
17. Filippova, E. & Astington, J. W. Further development in social reasoning revealed in discourse irony understanding. *Child Dev* **79,** 126–138 (2008).
18. Cantlon, J. F., Pinel, P., Dehaene, S. & Pelphrey, K. A. Cortical Representations of Symbols, Objects, and Faces Are Pruned Back during Early Childhood. *Cerebral Cortex* **21,** 191–199 (2010).
19. Dehaene, S. *et al.* How learning to read changes the cortical networks for vision and language. *Science* **330,** 1359–1364 (2010).

20. Peelen, M. V., Glaser, B., Vuilleumier, P. & Eliez, S. Differential development of selectivity for faces and bodies in the fusiform gyrus. *Dev Sci* **12,** F16–F25 (2009).
21. Golarai, G. *et al.* Differential development of high-level visual cortex correlates with category-specific recognition memory. *Nature Publishing Group* **10,** 512 (2007).
22. Carter, E. J. & Pelphrey, K. A. School-aged children exhibit domain-specific responses to biological motion. *Social Neuroscience* **1,** 396–411 (2006).
23. Shaw, P. *et al.* Neurodevelopmental Trajectories of the Human Cerebral Cortex. *Journal of Neuroscience* **28,** 3586–3594 (2008).
24. Giedd, J. N. *et al.* Brain development during childhood and adolescence: a longitudinal MRI study. *Nat Neurosci* **2,** 861–863 (1999).
25. Gogtay, N. *et al.* Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences* **101,** 8174–8179 (2004).
26. Overgaauw, S., van Duijvenvoorde, A. C., Moor, B. G. & Crone, E. A. A longitudinal analysis of neural regions involved in reading the mind in the eyes. *Social Cognitive and Affective Neuroscience* **10,** 619–627 (2015).
27. Pfeifer, J. H. *et al.* Entering adolescence: resistance to peer influence, risky behavior, and neural changes in emotion reactivity. *Neuron* **69,** 1029–1036 (2011).
28. Pfeifer, J. H. *et al.* Longitudinal change in the neural bases of adolescent social self-evaluations: effects of age and pubertal development. *J. Neurosci.* **33,** 7415–7419 (2013).
29. Watson, A. C., Nixon, C. L., Wilson, A. & Capage, L. Social interaction skills and theory of mind in young children. *Developmental Psychology* **35,** 386 (1999).
30. Astington, J. W. & Jenkins, J. M. Theory of mind development and social understanding. *Cognition & Emotion* **9,** 151–165 (1995).
31. Wellman, H. M., Lopez-Duran, S., LaBounty, J. & Hamilton, B. Infant attention to intentional action predicts preschool theory of mind. *Developmental Psychology* **44,** 618–623 (2008).
32. Yamaguchi, M., Kuhlmeier, V. A., Wynn, K. & vanMarle, K. Continuity in social cognition from infancy to childhood. *Dev Sci* **12,** 746–752 (2009).
33. Wellman, H. M., Fang, F. & Peterson, C. C. Sequential progressions in a theory-of-mind scale: longitudinal perspectives. *Child Dev* **82,** 780–792 (2011).
34. Wellman, H. M. & Liu, D. Scaling of theory-of-mind tasks. *Child Dev* **75,** 523–541 (2004).
35. Nadig, A. S. & Sedivy, J. C. Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science* **13,** 329–336 (2002).
36. Dunn, L. M., Dunn, L. M., Bulheller, S. & Häcker, H. *Peabody picture vocabulary test.* (American Guidance Service Circle Pines, MN, 1965).
37. Kaufman, A. S. KBIT-2: Kaufman Brief Intelligence Test. Minneapolis, MN: NCS Pearson. (1997).
38. Fan, J., McCandliss, B. D., Sommer, T., Raz, A. & Posner, M. I. Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience* **14,** 340–347 (2002).
39. Carlson, S. M. & Moses, L. J. Individual differences in inhibitory control and children's theory of mind. *Child Dev* **72,** 1032–1053 (2001).
40. Zelazo, P. D. The Dimensional Change Card Sort (DCCS): a method of assessing executive function in children. *Nat Protoc* **1,** 297–301 (2006).

41. Keil, B. *et al.* Size-optimized 32-channel brain arrays for 3 T pediatric imaging. *Magn. Reson. Med.* **66,** 1777–1787 (2011).
42. Thesen, S., Heid, O., Mueller, E. & Schad, L. R. Prospective acquisition correction for head motion with image-based tracking for real-time fMRI. *Magn. Reson. Med.* **44,** 457–465 (2000).
43. Asendorpf, J. B. *et al.* Recommendations for increasing replicability in psychology. *European Journal of Personality* **27,** 108–119 (2013).
44. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nat. hum. behav.* **1,** 0021 (2017).
45. Whitfield-Gabrieli, S., Nieto-Castanon, A. & Ghosh, S. Artifact Detection Tools (ART). *Cambridge, MA. Release version* **7,** 11 (2011).
46. Carp, J. Optimizing the order of operations for movement scrubbing: Comment on Power et al. *NeuroImage* **76,** 436–438 (2013).
47. Hallquist, M. N., Hwang, K. & LUNA, B. The nuisance of nuisance regression: spectral misspecification in a common approach to resting-state fMRI preprocessing reintroduces noise and obscures functional connectivity. *NeuroImage* **82,** 208–225 (2013).
48. Behzadi, Y., Restom, K., Liau, J. & Liu, T. T. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* **37,** 90–101 (2007).
49. Sabbagh, M. A., Bowman, L. C., Evraire, L. E. & Ito, J. M. B. Neurodevelopmental correlates of theory of mind in preschool children. *Child Dev* **80,** 1147–1162 (2009).
50. Dufour, N. *et al.* Similar Brain Activation during False Belief Tasks in a Large Sample of Adults with and without Autism. *PLoS ONE* **8,** e75468 (2013).
51. Galton, F. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland* **15,** 246–263 (1886).
52. Pascual-Marqui, R. D. Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. *Methods Find Exp Clin Pharmacol* **24,** 5–12 (2002).
53. Wiesmann, C. G., Schreiber, J., Singer, T., Steinbeis, N. & Friederici, A. D. White matter maturation is associated with the emergence of Theory of Mind in early childhood. *Nature Communications* **8,** 14692 (2017).
54. Richardson, H., Lisandrelli, G., Riobueno-Naylor, A. & Saxe, R. Development of the social brain from age three to twelve years. *Nature Communications* **9,** 1027 (2018).
55. Norman, K. A., Polyn, S. M., Detre, G. J. & Haxby, J. V. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences* **10,** 424–430 (2006).
56. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2,** 4 (2008).
57. Coutanche, M. N., Thompson-Schill, S. L. & Schultz, R. T. Multi-voxel pattern analysis of fMRI data predicts clinical symptom severity. *NeuroImage* **57,** 113–123 (2011).
58. Koster-Hale, J., Saxe, R., Dungan, J. & Young, L. L. Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences* **110,** 5648–5653 (2013).

Supplementary Materials

1. Group ROI Analyses

We conducted the primary fMRI analyses in group regions of interest, as supplementary to the main analyses conducted in individually defined ROIs. Group ROIs are a noisier measurement in each individual subject, because they are not tailored to each individual's functional response profile[1,2]. However, group ROIs enable studying responses in all individuals at both timepoints, as opposed to only individuals in whom individual functional ROIs are successfully defined. This has two implications which could be important: first, group ROI analyses include more participants: in the combined dataset, 47/58 participants have individual RTPJ ROIs at both visits, compared to 56/58 participants with longitudinal data from group ROIs, and 42/58 participants have individual DMPFC ROIs at both visits, compared to 55/58 participants with group ROIs. Second, group ROI analyses will necessarily include participants whose responses were not sufficiently selective for inclusion in individual ROI analyses. Because group ROI analyses include more participants and participants who have less selective responses, these analyses could be more sensitive to neural developmental change with age. Additionally, unlike the individual ROIs defined here, group ROIs enable independently estimating responses to the Mental and Physical conditions, because responses to these conditions are not used for ROI definition.

Group ROIs were 10mm sphere ROIs drawn around peak coordinates to a False Belief > False Photograph contrast in a group of 462 neurotypical adults[3]. Selectivity in group ROIs was calculated as the difference in beta values to Mental – Social conditions * 100. Based on previous analyses, we expected the selectivity measure to be between -50 and 100 in group ROIs. Selectivity values that fell outside of this range were excluded from all analyses (Study 1: 9 participants excluded from group ROI analyses; Study 2: 3 participants excluded from group ROI analyses). This exclusion criterion was pre-specified in our analysis plan (https://osf.io/jh68b/).

See Supplementary Tables 5 and 6 for full regression results from group ROI analyses.

1.1 fMRI Results: Developmental Change in Selectivity of Group ROIs

We tested for significant increases in selectivity with age, and for a significant positive correlation between selectivity and ToM behavior. In **Study 1** (ages 5-12), selectivity did not increase with age, cross-sectionally or longitudinally (ps>.3; Supplementary Table 5 and Supplementary Figure 3). In a regression simultaneously testing for effects of within- and between-subject differences in age, neither variable had a significant effect on selectivity (ps>.10). Descriptively, selectivity was similar across visits in both ROIs (M(SE) RTPJ: V1: 34.9(7.1), V2: 26(5.04); DMPFC: V1: 32.9(6.4), V2: 25.3(6.7)). In **Study 1**, within-subject change in ToM behavior had a negative effect on selectivity, such that individuals who underwent less change in ToM over time (e.g., because they performed well at Visit 1), had more selective responses (negative effect of within-individual ToM variable: p=.01).

In **Study 2** (ages 5-7), selectivity increased with age cross-sectionally and longitudinally (ps=.03; Supplementary Table 5). In a regression simultaneously testing for effects of within- and between-subject differences in age, only within-subject change in age had a marginal positive effect on selectivity (NS effect of between-subject age variable: p=.30; marginal effect of within-

individual age variable: p=.08). Descriptively, selectivity increased across visits in both regions (RTPJ: V1: 27.9(4.0), V2: 24.9(6.0); DMPFC: V1: 14.1(4.4), V2: 32.7(4.1)). There was no evidence for a significant relationship between selectivity and within- or between-subject differences in ToM behavior (ps>.2).

1.2 fMRI Results: Stable Individual Differences in Group ROIs

Visit 1 selectivity did not predict Visit 2 selectivity in either study (ps>.2; see Supplementary Table 5).

1.3 Predictive Relationships between behavioral and neural ToM in Group ROIs

We tested if behavioral ToM ability 1) predicts later measures of selectivity in ToM brain regions, 2) is predicted by earlier measures of selectivity in ToM brain regions, and 3) improves with increases in selectivity across the two visits. Early ToM ability did not predict later selectivity (ps>.05). Early selectivity did not predict later ToM ability in RTPJ or in DMPFC (ps>.1). In **Study 1,** ToM development had a significant negative effect on amount of change in selectivity between visits (p=.03). In **Study 2**, there was no effect of ToM development on amount of change in selectivity (p=.73). See Supplementary Table 6.

1.4 Unplanned Combined Analyses in Group ROIs

1.4.1 Developmental Change in selectivity of Group ROIs (Combined Analysis)

In the combined dataset, we did not find evidence for an increase in selectivity with age, cross-sectionally or longitudinally (ps>.6, see Supplementary Table 5). In a regression simultaneously testing for effects of within- and between-subject differences in age, neither variable had a significant effect on selectivity (ps>.2). There was no evidence for a significant relationship between selectivity and within- or between-subject differences in ToM behavior (ps>.4).

1.4.2 Stable Neural Individual Differences in Group ROIs (Combined Analysis)

In the combined sample, Visit 1 selectivity did not predict Visit 2 selectivity (p=.98, see Supplementary Table 5).

1.4.3 Predictive Relationships between behavioral and neural ToM in group ROIs (Combined Analysis)

In the combined sample, early ToM ability did not predict later selectivity (p=.19), and early selectivity in RTPJ did not predict later ToM ability (p=.21). However, early selectivity in DMPFC significantly predicted later behavioral ToM score (p=.02). Finally, theory of mind development ($V2_{Matched}$ – V1 ToM score) had a very marginal negative effect on amount of change in selectivity between visits (p=.098). See Supplementary Table 6.

1.5 Brief Summary of Group ROI Results

The results from group ROIs largely corresponded with the results from individual functional regions of interest: there was evidence for developmental increases in **Study 2**, but not **Study 1**, suggesting that there is little developmental change in response selectivity after approximately age seven years. There was no evidence for stable neural individual differences: selectivity at Visit 1 did not predict selectivity at Visit 2 in either study, or in the combined dataset. There was also little evidence to suggest a relationship between response selectivity and theory of mind behavior: there was no evidence for such a relationship in either study alone. When combined,

there was a significant positive relationship such that early selectivity in DMPFC predicted later behavioral ToM. Though this result is compatible with previous work[4], it should be interpreted with caution: it was the result of an unplanned analysis of the combined dataset, in group regions of interest. Future work is necessary to determine the robustness of the predictive relationship between selectivity in DMPFC and ToM.

2. Planned Exploratory Analyses in Individual ROIs

Given hints in Study 1 that RTPJ is marginally more selective than DMPFC, and that RTPJ undergoes marginally less change in selectivity across visits, we tested if early RTPJ selectivity predicts later DMPFC selectivity. This planned exploratory analysis was initially motivated by evidence from anatomical studies that suggest that parietal cortex undergoes cortical thinning earlier in developmental than prefrontal cortices[5,6], and by previous developmental studies that suggest that RTPJ shows functionally selective responses earlier than DMPFC (in Gweon et al., 17/20 children have functionally selective RTPJ ROIs, compared to 10/20 DMPFC ROIs). However, there was no evidence that early selectivity in RTPJ predicted later selectivity in DMPFC in either longitudinal study (ps>.2, see Supplementary Table 7).

3. Whole-Brain Random Effects Analysis

Whole-brain analyses were used to examine the main contrast of interest (Mental > Physical) within each visit as well as between visits. These analyses were corrected for multiple comparisons by estimating the false-positive rate via 5,000 Monte Carlo permutations using the SnPM toolbox for SPM5 (http://www.fil.ion.ucl.ac.uk/spm/software/spm5/; p<.05). To view the difference in response to this contrast across visits, we ran a corrected random effects analysis on the within-subject contrast difference (V2 Mental > Physical – V1 Mental > Physical). See Supplementary Figure 6.

Supplementary Table 1

SubID	V1 Age	V2 Age	V1 ToM	V2 ToM (M)	V2 ToM (C)	Handedness	Gender	NVIQ (KBIT)	VIQ (PPVT)	Resp. Inhibition	V1 Motion	V2 Motion
Study1_S01	8.68	11.00	1.000	1.000	0.963	R	M	101	130	0.000	60	117
Study1_S02	8.93	11.15	0.872	1.000	0.981	R	M	126	118	-0.050	102	81
Study1_S03	7.01	9.23	0.872	1.000	0.981	R	M	132	92	-0.147	107	93
Study1_S04	10.75	12.97	0.923	1.000	1.000	R	M	126	98	0.000	130	77
Study1_S05	8.92	11.07	0.897	1.000	0.962	R	M	110	124	0.000	94	130
Study1_S06	8.49	10.73	0.947	1.000	1.000	R	M	127	148	-0.036	40	16
Study1_S07	9.76	11.74	0.947	1.000	0.944	R	M	128	123	-0.033	35	81
Study1_S08	8.41	10.38	0.846	1.000	0.926	R	M	114	120	-0.067	76	19
Study1_S09	9.15	11.43	0.872	0.958	0.833	R	M	105	151	0.033	101	43
Study1_S10	5.26	7.55	0.816	0.957	0.941	R	M	107	131	-0.067	144	119
Study1_S11	8.41	11.73	0.914	1.000	1.000	R	F	124	123	0.033	21	8
Study1_S12	5.38	8.92	0.743	0.913	0.923	R	F	103	119	0.000	18	39
Study1_S13	7.53	9.29	0.895	1.000	0.944	R	M	110	143	-0.033	132	41
Study1_S14	7.38	9.12	0.811	0.957	0.878	R	F	118	123	-0.100	15	97
Study1_S15	7.02	9.99	0.921	1.000	1.000	R	F	109	121	0.000	5	27
Study1_S16	8.68	11.14	0.974	1.000	0.943	R	M	96	126	-0.069	163	83
Study1_S17	7.93	9.87	0.676	1.000	0.885	R	M	125	124	-0.066	123	122
Study1_S18	9.83	11.78	0.846	1.000	0.944	R	M	128	142	0.034	104	49
Study1_S19	7.51	9.46	0.949	1.000	0.963	R	F	130	140	-0.001	65	53
Study1_S20	6.12	9.58	0.757	0.917	0.906	R	M	112	138	0.000	38	40
Study1_S21	9.42	11.24	0.872	1.000	1.000	R	M	109	80	0.033	125	29
Study1_S22	9.20	12.51	0.941	1.000	0.942	R	F	132	133	0.000	136	12
Study1_S23	7.61	9.43	0.838	1.000	1.000	R	M	122	121	-0.077	100	119
Study1_S24	10.12	12.41	0.921	1.000	0.943	R	M	124	146	-0.036	122	95
Study1_S25	6.23	8.51	0.853	1.000	0.923	R	F	99	137	0.000	169	40
Study1_S26	5.84	9.21	0.641	1.000	0.887	R	F	115	121	-0.133	19	176
Study1_S27	9.97	11.83	0.974	1.000	0.981	R	M	119	125	-0.033	141	69
Study1_S28	7.73	9.59	0.897	1.000	0.962	R	M	115	109	-0.036	116	121
Study1_S29	8.59	11.43	0.949	1.000	0.963	R	M	119	123	-0.033	46	48
Study1_S30	8.09	9.86	0.889	1.000	1.000	L	M	105	120	0.000	211	145
Study1_S31	6.04	7.80	0.842	0.917	0.925	R	M	143	123	-0.083	84	134
Study2_S01	5.43	6.43	0.667	NA	NA	R	F	112	113	2	53	6
Study2_S02	5.81	6.8	0.769	NA	NA	R	M	92	118	2	63	36
Study2_S03	5.99	7.04	0.923	1.000	0.982	R	M	99	120	3	55	31
Study2_S04	5.39	6.38	0.795	0.958	0.855	R	F	100	124	3	55	66
Study2_S05	5.52	6.55	0.897	0.875	0.818	R	M	106	119	2	94	52
Study2_S06	5.55	6.53	0.795	1.000	0.945	R	F	109	129	2	46	9
Study2_S07	5.49	6.52	0.821	1.000	0.909	R	M	121	134	2	20	68
Study2_S08	5.79	6.81	0.923	0.958	0.945	Ambi (V1); L (V2)	M	141	130	3	0	15
Study2_S09	5.76	6.76	0.744	0.958	0.927	R	M	92	123	2	55	1
Study2_S10	5.97	6.95	0.641	0.917	0.909	R	F	108	124	3	56	27
Study2_S11	5.47	6.47	0.821	1.000	0.927	R	F	92	109	2	53	8
Study2_S12	5.99	7.03	0.872	1.000	0.927	R	M	120	122	2	91	18
Study2_S13	5.23	6.24	0.615	0.875	0.782	R	M	107	134	3	20	75
Study2_S14	5.38	6.37	0.846	NA	NA	R	M	109	128	2	74	64
Study2_S15	5.46	6.46	0.769	1.000	0.945	R	F	130	121	2	40	12
Study2_S16	5.46	6.46	0.821	0.875	0.818	R	M	114	121	2	49	6
Study2_S17	5.55	6.54	0.692	0.792	0.815	R	F	85	114	2	14	95
Study2_S18	5.82	6.84	0.821	0.958	0.873	L	M	108	135	3	40	41
Study2_S19	5.26	6.33	0.590	0.917	0.855	R	F	113	122	2	8	101
Study2_S20	5.66	6.65	0.821	0.917	0.855	R	F	112	117	2	62	69
Study2_S21	5.13	6.14	0.487	0.958	0.891	R	F	97	119	2	24	11
Study2_S22	5.51	6.53	0.769	0.958	0.873	L	F	106	129	3	72	26
Study2_S23	5.12	6.12	0.897	1.000	0.855	R	F	100	119	3	9	35
Study2_S24	5.56	6.56	0.718	0.917	0.873	Ambi	F	114	109	2	71	94
Study2_S25	5.33	6.33	0.333	0.667	0.709	L	M	104	123	2	18	46
Study2_S26	5.44	6.44	0.538	0.958	0.800	R	M	127	119	3	88	16
Study2_S27	5.02	6.05	0.897	1.000	0.891	R	F	114	120	2	71	38
Study 1	**8.06 (1.45)**	**10.39 (1.39)**	**.87 (.08)**	**.99 (.03)**	**.95 (.04)**	**1 L**	**8 F**	**117.19 (11.35)**	**124.90 (15.52)**	**-.03 (.05)**	**91.68 (52.20)**	**74.94 (44.82)**
Study 2	**5.52 (.26)**	**6.53 (.26)**	**.75 (.14)**	**.94 (.08)**	**.87 (.06)**	**4 L; 1 Ambi (V2)**	**14 F**	**108.59 (12.49)**	**122.04 (6.89)**	**2.33 (.48)**	**48.19 (26.54)**	**39.48 (30.10)**

Supplementary Table 1. Participant Demographics. V1: Visit 1; V2: Visit 2; ToM: performance on ToM behavioral battery (proportion of questions answered correct); Response inhibition is the difference in accuracy between congruent – incongruent trials for the flanker task (Study 1), and the Dimensional Change Card Sort summary score (Study 2); Motion is number of artifact timepoints. Bottom two rows show M(SD) and summaries per measure per study.

Supplementary Table 2

STUDY 1	# Identified		Peak Coordinate		N voxels M(SD)		Peak T-Value M(SD)	
Regions of Interest	**Visit 1**	**Visit 2**	**Visit 1**	**Visit 2**	**Visit 1**	**Visit 2**	**Visit 1**	**Visit 2**
RTPJ	29/31	29/31	[54,-49,18]	[55,-51,20]	140 (85)	169 (97)	5.7 (1.3)	5.8 (1.3)
DMPFC	26/31	29/31	[-4,54,30]	[0,52,30]	65 (42)	92 (68)	4.9 (.69)	5.4 (1.2)
Other ToM Regions								
LTPJ	28/31	29/31	[-51,-56,25]	[-51,-56,22]	142 (101)	158 (71)	5.6 (1.1)	5.8 (1.4)
PC	27/31	26/31	[-1,-52,35]	[0,-52,33]	133 (94)	166 (85)	5.4 (1.0)	5.8 (1.2)
MMPFC	28/31	26/31	[2,56,13]	[4,56,14]	89 (69)	104 (70)	5.1 (.9)	5.4 (1.1)
VMPFC	18/31	22/31	[-3,52,-10]	[2,53,-15]	66 (39)	71 (53)	5.0 (1.1)	5.1 (1.1)
RSTS	29/31	30/31	[52,-8,20]	[51,-6,20]	110 (76)	122 (78)	5.6 (1.1)	5.9 (.9)
STUDY 2	**# Identified**		**Peak Coordinate**		**N voxels M(SD)**		**Peak T-Value M(SD)**	
Regions of Interest	**Visit 1**	**Visit 2**	**Visit 1**	**Visit 2**	**Visit 1**	**Visit 2**	**Visit 1**	**Visit 2**
RTPJ	21/27	25/27	[56,-50,20]	[54,-51,20]	111 (92)	132 (66)	5.3 (1.3)	6.0 (1.0)
DMPFC	19/27	26/27	[-1,53,28]	[-2,55,28]	84 (49)	97 (68)	5.2 (.8)	5.7 (1.3)
Other ToM Regions								
LTPJ	23/27	26/27	[-53,-54,21]	[-53,-57,21]	138 (107)	169 (87)	5.7 (1.5)	6.2 (1.5)
PC	24/27	24/27	[-1,-55,33]	[-2,-52,34]	126 (104)	176 (79)	5.6 (1.7)	6.5 (1.3)
MMPFC	20/27	24/27	[3,55,13]	[1,58,16]	104 (79)	113 (76)	5.3 (.9)	5.9 (1.1)
VMPFC	14/27	19/27	[2,51,-13]	[-1,53,-13]	64 (59)	114 (85)	4.8 (.9)	5.7 (1.3)
RSTS	26/27	26/27	[58,-16,-14]	[57,-13,-16]	87 (71)	138 (85)	5.0 (.9)	5.7 (1.3)
Group ROIs	**# Included**		**Peak Coordinate**		**N Voxels**			
STUDY 1	**Visit 1**	**Visit 2**						
RTPJ	28/31	31/31	[54,-52,23]		463			
DMPFC	27/31	29/31	[-1,53,29]		455			
STUDY 2	**Visit 1**	**Visit 2**						
RTPJ	27/27	27/27	[54,-52,23]		463			
DMPFC	25/27	26/27	[-1,53,29]		455			

Supplementary Table 2. Regions of Interest # Identified is number of participants in whom an ROI was successfully identified at p<.001, k=10 thresholds, to the Mental > Physical contrast. Peak coordinates are in mm space. For group ROIs, # Included is the number of participants whose selectivity values fell within the pre-defined range of reasonable values.

Supplementary Table 3

Predictive Relationships (Section 2.3 in Results of Main Text)	Study	Predictor	Beta	T-value	p-value
Does behavioral ToM at V1 predict selectivity at V2?	**Study 1**	ToM_{V1}	0.14	0.9	0.38
$lme(Sel_{V2} \sim ToM_{V1} + Sel_{V1} + Age_{Av} + ROI + Motion + 1\|SubID)$		Sel_{V1}	0.24	1.8	0.09
		Age_{Av}	-0.22	-1.4	0.19
		ROI	0.43	1.6	0.13
		Motion	-0.10	-0.7	0.48
	Study 2	ToM_{V1}	0.15	0.8	0.47
		Sel_{V1}	0.12	0.6	0.55
		Age_{Av}	-0.04	-0.2	0.86
		ROI	-0.38	-1.3	0.22
		Motion	0.14	0.7	0.49
	Combined	ToM_{V1}	0.15	1.2	0.25
		Sel_{V1}	0.22	2.0	0.05
		Age_{Av}	-0.26	-1.8	0.08
		ROI	0.07	0.4	0.72
		Motion	-0.03	-0.2	0.81
Is behavioral ToM at V2 predicted by RTPJ selectivity at V1?	**Study 1**	RTPJ Sel_{V1}	0.06	0.3	0.73
$lm(ToM_{V2C} \sim RTPJ\ Sel_{V1} + ToM_{V1} + Age_{Av} + Motion)$		ToM_{V1}	0.58	2.8	**0.009**
		Age_{Av}	-0.03	-0.1	0.91
		Motion	0.05	0.3	0.76
	Study 2	RTPJ Sel_{V1}	0.09	0.5	0.64
		ToM_{V1}	0.33	1.5	0.17
		Age_{Av}	0.30	1.6	0.13
		Motion	-0.30	-1.8	0.10
	Combined	RTPJ Sel_{V1}	0.06	0.6	0.54
		ToM_{V1}	0.46	3.3	**0.002**
		Age_{Av}	0.29	2.1	**0.04**
		Motion	0.001	0.01	1
Is behavioral ToM at V2 predicted by DMPFC selectivity at V1?	**Study 1**	DMPFC Sel_{V1}	0.28	1.5	0.15
$lm(ToM_{V2C} \sim DMPFC\ Sel_{V1} + ToM_{V1} + Age_{Av} + Motion)$		ToM_{V1}	0.65	3.0	**0.006**
		Age_{Av}	-0.11	-0.4	0.69
		Motion	-0.02	-0.1	0.91
	Study 2	DMPFC Sel_{V1}	0.06	0.3	0.80
		ToM_{V1}	0.56	2.7	**0.02**
		Age_{Av}	0.26	1.1	0.28
		Motion	-0.18	-0.9	0.40
	Combined	DMPFC Sel_{V1}	0.17	1.6	0.12
		ToM_{V1}	0.58	4.7	**3.6x10-5**
		Age_{Av}	0.22	1.5	0.15
		Motion	-0.06	-0.5	0.65
Is ToM development related to increases in selectivity?	**Study 1**	$ToM_{V2\text{-}V1}$	-0.16	-1.0	0.33
$lme(Sel_{V2\text{-}V1} \sim ToM_{V2\text{-}V1} + Age_{V1} + ROI + Motion + 1\|SubID)$		Age_{V1}	-0.20	-1.2	0.25
		ROI	0.22	0.8	0.45
		Motion	0.02	0.2	0.89
	Study 2	$ToM_{V2\text{-}V1}$	-0.43	-2.0	0.07
		Age_{V1}	-0.23	-1.1	0.29
		ROI	-0.55	-2.2	**0.048**
		Motion	0.07	0.4	0.72
	Combined	$ToM_{V2\text{-}V1}$	-0.26	-2.0	0.052
		Age_{V1}	-0.38	-2.5	**0.02**
		ROI	-0.08	-0.4	0.69
		Motion	0.04	0.26	0.80

Supplementary Table 3. Predictive Relationships. Linear regression equations are shown in the left column; statistical results are shown in the right column. Section headers correspond to those used in the results section

(section 2.3 in Results of main text). Abbreviations: Sel: Selectivity index: (Mental-Social)/(Mental-Physical)*100; ToM: Proportion correct on ToM battery (matched score unless otherwise specified (e.g. ToM_{V2C})); Age_{Av}: average age per participant, across the two visits (between-subject age differences); Age_{V1}: chronological age at Visit 1; ROI: Region of interest (RTPJ or DMPFC); Motion: Number of artifact timepoints; V1: Visit 1; V2: Visit 2; V2-V1: difference between two visits; 1|SubID: random effect of subject. P-values of significant results (p<.05) are in bold.

Supplementary Table 4

Stable Individual Differences in RTPJ & Relationship to ToM (Section 3.1 in Results of Main Text)	Study	Predictor	Beta	T-value	p-value
lme(RTPJ Sel_{V2} ~ RTPJ Sel_{V1} + Age_{Av} + 1|SubID)	**Study 1**	RTPJ Sel_{V1}	0.41	2.2	**0.04**
		Age_{Av}	-0.07	-0.4	0.72
		Motion	-0.11	-0.6	0.58
	Study 2	RTPJ Sel_{V1}	0.22	0.8	0.42
		Age_{Av}	0.08	0.3	0.74
		Motion	0.01	0.03	0.97
	Combined	RTPJ Sel_{V1}	0.33	2.3	**0.03**
		Age_{Av}	0.02	0.1	0.89
		Motion	-0.06	-0.4	0.71
Relationship between selectivity in RTPJ and ToM	**Study 1**	ToM_{Av}	0.002	0.01	0.99
lme(RTPJ Sel ~ ToM_{Av} + $ToM_{w/i\text{-}sub}$ + Motion + 1|SubID)		$ToM_{w/i\text{-}sub}$	0.01	0.1	0.93
		Motion	-0.09	0.6	0.58
	Study 2	ToM_{Av}	-0.07	-0.4	0.70
		$ToM_{w/i\text{-}sub}$	-0.05	-0.4	0.71
		Motion	0.1	0.5	0.61
	Combined	ToM_{Av}	0.004	0.03	0.97
		$ToM_{w/i\text{-}sub}$	-0.03	-0.3	0.74
		Motion	0.03	0.2	0.81
Is RTPJ selectivity predicted by earlier behavioral ToM?	**Study 1**	ToM_{V1}	-0.04	-0.2	0.86
lme(RTPJ Sel_{V2} ~ ToM_{V1} + RTPJ Sel_{V1} + Age_{Av} + Motion + 1|SubID)		RTPJ Sel_{V1}	0.41	2.1	**0.046**
		Age_{Av}	-0.05	-0.2	0.84
		Motion	-0.11	-0.6	0.58
	Study 2	ToM_{V1}	0.15	0.5	0.59
		RTPJ Sel_{V1}	0.25	0.9	0.38
		Age_{Av}	0.03	0.1	0.91
		Motion	-0.001	-0.006	0.99
	Combined	ToM_{V1}	0.09	0.5	0.60
		RTPJ Sel_{V1}	0.33	2.3	**0.03**
		Age_{Av}	-0.03	-0.2	0.87
		Motion	-0.06	0.3	0.74
Does ToM development predict amount of change in selectivity in RTPJ?	**Study 1**	$ToM_{V2\text{-}V1}$	0.02	0.1	0.92
lme(RTPJ $Sel_{V2\text{-}V1}$ ~ $ToM_{V2\text{-}V1}$ + Age_{V1} + Motion + 1|SubID)		Age_{V1}	-0.1	-0.4	0.67
		Motion	0.01	0.1	0.96
	Study 2	$ToM_{V2\text{-}V1}$	-0.53	-2.1	0.057
		Age_{V1}	-0.32	-1.2	0.24
		Motion	-0.03	-0.1	0.89
	Combined	$ToM_{V2\text{-}V1}$	-0.26	-1.5	0.13
		Age_{V1}	-0.25	-1.3	0.21
		Motion	0.05	0.3	0.80
Predictive Relationships between Neural and Behavioral ToM (Section 3.2 in Results of Main Text)	**Study**	**Predictor**	**Beta**	**T-value**	**p-value**
Does early ToM predict amount of change in selectivity?	**Study 1**	ToM_{V1}	0.10	0.7	0.47
lme($Sel_{V2\text{-}V1}$ ~ ToM_{V1} + Age_{V1} + Sel_{V1} + Motion + 1|SubID)		Age_{V1}	-0.15	-1.0	0.31
		Sel_{V1}	-0.62	-5.4	**0**
		ROI	0.36	1.6	0.13
		Motion	-0.07	-0.6	0.56
	Study 2	ToM_{V1}	0.12	0.8	0.46
		Age_{V1}	-0.03	-0.2	0.85
		Sel_{V1}	-0.61	-3.9	**0.002**
		ROI	-0.31	-1.3	0.21
		Motion	0.11	0.7	0.49
	Combined	ToM_{V1}	0.12	1.1	0.27
		Age_{V1}	-0.20	1.7	0.10
		Sel_{V1}	-0.59	-6.8	**0**
		ROI	0.06	0.3	0.73
		Motion	-0.02	-0.2	0.81
Does early selectivity in RTPJ predict amount of ToM improvement?	**Study 1**	RTPJ Sel_{V1}	-0.002	-0.03	0.97
lm($ToM_{V2\text{-}V1}$ ~ RTPJ Sel_{V1} + ToM_{V1} + Age_{V1} + Motion)		ToM_{V1}	-1.05	-14.3	**3.16x10-13**
		Age_{V1}	0.15	1.8	0.08
		Motion	0.06	1.0	0.33
	Study 2	RTPJ Sel_{V1}	-0.09	-0.9	0.36
		ToM_{V1}	-1.06	-9.4	**3.5x10-7**
		Age_{V1}	-0.01	-0.1	0.91
		Motion	0.09	0.9	0.38
	Combined	RTPJ Sel_{V1}	-0.03	-0.8	0.45
		ToM_{V1}	-1.15	-18.1	**<2.0x10-16**
		Age_{V1}	0.10	1.6	0.11
		Motion	0.05	1.0	0.33
Does early selectivity in RTPJ predict amount of ToM improvement?	**Study 1**	DMPFC Sel_{V1}	0.0005	0.008	0.99
lm($ToM_{V2\text{-}V1}$ ~ DMPFC Sel_{V1} + ToM_{V1} + Age_{V1} + Motion)		ToM_{V1}	-1.03	-14.6	**1.79x10-12**
		Age_{V1}	0.10	1.1	0.28
		Motion	0.07	0.8	0.43
	Study 2	DMPFC Sel_{V1}	-0.18	-1.3	0.23
		ToM_{V1}	-0.80	-6.5	**4.29x10-5**
		Age_{V1}	-0.06	-0.5	0.66
		Motion	0.16	1.1	0.29
	Combined	DMPFC Sel_{V1}	-0.05	-0.8	0.43
		ToM_{V1}	-0.91	-12.0	**2.84x10-14**
		Age_{V1}	-0.03	-0.4	0.72
		Motion	0.12	1.3	0.21

Supplementary Table 4. Full Regression Statistics for fMRI Results of Exploratory Analyses. Linear regression equations are shown in the left column; statistical results are shown in the right column. Section headers correspond to those used in the results section (sections 3.1 and 3.2 of Results of main text). Abbreviations: Sel: Selectivity index: (Mental-Social)/(Mental-Physical)*100; ROI: Region of interest (RTPJ or DMPFC); Motion: Number of artifact timepoints; Age_{Av}: average age per participant, across the two visits (between-subject age differences); Age_{V1}: chronological age at Visit 1; ToM_{Av}: average (matched) ToM score per participant, across the two visits (between-subject differences in ToM); $ToM_{w/i\text{-}sub}$: difference between participant's average ToM and the ToM score at each visit (within-subject change in ToM); V1: Visit 1; V2: Visit 2; 1|SubID: random effect of subject. P-values of significant results (p<.05) are in bold.

Supplementary Table 5

Developmental Change in Response Selectivity in Group ROIs	Study	Predictor	Beta	T-value	p-value
Cross-sectionally: lm(Sel ~ Age + ROI + Motion)	**Study 1**	Age	-0.08	-0.8	0.43
		ROI	0.07	0.4	0.71
		Motion	-0.006	-0.1	0.95
	Study 2	Age	0.22	2.2	**0.03**
		ROI	0.09	0.5	0.63
		Motion	0.12	1.2	0.22
	Combined	Age	0.03	0.4	0.67
		ROI	0.07	0.5	0.60
		Motion	0.05	0.7	0.46
Longitudinally: lme(Sel ~ Age + ROI + Motion + 1|SubID)	**Study 1**	Age	-0.10	-1.0	0.34
		ROI	0.07	0.4	0.69
		Motion	-0.02	-0.3	0.81
	Study 2	Age	0.22	2.3	**0.03**
		ROI	0.09	0.5	0.63
		Motion	0.12	1.3	0.21
	Combined	Age	0.02	0.3	0.77
		ROI	0.07	0.6	0.58
		Motion	0.04	0.6	0.57
Simultaneous test of within- and between-subject age differences:	**Study 1**	Age_{Av}	0.03	0.3	0.73
lme(Sel ~ Age_{Av}+ $Age_{w/i\text{-}sub}$ + ROI + Motion+ 1|SubID)		$Age_{w/i\text{-}sub}$	-0.15	-1.6	0.11
		ROI	0.07	0.4	0.72
		Motion	0.08	0.8	0.45
	Study 2	Age_{Av}	0.12	1.1	0.30
		$Age_{w/i\text{-}sub}$	0.16	1.8	0.08
		ROI	0.08	0.5	0.65
		Motion	0.08	0.7	0.50
	Combined	Age_{Av}	0.05	0.6	0.56
		$Age_{w/i\text{-}sub}$	-0.08	-1.2	0.25
		ROI	0.07	0.6	0.58
		Motion	0.08	1.0	0.32
Simultaneous test of within- and between-subject ToM differences:	**Study 1**	ToM_{Av}	0.01	0.1	0.88
lme(Sel ~ ToM_{Av}+ $ToM_{w/i\text{-}sub}$ + ROI + Motion+ 1|SubID)		$ToM_{w/i\text{-}sub}$	-0.23	-2.5	**0.01**
		ROI	0.07	0.4	0.71
		Motion	0.07	0.7	0.46
	Study 2	ToM_{Av}	0.02	0.1	0.89
		$ToM_{w/i\text{-}sub}$	0.12	1.2	0.22
		ROI	0.07	0.4	0.73
		Motion	0.02	0.2	0.86
	Combined	ToM_{Av}	0.04	0.5	0.65
		$ToM_{w/i\text{-}sub}$	-0.05	-0.7	0.46
		ROI	0.07	0.5	0.63
		Motion	0.10	1.3	0.21
Stable Neural Individual Differences in Group ROIs	**Study**	**Predictor**	**Beta**	**T-value**	**p-value**
lme(Sel_{V2} ~ Sel_{V1} + Age_{Av} + ROI + 1|SubID)	**Study 1**	Sel_{V1}	-0.01	-0.1	0.92
		Age_{Av}	-0.002	-0.01	0.99
		ROI	0.05	0.2	0.83
		Motion	0.05	0.3	0.78
	Study 2	Sel_{V1}	0.16	1.1	0.29
		Age_{Av}	0.10	0.7	0.50
		ROI	-0.41	-1.4	0.17
		Motion	0.16	1.2	0.26
	Combined	Sel_{V1}	-0.003	-0.03	0.98
		Age_{Av}	-0.05	-0.4	0.66
		ROI	-0.11	-0.6	0.54
		Motion	0.07	0.6	0.59

Supplementary Table 5. Full Regression Statistics for fMRI Results in Group ROIs: Developmental Change and Stable Individual Differences. Linear regression equations are shown in the left column; statistical results are shown in the right column. Section headers correspond to those used in the results section (sections 1.1, 1.2, 1.4.1, and 1.4.2 of Supplementary Information). Abbreviations: Sel: Selectivity index: (Mental-Social)*100; ROI: Region of interest (RTPJ or DMPFC); Motion: Number of artifact timepoints; Age: chronological age per participant per visit; Age: chronological age per participant per visit; Age_{Av}: average age per participant, across the two visits (between-subject age differences); $Age_{w/i\text{-}sub}$: difference between participant's average age and their age at each visit (within-subject change in age); ToM_{Av}: average (matched) ToM score per participant, across the two visits (between-subject differences in ToM); $ToM_{w/i\text{-}sub}$: difference between participant's average ToM and the ToM score at each visit (within-subject change in ToM); V1: Visit 1; V2: Visit 2; 1|SubID: random effect of subject. P-values of significant results ($p<.05$) are in bold.

Supplementary Table 6

Predictive Relationships between Neural and Behavioral ToM in Group ROIs	Study	Predictor	Beta	T-value	p-value
Does behavioral ToM at V1 predict selectivity at V2?	**Study 1**	ToM_{V1}	0.35	1.7	0.09
$lme(Sel_{V2} \sim ToM_{V1} + Sel_{V1} + Age_{Av} + ROI + Motion + 1\|SubID)$		Sel_{V1}	0.03	0.3	0.80
		Age_{Av}	-0.20	-1.0	0.31
		ROI	0.08	0.3	0.74
		Motion	0.05	0.3	0.73
	Study 2	ToM_{V1}	0.05	0.3	0.74
		Sel_{V1}	0.16	1.1	0.29
		Age_{Av}	0.08	0.6	0.59
		ROI	-0.41	-1.4	0.18
		Motion	0.16	1.1	0.28
	Combined	ToM_{V1}	0.18	1.3	0.19
		Sel_{V1}	0.009	0.1	0.93
		Age_{Av}	-0.15	-1.1	0.28
		ROI	-0.11	-0.6	0.55
		Motion	0.06	0.5	0.60
Is behavioral ToM at V2 predicted by RTPJ selectivity at V1?	**Study 1**	RTPJ Sel_{V1}	0.13	0.8	0.44
$lm(ToM_{V2C} \sim RTPJ\ Sel_{V1} + ToM_{V1} + Age_{Av} + Motion)$		ToM_{V1}	0.62	3.1	**0.005**
		Age_{AV}	-0.07	-0.33	0.74
		Motion	0.04	0.3	0.79
	Study 2	RTPJ Sel_{V1}	0.14	0.9	0.36
		ToM_{V1}	0.49	3.2	**0.005**
		Age_{AV}	0.29	1.8	0.09
		Motion	-0.30	-2.0	0.06
	Combined	RTPJ Sel_{V1}	0.12	1.3	0.21
		ToM_{V1}	0.53	4.6	**2.6x10-5**
		Age_{Av}	0.27	2.2	**0.03**
		Motion	0.0009	0.008	0.99
Is behavioral ToM at V2 predicted by DMPFC selectivity at V1?	**Study 1**	DMPFC Sel_{V1}	0.25	1.6	0.12
$lm(ToM_{V2C} \sim DMPFC\ Sel_{V1} + ToM_{V1} + Age_{Av} + Motion)$		ToM_{V1}	0.62	3.3	**0.003**
		Age_{Av}	-0.05	-0.3	0.80
		Motion	0.05	0.3	0.77
	Study 2	DMPFC Sel_{V1}	0.15	0.9	0.41
		ToM_{V1}	0.47	3.0	**0.007**
		Age_{Av}	0.32	2.0	0.055
		Motion	-0.27	-1.7	0.11
	Combined	DMPFC Sel_{V1}	0.22	2.3	**0.02**
		ToM_{V1}	0.52	4.8	**1.71x10-5**
		AgeAv	0.24	2.0	0.054
		Motion	-0.004	-0.04	0.97
Is ToM development related to increases in selectivity?	**Study 1**	$ToM_{V2\text{-}V1}$	-0.40	-2.3	**0.03**
$lme(Sel_{V2\text{-}V1} \sim ToM_{V2\text{-}V1} + Age_{V1} + ROI + Motion + 1\|SubID)$		Age_{V1}	-0.21	-1.3	0.22
		ROI	0.03	0.12	0.90
		Motion	0.002	0.02	0.99
	Study 2	$ToM_{V2\text{-}V1}$	-0.05	-0.4	0.73
		Age_{V1}	-0.01	-0.09	0.93
		ROI	-0.75	-2.7	**0.01**
		Motion	0.19	1.3	0.20
	Combined	$ToM_{V2\text{-}V1}$	-0.20	-1.7	0.098
		Age_{V1}	-0.22	-1.7	0.0997
		ROI	-0.26	-1.4	0.16
		Motion	-0.05	-0.4	0.68

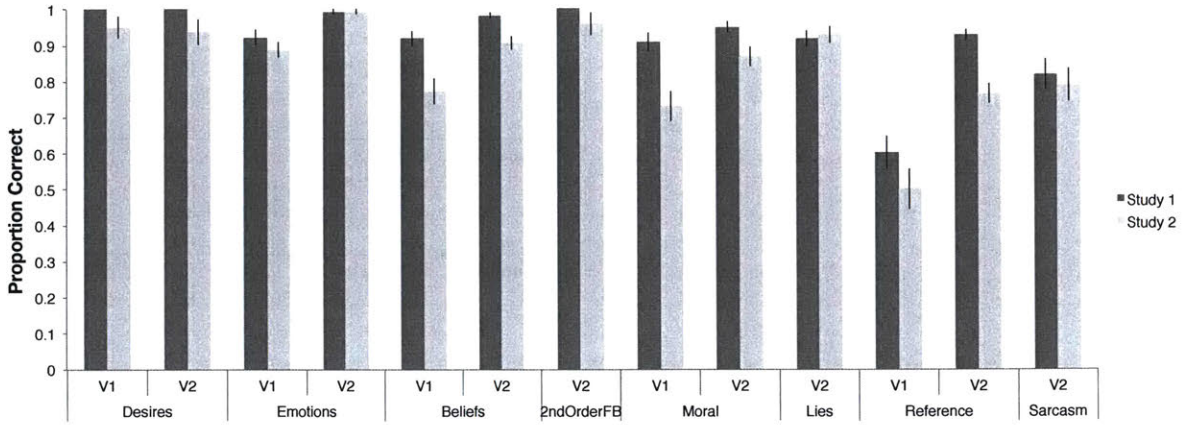
Supplementary Table 6. Full Regression Statistics for fMRI Results in Group ROIs: Predictive Relationships. Linear regression equations are shown in the left column; statistical results are shown in the right column. Section headers correspond to those used in the results section (sections 1.3 and 1.4.3 of Supplementary Information). Abbreviations: Sel: Selectivity index: (Mental-Social)*100; ROI: Region of interest (RTPJ or DMPFC); Motion: Number of artifact timepoints; Age_{Av}: average age per participant, across the two visits (between-subject age differences); ToM: proportion correct on ToM behavioral task; matched score is used unless otherwise specified (ToM_{V2C}: Visit 2 "complete" score, which uses all items instead of only items that were matched across visits); V1: Visit 1; V2: Visit 2; V2-V1: difference between two visits; 1|SubID: random effect of subject. P-values of significant results ($p<.05$) are in bold.

Supplementary Table 7

Does early RTPJ selectivity predict later DMPFC selectivity?	Study	Predictor	Beta	T-value	p-value
lm(DMPFC Sel_{V2} ~ RTPJ Sel_{V1} + Age_{Av} + Motion)	**Study 1**	RTPJ Sel_{V1}	-0.13	-0.6	0.55
		DMPFC Sel_{V1}	0.10	0.4	0.67
		Age_{Av}	-0.29	-1.1	0.30
		Motion	-0.11	-0.5	0.66
	Study 2	RTPJ Sel_{V1}	0.40	1.3	0.23
		DMPFC Sel_{V1}	-0.32	-1.0	0.34
		Age_{Av}	-0.07	-0.3	0.79
		Motion	0.16	0.8	0.47
	Combined	RTPJ Sel_{V1}	0.03	0.2	0.86
		DMPFC Sel_{V1}	0.05	0.3	0.79
		Age_{Av}	-3.72	-1.1	0.28
		Motion	-0.1	-0.5	0.59

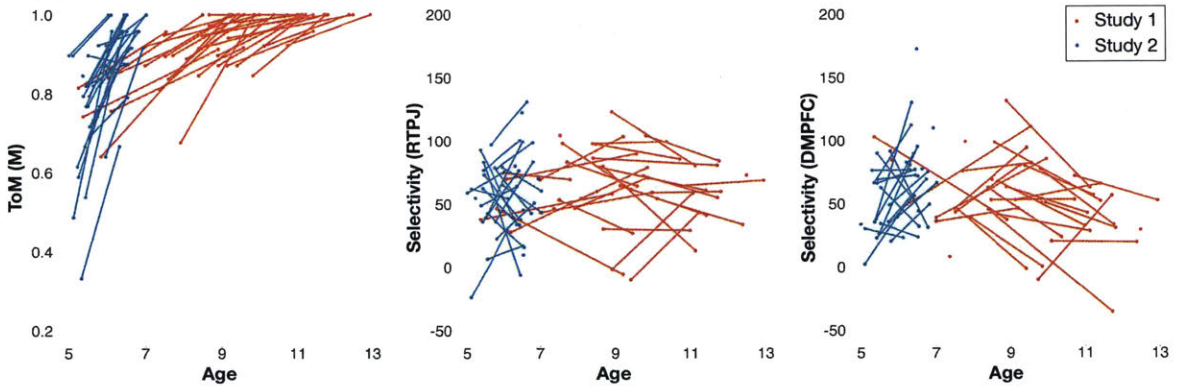
Supplementary Table 7. Planned exploratory analysis of predictive relationships between RTPJ and DMPFC. Linear regression equations are shown in the left column; statistical results are shown in the right column. Section headers correspond to those used in the results section (section 2 of Supplementary Information). Abbreviations: Sel: Selectivity index: (Mental-Social)/(Mental-Physical)*100; Motion: Number of artifact timepoints; Age_{Av}: Average age per participant, across the two visits (between-subject age differences); V1: Visit 1; V2: Visit 2. P-values of significant results ($p<.05$) are in bold.

Supplementary Figure 1



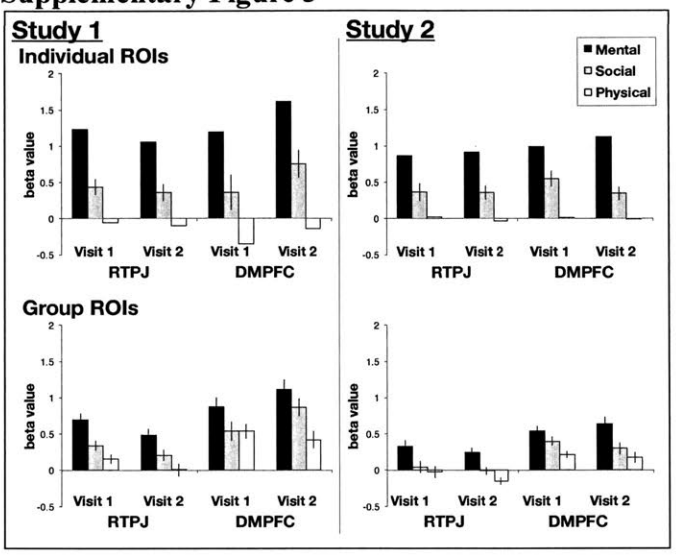
Supplementary Figure 1. ToM Behavioral Performance by Category. Bars show average and standard error for proportion correct across participants, per visit and per question category. Study 1 is shown in dark grey; Study 2 in light grey.

Supplementary Figure 2



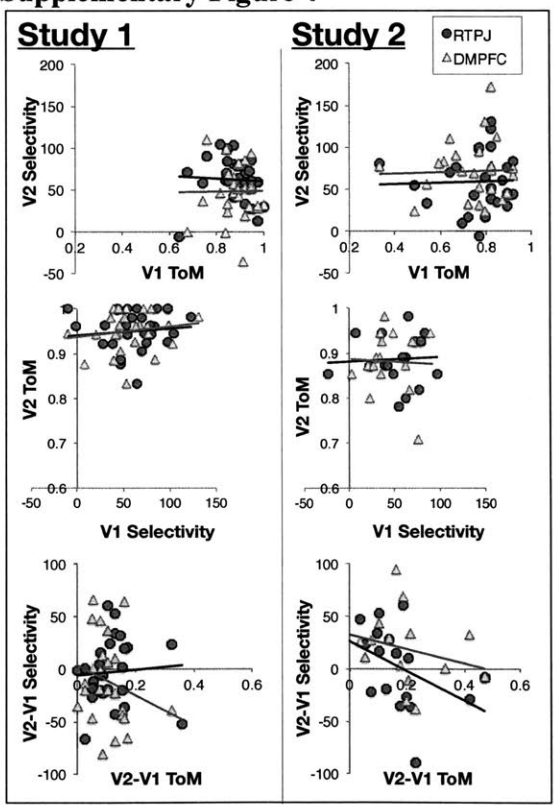
Supplementary Figure 2. Developmental Trajectories for ToM Behavior and Selectivity. Each line connects the two data points from each participant (one data point per visit), in order to show amount and rate of developmental change in behavioral theory of mind (left), and selectivity in RTPJ (middle) and DMPFC (right). Study 1 participants are shown in red; Study 2 participants are shown in blue.

Supplementary Figure 3



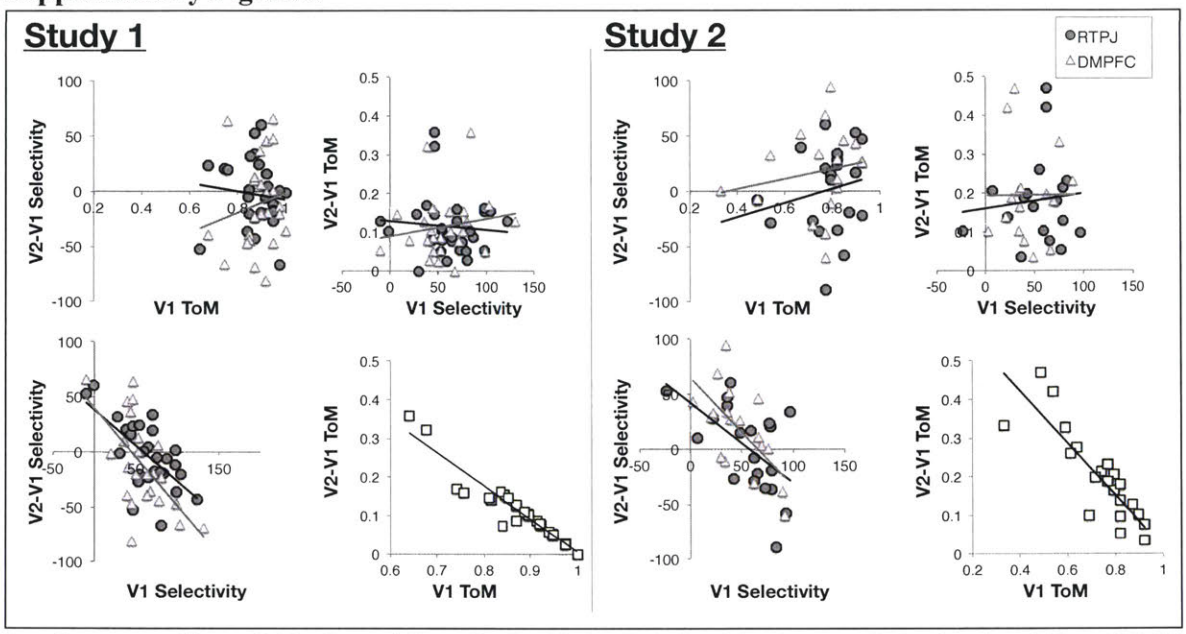
Supplementary Figure 3. Average beta values per condition. Bars show average and standard error for beta values per condition, for each visit, ROI, and Study. The top row shows data from individual ROIs defined based on the Mental and Physical conditions. Because these conditions are non-independent from ROI definition, they are plotted for visualization purposes only (and therefore do not have standard error bars). The bottom row shows data from group ROIs, in which every condition is independent from ROI definition.

Supplementary Figure 4



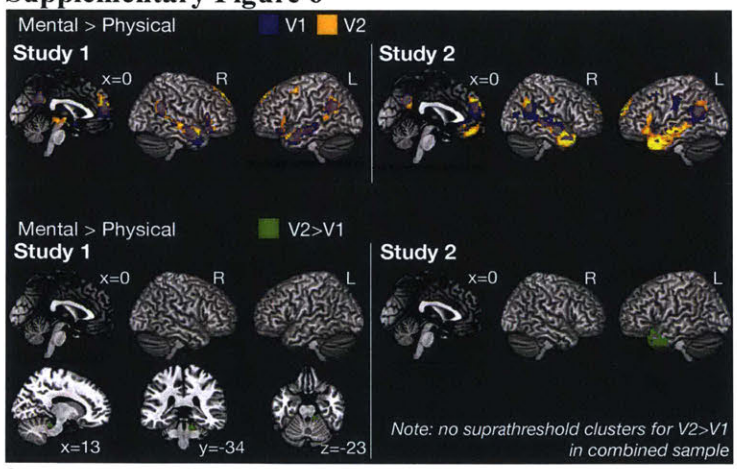
Supplementary Figure 4. Planned Tests for Predictive Relationships between Behavioral and Neural ToM. We did not find evidence for predictive relationships between behavioral Theory of Mind score and response selectivity. Theory of mind behavior does not predict and is not predicted by selectivity. Developmental improvement in ToM is not significantly correlated with change in response selectivity in RTPJ or DMPFC. Abbreviations: V1: Visit 1; V2: Visit 2; V2-V1: difference between Visit 1 and Visit 2; ToM is proportion correct on the ToM booklet task; V2 ToM is the complete score, whereas V2-V1 ToM uses the matched scores. Selectivity is calculated using beta values estimated per condition, per ROI, per participant: Selectivity index = (Mental-Social/Mental-Physical)*100).

Supplementary Figure 5



Supplementary Figure 5. Unplanned Tests for Predictive Relationships between Behavioral and Neural ToM. Amount of change in selectivity and ToM behavior between the two visits was predicted by Visit 1 selectivity and ToM score, respectively: participants who were the least selective responses at Visit 1 showed the most developmental change between visits, and participants who had the lowest scores on the ToM task at Visit 1 showed the most improvement between visits. Abbreviations: V1: Visit 1; V2: Visit 2; V2-V1: difference between Visit 1 and Visit 2; ToM is proportion correct on the ToM booklet task; V2-V1 ToM uses the matched scores. Selectivity is calculated using beta values estimated per condition, per ROI, per participant: Selectivity index = (Mental-Social/Mental-Physical)*100).

Supplementary Figure 6



Supplementary Figure 6. Whole-Brain Random Effects Analysis. The top row shows the response to Mental > Physical per visit (V1 in blue; V2 in orange), per study (corrected for multiple comparisons; $p<.05$). The bottom row shows surviving clusters for the V2 Mental > Physical – V1 Mental > Physical difference (green).

Supplementary References

1. Fedorenko, E., Hsieh, P. J., Nieto-Castanon, A., Whitfield-Gabrieli, S. & Kanwisher, N. New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *Journal of Neurophysiology* **104,** 1177–1194 (2010).
2. Saxe, R., Brett, M. & Kanwisher, N. Divide and conquer: a defense of functional localizers. *NeuroImage* **30,** 1088–1096 (2006).
3. Dufour, N. *et al.* Similar Brain Activation during False Belief Tasks in a Large Sample of Adults with and without Autism. *PLoS ONE* **8,** e75468 (2013).
4. Sabbagh, M. A., Bowman, L. C., Evraire, L. E. & Ito, J. M. B. Neurodevelopmental correlates of theory of mind in preschool children. *Child Dev* **80,** 1147–1162 (2009).
5. Giedd, J. N. *et al.* Brain development during childhood and adolescence: a longitudinal MRI study. *Nat Neurosci* **2,** 861–863 (1999).
6. Gogtay, N. *et al.* Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences* **101,** 8174–8179 (2004).

Chapter 2: Development of the social brain from age three to twelve years

Human adults recruit distinct networks of brain regions to think about the bodies and minds of others. This study characterizes the development of these networks, and tests for relationships between neural development and behavioral changes in reasoning about others' minds ("theory of mind", ToM). A large sample of children (n=122, 3-12 years), and adults (n=33), watched a short movie while undergoing fMRI. The movie highlights the characters' bodily sensations (often pain) and mental states (beliefs, desires, emotions), and is a feasible experiment for young children. Here we report three main findings: 1) ToM and pain networks are functionally distinct by age three years, 2) functional specialization increases throughout childhood, and 3) functional maturity of each network is related to increasingly anti-correlated responses between the networks. Furthermore, the most studied milestone in ToM development, passing explicit false-belief tasks, does not correspond to discontinuities in the development of the social brain.

Note: A version of this chapter appeared as:
Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. Nature communications, 9(1), 1027.

Manuscript is available at: http://rdcu.be/IRh8

Introduction

Over the past decade, fMRI research has made significant progress identifying functional divisions of labor within the adult social brain[1]. For example, while many areas of human cortex show elevated responses while looking at, listening to, or thinking about other people, studies of these cortical responses suggest a striking division between regions responding preferentially to internal states of others' bodies, versus internal states of others' minds[2,3,4,5,6]. Both bodily sensations, like hunger and pain, and mental states, like beliefs and desires, are internal states of other people; both are important for observers' reasoning about others' actions and reactions, to facilitate the observer's own prosocial (e.g. helping) or antisocial (e.g. competing) choices. In spite of these similarities, a robust dissociation between responses to others' bodies and minds has been replicated across a wide range of paradigms: when human adults think about other people, our cortical responses are surprisingly dualist[7].

An important extension of this work is to study the emergence of these functionally specialized brain regions during development. The current study investigates the developmental origins of the cortical dissociation between others' bodies and minds, and the links between cortical and cognitive changes in children's social development.

Although children's developing understanding of others' minds (their "theory of mind" (ToM)) has been studied intensively[8], we know very little about the neural changes that support this development. One cause of this gap in knowledge is that most behavioral studies on ToM focus on children younger than five years old[9,10]. For example, one active debate in developmental psychology concerns children and infants' ability to reason about false beliefs[11]. Children's ability to explicitly predict or explain another person's actions based on her false beliefs has been interpreted as depending on a conceptual leap occurring around age 4 years[12-14]. However, recent measures of spontaneous looking and helping suggest that even toddlers may be sensitive to others' false beliefs[15,16]. By contrast, fMRI studies of ToM reasoning have focused on children older than five years old[17-23], adolescents[24,25], and adults[26-28]. Prior neuroimaging studies thus leave open questions of core interest concerning early stages of theory of mind development.

Based on theories in developmental psychology, we derive three predictions for observations in the social brain regions of young children. First, success on explicit false-belief tasks could reflect an important conceptual leap or discontinuity in ToM development, as theories of others' internal states are dramatically altered by insight into the representational nature of mental states[29,30]. According to this view, the division between cortical responses to others' bodies versus minds might emerge concurrently with childrens' explicit understanding of false beliefs. Second, success on explicit false-belief tasks could reflect development in other domain-general brain regions, removing earlier performance limitations (such as response inhibition and selection, and production of verbal response)[31-33]. According to this view, spontaneous processing of others' mental states within domain-specific regions for ToM might be similar in children who pass and fail explicit false-belief tasks. Third, success on explicit false-belief tasks could be a single step in the ongoing conceptual development of ToM, which begins before – and continues after – false-belief reasoning[34-37]. According to this view, change within ToM brain regions might occur both before and after children explicitly reason about false-beliefs. Of course, these predictions only reflect a subset of those that could be derived from each theoretical perspective, and are not mutually exclusive; reality could include a mixture of these three views.

The present study characterizes development of brain regions recruited for reasoning about others' minds and bodies, in a large, cross-sectional sample of children between the ages of 3-12 years old. These 122 children and a reference group of 33 adults, watched a short, animated movie that included events evoking the mental states and physical sensations of the characters, while undergoing fMRI. Watching this movie is feasible for young children – it is short, engaging, and does not require learning a task. This movie has been validated as activating ToM brain regions and the pain matrix in adults[38]. ToM brain regions include bilateral temporoparietal junction, precuneus, and dorso-, middle-, and ventromedial prefrontal cortex[26-28]. The pain matrix includes brain regions recruited when perceiving the physical pain and bodily sensations of others: bilateral medial frontal gyrus, insula, and secondary sensory cortex, and dorsal anterior middle cingulate cortex[39]. Within both functional networks, individual regions have been implicated with specific functions (for example, insula and cingulate cortex for nociceptive pain[39], and prefrontal cortex for reasoning about emotions and preferences[40]). Here, we collapse across specific functions, and operationalize ToM and pain networks recruited generally for reasoning about others' internal mental and physical states, respectively[38].

We measured three features of children's hemodynamic responses during the movie. First, we conducted inter-region correlation analyses to test the degree to which ToM and pain brain regions operate as functionally distinct networks (i.e. high within-network, and low between-network correlations)[41,42]. Because results suggested that networks for ToM and pain are distinct even in the youngest children, we used the average response of each network in the next two analyses. Second, we measured the magnitude of evoked response, in children, to the events in the movie that evoke peak responses in adults (identified by reverse correlation analyses). Third, we measured the functional maturity (i.e. similarity to adults) of each network's entire timecourse[43]. All child participants additionally completed an assessment of explicit ToM after the scan, to measure overall theory of mind reasoning, including performance on explicit false-belief tasks. We tested whether each of the three neural measures was related to children's age, to children's explicit performance on ToM tasks, and to one another.

We report evidence that ToM and pain networks are functionally distinct by three years of age, and become increasingly specialized between the ages of three and twelve years. Functional maturity of each network is related to increasingly anti-correlated responses between the two networks. Finally, we find that a distinct neural response to others' minds and bodies is present before - and continues to develop after - children pass explicit false-belief tasks.

Results

Behavioral Results

All children completed a behavioral battery after completing the fMRI scan, which included a custom-made explicit ToM task (see Methods)[21]. Three to five-year-old children (n=65) additionally completed a measure of response inhibition (Dimensional Change Card Sort task (DCCS)[44]). Performance on the ToM task (proportion correct) and DCCS were both positively correlated with age (ToM (kendall tau correlation test (n=122)): $r_k(120)=.66$, $p<.00001$; DCCS (kendall tau correlation test (n=64)): $r_k(62)=.20$, $p=.049$); see Figure 1a. In the three to five-year-old subset of children who completed both measures, ToM and DCCS scores were positively

correlated (partial kendall tau correlation test (n=64), controlling for age: $r_k(61)=.19$, p=.03). See Supplementary Table 1 for behavioral data and participant demographics.

Figure 1

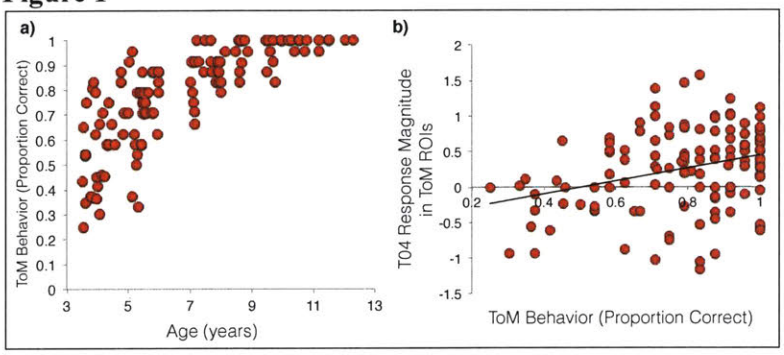


Figure 1. Theory of mind behavioral performance. a) Theory of mind behavioral performance (proportion correct; y-axis) of all children (n=122) by age in years (x-axis). **b)** Average response magnitude in ToM network to peak timepoint of event T04 (Peck returning to Gus, donning protective gear), per child (y-axis), by theory of mind behavioral performance (proportion correct; x-axis).

For three to five-year-old children, an explicit false-belief composite score was calculated based on responses to six explicit false-belief questions embedded within the ToM measure; this composite measure was used to categorize these children as false-belief passers (5-6 FB questions correct; n=30 (15 female)), inconsistent performers (3-4 FB questions correct; n=20 (13 female)), and false-belief failers (0-2 FB questions correct; n=15 (6 female)). False-belief task failers and inconsistent performers did worse on the remaining ToM items than passers (Fail M(SE)=.55(.04), Inc M(SE)=.57(.03), Pass M(SE)=.75(.02); Tukey Honest Significant Difference (HSD) test of ToM*FB-Group ANOVA: Pass-Fail: diff=1.2, p<.00005; Pass-Inc: diff=1.08, p<.0001; Inc-Fail: diff=.16, p=.8; Kruskal-Wallis rank sum test of ToM*FB-Group (for non-normal distributions; 3 groups: Pass (n=30), Inc (n=20), Fail (n=15)): $H(2)=22.96$, p<.0001). False-belief task failers were on average younger than passers and inconsistent performers (Fail M(SD)=4.1(.56) years; Inc M(SD)=4.8(.73) years; Pass M(SD)=5.2(.70) years; Tukey HSD test of Age*FB-Group ANOVA: Pass-Fail: diff=1.4, p<.00001; Inc-Fail: diff=.83, p=.01; Pass-Inc: diff=.59, p=.047). Similarly, failers demonstrated worse response inhibition than the other two groups (DCCS Summary score: Fail M(SE)=1.73(.21), Inc M(SE)=2.26(.17), Pass M(SE)=2.33(.09); Tukey HSD test of DCCS*FB-Group ANOVA: Pass-Fail: diff=.88, p=.01; Inc-Fail: diff=.78, p=.052; Pass-Inc: diff=.1, p=.9; Kruskal-Wallis rank sum test of DCCS*FB-Group (for non-normal distributions; 3 groups: Pass (n=30), Inc (n=19), Fail (n=15)): $H(2)=7.56$, p=.02).

Inter-region Correlation Analysis

Inter-region correlation analyses reveal the extent to which a group of brain regions operate as a network with synchronized responses. We conducted inter-region correlation analyses (see Methods)[42], in order to test three hypotheses about the development of ToM and pain brain regions: 1) that adults exhibit greater within-network correlations and greater anti-correlations between ToM and pain networks, compared to children, 2) that by age three, ToM and pain brain regions operate as specialized networks with synchronized responses, and 3) that maturity of the within- and across- network correlations is related to ToM task performance in childhood.

In adults, each network exhibited strong positive correlations within-network, and strong negative correlations across network (within-ToM correlation M(SE)=.48(.02); within-Pain

correlation M(SE)=.35(.02); across-network M(SE)=-.17(.02); paired sample two-tailed t-tests (n=33): within-ToM vs. across-network: t(32)=19.1, $p<2.2x10^{-16}$; within-Pain vs. across-network: t(32)=23.2, $p<2.2x10^{-16}$). See Methods, Supplementary Fig. 1, and Supplementary Table 2 for details about the regions of interest.

This pattern of network correlations strengthened substantially between the ages of three and twelve years (Figure 2; Supplementary Fig. 2 & 3). Among children, within-ToM and within-Pain network correlations increased significantly with age (spearman partial correlation test, including motion (number of artifact timepoints) as a covariate (n=122): within-ToM: $r_s(119)=.37$, p<.00005; within-Pain: $r_s(119)=.28$, p=.002). Across-network correlations decreased significantly with age (spearman partial correlation test, including motion as a covariate (n=122): $r_s(119)=-.35$, p<.0001). Within and across-network correlations were significantly greater in adults, compared to children (linear regression testing for effects of age group and motion on within-ToM correlation: effect of group (child (n=122) vs. adult (n=33)): b=-.97, t=-5.7, $p<6.2x10^{-8}$, effect of motion: b=-.3, t=-4.3, p<.0001; linear regression testing for effects of age group and motion on within-Pain correlation: effect of group (child (n=122) vs. adult (n=33)): b=-.75, t=-3.8, p=.0002, effect of motion: b=-.03, t=-.31, p=.8; linear regression testing for effects of age group and motion on across-network correlation: effect of group (child (n=122) vs. adult (n=33)): b=1.26, t=7.2, $p=2.2x10^{-11}$, effect of motion: b=.07, t=.94, p=.4). To ensure that developmental changes in correlation strength is not driven by various aspects of data quality (such as improved co-registration with age), we conducted inter-region correlation analyses on face and scene brain regions as well as bilateral primary motor and visual cortices; see Supplementary Fig. 3. These analyses showed that inter-region correlations in other networks (e.g. the face network and primary visual areas) do not show age-related change.

Figure 2

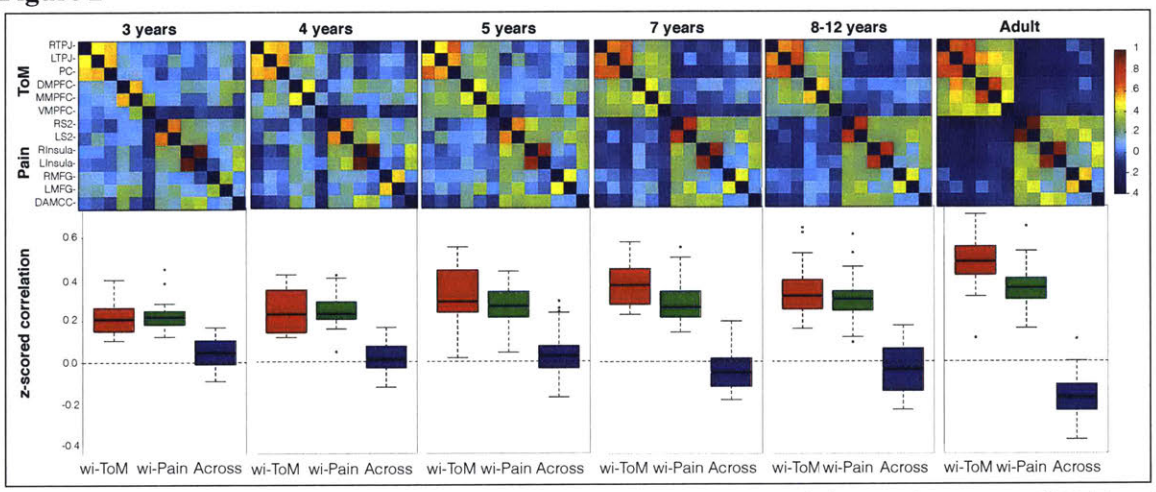


Figure 2. Inter-region correlation analysis. Top row: Average z-scored correlation matrices across all ToM and pain brain regions of interest (see Y-axis) per age group (3yo: n=17; 4yo: n=14; 5yo: n=34; 7yo: n=23; 8-12yo: n=34; adults: n=33). Regions are in the same order along the X- and Y-axes. Bottom row: Boxplots of the within-ToM (red), within-Pain (green), and across-network (blue) z-scored correlation values per age group. Note that while data are binned into age groups here, age is a continuous variable in statistical tests.

Nevertheless, the two networks were already functionally distinct in the youngest group of children we tested. In three-year-old children only (n=17), both ToM and pain networks had positive within-network correlations (within-ToM correlation M(SE)=21(.02); within-Pain correlation M(SE)=.23(.02)). Within-network correlations were higher than the across-network correlation (paired sample two-tailed t-tests (n=17): within-ToM vs. across-network: t(16)=6.2, p <.00005, within-Pain vs. across-network: t(16)=6.9, p<.00001). By contrast, unlike adults, ToM and pain networks were not anti-correlated in three year olds (across-network correlation M(SE)=.05(.02)). However, significantly greater within- than across- network correlations suggests that ToM and pain networks are functionally distinct by age three years. The strongest within-network correlations in the three year olds were between homologous pairs of regions in opposite hemispheres, such as right and left TPJ (ToM), and the right and left insula (Pain). These strong correlations, between pairs of regions that are functionally homologous but physically distant, suggest that even the data from three year old children are of high enough quality to detect inter-region correlations when they exist; and therefore that changes with age in other inter-region correlations reflect real changes in the functional relationships between those regions. However, the functional separation of the two networks was not fully explained by the strong correlations between bilateral pairs (Within-non-bilateral-ToM correlation M(SE) =.20(.02), Within-non-bilateral-Pain correlation M(SE) =.17(.02); paired sample two-tailed t-tests (n=17): within-non-bilateral-ToM vs. across-network: t(16)=5.1, p=.0001, within-non-bilateral-Pain vs. across-network: t(16)=4.4, p=.0005).

In children the strength of inter-region correlations within the ToM network was positively correlated with behavioral performance on the ToM battery outside the scanner (kendall tau partial correlation test, including motion as a covariate (n=122): r_k(119)=.23, p=.0002). The anti-correlation of ToM and pain networks was also correlated with ToM score (kendall tau partial correlation test, including motion as a covariate (n=122): r_k(119)=-.20, p=.001). However, there was no relationship between within-ToM or across-network correlations and ToM score when controlling for age in addition to motion (linear regressions testing for effect ToM score on within-ToM and across-network correlation, including age and motion as additional predictors (n=122): NS effects of ToM score: ts<1, ps>.3).

We additionally tested for neural differences based on performance on explicit false-belief questions, among 3- to 5-year-old children. These questions were a subset of the questions in the ToM behavioral battery (see Methods). There was a significant difference in within-ToM network correlation between explicit false-belief task passers and failers (Within-ToM: Passers M(SE)=.29(.02), Failers M(SE)=.25(.03); linear regression testing for effects of FB-Group (pass vs. fail), age, and motion on within-ToM network correlation: effect of FB-Group (pass (n=30) vs. fail (n=15)): b=-.70, t=-2.06, p=.046, effect of age: b=.73, t=4.4, p<.0005, effect of motion: b=-.34, t=-2.7, p=.009). This group difference becomes marginal when response inhibition (DCCS summary score) is additionally included in the regression (effect of FB-Group (pass (n=30) vs. fail (n=15)): b=-.64, t=-1.80, p=.079, effect of age: b=.74, t=4.4, p<.0001, effect of motion: b=-.33, t=-2.5, p=.02, NS effect of DCCS (response inhibition): b=-.08, t=-.59, p=.56). There was no difference in across-network correlation between these two groups (Passers M(SE)=.04(.02), Failers M(SE)=.03(.03); linear regression testing for effects of FB-Group (pass vs. fail), age, and motion on across-network correlation: NS effect of FB-group (pass (n=30) vs.

fail (n=15)): b=.51, t=1.2, p=.23, NS effect of age: b=-.29, t=-1.4, p=.16, NS effect of motion: b=-.004, t=-.02, p=.98). See Figure 3a-b.

Figure 3

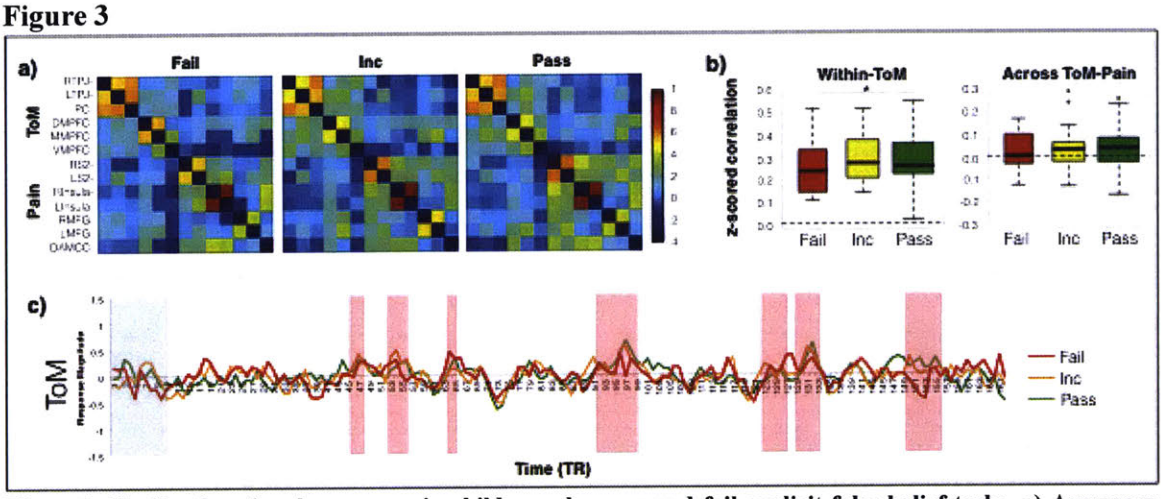


Figure 3. Similar functional responses in children who pass and fail explicit false-belief tasks. a) Average z-scored correlation matrices for three to five-year-old children who pass (n=30), fail (n=15), or perform inconsistently on (n=20) explicit false-belief tasks. Regions are in the same order along the x- and y- axes. **b)** Boxplots of z-scored correlation values within-ToM and across-ToM-Pain brain regions, based on false-belief task performance. Asterisk denotes a significant effect of false-belief task group (pass vs. fail) in a linear regression that also includes age and amount of motion (number of artifact timepoints) as covariates ($p<.05$); this group effect becomes marginal when additionally including a measure of response inhibition (DCCS). **c)** Average timecourse of response in the ToM network for false-belief passers (green), failers (red), and inconsistent performers (orange), during viewing of "Partly Cloudy."[61]

Reverse Correlation Analysis

Reverse correlation analyses are data-driven analyses used to identify events (>4 sec) in a continuous naturalistic stimulus that evoke reliable positive hemodynamic responses in the same region across subjects[41]. Here, we first use reverse correlation analyses to identify events that drive activity in ToM and pain brain regions, and subsequently test for developmental change in the magnitude of response to these events in children. As a first step, we successfully replicated previous results that responses in the fusiform gyrus are driven by face stimuli[41]; see Supplementary Fig. 4. Given these analyses have not yet been applied to pediatric data, this replication enabled us to be more confident in our analysis stream, the use of group regions of interest (ROIs), as opposed to individually defined ROIs, and the quality of our fMRI data (especially in young children, using a relatively short movie).

We applied reverse correlation analyses to the average response timecourses in the ToM network and pain matrix in adult participants. Because the inter-region correlation analysis suggested that ToM and pain regions comprise two functionally distinct networks by age three, we calculated the average timecourse across ROIs within each network. After identifying events based on the timecourse data from ToM and pain networks in adults, we extracted the response magnitude of each event from all child participants (see Methods). This analysis was used to determine 1) which events in the movie elicit the highest responses from ToM and pain regions in adults, 2) whether responses in ToM and pain regions in three-year-old children are driven by the same

events that drive corresponding responses in adults, and 3) the extent to which the responses to these events changes with age or ToM development in childhood.

In adults, the reverse correlation analysis produced seven theory of mind events (68s total, M(SD) length 9.7(4.2)s) and twelve pain events (86s total, M(SD) length 7.2(4.7)s); see Figure 4. All seven peak "mind" events depict (changes in) the characters' beliefs, desires, and/or emotions: e.g. Gus is afraid that Peck will abandon him, Peck is embarrassed when Gus catches him gazing at another cloud. A majority of the "body" events (8/12) depict the characters' physical pain (e.g. Peck being bitten by a crocodile) or transformations to the body (e.g. electricity changing a ball of cloud into a ram). See online manuscript (http://rdcu.be/IRh8) for versions of Figure 4 and Supplementary Fig. 5 that include thumbnails images of events. Additionally see Supplementary Table 3 for full descriptions of events and timing and duration information, Supplementary Fig. 6 for a replication in an independent sample of adults, and Supplementary Fig. 7 for correspondence between these events and previously used hand-coded events.

Figure 4

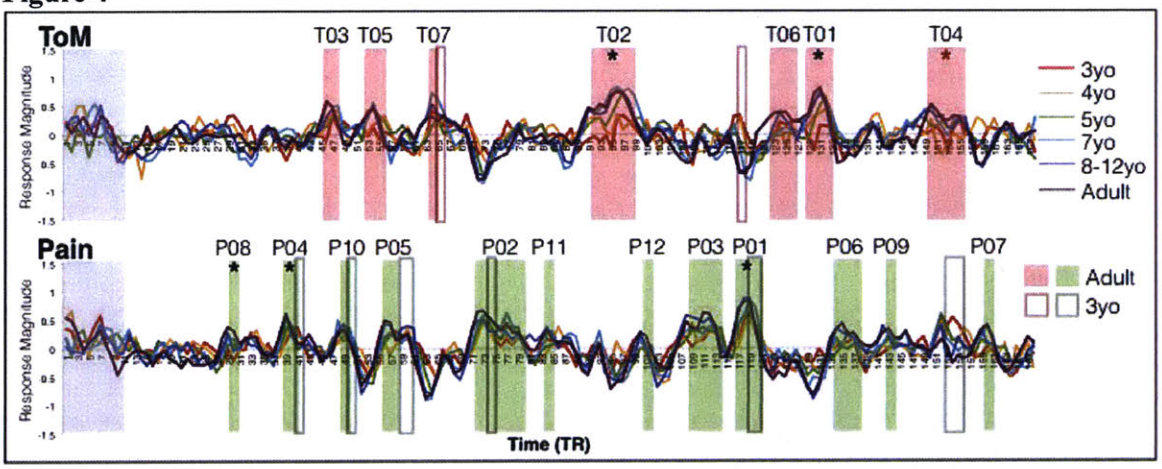


Figure 4. Reverse Correlation Analysis. a. The average timecourse per age group for the ToM network (top) and Pain matrix (bottom), during viewing of "Partly Cloudy"[61]. Each timepoint along the x-axis corresponds to a single TR (2 seconds); the entire movie was 168 TRs (<6 minutes). Shaded blocks show timepoints identified as ToM (red) and Pain (green) events in a reverse correlation analysis conducted on adults (n=33); timepoints within the grey block correspond to the opening logos of the movie and were not analyzed. Dark red and green borders show timepoints identified as ToM and pain events, respectively, in three-year-old children (n=17). Event labels (e.g. T01, P01) indicate ranking of average magnitude of response in adults. Asterisks indicate significant positive correlations between peak magnitude of response and age (continuous variable; black) and ToM behavioral score (continuous variable; red), after correcting for multiple comparisons (age: 19 ToM/Pain events, α=.0026; ToM: 7 ToM events, α=.007). See online manuscript for thumbnail images of events (http://rdcu.be/IRh8).

The timepoints that exceeded baseline for ToM and pain networks were almost entirely non-overlapping, with the exception of a single timepoint (2s). This timepoint is the last timepoint of event T05, and the first timepoint of event P05; the response magnitude of both networks is significantly above baseline during this timepoint; see Figure 4a. This extent of overlap is significantly less than that that would occur by chance (5/1000 random timecourse permutations with the same number and duration of ToM and Pain events have at most one timepoint of

overlap; p=.005), and is present despite not regressing out a global signal from the timecourses of each network. See Supplementary Note 1 for a similar overlap analysis between face and ToM, and face and pain, events. These results converge with previous evidence for a similar functional division when participants read short verbal narratives, or when participants endogenously shift their attention to bodily versus mentalistic aspects of one movie or picture[2,3,4,5,38].

The average timecourse in ToM and pain regions in children was highly correlated with that of adults (Pearson correlation tests between adult average timecourse and child average timecourse, TRs 11:168, for each child age bin: ToM: 3yo: r=.28, 4yo: r=.31, 5yo: r=.60, 7yo: r=.72, 8-12yo: r=.82 (all ps <.0005; Bonferroni correction for multiple comparisons α =.01, for five age bins); Pain: 3yo: r=.60, 4yo: r=.56, 5yo: r=.73, 7yo: r=.83, 8-12yo: r=.89 (all ps<$1.0\text{x}10^{-13}$; α =.01); see Supplementary Table 4). Nevertheless, we observed evidence of developmental change. Among children, three pain events (P01, P04, P08) and two ToM events (T01, T02) evoked significantly greater responses with age (spearman partial correlation tests, including motion as a covariate (n=122); Pain: ps<.002, r_ss>.29; ToM: ps<.0026, r_ss>.28; Bonferroni correction for multiple comparisons α =.0026, correcting for 19 events/tests). The two ToM events that showed greater responses with age are longer events that involve multiple and more complicated mental states (Supplementary Table 3). Responses in ToM regions during a third ToM event (T04) were significantly positively correlated with ToM score, controlling for age and motion (linear regression testing for effects of ToM score, age, and motion on T04 response magnitude (n=122): effect of ToM score: b=.4, t=2.98, p=.0035, NS effect of age: b=-.14, t=-.99, p=.32, NS effect of motion: b=-.07, t=-.77, p=.45; MC α =.007, correcting for 7 ToM events/tests); see Figure 1b. Response magnitude during ToM events did not differ significantly between children who pass and fail explicit false-belief tasks (all ps>.08; linear regressions testing for effects of FB-Group (pass (n=30) vs. fail (n=15)), including age and motion as covariates); see Figure 3c.

We next examined just the youngest children. As reported above, the overall timecourse of each network in three year olds (n=17) was highly correlated with the average adult timecourses (Pearson correlation test between adult average timecourse and average three year old timecourse, TRs 11:168: ToM: r=.28 p=.00046; Pain: r=.60, p<$1.0\text{x}10^{-15}$). Reverse correlation analysis conducted on the three year olds' data alone identified four of the twelve pain events and one of the seven ToM events discovered in the adult sample. These events correspond to a subset of the timepoints that were identified as ToM or pain events in three year olds (Pain: 14/32s, ToM: 4/8s). Interestingly, 8 of the remaining 18s identified as a pain event in three-year-old children corresponds to a ToM event (T04) in adults, and the remaining 4s identified as a ToM event corresponds to a pain event (P01) in adults (Figure 4). The remaining 10s identified as pain events occurred immediately after adult pain event timepoints.

Relating Functional Maturity to Inter-region Correlations

We tested whether the functional maturity (i.e. similarity to adults) of a child's movie-driven timecourse was related to the inter-region correlations measuring the child's network properties. Functional maturity was quantified by correlating each child's timecourse with the average adult timecourse. We found that the maturity of the movie-driven timecourse in both ToM and Pain networks was predicted by the anti-correlation of regions across networks (linear regressions testing for effects of across-network correlation, within- network correlation, age, and motion on

functional maturity measure (n=122): ToM: effect of across-network correlation: b=-.4, t=-5.5, $p=2.2x10^{-7}$, NS effect of within-ToM correlation: b=.1, t=1.5, p=.14, effect of age: b=.4, t=5.3, $p=5.7x10^{-7}$, NS effect of motion: b=-.1, t=-1.5, p=.14; Pain: effect of across-network correlation: b=-.51, t=-7.4, $p=2.8x10^{-11}$, NS effect of within-Pain correlation: b=.13, t=1.9, p=.06, effect of age: b=.3, t=4.6, $p=1.3x10^{-5}$, NS effect of motion: b=-.08, t=-1.3, p=.2); see Figure 5.

Figure 5

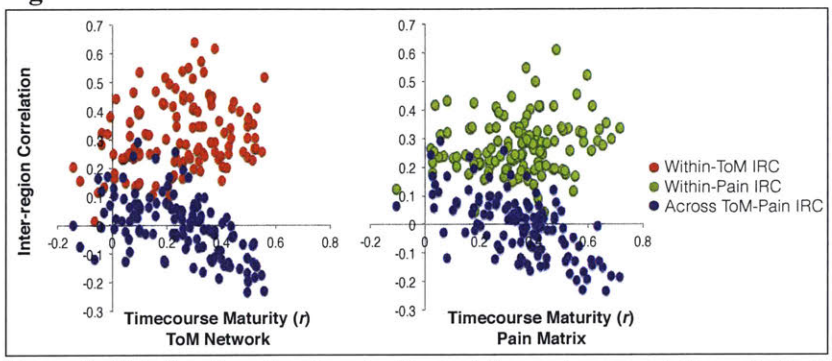


Figure 5. Relating functional maturity to inter-region correlations. In both networks, timecourse maturity (i.e. how correlated each child's timecourse is to the average adult timecourse (pearson's *r*), x-axis) is predicted by the extent to which the responses in ToM and Pain networks are anti-correlated (z-scored correlation values, y-axis). Both scatterplots show these values for all children (n=122).

That is, for children whose regions across the two networks showed more distinct responses, the average response within each network to the movie was more adult-like. This same pattern did not hold for within-network correlations. Greater within-network correlations were strongly associated with age, but not with timecourse maturity (linear regressions testing for effects of timecourse maturity, age, and motion on within-network correlations (n=122): ToM: NS effect of functional maturity: b=.05, t=.48, p=.63, effect of age: b=.3, t=2.5, p=.01, effect of motion: b=-.3, t=-3.56, p=.0005; Pain: NS effect of functional maturity: b=.07, t=.61, p=.55, marginal effect of age: b=.2, t=1.96, p=.05, NS effect of motion: b=-.005, t=-.05, p=.96). Additionally, while having an adult-like ToM timecourse was positively correlated with ToM behavior (spearman partial correlation test, including motion as a covariate (n=122): $r_s(119)=.54$, $p=1.3x10^{-10}$), this relationship did not remain significant in a regression including age as an additional predictor (linear regression testing for effects of age, ToM score, and motion on functional maturity of ToM timecourse (n=122): effect of Age: b=.5, t=4.2, p<.00005, NS effect of ToM score: b=.1, t=1.2, p=.3, effect of motion: b=-.15, t=-2.1, p=.04). Functional maturity of the ToM timecourse did not differ based on explicit false-belief task performance (linear regression testing for effects of FB-Group (pass vs. fail), age, and motion on functional maturity of ToM timecourse: NS effect of group (pass (n=30) vs. fail (n=15)): b=-.12, t=-.35, p=.73, effect of age: b=.49, t=2.9, p=.005, effect of motion: b=-.43, t=-3.4, p=.002). Thus, among children, having functionally mature, task-driven responses is predicted by a child's anti-correlated responses in regions of the ToM and Pain networks.

Discussion

Children's brains and their cognitive abilities undergo dramatic development in early childhood. In social cognition, for example, young children develop a remarkably sophisticated understanding of others' desires, thoughts and emotions, as distinct from their bodily reflexes, pains, and illnesses; much of this development occurs before children begin formal schooling at six years old[45,46,47]. Although brain regions involved in ToM have been extensively studied in

adults, adolescents, and older children, fMRI experiments present serious obstacles for very young children. By using a short, engaging and naturalistic movie stimulus, we were able to collect functional data from a large sample of children (n=122), including 65 children between the three and six years of age. The movie stimulus, Pixar's "Partly Cloudy," depicts multiple events that focus on two aspects of the main characters (a cloud named Gus, and his stork friend Peck): their bodily sensations (often physical pain) and their mental states (beliefs, desires, and emotions). We measure developmental change in cortical networks recruited for reasoning about bodies (the pain matrix) and minds (the theory of mind network), and relate development in the ToM network to behavioral changes in theory of mind – bridging the gap between previous fMRI studies in older children, and a large behavioral literature on early ToM development.

The first goal of this project was to measure developmental change in the pain matrix and theory of mind network. A key result emerged from multiple different analysis approaches: a core aspect of development in the social brain is the differentiation of spontaneous cortical responses to depictions of others' bodies versus minds. First, anti-correlations between the ToM and pain networks showed particularly dramatic change with age: regions in these two networks were uncorrelated in three year olds, but robustly anti-correlated in older children and adults. This anti-correlation predicted the maturity (i.e. similarity to adults) of each network's timecourse of response evoked by the movie. Second, while activity in ToM and pain networks in adults is driven by non-overlapping mentalistic and bodily events, respectively, in three year olds some events led to increased activity in the opposite network: the adult pain event P01 elicited activity in the ToM network, and the adult ToM event T04 elicited activity in the pain network of three-year-old children. These results are in line with previous evidence that functionally selective brain regions respond less to non-preferred categories with age,[20,21,48,49] and suggest that development of functionally selective brain regions for reasoning about others' internal states involves increasingly accurate application of specific neural resources (i.e. distinct groups of brain regions) to specific inputs (events depicting others' mental states versus physical sensations).

Almost all previous publications of timecourse data in young children describe analyses of resting state data: fMRI data collected while participants are not performing any particular cognitive task, or in some cases, while participants are asleep[50]. One advantage of measuring inter-region correlations during a movie, as we did here, is that children's psychological state (e.g. attention, anxiety, alertness) is likely more similar, across ages. On the other hand, a disadvantage is that we cannot distinguish between intrinsic and task-driven contributions to the inter-region correlations[51]. For example, the development of anti-correlations between ToM and pain networks may reflect a combination of both intrinsic changes in network structure, and increasing functional selectivity of the movie-driven response in individual regions[52]. Future studies could tease apart contributions of intrinsic and task-driven connectivity by collecting both resting-state and functional task data from the same child; however, for three-year-old children any additional data collection within a session would be challenging.

The second goal of this project was to ask how change in the ToM network relates to children's theory of mind cognitive abilities. All children were asked questions about other people's actions, beliefs, desires, expectations, and moral blameworthiness. Within this set of questions, six questions focused specifically on predicting and explaining actions based on false beliefs. The

transition from failure to success on the false-belief task has sometimes been interpreted as evidence of discontinuity in development around age four years: the emergence of a new theory, or cognitive mechanism, that did not exist earlier[12-14]. A second possibility is that changes in executive function (e.g. response inhibition) unmask children's previously existing ToM[31-33]. A third possibility is that children's theory of mind itself undergoes continuous and gradual development, from relatively simple concepts of perceptions and goals in two year olds to a sophisticated understanding of negligence and irony in early adolescence[34-37,53]. Each of these possibilities makes different predictions for the patterns of neural data we measured here. Unlike any previous fMRI study of ToM, our sample included a substantial number of children who systematically failed explicit false-belief tests. This enabled us to test for signatures of neural responses that predict improved performance on false-beliefs tasks, in addition to ToM reasoning more generally.

Our data were most inconsistent with the prediction of a robust discontinuity in response, associated with the transition from failure to success on explicit false-belief tasks. In the profiles of neural responses, we saw no major discontinuity when children begin to systematically pass false-belief tasks. Brain regions involved in ToM in adulthood already constitute a distinct network in three-year-old children, which gradually becomes more integrated and distinct from other networks over the next decade. Similarly, the timecourse of response in the ToM network in response to a social movie is strongly positively correlated, even between three year olds and adults. The timecourse and peak event responses show gradual continuous development over childhood. Focusing specifically on three- to five-year-old children, the neural responses to social movies in children who systematically fail versus pass explicit false-belief tasks were similar: there were no differences in the magnitude of response to the seven ToM events identified using reverse correlation analyses, and no difference in the extent of anti-correlation of the responses in ToM and pain networks. Consistent with recent evidence that false-belief passers have increased structural connectivity between ToM brain regions, compared to failers[54], we find that passing false-belief tasks was associated with increased functional correlations among regions in the ToM network, but this group difference became marginal when taking response inhibition abilities into account, and the same neural measure was also associated with age in the full sample.

Our data were partially consistent with the prediction that spontaneous processing of others' mental states within domain-specific regions for ToM is similar, regardless of performance on explicit false-belief tasks. Research in adults suggests that the same ToM brain regions are recruited to reason about mental state content, regardless of whether the stimulus is verbal or nonverbal, instructed or spontaneous.[19,38,55,56] Spontaneously generating mentalistic descriptions of actions is a precursor of performance on explicit tasks[57], and is correlated with cortical thinning of ToM regions in adults[58]. In the current study, three-year-old children who systematically fail false-belief tasks nevertheless recruited ToM brain regions at similar times in the movie and as a distinct network from the pain matrix. On the other hand, we did observe significant change within ToM brain regions, and in the dissociation between ToM and pain networks, which is not predicted by the view that explicit ToM tasks measure change in domain general performance limitations.

Overall, our results seem most consistent with the prediction that a distinct neural response to others minds versus bodies is already beginning to develop well before children explicitly pass false-belief tasks, and continues to develop well after[7,8,47]. For example, for one event in the movie, the magnitude of response in the ToM network correlated with the child's score on the full ToM battery (not limited to false belief items). This event (T04) shows Peck donning protective football gear in front of Gus. In context, this event depicts Gus revising previous beliefs and emotions (because Gus believed that Peck had abandoned him, Gus had been furious and devastated; once Peck shows Gus the helmet and pads, Gus realizes that Peck has not abandoned him and indeed never intended to abandon him, and Gus feels happy and relieved). Increased activity in ToM regions during this event may reflect children's improved ability to consider the relevance of the current event for (past) beliefs or emotions that are not explicitly depicted[59].

These fMRI results are thus consistent with evidence in developmental psychology for slow, continuous development of theory of mind. In individual children, the transition from failing to passing explicit false-belief tasks occurs gradually and noisily: children who begin to answer explicit false-belief questions correctly often subsequently fall back to incorrect responses[57]. Improvement is boosted by explicit explanatory practice and feedback over a relatively long period of time. The noisiness of development is visible in the current dataset: twenty children answered three or four out of six explicit false-belief questions correctly, within a single testing session. Also, mastering explicit false-belief tasks is not equivalent to having a fully mature theory of mind[60]; older children are still learning to infer hidden emotions[34], discriminate degrees of moral blameworthiness[53], and understand non-literal speech like sarcasm and irony[37]. On this view false-belief task performance is likely just one step along a long trajectory of increasingly sophisticated understanding of other minds.

In sum, we report evidence that when people spontaneously watch an animated movie evoking the internal states of others, distinct networks of cortical regions are recruited for events that make salient internal states of the mind versus of the body. These networks are already functionally distinct in three-year-old children, but show increasing within-network and decreasing across-network correlations throughout childhood. The anti-correlation of the two networks strongly predicts the maturity of each network, in response to the movie. Specific peak events within the movie evoke activity that increases with age, and with theory of mind reasoning ability. On the other hand, the most famous milestone in ToM behavioral development, passing explicit false-belief tasks, does not correspond with a discontinuity in the neural basis for reasoning about the minds of others.

Methods

Participants

122 3.5-12 year-old children (M(SD)=6.7(2.3); 64 females) participated in the study. 110 children were right-handed and 3 were ambidextrous (as indicated by parent or legal guardian). This sample includes 65 children under the age of six (M(SD)=4.82(.81) years; 34 females; 54 RH/3 Ambi); this subset of children were used to test for neural differences between children who pass (n=30; M(SD)=5.2(.70); 15 females; 26 RH/2 Ambi) and fail (n=15; M(SD)=4.08(.56); 6 females; 11 RH/4 LH) false-belief tasks. 20 children in this subset responded inconsistently to false-belief tasks (M(SD)=4.75(.73); 13 female; 17 RH/1 Ambi). An additional 19 children were

recruited to participate and excluded from all analyses for not completing or participating in the study (n=12), language delays (n=2), and excessive motion during the fMRI scan (n=5; see fMRI Data Analysis for details). 33 adult participants (ages 18-39 years; M(SD)=24.8(5.3); 20 females; 32 RH/1 LH) additionally participated in the fMRI portion of the study. Child and adult participants were recruited from the local community. All adult participants gave written consent; parent/guardian consent and child assent was received for all child participants. Recruitment and experiment protocols were approved by the Committee on the Use of Humans as Experimental Subjects (COUHES) at the Massachusetts Institute of Technology.

FMRI Stimuli

Participants watched a silent version of "Partly Cloudy,"[61], a 5.6-minute animated movie.[38] A short description of the plot can be found online (https://www.pixar.com/partly-cloudy#partly-cloudy-1). Previous research suggests that pediatric populations move significantly less during fMRI scans using movie stimuli[62]. The stimulus was preceded by 10s of rest, and participants were instructed to watch the movie and remain still. Participants aged five and older completed additional tasks prior to viewing this stimulus; these tasks largely involved listening to (children) or reading (adults) stories.

FMRI Data Acquisition

Prior to the scan, child participants completed a mock scan in order to become acclimated to the scanner environment and sounds, and to learn how to stay still. Children were given the option to hold a large stuffed animal during the fMRI scan in order to feel calm and to prevent fidgeting. An experimenter stood by child participants' feet, near the entrance of the MRI bore, to ensure that the participant remained awake and attentive to the movie. If this experimenter noticed participant movement, she placed her hand gently on the participant's leg, as a reminder to stay still.

Whole-brain structural and functional MRI data were acquired on a 3-Tesla Siemens Tim Trio scanner located at the Athinoula A. Martinos Imaging Center at MIT. Children under age five years used one of two custom 32-channel phased-array head coils made for younger (n=3, M(SD)=3.91(.42) years) or older (n=28, M(SD)=4.07(.42) years) children[63]; all other participants used the standard Siemens 32-channel head coil. T1-weighted structural images were collected in 176 interleaved sagittal slices with 1mm isotropic voxels (GRAPPA parallel imaging, acceleration factor of 3; adult coil: FOV: 256mm; kid coils: FOV: 192mm). Functional data were collected with a gradient-echo EPI sequence sensitive to Blood Oxygen Level Dependent (BOLD) contrast in 32 interleaved near-axial slices aligned with the anterior/posterior commissure, and covering the whole brain (EPI factor: 64; TR: 2s, TE: 30 ms, flip angle: 90°). As participants were initially recruited for different studies, there are small differences in voxel size and slice gaps across participants (3.13 mm isotropic with no slice gap (n=5 adults, n=3 7yos, n=20 8-12yo); 3.13 mm isotropic with 10% slice gap (n=28 adults), 3 mm isotropic with 20% slice gap (n=1 3yo, n=3 4yo, n=2 7yo, n=1 9yo); 3 mm isotropic with 10% slice gap (all remaining participants)); all functional data were subsequently upsampled in normalized space to 2mm isotropic voxels. Prospective acquisition correction was used to adjust the positions of the gradients based on the participant's head motion one TR back[64]. 168 volumes were acquired in each run; children under age five completed two functional runs, while older participants

completed only one run. For consistency across participants, only the first run of data was analyzed. Four dummy scans were collected to allow for steady-state magnetization.

FMRI Data Analysis

FMRI data were analyzed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm)[65] and custom software written in Matlab and R. Functional images were registered to the first image of the run; that image was registered to each participant's anatomical image, and each participant's anatomical image was normalized to the Montreal Neurological Institute (MNI) template. This enabled us to use group regions of interest (ROIs) and hypothesis spaces created in adult datasets, and to directly compare responses between child and adult participants. Previous research has suggested that anatomical differences between children as young as seven years are small relative to the resolution of fMRI data, which supports usage of a common space between adults and children of this age (for similar procedures with children under age seven, see[66,21,67]; for methodological considerations, see[68]). Registration of each individual's brain to the MNI template was visually inspected, including checking the match of the cortical envelope and internal features like the AC-PC and major sulci. All data were smoothed using a Gaussian filter (5mm kernel).

Artifact timepoints were identified via the ART toolbox (https://www.nitrc.org/projects/artifact_detect/)[69] as timepoints for which there was 1) more than 2mm composite motion relative to the previous timepoint or 2) a fluctuation in global signal that exceeded a threshold of three standard deviations from the mean global signal. Participants were dropped if one-third or more of the timepoints collected were identified as artifact timepoints; this resulted in dropping five child participants from the sample (see Participants). Number of artifact timepoints differed significantly between child and adult participants (Child (n=122): M(SD)=10.5(10.6), Adult (n=33): M(SD)=2.8(4), Welch two-sample t-test: t(137.7)=6.49, p<.000001). Among children, number of motion artifact timepoints was not correlated with age (spearman correlation test (n=122): $r_s(120)=.02$, p=.86) or ToM score (kendall tau correlation test (n=122): $r_k(120)=-.005$, p=.94). Number of artifact timepoints did not differ between young (3-5 year old) children based on false-belief task performance (linear regression tests for effect of FB-Group on number of motion artifact timepoints: NS effect of FB-group (Pass (n=30) vs. Fail (n=15)): b=-.04, t=-.12, p=.9; NS effect of FB-group (Pass (n=30), Inc (n=20), or Fail (n=15)): bs<.05, ps>.9) or response inhibition (linear regression test for effect of DCCS on number of motion artifact timepoints (n=64): NS effect of DCCS summary score: b=.16, t=1.18, p=.25). See Supplementary Fig. 8 for visualization of the amount of motion per age group. Despite amount of motion being matched across children, and therefore likely not driving developmental effects within the child sample, we include number of motion artifact timepoints as a covariate in all analyses. Number of artifact timepoints is highly correlated with measures of mean translation, rotation, and distance ($rs>.8$). Because this measure is not normally distributed, spearman correlations were used when including amount of motion as a covariate in partial correlations.

Region of interest (ROI) analyses were conducted using group ROIs. ToM and pain matrix group ROIs were created in an independent group of adults (n=20), scanned by Evelina Fedorenko and colleagues. These data were preprocessed and analyzed with procedures identical to those used for participants in the current study. Reverse correlation analyses were conducted in this separate group of adults, using 10mm group ROIs surrounding peaks reported in previous publications

(ToM regions[70]; Pain matrix[71]). Seven ToM and nine pain events were identified (ToM: 60s total, M(SD) length: 8.6(4.6)s, Pain: 66s total, M(SD) length: 7.3(4.4)s). We subsequently used a general-linear model to analyze BOLD activity of these participants as a function of condition, using these events. Second-level random effects analyses were used to examine the group-level response to Mental > Pain and Pain> Mental ($p<.001$, $k=10$, uncorrected). We then drew 9mm spheres surrounding the peak activation in each region, to create new group ROIs that were tailored to the stimulus, but defined in an independent sample of adults (see Supplementary Fig. 1 and Supplementary Table 2 for more information on all group ROIs, and Supplementary Fig. 7 for details of the convergence between events across the two adult samples and ROIs).

All timecourse analyses were conducted by extracting the scaled preprocessed timecourse from each voxel per group ROI. We applied nearest neighbor interpolation over artifact timepoints (for methodological considerations on interpolating over artifacts before applying temporal filters, see[72,73]), and regressed out two kinds of nuisance covariates to reduce the influence of motion artifacts: 1) motion artifact timepoints, and 2) five principle component analysis (PCA)-based noise regressors generated using CompCor within individual subject white matter masks[74]. White matter masks were eroded by two voxels in each direction, in order to avoid partial voluming with cortex. CompCor regressors were defined using scrubbed data (e.g. artifact timepoints were identified and interpolated over prior to running CompCor).

For inter-region correlation analyses only, we additionally regressed out the raw timecourse extracted from bilateral primary motor cortex (M1). Primary motor cortex ROIs were 10mm spheres drawn around peak coordinates generated with Neurosynth (http://neurosynth.org/; search term: "primary motor," forward inference from 273 studies; coordinates: [38,-24,58], [-38,-20,58]). These ROIs are included in the expanded inter-region correlation analysis shown in Supplementary Figure 4; the bilateral M1 timecourse was not regressed out for this supplementary analysis. However, because this analysis showed that the within-M1 inter-region correlation increases with age among children, we regressed the bilateral M1 timecourse from the ToM and Pain timecourses for the inter-region correlation analyses reported in the main text, to ensure that the age effects in the ToM and pain networks are above and beyond developmental effects present in regions like primary motor cortex, and that within-network correlations are not falsely inflated by commonalities in signal fluctuation across the brain.

The residual timecourses were then high-pass filtered with a cutoff of 100 seconds. Timecourses from all voxels within an ROI were averaged, creating one timecourse per group ROI, and artifact timepoints were subsequently excluded (NaNed).

In inter-region correlation analyses, each ROI timecourse was correlated with every other ROI's timecourse, per subject, and these correlation values were Fisher z-transformed. Within-ToM correlations were the average correlation from each ToM ROI to every other ToM ROI, within-Pain correlations were the average correlation from each Pain ROI to every other Pain ROI, and across-network correlations was the average correlation from each ToM ROI to each Pain ROI. This procedure is similar to that used by[42]. In order to test for developmental change in within- and across-network correlations, we conduct linear regressions to test for 1) significant differences between adults and children, in regressions that include group (child vs. adult) and number of artifact timepoints as predictors, and 2) significant effects of age (as a continuous

variable), ToM performance, and number of artifact timepoints among children, and 3) significant group differences between children who pass and fail explicit false-belief tasks, including number of artifact timepoints and age as predictors. In order to test whether ToM and pain networks are coherent and specialized early in childhood, we use t-tests to compare within- versus across-network correlations in three-year-old children (n=17). Within- and across-network correlation measures were normally distributed (ps>.22, one-sample Kolmogorov-Smirnov tests), and variance in within-ToM, within-Pain, and across-network correlations did not differ across children and adults, or false-belief passers vs. failers (F-tests to compare two variances: children (n=122) vs. adults (n=33): Fs(32,121)>1.1, ps>.66; pass (n=30) vs. fail (n=15): Fs(14,29)>.78, ps>.65).

Initial reverse correlation analyses were conducted on adult participants only. Each ROI timecourse was z-normalized, and timecourses within each network were averaged across ROIs, resulting in one timecourse for face regions, ToM regions, and the pain matrix per adult participant. Except for the first ten timepoints (5 TRs rest, followed by 5 TRs of the movie introduction (Disney castle and Pixar logos)), the residual signal values across adult subjects for each timepoint were tested against baseline (0) using a one-tailed t-test. This procedure is similar to that used by[41]. Events were defined as two or more consecutive significantly positive timepoints within each network. Events were rank-ordered according to the average magnitude of response to the peak timepoint in adults, and labeled according to the ordering (e.g. event T01 is the ToM event that evoked the highest magnitude of response in the ToM network).

In adults, we conducted an overlap analysis to determine whether the number of timepoints labeled as both ToM and pain events was statistically fewer than would occur by chance. We constructed 1000 permutations of ToM and pain timecourses, which had the same number and duration of events. The constructed timecourses were 158 TRs in length (the experiment was 168 TRs; the first 10 TRs were excluded from the reverse correlation analysis because the movie started on TR 11). For each permutation, we randomly scrambled the order of ToM and pain events. We then filled in the timepoints between events with zeros, with a random proportion of zeros between events such that the total number of zeros was equal to the total number of non-event timepoints in the original timecourses (ToM: 125 TRs; Pain: 116 TRs). Events within a timecourse (ToM or Pain) necessarily had to be separated by at least one timepoint, since they would otherwise be counted as a single event. The first event of each timecourse could be preceded by zero zeros, and the last event of each timecourse could be followed by zero zeros. We calculated the sum of the number of timepoints tagged as ToM and pain events in each pair of permutations (ToM and pain timecourses), and subsequently calculated the proportion of permutations that resulted in the same or a smaller amount of overlap as observed in the reverse correlation analysis.

In order to test for developmental effects in the magnitude of response to ToM and pain events, we defined a peak timepoint per event as the timepoint with the highest average signal value in adults, and tested for significant correlations between magnitude of response at peak timepoints and age (as a continuous variable), including amount of motion (number of artifact timepoints) as a covariate. Because this measure of motion is non-normally distributed, we employed spearman correlations. For ToM regions only, we used linear regressions to test for a significant relationship between peak magnitude of response and theory of mind behavior (overall, in all

children), and to test whether responses at peak timepoints differed between children who pass (n=30) and fail (n=15) explicit false-belief tasks. Response magnitude at all peak events was normally distributed (all ps>.23, one-sample Kolmogorov-Smirnov test). Response magnitudes showed similar variance across false-belief task passers (n=30) and failers (n=15) (F-tests to compare two variances: all Fs(13,28)>.7, ps>.07), with the exception of one event (T03: F(14,28)=.30, p=.02). A permutation test was used to test for group differences in magnitude of response to this event[75]. We ran the reverse correlation analysis in three-year-old participants only (n=17), in order to examine response specificity at this young age, and to better understand developmental differences.

Finally, we tested whether the functional maturity of each child's timecourse responses (i.e. similarity to adults) was related to the inter-region network correlations. We calculated the Pearson correlation between each child's ToM timecourse (averaged across ROIs) and the average adult ToM timecourse; we similarly calculated the Pearson correlation between each child's pain matrix timecourse and the average adult pain matrix timecourse. The timecourses used for this analysis were the same as those used for the reverse correlation analysis, prior to z-normalization (TRs 11:168). We tested if, across children, this measure of functional maturity per network was correlated with within-network and across-network inter-region correlations, or related to ToM behavior. The neural maturity measure was normally distributed in both networks (ps>.29, one-sample Kolmogorov-Smirnov test). Variance in this measure in the ToM network did not differ between children who pass (n=30) and fail (n=15) false-belief tasks (F test to compare two variances: F(14,29)>1.00, ps>.95). We additionally calculated and report the Pearson correlation between the average timecourse of children in each age group and the average adult timecourse.

All of the analyses reported in this manuscript should be considered exploratory, not confirmatory, in that the analyses described here were not chosen prior to data collection, and data collection was not completed with this specific set of analyses in mind. While we deliberately chose this stimulus in order to measure neural responses in very young children (ages 3-4 years), older children visited the lab to participate in a different study, and additionally completed the protocol of the current study. We then recognized the opportunity of analyzing the full cross-sectional dataset, and chose analyses based on the stimulus (time series analyses seemed to utilize more data and be more sensitive than previous analysis methods[38]), and on recent relevant progress in the field.[42,76,77]

Behavioral Battery

After the scan, all children completed a behavioral task battery including (in order) an explicit theory of mind battery and a measure of nonverbal IQ (under 5 years: WPPSI block design[78], over 5 years: nonverbal KBIT-II[79]). Children under age seven then completed a computerized version of the Dimensional Change Card Sort task as a measuring of response inhibition. Performance on DCCS was captured using the summary score[44]; one child (an inconsistent FB task performer) failed to complete the DCCS task.

Explicit ToM Task and False-Belief Composite Score

All children completed a custom-made explicit ToM battery[21] (https://osf.io/G5ZPV/), which involved listening to an experimenter tell a story and answering prediction and explanation

questions that required reasoning about the mental states of the characters. Because this task was designed to capture variability in ToM reasoning across a wide age-range of children, the questions varied in difficulty. Easier items involved reasoning about similar and diverse desires, true beliefs, and emotion prediction; harder items included reasoning about false beliefs, moral blame-worthiness, and second-order false beliefs. Two analogous booklets were used; children ages 3-4 and 10-12 years old listened to a story about students finding snacks, and five-year-old children listened to a story about students finding books; 7-9 year-old-children were split among the books (snacks: n=16; books: n=33). Different booklets were used across children because children of different ages participated in different studies that all involved the current protocol. However, the two booklets were designed for repeated measures designs: analogous stories and questions across the two booklets had identical syntax, but different semantic content: one story was about helping children find their books, the other was about finding snacks. A previous study using the "finding books" booklet suggests the validity of this task to capture theory of mind development in children ages five to twelve years old[21]. These booklet tasks and instructions are available on the Open Science Framework (https://osf.io/G5ZPV/; DOI: 10.17605/OSF.IO/G5ZPV; ARK: c7605/osf.io/g5zpv).

Each child's performance on the ToM battery was summarized as the proportion of questions answered correctly, out of 24 matched items (14 prediction items and 10 explanation items). An additional two control items were asked to ensure that children were paying attention; after ensuring all children answered these questions correctly, these items were not further analyzed. Children ages 3-5 years old were also categorized based on their performance on a false-belief composite score based on six explicit false-belief questions (4 prediction, 2 explanation) within the ToM booklet. These six questions were chosen because they were canonical explicit false-belief questions describing changes in location or unexpected contents[11,12,80]. The composite score demonstrated acceptable reliability (Cronbach's alpha=.71). Children were categorized as explicit false-belief "passers" if they answered five or six out of six false-belief questions correct, "inconsistent performers" if they answered three or four questions correct, and "failers" if they answered zero to two questions correct.

We tested for significant correlations between age, DCCS and ToM, and for differences in these scores between children who pass and fail false-belief tasks. We use Kendall's rank correlation tau, given non-normal distributions of ToM score (Shapiro-Wilk normality test: $w=.9$, $p<.00001$) and DCCS score ($w=.75$, $p<.00001$), and given the frequency of ties in both of these measures.

Data Availability

The fMRI and behavioral data collected and analyzed during the current study are available through the OpenfMRI project (https://openfmri.org/; Link: https://www.openfmri.org/dataset/ds000228/ DOI: 10.5072/FK2V69GD88). The ToM behavioral battery is additionally available through OSF (https://osf.io/G5ZPV/; DOI: 10.17605/OSF.IO/G5ZPV; ARK: c7605/osf.io/g5zpv).

Acknowledgements

We thank the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT, Jorie Koster-Hale, Natalia Velez-Alicea, Mika Asaba, and Nir Jacoby for help with data collection, and Stefano Anzellotti, Dorit Kliemann, Julia Leonard, and Lindsey Powell

for helpful feedback and discussion. We thank Hyowon Gweon for development of the theory of mind behavioral battery, and Todd Thompson for helping to make the data available. We thank members of the Fedorenko lab for providing the data for the replication experiment. In particular, Alex Paunov and Zach Mineroff led the data collection effort, with help from Caitlyn Hoeflin, Amaya Arcelus, Brianna Pritchett, Idan Blank, and Cara Borelli. We also gratefully acknowledge support of this project by a NSF Graduate Research Fellowship (#1122374 to HR), and an NSF CAREER award (#095518 to RS), NIH R01-MH096914-05, a Middleton Chair grant (RS), and support from the David and Lucile Packard Foundation (#2008-333024 to RS).

References

1. Adolphs, R. The Social Brain: Neural Basis of Social Knowledge. *Annu. Rev. Psychol.* **60,** 693–716 (2009).
2. Lombardo, M. V. *et al.* Shared neural circuits for mentalizing about the self and others. *Journal of Cognitive Neuroscience* **22,** 1623–1635 (2010).
3. Bruneau, E. G., Pluta, A. & Saxe, R. Distinct roles of the 'shared pain'and "theory of mind"networks in processing others' emotional suffering. *Neuropsychologia* **50,** 219–231 (2012).
4. Morelli, S. A., Rameson, L. T. & Lieberman, M. D. The neural components of empathy: predicting daily prosocial behavior. *Social Cognitive and Affective Neuroscience* **9,** 39–47 (2014).
5. Spunt, R. P., Kemmerer, D. & Adolphs, R. The neural basis of conceptualizing the same action at different levels of abstraction. *Social Cognitive and Affective Neuroscience* nsv084 (2015).
6. Kanske, P., Böckler, A., Trautwein, F.-M. & Singer, T. Dissecting the social brain: Introducing the EmpaToM to reveal distinct neural networks and brain–behavior relations for empathy and Theory of Mind. *NeuroImage* **122,** 6–19 (2015).
7. Bloom, P. *Descartes' baby: How the science of child development explains what makes us human.* (Basic Books, 2009).
8. Wellman, H. M. *Making minds: How theory of mind develops.* (Oxford University Press, 2014).
9. Astington, J. W. & Edward, M. J. The development of theory of mind in early childhood. *Social Cognition in Infancy* **5,** 16 (2010).
10. Bartsch, K. & Wellman, H. M. *Children talk about the mind.* (Oxford university press, 1995).
11. Wellman, H. M., Cross, D. & Watson, J. Meta-analysis of theory-of-mind development: the truth about false belief. *Child Dev* **72,** 655–684 (2001).
12. Wimmer, H. & Perner, J. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *COGNITION* **13,** 103–128 (1983).
13. Perner, J., Leekam, S. R. & Wimmer, H. Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology* **5,** 125–137 (1987).
14. Callaghan, T. *et al.* Synchrony in the onset of mental-state reasoning: Evidence from five cultures. *Psychological Science* **16,** 378–384 (2005).
15. Knudsen, B. & Liszkowski, U. 18-Month-Olds Predict Specific Action Mistakes Through Attribution of False Belief, Not Ignorance, and Intervene Accordingly. *Infancy* **17,** 672–691 (2012).
16. Knudsen, B. & Liszkowski, U. Eighteen-and 24-month-old infants correct others in anticipation of action mistakes. *Dev Sci* **15,** 113–122 (2012).
17. Ohnishi, T. *et al.* The neural network for the mirror system and mentalizing in normally developed children: an fMRI study. *NeuroReport* **15,** 1483–1487 (2004).
18. Moriguchi, Y., Ohnishi, T., Mori, T., Matsuda, H. & Komaki, G. Changes of brain activity in the neural substrates for theory of mind during childhood and adolescence. *Psychiatry Clin. Neurosci.* **61,** 355–363 (2007).

19. Kobayashi, C., Glover, G. H. & Temple, E. Children's and adults' neural bases of verbal and nonverbal "theory of mind". *Neuropsychologia* **45,** 1522–1532 (2007).
20. Saxe, R. R., Whitfield-Gabrieli, S., Scholz, J. & Pelphrey, K. A. Brain regions for perceiving and reasoning about other people in school-aged children. *Child Dev* **80,** 1197–1209 (2009).
21. Gweon, H., Dodell-Feder, D., Bedny, M. & Saxe, R. Theory of Mind Performance in Children Correlates With Functional Specialization of a Brain Region for Thinking About Thoughts. *Child Dev* **83,** 1853–1868 (2012).
22. Decety, J., Michalska, K. J. & Akitsuki, Y. Who caused the pain? An fMRI investigation of empathy and intentionality in children. *Neuropsychologia* **46,** 2607–2614 (2008).
23. Decety, J., Michalska, K. J. & Kinzler, K. D. The contribution of emotion and cognition to moral sensitivity: a neurodevelopmental study. *Cerebral Cortex* **22,** 209–220 (2012).
24. Blakemore, S.-J. The social brain in adolescence. *Nat Rev Neurosci* **9,** 267–277 (2008).
25. Burnett, S., Sebastian, C., Kadosh, K. C. & Blakemore, S.-J. The social brain in adolescence: Evidence from functional magnetic resonance imaging and behavioural studies. *Neuroscience and Biobehavioral Reviews* **35,** 1654–1664 (2011).
26. Saxe, R. & Kanwisher, N. People thinking about thinking people: the role of the temporo-parietal junction in 'theory of mind'. *NeuroImage* **19,** 1835–1842 (2003).
27. Gallagher, H. L. & Frith, C. D. Functional imaging of 'theory of mind'. *Trends in Cognitive Sciences* **7,** 77–83 (2003).
28. Saxe, R. & Wexler, A. Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia* **43,** 1391–1399 (2005).
29. Carey, S. Conceptual change in childhood. (1985).
30. Gopnik, A., Meltzoff, A. N. & Bryant, P. *Words, thoughts, and theories.* **1,** (Mit Press Cambridge, MA, 1997).
31. Baillargeon, R., Scott, R. M. & He, Z. False-belief understanding in infants. *Trends in Cognitive Sciences* **14,** 110–118 (2010).
32. Scott, R. M. & Baillargeon, R. Early False-Belief Understanding. *Trends in Cognitive Sciences* (2017).
33. Carlson, S. M., Moses, L. J. & Hix, H. R. The role of inhibitory processes in young children's difficulties with deception and false belief. *Child Dev* **69,** 672–691 (1998).
34. Wellman, H. M. & Liu, D. Scaling of theory-of-mind tasks. *Child Dev* **75,** 523–541 (2004).
35. Filippova, E. & Astington, J. W. Further development in social reasoning revealed in discourse irony understanding. *Child Dev* **79,** 126–138 (2008).
36. Wellman, H. M., Fang, F. & Peterson, C. C. Sequential progressions in a theory-of-mind scale: longitudinal perspectives. *Child Dev* **82,** 780–792 (2011).
37. Peterson, C. C., Wellman, H. M. & Slaughter, V. The mind behind the message: Advancing theory-of-mind scales for typically developing children, and those with deafness, autism, or Asperger syndrome. *Child Dev* **83,** 469–485 (2012).
38. Jacoby, N., Bruneau, E., Koster-Hale, J. & Saxe, R. Localizing Pain Matrix and Theory of Mind networks with both verbal and non-verbal stimuli. *NeuroImage* **126,** 39–48 (2016).
39. Zaki, J., Wager, T. D., Singer, T., Keysers, C. & Gazzola, V. The anatomy of suffering: understanding the relationship between nociceptive and empathic pain. *Trends in Cognitive Sciences* **20,** 249–259 (2016).
40. Amodio, D. M. & Frith, C. D. Meeting of minds: the medial frontal cortex and social

cognition. *Nat Rev Neurosci* **7,** 268–277 (2006).
41. Hasson, U. Intersubject Synchronization of Cortical Activity During Natural Vision. *Science* **303,** 1634–1640 (2004).
42. Blank, I., Kanwisher, N. & Fedorenko, E. A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *Journal of Neurophysiology* **112,** 1105–1118 (2014).
43. Cantlon, J. F. & Li, R. Neural activity during natural viewing of Sesame Street statistically predicts test scores in early childhood. *Plos Biol* **11,** e1001462 (2013).
44. Zelazo, P. D. The Dimensional Change Card Sort (DCCS): a method of assessing executive function in children. *Nat Protoc* **1,** 297–301 (2006).
45. Schult, C. A. & Wellman, H. M. Explaining human movements and actions: Children's understanding of the limits of psychological explanation. *COGNITION* **62,** 291–324 (1997).
46. Schulz, L. E., Bonawitz, E. B. & Griffiths, T. L. Can being scared cause tummy aches? Naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology* **43,** 1124 (2007).
47. Cohen, E., Burdett, E., Knight, N. & Barrett, J. Cross-Cultural similarities and differences in person-body reasoning: Experimental evidence from the United Kingdom and Brazilian Amazon. *Cogn Sci* **35,** 1282–1304 (2011).
48. Carter, E. J. & Pelphrey, K. A. School-aged children exhibit domain-specific responses to biological motion. *Social Neuroscience* **1,** 396–411 (2006).
49. Cantlon, J. F., Pinel, P., Dehaene, S. & Pelphrey, K. A. Cortical Representations of Symbols, Objects, and Faces Are Pruned Back during Early Childhood. *Cerebral Cortex* **21,** 191–199 (2010).
50. Menon, V. Developmental pathways to functional brain networks: emerging principles. *Trends in Cognitive Sciences* 1–14 (2013). doi:10.1016/j.tics.2013.09.015
51. Simony, E. *et al.* Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications* **7,** (2016).
52. Chai, X. J., Ofen, N., Gabrieli, J. D. & Whitfield-Gabrieli, S. Selective development of anticorrelated networks in the intrinsic functional organization of the human brain. *Journal of Cognitive Neuroscience* **26,** 501–513 (2014).
53. Cushman, F., Sheketoff, R., Wharton, S. & Carey, S. The development of intent-based moral judgment. *COGNITION* **127,** 6–21 (2013).
54. Wiesmann, C. G., Schreiber, J., Singer, T., Steinbeis, N. & Friederici, A. D. White matter maturation is associated with the emergence of Theory of Mind in early childhood. *Nature Communications* **8,** 14692 (2017).
55. Gallagher, H. L. *et al.* Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia* **38,** 11–21 (2000).
56. Schneider, D., Slaughter, V. P., Becker, S. I. & Dux, P. E. Implicit false-belief processing in the human brain. *NeuroImage* 1–8 (2014). doi:10.1016/j.neuroimage.2014.07.014
57. Amsterlaw, J. & Wellman, H. M. Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Journal of Cognition and Development* **7,** 139–172 (2006).
58. Rice, K. & Redcay, E. Spontaneous mentalizing captures variability in the cortical thickness of social brain regions. *Social Cognitive and Affective Neuroscience* **10,** 327–334 (2015).

59. Lagattuta, K. H., Wellman, H. M. & Flavell, J. H. Preschoolers' understanding of the link between thinking and feeling: Cognitive cuing and emotional change. *Child Dev* **68,** 1081–1104 (1997).
60. Blijd-Hoogewys, E. & van Geert, P. L. Non-linearities in Theory-of-Mind Development. *Frontiers in Psychology* **7,** 1970 (2017).
61. Reher, K. (Producer), & Sohn, P. (Director). *Partly Cloudy* [Motion Picture]. United States: Pixar Animation Studios and Walt Disney Pictures (2009).
62. Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L. C. & Castellanos, F. X. Inscapes: A movie paradigm to improve compliance in functional magnetic resonance imaging. *NeuroImage* **122,** 222–232 (2015).
63. Keil, B. *et al.* Size-optimized 32-channel brain arrays for 3 T pediatric imaging. *Magn. Reson. Med.* **66,** 1777–1787 (2011).
64. Thesen, S., Heid, O., Mueller, E. & Schad, L. R. Prospective acquisition correction for head motion with image-based tracking for real-time fMRI. *Magn. Reson. Med.* **44,** 457–465 (2000).
65. Neuroimaging, W. T. C. F. SPM.
66. Cantlon, J. F., Brannon, E. M., Carter, E. J. & Pelphrey, K. A. Functional Imaging of Numerical Processing in Adults and 4-y-Old Children. *Plos Biol* **4,** e125–11 (2006).
67. Bedny, M., Richardson, H. & Saxe, R. 'Visual' Cortex Responds to Spoken Language in Blind Children. *Journal of Neuroscience* **35,** 11674–11681 (2015).
68. Burgund, E. D. *et al.* The Feasibility of a Common Stereotactic Space for Children and Adults in fMRI Studies of Development. *NeuroImage* **17,** 184–200 (2002).
69. Whitfield-Gabrieli, S., Nieto-Castanon, A. & Ghosh, S. Artifact Detection Tools (ART). *Cambridge, MA. Release version* **7,** 11 (2011).
70. Dufour, N. *et al.* Similar Brain Activation during False Belief Tasks in a Large Sample of Adults with and without Autism. *PLoS ONE* **8,** e75468 (2013).
71. Bruneau, E. G., Jacoby, N. & Saxe, R. Empathic control through coordinated interaction of amygdala, theory of mind and extended pain matrix brain regions. *NeuroImage* **114,** 105–119 (2015).
72. Carp, J. Optimizing the order of operations for movement scrubbing: Comment on Power et al. *NeuroImage* **76,** 436–438 (2013).
73. Hallquist, M. N., Hwang, K. & LUNA, B. The nuisance of nuisance regression: spectral misspecification in a common approach to resting-state fMRI preprocessing reintroduces noise and obscures functional connectivity. *NeuroImage* **82,** 208–225 (2013).
74. Behzadi, Y., Restom, K., Liau, J. & Liu, T. T. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* **37,** 90–101 (2007).
75. Wheeler, B. lmPerm: Permutation tests for linear models. *R package version* **1,** 1–2 (2010).
76. Wagner, D. D., Kelley, W. M., Haxby, J. V. & Heatherton, T. F. The Dorsal Medial Prefrontal Cortex Responds Preferentially to Social Interactions during Natural Viewing. *Journal of Neuroscience* **36,** 6917–6925 (2016).
77. Adolphs, R., Nummenmaa, L., Todorov, A. & Haxby, J. V. Data-driven approaches in the investigation of social perception. *Phil. Trans. R. Soc. B* **371,** 20150367 (2016).
78. Wechsler, D. Manual for the WPPSI-R. *New York: The Psychological Co* (1989).
79. Kaufman, A. S. KBIT-2: Kaufman Brief Intelligence Test. Minneapolis, MN: NCS

Pearson. (1997).
80. Gopnik, A. & Astington, J. W. Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Dev* 26–37 (1988).

Supplementary Materials

Supplementary Note 1: Face events overlap analysis

We conducted an overlap analysis to determine whether the amount of overlap between timepoints identified as face events and timepoints identified as ToM or pain events was significantly different from that expected by chance. The overlap analysis was identical to the analysis used to determine whether ToM and pain events were significantly non-overlapping, described in Methods. The permuted face timecourses included seven face events, with durations of 16, 4, 5, 7, 4, 4, 4 TRs (see Supplementary Figure 4). Face and ToM events had 4 TRs of overlap in the actual timecourses; 111/1000 random permutation tests showed the same or smaller amount of overlap ($p=.11$). Face and pain events had 15 TRs of overlap in the actual timecourses; 928/1000 random permutation tests showed the same or smaller amount of overlap ($p=.93$). Thus, the amount of overlap between face and ToM events, and face and pain events, did not differ from that expected by chance.

Supplementary Table 1

Age Group	N	Age Range M (SD)	Gender (#F)	Handedness (R/L/Ambi)	Raw IQ M (SD)	Scaled/ Standard IQ M (SD)	DCCS Summary M (SD)	ToM Score M (SD)	Explicit FB Groups (P/I/F)
3yo	17	3.52-3.99 3.75 (.18)	10	15/2/0	15.5 (3.4)	10.4 (2.3)	1.75 (.93)	.54 (.18)	4/4/9
4yo	14	4.06-4.86 4.43 (.29)	8	13/0/1	16.9 (3.9)	9.64 (3.4)	2.29 (.61)	.63 (.15)	3/7/4
5yo	34	5.01-5.99 5.51 (.29)	16	26/6/2	20.3 (5.2)	111.7 (13.3)	2.32 (.47)	.73 (.14)	23/9/2
7yo	23	7-7.96 7.54 (.37)	11	23/0/0	29.4 (6.8)	116.7 (16.8)	NA	.88 (.09)	20/3/0
8-12yo	34	8-12.3 9.77 (1.18)	19	33/1/0	35.6 (3.9)	120 (11.7)	NA	.96 (.06)	34/0/0
Adult	33	18-39 24.8 (5.3)	20	32/1/0	NA	NA	NA	NA	NA

Supplementary Table 1. Demographic information and behavioral data by age group. Number of participants (N), age range and average and standard deviation of age (years), gender, handedness, raw and standardized measures of nonverbal IQ, DCCS summary score (possible range: 0-3)[6], ToM score (proportion of all questions answered correctly; possible range: 0-1), and number of children in each explicit false belief task group (pass, inconsistent, fail), per age group. Nonverbal IQ was measured via the WPPSI block design task for children ages 3-4 years[7], and via the KBIT-2 matrices task for children ages 5-12 years[8]. Children ages 7 and older did not complete the DCCS task. False belief task passers answered 5 or 6 of 6 questions correctly, inconsistent performers answered 3-4 questions correctly, failers answered at most 2 of 6 questions correctly.

Supplementary Table 2

Network Contrast	ROI	Center Coordinate	Size (voxels)
ToM			
ToM > Pain	RTPJ	[48 -60 30]	376
	LTPJ	[-48 -62 30]	368
	PC	[0 -54 34]	382
	DMPFC	[-6 54 36]	217
	MMPFC	[-4 58 16]	275
	VMPFC	[-4 56 -16]	198
Pain			
Pain > ToM	RS2	[60 -28 38]	368
	LS2	[-62 -32 34]	269
	Rinsula	[42 6 -6]	309
	Linsula	[-42 -2 -4]	240
	RMFG	[50 42 12]	142
	LMFG	[-46 36 14]	256
	AMCC	[0 2 42]	249
Face			
Face > Object	RFFA	[38 -42 -22]	1019
	LFFA	[-40 -52 -18]	531

Supplementary Table 2. Group regions of interest. Contrast used, regions identified, peak/center coordinate [x y z], and size (number of voxels) for each region of interest in the ToM network and Pain matrix, and for the bilateral fusiform regions used in Supplementary Note 1 and Supplementary Figure 4. See Supplementary Figure 1 for a visualization of these regions of interest.

Supplementary Table 3

	Even	Time	Duration (s)	Peak Timepoint (TR)	Description
ToM Events	T01*	4:00-4:10	10	131	Peck flies away from Gus after seeing the baby shark (T06), landing on another (happier) cloud. Peck and the happy cloud seemingly laugh together about Gus.
	T02*	2:46-3:02	16	96	Peck stares longingly at a happy cloud who is making puppies. Gus notices this, and looks worried. Peck notices that Gus caught him looking longingly, and feels bashful.
	T03*	1:14-1:20	6	46	Baby crying, then becomes happy when given a helmet.
	T04*	4:42-4:56	14	150	Peck dons football gear, to explain to Gus that he did not abandon him, but rather was acquiring protective equipment such that he could continue to deliver Gus's (dangerous) babies.
	T05*	1:28-1:36	8	54	Pan from happy clouds to Gus, who expresses loneliness.
	T06*	3:48-3:58	10	124	Peck is startled by the baby shark Gus has made. He notices a happy cloud who is making chicks.
	T07*	1:50-1:54	4	64	Peck and Gus greet each other happily (they are friends).
Pain Events	P01*	3:36-3:46	10	118	Gus pulls porcupine spines out of Peck's head.
	P02*	2:06-2:24	18	73	Alligator baby is biting Peck's head repeatedly.
	P03*	3:20-3:32	12	111	Peck tosses porcupine baby; expressing pain.
	P04*	1:00-1:06	6	39	Cloud makes baby animals (lightning).
	P05*	1:34-1:40	6	56	Gus makes baby alligator (lightning).
	P06*	4:10-4:20	10	135	Gus expresses anger (lightning).
	P07*	5:02-5:06	4	160	Peck is electrocuted by baby eel (lightning).
	P08	0:42-0:46	4	29	Flock of birds fly through clouds.
	P09	4:28-4:32	4	143	Gus begins to cry heavily (rain shower).
	P10*	1:20-1:24	4	49	Baby reaches for and is given helmet.
	P11	2:30-2:34	4	84	Peck's feathers fall off; Gus brushes Peck off.
	P12*	3:04-3:08	4	100	Peck is hit by baby ram in bundle he is trying to carry.

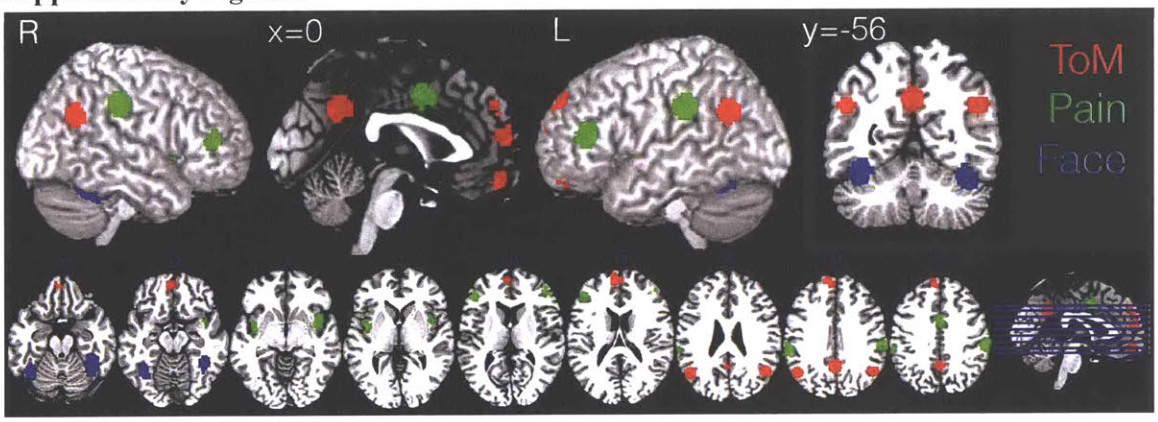
Supplementary Table 3. ToM and Pain Event details. Time (in stimulus), duration (seconds), and peak timepoint (TR) and description for each ToM and Pain event.[2] Peak timepoint is the timepoint with the greatest average response magnitude in adult participants. Event labels (T01, P01) reflect rank order of average response magnitude in adults. *Asterisks indicate events replicated in reverse correlation analysis of an independent sample of adults (Supplementary Figure 6).

Supplementary Table 4

Age Group	ToM	Pain	Face	Scene	M1	V1	Bi-FFA
3yo	0.28	0.60	0.75	0.61	0.11	0.53	0.72
4yo	0.31	0.56	0.59	0.67	0.06	0.61	0.60
5yo	0.60	0.73	0.71	0.77	0.35	0.78	0.68
7yo	0.72	0.83	0.84	0.80	0.44	0.74	0.82
8-12yo	0.82	0.89	0.86	0.85	0.53	0.76	0.83

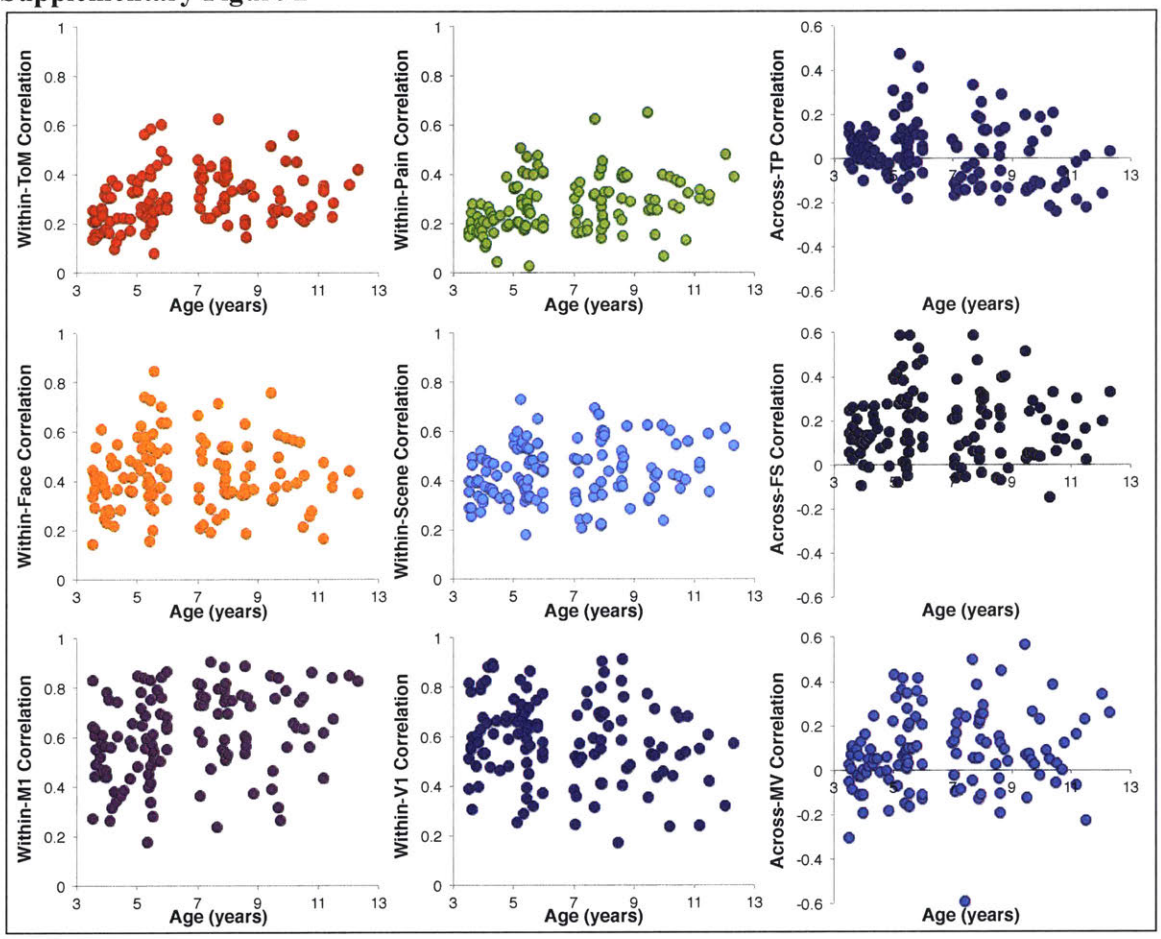
Supplementary Table 4. Average timecourse correlations. This table provides the Pearson correlation value (*r*) between the average timecourse of response in each network included in the expanded IRC analysis, for each age group, and the corresponding average timecourse of response in adults. These timecourses are the same as those used for the reverse correlation analysis (the M1 timecourse is not included as a regressor), prior to z-normalization. All correlations are significantly positive (ps<.0005) except those shaded in grey (3yo M1: p=.16; 4yo M1: p=.46).

Supplementary Figure 1



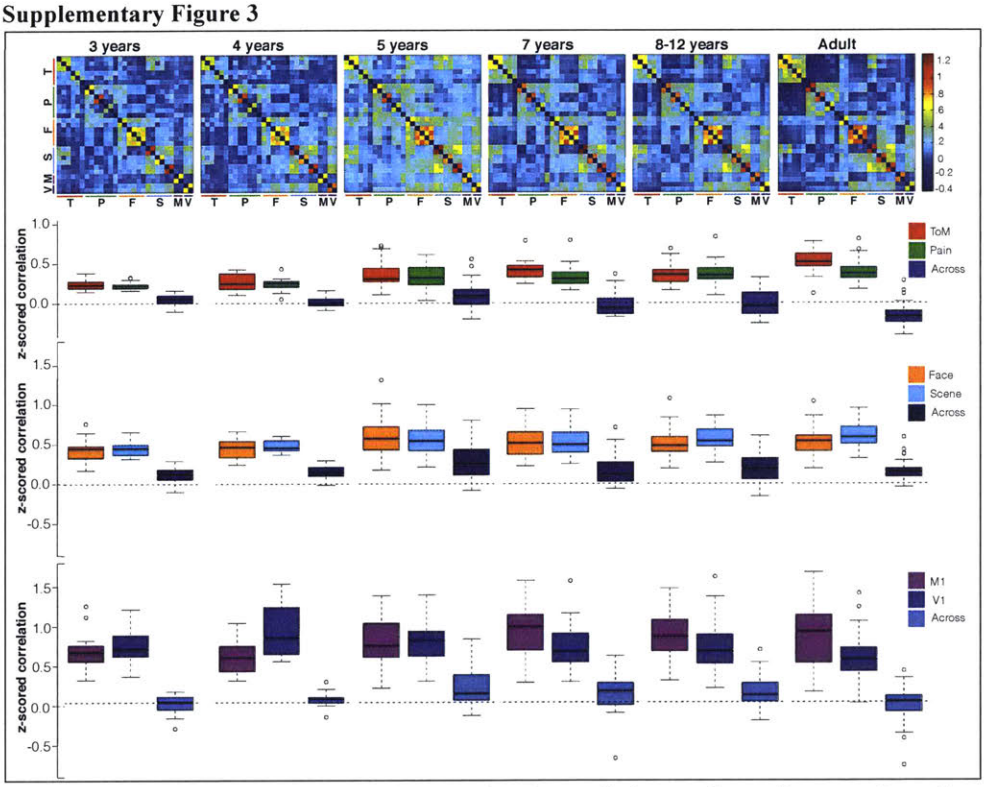
Supplementary Figure 1. Group Regions of Interest. ToM (red) and pain (green) regions were defined based on group-level contrast images in n=20 adults scanned by Evelina Fedorenko and colleagues (see Methods and Supplementary Figure 6). Bilateral fusiform regions (blue) were created by and described in[1]. See Supplementary Table 2 for ROI center coordinates and size.

Supplementary Figure 2



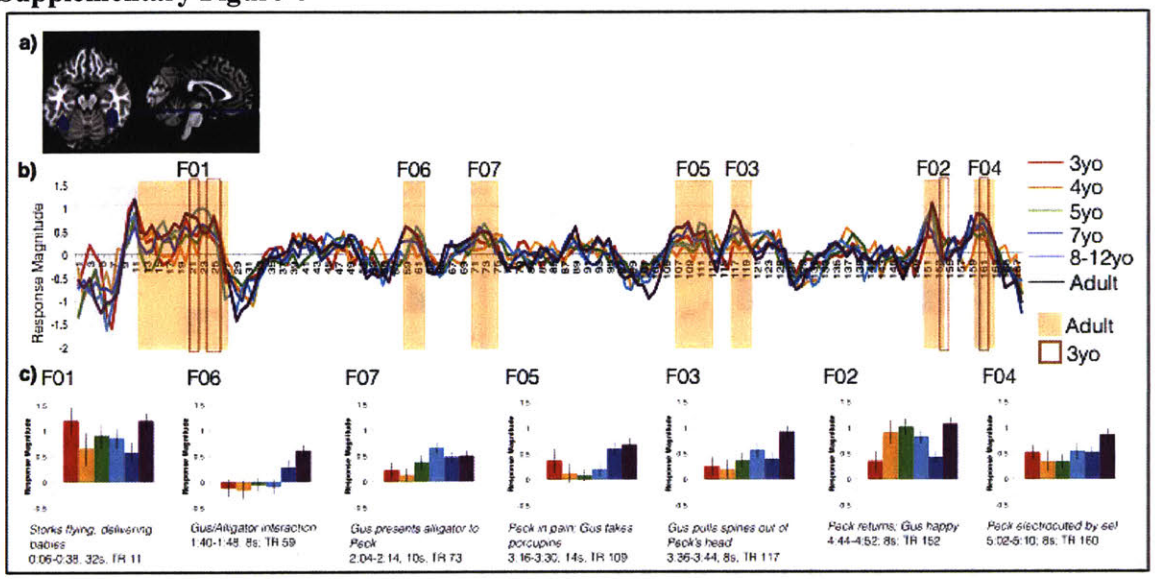
Supplementary Figure 2. Inter-regional correlations by age. Correlations are the raw (non z-scored) *r*-values (y-axis), calculated on the "raw" timecourses (without regression of the bilateral-M1 timecourse). Correlation values are shown for all children (n=122), with age on the x-axis. **Top row:** Within-ToM (red), Within-Pain (green), and across-ToM-Pain (dark blue) network correlations. **Middle row:** Within-Face (orange), Within-Scene (light blue), and across-Face-Scene (navy) network correlations. **Bottom row**: Within-M1 (purple), Within-V1 (bright blue), and across-M1-V1 (light blue) network correlations. See Results (main text) and Supplementary Figure 3 for statistics on change with age (all statistical tests used z-scored correlation values).

Supplementary Figure 3



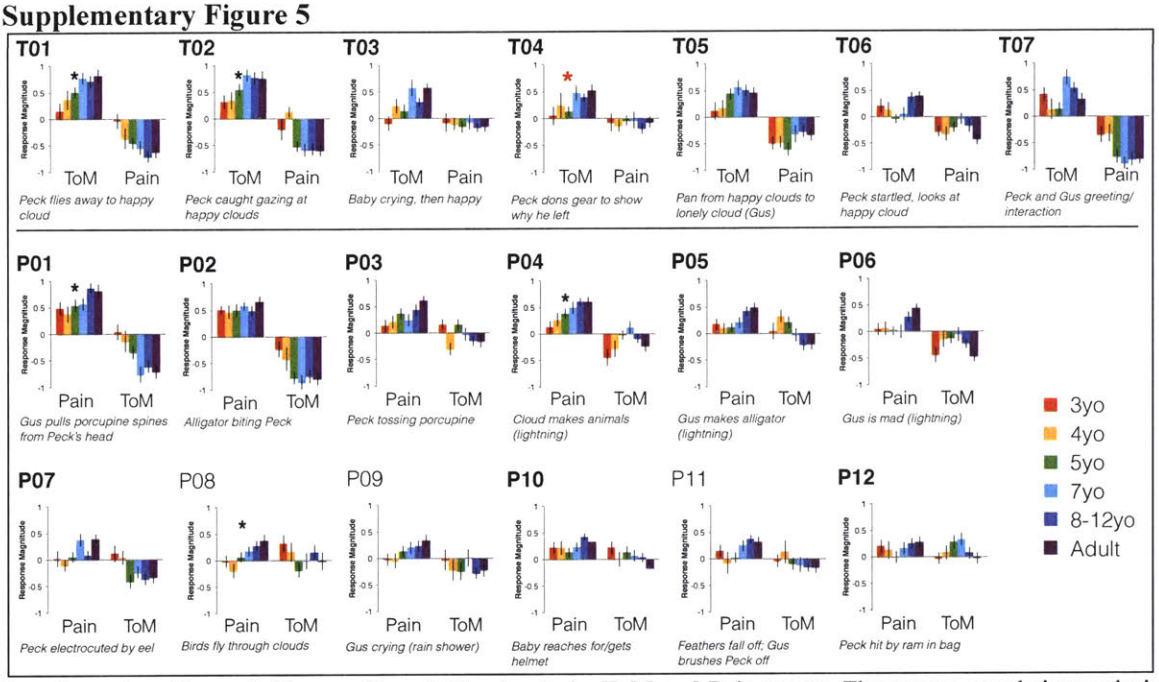
Supplementary Figure 3. Expanded inter-regional correlation analyses. Top row shows interregional z-scored correlation matrices for an expanded list of brain regions in the following order (ToM regions: RTPJ, LTPJ, PC, DMPFC, MMPFC, VMPFC, Pain regions: RS2, LS2, RInsula, LInsula, RMFG, LMFG, daMCC, Face regions: RSTS, LSTS, ROFA, LOFA, RFFA, LFFA; parcels from[1], Scene regions: RRSC, LRSC, RTOS, LTOS, RPPA, LPPA; parcels from[1], primary motor cortex: RPM, LPM, primary visual cortex: R Calcarine Sulcus, LCalcSulc. Primary motor and visual cortex ROIs are 10mm spheres drawn around peak coordinates generated with Neurosynth (http://neurosynth.org/; M1 coordinates: [38,-24,58], [-38,-20,58]; V1 coordinates: [-10 -86 2], [10 -86 2], see Methods). Boxplots show within- and across-network z-scored correlation values for all participants (n=122 children, n=33 adults), binned by age group, for ToM and Pain networks (top row of boxplots), Face and Scene networks (middle row) and bilateral primary motor and visual cortex regions (bottom row). All age correlation tests were spearman partial correlation tests, including amount of motion (number of artifact timepoints) as a covariate. Significant positive age correlations among children (n=122) are present for within-ToM (r_s=.39, p<.0001), within-Pain (r_s=.38, p<.0001), within-Scene (r_s=.23, p=.01), and within-M1 regions (r_s=.30, p=.001; within-Face: r_s=.06, p=.53, within-V1: r_s=-.16, p=.09). Because the within-M1 correlation increases with age, including it as a regressor in the interregional correlation analyses in the main text ensures that reported age effects in the ToM and pain networks are above and beyond developmental effects present in regions like primary motor cortex. The M1 timecourse is not regressed out from the timecourses analyzed for this figure/the expanded IRC analysis. Across-ToM-Pain network correlations decrease with age (e.g., become more anti-correlated: r_s=-.26, p=.005). Across-Face-Scene and Across-M1-V1 correlations do not show significant change with age: Across-Face-Scene: r_s=.02, p=.8; Across-M1-V1: .18, p=.05). Positive correlations between within-ToM and within-Pain correlations and age were significantly stronger than within-Face and within-V1 correlations, but not significantly stronger than within-Scene and within-M1 correlations (Williams' test of differences in age correlations: Within-Face: vs. within-ToM: z=2.69, p=.01, vs. within-Pain: z=2.66, p=.01; within-Scene: vs. within-ToM: z=1.33, p=.18, vs. within-Pain: z=1.3, p=.19; M1: vs. within-ToM: z=.76, p=.44, vs. within-Pain: z=.73, p=.47; V1: vs. within-ToM: z=4.35, p=0 , vs. within-Pain: z=4.32, p=0). The across ToM-Pain anti-correlation was significantly stronger than the across Face-Scene and across M1-V1 anti-correlations (Face-Scene: z(122)=2.2, p=.03; M1-V1: z(122)=3.39, p=0). See Supplementary Figure 2 for scatter plots of raw correlation values by age, and Supplementary Table 4 for correlations between the average timecourse of each age group and adults, for these additional networks.

Supplementary Figure 4



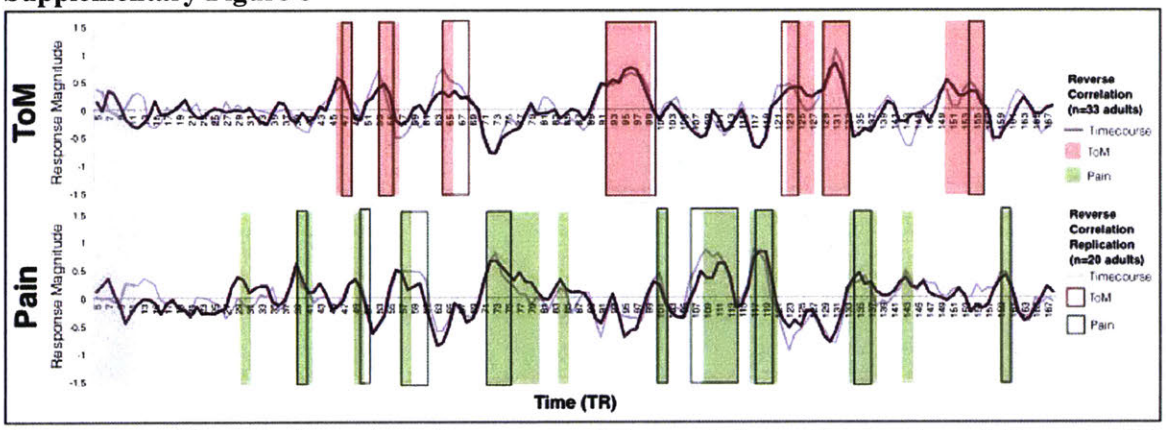
Supplementary Figure 4. Bilateral Fusiform Reverse Correlation Analysis. a) Bilateral fusiform regions of interest (ROIs). Regions are face parcels created with the group-constrained subject specific (GSS) method applied in n=30 adults, using a faces > objects contrast[1], and made publically available (http://web.mit.edu/bcs/nklab/GSS.shtml). A subset of participants (n=17 total; n=2 adults, n=12 8-12yos, n=1 7yo, 4yo, 3yo) had incomplete coverage of the ventral visual stream; however, all participants had measurable neural responses in at least 100 voxels in all face ROIs. **b)** Average timecourse of response extracted from bilateral fusiform ROIs, per age group. The average timecourse in bilateral fusiform ROIs in children was highly correlated with that of adults (Pearson correlation: r=.86, p<1.0x10^{-}47). This correlation remained high when comparing adults to three-year-old children alone (Pearson correlation: r=.72, p<1.0x10^{-26}). Shaded light orange blocks denote 7 events (88s total, M(SD) length 12.6(8.8)s) identified in a reverse correlation analysis of the timecourse of response in adult participants (n=33; see Methods); dark orange outlines denote 4 events (18s total M(SD) length 4.5(1)s) identified in a reverse correlation analysis of the timecourse of response in three-year-old children (n=17). Event labels (e.g. F01, F02) reflect rank order of magnitude of response in adults. A majority of the timepoints identified by the reverse correlation analysis in three year olds fall within adult events F02, F03, and F07 (8/9 TRs); the remaining timepoint immediately follows adult event F07. **c)** Short description, timing and duration, timepoint of peak response, and response magnitude by age group for each event identified in the reverse correlation analysis. Error bars represent standard error. Peak timepoints were chosen based on the adult data, included here for illustration. Statistical tests of age-related change were computed only on data from children (n=122). The magnitude of response in bilateral fusiform does not change with age among children (spearman partial correlation including motion as covariate; Bonferroni correction for multiple comparisons α=.0071, correcting for 7 events/tests; $|r|$s<.24, ps>.01). All events included at least one face, and close-ups of faces; some events featured particularly salient faces (e.g. faces with painful expressions: F01, F02, F05). See online version for thumbnail images of events (http://rdcu.be/IRh8).

Supplementary Figure 5



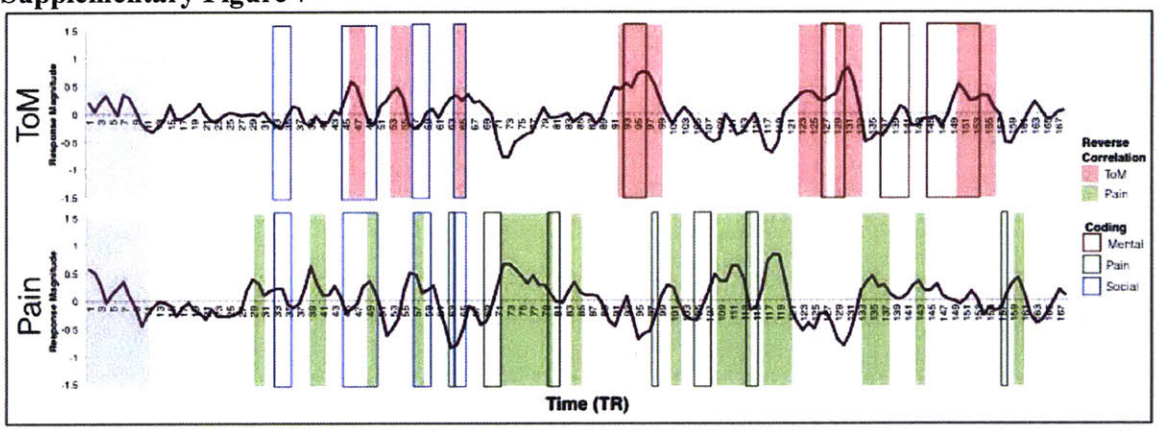
Supplementary Figure 5. Reverse Correlation Analysis: ToM and Pain events. The reverse correlation analysis in adults identified seven ToM events (top) and twelve pain events (bottom). A description is provided for each event; see online version of this figure for thumbnail images (http://rdcu.be/IRh8). The bar graphs show the average response magnitude of response per age group in the ToM and Pain networks, for each event. Peak timepoints were chosen based on the adult data, included here for illustration. Statistical tests of age-related change were computed only on data from children (n=122). Asterisks denote events that evoke significantly greater responses with age (black; partial spearman correlation controlling for motion and correcting for multiple comparisons (MC) (19 events, α=.0026)), or ToM behavioral performance (red; linear regression including age and motion as additional predictors, and correcting for MC (7 events, α=.007)). Event labels (e.g. T01, T02) in bold type are those that were replicated in an independent sample of adults (n=20; Supplementary Figure 6). See Supplementary Table 3 for event timing, duration, and full descriptions.

Supplementary Figure 6



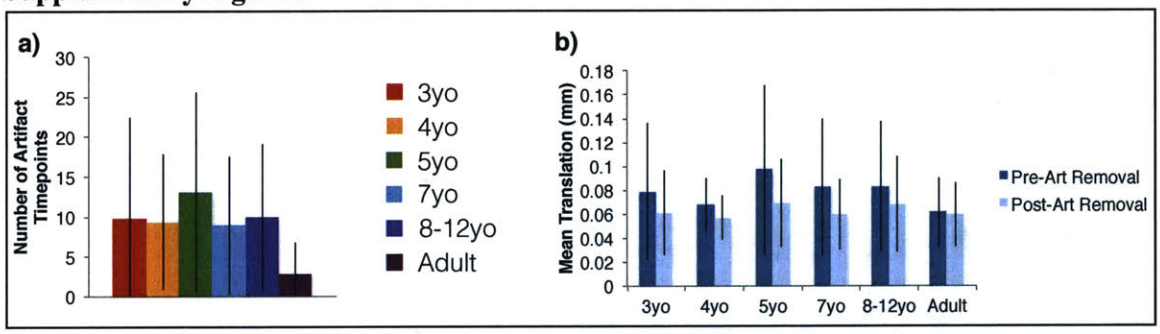
Supplementary Figure 6. Comparison of reverse correlation analysis results across two adult samples. We analyzed fMRI data from an independent sample of adults (n=20), collected by Evelina Fedorenko's lab, who viewed "Partly Cloudy" [2] in the scanner. We used this sample to create independent, stimulus-tailored group ROIs (see Methods). We also used this sample to test whether the events identified by the reverse correlation analysis in our adult sample (n=33) were replicated in an independent sample of adults. Because we used this independent sample of adults to create the group ROIs used for our sample of interest, we created a different set of group ROIs for the reverse correlation replication analysis of these participants (to avoid using non-independent ROIs). The group ROIs used in the independent sample were 10mm spheres surrounding peak coordinates reported in previous publications (ToM regions[3]; Pain matrix[4]). This figure shows the average timecourse of response in the primary adult sample (n=33, dark purple) and the independent replication sample (n=20, light purple), in each network. Shaded blocks indicate events identified by reverse-correlation in the primary adult sample; dark borders indicate events identified by reverse-correlation in the replication sample (ToM: red, Pain: green). Seven ToM and nine pain events were identified in the reverse correlation analysis of this independent sample of adults (ToM: 60s total, M(SD) length: 8.6(4.6)s, Pain: 66s total, M(SD) length: 7.3(4.4)s). All events identified in the independent group of adults (using group ROIs[3,4]) were also identified in our adult participants. Three pain events that were identified in our primary sample of adults were not labeled as events in the independent sample of adults: P08, P09, and P11 (9/12 overlapping pain events comprised of 52s of overlap, and 7/7 overlapping ToM events comprised of 54s of overlap). Thus, the reverse correlation analysis approach successfully identifies events that reliably evoke responses in ToM and pain brain regions across adult subjects, and across two independent adult samples. This suggests that this approach is particularly well suited for identifying events for further analyses of changes in neural responses with development.

Supplementary Figure 7



Supplementary Figure 7. Comparison of Reverse Correlation analysis and original event coding. A previous study coded "mental," "social," and "pain" events of the movie stimulus, in order to compare the magnitude of response across conditions and localize ToM and pain brain regions using contrasts (ToM regions: Mental > Pain; Pain matrix: Pain > Mental)[5]. We compared the coding created by the experimenters to the event labels suggested by the reverse correlation analysis in our adult participants (n=33). This figure shows the average timecourse of response in adult participants in ToM (top) and pain (bottom) networks. Shaded blocks indicate events identified by reverse correlation analysis (ToM: red, Pain: green). Colored borders indicate condition labels constructed by previous experimenters for the purpose of using the movie stimulus as a functional localizer for identifying ToM and pain brain regions[5]. While most ToM events identified by the reverse correlation analysis were at least partially included in the original coding (6/7 ToM events labeled as Mental or Social), only two of twelve pain events were included in the coding. This lends support to the use of reverse correlation analysis for identifying reliable events that evoke responses in particular regions, rather than experimenter-based coding, for further study. The reverse correlation analysis approach may be useful for refining theories about the function of these networks of brain regions.

Supplementary Figure 8



Supplementary Figure 8. Amount of motion in fMRI data. a) Number of artifact timepoints identified in the timecourse of response (one run per participant, 168 timepoints total), by age group (3yo: n=17, 4yo: n=14, 5yo: n=34, 7yo: n=23, 8-12yo: n=34, adult: n=33). Artifact timepoints are timepoints in which there is 2mm motion and/or a global signal change greater than three standard deviations from the mean, relative to the previous timepoint. Error bars show standard deviation from the mean. **b)** Mean translation (motion in x, y, z directions) in millimeters per age group, including (dark) and excluding (light) artifact timepoints. Error bars show standard deviation from the mean.

Supplementary References

1. Julian, J. B., Fedorenko, E., Webster, J. & Kanwisher, N. An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *NeuroImage* **60,** 2357–2364 (2012).
2. Reher, K. (Producer), & Sohn, P. (Director). *Partly Cloudy* [Motion Picture]. United States: Pixar Animation Studios and Walt Disney Pictures (2009).
3. Dufour, N., Redcay, E., Young, L., Mavros, P.L., Moran, J.M., Triantafyllou, C., Gabrieli, J, Saxe, R.. Similar Brain Activation during False Belief Tasks in a Large Sample of Adults with and without Autism. *PLoS ONE* **8,** e75468 (2013).
4. Bruneau, E. G., Jacoby, N. & Saxe, R. Empathic control through coordinated interaction of amygdala, theory of mind and extended pain matrix brain regions. *NeuroImage* **114,** 105–119 (2015).
5. Jacoby, N., Bruneau, E., Koster-Hale, J. & Saxe, R. Localizing Pain Matrix and Theory of Mind networks with both verbal and non-verbal stimuli. *NeuroImage* **126,** 39–48 (2016).
6. Zelazo, P. D. The Dimensional Change Card Sort (DCCS): a method of assessing executive function in children. *Nat Protoc* **1,** 297–301 (2006).
7. Wechsler, D. Manual for the WPPSI-R. *New York: The Psychological Co* (1989).
8. Kaufman, A. S. KBIT-2: Kaufman Brief Intelligence Test. Minneapolis, MN: NCS Pearson. (1997).

Chapter 3: Development of Brain Networks for Social Functions: Confirmatory Analyses in a Large Open Source Dataset

During natural viewing of movies, human observers show robust activity in distinct brain networks, driven by the content of the movies. For example, scenes that emphasize characters' thoughts and feelings evoke activity in the "Theory of Mind" (ToM) network, whereas scenes that emphasize characters' bodily and physical states evoke activity in the "Pain Matrix". The current study investigates the developmental origins of the cortical dissociation between these networks, and the links between cortical and cognitive changes in children's social development. In particular, we sought to confirm results of a previous exploratory study on children (n=122, 3-12 years) and adults (n=33) who watched Pixar's animated short "Partly Cloudy" while undergoing fMRI. The previous results found that 1) ToM and pain networks are functionally distinct by age three years, 2) functional selectivity increases throughout childhood, 3) the magnitude of response during one scene was correlated with cognitive performance on a behavioral test of ToM, and 4) the "functional maturity" of the response timecourse was linked to the inter-region correlations within and between the two networks. We analyzed a large independent publicly available dataset of children, adolescents, and young adults (n=241, ages 5-20 years) who viewed Jacob Frey's "The Present"[1] while undergoing fMRI. Participants additionally completed a resting state scan (n=200), enabling us to further characterize the link between inter-region correlations and stimulus-driven responses. We find confirmatory evidence for an early functional dissociation between ToM and pain brain regions (by age five years), and for developmental increases in functional selectivity with age, and with a behavioral index of social reasoning. We additionally provide evidence that the relationship between the stimulus-driven response and inter-region correlations within and between ToM and pain networks during movie viewing is specific to network properties measured during the movie; inter-region correlations measured at rest were not related to the functional maturity of the response. Given the intense financial and time investments required to collect large samples of fMRI data in young children, the availability of the public dataset is critical to strengthen and enrich the results from the in-house dataset. This study thus provides insight into the scientific benefits of open source data in developmental cognitive neuroscience.

Introduction

Evidence from fMRI studies has suggested a striking cortical division between brain regions that process information about others' minds (the Theory of Mind (ToM) network), and those that process information about others' bodies (the Pain Matrix). This cortical dissociation has been studied across multiple experimental contexts in adults[2-5], and has recently been characterized in an exploratory study of children as young as three years old[6]. In order to measure functional responses in such young children, the prior study used a naturalistic movie-viewing paradigm: children (n=122, 3-12 years) and adults (n=33) viewed Disney Pixar's "Partly Cloudy"[7] while undergoing fMRI. This movie includes scenes that emphasize characters' thoughts and feelings, and scenes that emphasize characters' bodily and physical states, making it an ideal stimulus for measuring functional responses in ToM and pain brain regions[5,6].

The results of this prior study provided insights into the development of the functional division between these two cortical networks. A key result was that signatures of the functional division between ToM and pain brain regions were present in three-year-old children: responses in ToM brain regions were more correlated with other ToM regions, than with regions in the Pain Matrix, and vice versa. Additionally, responses in three-year-old children were significantly correlated with the average adult timecourse, suggesting early "functional maturity." This study also found significant developmental change in inter-region correlations and functional maturity throughout childhood, suggesting continual development and refinement of the functional responses in both networks. For example, while three-year-old timecourses generally looked similar to those of adults, some ToM scenes evoked responses in the Pain Matrix, and some pain scenes evoked responses in the ToM network. The responses in these networks became increasingly distinct (anti-correlated) over childhood. Additionally, across all children, responses to a particular ToM scene were correlated with cognitive performance on an independent behavioral test of ToM (https://osf.io/g5zpv/). If robust, such a neural marker could be a useful index for designing and evaluating interventions aimed to improve social cognitive abilities. Finally, this prior study provided evidence for a relationship between inter-region correlations within and between ToM and pain brain regions, and the development of stimulus-driven functional responses within each network.

Here, we sought to provide confirmatory evidence for these results by analyzing a large, publicly available dataset of five to twenty year olds (n=241) who viewed Jacob Frey's "The Present"[1] while undergoing fMRI, and additionally completed a behavioral metric of social reasoning: the Social Communication Questionnaire (SCQ[8]). Confirmatory evidence strengthens the confidence in results based on exploratory analyses, and in this case, tests the generalizability of the results to a more diverse sample of participants and under a new experimental context (i.e., a different movie paradigm and behavioral measure of social reasoning). Developing robust generalizable neural markers of social cognitive behaviors is critical for developing and testing the effectiveness of social cognitive training paradigms and clinical interventions.

The current study additionally aimed to clarify the link between the development of stimulus-driven functional responses and inter-region correlations. Because participants did not complete a resting state scan, the previous study could not determine the extent to which inter-region correlations during movie viewing reflected stimulus-driven (functional) responses vs. intrinsic network properties. Intrinsic networks are cortical regions that have correlated (and anti-

correlated) timecourses of activity at rest[9,10], i.e., in absence of stimuli. These intrinsic networks largely correspond to the functional divisions in cortex: brain regions that are correlated during cognitive tasks are also correlated at rest[9,11,12].

One hypothesis is that the relationship between functional maturity and inter-region correlations is driven by the *stimulus driven response* in these two networks during movie viewing. That is, systematic functional responses to stimuli organize these brain regions into two functionally distinct, anti-correlated networks. Inter-region correlations during functional tasks could subsequently shape intrinsic inter-region correlations at rest. Previous work has found that stimulus-elicited connectivity predicts resting state connectivity patterns longitudinally[13], and resting state connectivity can be altered via intensive exposure to particular cognitive tasks[14]. Thus, engaging specific brain regions via functionally specific tasks could drive regions within a network to become correlated with one another, and anti-correlated with regions in other networks, and these functional dissociations may influence the intrinsic connectivity between brain regions at rest. Alternatively, this relationship could be driven by the *intrinsic properties* of the ToM and pain networks. Intrinsic networks are apparent by the end of the first year of life[15], if not earlier[16], and become more distinct over childhood[17]. Development of intrinsic networks could plausibly precede and influence the emergence of systematic functional responses. While the current cross-sectional study cannot determine predictive relationships or causal order of development between intrinsic and functional networks, it can test whether the functional maturity of responses in ToM and pain brain regions are specifically related to inter-region correlations during movie viewing, or are more generally related to inter-region correlations that are intrinsic, i.e., present at rest.

Thus, the current study was conducted with two goals. First, we sought to confirm the previous results by analyzing a publicly available dataset of 5 – 12 year old children who viewed Jacob Frey's "The Present"[1] while undergoing fMRI. Second, because many participants additionally completed a resting state scan, we used this sample to characterize the nature of the link between stimulus-driven responses and inter-region correlations within and between the ToM and Pain networks during movie viewing.

Results

We sought to confirm results of a previous exploratory study on children (n=122, 3-12 years) and adults (n=33) who watched Pixar's animated short "Partly Cloudy" while undergoing fMRI. The previous results found that 1) ToM and pain networks are functionally distinct by age three years, 2) network differentiation increases throughout childhood, 3) the magnitude of response during one scene is correlated with cognitive performance on a test of ToM, and 4) the "functional maturity" of the response timecourse is linked to the inter-region correlations of the networks. We attempted to replicate these results in an independent, large, and diverse sample of participants who viewed a different movie (Jacob Frey's "The Present"[1]) while undergoing fMRI.

Replication: Inter-region Correlation Analyses

As in the original study, we first confirmed that ToM and Pain brain regions were significantly more correlated with within-network brain regions, compared to brain regions in the opposite network. Higher within – across network correlations are one indication of functional

specialization. Among teenagers and young adults, within-network correlations (M(SE) Wi-ToM: .34(.02), Wi-Pain: .23(.01)) were significantly higher than across network correlations (M(SE) ac-TP: -.15(.01); within vs. across-network two-tailed paired t-tests: ToM: t(52)=21.4, $p<2.2x10^{-16}$; Pain: t(52)=22.9, $p<2.2x10^{-16}$).

We then tested for developmental change in inter-region correlations. Because the age range of the current sample (ages 5 – 20 years) differs from that of the previous study (ages 3 – 12 years, and adults), we conducted primary inter-region correlation analyses in the full sample, and additionally report evidence from age-matched child samples (5 – 12 year olds) from the two studies (see Supplementary Materials for the results of age-matched analyses in the prior study).

Consistent with the previous results, within-network inter-region correlations increased significantly with age (linear regression testing for effects of age and motion on within-ToM correlation: effect of age: b=.15, t=2.4, p=.02, effect of motion: b=-.22, t=-3.4, p=.0007; on within-Pain correlation: effect of age: b=.17, t=2.6, p=.009, NS effect of motion: b=-.12, t=-1.8, p=.08), and across network inter-region correlations decreased significantly with age (effect of age: b=-.20, t=-3.2, p=.001, effect of motion: b=.31, t=5.0, $p=9.9x10^{-7}$); see Figure 1 and Supplementary Figure 1. In both studies, developmental change with age was less apparent in the narrower, matched age range (5 – 12 year old children). In the current sample, we did not find significant developmental change with age in within-network correlations (spearman partial correlations including motion as covariate: ToM: $r_s(182)=.14$, p=.055; Pain: $r_s(182)=.11$, p=.12), or in across-network correlations ($r_s(182)=-.12$, p=.10), in 5 – 12 year old children, but all correlations showed developmental trends in the predicted direction.

Figure 1

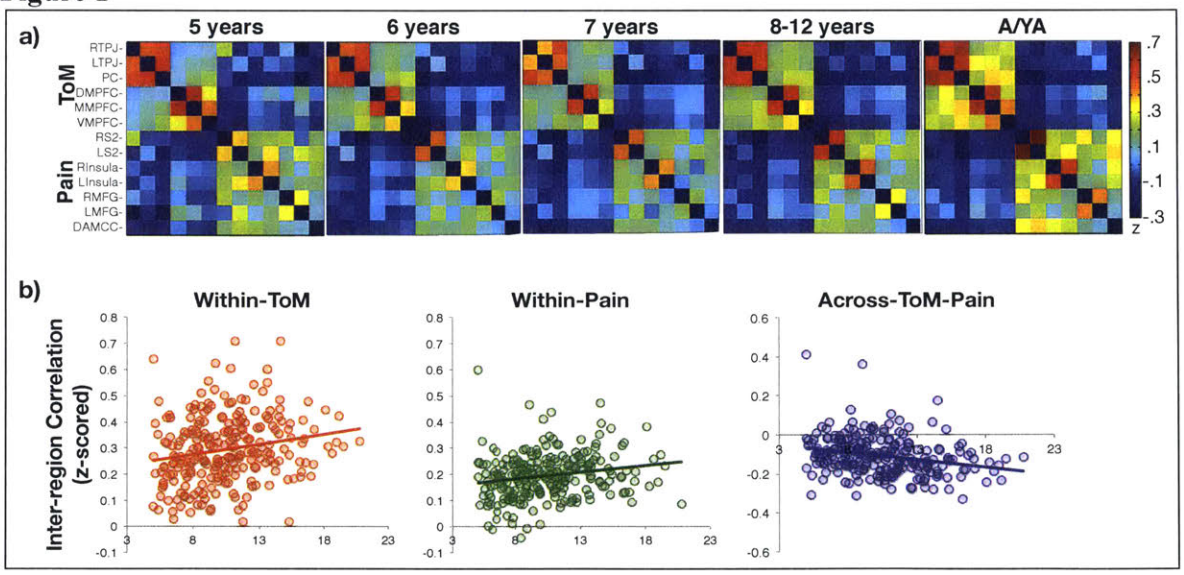


Figure 1. Developmental Change in Inter-region Correlations. a) Average z-scored correlation matrices across all ToM and pain brain regions of interest (see y-axis) per age group (5yo: n=16; 6yo: n=21; 7yo: n=26; 8-12yo: n=123; adolescents/young adults (YA; 13-20 years): n=55). Regions are in the same order along the X-axes and Y-axes. **b)** Z-scored inter-region correlations (y-axis) by age (x-axis) within the ToM network (left, red), within the Pain network (middle, green), and across the ToM-Pain networks (right, blue).

The previous study found evidence for functional network differentiation in children as young as three years old. In the current sample, the youngest children scanned were five years old. In these children, responses in ToM and Pain brain regions were more correlated with within-network brain regions than brain regions in the opposite network (n=16 5yo; M(SE) within-ToM: .25(.04), within-Pain: .19(.03), across-ToM-Pain: -.07(.04); within vs. across-network correlation paired two-tailed t-test: ToM: $t(15)=7.3$, $p=2.7x10^{-6}$, Pain: $t(15)=7.0$, $p=4.5x10^{-6}$).

We then tested for significant correlations between inter-region correlations in ToM brain regions and scores on the Social Communication Questionnaire (SCQ[8]), a parent report questionnaire that measures social cognitive reasoning. The previous study found a significant relationship between performance on a ToM behavioral battery and within-ToM and across-ToM-Pain inter-region correlations, but these relationships did not remain significant when additionally controlling for age. In the current sample, there were no significant correlations between within-ToM or across-ToM-Pain inter-region correlations and SCQ scores among children (partial correlations including motion as covariate: rs<|.11|, ps>.2), or in the full sample (rs<|.06|, ps>.4).

Replication: Reverse Correlation Analyses

Reverse-correlation analyses offer a data-driven way to determine what kinds of stimuli drive responses in particular brain regions. We conducted reverse correlation analyses on the neural responses in adolescent and young adult participants (n=55) while they watched "The Present."[1] Reverse correlation analyses of this stimulus produced seven ToM events (40 seconds total, M(SD) length: 5.7(3.0) seconds) and three Pain events (21.6 seconds total, M(SD) length: 7.2(1.4) seconds); see Figure 2. Six of the seven ToM events clearly depicted moments that involved reasoning about mental states (beliefs, goals, emotions) of the characters (e.g., boy curiously opening present, boy expressing annoyance, and gaining a new understanding of the boy, upon realizing that he, like the puppy, has lost his leg). The remaining ToM event introduced the boy character and showed him playing video games. The three Pain events depicted moments involving physical pain or clumsiness (due to the missing leg). See Supplementary Table 1 for more information about the timing and content of the events. Out of the 245 timepoints tested (all but the first 5 TRs (4s)), there were zero timepoints that reliably evoked significantly positive responses in both ToM and Pain events.

Responses to one ToM event (T01) increased significantly with age among 5-12 year old children (partial spearman correlation test, including motion as a covariate: $r_s(175)=.23$, $p=.0024$; Bonferroni correction for multiple comparisons $\alpha=.005$, for ten events; all other events: ps>.009). Responses to a second ToM event (T02) were positively correlated with SCQ score in a linear regression that included age and motion as covariates (effect of SCQ: -.21, $t=-2.4$, $p=.018$, effect of age: $b=.19$, $t=2.3$, $p=.02$, NS effect of motion: $b=-.05$, $t=-.57$, $p=.57$; all other events SCQ ps>.12), but this relationship did not survive Bonferroni correction for multiple comparisons ($\alpha=.007$, for seven ToM events).

As in the previous study, we additionally conducted reverse correlation analyses on the youngest participants scanned (age 5 years old, n=16). Responses in five year olds were generally highly correlated with the average adolescent/young adult timecourse (M(SE) of Pearson correlation: ToM: .23(.04), Pain: .22(.04); one sample t-tests against zero: $ts(15)=6.1$, $ps<2.2x10^{-5}$). In five

year olds, reverse correlation analyses identified two of the seven ToM events and one of the three Pain events defined in the adolescent/young adult participants; see Figure 2. These events made up a majority of the timepoints identified as events in the five year olds (18/32 TRs). Three of the remaining 14 TRs immediately preceded or followed these events. The remaining 11 TRs comprised one ToM event and one Pain event, which shared a single timepoint (6 TRs each); neither of these events were identified in the adolescent/young adult sample. See Supplementary Figure 2 and Supplementary Table 1 for more information about all events.

Figure 2

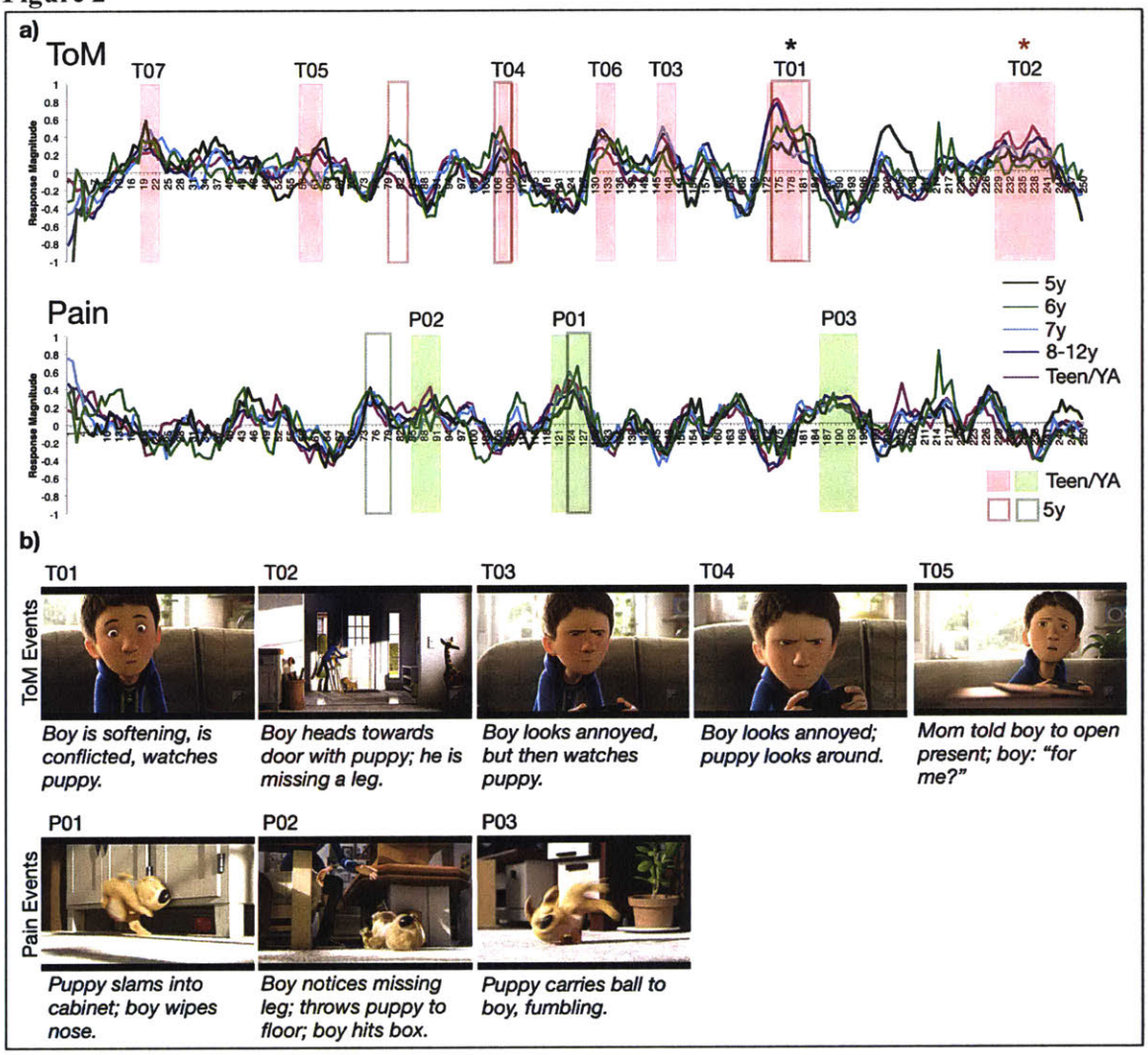


Figure 2. Functional Timecourses during "The Present." a) The average timecourse per age group for the ToM network (top) and Pain matrix (bottom), during viewing of Jacob Frey's "The Present."[1] Each timepoint along the x-axis corresponds to a single TR (800ms); the entire movie was 250 TRs (<4 min). Shaded blocks show timepoints identified as ToM (red) and Pain (green) events in a reverse correlation analysis conducted on adolescent/young adult participants (13-20 year olds; n=55); timepoints within the gray block were not analyzed. Dark red and green borders show timepoints identified as ToM and pain events, respectively, in 5-year-old children (n=16). Event labels (e.g., T01, P01) indicate ranking of average peak magnitude of response in adolescents/young adults. Black asterisk indicates significant positive correlation between peak magnitude of response and age (continuous variable) among

children, after correcting for multiple comparisons (10 ToM/Pain events, α=.005). Red asterisk indicates significant positive correlation between peak magnitude of response and SCQ score (continuous variable) among children; this correlation does not survive correcting for multiple comparisons (7 ToM events; α=.007, p=.02). **b)** Example frames and descriptions for the five events with the highest magnitude of response in adolescents/young adults, per network (see Supplementary Fig. 2 for all events, and Supplementary Table 1 for full event descriptions and timing and duration information). Thumbnail images used with permission from Jacob Frey.

Replication: Functional Maturity

The current dataset included adolescents/young adults (13-20 years old), rather than adults (in the previous study: ages 18-39 years old). Response timecourses among 5 – 12 year old children were generally positively correlated with the average timecourse of adolescents and young adults (n=186 5-12yo: M(SE) Pearson correlation value (*r*): ToM: .30(.01), Pain: .27(.01)). However, as in the previous study, functional maturity (i.e., similarity to responses in adolescents/young adults) in ToM and Pain networks increased with age among 5-12 year old children (spearman partial correlations including motion as a covariate: ToM: $r_s(182)$=.20, p=.006; Pain: $r_s(182)$=.19, p=.01). Functional maturity in the ToM network was not significantly correlated with SCQ score (spearman partial correlation including motion as a covariate: $r_s(150)$=.08, p=.35).

Figure 3

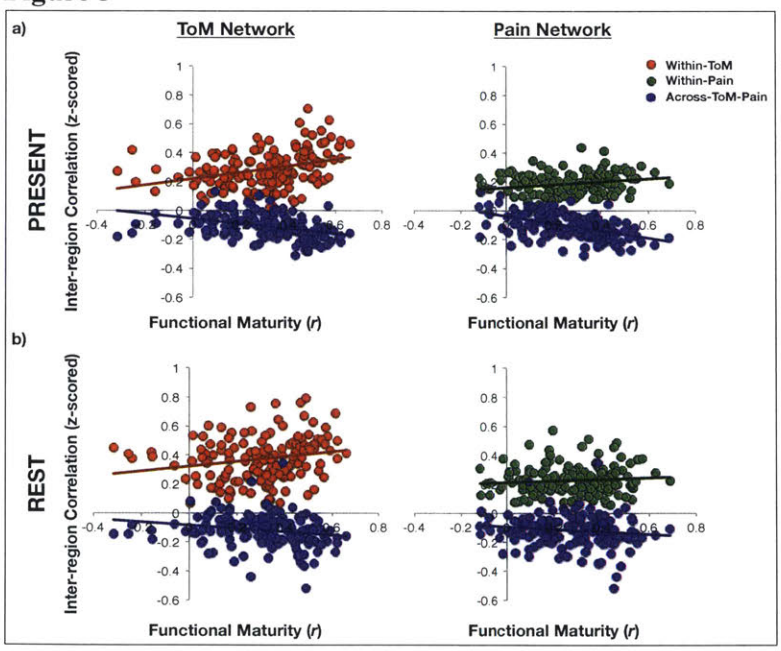


Figure 3. Relating Functional Maturity to Inter-region Correlations. Scatterplots show timecourse maturity (i.e., how correlated each child's timecourse is to the average adolescent/young adult timecourse (Pearson's *r*, x-axis) while viewing Jacob Frey's "The Present."[1] The y-axis shows z-scored inter-region correlation values within-ToM (red), within-Pain (green), and across-ToM-Pain (blue) networks, as measured while viewing **a)** "The Present", or **b)** at rest.

There were significant effects of within- and across-network correlations on functional maturity in the ToM network (effect of across-ToM-Pain correlation: b=-.17, t=-2.2, p=.03, effect of within-ToM correlation: b=.25, t=3.4, p=.0008, NS effect of age: b=.13, t=1.8, p=.07, NS effect of motion: b=-.09, t=-1.3, p=.21); see Figure 3a. In the Pain network, only the across-network correlation significantly predicted functional maturity (effect of across-ToM-Pain correlation: b=-.41, t=-5.9, p=$1.5x10^{-8}$, NS effect of within-Pain correlation: b=.11, t=1.6, p=.11, effect of age: b=.13, t=1.9, p=.06, NS effect of motion: b=-.07, t=-1.0, p=.30). For subsequent comparison to the resting state data, we confirmed that this same pattern of evidence was apparent in the low/matched motion subset of participants who contributed fMRI data to the movie and resting state scans (n=106; including

n=75 5-12yo). In this subset, functional maturity in both networks was predicted by the anti-correlation between the two networks (ToM: effect of across-ToM-Pain correlation: b=-.39, t=-2.8, p=.007, NS effect of within-ToM correlation: b=.10, t=.75, p=.46, NS effect of age: b=.03, t=.28, p=.78, NS effect of motion: b=-2.0, t=-1.9, p=.07; Pain: effect of across-ToM-Pain correlation: b=-.43, t=-3.9, p=.0002, NS effect of within-Pain correlation: b=.08, t=.74, p=.46, NS effect of age: b=.01, t=.93, p=.36, NS effect of motion: b=-.05, t=-.50, p=.62).

Extension: Inter-region Correlations During Resting State
A subset of the sample completed a resting state scan (n=200), enabling us to test if the pattern of results from the inter-region correlation analyses was specific to functional responses during a social, naturalistic movie-viewing paradigm. All inter-region correlation measures were highly correlated across movie-viewing and rest scans, even when controlling for age and motion (within-ToM: b=.50, t=5.2, p=1.04 x 10^{-6}, NS effect of age and motion: ps>.7; within-Pain: b=.49, t=5.2, p=9.3 x 10^{-7}, NS effect of age and motion: ps>.4; across-ToM-Pain: b=.47, t=4.5, p=1.7 x 10^{-5}, NS effect of age and motion: ps>.6); see Supplementary Figure 3.

Within-network correlations (M(SE) within-ToM: .51(.02), within-Pain: .29(.02)) were higher than across-network correlations (M(SE) across-ToM-Pain:-.23(.02)) during rest in adolescents/young adults (within vs. across-network correlation paired two-tailed t-test: ToM: t(48)=21, p<2.2x10^{-16}; Pain: t(48)=19.2, p<2.2x10^{-16}).

In the full sample (ages 5 – 20 years), within-network inter-region correlations increased significantly with age (linear regression testing for effects of age and motion on within-ToM correlation: effect of age: b=.37, t=5.6, p=6.7x10^{-8}, effect of motion: b=-.20, t=-3.1, p=.002; on within-Pain correlation: effect of age: b=.26, t=3.7, p=.0003, NS effect of motion: b=-.002, t=-.03, p=.97), and across network inter-region correlations decreased significantly with age (effect of age: b=-.45, t=-6.9, p=5.4x10^{-11}, NS effect of motion: b=.04, t=.60, p=.55); see Figure 4. Interestingly, while within-ToM and within-Pain network correlations did not increase significantly with age among 5 – 12 year old children (spearman partial correlations including motion as covariate: within-ToM: r_s(148)=.15, p=.07; within-Pain: r_s(148)=.07, p=.41), consistent with the results from the movie viewing task, across-network correlations during rest decreased with age (r_s(148)=-.28, p=.0005). There were no significant correlations between within-ToM or across-ToM-Pain inter-region correlations measured at rest and SCQ scores among children (partial correlations including motion as covariate: *r*s<|.08|, ps>.4), or in the full sample (*r*s<|.04|, ps>.6).

Very few five year olds were included in the resting state sample (n=7). However, even in this small sample, within-network correlations (M(SE) within-ToM: .32(.03), within-Pain: .26(.04)) were significantly higher than across network correlations (M(SE): -.08(.08); within vs. across-network correlation paired two-tailed t-test: ToM: t(6)=4.0, p=.008; Pain: t(6)=4.0, p=.007).

Next, we tested if the developmental separation of functional responses in the ToM and pain networks differed by task ("The Present" vs. resting state). We conducted this analysis on a subset of participants (n=106; including n=75 5-12yo) who had low and a matched amount of motion in both scans. We used a mixed effects regression to test for main effects of age, task (movie vs. rest), motion, and a task-by-age interaction, on the within – across network

correlation difference, per network. If the task-by-age interaction was non-significant, we repeated the regression without the interaction term and considered statistical evidence from the second regression only. Regressions included a subject identifier as a random effect in order to account for non-independence of data across the two tasks. In the full sample, the within – across network correlation difference in both networks was larger during the resting state scan, and there was a significant task-by-age interaction such that the positive effect of age on the within – across difference was stronger in the resting state (ToM: effect of task (rest > movie): b=.50, t=5.5, p=0, NS effect of age: b=.15, t=1.8, p=.08, effect of motion: b=-.19, t=-2.9, p=.005, significant task-by-age interaction: b=.26, t=2.9, p=.005; Pain: effect of task: b=.39, t=4.1, p=.0001, NS effect of age: b=.17, t=1.9, p=.06, effect of motion: b=-.15, t=-2.2, p=.03, significant task-by-age interaction: b=.27, t=2.8, p=.006). Among 5 – 12 year old children, the within – across network correlation difference in both networks did not increase with age, and was larger during the resting state scan (ToM: effect of task (rest > movie): b=.39, t=3.6, p=.0006, NS effect of age: b=.15, t=1.6, p=.11, effect of motion: b=-.21, t=-2.7, p=.008, no significant interaction; Pain: significant effect of task (rest > movie): b=.26, t=2.3, p=.03, NS effect of age: b=.17, t=1.8, p=.07, effect of motion: b=-.16, t=-2.0, p=.049, no significant interaction); see Figure 4.

Figure 4

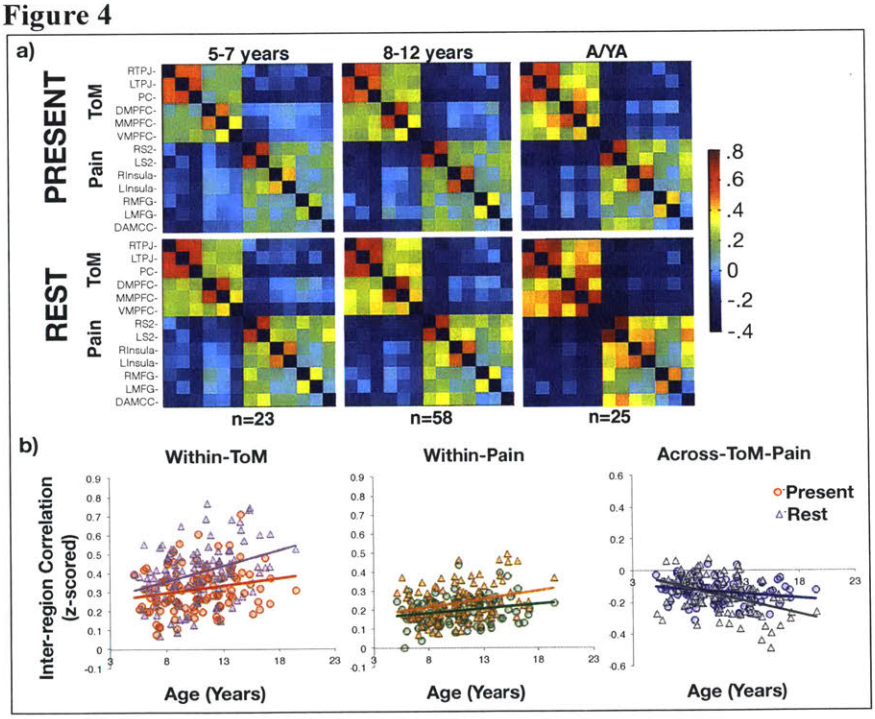


Figure 4. Inter-region Correlations during Movie Viewing and at Rest.
a) Average z-scored correlation matrices across all ToM and pain brain regions of interest (see y-axis) in low/matched motion participants, as measured while viewing Jacob Frey's "The Present"[1] (top row), or at rest (bottom row), by age group (5-7 years: n=23; 8-12 years; n=58; adolescents/young adults (A/YA): n=25). **b)** Z-scored inter-region correlations (y-axis) by age (x-axis) within the ToM network (left, red/purple), within the Pain network (middle, green/orange), and across the ToM-Pain networks (right, blue/grey). Circles show inter-region correlations as measured during Jacob Frey's "The Present"[1]; triangles show inter-region correlations as measured during rest.

Finally, we tested if inter-region correlations measured at rest were similarity correlated with the "functional maturity" of the response, as measured during movie viewing. In both networks, neither within-network or across-network inter-region correlations were significantly correlated with functional maturity (ToM: NS effect of across-TP: b=-.07, t=-.66, p=.51, NS effect of within-ToM: b=.10, t=.92, p=.36, NS effect of age: b=.15, t=1.7, p=.10, NS effect of motion: b=-.13, t=-1.4, p=.16; Pain: NS effect of across-TP: b=-.03, t=-.35, p=.73, NS effect

of within-Pain: b=.07, t=.80, p=.43, NS effect of age: b=.12, t=1.4, p=.16, effect of motion: b=-.18, t=-2.2, p=.03). The same pattern of evidence was apparent in the low/matched subset of participants (ToM: NS effect of across-TP: b=-.12, t=-.84, p=.40, NS effect of within-ToM: b=-.06, t=-.41, p=.68, NS effect of age: b=.07, t=.66, p=.51, effect of motion: b=-.25, t=-2.1, p=.04; Pain: NS effect of across-TP: b=-.06, t=-.46, p=.64, NS effect of within-Pain: b=.11, t=.81, p=.42, NS effect of age: b=.18, t=1.6, p=.12, NS effect of motion: b=-.06, t=-.52, p=.61); see Figure 3b.

We conducted lasso regressions in the low-motion subset of participants (n=106) to simultaneously test for effects of inter-region correlations as measured at rest and as measured during movie viewing on functional maturity, per network. The predictors included in these regressions were: across-TP-movie, across-TP-rest, within-[ToM or Pain]-movie, within-[ToM or Pain]-rest, age, and motion (average number of artifact timepoints across the movie and resting state scans). As determined by minimizing Mallow's Cp[18], in the ToM network, functional maturity was best predicted by a model that included across-TP-movie (b=-.07), wi-ToM-movie (b=.30), motion (b=-.01), and within-ToM-rest (-.24) predictors, in that order. In the Pain network, functional maturity was best predicted by a model that included all predictors, in the following order: across-TP-movie (b=-.09), wi-Pain-movie (b=.23), age (b=.008), across-TP-rest (b=.06), motion (b=-.004), within-Pain-rest (b=.18).

Discussion

One challenge in developmental cognitive neuroscience, developmental psychology, and cognitive neuroscience is to design and execute studies that are easily replicable[19] as well as generalizable to diverse samples[20]. "Big Data" offers one way to address this challenge, by providing large, diverse datasets that enable discovery and replication of generalizable patterns or principles of brain development. In the current study, we analyzed a diverse, publicly available fMRI dataset to replicate and extend the results of a previous exploratory fMRI study. A key goal was to determine the robustness of previously identified neural markers of brain development that relate to behavioral measures of social cognition.

We provide confirmatory evidence that ToM and pain brain regions are functionally distinct in children as young as five years of age. Inter-region correlation analyses revealed strong, positive correlations between brain regions within each network, and anti-correlated responses across the two networks. Reverse correlation analyses identified distinct events that evoked responses in each of these networks; these events were consistent with previous evidence that ToM brain regions preferentially respond to scenes that highlight mental states (beliefs, desires, emotions), and "Pain Matrix" regions preferentially respond to scenes that highlight bodily sensations (physical pain).

While responses in both networks in children were generally highly correlated with the average timecourse of responses in adolescents and young adults, we also observed significant developmental change in functional responses to the movie. Responses to one ToM event (T01) increased significantly with age. This event showed the boy, who had previously expressed annoyance at the three-legged puppy his mother gave him, softening, and feeling conflicted about softening, while watching the puppy. As in the previous study, the event that showed

change with age was a relatively long event involving complicated mental state reasoning, and was the event with the highest response magnitude in adults.

One benefit of conducting confirmatory analyses in publicly available datasets is that it tests the generalizability of results to samples that are more diverse than those typically acquired by a single lab. Indeed, the current sample had a large range of SCQ scores, and included participants whose scores are above typical cut-offs indicating social difficulties. Given the range and variability of SCQ scores, this dataset could offer a more sensitive test case for how real world variability relates to variability in neural responses. In fact, just like the in previous study, we found that magnitude of responses to a particular ToM event (T02) were correlated with SCQ score. While this result does not survive correcting for multiple comparisons across all seven ToM events, event T02 bears the most resemblance to the kind of event that was related to ToM behavior in the prior study[6]. Event T02 involves the revelation (for the audience members) that the boy, too, is missing a leg. In the context of the movie, this scene provides insight into the boy's behavior: he was initially put off by the puppy's missing leg, because he is adapting to his own new physical limitations, but eventually warms up to the puppy and feels encouraged to play outside rather than sit inside and play video games all day. As in the previous study, increased activity in ToM regions during this event may reflect children's improved ability to spontaneously consider the relevance of the current scene for past beliefs or emotions that are not explicitly marked. Together, these results signal that tasks of social cognition that are enriched for this particular demand may be ideal for relating behavioral and neural measures of ToM.

Given the large range and variability of SCQ scores, why isn't this measure more sensitive to other aspects of the functional response in ToM regions? In the previous study, proportion correct on a ToM task was correlated with inter-region correlations, functional maturity, and response magnitude to a ToM event in the ToM network; the correlation with response magnitude remained significant when additionally controlling for age. In the current study, SCQ score was correlated with the response magnitude to one ToM event (described above), but uncorrelated with the other neural measures. One possible explanation for the overall weak relationship is that the SCQ measure is not optimal for measuring individual differences in social cognition that are relevant for these neural responses. The SCQ is a parent questionnaire comprised of Yes/No questions about their child's social and communication skills[8]. Many of these questions ask parents to "focus on the time period between the child's fourth and fifth birthday," which can be challenging, especially for parents of the oldest participants. The previous study used a publicly available ToM behavioral battery to measure ToM reasoning (https://osf.io/g5zpv/), which requires children to answer prediction and explanation questions that draw on a large number of concepts in ToM (e.g., similar/diverse desires, true/false beliefs, knowledge access, moral blameworthiness, mistaken referents, non-literal speech). A second possible explanation is that apparent deficits captured by the SCQ are not caused by differences in basic processing of social stimuli, as reflected by inter-region correlations and properties of the functional response in ToM brain regions[21]. Instead, these deficits may be better captured by measures of other neural systems, like those underlying social motivation, or by measures of the interactions between different neural systems[22].

The results of the current study provide several new insights into the results of the previous study. First, while we generally replicate evidence for developmental change with age, these

trends are most apparent in a wide age range of children. For example, in the previous sample as well as in the current sample, within-network correlations showed moderate (non-significant) developmental increases between ages five to twelve years. However, expanding the age range to include younger participants (as in the previous study) and older participants (as in the current study) revealed strong evidence for developmental change with age. Thus, measuring developmental change in ToM and pain brain regions may require large samples that utilize wide age ranges. One challenge for this kind of research is designing an experimental paradigm that is suited for such wide age ranges. Movie viewing paradigms offer one promising solution to this challenge, as they are generally engaging for participants of many ages.

Second, while functional maturity (i.e., similarity to the average "adult" timecourse) was significantly correlated with the extent to which the ToM and pain brain regions were *anti-correlated* (as reported by the previous study), this measure was also significantly positively correlated with the extent to which brain regions *within* each network were correlated, in the ToM network. This pattern of results was also true in the previous study, when analyzing inter-region correlations in raw timecourses (see Supplementary Materials). Thus, it is likely that both within-network and across-network correlations contribute to the maturity of the functional response in ToM and Pain brain regions.

What is the nature of the link between the stimulus driven timecourse in ToM and pain brain regions, and the inter-region correlations within and between these two networks? We measured inter-region correlations in ToM and pain brain regions while at rest, in order to determine whether the link between functional maturity and inter-region correlations was specific to stimulus-driven responses, or reflective of intrinsic properties of these two networks. In the current dataset, the responses in the ToM and pain brain regions showed high within-network correlations and negative across-network correlations at rest, and inter-region correlations measured at rest were significantly correlated with those measured during naturalistic movie viewing. Interestingly, within-network correlations were *higher* in absence of stimuli, relative to during movie viewing. Previous studies have suggested that the extent of the correlation within- and across- ToM and pain brain regions varies by task. However, evidence regarding the direction of the effect of tasks on intrinsic correlations is mixed. Some studies report enhanced inter-region correlations during tasks, relative to rest[23]. Others, like ours, show reduced inter-region correlations during task, relative to rest[24,25]. One possibility is that the direction of this effect depends on the relevance of the content of the movie for the functional regions examined[26,27].

Critically, only inter-region correlations measured during movie viewing were significantly correlated with the functional maturity of the responses in ToM and pain networks. This is consistent with previous evidence that the correlations between "default mode" brain regions are altered during narrative processing[28]. Importantly, this result suggests that, despite the high correlation between the two measures, the differences between inter-region correlations measured during a task versus at rest are relevant for relating these measures to functional properties of the neural response.

Inter-region correlations measured during the task and at rest also had different rates of developmental change with age. In the full sample, there was a significant age by task (movie vs.

resting) interaction such that the within – across network correlation difference showed more developmental change as measured during rest, compared to during movie viewing. This is suggestive of different developmental trajectories of functional and intrinsic properties of these two networks: functional responses may develop early, and subsequently undergo more gradual change in childhood, whereas "intrinsic" properties may undergo more rapid change during childhood. While consistent with previous evidence that inter-region correlations during functional tasks can influence later correlations at rest[13,14], future longitudinal or training studies are necessary to clarify the causal order of development of functional and intrinsic network properties in ToM and pain brain regions.

In sum, the current study used strategies for analyzing functional responses during naturalistic movie viewing to replicate previous evidence concerning developmental changes in functional responses in social brain regions, and to better understand the relationship between inter-region correlations during movie viewing, "intrinsic" inter-region correlations present at rest, and functionally selective responses. Further, we demonstrate the promise of naturalistic movie viewing experiments for replicating results across research sites and samples, and for future studies of pediatric and clinical populations.

Methods

Participants

Participants were a subset of participants recruited by the Child Mind Institute. The final sample included 241 participants, 200 of whom additionally had usable resting state data. We downloaded participants from Data Releases 1.1 and 2.1 who completed both "The Present" in addition to an anatomical (T1) scan (n=322); 314 of these participants additionally completed a resting state scan. Participants were excluded from analyses for excessive motion during the scan (Present: n=7; Rest: n=45) or failed registration/lack of sufficient coverage (Present: n=74; Rest: n=66, see fMRI Data Analysis for detailed exclusion criterion). For inter-region correlation analyses, three additional participants were subsequently excluded for having outlier correlation values (see Inter-region Correlation section of Methods), leaving n=238 ("Present") and n=200 (Resting) participants for these analyses.

In order to make comparisons between this sample and a previous study[6], we grouped participants by age as children (n=186 5-12 year old participants; M(SD) age: 9.1(2.1) years, 60 females, n=153 right-handed; resting state subset: n=151, M(SD) age: 9.3(2.1) years, 50 females, n=124 right-handed) and adolescents/young adults (n=55 13-20 year old participants: M(SD) age: 15.3(1.9) years, 26 females, n=49 right-handed; resting state subset: n=51, M(SD) age: 15.4(1.8) years, 25 females, n=45 right-handed). We additionally created a low-motion subset of participants in order to directly compare response timecourses during "The Present" to those at rest (children (ages 5-12 years): n=81, M(SD) age: 9.4(2.1); full sample (ages 5-19 years: n=106, M(SD) age: 10.8(3.1) years; see Methods for more details about this subset).

All participants were recruited by the Child Mind Institute via a community-referred recruitment model[29]. All adult participants gave written consent; parent/guardian consent and child assent was received for all child participants. Recruitment and experiment protocols were approved by the Chesapeake Institutional Review Board; the Committee on the Use of Humans as

Experimental Subjects (COUHES) at the Massachusetts Institute of Technology and the Child Mind Institute approved data access and analyses.

FMRI Stimuli

During the functional MRI scan, participants watched Jacob Frey's "The Present,"[1] a 3.5-minute animated movie (https://vimeo.com/152985022). During the resting state scan, participants were instructed to keep their eyes open and fixate on a crosshair in the middle of the screen.

FMRI Data Acquisition

Prior to the scan, participants completed a mock scan in order to become acclimated to the scanner environment, and to learn how to stay still.

Whole-brain structural and functional MRI data were acquired on a 3-Tesla Siemens Tim Trio scanner located at the Rutgers University Brain Imaging Center, using the standard Siemens 32-channel head coil and CMRR simultaneous multi-slice echo planar imaging sequence. T1-weighted structural images were collected in 224 sagittal slices with .8mm isotropic voxels (%FOV Phase: 100%). Functional data were collected with a gradient-echo EPI sequence sensitive to Blood Oxygen Level Dependent (BOLD) contrast in 60 slices covering the whole brain (TR: 800ms, TE: 30ms, flip angle: 31°, multi-band acceleration: 6). Functional data during "The Present" were acquired in a single 3.5-minute run (250 volumes); resting state data were collected across two 5.1-minute runs (375 volumes per run).

FMRI Data Analysis

FMRI data were analyzed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm)[30] and custom software written in Matlab and R, using identical procedures to those used in the prior study that was the target for replication study[6]. Functional images were registered to the first image of the run; that image was registered to each participant's anatomical image, and each participant's anatomical image was normalized to the Montreal Neurological Institute (MNI) template. Registration of each individual's brain to the MNI template was visually inspected, including checking the match of the cortical envelope and internal features like the AC-PC and major sulci. All data were smoothed using a Gaussian filter (5mm kernel).

Artifact timepoints were identified via the ART toolbox (https://www.nitrc.org/projects/artifact_detect/)[31] as timepoints for which there was 1) more than 2mm composite motion relative to the previous timepoint or 2) a fluctuation in global signal that exceeded a threshold of three standard deviations from the mean global signal. Data were excluded from analyses if one-third or more of the timepoints collected (per scan type) were identified as artifact timepoints (Present: 83 TRs, n=7 participants excluded; Resting: 250 TRs, n=43 participants excluded; n=2 additional participants excluded for >83 TRs motion in truncated Resting scan). For subsequent analyses of the resting state scan, we used the first 250 TRs only, in order to match amount of data analyzed across tasks. We included number of motion artifact timepoints as a covariate in all analyses. In the current dataset, number of artifact timepoints was highly correlated with mean translation during both scans ($rs>.62$; $ps<2.2x10^{-12}$). Because this measure was not normally distributed ($ps<3.5x10^{-16}$), spearman correlations were used when including amount of motion as a covariate in partial correlations. Number of artifact timepoints (henceforth, "Motion") during "The Present" decreased significantly with age

in the full sample (Child (n=186): M(SD)=14.6(15), Adolescents/Young Adults (n=55): M(SD)=9.0(9.4), linear regression on motion: effect of age: b=-.21, t=-3.2, p=.001); the effect of age was marginal during the truncated resting state scan (Child (n=151) M(SD)=10.3(20.4), A/YA (n=51) M(SD)=8.3(24.8), linear regression on motion: effect of age: b= -.13, t=-1.9, p=.06). Among 5 – 12 year old children, motion was significantly negatively correlated with age in the resting state scan only (spearman correlation test: Present: $r_s(184)$=-.11, p=.13; Resting: $r_s(149)$=-.17, p=.04). SCQ score was not correlated with motion during either scan (spearman correlation test: Present: r_s=-.05, p=.51; Resting: r_s=-.01, p=.88). See Supplementary Figure 4 for visualization of the amount of motion in this sample.

In order to directly compare inter-region correlations during "The Present" and at rest, we created a low- and matched-motion subset of participants (n=106 participants, including n=81 5 – 12 year old children). To create this subset, we first selected participants who had fewer than 10% of timepoints identified as motion artifact in both scans (<25 timepoints); participants were subsequently excluded based on the difference in motion between the two scans, until a motion-matched sample was obtained (two-tailed paired t-test on number of artifact timepoints: children: t(80)=-.46, p=.64; full sample: t(105)=-.34, p=.74). Then, because this sample was specifically created to test for significant task-by-age interactions on inter-region correlations, we excluded the four oldest participants with the largest difference in motion between the two scans, such that the task-by-age interaction on amount of motion was non-significant (children: NS effect of task-by-age interaction: p=.12; regression on motion without interaction: NS effect of task: p=.65, NS effect of age: p=.07; full sample: NS effect of task-by-age interaction: p=.18; regression on motion without interaction: NS effect of task: p=.74, effect of age: p=.001).

Region of interest (ROI) analyses were conducted using group ROIs. ToM and pain matrix group ROIs were created in an independent group of adults (n=20), scanned by Evelina Fedorenko and colleagues, as previously described[6]. We used these group ROIs for easy comparison to the previous study.

All timecourse analyses were conducted by extracting the scaled, preprocessed timecourse from each voxel per group ROI. We applied nearest neighbor interpolation over artifact timepoints (for methodological considerations on interpolating over artifacts before applying temporal filters, see[32,33]), and regressed out two kinds of nuisance covariates to reduce the influence of motion artifacts: 1) motion artifact timepoints, and 2) five principle component analysis (PCA)-based noise regressors generated using CompCor within individual subject white matter masks[34]. White matter masks were eroded by two voxels in each direction, in order to avoid partial voluming with cortex. CompCor regressors were defined using scrubbed data (i.e. artifact timepoints were identified and interpolated over prior to running CompCor). The residual timecourses were then high-pass filtered with a cutoff of 100 seconds. Timecourses from all voxels within an ROI were averaged, creating one timecourse per group ROI, and artifact timepoints were subsequently excluded (NaNed).

Inter-region Correlation Analyses

In inter-region correlation analyses, each ROI timecourse (excluding the first three timepoints) was correlated with every other ROI's timecourse, per subject, and these correlation values were Fisher z-transformed. Within-ToM correlations were the average correlation from each ToM

ROI to every other ToM ROI, within-Pain correlations were the average correlation from each Pain ROI to every other Pain ROI, and across-network correlations was the average correlation from each ToM ROI to each Pain ROI. Based on the previous study, we defined a range of expected values for inter-region correlations. We calculated this range as the average within-ToM, within-Pain, and across-ToM Pain correlation in the 5-12 year old and adult participants from the original study, plus or minus three standard deviations (wi-ToM: -.03 – .83; wi-Pain: -.05 – .75; ac-ToM-Pain: -.55 – .51). We included adults as well as 5-12 year old children to calculate these values, in order to better suit the current sample (ages 5-20 years old). Data points that fell outside of this range were considered outliers and were excluded from inter-region correlation analyses (Present: n=3; Resting: n=11).

In order to test for developmental change in within- and across-network correlations, we conducted linear regressions to test for 1) significant effects of age (as a continuous variable) in the full sample (ages 5 – 20 years), in regressions that additionally included number of artifact timepoints as a predictor, and 2) significant effects of age (as a continuous variable), SCQ, and number of artifact timepoints among children. In order to test whether ToM and pain networks are functionally dissociated early in childhood, we used t-tests to compare within- versus across-network correlations in five-year-old children (n=16). During both types of scans, within-ToM correlations were normally distributed (Present: p=.06; Rest: p=.10; Shapiro-Wilk normality test), but within-Pain and across-network correlation measures were not (Present: ps<.0002; Rest: ps<.00005).

Reverse Correlation Analyses

Initial reverse correlation analyses of "The Present" task were conducted on adolescent/young adult participants only (n=55). Each ROI timecourse was z-normalized, and timecourses within each network were averaged across ROIs, resulting in one timecourse for the ToM network and one timecourse for the pain matrix per participant. Except for the first five timepoints (4s), the residual signal values across adult subjects for each timepoint were tested against baseline (0) using a one-tailed t-test. Events were defined as five or more consecutive significantly positive timepoints within each network (i.e., as in the previous study[6], events were at least 4s in duration). Events were rank-ordered according to the average magnitude of response to the peak timepoint, and labeled according to the ordering (i.e. event T01 is the ToM event that evoked the highest magnitude of response in the ToM network).

In order to test for developmental effects in the magnitude of response to ToM and pain events, we defined a peak timepoint per event as the timepoint with the highest average signal value in adults, and tested for significant correlations between magnitude of response at peak timepoints and age (as a continuous variable), including amount of motion (number of artifact timepoints) as a covariate. Because this measure of motion is non-normally distributed, we employed spearman correlations. For ToM regions only, we used linear regressions to test for a significant relationship between peak magnitude of response and score on the Social Communication Questionnaire (SCQ). Response magnitude to eight of ten events was normally distributed (all ps>.06, Shapiro-Wilk normality test); response magnitude to events T03 and P02 were non-normally distributed among children (ps<.02). As in the previous study[6], we additionally ran the reverse correlation analysis in the youngest children scanned (five-year-old participants; n=16).

Functional Maturity
Finally, we tested whether the functional maturity of each child's timecourse responses (i.e. similarity to adolescents/young adults) during "The Present" was related to the inter-region network correlations. We calculated the Pearson correlation between each child's ToM timecourse (averaged across ROIs) and the average adult ToM timecourse; we similarly calculated the Pearson correlation between each child's pain matrix timecourse and the average adult pain matrix timecourse. The timecourses used for this analysis were the same as those used for the reverse correlation analysis, prior to z-normalization (TRs 6:250). We tested if, across children, this measure of functional maturity per network was correlated with within-network and across-network inter-region correlations, or related to SCQ score. The neural maturity measure was normally distributed in the Pain ($p=.10$, Shapiro-Wilk normality test) but not ToM network ($p=.004$). We additionally calculated the Pearson correlation between the average timecourse of children (all children, and five year olds separately) and the average adolescent/young adult timecourse.

Comparison to Previous Results
For easy comparison to the results of the prior study, we reanalyzed the previous sample to include 5 – 12 year old children only (i.e., excluding 3 – 4 year olds). The analysis procedures were identical to those described above; information about the participants and experimental paradigm were described previously[6]. The results of these analyses are included in the Supplementary Materials.

Behavioral Measures
We used the Social Communication (SCQ[8]) score as a measure of individual differences in social cognition. We initially downloaded two phenotypic measurements collected by the Child Mind Institute that characterize social behavior: the SCQ and the Social Responsiveness Scale (SRS[35]). While the Child Mind Institute is additionally conducting ADOS screening[36], these data are not yet available for download. The SRS and SCQ measures were significantly positively correlated in the current sample ($r_s(191)=.70$, $p<2.2\text{x}10^{-16}$), even when including age as a covariate ($p<2\text{x}10^{16}$). Neither of these measures were normally distributed (Shapiro-Wilk normality test: $ps<3.4\text{x}10^{-7}$). Because these measures were highly correlated, and the SCQ task is identical for all participants (whereas younger participants complete a different version of the SRS), we used SCQ scores as the primary behavioral measure of individual differences in social cognition. Scores on the Social Communication Questionnaire were significantly correlated with age ($r_s(193)=.19$, $p=.01$), reflecting more variable and high scores among older children who contributed fMRI data to the current study.

Data availability
The fMRI and behavioral (phenotypic) data analyzed in the current study were made publicly available by and downloaded from the Child Mind Institute (http://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network/index.html; DOI: 10.1038/sdata.2017.181)[29].

Acknowledgements
We would like to thank the researchers at the Child Mind Institute for collecting, organizing, and sharing this incredibly valuable dataset, Dima Ayyash, AJ Haskins, and Lyneé Alves for helping

with visual inspection of registrations, and Todd Thompson for technical support and advice. We gratefully acknowledge support for this project by a Whitaker Health Sciences Fund Fellowship (HR).

References

1. Frey, J. *The Present* [Motion Picture]. 3:23 min. (2014).
2. Lombardo, M. V. *et al.* Shared neural circuits for mentalizing about the self and others. *Journal of Cognitive Neuroscience* **22,** 1623–1635 (2010).
3. Bruneau, E. G., Pluta, A. & Saxe, R. Distinct roles of the 'shared pain' and "theory of mind"networks in processing others' emotional suffering. *Neuropsychologia* **50,** 219–231 (2012).
4. Spunt, R. P., Kemmerer, D. & Adolphs, R. The neural basis of conceptualizing the same action at different levels of abstraction. *Social Cognitive and Affective Neuroscience* nsv084 (2015).
5. Jacoby, N., Bruneau, E., Koster-Hale, J. & Saxe, R. Localizing Pain Matrix and Theory of Mind networks with both verbal and non-verbal stimuli. *NeuroImage* **126,** 39–48 (2016).
6. Richardson, H., Lisandrelli, G., Riobueno-Naylor, A. & Saxe, R. Development of the social brain from age three to twelve years. *Nature Communications* **9,** 1027 (2018).
7. Reher, K. (Producer), & Sohn, P. (Director). *Partly Cloudy* [Motion Picture]. United States: Pixar Animation Studios and Walt Disney Pictures (2009).
8. Rutter, M., Bailey, A. & Lord, C. *The social communication questionnaire: Manual.* (Western Psychological Services, 2003).
9. Greicius, M. D., Krasnow, B., Reiss, A. L. & Menon, V. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences* **100,** 253–258 (2003).
10. Fox, M. D. *et al.* The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences* **102,** 9673–9678 (2005).
11. Miall, R. C. & Robertson, E. M. Functional imaging: is the resting brain resting? *Current Biology* **16,** R998–R1000 (2006).
12. Cole, M. W., Bassett, D. S., Power, J. D., Braver, T. S. & Petersen, S. E. Intrinsic and task-evoked network architectures of the human brain. *Neuron* **83,** 238–251 (2014).
13. Gabard-Durnam, L. J. *et al.* Stimulus-elicited connectivity influences resting-state connectivity years later in human development: a prospective study. *Journal of Neuroscience* **36,** 4771–4784 (2016).
14. Mackey, A. P., Singley, A. T. M. & Bunge, S. A. Intensive reasoning training alters patterns of brain connectivity at rest. *Journal of Neuroscience* **33,** 4796–4803 (2013).
15. Yeo, B. T. *et al.* The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology* **106,** 1125–1165 (2011).
16. van den Heuvel, M. I. & Thomason, M. E. Functional connectivity of the human brain in utero. *Trends in Cognitive Sciences* **20,** 931–939 (2016).
17. Chai, X. J., Ofen, N., Gabrieli, J. D. & Whitfield-Gabrieli, S. Selective development of anticorrelated networks in the intrinsic functional organization of the human brain. *Journal of Cognitive Neuroscience* **26,** 501–513 (2014).
18. Mallows, C. L. Some comments on C p. *Technometrics* **15,** 661–675 (1973).
19. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nat. hum. behav.* **1,** 0021 (2017).
20. Falk, E. B. *et al.* What is a representative brain? Neuroscience meets population science. *Proceedings of the National Academy of Sciences* **110,** 17615–17622 (2013).
21. Kliemann, D. *et al.* Cortical responses to dynamic emotional facial expressions generalize

across stimuli, and are sensitive to task-relevance, in adults with and without Autism. *Cortex* **103,** 24–43 (2018).
22. Kennedy, D. P. & Adolphs, R. The social brain in psychiatric and neurological disorders. *Trends in Cognitive Sciences* **16,** 559–572 (2012).
23. Vanderwal, T. *et al.* Individual differences in functional connectivity during naturalistic viewing conditions. *NeuroImage* **157,** 521–530 (2017).
24. Greene, D. J. *et al.* Behavioral interventions for reducing head motion during MRI scans in children. *NeuroImage* (2018).
25. Betti, V. *et al.* Natural scenes viewing alters the dynamics of functional connectivity in the human brain. *Neuron* **79,** 782–797 (2013).
26. Gratton, C., Laumann, T. O., Gordon, E. M., Adeyemo, B. & Petersen, S. E. Evidence for two independent factors that modify brain networks to meet task goals. *Cell reports* **17,** 1276–1288 (2016).
27. Dixon, M. L. *et al.* Interactions between the default network and dorsal attention network vary across default subsystems, time, and cognitive states. *NeuroImage* **147,** 632–649 (2017).
28. Simony, E. *et al.* Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications* **7,** (2016).
29. Alexander, L. M. *et al.* The healthy brain network biobank: an open resource for transdiagnostic research in pediatric mental health and learning disorders. bioRxiv. *bioRxiv* (2017).
30. Friston, K. J. Statistical parametric mapping. (1994).
31. Whitfield-Gabrieli, S., Nieto-Castanon, A. & Ghosh, S. Artifact Detection Tools (ART). *Cambridge, MA. Release version* **7,** 11 (2011).
32. Carp, J. Optimizing the order of operations for movement scrubbing: Comment on Power et al. *NeuroImage* **76,** 436–438 (2013).
33. Hallquist, M. N., Hwang, K. & LUNA, B. The nuisance of nuisance regression: spectral misspecification in a common approach to resting-state fMRI preprocessing reintroduces noise and obscures functional connectivity. *NeuroImage* **82,** 208–225 (2013).
34. Behzadi, Y., Restom, K., Liau, J. & Liu, T. T. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* **37,** 90–101 (2007).
35. Constantino, J. N. & Gruber, C. P. *Social responsiveness scale (SRS).* (Western Psychological Services Torrance, CA, 2012).
36. Lord, C. *et al.* The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *J Autism Dev Disord* **30,** 205–223 (2000).

Supplementary Materials

Reanalysis of Previous Study: 5 – 12 Year Old Participants Only

For the closest comparison of the results of the previous study to the current results, we re-analyzed the previous study excluding the three and four year old children.

Developmental Change in Inter-Region Correlations

In the previous study, evidence for developmental increases in within-network inter-region correlations depended on including the youngest children scanned (ages 3-4 years old); within-network correlations were not significantly correlated with age among the 5-12 year old children (spearman partial correlations including motion as covariate: ToM: $r_s(88)$=-.03, p=.75; Pain: $r_s(88)$=.08, p=.45). However, evidence for decreases in across-network correlations with age remained significant in 5-12 year old children ($r_s(88)$=-.35, p=.0007).

Functional Maturity

In the previous study, "functional maturity" (i.e., similarity to responses in adults) among 5 – 12 year olds increased with age in both networks, and was significantly positively correlated with the extent to which the ToM and Pain networks were *anti-correlated* during the task (controlling for age, within-network correlations, and motion). That is, children who had more anti-correlated ToM and Pain response timecourses also had timecourses that were more similar to adult participants. These results remained significant among 5 – 12 year old participants (excluding 3 – 4 year olds): functional maturity increased with age (spearman partial correlations including motion as a covariate: ToM: $r_s(91)$=.38, p=.0002; Pain: $r_s(91)$=.48, p=$1.5x10^{-6}$) and was predicted by the anti-correlation of responses in the two networks (linear regressions testing for effects of across-network correlation, within-network correlation, age, and motion on functional maturity measure (n=91): ToM: effect of across-network correlation: b=-.7, t=-7.3, p=$1.4x10^{-10}$, effect of within-ToM correlation: b=.02, t=.28, p=.007, NS effect of age: b=.13, t=1.6, p=.10, effect of motion: b=-1.7, t=-2.1, p=.04; Pain: effect of across-network correlation: b=-.69, t=-6.7, p=$1.7x10^{-9}$, NS effect of within-Pain correlation: b=.22, t=2.3, p=.03, effect of age: b=.22, t=2.6, p=.01, NS effect of motion: b=-.10, t=-1.4, p=.17). However, these results show that within-network correlations also positively predicted functional maturity.

Reanalysis of Previous Study: 5 – 12 Year Olds Only & M1 Timecourse Regression

In the previous study, primary inter-region correlation analyses were conducted on residual response timecourses, after regressing out the average bilateral primary motor (M1) cortex timecourse. Inter-region correlation analyses of the raw timecourses were included as supplementary analyses. In the current study, we conducted inter-region correlation analyses on the raw timecourses (without regressing out the M1 timecourse). Below we report the re-analysis of the original sample, excluding three and four year old participants, using residual timecourses (with M1 regressed out).

Developmental Change in Inter-Region Correlations

As in the analyses of the raw timecourse (above), evidence for developmental increases in within-network inter-region correlations in the previous dataset depended on including the youngest children scanned (ages 3-4 years old); within-network correlations did not change with age among the 5-12 year old children (spearman partial correlations including motion as

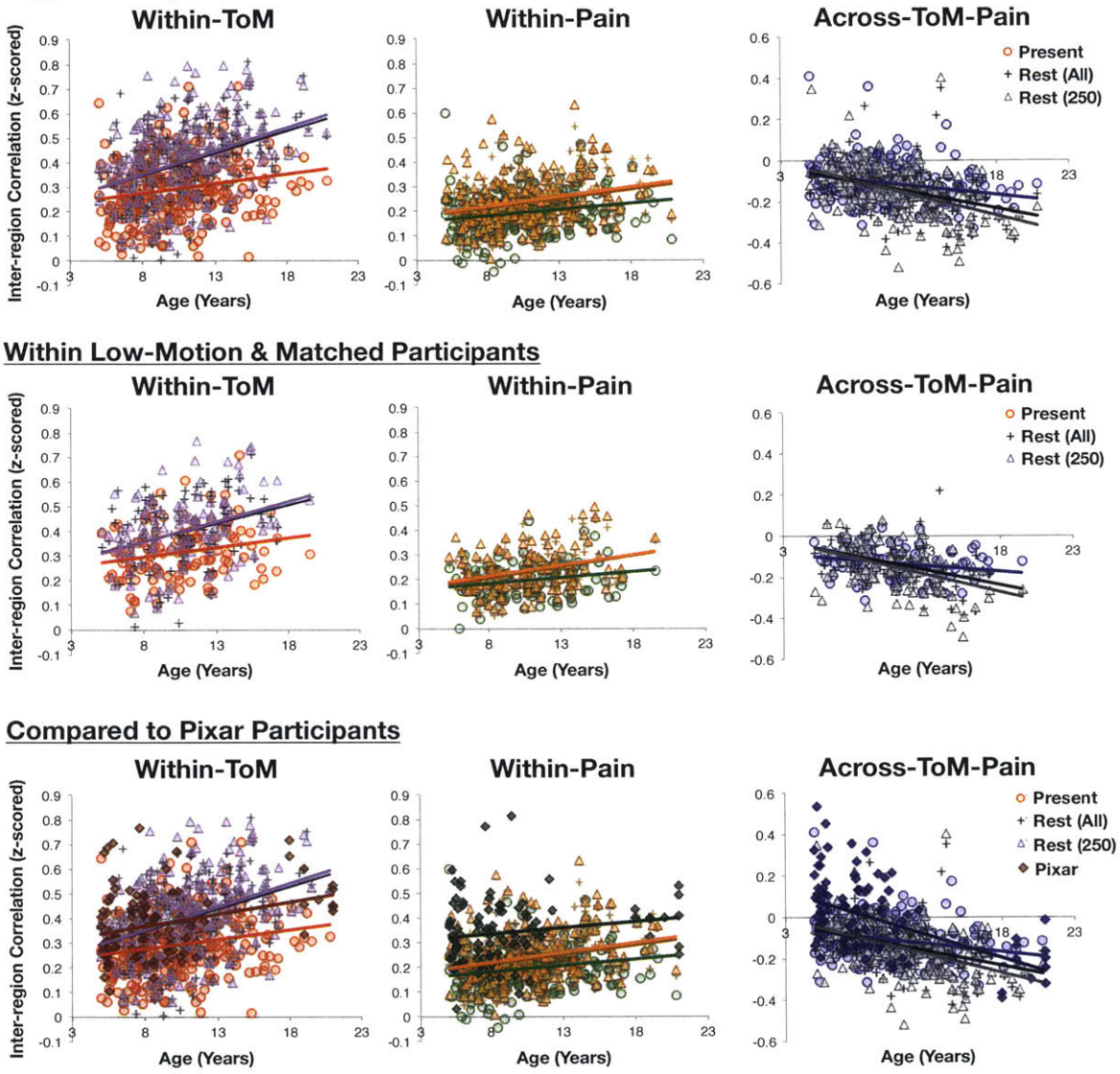
covariate: ToM: $r_s(88)$=-.03, p=.75; Pain: $r_s(88)$=.13, p=.21). However, evidence for decreases in across-network correlations with age remained significant in 5-12 year old children ($r_s(88)$=-.35, p=.0007).

Functional Maturity Analysis

The previous study suggested that across-network inter-region correlations (and not within-network inter-region correlations) were significantly correlated with functional maturity. This result remains the same in an analysis of 5 – 12 year old participants only (linear regressions testing for effects of across-network correlation, within-network correlation, age, and motion on functional maturity measure (n=91): ToM: effect of across-network correlation: b=-.5, t=-6.3, $p=1.1x10^{-8}$, NS effect of within-ToM correlation: b=.02, t=.22, p=.82, effect of age: b=.2, t=2.1, p=.04, effect of motion: b=-.2, t=-2.6, p=.01; Pain: effect of across-network correlation: b=-.57, t=-7.1, $p=3.1x10^{-10}$, NS effect of within-Pain correlation: b=.09, t=1.2, p=.22, effect of age: b=.3, t=3.3, p=.001, NS effect of motion: b=-.08, t=-1.0, p=.3).

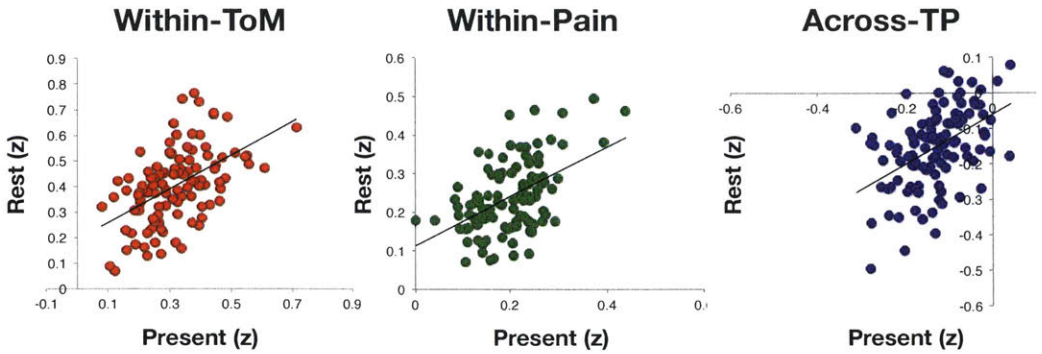
Supplementary Figures and Tables

Supplementary Figure 1



Supplementary Figure 1. Inter-region Correlations during Movie Viewing and at Rest. All scatter plots show z-scored inter-region correlations (y-axis) by age (x-axis) within the ToM network (left, red/purple), within the Pain network (middle, green/orange), and across the ToM-Pain networks (right, blue/grey). Circles show inter-region correlations as measured during Jacob Frey's "The Present"[1]; triangles show inter-region correlations as measured during the length-matched resting state scan (250 TRs; included in all main analyses); plus signs show inter-region correlations as measuring during the full resting state scan (750 TRs). The top row includes data from all participants (n=238 for "The Present", n=200 for resting). The middle row includes data from the low/matched motion subset of participants (n=106). The bottom row shows all data from the current sample (identical to the top row) as well as data from the prior study[6] ("Pixar", diamond data points). Pixar participants include n=91 children (5 – 12 years old) and n=11 adults (18-21 years old). Older adults (n=22, ages 22-39 years) and younger children (n=31 3 – 4 year olds) were excluded from these plots in order to better match the age range of the current sample.

Supplementary Figure 2

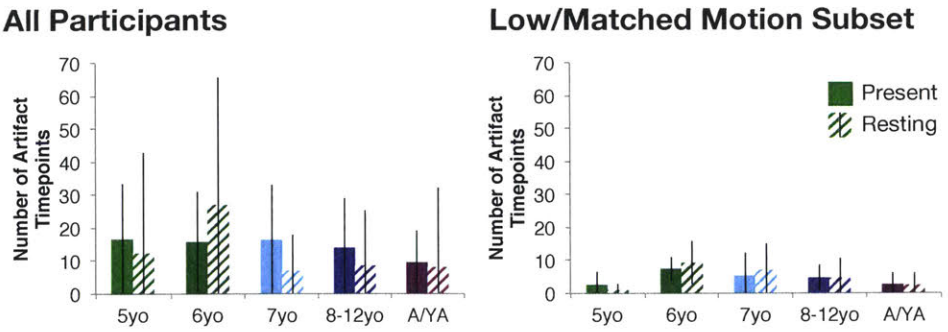


Supplementary Figure 2. Inter-region Correlations during Movie Viewing and at Rest are Correlated. Scatter plots show z-scored inter-region correlations as measured during "The Present" (x-axis), by those measured during rest (y-axis) within the ToM network (left, red), within the Pain network (middle, green), and across the ToM-Pain networks (right, blue), in the low/matched motion subset of participants (n=106).

Supplementary Figure 3

Supplementary Figure 3. Reverse Correlation Events. Thumbnail images of each event identified by the reverse correlation analysis of the timecourse of response during "The Present" in adolescent and young adult participants (ages 13-20 years, n=55). Thumbnails are shown in order of presentation during the movie; event names reflect the rank order of average response magnitude to the peak timepoint during each event. See Supplementary Table 1 for descriptions and timing information. Thumbnail images used with permission from Jacob Frey.

Supplementary Figure 4



Supplementary Figure 4. Amount of Motion during FMRI Scans. Plots show mean number of artifact timepoints (>2mm motion, > 3 standard deviation from average global signal) per age group during "The Present" (solid bars) and during the resting state scan (striped bars). The plot on the left includes all participants (n=241); the plot on the right includes the low/matched motion subset of participants (n=106). Error bars show standard deviation from the mean.

Supplementary Table 1

	Event	Time in Movie (m:s:ms)	Duration (s)	Peak Timepoint (TR)	Description
ToM Events	T01	2:13:80 - 2:21:00	140	175	Boy is softening, seems conflicted. Watches puppy.
	T02	2:58:60 - 3:10:60	190.4	238	Boy heads towards door. It becomes clear that he is missing a leg, like the puppy. Puppy and boy go outside together.
	T03	1:52:20 - 1:56:20	117.6	147	Boy looks annoyed; puppy heads towards ball in box.
	T04	1:21:00 - 1:25:00	86.4	108	Boy looks annoyed; puppy looks around.
	T05	0:41:80 - 0:46:60	45.6	57	[Mom just told boy to open present]. Boy says "for me?" and looks at box.
	T06	1:40:20 - 1:44:20	105.6	132	Boy notices ball; puppy approaches boy expectantly.
	T07	0:10:60 - 0:14:60	17.6	22	Boy playing video game.
Pain Events	P01	1:31:40 - 1:39:40	99.2	124	Puppy slams into cabinet. Boy wipes nose.
	P02	1:04:20 - 1:09:80	72	90	Boy notices missing leg and tosses puppy to floor. Boy hits present box.
	P03	2:24:20 - 2:32:20	154.4	193	Puppy carries ball over to boy, fumbling a bit because of his missing leg.

Supplementary Table 1. Reverse Correlation Events. Table includes the name, time, duration, peak timepoint, and description of each event identified by the reverse correlation analysis of the timecourse of response during "The Present" in adolescent/young adult participants (n=55). Event name indicates the rank of the event (T01 is the event with the highest peak magnitude of response, T02 the second highest, etc.). See Supplementary Figure 3 for thumbnail images of each event.

Chapter 4: Language Facilitates Theory of Mind Development: Behavioral and Neural Evidence from Individuals with Delayed Access to Language

Language abilities are clearly related to performance on theory of mind (ToM) tasks in childhood, but the precise role of language in ToM development continues to be debated. Language could play a causal role in ToM development, either by facilitating conversations about the minds of others and/or by enabling sophisticated representations of mental states. Alternatively, it is possible that language proficiency is simply a task demand of most ToM measures and is not otherwise required for ToM development. One difficulty in teasing apart these alternatives is that language and ToM develop in tandem for most children. Deaf children who use sign language offer a way to address this question because while they are otherwise neurotypical, they have varying ages of exposure to an accessible language, corresponding to when they were first exposed to sign language. Indeed, deaf children with delayed exposure to language have delayed ToM development, specifically false-belief understanding, compared to both hearing children and to deaf children exposed to sign language from infancy. Here, we used verbal and non-verbal measures of behavioral ToM alongside fMRI to investigate ToM development in signing children (4-12 years old) as a function of the age at which they were first exposed to a sign language (birth-7 years). In addition, we investigated the effect of age of sign language exposure on the neural signature of ToM processing in d/Deaf adults. Among children, delayed access to language was associated with reduced scores on verbal tests of advanced ToM reasoning (e.g., second order false belief, moral blameworthiness). The effect of delayed language was limited to verbal tasks of advanced ToM reasoning, and was not observed in conceptually analogous non-verbal ToM tasks. Responses in brain regions associated with ToM showed reduced selectivity to mental state stimuli in children who had delayed access to language in a verbal story task, but were indistinguishable from those of native signing participants during a nonverbal naturalistic viewing task. The difference in response selectivity during the verbal task was not present in adults. While the effects of delayed access to language are most apparent in verbal contexts, they likely reflect a facilitative role for language on expression and development of ToM reasoning. Importantly, effects of delayed access to language on ToM are no longer present in adults who gain early access to and proficiency in sign language.

Note: This study is in collaboration with Jorie Koster-Hale, Naomi Caselli, Rachel Magid, Rachel Benedict, Halie Olson, Jennie Pyers, and Rebecca Saxe.

Introduction

The human ability to represent and reason about the internal mental states of others is described as having a "Theory of Mind" (ToM): a rich, structured theory that enables us to link observable behaviors to unobservable beliefs, desires, and emotions. Like many cognitive capacities, this theory improves dramatically during early childhood. While there is evidence that this improvement is domain-specific[1-3], environmental and experiential factors contribute to social cognitive change. For example, language abilities in childhood are strongly correlated with and predictive of performance on early ToM milestones[4-6]. But the precise role of language on theory of mind development continues to be debated. Language could play a causal role in ToM development, either by facilitating conversations about the minds of others[7,8] and/or by enabling sophisticated representations of mental states[9-11]. Alternatively, it is possible that language proficiency is simply a task demand of most ToM measures (e.g., for expression of ToM competence) but is not otherwise required for ToM development. Evidence in support of this view comes largely from studies showing simultaneous success on non- or low-verbal ToM tasks and failure on verbal ToM tasks[12,13] or early success on versions of ToM tasks with reduced linguistic demands[14-16].

One challenge in teasing apart these alternatives is that language and ToM develop in tandem for most children. Behavioral studies have made progress by studying ToM development in children with varying amounts of linguistic input. For example, maternal use of mental state vocabulary predicts early false-belief reasoning in hearing children[17] and in children who are d/Deaf[18]. These studies suggest that the amount of linguistic input directly impacts – and accelerates – early ToM development. A second strategy is to study individuals who have varying ages of exposure to an accessible language. For example, while d/Deaf individuals are otherwise neurotypical, many deaf people experience language deprivation: limited access to any language– spoken or signed – during childhood. Indeed, deaf children with delayed exposure to language have delayed ToM development, specifically false-belief understanding, compared to both hearing children and to deaf children exposed to sign language from infancy[8,19-21].

Using neuroimaging to measure the functional responses in "social" brain regions in d/Deaf individuals could offer key insight into this debate. Human adults and children recruit a specific network of brain regions when reasoning about the minds of others[22] (for reviews, see[23,24]). These brain regions respond preferentially to mental state content, and response selectivity (e.g., the specificity and extent of this preference) increases in early childhood[25-27]. The two hypotheses described above – that language is either critical for or superficially related to ToM development – make distinct predictions about the neural responses in these brain regions. For example, if early language input is critical to normal theory of mind development, a strong interpretation would predict that brain regions specialized for ToM would not develop without this input. Correspondingly, delayed access to language would result in temporary developmental delays or even lifelong differences in the preferential responses to mental state content in these regions. On the other hand, if language is only superficially related to theory of mind development, responses in ToM brain regions should be similar despite difficulties on verbal tasks of ToM for individuals with delayed access to language.

The current study uses fMRI in addition to behavioral testing to characterize the impact of delayed access to language on ToM development, behaviorally and neurally. In addition to

providing novel neuroimaging evidence, the current study builds on previous research in two key ways. First, we focus on children who are currently fluent and proficient in ASL. Additionally, non-native signing children were exposed to sign language relatively early in development (before age 7 years). This focus makes potential deficits in ToM easier to interpret: measured deficits in ToM reasoning are less likely to be driven by linguistic task demands in a sample that is matched on current ASL proficiency. This focus additionally provides important information to clinicians and educators about the potential impact of early exposure to ASL on social development in childhood.

Second, we measure behavioral ToM reasoning using tasks that include relatively advanced ToM concepts, such as reasoning about moral blameworthiness during accidents, non-literal speech, and second-order false-beliefs. Almost all prior studies on ToM reasoning in d/Deaf children with delayed access to language focuses on false-belief task performance. The false-belief task was developed to be a diagnostic measure for whether an individual is capable of representing the contents of another mind[28-30], and has proven to be a useful tool for studying developmental change in theory of mind in typical development[31], in clinical populations[32,33], and across cultures[34]. However, reasoning about others' minds includes progressively more sophisticated concepts than (false) beliefs that children go on to master after age five years[35,27]. The current study uses behavioral and neural measurements of ToM reasoning, including concepts that develop before and after false-belief reasoning, in verbal and non-verbal contexts, in order to characterize the impact of delayed access to language on ToM reasoning.

Results

We used verbal and non-verbal tasks to measure ToM, behaviorally and neurally, in native signing (NS; n=21, 4-12.7 years old, M(SD) = 8.19 (2.2) years, 10 female), and early signing (ES; n=12 6.2-12.1 years old, M(SD) = 9.29 (1.9) years, 5 female) children. Native signing children received exposure to ASL from birth, whereas ES children received exposure to ASL after an initial delay of .25 – 7 years (M(SD) = 2.9 (2.2) years). We additionally measured neural responses in native signing (n=20, 20-54 years old, M(SD) = 30.2(9.5), 13 female) and non-native signing (n=16, 21-64 years old, M(SD) = 38.1(12.7) years, 4 female) adults (see Methods and Supplementary Table 1 for additional details).

Behavioral Results

All participants were recruited via a screening process that ensured proficiency in ASL. Child participants additionally completed a measure of receptive ASL (the ASL-RST). Receptive ASL proficiency increased with age ($r(29)=.54$, $p=.002$), but there was no difference in receptive ASL as a function of the age of ASL onset (NS effect of age of ASL onset: b=-.09, t=-.55, p=.59, positive effect of age: b=.57, t=3.4, p=.002).

Given that some participants experienced delayed access to language, we measured ToM reasoning using both verbal (ToM_V) and non-verbal (ToM_{NV}) tasks. The verbal task involved the experimenter telling a story in ASL, and required children to answer prediction and explanation questions about the mental states of the characters. This task was largely based on an English version used to measure ToM behavior in hearing children (https://osf.io/g5zpv/)[27]. The non-verbal task involved watching an experimenter place a series of three to five pictures on a board, which presented characters undergoing a sequence of events. After the initial sequence, the

experimenter would place two pictures side by side and use the prompt "What comes next?" for action or emotion prediction items, or say "You decide- is this good (pointing to thumb up), bad (pointing to thumb down), or okay (point to neutral thumb)?" for moral reasoning items. Children responded via pointing to one of two pictures (action/emotion prediction items), or to a picture of a thumbs up, thumbs down, or neutral thumb (moral reasoning items). The script and materials for the verbal and non-verbal ToM tasks are publicly available via the Open Science Framework (OSF; https://osf.io/kyu3f/)

Across all children, performance on both tasks was significantly positively correlated with age (ToM_V: $r(26)=.40$, $p=.04$; ToM_{NV}: $r(31)=.63$, $p=9.3x10^{-5}$). Additionally, performance on the verbal and nonverbal ToM tasks was significantly positively correlated ($r(26)=.63$, $p=.0003$)). This relationship remained significant when controlling for age (effect of ToM_{NV}: $p=.004$; NS effect of age: $p=.61$) and age and receptive ASL score (effect of ToM_{NV}: $p=.007$; NS effect of age: $p=.61$; NS effect of ASL: $p=.87$).

Critically, we tested if performance on either ToM task varied based on age of ASL onset. There was a significant negative effect of age of ASL onset on verbal ToM performance, such that children who experienced longer delays before exposure to ASL performed worse on the verbal ToM task (negative effect of age of ASL onset: $b=-.54$, $t=-3.4$, $p=.002$; positive effect of age: $b=.56$, $t=3.5$, $p=.002$; see Figure 1a, and Supplementary Figure 1). This effect remained significant when additionally including ASL proficiency as a covariate (negative effect of age of ASL onset: $b=-.52$, $t=-3.2$, $p=.004$, positive effect of age: $b=.51$, $t=2.8$, $p=.009$, NS effect of receptive ASL: $b=.10$, $t=.59$, $p=.56$). In follow-up analyses, there was no effect of ASL-onset on control items (NS effect of age of ASL onset: $b=-.12$, $t=-.66$, $p=.52$; positive effect of age: $b=.48$, $t=2.6$, $p=.02$). This effect appeared to be driven by more advanced items (conceptually and linguistically, see Figure 1b). There was no such effect of age of ASL onset on non-verbal ToM task performance (NS effect of age of ASL onset: $b=-.21$, $t=-1.5$, $p=.15$; positive effect of age: $b=.70$, $t=4.8$, $p=4.1x10^{-5}$). See Figure 1c-d.

We additionally tested for differences in non-verbal IQ, spatial working memory, and response inhibition as a function of age of ASL onset. Age of ASL onset was not correlated with spatial working memory span (NS effect of age of ASL onset: $b=-.27$, $t=-1.4$, $p=.18$; positive effect of age: $b=.61$, $t=3.2$, $p=.004$), but was significantly negatively correlated with standardized non-verbal IQ score in this sample ($b=-.38$, $t=2.3$, $p=.03$). The difference based on age of ASL onset in the verbal ToM task remained significant when additionally including standardized non-verbal IQ as a covariate (negative effect of age of ASL onset: $b=-.33$, $t=-2.2$, $p=.04$; positive effect of age: $b=.38$, $t=2.7$, $p=.01$; positive effect of non-verbal IQ: $b=.42$, $t=2.8$, $p=.009$).

Behavioral Results: Exploratory Analyses

Given reduced performance in early signers during verbal but not non-verbal ToM tasks, we conducted follow-up analyses to test for effects of age of ASL onset on ToM performance by concept category. Verbal and non-verbal tasks both included (1) "easy" items, which involved reasoning about desires, emotions, and true beliefs, (2) false belief items, and (3) moral judgment items. The verbal task additionally included items that involved reasoning about (4) lies and second-order false beliefs, (5) mistaken referents, and (6) non-literal speech (e.g., sarcasm).

Figure 1

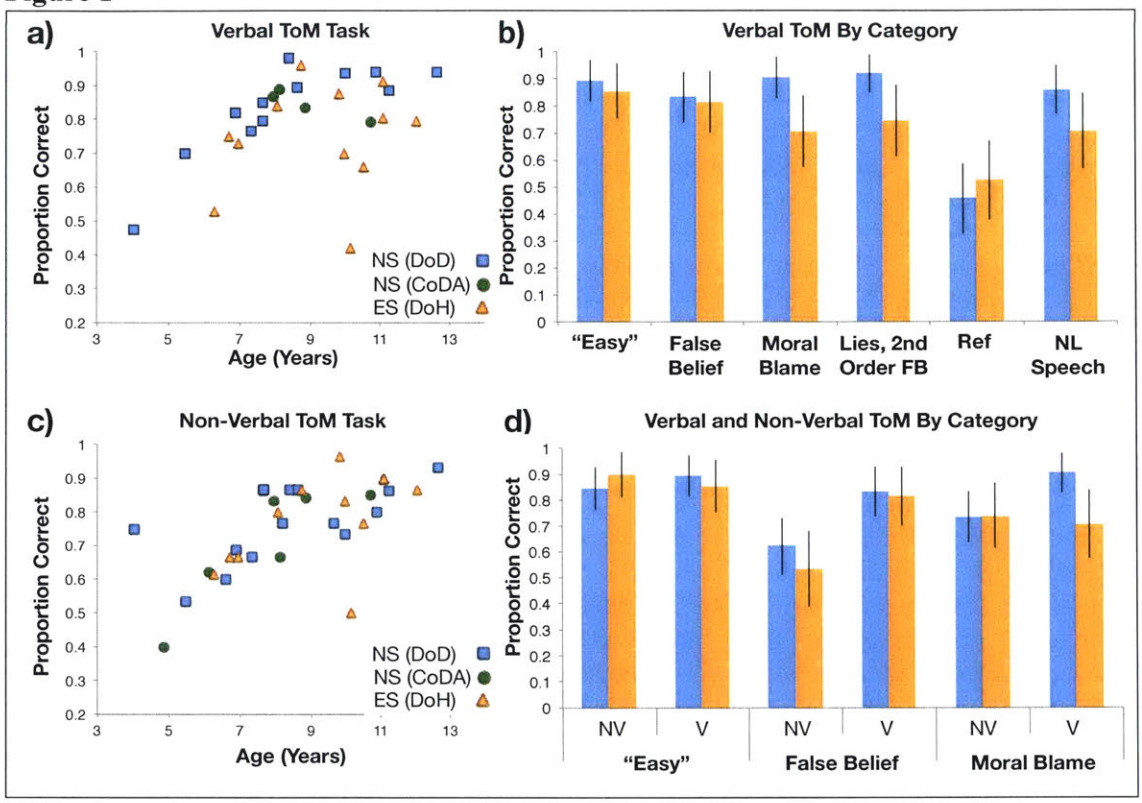


Figure 1. Theory of Mind Behavior a) Proportion correct on verbal (ASL) ToM task (y-axis) by age (x-axis). Blue squares show deaf native signers who were born to d/Deaf parents (DoD); Green circles show hearing native signers who were born to d/Deaf parents (CoDA); Orange triangles show deaf early signers, all of whom were born to hearing parents (DoH). **b)** Mean proportion correct for native (blue) and early (orange) signers, on verbal (ASL) ToM task (y-axis), by question category. Error bars show standard error from the mean. **c)** Proportion correct on non-verbal ToM task (y-axis) by age (x-axis). Blue squares show deaf native signers who were born to d/Deaf parents (DoD); Green circles show hearing native signers who were born to d/Deaf parents (CoDA); Orange triangles show deaf early signers, all of whom were born to hearing parents (DoH). **d)** Mean proportion correct for native (blue) and early (orange) signers on non-verbal (NV) and verbal (V) ToM tasks, by (analogous) question categories. Error bars show standard error from the mean.

These categories were created to sensitively capture differences in conceptual content, while minimizing the total number of categories (maximizing items per category). We plotted proportion correct on these item categories per group (native signers and early signers), in order to visualize which categories contributed to the performance difference by age of ASL onset in the verbal task. This visualization suggested that reduced performance on the verbal task was driven by more advanced ToM items: those that involved making moral judgments based on intent, and reasoning about lies, second-order false-beliefs, and non-literal speech (see Figure 1b).

Next, we conducted post-hoc linear regressions to test for effects of age of ASL onset and age on false-belief items and moral judgment items. There was no difference in proportion correct for false-belief items based on age of ASL onset in either task (Verbal FB: NS effect of ASL onset: b=-.27, t=-1.7, p=.10, effect of age: b=.65, t=3.9, p=.0007; Non-Verbal FB: NS effect of ASL

onset: b=.005, t=.03, p=.98, marginal effect of age: b=.36, t=2.0, p=.057). In a mixed effect linear regression that additionally tested for an effect of task (including subject as a random identifier), and a task-by-age of ASL onset interaction, there was an effect of task such that performance was higher during the verbal task, in addition to a positive effect of age, and no effect of age of ASL onset (NS effect of ASL onset: b=-.02, t=-.11, p=.91, effect of task: b=.79, t=3.7, p=.001, effect of age: b=.20, t=3.7, p=.0009, NS task*ASL-onset interaction: b=-.17, t=-.80, p=.43; see Figure 1d).

For moral items, age of ASL onset was negatively correlated with performance in the verbal task (effect of ASL onset: b=-.39, t=-2.1, p=.04, NS effect of age: b=.18, t=.90, p=.38), but there were no differences based on age of ASL onset in the non-verbal task (NS effect of age of ASL onset: b=-.14, t=-.76, p=.45, effect of age: b=.39, t=2.2, p=.04). In the subsequent mixed effects linear regression, age and the task-by-age of ASL onset interaction were marginal predictors of performance on moral items (NS effect of ASL onset: b=-.11, t=-.60, p=.55, effect of task: b=.30, t=1.5, p=.15, marginal effect of age: b=.29, t=-1.7, p=.057, marginal task*ASL-onset interaction: b=-.35, t=-1.7, p=.09; see Figure 1d).

FMRI Results: ASL Story Task

During the ASL Story task, participants watched movies of a woman telling stories in ASL, which involved characters and their mental states (Mental condition), characters and their physical appearance or social relationships (Social condition), or descriptions of physical objects and events in the world (Physical condition). Based on behavioral ratings, stories were matched across conditions for linguistic features (e.g., syntactic complexity, number of signs, number of verbs), psychological features (e.g., how easy to understand, how interesting), and imageability. Stories were told using simple language, in an enthusiastic, narrative way. Participants would see the main story, followed by a pause, and then a final sentence that was either a natural continuation of the story, or an ending drawn from a different story stimulus. Participants used a button box to indicate whether the ending followed the initial story segment or not. We measured three properties of the neural response in ToM brain regions during the story task: response selectivity in ToM brain regions, inter-region correlations between ToM and language brain regions, and response lateralization in the temporal lobe. Story task analyses were pre-registered (https://osf.io/kyu3f/).

First, we measured response selectivity. Previous work suggests that ToM brain regions become increasingly functionally selective during early childhood[25-27]. We tested if delayed access to linguistic input results in delayed or disrupted functional specialization of ToM brain regions. Among child participants, there was a significant negative effect of age of ASL onset, such that children who had a longer delay before accessing ASL have less selective responses in RTPJ and DMPFC (negative effect of age of ASL onset: b=-.34, t=-2.1, p=.049; marginal positive effect of age: b=.31, t=2.0, p=.06; NS effect of ROI: b=.02, t=.08, p=.94; NS effect of motion: b=.20, t=1.3, p=.21; See Figure 2a and Supplementary Figure 2). While no variables predicted selectivity in a standard mixed effects regression additionally including non-verbal IQ as a covariate (all ps>.15), a lasso regression (which excludes predictors in order to determine the best model fit), suggested that a model with all predictors best explained response selectivity (in order of inclusion: nonverbal IQ: b=6.40; age of ASL onset: b=-5.83; motion: b= 6.76; age: b=5.95). The negative effect of age of ASL onset was significant in the full sample (children and

adults; negative effect of age of ASL onset: b=-.30, t=-2.2, p=.03; NS effect of age group: b=-.18, t=-.85, p=.40; NS effect of ROI: b=-.15, t=-.84, p=.40; positive effect of motion: b=.29, t=2.1, p=.03), but there was no effect of delayed access to language on response selectivity among adults alone (NS effect of age of ASL: b=-.24, t=-1.23, p=.23; NS effect of ROI: b=-.24, t=-1.0, p=.32; NS effect of motion: b=.22, t=1.2, p=.26). We did not find evidence for a relationship between response selectivity and performance on either ToM behavioral task (ToM_V: NS effect of ToM_V: b=.18, t=.97, p=.35, NS effect of ROI: b=-.10, t=-.34, p=.74, NS effect of motion: b=.14, t=.78, p=.45; ToM_{NV}: NS effect of ToM_{NV}: b=.29, t=1.8, p=.08, NS effect of ROI: b=-.03, t=-.12, p=.91, NS effect of motion: b=.29, t=1.9, p=.08). Supplementary analyses of group regions of interest show a similar pattern of results (see Supplementary Materials and Supplementary Figure 2).

Second, we measured the lateralization of the suprathreshold response (p<.001) to the Mental > Physical contrast in the temporal lobe. In children and adults, the response to Mental > Physical was not lateralized to either hemisphere (M(SE) laterality index: adults: .10(.04), children: -.08(.05); t-test (mu=0): ps>.06). Responses in children were marginally less left-lateralized than adults (marginal effect of age group: b=-.48, t=-1.9, p=.07, NS effect of motion: b=.05, t=.42, p=.68). Among children, there was no effect of age (NS effect of age: b=.02, t=.09, p=.93, NS effect of motion: b=-1.7, t=-.78, p=.44) or ToM (ToM_V: NS effect of ToM_V: b=.07, t=.38, p=.71, NS effect of motion: b=-.17, t=-.80, p=.43; ToM_{NV}: NS effect of ToM_{NV}: b=.21, t=.99, p=.33, NS effect of motion: b=-.16, t=-.75, p=.46) on response lateralization. Critically, response lateralization did not differ as a function of age of ASL onset, either in the full group (NS effect of age of ASL onset: b=.15, t=.85, p=.40, NS effect of age group: b=-.41, t=-1.5, p=.15, NS effect of motion: b=-.05, t=-.27, p=.79), or among children (NS effect of age of ASL onset: b=.15, t=.68, p=.51, NS effect of age: b=-.87, t=-.04, p=.97, NS effect of motion: b=-.14, t=-.63, p=.54) or adults (NS effect of age of ASL onset: b=.006, t=.02, p=.98, NS effect of motion: b=.16, t=.60, p=.56; see Figure 2b). Analyses conducted at a more lenient threshold (p=.01) yielded the same pattern of results.

Finally, we measured the extent to which responses in the ToM network and the language network were correlated during this task. Correlated response timecourses among brain regions could reflect similar functional selectivity profiles (two regions activate and deactivate to the same content within the stimulus), information transfer or division of labor between regions (two regions work concurrently to process different aspects of the same stimulus), and/or intrinsic network properties (two regions activate and deactivate together regardless of stimulus).

We measured inter-region correlations in ToM and language brain regions, in both age groups. Because of paradigm differences between children and adults, inter-region correlation analyses were conducted in each age group separately. Within both age groups, responses of brain regions within each network were highly correlated (M(SE) of within-ToM correlations: children: .19(.03), adults: .36(.03); M(SE) of within-Language correlations: children: .25(.03), adults: .31(.02)). Response timecourses in language and ToM brain regions were also significantly positively correlated across-network during this task, in both age groups (M(SE) of across-ToM-Lang correlations: children: .15(.02); adults: .17(.02); t-test against zero (mu=0): children: $t(23)=7.2$, $p=2.6\times10^{-7}$; adults: $t(35)=10.5$, $p=2.1\times10^{-12}$). Language brain regions were

significantly more correlated with other regions within the language network, relative to regions in the ToM network, in both age groups (within vs. across-network correlations: adults:

Figure 2

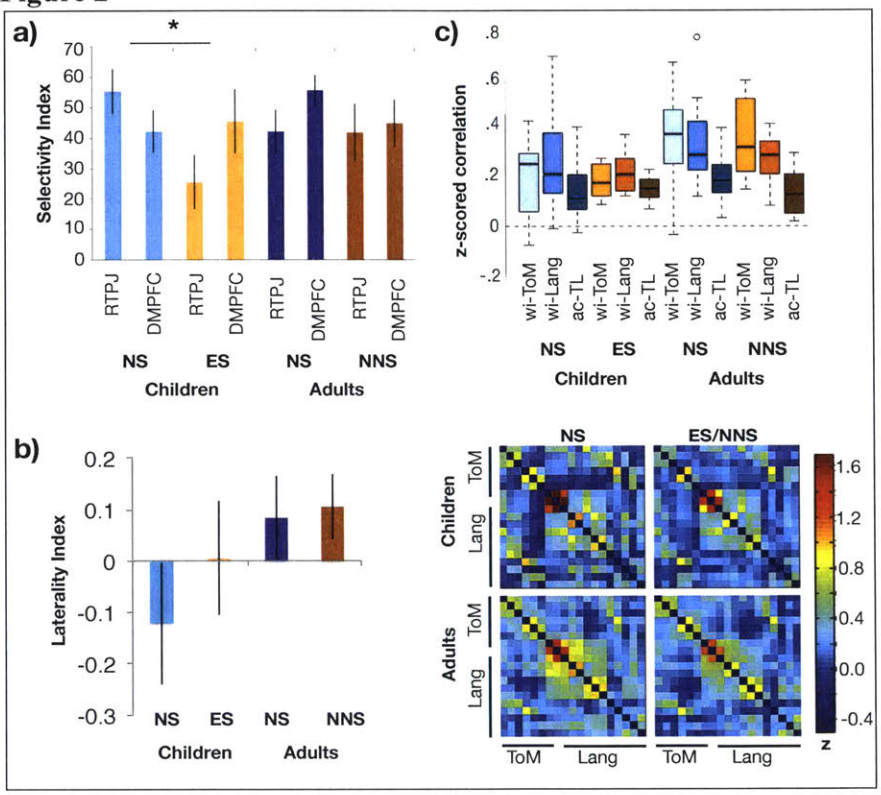


Figure 2. Neural Responses to Story Task a) Bars show the mean selectivity index for the response in RTPJ and DMPFC, in native signing (NS; light blue) and early signing (ES; orange) children, and native signing (NS; dark blue) and non-native signing (NNS; brown) adults. Selectivity index was calculated as the average beta estimate to (Mental – Social / Mental – Physical)*100. Error bars show standard error from the mean. Asterisk indicates a significant effect of age of ASL onset on selectivity among children, such that early signers have less selective responses. **b)** Mean laterality index for NS and ES children, and NS and NNS adults. Laterality index was calculated as ($NumVox_L$ - $NumVox_R$)/($NumVox_L$ + $NumVox_R$), where NumVox is number of suprathreshold voxels to the Mental > Physical contrast within the temporal lobe ($p<.001$; results unchanged at $p<.01$). Error bars show standard error from the mean. **c)** Proportion correct on non-verbal ToM task (y-axis) by age (x-axis). **c)** Box plots (top) show z-scored inter-region correlations within the ToM network, within the language network, and across the ToM-language networks, for native signing (blue) and early signing (orange) children, and native signing (blue) and non-native signing adults (orange). Correlation matrices (bottom) show average z-scored correlation values across all ToM and language brain regions of interest, for native signing (left) and non-native signing (right) children (top) and adults (bottom). Regions are in the same order along the x- and y-axes: R/LTPJ, PC, D/M/VMPFC, RSTS, RSTS/RMidAntTemp (overlap), R/LMidAntTemp, LAntTemp, R/LMidPostTemp, LPostTemp, LAngGyrus, LSFG, LMFG, LIFGOrb, LIFG).

$t(35)=8.7$, $p=2.7x10^{-10}$; children: $t(23)=4.4$, $p=.0002$). Responses in ToM brain regions were similarly significantly more correlated with other regions within their own network, relative to regions in the language network in adults (within vs. across-network correlations: adults: $t(35)=7.1$, $p=3.2x10^{-8}$), but this effect was not significant among children ($t(23)=1.7$, $p=.11$). Among children, there was no significant correlation between inter-region correlations and age (all $ps>.3$), and inter-region correlations were not correlated with performance on either ToM task (ToM_V: NS effect of wi-ToM: $b=.09$, $t=.31$, $p=.76$, NS effect of wi-Lang: $b=-.21$, $t=-.51$, $p=.62$, NS effect of ac-ToM-Lang: $b=.27$, $t=.69$, $p=.50$, NS effect of motion: $b=.09$, $t=.34$, $p=.74$; ToM_{NV}: NS effect of wi-ToM: $b=.09$, $t=.32$, $p=.75$, NS effect of wi-Lang: $b=.05$, $t=.13$,

p=.90, NS effect of ac-ToM-Lang: b=-.20, t=-.57, p=.58, NS effect of motion: b=-.07, t=-.28, p=.79; see Figure 2c).

The key goal was to determine if delayed access to language impacts the functional correlations within or between ToM and language networks. We found no evidence for an effect of age of ASL onset on functional correlations within ToM brain regions (children: NS effect of age of ASL onset: b=-.06, t=-.30, p=.77, NS effect of age: b=.02, t=.72, p=.48, NS effect of motion: b=-.03, t=-1.5, p=.14; adults: NS effect of age of ASL onset: b=.40, t=1.6, p=.12, effect of motion: b=-.67, t=-2.7, p=.01), within language brain regions (children: NS effect of age of ASL onset: b=-.01, t=-.04, p=.97, NS effect of age: b=.05, t=.23, p=.82, NS effect of motion: b=-.32, t=-1.5, p=.15; adults: NS effect of age of ASL onset: b=-.17, t=-.65, p=.52, effect of motion: b=-.26, t=-1.0, p=.32), or across the two networks (children: NS effect of age of ASL onset: b=.09, t=.42, p=.68, NS effect of age: b=.17, t=.79, p=.44, NS effect of motion: b=-.23, t=-1.1, p=.29; adults: NS effect of age of ASL onset: b=-.46, t=-1.9, p=.06, NS effect of motion: b=-.09, t=-.37, p=.71). In a subsequent regression excluding the four late signing adults, the marginal effect of age of ASL onset was non-significant (NS effect of age of ASL onset: b=-.33, t=-1.5, p=.14, NS effect of motion: b=-.03, t=-.12, p=.90; see Figure 2c).

FMRI Results: Non-Verbal Movie Viewing

We additionally measured neural responses during a non-verbal naturalistic movie-viewing task: Disney Pixar's "Partly Cloudy"[36]. This movie has been shown to drive responses in ToM brain regions as well as the "extended Pain Matrix,"[37] and has previously been used to measure developmental change in these regions in children[27]. Based directly on this previous work, we measured four properties of the neural response in ToM brain regions during this task. Three properties arguably reflect functional maturation of the ToM network: (1) inter-region correlations (within the ToM network, and across ToM and Pain networks), (2) the extent to which the timecourse of response in ToM brain regions is similar to (positively correlated with) an average adult response timecourse, and (3) the response magnitude to three particular events embedded within the movie. Previous work using a large, cross-sectional sample of children found a significant positive correlation between response magnitude during these events and age (T01, T02) and performance on a verbal ToM behavioral battery (T04)[27].

Interregional correlations within the ToM network were significantly higher than across ToM-Pain network correlations in children ($t(53.6)=12.3$, $p<2.2x10^{-16}$) and in adults ($t(55.3)=14.8$, $p<2.2x10^{-16}$), and, critically, similarly high despite delayed access to language in children (NS effect of age of ASL onset: b=.17, t=1.1, p=.26, positive effect of age: b=.60, t=4.1, p=.0004, NS effect of motion: b=-.25, t=-1.8, p=.08), and adults (NS effect of ASL onset: b=.23, t=.80, p=.43, NS effect of motion: b=-.31, t=-1.1, p=.29). In the full sample, within-ToM correlations were *higher* in individuals with a longer delay before ASL exposure (marginal positive effect of age of ASL onset: b=.34, t=2.0, p=.047, negative effect of age group (child): b=-.59, t=-2.3, p=.02, negative effect of motion: b=-.39, t=2.4, p=.02); this effect was marginal upon excluding late signing adults (NS effect of age of ASL onset: b=.28, t=2.0, p=.052, negative effect of age group(child): b=-.63, t=-2.5, p=.02, negative effect of motion: b=-.30, t=-2.1, p=.04). Across-ToM-Pain network correlations did not differ based on ASL onset in children (NS effect of age of ASL onset: b=.24, t=1.2, p=.23, NS effect of age: b=-.31, t=-1.5, p=.14, NS effect of motion: b=.22, t=1.1, p=.26), adults (NS effect of ASL onset: b=-.41, t=-1.6, p=.12, positive effect of

motion: b=.68, t=2.7, p=.01) or in the full sample (NS effect of age of ASL onset: b=-.21, t=-1.2, p=.24, NS effect of age group: b=.20, t=.78, p=.44, effect of motion: b=.46, t=2.7, p=.009; see Figure 3a).

Figure 3

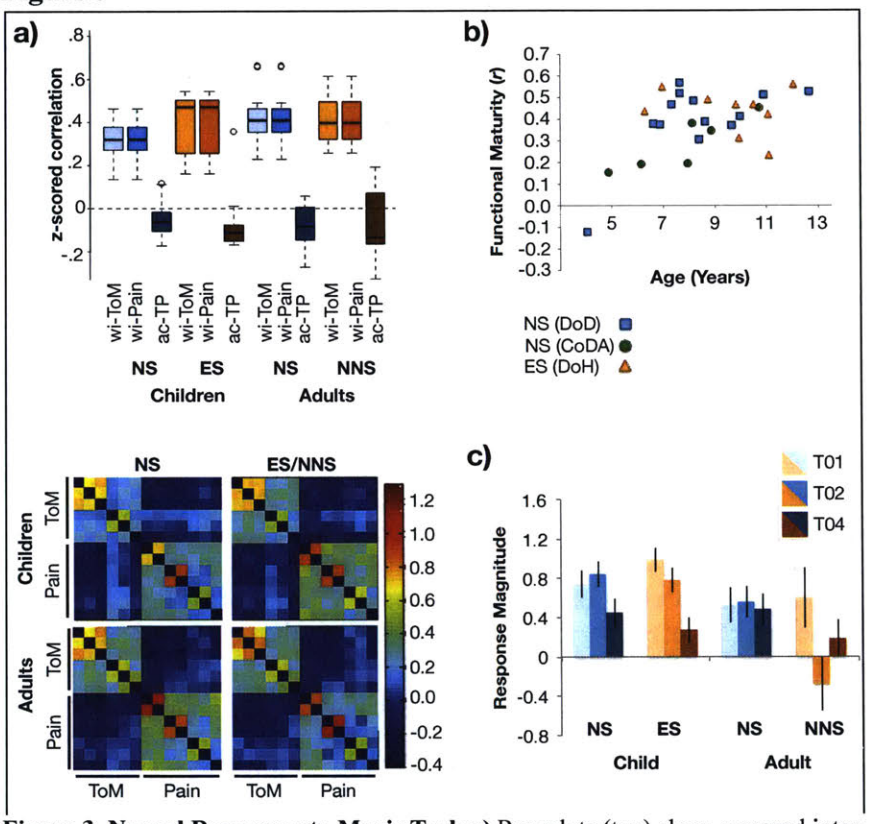


Figure 3. Neural Responses to Movie Task a) Box plots (top) show z-scored inter-region correlations within the ToM network, within the Pain matrix, and across the ToM-Pain networks, for native signing (blue) and early signing (orange) children, and native signing (blue) and non-native signing adults (orange). Correlation matrices (bottom) show average z-scored correlation values across all ToM and Pain brain regions of interest, for native signing (left) and non-native signing (right) children (top) and adults (bottom). Regions are in the same order along the x- and y-axes: R/LTPJ, PC, D/M/VMPFC, R/LS2, R/LInsula, R/LMFG, dAMCC. **b)** Scatterplot shows the functional maturity (i.e. similarity to average adult timecourse, Pearson's *r*) (y-axis) by age (x-axis) among child participants. Native signing participants are shown as blue squares (deaf children of d/Deaf adults) and green circles (hearing children of d/Deaf adults), early signing participants are shown as orange triangles. **c)** Bars show mean response magnitude in the ToM network in native signing (NS, blue) and early or non-native singing (ES, NNS, orange) children and adults, to three ToM events (T01, T02, T04), which previously showed developmental change with age (T01, T02) and ToM score (T04).

Response timecourses to the movie were generally highly correlated with an average adult response timecourse (M(SE) *r*-value: children: .39(.03), adults: .29 (.04)). Among children, there was a significant ASL-onset-by-age interaction such that the effect of age on "functional maturity" was smaller in children who experienced a longer delay before exposure to ASL, but no significant effect of age of ASL onset (NS effect of age of ASL onset: b=.42, t=1.5, p=.14, NS effect of age: b=.25, t=1.3, p=.21, negative effect of motion: b=-.31, t=-2.1, p=.048, age of ASL onset-by-age interaction: b=-.73, t=-2.1, p=.04; see Figure 3b and Supplementary Figure 3). The interaction may be driven by a moderate correlation between age and age of ASL onset among early signing children who completed this task (*r*=.31). In a follow-up regression with the interaction term removed, the effect of ASL onset remained non-significant, and age became a significant predictor of functional maturity (NS effect of age of ASL onset: b=-.08, t=-.48, p=.64, positive effect of age: b=.49, t=3.0,

p=.007, negative effect of motion: b=-.39, t=-2.5, p=.02). There was no effect of age of ASL onset in the full sample (NS effect of age of ASL onset: b=.003, t=.02, p=.99, NS effect of age group: b=.40, t=1.5, p=.14, NS effect of motion: b=-.25, t=-1.5, p=.15; NS ASL-onset-by-age interaction), or in adults alone (NS effect of age of ASL onset: b=-.16, t=-.56, p=.58; NS effect of motion: b=-.09, t=-.31, p=.76).

Across all children, functional maturity was significantly positively correlated with performance on the non-verbal ToM task (effect of non-verbal ToM: b=.34, t=2.1, p=.045; effect of motion: b=-.49, t=-3.02, p=.006), but this relationship did not remain significant when additionally controlling for age (effect of non-verbal ToM task: b=.04, t=.16, p=.87, marginal effect of age: b=.44, t=2.0, p=.06, effect of motion: b=-.40, t=-2.5, p=.02). Performance on the verbal ToM task was not correlated with functional maturity (effect of verbal ToM: b=.10, t=.52, p=.61; effect of motion: b=-.61, t=-3.2, p=.005).

Finally, there was no difference in the response magnitude during the ToM events of interest (events T01, T02, and T04) as a function of age of ASL onset in children (NS effect of age of ASL onset: b=-.05, t=-.53, p=.60, positive effect of age: b=.11, t=2.24, p=.03, NS effect of event (T06): b=-.002, t=-.009, p=.99, negative effect of event (T07): b=-.79, t=-3.4, p=.001, negative effect of motion: b=-.30, t=-3.1, p=.005). There was no effect of age of ASL onset in the full sample (NS effect of age of ASL onset: b=-.03, t=-.23, p=.82, NS effect of age group: b=.36, t=1.9, p=.06, NS effect of event (T06): b=-.22, t=-1.4, p=.16, negative effect of event (T07): b=-.43, t=-2.7, p=.008, NS effect of motion: b=-.17, t=-1.4, p=.17), or among adults (NS effect of age of ASL onset: b=-.15, t=-.70, p=.49, NS effect of event (T06): b=-.38, t=-1.8, p=.08, NS effect of event (T07): b=-.22, t=-1.0, p=.31, NS effect of motion: b=-.04, t=-.21, p=.83; see Figure 3c and Supplementary Figure 3). Among children, response magnitude to these events was not correlated with either ToM task (ToM_V: p=.33; ToM_{NV}: p=.06).

Exploratory Analyses

The results of the planned fMRI analyses found evidence for reduced selectivity in children with delayed exposure to language during the story task, but no neural differences between native and early signers during the movie task. One difference between the analyses is that the story task analyses focused on responses in specific regions of interest (RTPJ and DMPFC, defined individually), whereas the movie analyses used the average response across multiple ToM brain regions (bilateral TPJ, precuneus, D/M/VMPFC, group ROIs). In exploratory analyses, we examined the effect size of age of ASL onset on selectivity during the story task in all individually defined ToM ROIs. Among children, the effect of age of ASL onset on selectivity was significant in RTPJ alone (effect of age of ASL onset: p=.03; all other ROIs, ps >.2). Age of ASL onset did not have a significant effect on selectivity in any ROI among adults (ps>.2; see Table 1).

We subsequently analyzed responses to the movie-viewing task in these individual RTPJ ROIs. While inter-region correlations require the analysis of multiple brain regions, functional maturity and response magnitude to particular events can be measured in individual regions. There was no effect of age of ASL onset on functional maturity in RTPJ among children (NS effect of age of ASL onset: b=.08, t=.34, p=.74, NS effect of age: b=.10, t=.35, p=.73, NS effect of motion: b=-.29, t=-1.2, p=.25), in the full sample (NS effect of age of ASL onset: b=-.17, t=-.93, p=.36, NS

effect of age group: b=-.05, t=-.16, p=.88, NS effect of motion: b=-.09, t=-.50, p=.62), or among adults (NS effect of age of ASL onset: b=-.34, t=-1.2, p=.23, NS effect of motion: b=.04, t=.14, p=.89). Similarly, there was no effect of age of ASL onset on the magnitude of response to events T01, T02, and T04 among children (NS effect of age of ASL onset: b=-.18, t=-1.6, p=.13, NS effect of age: b=.11, t=1.6, p=.12, NS effect of event (T06): b=.01, t=.03, p=.98, negative effect of event (T07): b=-1.04, t=-4.1, p=.0002, NS effect of motion: b=.11, t=.99, p=.34), in the full sample (NS effect of age of ASL onset: b=-.13, t=-1.07, p=.29, NS effect of age group: b=-.04, t=-.26, p=.79, NS effect of event (T06): b=-.10, t=-.58, p=.56, negative effect of event (T07): b=-.83, t=-4.7, p=0, NS effect of motion: b=.07, t=.62, p=.54), or among adults (NS effect of age of ASL onset: b=-.08, t=-.42, p=.68, NS effect of event (T06): b=-.18, t=-.74, p=.46, negative effect of event (T07): b=-.65, t=-2.7, p=.009, NS effect of motion: b=.01, t=.06, p=.95). The same pattern of results was obtained in identical analyses of group RTPJ ROIs.

Table 1

	ROI	Predictor	Children	Adults
Planned	RTPJ	ASL-onset	**b=-.47, t=-2.3, p=.03**	b=-.30, t=-1.1, p=.28
		Age	b=.28, t=1.4, p=.18	
		Motion	b=.11, t=.55, p=.59	b=.28, t=1.0, p=.31
	DMPFC	ASL-onset	b=-.09, t=-.39, p=.71	b=-.17, t=-.60, p=.55
		Age	b=.34, t=1.6, p=.13	
		Motion	b=.43, t=1.6, p=.14	b=.15, t=.54, p=.59
Exploratory	LTPJ	ASL-onset	b=-.17, t=-.81, p=.43	b=-.13 t=-.46, p=.65
		Age	b=.39, t=1.9, p=.07	
		Motion	b=-.03, t=-.13, p=.90	b=.22, t=.82, p=.42
	MMPFC	ASL-onset	b=-.27, t=-1.1, p=.27	b=.05, t=.17, p=.87
		Age	b=.25, t=1.1, p=.31	
		Motion	b=.13, t=.55, p=.59	b=.01, t=.02, p=.99
	VMPFC	ASL-onset	b=-.30, t=-1.1, p=.3	b=.04, t=.13, p=.90
		Age	b=.17, t=.62, p=.55	
		Motion	b=.23, t=.84, p=.42	b=-.04, t=-.15, p=.88
	PC	ASL-onset	b=-.03, t=-.17, p=86	b=.20, t=.71, p=.48
		Age	b=-.06, t=-32, p=.76	
		Motion	**b=.65, t=3.3, p=.005**	b=.02, t=.09, p=.93

Table 1. Selectivity of Responses to ASL Story Task by ROI. Full statistics for results of linear regressions testing for significant effects of age of ASL onset (ASL-onset), age, and motion on response selectivity, by ROI. Planned tests in the main analyses focused on responses in RTPJ and DMPFC; analyses of responses in other ToM ROIs (LTPJ, MMPFC, VMPFC, PC) were exploratory. Significant effects are in bold: selectivity of the RTPJ is reduced based on age of ASL onset, among children.

Finally, we tested for differences in response selectivity in RTPJ during the movie, based on age of ASL onset. Selectivity was calculated as the average response to the seven ToM event peaks, minus the average response to the twelve Pain event peaks. We found no differences in the selectivity of RTPJ during the movie task, among children (NS effect of age of ASL onset: b=.06, t=.28, p=.79, NS effect of age: b=.36, t=1.3, p=.22, NS effect of motion: b=-.12, t=-.50, p=.62), adults (NS effect of age of ASL onset: b=-.24, t=-.85, p=.40, NS effect of motion: b=.03, t=.10, p=.92), or in the full sample (NS effect of age of ASL onset: b==.13, t=-.67, p=.51, NS effect of age group: b=-.22, t=-.75, p=.46, NS effect of motion: b=-.05, t=-.28, p=.78).

Discussion

Social cognitive and language abilities undergo dramatic development in childhood. A key debate concerning Theory of Mind development is the extent to which language plays a causal role in ToM development, versus a superficial role in the expression of ToM competence during verbal tasks. In order to provide insight into this debate, we measured multiple aspects of ToM reasoning, behaviorally and neurally, in children and adults who experienced delayed access to language. Among children, we find evidence for neural and behavioral delays based on age of ASL onset in verbal, but not non-verbal contexts. There were no neural differences based on age of ASL onset among adults, in either task. Though differences based on age of ASL onset were

most apparent in verbal contexts, our results are most consistent with the hypothesis that language facilitates theory of mind development, in addition to expression of ToM competence.

We measured ToM reasoning among child participants via verbal and non-verbal behavioral tasks. Children who experienced delayed access to language performed similarly to their native signing peers on verbal and non-verbal versions of the false-belief task, but showed behavioral ToM delays on more advanced (verbal) ToM questions (e.g., reasoning about the moral blameworthiness of individuals who cause harm accidentally, lies and second-order false-beliefs, and non-literal speech (sarcasm)). Interestingly, performance differences on questions concerning moral blameworthiness were only apparent in the verbal task; there were no differences based on age of ASL onset on analogous non-verbal questions concerning moral blameworthiness.

We additionally used fMRI to measure multiple properties of the neural response while participants either watched stories (in ASL) or watched a non-linguistic social movie. During the story task, we measured response selectivity, response lateralization, and inter-region correlations between ToM and language brain regions. We observed reduced selectivity in ToM brain regions based on the age of ASL onset: children who experienced a longer delay prior to ASL exposure had less selective responses in the right temporoparietal junction (RTPJ). Reduced selectivity was not a product of altered response lateralization: there were no differences in response lateralization based on age of ASL onset. Additionally, reduced selectivity was not related to altered inter-region correlations within- and across- ToM and language networks: within-network correlations of ToM and language regions and across-network correlations between ToM and language regions were high regardless of age of ASL onset. In contrast to the fMRI results from the story task, there were no differences based on age of ASL onset in ToM neural responses to the non-verbal movie task. We measured inter-region correlations between ToM and pain brain regions, and the extent to which functional responses in the ToM network were similar to an average adult timecourse ("functional maturity"). Like response selectivity, these two measures increase with age during childhood[27]. We additionally measured the magnitude of response to particular events in the movie (events T01, T02, and T04). Responses to these events were previously found to be positively correlated with age (T01, T02) and theory of mind behavior (T04)[27]. None of these measures appeared to be affected by age of ASL onset, even when focusing specifically on responses in RTPJ. Responses to the movie in early signers were indistinguishable from those of native signers, among children and adults.

Given these results, does language facilitate conceptual change in ToM, or does language simply enable the expression of ToM competence? The behavioral and neural data appear to converge and suggest that delayed access to language leads to delays in ToM reasoning in verbal but not non-verbal contexts. At first glance, this pattern of results could suggest that the primary role of language is to facilitate expression of ToM competence in verbal contexts. However, when considering the aspects of ToM measured in each task, these results are most consistent with the hypothesis that language plays a facilitative role in theory of mind development, in addition to expression.

Multiple behavioral studies have reported delays in ToM reasoning – and specifically, reasoning about false-beliefs – in children who experience delayed access to language[8,20,21]. In contrast to

these studies, we saw similar performance on false-belief items in verbal and non-verbal tasks, regardless of age of ASL exposure. Interestingly, across all participants, performance on false-belief items was higher in the verbal task relative to the analogous false-belief items in the non-verbal task. It is unlikely that children's actual false belief understanding varies across these analogous tasks; rather, this suggests that the format of the task matters for measuring ToM competence. Verbalizing ToM concepts could boost performance by facilitating the representations of those ToM concepts during the task, by providing representational structure[9,10,38], and/or by providing specific mental state verbs[39]. Or, verbalizing ToM concepts could boost performance by mitigating executive function and working memory demands[40]. Any of these mechanisms could be impacted by delayed access to or low proficiency in language[41], which could contribute to the delays in false belief task performance in prior studies and to the success of the children in our sample. In the current study, all non-native signing child participants were proficient in ASL and received access to sign language relatively early in development ("early signers," having received access to sign language between ages .25 – 7 years). Prior studies suggest that the age of exposure to a sign language predicts the extent of delay on many aspects of language development[42]. Early exposure and proficiency in sign language could reduce delays on false-belief tasks by providing earlier access to the facilitative benefits of language for ToM.

Though we did not find evidence for a delay in false-belief reasoning, we did observe behavioral ToM delays on more advanced ToM questions, involving reasoning about the moral blameworthiness of individuals who cause harm accidentally, lies and second-order false-beliefs, and non-literal speech (sarcasm). One possibility is that reduced performance on these items reflects ongoing delays in ToM development, even after successfully "catching up" on false-belief reasoning. In favor of this interpretation, early signers did not perform worse on control items in the verbal task, which were designed to ensure that children could follow the narrative and provide simple (linguistic) responses. A second possibility is that the observed effects of delayed access to language on these items are task-related, i.e., related to expression of ToM understanding. Most advanced ToM categories (lies/second-order false beliefs, mistaken referents, and non-literal speech) did not have analogous items in the non-verbal ToM task, due to the difficulty in communicating these concepts clearly in a non-verbal context. These items involved complicated ToM concepts embedded within linguistically complicated stories, making it difficult to tease apart whether performance deficits in these categories reflect delays in ToM per se, or delays in language. However, moral judgment questions, which involved assigning moral blame by reasoning about the intentions of the character, were included in both non-verbal and verbal tasks. Early signers performed worse than native signers on the moral judgment questions in the verbal task only, consistent with the interpretation that observed differences based on age of ASL onset were task-related. Still, it remains possible that early signers can catch up to their native signing peers on earlier developing aspects of ToM development, and simultaneously shown ongoing delays on other, later developing aspects.

The results from the behavioral tasks certainly suggest a role for language in the expression of ToM competence, but the extent to which language plays a role in ToM development is unclear based on the behavioral data alone. The results from the neuroimaging data are particularly suggestive of a role for language in facilitating ToM development per se. A key result of the current study is that the response selectivity of RTPJ was reduced in children with delayed

access to language in the verbal story task. The response selectivity measure indexes the preferential neural response to mental state information relative to general social information (e.g., physical appearance or social relationships). Previous studies have found that ToM brain regions become increasingly functionally selective during childhood, via decreases in responses to non-preferred stimuli[25-27]. The RTPJ in particular has a highly selective response profile in adults[22], and functional selectivity of the RTPJ is significantly correlated with behavioral ToM reasoning in children[26]. Increasingly functionally selective responses reflect more refined boundaries between preferred and non-preferred stimuli, as a brain region becomes specialized for the particular computational demands of ToM processes. Thus, early and extensive exposure to language may facilitate the discrimination of concepts relevant for ToM processing, and for the development of functionally selective responses in the RTPJ.

In contrast to the story task, the selectivity of the RTPJ was not reduced based on age of ASL onset during the non-verbal movie task. What aspect of ToM drives the difference between these two tasks? An obvious possibility is that these results reflect the stimulus modality of the experiments. That is, the stories were presented in ASL whereas the movie was non-linguistic. In our sample, all children were equally proficient on an existing test of ASL (the ASL-RST). Nevertheless, sentences about mental states have unusually complex syntax, whose comprehension might not be fully captured by basic language comprehension scores. Delayed access to language could specifically affect transforming complex syntactic contractions into conceptual representations. Perhaps extensive developmental experience with these sentences facilitates the extraction of complex meaning from linguistic inputs[39]. Children with delayed linguistic experience would then be inefficient at this transformation, reflected by less selective neural activity. One puzzle for this view, though, is why the change in selectivity occurs by *decreasing* the response to social stories, not by *increasing* the response to mental stories.

A second possibility is that the relevant difference between the stories and the movie is not the stimulus modality but the content, specifically of the Social control condition. Both the stories and movie tasks evoked mental states, including mistaken beliefs and changing emotions; the Social stories described the physical appearance and enduring relationships of characters, whereas the control (Pain) events in the movie depicted physically painful experiences and bodily transformations. It is possible that selectivity in the stories task relies more on the distinction between mental states vs. social traits, and that differentiation of these two categories is more impacted by developmental experience than the differentiation of mental vs. bodily states. Individuals with delayed access to ASL may still be refining the distinction between mental and social content, despite having developed other distinctions between conceptual categories.

Critically, the reduction in selectivity during the story task was not observed among adults: neural responses to mental state stimuli in d/Deaf adults exposed to sign language after a delay were indistinguishable from those observed in native signers, regardless of task. This suggests that neural differences early in development between native signing and early signing children are ultimately resolved in adults who have prolonged exposure to and proficiency in a sign language. One open question for future research is whether there is a critical or sensitive period during which linguistic input must be received in order to develop the typical neural profile in ToM regions[43].

In sum, the current study suggests that language plays a facilitative role for forming, manipulating, and/or discriminating ToM concepts in childhood. Children with delayed access to language show typical profiles of ToM development in non-verbal contexts, neurally and behaviorally. However, all children receive a performance boost on behavioral measures of analogous verbal ToM questions, suggesting that language is facilitative for expression of ToM competence. Similarly, cortical regions for thinking about others' thoughts become functionally specialized in childhood despite delayed access to language (as shown with the non-verbal movie task), but the extent of specialization is reduced in proportion to the extent of the language delay. Reduced selectivity in ToM brain regions reflects delays in the refinement of conceptual categories that distinguish between preferred and non-preferred stimuli, and corresponding deployment of ToM-specific computational processes in ToM brain regions. Importantly, differences in functional specialization based on age of ASL onset are resolved in adults who received early and prolonged exposure to a sign language.

Taken together, these results point to an important facilitative role for language on the development and expression of ToM reasoning in childhood. Early exposure and proficiency in ASL provides earlier access to these facilitative benefits. Language deprivation has behavioral consequences for social development: late signing and oral deaf children consistently perform worse on standard tasks of ToM[8,44,45]. The current results suggest that early access to and proficiency in sign language facilitates social development in childhood, and protects against extensive or permanent delays in social development associated with language deprivation.

Methods

Participants

Child participants were 21 native signers (NS; 4-12.7 years old, M(SD) = 8.19 (2.2) years, 10 female), who received exposure to ASL from birth (15 deaf children and 6 hearing children of d/Deaf adults), and 12 "early signers" (ES; 6.2-12.1 years old, M(SD) = 9.29 (1.9) years, 5 female), who were born to hearing parents and received exposure to ASL after an initial delay of .25 – 7 years (M(SD) = 2.9 (2.2) years).

Adult participants included 20 NS (20-54 years old, M(SD) = 30.2(9.5), 13 female). In contrast to the native signing children, who were all born to d/Deaf parents, three native signing adults were born to hearing parents (n=10 NS d/Deaf born to d/Deaf parents, n=7 NS children of d/Deaf parents). All non-native signing adults were born to hearing parents (n=16, 21-64 years old, M(SD) = 38.1(12.7) years, 4 female). Whereas all non-native signing children were early signers, non-native signing adult participants included 12 ES and 4 "late signers" (LS). LS adults received exposure to ASL at ages 11, 15, 18, and 20 years; early signers received exposure to ASL by age seven years (12 ES adults: 1.5-7 year delay, M(SD) = 3.3 (1.9) years; combined ES and LS delay M(SD) = 6.5 (6.2) years).

Participants were recruited via the researchers' social networks and by snowball sampling. Child participants were also recruited with help from several schools for the deaf. All participants were screened by a native ASL signer to determine current ASL fluency, and only fluent signers were recruited to participate in the study. Child participants signed an assent form; adult participants and parents of child participants signed a consent form. All assent and consent forms and

experimental protocols were approved by the Committee on the Use of Humans as Experimental Subjects (COUHES) at MIT. See Supplementary Table 1 for additional information about participants.

Behavioral Battery

The custom-made verbal ToM task was an ASL-adapted version of a battery previously used to measure ToM in hearing children (https://osf.io/g5zpv/)[27]. The task involved watching an experimenter tell a story, and answering 26 prediction and 24 explanation questions about the mental states of the characters, in the context of helping them find their snacks. 12 additional control questions were asked and used to ensure task comprehension; these items were not included in the summary score. The script and materials for this task are publicly available via the Open Science Framework (OSF; https://osf.io/kyu3f/). The summary score was calculated as the proportion of questions answered correctly (ToM_V); for follow-up analyses we additionally calculated summary scores for the control, false-belief, and moral judgment items.

The custom-made non-verbal ToM task involved watching an experimenter place a series of three to five pictures on a board, which presented characters undergoing a sequence of events. In the first part of the task, the experimenter would then place two pictures side by side and use the prompt "What comes next?" Children responded by pointing to the picture that best completed the series (19 items). The second part of the task focused on moral reasoning (11 items). Before these items, the experimenter said "You decide- is this good (pointing to thumb up), bad (pointing to thumb down), or okay (point to neutral thumb)?" They then showed a series of pictures for each item, ending with a single picture of a character who inflicted harm either accidentally or intentionally. Children responded by pointing to the thumbs up, thumbs down, or neutral thumb picture, suggesting that the character was "good," "bad," or "okay." Children completed 6 practice trials before the initial sequence-completion questions, and an additional 3 practice trials immediately before the moral reasoning questions. Practice trials ensure that children understood the task instructions, but were otherwise not analyzed. The summary score was calculated as the proportion of questions answered correctly (ToM_{NV}); for follow-up analyses we additionally calculated summary scores for the false-belief and moral judgment items. The protocol and materials for this task are also available via OSF (https://osf.io/kyu3f/).

Children additionally completed the American Sign Language Receptive Skills Test (ASL-RST)[46]. After completing an initial vocabulary check (n=20 trials), children watched an adult signing in a movie, and responded by pointing to the picture (out of a 4-picture array) that corresponded to the sign. Children completed three practice trials after the vocabulary check and prior to the receptive skills test. Two items were ultimately excluded from analysis (item 37: BOX DOG-IN-FRONT and item 42: INTERSECTION HOUSE-TOP-RIGHT), because more than 75% of participants answered these items incorrectly.

Finally, children completed a standardized task of nonverbal IQ (KBIT-II[47]), and, when possible, a measure of spatial working memory (computerized CORSI task[48,49]; n=24).

FMRI: ASL Story Task

During the fMRI scan, participants watched movies of a woman telling stories in ASL, which involved characters and their mental states (Mental condition), characters and their physical

appearance or social relationships (Social condition), or descriptions of physical objects and events in the world (Physical condition). A subset of stimuli (24/42) were English stories previously used to measure neural responses in hearing children and adults[25,26], translated into ASL. The child paradigm included 14 of these 24 stimuli (4 Mental, 5 Social, 5 Physical), and ten novel stimuli (see https://osf.io/kyu3f/ for all stimuli). All 42 stories were normed by 10 naïve, Deaf native signers. Based on behavioral ratings, stories were matched across conditions for linguistic features (e.g., syntactic complexity, number of signs, number of verbs), psychological features (e.g., how easy to understand, how interesting), and imageability. Stories were told using simple language, in an enthusiastic, narrative way.

To encourage engagement during the story task, stories were presented in two consecutive segments: the main story (29-41s) and a final sentence containing the story ending or the ending of an unrelated story (4-8s). The stimuli were followed by a 3s pause during which participants responded to the prompt by pushing one of two buttons ("Yes" or "No"), and a rest period (8-24s; such that each block lasted 60s). Half of the presented stories were followed by the correct ending ("Yes" response). Incorrect responses were drawn randomly from another story. The story ending was not included in subsequent analyses. During the prompt, story ending, and response portion of the experiment, an image of a check (left) and an "X" (right) was displayed to encourage participants to answer the question, and remind them which buttons corresponded to "yes" and "no" answers. Child participants were introduced to the task and completed four practice trials prior to the scan.

Stimuli were presented in Matlab 2010a running on an Apple MacBook Pro. Child participants heard 24 stories (8 per condition) across four 8.3-minute runs. Adult participants heard 30 stories (10 per condition) across five 10.3-minute runs. All children saw the same 8 stories per condition; each adult participant saw 10 of 14 stories per condition. Stories were counterbalanced across runs and participants. Participants also saw 8 (child) or 20 (adult) clips of non-signs; the non-sign stimuli were excluded from the present analyses. Each run included six 60-second blocks (2 per condition), as well as 10 seconds of rest at the beginning and end of each run. The order of conditions in each run was palindromic (e.g., A B C C B A) and counterbalanced across runs.

An experimenter in the control room monitored participants during the scan. For child participants, a second experimenter stood in the MRI room near the participant's feet. If the participant moved noticeably during the scan, this experimenter would place her hand on the child's leg, as a reminder to stay still. In between functional runs of the scan, the experimenter in the control room communicated with participants by signing via live video. Participants used the button box to response to questions like "Are you okay?" and "Are you ready to continue?" Participants were also given a squeeze ball that would alert the experimenters in the control room if they wanted to stop the scan session.

Behavioral performance on the task was measured via accuracy (proportion of questions answered correctly) on trials from included functional runs only; trials from runs that were excluded due to excessive motion were not analyzed. Participants generally performed well above chance on this task (M(SD) per condition: Children: Mental: .80(.23), Social: .79(.19), Physical: .68(.23); Adults: Mental: .87(.15), Social: .87(.17), Physical: .83(.17)). Among

children, there was no effect of age of ASL onset on accuracy, but older children were more accurate, and on average, children performed worse on Physical condition trials (effect of age: b=.38, t=2.5, p=.02, NS effect of age of ASL onset: b=-.21, t=-1.4, p=.18, negative effect of Physical condition (compared to Mental): b=-.54, t=-2.7, p=.009, NS effect of social condition: b=-.03, t=-.13, p=.89; no significant age*ASL-onset or ASL-onset*condition interactions). In a second regression that additionally included receptive ASL score, the results were unchanged, and there was no effect of ASL proficiency (effect of age: b=.37, t=2.3, p=.04, NS effect of age of ASL proficiency: b=.15, t=.91, p=.37, NS effect of ASL onset: b=-.19, t=-1.3, p=.23, negative effect of Physical condition (compared to Mental): b=-.53, t=-2.6, p=.01, NS effect of social condition: b=.07, t=.36, p=.72). Among adults, there was an effect of age of ASL onset such that adults who experienced a longer delay before exposure to language performed worse overall on the task, and an effect of condition such that adults performed marginally worse on Physical condition trials (negative effect of age of ASL onset: b=-.53, t=-4.5, p=.0001, marginal negative effect of Physical condition (compared to Mental): b=-.24, t=-1.8, p=.08, NS effect of social condition: b=.002, t=.01, p=.99; no significant ASL-onset*condition interactions). This pattern of results remained the same when excluding the four adults who received relatively late exposure to ASL (after age 10; negative effect of age of ASL onset: b=-.34, t=-2.7, p=.01, marginal negative effect of Physical condition (compared to Mental): b=-.32, t=-1.8, p=.08, NS effect of social condition: b=-.04, t=-.22, p=.83; no significant ASL-onset*condition interactions).

Six children did not complete more than one run of the story task experiment and were therefore excluded from analyses (see Supplementary Table 1).

FMRI: Movie Viewing

After the story task, participants watched a silent version of "Partly Cloudy," a 5.6-minute non-verbal animated movie. A short description of the plot can be found online (https://www.pixar.com/partly-cloudy#partly-cloudy-1). The stimulus was preceded by 10s of rest, and participants were instructed to watch the movie and remain still. Previous work has suggested that this movie stimulus can be used to localize ToM brain regions[37], and to study developmental change in the response of ToM brain regions in children[27]. Seven adult and two child participants did not complete the movie viewing (see Supplementary Table 1).

FMRI Data Acquisition

Prior to the fMRI scan, children watched a movie of their choice in a mock scanner while practicing lying still on their back for 10-15 minutes. Hearing children (native signing children of d/Deaf adults) listened to a recording of scanner sounds during the mock scan. If participants moved during the mock scan, their movie paused for three seconds, reminding and training them to stay still.

Whole-brain structural and functional MRI data were acquired on a 3-Tesla Siemens Tim Trio scanner located at the Athinoula A. Martinos Imaging Center at MIT, using one of two custom 32-channel phased-array head coils made for children[50] or the standard Siemens 32-channel head coil; all adult participants were scanned using the standard 32-channel coil. T1-weighted structural images were collected in 176 interleaved sagittal slices with 1mm isotropic voxels (GRAPPA parallel imaging, acceleration factor of 3; adult coil: FOV: 256mm; pediatric coils:

FOV: 192mm). Functional data were collected with a gradient-echo EPI sequence sensitive to Blood Oxygen Level Dependent (BOLD) contrast in 3 mm isotropic voxels with a 20% slice gap (n=7 adults, n=28 children) or 3.13 mm isotropic voxels with no slice gap (n=29 adults, n=1 child) in 32 interleaved near-axial slices aligned with the anterior/posterior commissure, and covering the whole brain (EPI factor: 64; TR: 2s, TE: 30ms, flip angle: 90°); all functional data were subsequently upsampled in normalized space to 2mm isotropic voxels. Prospective acquisition correction was used to adjust the positions of the gradients based on the participant's head motion one TR back[51]. 310 (adults) or 250 (children) volumes were acquired in each run of the story task, and functional data was acquired across five (adults) or four (children) runs. 155 volumes were acquired during the single run of the movie viewing. Four dummy scans were collected and excluded to allow for steady-state magnetization in each run.

FMRI Data Analysis

All analysis decisions (including preprocessing, region of interest selection and definition, motion exclusion and treatment procedures, calculation of selectivity indices) and planned analyses for the stork task were published via OSF (https://osf.io/kyu3f/)[52,53]. Story and movie task analyses were constrained by methods used in prior studies, in order to facilitate comparisons across studies[27]. Exploratory analyses are specifically marked as such.

FMRI data were analyzed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm) and custom software written in Matlab. Functional images were registered to the first image of the first run; that image was registered to each child's anatomical scan, and each child's anatomical scan was normalized to a common brain space (Montreal Neurological Institute (MNI) template). All data were smoothed using a Gaussian filter (5mm kernel).

Motion artifact timepoints were identified using the ART toolbox (https://www.nitrc.org/projects/artifact_detect/)[54] as timepoints for which there was 1) more than 2mm of motion in any direction relative to the previous timepoint or 2) a fluctuation in global signal that exceeded a threshold of three standard deviations from the mean global signal. Runs were excluded from analyses if one-third or more of the timepoints collected were identified as motion artifact timepoints (Story task: n=83 (child) or 103 (adult) timepoints; Movie: n=56 timepoints). Participants were excluded from analyses of the story task if they had fewer than two runs of usable data (n=3 child and 0 adult participants). The movie task consisted of a single run; 3 child participants were excluded for excessive motion during the movie task. In previous work the total number of included timepoints has been highly correlated with mean translation (e.g., r>.5). In the present dataset, this measure was not significantly correlated with mean translation in either task, in children or adults (rs<.31, ps>.15). Because mean translation is a direct measure of the amount of motion between analyzed functional images, we used mean translation to test for differences in motion based on ASL onset or age, and included this measure in all linear regressions including neural measures (as pre-specified in the analysis plan). Mean translation during the story task was not significantly correlated with age or ToM behavior, among children (Age: $r_s(22)$=-.05, p=.82; ToM_V: $r_k(19)$=.05, p=.82; ToM_{NV}: $r_k(22)$=-.06, p=.79), or with age of ASL onset in children ($r_s(22)$=-.20, p=.34). Among adults, age of ASL onset was positively correlated with mean translation during the story task ($r_s(34)$=.78, p=2.3×10^{-8}). This correlation remained significant when excluding the four adults who received exposure to ASL relatively late (after age 10; $r_s(30)$=.59, p=.0004). Among children, mean translation

during movie viewing was not significantly correlated with age or nonverbal ToM score (Age: $r_s(26)$=-.15, p=.44; ToM_{NV}: $r_k(26)$=.08, p=.70), but was significantly negatively correlated with performance on the verbal ToM task (ToM_V: $r_k(21)$=-.49, p=.02). Mean translation during the movie task was not correlated with age of ASL onset in children ($r_s(26)$=.03, p=.89); but this relationship was significantly positive in adults ($r_s(27)$=.74, p=$5.5x10^{-6}$), even when excluding adults who received access to ASL relatively late ($r_s(24)$=.63, p=.0005). See Supplementary Table 1 for amount of motion per participant, and Supplementary Figure 4 for a visualization of motion in each task per signing and age group. Data were high-pass filtered with a cutoff of 500 (story task) or 100 seconds (inter-region correlation analyses for both tasks, see below), in order to remove low-frequency noise, after interpolating over artifact timepoints[55,56]. We additionally implemented SPM's image scaling.

Analyses of FMRI Story Task Data

We used a general-linear model to analyze BOLD activity of each participant as a function of condition. Data were modeled in SPM8 using a standard hemodynamic response function (HRF). Boxcar regressors for each condition and the response period were convolved with the standard HRF, and nuisance covariates were included for run effects, motion artifact timepoints, and signal of no interest (five PCA-based regressors generated with CompCor[57] from timecourses extracted from individual eroded white matter masks).

We conducted Region of Interest (ROI) analyses on two ROIs: the right temporoparietal junction (RTPJ) and dorsal medial prefrontal cortex (DMPFC). Previous work has suggested that development of these regions is related to behavioral theory of mind abilities in childhood[58]. Detailed results from exploratory analyses of other ToM ROIs (left temporoparietal junction, middle medial prefrontal cortex, and precuneus) are reported in supplementary materials. Individual ROIs were defined as contiguous (minimum k=10) suprathreshold (p<.001) voxels within a 9mm radius sphere of the peak voxel to the Mental > Physical contrast, within previously defined region search spaces. Region search spaces were defined based on a random effects analysis using a False Belief > False Photograph contrast in a separate group of 462 typically developing adults[59]. We extracted the mean beta value per condition from these two regions, and calculated selectivity as (Mental – Social) / (Mental – Physical)*100. This calculation has previously been used in previous studies[26] (https://osf.io/jh68b/). Because the Mental and Physical difference is used to identify ROIs, this measure focuses on the relative difference between Mental and Social conditions. In supplementary analyses, we additionally measure selectivity in group ROIs which were 10mm spheres drawn around the peak coordinates of the random effects analysis used to create search spaces[59], excluding voxels that overlapped with language group ROIs. We used these group ROIs for easy comparison of results to other studies (e.g., https://osf.io/jh68b/). Unlike individual ROIs, the voxels analyzed in group ROIs did not necessarily respond more to the Mental condition compared to the Physical condition (voxels in group ROIs are not selected based on their functional response profile). Thus, we extracted average beta values and calculated selectivity as (Mental – Social) * 100. See Supplementary Table 2 for additional information about individual and group regions of interest.

Based on previous analyses, we expected the selectivity measure to be between -50 and 200 in individual ROIs, and excluded participants whose selectivity values fell outside of this range

(https://osf.io/kyu3f/). This resulted in excluding a single native-signing adult participant from analyses of the DMPFC ROI (selectivity value= -64.2).

For inter-region correlation analyses, preprocessed, scaled timecourses were extracted from each voxel per group ROI. The five principle component analysis (PCA)-based noise regressors and motion artifact timepoint regressors (included as nuisance regressors in the story task) were regressed from these timecourses, and the residual timecourses were high-pass filtered with a cut-off of 100 seconds. Timecourses from all voxels within an ROI were averaged, creating one timecourse per group ROI, and artifact timepoints were subsequently NaNed. Each ROI timecourse was correlated with every other ROI timecourse, per subject, and these correlation values were Fisher z-transformed. Within-ToM, within-Lang, and within-Pain network correlations were calculated as the average correlation value between brain regions within each of these networks. Similarly, across-ToM-Lang and across-ToM-Pain correlations were calculated as the average correlation value between ToM and Language, or ToM and Pain, brain regions, respectively. This exact procedure was used in a previous study[27]. In order to test if different brain networks (ToM-Lang, ToM-Pain) are significantly correlated with one another, we use t-tests to compare within- versus across-network correlations.

We additionally measured the extent of response lateralization to the Mental > Physical contrast in a large ROI encompassing the bilateral temporal lobe. The ROI was created from publicly available right hemisphere search spaces (http://saxelab.mit.edu/ToMgroupMaps.php)[59]; the right hemisphere was flipped to create the left hemisphere ROI. The lateralization index (LI) was calculated as the number of suprathreshold voxels in the left hemisphere minus the number of suprathreshold voxels in the right hemisphere, divided by the sum of the number of suprathreshold voxels in the left and right hemispheres[60]. We used a threshold of $p<.001$, uncorrected, and confirmed that results were not threshold dependent by repeating analyses at $p<.01$. We planned to exclude participants if the denominator was smaller than 20 (indicating fewer than 20 suprathreshold voxels, bilaterally); zero participants fit this exclusion criterion. Using this measure, large positive LI values indicate strong left lateralization, whereas an LI of zero indicates no response lateralization.

Analyses of FMRI Movie Viewing Data

As in the story task, we conducted inter-region correlation analyses using the response timecourses during movie viewing. Based on previous work, we analyzed responses from ToM brain regions (the same group ROIs used for the story task), and the extended "Pain Matrix." A previous study found that responses in these two networks were driven by this movie, and that brain regions within these two networks become increasingly correlated within-network, and increasingly anti-correlated across-network, during childhood. Aside from the networks analyzed, inter-region analysis procedures were identical to those described above for the story task.

Second, we tested whether the functional maturity of each participant's timecourse responses (i.e. similarity to adults) varied as a function of the age at which they were first exposed to language. We calculated the Pearson correlation between each participant's ToM timecourse (averaged across ROIs) and the average adult ToM timecourse, derived from a previous study[27]. Participants saw a truncated version of the movie, compared to the stimulus used in the previous

paper, such that the task ended with the end of the movie, as opposed to including movie credits. Thus, we used TRs 11:155 (instead of TRs 11:168) for all inter-region correlation analyses, and adapted the average adult timecourse from the previous study accordingly.

Finally, we tested for differences in the magnitude of response in ToM brain regions to three specific events in the movie. Response magnitude to two of these events was previously found to increase with age in three to twelve year old children (events T01 and T02[27]). The third event depicts Peck, the stork character, putting on football gear in front of Gus, the cloud character. In the context of the movie, this action indicates to Gus that Peck had not previously abandoned him, as he had feared, and indeed never intended to. In a previous study with three to twelve year old children, the average magnitude of response in ToM brain regions to this event (referred to as event T04) was significantly positively correlated with performance on a verbal ToM behavioral battery, controlling for age and motion (and correcting for multiple comparisons)[27]. Thus, neural responses to this event could be considered a sophisticated measure of cognitive ToM.

Statistical Regressions

We used linear regressions to test if properties of the neural response in ToM brain regions differed as a function of age of first exposure to language. For the story task, we measured response selectivity of ToM brain regions (RTPJ and DMPFC), inter-region correlations (within- and between- ToM and language brain regions), and response lateralization. For the movie-viewing task, we measured inter-region correlations (within- and between- ToM and pain brain regions), functional maturity, and response magnitude in the ToM network to events T01, T02, and T04 of the movie. We conducted regressions on each of these measures within children, within adults, and, when possible (due to identical experimental procedures), across the full sample. We tested for effects of the age of ASL onset, a continuous variable ranging between .25-7 years in early signing children, and 1.5-20 years in early/late signing adults, on each of these neural properties. Age of ASL onset for children of d/Deaf adults was zero. We additionally included age group (child vs. adult) or age (within children only) as a covariate. As specified in the analysis plan, we first tested for significant Age (or Age-Group)*ASL-onset interactions, and if the interaction term was not significant, removed it from the regression. All regressions on story task data included data from both ROIs, and tested for a significant effect of ROI. Regressions on the response magnitude to the three ToM events similarly included data from all events, and tested for a significant effect of Event. We included mean translation as a between-subject predictor in all regressions, and a subject identifier as a random effect in regressions that included non-independent measurements (e.g., data from two ROIs, or three ToM events, per subject). Continuous regression variables were standardized, such that the units of the regression beta coefficients are the same. Among children, we additionally tested for significant correlations between neural measures and performance on the verbal and non-verbal ToM tasks.

Acknowledgements

We gratefully acknowledge the families and participants who made this research possible, Amy Wilson and Kriston Pumphrey for recruiting and scheduling participants, Jenny Lu, Daniel DiDonna for help with data collection, Alexa Riobueno-Naylor, Lyneé Alves, Kary Richardson, and Lauren Berger for help with data organization and coding, and Wanda Riddle for performing

the ASL stories. We would also like to thank the American School for the Deaf (Hartford, CT), The Learning Center for the Deaf (Framingham, MA), New York School for the Deaf (White Plains, NY), University of Vermont American Sign Language Program, Boston University Deaf Studies Program, American Society for Deaf Children, and Hands & Voices for distributing information about this study to potential participants. Finally, we would like to thank the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT, and gratefully acknowledge the support of this project by a Whitaker Health Sciences Fund Fellowship (to HR), and an NSF Career award (#1122374 to RS).

References

1. Wellman, H. M., Lopez-Duran, S., LaBounty, J. & Hamilton, B. Infant attention to intentional action predicts preschool theory of mind. *Developmental Psychology* **44,** 618–623 (2008).
2. Yamaguchi, M., Kuhlmeier, V. A., Wynn, K. & vanMarle, K. Continuity in social cognition from infancy to childhood. *Dev Sci* **12,** 746–752 (2009).
3. Wellman, H. M., Fang, F. & Peterson, C. C. Sequential progressions in a theory-of-mind scale: longitudinal perspectives. *Child Dev* **82,** 780–792 (2011).
4. Astington, J. W. & Jenkins, J. M. A longitudinal study of the relation between language and theory-of-mind development. *Developmental Psychology* **35,** 1311 (1999).
5. Astington, J. W. The developmental interdependence of theory of mind and language. *The roots of human sociality: Culture, cognition, and human interaction* 179–206 (2006).
6. Milligan, K., Astington, J. W. & Dack, L. A. Language and theory of mind: meta-analysis of the relation between language ability and false-belief understanding. *Child Dev* **78,** 622–646 (2007).
7. Siegal, M. & Peterson, C. C. Children's theory of mind and the conversational territory of cognitive development. *Children's early understanding of mind: Origins and development* 427–455 (1994).
8. Peterson, C. C. & Siegal, M. Representing inner worlds: Theory of mind in autistic, deaf, and normal hearing children. *Psychological Science* **10,** 126–129 (1999).
9. de Villiers, J. G. & Pyers, J. E. Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cognitive Development* **17,** 1037–1060 (2002).
10. Hale, C. M. & Tager-Flusberg, H. The influence of language on theory of mind: a training study. *Dev Sci* **6,** 346–359 (2003).
11. de Villiers, J. G. & de Villiers, P. A. in *Language acquisition* 169–195 (Springer, 2009).
12. Clements, W. A. & Perner, J. Implicit understanding of belief. *Cognitive Development* **9,** 377–395 (1994).
13. Geren, J., Snedeker, J., Shafto, C. L. & Geren, J. The Link between Language and Theory of Mind: Evidence from Internationally-Adopted Children. (2009).
14. Onishi, K. H. & Baillargeon, R. Do 15-month-old infants understand false beliefs? *Science* **308,** 255–258 (2005).
15. He, Z., Bolz, M. & Baillargeon, R. False-belief understanding in 2.5-year-olds: evidence from violation-of-expectation change-of-location and unexpected-contents tasks. *Dev Sci* **14,** 292–305 (2011).
16. Scott, R. M., He, Z., Baillargeon, R. & Cummins, D. False-belief understanding in 2.5-year-olds: Evidence from two novel verbal spontaneous-response tasks. *Dev Sci* **15,** 181–193 (2012).
17. Ruffman, T., Slade, L. & Crowe, E. The relation between children's and mothers' mental state language and theory-of-mind understanding. *Child Dev* **73,** 734–751 (2002).
18. Moeller, M. P. & Schick, B. Relations between maternal input and theory of mind understanding in deaf children. *Child Dev* **77,** 751–766 (2006).
19. Schick, B. & Hoffmeister, R. ASL skills in deaf children of deaf parents and of hearing parents. in (2001).
20. Schick, B., De Villiers, P., De Villiers, J. & Hoffmeister, R. Language and theory of mind: A study of deaf children. *Child Dev* **78,** 376–396 (2007).

21. Woolfe, T., Want, S. C. & Siegal, M. Signposts to development: Theory of mind in deaf children. *Child Dev* **73,** 768–778 (2002).
22. Saxe, R. & Kanwisher, N. People thinking about thinking people: the role of the temporo-parietal junction in 'theory of mind'. *NeuroImage* **19,** 1835–1842 (2003).
23. Adolphs, R. The Social Brain: Neural Basis of Social Knowledge. *Annu. Rev. Psychol.* **60,** 693–716 (2009).
24. Koster-Hale, J. & Saxe, R. Functional neuroimaging of theory of mind. *Understanding other minds: Perspectives from developmental social neuroscience* 132–163 (2013).
25. Saxe, R. R., Whitfield-Gabrieli, S., Scholz, J. & Pelphrey, K. A. Brain regions for perceiving and reasoning about other people in school-aged children. *Child Dev* **80,** 1197–1209 (2009).
26. Gweon, H., Dodell-Feder, D., Bedny, M. & Saxe, R. Theory of Mind Performance in Children Correlates With Functional Specialization of a Brain Region for Thinking About Thoughts. *Child Dev* **83,** 1853–1868 (2012).
27. Richardson, H., Lisandrelli, G., Riobueno-Naylor, A. & Saxe, R. Development of the social brain from age three to twelve years. *Nature Communications* **9,** 1027 (2018).
28. Premack, D. & Woodruff, G. Does the chimpanzee have a theory of mind? *Behav Brain Sci* **1,** 515–526 (1978).
29. Wimmer, H. & Perner, J. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *COGNITION* **13,** 103–128 (1983).
30. Dennett, D. C. Beliefs about beliefs [P&W, SR&B]. (1978).
31. Wellman, H. M., Cross, D. & Watson, J. Meta-analysis of theory-of-mind development: the truth about false belief. *Child Dev* **72,** 655–684 (2001).
32. Baron-Cohen, S., Leslie, A. M. & Frith, U. Does the autistic child have a 'theory of mind'? *COGNITION* **21,** 37–46 (1985).
33. Baron-Cohen, S. The autistic child's theory of mind: A case of specific developmental delay. *Journal of Child Psychology and Psychiatry* **30,** 285–298 (1989).
34. Sabbagh, M. A., Xu, F., Carlson, S. M., Moses, L. J. & Lee, K. The development of executive functioning and theory of mind a comparison of Chinese and US preschoolers. *Psychological Science* **17,** 74–81 (2006).
35. Peterson, C. C., Wellman, H. M. & Slaughter, V. The mind behind the message: Advancing theory-of-mind scales for typically developing children, and those with deafness, autism, or Asperger syndrome. *Child Dev* **83,** 469–485 (2012).
36. Reher, K. (Producer), & Sohn, P. (Director). *Partly Cloudy* [Motion Picture]. United States: Pixar Animation Studios and Walt Disney Pictures (2009).
37. Jacoby, N., Bruneau, E., Koster-Hale, J. & Saxe, R. Localizing Pain Matrix and Theory of Mind networks with both verbal and non-verbal stimuli. *NeuroImage* **126,** 39–48 (2016).
38. de Villiers, J. G. IQ Can Language Acquisition Give Children a Point ofView. *Why language matters for theory of mind* **186,** (2005).
39. De Villiers, J. & Pyers, J. Complementing cognition: The relationship between language and theory of mind. in **1,** 136 (Cascadilla Press, 1997).
40. San Juan, V. & Astington, J. W. Bridging the gap between implicit and explicit understanding: How language development promotes the processing and representation of false belief. *British Journal of Developmental Psychology* **30,** 105–122 (2011).
41. Hall, M. L., Eigsti, I.-M., Bortfeld, H. & Lillo-Martin, D. Auditory deprivation does not

impair executive function, but language deprivation might: evidence from a parent-report measure in deaf native signing children. *The Journal of Deaf Studies and Deaf Education* **22,** 9–21 (2016).
42. Mayberry, R. I. & Eichen, E. B. The long-lasting advantage of learning sign language in childhood: Another look at the critical period for language acquisition. *Journal of Memory and Language* **30,** 486–512 (1991).
43. Newman, A. J., Bavelier, D., Corina, D., Jezzard, P. & Neville, H. J. A critical period for right hemisphere recruitment in American Sign Language processing. *Nature Publishing Group* **5,** 76 (2002).
44. Peterson, C. C. Theory-of-mind development in oral deaf children with cochlear implants or conventional hearing aids. *Journal of Child Psychology and Psychiatry* **45,** 1096–1106 (2004).
45. Courtin, C. & Melot, A.-M. Development of theories of mind in deaf children. *Psychological perspectives on deafness* **2,** (1998).
46. Enns, C. J. & Herman, R. C. Adapting the assessing british sign language development: Receptive skills test into American sign language. *Journal of deaf studies and deaf education* **16,** 362–374 (2011).
47. Kaufman, A. S. KBIT-2: Kaufman Brief Intelligence Test. Minneapolis, MN: NCS Pearson. (1997).
48. Vandierendonck, A., Kemps, E., Fastame, M. C. & Szmalec, A. Working memory components of the Corsi blocks task. *British Journal of Psychology* **95,** 57–79 (2004).
49. Kessels, R. P., Van Zandvoort, M. J., Postma, A., Kappelle, L. J. & De Haan, E. H. The Corsi block-tapping task: standardization and normative data. *Applied neuropsychology* **7,** 252–258 (2000).
50. Keil, B. *et al.* Size-optimized 32-channel brain arrays for 3 T pediatric imaging. *Magn. Reson. Med.* **66,** 1777–1787 (2011).
51. Thesen, S., Heid, O., Mueller, E. & Schad, L. R. Prospective acquisition correction for head motion with image-based tracking for real-time fMRI. *Magn. Reson. Med.* **44,** 457–465 (2000).
52. Asendorpf, J. B. *et al.* Recommendations for increasing replicability in psychology. *European Journal of Personality* **27,** 108–119 (2013).
53. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nat. hum. behav.* **1,** 0021 (2017).
54. Whitfield-Gabrieli, S., Nieto-Castanon, A. & Ghosh, S. Artifact Detection Tools (ART). *Cambridge, MA. Release version* **7,** 11 (2011).
55. Carp, J. Optimizing the order of operations for movement scrubbing: Comment on Power et al. *NeuroImage* **76,** 436–438 (2013).
56. Hallquist, M. N., Hwang, K. & LUNA, B. The nuisance of nuisance regression: spectral misspecification in a common approach to resting-state fMRI preprocessing reintroduces noise and obscures functional connectivity. *NeuroImage* **82,** 208–225 (2013).
57. Behzadi, Y., Restom, K., Liau, J. & Liu, T. T. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* **37,** 90–101 (2007).
58. Sabbagh, M. A., Bowman, L. C., Evraire, L. E. & Ito, J. M. B. Neurodevelopmental correlates of theory of mind in preschool children. *Child Dev* **80,** 1147–1162 (2009).
59. Dufour, N. *et al.* Similar Brain Activation during False Belief Tasks in a Large Sample of

Adults with and without Autism. *PLoS ONE* **8,** e75468 (2013).
60. Desmond, J. E. *et al.* Functional MRI measurement of language lateralization in Wada-tested patients. *Brain* **118,** 1411–1419 (1995).

Supplementary Materials

Analysis of Response Selectivity in RTPJ and DMPFC Group ROIs (ASL Story Task)

We tested if delayed access to linguistic input results in delayed or disrupted functional specialization of ToM brain regions using pre-specified group regions of interest. Among child participants, there was a significant negative effect of age of ASL onset (negative effect of age of ASL onset: b=-.50, t=-2.1, p=.048; positive effect of age: b=.41, t=2.3, p=.03; NS effect of ROI: b=.32, t=1.3, p=.19; NS effect of motion: b=.08, t=.48, p=.64; marginal age of ASL-onset by age interaction: b=.59, t=1.9, p=.08). The effect of age of ASL onset remained significant when additionally including non-verbal IQ as a covariate (negative effect of age of ASL onset: b=-.53, t=-2.2, p=.045; positive effect of age: b=.40, t=2.1, p=.049; NS effect of ROI: b=.36, t=1.4, p=.16; NS effect of motion: b=.06, t=.36, p=.73; NS effect of non-verbal IQ: b=-.15, t=-.76, p=.46, marginal age of ASL-onset by age interaction: b=.56, t=1.8, p=.09).

The negative effect of age of ASL onset was not significant in the full sample (children and adults; NS effect of age of ASL onset: b=-.17, t=-1.1, p=.26; NS effect of age group: b=.19, t=.80, p=.43; NS effect of ROI: b=.03, t=.22, p=.83; NS effect of motion: b=.15, t=1.0, p=.32), and there was no effect of delayed access to language on response selectivity in among adults (NS effect of age of ASL: b=-.11, t=-.50, p=.62; NS effect of ROI: b=-.18, t=-.83, p=.41; NS effect of motion: b=.07, t=.32, p=.75).

We did not find evidence for a relationship between response selectivity and performance on either ToM behavioral task (Verbal: NS effect of verbal ToM: b=-.004, t=-.02, p=.98, NS effect of ROI: b=.38, t=1.7, p=.11, NS effect of motion: b=.11, t=.52, p=.61; Non-verbal: NS effect of non-verbal ToM: b=.05, t=.26, p=.80, NS effect of ROI: b=.32, t=1.4, p=.18, NS effect of motion: b=.16, t=.90, p=.38).

Exploratory Analysis of Response Selectivity in Other ToM ROIs (ASL Story Task)

We repeated the analysis of the selectivity of responses during the story task in other regions recruited for ToM reasoning: left temporoparietal junction (LTPJ), middle medial prefrontal cortex (MMPFC), ventro-medial prefrontal cortex (VMPFC) and precuneus (PC). As specified in the analysis plan, analysis of the responses in these regions was considered exploratory because there is no previous evidence for a relationship between selectivity in these regions and ToM behavioral development in childhood (https://osf.io/kyu3f/). Individual ROIs were created using the same procedure as described for RTPJ and DMPFC ROIs.

We tested if delayed access to linguistic input results in delayed or disrupted functional specialization of these ToM brain regions using individual ROIs and pre-specified group ROIs. Among child participants, there was no effect of age of ASL onset on response selectivity, in individual ROIs (NS effect of age of ASL onset: b=-.21, t=-1.47, p=.16, NS effect of age: b=.19, t=1.4, p=.19, NS effect of ROI (MMPFC): b=-.07, t=-.26, p=.80, NS effect of ROI (PC): b=-.24, t=-.90, p=.37, NS effect of ROI (VMPFC): b=-.08, t=-.26, p=.80, NS effect of motion: b=.19, t=1.4, p=.18).

There was similarly no effect of age of ASL onset in the full sample (children and adults; NS effect of age of ASL onset: b=-.08, t=-.64, p=.53, NS effect of age group: b=.24, t=1.2, p=.24, NS effect of ROI (MMPFC): b=-.19, t=-1.2, p=.24, effect of ROI (PC): b=-.32, t=-2.1, p=.04, NS

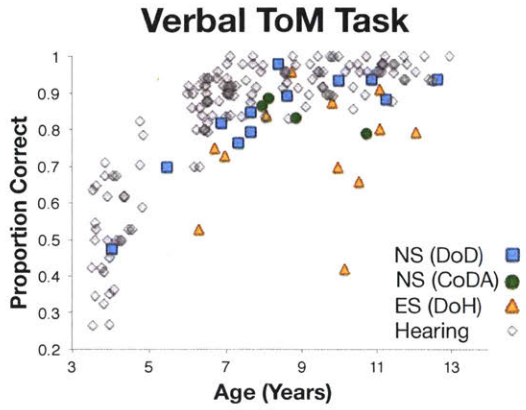
effect of ROI (VMPFC): b=-.21, t=-1.3, p=.19, NS effect of motion: b=.16, t=1.3, p=.21), or among adults alone (NS effect of age of ASL onset: b=.05, t=.23, p=.82, NS effect of ROI (MMPFC): b=-.25, t=-1.3, p=.21, NS effect of ROI (PC): b=-.37, t=-1.9, p=.06, NS effect of ROI (VMPFC): b=-.29, t=-1.4, p=.15, NS effect of motion: b=.05, t=.26, p=.80).

Random Effects Analysis

Whole-brain analyses were used to examine the main contrast of interest (Story: Mental > Physical; Movie: Mental > Pain), per group (native and early/non-native signers, in children and adults). These analyses were corrected for multiple comparisons by estimating the false-positive rate via 5,000 Monte Carlo permutations using the SnPM toolbox for SPM5 (http://www.fil.ion.ucl.ac.uk/spm/software/spm5/; $p<.05$). To view the difference in response to this contrast across visits, we additionally ran a corrected random effects analysis on the between-group contrast difference (NS > NNS Mental > Physical and NS > NNS Mental > Pain, per age group). No significant clusters were identified in the between group analysis. The random effects analysis results are shown in Supplementary Figure 5. In some cases, the corrected analysis prevented visualization of (subthreshold) activation. For visualization purposes, activation is shown at the same thresholds in native vs. non-native signing groups. The child story task activations and adult movie task activations are shown at $p<.01$, $k=50$, uncorrected. The child movie task activations and adult story task activations show the analyses corrected for multiple comparisons ($p<.05$).

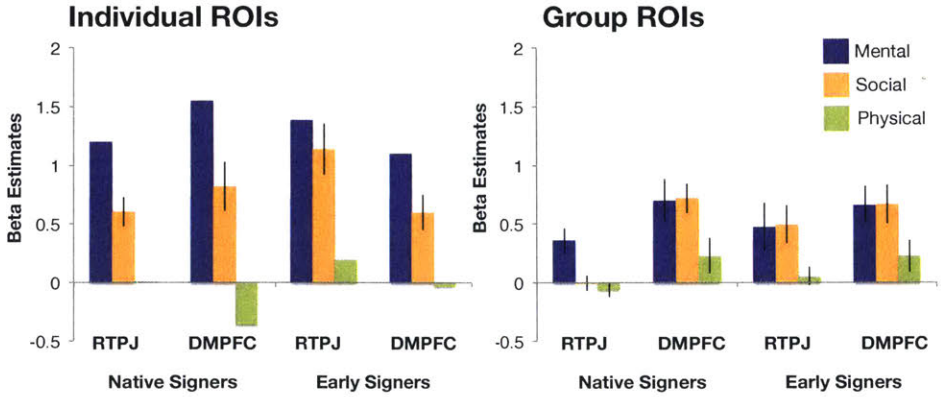
Supplementary Figures and Tables

Supplementary Figure 1



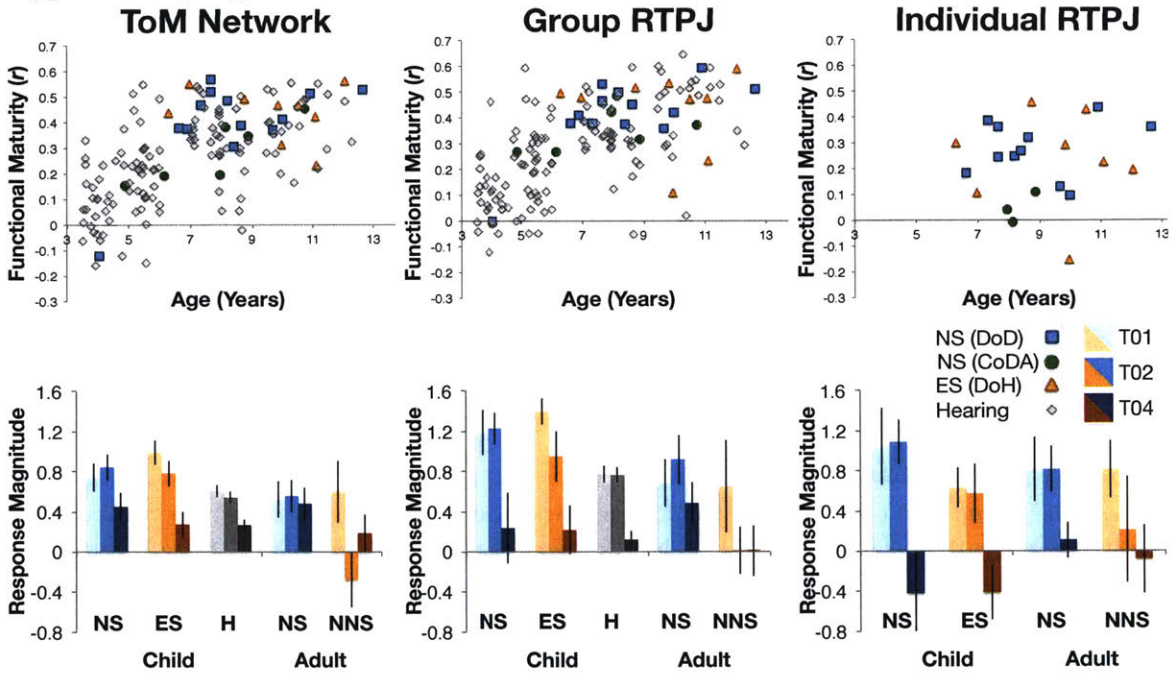
Supplementary Figure 1. Verbal ToM Task in Larger Participant Pool. Scatterplot shows proportion correct on the verbal ToM task (y-axis) by age (x-axis) in the current participants (blue squares: native signing deaf children born to d/Deaf parents, green circles: native signing hearing children born to d/Deaf parents, orange triangles: early signing children (born to hearing parents)), and in hearing participants from other studies who completed an analogous task in English (grey diamonds).

Supplementary Figure 2



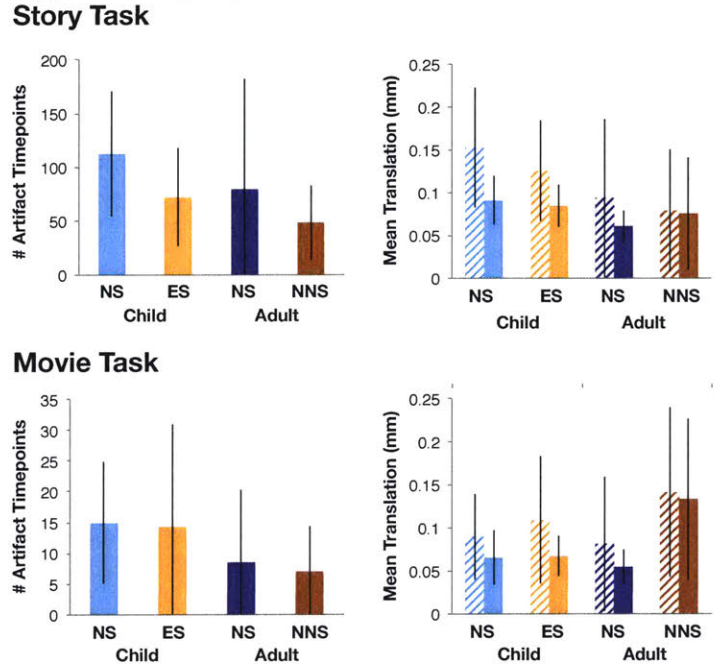
Supplementary Figure 2. Average beta values per condition, in child participants. Bars show average and standard error for beta values per condition, for each ROI (RTPJ and DMPFC), by group (native and early signing children). The plot on the left shows data from individual ROIs defined based on the Mental and Physical conditions. Because these conditions are non-independent from ROI definition, they are plotted for visualization purposes only (and therefore do not have standard error bars). The plot on the right shows data from group ROIs, in which every condition is independent from ROI definition.

Supplementary Figure 3



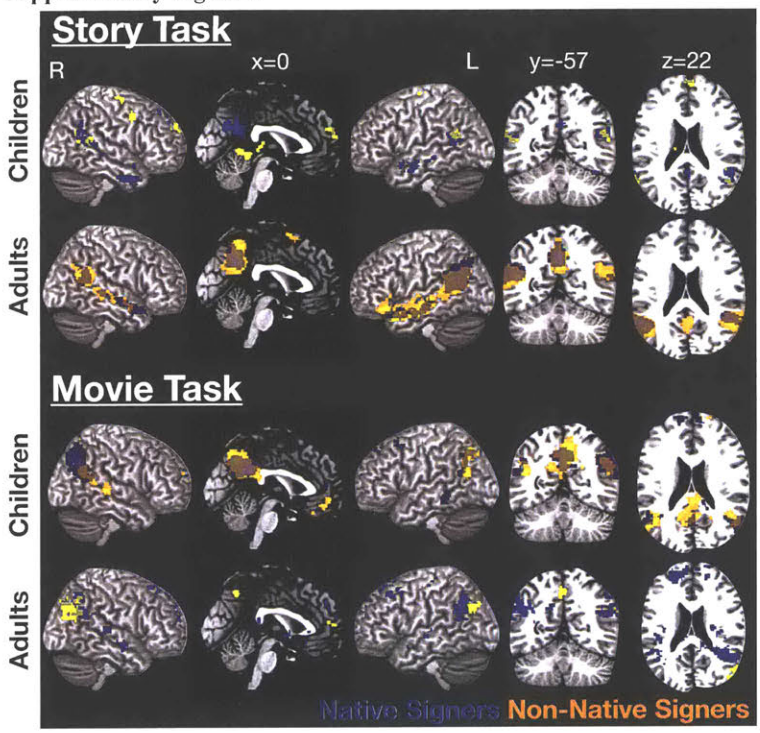
Supplementary Figure 3. Neural Responses to Movie in Larger Participant Pool. Scatterplots (top row) show functional maturity (i.e., similarity to average adult timecourse, Pearson's *r*) on the y-axis, and age on the x-axis in the current participants (blue squares: native signing deaf children born to d/Deaf parents, green circles: native signing hearing children born to d/Deaf parents, orange triangles: early signing children (born to hearing parents)), and in hearing participants from other studies who completed this fMRI task (grey diamonds). The left scatterplot shows functional maturity of the ToM network (averaging across all ToM regions), the middle scatterplot shows functional maturity in the group RTPJ ROI, and the right scatterplot shows functional maturity in individual RTPJ ROIs (which were not available for hearing participants). The bottom row shows mean response magnitude in the ToM network (left), in group RTPJ ROIs (middle) and in individual RTPJ ROIs (right) in native signing (NS, blue), early or non-native singing (ES, NNS, orange) children and adults, and hearing children (H, grey), to three ToM events (T01, T02, T04), which previously showed developmental change with age (T01, T02) and ToM score (T04).

Supplementary Figure 4



Supplementary Figure 4. Amount of motion in fMRI data. Left: Number of artifact timepoints identified in the timecourse of response (Top: Story task, Bottom: Movie task), by group (native signing (NS), and early signing (ES) children, and NS and non-native signing (NNS) adults). Artifact timepoints are timepoints in which there is 2mm motion and/or a global signal change greater than three standard deviations from the mean, relative to the previous timepoint. Error bars show standard deviation from the mean. **Right:** Mean translation (motion in x, y, z directions) in millimeters per group, including (striped) and excluding (solid) artifact timepoints. Error bars show standard deviation from the mean.

Supplementary Figure 5



Supplementary Figure 5. Random Effects Analysis. Whole-brain analyses were used to examine the main contrast of interest (Story: Mental > Physical; Movie: Mental > Pain), per group (native and early/non-native signers, in children and adults). For visualization purposes, activation is shown at the same thresholds in native (blue) vs. non-native signing groups (orange). The child story task activations and adult movie task activations are shown at $p<.01$, $k=50$, uncorrected. The child movie task activations and adult story task activations show the analyses corrected for multiple comparisons ($p<.05$). Subtraction analyses (NS > NNS) revealed no significant clusters.

Supplementary Table 1

Subject	Age	Gender	Hand	Age of ASL Onset	Parent Hearing Status	Native vs. Non-Native Signer	ASL-Rec	KBIT-Std	CORSI	ToM-V	ToM-NV	Num Art (Story)	Mean Trans (Story)	Num Art (Movie)	Mean Trans (Movie)	fMRI Data (Story)	fMRI Data (Movie)
C_01	11.25	F	R	0	DOD	NS	0.98	115	NA	0.884	0.862	31	0.0985782	NA	NA	1	0
C_02	6.29	M	L	0.25	DOH	ES	0.83	NA	NA	0.529	0.615	51	0.0898757	2	0.0754022	1	1
C_03	8.12	M	R	0	CODA	NS	0.75	116	NA	0.889	0.667	106	0.0597092	1	0.0273287	1	1
C_04	4.87	M	R	0	CODA	NS	0.50	104	NA	NA	0.400	NA	NA	21	0.0383952	0	1
C_05	10.73	F	R	0	CODA	NS	0.88	120	5	0.791	0.852	99	0.0675837	22	0.0416039	1	1
C_06	7.96	F	R	0	CODA	NS	0.65	116	5	0.867	0.833	106	0.144713	4	0.112514	1	1
C_07	9.97	M	R	4.5	DOH	ES	0.83	102	NA	0.700	0.833	137	0.0993137	4	0.0901133	1	1
C_08	8.08	F	R	1	DOH	ES	0.83	115	5	0.841	0.800	NA	NA	NA	NA	0	0
C_09	6.13	F	R	0	CODA	NS	0.40	86	NA	NA	0.621	NA	NA	2	0.048429	0	1
C_10	8.87	F	R	0	CODA	NS	0.68	119	5	0.833	0.842	196	0.0933589	17	0.07862	1	1
C_11	8.20	M	R	0	DOD	NS	NA	143	5	NA	0.767	102	0.113929	27	0.0583015	1	1
C_12	6.61	M	R	0	DOD	NS	0.55	109	5	NA	0.600	198	0.0757464	2	0.0526384	1	1
C_13	9.67	M	L	0	DOD	NS	NA	122	5	NA	0.767	186	0.142535	16	0.114878	1	1
C_14	7.32	M	R	0	DOD	NS	0.43	87	4	0.766	0.667	111	0.103322	21	0.109	1	1
C_15	9.99	M	R	0	DOD	NS	0.68	128	5	0.936	0.733	51	0.124671	17	0.0693783	1	1
C_16	12.66	M	R	0	DOD	NS	0.88	117	5	0.938	0.931	91	0.0716908	14	0.0504281	1	1
C_17	7.65	F	R	0	DOD	NS	0.88	132	4	0.848	0.867	89	0.0607219	14	0.0468602	1	1
C_18	6.89	M	R	0	DOD	NS	0.80	124	4	0.818	0.688	NA	NA	25	0.0631707	0	1
C_19	8.40	F	R	0	DOD	NS	0.90	111	4	0.979	0.867	175	0.0902428	13	0.0462793	1	1
C_20	4.03	F	R	0	DOD	NS	0.56	101	NA	0.474	0.750	NA	NA	40	0.141186	0	1
C_21	5.48	M	R	0	DOD	NS	0.53	96	3	0.700	0.533	NA	NA	NA	NA	0	0
C_22	7.65	M	R	0	DOD	NS	0.78	130	5	0.795	0.867	182	0.0832855	10	0.0478129	1	1
C_23	8.62	F	L	0	DOD	NS	0.88	124	5	0.894	0.867	55	0.0641405	11	0.0546203	1	1
C_24	10.90	F	L	0	DOD	NS	0.65	129	6	0.939	0.800	26	0.0611548	7	0.0428594	1	1
C_25	6.71	F	R	1.83	DOH	ES	0.80	135	NA	0.750	0.667	NA	NA	NA	NA	0	0
C_26	11.09	M	R	2	DOH	ES	0.83	110	5	0.911	0.900	95	0.134981	12	0.108593	1	1
C_27	12.06	M	R	0.75	DOH	ES	0.78	118	NA	0.796	0.867	18	0.069598	3	0.0421477	1	1
C_28	10.52	M	R	7	DOH	ES	0.73	95	5	0.660	0.767	5	0.0565136	7	0.0680672	1	1
C_29	6.96	F	R	1.5	DOH	ES	0.80	104	5	0.729	0.667	108	0.0861463	36	0.0516144	1	1
C_30	8.74	F	R	4	DOH	ES	0.65	111	4	0.959	0.867	68	0.0628953	5	0.0362975	1	1
C_31	10.15	M	R	6	DOH	ES	0.65	60	4	0.422	0.500	NA	NA	NA	NA	0	0
C_32	9.83	F	R	1.5	DOH	ES	0.83	142	6	0.875	0.967	96	0.0778488	11	0.0708923	1	1
C_33	11.10	M	R	4	DOH	ES	0.90	132	5	0.804	0.900	NA	NA	49	0.060906	0	1
A_01	22	F	R	0	DOD	NS						52	0.0533963	NA	NA	1	0
A_02	39	M	R	2	DOH	ES						7	0.107568	NA	NA	1	0
A_03	33	F	R	2	DOH	ES						112	0.129437	NA	NA	1	0
A_04	29	M	R	1.5	DOH	ES						11	0.0790244	NA	NA	1	0
A_05	22	F	R	2.5	DOH	ES						11	0.0445666	NA	NA	1	0
A_06	64	M	R	11	DOH	LS						47	0.198658	3	0.146042	1	1
A_07	20	F	R	0	DOH	NS						210	0.086899	7	0.0742698	1	1
A_08	32	M	L	0	DOH	NS						31	0.040725	4	0.0391772	1	1
A_09	34	F	R	0	CODA	NS						30	0.0565903	2	0.0350653	1	1
A_10	27	M	L	0	DOD	NS						23	0.068427	15	0.072192	1	1
A_11	31	F	R	1.5	DOH	ES						67	0.0727897	17	0.0806826	1	1
A_12	29	M	L	4	DOH	ES						52	0.0495442	7	0.0668267	1	1
A_13	24	F	R	0	DOH	NS						114	0.0660962	NA	NA	1	0
A_14	29	M	R	3	DOH	ES						98	0.0579698	6	0.0790627	1	1
A_15	22	F	R	0	CODA	NS						101	0.0598114	0	0.0660624	1	1
A_16	32	M	R	0	DOD	NS						53	0.0665273	1	0.0524593	1	1
A_17	23	M	R	0	CODA	NS						275	0.0681632	48	0.0421734	1	1
A_18	45	M	L	7	DOH	ES						46	0.101007	23	0.0782704	1	1
A_19	40	F	R	0	DOD	NS						21	0.0419209	2	0.0236634	1	1
A_20	47	F	R	0	CODA	NS						27	0.0627191	11	0.0729769	1	1
A_21	23	M	R	0	DOD	NS						31	0.0642784	4	0.0625428	1	1
A_22	25	M	R	0	DOD	NS						360	0.113317	27	0.0939957	1	1
A_23	21	F	R	0	DOD	NS						44	0.0597579	4	0.0532892	1	1
A_24	54	F	R	0	DOD	NS						27	0.0419381	11	0.0359488	1	1
A_25	29	M	R	2	DOH	ES						25	0.0764344	7	0.0700767	1	1
A_26	62	M	R	18	DOH	LS						78	0.167055	NA	NA	1	0
A_27	31	F	L	0	DOD	NS						24	0.0396588	2	0.0345828	1	1
A_28	24	M	R	0	CODA	NS						67	0.0462784	3	0.0257655	1	1
A_29	24	F	R	0	CODA	NS						27	0.0444459	2	0.0480092	1	1
A_30	37	F	R	0	CODA	NS						6	0.0543906	4	0.0578326	1	1
A_31	45	M	R	3	DOH	ES						2	0.0768339	1	0.0806965	1	1
A_32	40	M	R	20	DOH	LS						0	0.175384	0	0.189842	1	1
A_33	48	M	R	15	DOH	LS						51	0.255586	10	0.275847	1	1
A_34	44	M	NA	7	DOH	ES						44	0.244316	2	0.332669	1	1
A_35	21	F	NA	4	DOH	ES						63	0.0644823	1	0.067037	1	1
A_36	41	F	R	0	DOD	NS						20	0.0753081	6	0.0782731	1	1

Supplementary Table 1. Participant Demographics. Subjects C_# indicate children, subjects A_# indicate adults. Age and age of ASL onset are in years. Parent hearing status indicates participants who are deaf children of d/Deaf adults (DoD), hearing children of d/Deaf adults (CoDA), and d/Deaf children of hearing adults (DoH). Native vs. non-native signer indicates native signers (NS), and signers who received exposure to ASL by age seven years (early signers, ES) or later (late signers, LS). ASL-Rec indicates score on a test of receptive vocabulary (ASL-RST), KBIT-Std indicates standardized score on the non-verbal IQ subtest, CORSI indicates blocks passed on computerized working memory task. ToM-V is proportion correct on the verbal ToM task; ToM-NV is proportion correct on the non-verbal ToM task. NumArt (Story/Movie) indicates the number of artifact timepoints during the story and movie fMRI tasks, respectively; MeanTrans (Story/Movie) indicates the mean amount of translation (movement in x-, y-, and z- axes) between timepoints during the story and movie fMRI tasks, respectively. fMRI Data (Story/Movie) indicates whether a participant contributed usable data to the story and movie tasks, respectively.

Supplementary Table 2

INDIVIDUAL Regions of Interest						
Age Group	**ROI**	**ASL Group**	**# Identifed**	**Peak Coordinate**	**M(SD) N Voxels**	**M(SD) Peak T**
Child	RTPJ	NS	15/16	[53 -47 18]	161 (105)	6.9 (2.3)
		ES	8/8	[57 -49 21]	133 (103)	6.6 (1.7)
	DMPFC	NS	12/16	[-2 54 30]	130 (90)	6.7 (1.4)
		ES	6/8	[4 56 28]	156 (90)	6.3 (1.3)
	LTPJ	NS	16/16	[-53 -58 20]	138 (99)	6.5 (1.9)
		ES	8/8	[-54 -57 29]	124 (74)	7.1 (1.5)
	MMPFC	NS	12/16	[4 54 14]	128 (78)	6.7 (1.3)
		ES	7/8	[-1 58 12]	147 (112)	6.6 (1.9)
	VMPFC	NS	8/16	[2 56 -9]	112 (103)	5.6 (1.4)
		ES	7/8	[-1 45 -11]	82 (76)	5.5 (1.4)
	PC	NS	14/16	[0 -54 33]	170 (119)	6.3 (2.0)
		ES	6/8	[3 -55 37]	117 (127)	5.8 (1.7)
Adult	RTPJ	NS	20/20	[54 -51 20]	179 (91)	8.0 (1.9)
		NNS	16/16	[56 -47 19]	179 (99)	7.9 (2.1)
	DMPFC	NS	19/20	[-4 55 32]	88 (54)	6.9 (1.8)
		NNS	16/16	[-1 54 30]	116 (88)	7.0 (2.9)
	LTPJ	NS	20/20	[-53 -59 22]	221 (89)	8.9 (1.6)
		NNS	16/16	[-52 -57 21]	210 (103)	8.5 (2.4)
	MMPFC	NS	18/20	[-1 58 16]	86 (77)	6.2 (1.9)
		NNS	14/16	[3 56 14]	134 (99)	7.3 (2.1)
	VMPFC	NS	16/20	[0 56 -15]	104 (52)	7.3 (2.2)
		NNS	16/16	[-1 54 -14]	99 (76)	6.3 (1.8)
	PC	NS	20/20	[-1 -57 39]	203 (105)	7.3 (1.6)
		NNS	15/16	[-5 -55 37]	199 (99)	7.7 (2.4)
GROUP Regions of Interest						
STORY	**ROI**			**Center Coordinate**	**N Voxels**	
ToM	RTPJ			[54 -52 23]	463	
	LTPJ			[-52 -58 25]	379	
	PC			[1 -56 34]	498	
	DMPFC			[-1 53 29]	455	
	MMPFC			[1 54 12]	498	
	VMPFC			[1 50 -12]	498	
	RSTS			[55 -10 -16]	172	
Overlap	RSTS/RMidAntTemp			N/A	326	
Language	RMidAntTemp			[55 -14 -13]	210	
	LMidAntTemp			[-55 -18 -13]	536	
	LAntTemp			[-52 2 -18]	515	
	RMidPostTemp			[58 -45 10]	463	
	LMidPostTemp			[-56 -40 10]	515	
	LPostTemp			[-48 -62 15]	379	
	LAngGyrus			[-37 -76 30]	498	
	LSFG			[-7 50 41]	461	
	LMFG			[-40 -2 53]	498	
	LIFGOrb			[-48 33 -4]	498	
	LIFG			[-48 16 24]	515	
MOVIE	**ROI**			**Center Coordinate**	**N Voxels**	
ToM	RTPJ			[48 -60 30]	376	
	LTPJ			[-48 -62 30]	368	
	PC			[0 -54 34]	382	
	DMPFC			[-6 54 36]	217	
	MMPFC			[-4 58 16]	275	
	VMPFC			[-4 56 -16]	198	
Pain	RS2			[60 -28 38]	368	
	LS2			[-62 -32 34]	269	
	Rinsula			[42 6 -6]	309	
	Linsula			[-42 -2 -4]	240	
	RMFG			[50 42 12]	142	
	LMFG			[-46 36 14]	256	
	AMCC			[0 2 42]	249	

Supplementary Table 2. Individual and Group Regions of Interest Top half of table summarizes information about individual regions of interest, identified functionally in individual participants. ASL group indicates native (NS) vs. early or non-native (ES, NNS) signing participants. # Identified is number of participants in whom an ROI was successfully identified at $p<.001$, $k=10$ thresholds, to the Mental > Physical contrast. Peak coordinates are in mm space. Bottom half of table provides information about group ROIs.

Chapter 5: Conducting Pediatric fMRI Experiments: Challenges and Strategies

Conducting pediatric neuroimaging experiments to learn about cognitive development involves addressing many of the same challenges that cognitive neuroscientists and cognitive scientists face: designing elegant paradigms that address the research question at hand, recruiting participants, controlling for confounds, and ruling out competing explanations. However, each challenge is somewhat exaggerated in developmental research. Paradigms must not only address the research question at hand and control for confounds; they must also be feasible for the participants to complete. Recruitment of participants involves engaging with the community, and communicating about research protocols and results with parents and children. And ruling out competing explanations is complicated by the fact that individual child participants may not be able to complete all parts of an ideal study, or any parts of an ideal study. Any issue of interpreting what adults *mean* when they provide a behavioral response or explanation is made more complicated in child research by greater variability in introspective or expressive ability among children. And of course, studying the development of a particular cognitive ability is complicated by the fact that so many cognitive abilities undergo dramatic change with age. Throughout this thesis I have touched on various challenges to using cognitive neuroscience tools, and specifically MRI, to test hypotheses in cognitive development. Here, I focus on those challenges in depth, because in some cases, the cognitive implications of neuroimaging studies are most limited by our ability to make meaningful measurements.

Data Collection

Collecting sufficient, high quality MRI data from children is very difficult. Participating in an MRI experiment can be a stressful experience. The experiments require children to lie on their back in a dark, noisy tube, alone, and hold completely still (<2 mm motion) for a long time (typically 30 to 120 minutes). To collect stable measurements of neural responses, given all of the other sources of noise in MRI, conditions must be repeated many times (>6 measurements per condition, typically), so the experiments are often repetitive and boring. These demands are challenging for anyone, but can prove insurmountable for many populations, including very young children (e.g. four years and under) and children with developmental disorders.

Labs that scan children have developed many techniques to address these challenges. In the Saxe lab, for example, children prepare for a visit to the lab in advance. We send participants a storybook that includes pictures of the researchers and the testing environment, and a mnemonic device to remember to lie still (the three Ss: still, soft, and super-duper!). Once at MIT, children practice being scanned in a "mock scanner", which is designed to look, feel, and sound like a real scanner. We play the noises of the scanner over speakers, and provide feedback on whether or not they are laying still enough. Mock scans have been shown to reduce participant movement and increase rates of scan completion and data retention[1,2]. In the "real" scanner, children can choose to be scanned hugging our scanner buddy, a large plush dog. This scanner buddy not only helps children to feel comfortable and calm, but also (anecdotally) prevents children from fidgeting with their hands or touching their face. An experimenter also stands next to the child throughout the scan, and uses a gentle touch on the leg as a reminder to stay still. Children lie with their head held snugly in a custom-made child-sized head coil[3], surrounded by soft padding held in place by medical tape.

Finally, we aim to make our experimental paradigms easy and "naturalistic". Children typically listen to short stories presented in child-directed speech, or watch short, animated movies. Movie paradigms are particularly suitable for studying children under age five years: they can be tailored in length, are engaging, and do not require completing potentially discouraging tasks. Additionally, participant motion tends to be reduced during movie paradigms[4].

Even with all of these strategies, the data we collect from children are often noisier than data collected in comparable experiments with adults. The lower data quality poses many challenges for analysis. For example, in the analyses described in this thesis, I detected and excluded data from individual timepoints when the child moved more than 2mm, and excluded all of the data from any participant whose dataset is less than 65% complete after excluding these timepoints. Nevertheless, there is a trade-off: excluding too little leaves the data extremely noisy and uninterpretable, but excluding too much may result in a non-representative sample (e.g. excluding all of the younger participants) and/or leave the dataset underpowered to detect real effects. Furthermore, setting each of the many analysis parameters allows for high "researcher degrees of freedom"[5], which is particularly threatening since neuroimaging studies of children are often already under-powered (due to smaller sample sizes, fewer experimental trials, and less data retention) and facing a multiple-comparisons problem[6,7]. One challenge for developmental cognitive neuroscience will therefore be to develop techniques and standards that support strong confirmatory tests in independent data.

A second challenge for pediatric neuroimaging research is obtaining large, nationally representative samples. In behavioral research, online platforms have been developed in order to obtain these kinds of samples (e.g., Amazon's Mechanical Turk: https://www.mturk.com/mturk/welcome, or for developmental studies, Lookit: https://lookit.mit.edu/[8,9]). Scaling up pediatric neuroimaging studies in this way is more difficult. In the Saxe lab, children visit with at least one parent or legal guardian, and the testing day (usually a weekend, given school schedules) takes two to four hours. Though all studies are ultimately comprised of participants who have the time, ability, and interest to participate, the increased requirements of pediatric neuroimaging studies often result in samples that are not diverse in race or socioeconomic status. Recently, larger organizations have committed to collecting and sharing large, nationally representative neuroimaging databases[10-12]. This effort is critical for developing robust measures of brain development that are generalizable to diverse samples.

Relating Neural Measures to Cognitive Change

Given an aspect of the neural response that changes with development, how do we assess its causal link to behavioral change? The results of studies that relate neural and behavioral change depend in part on the strength of the behavioral measures used. Behavioral performance on the cognitive task of interest may best be measured independently, outside of the scanner. Ideally these behavioral measures control for other important factors in performance across conditions, such as difficulty or language demands; at a minimum these factors should be assessed separately and included in regression analyses.

Measuring behavior independently prevents issues that arise when trying to interpret neural signatures of a cognitive task when children are not completing that task successfully. For example, reduced activity in young children, compared to older children, during a hard

mentalizing task could reflect less mature mentalizing brain regions, which caused children to struggle with the task. Alternatively, children who do not yet have the required concepts may not be completing the same task at all; acquiring the harder mental concepts may cause children to engage in more complex cognitive processes, reflected in increased recruitment of mentalizing brain regions.

By contrast, behavioral measures collected at the time of imaging are ideally orthogonal to the cognitive behavior of interest. For example, in the Saxe lab, the behavioral task during the fMRI scan might ask children to report if the story ending is a natural continuation of the story beginning ("Does this come next?"), while our experimental conditions of interest are manipulated within the beginning portion of the story. This provides a way to check that children paid attention during the scan, while not burdening them with a potentially difficult task. Other strategies for ensuring that children are attending to the stimulus during the scan include having an experimenter watch their eyes, collecting eyetracking data, and asking comprehension questions about the stimuli immediately after the scan.

Intervening on Development via Training Studies

A majority of developmental studies, and almost all developmental cognitive neuroscience studies, are correlational. Correlational relationships between neural and cognitive change can be difficult to interpret, and cannot provide information about causality. As discussed in Chapter 1, longitudinal studies that study developmental change within individual participants can reveal predictive relationships in development, and distinguish between neural measures that reflect stable individual differences (demonstrating continuity across time), and those that support behavioral improvements (showing developmental change across time). Measuring the behavioral and neural consequences of specific, short-term experience is another way to learn about origins of knowledge and mechanisms of developmental change.

Comparative research often uses controlled rearing studies to discover what kinds of knowledge require no relevant experience at all[13]. In the same vein, developmental psychologists employ training studies to measure effects of very particular experiences on knowledge and behavior[14-16]. By holding age relatively constant and providing specific experiences, controlled training studies can tease apart the relative influences of maturation and experience on developmental change. Training studies that include neuroimaging measures can additionally test (1) whether neural markers of change can similarly be influenced by experience, and (2) if an experience that is sufficient to change behavior is also sufficient to change the neural marker. This experimental design may additionally help distinguish between causal predictors and correlates of conceptual change; if conceptual change occurs in absence of a change in neural response patterns, then the neural marker is likely not necessary for conceptual change. By establishing conditions and limits of experience-driven neural change, training studies will be important for learning about neural plasticity and critical periods of development.

Constraining Hypotheses for and Interpretations of Neuroimaging Data

The deepest challenge for developmental cognitive neuroscience is not methodological, but conceptual. How should we interpret observed neural differences and similarities across development[17]? How can we tease apart whether neural differences reflect change in the cognitive task completed, the process being used to complete the task, or the metabolic cost incurred[17,18]? Below, we argue that bolstering developmental cognitive neuroscience research

with other methods may help to constrain hypotheses about neural and cognitive change; constrained hypotheses in turn produce results that are more straightforward to interpret.

Studying the neural basis of mature behavior in adults is an important precursor to developmental neuroimaging studies. All of the developmental cognitive neuroscience studies in this thesis were preceded by similar studies conducted with adults, or included adults as a comparison group. Similarly, utilizing knowledge of developmental trends observed via behavioral studies, and collecting rigorous, independent behavioral data in addition to neural measures within child neuroimaging projects will help to formulate and test constrained hypotheses. Prior to developmental neuroimaging work, the neural basis of theory of mind had been studied intensively in adults, and behavioral development of theory of mind abilities had been intensively studied in children and infants. This work enabled the formation of testable hypotheses about where in the brain to expect theory of mind processing to occur, suggested plausible methods for isolating and studying mental state reasoning processes, and provided time windows during which to expect relevant development to occur. For example, by including adult data in the cross-sectional experiment described in Chapter 2, we were able to test for developmental change in the response magnitude of ToM brain regions at specific moments of the movie that evoked high responses in adults. Without the adult sample, we would have either had to try to identify these moments in the child sample (iteratively, to avoid non-independence issues), or we would have had to test for all possible timepoints, creating a multiple comparisons problem[6,7]. Of course, one downside of constraining hypotheses based on adult data is the possibility of failing to detect meaningful, but unpredicted, patterns of data. Using exploratory analyses to detect these unpredicted patterns, and designing independent studies specifically to test these new hypotheses, mitigates this potential cost and will help prevent ad-hoc interpretations of unpredicted results.

In addition to testing hypotheses based on or inspired by relevant research in adult neuroimaging or developmental psychology, developmental cognitive neuroscience studies can use multivariate methods to directly test competing cognitive hypotheses. In contrast to univariate analyses, which provide information about which brain regions are involved in a cognitive process, multivariate analyses provide information about the internal structure of the representations within a given brain region. Multivariate analyses of the responses in theory of mind brain regions have (1) provided evidence that particular abstract features guide response similarity in ToM brain regions regardless of personal experience with those features[19], (2) suggested divisions of labor between ToM brain regions[20], and (3) demonstrated that ToM brain regions contain abstract representations of emotions[21]. By constructing dissimilarity matrices of the neural responses to various emotions, and comparing these matrices to dissimilarity matrices representing competing models about how emotion information is encoded and organized, Skerry & Saxe (2015) provided evidence that ToM brain regions represent emotions as abstract attributions, rather than as combinations of simple features like valence and arousal[21]. Developmental cognitive neuroscience studies could similarly construct models corresponding to distinct hypotheses about the internal structure of representations, and test these models against one another in order to evaluate how well they each explain human behavior. Testing for developmental change in the fit of particular models to neural responses could reveal corresponding developmental change in these representations. This approach requires explicitly stating alternative hypotheses about the plausible organizing features of information in a given

brain region, and formulating those hypotheses in a format that is directly comparable to the observed neural responses.

Over the past fifteen years, methods for collecting pediatric neuroimaging data, extracting more and finer-grained information from fMRI data, and for relating neuroimaging data to cognitive theories, have improved rapidly. I believe that the challenges described above will continue to push the field forward in exciting ways, and that we will see continued progress as a result.

References

1. de Bie, H. M. A. *et al.* Preparing children with a mock scanner training protocol results in high quality structural and functional MRI scans. *Eur J Pediatr* **169,** 1079–1085 (2010).
2. Slifer, K. J., Koontz, K. L. & Cataldo, M. F. Operant-contingency-based preparation of children for functional magnetic resonance imaging. *J Appl Behav Anal* **35,** 191–194 (2002).
3. Keil, B. *et al.* Size-optimized 32-channel brain arrays for 3 T pediatric imaging. *Magn. Reson. Med.* **66,** 1777–1787 (2011).
4. Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L. C. & Castellanos, F. X. Inscapes: A movie paradigm to improve compliance in functional magnetic resonance imaging. *NeuroImage* **122,** 222–232 (2015).
5. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* **22,** 1359–1366 (2011).
6. Vul, E., Harris, C., Winkielman, P. & Pashler, H. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science* **4,** 274–290 (2009).
7. Hsu, J. *Multiple comparisons: theory and methods.* (CRC Press, 1996).
8. Scott, K., Chu, J. & Schulz, L. Lookit (Part 2): Assessing the Viability of Online Developmental Research, Results From Three Case Studies. *Open Mind* **1,** 15–29 (2017).
9. Scott, K. & Schulz, L. Lookit (part 1): A new online platform for developmental research. *Open Mind* **1,** 4–14 (2017).
10. Alexander, L. M. *et al.* The healthy brain network biobank: an open resource for transdiagnostic research in pediatric mental health and learning disorders. bioRxiv. *bioRxiv* (2017).
11. Satterthwaite, T. D. *et al.* The Philadelphia Neurodevelopmental Cohort: a publicly available resource for the study of normal and abnormal brain development in youth. *NeuroImage* **124,** 1115–1119 (2016).
12. Marcus, D. S. *et al.* Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience* **19,** 1498–1507 (2007).
13. Chiandetti, C. & Vallortigara, G. Intuitive physical reasoning about occluded objects by inexperienced chicks. *Proceedings of the Royal Society of London B: Biological Sciences* **278,** 2621–2627 (2011).
14. Appleton, M. & Reddy, V. Teaching Three Year-Olds to Pass False Belief Tests: A Conversational Approach. *Social Development* (1996).
15. Slaughter, V. & Gopnik, A. Conceptual coherence in the child's theory of mind: training children to understand belief. *Child Dev* **67,** 2967–2988 (1996).
16. Sommerville, J. A., Woodward, A. L. & Needham, A. Action experience alters 3-month-old infants' perception of others' actions. *COGNITION* **96,** B1–B11 (2005).
17. Poldrack, R. A. Interpreting developmental changes in neuroimaging signals. *Hum. Brain Mapp.* **31,** 872–878 (2010).
18. Poldrack, R. A. Is 'efficiency' a useful concept in cognitive neuroscience? *Accident Analysis and Prevention* **11,** 12–17 (2015).
19. Koster-Hale, J., Bedny, M. & Saxe, R. Thinking about seeing: Perceptual sources of knowledge are encoded in the theory of mind brain regions of sighted and blind adults.

COGNITION **133,** 65–78 (2014).
20. Koster-Hale, J. *et al.* Mentalizing regions represent distributed, continuous, and abstract dimensions of others' beliefs. *NeuroImage* **161,** 9–18 (2017).
21. Skerry, A. E. & Saxe, R. Neural Representations of Emotion Are Organized around Abstract Event Features. *Current Biology* **25,** 1945–1954 (2015).

General Discussion and Conclusion

One of the most critical components of cognitive development in childhood is *social* cognitive development. Children's "Theory of Mind" (ToM)- their ability to infer and predict other people's beliefs, desires, and emotions- undergoes drastic changes during childhood, and is the basis for building and maintaining social relationships. A specific network of brain regions is recruited for ToM reasoning in both children and adults. What kinds of neural changes support the development of the remarkable cognitive capacity to consider the minds of others, and what novel insights about ToM development can we gain by using neuroimaging measures? Below, I briefly review the results of the experiments included in this thesis, provide a discussion of the development of functionally selective responses in ToM brain regions, and describe limitations and areas for future research.

Brief Review of Results

Chapter 1 described results from two longitudinal experiments on developmental change and stable individual differences in response selectivity of ToM brain regions, and behavioral performance on ToM tasks. Consistent with the hypothesis that mastering relatively later developing components of ToM involves domain-specific change in ToM, across two longitudinal studies, earlier behavioral ToM performance predicted later ToM performance within individual children ages 5-12 years old. We also found developmental change in selectivity between five and seven years of age. However, contrary to the two prior cross-sectional studies, we did not replicate the relationship between behavioral ToM and response selectivity in ToM brain regions[1,2]. I provided a discussion of the potential limitations of selectivity as a neural marker of developmental change in ToM, and directly addressed the discrepancy between the results of these experiments and the previous cross-sectional studies.

Chapter 2 described results from a large cross-sectional experiment that involved measuring functional responses in ToM brain regions in children as young as three years of age. By using a naturalistic movie paradigm, we obtained high quality fMRI data from very young children, potentially addressing one of the limitations of the longitudinal experiments in Chapter 1. We found early signatures of a functional dissociation between brain regions recruited for reasoning about others' mental states (the "theory of mind" network) and physical pain (the "pain matrix"), by age three years. Additionally, we observed developmental change such that the responses of these two networks become more functionally selective, and increasingly distinct, throughout childhood. Finally, by comparing functional responses in children who pass and fail false-belief tasks, this study provided evidence that this well-studied behavioral milestone is not accompanied by drastic changes in social brain regions. While the analyses conducted in this experiment were largely exploratory, Chapter 3 described confirmatory evidence for the observed patterns of developmental change in a large, publicly available pediatric dataset. Chapter 3 additionally provided new insight into the link between inter-region correlations and functionally selective responses: inter-region correlations measured during a functional task primarily reflect the stimulus-driven response, rather than the intrinsic correlations within and between ToM and pain brain regions that are also present at rest.

In Chapter 4, I described evidence from neural and behavioral measurements of ToM in children who experienced delayed access to language, which suggests a key role of early and extensive language experience for refining the functional response in RTPJ. This study additionally

suggests that language facilitates expression and development of ToM reasoning, and may be a protective factor against prolonged or permanent delays in ToM development.

Finally, in Chapter 5 I discussed challenges and limitations of developmental cognitive neuroscience studies of theory of mind. Despite the challenges and limitations, the experiments in this thesis have provided key insights into theory of mind development, neurally and behaviorally. In particular, the evidence presented in this thesis suggests that one key aspect of ToM development in childhood is the refinement of conceptual distinctions between mental state content and other relevant social information (e.g., bodily sensations, physical appearance, knowledge about enduring social relationships).

What kinds of neural changes support Theory of Mind Development?

Development of Increasingly Functionally Selective Responses

A key neural signature of theory of mind development is increasingly functionally selective responses in ToM brain regions. ToM brain regions become increasingly sensitive to categorical divisions between mentalistic and non-mentalistic social content as children get older, perhaps in order to better handle the specific computational demands of ToM processes. Development of functionally selective responses corresponds to improvements and delays in ToM development.

Developmental increases in response selectivity could be driven by increased responses to preferred stimuli, reduced responses to non-preferred stimuli, or both. Consistent with prior studies of ToM[1,2] as well as evidence from other cognitive domains[3,4], the evidence in this thesis suggests that increases in response selectivity in ToM brain regions correspond to reduced responses to non-preferred stimuli. In these experiments, non-preferred stimuli include descriptions of enduring social relationships and physical appearances of characters (Chapters 1 & 4), and physical (bodily) sensations (Chapters 2 – 3).

Prior work in other domains has suggested that microproliferation of synapses and/or dendritic arbors may drive these kinds of functional changes[5], by increasing the spatial extent over which preferred information is stored and processed, and/or over which non-preferred responses can be inhibited[5,6,7]. Future work is necessary to determine the relative role of microproliferation (and other structural changes, like axonal myelination, synaptic pruning, and improved potentiation) in the development of increasingly selective responses in ToM brain regions. However, increases in selectivity in ToM brain regions may involve strengthening of local inhibitory responses that support cross-category discriminations: ToM brain regions are activated in response to scenes that highlight mental states, and correspondingly *deactivated* in response to scenes that highlight physical (bodily) sensations (Chapters 2 & 3). As activity to ToM scenes increases throughout childhood, activity to Pain scenes decreases. This developmental change could plausibly occur via strengthening of local inhibitory responses within ToM regions to Pain scenes[8]. Activation in ToM brain regions to mental state scenes is also coupled with corresponding deactivation of adjacent regions in the Pain Matrix (and vice versa). Similar symmetric cross-category inhibitory relationships exist between adjacent cortical regions in extrastriate cortex, and have been hypothesized to act as a higher-level mechanism for "sharpening" functionally selective responses[6,7]. Responses in ToM and pain networks become increasingly anti-correlated across development, and children with more "adult-like" functional response timecourses also have more anti-correlated responses across these two networks. Thus, increases in functional

selectivity may also be related to the strengthening of a mutual inhibitory relationship between ToM brain regions and those in the Pain Matrix. These features of the neural response additionally highlight the categorical division between minds and bodies as particularly relevant for ToM processes.

Role of Developmental Experience on Functional Selectivity

To what extent are developmental increases in functional selectivity driven by maturational vs. experiential factors? By one extreme, evolutionary pressures and genetic makeup could drive ToM brain regions to develop increasingly functionally selective responses throughout childhood, regardless of developmental experience. On the other extreme, the development of functionally selective responses could be quite fragile or flexible: lack of a necessary input or experience during a critical or sensitive time could preclude the development of brain regions selective for ToM processes. Chapter 4 provides evidence that delayed access to language does not preclude the eventual development of functionally selective ToM brain regions, but does result in delayed increases in selectivity in childhood. While neural responses among adults were indistinguishable regardless of age of linguistic exposure, the relative difference between Mental and Social content was reduced in children who experienced delayed access to language. Interestingly, this delay was most apparent in the responses of the RTPJ, which is typically quite selective by approximately age seven years (Chapter 1). Thus, developmental experience, and in particular, early and extensive exposure to language, is important for refining the functional response of RTPJ in childhood. Whether there is a sensitive or critical period during which linguistic input must be provided for the development of highly functionally selective responses in RTPJ remains unknown.

There is significant need for more research on the developmental factors that drive functional responses to become increasingly selective. One open question concerns the *extent* to which developmental experiences are necessary for the maintenance and development of selective responses. Interestingly, children with delayed access to language did not show reduced response selectivity during the non-verbal movie task. ToM brain regions respond preferentially to minds, and deactivate in response to bodies, during this experimental context (Chapters 2 – 3), regardless of age of exposure to language (Chapter 4). The functional dissociation between brain regions that respond to minds and bodies is apparent by age three years (Chapter 2). Together, these data suggest that ToM brain regions are sensitive to at least an approximation of the categorical boundary between minds and bodies early in development, and the refinement of this boundary during childhood is less dependent on developmental experience (early/prolonged exposure to language).

A second question is whether early linguistic input is *sufficient* for typical development of brain regions specialized for ToM reasoning. Studies of children who are congenitally blind could help to address this question. Blind children have typical linguistic input, but reduced access to information about minds that is conveyed through vision. Vision provides a way to perceive consequences of mental states (e.g. if a person reaches for a teddy bear, she prefers it to the ball; if a person expresses sadness upon seeing a puppy, she's remembering when the puppy stole her snack), and facilitates early interactions and social bonding (e.g. through eye contact, joint attention, and attention to facial expressions). While previous neuroimaging research with adults suggests that by adulthood, blindness has no effect on the functional responses in theory of mind

brain regions[9,10], behavioral studies find some evidence for delayed ToM development in children who are blind[11-14]. Future work investigating the development of functionally selective responses in children who are congenitally blind could clarify whether linguistic input is particularly important for refining functional responses, and if visual input during development plays a similar role in refining the functional responses in RTPJ.

Relationship between Functionally Selective Responses and ToM Development

Are increases in functionally selective responses related to ToM development in childhood? The evidence presented in this thesis suggests that increases in functionally selective responses are related to the appropriate application of ToM processing in absence of explicit cues to do so. Across two experiments, selective responses to events that involve spontaneously considering the relevance of the current event for (past) beliefs or emotions that are not explicitly marked were related to behavioral measures of ToM (Chapters 2 & 3). Children who have larger selective responses for storing and processing preferred information, and/or for inhibiting non-preferred responses, may have more refined ToM concepts. This may enable these children to use ToM concepts flexibly, and to more easily recognize the relevance of particular concepts across different contexts. Future experiments targeting this aspect of ToM reasoning may be particularly important for linking neural and behavioral ToM measures.

Flexible use of ToM concepts could also be supported by faster and more efficient communication between ToM brain regions with distinct computational roles. Inter-region correlations between ToM brain regions during a functional task are related to the development of the stimulus-driven functional response in ToM brain regions. Children with more "adult like" functional responses in ToM brain regions also had higher inter-region correlations within the ToM network during movie viewing (Chapters 2 & 3). Inter-region correlations within the ToM network are also related to ToM behavior: young children who "passed" explicit false-belief tasks had higher within-ToM network inter-region correlations than children who failed, even when controlling for age (Chapter 2). Within-ToM IRCs were correlated with overall ToM behavioral score in 3 – 12 year old children, but this relationship did not remain significant when controlling for age (Chapter 2). Developing stronger inter-region correlations within the ToM network could be particularly relevant to or reflective of ToM development in early childhood. This hypothesis is consistent with prior evidence linking white matter tract development between ToM brain regions to false-belief task performance[15]. As discussed in the introduction, future work simultaneously measuring white matter development and the development of functional responses will be important for clarifying causal order of development, and relative contributions of each to ToM behavioral improvement.

The relationship between neural changes in ToM brain regions and behavioral measures of ToM reasoning provides evidence for domain-specific developmental change in theory of mind. Two experiments in this thesis (Chapters 2 & 4) used neuroimaging measures to inform debates about the extent to which ToM development is domain-specific. In Chapter 2, the debate centers around the nature of the transition from failing to passing false-belief tasks: does passing false-belief tasks involve domain-specific change in ToM, or development of executive functions alone? In Chapter 4, the debate focuses on the role of language in ToM development: does language facilitate ToM development per se, or simply enable expression of ToM understanding? In both experiments, the neuroimaging measures provided evidence for domain-

specific change in ToM in childhood. Developmental improvements in ToM reasoning are accompanied by neural changes in brain regions recruited selectively to reason about other minds, and linguistic experience in childhood impacts the development of functionally selective responses in these same brain regions. The results of these experiments not only offer converging evidence for domain-specific ToM development from a different level of analysis, but also suggest a good fit between the question about domain-specificity originally posed by cognitive and developmental psychologists, and the methods and measures used in cognitive neuroscience.

Potential Limitations & Challenges for Measuring Developmental Change in Selectivity

The evidence presented in these chapters suggests potential limitations to using response selectivity as a measure of ToM development. Primarily, it is unclear if response selectivity is related to *stable individual differences* in ToM, even among young children in whom response selectivity is clearly increasing with age (ages 5 – 7 years; Chapter 1 Exp. 2). In both longitudinal studies, we found a significant relationship between a child's *behavioral* ToM score at visit one and their score at visit two: children who performed better on the ToM task at visit one, relative to other participants, also performed better on the task at visit two, controlling for other variables that predicted ToM performance. There was less clear evidence for stable *neural* individual differences in these studies. In Experiment 2, response selectivity in RTPJ at visit one predicted response selectivity at visit two, but this result was not replicated in Experiment 1. Moreover, across both studies, we did not find a significant correlation between response selectivity and ToM behavior, which did vary stably across individuals. Thus, while this thesis provides evidence for neural markers that relate to ToM behavior, cross-sectionally (Chapters 2 & 3), this thesis also suggests that measuring the relationship between response selectivity and ToM behavior is challenging. Robust neural markers that measure stable neural individual differences and predict ToM behavior are critical for designing and testing effectiveness of clinical interventions and/or training paradigms that aim to improve social cognitive abilities.

Null results are difficult to interpret: they could reflect the true absence of a relationship, or they could reflect the failure of an experiment to measure the presence of a relationship. The evidence reviewed in this thesis suggests at least two hypotheses about why relating response selectivity to improvements in ToM is challenging. First, capturing this relationship (cross-sectionally and longitudinally) may require studying a large age range of children that includes a substantial number of young children (<7 years). The observed relationships between the functional response and ToM behavior were from cross-sectional samples that included participants aged 3 – 12 years and 5 – 12 years old (in addition to adults, and 13 – 20 year olds), respectively (Chapters 2 & 3). Chapter 3 provided direct evidence that developmental change with age was more apparent in samples with large age ranges (e.g., 3 – 12 and 5 – 20 years, compared to 5 – 12 years). Thus, while longitudinal studies are more sensitive to developmental change within individuals, a large age range may still be important for capturing meaningful change in response selectivity, which undergoes slow, gradual change throughout childhood. The longitudinal studies described in Chapter 1 may not have included enough young children (Study 1, ages 5 – 12 years, including six children initially younger than age 7 years), or may not have included a wide enough age range (Study 2, ages 5 – 7 years) to measure the relationship between response selectivity and ToM behavior.

A second possible explanation for why it is particularly tricky to relate response selectivity to behavioral measures of ToM over time is that doing so requires making a developmentally relevant comparison between the neural response to preferred stimuli (mental states) and non-preferred stimuli (e.g., bodies, non-mentalistic social information) at each timepoint. For example, if the RTPJ is maximally selective for mental states *compared to social information* at timepoint one, then measuring relative responses to those same conditions at visit two will at best provide a better approximation of the magnitude of the selective response (e.g., regression to the mean[16]). This comparison won't be very sensitive to developmental change, because the neural response to non-preferred stimuli was low to begin with. Behavioral tasks that successfully measure developmental change in ToM measure responses to different types of ToM questions[2,17,18]. Developmental improvements in ToM include becoming sensitive to the difference between emotions that are hidden vs. expressed[19], harm that is accidental vs. intentional[20], beliefs that are justified vs. suspect[21-23], and speech that is literal vs. sarcastic[24,25]. Future pediatric imaging studies may similarly need to measure developmental change and individual differences in within-category discriminations, instead of (or in addition to) across-category boundaries, in neural responses of ToM brain regions. These neural changes may be best measured via multivariate approaches. In adults, multivariate pattern analyses have provided evidence for distinct patterns of response in TPJ depending on the source modality of a person's belief[10,26], and for the justification of evidence for a given belief[26]. Similar methods could plausibly be used to capture developmental changes in ToM that involve increased sensitivity to distinctions among mental states. These methods are additionally generally sensitive to across-category boundaries[27,28], and have provided evidence for distinct functional roles of different ToM brain regions[26]. Future work measuring the pattern of responses to different features of mental state stimuli will be useful for relating neural responses to specific aspects of conceptual change in ToM, and will also provide insight into the developmental origins of the distinct functional roles of brain regions within the ToM network.

A final limitation of response selectivity, which will also constrain the benefits of (current) multivariate approaches, is that these measures do not currently capture the causal structure that is inherent in theory of mind reasoning. These methods quantify the sensitivity of neural responses to the difference between two conceptual categories, or along a particular dimension within a category, which is informative for determining which features of mental states are peripheral vs. relevant to category membership, and which features constrain similarity. But they don't yet describe the explanatory structure that even children use to decide which features are relevant in the first place[29-31]. Conceptual knowledge and relevant experiences both need a framework, theory, or structure to be incorporated into, in order to be useful for theory of mind development[32,33]. Future work is necessary to describe the way that the developing brain supports and utilizes such a structure[34].

Future Directions

Throughout this thesis, I have described several areas for future research, including longitudinal studies, training studies, studies that measure physical properties of the brain in addition to functional responses, and studies that utilize multivariate approaches to measure the development of fine-grained conceptual distinctions in ToM. In this final section, I discuss two additional areas for future research: studying the neural basis of ToM in social deprivation, and in young infants.

Early social deprivation has vast consequences for cognitive, social, and emotional development[35]. A study of institutionalized children in Romania has suggested that age two years is a particularly important marker for subsequent recovery: children who were placed in foster care by age two years recovered significantly, whereas children who were placed in foster care after age two showed extensive and ongoing developmental delays[36]. Interventions, and studies of the effectiveness of interventions, are the primary concern for these children. However, these children also offer a way to understand the development of functionally selective ToM brain regions: is there a critical and/or sensitive period for the maintenance and further development of functionally selective response in ToM brain regions? Relatedly, in absence of preferred stimuli, what kinds of cognitive processes do ToM brain regions perform? Prior work has suggested that cortex is incredibly "pluripotent." That is, despite remarkable consistency in the location and functions of cortical regions across individuals, developmental input drives functional specialization of cortex[37]. Auditory cortex can process visual inputs[38] and, in congenitally blind individuals, visual cortex can take on language processing[39,40]. Are ToM brain regions similarly pluripotent, like visual cortex, or does a lack of preferred input simply result in cortical atrophy[41]?

A final area for future research concerns characterizing the neural responses of ToM brain regions in infancy. While there has been recent success measuring functional responses in awake infants using fMRI in other cognitive domains, this method remains challenging for theory of mind research. In order to measure functional responses to visual or auditory categories, previous studies have utilized brief movie or sound clips, neither of which require particularly prolonged visual attention, or linguistic understanding[42,43]. By contrast, stimuli used to evoke responses in ToM brain regions typically involve telling a story or showing a movie about a character in some context, in order to evoke mental state inferences. The evidence reviewed in this thesis provides relatively "early" markers of functional specialization in three year olds, who have yet to pass false-belief tasks. But a typically developing three-year-old has had a myriad of social experiences to learn from. If the goal is to use neuroimaging methods to clarify the "starting state" of brain regions that are eventually functionally specialized to reason about the minds of others, to describe the conceptual repertoire of infants, or to identify early clinical markers of social disorders, functional responses need to be measured in infancy.

Methods that are more tolerant to participant motion, like EEG and fNIRS, have provided key insights into ToM development in young children and infants[44-46], and are promising for future research in infants. The "functional channel of interest" (fCOI) approach is particularly promising for dealing with uncertainty regarding the source location of measured activity, and for relating neural responses in infants to the mature profile of functional responses in adults[47]. As with the research described in this thesis, the main challenge to this research program will concern the interpretation of results. For example, finding activation to specific content in similar cortical locations in infants and adults does not constitute evidence for similar *cognitive processes* within those cortical locations in infants and adults. Of course, this kind of result is still incredibly useful: these studies place a previously unknown constraint on the timeline of the development of functionally selective responses, which in turn constrains hypotheses about the role of experience, evolutionary pressures, and biological maturation on brain development. And,

critically, having pre-specified regions of interest enables the articulation of precise, clear, and specific hypotheses in subsequent studies of cognitive development.

The Promise of Developmental Cognitive Neuroscience

A primary benefit of conducting cognitive neuroscience studies with children is the opportunity to learn about the developing brain. In this thesis, I have argued that conducting cognitive neuroscience studies with children can also inform our theories about the developing mind. One of the main promises of cognitive neuroscience research is that it offers a way to "look under the hood"- providing a window into the previously unobservable cognitive states and mechanisms that give rise to behavior. In the domain of theory of mind, this promise has been put to practice: neuroimaging studies have begun to provide novel evidence to help inform developmental hypotheses about how and when we come to understand the minds of others.

I have provided initial evidence that neuroimaging studies can be used to test previously untested hypotheses about neural correlates of conceptual development, and have suggested methods for tackling questions of conceptual continuity and change in theory of mind. Moving forward, studies that integrate multiple kinds of measures (e.g. measures of connectivity, functional responses, and behavior), longitudinal studies that measure within-subject change, and training studies that measure consequences of specific instances of conceptual change will help to clarify the nature of these developmental accomplishments.

A critical goal of this kind of research is to inform the theories that motivate it. Children are undergoing the very processes that cognitive scientists and neuroscientists seek to understand: they are acquiring new facts, memories and skills, evaluating and revising intuitive theories about people, objects, and space, and building and fostering social relationships with others. All of these processes involve change in their brains, of course; but by linking specific cognitive achievements to particular aspects of neural change, scientists will get a whole new window on the development of cognition.

References

1. Saxe, R. R., Whitfield-Gabrieli, S., Scholz, J. & Pelphrey, K. A. Brain regions for perceiving and reasoning about other people in school-aged children. *Child Dev* **80,** 1197–1209 (2009).
2. Gweon, H., Dodell-Feder, D., Bedny, M. & Saxe, R. Theory of Mind Performance in Children Correlates With Functional Specialization of a Brain Region for Thinking About Thoughts. *Child Dev* **83,** 1853–1868 (2012).
3. Cantlon, J. F., Pinel, P., Dehaene, S. & Pelphrey, K. A. Cortical Representations of Symbols, Objects, and Faces Are Pruned Back during Early Childhood. *Cerebral Cortex* **21,** 191–199 (2010).
4. Dehaene, S. *et al.* How learning to read changes the cortical networks for vision and language. *Science* **330,** 1359–1364 (2010).
5. Gomez, J. *et al.* Microstructural proliferation in human cortex is coupled with the development of face processing. *Science* **355,** 68–71 (2017).
6. Allison, T., Puce, A. & McCarthy, G. Category-sensitive excitatory and inhibitory processes in human extrastriate cortex. *Journal of Neurophysiology* **88,** 2864–2868 (2002).
7. Pelphrey, K. A., Mack, P. B., Song, A., Güzeldere, G. & McCarthy, G. Faces evoke spatially differentiated patterns of BOLD activation and deactivation. *NeuroReport* **14,** 955–959 (2003).
8. Tsao, D. Y. & Livingstone, M. S. Mechanisms of face perception. *Annu. Rev. Neurosci.* **31,** 411–437 (2008).
9. Bedny, M., Pascual-Leone, A. & Saxe, R. R. Growing up blind does not change the neural bases of Theory of Mind. *Proceedings of the National Academy of Sciences* **106,** 11312–11317 (2009).
10. Koster-Hale, J., Bedny, M. & Saxe, R. Thinking about seeing: Perceptual sources of knowledge are encoded in the theory of mind brain regions of sighted and blind adults. *COGNITION* **133,** 65–78 (2014).
11. Brambring, M. & Asbrock, D. Validity of False Belief Tasks in Blind Children. *J Autism Dev Disord* **40,** 1471–1484 (2010).
12. Brown, R., Hobson, R. P., Lee, A. & Stevenson, J. Are There 'Autistic-like' Features in Congenitally Blind Children? *Journal of Child Psychology and Psychiatry* **38,** 693–703 (1997).
13. Peterson, C. C., Peterson, J. L. & Webb, J. Factors influencing the development of a theory of mind in blind children. *British Journal of Developmental Psychology* **18,** 431–447 (2000).
14. Minter, M., Hobson, R. P. & Bishop, M. Congenital visual impairment and 'theory of mind'. *The British Journal of Developmental Psychology* **16,** 183 (1998).
15. Wiesmann, C. G., Schreiber, J., Singer, T., Steinbeis, N. & Friederici, A. D. White matter maturation is associated with the emergence of Theory of Mind in early childhood. *Nature Communications* **8,** 14692 (2017).
16. Galton, F. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland* **15,** 246–263 (1886).
17. Wellman, H. M. & Liu, D. Scaling of theory-of-mind tasks. *Child Dev* **75,** 523–541 (2004).
18. Richardson, H., Lisandrelli, G., Riobueno-Naylor, A. & Saxe, R. Development of the

social brain from age three to twelve years. *Nature Communications* **9,** 1027 (2018).
19. Harris, P. L. *Children and emotion: The development of psychological understanding.* (Basil Blackwell, 1989).
20. Cushman, F., Sheketoff, R., Wharton, S. & Carey, S. The development of intent-based moral judgment. *COGNITION* **127,** 6–21 (2013).
21. Astington, J. W., Pelletier, J. & Homer, B. Theory of mind and epistemological development: The relation between children's second-order false-belief understanding and their ability to reason about evidence. *New ideas in Psychology* **20,** 131–144 (2002).
22. Pillow, B. H. & Henrichon, A. J. There's more to the picture than meets the eye: Young children's difficulty understanding biased interpretation. *Child Dev* **67,** 803–819 (1996).
23. Pillow, B. H., Hill, V., Boyce, A. & Stein, C. Understanding inference as a source of knowledge: Children's ability to evaluate the certainty of deduction, perception, and guessing. *Developmental Psychology* **36,** 169–179 (2000).
24. Happé, F. G. An advanced test of theory of mind: understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *J Autism Dev Disord* **24,** 129–154 (1994).
25. Peterson, C. C., Wellman, H. M. & Slaughter, V. The mind behind the message: Advancing theory-of-mind scales for typically developing children, and those with deafness, autism, or Asperger syndrome. *Child Dev* **83,** 469–485 (2012).
26. Koster-Hale, J. *et al.* Mentalizing regions represent distributed, continuous, and abstract dimensions of others' beliefs. *NeuroImage* **161,** 9–18 (2017).
27. Haxby, J. V. *et al.* Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293,** 2425–2430 (2001).
28. Spiridon, M. & Kanwisher, N. How distributed is visual category information in human occipito-temporal cortex? An fMRI study. *Neuron* **35,** 1157–1165 (2002).
29. Ahn, W.-K., Kim, N. S., Lassaline, M. E. & Dennis, M. J. Causal status as a determinant of feature centrality. *Cognitive Psychology* **41,** 361–416 (2000).
30. Ahn, W.-K., Gelman, S. A., Amsterlaw, J. A., Hohenstein, J. & Kalish, C. W. Causal status effect in children's categorization. *COGNITION* **76,** B35–B43 (2000).
31. Ahn, W.-K. & Kim, N. S. in *Psychology of learning and motivation* **40,** 23–65 (Elsevier, 2000).
32. Gopnik, A. & Wellman, H. M. 10 The theory theory. *Mapping the mind: Domain specificity in cognition and culture* 257 (1994).
33. Gopnik, A. & Wellman, H. M. Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin* **138,** 1085–1108 (2012).
34. Saxe, R. Seeing Other Minds in 3D. *Trends in Cognitive Sciences* **22,** 193–195 (2018).
35. Nelson, C. A., Fox, N. A. & Zeanah, C. H. *Romania's Abandoned Children.* (Harvard University Press, 2014). doi:10.4159/harvard.9780674726079
36. Nelson, C. A. *et al.* Cognitive recovery in socially deprived young children: The Bucharest Early Intervention Project. *Science* **318,** 1937–1940 (2007).
37. Bedny, M. Evidence from Blindness for a Cognitively Pluripotent Cortex. *Trends in Cognitive Sciences* **21,** 637–648 (2017).
38. Melchner, Von, L., Pallas, S. L. & Sur, M. Visual behaviour mediated by retinal projections directed to the auditory pathway. *Nature* **404,** 871 (2000).
39. Bedny, M., Pascual-Leone, A., Dodell-Feder, D., Fedorenko, E. & Saxe, R. Language

processing in the occipital cortex of congenitally blind adults. *Proceedings of the National Academy of Sciences* **108,** 4429–4434 (2011).
40. Bedny, M., Richardson, H. & Saxe, R. 'Visual' Cortex Responds to Spoken Language in Blind Children. *Journal of Neuroscience* **35,** 11674–11681 (2015).
41. Sheridan, M. A., Fox, N. A., Zeanah, C. H., McLaughlin, K. A. & Nelson, C. A. Variation in neural development as a result of exposure to institutionalization early in childhood. *Proceedings of the National Academy of Sciences* **109,** 12927–12932 (2012).
42. Deen, B. *et al.* Organization of high-level visual cortex in human infants. *Nature Communications* **8,** 13995 (2017).
43. Shultz, S., Vouloumanos, A., Bennett, R. H. & Pelphrey, K. Neural specialization for speech in the first months of life. *Dev Sci* n/a–n/a (2014). doi:10.1111/desc.12151
44. Bowman, L. C., Kovelman, I., Hu, X. & Wellman, H. M. Children's belief-and desire-reasoning in the temporoparietal junction: evidence for specialization from functional near-infrared spectroscopy. *Front. Hum. Neurosci.* **9,** 560 (2015).
45. Hyde, D. C., Simon, C. E., Ting, F. & Nikolaeva, J. Functional organization of the temporal-parietal junction for theory of mind in preverbal infants: A near-infrared spectroscopy study. *Journal of Neuroscience* 0264–17 (2018).
46. Sabbagh, M. A., Bowman, L. C., Evraire, L. E. & Ito, J. M. B. Neurodevelopmental correlates of theory of mind in preschool children. *Child Dev* **80,** 1147–1162 (2009).
47. Powell, L. J., Deen, B. & Saxe, R. Using individual functional channels of interest to study cortical development with fNIRS. *Dev Sci* (2017).

Publicly Available Resources

Pre-Registered Analysis Plans & FMRI Stimuli
Chapter 1: https://osf.io/jh68b/
Chapter 4: https://osf.io/mhgp8/
Note: Chapter 2 describes exploratory analyses; Chapter 3 provides confirmatory evidence using identical analysis procedures.

Behavioral Tasks
Theory of Mind Behavioral Batteries (Chapters 1, 2): https://osf.io/g5zpv/
ASL and Non-Verbal Theory of Mind Behavioral Batteries (Chapter 4): https://osf.io/mhgp8/

FMRI Data
Chapter 2: https://openfmri.org/dataset/ds000228/
Chapter 3: http://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network/index.html
(Chapter 3 data was made publicly available by the Child Mind Institute)