Implicit Theory of Mind in Neurotypical and Neurodivergent Social Cognitive Development

Inaugural-Dissertation

zur Erlangung des Doktorgrades der Philosophie
der Ludwig-Maximilians-Universität München

vorgelegt von

Lucie Zimmer

aus

Tegernsee

Juni 2025

Erstgutachter:

PD Dr. Tobias Schuwerk, Ludwig-Maximilians-Universität München

Zweitgutachter:

Prof. Dr. Hannes Rakoczy, Georg-August-Universität Göttingen

Drittes Mitglied der Prüfungskommission:

Prof.^{in} Dr.^{in} Corinna Reck, Ludwig-Maximilians-Universität München

Tag der mündlichen Prüfung: 27.10.2025

Danksagung / Acknowledgements

Ich möchte mich herzlich bei allen bedanken, die mich während meiner Promotion begleitet und unterstützt haben!

Ein besonderer Dank geht an PD Dr. Tobias Schuwerk für die Betreuung meiner Dissertation. Danke, Tobi, für deine gelassene, wohlwollende Art, deine Unterstützung und Offenheit während meiner Promotion und für all die Erfahrung, die ich in den vielen spannenden Projekten sammeln durfte.

Prof. Dr. Hannes Rakoczy danke ich für die Übernahme der Zweitbetreuung. Danke, Hannes, für deine Unterstützung, die tolle Zusammenarbeit und dein Vertrauen, dass ich Teil des ManyBabies 2 Projekt sein durfte.

Außerdem danke ich Prof.^{in} Dr.^{in} Corinna Reck. Danke, Corinna, für deine herzliche Aufnahme im LFE-Team und deine wohlwollende Unterstützung.

Ebenso danke ich Prof.^{in} Dr.^{in} Beate Sodian für die Möglichkeit in das Crossing-Projekt miteinsteigen zu dürfen und für die gute und lehrreiche Zusammenarbeit.

Danke an alle Ko-Autor:innen. A big thank you to Dr.^{in} Hilary Richardson and Prof.^{in} Dr.^{in} Nivedita Mani for your great support with the data analysis, your valuable feedback and everything I learned from you. Danke Adrian Steffan für die tolle Zusammenarbeit und deine zuverlässige technische Unterstützung. Danke Prof. Dr. Markus Paulus und Dr.^{in} Carolina Pletti für euer wertvolles Feedback. Thank you to all co-authors around the world contributing to ManyWebcams and ManyBabies 2.

Danke an alle studentischen Hilfskräfte, die die Projekte unterstützt haben. Danke Sarina Drexler, Paula Kirchner und Rebekka Mattes für eure verlässliche Unterstützung und einfühlsame Durchführung der Studien.

Danke an meine Freundinnen und Freunde für euer offenes Ohr und eure stärkende Unterstützung, besonders dafür nicht (ganz) den Blick für die Realität zu verlieren. Danke, Kathi, dass du schon so lange an meiner Seite bist—als beste Freundin und Studienteilnehmerin.

Danke, Mama und Papa, für euren unerschütterlichen Glauben an mich und, dass ihr mir von klein auf das Gefühl gegeben habt, dass alles möglich ist. Danke, Benni, dass du mich mit in deine Welt nimmst und ich so viel von dir lernen darf.

Danke, Moritz, für deine liebevolle und fürsorgliche Unterstützung und deinen sicheren Rückhalt. Mit dir fühlt sich alles so viel leichter und schöner an.

Und zuletzt danke ich allen kleinen und großen Studienteilnehmer:innen, ohne diese die Forschung gar nicht möglich gewesen wäre.

Table of Contents

Abstract

In social interactions, people reason about their own mental states—such as beliefs, desires, or intentions—and ascribe mental states to others in order to predict or explain their behavior. This ability, referred to as Theory of Mind, is central to successful social interaction and develops in childhood. Explicit Theory of Mind, a conscious and verbally expressible understanding of mental states, typically emerges around age 4. In contrast, the earlier development of implicit Theory of Mind, an unconscious, non-verbal sensitivity to others' mental states, remains debated due to inconsistent findings across various experimental paradigms. Theory of Mind abilities are closely associated with both the use and understanding of mental state language. In mental state language understanding research, empirical evidence demonstrates that children verbally distinguish between epistemic verbs—such as know, think or, believe—from around the ages of 4 to 5. However, it remains unclear whether implicit mental state language understanding is already present earlier due to limited research. In neurodivergent social cognitive development, such as in autism, recent predictive coding accounts link difficulties in Theory of Mind reasoning to attenuations in predictive processing. Since implicit Theory of Mind relies on unconscious predictions about others' mental states, differences between autistic and neurotypical individuals may reflect alterations in the underlying predictive mechanisms, thereby offering a potential explanation for challenges in social interaction. However, neural findings remain elusive. Based on these considerations, the following key questions arise: How robust are existing paradigms for measuring implicit Theory of Mind across the lifespan? Does implicit mental state language understanding developmentally precede an explicit one? Does predictive processing in Theory of Mind brain regions differ between autistic and neurotypical individuals?

The aim of this dissertation was to systematically investigate these questions. To this end, four empirical studies were conducted, each addressing different aspects of implicit Theory of Mind within neurotypical and/or neurodivergent social cognitive development. Study 1 validated a novel web-based eye-tracking method with toddlers aged 18 to 27 months (*N* = 125). Within an anticipatory looking paradigm, goal-based action anticipation was measured and compared with findings from a laboratory study. The results showed that with our web-based setting goal-based action anticipation can be successfully measured, although the proportion of anticipatory looking was slightly lower, and exclusion rate was higher. Study 2 also used the anticipatory looking paradigm and tested whether toddlers (*N* = 521) and adults (*N* = 703) differ in their action anticipation based on epistemic states such as knowledge and ignorance. Adults clearly distinguished between knowledge and ignorance, as indicated by their anticipatory looking behavior. Unexpectedly, toddlers did not show this differentiation, highlighting the need for further research. Study 3 investigated the implicit understanding of the epistemic verbs "know" and "think" in toddlers aged 27 (*N* = 199) and 36 months (*N* = 131). The results revealed that toddlers as young as 27 months were able to differentiate between these verbs, with a spontaneous preference for speaker certainty (i.e., "know"). This indicates that implicit mental state language understanding seems to precede an explicit one. Study 4 investigated predictive processing in the Theory of Mind network of non-autistic (*N* = 61 in Experiment 1; *N* = 30 in Experiment 2) and autistic (*N* = 30 in Experiment 2) adults using functional magnetic resonance imaging. Contrary to our expectations, predictive processing was absent in the Theory of Mind network in either group. However, in autistic (in comparison to non-autistic) adults a reduced repetition suppression was observed in a specific scene involving complex Theory of Mind processes. This provides preliminary neural evidence for the predictive coding theory in autism. Together, the four studies contribute to a more differentiated understanding of implicit Theory of Mind abilities across development and neurodiversity. They offer methodological innovations for implicit measurement of early social cognitive abilities, highlight both the potential and limitations of current theoretical accounts and empirical paradigms, and point to directions for future research.

Deutsche Zusammenfassung

In sozialen Interaktionen denken Menschen über ihre eigenen mentalen Zustände —wie Überzeugungen, Wünsche oder Absichten —nach und schreiben auch anderen Personen mentale Zustände zu, um deren Verhalten vorherzusagen oder zu erklären. Diese Fähigkeit, die als Theory of Mind bezeichnet wird, ist zentral für gelingende soziale Interaktion und entwickelt sich im Kindesalter. Explizite Theory of Mind, ein bewusstes und verbal ausdrückbares Verständnis mentaler Zustände, entwickelt sich typischerweise im Alter von etwa 4 Jahren. Im Gegensatz dazu ist die frühere Entwicklung der impliziten Theory of Mind, einem unbewussten, nicht-verbalen Zugang zu mentalen Zuständen anderer, weiterhin umstritten, da Befunde aus verschiedenen experimentellen Paradigmen inkonsistent sind. Theory of Mind-Fähigkeiten stehen in engem Zusammenhang sowohl mit der Verwendung als auch mit dem Verstehen mentaler Zustandssprache. Die Forschung im Bereich des Verstehens mentaler Zustandssprache zeigt, dass Kinder ab etwa 4 bis 5 Jahren verbal zwischen epistemischen Verben—wie wissen, glauben oder denken—unterscheiden können. Es bleibt jedoch unklar, ob diese Fähigkeit bereits früher vorhanden ist, da Forschung in diesem Bereich bislang begrenzt ist. Bei neurodivergenter sozial-kognitiver Entwicklung, etwa bei Autismus, bringen aktuelle Predictive Coding-Ansätze Schwierigkeiten in Theory of Mind-Fähigkeiten mit einer Abschwächung in der prädiktiven Verarbeitung in Verbindung. Da implizite Theory of Mind auf unbewussten Vorhersagen über mentale Zustände anderer beruht, könnten Unterschiede zwischen autistischen und neurotypischen Personen auf Veränderungen in zugrunde liegenden prädiktiven Mechanismen hinweisen und somit eine mögliche Erklärung für Herausforderungen in der sozialen Interaktion liefern. Dennoch bleiben eindeutige neuronale Befunde bislang aus. Vor diesem Hintergrund ergeben sich folgende zentrale Fragen: Wie robust sind die bestehenden Paradigmen zur Erfassung impliziter Theory of Mind über die Lebensspanne hinweg? Entwickelt sich das implizite Verständnis epistemischer Sprache vor dem expliziten? Unterscheidet sich die prädiktive

Verarbeitung in Theory of Mind-relevanten Gehirnregionen zwischen neurotypischen und autistischen Personen?

Ziel dieser Dissertation war es, diese Fragen systematisch zu untersuchen. Zu diesem Zweck wurden vier empirische Studien durchgeführt, die jeweils unterschiedliche Aspekte impliziter Theory of Mind im Rahmen neurotypischer und/oder neurodivergenter Entwicklung beleuchteten. In Studie 1 wurde ein neues web-basiertes Eye-Tracking-Verfahren bei Kleinkindern im Alter von 18 bis 27 Monaten (*N* = 125) validiert. Im Rahmen eines antizipatorischen Blickverhalten-Paradigmas (engl. *anticipatory looking paradigm*) wurde zielgerichtete Handlungsantizipation erfasst und mit Ergebnissen einer Laborstudie verglichen. Die Ergebnisse zeigten, dass zielgerichtete Handlungsantizipation in unserem web-basierten Setting zuverlässig messbar ist, wenngleich der Anteil des antizipatorischen Blickverhaltens etwas geringer und die Ausschlussrate höher war. Studie 2 nutzte ebenfalls das Paradigma aus Studie 1, das antizipatorisches Blickverhalten misst, und untersuchte, ob sich die Handlungsantizipation von Kleinkindern (*N* = 521) und Erwachsenen (*N* = 703) basierend auf den epistemischen Zuständen Wissen und Unwissen unterscheiden. Erwachsene unterschieden klar zwischen Wissen und Unwissen, was sich in ihrem antizipatorischen Blickverhalten widerspiegelte. Unerwarteterweise zeigten Kleinkinder diese Differenzierung nicht, was die Notwendigkeit weiterer Forschung unterstreicht. Studie 3 untersuchte das implizite Verstehen der epistemischen Verben „wissen“ und „glauben“ bei 27-monatigen (*N* = 199) und 36-monatigen Kindern (*N* = 131). Die Ergebnisse ergaben, dass bereits 27 Monate alte Kinder zwischen diesen Verben differenzieren konnten und spontan das Verb „wissen“ präferierten. Dies deutet darauf hin, dass ein implizites Verständnis epistemischer Verben dem expliziten vorauszugehen scheint. Studie 4 untersuchte prädiktive Prozesse im Theory of Mind-Netzwerk bei nicht-autistischen (*N* = 61 in Experiment 1; *N* = 30 in Experiment 2) und autistischen Erwachsenen (*N* = 30 in Experiment 2) mittels funktioneller Magnetresonanztomographie. Entgegen den Erwartungen fanden sich in keiner der beiden Gruppen Hinweise auf prädiktive Verarbeitung im Theory of Mind-Netzwerk.

Allerdings zeigte sich bei autistischen (im Vergleich zu nicht-autistischen) Erwachsenen eine reduzierte Abschwächung der neuronalen Antwort bei wiederholter Präsentation (engl. *repetition suppression*) in einer Szene, die komplexe Theory of Mind-Prozesse erforderte. Diese ersten neuronale Hinweise stützen die Theorie der prädiktiven Kodierung bei Autismus. Gemeinsam tragen die vier Studien zu einem differenzierteren Verständnis impliziter Theory of Mind-Fähigkeiten über Entwicklungsverläufe und Neurodiversität hinweg bei. Sie bieten methodische Innovationen zur impliziten Erfassung früher sozial-kognitiver Fähigkeiten, zeigen sowohl Potenziale als auch die Grenzen bestehender theoretischer Ansätze und empirischer Paradigmen auf und weisen auf mögliche Richtungen für zukünftige Forschung hin.

List of scientific publications for the cumulative thesis

The present thesis is based on four original articles: three are published in peer-reviewed journals, one is under review as a Stage 2 Registered Report and available as a preprint on OSF.

1) Steffan, A.*, **Zimmer, L.***, Arias-Trejo, N., Bohn, M., Dal Ben, R., Flores-Coronado, M. A., Franchin, L., Garbisch, I., Grosse Wiesmann, C., Hamlin, J. K., Havron, N., Hay, J. F., Hermansen, T. K., Jakobsen, K. V., Kalinke, S., Ko, E.-S., Kulke, L., Mayor, J., Meristo, M., ... Schuwerk, T. (2024). Validation of an open source, remote web-based eye-tracking method (WebGazer) for research in early childhood. *Infancy, 29*(1), 31–55. https://doi.org/10.1111/infa.12564 (* shared first authorship)

2) Schuwerk, T.*, Kampis, D.*, Alessandroni, N., Altvater-Mackensen, N., Arias-Trejo, N., Axelsson, E. L., Baillargeon, R., Baumann, A.-E., Bernard, C., Biro, S., Blankenship, T. L., Blomberg, I., Bohn, M., Bradford, E. E. F., Byers-Heinlein, K., Canudas Grabolosa, I., Chen, E. M., Chen, X., Corbit, J., … **Zimmer, L.**, … Rakoczy, H. (2025, under review). *Action anticipation based on an agent's epistemic state in toddlers and adults.* [Stage 2 Registered Report, following in-principle acceptance at Stage 1] Child Development. Preprint: https://doi.org/10.31234/osf.io/x4jbm (*shared first authorship)

3) **Zimmer, L.**, Sodian, B., Mani, N., Grosso, S. S., Kristen-Antonow, S., & Schuwerk, T. (2025). Two- to three-year-old toddlers differentiate the epistemic verbs "know" and "think" in a preferential looking eye-tracking paradigm. *Developmental Psychology*. https://doi.org/10.1037/dev0001933

4) **Zimmer, L.**, Richardson, H., Pletti, C., Paulus, M., & Schuwerk, T. (2025). Predictive responses in the theory of mind network: A comparison of autistic and non-autistic adults. *Cortex, 187*, 159–171. https://doi.org/10.1016/j.cortex.2025.04.006

1. General Introduction

Mentalizing, the ability to attribute mental states, such as beliefs, desires, and intentions, to oneself and others, is fundamental to social interaction (Quesque et al., 2024). In everyday social exchanges, individuals think about their own mental states and ascribe mental states to other people in order to predict and/or explain their behavior—a process known as Theory of Mind (Premack & Woodruff, 1978; Wimmer & Perner, 1983). Gaining insight into how individuals reason about mental states of others is essential for advancing social cognition research, especially in early development and within the context of neurodiversity. Neurodiversity describes the range of natural variation in neurological development and recognizes that conditions such as autism spectrum disorder represent one variation of human diversity (Pellicano & den Houting, 2022). Hence, neurotypical refers to the neurological development of individuals that falls within the range generally considered to be "typical", while individuals outside this range are considered neurodivergent. Yet, because mental states are inherently unobservable, they pose a dual challenge: for individuals as they develop the capacity to reason about them, and for researchers attempting to study these internal cognitive processes. Research distinguishes between explicit and implicit Theory of Mind: explicit Theory of Mind refers to conscious, verbally expressed reasoning about others' mental states. It draws on explicit knowledge, where the inference from belief to action is guided by a consciously represented conditional. Implicit Theory of Mind, in contrast, refers to early, non-verbal and automatic sensitivity to others' mental states. It builds on implicit knowledge, where such conditionals unconsciously guide behavior without being mentally represented (cf. Perner & Roessler, 2012).

In my dissertation, I present work investigating different aspects of implicit Theory of Mind, ranging from goal-based as well as epistemic state-based action anticipation to mental state language understanding and neural Theory of Mind processing in neurotypical and neurodivergent individuals, utilizing eye-tracking and functional magnetic resonance imaging (fMRI). First, I will introduce the current state of research on implicit Theory of Mind, integrating findings from both neurotypical and neurodivergent social cognitive development. In addition, I will outline paradigms and methods used to measure implicit Theory of Mind. Second, I will summarize four empirical studies conducted to test different aspects of implicit Theory of Mind reasoning across two age ranges (toddlerhood and adulthood), including neurodiversity (neurotypical and autistic[1] individuals). One aim of these studies was to shed light on when mental state reasoning develops and how Theory of Mind processes may differ between neurotypical and neurodivergent social cognitive development. The studies incorporate a variety of research contexts (measurement methods, lab contributions and study settings) to explore how implicit Theory of Mind can be measured in a way that is comfortable for participants, feasible for labs, and valid for research. Third, based on the results of these studies, I will discuss the developmental trajectory of implicit Theory of Mind reasoning across age and neurodiversity, taking methodological considerations into account.

1.1. Theory of Mind

1.1.1. Explicit False Belief Understanding

More than 40 years ago, Theory of Mind research gained momentum when Premack and Woodruff (1978) published their seminal study on whether chimpanzees have a Theory of Mind and found that they can learn to infer a human's intention. This idea was further developed, ensuring that actions were not merely based on interpretation of another's behavior but rather on the individual's reasoning about the other's mental states (Bennett, 1978; Dennett, 1978; Harman, 1978). These theoretical considerations led to the proposal of the first study using the *location-change task* to measure false belief understanding as an indicator of Theory of Mind in children: "[…] Maxi puts chocolate into a cupboard x. In his absence his mother displaces the chocolate from x into cupboard y. Subjects have to indicate the box where Maxi will look for the chocolate when he returns." (Wimmer & Perner, 1983, p. 106). In order to solve this task children have to

[1] In line with preferences within the autistic community (Kenny et al., 2016), identity-first language (e.g., autistic individuals) is used throughout this dissertation.

take into account that Maxi's belief about the location of the chocolate differs from both their own mental representation and reality[2]. Between the ages of 4 and 6, children begin to demonstrate explicit Theory of Mind skills, as evidenced by their correct verbal prediction that Maxi will search for his chocolate in cupboard x, despite its actual location in cupboard y. This prediction reflects explicit knowledge in the sense that the inference from belief to action is guided by a consciously accessible representation of the underlying conditional ("if someone believes the chocolate is in x, they will look in x"; Perner & Roessler, 2012). None of the 3-year-olds successfully passed this experiment. To test false belief understanding in younger children, Perner et al. (1987) developed an *unexpected-contents task*, in which children were shown a familiar Smarties box which contained a pencil rather than sweets. They were then asked about their own prior belief and what a friend, who hadn't seen inside, would think is in the box, testing false belief understanding. However, this simpler task remained difficult for 3-year-olds, leading the authors to conclude that a conceptual limitation underlies their difficulty with false belief understanding. This interpretation informed the development of *conceptual-change accounts* that posit that explicit false belief understanding—requiring children to provide verbal responses—only becomes possible once the conceptual capacity for belief understanding emerged.

Numerous successful replications of explicit Theory of Mind tasks have confirmed the original findings, with a seminal meta-analysis by Wellman et al. (2001) showing that less than 20% of children at 2.5 years were able to correctly solve false belief tasks. Around one year later, approximately half of the children provided correct responses, with performance on false belief tasks improving substantially with increasing age. Several studies examined potential reasons for the emergence of Theory of Mind abilities around the age of 4, highlighting associations with advanced language abilities and executive functions (Devine & Hughes, 2014; Milligan et al.,

[2] False belief understanding is commonly distinguished into first-order (i.e., recognizing that another person holds a false belief about owns mental representations and reality) and second-order (i.e., understanding that someone holds a false belief about another's belief), with the latter typically emerging between ages 6 and 8 (Perner & Wimmer, 1985). The framework has since been extended to higher-order beliefs involving recursive reasoning that develops into adulthood (Rakoczy, 2022). This dissertation focuses on first-order false belief understanding.

2007). In sum, the results of explicit (i.e., prompted by direct mental state questions) measures of false belief understanding are largely robust and consistent, providing evidence for developmental changes in explicit Theory of Mind between the ages of 3 and 4.

Yet, studies using non-verbal tasks suggest that an understanding of false belief may already be present before children are able to express it explicitly. In an implicit location-change task, Clements and Perner (1994) contrasted results with explicit answers. They showed that children between 3 and 4.5 years of age correctly looked toward the location where an agent falsely believed an object to be. However, more than half of them failed to indicate this location when explicitly asked. Thus, children younger than 4 years of age may already be able to ascribe false beliefs to others', although they cannot express this understanding verbally. In addition, Onishi and Baillargeon (2005) developed a way to investigate false belief understanding in pre-verbal children using the location-change task. Through looking time measures they demonstrated that even infants are sensitive to others' false belief understanding. Based on their finding they proposed—in contrast to the conceptual change accounts—early-emerging and gradually developing Theory of Mind abilities. In contrast to this finding further studies on false belief understanding in infancy report mixed results: while some support the original finding (e.g., Southgate et al., 2007; Surian et al., 2007), others have failed to replicate it (e.g., Kampis et al., 2021; Kulke, Reiß et al., 2018; Schuwerk et al., 2018).

Before addressing implicit false belief understanding in more detail, I summarize the development of precursors and more basic forms of implicit Theory of Mind. These findings are less debated and help to better understand what remains unresolved.

1.1.2. Goal-Based Action Prediction

A seminal study suggests that infants as young as 6 months of age already engage in simple *goal-based action* expectations. As evidenced by their looking behavior, they seem to form expectations about simple and familiar action goals, such as grasping for an object (Woodward, 1998). In another study, 12-month-olds (and adults) —but not 6-month-olds—anticipated the goal of a simple manual reach-and-transport action, in which an agent reached for an object and placed it into a bucket (Falck-Ytter et al., 2006). In an extension of this paradigm, in which children observed failed goal-based reaching actions, 10-month-olds were still able to anticipate the actor's goal (Brandone et al., 2014). The type of action and goal modulates infants' anticipatory gaze shifts, with greater anticipation occurring when objects are placed into containers rather than merely displaced (Gredebäck et al., 2009). To test the ability to predict more complex goal-based actions at this age, Cannon and Woodward (2012) adapted the paradigm of Woodward (1998): 11-month-old infants were familiarized with movie clips showing a hand grasping one of two objects. In the test trial, the object locations were swapped, and the hand made an incomplete reach between the two objects. Results revealed that infants preferred the familiarized object in the new location, leading the authors to conclude that infants encode goals behind others' actions. This sensitivity, however, seems limited to actions performed by human agents (Cannon & Woodward, 2012), unless non-human agents (e.g., mechanical claws) display agency cues (for a review, see Elsner & Adam, 2020). Recent studies, however, failed to replicate results of the Cannon and Woodward (2012) paradigm, showing that 1- to 3-year-old children were more likely to anticipate the action based on movement rather than the goal (Ganglmayer et al., 2019). Moreover, preschool children predicted the movement trajectory even when it led to a new goal, indicating that they expected the agent to follow a familiar path rather than pursue the previously established goal (Gönül et al., 2024). However, when the goal location varied across trials, children prioritized the goal object over the movement path (e.g., Ganglmayer et al., 2020; Paulus et al., 2017). By the second year of life, children not only consider the intentions behind adults' actions but also begin to imitate them based on the predicted goals (Carpenter et al., 2005). Interestingly, children younger than 3 years of age were not yet able to use verbally provided (i.e., explicit) information to visually anticipate others' actions, indicating limitations in the integration of explicit and implicit goal-based action prediction at this age (Paulus et al., 2017). Nonetheless, a longitudinal study documented relations between the ability to encode the goal of an action using the Cannon and Woodward paradigm and later Theory of Mind abilities (Aschersleben et al., 2008), indicating a developmental link between goal-based action predictions and later false belief understanding.

1.1.3. Epistemic State-Based Action Prediction

Beyond the understanding what others want (i.e., goals and intentions), children also become sensitive to what others perceive, know or belief (i.e., mental states such as attention and knowledge). With the beginning of the second year of life children start engaging in basic forms of *epistemic state-based action* considerations. From their first birthday onwards, children engage in declarative joint attention by showing sensitivity to the knowledge states of others and communicating it non-verbally with pointing gestures. They recognize that certain objects may be unfamiliar to others—even if these objects are already familiar to themselves—when they are actively engaged in joint attention with another person (Tomasello & Haberl, 2003). In contrast, they are not able to make this distinction when observing someone perceive these objects from a third-person perspective (Moll et al., 2007). Similarly, it has been argued that only children starting with the age of 2 years have the capacity to differentiate what another person does and does not see, the so-called *Level 1 visual perspective-taking* (Moll & Tomasello, 2006). This is followed by the more advanced *Level 2 visual perspective-taking* at around age 4, that is the ability to understand that another person can see the same object differently from oneself (Flavell et al., 1981). Moreover, infants are able to distinguish between knowledgeable and ignorant partners not only when helping others, but also when seeking information for themselves. They prefer strangers over their caregivers, when they expect strangers and not their caregivers to possess the respective information (Stenberg, 2009). Additionally, they help an ignorant but not a knowledgeable person (Dunham et al., 2000; Liszkowski et al., 2008; see O'Neill, 1996 for similar findings in 2-year-olds). Thus, 1-year-old children seem to have an implicit representation of their communication partner's knowledge states and informational needs.

1.1.4. Implicit False Belief Understanding

Thus, as aforementioned, the past two decades have seen a growing number of studies employing spontaneous-response measures to investigate whether even preverbal children possess an implicit Theory of Mind. In their pivotal violation of expectation study, Onishi and Baillargeon (2005) first familiarized infants with an agent who played with a toy and hid it in one of two boxes and, after a brief pause, reached into the same box in which they had previously placed it. This was followed by belief induction trials, during which the infants observed a change in the toy's location while the agent either held a true or a false belief about its new location. In the test trial, the agent reached into one of the two boxes and paused. Fifteen-month-old infants looked longer when the agent reached toward the box where the toy really was (i.e., the agent's behavior was incongruent with their false belief) than when the agent reached toward the box where the agent falsely believed the toy to be (i.e., the agent's behavior was congruent with their false belief). This was interpreted as infants' sensitivity to the agent's mental state.

In their view, Southgate et al. (2007) considered this interpretation as possibly reflecting an attribution of ignorance rather than a true understanding of belief. To provide an even stronger test of implicit false belief understanding, they developed an anticipatory looking paradigm. In this paradigm, children were first familiarized with two events, in which a puppet bear hid a ball in one of two boxes, and then an actor (1) reached to the box to retrieve the ball and (2) only looked at the box with the ball in it[3]. In two false belief conditions, the actor witnessed the ball being hidden in one of the two boxes (i.e., the original box). Depending on the condition, the actor either saw the ball transferred from the original to the other box (i.e., false belief 1 condition) or got distracted and did not witness the location-change (i.e., false belief 2 condition). In each condition, the bear took the ball away from the scene while the actor was facing away. When the actor turned back, anticipatory looking was measured. In both conditions,

[3] Only children who correctly anticipated the outcome of the second familiarization trial were included in the analysis. This criterion is often referred to as the "Southgate criterion" in subsequent research.

twenty-five-month-old infants correctly anticipated that the actor would reach for the location consistent with the actor's false-belief (i.e., original location), indicating a false belief attribution.

These findings from non-traditional false belief understanding studies in children in their second and third year of life have been supported by several subsequent studies (Surian et al., 2007; Surian & Geraci, 2012; for results under age 1 year, see Kovács et al., 2010), supporting the early developing, full mentalistic capacity accounts. Researchers supporting this theoretical perspective explain young children's failure in explicit Theory of Mind tasks for instance by processing difficulties (Carruthers, 2013; for a review see Scott & Baillargeon, 2017).

In contrast, traditional research explains the absence of finding explicit Theory of Mind abilities in children under 3 to 4 years of age by their conceptual limitations. Consequently, findings on early false belief understanding are interpreted as reflecting simple behavioral rules: rather than reasoning about mental states, infants may merely link an agent's perception to their subsequent behavior. Accordingly, explanations such as "people tend to search for objects where they last saw them" have been proposed (Perner & Ruffman, 2005). Additionally, these accounts highlight the role of infants' attentional processes and sensitivity to behavioral patterns (Ruffman, 2014).

Another theory, the dual-system theory, explains the contradictory findings on Theory of Mind abilities by proposing two distinct and parallel systems. On the one hand, this theory proposes an early-developing minimal Theory of Mind, which is cognitively efficient but inflexible. This system allows infants to pass spontaneous false belief tasks. On the other hand, it describes a later-developing full-blown Theory of Mind, which is cognitively demanding but flexible. This second system enables successful completion of explicit false belief tasks around the age of 4, supported by the development of language and executive functions (Apperly & Butterfill, 2009; Butterfill & Apperly, 2013). Similarly, the low-level novelty hypothesis explains successful false belief understanding studies in infants due to the novelty of low-level properties (such as color, shape, and movement) relative to earlier encoded events (Heyes, 2014).

Parallel to these theoretical discussions, failed replications of early false belief understanding studies raised doubts about an early implicit Theory of Mind. Several replication attempts failed to reproduce the original findings of two influential studies on early false belief understanding: Onishi and Baillargeon's (2005) violation of expectation study (failed replications: Poulin-Dubois & Yott, 2018; Powell et al., 2018) and Southgate et al.'s (2007) anticipatory looking study (failed replications: Dörrenberg et al., 2018; Grosse Wiesmann et al., 2018; Kampis et al., 2021; Kulke, Reiß et al., 2018; Schuwerk et al., 2018). These repeated replication failures contributed to what is now referred to as a replication crisis in developmental cognitive science (Poulin-Dubois et al., 2018; but see Baillargeon et al., 2018). In a review and meta-analysis, Barone et al. (2019) underpinned doubts about the robustness of findings on implicit Theory of Mind by revealing an asymmetric distribution of studies testing false belief understanding in children younger than 2 years. The findings suggest a publication bias, with studies reporting larger effect sizes and smaller sample sizes being more likely to be published. They also highlight that success rates vary by paradigm, with higher success rates found in violation-of-expectation tasks. Similarly, Kulke and Rakoczy (2017) provided a qualitative summary of both published and unpublished studies on false belief understanding. They reported that when unpublished studies are considered, non-replications and partial replications outnumber successful replications.

In adults, a substantial body of literature suggests that Theory of Mind reasoning occurs spontaneously, automatically and without conscious effort (for a review, see Schneider et al., 2017). Consistent with this view, some studies documented successful application of anticipatory looking paradigms in adult false belief tasks (Schuwerk et al., 2018; Senju et al., 2009). In contrast, other studies found weaker anticipatory false belief reasoning in adults (Burnside et al., 2018; Kulke, Reiß et al., 2018), thereby further challenging the notion of early implicit false belief understanding. This raises the question of whether implicit Theory of Mind abilities are less robust than previously assumed or simply more difficult to detect. Thus, systematically evaluating the validity and reliability of implicit Theory of Mind findings across the lifespan—irrespective of underlying theoretical frameworks—is essential for advancing a sound understanding of early social cognitive development.

1.2. Mental State Language

1.2.1. Mental State Language Acquisition

Language plays a fundamental role in human interaction, enabling individuals to convey mental states such as thoughts, beliefs and knowledge. Research shows that advanced mental state language in 2.5-year-olds predicts Theory of Mind development by age 4 (Brooks & Meltzoff, 2015; Olineck & Poulin-Dubois, 2007). Moreover, training in mental state language promotes understanding of concepts like knowledge and ignorance already before age 3 (Kaltefleiter et al., 2022). Thus, mental state language appears to be related to the understanding of epistemic states and false beliefs.

The frequency of mental state words (e.g., think, know, believe) used by children is associated with the overall number of words produced, indicating a relation between cognitive word usage and broader cognitive processes (Booth et al., 1997). Children first learn to verbalize their own mental states before expressing those of others (Gonzales et al., 2018; Kaltefleiter et al., 2021). At around 18 months of age, children typically acquire a productive vocabulary of approximately 50 words (Kauschke & Hofmeister, 2002). This milestone coincides with the first verbal expressions of desire and emotion (Bartsch & Wellman, 1995; Bretherton & Beeghly, 1982). Around six months later, they start talking about epistemic states such as knowledge and ignorance ("know" and "don't know"), albeit references to beliefs ("think") remain rare at this age (Shatz et al., 1983). Early analyses of toddlers' spontaneous mental state language claimed that between the ages of 2 and 3, mental terms primarily serve conversational purposes (e.g., "you know") before toddlers are able to clearly use them to refer to mental states (Bartsch & Wellman, 1995; Shatz et al., 1983). In contrast, more recent studies challenged this claim by showing that 2-year-olds intentionally reference epistemic states in conversations, spontaneously relating to the ongoing topic. They appropriately affirm their own knowledge and, to a lesser extent that of their conversation partner, as well as express denial of their own knowledge (Harris, Yang & Cui, 2017). Moreover, in an experimental study assessing children's awareness of own ignorance, 28- to 37-month-old children said "I don't know" or asked for information more frequently when asked to name fictitious (i.e., unknown) compared to real objects. Even earlier, at 16 to 27 months, toddlers tended to non-verbally express their ignorance in this task (Harris, Ronfard & Bartz, 2017). Thus, contrary to earlier studies that attributed limited mental state language abilities to young children, more recent research indicates that 2-year-olds already distinguish "know" from other epistemic verbs in their spontaneous speech.

1.2.2. Mental State Language Understanding

While empirical evidence revealed that children begin using mental state terms in their second year, their understanding of these terms seems to develop later. A longitudinal study found that use of desire terms at 2 years was linked to belief term use at 3 years and understanding at 4 years (Moore et al., 1994). In preschoolers, understanding of epistemic verbs has been linked to false belief understanding (Matsui et al., 2006). These studies further inform a developmental pattern from early epistemic verb use to understanding and later to false belief reasoning. In a pioneering study on the development of mental state language, Moore et al. (1989) tested children's ability to distinguish between verbs with varying degrees of speaker certainty, such as "know", "think", and "guess". They used a *conflicting sources task*, in which children had to choose which statement to follow based on two conflicting sources. The children could not verify the assertions against a state of reality. While 3-year-olds were at chance, by age 4 children preferred the "know" statement over the others, with further improvement apparent between the ages of 4 and 5. Kristen-Antonow et al. (2019) replicated the study with German-speaking children and found weaker performance overall. Children only reached above-chance competence in either one of the "know-guess" and "know-think" contrasts at age 5, with more than 50% reaching full competence by age 7. In sum, these studies suggest that a full understanding of "know", "think", and "guess" is attained in school age. By ages 4 to 5, children show a basic grasp of the "know–think" contrast, with understanding of high speaker certainty (i.e., "know") developing earlier than that of lower speaker certainty (i.e., "think").

Verbs like "know" link an agent to the truth, implying certainty about the truth of the embedded statement and are called factive mental state verbs. In contrast, verbs like "think" or "guess" can link an agent to either truth or falsehood, indicating uncertainty about the truth of the statement and are called non-factive mental state verbs (Abbeduto & Rosenberg, 1985; Nagel, 2017). Dudley et al. (2015) also used Moore et al. (1989)'s interactive game paradigm, but without presenting contrasting statements. They tested 3-year-olds' understanding of negated "know" and "think" sentences in order to gain insights into their understanding of facticity. While children recognized that "doesn't think" is unreliable, they did not understand the implications of negated "know"-statements about a third party's knowledge (e.g., "the agent doesn't know that it's in the x box"), thus treating both verbs as non-factive. The protracted development implies that young children differentiate "know" from "think" based on cues to speaker certainty, rather than an understanding of factive and non-factive verbs (Kristen-Antonow et al., 2019; Moore et al., 1989).

1.3. Measuring Implicit Theory of Mind

1.3.1. Paradigms

Interactive Paradigms. In previous decades, researchers used a variety of approaches to explore how children make sense of others' beliefs, desires, and intentions. Among these, measures that require verbal responses were widely used to study children's cognitive development. For instance, the location-change task by Wimmer and Perner (1983) as well as the unexpected-content task (Perner et al., 1987) were established to measure false belief understanding. Moreover, the conflicting sources task by Moore et al. (1989) was used to measure mental state language understanding. However, all these paradigms rely on the child's advanced language skills, making them unsuitable for examining cognitive processes in younger and/or pre-verbal children. To address this limitation, researchers increasingly turned to methods that rely on non-verbal responses to explore early implicit mental state reasoning. For instance, in non-verbal interaction studies, experimenters engage with children in a playful way, motivating them to proactively intervene. In these studies, non-verbal behavioral responses, such as actively helping or pointing, are measured to determine whether children are able to take the experimenter's mental states into account (e.g., Buttelmann et al., 2009; Liszkowski et al., 2008). For example, in a study by Buttelmann et al. (2009), an experimenter struggled to open a box in which they either truly or falsely believed a toy was hidden. Children's behavioral responses, namely opening or at least touching the correct box, were measured as an indicator of their ability to take the adult's belief into account to achieve their goal. However, failed replication studies suggest that the observed effects may instead be driven by for instance general prosocial tendencies rather than reasoning about other's mental states (Poulin-Dubois & Yott, 2018; Priewasser et al., 2018).

Violation of Expectation Paradigms. Moreover, research using (spontaneous) response measures provides a valuable alternative for exploring cognitive development in younger children. Habituation-based approaches, such as violation of expectation paradigms, measure for instance looking time in response to a participant's observation of a completed event. Events that contradict participants' expectations are assumed to elicit longer looking times than expected events (e.g., Onishi & Baillargeon, 2005; Surian et al., 2007; Woodward, 1998). While some argue that violation of expectation paradigms, with their low task demands, are especially suitable for children (Stahl & Kibbe, 2022), others suggest that these paradigms rely on too many assumptions, making the findings difficult to interpret. As an alternative, preferential looking methods are recommended in which children are presented with two stimuli simultaneously and are expected to prefer one over the other (Paulus, 2022).

Anticipatory Looking Paradigms. Similarly, anticipatory looking paradigms, such as measurements of first look or anticipatory looking time, have been established as spontaneous response methods for measuring mental state reasoning across the lifespan. Anticipatory looking refers to a participant's gaze (shift) toward an expected outcome before it occurs. Such methods measure gaze within milliseconds while the action is still incomplete. It serves as an indicator of participants' predictions about an agent's upcoming action (e.g., Southgate et al., 2007; Surian & Geraci, 2012). During the second half of their first year, infants begin to produce proactive, goal-directed eye movements (Falck-Ytter et al., 2006). In the first study assessing implicit false belief understanding, anticipatory looking was used in a location-change task (Clements & Perner, 1994). Since then, it has become a key method for studying implicit Theory of Mind. Unlike violation of expectation paradigms, which assess reactions after an action unfolds, it offers a more spontaneous assessment of participants' predictions. In recent years, however, the reliability and robustness of anticipatory looking paradigms have been increasingly questioned, as studies reported mixed findings in false belief understanding across the lifespan (e.g., Kampis et al., 2021; Kulke, von Duhn et al., 2018; Schneider et al., 2017; Southgate et al., 2007). Notably, some participants failed to anticipate the agent's action even during familiarization trials—designed to convey the agent's goal without requiring belief reasoning (i.e., goal-based action anticipation)—resulting in high exclusion rates and limiting the interpretability of findings (Kampis et al., 2021; Kulke, Reiß et al., 2018; Schuwerk et al., 2018). These studies applied stimuli and exclusion criteria established by original studies. These exclusion criteria were based on the assumption that only participants who correctly predicted the agent's action in the familiarization trials could be considered to have understood the agent's goal and to be sufficiently motivated to engage in visual action anticipation (e.g., Senju et al., 2009; Southgate et al., 2007). However, for instance in a failed replication study no performance differences were found between children who passed or failed Southgate et al.'s (2007) criterion (Grosse Wiesmann et al., 2018). Altogether, these findings underscore the need for further research to better understand the reliability and robustness of anticipatory looking measures in Theory of Mind research.

1.3.2. Methods

In-Lab Eye-Tracking. Traditionally, research relying on measures such as looking behavior required manual coding by human observers. However, manual coding is labor-intensive, inaccurate, and prone to observer bias, requiring careful training and establishment of inter-rater reliability (Oakes, 2012). Considering the advancements in current technology, more fine-grained techniques such as pupil-corneal reflection eye-tracking have become widely used to draw inferences about children's cognitive processes. By directing infrared light at the participant's eye, the reflections on the cornea relative to the center of the pupil provide an estimate of where the participant's gaze is fixated. Eye-tracking enables a more accurate and unbiased assessment of measurements that were previously coded manually. The improved spatiotemporal resolution of eye-tracking technology allows for the analysis of the microstructure of gaze behavior and enables the examination of for instance the frequency, discrete fixations within areas of interests (AOIs) and patterns of gaze behavior (Aslin, 2007; Oakes, 2012). In preferential looking paradigms proportional looking time scores can be calculated by dividing the fixations or time children looked at a target by the overall fixations or time children looked at the target and distractor, excluding looks outside of either AOI. For example, measurements like the proportion of target looking time (PTL) in which fixations are calculated, are commonly used in language development research (e.g., Mani & Plunkett, 2007, 2008). In addition, the proportion of differential looking score (DLS), in which looking time is calculated, is used as anticipatory looking measures in Theory of Mind research (e.g., Senju et al., 2009). In sum, eye-tracking enables the investigation of new research questions and allows for a broad range of dependent variables.

Web-Based Eye-Tracking. Although progress in web-based eye-tracking had already begun prior to the Covid-19 pandemic, related restrictions accelerated the shift away from laboratory-based studies using commercial eye-tracking systems. As a result, remote web-based studies became increasingly popular. Typical developmental patterns of false belief understanding in explicit tasks were for instance successfully replicated in moderated video-call sessions with toddlers (Schidelko et al., 2021). Preferential looking behavior in infants was replicated in both moderated video-call studies (Smith-Flores et al., 2021) and unmoderated online studies (Bánki et al., 2022), although these studies still required manual coding by human observers. Platforms such as Lookit (Scott & Schulz, 2017) enabled large-scale webcam-based testing of children, but typically also still rely on manual frame-by-frame annotation. Together, these studies support the feasibility of collecting valid data online, however, manual coding seems to remain necessary.

To address this limitation, new tools were developed that use machine learning techniques to automate gaze classification. For instance, iCatcher+ (Erel et al., 2023) uses deep learning trained on hand-labeled infant webcam footage to classify gaze behavior into categories such as left, right, or away with human-level accuracy. Similarly, both WebGazer (Papoutsaki et al., 2016) and OWLET (Werchan et al., 2022) employ machine learning techniques to estimate continuous gaze coordinates. WebGazer infers gaze locations in real time within the browser, while OWLET processes recorded gaze data post hoc. Given the limitations of manual coding, such automated approaches offer a resource-saving alternative for at-home testing (Ozkan, 2018). Their scalability promises greater access to larger, more diverse samples and may facilitate international collaboration. Nevertheless, these methods are still under development and come with challenges, including poorer image quality, uncontrolled experimental conditions, and reduced sampling rates compared to in-lab setups (Wass, 2016; Zaadnoordijk et al., 2021). Hence, studies are needed to validate automated web-based eye-tracking and refine these tools for remote use in developmental research.

Functional Magnetic Resonance Imaging (fMRI). To complement behavioral and gaze-based measures, neuroimaging methods like fMRI can provide insights into the neural mechanisms underlying cognitive processes. fMRI is a non-invasive neuroimaging technique used to localize brain areas of cognitive activation. During cognitive activation of a brain region, blood flow increases to a greater extent than oxygen extraction, leading to a rise of oxygenated hemoglobin in active brain regions. This change in oxygen levels alters the blood properties, detected as the BOLD (for blood oxygenation level dependent) signal (American Psychological Association, 2018). The technique provides good spatial resolution but its temporal resolution is limited due to the time lag in oxygen extraction.

In 2003, Saxe and Kanwisher first identified the temporo-parietal-junction (TPJ) as a key region involved in processing third-person mental states such as goals and beliefs. In research of the following decade, overlap in activation was detected in the medial prefrontal cortex (MPFC) and bilateral posterior TPJ across different tasks (Schurz et al., 2014). This supports the idea of a core Theory of Mind network that is activated during reasoning about mental states, regardless of task or stimulus (Mar, 2011). Thus, fMRI has become a valuable tool in mapping the neural architecture underlying social cognitive processes.

Multi-Lab Collaboration Studies. One promising response to the replication crisis in psychological science was the implementation of multi-lab (or multi-site) collaboration studies. These studies involve coordinated efforts across multiple independent research labs, all using the same protocols to investigate the same phenomena. By pooling data across sites, they help to overcome limitations of traditional single-lab studies, such as small, homogeneous samples (e.g., an overreliance on WEIRD populations: Western, Educated, Industrialized, Rich, Democratic; Henrich et al., 2010), as well as issues of limited reproducibility, generalizability, and robustness (Nosek et al., 2022; Visser et al., 2022). Recent multi-lab studies in psychological science replicated approximately 50% of previously published original findings (Open Science Collaboration, 2015; for a review, see Nosek et al., 2022), highlighting the importance of accumulating evidence through systematic multi-lab replication. In developmental psychology, the *ManyBabies* project exemplifies this collaborative approach by providing more precise estimates of key developmental phenomena and generating new theoretical insights into how these vary across age groups, linguistic communities, and measurement methods (Frank et al., 2017; Visser, 2022). For example, the first ManyBabies study (i.e., ManyBabies 1) successfully replicated previous findings on infant-directed speech and extended these results to diverse populations across North America, Europe, Australia, and Asia (ManyBabies Consortium, 2020).

Such collaborative initiatives underscore the potential of large-scale, cross-cultural research to produce more reliable and generalizable insights into early cognitive development.

Multi-lab collaborations mark a positive structural shift in psychological science. By promoting transparency, openness, and collective responsibility, they exemplify the values of the ongoing credibility revolution, which reframes the replication crisis as an opportunity for lasting reform in psychological science (Korbmacher et al., 2023).

1.4. Theory of Mind in Neurodivergent Development

1.4.1. Autism Spectrum Disorder

According to the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5), autism spectrum disorder (hereafter *autism*), is a neurodevelopmental condition characterized by two core symptoms: difficulties in social interaction and communication and restricted, repetitive behaviors (American Psychiatric Association, 2013). The difficulties in social interaction and communication can manifest in various ways. For instance, autistic individuals often report challenges in initiating conversations, managing unstructured dialogues, interpreting gestures and tone of voice as well as understanding of concepts such as irony or sarcasm, and drawing appropriate social and emotional inferences; they may also struggle with understanding both implicit and explicit messages in social interactions (Müller et al., 2008). Autism is conceptualized as a spectrum, reflecting the wide variability in symptom representation. While the ICD-10 distinguished several types of autism (Dilling et al., 2015), the DSM-5 and ICD-11 classify autism as a single spectrum disorder, including distinctions based on severity, intellectual disability, and the presence or absence of functional language impairment (American Psychiatric Association, 2013; World Health Organization., 2022). Prevalence estimates are around 1% (Zeidan et al., 2022), although rates appear to be increasing (e.g., in the US, about 3.2%, are diagnosed with autism; Shaw et al., 2025).

1.4.2. (Traditional) Theory of Mind Research in Autism

Beyond its relevance for neurotypical development, Theory of Mind is also considered highly relevant for understanding neurodivergent trajectories, especially in autism. Since Theory of Mind abilities are fundamental for understanding and responding appropriately in social interactions, it has often been linked to one of the core symptoms of autism: the difficulties in social interaction and communication. This led researchers to examine false belief understanding in autistic individuals in order to identify potential differences in Theory of Mind reasoning compared to neurotypical individuals.

Baron-Cohen et al. (1985) adapted the puppet play introduced by Wimmer and Perner (1983) and developed the *Sally-Anne false belief task* to test false belief understanding in autistic children. In this task, Sally places a marble in her basket and leaves. In her absence, Anne moves it into a box. Children were then asked where Sally would look for her marble. Autistic children, but not neurotypical children and children with Down's syndrome, pointed to the marble's actual location. This was interpreted as a failure to take the other's false beliefs into account. Further studies using explicit Theory of Mind tasks in children supported these findings, leading early research to attribute the social difficulties observed in autism to a Theory of Mind deficit (for a review, see Baron-Cohen, 2000). However, in each of these studies, at least some autistic children passed false belief tasks (Tager-Flusberg, 2007), although they sometimes required a higher verbal mental age than non-autistic children (Happé, 1995). This has challenged the notion of a universal Theory of Mind deficit in autism. To further examine differences in Theory of Mind reasoning between autistic and non-autistic individuals, research in adulthood might provide valuable insights. Assessing explicit Theory of Mind in adults is challenging due to the limited availability of suitable tasks (for a review, see Livingston et al., 2019). Commonly used measures for distinguishing autistic from non-autistic individuals include the *Strange Stories Task*, the *Movie for the Assessment of Social Cognition (MASC),* and the *Reading the Mind in the Eyes Test*. In the Strange Stories Task, the so-called higher-order Theory of Mind is assessed by presenting short vignettes involving social scenarios requiring the participants to infer the character's (social) intentions or false beliefs to explain their behavior (Happé, 1994). In the MASC, participants watch a short movie and answer questions about the characters' mental states (Dziobek et al., 2006). Another, now-debated, task is the Reading the Mind in the Eyes Test, which requires participants to infer mental/emotional states from photographs of the eye region (Baron-Cohen et al., 2001; but see Oakley et al., 2016 for critique). Using these methods, multiple studies found that autistic adults perform equally well in explicit Theory of Mind tasks (e.g., Bowler, 1992; Schneider et al., 2013; Schuwerk et al., 2015; Senju et al., 2009; for school-aged children, see Scheeren et al., 2013). Nevertheless, they experience profound difficulties in social interactions in everyday life. It has been argued that this discrepancy may be due to compensatory strategies (Senju, 2012), which may not be flexible enough to use them in complex, social interaction (Livingston et al., 2019). However, these findings highlight the need for more nuanced accounts of Theory of Mind in autism.

With the development of implicit Theory of Mind measures in developmental psychology, these paradigms were also applied to autism research. Senju et al. (2009) conducted an eye-tracking study similar to that of Southgate et al. (2007) and argued that autistic adults lack spontaneous Theory of Mind, as reflected in reduced anticipatory looking in response to an actor's false belief. Additionally, in a multiple-trial implicit false belief paradigm it has been shown that learning processes do not mitigate this impairment (Schneider et al., 2013). However, when autistic individuals are provided with the outcome, they behave as neurotypical individuals in subsequent trials, suggesting that they quickly learn from action-outcome contingencies (Schuwerk et al., 2015). Consequently, experience leads to no differences between autistic and neurotypical adults. Other studies found no differences in an implicit Theory of Mind task in autism (Nijhof et al., 2018) but revealed that higher social symptomatology was associated with lower false belief performance (Deschrijver et al., 2016). These mixed findings from implicit Theory of Mind tasks together with evidence that autistic individuals perform well on explicit

Theory of Mind tasks, indicate that the relationship between such task performance and social communicative difficulties may be more nuanced than previously assumed. Additionally, in their review, Gernsbacher and Yergeau (2019) cast serious doubt on the claim of a general Theory of Mind deficit in autism, highlighting the failure to replicate seminal findings, the inability of Theory of Mind tasks to predict autistic traits or social interaction, and the lack of convergence among these tasks. These challenges highlight the need for new methodological approaches that more accurately capture the potentially nuanced relationship between Theory of Mind performance and difficulties in social interaction and communication in autism.

1.4.3. Neural Correlates of Theory of Mind in Autism

To further inform behavioral Theory of Mind research in autism, neuroimaging studies investigate whether observed and experienced differences in social interactions in autism can be linked to distinct neural mechanisms or processing. Several studies reported reduced activation or hypoactivation (Ciaramidaro et al., 2015; Kana et al., 2009; Kana et al., 2014; Nijhof et al., 2018; for a meta-analysis, see Sugranyes et al., 2011) as well as less functional connectivity (Ciaramidaro et al., 2015; Kana et al., 2009; Kana et al., 2014) in autistic individuals in brain regions that support Theory of Mind reasoning (e.g., TPJ, MPFC, superior temporal sulcus) during social cognitive tasks. In line with this, a recent review found reduced activation during Theory of Mind tasks, particularly in the TPJ, MPFC, and anterior cingulate cortex (ACC) in autistic individuals (Duvall et al., 2023). Cross-sectional comparisons of brain activity across development revealed both hypo- and hyperactivation in fronto-temporal structures in autistic children compared to autistic adults during social tasks (Dickstein et al., 2013). In contrast, large-scale studies found similar brain activation in autistic and neurotypical individuals in the Theory of Mind network during Theory of Mind tasks (Dufour et al., 2013; Moessnang et al., 2020), as well as during passive watching of a movie including social scenes (Mangnus et al., 2024). In sum, research on Theory of Mind reasoning and related neural activity in associated brain regions in autism yields conflicting findings, leaving robust and reliable neural evidence elusive.

1.4.4. Predictive Coding Theory

The inconsistent evidence regarding social cognition in autism highlights the need for alternative explanatory approaches. The *predictive coding theory* proposes that the human brain constantly attempts to match incoming sensory input with prior predictions or expectations about the world. The neural processes thereby aim to minimize the difference between predicted and actual input (i.e., prediction errors) by updating priors (Clark, 2013). In the context of social interaction, this means that others' actions can be predicted from mental states ascribed based on available prior information about the acting person, the corresponding situation, and/or people in general. Thus, mental states of others can be actively predicted (Koster-Hale & Saxe, 2013). Following this hypothesis, social difficulties in autism may arise from weakened social cognitive predictions and a stronger reliance on sensory input, leading to a more accurate —but less experience-modulated—perception of the world (Pellicano & Burr 2012). As a consequence, social interactions may appear unpredictable and stressful to autistic individuals, leading to misalignments between their behavior and the expectations of interaction partners (Bolis et al., 2017). Given that Theory of Mind tasks inherently involve prediction and extend beyond immediate sensory input, differences observed in autism can be understood within the predictive coding framework (Sinha et al., 2014). Empirical evidence supports this account: autistic individuals compared to non-autistic individuals show distinct differences in predictive learning and responses, especially when predictive cues are weak or inconsistent (Cannon et al., 2021). While autistic individuals are, for example, able to predict simple action goals of others (Schuwerk & Paulus, 2018), they tend to rely less on prior information and therefore often require more time than non-autistic individuals (Ganglmayer et al., 2020). This has been linked to an expectation of high precision between predicted and observed outcomes, as well as a tendency to overweight prediction errors (Van de Cruys et al., 2014). These theoretical assumptions and behavioral findings raise the question of whether similar predictive differences can also be observed at the neural level.

Recent neuroimaging studies in neurotypical individuals using movie-viewing paradigms provide initial evidence for testing predictive coding accounts at the neural level. Neurotypical adults' prior knowledge of a narrative seems to enable neural anticipation of event patterns. Brain regions within the Theory of Mind network were recruited earlier in time, indicating anticipation of the narrative during repeated exposure (Baldassano et al., 2017; Lee et al., 2021). In a similar paradigm using a movie known to elicit responses in Theory of Mind brain regions (Richardson et al., 2018), Richardson and Saxe (2019) reported a narrative anticipation effect in children, with predictive responses increasing between the ages of 3 and 7 years. This finding suggests that Theory of Mind brain regions are involved in the active prediction of others' mental states during childhood. Moreover, in adults, Theory of Mind regions use current mental state information to predict future social states. Unpredictable sequences of mental states evoke stronger neural responses compared to predictable ones (Thornton et al., 2019). Similarly, unexpected outcomes, compared to expected outcomes, elicit a stronger response in Theory of Mind regions when prior information about an agent's behavior is available (Heil et al., 2019). The magnitude of this effect was inversely related to autistic-like traits, indicating that individuals with more autistic-like traits show a decreased response to unexpected outcomes (Dungan et al., 2016). However, neuroimaging studies examining predictive coding mechanisms in autistic individuals remain limited.

1.5. Summary: Theory of Mind Across Development and Neurodiversity

In neurotypical development, children demonstrate explicit Theory of Mind reasoning, such as understanding of false beliefs, around the age of 4 (Wellman et al., 2001; Wimmer & Perner, 1983). However, already in their first year of life, infants show goal-based action predictions and one year later basic forms of epistemic state-based action predictions, including sensitivity to others' knowledge states. Early evidence comes from violation of expectation paradigms measuring gaze behavior, as well as interactive tasks (Cannon & Woodward, 2012; Liszkowski et al., 2008; Tomasello & Haberl, 2003). Researchers aiming to test early emerging, spontaneous Theory of Mind, applied non-verbal paradigms, including gaze-based violation of expectation (Onishi & Baillargeon, 2005) and anticipatory looking tasks (Southgate et al., 2007) to infants, and found that even infants appear to be sensitive to others' (false) beliefs. However, subsequent replication attempts yielded mixed results (e.g., Barone et al., 2019; Kampis et al., 2021), raising concerns about the robustness of the paradigms used and consequently the interpretability of early implicit/spontaneous Theory of Mind claims.

The development of mental state language, such as terms for knowledge and beliefs, predicts Theory of Mind development (Brooks & Meltzoff, 2015; Olineck & Poulin-Dubois, 2007). Children first use these terms to describe their own mental states and then apply them to others (Gonzales et al., 2018). By ages 2-3, they begin using mental state language (Harris, Yang & Cui, 2017; Shatz et al., 1983), while understanding different degrees of speaker certainty, such as "know" or "think", develops later, by ages 4-5, as shown by verbal explicit measures (Kristen-Antonow et al., 2019; Moore et al., 1989).

In autism, a Theory of Mind deficit was initially proposed (Baron-Cohen, 2000); however, findings from explicit, implicit, and neuroimaging studies have since yielded inconsistent results, challenging the universality of this claim (e.g., Duvall et al., 2023; Gernsbacher & Yergeau, 2019; Moessnang et al., 2020; Nijhof et al., 2018; Senju et al., 2009). Despite this, the mechanisms underlying social difficulties in autism remain insufficiently understood. Predictive coding accounts offer an explanation, suggesting that autistic individuals may use prior information less which leads to weaker social predictions and mismatches during social interaction (Clark, 2013; Sinha et al., 2014). In neurotypicals, predictive coding in Theory of Mind brain regions has been documented in movie viewing paradigms (Baldassano et al., 2017; Lee et al., 2021). In children, such predictive responses increase with age, reflecting growing involvement of these regions in anticipating others' mental states (Richardson & Saxe, 2019).

1.6. Open Questions

Despite significant progress in the field of social cognitive development—particularly in Theory of Mind research—over the past 40 years, several core questions remain unresolved and warrant further empirical investigation.

One key challenge concerns the development and validity of novel eye-tracking methodologies for assessing early social cognitive processes outside traditional laboratories: Can remote webcam-based eye-tracking serve as a valid alternative to traditional in-lab methods for assessing early social cognitive processes in anticipatory looking paradigms? Successfully addressing this question would enhance the recruitment and testing of larger, more diverse samples to identify more generalizable effects, thereby advancing social cognitive research in young children.

Second, among the three main paradigms used to investigate implicit false belief understanding, anticipatory looking has emerged as a promising tool. However, there are differences in the success of replication attempts of earlier studies challenging the robustness, reliability, and replicability of anticipatory looking for Theory of Mind research: Can the anticipatory looking paradigm distinguish between basic epistemic states, such as knowledge and ignorance, in toddlers and adults? Determining whether anticipatory looking captures basic epistemic state-based action anticipation would help to clarify the robustness and interpretive scope for following implicit Theory of Mind studies.

A third unresolved question concerns the early understanding of mental state language. Specifically, it remains unclear whether toddlers can distinguish between varying degrees of speaker certainty as conveyed by epistemic verbs even before this ability can be assessed through explicit verbal tasks: Can toddlers implicitly differentiate between different degrees of speaker (un-)certainty in mental state language? Addressing this question would provide insight into the possible developmental trajectory from implicit to explicit mental state language understanding and contribute to the current Theory of Mind literature.

Finally, in neurodivergent social cognitive development—especially in autism research—key questions remain about the underlying processes driving differences in social communication and interaction, and the role of Theory of Mind: Can neuroimaging paradigms targeting predictive coding in Theory of Mind brain regions be extended to autistic adults to identify possible alterations in predictive processing relative to non-autistic adults? Addressing this question would deepen our understanding of the neural mechanisms underlying social cognition in autism.

2. Exploring Implicit Theory of Mind

As reviewed in the previous section, there is a considerable debate within the scientific community on when children develop a Theory of Mind and how to measure it. Additionally, research on Theory of Mind in autism presents divergent evidence regarding whether and how Theory of Mind reasoning differs from that of neurotypical individuals. To comprehensively explore implicit Theory of Mind and contribute to the ongoing debates in Theory of Mind research, we conducted four studies with differing study focus, age groups, social cognitive development (see Figure 1), methods, settings and contributions (see Figure 2). While Studies 1, 2 and 3 focused on neurotypical social cognitive development, Study 4 examined neurodivergent social cognitive development.

Figure 1. Graph depicting the content of Studies 1 to 4 on three dimensions: Age group (x-axis), study focus (y-axis), and social cognitive development (z-axis).

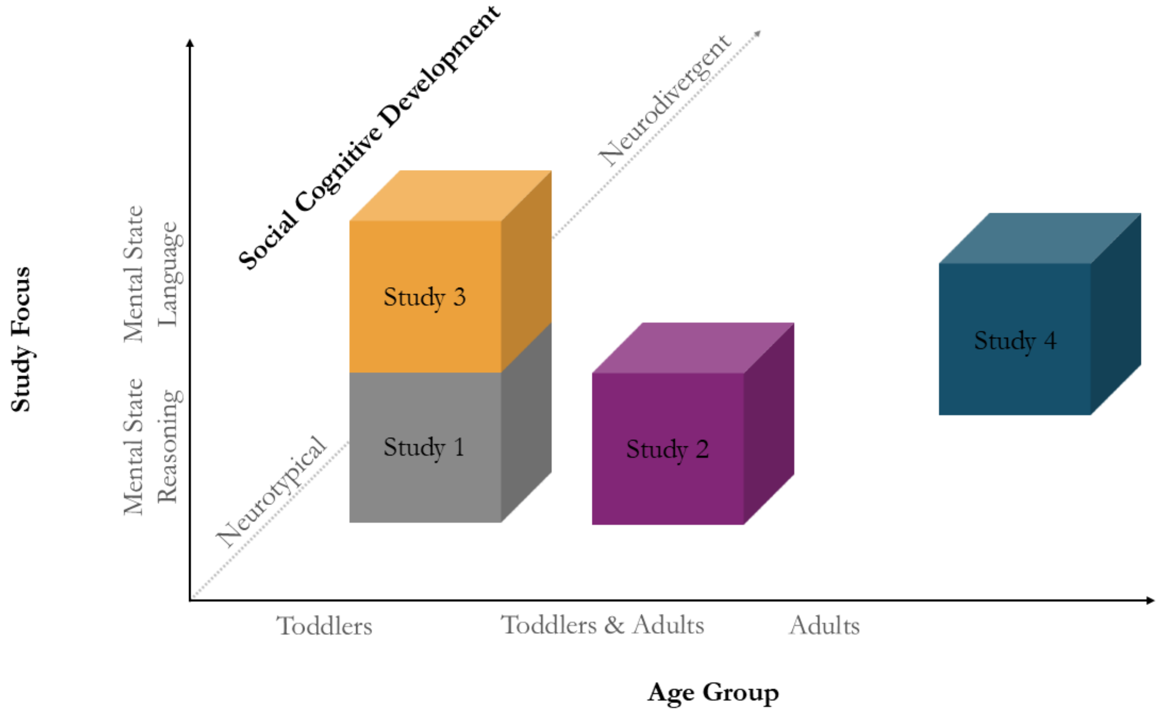
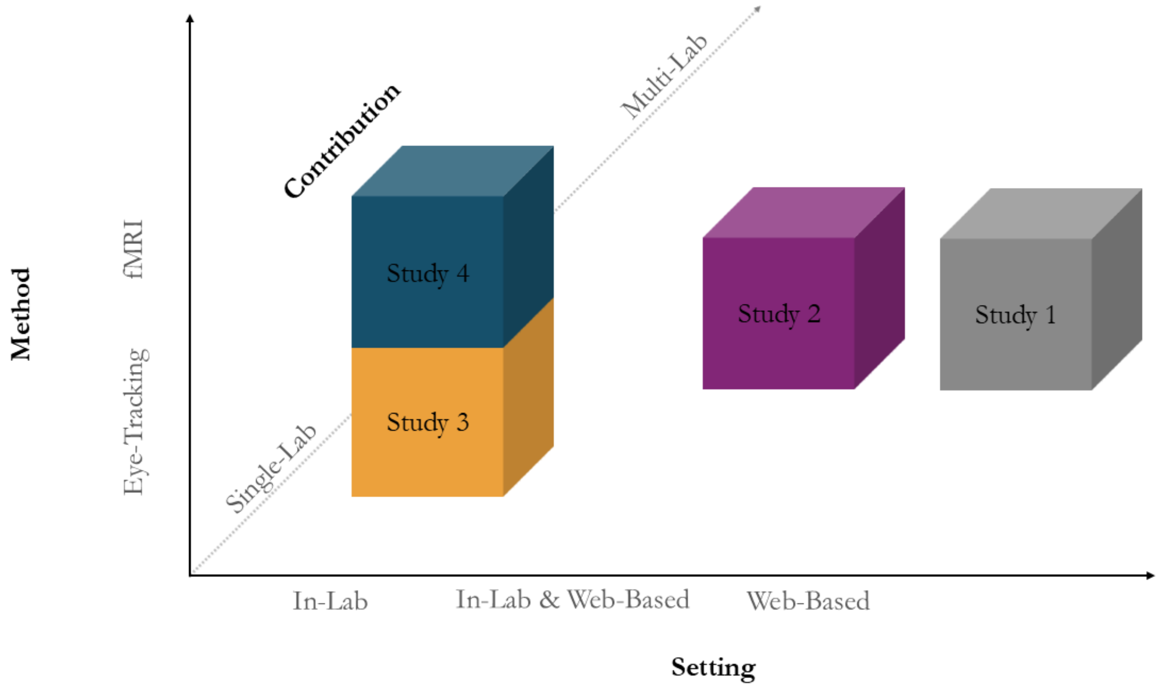


Figure 2. Graph depicting the method of Studies 1 to 4 on three dimensions: Setting (x-axis), method (y-axis), and contribution (z-axis).



In Study 1, we focused on the methodological aspect by testing a novel remote, web-based eye-tracking method in a multi-lab study to capture goal-based action anticipation in toddlers. In Study 2, we built upon the methodological insights gained in Study 1. Thus, in Study 2, we employed an anticipatory looking paradigm and examined toddlers' and adults' basic epistemic state-based action anticipation in a multi-lab study using both web-based and in-lab eye-tracking. In Study 3, we again used eye-tracking in toddlers but this time in a single, in-lab study to implicitly measure their understanding of mental state language. In Study 4, we tested the predictive coding theory in the Theory of Mind network using fMRI in autistic and non-autistic adults.

Below, I summarize the four studies conducted over the course of my dissertation. For additional details and information, please refer to the original manuscripts (see Appendices A-D).

2.1. Neurotypical Social Cognitive Development

2.1.1. Study 1. Web-Based, Remote Eye-Tracking of Goal-Based Action Anticipation in Toddlers

Steffan, A*., **Zimmer, L.***, Arias-Trejo, N., Bohn, M., Dal Ben, R., Flores-Coronado, M. A., Franchin, L., Garbisch, I., Grosse Wiesmann, C., Hamlin, J. K., Havron, N., Hay, J. F., Hermansen, T. K., Jakobsen, K. V., Kalinke, S., Ko, E.-S., Kulke, L., Mayor, J., Meristo, M., … Schuwerk, T. (2023). Validation of an open source, remote web-based eye-tracking method (WebGazer) for research in early childhood. *Infancy*,*79*(1), 31–55. https://doi.org/10.1111/infa.12564 (*shared first authorship)

Eye-tracking technology enables researchers to investigate young children's cognitive processes as they interact with the world. While traditional in-lab setups have been widely used, recent advances introduced automated, web-based applications which are already well established in adult research (e.g., Bogdan et al., 2024; Semmelmann & Weigelt, 2018; Yang & Krajbich, 2021). However, validated paradigms for young children are still lacking. The need for remote eye-tracking has increased as it can be conducted in the family's natural environment, making participation in studies less time-consuming and more comfortable. In addition, it offers a cost-efficient alternative that facilitates multi-lab collaborations and supports the recruitment of larger and more diverse samples. This would be particularly valuable for social cognitive research in children given recent discussions in the field.

In a multi-lab study within the framework of ManyBabies, we evaluated the precision of a novel remote web-based eye-tracking method using WebGazer.js and jsPsych with the webcams of participating families. We tested children aged 18 to 27 months by aiming to replicate an anticipatory looking task from the ManyBabies 2 project (Schuwerk, Kampis et al., 2025)[4], in which goal-based action anticipation was measured while participants watched a hiding game

[4] This task refers to the familiarization trials that were initially tested during the pilot phase of the first study of the ManyBabies 2 project. In the article of Study 1 (Steffan, Zimmer et al., 2024), an earlier version of the Registered Report was cited (Schuwerk, Kampis et al., 2022). This was the current version at the time. Here, in this dissertation, I cite the most recent version of the same study (Schuwerk, Kampis et al., 2025; Study 2 in this dissertation), which reflects the post–data collection stage and is currently under review. Thus, both references refer to the same study at different stages of the Registered Report process.

between two agents. We expected that (1) children in the web-based setting would engage in goal-based action anticipation, indicated by above-chance anticipatory looking toward the location that matches the outcome of the agent's action goal, (2) the eye-tracking method might affect proportional looking scores, however, we were unsure about the direction of the effect, and (3) fewer children would contribute usable data in the remote setting compared to in-lab eye-tracking.

Results of our remotely tested sample of 18-27-month-old toddlers (*N* = 125) revealed that using WebGazer, web-based eye-tracking successfully captured goal-based action anticipations. However, the proportion of goal-based anticipatory looking was lower compared to the in-lab sample (*N* = 70). As expected, exclusion rate was substantially higher in the web-based (42%) than the in-lab sample (10%). To reduce noise in the data, we added an important preprocessing step: we excluded trials based on visual inspection of the match of time-locked gaze coordinates and the participant's webcam video overlayed on the stimuli. In sum, our study demonstrates that remote web-based eye-tracking can be a useful tool for testing toddlers, facilitating recruitment of larger and more diverse samples; a caveat to consider is the larger drop-out rate. For experiments in which the benefits of remote testing are substantial, such as with children, and a reduced spatial resolution can be tolerated, web-based webcam eye-tracking using WebGazer is a promising method.

Lucie Zimmer's contributions to the article (CRediT report): Leading: Formal analysis, Investigation, Project administration, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing; Supporting: Conceptualization, Data curation, Methodology, Resources.

2.1.2. Study 2. Large-Scale, Multi-Lab Eye-Tracking Study on Epistemic State-Based Action Anticipation

Schuwerk, T.*, Kampis, D.*, Alessandroni, N., Altvater-Mackensen, N., Arias-Trejo, N., Axelsson, E. L., Baillargeon, R., Baumann, A.-E., Bernard, C., Biro, S., Blankenship, T. L., Blomberg, I., Bohn, M., Bradford, E. E. F., Byers-Heinlein, K., Canudas Grabolosa, I., Chen, E. M., Chen, X., Corbit, J., … **Zimmer, L.**, … Rakoczy, H. (2025, under review). *Action anticipation based on an agent's epistemic state in toddlers and adults.* [Stage 2 Registered Report, following in-principle acceptance at Stage 1] Child Development. Preprint: https://doi.org/10.31234/osf.io/x4jbm (*shared first authorship)

Recent research, including studies using the anticipatory looking paradigm, challenged assumptions about Theory of Mind abilities by suggesting that even infants can track others' false beliefs (e.g., Southgate et al., 2007). However, recent replication failures cast doubt on the robustness of this paradigm (Barone et al., 2019). To address this, we, the ManyBabies 2 collaboration, conducted a multi-lab, large-scale conceptual replication study of the anticipatory looking paradigm. First, we attempted to replicate goal-based action anticipation that were reported in our pilot study. Second, we tested basic forms of epistemic state-based (knowledge vs. ignorance) action anticipations as a first step to assess the reliability and robustness of spontaneous Theory of Mind measures.

Eighteen- to 27-month-old toddlers and adults watched animations of a hide-and-seek scenario. Participants first saw familiarization trials (as in the pilot study), followed by test trials in which the key manipulation varied depending on whether the chaser had knowledge of the chasee's hiding location or was ignorant of it. Including adults would help to clarify the validity of anticipatory looking measures of Theory of Mind across the lifespan. We expected (1) to replicate the results of the pilot study with participants robustly anticipating the agent's action based on their goal in familiarization trials, and (2) that if participants took the agent's epistemic state into account, they would anticipate the agent's action, accordingly, indicated by differences in their anticipatory looking between the knowledge and ignorance conditions. That is, they were expected to show greater anticipatory looking toward the exit where the chasee was hiding in the knowledge condition, whereas this preference was expected to be absent or reduced in the ignorance condition.

In adults (*N* = 703), we found clear support for both goal-based action anticipation and epistemic state-based action anticipation. Specifically, adults clearly differentiated between knowledge and ignorance as predicted, aligning with epistemic sensitivity. In toddlers (*N* = 521), in contrast, the results were less clear. Toddlers engaged in simple goal-based action anticipation (in both the pilot and main study) but in the main study, they did not show the predicted differentiation between knowledge and ignorance conditions. Instead, they tended to anticipate the agent's action toward the actual location of the chasee in both conditions. However, quantitatively, their proportion of anticipatory looking was substantially higher in the ignorance condition than in the knowledge condition. These unexpected patterns challenge existing assumptions about early Theory of Mind and point to new directions for research on its developmental trajectory.

Lucie Zimmer's contributions to the article (CRediT report): Leading: Investigation, Project administration; Supporting: Data curation, Methodology, Software, Supervision, Writing - review & editing.

2.1.3. Study 3. Mental State Language Understanding in Toddlers

Zimmer, L., Sodian, B., Mani, N., Grosso, S. S., Kristen-Antonow, S., & Schuwerk, T. (2025). Two- to three-year-old toddlers differentiate the epistemic verbs "know" and "think" in a preferential looking eye-tracking paradigm. *Developmental Psychology*. https://doi.org/10.1037/dev0001933

Language plays a crucial role in children's understanding of mental states (Harris et al., 2005). Research has shown that pre-verbal infants already exhibit sensitivity to others' knowledge states, as demonstrated through joint attention and communicative gestures (Moll et al., 2007; Stenberg, 2009). However, while toddlers begin to use epistemic verbs in the third year of life, their comprehension and differentiation of epistemic verbs, that express varying degrees of speaker certainty, such as "know" and "think" remain underexplored. Previous research suggests that children as young as 4 to 5 years begin to verbally distinguish between these verbs (Kristen-Antonow et al., 2019; Moore et al., 1989), leaving open whether an implicit understanding may emerge earlier.

In this study we investigated whether toddlers implicitly differentiate between the epistemic state verbs "know" and "think" using a novel eye-tracking paradigm. A longitudinal design was employed to assess the same children at 27 months and 36 months. In our preferential looking task, toddlers were faced with two animated agents who indicated the location of a hidden object (right vs. left box). A narrator attributed contrasting degrees of certainty to the agents' statement: one expressing certainty ("He knows it is in there.") and the other expressing uncertainty ("He thinks it is in there."). After the introduction of the agents the toddlers were asked about the object's location and preference indicated by their looking time was measured. We expected that (1) a difference between "know" and "think" would already be observable in 27-month-old toddlers during at least part of the response phase and, that (2) 36-month-olds would reliably differentiate between "know" and "think" during the whole response phase.

Results showed that toddlers at both 27 months (*N* = 199) and 36 months (*N* = 131) show systematic differences in their preferential looking in at least one of the response phases. They exhibited a spontaneous preference for the box associated with the agent who "knew" the object's location already before the narrator's question was asked (i.e., pre-questioning phase). Their preference switched in the post-questioning phase; however, this effect was smaller. These findings suggest that children as young as 2 years differentiate between epistemic verbs based on speaker certainty, even before they can explicitly express this distinction. The study contributes to the growing body of evidence that mental state language understanding emerges in early childhood and highlights the importance of implicit measures in assessing young children's cognitive development.

Lucie Zimmer's contributions to the article (CRediT report): Leading: Data curation, Formal analysis, Software, Validation, Visualization, Writing - original draft, Writing - review & editing.

2.2. Neurodivergent Social Cognitive Development

2.2.1. Study 4. Predictive Responses in the Theory of Mind Network in Autism

Zimmer, L., Richardson, H., Pletti, C., Paulus, M., & Schuwerk, T. (2025). Predictive responses in the theory of mind network: A comparison of autistic and non-autistic adults. *Cortex, 187*, 159–171. https://doi.org/10.1016/j.cortex.2025.04.006

The predictive coding theory (Clark, 2013) offers a framework to understand social difficulties experienced in autistic individuals. It suggests that autistic individuals struggle with social interaction (and some Theory of Mind tasks) due to an attenuated use of prior information about others' mental states, making social interactions less predictable and more stressful (Bolis et al., 2017; Pellicano & Burr, 2012; Sinha et al., 2014). Despite theoretical claims, robust and replicable neural differences in Theory of Mind-related brain regions remain elusive.

To test the predictive coding theory in autism, we applied a fMRI paradigm, which was previously tested in neurotypical children, to autistic and non-autistic adults. Participants were presented with a short movie featuring mental state reasoning twice while undergoing fMRI. In two experiments we examined whether non-autistic and autistic adults recruit the Theory of Mind network earlier during the second viewing of a short movie compared to the first (i.e., narrative anticipation effect) and whether this effect was altered in autistic adults. We expected that (1) non-autistic adults would show predictive coding processes, indicated by a narrative anticipation effect (Experiment 1), (2) this effect would be replicable in a new sample of non-autistic adults, and (3) autistic adults would show an attenuation of this effect, in the form of reduced or absent narrative anticipation (Experiment 2).

In contrast to our expectations, we found no evidence for a narrative anticipation effect in Theory of Mind regions in both non-autistic (*N* = 61 in Experiment 1; *N* = 30 in Experiment 2) and autistic adults (*N* = 30 in Experiment 2). These findings may indicate either that a narrative anticipation effect exists, but our task is not sensitive enough to capture it in adults or that adults do not show a narrative anticipation effect when watching this movie. However, in exploratory reverse correlation analyses we identified a scene in both non-autistic samples, but not in the autistic sample. This scene evoked a reduced repetition suppression (i.e., smaller difference in response between first and second viewing) in autistic compared to non-autistic adults. In contrast to other social scenes in the movie, this key scene requires complex reasoning about the false beliefs of the characters. This preliminary subtle difference in processing a complex Theory of Mind scene opens a promising avenue for future research to better understand the nature of differences in social interaction between autistic and non-autistic adults.

Lucie Zimmer's contributions to the article (CRediT report): Leading: Data curation, Formal analysis, Investigation, Project administration, Validation, Visualization, Writing – original draft, Writing - review & editing.

3. General Discussion

The present set of studies was designed to address implicit Theory of Mind across neurodiversity by using novel methodological approaches and targeting core theoretical debates. Below, I first summarize the main findings of each study. I then integrate these results into the broader research context by discussing developmental trajectories and methodological considerations related to implicit Theory of Mind, similarities and differences in Theory of Mind in autism, as well as the studies' strengths, limitations, and implications for future research.

3.1. Summary of Main Findings

First, in a multi-lab study, we examined whether remote, webcam-based eye-tracking using WebGazer can complement traditional in-lab paradigms. We therefore assessed goal-based action anticipation in 18-27-month-old toddlers. The results of Study 1 showed that toddlers tested remotely did engage in goal-based action anticipation measured via anticipatory looking. Although in comparison to in-lab testing, the data quality was lower and the attrition rate was substantially higher, the web-based method proved viable for capturing early social cognitive processes.

Second, in another multi-lab study, we aimed to replicate findings of a pilot study showing goal-based action anticipation in the familiarization trials and investigated whether anticipatory looking also reflects epistemic state-based action anticipation. Specifically, we tested the ability to distinguish between knowledge and ignorance in 18-27-month-old toddlers and adults using both in-lab and the web-based eye-tracking method validated in Study 1. Study 2 replicated findings of the pilot study, revealing that both adults and toddlers engage in goal-based action anticipation in familiarization trials. Further, Study 2 showed that adults clearly distinguished between these epistemic states when predicting others' actions, displaying a preference in the knowledge condition but not in the ignorance condition. In contrast, toddlers' anticipatory looking pattern was unexpected: they anticipated the agent's action in the knowledge condition but showed even stronger anticipation toward the same location in the ignorance condition, in which no preference was expected.

Third, using a preferential looking paradigm, we tested epistemic states with a focus on language understanding. Specifically, we tested whether toddlers implicitly differentiate between epistemic verbs such as "know" and "think" as markers of speaker (un-)certainty. In Study 3, toddlers at both 27 and 36 months, showed a spontaneous preference for agents who were described as knowing the object's location, as indicated by their preferential looking behavior. However, subsequently, after a few seconds, when being explicitly asked about the object's location, children's preference switched.

Fourth, turning to neurodivergent social cognitive development, we investigated whether predictive processing can be measured in the Theory of Mind network in autistic and non-autistic adults and whether it is attenuated in autistic compared to non-autistic adults. We tested this by examining a narrative anticipation effect previously reported in children. No general narrative anticipation effect was observed in Theory of Mind brain regions in both autistic and non-autistic adults. However, exploratory analyses identified a specific scene involving complex mental state reasoning that was processed differently by autistic and non-autistic adults. Specifically, autistic adults (compared to non-autistic adults) showed reduced repetition suppression to this specific scene.

Taken together, these studies demonstrate that remote, web-based eye-tracking is a feasible tool for studying early social cognition, especially in multi-lab collaborations, despite some limitations in data quality. Adults seem to be sensitive to basic forms of epistemic states and goal-based predictions in an anticipatory looking paradigm, whereas spontaneous epistemic state-based action anticipation in 1.5- to 2.5-year-old toddlers appear less robust. Although toddlers accounted for the epistemic status of knowledgeable agents, they did not differentiate their predictions toward ignorant agents. By their third year of life, toddlers show developmental progress in understanding epistemic state verbs, distinguishing varying degrees of speaker certainty. In neurodivergent development, no broad differences in processing mental state narratives were observed between autistic and neurotypical adults. However, subtle differences emerged in predictive processing during complex mental state reasoning, indicating subtle rather than generalized alterations in autism.

3.2. Developmental Trajectory of Implicit Theory of Mind

A discrepancy between early findings on infants' implicit false belief understanding and later replication failures has raised the ongoing, debated question: do young children possess an implicit Theory of Mind? To contribute to this debate, one aim of this dissertation was to investigate more basic forms of Theory of Mind reasoning. This was intended as a systematic approach and as a basis for future research.

Among the abilities thought to precede false belief understanding is goal-based action prediction, which has been associated with later Theory of Mind development (Aschersleben et al., 2008). Infants typically begin to anticipate others' goals within their first year of life, as indicated in predictive gaze shifts (e.g., Cannon & Woodward, 2012; Falck-Ytter et al., 2006). Consistent with these findings, Study 1 and Study 2 demonstrated reliable goal-based action anticipation in 1.5- to 2.5-year-old toddlers. Study 2 further confirmed the robustness of this ability in adults. Thus, the ability to anticipate others' action goals seems to emerge early in development and remains robust in adulthood.

Beyond goal-based predictions, children in their second year of life begin to show sensitivity to others' basic epistemic states such as knowledge (Stenberg, 2009; Tomasello & Haberl, 2003). They show awareness of what others know or do not know (Dunham et al., 2000; Liszkowski et al., 2008; O'Neill, 1996). These early forms of epistemic understanding are considered more fundamental than belief representations (Phillips et al., 2020). Building on this line of research, Study 2 examined whether children and adults use others' epistemic access—specifically, whether others are knowledgeable or ignorant—to predict their actions. Adults reliably anticipated an agent's action in accordance with their epistemic state. This is consistent with findings on false belief reasoning in adulthood (Schneider et al., 2017). Unexpectedly, toddlers, did not show the expected differentiation between knowledge and ignorance in Study 2 (as described in section 3.1.). The observed pattern appears independent of both age and (prior) goal-based action anticipation performance. Hence, the results of Study 2 may indicate that toddlers either cannot yet differentiate between knowledge and ignorance or have interpreted the task differently. Thus, the stimuli may not have clearly isolated the epistemic contrast, due to either timing differences between conditions or high cognitive demands. Given empirical evidence reviewed above (Dunham et al., 2000; Liszkowski et al., 2008; O'Neill, 1996), a stimulus-related explanation seems plausible. This issue should be more thoroughly addressed (see 3.6. Limitations and Future Directions) before drawing firm conclusions about toddlers' ability of epistemic state-based action prediction. However, if stimulus-related factors can be ruled out, the findings of Study 2 may further challenge the robustness of claims regarding early, implicit false belief understanding (Onishi & Baillargeon, 2005; Southgate et al., 2007). Large-scale replication efforts, such as those planned within the ManyBabies 2 project, are needed to determine when and how epistemic state reasoning emerges in early development.

Children's early sensitivity to epistemic states is closely linked to their developing use and understanding of mental state language, offering an informative perspective on the emergence of implicit Theory of Mind. A developmental pattern has been proposed in which early productive use of epistemic verbs between ages 2 and 3 is linked to later understanding of these verbs (Moore et al., 1994) as well as Theory of Mind performance (Brooks & Meltzoff, 2015; Olineck & Poulin-Dubois, 2007). Verbal expressions of own knowledge and ignorance typically emerge between 2.5 and 3 years (Harris, Ronfard & Bartz, 2017). An explicit understanding and differentiation of epistemic verbs such as know and think typically develops between the ages of 4 and 5 (Kristen-Antonow et al., 2019; Moore et al., 1989). What remained unclear, however, was whether implicit understanding of epistemic verbs begins to emerge between the early usage around age 2–3 and the explicit understanding typically observed by age 4–5. To address this gap, Study 3 assessed toddlers' implicit understanding of epistemic verbs. Study 3 revealed that children in their third year of life are already sensitive to differences in speaker certainty. At both 27 and 36 months of age, children spontaneously preferred agents described as knowledgeable over those described as uncertain. Thus, at the age when toddlers begin using epistemic verbs to express their own—and to a lesser extent others' — certainty in social interactions (Harris, Yang & Cui, 2017), they already appear to understand varying degrees of certainty in others (Study 3). More broadly, results of Study 3 support the view that implicit understanding can precede explicit understanding (Clements & Perner, 1994; Kloo et al., 2020). Moreover, this developmental trajectory is further refined by cross-linguistic evidence. Japanese-speaking children, for example, show sensitivity to sentence-final particles. As early as age 3, they prefer those speakers who use a language-specific grammatical marker expressing epistemic certainty over those expressing uncertainty (Matsui et al., 2016), mirroring findings of Study 3. In contrast, less frequently used lexical markers such as "know" and "think" are understood later—underscoring the importance of language-specific grammatical features in the development of mental state language understanding.

To conclude, the results of this dissertation trace a developmental trajectory in which robust goal-based action prediction emerges early (Study 1; Study 2), while basic forms of epistemic state sensitivity remain unclear in toddlers but are well established in adults (Study 2), with first evidence of mental state language understanding emerging in toddlerhood and preceding explicit one (Study 3).

3.3. Methodological Considerations: Eye-Tracking in Theory of Mind Research

In past decades, researchers commonly used verbal-response tasks to study children's understanding of epistemic states as well as epistemic state language (Moore et al., 1989; Perner et al., 1987; Wimmer & Perner, 1983). While these paradigms provided valuable insights into cognitive development, their reliance on advanced language skills limits their applicability for investigating mental state reasoning in younger or pre-verbal children.

To overcome these limitations, recent research has increasingly turned to non-verbal methods. Study 3 contributes to this methodological shift by employing a preferential looking eye-tracking paradigm to assess toddlers' understanding of epistemic verbs at an age at which they typically fail explicit tasks. Within this paradigm children differentiate between epistemic verbs around the age of 3, establishing a spontaneous preference. Interestingly, this preference switched after a narrator's prompt. This unexpected pattern may indicate that after the first trial children anticipated the narrator's question. Consequently, they responded pre-emptively in the following trials. The switch in preference may reflect an exploratory re-evaluation due to curiosity, uncertainty due to lack of feedback or attentional dynamics such as inhibition of return, which is the tendency for a counteractive response to occur following an initial preference (Klein, 1988; Klein & MacInnes, 1999). These factors should be considered in future studies to refine preferential looking paradigms and thus for instance examine mental state language understanding more precisely (see 3.6. Limitations and Future Directions).

In addition, anticipatory looking paradigms have become increasingly important in developmental psychology research, particularly for studying expectations or predictions spontaneously. In recent goal-based action anticipation studies, it was questioned whether the observed predictive gaze shifts reflect another person's goal understanding or mere movement path anticipation (Ganglmayer et al., 2019; Gönül et al., 2024). However, studies showed that when participants are provided with sufficient information to learn about the agent's goal, they prioritize goals over movement paths (Ganglmayer et al., 2020; Paulus et al., 2017). Similarly, the hide-and-seek scenario used in Study 1 and Study 2 minimized path learning by randomizing the agent's starting positions and goal locations across trials.

Moreover, previous studies using anticipatory looking eye-tracking paradigms in false belief research reported difficulties in consistently eliciting reliable goal-based action anticipation during familiarization trials (e.g., Grosse Wiesmann et al., 2018; Kampis et al., 2021; Kulke, Reiß et al., 2018; Schuwerk et al., 2018). In contrast, Study 1 and Study 2, demonstrated that spontaneous goal-based action anticipation can be robustly measured in both toddlers and adults using newly developed, engaging stimuli within an anticipatory looking paradigm. Thus, these stimuli offer a viable method for establishing predictions of action goals in participants across different age groups and cultural contexts. Moreover, earlier studies applied strict exclusion criteria based on performance in familiarization trials (e.g., Senju et al., 2009; Southgate et al., 2007). However, this approach has been challenged by findings showing no significant performance differences between children who passed or failed such criteria (Grosse Wiesmann et al., 2018). Consistent with this, Study 2 found no differences in epistemic state-based action anticipation based on performance in familiarization trials. The anticipatory looking paradigm further proved effective in measuring a basic form of epistemic state-based action anticipation in adults in Study 2. However, the unexpected results observed in toddlers in Study 2 (as already discussed in section 3.2.) preclude definitive conclusions regarding applicability of anticipatory looking paradigm in studying this epistemic state-based action anticipation in young childhood.

Together, this dissertation demonstrates that anticipatory and preferential looking paradigms offer reliable, age-appropriate methods for capturing children's implicit understanding of mental state language (Study 3) and others' action goals (Study 1; Study 2). In contrast, the suitability of anticipatory looking paradigms in assessing early sensitivity to epistemic states in toddlers remains uncertain and needs further investigation (Study 2).

Moreover, recent advances in eye-tracking opened new possibilities for conducting studies remotely using web-based methods. However, these approaches are still in the early stages of validation within cognitive developmental psychology. Previous online studies successfully replicated developmental patterns observed in in-lab research, such as false belief understanding (Schidelko et al., 2021). Similarly, comparable looking behaviors were reported in infants using traditional paradigms like preferential looking (Bánki et al., 2022; Smith-Flores et al., 2021) and violation of expectation tasks (Raz et al., 2024). Building on this work, Study 1 replicated goal-based action anticipation in toddlers within an anticipatory looking paradigm,using a novel remote, web-based setup. As an extension to moderated studies (e.g., Schidelko et al., 2021; Smith-Flores et al., 2021; Study 1), unmoderated studies using iCatcher+ replicated similar effects with slightly reduced effect sizes and emphasized the need to adapt tasks for remote assessment of early social cognition (Raz et al., 2024; Tenenbaum et al., 2025). These findings are further corroborated by a recent meta-analysis. This meta-analysis reports that web-based developmental studies, regardless of whether they are conducted under moderated or unmoderated conditions, produce results largely comparable to those obtained in laboratory settings, albeit with somewhat reduced effect sizes (Chuey et al., 2024).

To date, WebGazer has not been applied to infant or toddler studies, whereas deep learning-based tools like iCatcher+ are more commonly used in child research (Raz et al., 2024; Tenenbaum et al., 2025). Compared to iCatcher+, which classifies gaze into discrete categories, WebGazer estimates continuous gaze coordinates, offering greater flexibility for capturing dynamic gaze behavior. In adult studies, WebGazer has proven feasible, although it yields lower data quality and effect sizes compared to traditional commercial in-lab systems (Bogdan et al., 2024; Semmelmann & Weigelt, 2018; Yang & Krajbich, 2021). Specifically, in anticipatory looking paradigms, WebGazer has been shown to reliably detect broad gaze patterns, while its sensitivity to brief or subtle effects may be limited (Slim et al., 2024). Study 1 and Study 2 provide further evidence that WebGazer can successfully capture broad gaze patterns in toddlers using an anticipatory looking paradigm. While in Study 1 the effect size was lower than in the in-lab sample—likely due to reduced spatial resolution and sampling rate—the key anticipatory effects were still detectable. Notably, web-based samples usually show higher attrition rates than samples tested in the laboratory (Bánki et al., 2022; Semmelmann & Weigelt, 2018; Study 1; Yang & Krajbich, 2021). However, Study 1's successful replication of in-lab findings under these constraints supports the validity of WebGazer as a tool for remote cognitive developmental research.

In sum, this dissertation demonstrates that web-based eye-tracking using WebGazer can effectively complement in-lab approaches (Study 1) and offers a promising tool for testing toddlers' anticipatory looking (Study 1; Study 2), particularly when reduced spatial resolution is acceptable.

3.4. Theory of Mind in Autism: Similarities and Differences

In neurodivergent social cognitive development, such as in autism, individuals often experience social interactions as challenging. These challenges have frequently been linked to differences in Theory of Mind processing, though some traditional accounts frame these as deficits. However, consistent findings remain elusive. Predictive coding accounts offer a potential explanation. According to this view, autistic individuals may form less flexible predictions which can lead them to perceive social interactions as unpredictable and, consequently, cause interactive mismatches (Pellicano & Burr, 2012; Sinha et al., 2014). Thus, predictive coding accounts in autism suggest that difficulties in social interaction may arise not from a lack of Theory of Mind, but from altered predictive processing of social information such as mental states. Neural predictive Theory of Mind processing has been observed in neurotypical adults (Baldassano et al., 2017; Lee et al., 2021) and children (Richardson & Saxe, 2019). Unlike these findings, Study 4 found no evidence for narrative anticipation in the Theory of Mind network in adults. Importantly, alternative explanations for this null finding were ruled out: neither faster nor more localized neural anticipation were observed. Additionally, no group difference emerged between autistic and non-autistic adults. In both autistic and non-autistic adults, Study 4 revealed comparable processing of movie stimuli across both viewings. Like Study 4, highly similar neural responses in autistic and non-autistic adults were documented during movie-watching (Mangnus et al., 2024). Further, large-scale studies found no measurable differences in Theory of Mind network activation between autistic and non-autistic adults (Dufour et al., 2013; Moessnang et al., 2020). However, findings of Study 4 appear to contrast with studies reporting altered activation in core Theory of Mind regions during social cognitive tasks in autism (e.g., Ciaramidaro et al., 2015; Duvall et al., 2023; Kana et al., 2009; Kana et al., 2014; Nijhof et al., 2018; Sugranyes et al., 2011). A key difference lies in the methodological approach: while those studies used more structured tasks, the paradigm in Study 4 involved a naturalistic third-person narrative. This probably facilitated the investigation of neural responses under more ecologically valid conditions. In sum, from a theoretical perspective, recent research (Dufour et al., 2013; Gernsbacher & Yergeau, 2019; Mangnus et al., 2024; Moessnang et al., 2020; Study 4) challenges the view that a circumscribed and profound Theory of Mind deficit underlies interaction and communication difficulties in autistic adults.

Recently, it was reported that differences in autism may only emerge when Theory of Mind reasoning becomes demanding (Schuwerk & Sodian, 2023). Similarly, in Study 4, a scene involving complex false belief reasoning elicited reduced repetition suppression in autistic compared to non-autistic adults. Repetition suppression, which is the attenuation of neural responses upon repeated exposure, is shaped by expectations: expected repetitions elicit stronger suppression than unexpected ones (Summerfield et al., 2008). In line with this, neurotypical individuals typically show heightened neural responses to unpredictable mental states in comparison to predictable ones (Heil et al., 2019; Thornton et al., 2019). In Study 4, neurotypical adults may have benefit from repeated viewing, leading to higher predictability and lower neural response. In contrast, as proposed by predictive coding theories, autistic individuals may experience difficulties when using prior information (Pellicano & Burr, 2012). Potentially, the similar neural responses upon repetition observed in autistic adults in Study 4 may suggest reduced flexibility in updating predictions and sustained cognitive effort. Notably, Study 4 only found exploratory evidence which is not supported by the main analysis. This may either indicate a subtle difference or no difference, thereby highlighting the need for future research to better understand the nature of this potential alteration.

Taken together, the findings of this dissertation reveal an absence of overall predictive processing in Theory of Mind brain regions, indicating comparable processing patterns in autistic and non-autistic individuals (Study 4). Subtle differences in predictive processing in the Theory of Mind network appear to be content-specific and emerge during more complex Theory of Mind reasoning (Study 4).

3.5. Strengthening Theory of Mind Research: Replication and Collaboration

In recent years concerns about the replicability of key findings in developmental cognitive science have increased, as for instance in the domain of implicit Theory of Mind research. Several replication failures raised doubts about the robustness of early findings, contributing to what has been termed a replication crisis (Barone et al., 2019; Kulke & Rakoczy, 2017; Poulin-Dubois et al., 2018). Study 1, Study 2, and Study 4 addressed this issue by emphasizing replications. Study 1 replicated the familiarization trials from the pilot of Study 2. The main study of Study 2 provided a second confirmation of their reliability. These multiple replications demonstrate that our paradigm robustly elicits goal-based anticipatory looking responses, addressing previous challenges in achieving consistent participant anticipation (e.g., Kampis et al., 2021; Kulke, Reiß et al., 2018).

Moreover, Study 4 represents a failed extension attempt. The narrative anticipation effect previously reported in children by Richardson and Saxe (2019) could not be replicated in an adult sample, even though the identical paradigm, stimuli, and analysis procedures as the original study were employed. Such null findings and their transparent reporting, particularly when testing the generalizability of prior effects, are equally important for refining the scope and boundaries of theoretical claims (Heene & Ferguson, 2017). It aligns with a current perspective that views the so-called replication crisis not as a failure, but as a driver of structural and cultural reform in science. From this viewpoint, the transparent reporting of null results contributes to the ongoing credibility revolution by challenging publication bias and promoting theoretical refinement (Korbmacher et al., 2023).

Another promising response to the replication crisis has been the rise of multi-lab collaboration studies. Study 1 and Study 2 followed a coordinated multi-lab design, enhancing the robustness and generalizability of findings. Samples within multi-lab studies are more diverse than those in most single-lab, in-person developmental studies (Singh et al., 2023; Study 1; Study 2), highlighting that web-based methods offer the opportunity to test demographically diverse, large international samples under comparable conditions. However, samples in Study 1 and Study 2 were still predominantly WEIRD: in Study 1, only 11.2% of the data included came from participants outside Western Europe or North America. Notably, the precondition to possess or have access to a computer already excludes large portions of the world's population. In Study 2, which was a larger multi-lab study than Study 1, 16.3% of the included data came from labs outside Western Europe, North America, or Australia (with similar percentages across age groups). These proportions are comparable to those reported in other multi-lab studies, such as for instance ManyBabies 1 (Byers-Heinlein et al., 2020) or Many Labs 2 (Schimmelpfennig et al., 2025). New methodological approaches can help select samples exhibiting significant cultural variation (e.g., Muthukrishna et al., 2020). Yet, adult studies show that samples are often homogeneous even within countries, masking important diversity (Ghai et al., 2024). Therefore, for future multi-lab studies, recruiting representative and heterogeneous samples within countries while carefully considering diverse cultural contexts remains crucial.

In sum, this dissertation highlights the value of replication (Study 1; Study 2; Study 4) and collaboration (Study 1; Study 2) for more diverse and transparent Theory of Mind research, supporting the view that the replication crisis serves as a catalyst for an enduring reform in social cognitive psychology.

3.6. Limitations and Future Directions

In Study 1, we employed a visual inspection process as a preprocessing step to efficiently reduce data noise and successfully applied the same approach in Study 2. We created overlays of time-locked gaze coordinates alongside the participant's webcam video overlaid on the stimuli. Unlike detailed manual coding, this method involved holistic and rapid evaluation. For instance, coders could determine at a glance whether gaze data were clearly invalid, which greatly improved efficiency. As a next step, developing an automated pre-selection method to flag trials for potential exclusion would help to further minimize the amount of video data requiring visual inspection. Moreover, future research can benefit from adult datasets that simulate infant webcam noise. Training algorithms using these datasets can improve the quality and robustness of automated gaze tracking in our WebGazer setup and address high dropout rates. This has already been successfully demonstrated with tools like OWLET and iCatcher+ (Hagihara et al., 2024).

Following adults' successful epistemic-based anticipatory looking in Study 2, the next step in the ManyBabies 2 project is to investigate whether the anticipatory looking paradigm can reliably capture adults' spontaneous true/false belief-based action predictions. Proceeding with this planned true/false belief understanding study in children would not be meaningful at this stage. Instead, future studies should first address stimulus-based explanations for the unexpected pattern of results by aligning timing across conditions and reducing processing demands by for instance minimizing the chaser's back-and-forth movement in the knowledge condition. Additionally, it should be ensured that the bear is no longer visible in the ignorance condition (e.g., through the use of occluders), since this might have confused/distracted children. Second, other measures such as looking behavior (e.g., Daum et al., 2012) or pupil dilation (as currently done in a ManyBabies 2 spin-off study) should be used to thoroughly investigate this basic form of epistemic state-based action prediction. Third, older children should be tested to ensure that the toddlers in our study were not simply too young to show this basic form of epistemic state-based action prediction.

Regarding Study 3, future research should shorten the test phase and reduce the number of trials to better accommodate children's limited attention span. Specifically, posing the narrator's prompt immediately after the introduction of the agent's degree of (un-)certainty and limiting the display time the objects, could promote more spontaneous and naturalistic responses. These developmentally appropriate conditions are expected to improve attentiveness and lower the high trial exclusion rates observed in this study. Subsequently, with these adjustments, mental state language understanding using our paradigm should be tested again within the age group of our study to assess whether the preferential looking patterns can be reliably replicated. In addition, this new task procedure should also be validated in older children and/or adults who already have an advanced understanding of epistemic verbs.

In Study 4, we did not observe a narrative anticipation effect in adults, despite using the same paradigm as a previous study that showed this effect increases with age in children (Richardson & Saxe, 2019). Future research should consider employing age-appropriate stimuli for adult samples (e.g., Baldassano et al., 2017; Lee et al., 2021) to gain a deeper understanding of the development of predictive processing and how it may differ across age groups and neurodiverse populations. Study 4 enhances ecological validity of structured Theory of Mind tasks by using naturalistic, emotionally rich movie-stimuli from a third-person perspective (Sonkusare et al., 2019). A valuable next step would be to integrate second-person paradigms, which better capture the interactive nature of social cognition (Bolis et al., 2023). Rather than viewing autism solely as a disordered function within individual brains, these approaches emphasize dynamic mismatches in neurodivergent–neurotypical interactions. Focusing on the interpersonal level moves beyond binary group comparisons and helps challenge neurocentric biases (Bolis et al., 2017).

3.7. Conclusion

This work advances the understanding of implicit Theory of Mind by integrating developmental insights, methodological innovation, and neurodivergent perspectives. Across studies, we demonstrated that web-based remote eye-tracking proved feasible for testing toddlers' action anticipation, enabling more diverse and scalable research despite technical limitations. Within multi-lab frameworks we confirmed that goal-based predictions can be reliably measured within an anticipatory looking paradigm in both toddlers and adults. While adults' anticipatory looking also reflected epistemic sensitivity, toddlers did not consistently differentiate between knowledge and ignorance, underscoring the need for further refinement of our paradigm. At the same time, toddlers spontaneously distinguished epistemic verbs, suggesting that an implicit sensitivity to speaker certainty emerges well before explicit verbal mastery. Finally, in an adult neuroimaging study we found no evidence of predictive Theory of Mind processing in autism. However, we revealed preliminary subtle group differences when complex belief reasoning is required, pointing to nuanced effects consistent with predictive coding accounts.

Together, the findings suggest that implicit Theory of Mind does not emerge as a single unified ability, but rather in a piecemeal fashion, with different components—such as action prediction, language understanding, and neural processing—following distinct but related developmental and neurocognitive trajectories. This fragmentation may reflect the diversity of social cognitive mechanisms involved in mental state understanding. For research, it underscores the importance of integrating neurodiversity and employing multiple methods to capture this complexity. Thus, this work outlines a developmental and neurodiverse pathway of implicit Theory of Mind and offers a foundation for more robust, inclusive, and theoretically grounded research in social cognition.

4. References

Abbeduto, L., & Rosenberg, S. (1985). Children's knowledge of the presuppositions of know and other cognitive verbs. *Journal of Child Language, 12*(3), 621–641. https://doi.org/10.1017/s0305000900006693

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed.). American Psychiatric Publishing.

American Psychological Association. (2018). *Functional magnetic resonance imaging.* APA Dictionary of Psychology. https://dictionary.apa.org/functional-magnetic-resonance-imaging

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*(4), 953–970. https://doi.org/10.1037/a0016923

Aschersleben, G., Hofer, T., & Jovanovic, B. (2008). The link between infant attention to goal-directed action and later theory of mind abilities. *Developmental Science, 11*(6), 862–868. https://doi.org/10.1111/j.1467-7687.2008.00736.x

Aslin, R. N. (2007). What's in a look? *Developmental Science, 10*(1), 48–53. https://doi.org/10.1111/j.1467-7687.2007.00563.x

Baillargeon, R., Buttelmann, D., & Southgate, V. (2018). Invited commentary: Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. *Cognitive Development, 46*, 112–124. https://doi.org/10.1016/j.cogdev.2018.06.001

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron, 95*(3), 709–721. https://doi.org/10.1016/j.neuron.2017.06.041

Bánki, A., de Eccher, M., Falschlehner, L., Hoehl, S., & Markova, G. (2022). Comparing online webcam- and laboratory-based eye-tracking for the assessment of infants' audio-visual synchrony perception. *Frontiers in Psychology*, *12*, Article 733933. https://doi.org/10.3389/fpsyg.2021.733933

Baron-Cohen, S. (2000). Theory of mind and autism: A review. *International Review of Research in Mental Retardation, 23,* 169–184. https://doi.org/10.1016/S0074-7750(00)80010-5

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*(1), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). 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*(2), 241–251. https://doi.org/10.1111/1469-7610.00715

Barone, P., Corradi, G., & Gomila, A. (2019). Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis. *Infant Behavior & Development*, *57*, Article 101350. https://doi.org/10.1016/j.infbeh.2019.101350

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind*. Oxford University Press.

Bennett, J. (1978). Some remarks about concepts. *Behavioral and Brain Sciences, 1*(4), 557–560. https://doi.org/10.1017/S0140525X00076573

Bogdan, P. C., Dolcos, S., Buetti, S., Lleras, A., & Dolcos, F. (2024). Investigating the suitability of online eye tracking for psychological research: Evidence from comparisons with in-person data using emotion–attention interaction tasks. *Behavior Research Methods, 56*, 2213–2226. https://doi.org/10.3758/s13428-023-02143-z

Bolis, D., Balsters, J., Wenderoth, N., Becchio, C., & Schilbach, L. (2017). Beyond autism: Introducing the dialectical misattunement hypothesis and a bayesian account of intersubjectivity. *Psychopathology*, *50*(6), 355–372. https://doi.org/10.1159/000484353

Bolis, D., Dumas, G., & Schilbach, L. (2023). Interpersonal attunement in social interactions: From collective psychophysiology to inter-personalized psychiatry and beyond. *Philosophical Transactions of the Royal Society B, 378*(1870), Article 20210365. https://doi.org/10.1098/rstb.2021.0365

Booth, J. R., Hall, W. S., Robison, G. C., & Kim, S. Y. (1997). Acquisition of the mental state verb *know* by 2- to 5-year-old children. *Journal of Psycholinguistic Research, 26*(6), 593–609. https://doi.org/10.1023/A:1025093906884

Bowler D. M. (1992). "Theory of mind" in Asperger's syndrome. *Journal of Child Psychology and Psychiatry*, *33*(5), 877–893. https://doi.org/10.1111/j.1469-7610.1992.tb01962.x

Brandone, A. C., Horwitz, S. R., Aslin, R. N., & Wellman, H. M. (2014). Infants' goal anticipation during failed and successful reaching actions. *Developmental Science*, *17*(1), 23–34. https://doi.org/10.1111/desc.12095

Bretherton, I., & Beeghly, M. (1982). Talking about internal states: The acquisition of an explicit theory of mind. *Developmental Psychology, 18*(6), 906–921. https://doi.org/10.1037/0012-1649.18.6.906

Brooks, R., & Meltzoff, A. N. (2015). Connecting the dots from infancy to childhood: A longitudinal study connecting gaze following, language, and explicit theory of mind. *Journal of Experimental Child Psychology, 130*, 67–78. https://doi.org/10.1016/j.jecp.2014.09.010

Burnside, K., Ruel, A., Azar, N., & Poulin-Dubois, D. (2018). Implicit false belief across the lifespan: Non-replication of an anticipatory looking task. *Cognitive Development, 46,* 4–11. https://doi.org/10.1016/j.cogdev.2017.08.006

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition, 112*(2), 337–342. https://doi.org/10.1016/j.cognition.2009.05.006

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language, 28*(5), 606–637. https://doi.org/10.1111/mila.12036

Byers-Heinlein, K., Bergmann, C., Davies, C., Frank, M. C., Hamlin, J. K., Kline, M., Kominsky, J. F., Kosie, J. E., Lew-Williams, C., Liu, L., Mastroberardino, M., Singh, L., Waddell, C. P. G., Zettersten, M., & Soderstrom, M. (2020). Building a collaborative psychological science: Lessons learned from ManyBabies 1. *Canadian Psychology, 61*(4), 349–363. https://doi.org/10.1037/cap0000216

Cannon, E. N., & Woodward, A. L. (2012). Infants generate goal-based action predictions. *Developmental Science, 15*(2), 292–298. https://doi.org/10.1111/j.1467-7687.2011.01127.x

Cannon, J., O'Brien, A. M., Bungert, L., & Sinha, P. (2021). Prediction in autism spectrum disorder: A systematic review of empirical evidence. *Autism Research, 14*(4), 604–630. https://doi.org/10.1002/aur.2482

Carpenter, M., Call, J., & Tomasello, M. (2005). Twelve-and 18-month-olds copy actions in terms of goals. *Developmental Science, 8*(1), F13–F20. https://doi.org/10.1111/j.1467-7687.2004.00385.x

Carruthers, P. (2013). Mindreading in infancy. *Mind & Language, 28*(2), 141–172. https://doi.org/10.1111/mila.12014

Ciaramidaro, A., Bölte, S., Schlitt, S., Hainz, D., Poustka, F., Weber, B., Bara, B. G., Freitag, C., & Walter, H. (2015). Schizophrenia and autism as contrasting minds: Neural evidence for the hypo-hyper-intentionality hypothesis. *Schizophrenia Bulletin*, *41*(1), 171–179. https://doi.org/10.1093/schbul/sbu124

Chuey, A., Boyce, V., Cao, A., & Frank, M. C. (2024). Conducting developmental research online vs. in-person: A meta-analysis. *Open Mind*, *8*, 795–808. https://doi.org/10.1162/opmi_a_00147

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development, 9*(4), 377–395. https://doi.org/10.1016/0885-2014(94)90012-4

Daum, M. M., Attig, M., Gunawan, R., Prinz, W., & Gredebäck, G. (2012). Actions seen through babies' eyes: A dissociation between looking time and predictive gaze. *Frontiers in Cognition, 3*, Article 370. https://doi.org/10.3389/fpsyg.2012.00370

Dennett, D. C. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences, 1*(4), 568–570. https://doi.org/10.1017/S0140525X00076664

Deschrijver, E., Bardi, L., Wiersema, J. R., & Brass, M. (2016). Behavioral measures of implicit theory of mind in adults with high functioning autism. *Cognitive Neuroscience*, *7*(1–4), 192–202. https://doi.org/10.1080/17588928.2015.1085375

Devine, R. T., & Hughes, C. (2014). Relations between false belief understanding and executive function in early childhood: A meta-analysis. *Child Development, 85*(5), 1777–1794. https://doi.org/10.1111/cdev.12237

Dickstein, D. P., Pescosolido, M. F., Reidy, B. L., Galvan, T., Kim, K. L., Seymour, K. E., Laird, A. R., Di Martino, A., & Barrett, R. P. (2013). Developmental meta-analysis of the functional neural correlates of autism spectrum disorders. *Journal of the American Academy of Child and Adolescent Psychiatry*, *52*(3), 279–289, Article e16. https://doi.org/10.1016/j.jaac.2012.12.012

Dilling, H., Mombour, W., Schmidt, M. H., Schulte-Markwort, E., Remschmidt, H., & Weltgesundheitsorganisation (Hrsg.). (2015). *Internationale Klassifikation psychischer Störungen: ICD-10 Kapitel V (F) klinisch-diagnostische Leitlinien* (10. Aufl.). Hogrefe.

Dörrenberg, S., Rakoczy, H., & Liszkowski, U. (2018). How (not) to measure infant theory of mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development, 46,* 12–30. https://doi.org/10.1016/j.cogdev.2018.01.001

Dudley, R., Orita, N., Hacquard, V., & Lidz, J. (2015). Three-year-olds' understanding of *know* and *think*. In F. Schwarz (Ed.), *Experimental perspectives on presuppositions* (Vol. 45, pp. 241–262). Springer. https://doi.org/10.1007/978-3-319-07980-6_11

Dufour, N., Redcay, E., Young, L., Mavros, P. L., Moran, J. M., Triantafyllou, C., Gabrieli, J. D., & Saxe, R. (2013). Similar brain activation during false belief tasks in a large sample of adults with and without autism. *PLoS ONE, 8*(9), Article e75468. https://doi.org/10.1371/journal.pone.0075468

Dungan, J. A., Stepanovic, M., & Young, L. (2016). Theory of mind for processing unexpected events across contexts. *Social Cognitive and Affective Neuroscience*, *11*(8), 1183–1192. https://doi.org/10.1093/scan/nsw032

Dunham, P., Dunham, F., & O'Keefe, C. (2000), Two-year-olds' sensitivity to a parent's knowledge state: Mind reading or contextual cues?. *British Journal of Developmental Psychology, 18*(4), 519–532. https://doi.org/10.1348/026151000165832

Duvall, L., May, K. E., Waltz, A., & Kana, R. K. (2023). The neurobiological map of theory of mind and pragmatic communication in autism. *Social Neuroscience*, *18*(4), 191–204. https://doi.org/10.1080/17470919.2023.2242095

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., Kessler, J., Woike, J. K., Wolf, O. T., & Convit, A. (2006). Introducing MASC: A movie for the assessment of social cognition. *Journal of Autism and Developmental Disorders, 36*(5), 623–636. https://doi.org/10.1007/s10803-006-0107-0

Elsner, B., & Adam, M. (2021). Infants' goal prediction for simple action events: The role of experience and agency cues. *Topics in Cognitive Science*, *13*(1), 45–62. https://doi.org/10.1111/tops.12494

Erel, Y., Shannon, K. A., Chu, J., Scott, K., Struhl, M. K., Cao, P., Tan, X., Hart, P., Raz, G., Piccolo, S., Mei, C., Potter, C., Jaffe-Dax, S., Lew-Williams, C., Tenenbaum, J., Fairchild, K., Bermano, A., & Liu, S. (2023). iCatcher+: Robust and automated annotation of infants' and young children's gaze behavior from videos collected in laboratory, field, and online studies. *Advances in Methods and Practices in Psychological Science*, *6*(2), Article 25152459221147250. https://doi.org/10.1177/25152459221147250

Falck-Ytter, T., Gredebäck, G., & von Hofsten, C. (2006). Infants predict other people's action goals. *Nature Neuroscience, 9*(7), 878–879. https://doi.org/10.1038/nn1729

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1-Level 2 distinction. *Developmental Psychology, 17*(1), 99–103. https://doi.org/10.1037/0012-1649.17.1.99

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S., & Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. https://doi.org/10.1111/infa.12182

Ganglmayer, K., Attig, M., Daum, M. M., & Paulus, M. (2019). Infants' perception of goal directed actions: A multi-lab replication reveals that infants anticipate paths and not goals. *Infant Behavior and Development, 57*, Article 101340. https://doi.org/10.1016/j.infbeh.2019.101340

Ganglmayer, K., Schuwerk, T., Sodian, B., & Paulus, M. (2020). Do children and adults with autism spectrum condition anticipate others' actions as goal-directed? A predictive coding perspective. *Journal of Autism, 50*, 2077–2089. https://doi.org/10.1007/s10803-019-03964-8

Gernsbacher, M. A., & Yergeau, M. (2019). Empirical failures of the claim that autistic people lack a theory of mind. *Archives of Scientific Psychology, 7*(1), 102–118. https://doi.org/10.1037/arc0000067

Ghai, S., Forscher, P. S., & Chuan-Peng, H. (2024). Big-team science does not guarantee generalizability. *Nature Human Behaviour*, *8*(6), 1053–1056. https://doi.org/10.1038/s41562-024-01902-y

Gonzales, C. R., Fabricius, W. V., & Kupfer, A. S. (2018). Introspection plays an early role in children's explicit theory of mind development. *Child Development, 89*(5), 1545–1552. https://doi.org/10.1111/cdev.12876

Gönül, G., Kammermeier, M., & Paulus, M. (2024). What is in an action? Preschool children predict that agents take previous paths and not previous goals. *Developmental Science*, *27*(3), Article e13466. https://doi.org/10.1111/desc.13466

Gredebäck, G., Stasiewicz, D., Falck-Ytter, T., von Hofsten, C., & Rosander, K. (2009). Action type and goal type modulate goal-directed gaze shifts in 14-month-old infants. *Developmental Psychology*, *45*(4), 1190–1194. https://doi.org/10.1037/a0015667

Grosse Wiesmann, C., Friederici, A. D., Disla, D., Steinbeis, N., & Singer, T. (2018). Longitudinal evidence for 4-year-olds' but not 2- and 3-year-olds' false belief-related action anticipation. *Cognitive Development*, *46*, 58–68. https://doi.org/10.1016/j.cogdev.2017.08.007

Hagihara, H., Zaadnoordijk, L., Cusack, R., Kimura, N., & Tsuji, S. (2024). Exploration of factors affecting webcam-based automated gaze coding. *Behavior Research Methods*, *56*(7), 7374–7390. https://doi.org/10.3758/s13428-024-02424-1

Happé F. G. E. (1994). An advanced test of theory of mind: understanding of story Characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, *24*(2), 129–154. https://doi.org/10.1007/BF02172093

Happé, F. G. E. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development*, *66*(3), 843–855. https://doi.org/10.2307/1131954

Harman, G. (1978). Studying the chimpanzee's theory of mind. *Behavioral and Brain Sciences, 1*(4), 576–577. https://doi.org/10.1017/S0140525X00076743

Harris, P. L., de Rosnay, M., & Pons, F. (2005). Language and children's understanding of mental states. *Current Directions in Psychological Science, 14*(2), 69–73. https://doi.org/10.1111/j.0963-7214.2005.00337.x

Harris, P. L., Ronfard, S., & Bartz, D. (2017). Young children's developing conception of knowledge and ignorance: Work in progress. *European Journal of Developmental Psychology, 14(*2), 221–232. https://doi.org/10.1080/17405629.2016.1190267

Harris, P. L., Yang, B., & Cui, Y. (2017). 'I don't know': Children's early talk about knowledge. *Mind & Language, 32*(3), 283–307. https://doi.org/10.1111/mila.12143

Heene, M., & Ferguson, C. J. (2017). Psychological science's aversion to the null, and why many of the things you think are true, aren't. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 34–52). Wiley Blackwell. https://doi.org/10.1002/9781119095910.ch3

Heil, L., Colizoli, O., Hartstra, E., Kwisthout, J., van Pelt, S., van Rooij, I., & Bekkering, H. (2019). Processing of prediction errors in mentalizing areas. *Journal of Cognitive Neuroscience*, *31*(6), 900–912. https://doi.org/10.1162/jocn_a_01381

Henrich, J., Heine, S., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature, 466,* 29 https://doi.org/10.1038/466029a

Heyes, C. (2014). False belief in infancy: A fresh look. *Developmental Science, 17*(5), 647–659. https://doi.org/10.1111/desc.12148

Kaltefleiter, L., Sodian, B., Kristen-Antonow, S., Grosse Wiesmann, C., & Schuwerk, T. (2021). Does syntax play a role in theory of mind development before the age of 3 years? *Infant Behavior and Development, 64*, Article 101575. https://doi.org/10.1016/j.infbeh.2021.101575

Kampis, D., Karman, P., Csibra, G., Southgate, V., & Hernik, M. (2021). A two-lab direct replication attempt of Southgate, Senju and Csibra (2007). *Royal Society Open Science, 8*(8), Article 210190. https://doi.org/10.1098/rsos.210190

Kana, R. K., Keller, T. A., Cherkassky, V. L., Minshew, N. J., & Just, M. A. (2009). Atypical frontal-posterior synchronization of theory of mind regions in autism during mental state attribution. *Social Neuroscience, 4*, 135–152. http://doi.org/10.1080/17470910802198510

Kana, R. K., Libero, L. E., Hu, C. P., Deshpande, H. D., & Colburn, J. S. (2014). Functional brain networks and white matter underlying theory-of-mind in autism. *Social Cognitive and Affective Neuroscience, 9*(1), 98–105. https://doi.org/10.1093/scan/nss106

Kauschke, C., & Hofmeister, C. (2002). Early lexical development in German: A study on vocabulary growth and vocabulary composition during the second and third year of life. *Journal of Child Language, 29*(4), Article 735757. https://doi.org/10.1017/S0305000902005330

Kenny, L., Hattersley, C., Molins, B., Buckley, C., Povey, C., & Pellicano, E. (2016). Which terms should be used to describe autism? Perspectives from the UK autism community. *Autism*, *20*(4), 442–462. https://doi.org/10.1177/1362361315588200

Klein, R. (1988). Inhibitory tagging system facilitates visual search. *Nature, 334*(6181), 430–431. https://doi.org/10.1038/334430a0

Klein, R. M., & MacInnes, W. J. (1999). Inhibition of return is a foraging facilitator in visual search. *Psychological Science, 10*(4), 346–352. https://doi.org/10.1111/1467-9280.00166

Kloo, D., Kristen-Antonow, S., & Sodian, B. (2020). Progressing from an implicit to an explicit false belief understanding: A matter of executive control? *International Journal of Behavioral Development, 44*(2), 107–115. https://doi.org/10.1177/0165025419850901

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron, 79*(5), 836–848. https://doi.org/10.1016/j.neuron.2013.08.020

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science, 330*(6012), 1830–1834. https://doi.org/10.1126/science.1190792

Korbmacher, M., Azevedo, F., Pennington, C. R., Hartmann, H., Pownall, M., Schmidt, K., Elsherif, M., Breznau, N., Robertson, O., Kalandadze, T., Yu, S., Baker, B. J., O'Mahony, A., Olsnes, J. Ø., Shaw, J. J., Gjoneska, B., Yamada, Y., Röer, J. P., Murphy, J., Alzahawi, S., … Evans, T. (2023). The replication crisis has led to positive structural, procedural, and community changes. *Communications Psychology*, *1*(1), Article 3. https://doi.org/10.1038/s44271-023-00003-2

Kristen-Antonow, S., Jarvers, I., & Sodian, B. (2019). Preschoolers' developing understanding of factivity in mental verb comprehension and its relation to first-and second-order false belief understanding: A longitudinal study. *Journal of Cognition and Development, 20*(3), 354–369. https://doi.org/10.1080/15248372.2019.1586710

Kulke, L., & Rakoczy, H. (2018). Implicit theory of mind–an overview of current replications and non-replications. *Data in Brief, 16*, 101–104. https://doi.org/10.1016/j.dib.2017.11.016

Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (2018). How robust are anticipatory looking measures of theory of mind? Replication attempts across the life span. *Cognitive Development, 46,* 97–111. https://doi.org/10.1016/j.cogdev.2017.09.001

Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018). Is implicit theory of mind a real and robust phenomenon? Results from a systematic replication study. *Psychological Science, 29*(6), 888–900. https://doi.org/10.1177/0956797617747090

Lee, A., Aly, M., & Baldassano, C. (2021). Anticipation of temporally structured events in the brain. *eLife, 10*, Article e64972. https://doi.org/10.7554/eLife.64972

Liszkowski, U., Carpenter, M., & Tomasello, M. (2008). Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. *Cognition, 108*(3), 732–739. https://doi.org/10.1016/j.cognition.2008.06.013

Livingston, L. A., Carr, B., & Shah, P. (2019). Recent advances and new directions in measuring theory of mind in autistic adults. *Journal of Autism and Developmental Disorders*, *49*(4), 1738–1744. https://doi.org/10.1007/s10803-018-3823-3

Mangnus, M., Koch, S. B. J., Cai, K., Greidanus Romaneli, M., Hagoort, P., Bašnáková, J., & Stolk, A. (2024). Preserved spontaneous mentalizing amid reduced intersubject variability in autism during a movie narrative. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. Advance online publication. https://doi.org/10.1016/j.bpsc.2024.10.007

Mani, N., & Plunkett, K. (2007). Phonological specificity of vowels and consonants in early lexical representations. *Journal of Memory and Language, 57*(2), 252–272. https://doi.org/10.1016/j.jml.2007.03.005

Mani, N., & Plunkett, K. (2008). Fourteen-month-olds pay attention to vowels in novel words. *Developmental Science, 11*(1), 53–59. https://doi.org/10.1111/j.1467-7687.2007.00645.x

ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science*, 3(1), 24–52. https://doi.org/10.1177/2515245919900809

Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annual Review of Psychology*, *62*, 103–134. https://doi.org/10.1146/annurev-psych-120709-145406

Matsui, T., Yamamoto, T., & McCagg, P. (2006). On the role of language in children's early understanding of others as epistemic beings. *Cognitive Development*, *21*, 158–173. https://doi.org/10.1016/j.cogdev.2005.10.001

Matsui, T., Yamamoto, T., Miura, Y., & McCagg, P. (2016). Young children's early sensitivity to linguistic indications of speaker certainty in their selective word learning. *Lingua, 175–176*, 83–96. https://doi.org/10.1016/j.lingua.2015.10.007

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development, 78*(2), 622–646. https://doi.org/10.1111/j.1467-8624.2007.01018.x

Moessnang, C., Baumeister, S., Tillmann, J., Goyard, D., Charman, T., Ambrosino, S., Baron-Cohen, S., Beckmann, C., Bölte, S., Bours, C., Crawley, D., Dell'Acqua, F., Durston, S., Ecker, C., Frouin, V., Hayward, H., Holt, R., Johnson, M., Jones, E., Lai, M. C., … EU-AIMS LEAP group (2020). Social brain activation during mentalizing in a large autism cohort: The longitudinal European Autism Project. *Molecular Autism*, *11*(1), Article 17. https://doi.org/10.1186/s13229-020-0317-x

Moll, H., Carpenter, M., & Tomasello, M. (2007). Fourteen-month-olds know what others experience only in joint engagement. *Developmental Science, 10*(6), 826–835. https://doi.org/10.1111/j.1467-7687.2007.00615.x

Moll, H., & Tomasello, M. (2006). Level 1 perspective-taking at 24 months of age. *British Journal of Developmental Psychology, 24*(3), 603–613. https://doi.org/10.1348/026151005X55370

Moore, C., Bryant, D., & Furrow, D. (1989). Mental terms and the development of certainty. *Child Development, 60*(1), 167–171. https://doi.org/10.2307/1131082

Moore, C., Furrow, D., Chiasson, L., & Patriquin, M. (1994). Developmental relationships between production and comprehension of mental terms. *First Language, 14*(42–43), 1–17. https://doi.org/10.1177/014272379401404201

Müller, E., Schuler, A., & Yates, G. B. (2008). Social challenges and supports from the perspective of individuals with Asperger syndrome and other autism spectrum disabilities. *Autism, 12*(2), 173–190. https://doi.org/10.1177/1362361307086664

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond western, educated, industrial, rich, and democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological Science*, *31*(6), 678–701. https://doi.org/10.1177/0956797620916782

Nagel, J. (2017). Factive and non-factive mental state attribution. *Mind & Language, 32*(5), 525–544. https://doi.org/10.1111/mila.12157

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Nijhof, A. D., Bardi, L., Brass, M., & Wiersema, J. R. (2018). Brain activity for spontaneous and explicit mentalizing in adults with autism spectrum disorder: An fMRI study. *NeuroImage. Clinical*, *18*, 475–484. https://doi.org/10.1016/j.nicl.2018.02.016

Oakes, L. M. (2012). Advances in eye tracking in infancy research. *Infancy, 17*(1), 1–8. https://doi.org/10.1111/j.1532-7078.2011.00101.x

Oakley, B. F. M., Brewer, R., Bird, G., & Catmur, C. (2016). Theory of mind is not theory of emotion: A cautionary note on the Reading the Mind in the Eyes test. *Journal of Abnormal Psychology, 125*(6), 818–823. https://doi.org/10.1037/abn0000182

Olineck, K. M., & Poulin-Dubois, D. (2007). Imitation of intentional actions and internal state language in infancy predict preschool theory of mind skills. *European Journal of Developmental Psychology, 4*(1), 14–30. https://doi.org/10.1080/17405620601046931

O'Neill, D. K. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child Development, 67*(2), 659–677. https://doi.org/10.2307/1131839

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs?. *Science, 308*(5719), 255–258. https://doi.org/10.1126/science.1107621

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Ozkan, A. (2018). Using eye-tracking methods in infant memory research. *The Journal of Neurobehavioral Sciences, 5,* 62–66. https://doi.org/10.5455/JNBS.1516325266

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer: Scalable webcam eye tracking using user interactions. In *Proceedings of the 25th international joint conference on artificial intelligence* (pp. 3839–3845).

Paulus, M. (2022). Should infant psychology rely on the violation-of-expectation method? Not anymore. *Infant and Child Development, 31*(1), Article e2306. https://doi.org/10.1002/icd.2306

Paulus, M., Schuwerk, T., Sodian, B., & Ganglmayer, K. (2017). Children's and adults' use of verbal information to visually anticipate others' actions: A study on explicit and implicit social-cognitive processing. *Cognition*, *160*, 145–152. https://doi.org/10.1016/j.cognition.2016.12.013

Pellicano, E., & Burr, D. (2012). When the world becomes 'too real': A bayesian explanation of autistic perception. *Trends in Cognitive Sciences, 16*(10), 504–510. https://doi.org/10.1016/j.tics.2012.08.009

Pellicano, E., & den Houting, J. (2022). Annual Research Review: Shifting from 'normal science' to neurodiversity in autism science. *Journal of Child Psychology and Psychiatry*, *63*(4), 381–396. https://doi.org/10.1111/jcpp.13534

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology, 5*(2), 125–137. https://doi.org/10.1111/j.2044-835X.1987.tb01048.x

Perner, J., & Roessler, J. (2012). From infants' to children's appreciation of belief. *Trends in Cognitive Sciences*, *16*(10), 519–525. https://doi.org/10.1016/j.tics.2012.08.004

Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How deep?. *Science,* 308, 214–216. https://doi.org/10.1126/science.1111656

Perner, J., & Wimmer, H. (1985). "John *thinks* that Mary *thinks* that…": Attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology, 39*(3), 437–471. https://doi.org/10.1016/0022-0965(85)90051-7

Phillips, J., Buckwalter, W., Cushman, F., Friedman, O., Martin, A., Turri, J., Santos, L., & Knobe, J. (2020). Knowledge before belief. *The Behavioral and Brain Sciences*, *44*, Article e140. https://doi.org/10.1017/S0140525X20000618

Poulin-Dubois, D., Rakoczy, H., Burnside, K., Crivello, C., Dörrenberg, S., Edwards, K., Krist, H., Kulke, L., Liszkowski, U., Low, J., Perner, J., Powell, L., Priewasser, B., Rafetseder, E., & Ruffman, T. (2018). Do infants understand false beliefs? We don't know yet – a commentary on Baillargeon, Buttelmann and Southgate's commentary. *Cognitive Development, 48*, 302–315. https://doi.org/10.1016/j.cogdev.2018.09.005

Poulin-Dubois, D., & Yott, J. (2018). Probing the depth of infants' theory of mind: Disunity in performance across paradigms. *Developmental Science*, *21*(4), Article e12600. https://doi.org/10.1111/desc.12600

Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development, 46,* 40–50. https://doi.org/10.1016/j.cogdev.2017.10.004

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*, 515–526. https://doi.org/10.1017/S0140525X00076512

Priewasser, B., Rafetseder, E., Gargitter, C., & Perner, J. (2018). Helping as an early indicator of a theory of mind: Mentalism or teleology? *Cognitive Development, 46*, 69–78. https://doi.org/10.1016/j.cogdev.2017.08.002

Quesque, F., Apperly, I., Baillargeon, R., Baron-Cohen, S., Becchio, C., Bekkering, H., Bernstein, D., Bertoux, M., Bird, G., Bukowski, H., Burgmer, P., Carruthers, P., Catmur, C., Dziobek, I., Epley, N., Erle, T. M., Frith, C., Frith, U., Galang, C. M., Gallese, V., … Brass, M. (2024). Defining key concepts for mental state attribution. *Communications Psychology*, *2*(1), Article 29. https://doi.org/10.1038/s44271-024-00077-6

Rakoczy, H. (2022). Foundations of theory of mind and its development in early childhood. *Nature Reviews Psychology, 1*, 223–235. https://doi.org/10.1038/s44159-022-00037-z

Raz, G., Piccolo, S., Medrano, J., Liu, S., Lydic, K., Mei, C., Nguyen, V., Shu, T., & Saxe, R. (2024). An asynchronous, hands-off workflow for looking time experiments with infants. *Developmental Psychology, 60*(8), 1447–1456. https://doi.org/10.1037/dev0001791

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), Article 1027. https://doi.org/10.1038/s41467-018-03399-2

Richardson, H., & Saxe, R. (2019). Development of predictive responses in theory of mind brain regions. *Developmental Science*, *23*(1), Article e12863. https://doi.org/10.1111/desc.12863

Ruffman, T. (2014). To belief or not belief: Children's theory of mind. *Developmental Review, 34*(3), 265–293. https://doi.org/10.1016/j.dr.2014.04.001

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *NeuroImage*, *19*(4), 1835–1842. https://doi.org/10.1016/s1053-8119(03)00230-1

Scheeren, A. M., de Rosnay, M., Koot, H. M., & Begeer, S. (2013). Rethinking theory of mind in high-functioning autism spectrum disorder. *Journal of Child Psychology and Psychiatry*, *54*(6), 628–635. https://doi.org/10.1111/jcpp.12007

Schidelko, L. P., Schünemann, B., Rakoczy, H., & Proft, M. (2021). Online testing yields the same results as lab testing: A validation study with the false belief task. *Frontiers in Psychology*, *12*, Article 703238. https://doi.org/10.3389/fpsyg.2021.703238

Schimmelpfennig, R., Spicer, R., White, C. J. M., Gervais, W., Norenzayan, A., Heine, S., Henrich, J., & Muthukrishna, M. (2024). The moderating role of culture in the generalizability of psychological phenomena. *Advances in Methods and Practices in Psychological Science, 7*(1), Article 25152459231225163. https://doi.org/10.1177/25152459231225163

Schneider, D., Slaughter, V. P., Bayliss, A. P., & Dux, P. E. (2013). A temporally sustained implicit theory of mind deficit in autism spectrum disorders. *Cognition*, *129*(2), 410–417. https://doi.org/10.1016/j.cognition.2013.08.004

Schneider, D., Slaughter, V. P., & Dux, P. E. (2017). Current evidence for automatic theory of mind processing in adults. *Cognition*, *162*, 27–31. https://doi.org/10.1016/j.cognition.2017.01.018

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, *42*, 9–34. https://doi.org/10.1016/j.neubiorev.2014.01.009

Schuwerk, T., & Paulus, M. (2018). Action prediction in autism. In F. R. Volkmar (Ed.), *Encyclopedia of autism spectrum disorders*. Springer. https://doi.org/10.1007/978-1-4614-6435-8_102206-1

Schuwerk, T., Priewasser, B., Sodian, B., & Perner, J. (2018). The robustness and generalizability of findings on spontaneous false belief sensitivity: A replication attempt. *Royal Society Open Science*, *5*(5), Article 172273. https://doi.org/10.1098/rsos.172273

Schuwerk, T., & Sodian, B. (2023). Differences in self-other control as cognitive mechanism to characterize theory of mind reasoning in autistic and non-autistic adults. *Autism Research, 16*(9), 1728–1738. https://doi.org/10.1002/aur.2976

Schuwerk, T., Vuori, M., & Sodian, B. (2015). Implicit and explicit theory of mind reasoning in autism spectrum disorders: The impact of experience. *Autism*, *19*(4), 459–468. https://doi.org/10.1177/1362361314526004

Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences, 21*(4), 237–249. https://doi.org/10.1016/j.tics.2017.01.012

Scott, K., & Schulz, L. (2017). Lookit (Part 1): A new online platform for developmental research. *Open Mind, 1*(1), 4–14. https://doi.org/10.1162/OPMI_a_00002

Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods*, *50*(2), 451–465. https://doi.org/10.3758/s13428-017-0913-7

Senju, A. (2012). Spontaneous theory of mind and its absence in autism spectrum disorders. *The Neuroscientist*, *18*(2), 108–113. https://doi.org/10.1177/1073858410397208

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An absence of spontaneous theory of mind in Asperger syndrome. *Science*, *325*(5942), 883–885. https://doi.org/10.1126/science.1176170

Shatz, M., Wellman, H. M., & Silber, S. (1983). The acquisition of mental verbs: A systematic investigation of the first reference to mental state. *Cognition, 14*(3), 301–321. https://doi.org/10.1016/0010-0277(83)90008-2

Shaw, K. A., Williams, S., Patrick, M. E., Valencia-Prado, M., Durkin, M. S., Howerton, E. M., Ladd-Acosta, C. M., Pas, E. T., Bakian, A. V., Bartholomew, P., Nieves-Muñoz, N., Sidwell, K., Alford, A., Bilder, D. A., DiRienzo, M., Fitzgerald, R. T., Furnier, S. M., Hudson, A. E., Pokoski, O. M., Shea, L., … Maenner, M. J. (2025). Prevalence and early identification of autism spectrum disorder among children aged 4 and 8 years - Autism and developmental disabilities monitoring network, 16 Sites, United States, 2022. *Morbidity and Mortality Weekly Report. Surveillance Summaries, 74*(2), 1–22. https://doi.org/10.15585/mmwr.ss7402a1

Singh, L., Cristia, A., Karasik, L. B., Rajendra, S. J., & Oakes, L. M. (2023). Diversity and representation in infant research: Barriers and bridges toward a globalized science of infant development. *Infancy*, *28*(4), 708–737. https://doi.org/10.1111/infa.12545

Sinha, P., Kjelgaard, M. M., Gandhi, T. K., Tsourides, K., Cardinaux, A. L., Pantazis, D., Diamond, S. P., & Held, R. M. (2014). Autism as a disorder of prediction. *Proceedings of the National Academy of Sciences, 111*(42), 15220–15225. https://doi.org/10.1073/pnas.1416797111

Slim, M. S., Kandel, M., Yacovone, A., & Snedeker, J. (2024). Webcams as windows to the mind? A direct comparison between in-lab and web-based eye-tracking methods. *Open Mind, 8*, 1369–1424. https://doi.org/10.1162/opmi_a_00171

Smith-Flores, A. S., Perez, J., Zhang, M. H., & Feigenson, L. (2022). Online measures of looking and learning in infancy. *Infancy*, *27*(1), 4–24. https://doi.org/10.1111/infa.12435

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic stimuli in neuroscience: Critically acclaimed. *Trends in Cognitive Sciences, 23*(8), 699–714. https://doi.org/10.1016/j.tics.2019.05.004

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science, 18*(7), 587–592. https://doi.org/10.1111/j.1467-9280.2007.01944.x

Stahl, A. E., & Kibbe, M. M. (2022). Great expectations: the construct validity of the violation-of-expectation method for studying infant cognition. *Infant and Child Development, 31*(6), Article e2359. https://doi.org/10.1002/icd.2359

Stenberg, G. (2009). Selectivity in infant social referencing. *Infancy, 14*(4), 457–473. https://doi.org/10.1080/15250000902994115

Sugranyes, G., Kyriakopoulos, M., Corrigall, R., Taylor, E., & Frangou, S. (2011). Autism spectrum disorders and schizophrenia: Meta-analysis of the neural correlates of social cognition. *PLoS ONE, 6*(10), Article e25322. https://doi.org/10.1371/journal.pone.0025322

Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M. M., & Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. *Nature Neuroscience*, *11*(9), 1004–1006. https://doi.org/10.1038/nn.2163

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science, 18*(7), 580–586. https://doi.org/10.1111/j.1467-9280.2007.01943.x

Surian, L., & Geraci, A. (2012). Where will the triangle look for it? Attributing false beliefs to a geometric shape at 17 months. *The British Journal of Developmental Psychology*, *30*(1), 30–44. https://doi.org/10.1111/j.2044-835X.2011.02046.x

Tager-Flusberg, H. (2007). Evaluating the theory-of-mind hypothesis of autism. *Current Directions in Psychological Science, 16*(6), 311–315. https://doi.org/10.1111/j.1467-8721.2007.00527.x

Tenenbaum, E. J., Stone, C., Vu, M. H., Hare, M., Gilyard, K. R., Arunachalam, S., Bergelson, E., Bishop, S. L., Frank, M. C., Hamlin, J. K., Kline Struhl, M., Landa, R. J., Lew-Williams, C., Libertus, M. E., Luyster, R. J., Markant, J., Sabatos-DeVito, M., Sheinkopf, S. J., Wagner, J. B., Park, K., … Jeste, S. (2025). Remote infant studies of early learning (RISE): Scalable online replications of key findings in infant cognitive development. *Developmental Psychology*, *61*(1), 151–167. https://doi.org/10.1037/dev0001849

Thornton, M. A., Weaverdyck, M. E., & Tamir, D. I. (2019). The social brain automatically predicts others' future mental states. *The Journal of Neuroscience, 39*(1), 140–148. https://doi.org/10.1523/JNEUROSCI.1431-18.2018

Tomasello, M., & Haberl, K. (2003). Understanding attention: 12-and 18-month-olds know what is new for other persons. *Developmental Psychology, 39*(5), 906–912. https://doi.org/10.1037/0012-1649.39.5.906

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., &

Wagemans, J. (2014). Precise minds in uncertain worlds: Predictive coding in autism. *Psychological Review*, *121*(4), 649–675. https://doi.org/10.1037/a0037665

Venker, C. E., & Kover, S. T. (2015). An open conversation on using eye-gaze methods in studies of neurodevelopmental disorders. *Journal of Speech, Language, and Hearing Research, 58*(6), 1719–1732. https://doi.org/10.1044/2015_JSLHR-L-14-0056

Visser, I., Bergmann, C., Byers-Heinlein, K., Dal Ben, R., Duch, W., Forbes, S., Franchin, L., Frank, M. C., Geraci, A., Hamlin, J. K., Kaldy, Z., Kulke, L., Laverty, C., Lew-Williams, C., Mateu, V., Mayor, J., Moreau, D., Nomikou, I., Schuwerk, T., … Zettersten, M. (2022). Improving the generalizability of infant psychological research: The Many Babies model. *Behavioral and Brain Sciences, 45*, Article e35. https://doi.org/10.1017/S0140525X21000455

Wass, S. V. (2016). The use of eye-tracking with infants and children. In J. Prior & J. Van Herwegen (Eds.), *Practical research with children* (1st ed., pp. 24–45). Routledge. https://doi.org/10.4324/9781315676067

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655–684. https://doi.org/10.1111/1467-8624.00304

Werchan, D. M., Thomason, M. E., & Brito, N. H. (2022). OWLET: An automated, open-source method for infant gaze tracking using smartphone and webcam recordings. *Behavior Research Methods, 55*, 3149–4163. https://doi.org/10.3758/s13428-022-01962-w

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103-128. https://doi.org/10.1016/0010-0277(83)90004-5

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69*, 1–34. https://doi.org/10.1016/S0010-0277(98)00058-4

World Health Organization. (2022). *International classification of diseases for mortality and morbidity statistics (11th revision)*. https://icd.who.int/

Yang, X., & Krajbich, I. (2021). Webcam-based online eye-tracking for behavioral research. *Judgment and Decision Making, 16*(6), 1485–1505. https://doi.org/10.1017/S1930297500008512

Zaadnoordijk, L., Buckler, H., Cusack, R., Tsuji, S., & Bergmann, C. (2021). A global perspective on testing infants online: Introducing ManyBabies-AtHome. *Frontiers in Psychology*, *12*, Article 703234. https://doi.org/10.3389/fpsyg.2021.703234

Zeidan, J., Fombonne, E., Scorah, J., Ibrahim, A., Durkin, M. S., Saxena, S., Yusuf, A., Shih, A., & Elsabbagh, M. (2022). Global prevalence of autism: A systematic review update. *Autism Research*, *15*(5), 778–790. https://doi.org/10.1002/aur.2696

5. Appendices

Appendix A. Manuscript Study 1

DOI: 10.1111/infa.12564

RESEARCH ARTICLE

INFANCY THE OFFICIAL JOURNAL OF THE INTERNATIONAL CONGRESS OF INFANT STUDIES WILEY

Validation of an open source, remote web-based eye-tracking method (WebGazer) for research in early childhood

Adrian Steffan[1] | Lucie Zimmer[1] | Natalia Arias-Trejo[2] | Manuel Bohn[3,4] | Rodrigo Dal Ben[5] | Marco A. Flores-Coronado[2] | Laura Franchin[6] | Isa Garbisch[7] | Charlotte Grosse Wiesmann[8] | J. Kiley Hamlin[9] | Naomi Havron[10] | Jessica F. Hay[11] | Tone K. Hermansen[12] | Krisztina V. Jakobsen[13] | Steven Kalinke[3] | Eon-Suk Ko[14] | Louisa Kulke[15] | Julien Mayor[12] | Marek Meristo[16] | David Moreau[17] | Seongmin Mun[14] | Julia Prein[3] | Hannes Rakoczy[7] | Katrin Rothmaler[8] | Daniela Santos Oliveira[11] | Elizabeth A. Simpson[18] | Sylvain Sirois[19] | Eleanor S. Smith[20] | Karin Strid[16] | Anna-Lena Tebbe[8] | Maleen Thiele[3] | Francis Yuen[9] | Tobias Schuwerk[1]

Correspondence
Lucie Zimmer, Ludwig-Maximilians-Universität, Leopoldstr. 13, 80802 München, Germany.
Email: lucie.zimmer@psy.lmu.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: DFG RA 2155/7-1, DFG SCHU 3060/2-1; National Institutes of Health, Grant/Award Numbers: NIH 1R01HD083312, NIH 1R15HD099706; National Research Foundation of Korea, Grant/Award Number: NRF-2021R1I1A2051993; National Science Foundation, Grant/Award Number: NSF CAREER 1653737

Abstract

Measuring eye movements remotely via the participant's webcam promises to be an attractive methodological addition to in-person eye-tracking in the lab. However, there is a lack of systematic research comparing remote web-based eye-tracking with in-lab eye-tracking in young children. We report a multi-lab study that compared these two measures in an anticipatory looking task with toddlers using WebGazer.js and jsPsych. Results of our remotely tested sample of 18-27-month-old toddlers (N = 125) revealed that web-based eye-tracking successfully captured

Adrian Steffan and Lucie Zimmer shared first-authorship.

goal-based action predictions, although the proportion of the goal-directed anticipatory looking was lower compared to the in-lab sample ($N = 70$). As expected, attrition rate was substantially higher in the web-based (42%) than the in-lab sample (10%). Excluding trials based on visual inspection of the match of time-locked gaze coordinates and the participant's webcam video overlayed on the stimuli was an important preprocessing step to reduce noise in the data. We discuss the use of this remote web-based method in comparison with other current methodological innovations. Our study demonstrates that remote web-based eye-tracking can be a useful tool for testing toddlers, facilitating recruitment of larger and more diverse samples; a caveat to consider is the larger drop-out rate.

1 | INTRODUCTION

Eye-tracking technology allows researchers to better understand childrens' interactions with the world. Compared to the manual coding of gaze behaviors, eye-tracking can automatically and accurately track gaze patterns on more complex stimuli with higher spatial and temporal resolution (Oakes, 2012; Wass et al., 2013). Best practices for using in-person eye-tracking with young children have been outlined (Oakes, 2012); however, to date, eye-tracking with children has required in-person testing using a commercial eye-tracking system. In adults, remote automated web-based eye-tracking methods have been established in both computational (Valliappan et al., 2020; Xu et al., 2015) and behavioral research (Bogdan et al., 2023; Schneegans et al., 2021; Semmelmann & Weigelt, 2018; Yang & Krajbich, 2021). So far, to our knowledge, none of these systems have been validated in an interactive paradigm for use with young children (for automated gaze coding of already recorded videos, see, Erel et al., 2022; Werchan et al., 2022; for an overview, see, Kominsky et al., 2021; for in-person vs. remote web-based eye-tracking comparison in a looking time paradigm in infants, see, Bánki et al., 2022). Yet, remote automated web-based eye-tracking has become increasingly important in developmental research due to the growing need for testing children at home. During the Covid-19 pandemic, many labs around the world were unable to conduct in-person studies. Remote web-based studies have thus become more popular in recent years (Kominsky et al., 2021; Leshin et al., 2021; Rhodes et al., 2020; Sheskin et al., 2020; Su & Ceci, 2021), with new tools and techniques for moderated versus unmoderated remote studies emerging in developmental psychology (Lo et al., 2021; Oliver & Pike, 2021; Rhodes et al., 2020; Schidelko et al., 2021; Su & Ceci, 2021).

While some of these projects measure children's looking behavior, they still require manual coding from human observers (e.g., Bacon et al., 2021; Bánki et al., 2022; Nelson & Oakes, 2021; Scott & Schulz, 2017). Manual video-coding is still considered, among many researchers, the gold standard. However, it is labor-intensive making it impractical for studies with a large sample size, and requires comprehensive training (Venker & Kover, 2015). To maintain the reliability of manual coding, it is common for coders to participate in lab-wide reliability checks (Yoder et al., 2018) and to report inter-coder agreement for a subset of videos (Fernald et al., 2008), which - though key to coding reliability and replicability of results - even further exacerbates the problem of significantly greater number of hours spent on manual annotation than on running machine algorithms. In contrast, automated web-based eye-tracking provides a resource-saving alternative. It is more efficient and has –compared to manual coding of gaze direction from video replays– a relatively high temporal and spatial resolution. As a result, automated coding methods are capable of capturing dependent variables that manual coding cannot (e.g., pupil size, or discrete fixations within an AOI), providing, in conjunction with the technology allowing for at-home testing, exciting new areas of exploration (Ozkan, 2018). Additional advantages of conducting eye-tracking studies remotely compared to traditional one-lab in-person studies are that they (1) make it easier to scale up for large samples; (2) enable researchers to reach a more demographically diverse cohort (e.g., linguistic diversity, racial/ethnic/cultural backgrounds, socio-economic status) as remote web-based studies can be performed from around the world, improving generalizability (Byers-Heinlein et al., 2020; Visser et al., 2022; (3) can potentially reduce costs associated with renting lab space, buying expensive equipment, and other expenses associated with in-person studies; (4) are less time-consuming for participants and more comforting as they can do the testing in their natural environment; (5) offer greater flexibility in terms of scheduling and the ability to collect data from participants in different time zones and (6) have the potential to facilitate international collaborations among research groups, as they are more easily reproducible and less subjective.

Despite these clear advantages, the new remote web-based eye-tracking methods are still undergoing development and involve limitations such as poorer image quality and uncontrolled experimental conditions when compared to their in-lab counterparts (i.e., infant positioning, lighting in the room, and presence of distractors; Wass, 2016; Zaadnoordijk et al., 2021). In a traditional lab, the researcher can ensure that participants are following the instructions of the study, whereas in a remote setting, the researcher may not be able to monitor the participant as closely, and the quality of the setup often varies. Additionally, commercial eye-trackers have a higher sampling rate (one sample per two or four milliseconds) compared to the average webcams available to participants taking about one sample each 30 ms, leaving the data more noisy.

Here, we aimed to test the precision of a web-based eye-tracking system that uses the participant's webcam. Our experiment is based on jsPsych and WebGazer.js (de Leeuw, 2015; Papoutsaki et al., 2016). jsPsych is a javascript framework used to create behavioral experiments that run in a web browser. It was used, in this instance, to control the content the participants interacted with during the experiment but cannot collect eye tracking data in isolation. Thus, it was combined with the WebGazer plugin to produce the present paradigm and data collection set-up. WebGazer captures gaze coordinates by predicting the participant's gaze location on the screen from the head and eyes position recorded via webcam, relative to the displayed stimuli. To evaluate whether this web-based eye-tracking method is comparable to lab-based eye-tracking, we aimed to replicate findings of an in-lab paradigm of the ManyBabies2 project, which revealed spontaneous goal-directed action anticipation measured by anticipatory looking using commercial eye-tracking systems (Schuwerk et al., 2022). The paradigm involves two agents, one who moves through an opaque tunnel and hides from the other in one of two locations and a chaser who also enters the tunnel and seeks the agent who is hiding. A goal of the ManyBabies2 project is to replicate the finding that infants and toddlers visually anticipate an agent's action which is based on a false belief (Southgate et al., 2007). Action prediction, measured by anticipatory looking toward the outcome of that action, is a strong indicator of infant's cognitive reasoning that drives these predictions (Falck-Ytter et al., 2006). In the employed anticipatory looking paradigm, before presenting a false belief-based action, simple goal-directed actions are presented to familiarize toddlers with the set-up. Showing that they anticipate a goal-directed action in these trials, something that toddlers at that age are capable of (e.g., Liszkowski et al., 2007; Luo & Baillargeon, 2007), is an important validity check of this paradigm. We expected participants to anticipate where the chaser will seek the hiding agent. We compared this anticipatory looking behavior recorded in-lab with anticipatory looking behaviors recorded remotely via webcam in 18- to 27-month-old children.

Following the ManyBabies collaborative framework (Frank et al., 2017; Visser et al., 2021), we conducted a cross-sectional web-based eye-tracking experiment with participants recruited and tested across 16 different labs globally. Labs contributed to recruitment, data collection, data analyses, and other related tasks.

The hypotheses of the present study were the following: First, we expected 18- to 27-month-old children in our web-based eye-tracking sample to engage in goal-based action predictions, indicated by above-chance looking toward the location that matches the outcome of an agent's action goal (i.e., finding the hiding agent). This would replicate Schuwerk et al.'s (2022) results obtained using in-lab commercial eye-tracking systems. Second, we then tested whether the eye-tracking method had an effect on the measured proportional looking score, but had no strong directional hypothesis either way. It could have been that due to the reduced accuracy of remote web-based eye-tracking and increased noise of the at-home test setting, the proportional looking score indicating goal-directed action prediction is smaller in remote web-based than in in-lab eye-tracking. Alternatively, the proportional looking score obtained via remote web-based eye-tracking could have been larger, potentially due to beneficial effects of the familiar environment at home, the increased scheduling flexibility to match children's most attentive times, and the lack of an exhausting trip to a lab. It could also have been that the method would have no effect on the proportional looking score—as these two trends might pull in opposite directions. Third, we expected that the proportion of children who contribute useable data would be lower in the remote web-based setting as compared to in-lab eye-tracking.

A successful replication of in-lab results with our remotely tested sample would render remote automated web-based eye-tracking via the participant's webcam an attractive alternative to in-lab eye-tracking for research on cognitive development. Moreover, our open-source tool would provide the community with a free and powerful method for future research.

2 | METHODS

The study was pre-registered on Open Science Framework (OSF).[1] All materials, data, and the analytic codes are also available on OSF.[2] The software implementing the experiment can be found on GitHub.[3]

2.1 | Participation details

In this multi-lab study, participants were recruited by 16 different labs. For feasibility and data protection reasons, only 11 of these 16 labs were involved in testing. The labs were located in Austria ($n = 1$), Canada ($n = 1$), Germany ($n = 5$), Israel ($n = 1$), Italy ($n = 1$), Mexico ($n = 1$), Norway ($n = 1$), United Kingdom ($n = 1$), United States ($n = 2$), South Korea ($n = 1$), and Sweden ($n = 1$). As participants were recruited and tested by several labs, differing recruitment methods were used (e.g., internal database of laboratories, selected kindergartens, online via social media, birth registries from local registration offices). Participants were compensated according to each individual lab policy (e.g., by gifts, cash). The present study was conducted according to guidelines laid down in the Declaration of

[1] permanent link to pre-registration: https://doi.org/10.17605/OSF.IO/SMYA4.

[2] https://osf.io/p3f67/.

[3] https://github.com/adriansteffan/manywebcams-eyetracking/tree/848504f07fa8c25eb3f28444349a4d60151a7895.

Helsinki, with written informed consent obtained from a parent or guardian for each child before any assessment or data collection. All procedures involving human subjects in this study were approved by the respective Institutional Review Boards (IRBs; for a full list see Supplementary Table 1).

2.1.1 | Time-frame

On September 27th, 2021 we sent an email to the ManyBabies mailing list inviting labs to join the project. Three months later, in January 2022, data collection began and ended in August 2022.

2.1.2 | Lab participation criterion

Participation was open to all labs. However, there were some requirements to participate in data collection or recruitment. Labs needed to: 1) provide ethics approval from their local ethics committee by the start of data collection, 2) be able to actively recruit at least 10 participants and/or be able to test them using either their own WebGazer setup or the one provided by LMU Munich, 3) read the ManyWebcams Manual and comply with the ManyBabies code of conduct (for details see OSF). Note that labs did not have to contribute 10 included participants. Each number of finally useable datasets was included in the overall sample.

2.2 | Participants

The final remotely tested sample consisted of 125 participants (67 girls, 58 boys) aged 18–27 months (548–822 days, M_{age} = 21.83 months, SD_{age} = 2.45 months). All toddlers were born full-term (>37 weeks gestation) and had no reported cognitive, visual, or hearing impairments. Since multiple labs around the world collected data, the participants' places of residence were diverse: Germany (n = 52), Norway (n = 11), Italy (n = 10), United States (n = 10), Sweden (n = 9), United Kingdom (n = 8), Canada (n = 6), Austria (n = 5), Israel (n = 5), South Korea (n = 5), and Mexico (n = 4) (see Figure 1). For most of the participants, at least one parent had an educational degree comparable to a bachelor or higher (n = 105). The parent with the higher educational degree spent on average 17.70 years in education. Among the participants, 26% were raised with a second language (n = 32), and 6% with a third language (n = 7). Regarding the number of siblings, 54% of participants had no siblings (n = 68), 35% had one sibling (n = 44), 10% had two siblings (n = 12) and 1% had three siblings (n = 1). The majority of participants were going to daycare (n = 87) and spent there 31 h per week on average. An additional 118 participants were tested but excluded from the analysis. There was no indication of any systematic differences between the included and excluded participants except for residence country (for more details see Figure 1 and Supplementary Table 2 and 3). Participants were excluded for three main reasons (see Supplementary Figure 1): participant-related exclusions (n = 27), technical-related exclusions (n = 52), or exclusions after visual inspection (n = 39). Participant-related exclusions were due to a mismatch between participants' age and our predefined age range (n = 9), prematurity (n = 8), reported cognitive (n = 8) or vision (n = 2) impairments. Technical-related exclusions and exclusions after the visual inspection process are described in more detail in the results section.

The lab-based sample consisted of 70 toddlers (39 girls, 31 boys) aged between 18 and 27 months (552–812 days, M_{age} = 22.92 months, SD_{age} = 2.62 months). This sample was collected in seven labs across the world. Note that for the analyses of the current study we were able to use data from 70 toddlers tested for a pilot study of the ManyBabies2 project (for the original analysis stricter criteria were applied which led to a final sample of 65 included toddlers; for further details, including further

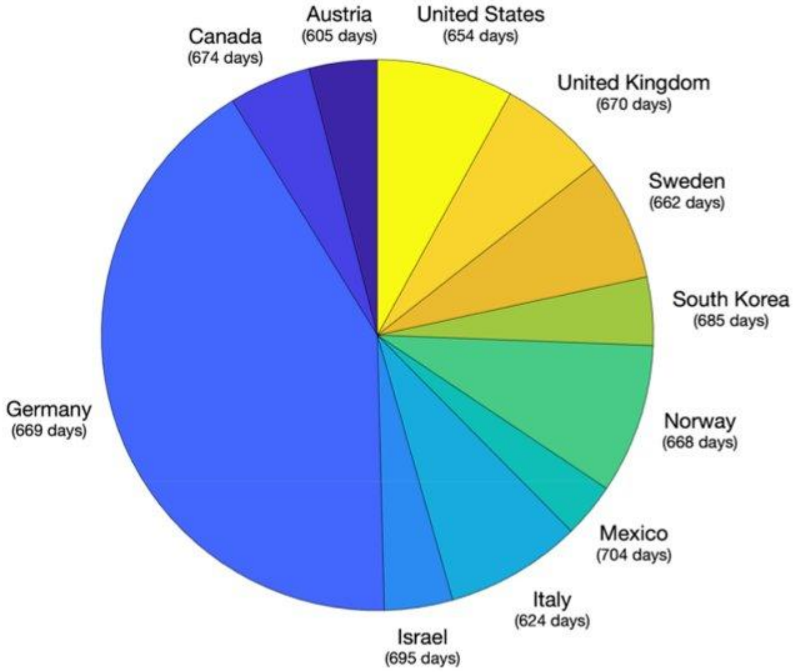


FIGURE 1 Pie chart of the different residence countries of the included participants alongside with their mean age in days in brackets.

information on participating labs, see in Schuwerk et al., 2022). In this pilot study, the appropriateness of the newly developed paradigm was measured. In particular, it was tested if toddlers engage in goal-based action predictions when watching the stimuli.

2.3 | Sample size

Our sample size rationale was based on two effect sizes: Using the same paradigm with in-lab eye-tracking, Schuwerk et al. (2022) observed an effect-size of Cohen's $d = 1.03$ in a sample of 65 toddlers (one sample t test of proportional looking score against chance level). In a pilot study for the current remote web-based version, we tested 40 adults ($M_{age} = 30.10$ years, $SD_{age} = 14.35$ years) and 15 children ($M_{age} = 23.25$ months, $SD_{age} = 10.48$ months). We observed an effect size of Cohen's $d = 0.56$ in a sample of 20 adults who were included in the final analysis, and we did not find a statistically significant effect from the 8 children that were included in the final analysis.

We anticipated two major sources of noise in our data: poorer accuracy of remote web-based eye-tracking as compared to in-lab eye-tracking (Semmelmann & Weigelt, 2018) and more movements artifacts and inattentiveness in toddlers compared to adults (Dalrymple et al., 2018). Based on the observed effect sizes and these considerations, we performed a power analysis for our main hypothesis with the conservative effect size estimate of Cohen's $d = 0.3$. To detect such an effect with a power (1-beta) of 0.95 (using a one sample t test against chance, one-tailed, alpha = 0.05), a minimal sample of 122 toddlers was required. Because in this multi-lab study the exact number of tested participants could not be determined before the end of data collection, we set $N = 122$ as the minimal sample size of included participants.

2.4 | Materials and design

The experimental design was identical to the familiarization phase of the paradigm previously developed for ManyBabies2.[4]

2.5 | Stimuli

2.5.1 | General scene setup

We used 3D animations representing a chasing scenario between two agents (chaser and chasee; Figure 2). The scene depicted an open, blue-colored room divided into two sections by a horizontal brown picket fence: an upper section, which was about one-third of the height of the room, and a lower section, which was about 2/3 of the height of the room. At the beginning of the scene, two animated agents of the same size were visible in the upper section: a brown bear (chaser) and a yellow mouse (chasee). The agents communicated briefly with pseudo statements. When they moved, one could hear their footsteps. The fence dividing the room was interrupted in the middle by a white inverted Y-shaped tunnel through which the agents could pass from one section to the other. One exit of the tunnel led to the upper section and two identical exits to the lower section of the room, one on the right- and one on the left-hand side. In front of the tunnel exits in the lower section of the room, there were two identical brown boxes with a movable lid, one box in front of each exit.

2.5.2 | Test trials

All participants viewed four trials, with each trial lasting 38s (for a detailed description see Schuwerk et al., 2022). Each trial started with a brief game of tag between two agents, the chaser and the chasee, in which the chasee started either on the left or on the right side. After chasing each other, they stopped, did a high five, and ended up standing side by side in front of the tunnel entrance (left or right position counterbalanced). Both chasee and chaser looked at each other briefly. The chaser continued watching as the chasee headed to the tunnel and entered it. After the chasee disappeared in the tunnel, the chaser moved to the tunnel entrance and remained there until the chasee exited the tunnel (left or right, counterbalanced). During this time, only the sound of footsteps indicated that the chasee was moving through the tunnel. After leaving the tunnel, the chasee turned back, implying eye contact with the chaser, to which the chaser responded by raising their hands, and jumped into the opaque box, which was positioned behind the tunnel exit. The chaser also entered the tunnel and, again, the sound of footsteps indicated their walking through the tunnel (anticipatory period, i.e., 4000 ms). The chaser exited the tunnel on the same side the chasee was hiding. Then, the chaser knocked on the box, the chasee jumped out and, again, the agents did a high five. See the OSF repository for the full animations.

[4] https://manybabies.org/MB2/.

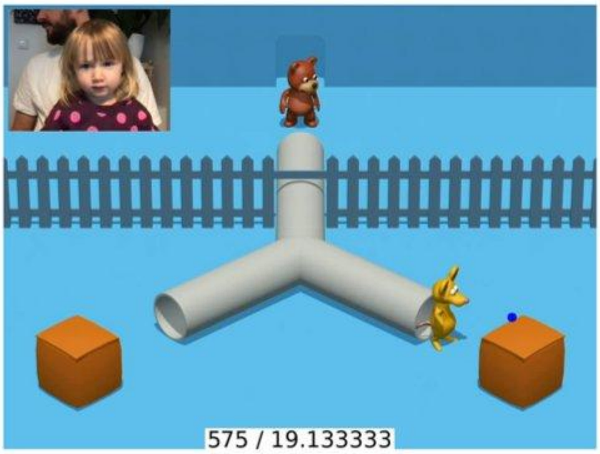


FIGURE 2 A still frame from an overlay of the normalized predictions of gaze location (indicated by the blue dot), the stimuli, and the synchronized webcam video. At the bottom, the current duration (left: frames, right: seconds) is displayed. These overlays were used for the visual inspection process.

2.5.3 | Trial randomization

We used two factors for balancing in the study. First, the location from which the chasee started in the upper section of the room left (L) versus right (R) and second, the box in which the chasee eventually hid (L vs. R). This resulted in four trials: chasee started from the right and ended up in right box (RR); started from the right and ended up in left box (RL); started from the left and ended up in right box (LR); and started from the left and ended up in left box (LL). The order of the four test trials was counterbalanced across participants using two pre-specified pseudo-randomized orders to which they were randomly assigned: LR, LL, RR, RL (Order A); RL, RR, LL, LR (Order B). The trial orders were identical to those used in the in-lab study.

2.6 | Apparatus and procedure

2.6.1 | Testing procedure

Participants met the researcher via a video conference software (e.g., Zoom). Before the test session, the caregiver provided informed written consent via an online survey tool offered by their institution or other third-party software solutions, given the use was covered by their local ethics approval. Subsequently, caregivers completed a demographic questionnaire, which included questions about linguistic and racial/ethnic background, resident country, socio-economic status, caregivers characteristics, and family characteristics. After explaining the general procedure, the researcher offered the caregiver the following instructions. Caregivers were asked to have the child sit in front of a laptop or desktop computer screen with a horizontal screen orientation at a distance of approximately 40 cm. The child could be seated either on their caregiver's lap or in a highchair. Then, the experimenter guided the caregiver to obtain suitable lighting and webcam positioning: If a laptop was used, the caregiver was asked to place it on top of a table and have the child sit in front of it. If a light source (e.g., a window) caused backlight, the experimenter asked the caregiver to reposition the computer to reach an appropriate angle toward the light source or asked the caregiver to cover it. Caregivers adjusted the angle of the webcam/laptop screen, so that the child's head was centered on the screen, and the caregiver's head was outside of the camera's scope. Alternatively, caregivers were advised to obstruct, close, or move their eyes away from the range of the camera during the experiment, as to not interfere with the eye-tracking procedure. The experimenter then provided the caregiver with a link to access the experimental task and reminded the caregiver to rejoin the video conference after the end of the experiment. Subsequently, the caregiver left the video conference session and accessed the experiment on a browser of their choice (Google Chrome and Firefox were recommended) and started the experiment. During the experiment, the participant's webcam was used to record the child's gaze locations. We also saved the webcam video, which recorded the child's behavior while watching the stimuli. We used a modified version of jsPsych v6.3.1 (de Leeuw, 2015) to control the experimental procedure and stimuli video presentation. To infer the participant's gaze location during the video stimulus presentation, we used WebGazer.js (Papoutsaki et al., 2016). WebGazer is a browser-based eye-tracking library that uses webcam video to infer the participant's gaze locations. It approximates gaze location using a regression model that learns the mapping from pupil positions and eye features to screen coordinates. During the initialization of the eye-tracking procedure, the software also controlled for the distance of the participant in relation to the monitor. To satisfy the headpose requirements enforced by WebGazer, the experiment proceeded only if both eyes were detected within a rectangle (with dimensions equivalent to ⅔ of the webcam feed's height) which was displayed on the screen. Following this requirement, the distance range accepted by the experiment's software spanned 40–130 cm (i.e., 15.7–51.2in). Distances outside of this range caused the program to prompt the participant to move closer or further away from the screen.

At the beginning of the experimental task, a 9-point calibration of the eye-tracking software was displayed, each point appearing for 3 s. During this calibration procedure, a looping animation of a dancing teddy bear was presented as an attention-getter at each calibration point (coordinates in screen percentage [width, height] in order: ([50,50], [50,12], [12,12], [12,50], [12,88], [50,88], [88,88], [88,50], [88,12]) along with an audio cue to attract the participant's attention. This combination of a 9-point calibration procedure and child-friendly attention-getter was used to enhance data accuracy (Zeng et al., 2023). We assessed the quality of the calibration twice: once after the calibration procedure and once after the stimulus display (the second assessment quantified the decrease in eye-tracking quality over time). An attention getter appeared in the middle of the screen for 5 s, and we recorded the average x/y deviations of inferred gaze locations from the center of the screen in pixels during this time. Even though there was no ground truth to compare these values against (making the absolute values difficult to interpret), comparing the average deviations at the two measuring times with each other provides an estimate of the deterioration in eye-tracking quality.

After completion of the experimental task, which lasted approximately 6 min, the experiment software transmitted the data to the experimenter's server for storage and the caregivers returned to the video conference. Caregivers were debriefed on the purpose of the experiment and were given a chance to report any issues faced during the test. The whole experiment lasted approximately 20 min.

2.6.2 | Software setup

The experiment was implemented as a webpage using a modified version of the jsPsych framework v6.3.1 (de Leeuw, 2015). To deliver this page to the participants' machines, we hosted the webpage on an Apache HTTP Server (Version 2.4; Apache Software Foundation, 2012) on a virtual machine running Ubuntu 18.04 LTS (Canonical Ltd, 2018). The participant's browser ran the code controlling the experiment to present stimuli and record the participant through the webcam. Eye-tracking was performed in real-time on the participant's device. After completing an experiment, the browser sent the data back to the Apache server, where the data was processed and saved using a script written in PHP (Version 8.0; The PHP Group, 2020).

Participating labs had the option of hosting the software on a server of their own using a comparable setup. Alternatively, they could test their participants using the preconfigured server provided by the LMU Munich lab. If they chose to do so, the experiments' software used the ManyKeys library (Steffan & Müller, 2021) to apply end-to-end encryption to the participants' data before transmitting it to the server. This step ensured that only the lab responsible for handling the specific participant's data could access the webcam recordings, enabling different labs to use the same infrastructure for testing while still keeping their participants' data fully private.

2.6.3 | General procedure

We compared the data in the current study to the data collected by Schuwerk et al. (2022). Additionally, data from our pilot study was only used to test our remote web-based eye-tracking paradigm, method feasibility, and sample size rationale, and was not included in the final data analysis.

As WebGazer runs on the participant's device, the achievable sampling rate depends on the participant's hardware capacity. Thus, the sampling rate could not be manipulated but was recorded with our setup for reporting. While we expected a sampling rate of up to 30 Hz for commonly used consumer hardware, our pilot study showed that 15–25 Hz was a more realistic estimate for most devices. Experiments with similar setups reported ranges of 4.50–25.69 Hz (Semmelmann & Weigelt, 2018).

For all videos, we defined two rectangular areas of interest (AOI) around both tunnel exits. We labeled the AOI covering the tunnel exit where the chaser will reappear according to their goal "target AOI" and the other one "distractor AOI". The software tracked whether the child's gaze fell into the left, the right, or neither AOI (Figure 3). According to tests conducted using an adult sample, a gaze

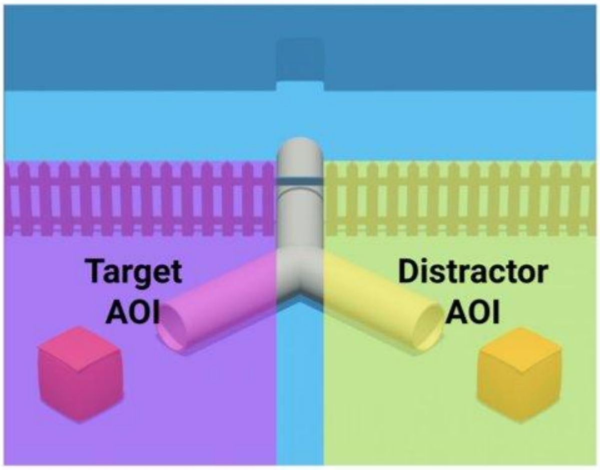


FIGURE 3 Illustration of the scene during the anticipatory period. Colored regions display AOI dimensions we used for our analyses of the web-based eye-tracking data. "Target AOI ″ was the region where the chaser reappeared according to their action goal. "Distractor AOI" was the region covering the other tunnel exit and its surroundings (Dimensions relative to the stimulus video: Left AOI: x: 0%–45%, y: 0%–66%; Right AOI: x: 55%–100%, y: 0%–66%).

point collected with WebGazer has an area of uncertainty of about 100–200 pixels on 1920 × 1080 screens in a practical setting (Papoutsaki et al., 2016). We assumed a similar area of uncertainty for our setup, which is our rationale for choosing AOIs this large (as compared to in-lab data from Schuwerk et al., 2022) for our main analysis. This constituted a necessary trade-off given the technical limitations of our approach. The child's gaze-coordinates, AOI hits, webcam videos, and miscellaneous data (screen size, browser and system information) were submitted to the experimenters' server once the trials concluded.

2.7 | Measures

The experiment consisted of only one trial type in which we manipulated the action sequences of two agents to measure goal-based action predictions via anticipatory looking. We measured the duration of children's gazes toward the target and distractor AOIs between the time the chaser entered the tunnel (first frame the chaser completely disappeared in the tunnel) and the time the chaser exited the tunnel (last frame in which the chaser was entirely inside the tunnel and not yet visible at the tunnel exit). During stimulus playback, the experiment's software sampled gaze predictions as fast as the user's device allowed for, producing the following raw data for every participant/stimulus combination: Per update of the gaze prediction, it included X and Y pixel-coordinates of the estimated gaze location on the screen, which AOIs the gaze fell into (left rectangle, right rectangle, none), and a timestamp specifying how many milliseconds had passed since the stimulus playback started. Using the height and width of the user's browser window, these data were normalized to be relative to the stimulus dimensions. Combining these normalized predictions with the stimulus and webcam video, a replay was created that overlaid the gaze location over the stimulus videos and added the synchronized webcam video in the upper-left corner. These videos were visually inspected to identify trials that had to be excluded (see exclusion criteria below). These trials were omitted from the following pre-processing steps. Participants with a sampling rate below our defined threshold (see Data exclusion) also were excluded. Using information about which AOI is defined as the "target" or "distractor" AOI for a given stimuli version (LR, LL, RR, RL), every captured gaze was classified to fall into one of three categories: "target AOI", "distractor AOI", or "no AOI" (Figure 3). We only included samples with timestamps that fell into the anticipatory period, that is, 4000 ms preceding the frame in which the chaser exited the tunnel. We then calculated for each participant what percentage of gazes fell into each category during this critical time frame aggregated across all trials (for the main hypothesis) and aggregated by trial (for the second hypothesis and the exploratory analyses). This relative percentage was necessary, as sampling rates differed between participants. We computed the proportion of looking toward the target AOI by dividing the number of samples spent looking at the target AOI by the number of samples spent looking at the target plus distractor AOIs (also referred to as total relative looking time; Senju et al., 2009): Proportional looking score = target/(target + distractor).

The score ranged between 0 and 1, whereby a score of 0 meant that the participant had exclusively looked at the distractor, a score of 1 meant that they exclusively looked at the target, and a score of 0.5 meant that they looked for an equally long duration at both AOIs (no preference). By using this proportional score, we were able to compare data across different sampling rates from individual webcams. Further, using this score we could statistically compare the web-based eye-tracking data with in-lab data by Schuwerk et al. (2022), for which we computed the same proportional differential looking score. The resulting data, which now assigned a percentage value to each participant/stimulus/AOI category combination, were used for further statistical analysis. For visualization purposes (beeswarm plots, available on OSF), the gaze data were also resampled to 15 Hz; however, the resampled data were not used to run statistical analysis.

The collected data points and the processing for the in-lab data by Schuwerk et al. (2022) were comparable, with two differences: First, gaze points were collected using dedicated eye-tracking hardware and resampled to a sampling rate of 40 Hz. Second, the AOIs for the target and distractor were defined to be smaller, as they were not subject to the size increase, we later applied to account for the lower accuracy of webcam-based eye-trackers (see Schuwerk et al., 2022 for more details).

2.8 | Data exclusion

Participants were excluded from analyses if technical problems occurred or if participants did not provide at least one useable trial after the visual inspection. Technical problems included browser freezes that halted the stimulus presentation completely (as reported by the caregiver), crashes due to the hardware being unable to handle real-time eye-tracking, issues with transmitting the data to the experimenters, corrupted data as a result of software failure, and other technical difficulties that can appear in browser-based study setups. As pre-registered, participants providing a sampling rate of 10 Hz or below were also excluded. We chose this cut-off at 1/3rd of the maximum achievable sampling rate of 30 Hz because our pilot data showed that most participants providing sample rates of 10 Hz or lower had very weak hardware, resulting in low refresh rates (around 1–2 Hz). A previous study reported a cut-off at ≤5 Hz (Yang & Krajbich, 2021), but no formal rationale for this cut-off was provided. All webcam video/gaze plot overlays (see Figure 2) were manually checked and individual trials were excluded if: (1) the caregiver interfered with the procedure (e.g., by pointing at stimuli or talking to their toddler), (2) if more than 50% of the gaze data is missing due to inattentiveness of the toddler, and/or (3) the toddler's gaze direction, judged from visual inspection of the webcam video, did not match the recorded gaze coordinates, displayed on the stimulus material as a gaze plot. Reasons for such a mismatch could include: visual properties of the environment (e.g., suboptimal lighting, movements in the background), toddler was looking away, and the gaze coordinates froze at the last location at which the toddler was looking, and/or the toddler attended to the screen, but the gaze coordinates (locations and trajectories) did not match the head and eye movements of the webcam video. Trials were also excluded if a mismatch could not be properly checked due to webcam video and recorded gaze coordinates stemming from two different webcams. WebGazer ensured during initialization that the front-facing webcam was used, but the part of the software responsible for recording the webcam footage for manual checking chose the first available connected webcam, which sometimes resulted in this mismatch in cases when two or more webcams were connected.

A third of all participants were randomly chosen and coded by a second naive rater to obtain interrater reliability (IRR). Cohen's kappa resulted in $\kappa = 0.74$, indicating a substantial inter-rater agreement. Since in our study IRR varies considerably across labs we suggest providing additional guidance for labs demonstrating low IRR in future studies.

2.9 | Statistical analyses

2.9.1 | Confirmatory analysis

All statistical analyses were carried out in R (version 4.1.1, R Core Team, 2021). To test whether participants anticipated goal-directed action outcomes in the web-based method, we measured above-chance looking toward the location that matched the outcome of the agent's action goal using a one sample *t* test. To test whether the eye-tracking method influenced the measured proportional looking score, we compared web-based eye-tracking data from the current study to lab-based eye-tracking data from the study by Schuwerk et al. (2022) in a generalized linear mixed effects model using the glmmTMB package for R (Brooks et al., 2017). This model was set to predict the proportional looking score based on the fixed effect method (web-based vs. lab-based) and a random effect for labs and participants. We also included trial number (z-transformed) as a control predictor—both as a fixed effect and a random slope within participant. Because proportions are naturally bound to be between 0 and 1, we modeled the data using a beta distribution. The model specification was:

Proportional looking score ~ method + *z*_trial + (1|lab) + (*z*_trial|participant)

A main effect of method would indicate that the way gaze data is sampled in this paradigm has an effect on the proportional looking score, suggesting that this measure of goal-directed anticipatory looking is dependent on the eye-tracking method.

To check whether exclusion rates differed between web-based and in-lab eye-tracking, we computed a Chi-square test on the 2 (web-based vs. in-lab) x 2 (percentage included vs. percentage excluded) contingency table.

2.9.2 | Exploratory analysis

To investigate potential effects of age on the proportional looking score, standardized age and trial (z-scores) were added to the model as fixed effects. Lab was included as a random effect with z_age as a random slope within lab. Participant was included as a random effect with *z*_trial as a slope within participant. The model specification was:

Proportional looking score ~ method + *z*_age + *z*_trial + (*z*_age|lab) + (*z*_trial|participant)

In addition, we analyzed the effect of the recording's sampling rate in the web-based sample on the proportional looking score in an additional model. In this model, we added age, trial and the sampling rate as fixed effects. Lab and participant were included as random effects, with *z*_age and *z*_sampling_rate as random slopes within lab and *z*_trial as a random slope within participant. The model specification was:

Proportional looking score ~ *z*_sampling_rate + *z*_trial + *z*_age + (*z*_age + *z*_sampling_rate|lab) + (*z*_trial|participant)

3 | RESULTS

3.1 | Confirmatory analysis

3.1.1 | Anticipatory looking behavior

In our web-based sample, the relative looking time toward the location that matched the outcome of the agent's action goal (target AOI; $M = 0.62$, $SD = 0.18$; Figure 4) was significantly different from chance level (0.5), $t\ (124) = 7.34$, $p < 0.001$, indicating that the participants anticipated the goal-directed action outcome. In the in-lab sample (Schuwerk et al., 2022), the average proportional looking score was 0.73 ($SD = 0.22$) and participants also showed above-chance looking toward the target AOI, $t\ (69) = 8.80$, $p < 0.001$.

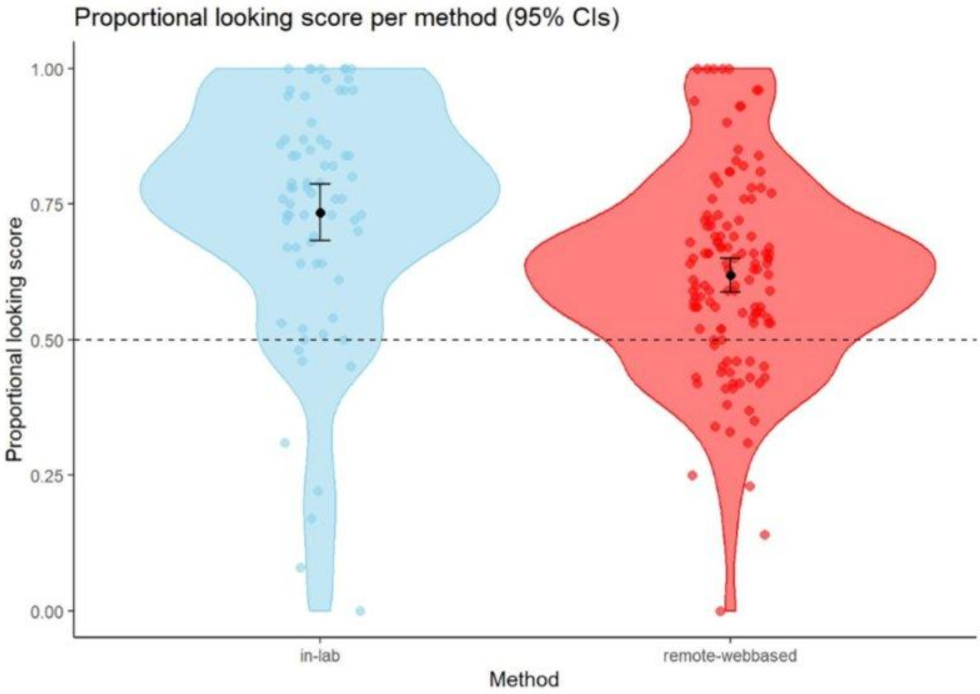


FIGURE 4 Graph depicting the proportional looking score (looking time to target AOI/looking time to target + distractor AOI) (*y* Axis) per method, remote web-based and in-lab eye-tracking (*x* Axis). The error bars represent the 95% confidence intervals (CIs).

In our web-based sample, we observed an effect-size of Cohen's $d = 0.66$ (95% confidence interval: 0.29–1.02) in the one sample directed *t* test contrasting the proportional looking score against chance level. Schuwerk et al. (2022) observed an effect size of Cohen's $d = 1.03$ (95% confidence interval: 0.50–1.56).

3.1.2 | Comparison of remote web-based versus in-lab eye-tracking in toddlers

To test whether the method had an effect on the proportional looking score, we fit a generalized linear mixed model and found a significant main effect of method ($\beta = 0.52$, $z = 4.46$, $p < 0.001$), reflecting the fact that the proportion of goal-directed anticipatory looking was higher in the in-lab sample (Figure 4).

3.1.3 | Rate of exclusion

In our web-based sample, 125 out of 216 tested participants (58%), that matched our predefined eligibility requirements, were included in the final sample. Thus, 91 participants (42%) were excluded. From these, 52 toddlers (57% of excluded participants) were excluded due to technical reasons. Caregivers had the chance to report any technical issues after completing the experimental task when they returned to the video conference meeting with the experimenter for their debriefing. Technical problems mainly occurred during the stimulus presentation or during data transmission from the participating families to the experimenters (n = 35, 67% of technical errors). Other technical reasons for exclusions were a sampling rate below our predefined threshold (n = 8, 15% of technical errors), experimenter error (n = 2, 4% of technical errors) or technical error without further information (n = 7, 13% of technical errors). As a result of the visual inspection process, a total of 39 toddlers were excluded (43% of excluded participants). They were excluded due to a mismatch between gaze coordinates and their head/eyes movement (n = 20, 51% of visual exclusions), interference by caregiver (n = 4, 10% of visual exclusions), inattentiveness of the toddler (n = 4, 10% of visual exclusions), two different active webcams (n = 6, 15% of visual exclusions), suboptimal positioning of the toddler (n = 1, 3% of visual exclusions) and error without further information (n = 4, 10% of visual exclusions). In contrast, in the in-lab sample, 70 out of 78 tested participants were included, which results in an exclusion rate of 10%. Reasons for exclusion were early termination of the experiment (n = 6) and technical problems with data collection (n = 2; Schuwerk et al., 2022). We compared web-based and in-lab exclusion rates and found a statistically significant difference, χ^2 (1, n = 294) = 24.65, $p < 0.001$. See Figure 5 for a comparison of exclusions for in-lab versus web-based methods.

3.2 | Exploratory analysis

3.2.1 | Change in tracking quality for the web-based sample

We ran calculations for x/y deviations during validation trials for all included participants (n = 125). To adjust for different screen resolutions, all values are reported as percentages relative to the screens' width and height. Across all validation trials, we found a mean deviation of 10.51% (SD = 10.18%) for x coordinates and a mean deviation of 11.83% (SD = 12.59%) for y coordinates. We performed a two-tailed t test for paired samples to compare both validation time points. We found no significant difference for either coordinate (X differences: M = 0.196%, SD = 12.539%, t (124) = 0.175, p = 0.861, delta = 0.016; Y differences: M = 1.991%, SD = 16.898%, t (124) = 1.317, p = 0.190, delta = 0.118). We thus assume that tracking quality did not deteriorate significantly during the trials (for more details

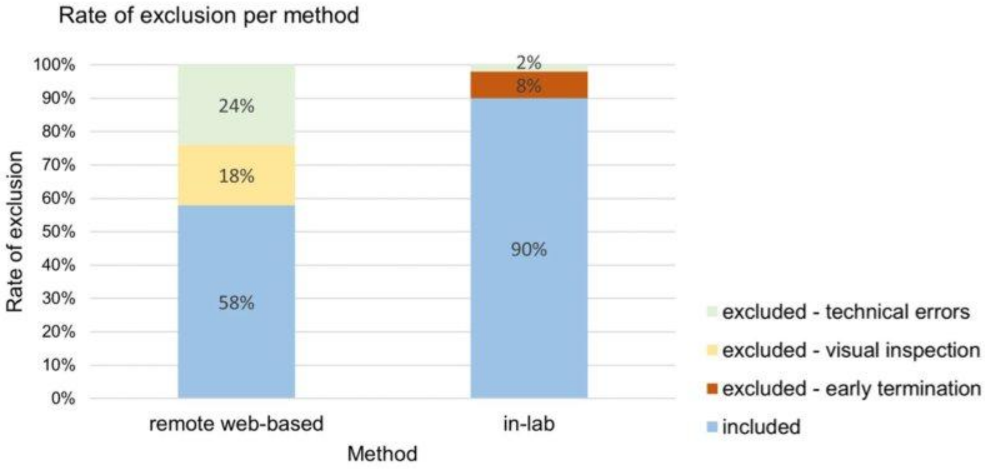


FIGURE 5 The graph depicts the rate of exclusion and reasons for exclusion (y-Axis) per method, remote web-based and in-lab eye-tracking (x-Axis).

about the participants' technical specifications see Supplementary Figure 2 and for visualizations of the first and second validation see Supplementary Figure 3).

3.2.2 | Age analysis

Using the previously described generalized linear mixed effects model, we did not find a statistically significant effect of age on the proportional looking score ($\beta = -0.05$, $z = -0.88$, $p = 0.379$), meaning that in our sample the toddler's age had no influence on anticipatory looking in the web-based task.

3.2.3 | Sampling rate analysis

We observed sampling rates between 10.42 and 40.10 Hz, resulting in a mean sampling rate of 22 Hz (SD = 7.3 Hz) in our web-based sample after exclusions. We did not find a statistically significant effect of the sampling rate on the proportional looking score ($\beta = -0.005$, $z = 0.08$, $p = 0.554$), meaning that the sampling rate had no effect on anticipatory looking in our remote sample.

4 | DISCUSSION

In the present study, we validated an open-source, remote web-based eye-tracking method for young children by replicating an anticipatory looking paradigm designed for commercial in-lab eye-trackers (Schuwerk et al., 2022). We measured anticipatory looking behavior via participants' webcams and compared our findings with results of an in-lab study. Although the eye-tracking performance in our remote web-based sample was lower and attrition rate was higher than in the in-lab sample, we successfully replicated in-lab findings, which demonstrates that remote web-based eye-tracking in toddlers is feasible. The fact that we were able to replicate the effect of goal-based anticipatory looking in multiple labs with multiple experimenters, introducing substantial variability in the data collection procedure, strengthens this conclusion. By testing children remotely and collaboratively, we were able to access participants from diverse parts of the world (Asia, Europe, North America and South America) and thus contributed an important first step in reaching more diversity in developmental research, especially in terms of a diverse cultural background.

4.1 | Measuring goal-based action prediction using remote web-based eye-tracking

We found that 18- to 27-month-olds' goal-based action predictions—reflected in above-chance looking toward the location that matches the outcome of an agent's action goal—occurred in our remotely tested sample, replicating results obtained with in-lab commercial eye-tracking systems (Schuwerk et al., 2022). This finding shows that web-based eye tracking can be used successfully to assess children's goal-based action predictions and is in line with previous studies reporting that moderated web-based test sessions with children are comparable to in-lab sessions (Chuey et al., 2021, 2022; Prein et al., 2022; Schidelko et al., 2021). Also, in line with previous remote studies in children, we found no statistically significant age effect (Chuey et al., 2022), suggesting that our web-based eye-tracking method may capture anticipatory looking behavior equally well among 18- to 27-month-olds.

4.2 | Comparing performance of web-based versus in-lab eye-tracking

We found that the eye-tracking method influenced the measured proportional looking score: the in-lab sample's mean proportional looking score toward the target location was higher than the web-based sample's score. This suggests that there may be limitations to remote web-based eye-tracking. Two main limitations of the web-based eye-tracking we used here are lower sampling rate and lower accuracy as compared to when using commercial eye-tracking systems in the lab. In the in-lab data we used for a comparison, the eye-trackers had sampling rates ranging from 60 to 500 Hz. Further, pupil-corneal reflection eye-tracking has a much higher accuracy in measuring x/y-coordinates of gaze points than the regression model WebGazer uses based on webcam videos. Although we took both these limitations into account and adjusted the AOIs in our web-based sample, we unsurprisingly still were not able to track the gaze behavior as fine-grained as in the lab. We assume that lower sampling rate and accuracy in the web-based sample led to noisier data which drove the proportional looking score toward chance-level. Replicating main findings from the lab given such added noise is thus a robust demonstration of the method.

4.3 | Comparing data quality of web-based and in-lab eye-tracking

We found support for our hypothesis that the proportion of children who contributed useable data was lower in web-based as compared to in-lab eye-tracking; this is likely largely due to poorer data quality and/or technical challenges with the remote web-based approach. Because the participating families were responsible for allowing data transmission to our servers, the dropout due to transmitting failures were particularly high. For instance, if the caregiver accidently closed the experiment's browser window after completing the last trial but before the process of data transmission was finished, the data transmission to our servers stopped. In future studies, this source of data loss could be minimized by explicitly instructing participants to keep the browser open for a longer time during the instruction, as well as using clearly visible warning displays during the data transfer. Programming a regular backup of the data during the experiment can be useful if this does not interrupt the experimental flow. Our high attrition rate in the web-based sample is in line with results of previous web-based eye-tracking studies with infants using a commercial eye-tracking platform (52% in Bánki et al., 2022), but also with adults using automated gaze coding (62% in Yang & Krajbich, 2021; 66% in Semmelmann & Weigelt, 2018). Interestingly, attrition rates in child and adult samples seem to converge when testing remotely, despite the fact that higher attrition rates are usually observed in young children compared to adults in in-lab studies using commercial eye-tracking systems (Holmqvist et al., 2023).

4.4 | Limitations

While this study examined the replicability of an in-lab paradigm, we did not explicitly measure the accuracy of WebGazer for toddlers. Using an in-lab eye-tracker concurrently while running a WebGazer experiment could provide us with a proper benchmark to compare against the inferred gaze coordinates. These data points would allow us to create accuracy measures that are directly comparable to the measures reported by Papoutsaki et al. (2016), thus providing a better idea of how the noise levels differ between infant and adult data for webcam eye-tracking.

Even though WebGazer estimates x/y gaze coordinates, the reliability of these measurements is greatly reduced by the noisy nature of the prediction. Therefore, WebGazer could suffer in designs with a larger number of AOIs, limiting the kind of studies it can be deployed in. As our present design only included two AOIs, we cannot make claims about the performance of WebGazer in these more complex scenarios. Further experiments with a larger number of AOIs need to be conducted to make definitive statements about the general usefulness of WebGazer.

To make the data of this study comparable to the in-lab sample, we used the same 4:3 aspect ratio for the stimulus material. As most computer screens today have a widescreen aspect ratio of 16:9, the stimulus material did not fill the screen's full width but left borders on both sides of the video. We replicated the findings of Schuwerk et al. (2022) under these conditions. Still, paradigms that use the full width of the screen (33% increase in presentation space) should be even less affected by the accuracy drop from using WebGazer as opposed to in-lab eye-tracking, as the horizontal eye movements would occur more clearly when a larger area is used to display the stimuli.

The range of head-to-screen distances accepted by WegGazer was large, spanning 40–130 cm (i.e., 15.7–51.2in). Calibration procedures like the one employed by WebGazer help to normalize gaze location estimations across different distances, but it is possible that children sitting particularly close to or far away from the screen could have exhibited worse tracking performance. As the head-to-screen distance was not recorded, the current study cannot determine if the highly variable distances affected the tracking quality. Additionally, this variability in distances also prevented us from calculating visual angles for our validation, which limits comparisons to other eye-tracking tools.

We decided to place two calibration quality checks in the experiment - one directly after calibration and one after the stimulus presentation. The number and placement were chosen to not overly disturb the stimulus presentation and to stay as close to the in-lab experiment as possible. However, these checks on tracking quality only compare the quality at two discrete time points. If, say, tracking quality worsened during stimulus presentation but improved to normal levels toward the end, our validation data would not report a change, even though the resulting data may be affected negatively.

Remote testing comes with an inherently higher exclusion rate than in-lab data as additional sources of errors are introduced. While software improvements could aid in lowering the attrition rate, there are many variables to control for when testing on participants' devices, such as available hardware, software characteristics like operating system, Internet connection strength, or available webcams. In addition, we recommend systematically collecting parents' feedback on technical difficulties they experienced to gain more information about potential reasons for data loss. Thus, at this point, remote testing is unlikely to reach levels comparable to in-lab studies.

Our remote sample was more diverse and global than samples from most in-person developmental studies (Singh et al., 2021), but it was still primarily a WEIRD sample (Western, Educated, Industrialized, Rich, Democratic). Thus, it is far from representing a multifaceted set of different linguistic, cultural, ethnic or socio-economic backgrounds. For example, the fact that possessing or having access to a computer is a precondition to participation already excludes large parts of the world's population. Participants outside of Western Europe or North America comprised only 11.2% of our final data, and interestingly, exclusion rates were higher in that sample (53.33% vs. 21.62%). Were we to obtain a more geographically diverse sample, we could have tested the effect of background culture on different aspects of our analysis. For example, cultural context can affect children's visual perception of scenes (e.g., Nisbett & Miyamoto, 2005). Additionally, children from different cultures might differ in their performance on different tasks. For example, Callaghan et al. (2011) found that tasks that involved pretense or graphic symbols showed cultural differences. Canadian children developed such skills sooner than Indian and Peruvian children. In ManyBabies4, which tests infants' social evaluation development, there is an attempt to tackle such cross-cultural comparisons as a spin-off project which examines cultural values and behaviors and their relation to children's social evaluation development (Wang et al., 2023).

The method used here has potential to enable research outside privileged research environments: first, by providing researchers with a low-cost eye-tracking solution, and second, by the possibility to reach participants in their homes, leveraging burdens to participate such as geographical distance to the lab or lack of time or resources to get there.

Another factor that could have influenced our remote sample is the Covid-19 pandemic. The testing took place over a period (January-August 2022) of stay-at-home restrictions. The toddlers' experience with electronic devices, the time spent with these devices, and their screen time were most likely increased during that period, and higher than the in-lab sample. This difference between the samples could have partially affected the results. However, it is important to keep in mind that at least in the last two decades, and probably before, a rapid increase in the number of electronic toys, and toys linked to electronic media, marketed for infants and toddlers (e.g., Levin & Rosenquest, 2001), and the development of the interactive mobile media technology in general (e.g., Courage et al., 2021), has led to a remarkable and widespread increase of toddlers' experience with electronic devices and their time spent with these devices, even before the Covid-19 pandemic.

4.5 | Current method in the larger context of recently emerging technical approaches

Recently, online experiment platforms such as Lookit (Scott & Schulz, 2017) have enabled remote testing of infants and toddlers using webcam video. While these platforms make it easier for labs to collect data online, they currently require manual coding of video frames to derive dependent variables. This data coding method is time-consuming when dealing with large datasets and introduces objectivity issues, so employing automated methods is desirable. However, we implemented visual inspection only as a preprocessing step to efficiently reduce noise in the data. Unlike manual coding, it involves a holistic assessment (e.g., a general impression whether the look is completely off or not by reviewing the time-locked gaze coordinates and the participant's webcam video overlayed on the stimuli), making the process more efficient, which typically does not take longer than the duration of the trial itself. As a next step, implementing an automated pre-selection process to identify trials that potentially need to be excluded would be ideal to minimize the number of videos that require visual inspection. Also, the IRR should be more consistent across labs in future studies that also implement our visual inspection process. In our study, we observed an IRR ranging from 0.5 to 1, which demonstrates a high variability across labs and results in a substantial agreement. Given that there were some labs with very high IRR, we assume that the low IRR stems from simple misunderstandings. Thus, we suggest a thorough review of the coding process with labs demonstrating low IRR.

Currently, there are several commercial online webcam-based eye-tracking platforms (e.g., Finger et al., 2017; GazeRecorder, 2010; Lewandowska, 2019). Bánki et al. (2022) used LabVanced (Finger et al., 2017) for remote eye-tracking studies with infants, but in general, these platforms have yet to be widely validated for infant research. While these platforms allow for researchers to quickly set up experiments with little programming knowledge, free, open-source approaches such as WebGazer provide considerable advantages of their own. First, the transparency of open-source code is desirable in a research context, as it allows other researchers to verify the validity of the analysis and promotes openness and accessibility, which can help democratize the scientific process and make research more inclusive. Furthermore, due to the code being available and modifiable, scientists can change the software to fit specific research needs, like making the calibration procedure more infant-friendly. This can save time and resources, as researchers can build on existing code and incorporate it into their own work, rather than starting from scratch. Lastly, the low cost of the method enables labs with fewer resources to use eye-tracking, an important factor for promoting research outside of privileged research infrastructures.

The potential of webcam-based eye-tracking is further amplified through recent applications in both educational and clinical settings (Hutt & D'Mello, 2022; Wong et al., 2023).

4.5.1 | Post hoc gaze inference

WebGazer performs real-time gaze location prediction on the participant's device, which has at least two downsides. The achievable sampling rate depends on the participant's hardware capacity and thus varies among participants. Also, real-time gaze inference requires frequent updates, limiting the complexity of the predictive models. Using more sophisticated methods or computationally expensive deep learning models to capture the face's geometry, locate the pupil, and infer gaze locations is not currently feasible in a real-time setting (Erel et al., 2022; Valliappan et al., 2020).

An alternative approach is to capture webcam footage online but run the calculations to determine gaze locations after the experiment concluded. Doing so would lift the restrictions on inference speed, and the computation of gaze location would not need to be performed on the participants' hardware.

Werchan et al. (2022) recently presented OWLET, an infant-focused webcam eye-tracking system that follows this approach, performing gaze data processing post hoc. OWLET may outperform WebGazer on some dimensions. For instance, the best-performing inference models of WebGazer achieve an average error of 4.17° in an adult sample with a controlled calibration (Papoutsaki et al., 2016). OWLET reported mean absolute x/y calibration deviations of 3.36°/2.67° across infants with a simpler, infant-friendly calibration.

While our study validated WebGazer exclusively on PCs, OWLET can also infer gaze location from video captured on tablet computers and mobile devices. In a study testing the robustness of OWLET, the authors found higher socioeconomic and racial/ethnic diversity in their sample using mobile devices compared to laptops (Werchan et al., 2022). The ability to run eye-tracking studies on these devices would, therefore, be desirable for projects aiming to diversify samples, such as the ones under the ManyBabies framework (Frank et al., 2017; Visser et al., 2021).

On the other hand, our setup is more flexible and easier to use than the OWLET. Whereas WebGazer can be configured to allow any calibration scheme and webcam format, OWLET by default only allows a fixed four-point calibration and only processes 16:9 webcam videos with a framerate of 30 frames per second or higher. Moreover, WebGazer can be plugged into any online experiment set up with jsPsych to produce inferred gaze coordinates without additional post hoc processing through dedicated software. This advantage is important for big team science collaborations like ManyBabies, for example, by reducing the need for additional software installations for all participating labs. Furthermore, given that WebGazer provides real-time tracking, and assuming enough computational power, only WebGazer could be adapted to create infant-controlled experiments.

In sum, when choosing a web-based eye-tracking solution, researchers must consider these tradeoffs based on their resources and paradigm. With further work on streamlining the process, a system can be built that utilizes the improved accuracy of OWLET with the convenience and flexibility that WebGazer provides.

4.5.2 | Deep learning

While WebGazer and OWLET use traditional computer vision algorithms to extract facial information and map them to screen coordinates based on regressions and polynomial functions respectively, applying end-to-end deep learning algorithms trained on large datasets shows great potential for webcam eye-tracking. Valliappan et al. (2020) used deep-learning models to achieve gaze-tracking accuracy for adults comparable to specialized eye-tracking software using only a smartphone's front camera. Unfortunately, the software they developed is not openly available and needs to be reimplemented to be used in experiments. Furthermore, their training data exclusively consisted of adults, so the generalizability to infant footage remains unknown. Nonetheless, their results show the potential of webcam-based eye-tracking through deep learning algorithms.

iCatcher+ also uses deep learning algorithms to classify gaze points into either left, right, or away (Erel et al., 2022). The model was trained on a hand-labeled dataset of infant webcam footage. iCatcher+ reaches gaze coding accuracy comparable to that of human coders, making it a viable choice for paradigms with binary dependent variables. Until deep learning solutions for x/y coordinate inference from webcam footage are created, online studies that require more fine-grained paradigms have to rely on tools like OWLET or WebGazer.

5 | CONCLUSION

Web-based eye-tracking can be used to capture toddlers' goal-based action anticipation. Thus, in-lab findings can be replicated using remote webcam-based testing, which provides children and their caregivers with a more comfortable participation experience in their natural environment. In developmental research, eye-tracking is commonly performed using in-lab pupil-corneal reflection eye-tracking. While this specialized hardware enables high gaze tracking accuracy that software-only solutions cannot match, it comes with substantially higher costs and physical and/or social boundaries that are hard to overcome. Collecting eye-tracking data remotely using common computers and WebGazer substantially reduces the cost of running experiments, makes testing young participants less time-consuming and more flexible, while providing the opportunity to test demographically diverse, large international samples under comparable conditions. For experiments in which the benefits of remote testing are substantial, such as with children, and a reduced spatial resolution can be tolerated, web-based webcam eye-tracking using WebGazer is a promising method.

AFFILIATIONS

[1]Department of Psychology, Ludwig-Maximilians-Universität München, München, Germany

[2]Facultad de Psicología, Universidad Nacional Autónoma de México, Ciudad de México, México

[3]Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[4]Institute of Psychology, Leuphana University Lüneburg, Lüneburg, Germany

[5]Faculty of Arts & Science, Ambrose University, Calgary, Alberta, Canada

[6]Department of Psychology and Cognitive Science, University of Trento, Trento, Italy

[7]Department of Developmental Psychology, University of Göttingen, Göttingen, Germany

[8]Research Group Milestones of Early Cognitive Development, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

[9]Department of Psychology, The University of British Columbia, Vancouver, British Columbia, Canada

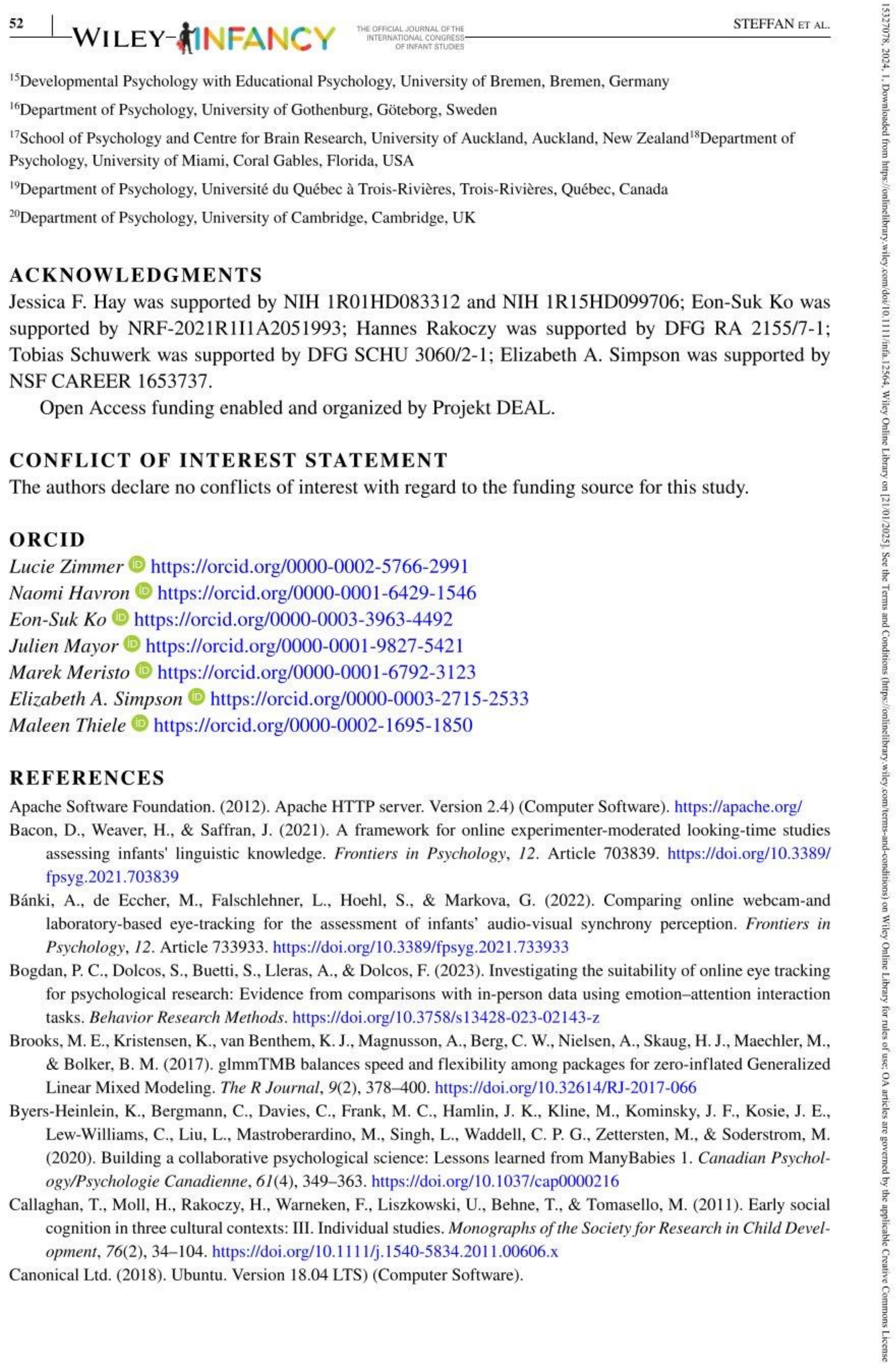
[10]School of Psychological Sciences & Center for the Study of Child Development, University of Haifa, Haifa, Israel

[11]Department of Psychology, University of Tennessee, Knoxville, Tennessee, USA

[12]Department of Psychology, University of Oslo, Oslo, Norway

[13]Department of Psychology, James Madison University, Harrisonburg, Virginia, USA

[14]Department of English Language and Literature, Chosun University, Gwangju, Korea



[15]Developmental Psychology with Educational Psychology, University of Bremen, Bremen, Germany

[16]Department of Psychology, University of Gothenburg, Göteborg, Sweden

[17]School of Psychology and Centre for Brain Research, University of Auckland, Auckland, New Zealand[18]Department of Psychology, University of Miami, Coral Gables, Florida, USA

[19]Department of Psychology, Université du Québec à Trois-Rivières, Trois-Rivières, Québec, Canada

[20]Department of Psychology, University of Cambridge, Cambridge, UK

ACKNOWLEDGMENTS

Jessica F. Hay was supported by NIH 1R01HD083312 and NIH 1R15HD099706; Eon-Suk Ko was supported by NRF-2021R1I1A2051993; Hannes Rakoczy was supported by DFG RA 2155/7-1; Tobias Schuwerk was supported by DFG SCHU 3060/2-1; Elizabeth A. Simpson was supported by NSF CAREER 1653737.

Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest with regard to the funding source for this study.

ORCID

Lucie Zimmer https://orcid.org/0000-0002-5766-2991
Naomi Havron https://orcid.org/0000-0001-6429-1546
Eon-Suk Ko https://orcid.org/0000-0003-3963-4492
Julien Mayor https://orcid.org/0000-0001-9827-5421
Marek Meristo https://orcid.org/0000-0001-6792-3123
Elizabeth A. Simpson https://orcid.org/0000-0003-2715-2533
Maleen Thiele https://orcid.org/0000-0002-1695-1850

REFERENCES

Apache Software Foundation. (2012). Apache HTTP server. Version 2.4) (Computer Software). https://apache.org/

Bacon, D., Weaver, H., & Saffran, J. (2021). A framework for online experimenter-moderated looking-time studies assessing infants' linguistic knowledge. *Frontiers in Psychology*, *12*. Article 703839. https://doi.org/10.3389/fpsyg.2021.703839

Bánki, A., de Eccher, M., Falschlehner, L., Hoehl, S., & Markova, G. (2022). Comparing online webcam-and laboratory-based eye-tracking for the assessment of infants' audio-visual synchrony perception. *Frontiers in Psychology*, *12*. Article 733933. https://doi.org/10.3389/fpsyg.2021.733933

Bogdan, P. C., Dolcos, S., Buetti, S., Lleras, A., & Dolcos, F. (2023). Investigating the suitability of online eye tracking for psychological research: Evidence from comparisons with in-person data using emotion–attention interaction tasks. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02143-z

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated Generalized Linear Mixed Modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.32614/RJ-2017-066

Byers-Heinlein, K., Bergmann, C., Davies, C., Frank, M. C., Hamlin, J. K., Kline, M., Kominsky, J. F., Kosie, J. E., Lew-Williams, C., Liu, L., Mastroberardino, M., Singh, L., Waddell, C. P. G., Zettersten, M., & Soderstrom, M. (2020). Building a collaborative psychological science: Lessons learned from ManyBabies 1. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 349–363. https://doi.org/10.1037/cap0000216

Callaghan, T., Moll, H., Rakoczy, H., Warneken, F., Liszkowski, U., Behne, T., & Tomasello, M. (2011). Early social cognition in three cultural contexts: III. Individual studies. *Monographs of the Society for Research in Child Development*, *76*(2), 34–104. https://doi.org/10.1111/j.1540-5834.2011.00606.x

Canonical Ltd. (2018). Ubuntu. Version 18.04 LTS) (Computer Software).

Chuey, A., Asaba, M., Bridgers, S., Carrillo, B., Dietz, G., Garcia, T., Leonard, J. A., Liu, S., Merrick, M., Radwan, S., Stegall, J., Velez, N., Woo, B., Wu, Y., Zhou, X. J., Frank, M. C., & Gweon, H. (2021). Moderated online data-collection for developmental research: Methods and replications. *Frontiers in Psychology*, *12*. Article 734398. https://doi.org/10.3389/fpsyg.2021.734398

Chuey, A., Boyce, V., Cao, A., & Frank, M. C. (2022). Conducting developmental research online vs. in-person: A meta-analysis. *PsyArXiv*. https://doi.org/10.31234/osf.io/qc6fw

Courage, M. L., Frizzell, L. M., Walsh, C. S., & Smith, M. (2021). Toddlers using tablets: They engage, play, and learn. *Frontiers in Psychology*, *12*. Article 564479. https://doi.org/10.3389/fpsyg.2021.564479

Dalrymple, K. A., Manner, M. D., Harmelink, K. A., Teska, E. P., & Elison, J. T. (2018). An examination of recording accuracy and precision from eye tracking data from toddlerhood to adulthood. *Frontiers in Psychology*, *9*. Article 803. https://doi.org/10.3389/fpsyg.2018.00803

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Erel, Y., Shannon, K. A., Chu, J., Scott, K. M., Kline Struhl, M., Cao, P., Tan, X., Hart, P., Raz, G., Piccolo, S., Mei, C., Potter, C., Jaffe-Dax, S., Lew-Williams, C., Tenenbaum, J., Fairchild, K., Bermano, A., & Liu, S. (2022). iCatcher+: Robust and automated annotation of infant's and young children's gaze direction from videos collected in laboratory, field, and online studies. *PsyArXiv*. https://doi.org/10.31234/osf.io/up97k

Falck-Ytter, T., Gredebäck, G., & von Hofsten, C. (2006). Infants predict other people's action goals. *Nature Neuroscience*, *9*(7), 878–879. https://doi.org/10.1038/nn1729

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In I. A. Sekerina, E. Fernandez, & H. Clahsen (Eds.), *Developmental Psycholinguistics: On-line methods in children's language processing* (pp. 97–135). John Benjamins.

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). *LabVanced: A unified JavaScript framework for online studies [conference paper]*. International Conference on Computational Social Science IC2S2S. (Germany).

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S., & Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. https://doi.org/10.1111/infa.12182

GazeRecorder (2010). GazeRecorder (Computer Software) https://gazerecorder.com/

Holmqvist, K., Örbom, S. L., Hooge, I. T. C., Niehorster, D. C., Alexander, R. G., Andersson, R., Benjamins, J. S., Blignaut, P., Brouwer, A.-M., Chuang, L. L., Dalrymple, K. A., Drieghe, D., Dunn, M. J., Ettinger, U., Fiedler, S., Foulsham, T., van der Geest, J. N., Hansen, D. W., Hutton, S. B., …, & Hessels, R. S. (2023). Eye tracking: Empirical foundations for a minimal reporting guideline. *Behavior Research Methods*, *55*(1), 364–416. https://doi.org/10.3758/s13428-021-01762-8

Hutt, S., & D'Mello, S. K. (2022). Evaluating calibration-free webcam-based eye tracking for gaze-based user modeling. In *Proceedings of the 2022 international conference on multimodal interaction* (pp. 224–235).

Kominsky, J. F., Begus, K., Bass, I., Colantonio, J., Leonard, J. A., Mackey, A. P., & Bonawitz, E. (2021). Organizing the methodological toolbox: Lessons learned from implementing developmental methods online. *Frontiers in Psychology*, *12*. Article 702710. https://doi.org/10.3389/fpsyg.2021.702710

Leshin, R., Leslie, S.-J., & Rhodes, M. (2021). Does it matter how we speak about social kinds? A large, pre-registered, online experimental study of how language shapes the development of essentialist beliefs. *Child Development*, *92*(4), e531–e547. https://doi.org/10.1111/cdev.13527

Levin, D. E., & Rosenquest, B. (2001). The increasing role of electronic toys in the lives of infants and toddlers: Should we be concerned? *Contemporary Issues in Early Childhood*, *2*(2), 242–247. https://doi.org/10.2304/ciec.2001.2.2.9

Lewandowska, B. (2019). RealEye eye-tracking system technology whitepaper. Retrieved https://support.realeye.io/realeye-accuracy/ 19 December 2022.

Liszkowski, U., Carpenter, M., & Tomasello, M. (2007). Pointing out new news, old news, and absent referents at 12 months of age. *Developmental Science*, *10*(2), F1–F7. https://doi.org/10.1111/j.1467-7687.2006.00552.x

Lo, C. H., Mani, N., Kartushina, N., Mayor, J., & Hermes, J. (2021). e-Babylab: an open-source browser-based tool for unmoderated online developmental studies. *PsyArXiv*. https://doi.org/10.31234/osf.io/u73sy

Luo, Y., & Baillargeon, R. (2007). Do 12.5-month-old infants consider what objects others can see when interpreting their actions? *Cognition*, *105*(3), 489–512. https://doi.org/10.1016/j.cognition.2006.10.007

Nelson, C. M., & Oakes, L. M. (2021). "May I grab your attention?": An investigation into infants' visual preferences for handled objects using Lookit as an online platform for data collection. *Frontiers in Psychology*, *12*. Article 3866. https://doi.org/10.3389/fpsyg.2021.733218

Nisbett, R. E., & Miyamoto, Y. (2005). The influence of culture: Holistic versus analytic perception. *Trends in Cognitive Sciences*, *9*(10), 467–473. https://doi.org/10.1016/j.tics.2005.08.004

Oakes, L. M. (2012). Advances in eye tracking in infancy research. *Infancy*, *17*(1), 1–8. https://doi.org/10.1111/j.1532-7078.2011.00101.x

Oliver, B. R., & Pike, A. (2021). Introducing a novel online observation of parenting behavior: Reliability and validation. *Parenting*, *21*(2), 168–183. https://doi.org/10.1080/15295192.2019.1694838

Ozkan, A. (2018). Using eye-tracking methods in infant memory research. *The Journal of Neurobehavioral Sciences*, *5*, 62–66.

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer: Scalable webcam eye tracking using user interactions. In *Proceedings of the 25th international joint conference on artificial intelligence* (pp. 3839–3845). IJCAI).

Prein, J. C., Bohn, M., Kalinke, S., & Haun, D. B. M. (2022). Tango: A reliable, open-source, browser-based task to assess individual differences in gaze understanding in 3 to 5-year-old children and adults. *PsyArXiv*. https://doi.org/10.31234/osf.io/vghw8

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., Benitez, J., & Ocampo, J. D. (2020). Advancing developmental science via unmoderated remote research with children. *Journal of Cognition and Development*, *21*(4), 477–493. https://doi.org/10.1080/15248372.2020.1797751

Schidelko, L. P., Schünemann, B., Rakoczy, H., & Proft, M. (2021). Online testing yields the same results as lab testing: A validation study with the false belief task. *Frontiers in Psychology*, *12*. Article 703238. https://doi.org/10.3389/fpsyg.2021.703238

Schneegans, T., Bachman, M. D., Huettel, S. A., & Heekeren, H. (2021). Exploring the potential of online webcam-based eye tracking in decision-making research and influence factors on data quality. *PsyArXiv*. https://doi.org/10.31234/osf.io/zm3us

Schuwerk, T., Kampis, D., Baillargeon, R., Biro, S., Bohn, M., Byers-Heinlein, K., Dörrenberg, S., Fisher, C., Franchin, L., Fulcher, T., Garbisch, I., Geraci, A., Grosse Wiesmann, C., Hamlin, K., Haun, D. B. M., Hepach, R., Hunnius, S., Hyde, D. C., Karman, P., …, & rakoczy, h. (2022). Action anticipation based on an agent's epistemic state in toddlers and adults. *PsyArXiv*. https://doi.org/10.31234/osf.io/x4jbm

Scott, K., & Schulz, L. (2017). Lookit (Part 1): A new online platform for developmental research. *Open Mind*, *1*(1), 4–14. https://doi.org/10.1162/OPMI_a_00002

Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods*, *50*(2), 451–465. https://doi.org/10.3758/s13428-017-0913-7

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An absence of spontaneous theory of mind in Asperger syndrome. *Science*, *325*(5942), 883–885. https://doi.org/10.1126/science.1176170

Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., Fei-Fei, L., Keil, F. C., Gweon, H., Tenenbaum, J. B., Jara-Ettinger, J., Adolph, K. E., Rhodes, M., Frank, M. C., Mehr, S. A., & Schulz, L. (2020). Online developmental science to foster innovation, access, and impact. *Trends in Cognitive Sciences*, *24*(9), 675–678. https://doi.org/10.1016/j.tics.2020.06.004

Singh, L., Cristia, A., Karasik, L. B., Rajendra, S. J., & Oakes, L. (2021). Diversity and Representation in Infant Research: Barriers and bridges towards a globalized science of infant development. *PsyArXiv*. https://doi.org/10.31234/osf.io/hgukc

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*(7), 587–592. https://doi.org/10.1111/j.1467-9280.2007.01944.x

Steffan, A., & Müller, T. (2021). ManyKeys. Version 1.0. Computer Software) https://github.com/adriansteffan/manykeys/tree/bed46cdaf3cb8a578c6277eff669b0abb36c3a26

Su, I. A., & Ceci, S. (2021). Zoom Developmentalists": *Home-based videoconferencing developmental research during COVID-19*. *PsyArXiv*. https://doi.org/10.31234/osf.io/nvdy6

The PHP Group. (2020). PHP. Version 8.0) (Computer Software).

Valliappan, N., Dai, N., Steinberg, E., He, J., Rogers, K., Ramachandran, V., Xu, P., Shojaeizadeh, M., Guo, L., Kohlhoff, K., & Navalpakkam, V. (2020). Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature Communications*, *11*(1), 4553. Article 4553. https://doi.org/10.1038/s41467-020-18360-5

Venker, C. E., & Kover, S. T. (2015). An open conversation on using eye-gaze methods in studies of neurodevelopmental disorders. *Journal of Speech, Language, and Hearing Research*, *58*(6), 1719–1732. https://doi.org/10.1044/2015_JSLHR-L-14-0304

Visser, I., Bergmann, C., Byers-Heinlein, K., Dal Ben, R., Duch, W., Forbes, S., Franchin, L., Frank, M. C., Geraci, A., Hamlin, J. K., Kaldy, Z., Kulke, L., Laverty, C., Lew-Williams, C., Mateu, V., Mayor, J., Moreau, D., Nomikou, I., Schuwerk, T., … Zettersten, M. (2022). Improving the generalizability of infant psychological research: The ManyBabies model. *Behavioral and Brain Sciences*, *45*, e35. Article e35. https://doi.org/10.1017/S0140525X21000455

Wang, Y., Alpcetin, B., Zhu, J., Buyurucu, G., Sancar, B. H., Kaya, M. E., Dresel, M., Exner, A., Hamlin, J. K., Havron, N., Henderson, A., Martin, A., Partridge, T. T., Schuwerk, T., Shainy, M. R., Su, Y., Tsang, C. K. A., Uzefovsky, F., Wong, T. T.-Y., … Lucca, K. (2023). *Individual differences in infants' social evaluations across cultures: A spin-off project of many babies*. CEO Conference on Cognitive Development. [Poster Presentation]. The 13th Budapest https://osf.io/jp532

Wass, S. V. (2016). The use of eye-tracking with infants and children. In J. Prior & J. Van Herwegen (Eds.), *Practical research with children* (1st ed., pp. 24–45). Routledge. https://doi.org/10.4324/9781315676067

Wass, S. V., Smith, T. J., & Johnson, M. H. (2013). Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behavior Research Methods*, *45*(1), 229–250. https://doi.org/10.3758/s13428-012-0245-6

Werchan, D. M., Thomason, M. E., & Brito, N. H. (2022). Owlet: An automated, open-source method for infant gaze tracking using smartphone and webcam recordings. *Behavior Research Methods*, *55*(6), 3149–3163. https://doi.org/10.3758/s13428-022-01962-w

Wong, A. Y., Bryck, R. L., Baker, R. S., Hutt, S., & Mills, C. (2023). Using a webcam based eye-tracker to understand students' thought patterns and reading behaviors in neurodivergent classrooms. In *LAK23: 13th international learning analytics and knowledge conference* (pp. 453–463).

Xu, P., Ehinger, K. A., Zhang, Y., Finkelstein, A., Kulkarni, S. R., & Xiao, J. (2015). TurkerGaze: Crowdsourcing saliency with webcam based eye tracking. arXiv http://arxiv.org/abs/1504.06755

Yang, X., & Krajbich, I. (2021). Webcam-based online eye-tracking for behavioral research. *Judgment and Decision Making*, *16*(6), 1485–1505. https://doi.org/10.1017/S1930297500008512

Yoder, P. J., Lloyd, B. P., & Symons, F. J. (2018). *Observational measurement of behavior* (2nd ed.). Paul H. Brookes.

Zaadnoordijk, L., Buckler, H., Cusack, R., Tsuji, S., & Bergmann, C. (2021). A global perspective on testing infants online: Introducing ManyBabies-AtHome. *Frontiers in Psychology*, *12*. Article 703234. https://doi.org/10.3389/fpsyg.2021.703234

Zeng, G., Simpson, E. A., & Paukner, A. (2023). Maximizing valid eye-tracking data in human and macaque infants by optimizing calibration and adjusting areas of interest. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-02056-3

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Steffan, A., Zimmer, L., Arias-Trejo, N., Bohn, M., Dal Ben, R., Flores-Coronado, M. A., Franchin, L., Garbisch, I., Grosse Wiesmann, C., Hamlin, J. K., Havron, N., Hay, J. F., Hermansen, T. K., Jakobsen, K. V., Kalinke, S., Ko, E.-S., Kulke, L., Mayor, J., Meristo, M., … Schuwerk, T. (2024). Validation of an open source, remote web-based eye-tracking method (WebGazer) for research in early childhood. *Infancy*, *29*(1), 31–55. https://doi.org/10.1111/infa.12564

Supporting information: **infa12564-sup-0001-suppl-data.docx**

Appendix B. Manuscript Study 2

Registered Report

Action Anticipation Based on an Agent's Epistemic State in Toddlers and Adults

Tobias Schuwerk[1†], Dora Kampis[2], Nicolás Alessandroni[3], Nicole Altvater-Mackensen[4], Natalia Arias-Trejo[5], Emma L. Axelsson[6], Renée Baillargeon[7], Anna-Elisabeth Baumann[8,3], Cyann Bernard[9], Szilvia Biro[10,11], Tashauna L. Blankenship[12], Isa Blomberg[13], Manuel Bohn[14,15], Elisabeth E. F. Bradford[16], Krista Byers-Heinlein[3], Irene Canudas Grabolosa[17], Emily M. Chen[18,19], Xiaoyun Chen[20], John Corbit[21], Cynthia Fisher[7], Samuel H. Forbes[22], Laura Franchin[23], Tess Fulcher[24], Alessandra Geraci[25], Nayeli Gonzalez-Gomez[26], Paul Grohmann[15], Charlotte Grosse Wiesmann[27,28], J. Kiley Hamlin[29], Daniel Haun[15], Naomi Havron[30,31], Robert Hepach[32], Tone K. Hermansen[33], Mikołaj Hernik[34], Michael Huemer[17], Sabine Hunnius[35], Daniel C. Hyde[36], Sagi Jaffe-Dax[37,38], Krisztina V. Jakobsen[39], Natalia Kartushina[40], Coral Kfir[37], Manar Khalaila[37], Osman S. Kingo[41], Mona M. Klau[42], Ada Koleini[33], Abhinav Kona[43], Shannon P. Kong[26], Heather L. Kosakowski[17,44], Ágnes M. Kovács[45], Anna Krämer[46], Judith L. Krief[2], Peter Krøjgaard[41], Louisa Kulke[47], Crystal Y. Lee[48], Edward W. Legg[49,50], Tiffany S. Leung[51], Casey Lew-Williams[48], Yixun Li[52], Ulf Liszkowski[53], Liquan Liu[54,55], Yiyu Liu[7], Alexander Mackiel[24], Kyle Mahowald[56], Milena Marx[57], Olivier Mascaro[9], Magda Matetovici[58], Marlena Mayer[53], Julien Mayor[33], Marek Meristo[59], Marlene Meyer[35], David Moreau[60], Hay Mar Myat Kyaw[52], Andreea A. Nistor[32], Josef Perner[46], Stefanie Peykarjou[57,61], Que Anh Pham[12], Diane Poulin-Dubois[3], Lindsey J. Powell[62], Julia C. Prein[14,15], Beate Priewasser[63], Marina Proft[13], Alyssa A. Quinn[6], Gal Raz[18], Peter J. Reschke[64], Josephine Ross[65], Audun Rosslund[33,66], Katrin Rothmaler[67], Petra Šarić[49], Rebecca Saxe[18], Karola Schlegelmilch[68,69], Dana Schneider[70], Melanie S. Schreiner[13], Verena Schuhmann[4], Lital Shapiro[37], Elizabeth A. Simpson[51], Lauren Smith[62], Trine Sonne[41], Victoria Southgate[2], Adrian Steffan[1], Yanjie Su[71], Luca Surian[23], Alvin W. M. Tan[19], Anna-Lena Tebbe[67,72], Angeline Sin Mei Tsui[19], Ingmar Visser[42], Yanwei Wang[71], Annie E. Wertz[73,69], Gert Westermann[20], Amanda L. Woodward[24], Yukun Yu[36], Francis Yuen[29], Amanda Rose Yuile[74,36], Zhen Zeng[75,76], Martin Zettersten[77], Lucie Zimmer[1], Michael C. Frank[19], Hannes Rakoczy[13]

[1]Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany, [2]Department of Psychology, University of Copenhagen, Copenhagen, Denmark, [3]Department of Psychology, Concordia University, Montréal, Canada, [4]Psycholinguistics, School of Humanities, University of Mannheim, Mannheim, Germany, [5]Facultad de Psicología, Universidad Autónoma de México, Mexico City, México, [6]School of Psychological Sciences, University of Newcastle, Australia, [7]Department of Psychology, University of Illinois Urbana-Champaign, Champaign, USA, [8]University of Calgary, Calgary, Canada, [9]Université Paris Cité, INCC UMR 8002, CNRS, F-75006 Paris, France, [10]Leiden Institute of Education and Child Studies, Leiden University, Leiden, The Netherlands, [11]Leiden Institute for Brain and Cognition, Leiden University, Leiden, The Netherlands, [12]Department of Psychology, University of Massachusetts Boston, Boston, USA, [13]Department of Psychology, University of Göttingen, Göttingen, Germany, [14]Institute of Psychology in Education, Leuphana University Lüneburg, Lüneburg, Germany, [15]Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, [16]School of Psychology & Neuroscience, University of St Andrews, Scotland, UK,

[17]Department of Psychology, Harvard University, Cambridge, USA, [18]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, USA, [19]Department of Psychology, Stanford University, Stanford, USA, [20]Department of Psychology, Lancaster University, Lancaster, UK, [21]Psychology Department, St. Francis Xavier University, Antigonish, Canada, [22]Department of Psychology, Durham University, Durham, UK, [23]Department of Psychology and Cognitive Sciences, University of Trento, Trento, Italy, [24]Department of Psychology, University of Chicago, Chicago, USA, [25]Department of Educational Sciences, University of Catania, Catania, Italia, [26]Centre for Psychological Research, Oxford Brookes University, Oxford, UK, [27]Department of Liberal Arts and Sciences, University of Technology Nuremberg, Nuremberg, Germany, [28]Research Group Milestones of Early Cognitive

Development, Max Planck Institute for Cognitive and Brain Sciences, Leipzig, Germany, [29]Department of Psychology, University of British Columbia, Vancouver, Canada, [30]School of Psychological Sciences, Faculty of Social Sciences, University of Haifa, Haifa, Israel, [31]The Institute of Information Processing and Decision Making, University of Haifa, Haifa, Israel, [32]Department of Experimental Psychology, University of Oxford, Oxford, UK, [33]Department of Psychology, University of Oslo, Oslo, Norway, [34]Department of Psychology, UiT The Arctic University of Norway, Tromsø, Norway, [35]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands, [36]Department of Psychology, University of Illinois Urbana-Champaign, Champaign, United States, [37]School of Psychological Sciences, Tel Aviv University, Tel Aviv, Israel, [38]Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel, [39]Department of Psychology, James Madison University, Harrisonburg, USA, [40]Institute for Linguistics and Scandinavian Studies, University of Oslo, Oslo, Norway, [41]Department of Psychology and Behavioural Sciences, Aarhus University, Aarhus, Denmark, [42]Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands, [43]Department of BioSciences, Rice University, Houston, USA, [44]Department of Psychology, University of Southern California, Los Angeles, USA, [45]Department of Cognitive Science, Central European University, Vienna, Austria, [46]Centre for Cognitive Neuroscience & Department of Psychology, University of Salzburg, Salzburg, Austria, [47]Department of Developmental Psychology with Educational Psychology, University of Bremen, Bremen, Germany, [48]Department of Psychology, Princeton University, Princeton, USA, [49]Division of Cognitive Sciences, University of Rijeka, Rijeka, Croatia, [50]Centre for Mind and Behaviour, University of Rijeka, Rijeka, Croatia, [51]Department of Psychology, University of Miami, Coral Gables, USA, [52]Department of Early Childhood Education, The Education University of Hong Kong, Hong Kong, [53]Institute of Psychology, Universität Hamburg, Hamburg, Germany, [54]Graduate School of Health, University of Technology Sydney, Sydney, Australia, [55]School of Psychology, Western Sydney University, Sydney, Australia, [56]Department of Linguistics, The University of Texas at Austin, Austin, USA,

[57]Institute of Psychology, Heidelberg University, Heidelberg, Germany, [58]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, [59]Department of Psychology, University of Gothenburg, Gothenburg, Sweden, [60]School of Psychology, University of Auckland, Auckland, New Zealand, [61]Developmental and Pedagogical Psychology, Charlotte Fresenius Hochschule Wiesbaden, Wiesbaden, Germany, [62]Department of Psychology, University of California, San Diego, La Jolla, USA, [63]Institute Early Life Care, Paracelsus Medical University, Salzburg, Austria, [64]School of Family Life, Brigham Young University, Provo, UT, USA, [65]Department of Psychology, University of Dundee, Scotland, UK, [66]Department of Linguistics and Scandinavian Studies, University of Oslo, Oslo, Norway, [67]Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany, [68]Institute of Biology, Department of Human Biology and Primate Cognition, Leipzig University, Leipzig, Germany, [69]Max Planck Research Group Naturalistic Social Cognition, Max Planck Institute for Human Development, Berlin, Germany, [70]Friedrich-Schiller-University Jena, Jena, Germany, [71]School of Psychological and Cognitive Sciences, Peking University, Beijing, China, [72]Department of Psychology, University of Florida, Gainesville, USA, [73]Department of Psychological & Brain Sciences, University of California, Santa Barbara, Santa Barbara, USA, [74]Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, USA, [75]Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Hong Kong SAR, China, [76]MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, Australia, [77]Department of Cognitive Science, University of California, San Diego, La Jolla, USA

Author Note

* Shared co-first authorship

The registered Stage 1 manuscript unchanged from the point of IPA can be found on the Open Science Framework: https://osf.io/7ubkw. We have no known conflicts of interests to disclose. Data are available at the following URL: https://osf.io/jmuvd/. The analytic code necessary to reproduce the preregistered analyses presented in this paper is available at https://github.com/manybabies/mb2-analysis. The materials necessary to attempt to replicate the findings presented here are publicly accessible at the following URL:https://osf.io/pd35f/files.

Correspondence concerning this article should be addressed to Tobias Schuwerk, Ludwig-Maximilians-Universität München, Leopoldstr. 13, 80802 München, Germany. Email: tobias.schuwerk@psy.lmu.de

Acknowledgements

K.V.J. was supported by Alvin V. Jr. and Nancy C. Baird Professorship; J.P. was supported by Austrian Science Fund FWF I 3518-G24; A.K. was supported by the Austrian Science Fund FWF W1233-B and FWF I 3518; M. Huemer was supported by CareyFund, Harvard University; N.A. was supported by Concordia Horizon Postdoctoral Fellowship, FRQSC Postdoctoral Fellowship; U.L. was supported by DFG FOR2253, LI 1989/3–2; H.R. was supported by DFG RA 2155/7-1; T. Schuwerk was supported by DFG SCHU 3060/2-1; G.W. was supported by ESRC ES/S007113/1; O.M. was supported by FoundTrust, ANR-21-CE28-0017-01; E.W.L. was supported by Marie Sklodowska-Curie grant agreement No 794270; T.S.L. and E.A.S. were supported by NSF CAREER Award 1653737; Y.S. and Y.W. were supported by National Natural Science Foundation of China 32071075; M.P. was supported by RA 2155/7-1; J.C. was supported by RGPIN-2023-05954; A.-E.B. was supported by SSHRC Insight grant (# 435-2017-0564); J.K.H. was supported by SSHRC Partnership Development Grant 890-2020-0059; K.B.-H. was supported by Social Sciences and Humanities Research Council of Canada award #890-2020-0059; D.P.-D. was supported by Social Sciences and Humanity Research Council of Canada , Grant #435-2022-0805; L.L. was supported by The Foundation of Graduates in Early Childhood Studies, Forest Hill Foundation; C.F. was supported by UIUC Campus Research Board grant RB21096; A.E.W. was supported by funding from the Max Planck Society.

We thank William Fabricius for his helpful comments on the results. We are also grateful to Adeline Depierreux, Cathy Dai, Rémi Hagedorn, Heleen Lange, Ariel Pinto, Anne Schuit, Amber

Welker, Giulia Vigna, and Yukun Yu for their support with data collection, and to Luise Diekmann (née Zellner) for her help with project administration and preparation for data collection at the ToM Lab Salzburg.

Abstract

Do toddlers and adults engage in spontaneous Theory of Mind (ToM)? Evidence from anticipatory looking (AL) studies suggests they do. But a growing body of failed replication studies raised questions about the paradigm's suitability, urging the need to test the robustness of AL as a spontaneous measure of ToM. In a multi-lab collaboration we examine whether 18- to 27-month-olds' and adults' anticipatory looks distinguish between two basic forms of epistemic states: knowledge and ignorance. In adults (n = 703 included, 68 % female), we found clear support for epistemic state-based action anticipation: they engaged in simple goal-based action anticipation in pilot studies, and clearly differentiated between knowledge and ignorance conditions in the main study as predicted. In toddlers (n = 521 included, 49 % female), in contrast, the results were less clear. They did engage in simple goal-based action anticipation in pilot studies, but did not show the clear differentiation between knowledge and ignorance conditions in the main study as predicted. Future research with adults can now move on to probe whether their spontaneous action anticipation is also sensitive to more complex kinds of epistemic states, such as true and false beliefs. Future research with toddlers will first need to investigate more systematically the source of the puzzling findings in the present study and clarify whether they indicate competence or mere performance limitations.

Keywords: anticipatory looking; spontaneous Theory of Mind; replication

Word count: 15907

Action Anticipation Based on an Agent's Epistemic State in Toddlers and Adults

The capacity to represent epistemic states, known as Theory of Mind (ToM) or mentalizing, plays a central role in human cognition (Dennett, 1989; Frith & Frith, 2006; Premack & Woodruff, 1978). Although ToM has been under intense scrutiny in the past decades, its nature and ontogeny are still the subjects of much controversy. At the heart of these debates are questions about the reliability of the paradigms used to measure ToM (Baillargeon et al., 2018; Poulin-Dubois et al., 2018), including, among others, anticipatory looking (AL) paradigms. To address this issue, in a collaborative long-term project we assess the robustness of infants' and adults' tendency to spontaneously take into account different kinds of epistemic states—what they perceive, know, think, or believe—when predicting others' behavior. This paper reports the first foundational step of this project, which focuses on the most basic epistemic state ascription: the capacity to distinguish between knowledgeable and ignorant individuals. Simple forms of knowledge attribution (such as tracking what other individuals have seen or experienced) are typically assumed to develop early and to operate spontaneously throughout the lifespan (Liszkowski et al., 2007; Luo & Baillargeon, 2007; O'Neill, 1996; Phillips et al., 2021). Thus, evaluating whether ToM measures are sensitive to the knowledge-ignorance distinction is a crucial test case to assess their robustness. The present paper investigates this question in an AL paradigm including 18-27-month-old infants and adults.

In the following sections we first establish the background and scientific context of this study, namely the reliability and replicability of spontaneous ToM measures. We then introduce a novel way to approach these issues: a large-scale collaborative project targeting the replicability of ToM findings. Finally, we outline the rationale of the present study which uses an AL paradigm to test whether infants and adults distinguish between two basic forms of an agent's epistemic state: knowledge and ignorance.

Spontaneous Theory of Mind tasks

Humans are proficient at interpreting and predicting others' intentional actions. Adults as well as infants expect agents to act persistently towards the goal they pursue (Csibra & Gergely, 2007; Gergely & Csibra, 2003; Gergely et al., 1995, Woodward & Sommerville, 2000), and they anticipate others' actions based on their goals even before goals are achieved - that is, humans engage in goal-based action anticipation (for review, see Elsner & Adam, 2021; but see Ganglmayer et al., 2019). To predict others' actions, however, it is essential to consider their epistemic state: what they perceive, know, or believe. A number of seminal studies using non-verbal spontaneous measures have suggested that infants, toddlers, older children, and adults show action anticipation and action understanding not only based on other agents' goals (what they want) but also on the basis of their epistemic status (what they perceive, know, or believe). These studies suggest that from infancy onwards, humans spontaneously engage in ToM or mentalizing. For example, studies using the violation of expectation paradigm have demonstrated that infants look longer in response to events in which an agent acts in ways that are incompatible with their (true or false) beliefs, compared to events in which they act in belief-congruent ways (Onishi & Baillargeon, 2005; Surian et al., 2007; Träuble et al., 2010). Other studies have employed more interactive tasks requiring the child to play, communicate, or cooperate with experimenters and, for example, give an experimenter one of several objects as a function of their epistemic status. Such studies have shown that toddlers spontaneously adjust their behavior to the experimenter's beliefs (Buttelmann et al., 2009; Király et al., 2018; Knudsen & Liszkowski, 2012; Southgate et al., 2010).

The largest body of evidence for spontaneous ToM comes from studies using AL tasks. In such tasks, participants see an agent who acts in pursuit of some goal (typically, to collect a certain object) and has either a true or a false belief (for example, regarding the location of the target object). A number of studies have shown that infants, toddlers, older children, neurotypical adults, and even non-human primates anticipate (as indicated by looks to the location in question) that an agent will go where it (truly or falsely) believes the object to be, irrespective of the actual location of the object (Gliga et al., 2014; Grosse Wiesmann et al., 2017; Hayashi et al., 2020; Kano et al., 2019; Krupenye et al., 2016; Meristo et al., 2012; Schneider et al., 2012; Schneider et al., 2013; Senju et al., 2009; Senju et al., 2010; Senju et al., 2011; Surian & Franchin, 2020; Thoermer et al., 2012). These studies have revealed converging evidence for spontaneous ToM across the human lifespan and even in other primate species.

Across the different measures, the majority of early works on spontaneous ToM in infants and toddlers have reported positive results in the second year of life, and a few studies even within the first year (Kovács et al., 2010; Luo, 2011; Southgate & Vernetti, 2014), yielding a rich body of coherent and convergent evidence (for reviews see e.g., Barone et al., 2019; Kampis et al., 2020; Scott & Baillargeon, 2017). This growing body of literature has led to a theoretical transformation of the field. In particular, findings with young infants have paved the way for novel accounts of the development and cognitive foundations of ToM. The previous consensus was that full-fledged ToM emerges only at around age 4, potentially as the result of developing executive functions, complex language skills and other factors (e.g., Perner, 1991; Wellman & Cross, 2001). In contrast, the newer accounts proposed that some basic forms of ToM may be phylogenetically more ancient and may develop much earlier in ontogeny (e.g., Baillargeon et al., 2010; Carruthers, 2013; Kovács, 2016; Leslie, 2005).

Recently, however, a number of studies have raised uncertainty regarding the empirical foundations of the early-emergence theories, as we review below. In the following sections, we present an overview of the current empirical picture of early understanding of epistemic states and then introduce ManyBabies2 (MB2), a large-scale collaborative project exploring the replicability of ToM in infancy, of which the current study constitutes the first step.

Replicability of Spontaneous Theory of Mind Tasks

A number of failures to replicate findings from spontaneous ToM tasks have recently been published with infants, toddlers, and adults (e.g., Burnside et al., 2018; Dörrenberg et al., 2018; Grosse Wiesmann et al., 2017; Grosse Wiesmann et al., 2018; Kampis et al., 2021; Kulke, von Duhn, et al., 2018; Kulke & Hinrichs, 2021; Kulke, Johannsen, & Rakoczy, 2019; Kulke & Rakoczy, 2017, 2019; Kulke, Reiß, et al., 2018; Kulke, Wübker, & Rakoczy, 2019; Powell et al., 2018; Priewasser et al., 2020; Priewasser et al., 2018; Schuwerk et al., 2018; for overviews, see Barone et al., 2019; Kulke & Rakoczy, 2018). Besides conceptual replications, many of these studies involve more direct replication attempts with the original stimuli and procedures. One of these was a two-lab replication attempt of one of the most influential AL studies (Southgate et al., 2007). This failure to replicate is especially notable not only because of the influence of the original finding of the field, but also because of the large sample size and the involvement of some of the original authors (Kampis et al., 2021). Additional unpublished replication failures have also been reported. Kulke and Rakoczy (2018) examined 65 published and non-published studies including 36 AL studies (replications of Low & Watts, 2013; Schneider et al., 2012; Southgate et al., 2007; Surian & Geraci, 2012), as well as studies using other paradigms, and classified them as a successful, partial, or non-replication, depending on whether all, some, or none of the original main effects were found. Although no formal analysis of effect size was carried out, overall, non-replications and partial replications outnumbered successful replications, regardless of the method used.

In addition to the failure to replicate spontaneous anticipation of agents' behaviors based on their beliefs, many of the replication studies revealed an even more fundamental problem of spontaneous AL procedures: a failure to adequately anticipate an agent's action in the absence of a belief. That is, researchers did not find evidence for spontaneous anticipation of agents' behaviors based on their goals, even in the initial familiarization trials of the experiments (e.g., Kampis et al., 2020; Kulke, Reiß, et al., 2018; Schuwerk et al., 2018). The familiarization trials are designed to convey the goal of the agent, as well as the general timing and structure of events, to set up participants' expectations in the test trials where the agent's epistemic state is then manipulated. Typically, the last familiarization trial can also be used to probe participants' spontaneous action anticipation; and test trials can only be meaningfully interpreted if there is evidence of above-chance anticipation in the familiarization trials. In several AL studies many participants had to be excluded from the main analyses for failing to demonstrate robust action anticipation during the familiarization trials (e.g., Kampis et al., 2020; Kulke, Reiß, et al., 2018; Schuwerk et al., 2018; Southgate et al., 2007). This raises the possibility that these AL paradigms may not be suitable for reliably eliciting spontaneous action prediction in the first place (for discussion see Baillargeon et al., 2018; Kampis et al., 2021). In light of the complex and mixed state of the evidence, researchers called for a systematic, large-scale, multi-lab study to stringently test the robustness, reliability, and replicability of AL and other spontaneous measures of ToM.

General Rationale of MB2

To this end, MB2 was established as an international consortium dedicated to investigating infants' and toddlers' ToM skills. The main aim is to test the replicability and thus reliability of findings from spontaneous ToM tasks. In the long-term, MB2 will build on the initial findings and the aim will be extended to include testing the validity of these experimental designs and addressing theoretical accounts of spontaneous ToM. MB2 operates under the general umbrella of ManyBabies (MB), a large-scale international research consortium founded with the aim of probing the reliability of central findings from infancy research. In particular, MB projects bring together large and theoretically diverse groups of researchers to tackle pressing questions of infant cognitive development, by collaboratively designing and implementing methodologies and pre-registered analysis plans (Frank et al., 2017). The MB2 consortium involves authors of original studies as well as authors of both successful and failed replication studies, and researchers from very different theoretical backgrounds. It thus presents a case of true "adversarial collaboration" (Mellers et al., 2001).

Rationale of the Present Study

Based on both theoretical and practical considerations, the current paper presents the first foundational step in MB2, focusing on AL measures. It investigates whether toddlers and adults anticipate (in their looking behavior) how other agents will act based on their goals (i.e., what they want) and epistemic status (i.e., what they know or do not know). From a practical perspective, we focus on AL since it is a child-friendly and widely used method that is also suitable for humans across the lifespan and even other species. Additionally, as AL is screen-based and standardizable, identical stimuli can be presented in different labs. From a theoretical perspective, given the mixed findings with AL tasks reviewed in the previous section, we take a systematic and bottom-up approach.

First, we probe whether AL measures are suitable for measuring spontaneous goal-directed action anticipation. With the aim of improving the low overall rates of anticipatory looks in recent studies, we designed new, engaging stimuli to test whether these are successful in eliciting spontaneous action anticipation. Second, in case reliably elicited action anticipation can be found: we probe whether toddlers and adults take into account the agent's epistemic status in their spontaneous goal-based action anticipation. That is, do they track whether the agent saw or did not see a crucial event, and therefore whether this agent does or does not know something? In the current study we focus on the most basic form of tracking the epistemic status of agents: considering whether they had access to relevant information, and whether they are thus *knowledgeable* or *ignorant*. We reasoned that only after establishing whether a context can elicit spontaneous tracking of an agent's epistemic status in a more basic sense (i.e., the agent's knowledge vs. ignorance) is it eventually meaningful to ask whether this context also elicits more complex epistemic state tracking (i.e., the agent's beliefs).

Answering these first two questions in the present study will allow us, in the long run, to address a third set of questions in subsequent studies, probing the nature of the representations and cognitive mechanisms involved in infant ToM. Do toddlers and adults engage in full-fledged belief-ascription in their spontaneous goal-based action anticipation? What *kind* of epistemic states do toddlers and adults spontaneously attribute to others in their action anticipation (e.g., Horschler et al., 2020; Phillips et al., 2021)? Do the results that prove replicable really assess

ToM, or can they be interpreted in alternative ways such as behavioral rules, associations, or simple perceptual preferences (see, e.g., Heyes, 2014; Perner & Ruffman, 2005)? The present study lays the foundation for investigating these questions.

Regarding the knowledge-ignorance distinction, many accounts in developmental and comparative ToM research have argued for the ontogenetic and evolutionary primacy of representing *what* agents witness and represent, relative to more sophisticated ways of representing *how* agents represent (and potentially mis-represent) objects and situations (e.g., Apperly & Butterfill, 2009; Flavell, 1988; Kaminski et al., 2008; Martin & Santos, 2016; Perner, 1991; Phillips et al., 2021). For example, it is often assumed that young children and non-human primates may be capable of so-called "Level I perspective-taking" (understanding *who* sees *what*) but only human children from around age 4 may finally develop capacities for "Level II perspective-taking" (understanding *how* a given situation may appear to different agents; Flavell et al., 1981). Empirically, many studies using verbal and/or interactive measures have indicated that children may engage in knowledge-ignorance and related distinctions before they engage in more complex forms of meta-representation (e.g., Flavell et al., 1981; Hogrefe et al., 1986; Moll & Tomasello, 2006; O'Neill, 1996; Buttelmann & Kovács, 2019; Buttelmann et al., 2015; Kampis et al., 2020; though for some findings indicating Level II perspective-taking at an early age see Forgács et al., 2019; Scott & Baillargeon, 2009; Scott et al., 2015), and that non-human primates seem to master knowledge-ignorance tasks while not demonstrating any more complex, meta-representational form of ToM (e.g., Hare et al., 2001; Kaminski et al., 2008; Karg et al., 2015). The knowledge-ignorance distinction thus appears to be an ideal candidate for assessing epistemic status-based action anticipation in a wide range of populations.

To date, however, no study has probed whether or how children's (and adults') spontaneous action anticipation, as indicated by AL, is sensitive to ascriptions of knowledge vs. ignorance. Most studies that have addressed ToM with AL measures have targeted the more sophisticated true/false belief contrast. As reviewed above, the results of those studies yield a mixed picture regarding replicability of the findings. It has been argued that tasks that reliably replicate are ones which can be solved with the more basic knowledge-ignorance distinction, whereas tasks that do not replicate require more sophisticated belief-ascription (Powell et al., 2018)[5], suggesting that only some findings might not be replicable. Based on these considerations, the present study tests whether toddlers and adults engage in knowledge- and ignorance-based AL to probe the most basic form of spontaneous, epistemic state-based action anticipation.

Design and Predictions of the Present Study

The current study presented 18- to 27-month-old toddlers and adults with animated scenarios while measuring their gaze behavior. Testing adults (and not just toddlers) is crucial to address debates about the validity and interpretation of AL measures of ToM throughout the lifespan (e.g., Schneider et al., 2017). Following the structure of previous AL paradigms, participants were first familiarized to an agent repeatedly approaching a target (familiarization trials). AL was measured during the familiarization trials to probe whether participants understood the agent's goal and spontaneously anticipate their actions. Subsequently, during the test trials the agent's visual access was manipulated, leading them to be either knowledgeable or ignorant about the location of the target. Participants' AL was measured during the test trials to determine whether or not they take into account the agent's epistemic access and adjust their action anticipation accordingly. Participants' looking patterns were recorded using either lab-based corneal reflection eye-tracking or online recording of gaze patterns. We chose to provide the online testing option to increase the flexibility for data collection given the disruption caused by the COVID-19 pandemic. This option also provided the opportunity to potentially compare in-lab and online testing procedures (Sheskin et al., 2020; see Section S4 of Supplemental Material).

[5] For example, some studies have found partial replication results, with patterns of the following kind: participants showed systematic anticipation (or appropriate interactive responses) in true belief trials but showed looking (or interactive responses) at chance level in the false belief trials (e.g., Dörrenberg et al., 2019; Kulke, Reiß, et al., 2018; Powell et al., 2018). Such a pattern remains ambiguous since it may merely reflect a knowledge-ignorance distinction.

Novel animated stimuli were collectively developed within the MB2 consortium on the basis of previous work (e.g., Clements & Perner, 1994) and based on input from collaborators with experience with both successful and failed replication studies (e.g., Grosse Wiesmann et al., 2017; Surian & Geraci, 2012). These animated 3D scenes featured a dynamic interaction aimed to optimally engage participants' attention: a chasing scenario involving two agents, a *chaser* and a *chasee* (see Figures 1 and 2). As part of the chase, the chasee enters from the top of an upside-down Y-shaped tunnel with two boxes at its exits. The tunnel is opaque so participants cannot see the chasee after it enters the tunnel, but can hear noises that indicate movement. The chasee eventually exits from one of the arms of the Y, and goes into the box on that side. The chaser observes the chasee exit the tunnel and go into a box, and then follows it through the tunnel. During familiarization trials, the chaser always exits the tunnel on the same side as the chasee, and approaches the box where the chasee is currently located. Thus, if participants engage in spontaneous action anticipation during familiarization trials, they should reliably anticipate during the period when the chaser is in the tunnel that it will emerge at the exit that leads to the box containing the chasee.

During test trials, the chasee always first hides in one of the boxes but shortly thereafter leaves its initial hiding place and hides in the box at the other tunnel exit. Critically, the chaser either does (*knowledge* condition) or does not (*ignorance* condition) have epistemic access to the chasee's location. During knowledge trials, the chaser observes all movements of the chasee. During ignorance trials, the chaser observes the chasee enter the tunnel, but then leaves and only returns after the chasee is hidden inside the second box. The event sequences in the two conditions are thus identical with the only difference between conditions pertaining to what the chaser has or has not seen. They were designed in this way with the long-term aim to implement, in a minimal contrast design, more complex conditions of false/true belief contrasts with the very same event sequences (true belief conditions will then be identical to the knowledge conditions here, but in false belief conditions the chaser witnesses the chasee's placement in the first box, but then fails to witness the re-location)[6].

Participants' AL (their gaze pattern indicating where they expect the chaser to appear) will be assessed during the anticipatory period - that is, the period during which the chaser is going through the tunnel and is not visible. There will be two main dependent measures: first looks, and a differential looking score (DLS). The first look measure will be binary, indicating which of the two tunnel exits participants fixate first: the exit where the chasee is actually hiding, or the other exit. DLS is a measure of the proportion of time spent looking at the correct tunnel exit during the entire anticipatory period.

In two pilot studies (see Methods section), we addressed the foundational question of the current study: whether these stimuli reveal spontaneous goal-directed action anticipation as measured by AL in the above-described familiarization trials (i.e., without a change of location by the chasee or manipulation of the chaser's epistemic state). We found that our paradigm indeed elicited action anticipation and exclusion rates due to lack of anticipation were significantly lower relative to previous (original and replication) AL studies. Both toddlers and adults showed reliable anticipation of the chaser's exit at the chasee's location, indicating that in contrast with many previous AL studies the current paradigm successfully elicits spontaneous goal-based action anticipation. Based on these pilot data we concluded that the paradigm is suitable for examining the second and critical question: whether toddlers and adults, in their spontaneous goal-based action anticipation, take into account the agent's epistemic state.

[6] There is thus a certain asymmetry with regard to the interpretation and the consequences of potentially positive and negative results of the present knowledge-ignorance contrast: in the case of positive results, we can conclude that participants spontaneously engage in basic epistemic state ascription and can move on to test, with the minimal contrast comparison of knowledge-ignorance vs. false belief-true belief, whether this extends to more complex forms of epistemic state attribution. In the case of negative results, though, we cannot draw firm conclusions to the effect that participants do not engage in spontaneous epistemic state ascription. More caution is in order since the present knowledge-ignorance contrast has been designed in order to be comparable to future belief contrasts rather than to be the simplest implementation possible. Simpler implementations would then need to be devised that involve fewer steps (i.e. the chasee just goes to one location and this is or is not witnessed by the chasee).

We predict that if participants track the chaser's perceptual access and resulting epistemic state (knowledge/ignorance) and anticipate their actions accordingly, they should look more in anticipation to the exit at the chasee's location than the other exit in the knowledge condition, but should not do so (or to a lesser degree; see below) in the ignorance condition. We anticipate three potential factors that could influence participant's gaze patterns: Keeping track of the chaser's epistemic status in the ignorance condition might either lead to no expectations as to where the chaser will look (resulting in chance level looking between the two exits) or (if participants follow an "ignorance leads to mistakes"-rule, see e.g., Ruffman, 1996) to an expectation that the chaser will go to the wrong location (longer looking to the exit with the empty box; e.g., Fabricius et al., 2010). Either way, participants may still show a 'pull of the real' even in the ignorance condition, i.e., reveal a default tendency to look to the side where the chasee is located. But if they truly keep track of the epistemic status of the chaser (knowledge vs. ignorance), they should show this tendency to look to the side where the chasee really is in the ignorance condition to a lesser degree than in the knowledge condition.

In sum, the research questions of the present study are the following: First, can we observe in a large sample that toddlers and adults robustly anticipate agents' actions based on their goals in this paradigm, as they did in our pilot study? Second, can we find evidence that they take into account the agent's epistemic access (knowledge vs. ignorance) and adjust their action anticipation accordingly? In addressing these questions, the present study will significantly contribute to our knowledge on spontaneous ToM. It will inform us whether the present paradigm and stimuli can elicit spontaneous goal-based and mental-state-based action anticipation in adults and toddlers, based on a large sample of 1224 participants in total from 47 labs. In the long run, the present study will lay the foundation for future work to address broader questions of what *kind* of epistemic states toddlers and adults spontaneously attribute to others in their action anticipation and what cognitive mechanisms allow them to do so.

Methods

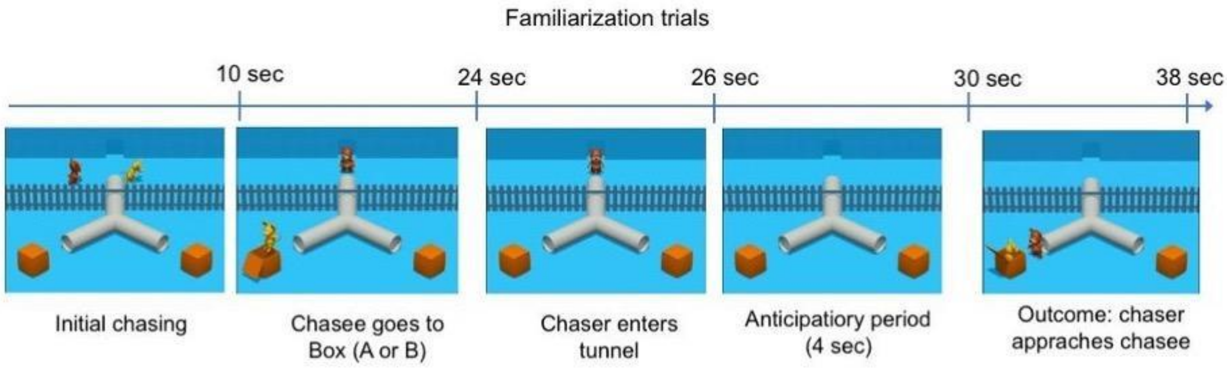
All materials, and the collected de-identified data, are provided on the Open Science Framework (OSF; https://osf.io/jmuvd/). All analysis scripts, including the pilot data analysis and simulations for the design analysis, can be found on GitHub (https://github.com/manybabies/mb2-analysis). We report how we determined our sample size and we report all data exclusions, all manipulations, and all measures in the study. Additional methodological details can be found in the Supplemental Material.

Stimuli

Figures 1 and 2 provide an overview of the paradigm. For the stimuli, 3D animations were created depicting a chasing scenario between two agents (chaser and chasee) who start in the upper part of the scene. At the very top of the scene a door leads to outside the visible scene. Below this area, a horizontal fence separates the space, and thus the lower part of the space can be reached by the Y-shaped tunnel only. Additional information on the general scene setup, events, and timings in the familiarization and the test trials, as well as trial randomization can be found in the Supplemental Material.

Figure 1

Timeline of the familiarization trials



Note. The last picture illustrates the reunion of the two agents, after the chaser left the tunnel, approached the closed box (with chasee still inside) and knocked on it.

Familiarization Trials

All participants viewed four familiarization trials (for an overview of key events see Figure 1). During the familiarization trials, after a brief chasing introduction, the chasee enters an upside-down Y-shaped tunnel with a box at both of its exits. The chasee then leaves the tunnel through one of the exits and hides in the box on the corresponding side. Subsequently, the chaser enters the tunnel (to follow the chasee), and participants' AL to the tunnel exits is measured before the chaser exits on the side the chasee is hiding, as an index of their goal-based action anticipation. In these familiarization trials, if participants engage in spontaneous action anticipation, they should reliably anticipate that the chaser should emerge at the tunnel exit that leads to the box where the chasee is. After leaving the tunnel, the chaser approaches the box in which the chasee is hiding and knocks on it. Then, the chasee jumps out of the box and the two briefly interact.

Familiarization Phase Pilot Studies. In a pilot study with 18- to 27-month-olds (*n* = 65) and adults (*n* = 42), seven labs used in-lab corneal reflection eye-tracking to collect data on gaze behavior in the familiarization phase using eight trials. A key desideratum of our paradigm was that it should produce sufficient AL, as a low rate of AL in previous studies had led to high exclusion rates. The goals of the pilot study were to 1) estimate the level of correct goal-based action predictions in the familiarization phase, 2) determine the optimal number of familiarization trials, 3) check for issues with perceptual properties of stimuli (e.g., distracting visual saliencies), and 4) test the general procedure including preprocessing and analyzing raw gaze data from different eye-tracking systems. We found that the familiarization stimuli elicited a relatively high proportion of goal-directed action anticipations, but we were concerned about the effects of some minor properties of the stimulus (in particular, a small rectangular window in the tunnel tube that allowed participants to see the agents at one point on their path to the tunnel exits).

In a second pilot study with 18- to 27-month-olds (*n* = 12, three participating labs), slight changes of stimulus features (the removal of the window in the tube; temporal changes of auditory anticipation cue) did not cause major changes in the AL rates.

Sixty-eight percent of toddlers' first looks in the first pilot, 69% of toddlers' first looks in the second pilot, and 69% of adults' first looks were toward the correct area of interest (AOI) during the anticipatory period. The average proportion of looking towards the correct AOI during the anticipatory period was 70.7% ($CI_{95\%}$ = 67.6% - 73.8%) in toddlers in the first pilot, 70.5% ($CI_{95\%}$ = 62.8% - 78.2%) in the second pilot for toddlers, and 75.3% ($CI_{95\%}$ = 71.0% - 79.5%) in adults. In Bayesian analyses, we found strong evidence that toddlers and adults looked more towards the target than towards the distractor during the anticipation period. Based on conceptual and practical methodological considerations while also considering previous studies, we decided to include four trials in the final experiment. The pilot data results of the toddlers supported this decision insofar as we observed a looking bias towards the correct location already in trials 1-4, without additional benefit of trials 5-8.

Further, prototypical analysis pipelines were established for combining raw gaze data from different eye-trackers. In short, we developed a way to resample gaze data from different eye-trackers to be at a common Hz rate and to define proportionally correct AOIs for different screen dimensions with the goal to merge all raw data into one data set for inferential statistics. The established analysis procedure is described further in the Data Preprocessing section below. In sum, we concluded that this paradigm sufficiently elicits goal-directed action predictions, an important prerequisite for drawing any conclusion on AL behavior in the test trials of this study. A detailed description of the two pilot studies can be found in the Supplemental Material.

Test Trials

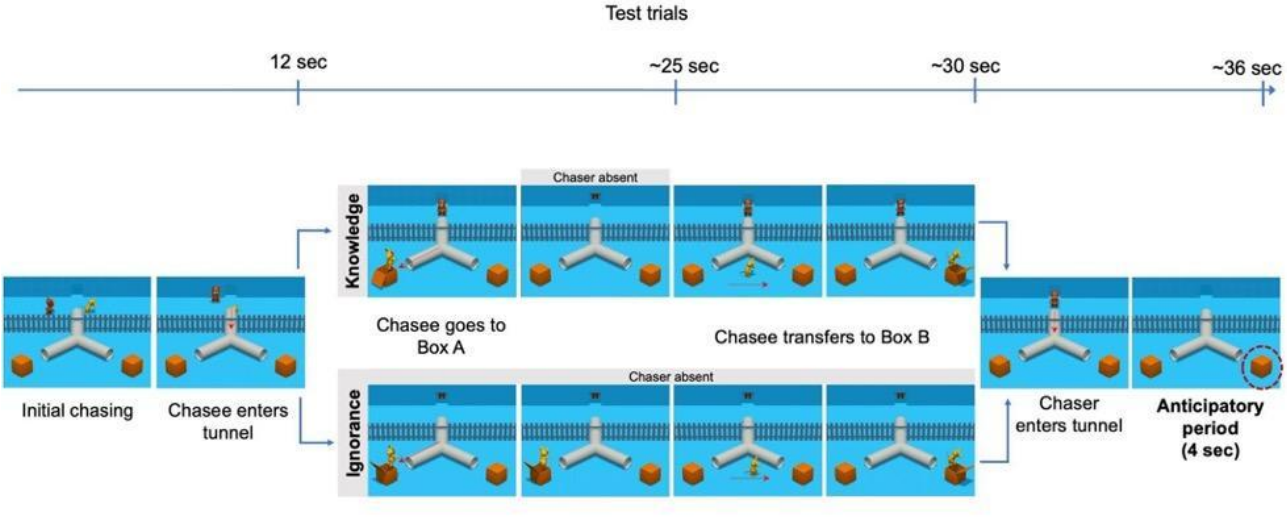
All participants saw two test trials, one knowledge and one ignorance trial. However, in line with common practice in ToM studies, the main comparison concerned the first test trial between-participants to avoid potential carryover effects. In addition, in exploratory analyses, we assessed whether results remain the same if both trials were taken into account and whether gaze patterns differed between the two trials (see Exploratory Analyses). If the results remained largely unchanged across the two trials, it might suggest that future studies could increase power by including multiple test trials.

In the test trials, the chasee first hid in one of the boxes, but shortly thereafter the chasee left this box and hid in the second box, at the other tunnel exit. Critically, the chaser either witnessed (knowledge condition) or did not witness (ignorance condition) from which tunnel exit the chasee exited and thus where the chasee was currently hiding (for an overview, see Figure 2). In the knowledge trial, the chaser observed all movements of the chasee. The chaser leaves for a brief period of time after the chasee exited the tunnel, but returns before the chasee changes location from one box to the other. Therefore, no events took place in the chaser's absence. In the ignorance trial, the chaser saw the chasee enter the tunnel, but then left. Therefore, the chaser did not see the chasee entering either box and only returned once the chasee was already hidden in the final location. Finally, the chaser entered the tunnel but did not appear in either exit. Rather, the scene "froze" for four seconds and participants' AL was measured. Thus, the knowledge and ignorance conditions were matched for the chaser leaving for a period of time, but they differed in whether they warranted the chaser's epistemic access to the location of the chasee. No outcome was shown in either test trial.

When designing the knowledge and ignorance conditions, we tried to keep all events and their timings parallel, except for the crucial manipulation. We showed the same events in both conditions. Where possible, all events also had the same duration. In the case of the chaser's absence in the knowledge condition, there were two main options, both with inevitable trade-offs. First, we could have increased the duration of the chaser's absence in the knowledge condition to equate the duration of the chaser's absence in both conditions. Yet, this would have potentially disrupted the flow of events, such as keeping track of the chasee's actions and the general scene dynamics, since nothing would have happened for a substantial amount of time. Second, the chaser could have been absent for a shorter time in the knowledge than in the ignorance condition, in which case the flow of events – the chasee's actions and the general scene dynamics – would remain natural. We chose the second option because we reasoned that the artificial break in the knowledge condition could have disrupted the participant's tracking of the chaser's epistemic state, thus being a confound that would have been more detrimental than the difference in the duration of absence. Further, the current contrast had the advantage that the chasee's sequence and timing of actions were identical in both conditions, thus minimizing the difference between conditions. Finally, with the current design, the duration of the chaser's absence is closely matched in the later planned false belief - true belief contrast, because in the future false belief condition, the chaser has to be absent for fewer events (because the chaser witnesses the first hiding events after the chasee reappeared at the other side of the tunnel).

Figure 2

Schematic overview of stimuli and conditions of the test trials



Note. After the familiarization phase, participants knew about the agent's goal (chaser wants to find chasee), perceptual access (chaser can see what happens on the other side of the fence), and situational constraints (boxes can be reached by walking through the forking tunnel). In the knowledge condition, the chaser witnessed the chasee walking through the tunnel and jumping in and out of the first box. While the chasee was in the box, the chaser briefly left the scene through the door in the back and returned shortly after. Subsequently, the chaser watched the chasee jumping out of the box again and hiding in the second box. In the ignorance condition, the chaser turned around and stood on the other side of the door in the back of the scene, thus unable to witness any of the chasee's actions. The chaser then returned and entered the tunnel to look for the chasee. During the test phase (4 seconds still frame), AL towards the end of the tunnels was measured.

Trial Randomization

We varied the starting location of the chasee (left or right half of the upper part of the scene) and the box the chasee ended up (left or right box) in both familiarization and test trials. The presentation of the familiarization trials was counterbalanced in two pseudo-randomized orders (LRRL and RLLR). Each lab signed up for one or two sets of 16-trial-combinations, for each of their tested age groups.

Lab Participation Details

Time-Frame

The contributing labs started data collection as soon as they were able to once our Registered Report received an in-principle acceptance. We originally anticipated submitting the study for Stage 2 review one year after in-principle acceptance. However, this timeline required adjustments due to several factors. First, the stimuli underwent additional modifications, which extended the preparation phase. Additionally, setting up the study for multiple eye-trackers and their respective software further contributed to delays. Furthermore, varying COVID-19 restrictions such as closures and reopening schedules across different labs introduced additional challenges and delayed the start of data collection. The most significant source of delay, however, was the data pre-processing phase, which took longer than expected. These combined factors necessitated a longer timeline before submission for Stage 2 review.

Participation Criterion

The participating labs were recruited from the MB2 consortium. In July 2020, we asked via the MB2 listserv which labs planned to contribute how many participants for the respective age group (toddlers and/or adults). Each lab made a commitment to collecting data from at least 16 participants (toddlers or adults), but we did not exclude any contributed data on the basis of the total sample size contributed by that lab. Labs were allowed to test using either in-lab eye-tracking or online methods.

Ethics

All labs were responsible for obtaining ethics approval from their appropriate institutional review board. The labs contributed de-identified data for central data analysis (i.e., eye-tracking raw data/coded gaze behavior, demographic information). Video recordings of the participants were stored at each lab according to the approved local data handling protocol. If allowed by the local institutional review board, video recordings were made available to other researchers via the video library Databrary (https://nyu.databrary.org/).

Participants

In a preliminary expression of interest, 26 labs signed up to contribute a minimal sample size of 16 toddlers and/or adults. Based on this information, we expected to recruit a total sample of 520 toddlers (ages 18-27 months) and 408 adults (ages 18-55 years). To avoid an unbalanced age distribution in the toddlers' sample, labs signed up for testing at least one of two age bins (bin 1: 18-22 months, bin 2: 23-27 months), and were asked to ensure approximately equal distribution of participants' age in their collected sample if possible. They were asked to try to ensure that the mean age of their sample lies in the middle of the range of the chosen bin and that participant ages were distributed across their whole bin. Both for adults and toddlers, basic demographic data was collected on a voluntary basis with a brief questionnaire (see Supplemental Material for details). The requested demographic information that was not used in the registered confirmatory and/or exploratory analyses of this study has been collected for further potential follow-up analyses in spin-off projects within the MB framework.

After completing the task, adult participants were asked to fill a funneled debriefing questionnaire. This questionnaire asked what the participant thought the purpose of the experiment was, whether the participant had any particular goal or strategy while watching the videos, and whether the participant consciously tracked the chaser's epistemic state. Additionally, we collected details regarding each testing session (see Supplemental Material S3).

Our final dataset consisted of 1224 participants, with an overall exclusion rate of 24.16% (toddlers: 35.60%, adults: 12.67%). Table 1 and Table 2 show the distribution of included participants across labs, eye-tracking methods, and ages. A final sample of 521 toddlers (49.14% female) that were tested in 37 labs (mean lab sample size = 14.08, SD = 5.56, range: 2 - 32) was analyzed. The average age of toddlers in the final sample was 22.49 months (SD = 2.53, range: 18 - 27.01). The final sample size of included adults was 703 (68.85% female), tested in 34 labs (mean lab sample size = 20.68, SD = 12.14, range: 8 - 65). Their mean age was 24.61 years (SD = 7.36, range: 18 - 55).

Table 1

Lab and Participant information for the adult age cohort

Lab	*n* collected	*n* included	Sex (*n* Female)	Mean Age (years)	Method
CogConcordia	21	16	11	22.12	In-lab
CorbitLab	16	15	14	19.87	In-lab
DevlabAU	20	20	15	25.15	In-lab
MEyeLab	53	53	39	24.47	In-lab
MiniDundee	15	13	10	30.23	In-lab
PKUSu	39	32	19	22.66	In-lab
SkidLSDLab	11	8	3	21.62	In-lab
ToMcdlSalzburg	33	31	22	27.23	In-lab
UIUCinfantlab	36	32	25	19.06	In-lab
WSUMARCS	18	13	8	29.85	In-lab

Lab	*n* collected	*n* included	Sex (*n* Female)	Mean Age (years)	Method
affcogUTSC	23	8	5	20.88	Web-based
babyLeidenEdu	20	16	12	23.31	In-lab
babylabAmsterdam	17	16	13	24.00	In-lab
babylabBrookes	67	65	49	21.78	In-lab
babylabINCC	18	18	12	31.00	In-lab
babylabMPIB	16	16	11	27.44	In-lab
babylabNijmegen	19	15	13	22.13	In-lab
babylabTrento	16	16	9	21.69	In-lab
babylabUmassb	33	11	10	19.00	In-lab
babyuniHeidelberg	16	16	14	22.06	In-lab
beinghumanWroclaw	19	16	9	32.75	Web-based
careylabHarvard	18	15	12	19.80	In-lab
cclUNIRI	32	32	17	30.53	In-lab
childdevlabAshoka	16	16	8	30.88	In-lab
collabUIOWA	16	16	10	19.19	In-lab
gaugGöttingen	30	28	18	31.71	In-lab
jmuCDL	32	32	22	18.81	In-lab
kidsdevUniofNewcastle	15	14	7	33.57	In-lab
labUNAM	20	11	8	22.45	In-lab
lmuMunich	31	30	23	22.53	In-lab
mecdmpihcbs	19	19	10	27.79	In-lab
socialcogUmiami	16	15	9	19.27	In-lab
sociocognitivelab	17	17	11	32.12	In-lab
tauccd	15	12	6	24.50	In-lab
Total	803	703	484	24.75	

Table 2

Lab and Participant information for the toddler age cohort

Lab	*n* collected	*n* included	Sex (*n* Female)	Mean Age (months)	Method
CogConcordia	21	8	4	22.92	Web-based
CorbitLab	11	10	5	22.77	In-lab
DevlabAU	18	17	8	19.00	In-lab
PKUSu	50	32	13	20.84	In-lab
SkidLSDLab	8	2	0	20.11	In-lab
ToMcdlSalzburg	17	12	6	22.20	In-lab
UIUCinfantlab	18	15	9	21.96	In-lab
babyLeidenEdu	18	12	8	22.59	In-lab
babylabAmsterdam	28	12	6	23.19	In-lab
babylabBrookes	17	12	7	22.15	In-lab
babylabChicago	17	13	4	20.10	In-lab
babylabINCC	16	9	6	23.40	In-lab

Lab	*n* collected	*n* included	Sex (*n* Female)	Mean Age (months)	Method
babylabNijmegen	19	10	3	23.52	In-lab
babylabOxford	25	19	8	23.42	In-lab
babylabPrinceton	17	11	7	22.15	In-lab
babylabTrento	18	17	10	22.72	In-lab
babylabUmassb	7	6	2	20.35	In-lab
babylingOslo	17	14	7	21.99	In-lab
babyuniHeidelberg	16	12	4	22.69	In-lab
beinghumanWroclaw	24	14	7	23.77	Web-based
careylabHarvard	17	12	5	21.99	In-lab
cecBYU	16	14	4	22.39	In-lab
childdevlabAshoka	16	10	6	22.44	In-lab
gaugGöttingen	28	15	9	23.06	In-lab
gertlabLancaster	21	17	8	23.03	In-lab
infantcogUBC	26	19	8	24.39	In-lab
irlConcordia	19	12	5	22.47	In-lab
kidsdevUniofNewcastle	16	14	9	22.36	In-lab
kokuHamburg	19	14	7	25.99	In-lab
labUNAM	18	12	7	22.68	In-lab
lmuMunich	48	24	16	22.68	In-lab
mecdmpihcbs	25	12	8	23.58	In-lab
mpievaCCP	22	18	10	23.33	In-lab
saxelab	31	15	2	23.13	Web-based
socallabUCSD	47	15	4	22.09	Web-based
tauccd	15	12	8	22.99	In-lab
unicph	43	29	16	21.50	In-lab
Total	809	521	256	22.48	

Apparatus and Procedure

Eye-tracking Methods

We expected that participating labs would use one of three types of eye-tracker brands to track the participant's gaze patterns: Tobii, EyeLink, or SMI. Thus, apparatus setup varied slightly across individual labs (e.g., different sampling rates and distances at which the participants were seated in front of the monitor). Participating labs reported their eye-tracker specifications and study procedure alongside the collected data. To minimize variation between labs, all labs using the same type of eye-tracker used the same presentation study file specific to that eye-tracker type. The Supplemental Material provides an overview of employed eye-trackers, stimulus presentation softwares, sampling rates, and screen dimensions.

Online Gaze Recording

To allow for the participation of labs that did not have access to an eye-tracker, or were not able to invite participants to their facilities due to current restrictions regarding the COVID-19 pandemic, labs could choose to collect data via online testing. We initially anticipated for labs to flexibly choose between different methods for coding gaze direction, including manual frame-by-frame coding from video recordings or using various online platforms for virtual data collection (e.g., LookIt, YouTube, Zoom, Labvanced). Additionally, we considered the possibility of webcam eye-tracking using tools such as WebGazer.js (Papoutsaki et al., 2016). However, we later decided that all labs would conduct only supervised testing, leading to the exclusive use of the MB-ManyWebcams WebGazer setup for data collection (Steffan & Zimmer et al., 2023; see also Supplemental Material S2). This ensured a standardized approach across all participating labs.

Testing Procedure

Toddlers were seated either on their caregiver's lap or in a highchair. The distance from the monitor depended on the data collection method. Caregivers were asked to refrain from interacting with their child and closed their eyes during stimulus presentation or wore a set of opaque sunglasses. Adult participants were seated on a chair at the appropriate distance from the monitor. Once the participant was seated, the experimenter initiated the eye-tracker-specific calibration procedure. Additionally, we presented another calibration stimulus before and after the presentation of the task. This allowed for evaluating the accuracy of the calibration procedure across labs (cf., Frank et al., 2012).

General Lab Practices

To ensure standardization of our experimental procedure, materials for testing practices and instructions were prepared and distributed to the participating labs. Each lab was responsible for maintaining these practices and reported all relevant details on testing sessions (for details see the Supplemental Material).

Videos of Participants

As with all MB projects, we strongly encouraged labs to record video data of their own lab procedures and each testing session, provided that this was in line with regulations of the respective institutional ethics review board and the given informed consent. Participating labs that could not contribute participant videos were asked to provide a video walk-through of their experimental set-up and procedure instead. If no institutional ethics review board restrictions occurred, labs were encouraged to share video recordings of the test sessions via Databrary.

Design Analysis

Here we provide a simulation of the predicted findings because a traditional frequentist power analysis was not applicable for our project for two reasons. First, we used Bayesian methods to quantify the strength of our evidence for or against our hypotheses, rather than assessing the probability of rejecting the null hypothesis. In particular, we computed a Bayes factor (BF; a likelihood ratio comparing two competing hypotheses), which allowed us to compare models. Second, because of the many-labs nature of the study, the sample size was not determined by power analysis, but by the amount of data that participating labs were able to contribute within the pre-established timeframe. Even if the effect size was much smaller than we anticipated (e.g., less than Cohen's $d = 0.20$), the results would have been informative as our study was expected to be dramatically larger than any previous study in this area. If, due to unforeseen reasons, the participating labs were not able to collect a minimum number of 300 participants per age group within the proposed time period, we planned to extend the time for data collection until this minimum number was reached. Conversely, if the effect size was large (e.g., more than Cohen's $d = 0.80$), the resulting increased precision of our model would allow us to test a number of other theoretically and methodologically important hypotheses (see Results section).

Although we did not determine our sample size based on a power analysis, here we provide a simulation-based design analysis to demonstrate the range of BFs we might have expected to see, given a plausible range of effect sizes and parameters. We focus on our key analysis of the test trials (as specified below), namely the difference in AL on the first test trial that participants saw. We describe below the simulation for the child sample, but based on our specifications, we expected that a design analysis for adult data would produce similar results.

We first ran a simulation for the first look analysis. In each iteration of our simulation, we used a set of parameters to simulate an experiment, using a first look (described below) as the key measure. For the key effect size parameter for condition (knowledge vs. ignorance), we sampled a range of effect sizes in logit space spanning from small to large effects (Cohen's *d* = 0.20 - 0.80; log odds from 0.36 - 1.45). For each experiment, the betas for age and the age x condition interaction were sampled uniformly between -0.20 and 0.20. The age of each participant was sampled uniformly between 18 and 27 months and then centered. The intercept was sampled from a normal distribution (1, 0.25), corresponding to an average looking proportion of 0.73. Lab intercepts and the lab slope by condition were set to 0.1, and other lab random effects were set to 0 as we do not expect them to be meaningfully non-zero. These values were chosen based on pilot data (average looking proportion), but also to have a large range of possible outcomes (lab intercept, age and age x condition interaction). We are confident that the results would be robust to different choices. We then used these simulated data to simulate an experiment with 22 labs and 440 toddlers and computed the resulting BFs, as specified in the analysis plan below. We adopted all of the priors specified in the results section below[7]. We ran 349 simulations and, in 72% of them, the BF showed strong evidence in favor of the full model (BF > 10); in 6% the BF showed substantial evidence (10 > BF > 3); it was inconclusive 14% of the time (1/10 < BF < 3), and in 8% of cases the null model was substantially favored (see Supplemental Material Figure

[7] After the design analysis, additional labs expressed their interest in contributing data, which is why the anticipated sample sizes and the numbers this design analysis is based on differ. Given the uncertainty in determining the final sample size in this project, we kept the design analysis as is to have a more conservative estimate of the study's power.

S5). The BF was not < 1/10 in any of the simulations. Thus, under the parameters chosen here for our simulations, it was likely that the planned experiment was of sufficient size to detect the expected effect.

We also ran a design analysis for the proportional looking analysis. We used the same experimental parameters (number of labs, participants, ages, etc.). For generating simulated data, we drew the condition effect from a uniform distribution between .05 and .20 (in proportion space). The age and age x condition interaction effects were drawn from uniform distributions between -.05 and .05. Sigma, the overall noise in the experiment, was drawn from a uniform distribution between .05 and .1. The intercept was drawn from a normal distribution with mean .65 and a standard deviation of .05. The by-lab standard deviation for the intercept and condition slope was set to .01. Priors were as described in the main text. We ran 119 simulations, and in all 119 we obtained a BF greater than 10, suggesting that, under our assumptions, the study was well-powered.

Data Preprocessing

Eye-tracking

Raw gaze position data (x- and y-coordinates) was extracted in the time window starting from the first frame at which the chaser entered the tunnel until the last frame before it exited the tunnel in the familiarization and test trials. For data collected from labs using a binocular eye-tracker, gaze positions of the left and the right eye were averaged.

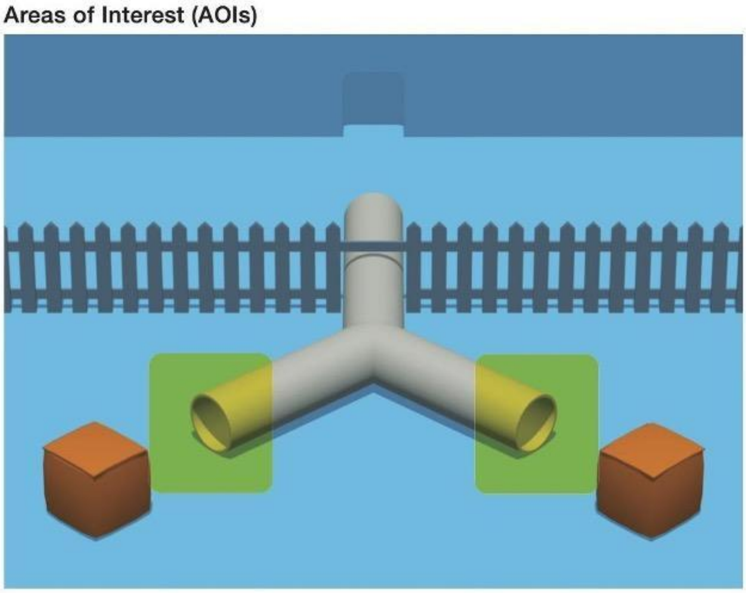
We used the peekds R package (https://github.com/peekbank/peekbankr) to convert eye-tracking data from disparate trackers into a common format. Because not all eye-trackers recorded data with the same frequency or regularity, we resampled all data to be at a common rate of 40 Hz (samples per second).

We excluded individual trials if more than 50% of the gaze data was missing (defined as off-screen or unavailable point of gaze during the whole trial, not just the anticipatory period). Applying this criterion would have caused us to exclude 4% of the trials in our pilot data, which inspection of our pilot data suggested was an appropriate trade-off between not excluding too much usable data and not analyzing trials which were uninformative.

For each monitor size, we determined the specific AOIs and computed whether the specific x- and y-position for each participant, trial, and time point fell within their screen resolution-specific AOIs. Our goal was to determine whether participants were anticipating the emergence of the chaser from one of the two tunnel exits. Thus, we defined AOIs on the stimulus by creating a rectangular region around the tunnel exit that was D units from the top, bottom, left, and right of the boundary of the tunnel exit, where D was the diameter of the tunnel exits. We then expanded the sides of the AOI rectangles by 25% in all directions to account for tracker calibration error. Our rationale was that, if we made the AOI too small, we might fail to capture anticipations by participants with poor calibrations. In contrast, if we made the regions too large, we might capture some fixations by participants looking at the box where the chasee actually was. On the other hand, these chasee looks would not be expected to vary between conditions and so would only affect our baseline level of looking. Thus, the chosen AOIs aimed at maximizing our ability to capture between-condition differences. For an illustration of the tunnel exit AOIs see Figure 3. We were not analyzing looks to the boxes, since they can less unambiguously be interpreted as epistemic state-based action predictions and because we observed few anticipatory looks to the boxes in the pilot studies. For more detailed information about the AOI definition process see the description of the pilot study results in the Supplemental Material.

Figure 3

Illustration of Areas of Interest (AOIs) for gaze data analysis during the anticipatory period



Note. The light green rectangles show the dimensions of the AOIs used for the analysis of AL during the test period.

Manual Coding

For data gathered without an eye-tracker (e.g., videos of participants gathered from online administration), precise estimation of looks to specific AOIs were not possible. Instead, videos were coded for whether participants were looking to the left or the right side of the screen (or "other/off screen"). In our main analysis, during the critical anticipatory window, we treated these looks identically to looks to the corresponding AOI. See exploratory analyses for analysis of data collected online.

Temporal Region of Interest

For familiarization trials, we defined the start of the anticipatory period (total length = 4000 ms) as starting 120 ms after the first frame after which the chaser had completely entered the tunnel and lasting until 120 ms after the first frame at which the chaser was visible again (we chose 120 ms as a conservative value for cutting off reactive saccades; cf., Yang et al., 2002). For test trials, we defined the start of the anticipatory period in the same way, with a total duration of 4000 ms.

Dependent Variables

We define two primary dependent variables:

1. First look. First saccades were determined as the first change in gaze occurring within the anticipatory time window that was directed towards one of the AOIs. The first look was then the binary variable denoting the target of this first saccade (i.e., either the correct or incorrect AOI) and was defined as the first AOI where participants fixated at for at least 150 ms, as in Rayner et al. (2009). The rationale for this definition was that, if participants were looking at a location within the tunnel exit AOIs before the anticipation period, they might have been looking there for other reasons than action prediction. We therefore counted only looks that started within the anticipation period because they more unambiguously reflected action predictions. This further prevented us from running into a situation where we would have included a lot of fixations on regions other than the tunnel exit AOIs because participants were looking somewhere else before the anticipation period began.
2. Proportion DLS (also referred to as total relative looking time; Senju et al., 2009). We computed the proportion looking (p) to the correct AOI during the full 4000 ms anticipatory window (correct looking time / (correct looking time + incorrect looking time)), excluding looks outside of either AOI.

Results

Confirmatory Analyses

Approach

We adopted a Bayesian analysis strategy to maximize our ability to make inferences about the presence or absence of a condition effect (i.e., our key effect of interest). In particular, we fitted Bayesian mixed effects regressions using the package brms in R (Bürkner, 2017). This framework allowed us to estimate key effects of interest while controlling for variability across grouping units (in our case, labs).

To facilitate interpretation of individual coefficients, we report means and credible intervals. For key inferences in our confirmatory analysis, we used the bridge sampling approach (Gronau et al., 2017) to compute BFs comparing different models. As the ratio of the likelihood of the observed data under two different models, BFs allowed us to quantify the evidence that our data provide with respect to key comparisons. For example, by comparing models with and without condition effects, we can quantify the strength of the evidence for or against such effects.

Bayesian model comparisons require the specification of proper priors on the coefficients of individual models. Here, for our first look analysis, we used a set of weakly informative priors that capture the expectation that the effects that we observe (of condition and, in some cases, trial order) are modest. For coefficients, we chose a normal distribution with mean of 0 and *SD* of 2. Based on our pilot testing and the results of MB1, we assumed that lab and participant-level variation would be relatively small, and so for the standard deviation of random effects (i.e., variation in effects across labs and, in the case of the familiarization trials, participants) we set a Normal prior with mean of 0 and *SD* of 0.1. We set an LKJ(2) prior on the correlation matrix in the random effect structure, a prior that is commonly used in Bayesian analyses of this type (Bürkner, 2017). Because the BF is sensitive to the choice of prior, we also ran a secondary analysis with a less informative prior: fixed effect coefficients chosen from a normal distribution with mean 0 and *SD* of 3, and random effect standard deviations drawn from a normal prior with a mean of 0 and *SD* of 0.5 (see Supplemental Material S4). With respect to the specification of random effects, we followed the approach advocated by Barr (2013), that is, specifying the maximal random effect structure justified by our design. Since we were interested in lab-level variation, we fitted random effect coefficients for fixed effects of interest within labs (e.g., condition within lab). Further, where there were participant-level repeated measure data (e.g., familiarization trials), we fitted random effects of participants.

For the proportion DLS analysis, we used a uniform prior on the intercept between -0.5 and 0.5 (corresponding to proportional looking scores between 0 and 1: the full possible range). For the priors on the fixed effect coefficients, we used a normal prior with a mean of 0 and an *SD* of 0.1. Because these regressions are in proportion space, 0.10 corresponds to a change in proportion of 10%. For the random effect priors, we used a normal distribution with mean 0 and standard deviation .05. The LKJ prior was specified as above.

All preregistered confirmatory models converged, with no divergent transitions, all Rhat values < 1.1, and an effective sample size greater than 20% of the total sample size.

Familiarization Trials

Figure 4 shows first look and proportion DLS (non-logit transformed) for toddlers and adults plotted across familiarization trials and test trials. Our first set of analyses examined data from the four familiarization trials and asked whether participants anticipated the chaser's reappearance at one of the tunnel exits. In our first analysis, we were interested in whether participants engaged in AL during the familiarization trials. To quantify the level of familiarization, we fitted Bayesian mixed effect models predicting target looks based on trial number (1-4) with random effects for lab and participants and random slopes for trial number for each.

In R formula notation (which we adopt here because of its relative concision compared with standard mathematical notation), our base model was as follows:

$$measure \sim 1 + trial_{number} + (trial_{number}|lab) + (trial_{number}|participant).$$

We fitted a total of four instances of this model, one for each age group (toddlers vs. adults) and dependent measure (first look vs. proportion DLS). First look models were fitted using Bernoulli family with a logit link function. The proportion DLS models were Gaussian and the dependent variable was centered by subtracting 0.5, such that 0 corresponded to chance-level performance.

Our key question of interest was whether overall anticipation was higher than chance levels on the familiarization trial immediately before the test trials, in service of evaluating the evidence that participants were attentive and making predictive looks immediately prior to test. To evaluate this question across the four models, we coded trial number so that the last trial before the test trials (trial 4) was set to the intercept, allowing the model intercept to encode an estimate of the proportion of correct anticipation immediately before test. We then fitted a simpler model for comparison

$$measure \sim 0 + trial\_number + (trial\_number \mid lab) + (trial\_number \mid participant),$$

which included no intercept term. We then computed the BF comparing this model to the full model. This BF quantified the evidence for an anticipation effect for each group and measure.

First Look.

Toddlers. Investigating first looks to the target location for toddlers, we used a Bayesian mixed effects model to predict whether toddlers' first look was to the target exit based on trial number (1-4), with random effects for lab and participants and random slopes for trial number for each. The Bayes factor comparing the full model to the simpler model was estimated to be BF > 1000, favoring the full model over the null model. This confirmed that toddlers showed above-chance looking to the target location during the anticipatory window of the last familiarization trial. The model also provided support for an effect of trial number on proportion of first looks, with the negative coefficient indicating a decrease in first looks to the target location across the familiarization trials.

Adults. Comparing the Bayesian mixed effects model of adults predicting proportion of first looks based on trial number (1-4), with random effects for lab and participants and random slopes for trial number for each with the simpler model without an intercept, we computed a Bayes factor of BF > 1000, strongly favoring the full model over the null model. Again, this suggested that also adults' anticipation was higher than chance levels on the last familiarization trial. In addition, there was support for an effect of trial number on proportion of first looks, with the positive coefficient indicating an increase in proportion of first target looks across the familiarization trials.

Proportion DLS.

Toddlers. We used a Bayesian mixed effects model to predict proportion DLS based on trial number (1-4) for toddlers, with random effects for lab and participants and random slopes for trial number for each. The Bayes factor comparing this model to the simpler null model without the intercept was estimated to be BF > 1000, strongly favoring the full model over the null model. See also Table 3 for regression coefficients for the full model. These results suggest a significant effect of trial number on proportion DLS, with the negative coefficient indicating a decrease in proportion DLS across the familiarization trials.

Adults. Next, we used a Bayesian mixed effects model to predict proportion DLS based on trial number (1-4) for adults, again with random effects for lab and participants and random slopes for trial number for each. The Bayes factor for the full model against the null model was BF > 1000, suggesting strong evidence for the full model. These results suggest a significant effect of trial number on proportion DLS, with the positive coefficient indicating an increase in target looks across the familiarization trials.

Test Trials

We focused our confirmatory analysis on the first test trial (see Exploratory Analysis section for an analysis of both trials). Our primary question of interest was whether AL differs between conditions (knowledge vs. ignorance, coded as -.5/.5) and by age (in months, centered). For child participants, we fitted models with the specification:

$$measure \sim 1 + condition + age + condition{:}age \\ + (1 + condition + age + condition{:}age|lab).$$

For adult participants, we fitted models with the specification:

$$measure \sim 1 + condition + (1 + condition|lab).$$

Again, we fitted models with a logistic link for first look analyses and with a standard linear link for proportion DLS.

In each case, our key BF was a comparison of this model with a simpler "null" model that did not include the fixed effect of condition but still included other terms. We took a BF > 3 in favor of a particular model as substantial evidence and a BF > 10 in favor of strong evidence. A BF < 1/3 was taken as substantial evidence in favor of the simpler model, and a BF < 1/10 as strong evidence in favor of the simpler model.

For the model of data from toddlers, we additionally were interested in whether the model showed changes in AL with age. We assessed evidence for this by computing BFs related to the comparison with a model that did not include an interaction between age and condition as fixed effects

$$measure \sim 1 + condition + age + (1 + condition + age + condition{:}\,age|lab).$$

These BFs captured the evidence for age-related changes in the difference in action anticipation between the two conditions.

It is important to note that in the case of a null effect, there are two main explanations: (1) toddlers and adults in our study do not distinguish between knowledgeable and ignorant agents when predicting their actions. (2) The method used is not appropriate to reveal knowledge/ignorance understanding. By using Bayesian analyses, we are able to better evaluate the first of these two possibilities: The BF provides a measure of our statistical confidence in the null hypothesis, i.e., no difference between experimental conditions, given the data in ways that standard null hypothesis significance testing does not. In other words, instead of merely concluding that we did not find a difference between conditions, we would be able to find no/anecdotal/moderate/strong/very strong/extreme evidence for the null hypothesis that our participants did not distinguish between knowledgeable and ignorant agents when predicting their actions (Schönbrodt & Wagenmakers, 2018). We therefore consider this analysis an important addition to our overall analysis strategy. Yet, even our Bayesian analyses are not able to rule out the second possibility that participants may well show such knowledge/ignorance differentiation with different methods, or that this ability may not be measurable with any methods available at the current time. Addressing this alternative explanation warrants follow up experiments.

First Look.

Toddlers. Investigating first looks for toddlers, we again used a Bayesian mixed effects model to predict target looks based on condition, with random effects for lab. The Bayes factor comparing the full model to the simpler model was estimated to be BF = 2.4, providing only anecdotal evidence in favor of the full model over the null model.

Again, we examined whether age influenced the difference in action anticipation between knowledge and ignorance trials. To do this, we compared the full model, which included an interaction between age and condition, with a simpler model without this interaction. The computed Bayes factor, BF < 0.01, extremely supports the simpler model, suggesting that the interaction term does not substantially improve the model's fit. This implies that age does not appear to significantly affect the difference in action anticipation between the two trial types.

Adults. We compared a Bayesian mixed-effects model predicting the proportion of first looks based on condition, including random effects for lab to a simpler model without the main effect of condition. The analysis yielded a Bayes factor of BF > 1000, providing strong evidence in favor of the full model over the null model. Results indicated that first looks to the target were significantly more frequent in the knowledge condition compared to the ignorance condition.

Proportion DLS. Figure 5 depicts mean proportion of looking to target and distractor during the anticipation window. The timecourse plot displays mean proportion of looking to target and distractor for both conditions, knowledge and ignorance, both age cohorts, toddlers and adults, as well as both test trials, first and second.

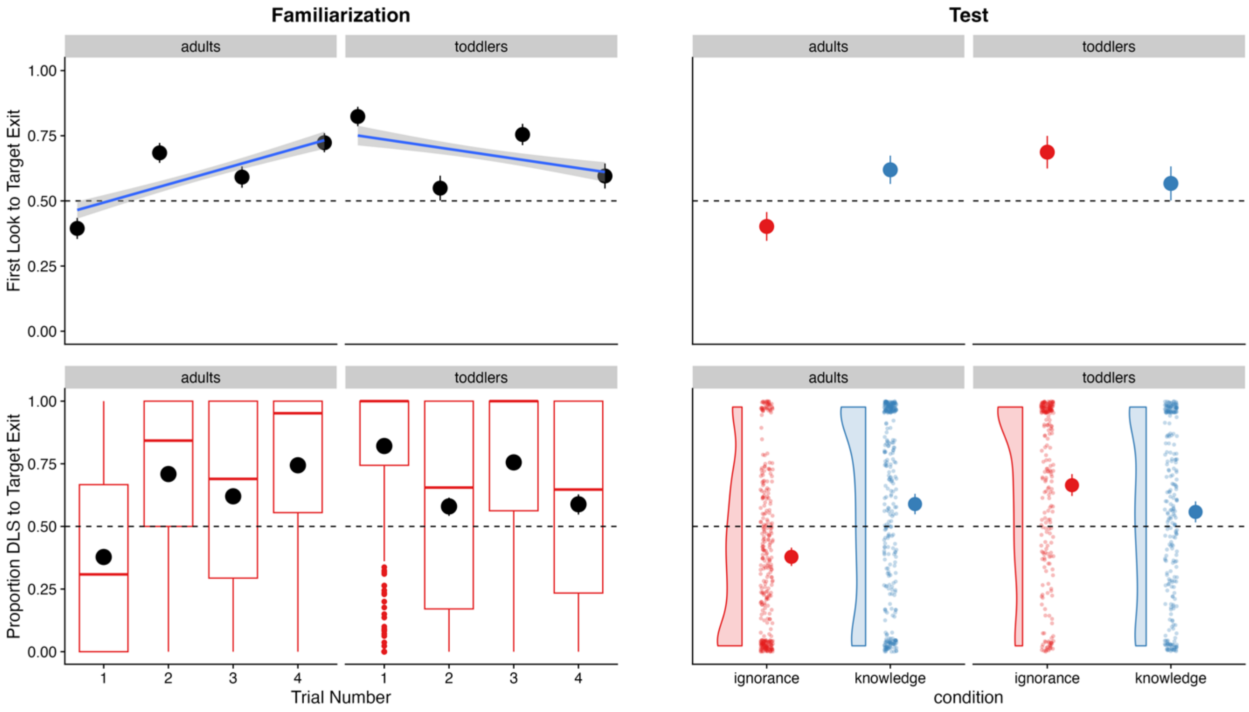
Toddlers. As a first model, we used a Bayesian mixed effects models to predict toddlers' proportion DLS based on condition, age, and the interaction of condition and age, while accounting for variability across labs. The Bayes factor comparing this model to the simpler null model without the interaction of condition was estimated to be BF = 21.2, favoring the full model over the null model. Table 4 shows the statistics for regression coefficients of the full model. These results suggest a significant effect of condition on proportion DLS, with the positive coefficient indicating higher proportion DLS for ignorance trials compared to knowledge trials. An exploratory test of above-chance looking in the knowledge condition, however, revealed that toddlers looked significantly above chance to the target, $t(255) = 2.64$, $p = 0.009$.

Additionally, we assessed whether toddlers' AL changed with age. Comparing our full model, which included an interaction between age and condition, with a simpler model without this interaction, yielded a Bayes factor, BF = 0.4, providing only anecdotal evidence supporting the simpler model. This result suggests that the interaction between age and condition might not be a necessary predictor, as it does not provide substantial additional explanatory power. Hence, our results do not provide sufficient evidence to determine whether age-related changes in AL are consistent across conditions or differ between them.

Adults. Next, we used a Bayesian mixed effects model to predict proportion DLS based on condition for adults, again with random effects for lab. The Bayes factor comparing this model to the simpler null model without the main effect of condition was estimated to be BF > 1000, strongly favoring the full model over the null model. These results suggest a significant main effect of condition on proportion DLS, with the negative coefficient indicating a higher number of target looks for knowledge than for ignorance trials.

Figure 4

First look and proportion DLS for toddlers and adults during familiarization and test phase



Note. For familiarization, first look and proportion DLS are shown for each of the four trials during the anticipatory window. For the test phase, first look and proportion DLS are shown for the anticipatory window of the first trial only. Error bars and error bands represent 95% CIs.

Table 3

Results of the Bayesian mixed effects models for the familiarization phase

model	term	β	*SE*	lower CrI95%	upper CrI95%	$\hat{R}$
first look toddlers	Intercept	0.44	0.09	0.27	0.61	1.00
	Trial Number	-0.22	0.05	-0.32	-0.12	1.00
first look adults	Intercept	1.03	0.09	0.86	1.20	1.00
	Trial Number	0.38	0.04	0.30	0.47	1.00
proportion DLS toddlers	Intercept	0.12	0.02	0.09	0.15	1.00
	Trial Number	-0.05	0.01	-0.06	-0.03	1.00
proportion DLS adults	Intercept	0.26	0.02	0.23	0.29	1.00
	Trial Number	0.10	0.01	0.09	0.11	1.00

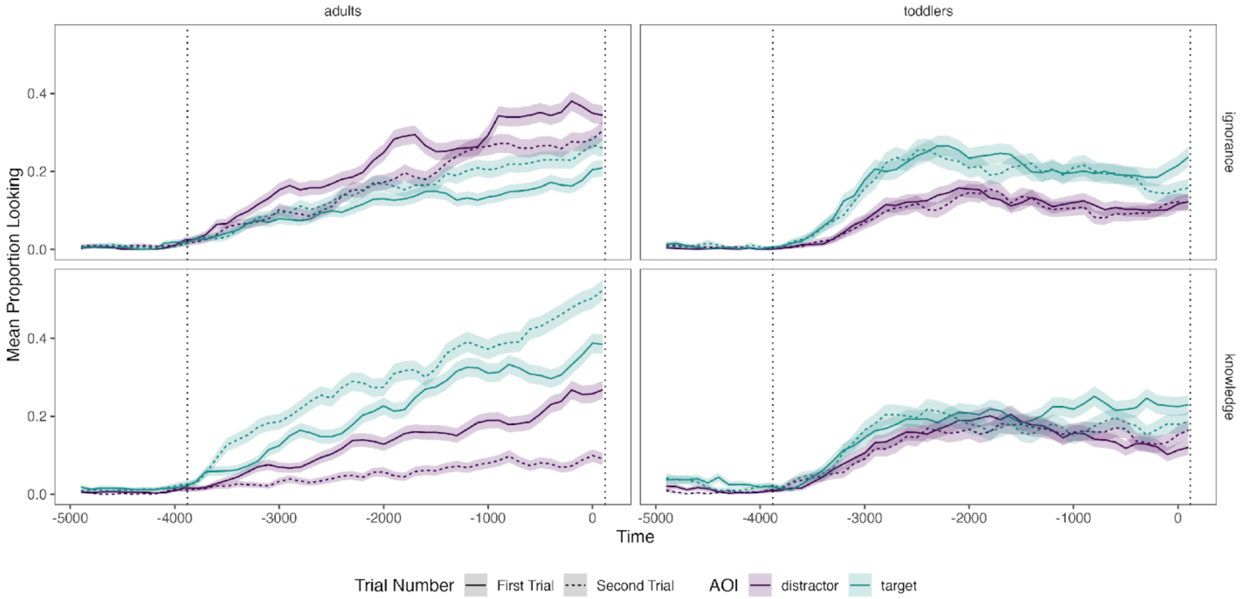
Table 4

Results of the Bayesian mixed effects models for the test phase

model	term	β	*SE*	lower	upper	$\hat{R}$
				CrI95%	CrI95%	
first look toddlers	Intercept	0.53	0.11	0.32	0.74	1.00
	Condition	0.53	0.21	0.13	0.93	1.00
	Age	0.06	0.05	-0.03	0.15	1.00
	Condition:Age	-0.13	0.09	-0.30	0.04	1.00
first look adults	Intercept	0.05	0.09	-0.12	0.22	1.00
	Condition	-0.89	0.17	-1.22	-0.56	1.00
proportion DLS toddlers	Intercept	0.61	0.02	0.58	0.65	1.00
	Condition	0.10	0.03	0.03	0.17	1.00
	Age	0.01	0.01	-0.01	0.02	1.00
	Condition:Age	-0.01	0.02	-0.04	0.02	1.00
proportion DLS adults	Intercept	0.48	0.02	0.45	0.51	1.00
	Condition	-0.20	0.03	-0.26	-0.15	1.00

Figure 5

Timecourse of mean proportion looking to target and distractor for the first and second test trial for toddlers and adults in the ignorance and knowledge condition



Note. The dotted vertical lines represent the onset and offset of the 4s anticipatory period. Time is centered such that 0 represents the moment the bear re-remerges from the tunnel. Error bands represent +1/-1 SEs.

Exploratory Analyses

Spill-over

We analyzed within-participants data from the second test trial that participants saw, using exploratory models to assess whether (1) findings are consistent when both trials are included (overall condition effect), (2) whether effects are magnified or diminished on the second trial (order main effect), and (3) whether there is evidence of "spillover"-dependency in anticipation on the second trial depending on what the first trial is (condition x order interaction effect; see Figure 5 and Figure 6).

For toddlers, we analyzed condition-effects of within-participants data for both test trials by fitting a Bayesian mixed-effects model with the dependent variable of proportion DLS and main effects of condition and age and their interaction. Comparing this full model to a null model that did not include the fixed effect of condition, we obtained a Bayes Factor of BF = 147.1, providing strong evidence in favor of the full model, suggesting no within-participants condition effect.

For adults, we also fitted a Bayesian mixed-effects model to predict their proportion DLS for both test trials with the main effect of condition and random effects for participant and lab. Again, the data provided very strong evidence for the inclusion of the main effect of condition with a Bayes Factor of BF > 1000. The effect of condition was negative and credible, suggesting that proportion DLS was significantly lower in the ignorance condition compared to the knowledge condition.

In order to investigate whether there was an interaction of condition and test trial number, we fitted Bayesian mixed-effects models to predict proportion DLS with fixed effects for condition, test trial number, and their interaction, along with random intercepts and slopes for these variables across labs, for toddlers and adults separately. For toddlers, the Bayes factor, BF = 0.4, provided anecdotal evidence for the simpler null model without the interaction term, indicating that the interaction between condition and test trial number does not add substantial explanatory power to the model. These results suggest that neither condition nor its interaction with test trial number significantly impacted proportion DLS in this sample.

For adults, the Bayes Factor, BF = 19.7, provided strong evidence for including the interaction of condition and test trial number as a fixed effect. These results indicate that while proportion DLS increased over trials, this effect was moderated by condition, with the ignorance condition showing a slower rate of increase compared to the knowledge condition.

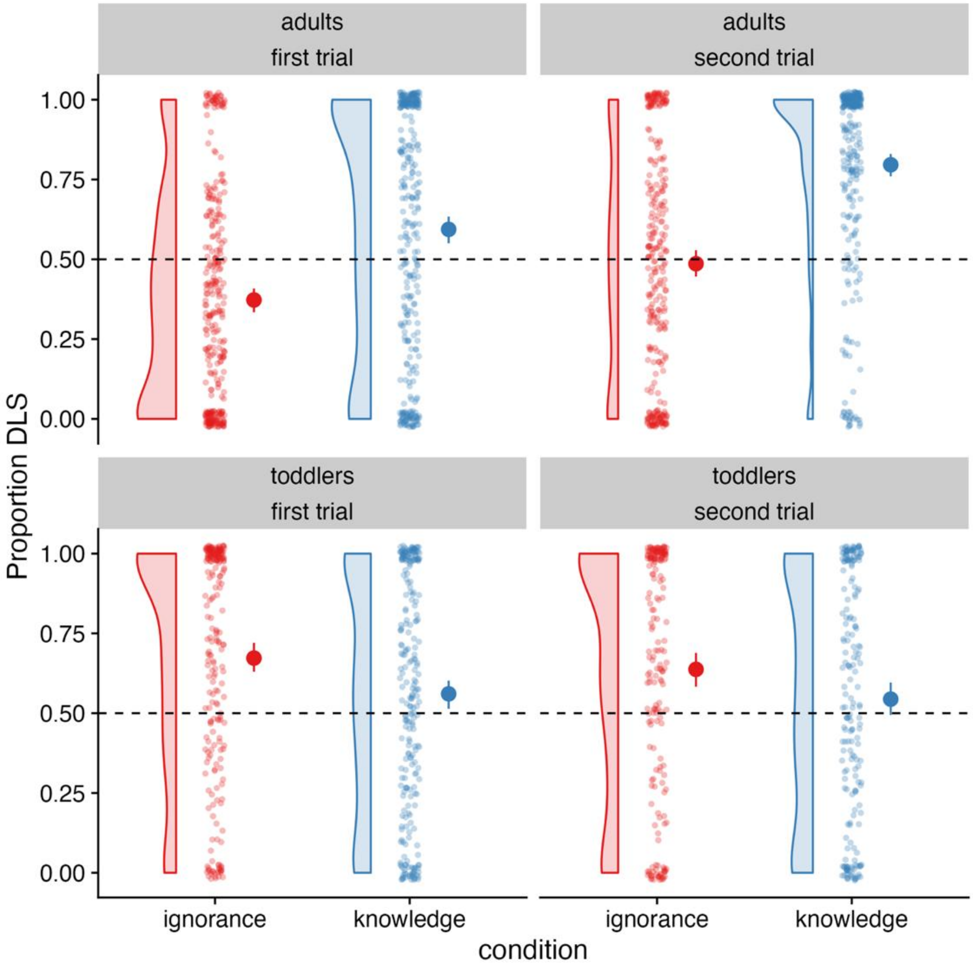
To examine whether anticipatory looking during the second test trial was influenced by condition and anticipatory looking during the first test trial, we fitted a Bayesian mixed-effects model for each age cohort separately. This model included fixed effects for condition, proportion of target looking during the first test trial, and their interaction. Random intercepts and slopes for these predictors were modeled at the lab level.

For toddlers, the Bayes factor, BF = 1.1, suggests anecdotal evidence in favor of including these predictors compared to the null model. In addition, we compared the full model to a model excluding the main effect of condition. The Bayes factor was BF = 2.3, again providing only anecdotal evidence in favor of the full model that included condition. Taken together, these results suggest that condition and its interaction with first trial anticipatory looking were not strong predictors of second trial anticipatory looking behavior in toddlers.

For adults, the Bayes factor comparing the full model including condition, first trial anticipatory looking, and their interaction to the null model with only the intercept was BF > 1000 providing extreme evidence for the full model. Comparing the full model with a model that excluded condition, the Bayes factor, BF > 1000, provided again extreme evidence in favor of the full model. Hence, the extremely large Bayes factors underscore the importance of considering these predictors in explaining second test trial AL behavior.

Figure 6

Proportional DLS for the first and second test trial for toddlers and adults in the ignorance and knowledge condition



Note. Error bars represent 95% CIs.

Relation between familiarization and test

We explored whether condition differences varied for participants who showed higher rates of anticipation during the four familiarization trials. To investigate whether participants who showed anticipatory looking during the familiarization phase also exhibit anticipatory looking during the test phase, we explored three different measures. First, we assessed anticipatory looking in participants who successfully anticipated during the final familiarization trial, defined as those whose first fixation was on the target. Second, we examined anticipatory looking in participants who consistently demonstrated anticipatory behavior across the last two familiarization trials before the test trials, operationalized as having a proportion DLS greater than 0.5 in each trial. Finally, we computed correlations to explore whether performance in the familiarization phase was related to performance in the test trials.

Relation between anticipatory looking during the first test trial and first look during the final familiarization trial. We fitted a main Bayesian hierarchical model testing the fixed effects of condition (ignorance vs. knowledge), first look during the final familiarization trial (target vs. distractor), and their interaction on first-trial proportion DLS during the anticipatory window for toddlers and adults separately. Random intercepts and slopes for all fixed effects and their interaction were included at the lab level, accounting for variability across different experimental settings.

For toddlers, the Bayes factor comparing this model to the simpler null model without the interaction of condition and first look during the final familiarization trial indicated that there was anecdotal evidence in favor of the simpler null model over the full model, BF = 0.7. The effect of condition was positive, but its confidence interval narrowly included zero, suggesting weak evidence for a condition effect (see Table 5). The effect of performance during the final familiarization trial was close to zero, indicating no substantial main effect of prior performance. Similarly, the interaction between condition and performance in the final familiarization trial was small and non-significant. These results suggest that while there was some weak evidence for a main effect of condition on anticipatory looking, neither performance during the final familiarization trial nor its interaction with condition substantially predicted anticipatory looking during the test trial. This result indicates that the relation between anticipatory looking during the first test trial and condition did not depend significantly on prior familiarization performance.

For adults, the Bayes factor comparing this model to the simpler null model without the main effect of condition was estimated to be BF > 1000, strongly favoring the full model over the null model. The regression coefficients showed a significant negative effect of condition, indicating that anticipatory looking was lower in ignorance trials compared to knowledge trials.

The decisive Bayes factor strongly favors the inclusion of condition and familiarization trial performance in the model, suggesting that these predictors are relevant for understanding anticipatory looking in adults. However, the small and non-significant estimates for the effects of familiarization trial performance and its interaction with condition imply that condition is the primary driver of anticipatory looking differences, with performance in familiarization trials contributing minimally.

Only above 50% looking to target during familiarization trials. To examine the effect of condition and successful anticipatory looking during familiarization (above 50% target looking during the last two familiarization trials before test) on anticipatory looking during the first test trial, we fitted Bayesian mixed-effects models for each age group separately. The models included fixed effects for condition, successful anticipatory looking during familiarization trials 3 and 4, and their interaction. Random intercepts and slopes for these predictors were included at the lab level.

Comparing the full model to the null model of toddlers revealed a Bayes Factor of BF = 17.9, providing strong evidence favoring the full model over a null model that excludes these predictors, suggesting that these factors contribute meaningfully to explaining the variance in test trial anticipatory looking. The regression analysis showed a positive main effect of condition, indicating higher anticipatory looking in one condition compared to the other (see Table 5). There was a small positive, but non-significant, effect of successful anticipatory looking during familiarization. The interaction between condition and successful anticipatory looking during familiarization was also small and non-significant. These results indicate that condition is a meaningful predictor of anticipatory looking during test trials in toddlers, with participants showing different levels of anticipatory looking based on condition. However, the successful anticipatory looking during familiarization trials and its interaction with condition appear to have minimal additional impact. The strong Bayes factor further supports the importance of including these predictors in the model but highlights that condition remains the primary driver of test trial differences.

The estimated Bayes factor in favor of the full model of adults over the null model was BF > 1000, indicating that the predictors substantially contribute to explaining test trial anticipatory looking. The regression coefficients revealed a significant main effect of condition, with participants showing lower anticipatory looking in the ignorance condition compared to the knowledge condition. There was a small, positive, and non-significant effect of successful anticipatory looking during familiarization and the interaction between condition and successful anticipatory looking during familiarization was negligible. These results indicate that condition has a substantial and meaningful impact on anticipatory looking during the first test trial in adults, while successful anticipatory looking in familiarization trials and its interaction with condition have limited additional influence. The extremely large Bayes factor highlights the strong explanatory power of including these predictors in the model, although condition remains the primary driver of the observed differences.

Table 5

Results of the Bayesian mixed effects models for the relation between familiarization and test

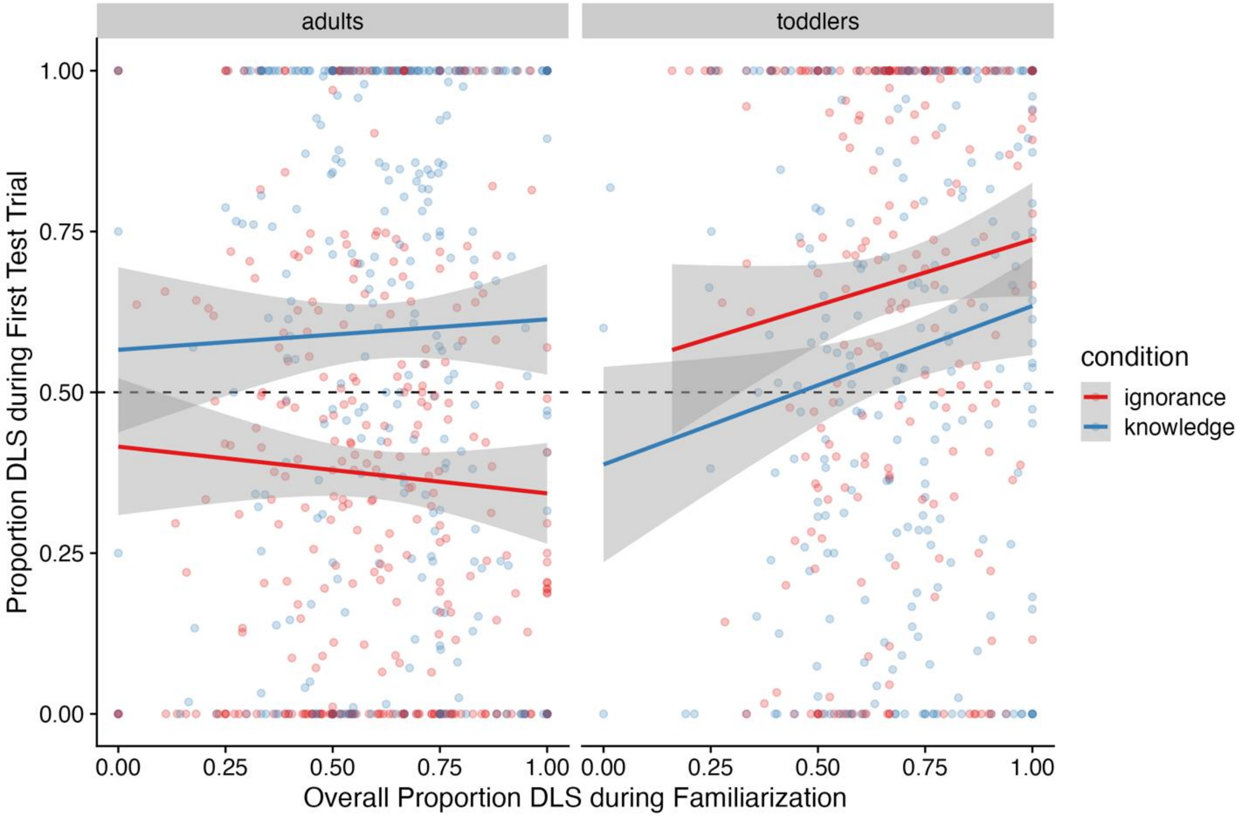
model	term	β	*SE*	lower CrI95 %	upper CrI95 %	$\hat{R}$
sufficient familiarization	Intercept	0.48	0.02	0.44	0.51	1.00
adults	Condition	-0.20	0.03	-0.26	-0.14	1.00
	Successful Fam Anticipation	0.04	0.03	-0.02	0.10	1.00
	Condition:Successful Fam Anticipation	0.00	0.05	-0.11	0.10	1.00
sufficient familiarization	Intercept	0.61	0.02	0.56	0.65	1.00
toddlers	Condition	0.10	0.04	0.02	0.17	1.00
	Successful Fam Anticipation	0.03	0.04	-0.04	0.11	1.00
	Condition:Successful Fam Anticipation	0.02	0.06	-0.11	0.14	1.00
correct first look	Intercept	0.48	0.02	0.45	0.52	1.00
adults	Condition	-0.20	0.03	-0.26	-0.13	1.00

	Correct First Look Final Fam	-0.04	0.03	-0.10	0.03	1.00
	Condition:Correct First Look Final Fam	-0.08	0.06	-0.19	0.04	1.00
correct first look toddlers	Intercept	0.62	0.02	0.58	0.65	1.00
	Condition	0.07	0.04	0.00	0.15	1.00
	Correct First Look Final Fam	0.00	0.04	-0.07	0.07	1.00
	Condition:Correct First Look Final Fam	-0.05	0.06	-0.17	0.07	1.00

Correlation between familiarization and test. We also examined the correlation between familiarization and test performance across the two age cohorts and conditions (see Figure 7). While no significant correlations were found for adults in either condition, toddlers in the knowledge condition exhibited a significant positive correlation of anticipatory looking in familiarization and test, $r = 0.15$, $t(254) = 2.35$, $p = 0.02$, suggesting that toddlers, who showed higher proportion DLS overall during familiarization, also showed higher proportion DLS in the first test trial. No such significant correlation was found for toddlers in the ignorance condition, $r = 0.11$, $t(234) = 1.68$, $p = 0.09$. However, there was also no significant interaction between condition and the correlation in anticipatory looking between familiarization and test for toddlers ($p = 0.79$).

Figure 7

Relation between anticipatory looking during familiarization and test for both age cohorts and conditions



Note. Error bars represent 95% CIs.

Looking patterns during mouse's change of location

To examine whether participants monitored both the bear and the mouse during the mouse's location change, and whether this influenced AL in the test phase, we defined new time windows of interest (TOIs) corresponding to the mouse's location change in each condition and new AOIs for both the mouse and bear. We hypothesized that participants who attended to both AOIs would exhibit greater AL compared to those who predominantly tracked the mouse during its location change. Specifically, we analyzed the frequency of gaze shifts between the mouse and bear mouse's location change. An additional exploratory analysis of differential gaze duration directed toward mouse and bear during the mouse's location change is provided in the Supplemental Material S4. Heatmap videos for knowledge and ignorance trials of both age cohorts can be found at https://drive.google.com/drive/folders/1vL-sobdpgdGAirqWkr_ZB3eEw6gApvL3.

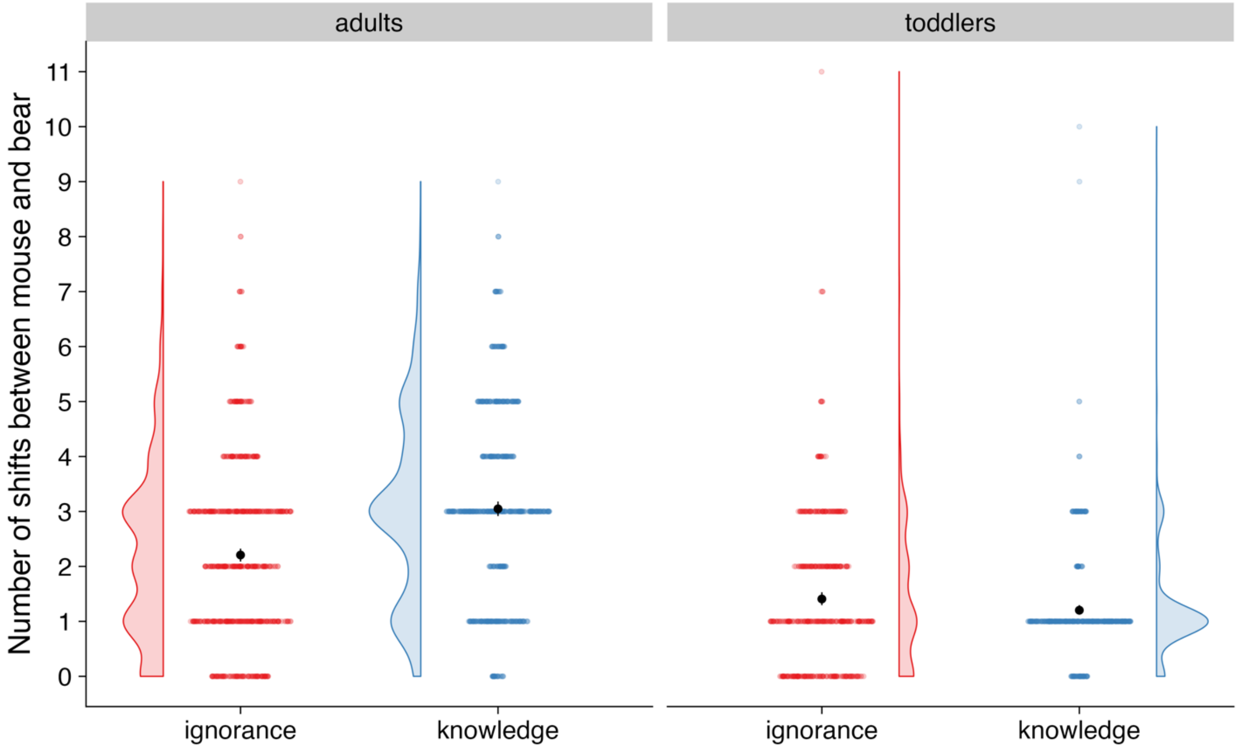
Comparing the number of shifts of toddlers and adults during the location change of the mouse. We fitted a Bayesian mixed-effects model using Poisson family to examine the relation between the number of shifts between mouse and bear and age cohort during location change of the mouse, while accounting for random effects by lab. The effect of condition was negative and approached significance, suggesting a potential reduction in the number of shifts for the ignorance condition compared to the knowledge condition. The main effect of age cohort was positive and credible, Estimate=0.56, indicating that the number of shifts was higher for adults than for toddlers. Importantly, the interaction between condition and age cohort was negative and credible, indicating that the negative effect of condition was more pronounced in the adult cohort (see Figure 8 and Table 6 for the results of Bayesian mixed effects models). Comparing this model to a simpler model without the interaction of condition and age cohort, a Bayes Factor of BF > 1000 was computed. This provides strong evidence in favor of including the interaction of condition and age cohort in the model. In order to interpret this interaction, we conducted follow-up analyses separately for toddlers and adults.

For toddlers, there was extreme evidence in favor of the null model compared to the full model that included condition as fixed effect, BF < 0.01, suggesting little to no reliable difference in the number of gaze shifts between the knowledge and ignorance condition.

In contrast, for adults, the Bayes Factor of BF > 1000 provided extreme evidence in favor of the full model that included condition as fixed effect. In the knowledge condition, adults showed more gaze shifts than in the ignorance condition. These findings demonstrate a robust condition effect in adults but not in toddlers.

Figure 8

Number of shifts between mouse and bear during location change of mouse in the test phase for toddlers and adults in the ignorance and knowledge condition



Note. Error bars represent 95% CIs.

Table 6

Results of the Bayesian mixed effects models for the number of shifts during location change of the mouse

model	term	β	*SE*	lower CrI95%	upper CrI95%	$\hat{R}$
shifts age cohort	Intercept	0.59	0.02	0.55	0.64	1.00
	Condition	-0.10	0.03	-0.16	-0.04	1.00
	Age Cohort	0.57	0.04	0.48	0.65	1.00
	Condition:Age Cohort	-0.33	0.06	-0.44	-0.22	1.00
shifts toddlers	Intercept	0.26	0.03	0.20	0.32	1.00
	Condition	0.12	0.05	0.02	0.22	1.00
shifts adults	Intercept	0.92	0.03	0.86	0.97	1.00
	Condition	-0.28	0.04	-0.35	-0.20	1.00

AL as a function of number of gaze shifts between mouse and bear during location change. In order to examine the effect of condition and the number of shifts between mouse and bear during location change of the mouse on anticipatory looking, we fitted Bayesian mixed-effects models for each age cohort separately (see Figure 9 and Table 7 for the results of Bayesian mixed effects models). The dependent variable was proportion DLS in the anticipation period. The fixed effects included the main effects of condition, the number of shifts, and their interaction. We also included random intercepts and slopes for number of shifts within each participant and within each lab, allowing us to account for the hierarchical structure of the data and potential variability between labs and participants.

For toddlers, comparing this model to a simpler model without the interaction of condition and number of shifts, a Bayes Factor of BF < 0.01 was computed, indicating that the data extremely favored the null model over the full model. A Bayes Factor comparison also provided extreme evidence against including the fixed effect of number of shifts, BF < 0.01. These results suggest that, for toddlers, condition influenced proportional DLS, but the number of location shifts—either alone or in interaction with condition—did not meaningfully contribute to explaining variation in their looking behavior.

For adults, the number of shifts showed a small but credible positive effect, suggesting that more shifts were associated with an increase in proportion DLS. The interaction between condition and the number of shifts was negative and credible, indicating that the effect of number of shifts on proportion looking at test was larger in the knowledge condition than in the ignorance condition. The Bayesian analysis of the interaction effect produced an extremely high Bayes Bactor of BF > 1000, strongly supporting the presence of this interaction. The effect of the number of shifts also received very strong evidence, BF = 35.5, indicating a meaningful contribution of this predictor to the model.

Table 7

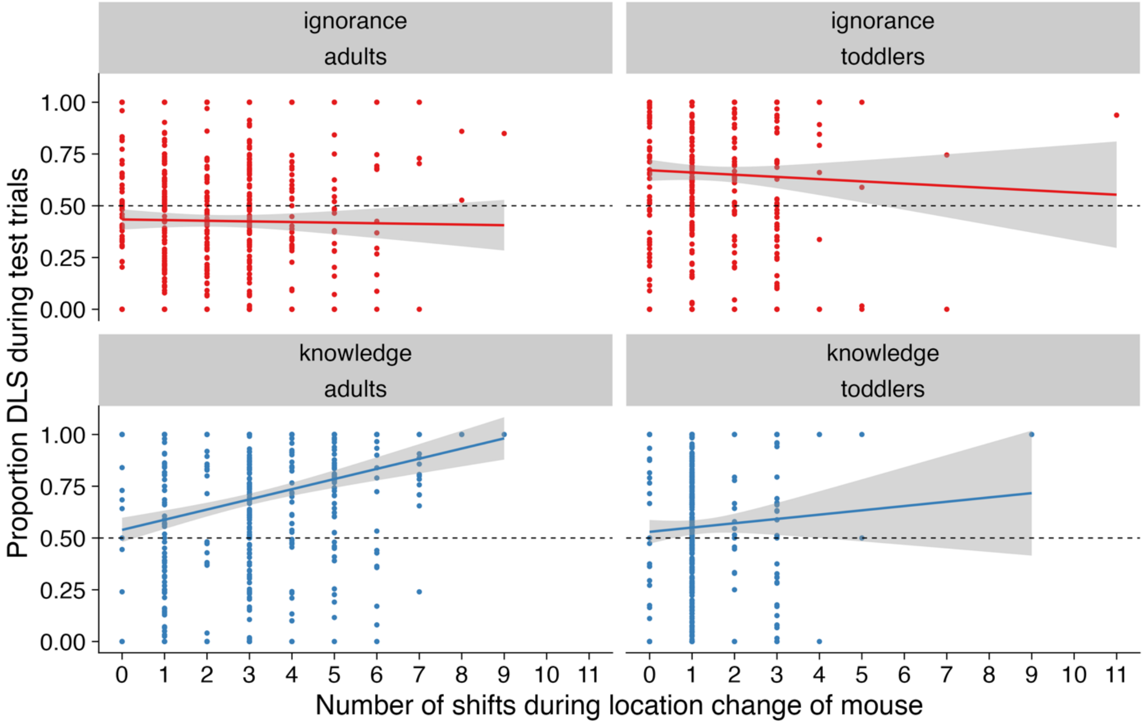
Results of the Bayesian mixed effects models for anticipatory looking as a function of the number of shifts during location change of the mouse

model	term	β	*SE*	lower CrI95%	upper CrI95%	$\hat{R}$
toddlers al shifts	Intercept	0.60	0.02	0.56	0.64	1.00
	Condition	0.12	0.04	0.04	0.19	1.00
	Number Shifts	0.00	0.01	-0.02	0.03	1.00

	Condition:Number Shifts	-0.02	0.02	-0.06	0.03	1.00
adults al shifts	Intercept	0.49	0.02	0.45	0.52	1.00
	Condition	-0.09	0.04	-0.16	-0.02	1.00
	Number Shifts	0.02	0.01	0.01	0.04	1.00
	Condition:Number Shifts	-0.06	0.01	-0.08	-0.03	1.00

Figure 9

AL as a function of number of shifts between mouse and bear during location change of mouse in the test phase for toddlers and adults



Note. Error bars represent 95% CIs.

General Discussion

The overarching aim of the MB2 consortium is to investigate the robustness and replicability of studies showing spontaneous Theory of Mind from infancy across the lifespan. The first project, the initial steps of which are reported here, takes a systematic, sequential bottom-up approach to address anticipatory looking as a measure of spontaneous Theory of Mind and pursues three objectives. First, we aim to develop stimulus material (i.e., videos) that reliably and generally elicits spontaneous, goal-based action anticipation – such that young children and adults look where an agent will go as a function of their goal. In this way, the problem of very high exclusion rates in previous studies (of children who did not spontaneously anticipate) could be overcome (e.g., Kampis et al., 2020; Kulke, Reiß, et al., 2018; Schuwerk et al., 2018; Southgate et al., 2007). If the first goal can be met, the second goal is to examine whether young children and adults demonstrate action anticipation that is sensitive to basic epistemic states of agents: do they anticipate as a function of whether the agent has or has not seen crucial events and is thus knowledgeable or ignorant of them? If the second goal can be fully met, the third goal, to be addressed in future work, is to test the replicability of the original findings of the false belief/true belief contrast. In this paper, we focused on the first two steps and laid the foundation for future work on the third.

Summary and evaluation of main findings

Concerning our first goal, the present work was successful. In two pilot studies (see Supplement S1) and the main study reported here, toddlers and adults reliably engaged in spontaneous goal-based action anticipation: they looked ahead of time towards the location where an agent would go, given their goal. This licensed the conclusion that the present stimulus material is suitable for studying spontaneous action anticipation and laid the foundation for addressing the second goal in the main study: is spontaneous action anticipation sensitive to the agent's epistemic status?

The results of this study were mixed, even with the large sample size of more than 500 toddlers and 700 adults tested in over 30 labs around the world. By accounting for random effects due to lab, our modeling approach allows us to generalize our findings to toddlers and adults across diverse socio-cultural and lab contexts.

Adults showed clear evidence of the anticipatory looking patterns that one would expect if they engaged in action anticipation that is sensitive to the agent's epistemic status. When the agent (chaser) witnessed the crucial events in the knowledge condition and thus knew where the target (chasee) was, adults looked in anticipation towards the corresponding location. This anticipation was indicated by both first looks and proportional looking time. When the chaser did not witness the crucial event in the ignorance condition and thus did not know where the chasee was, they did not show such a pattern of anticipatory looking and rather looked at the tunnel exit opposite of the chasee's actual location. Supporting this interpretation, there were clear condition differences in both the first-looks measure and proportion of anticipatory looking.

For toddlers, the findings were different and puzzling. They tended to show qualitatively the same anticipatory looking (both in first looks and in proportional looking time) towards the target location in both knowledge and ignorance trials. The qualitative result in the knowledge condition is consistent with adult behavior. Yet quantitatively, children anticipated substantially more in the ignorance compared to the knowledge condition in their proportion of anticipatory looking – precisely the reverse effect of what one would expect if children tracked the agent's epistemic status.

The overall picture that the present study presents is thus the following: The findings with adults were straightforward and in line with our predictions. In their anticipatory looking, adults engaged in spontaneous goal-based action anticipation (pilot studies and main study), and in doing so, they took into account the agent's epistemic status (main study). Based on this, the next step would be to pursue the third goal: testing whether adults take into account true/false beliefs of an agent in their spontaneous action anticipation. Specifically, adults are expected to anticipate that the agent will go to the actual location of the target in the true belief condition but to the location where the agent falsely believes the target to be in the false belief condition (e.g., Schneider et al., 2011; Senju et al., 2009).

In contrast, the findings with the toddlers are puzzling and not in line with our predictions. Although toddlers did engage in spontaneous goal-based action anticipation (pilot studies and main study), they did not show clear evidence of taking into account the agent's epistemic status (main study) in the way adults did.

Big open question: How can the puzzling looking patterns in toddlers be explained?

The puzzle is what to make of these findings with children. How did the surprising anticipatory looking pattern of children come about? Why did toddlers not anticipate more clearly the chaser's action in the knowledge condition? And why did they show anticipatory looking to the box with the chasee in the ignorance condition (where they should not do so, or at least to a lesser degree)? A number of initially plausible explanations could be ruled out via exploratory analyses. One such explanation was that behind the grouped data, more nuanced sub-group patterns were hidden. For example, older children could be performing as expected, while younger children were not; or children who anticipated strongly in the familiarization trials could be performing as expected, while children who showed little or no anticipation were not. However, corresponding exploratory analyses along these lines did not find compelling evidence for such sub-group patterns. Neither toddlers' age nor anticipatory looking in the familiarization trials had an effect on the pattern of test trial results. Another explanation was that perseveration from the last familiarization trial to the first test trials (such that children persevere in looking in anticipation to the location they have previously looked to) differentially affected knowledge and ignorance conditions and could thus account for at least parts of the puzzling pattern. But the relevant exploratory control analyses (for details see Supplemental Material) did not find any convincing evidence for such a possibility.

How then can these puzzling anticipatory looking patterns in the knowledge vs. ignorance conditions in toddlers be explained? More specifically, how can we explain why toddlers in the ignorance condition engaged in strong anticipatory looking towards the unpredicted location (where the chasee currently is, unbeknownst to the chaser)? And how can we explain why they showed only very weak correct action predictions in the knowledge condition – weaker than in the ignorance condition, and weaker than in the familiarization trials? We discuss several possibilities below. These are currently all, needless to say, post hoc speculations. But they may lay the foundation for testing them in future studies.

Timing differences between conditions

One possibility regarding the ignorance condition is that slight differences in timing between the conditions may have posed challenges for toddlers. Specifically, in the ignorance condition, the chaser hides at the back and fails to witness the key events where the chasee moves first to one box and then to another. In contrast, in the knowledge condition, the chaser observes the chasee going to one box, then leaves, returns and witnesses how the chasee moves between the boxes (see Figure 2). As a result, the conditions differed subtly in timing. In the ignorance condition, there was a slightly longer interval between the initial hiding event at location 1 and the anticipatory looking phase. This extended interval may have impaired children's memory of the event, making location 1 less salient and leading them to focus more on location 2 during the anticipation phase. Although the lack of an age effect between 18- and 27-month-olds slightly undermines the memory-capacity explanation (because increasing memory capacity in this age interval should have produced an effect of age), future studies could address this issue by equating the temporal structures of both conditions.

Attentional and other processing demands

A second possible explanation for the obtained pattern of results, and in particular, why toddlers did not anticipate more clearly and strongly in the knowledge condition, may relate to attentional and other processing demands of the knowledge and ignorance conditions.

Regarding attention, the knowledge condition raises particular demands of distributing and coordinating attentional focus. In the familiarization trials, children show clear and strong goal-based action anticipation. However, in these trials, the chaser remains present in the scene all the time (it never goes towards the back), and the chasee goes to one box in the chaser's presence but does not change to the other box. In the knowledge test trials, in contrast, there is much more going on: the chaser leaves towards the back and then returns, and the chasee first goes to one box, and then relocates to the other. Perhaps dividing attention between the relevant events (chaser is at the back, chasee at the same time in one box) and coordinating it over time (keeping track of what the chaser has witnessed when) was too demanding for toddlers, and as a result they lost track of the narrative structure of the events. Exploratory analyses of gaze shifts between chaser and chasee in toddlers vs. adults may be seen as an indication that there is something to this explanation: Adults seemed to track the chaser's perceptual access, as indicated by many gaze shifts between chaser and chase during the location change. In contrast, toddlers' attention remained largely on the moving chasee with fewer gaze shifts towards the chaser who was witnessing the relocation in the knowledge condition. Future studies could test more directly whether attentional demands made the present knowledge condition particularly demanding. The chaser could be continuously present all the time, for example, and never leave towards the back (which was introduced to keep the knowledge condition as similar as possible to the ignorance condition) – thus reducing the need to divide and coordinate attention between chaser and chasee. The corresponding ignorance condition could then be realized differently, not such that the chaser leaves, but, for example, such that their view becomes blocked by an occluder, or they fall asleep or are otherwise blindfolded.

Relatedly and more generally, the complexity of the event sequences to be followed and tracked in both conditions may pose excessive performance demands that mask children's competence to understand the agent's epistemic status. Overburdened by such processing demands, toddlers may revert to simpler cognitive strategies influenced by dynamic visual salience, for example. Future research should aim to address these cognitive constraints by further simplifying task demands and optimizing event timing.

Challenges of understanding the implementation of the ignorance condition

A third possibility is that toddlers' (and adults') anticipation is related to their differing conceptual understanding of the scenes in the ignorance trials. In the ignorance condition, the chaser leaves but its back is visible. This requires monitoring and understanding that this does *not* give the chaser epistemic access to the events, which may overburden toddlers. Additionally, the ignorance condition presents a challenge regarding what to anticipate: should participants expect that the chaser will come out at one of the two exits at random? Or that the chaser will go to the location where the chasee is not? Interestingly, adults' looking behavior in the ignorance condition suggests that they expected the chaser to go to the incorrect location, similar to what might be predicted in a False-Belief scenario or an ignorance-leads-to-error heuristic (e.g., Ruffman, 1996; but see Friedman & Petrashek, 2009). Adults and toddlers looked comparably in knowledge trials with slightly above chance anticipation (albeit stronger in adults), whereas the two groups showed entirely opposite patterns in the ignorance condition. This raises important questions about how toddlers interpreted the ignorance condition. Did they entirely lose track of the chaser's epistemic state? Did they show a 'pull of the real', focusing on the actual location of the chasee? If so, why did this not occur in the knowledge condition? Alternatively, they may have been governed by altogether different assumptions about the events in the scene. This raises the possibility that the ignorance condition may not be the most optimal comparison to the knowledge condition for toddlers. These concerns could be addressed by exploring alternative implementations of the ignorance condition, such as those proposed above.

Differential habituation and task construal across trials between the conditions

A fourth possibility, finally, is that toddlers habituated to and construed the events over time differently in the knowledge and ignorance conditions. In the knowledge condition, from toddlers' perspective the first test trial was the fifth trial (after the four familiarization trials) that was similar in the sense that the chaser went to look for the chasee after watching it hide. So they might have simply begun to lose interest in the task. In contrast, no such habituation may have taken place in the ignorance condition in which the first test trial did differ from the last familiarization trial in that the chaser did not witness the chasee hiding.

In addition, toddlers may have construed the first ignorance test trial in a particularly rich way: For the first time, the chaser did not know where to go—and the chasee knew that. Rather than not keeping track of the agents' perspectives on the scene (under-thinking), toddlers may have engaged in very complex reasoning (over-thinking) about such perspectives along the following lines: They may have looked toward the chasee's box because they expected the chasee to signal its hiding location in some way, to help the chaser find it and allow their cooperative game to continue. From this perspective, toddlers thus not only understood the chaser's ignorance, but they expected the chasee to understand it too and to act accordingly. This made the chasee, rather than the chaser, the focus of toddlers' anticipation: They anticipated that the chasee would somehow signal its location, to help the chaser find it. This interpretation of the ignorance condition also indicates that it may measure something different in toddlers and adults and that anticipatory looking in toddlers can have a different meaning than it has in adults – i.e., a lack of measurement invariance (Meredith, 1964). While such a very rich interpretation does not receive any direct support from the present data, it could be put to systematic test in future studies (for example, by having the two agents interact in less cooperative ways).

Conclusion and future directions

The current large-scale study examined the robustness and reliability of studies using anticipatory looking as a measure of spontaneous Theory of Mind in toddlers and adults. The novel stimuli designed for this study reliably elicited spontaneous goal-based action anticipation, as shown in two pilot studies and in the main study. Spontaneous anticipatory looking occurred to a higher degree and resulted in lower exclusion rates than in previous studies - confirming the importance of the baseline checks, especially in replication studies. In this sense, the current study provides robust stimulus material suitable to elicit goal-based AL in both toddlers and adults.

The main study tested whether toddlers and adults, in their spontaneous action anticipation, take into account the epistemic status of an agent who witnesses relevant events and knows where the target is (knowledge condition) or fails to do so (ignorance condition). Adults clearly did take into account the agent's epistemic status and distinguished between the knowledge and the ignorance condition: they anticipated that the agent would go to the target in the knowledge condition, but expected the opposite in the ignorance condition. In contrast, toddlers showed anticipatory looking to the target location in both conditions, but did so in quantitatively stronger ways in the ignorance than in the knowledge condition, was the opposite than expected.

Future research with adults could build on the present findings, for example, by testing whether adults engage in belief-based action anticipation and thus whether original findings of implicit Theory of Mind in adults can be replicated. Future research with children, in contrast, should first sort out the sources of the unexpected results found here. This will require systematic follow-up studies along several lines: First, the knowledge-ignorance condition contrast could be implemented in alternative, ideally simpler, ways as suggested in the above discussion. Second, an interesting extension would be to test whether young children do indicate some understanding of the present knowledge-ignorance contrasts in other measures. Anticipatory looking itself may be a demanding measure due to its predictive nature (cf. Johnson et al., 1991). Postdictive measures, in contrast, such as looking behavior and pupil dilation in response to events that are/are not expected given the agent's knowledge/ignorance may be more sensitive to uncover early competence (e.g., Daum et al., 2012). Currently, a spin-off project (https://manybabies.org/MB2P/) is running the first follow-up studies in this direction. Finally, looking at children at different ages, for example, older children approaching an age of verbal Theory of Mind reasoning, could shed light on whether their behavior in this task may be related to an underlying conceptual understanding (e.g., Grosse Wiesmann et al., 2018). Taken together, these future studies will hopefully shed more light on the reality and robustness of implicit Theory of Mind from infancy to adulthood.

To conclude, this study represents a critical step forward in understanding the development and robustness of spontaneous Theory of Mind across the lifespan. By developing novel, reliable stimuli and implementing a large-scale, multi-lab approach, it has laid the groundwork for replicable research in this domain. The findings demonstrate that adults' anticipatory looking aligns with epistemic sensitivity, while the unexpected, puzzling patterns in toddlers challenge existing assumptions and thus open up exciting new avenues for future research. By addressing these puzzles, this work paves the way for deeper insights into the developmental trajectory of Theory of Mind and the cognitive mechanisms underlying its expression in infancy, childhood, and beyond.

References

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953.

Baillargeon, R., Buttelmann, D., & Southgate, V. (2018). Invited commentary: Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. *Cognitive Development*, *46*, 112–124.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, *14*(3), 110–118.

Barone, P., Corradi, G., & Gomila, A. (2019). Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis. *Infant Behavior and Development*, *57*, 101350.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. Frontiers Media SA.

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*, 1–28.

Burnside, K., Ruel, A., Azar, N., & Poulin-Dubois, D. (2018). Implicit false belief across the lifespan: Non-replication of an anticipatory looking task. *Cognitive Development*, *46*, 4–11.

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, *112*(2), 337–342.

Buttelmann, F., & Kovács, Á. M. (2019). 14-month-olds anticipate others' actions based on their belief about an object's identity. *Infancy*, *24*(5), 738–751.

Buttelmann, F., Suhrke, J., & Buttelmann, D. (2015). What you get is what you believe: Eighteen-month-olds demonstrate belief understanding in an unexpected-identity task. *Journal of Experimental Child Psychology*, *131*, 94–103.

Carruthers, P. (2013). Mindreading in infancy. *Mind & Language*, *28*(2), 141–172.

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development*, *9*(4), 377–395.

Csibra, G., & Gergely, G. (2007). "Obsessed with goals": Functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychologica*, *124*(1), 60–78.

Daum, M. M., Attig, M., Gunawan, R., Prinz, W., & Gredebäck, G. (2012). Actions seen through babies' eyes: A dissociation between looking time and predictive gaze. *Frontiers in Psychology*, *3*, 370.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Dörrenberg, S., Rakoczy, H., & Liszkowski, U. (2018). How (not) to measure infant theory of mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development*, *46*, 12–30.

Dörrenberg, S., Wenzel, L., Proft, M., Rakoczy, H., & Liszkowski, U. (2019). Reliability and generalizability of an acted-out false belief task in 3-year-olds. *Infant Behavior and Development*, *54*, 13–21.

Elsner, B., & Adam, M. (2021). Infants' goal prediction for simple action events: The role of experience and agency cues. *Topics in Cognitive Science*, *13*(1), 45–62.

Fabricius, W. V., Boyer, T. W., Weimer, A. A., & Carroll, K. (2010). True or false: Do 5-year-olds understand belief? *Developmental Psychology*, *46*(6), 1402.

Flavell, J. H. (1988). *The development of children's knowledge about the mind: From cognitive connections to mental representations.*

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the level 1–level 2 distinction. *Developmental Psychology*, *17*(1), 99.

Forgács, B., Parise, E., Csibra, G., Gergely, G., Jacquey, L., & Gervain, J. (2019). Fourteen-month-old infants track the language comprehension of communicative partners. *Developmental Science, 22*, e12751.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., … Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435.

Frank, M. C., Vul, E., & Saxe, R. (2012). Measuring the development of social attention using free-viewing. *Infancy*, *17*(4), 355–375.

Friedman, O., & Petrashek, A. R. (2009). Children do not follow the rule "ignorance means getting it wrong." *Journal of Experimental Child Psychology*, *102*(1), 114–121.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron*, *50*(4), 531–534.

Ganglmayer, K., Attig, M., Daum, M. M., & Paulus, M. (2019). Infants' perception of goal-directed actions: A multi-lab replication reveals that infants anticipate paths and not goals. *Infant Behavior and Development*, *57*, 101340.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Gergely, G., Nádasdy, Z., Csibra, G., & Bı́ró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193.

Gliga, T., Jones, E. J., Bedford, R., Charman, T., & Johnson, M. H. (2014). From early markers to neuro-developmental mechanisms of autism. *Developmental Review*, *34*(3), 189–207.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., … Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97.

Grosse Wiesmann, C., Friederici, A. D., Singer, T., & Steinbeis, N. (2017). Implicit and explicit false belief development in preschool children. *Developmental Science*, *20*(5), e12445.

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour*, *61*(1), 139–151.

Hayashi, T., Akikawa, R., Kawasaki, K., Egawa, J., Minamimoto, T., Kobayashi, K., et al.others. (2020). Macaques exhibit implicit gaze bias anticipating others' false-belief-driven actions via medial prefrontal cortex. *Cell Reports*, *30*(13), 4433–4444.

Heyes, C. (2014). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science*, *9*(2), 131–143.

Hogrefe, G.-J., Wimmer, H., & Perner, J. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development*, 567–582.

Horschler, D. J., MacLean, E. L., & Santos, L. R. (2020). Do non-human primates really represent others' beliefs? *Trends in Cognitive Sciences*, *24*(8), 594–605.

Johnson, M. H., Posner, M. I., & Rothbart, M. K. (1991). Components of visual orienting in early infancy: Contingency learning, anticipatory looking, and disengaging. *Journal of Cognitive Neuroscience*, *3*(4), 335–344.

Kaminski, J., Call, J., & Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition*, *109*(2), 224–234.

Kampis, D., Buttelmann, F., & Kovács, Á. M. (2020). *Developing a theory of mind: Are infants sensitive to how other people represent the world?*

Kampis, D., Karman, P., Csibra, G., Southgate, V., & Hernik, M. (2021). A two-lab direct replication attempt of Southgate, Senju and Csibra (2007). *Royal Society Open Science*, *8*(8), 210190.

Kano, F., Krupenye, C., Hirata, S., Tomonaga, M., & Call, J. (2019). Great apes use self-experience to anticipate an agent's action in a false-belief test. *Proceedings of the National Academy of Sciences*, *116*(42), 20904–20909.

Karg, K., Schmelz, M., Call, J., & Tomasello, M. (2015). The goggles experiment: Can chimpanzees use self-experience to infer what a competitor can see? *Animal Behaviour*, *105*, 211–221.

Király, I., Oláh, K., Csibra, G., & Kovács, Á. M. (2018). Retrospective attribution of false beliefs in 3-year-old children. *Proceedings of the National Academy of Sciences*, *115*(45), 11477–11482.

Knudsen, B., & Liszkowski, U. (2012). 18-month-olds predict specific action mistakes through attribution of false belief, not ignorance, and intervene accordingly. *Infancy*, *17*(6), 672–691.

Kovács, Á. M. (2016). Belief files in theory of mind reasoning. *Review of Philosophy and Psychology*, *7*, 509–527.

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, *330*(6012), 1830–1834.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*(6308), 110–114.

Kulke, L., Duhn, B. von, Schneider, D., & Rakoczy, H. (2018). Is implicit theory of mind a real and robust phenomenon? Results from a systematic replication study. *Psychological Science*, *29*(6), 888–900.

Kulke, L., & Hinrichs, M. A. B. (2021). Implicit theory of mind under realistic social circumstances measured with mobile eye-tracking. *Scientific Reports*, *11*(1), 1215.

Kulke, L., Johannsen, J., & Rakoczy, H. (2019). Why can some implicit theory of mind tasks be replicated and others cannot? A test of mentalizing versus submentalizing accounts. *PloS One*, *14*(3), e0213772.

Kulke, L., & Rakoczy, H. (2017). How reliable and valid are anticipatory looking measures in theory of mind task? *Are Implicit Theory of Mind Findings Robust? Some Doubts from Converging Non-Replications Across the Lifespan*. Austin, Texas: Society for Research in Child Development Biennial Meeting.

Kulke, L., & Rakoczy, H. (2018). Implicit theory of mind–an overview of current replications and non-replications. *Data in Brief*, *16*, 101–104.

Kulke, L., & Rakoczy, H. (2019). Testing the role of verbal narration in implicit theory of mind tasks. *Journal of Cognition and Development*, *20*(1), 1–14.

Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (2018). How robust are anticipatory looking measures of theory of mind? Replication attempts across the life span. *Cognitive Development*, *46*, 97–111.

Kulke, L., Wübker, M., & Rakoczy, H. (2019). Is implicit theory of mind real but hard to detect? Testing adults with different stimulus materials. *Royal Society Open Science*, *6*(7), 190068.

Leslie, A. M. (2005). Developmental parallels in understanding minds and bodies. *Trends in Cognitive Sciences*, *9*(10), 459–462.

Liszkowski, U., Carpenter, M., & Tomasello, M. (2007). Pointing out new news, old news, and absent referents at 12 months of age. *Developmental Science*, *10*(2), F1–F7.

Low, J., & Watts, J. (2013). Attributing false beliefs about object identity reveals a signature blind spot in humans' efficient mind-reading system. *Psychological Science*, *24*(3), 305–311.

Luo, Y. (2011). Do 10-month-old infants understand others' false beliefs? *Cognition, 121*, 289-298.

Luo, Y., & Baillargeon, R. (2007). Do 12.5-month-old infants consider what objects others can see when interpreting their actions? *Cognition*, *105*(3), 489–512.

Martin, A., & Santos, L. R. (2016). What cognitive representations support primate theory of mind? *Trends in Cognitive Sciences*, *20*(5), 375–382.

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*(4), 269–275.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, *29*(2), 177–185.

Meristo, M., Morgan, G., Geraci, A., Iozzi, L., Hjelmquist, E., Surian, L., & Siegal, M. (2012). Belief attribution in deaf and hearing infants. *Developmental Science*, *15*(5), 633–640.

Moll, H., & Tomasello, M. (2006). Level 1 perspective-taking at 24 months of age. *British Journal of Developmental Psychology*, *24*(3), 603–613.

O'Neill, D. K. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child Development*, *67*(2), 659–677.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258.

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). *Proceedings of the 25th international joint conference on artificial intelligence (IJCAI).*

Perner, J. (1991). *Understanding the representational mind.* The MIT Press.

Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How deep? *Science*, *308*(5719), 214–216.

Phillips, J., Buckwalter, W., Cushman, F., Friedman, O., Martin, A., Turri, J., … Knobe, J. (2021). Knowledge before belief. *Behavioral and Brain Sciences*, *44*, e140.

Poulin-Dubois, D., Rakoczy, H., Burnside, K., Crivello, C., Dörrenberg, S., Edwards, K., et al.others. (2018). Do infants understand false beliefs? We don't know yet–a commentary on baillargeon, buttelmann and southgate's commentary. *Cognitive Development*, *48*, 302–315.

Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development*, *46*, 40–50.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526.

Priewasser, B., Fowles, F., Schweller, K., & Perner, J. (2020). Mistaken max befriends duplo girl: No difference between a standard and an acted-out false belief task. *Journal of Experimental Child Psychology*, *191*, 104756.

Priewasser, B., Rafetseder, E., Gargitter, C., & Perner, J. (2018). Helping as an early indicator of a theory of mind: Mentalism or teleology? *Cognitive Development*, *46*, 69–78.

Rayner, K., Smith, T. J., Malcolm, G. L., & Henderson, J. M. (2009). Eye movements and visual encoding during scene perception. *Psychological Science*, *20*(1), 6–10.

Ruffman, T. (1996). Do children understand the mind by means of simulation or a theory? Evidence from their understanding of inference. *Mind & Language*, *11*(4), 388–414.

Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2012). Eye movements reveal sustained implicit processing of others' mental states. *Journal of Experimental Psychology: General*, *141*(3), 433.

Schneider, D., Slaughter, V. P., Bayliss, A. P., & Dux, P. E. (2013). A temporally sustained implicit theory of mind deficit in autism spectrum disorders. *Cognition*, *129*(2), 410–417.

Schneider, D., Slaughter, V. P., & Dux, P. E. (2017). Current evidence for automatic theory of mind processing in adults. *Cognition*, *162*, 27–31.

Schuwerk, T., Priewasser, B., Sodian, B., & Perner, J. (2018). The robustness and generalizability of findings on spontaneous false belief sensitivity: A replication attempt. *Royal Society Open Science*, *5*(5), 172273.

Scott, R. M., & Baillargeon, R. (2009). Which penguin is this? Attributing false beliefs about object identity at 18 months. *Child Development*, *80*(4), 1172–1196.

Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences*, *21*(4), 237–249.

Scott, R. M., Richman, J. C., & Baillargeon, R. (2015). Infants understand deceptive intentions to implant false beliefs about identity: New evidence for early mentalistic reasoning. *Cognitive Psychology*, *82*, 32–56.

Senju, A., Southgate, V., Miura, Y., Matsui, T., Hasegawa, T., Tojo, Y., … Csibra, G. (2010). Absence of spontaneous action anticipation by false belief attribution in children with autism spectrum disorder. *Development and Psychopathology*, *22*(2), 353–360.

Senju, A., Southgate, V., Snape, C., Leonard, M., & Csibra, G. (2011). Do 18-month-olds really attribute mental states to others? A critical test. *Psychological Science*, *22*(7), 878–880.

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An absence of spontaneous theory of mind in asperger syndrome. *Science*, *325*(5942), 883–885.

Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al.others. (2020). Online developmental science to foster innovation, access, and impact. *Trends in Cognitive Sciences*, *24*(9), 675–678.

Southgate, V., Johnson, M. H., Karoui, I. E., & Csibra, G. (2010). Motor system activation reveals infants' on-line prediction of others' goals. *Psychological Science*, *21*(3), 355–359.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*(7), 587–592.

Southgate, V., & Vernetti, A. (2014). Belief-based action prediction in preverbal infants. *Cognition*, *130*(1), 1–10.

Steffan, A., Zimmer, L., Arias-Trejo, N., Bohn, M., Dal Ben, R., Flores-Coronado, M. A., Franchin, L., Garbisch, I., Grosse Wiesmann, C., Hamlin, J. K., Havron, N., Hay, J. F., Hermansen, T. K., Jakobsen, K. V., Kalinke, S., Ko, E.-S., Kulke, L., Mayor, J., Meristo, M., … Schuwerk, T. (2024). Validation of an open source, remote web-based eye-tracking method (WebGazer) for research in early childhood. *Infancy, 29*(1), 31–55.

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, *18*(7), 580–586.

Surian, L., & Franchin, L. (2020). On the domain specificity of the mechanisms underpinning spontaneous anticipatory looks in false-belief tasks. *Developmental Science*, *23*(6), e12955.

Surian, L., & Geraci, A. (2012). Where will the triangle look for it? Attributing false beliefs to a geometric shape at 17 months. *British Journal of Developmental Psychology*, *30*(1), 30–44.

Thoermer, C., Sodian, B., Vuori, M., Perst, H., & Kristen, S. (2012). Continuity from an implicit to an explicit understanding of false belief from infancy to preschool age. *British Journal of Developmental Psychology*, *30*(1), 172–187.

Träuble, B., Marinović, V., & Pauen, S. (2010). Early theory of mind competencies: Do infants understand others' beliefs? *Infancy*, *15*(4), 434–444.

Wellman, H. M., & Cross, D. (2001). Theory of mind and conceptual change. *Child Development*, *72*(3), 702–707.

Wiesmann, C. G., Friederici, A. D., Disla, D., Steinbeis, N., & Singer, T. (2018). Longitudinal evidence for 4-year-olds' but not 2-and 3-year-olds' false belief-related action anticipation. *Cognitive Development*, *46*, 58–68.

Woodward, A. L., & Sommerville, J. A. (2000). Twelve-month-old infants interpret action in context. *Psychological Science*, *11*(1), 73–77.

Yang, Q., Bucci, M. P., & Kapoula, Z. (2002). The latency of saccades, vergence, and combined eye movements in children and in adults. *Investigative Ophthalmology & Visual Science*, *43*(9), 2939–2949.

Appendix C. Manuscript Study 3

AMERICAN PSYCHOLOGICAL ASSOCIATION

Developmental Psychology

© 2025 The Author(s)
ISSN: 0012-1649

https://doi.org/10.1037/dev0001933

Two- to Three-Year-Old Toddlers Differentiate the Epistemic Verbs "Know" and "Think" in a Preferential Looking Eye-Tracking Paradigm

Lucie Zimmer[1], Beate Sodian[2], Nivedita Mani[3], Stella S. Grosso[2], Susanne Kristen-Antonow[2], and Tobias Schuwerk[1]

[1] Department of Psychology, Clinical Psychology of Childhood and Adolescence, Ludwig-Maximilians-Universität München
[2] Department of Psychology, Developmental Psychology, Ludwig-Maximilians-Universität München
[3] Psychology of Language, Georg-Elias-Müller-Institute for Psychology, Georg-August-Universität Göttingen

The acquisition of mental language understanding is crucial for social-cognitive development. While there is evidence for the production of epistemic terms in the third year of life, the comprehension of different degrees of speaker (un-)certainty has not yet been systematically investigated at this age. In the present longitudinal study, we developed an eye-tracking task and measured preferential looking as an indicator of an implicit understanding of the epistemic verbs "know" and "think" in toddlers twice at the ages of 27 ($N = 199$) and 36 months ($N = 131$). Toddlers were faced with two agents who indicated the location of a hidden object (right vs. left box), with a narrator attributing contrasting degrees of certainty to their statements ("know" vs. "think") before asking the toddlers about the object's location. We measured the extent to which children fixated the box associated with the agent described as knowing the target's location. At both 27 and 36 months of age, we observed systematic differences in their looking behavior toward this box across the trial. Children appeared to display a spontaneous preference for the box associated with the agent who knew the target's location, relative to the agent who only thought the target was in their box in the prequestioning phase. Subsequently, their preference switched in the postquestioning phase; however, this effect was smaller. These results indicate that toddlers in their third year of life distinguish different degrees of speaker (un-)certainty, expressed by the verbs "know" and "think."

Public Significance Statement

The present study found that toddlers both at 27 and 36 months of age distinguish between different degrees of speaker (un-)certainty, displaying a preference for "know" over "think" spontaneously, as indicated by their looking behavior. This suggests that an implicit differentiation between verbs expressing speaker certainty (i.e., "know") and uncertainty (i.e., "think") appears to developmentally precede an explicit one.

Keywords: eye tracking, preferential looking, toddlers, epistemic verbs, mental state language understanding

Supplemental materials: https://doi.org/10.1037/dev0001933.supp

Mikko Peltola served as action editor.

Lucie Zimmer https://orcid.org/0000-0002-5766-2991
Beate Sodian https://orcid.org/0000-0003-1844-5908
Nivedita Mani https://orcid.org/0000-0002-9843-4629
Tobias Schuwerk https://orcid.org/0000-0003-3720-7120

No potential conflict of interest was reported by the authors. This work was supported by the Deutsche Forschungsgemeinschaft (Grant SO 213/33-1,2) awarded to Beate Sodian. The project is part of the research group Crossing the Borders: The Interplay of Language, Cognition, and the Brain in Early Human Development.

Lucie Zimmer played a lead role in data curation, formal analysis, software, validation, visualization, writing–original draft, and writing–review and editing. Beate Sodian played a lead role in conceptualization, funding acquisition, methodology, project administration, resources, supervision, writing–original draft, and writing–review and editing. Nivedita Mani played a lead role in conceptualization, formal analysis, funding acquisition, methodology, software, supervision, visualization, and writing–review and editing. Stella S. Grosso played a lead role in data curation, investigation, and methodology and a supporting role in writing–original draft. Susanne Kristen-Antonow played a lead role in conceptualization and funding acquisition. Tobias Schuwerk played a lead role in conceptualization, funding acquisition, methodology, project administration, resources, supervision, and writing–review and editing and a supporting role in data curation, formal analysis, investigation, software, validation, visualization, and writing–original draft.

Correspondence concerning this article should be addressed to Lucie Zimmer, Department of Psychology, Clinical Psychology of Childhood and Adolescence, Ludwig-Maximilians-Universität München, Leopoldstraße 13, 80802 München, Germany. Email: lucie.zimmer@psy.lmu.de

For children, language is an important source of information about mental states (Harris et al., 2005). Typically, parents talk to toddlers in everyday situations about what the child or the parent "thinks" or "knows." Given that mental states are unobservable, the child cannot map a mental state term onto an external object that the parent may be pointing to. How then do young children figure out the meaning of epistemic verbs, such as "think" and "know"? And at what age do they understand these terms in an adult-like fashion? In the present article, we will briefly review the research literature on this issue, followed by an empirical study presenting a novel preferential-looking eye-tracking method to investigate toddlers' distinction between the verbs "know" and "think."

Recent research indicates that even before children begin to talk about mental states, they possess a rich preverbal understanding of mental states. Declarative joint attention, emerging in infants around the age of 12 months, involves the transmission of information and has, therefore, been interpreted as evidence for an implicit Theory of Mind, characterized by an implicit representation of the communication partner's knowledge states and need for information (see Camaioni et al., 2004; Sodian & Kristen-Antonow, 2015). Experimental evidence supports the view that 1-year-olds distinguish between what others know and don't know from past visual experience when selecting a novel object for their communication partner (Moll et al., 2007; Tomasello & Haberl, 2003). Infants also selectively provide informative pointing gestures to help ignorant but not knowledgeable partners (Dunham et al., 2000; Liszkowski et al., 2008; see O'Neill, 1996, for similar findings in 2-year-olds) and they tend to prefer looking at a stranger (experimenter) rather than their caregiver, when they are seeking information they expect the stranger—but not their caregiver—to possess (Stenberg, 2009). Two- and 3-year-olds also understand what others know based on previous auditory experience (Moll et al., 2014). Children's early understanding of others' epistemic states drives their learning behaviors, for instance when learning selectively from informants' behavior (Schütte et al., 2020). In sum, infants and toddlers appear to represent others' knowledge states appropriately in a wide range of situations. Thus, toddlers can build on these preverbal representations of epistemic states when acquiring verbal labels for knowledge and ignorance.

Research on toddlers' comprehension of mental terms relies mainly on spontaneous speech production data. Talk about mental states begins around the age of 18 months with desire and emotion terms (Bartsch & Wellman, 1995; Bretherton & Beeghly, 1982). References to epistemic states (knowing and believing) appear in the third year of life, primarily with the verbs "know" and "don't know," whereas other epistemic verbs such as "think" are uttered rarely by 2- and 3-year-olds (Shatz et al., 1983).

Early analyses of toddlers' spontaneous mental state language (Bartsch & Wellman, 1995; Shatz et al., 1983) pointed out that usage of epistemic terms does often not clearly refer to mental states but may be seen as conversational routines ("you know") or stereotyped repetitions. More recent analyses of toddlers' naturalistic epistemic state language revealed, however, that 2-year-olds use "know" and "don't know" in conversations spontaneously rather than simply repeating what their conversation partner says. Moreover, they almost always appropriately affirm their own knowledge and that of their conversation partner, and they appropriately express denial of their own knowledge (Harris, Yang, & Cui, 2017). In an experimental study assessing metacognition of own ignorance, 28- to 37-month-old children said "I don't know" or asked for information more frequently when asked to name fictitious, compared to real objects. Even at 16–27 months, toddlers tended to nonverbally express their ignorance in this task (Harris, Ronfard, & Bartz, 2017). Despite these early basic distinctions, metacognition of own ignorance undergoes protracted development in early childhood, with 4- and 5-year-olds failing partial ignorance tasks (Kloo et al., 2017), and often overestimating their own knowledge and reporting knowing something they clearly do not know (de Eccher et al., 2024). Similarly, 3-year-old children are not yet able to understand negations regarding a third-party's knowledge ("the agent doesn't know that it's in the x box"), but they do recognize that the expression "doesn't think" is unreliable (Dudley et al., 2015).

Although there is converging evidence from studies of both early understanding of epistemic states and early use of epistemic verbs indicating that 2- and 3-year-olds make some distinctions between knowledge and ignorance, both with respect to their own and others' knowledge, to date, there is no systematic study directly examining 2- to 3-year-olds' understanding of epistemic verbs. In particular, we do not know whether toddlers distinguish between epistemic verbs expressing varying degrees of certainty, such as "know" and "think."

Factive mental state verbs like "know" are verbs that link an agent to the truth, while nonfactive mental state verbs like "think" or "guess" are verbs that can link an agent to either truth or falsehood (Nagel, 2017). Factive verbs presuppose the truth of a statement, and they thus indicate certainty about the truth of the embedded statement. For instance, the statement "Laura knows that the cat is orange." implies that the cat is orange, whereas in the statement "Laura thinks that the cat is orange," what Laura thinks could be either true or false. Thus, in this context, "think" functions as a nonfactive verb that implies uncertainty about whether the embedded statement is true or not (Abbeduto & Rosenberg, 1985).

In a pioneering study, Moore et al. (1989) investigated 3- to 8-year-old children's ability to distinguish between the verbs "know," "think," and "guess." Children solved a "conflicting sources task," where unknowledgeable participants are given contrasting information by two different informants and are expected to endorse the opinion to which they ascribe more credibility, as indicated by the informant's (un)certainty. Children were presented with contrasting pairs of statements by two puppets about the location of a hidden object, each statement being introduced by an expression of speaker certainty ("I know/I think/I guess the object is in the x box"). Children were then required to locate the object in one of two boxes solely based on the puppets' statements. Children could not match the speakers' assertions against a state of reality. While 3-year-olds were at chance, by the age of 4 years, children were more likely to follow the "know" than the "think" or "guess" puppet, with further improvements apparent between the ages of 4 and 5 years. Kristen-Antonow et al. (2019) used the same paradigm on a larger sample of German-speaking children. Performance was somewhat weaker at all ages than in Moore et al.'s (1989) study. Only from 60 months onward, children reached above-chance competence in either one ofthe know/guess and know/think contrasts. Only at 94 months, more than 50% of the sample reached full competence in both the know/guess and know/think contrasts. The protracted course of development (parallel to that of metacognition) suggests that young children may differentiate "know" from "think" based on varying cues to speaker (un-)certainty rather than on a principled understanding of factive and nonfactive verbs.

In sum, the studies by Moore et al. (1989) and Kristen-Antonow et al. (2019) indicate that a full or explicit understanding of the semantics of "know," "think," and "guess" is only attained at school age, while a basic understanding of the know – think contrast is reached at the age of 4–5 years.

In recent research on the Theory of Mind, implicit understanding (e.g., of knowledge and ignorance or of false belief) has been found to precede full or explicit understanding (Clements & Perner, 1994; Kloo et al., 2020). It is likely that an implicit understanding of mental verbs may also developmentally precede an explicit one. Three-year-olds, who were at chance both in Moore et al.'s (1989) and Kristen-Antonow et al.'s (2019) studies, may not have been able to show their full competence due to task demands, such as the requirements involved in response selection in an explicit choice task (see Setoh et al., 2016). And even 2-year-olds, who appropriately talk about knowledge and ignorance, may possess some understanding of the privileged status of "to know" when evaluating the certainty of a speaker who "knows" against that of a speaker who "thinks."

The present study investigates an implicit understanding of the epistemic verbs "know" and "think" in 27- and 36-month-old children via eye tracking. In a longitudinal design, the same children were examined at two different time points. The longitudinal design allowed for the analysis of the stability of individual differences in looking-time patterns as well as for investigating concurrent and predictive relations of children's production of mental state vocabulary (assessed via parent questionnaire) and their performance in the eye-tracking task. We used a modified version of Moore et al.'s (1989) task in a preferential-looking paradigm. Eye tracking has been successfully used in research on word learning and recognition in young children (e.g., Golinkoff et al., 1987; Taxitari et al., 2020). Preferential looking is a well-established procedure in eye-tracking research to assess cognitive states during language processing in young children and allows for exploring mental phenomena without the need of elaborated language skills (e.g., Madhavan et al., 2024). Experimental word-learning research has not yet addressed the early development of mental state language.

Our study consisted of behavioral familiarization trials, in which toddlers learned to recognize and evaluate contrasting nonepistemic statements (based on Moore et al., 1989), and an experimental eye-tracking task, in which we investigated whether toddlers prefer speaker certainty (indicated by the factive verb "know") over uncertainty (indicated by the nonfactive verb "think"), when prompted to use a linguistic cue. In the eye-tracking task, toddlers were presented with a videotaped hiding game of two monkeys who indicated the location of the hidden object (right vs. left box) with a narrator subsequently attributing certainty versus uncertainty to their statements ("He knows it is in there (A)" vs. "He thinks it is in there (B)") before asking "Where is the sticker?" We measured the pattern of looking behavior to the target (the box associated with the agent described as knowing the sticker was in this box) compared to the distractor (the box associated with the agent described as thinking the sticker was in this box) across two phases of the trial: before the onset of the question and for a period of 2.5 s after the question. Looking reliably longer toward the target than the distractor was evidence for children's differentiation of "know" and "think." We expected this difference to be observable at 27 months of age, indicated by some distinction in at least one of the phases, since 2-year-olds distinguish between "know" and other epistemic verbs in spontaneous speech (Harris, Yang, & Cui, 2017). In 36-month-olds, who do not reliably differentiate "know" and "think" in explicit tasks (Moore et al., 1989), we expected to find reliable competence in all phases of the implicit task, similar to dissociations between implicit and explicit false belief understanding, that have been described in the Theory of Mind literature (e.g., Clements & Perner, 1994).

Method

The present study is part of a large longitudinal project examining the role of language in early Theory of Mind development (see, e.g., Kaltefleiter et al., 2021). Below, we describe only the tasks from the project that are pertinent to the present study.

Participants

The study was conducted at two measurement points: at 27 and 36 months of age. We had longitudinal data (i.e., two valid measurement points) of 121 children (59 girls, 62 boys). All children were born full-term and caregivers did not report any concerns about their (cognitive) development.

In the 27-month-olds sample, 199 children (101 girls, 98 boys) were included. Additional 11 children were tested but excluded due to contributing only one usable trial (n = 4), inattentiveness of the child (n = 3), technical error (n = 3), and interference by the caregiver (n = 1). Their mean age was 26.83 months (SD = 0.32) ranging from 26.33 to 28.20 months.

In the 36-month-olds sample, 131 children were included (67 girls, 64 boys). Additional nine children were tested but excluded due to contributing only one usable trial (n = 3), not conducting the eye-tracking task (n = 3), inattentiveness of the child (n = 2), and interference of the caregiver (n = 1). Their mean age was 35.61 months (SD = 0.44) ranging from 34.20 to 36.75 months.

At the first measurement point, at which the children were 27 months of age, the first caregivers were 36 years on average (n = 191, SD = 4.22) ranging from 24 to 49 years of age and 75% had at least a university degree. The second caregivers were 39 years on average (n = 191, SD = 5.95) ranging from 26 to 63 years of age and 72% had at least a university degree. Regarding the number of siblings, in our sample 49% of the children had no siblings (n = 93), 42% had one sibling (n = 80), and 9% had two siblings or more at the first measurement point.

The children were recruited via birth registries from the local registration office and the laboratory's database. The participating families received a gift and were compensated for their travel expenses at each measurement point. This study was approved by the Ethics Committee of the Department of Psychology and Education of the Ludwig-Maximilians-Universität München.

The sample size of the study was predetermined by the sample size rationale of the larger longitudinal project it was embedded in. To be sufficiently powered, this larger longitudinal project including several measurement points and measures required a respectively larger sample size. Because it was evident before data collection that the targeted sample size would consequently outnumber the required sample size also for this substudy, no additional power analysis was calculated.

Materials and Stimuli

Mental state language understanding was measured via eye-tracking, while children were watching a game including two characters on a screen. The eye-tracking test trials were preceded by behavioral familiarization trials. Here, we used two colored square boxes (9 × 9 × 3 cm, blue or green), a white hiding screen (60 × 50 cm), and two monkey puppets (height: 12 cm) each with a different shade of fur. Eight animal stickers (cat, dog, cow, sheep, duck, rabbit, pig, and chicken) were selected as objects to be hidden in one of the boxes during the game.

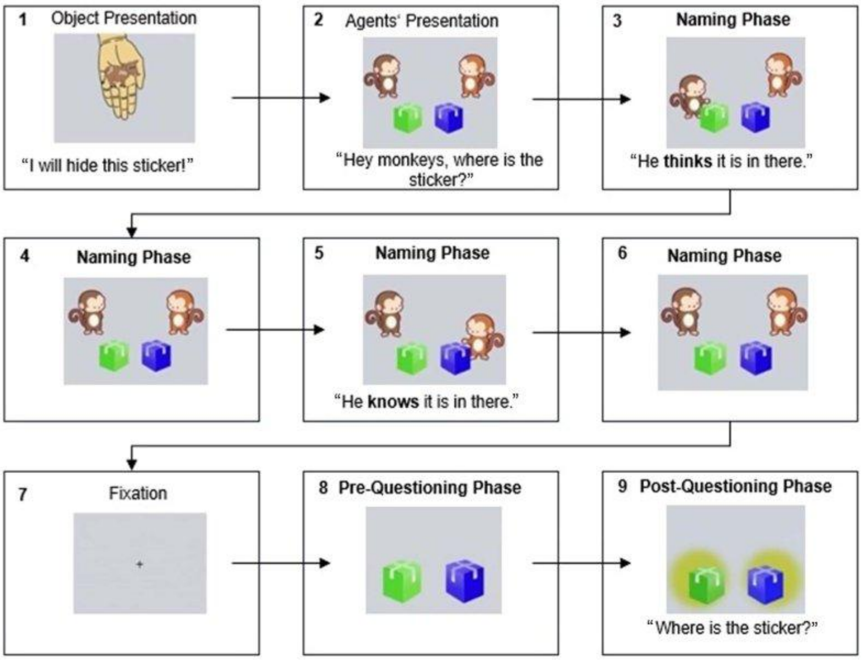
For the test trials of the hiding game, we adapted the paradigm used by Moore et al. (1989) to an eye-tracking setting. We created approximately 35-s-long cartoon movies animated by vector graphics elements on a gray-neutral background (see Figure 1 for a trial example including the time windows under analysis). Prior to the first trial, an introduction video was shown in which the narrator stated "Look at the monkeys! They will help you to find the sticker." At the beginning of each trial, a narrator (not pictured on screen) presented a sticker in her hand and expressed her intention ("I will hide this sticker!"). The animal stickers in the videos were the same the child played with during the behavioral familiarization trials. All voice-overs used in the videos were recorded in German by female experimenters in child-directed speech (for original stimuli in German see Supplemental Materials). The experiment was conducted at an average viewing distance of 60 cm from the screen, based on eye-tracking calibration preceding the experimental recording. Gaze data were recorded with a Tobii T60 eye tracker (Tobii Technology, Sweden) at 60 Hz. The stimuli were presented on a 17" display integrated into the eye tracker with a 1,280 × 1,024 video screen resolution. Both stimulus presentation and data acquisition were performed using the Tobii Studio software (Tobii Technology, Sweden).

Procedure

Sociodemographic information (e.g., age of child, caregiver, education of caregiver) of the participating families and children's understanding and usage of mental language (e.g., know, think,

Figure 1

Example of the Stimuli Presented During the Eye-Tracking Test Trials



Note. (1) Object presentation: first narrator (N1): "I will hide this sticker!" (3 s); (2) presentation of agents and locations: N1: "Hey monkeys, where is the sticker?" (3 s); (3) naming phase: Monkey 1: "Here!," second narrator (N2): "He thinks it is in there." (9 s); (4) second prompt: N1: "Hey monkeys, where is the sticker?" (3 s); (5) naming phase: Monkey 2: "Here!," N2: "He knows it is in there." (9 s); (6) still frame (2 s); (7) fixation screen; (8) response phase—prequestioning phase: both boxes visible (2.5 s); (9) response phase—postquestioning phase: boxes light up, N1: "Where is the sticker?" (2.5 s). See the online article for the color version of this figure.

guess, remember) were collected via caregiver self-report questionnaires at both measurement points.

Mental state language understanding was measured at 27 and 36 months of age and consisted of behavioral familiarization trials and eye-tracking test trials.

Behavioral Familiarization Trials

At the beginning of the session, the child sat alone or on the caregiver's lap at a table. A first experimenter sat in front of the child, while a second experimenter stood to the side. The first experimenter instructed the child that they would play a game, in which the child had to find a hidden sticker in one of two boxes. In the introductory trial, the experimenter showed a sticker to the child and hid it in plain sight. This introductory trial was used to assess the child's capacity to follow the game's instructions and find a hidden object. Once the sticker was hidden, the experimenter asked the child to point to the box in which the sticker was located. In case the child did not point or choose the wrong box, the trial was repeated until the child correctly answered the question.

Following the introductory trial, four practice trials were presented. The first experimenter explained to the child that she was going to hide a sticker in one of the two boxes behind a visual barrier. Children were told that the monkeys would help them find the sticker and that they had to listen carefully to what the monkeys said, because this information would help them find the sticker. After the sticker was hidden and the visual barrier removed, the first experimenter addressed the monkeys asking them where the sticker was. The two monkeys, maneuvered by the second experimenter, moved forward in turn and made a statement (uttering "Here!") indicating a box each. After each monkey's statement, the first experimenter commented either saying "It is really in there!" in case of the target box or "He just pretends that it is inside" in the case of the empty distractor box. At this point, the experimenter asked the child to look for the sticker (e.g., "Where is the sticker?"). After the child's response, the selected box was opened. If the chosen box was the incorrect one, feedback was given to the child by retrieving the sticker from the correct box. Every sticker used in the familiarization was given as a gift to the child, independently of success.

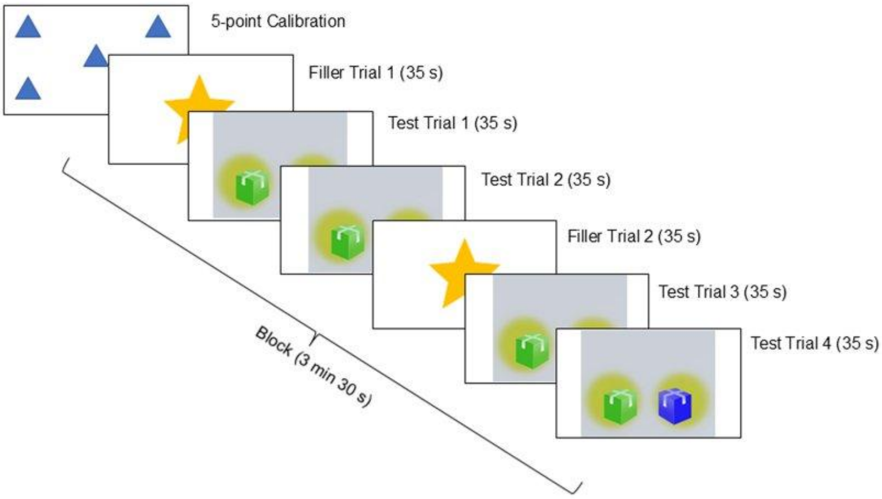
Each child was presented with a total of four practice trials, counterbalanced for color of the boxes, monkey speaking first, hidden location of the sticker, and statement associated with each monkey. The aim of the practice session was to familiarize the child with the goal of the task and the importance of paying attention to the experimenter's statements about the monkeys' (un-)certainty in each trial. From the practice sessions children learned that the monkeys were not always right, a point of fundamental importance for being able to assess uncertainty and discriminate between the epistemic verbs used in the test trials.

Eye-Tracking Test Trials

After the familiarization trials, the eye-tracking trials followed, which took approximately 10 min. The participants moved to the eye-tracking setup, where they sat comfortably in a child safety seat or on their caregiver's lap. The experimenter told the child that they were going to watch some animated movies where two monkeys would help them to find hidden stickers. No further instructions were given. After a 5-point calibration, children watched two blocks of two intermixed fillers and four test trials (see Figure 2). Each of the test and filler trials started with the presentation of an animal sticker (e.g., cow, duck, dog) held in a hand and the narrator saying, "I

Figure 2

Illustration of One of the Two Experimental Blocks (Duration: 3 Min 30 s)



Note. Each block consisted of four test trials and two filler trials (order: Filler Trial 1, Test Trial 1, Test Trial 2, Filler Trial 2, Test Trial 3, Test Trial 4). A trial's duration was 35 s. Prior to the first block, a 5-point calibration was performed. See the online article for the color version of this figure.

will hide this sticker!" Thereafter, the story differed between the filler and test trials. In filler trials, the narrator gave statements similar to the ones used in the familiarization trials ("He just pretends that it is inside" or "It is really in there!") after each monkey moved toward the respective box. At the end of the filler trials, the box opened revealing the real location of the sticker, in agreement with the statement about the monkey who was not pretending. Filler trials were introduced to reward children via feedback (the appearance of the sticker), in order to retain their motivation to look for the sticker throughout the whole block, even in test trials where they did not receive feedback about the location of the sticker.

Order of presentation of trials was randomized and the ordering of blocks across participants was counterbalanced to avoid possible order effects. A break was introduced in between the two blocks to ensure a resting pause for the child before the second part of the recording began. Calibration was performed a second time prior to the onset of the second block. A total of four attention grabbers were inserted across trials to prompt attention to the experimental videos.

Naming Phase. In the next scene, two monkeys standing next to one of the two boxes respectively were presented. The narrator asked: "Hey monkey, where is the sticker?" Following this question, one of the two monkeys started jiggling and moved toward the box next to him and replied: "Here!" After the first monkey spoke, a second narrator was introduced and gave a statement about the knowledge of the monkey, expressing a high ("He knows that it is in there") or low ("He thinks that it is in there") degree of certainty. After the first monkey returned to his original position, the first narrator asked again about the hidden object's location. Now, the second monkey jiggled and moved toward the box next to him. After the second monkey's reply ("Here!") the second narrator gave a statement about the degree of certainty of this monkey, always opposite to the statement about the first monkey's.

Response Phase. A screen with the same gray background and a centered fixation cross was inserted to refocus the attention of the child on the screen before the final part of the video. The screen was paced by the experimenter by mouse click, letting the video continue once the child focused on the fixation cross. In the final sequence, a steady image of the two boxes was presented for 2.5 s (i.e., prequestioning phase). The boxes were highlighted by two bright light circles flashing for 2.5 s during the final prompt question: "Where is the sticker?" (i.e., postquestioning phase). The monkeys were removed from the final scene to avoid biases in gaze behavior of the child due to the expectation, found in a previous pilot study of the experiment, that one of the monkeys would speak again. The color of the boxes, the order of statements referred to the monkey's degree of knowledge, as well as the order of which monkey was speaking first was counterbalanced across trials.

Measures

Behavioral Familiarization Trials

The toddler's pointing or verbal response when asked where the sticker was hidden was transcribed online by a second experimenter. An independent coder confirmed the accuracy of online coding by rescoring children's performance via video recordings of the session. Each child was given a score ranging from 0 (*no trials correct*) to 4 (*all four trials correct*). A score of two indicated performance at the chance level.

Proportion of Target Looking Time

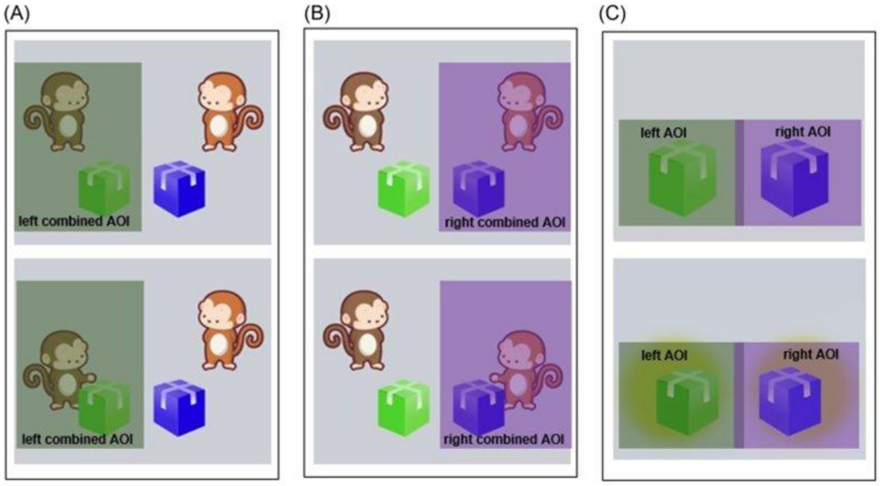
Mental state language understanding was measured in the eye-tracking test trials. Gaze behavior (i.e., fixations at the agents and the object's locations) was recorded across eight trials to compute the proportion of target looking time (PTL) across all phases (i.e., naming phase and response phase consisting of prequestioning and postquestioning phase).

The test trials were segmented into two phases: the naming phase, in which the monkeys and their degree of (un-)certainty were introduced, and the response phase with two time windows (i.e., prequestioning and postquestioning phase), during which only the boxes were presented. The naming phase began with a 2-s time window following the onset of the mental state verb (i.e., knows, defined as target, or thinks, defined as distractor) used by the narrator while the first monkey stood next to the box, to refer to each monkey's degree of knowledge. Then the first monkey returned to their original position and a 4-s time window followed with the second monkey moving to the box and speaking. The naming phase ended with the narrator referring to the second monkey's degree of knowledge and the second monkey returning to his original position. The prequestioning phase began after the fixation cross vanished, and the two boxes were presented for 2.5 s. Then, the postquestioning phase followed, which corresponded to a 2.5-s time window following the onset of the final question "Where is the sticker?"

The first 240 ms at the onset of the first fixation before the start of each phase were removed from the analysis to compensate for the processing time needed for saccade initiation in young children, due to the uncertain nature of the first fixation and the impossibility to clearly link it with task-related behavior (Swingley et al., 1999; Von Holzen & Mani, 2012). In a similar fashion, the last 40 ms of each test trial were removed from analyses to avoid skewed gaze patterns due to imperfect timing between the eye-tracking sampling rate and the segmentation of time windows.

Preprocessing of data was performed in R (Version 4.3.0, R Core Team, 2023). Relevant areas of interest (AOIs) for the analyses were defined as the region around the two boxes and the monkeys' upper body and face (see Figure 3). Four AOIs were defined: left and right monkeys in the naming phase (360 × 250 pixels, if the monkey was moving closer to the box; 300 × 300 pixels, if the monkey was standing still), left and right boxes in all phases (320 × 320 pixels). We added 50 pixels to the AOIs to compensate for inaccuracies of the eye tracker. Proportions of AOIs were chosen to cover equal surface areas. Because of some children's preference for the monkeys (as the most salient objects on the screen) and some toward the boxes (as objects of reference of monkeys' pointing behavior), we merged AOIs of monkeys and boxes on the same side of the screen for the naming phase. The combined AOIs of monkey + box (referred to as "combined AOIs") conveyed the highest amount of valid PTL data. For the naming phase, we then calculated the amount of toddler's fixations on target (the combined AOI of the monkey that knows) over the amount of toddler's fixations to the distractor (the combined AOI of the monkey that thinks) and target, resulting in a proportion of time toddlers spent looking at the target. The score for the naming phase was as follows: PTLNaming Phase = target/target +

Figure 3
Illustration of the AOI Dimensions in Each Phase



Note. Illustration of the AOI dimensions (colored regions) during the (A) naming phase with the left monkey up, (B) naming phase with the right monkey up, and (C) response phase. Dimensions in coordinates (begin, end): left combined AOI: *x*: (0, 460), *y*: (80, 1150) right combined AOI: *x*: (840, 1300), *y*: (80, 1150), left AOI: *x*: (240, 660), *y*: (530, 950), right AOI: *x*: (615, 1035), *y*: (530, 950). AOI = area of interest. See the online article for the color version of this figure.

distractor, see Mani and Plunkett (2007, 2008), for similar analyses.

For the response phase, the amount of toddlers' fixations to the target (the box associated with the monkey that knows) and distractor (the box associated with the monkey that thinks) was calculated in order to get the PTL: PTLResponse Phase = target/ target + distractor. The resulting values range from 1 (*looking only at the target*) to 0 (*looking only at the distractor*) allowing us to assess whether participants fixated the target in preference to the distractor. A score of 0.5 indicates an equal preference for the target and the distractor, whereas a higher score (>0.5) indicates a greater preference for the target.

Trial Exclusion

All participants' video recordings were visually inspected in Tobii Studio. Each trial, in which the participant did not have the chance to follow the action and especially to see the labeling of both monkeys, was excluded from the analysis after visual inspection. Reasons for exclusion after the visual inspection process were as follows: inattentiveness/looking away from the screen and/or interaction with the caregiver and/or experimenter. Participants remained included in analyses if they did provide at least two usable trials after the visual inspection. This resulted in a range from two to eight maximum trials available for analysis per participant. A third of all participants were randomly chosen and coded by a second naive rater to obtain interrater reliability. Cohen's κ resulted in κ = 0.83 for the 27-month-olds sample and κ = 0.76 for the 36-month-olds sample, indicating a substantial to near-perfect interrater agreement.

Statistical Models

All statistical analyses were carried out in R (Version 4.3.0, R Core Team, 2023). First, we tested whether participants looked above chance to the target AOI (i.e., the monkey and the box, that was associated with knowledge) in the naming phase to determine whether the children exhibited a systematic preference already during the introduction of the monkeys and their degree of (un) certainty. We therefore measured the PTL time for each participant averaged across trials in the naming phase and conducted a one-sample *t* test for each of the two age groups.

Second, we tested whether the participants had a preference in the response phase with linear mixed-effects modeling using the lme4 package 1.1-33 (Bates et al., 2015). We fit models for the response phase and included all participants of both measurement points. In each model, PTL at trial level was the dependent variable and we included the maximal random effects structure tolerated by the models (see Barr et al., 2013). We used the drop1 function from the stats package Version 4.3.0 to estimate the effect of removing individual predictors from the full model. Changes in model fit were evaluated using two times the change in the likelihood ratio, which follows a chi-square distribution with degrees of freedom equal to the number of parameters added for each comparison. Statistical significance (*p* values) for individual full models' parameter estimates was assessed using the normal approximation (treating the *t* value as a *z* value). Since our dependent variable was bound between 0 and 1, beta error distribution using glmmTMB (Brooks et al., 2017) would be recommended instead of Gaussian error distribution, but the models did not converge. The models that did converge produced results strikingly similar to the lmer models, thus we consequently opted for lmer models; however, it is essential to employ caution in their interpretation. Unless otherwise specified, we releveled phase, to make the prequestioning phase the baseline and, also age, to make the 27-month-olds sample the baseline. We tested toddlers' target looking in both phases (pre- vs. postquestioning phase) and across age (27 months vs. 36 months) by examining first the interaction between phase and age as fixed effects (Model 1) and second the effect of phase and age without interaction (Model 2). Third, we checked whether there were changes across trials by adding trial number as fixed effect (Model 3). Fourth, we tested whether the results remained the same when considering that some data were longitudinal by controlling for the longitudinal nature of the data (Models 4 and 5). Additionally, we tested the longitudinal stability and conducted an attrition analysis. Fifth, we did additional analyses for each phase and age group separately to further explore the effects (Models 6–9). Sixth, we tested the effect of children's utterances of mental words on toddlers' target looking (Models 10 and 11).

Transparency and Openness

Eye-tracking data, analysis code, and research materials are available at Open Science Framework (see https://osf.io/3gbu2/; Zimmer et al., 2023). Demographic data are not publicly available due to data privacy. This study's design and its analysis were not preregistered.

Results

Descriptives

Children's Utterances of Mental Words

In the 27-month-olds sample, the caregivers reported that 67% (n = 134) of the children used at least one of the mental words measured in this study (i.e., know or think). Sixty-six percent (n = 132) of the children used "know" and "don't know" spontaneously, 19% used "think" (n = 38), whereas 22% (n = 44) had never used "know" in any way.

In the 36-month-olds sample the caregivers reported that 90% (n = 118) of the children used at least "know" or "think". Ninety percent (n = 118) used "know"/"don't know", 57% used "think" (n = 75), and only 2% (n = 3) had never used "know" in any way.

Behavioral Familiarization Trials

On average, toddlers accurately identified the box containing the sticker in two out of four trials, with a mean of 2.25 (*SD* = 0.89) in the 27-month-olds sample and a mean of 2.36 (*SD* = 0.87) in the 36-month-olds sample. Thus, they passed 56% and 59% of the trials, respectively. Twelve and four toddlers, respectively, had no data in the behavioral familiarization trials due to noncompliance/fussiness, or experimenter error.

Proportion of Target Looking

In the naming phase, in which the monkeys were introduced, the average PTL was 0.50 (*SD* = 0.10) in the 27-month-olds and 0.50 (*SD* = 0.09) in the 36-month-olds (see Supplemental Figure 1).

In the prequestioning phase, the average PTL was 0.54 (*SD* = 0.14) in the 27-month-olds and 0.53 (*SD* = 0.16) in the 36-month-olds. For a comparison of both age groups see Figure 4.

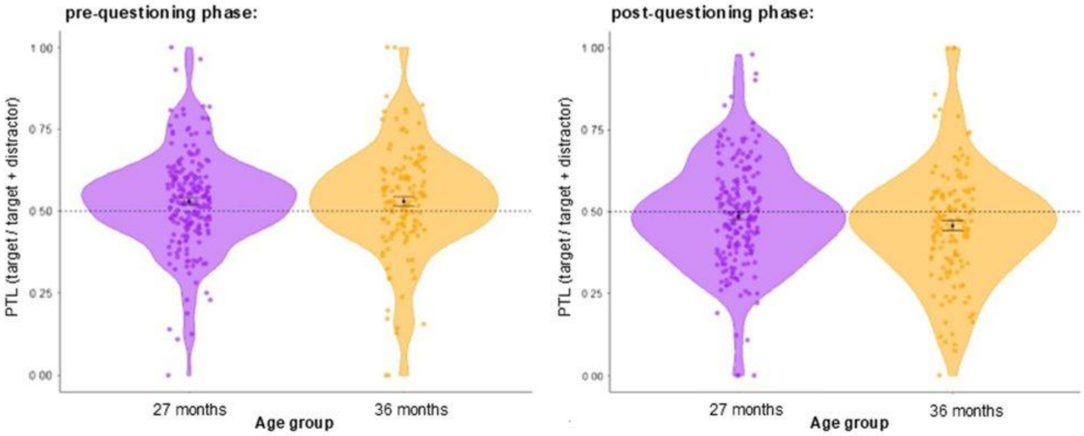
In the postquestioning phase, the average PTL was 0.49 (*SD* = 0.15) in the 27-month-olds and 0.46 (*SD* = 0.18) in the 36-month-olds. For a comparison of the dummy-coded preference of AOI (−1 = distractor, 1 = target) during the response phase see Figure 5 (for a detailed comparison of each trial see Supplemental Figure 2).

PTL in the Naming Phase

The PTL was not significantly different from the chance level (0.5) for both, the 27-month-olds with p = .874 and the 36-month-olds with

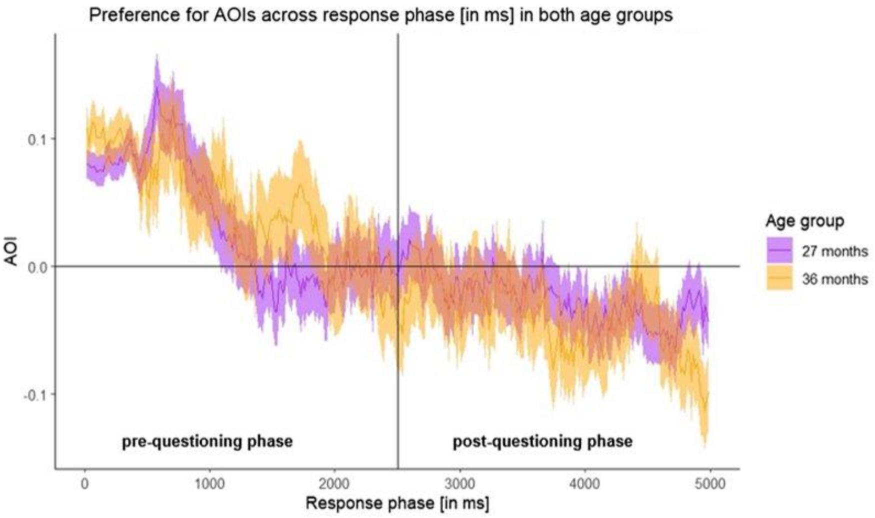
Figure 4

Proportion of Target Looking Time in the Response Phase



Note. Graphs depicting the proportion of target looking time (looking time to target AOI/looking time to target + distractor AOI) in the pre- and postquestioning response phase (y-axis) per age (x-axis) including error bars (upper: $M + SE$, lower: $M - SE$). PTL = proportion of target looking time; AOI = area of interest; SE = standard error. See the online article for the color version of this figure.

Figure 5
Preference for AOIs Across Response Phase



Note. Graph depicting the preference for the AOIs (AOIs were dummy coded: −1 = distractor, 1 = target, *y*-axis) during the response phase in ms (<2,500 ms: prequestioning phase, >2,500 ms postquestioning phase, *x*-axis) per age group averaged across all trials. AOI = area of interest. See the online article for the color version of this figure.

p = .836, reflecting that participants did not show above chance looking to the target AOI (i.e., the monkey that knows) in the naming phase. This indicates that during the introduction of the monkeys and their degree of (un)certainty, the children did not show any preferences.

PTL in the Response Phase

To evaluate toddlers' target looking in both phases (pre- vs. postquestioning phase) and across age (27 months vs. 36 months), we ran a model with PTL as the dependent variable examining the interaction between phase and age as fixed effects. As random effects, we included variation in the intercept and the effect of phase and trial number across participants, the effect of age across trial numbers and the effect of trial number and age across position of the target. The specification of Model 1 (full model) was as follows:

$$\text{ptl} \sim \text{phase} \times \text{age} + (1 + \text{phase} + \text{trial.no}||\text{id}) + (1 + \text{age}||\text{trial.no}) + (1 + \text{trial.no} + \text{age}||\text{target}).$$

In the null model we only included age as a fixed effect. The chi-squared statistic revealed a significant improvement in model fit within the full model, $\chi^2(2) = 24.451$, $p < .001$. The analysis revealed no significant interaction between age and phase ($p = .292$), but a significant difference of PTL from the pre- to the postquestioning phase ($p = .001$), indicating a significant reduction in looks to the target in the postquestioning phase relative to the prequestioning phase at 27 months of age (see Table 1).

Additionally, we fit a model without the interaction of phase and age, but with adding both as fixed effects. The specification of Model 2 (full model) was as follows:

$$\text{ptl} \sim \text{phase} + \text{age} + (1 + \text{phase} + \text{trial.no}||\text{id}) + (1 + \text{age}||\text{trial.no}) + (1 + \text{trial.no} + \text{age}||\text{target}).$$

Again, in the null model we only included age as a fixed effect. The chi-squared statistic revealed a significant improvement in model fit within the full model, $\chi^2(1) = 23.343$, $p < .001$. The analysis revealed no significant effect of age ($p = .627$), but a significant effect of phase across both ages, $p < .001$. Thus, the results revealed a significant difference of PTL from the pre- to the postquestioning phase, indicating a significant reduction in looks to the target in the postquestioning phase relative to the prequestioning phase across both age groups (see Table 2).

We added trial number as fixed effect to the no-interaction model (i.e., Model 3), which revealed no significant effect of trial

Table 1
Results for Model 1: LMEM Examining the Interaction of Age and Phase (i.e., Interaction Model)

Term	Estimate	*SE*	*t* value	*p*
Intercept	0.531	0.039	13.803	.030
PhasePost	−0.043	0.013	−3.188	.001
Age 36M	−0.003	0.025	−0.128	.913
PhasePost: Age 36M	−0.023	0.022	−1.053	.292

Note. LMEM = linear mixed-effects modeling; *SE* = standard error; M = month.

Table 2

Results for Model 2: LMEM Examining Age and Phase Without Interaction (i.e., No-Interaction Model)

Term	Estimate	*SE*	*t* value	*p*
Intercept	0.536	0.038	13.993	.032
PhasePost	−0.052	0.011	−4.869	<.001
Age 36M	−0.015	0.022	−0.662	.627

Note. LMEM = linear mixed-effects modeling; *SE* = standard error; M = month.

number (p = .842), indicating no evidence that the difference between pre- and postquestioning phase changes across trials (see Supplemental Table 1).

We also controlled for the longitudinal nature of data in both, the interaction model (i.e., Model 4) and the model without interaction (i.e., Model 5) and found that when we consider the fact that some data are longitudinal, the significant effect of phase does neither change in the interaction model (see Supplemental Table 2) nor in the no-interaction model (see Supplemental Table 3). To test the longitudinal stability of the response patterns, we fit linear models for each phase and found that in both phases the performance at 27 months of age did not predict performance at 36 months of age (see Supplemental Tables 4 and 5). Additionally, we conducted an attrition analysis testing whether there was a difference in 27 months performance between those who dropped out and those who attended again at 36 months of age and found no effect (see Supplemental Table 6).

Although we found no significant interaction between age and phase, we fit models to further explore the effect of phase in both age groups by rereferencing the model. The analysis revealed no difference between the age groups (see Models 6–9 in Supplemental Tables 7–10). Taken together, regardless of age group, children showed a decrease in looking at the target from the pre- to the postquestioning phase. This is indexed by the significant effect of phase in the models and the fact that this does not interact with age. This effect neither interacts with trial number nor does this effect change when we consider the fact that some data is longitudinal. Interestingly, the effect in the postquestioning phase is smaller than in the prequestioning phase.

Finally, we further explored the effect of children's utterances of mental words by adding the variables "know" (i.e., children use know/don't know) and "think" (i.e., children use think) as fixed effects. We analyzed these effects separately for both age groups (i.e., Model 10 for the 27-month-olds and Model 11 for the 36-month-olds). The model specification was as follows:

$$\text{ptl} \sim \text{phase} + \text{know} + \text{think} + (1 + \text{trial.no}||\text{id}) + (1 + \text{trial.no}||\text{target}).$$

The analyses for the 27-month-olds revealed no significant effect of know (p = .088), but additional to the remaining significant effect of phase (p = .001), a significant effect of think with p = .026 (see Table 3). Nevertheless, it is important to note that only 38 out of 199 toddlers (i.e., 19%) were already using the verb "think" at 27 months of age. We did not find a significant effect of mental utterances in the 36-month-olds (know: p = .840, think: p = .755; see Supplemental Table 11).

Discussion

In the present study, we measured toddlers' gaze behavior in a facilitated "conflicting sources task" adapted from Moore et al. (1989) designed to assess their comprehension of the epistemic verbs "know" and "think." In a longitudinal design, the eye-tracking task was administered twice, at the ages of 27 and 36 months. In both age groups, children distinguished between speaker certainty ("know") and uncertainty ("think"), with a significant preference for "know" at the beginning of the response phase (i.e., the prequestioning phase). Thus, an implicit differentiation between verbs expressing speaker certainty and uncertainty appears to developmentally precede an explicit distinction which has not been reported before the fourth birthday (Kristen-Antonow et al., 2019; Moore et al., 1989). These results also suggest that children are able to differentiate mental terms differing in certainty even before they can express these verbally which is consistent with research indicating that toddlers between 1 and 2.5 years differentiate between others' knowledge and ignorance in interactive as well as in looking-time tasks (Dunham et al., 2000; Liszkowski et al., 2008; Scott & Baillargeon, 2017, for a review).

The present findings are consistent with recent analyses of toddlers' early spontaneous use of the terms "know" and "don't know" (Harris, Yang, & Cui, 2017) and their performance in a test of metacognitive awareness of their own ignorance (Harris, Ronfard, & Bartz, 2017), indicating that 2- and 3-year-olds distinguish between knowledge and ignorance in their spontaneous speech, both with respect to their own and others' epistemic states.

Our exploratory analyses revealed that the usage of "think" at 27 months of age was related to a higher PTL time, while "know" usage did not show a significant association. This suggests that expressing a mental verb with a lower degree of speaker certainty at 27 months of age may relate to mastering the implicit conflicting source task by differentiating "know" from "think." These results should, however, be interpreted with caution as only 19% of the children in this age group used "think" (compared to 66% for "know"). At 36 months of age, the results did not reveal an association between children's target looking time and their use of mental words.

The present findings indicated that both age groups differentiated spontaneously between "know" and "think" in the eye-tracking task.

Table 3

Results for Model 10: LMEM Examining Mental Utterances of Know and Think in the Subset of the 27-Month-Olds

Term	Estimate	*SE*	*t* value	*p*
Intercept	0.545	0.042	12.871	.031
PhasePost	−0.047	0.014	−3.404	.001
Know 27M	−0.027	0.016	−1.721	.088
Think 27M	0.042	0.019	2.250	.026

Note. LMEM = linear mixed-effects modeling; *SE* = standard error; M = month.

However, the whole pattern of findings was more complex. In both age groups, children had a preference for "know" compared to "think" at the beginning of the response phase, when only the two boxes were presented (i.e., prequestioning phase), whereas they preferred "think" over "know" at the end of the response phase, after being prompted about the location of the sticker (i.e., in the postquestioning phase), indicated by a significant reduction in looks to the box associated with "know" in the postquestioning phase in comparison to the prequestioning phase. The preference for "know" in the prequestioning phase was, however, higher than the preference for "think" in the postquestioning phase, indicating a spontaneous preference for "know" over "think." On average, children of both age groups showed the same pattern, that is, a switching of preference from "know" to "think" in the middle of the 5-s-long eye-tracking interval, the "think"-preference being less distinct than the initial, spontaneous "know"-preference. This shift of attention suggests a potential recalibration of children's preferences; however, the reason for this shift remains unclear. One explanation could be an inhibition of return effect which is the tendency for a counteractive response to emerge after a preference has been established (Klein, 1988; Klein & MacInnes, 1999). Usually, this effect occurs with a delay, such that it is no longer directly related to the stimulus, that is, the presentation of the two boxes at the beginning of the response phase. This would mean that the children clearly preferred "know" at the beginning, but then simply considered the other option—in this case the box associated with "think"—which may no longer be related to the stimulus. Also, since nothing happened (i.e., the sticker did not reappear) at the location where children first spent most of their looking time during a time frame of 2.5 s, they might have become uncertain about their choice and/or they may have become curious about the alternative option and possible events at this other location, leading them to reassess their initial choice. This explanation is further underscored by the fact that the response phase was quite long (i.e., 5 s) compared to other eye-tracking test phases and thus children may have difficulty maintaining attention and looking only to one side during the whole phase.

Counter to expectation, there was no consistent response to the question about the location of the sticker by a subsequent look at the target box. We assume that after the first trial, children anticipated the question and immediately looked at the correct location upon the first presentation of the boxes, rather than waiting for the question to be asked. Thus, children may already have been cognitively engaged upon the presentation of the boxes, spontaneously looking toward the target, leaving them with insufficient capacity to provide a response when prompted again, since they had already focused on the target box.

Limitations

In interpreting the present findings, we have to consider that the task might have been too demanding for our age group. Gaze recording implies long sessions sitting still in front of a screen, where toddlers are required to keep 5–10 min of sustained attention. Through the introduction of pauses, attention grabbers, and subsequent recalibration procedures it was possible to record good-quality data. Because of this fragmented recording style, only a careful selection of trials (i.e., 65%–70% of all trials), where children continuously directed their gaze (recorded via an in-sync video camera) on the screen, could be included in the final analyses. Yet, this study has shown that eye tracking is a powerful tool to help detect implicit comprehension of epistemic verbs in early childhood: infants as early as 2.5 years of age show a striking sensitivity for the semantic distinction between "know" and "think."

Implications

Due to the novel nature of our task, future studies should validate the unexpected pattern of our study and focus on disentangling possible effects. First, future research should empirically test the interpretation of our results in the prequestioning phase, specifically whether children anticipate the narrator's location question after hearing it the first time (e.g., by conducting trials in which the box associated with an agent who "thinks" is the target). Second, as "know" is more frequent than "think" in both German (and English) children's input, future research should investigate a potential familiarity effect. In case of confirmatory evidence, we suggest investigating the developmental origins of the basic discrimination between epistemic verbs. This entails shedding further light onto whether the epistemic understanding of high degrees of certainty verbs such as know precedes the one of lower degrees of certainty; or whether an initial, full-fledged understanding of know is a prerequisite for learning about mental states expressing different shades of knowledge of the world. To reach this goal, methods from linguistic research and developmental psychology should be encouraged, as their synergistic use can provide a broader spectrum of insights in the developing mind. More specifically, in subsequent studies, the test trials could be enhanced by posing the location question immediately after introducing both agents (i.e., naming phase), followed by displaying the two boxes for a shorter period, which is less than 5 s. This adjustment would allow for an assessment of whether the identified pattern can be replicated, even with a more homogeneous and shorter testing phase that better fits to the attention span of this age group.

Conclusion

Our study of implicit epistemic verb comprehension demonstrates that toddlers as young as 27 months distinguish between degrees of speaker (un-)certainty, preferring "know" over "think" spontaneously. This challenges traditional assumptions about the development of mental state language understanding on the one hand and extends results from explicit measurement on the other hand that both propose a more protracted development. Adjustments to testing methodologies may enhance future investigations using our eye-tracking task within a preferential looking paradigm, providing a further elaboration of our findings and, more generally, a deeper understanding of early cognitive development.

References

Abbeduto, L., & Rosenberg, S. (1985). Children's knowledge of the presuppositions of know and other cognitive verbs. *Journal of Child Language*, *12*(3), 621–641. https://doi.org/10.1017/S0305000900006693

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind*. Oxford University Press. https://doi.org/10.1093/oso/9780195080056.001.0001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bretherton, I., & Beeghly, M. (1982). Talking about internal states: The acquisition of an explicit theory of mind. *Developmental Psychology*, *18*(6), 906–921. https://doi.org/10.1037/0012-1649.18.6.906

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.32614/RJ-2017-066

Camaioni, L., Perucchini, P., Bellagamba, F., & Colonnesi, C. (2004). The role of declarative pointing in developing a theory of mind. *Infancy*, *5*(3), 291–308. https://doi.org/10.1207/s15327078in0503_3

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development*, *9*(4), 377–395. https://doi.org/10.1016/0885-2014(94)90012-4

de Eccher, M., Mundry, R., & Mani, N. (2024). Children's subjective uncertainty-driven sampling behaviour. *Royal Society Open Science*, *11*(4), Article 231283. https://doi.org/10.1098/rsos.231283

Dudley, R., Orita, N., Hacquard, V., & Lidz, J. (2015). Three-year-olds' understanding of *know* and *think*. In F. Schwarz (Ed.), *Experimental perspectives on presuppositions* (Vol. 45, pp. 241–262). Springer. https://doi.org/10.1007/978-3-319-07980-6_11

Dunham, P., Dunham, F., & O'Keefe, C. (2000). Two-year-olds' sensitivity to a parent's knowledge state: Mind reading or contextual cues? *British Journal of Developmental Psychology*, *18*(4), 519–532. https://doi.org/10.1348/026151000165832

Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., & Gordon, L. (1987). The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of Child Language*, *14*(1), 23–45. https://doi.org/10.1017/S030500090001271X

Harris, P. L., de Rosnay, M., & Pons, F. (2005). Language and children's understanding of mental states. *Current Directions in Psychological Science*, *14*(2), 69–73. https://doi.org/10.1111/j.0963-7214.2005.00337.x

Harris, P. L., Ronfard, S., & Bartz, D. (2017). Young children's developing conception of knowledge and ignorance: Work in progress. *European Journal of Developmental Psychology*, *14*(2), 221–232. https://doi.org/10.1080/17405629.2016.1190267

Harris, P. L., Yang, B., & Cui, Y. (2017). "I don't know": Children's early talk about knowledge. *Mind & Language*, *32*(3), 283–307. https://doi.org/10.1111/mila.12143

Kaltefleiter, L. J., Sodian, B., Kristen-Antonow, S., Grosse Wiesmann, C., & Schuwerk, T. (2021). Does syntax play a role in theory of mind development before the age of 3 years? *Infant Behavior and Development*, *64*, Article 101575. https://doi.org/10.1016/j.infbeh.2021.101575

Klein, R. (1988). Inhibitory tagging system facilitates visual search. *Nature*, *334*(6181), 430–431. https://doi.org/10.1038/334430a0

Klein, R. M., & MacInnes, W. J. (1999). Inhibition of return is a foraging facilitator in visual search. *Psychological Science*, *10*(4), 346–352. https://doi.org/10.1111/1467-9280.00166

Kloo, D., Kristen-Antonow, S., & Sodian, B. (2020). Progressing from an implicit to an explicit false belief understanding: A matter of executive control? *International Journal of Behavioral Development*, *44*(2), 107–115. https://doi.org/10.1177/0165025419850901

Kloo, D., Rohwer, M., & Perner, J. (2017). Direct and indirect admission of ignorance by children. *Journal of Experimental Child Psychology*, *159*, 279–295. https://doi.org/10.1016/j.jecp.2017.02.014

Kristen-Antonow, S., Jarvers, I., & Sodian, B. (2019). Preschoolers' developing understanding of factivity in mental verb comprehension and its relation to first- and second-order false belief understanding: A longitudinal study. *Journal of Cognition and Development*, *20*(3), 354–369. https://doi.org/10.1080/15248372.2019.1586710

Liszkowski, U., Carpenter, M., & Tomasello, M. (2008). Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. *Cognition*, *108*(3), 732–739. https://doi.org/10.1016/j.cognition.2008.06.013

Madhavan, R., Malem, B., Ackermann, L., Mundry, R., & Mani, N. (2024). An examination of measures of young children's interest in natural object categories. *Cortex*, *175*, 124–148. https://doi.org/10.1016/j.cortex.2024.02.015

Mani, N., & Plunkett, K. (2007). Phonological specificity of vowels and consonants in early lexical representations. *Journal of Memory and Language*, *57*(2), 252–272. https://doi.org/10.1016/j.jml.2007.03.005

Mani, N., & Plunkett, K. (2008). Fourteen-month-olds pay attention to vowels in novel words. *Developmental Science*, *11*(1), 53–59. https://doi.org/10.1111/j.1467-7687.2007.00645.x

Moll, H., Carpenter, M., & Tomasello, M. (2007). Fourteen-month-olds know what others experience only in joint engagement. *Developmental Science*, *10*(6), 826–835. https://doi.org/10.1111/j.1467-7687.2007.00615.x

Moll, H., Carpenter, M., & Tomasello, M. (2014). Two- and 3-year-olds know what others have and have not heard. *Journal of Cognition and Development*, *15*(1), 12–21. https://doi.org/10.1080/15248372.2012.710865

Moore, C., Bryant, D., & Furrow, D. (1989). Mental terms and the development of certainty. *Child Development*, *60*(1), 167–171. https://doi.org/10.2307/1131082

Nagel, J. (2017). Factive and non-factive mental state attribution. *Mind & Language*, *32*(5), 525–544. https://doi.org/10.1111/mila.12157

O'Neill, D. K. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child Development*, *67*(2), 659–677. https://doi.org/10.2307/1131839

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Schütte, F., Mani, N., & Behne, T. (2020). Retrospective inferences in selective trust. *Royal Society Open Science*, *7*(2), Article 191451. https://doi.org/10.1098/rsos.191451

Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences*, *21*(4), 237–249. https://doi.org/10.1016/j.tics.2017.01.012

Setoh, P., Scott, R. M., & Baillargeon, R. (2016). Two-and-a-half-year-olds succeed at a traditional false-belief task with reduced processing demands. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(47), 13360–13365. https://doi.org/10.1073/pnas.1609203113

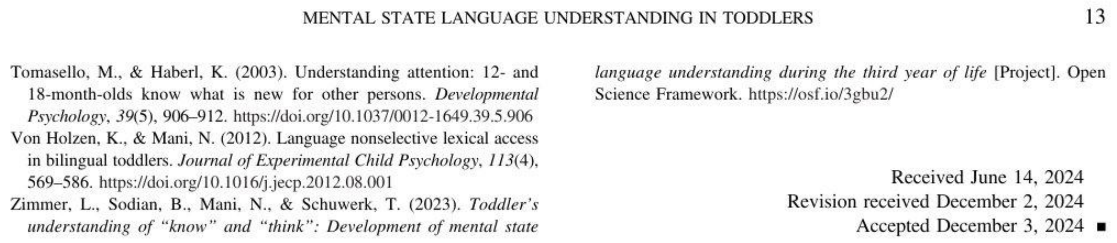
Shatz, M., Wellman, H. M., & Silber, S. (1983). The acquisition of mental verbs: A systematic investigation of the first reference to mental state. *Cognition*, *14*(3), 301–321. https://doi.org/10.1016/0010-0277(83)90008-2

Sodian, B., & Kristen-Antonow, S. (2015). Declarative joint attention as a foundation of theory of mind. *Developmental Psychology*, *51*(9), 1190–1200. https://doi.org/10.1037/dev0000039

Stenberg, G. (2009). Selectivity in infant social referencing. *Infancy*, *14*(4), 457–473. https://doi.org/10.1080/15250000902994115

Swingley, D., Pinto, J. P., & Fernald, A. (1999). Continuous processing in word recognition at 24 months. *Cognition*, *71*(2), 73–108. https://doi.org/10.1016/S0010-0277(99)00021-9

Taxitari, L., Twomey, K. E., Westermann, G., & Mani, N. (2020). The limits of infants' early word learning. *Language Learning and Development*, *16*(1), 1–21. https://doi.org/10.1080/15475441.2019.1670184



Tomasello, M., & Haberl, K. (2003). Understanding attention: 12- and 18-month-olds know what is new for other persons. *Developmental Psychology*, *39*(5), 906–912. https://doi.org/10.1037/0012-1649.39.5.906

Von Holzen, K., & Mani, N. (2012). Language nonselective lexical access in bilingual toddlers. *Journal of Experimental Child Psychology*, *113*(4), 569–586. https://doi.org/10.1016/j.jecp.2012.08.001

Zimmer, L., Sodian, B., Mani, N., & Schuwerk, T. (2023). *Toddler's understanding of "know" and "think": Development of mental state language understanding during the third year of life* [Project]. Open Science Framework. https://osf.io/3gbu2/

Received June 14, 2024
Revision received December 2, 2024
Accepted December 3, 2024 ■

Supporting information: **https://psycnet.apa.org/doi/10.1037/dev0001933.supp**

Appendix D. Manuscript Study 4

cortex 187 (2025) 159–171

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Research Report

Predictive responses in the Theory of Mind network: A comparison of autistic and non-autistic adults

Lucie Zimmer [a,*], *Hilary Richardson* [b], *Carolina Pletti* [c], *Markus Paulus* [a] *and Tobias Schuwerk* [a]

[a] *Department of Psychology, Ludwig-Maximilians-Universität München, München, Germany*
[b] *School of Philosophy, Psychology, and Language Sciences, The University of Edinburgh, Edinburgh, United Kingdom*
[c] *Department of Developmental and Educational Psychology, University of Vienna, Wien, Austria*

ARTICLE INFO

Article history:
Received 17 October 2024
Revised 1 April 2025
Accepted 1 April 2025
Action editor Punit Shah
Published online 30 April 2025

Keywords:
Autism spectrum disorder
Theory of mind
fMRI
Predictive coding
Narrative anticipation
Repetition suppression

ABSTRACT

Social cognitive processes, particularly Theory of Mind (ToM) reasoning, appear to differ between autistic and non-autistic individuals. This has been proposed to reflect the autistic core symptomatology of communication and social interaction difficulties. According to the predictive coding theory, autistic individuals' ToM reasoning difficulties arise from an attenuated use of prior information about others' mental states to explain and predict their behavior. This reduced use of prior assumptions makes the social world less predictable for autistic people, causing interactive mismatch and stress. Despite strong theoretical claims, robust and replicable neural differences in ToM brain regions remain elusive. Here, we investigated whether brain regions supporting ToM reasoning anticipate a narrative during repeated exposure (i.e., the narrative anticipation effect) in non-autistic adults (Experiment 1) and tested whether this effect was attenuated in autistic adults (Experiment 2). We presented a short movie with a plot including mental states with associated actions, twice, to 61 non-autistic adults who underwent functional magnetic resonance imaging [Experiment 1: $M(SD)_{age} = 25.9(4.4)$ years]. In Experiment 2, we used the same protocol with 30 autistic [$M(SD)_{age} = 32.4(10.7)$ years] and 30 non-autistic adults [$M(SD)_{age} = 33.2(10.1)$ years]. Analyses revealed no narrative anticipation effect in the ToM network in either group. Exploratory reverse correlation analyses identified a ToM scene that evoked a smaller difference in response between movie viewings (i.e., less repetition suppression) in autistic adults, compared to non-autistic adults. In sum, our study shows that predictive processing in the ToM network during a naturalistic movie-viewing experiment was absent in adults. Subtle differences in a key scene provide preliminary neural evidence for the predictive coding theory and open a promising avenue for future research to better understand the nature of differences in social interaction in autistic adults.

* *Corresponding author*. Ludwig-Maximilians-Universität, Leopoldstr. 13, 80802, München, Germany.
E-mail address: lucie.zimmer@psy.lmu.de (L. Zimmer).

https://doi.org/10.1016/j.cortex.2025.04.006

1. Introduction

According to DSM-5, Autism Spectrum Disorder (ASD, hereafter *autism*) is primarily characterized by two core symptoms: difficulties in social interaction and communication, and stereotyped behavior (American Psychiatric Association, 2013). Cognitive research focused on the social aspect of these core symptoms often examines differences in basic social cognitive processes between autistic and non-autistic individuals. For instance, autistic individuals can face difficulties in ascribing mental states, such as desires and beliefs, to explain and predict the behavior of others (Frith, 2012), the core of Theory of Mind (ToM) reasoning. It is believed that these difficulties are caused by altered social cognitive processes (Kennedy & Adolphs, 2012). However, difficulties of autistic individuals are observable in some but not all ToM tasks (Gernsbacher & Yergeau, 2019) indicating that applying typical ToM tasks may not fully reveal the origins of the challenges experienced by autistic individuals.

The predictive coding theory (Clark, 2013) offers a comprehensive framework for understanding the core autism symptoms, including social difficulties. In particular, others' actions can be predicted from mental states ascribed based on available prior information about the acting person, the corresponding situation and/or people in general. Similarly, mental states of others can be actively predicted (Koster-Hale & Saxe, 2013). Following this hypothesis, social difficulties in autism may arise from weakened social cognitive predictions, which lead autistic individuals to perceive social interactions as unpredictable, cause interactive mismatch (i.e., interactions in which behaviors and expectations of individuals involved do not align), and create stress (Bolis et al., 2017; Pellicano & Burr, 2012; Sinha et al., 2014).

While it has been suggested that predicting others' actions is generally atypical in autism, autistic children appear to attribute goals to others (e.g., Cattaneo et al., 2007; Somogyi et al., 2013). Recent research reported that autistic people do anticipate actions, make correct action predictions, and apply the same cognitive strategies (e.g., goal-directed eye movements) as individuals without autism (Falck-Ytter, 2010; Schuwerk & Paulus, 2018). Moreover, autistic adults learn quickly from action-outcome contingencies, allowing them to make accurate predictions upon subsequent encounters (Schuwerk et al., 2015). Although autistic individuals are able to anticipate another's action goal, evidence suggests that they use prior information less, thereby requiring more time compared to non-autistic individuals (Ganglmayer et al., 2020). Given these observations, the differences between autistic and non-autistic individuals appear to be subtler than previously assumed, indicating that the difficulties may lie in the nuanced use of prior experience and contextual cues. Investigations of predictive ToM reasoning in more complex and naturalistic situations may be important for capturing subtle differences in autistic individuals, but such investigations remain rare.

Following Clark's (2013) original neural examination of predictive coding, exploring the predictive coding theory in a naturalistic context at a neural level using functional magnetic resonance imaging (fMRI) could reveal underlying cognitive differences in autism. For instance, neurotypical adults' prior knowledge of a narrative seems to enable neural anticipation of event patterns (i.e., brain regions were recruited earlier in time indicating an anticipation of the narrative during repeated exposure; Baldassano et al., 2017); brain regions that show this effect include those in the *Theory of Mind network* (bilateral temporoparietal junction, precuneus, and medial prefrontal cortex; Tamir & Thornton, 2018). Anticipation effects reach further into the future in higher-order, anterior brain regions compared to lower-order, posterior brain regions - suggesting that higher-order brain regions may be involved in *narrative* anticipation, rather than anticipation of lower-level stimulus features (Lee et al., 2021). Further, ToM regions use current mental state information to predict future social states (Thornton et al., 2019). Predictive responses during short narratives that involve ToM reasoning have also been repeatedly evidenced in the right temporoparietal junction (rTPJ: Koster-Hale & Saxe, 2013) - a core brain region within the predictive social brain (Geng & Vossel, 2013; Saxe & Wexler, 2005; Schuwerk et al., 2017). In sum, these findings highlight the role of higher order brain regions (e.g., the ToM network) in social cognitive prediction processes in neurotypical adults.

To our knowledge, this kind of research has not yet been extended to autistic adults. This would provide a unique opportunity to test the predictive coding theory. In neurotypical participants, prior research found that unexpected outcomes compared to expected outcomes elicited a stronger response in ToM regions, given prior information about an agent's behavior (Heil et al., 2019), with the magnitude of this effect inversely related to autistic-like traits across participants (Dungan et al., 2016). Moreover, existing research on neural differences in the ToM network in autism yields conflicting results. For instance, studies have observed reduced TPJ activation and weaker functional connectivity during social cognition tasks (i.e., intentional causal attributions; Kana et al., 2014), as well as similar brain activation in ToM tasks (Dufour et al., 2013; Moessnang et al., 2020) and during passive viewing of movie scenes known to evoke ToM reasoning when presented once (Mangnus et al., 2024). Assessing predictive activity of the mentalizing network while autistic participants try to make sense of other's interactions is a promising way to increase ecological validity (Sonkusare et al., 2019) and test the predictive coding theory.

In developmental neuroscientific research this naturalistic approach has been applied using a novel paradigm in a fMRI study (Richardson & Saxe, 2019). Three-to-7-year-old neurotypical children were shown the Disney Pixar's movie *Partly Cloudy* twice in a row, while undergoing fMRI. The study tested the hypothesis that predictive responses in ToM brain regions of children might manifest as temporally earlier responses during the second viewing of the movie (reflecting, e.g., less information required to form predictions and/or less violation of expectations during the second viewing of the movie). Results demonstrated that as children got older, they recruited the ToM network (including bilateral temporoparietal junction, precuneus, medial prefrontal cortex) earlier during the second viewing compared to the first - suggesting that these brain regions increasingly anticipated the narrative of the movie (*narrative anticipation effect*). There was no such effect in a control network of brain regions recruited for reasoning about bodily sensations (the *pain matrix*). This exact approach has not yet been applied to adults. Here, we aim to investigate whether differences in social interactions between autistic and non-autistic individuals may arise from less reliable neural predictions, thereby testing the predictive coding theory using Richardson and Saxe's (2019) approach.

Thus, the aim of the present study was two-fold. First, we attempted to find predictive coding processes by extending Richardson and Saxe's (2019) findings of a narrative anticipation effect to adults, applying their paradigm with the exact same fMRI stimuli and analysis procedures. Second, we tested open questions concerning predictive processing in the ToM network of autistic individuals. In Experiment 1, we expected to find a narrative anticipation effect in the ToM network but not in the pain matrix control network, indicating predictive coding processes in non-autistic adults (Richardson & Saxe, 2019). In Experiment 2, we planned to replicate Experiment 1 and compare the narrative anticipation effect in autistic and non-autistic adults, hypothesizing that the narrative anticipation effect would be attenuated (i.e., no/less anticipation of the narrative in the ToM network) in autistic compared to non-autistic adults (cf., Pellicano & Burr, 2012).

2. Methods

2.1. *Participants*

Experiment 1 included 61 non-autistic adults (30 women, 31 men, $M_{age} = 25.9$ years, $SD_{age} = 4.4$ years). An additional 3 non-autistic participants were tested but excluded due to technical issues during data collection ($n = 2$) and data preprocessing ($n = 1$).

Table 1 – Demographic characteristics of the autistic and non-autistic group of Experiment 2.

	autistic group (*N* = 30)		non-autistic group (*N* = 30)		*Cohen's d*
	M	*SD*	*M*	*SD*	
Age	32.4	10.7	33.2	10.1	−.07
Verbal IQ (MWT)	111.0	14.6	115.5	18.0	−.28
Non-verbal IQ (CFT 20 R)	111.4	13.3	116.7	8.8	−.46
Autistic traits (AQ)	39.7	7.2	19.4	18.4	1.18 ***
Autistic traits (BAPQ)	4.5	.7	2.6	.7	1.63 ***

M, means; SD, standard deviation; MWT, Multiple Choice Vocabulary Intelligence Test (German: Mehrfachwahl-Wortschatz-Test); CFT 20 R, Culture Fair Test; AQ, Autism Quotient; BAPQ, Broad Autism Phenotype Questionnaire; ***, $p < .001$.

Experiment 2 included 30 autistic adults (16 women, 12 men, 2 non-binary, $M_{age} = 32.4$ years, $SD_{age} = 10.7$ years) and 30 non-autistic adults (18 women, 12 men, $M_{age} = 33.2$ years, $SD_{age} = 10.1$ years). All autistic participants had been diagnosed by a qualified clinical psychologist or psychiatrist on average at the age of 27.2 years (SD = 12.22, range = 5–52 years; age of diagnosis was self-reported), with over 70% ($n = 22$) receiving their autism diagnosis in adulthood; the subtle presentation of this group should be considered when interpreting the generalizability to populations diagnosed earlier, particularly in childhood. The autism diagnosis was verified through medical documentation provided by the participants. Participants specified their autism diagnosis via self-report according to the International Classification of Diseases– 10th Revision (ICD-10) criteria: Asperger's syndrome ($n = 23$), high-functioning autism ($n = 2$), atypical autism ($n = 1$), multiple autism diagnoses ($n = 3$), and no specification provided ($n = 1$). Group assignment was further validated by two additional self-assessment measures of autistic traits [*Autism Quotient* (AQ; Baron-Cohen et al., 2001) & *Broad Autism Phenotype Questionnaire* (BAPQ; Hurley et al., 2007)], which confirmed significant differences between the autistic and non-autistic groups. The groups were matched based on gender, chronological age, and verbal and nonverbal intelligence (see Table 1). In the autistic group 30% ($n = 9$) had at least a university degree; this proportion was higher in the non-autistic group 70% ($n = 21$).

Over 75% ($n = 23$) of the autistic participants reported having at least one comorbidity. According to ICD-10 criteria, half ($n = 16$) of the autistic participants reported comorbid depression, one-third reported Anxiety Disorder ($n = 10$) and Attention Deficit Hyperactivity Disorder ($n = 10$), and one-fifth ($n = 6$) Post-Traumatic Stress Disorder. Over 80% ($n = 25$) have been or are currently in psychotherapy. To account for high comorbidity rates in autistic participants (e.g., Mannion & Leader, 2013), non-autistic participants with a psychiatric condition were included in the sample to reach a closely matched comparison group (Schwartz & Susser, 2011). Because we did not collect detailed information about comorbid conditions in the comparison group, we are unfortunately unable to report this here. However, this does not compromise our matching procedure with respect to comorbid conditions.

Autistic adults were recruited via local networks including clinics, practitioners and autism organizations. Non-autistic adults from the comparison group were consecutively recruited to match the individuals in the autism group via Ludwig-Maximilians-Universität (LMU) München's mailing list and postings on social media. All of the participants gave written informed consent prior to their participation and received payment for participating. The study was approved by the Ethics Committee of the Department of Psychology and Education of the Ludwig-Maximilians-Universität München.

2.2. *Procedure and measures*

2.2.1. *Self-assessment measures*

Before coming to our lab the participants in Experiment 2 were asked to complete a demographic questionnaire (asking about age, gender, autism diagnosis, comorbid psychiatric diagnoses, etc.) via an online survey tool. To support group assignment, we additionally assessed autistic traits via the self-assessment measures *Autism Quotient* (AQ; Baron-Cohen et al., 2001) and *Broad Autism Phenotype Questionnaire* (BAPQ; Hurley et al., 2007). In both questionnaires (AQ and BAPQ), higher scores indicate more autistic traits, ranging from 0 to 50 in the AQ (cut-off criterion: score ≥32), and from 1 to 6 in the BAPQ (cut-off criterion: score ≥3.15).

2.2.2. *Assessment of verbal intelligence*

Verbal intelligence was measured with the German version of the *Multiple Choice Vocabulary Intelligence Test* (German: Mehrfachwahl-Wortschatz-Test MWT; Lehrl, 2005) and nonverbal intelligence with the *Culture Fair Test* (German: Grundintelligenztest CFT 20-R; Weiβ, 2019) after the fMRI scan.

2.2.3. *fMRI data acquisition*

The study was conducted at the NeuroImaging Core Unit Munich at LMU Munich using a 3-T MRI scanner. Participants used the standard Siemens 32-channel head coil. T1-weighted structural images were collected in 208 interleaved sagittal slices with isotropic voxels of .80 mm (GRAPPA parallel imaging, acceleration factor of 2; standard coil: FOV: 256 mm). Functional data were collected using a gradient-echo EPI sequence sensitive to Blood Oxygen Level Dependent (BOLD) contrast with 48 interleaved near-axial slices, 3 mm isotropic voxels, and a 10% slice gap, aligned with the anterior/posterior commissure, covering the entire brain (EPI factor: 70; TR: 1s, TE: 30 msec, flip angle: 45°). A total of 360 volumes were acquired per run, with the two movie viewings being collected across two separate runs.

In order to minimize stress and sensory overload in autistic participants, we adjusted the experimental setup (e.g., reduced number of staff involved, reduced waiting time, exact communication of time per scan and highlighting the remaining examination time between the scans, etc.) and customized the communication according to recent person-centered recommendations (Stogiannos et al., 2022).

2.2.4. *fMRI paradigm*

Following prior research (Richardson & Saxe, 2019), participants viewed the 5.6-min Disney Pixar's animated short movie *Partly Cloudy*, twice, while undergoing fMRI. The movie is about two main characters: a lonely grey cloud named Gus, who constantly creates dangerous baby animals like crocodiles, hedgehogs, and electric eels, and his loyal partner Peck, a stork, who delivers these animals to their parents. While the other storks deliver only sweet baby animals like puppies and kittens, Peck's job becomes increasingly challenging. When Peck flies away from Gus instead of delivering a baby shark, Gus becomes furious, believing Peck has abandoned him. To Gus's relief, Peck eventually returns with protective football gear that he had obtained during his absence to enable him to continue to work with Gus.

The movie was presented silently. Participants were asked to stay still and pay attention to the movie.

2.2.5. *fMRI data analysis*

Preprocessing procedures were identical to those used in Richardson and Saxe (2019), implemented using the same software (SPM8) and analysis scripts (Matlab 2017a). 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 Anterior-Posterior Commissures and major sulci. All data were smoothed using a Gaussian filter (5 mm kernel) and underwent SPM's global image scaling.

The realignment parameters were used to identify artifact timepoints, using the Artifact Detection Tool (Whitfield-Gabrieli et al., 2011). Artifact timepoints were defined as timepoints where composite motion exceeded 2 mm, relative to the previous time point, and/or the global signal deviated more than 3 standard deviations from the average global signal. Runs would have been excluded if one-third or more of the acquired timepoints were identified as artifacts; this resulted in zero exclusions (for our pre-registered exclusion criteria see https://osf.io/cqnmf).

Participant motion was relatively low [number of artifact timepoints: Experiment 1: M(SD) = 9.6(11.1); Experiment 2, autistic participants: M(SD) = 11.9(9.7); Experiment 2, non-autistic participants: M(SD) = 8.7(5.3); mean translation across all timepoints: Experiment 1: M(SD) = .03(.01) mm; Experiment 2, autistic participants: M(SD) = .04(.02) mm; Experiment 2, non-autistic participants: M(SD) = .03(.01)mm]. Number of artifact timepoints positively correlated with mean translation in the autistic sample [Experiment 2, autistic participants: r(28) = .38, p = .039], and did not correlate in the two non-autistic samples [Experiment 1: r(59) = .01, p = .941; Experiment 2, non-autistic participants: r(28) = .08, p = .672]. In Experiment 2, the size of effect of group (autistic versus non-autistic) on motion revealed a small effect in the number of artifact timepoints [Cohen's d = .41, 95% CI (−.11, .93)], and medium effect in the mean translation [Cohen's d = .52, 95% CI (−.01, 1.04)]. We included the amount of motion (i.e., mean translation) as a covariate of no interest in regressions that involve neural measures. We additionally defined five regressors using the CompCor method (Behzadi et al., 2007) in eroded white matter masks. These regressors were defined on white matter signal after interpolating over timepoints previously identified as artifact timepoints, such that these regressors were maximally independent from the artifact timepoint regressors.

Primary analyses were run on timecourses extracted from group functional regions of interest (ROIs) encompassing the ToM network [bilateral temporoparietal junction (R/LTPJ), precuneus (PC), and dorso-, middle- and ventromedial prefrontal cortex (D/M/VMPFC)] and the pain matrix (bilateral medial frontal gyrus, secondary sensory motor cortex, insula, and dorsal anterior cingulate cortex). Group ROIs were previously defined in neurotypical adults (n = 20; see Richardson et al., 2018 for details) and used in Richardson and Saxe (2019). ROIs are publicly available for download (https://openneuro.org/datasets/ds000228).

From each group ROI, we extracted the preprocessed timecourse from each voxel, applied nearest-neighbor interpolation over artifact timepoints, and regressed out (1) motion spikes and (2) five CompCor regressors (see above for details on motion artifact). Residual timecourses were high-pass filtered (threshold: 1 cycle/100s). Timecourses across voxels within each ROI were averaged and artifact timepoints were excluded (NaNed). ROI timecourses were averaged per network (ToM network, pain matrix), such that there was one timecourse per network, movie-viewing, and participant.

We then calculated the correlation between each participant's timecourses during the first and second viewings for the ToM network and pain matrix in two temporal shifting schemes. Since the narrative anticipation effect was strongest at 2s time difference in Richardson and Saxe (2019), we used this time lag between the *anticipation* and *no shift* schemes. Thus, in the anticipation scheme, we calculated the correlation between timepoints 3–360 in the first viewing to timepoints 1–358 in the second viewing. In the no shift scheme, we calculated the correlation between timepoints 1–360 in the first and second viewings.

3. Statistical analyses

3.1. *Confirmatory analyses*

All statistical analyses were carried out in R (version 4.1.1, R Core Team, 2021). The anticipation effect was assessed by the *anticipation-no shift* correlation difference (CD; for graphs depicting correlation in anticipation and no shift schemes, see Fig. 1). All measures were normally distributed.

To confirm that the expected ToM regions were recruited during the movie across groups and viewings, we used a general linear model to analyse BOLD activity of each participant and run (separately) as a function of scene-type (for details see Supplementary Material).

3.1.1. *Experiment 1: Narrative anticipation in non-autistic adults (extension of Richardson & Saxe, 2019)*

In Experiment 1, we first tested for a narrative anticipation effect in non-autistic adults using a one-tailed one sample *t*-test, testing the anticipation-no shift CD in the ToM network of non-autistic adults.

Second, we tested whether the narrative anticipation effect was larger in the ToM network as compared to the pain matrix using a one-tailed paired *t*-test. As a follow-up analysis we tested for a narrative anticipation effect in the pain matrix using a two-tailed one sample *t*-test.

3.1.2. *Experiment 2: Narrative anticipation in autistic adults (replication of experiment 1 and extension of Richardson & Saxe, 2019)*

In Experiment 2, we first repeated Experiment 1 analyses with non-autistic and autistic samples. We expected to replicate Experiment 1 results with the non-autistic sample in Experiment 2, and planned to test whether the anticipation effect was larger in the non-autistic adults, as compared to the autistic adults. We conducted a linear mixed effects model using the lme4 package 1.1–33 (Bates et al., 2015) to test for an effect of group (autistic versus non-autistic) on the anticipation-no shift CD. We also included motion as a fixed effect, to account for potential effects of data quality [i.e., lm(CD-ToM ~ group + motion)]. Finally, we planned to test whether such a group difference in narrative anticipation effect was specific to the ToM network, by using another linear regression model to test for a network (ToM network versus pain matrix)-by-group (autistic versus non-autistic) interaction [i.e., lm(CD ~ network + group + network*group + motion)].

3.2. *Exploratory analyses*

To foreshadow our results: our planned analyses did not reveal a narrative anticipation effect in the ToM network of non-autistic adults in either Experiment 1 or 2. Given this, we conducted a series of exploratory analyses to ensure that we did not miss the predicted narrative anticipation effect. For the sake of clarity, we include the plan for these additional analyses here. First, to ensure that we did not miss predicted narrative anticipation effects due to adults having faster/more efficient predictive responses (relative to children; Richardson & Saxe, 2019), we ran the same analyses using a different temporal shift (i.e., 1s rather than 2s), which was afforded by the faster acquisition time of the fMRI sequence used in the present study. Second, following Richardson and Saxe (2019; Supplementary Figure 6), we explored narrative anticipation effects in all six regions of interest of the ToM network separately (DMPFC, LTPJ, MMPFC, PC, RTPJ, VMPFC), to ensure that the effect was absent in every region (rather than present in a subset of regions but obscured by the network average timecourse). Third, following Richardson and Saxe (2019; Supplementary Figure 8), we tested whether a narrative anticipation effect was present in neural response patterns – which can at times be more sensitive than univariate approaches. Fourth, we tested for an overall repetition suppression effect (i.e., a lower response magnitude in second viewing compared to first viewing, on average across all timepoints in the response timecourse, following Richardson and Saxe (2019); Supplementary Figure 9].

We then conducted two exploratory analyses to test for a local – i.e., content-specific – narrative anticipation or repetition suppression effect – in case predictive effects in ToM brain regions were specific to scenes in the movie that evoke ToM reasoning. First, we tested for a narrative anticipation effect (i.e., earlier response during second movie-viewing) specifically during all ToM scenes (using a concatenated timecourse), as defined in Richardson et al. (2018; for an overview of corresponding timepoints see Supplementary Fig. 8). Second, we used data-driven reverse correlation analyses (see Hasson et al., 2004; Richardson et al., 2018) to identify scenes (>4s) in a continuous naturalistic stimulus, in which there was a reliable difference across participants in response magnitude across viewings, per network and sample (i.e., Experiment 1 non-autistic, Experiment 2 non-autistic, and Experiment 2 autistic, separately). Reduced responses during the second viewing of the movie were defined as repetition suppression effects. After identifying scenes that evoked a different response (positive or negative) during the second viewing in non-autistic adults of Experiment 1, we tested for a replication in non-autistic adults in Experiment 2, and compared results with autistic adults in Experiment 2. Specifically, we ran a linear regression with Experiment 2 data to test for group, viewing, and group-by-viewing interaction effects on the response magnitude to a scene that reliably showed repetition suppression in non-autistic adults [i.e., lm(response.magnitude ~ group + viewing + group* viewing + motion)].

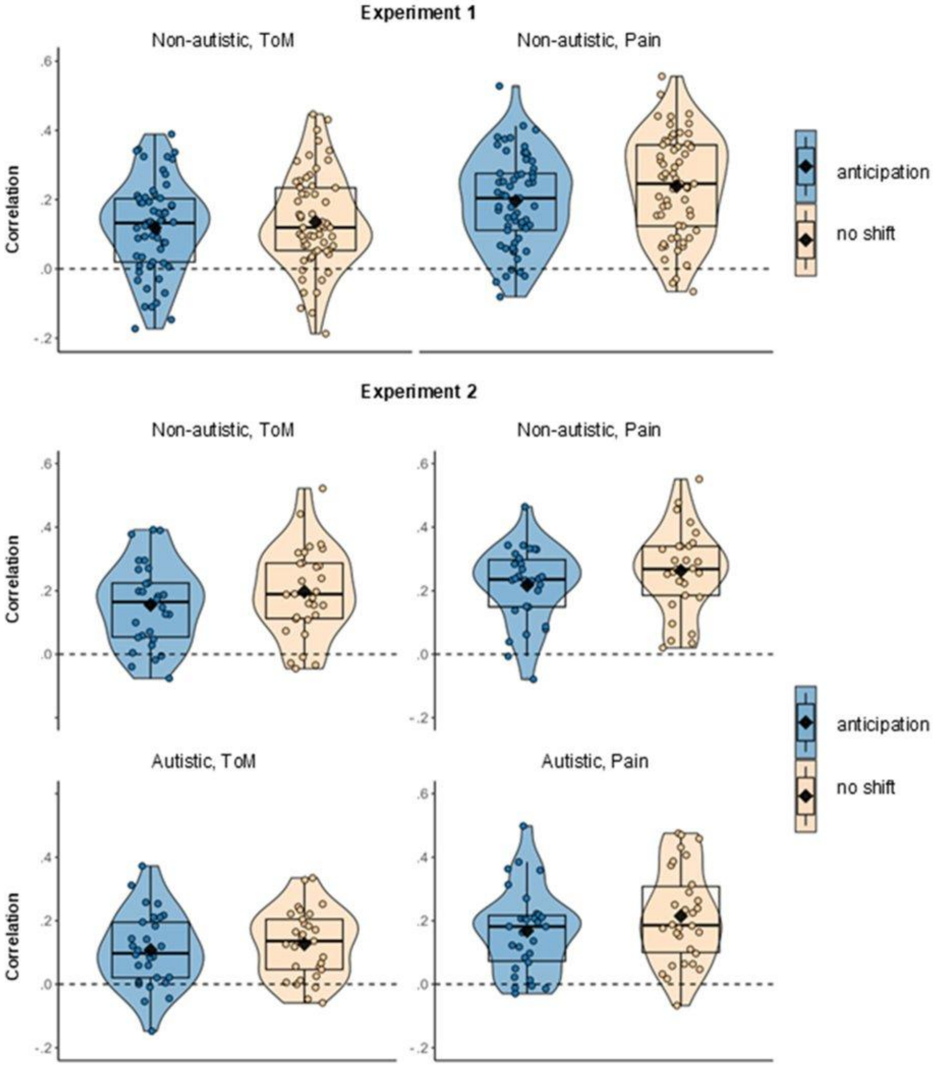


Fig. 1 – **Graph depicting the z-scored correlation (y-axis) in the anticipation (blue) and no shift (orange) correlation schemes for the ToM network (x-axis, left) and the pain matrix (x-axis, right) in Experiment 1 (upper plot) and in Experiment 2 (lower plot) for both groups (non-autistic: upper plot in Experiment 2 and autistic: lower plot in Experiment 2).**

4. Results

4.1. *Confirmatory analyses*

ToM movie scenes evoked responses in the ToM network in all groups (Experiment 1: non-autistic adults; Experiment 2: autistic and non-autistic adults) and in both movie viewings (see Supplementary Fig. 1).

4.1.1. *Experiment 1: Narrative anticipation in non-autistic adults (extension of Richardson & Saxe, 2019)*

In non-autistic adults, temporally misaligning the ToM time-courses from the first and second viewings into the anticipation scheme reduced the correlations between them [$M = -.02$, $SE = .01$; one-tailed one sample t-test (mu = 0): $t(60) = -2.03$, $p = .977$], indicating that the non-autistic participants did not show narrative anticipation in the ToM network during the second presentation of the movie. When comparing the anticipation effect in the ToM network to the pain matrix using a one-tailed paired t-test, the effect significantly differed between both networks, $t(60) = 2.50$, $p = .008$, such that the effect was smaller (more negative) in the pain matrix. In a follow-up analysis, a one sample two-way t-test revealed that the anticipation effect in the pain matrix ($M = -.04$, $SE = .01$) was significantly negative relative to 0, $t(60) = -5.25$, $p < .001$ (see Fig. 2; for graph depicting the average timecourses by network, and viewing, see Supplementary Fig. 2). This negative effect was predicted as any temporal shift in timecourses (in absence of a narrative anticipation effect) should lead to a reduced correlation between the timecourses.

4.1.2. *Experiment 2: Narrative anticipation in autistic adults (replication of experiment 1 and extension of Richardson & Saxe, 2019)*

First, we replicated the results above with our second non-autistic sample. In non-autistic participants, the anticipation effect in the ToM network ($M = -.04$, $SE = .01$) was again not significantly positive relative to 0, $t(29) = -3.02$, $p = .997$, indicating that adults did not show anticipation in the ToM network during the second presentation of the movie. Unlike in Experiment 1, when comparing the anticipation effect in the ToM network to the pain matrix using a one-tailed paired t-test, the effect did not significantly differ between the networks, $t(29) = .93$, $p = .180$. As in Experiment 1, a one sample two-way t-test revealed that the anticipation effect in the pain matrix ($M = -.06$, $SE = .01$) was significantly negative relative to 0, $t(29) = -3.93$, $p < .001$ (see Fig. 2; for graph depicting the average timecourses by network, and viewing, see Supplementary Fig. 2).

We observed the same pattern of results in autistic adults. Using a one-tailed one sample t-test, the anticipation effect in the ToM network ($M = -.02$, $SE = .01$) of the autistic participants was not significantly positive relative to 0, $t(29) = -1.65$, $p = .945$, indicating that the autistic participants did not show anticipation in the ToM network during the second presentation of the movie. When comparing the anticipation effect in the ToM network to the pain matrix using a one-tailed paired t-test, the effect significantly differed between both networks, $t(29) = 2.00$, $p = .027$, such that the effect was smaller (more negative) in the pain matrix. In a follow-up analysis, a one sample two-way t-test revealed that the anticipation effect in the pain matrix ($M = -.05$, $SE = .01$) was significantly negative relative to 0, $t(29) = -3.83$, $p < .001$ (see Fig. 2; for graph depicting the average timecourses by network, and viewing, see Supplementary Fig. 2).

To test for a group difference in the anticipation effect in the ToM network, we fit a linear mixed model and found no significant main effect of group ($ß = -.02$, $t = -1.28$, $p = .207$), indicating that the anticipation effect in the ToM network was not larger in the group of non-autistic participants compared to the autistic participants. Finally, we used a linear mixed effects model to simultaneously test for effects of brain network (ToM network versus pain matrix), group (autistic versus non-autistic), and the group-by-network interaction term on the correlation difference. We found no significant interaction effect of group-by-network ($ß = -.01$, $t = -.52$, $p = .610$), and also no significant main effects of network ($ß = .03$, $t = 1.64$, $p = .104$) or group ($ß = -.01$, $t = -.40$, $p = .688$).

4.2. *Exploratory analyses*

To ensure that predicted effects did not go unnoticed, we conducted a series of exploratory analyses, as detailed in the Methods and Supplementary Material. Briefly, the pattern of results remained unchanged with a 1s temporal shift (instead of 2s shift; see Supplementary Fig. 3). Second, the narrative anticipation effect was not present in any individual ToM

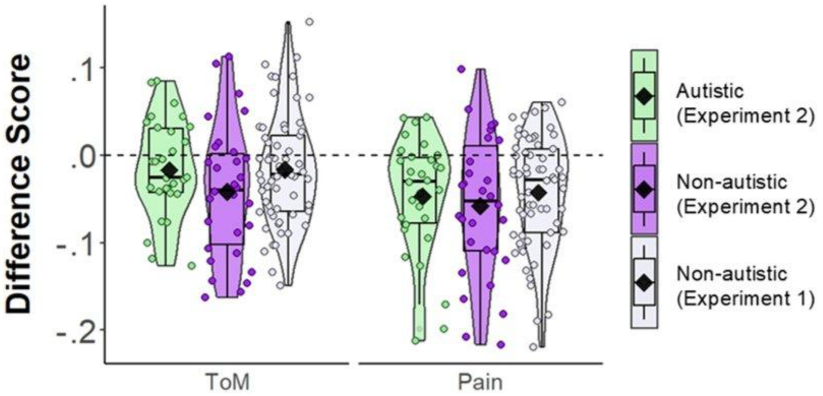


Fig. 2 – Graph depicting the difference score (y-axis) in the ToM network (x-axis, left) and pain matrix (x-axis, right) in Experiment 1 and 2 (green: autistic participants (Experiment 2); purple: non-autistic participants (Experiment 2); light purple: non-autistic participants (Experiment 1). A positive difference score would evidence a narrative anticipation effect.

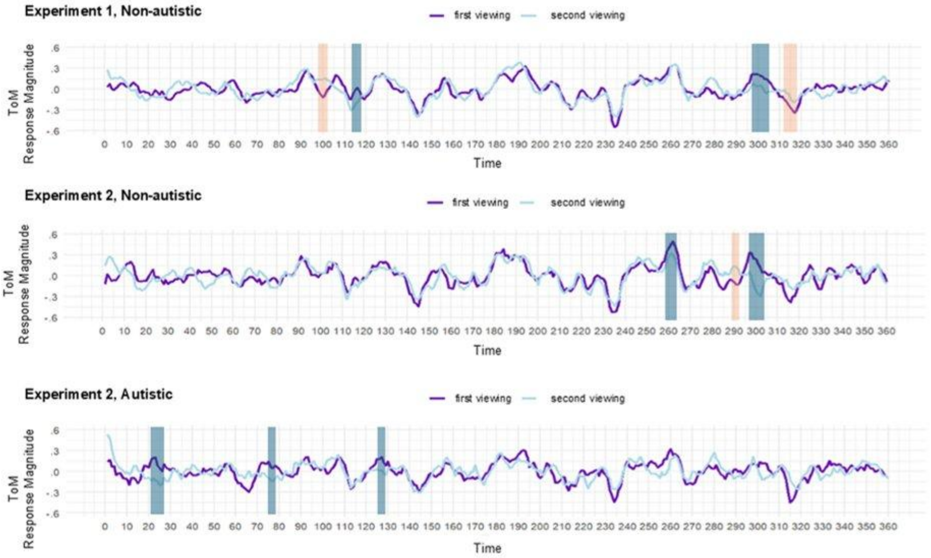


Fig. 3 – Average timecourses in the ToM network for each experiment (Experiment 1 versus Experiment 2), group (in Experiment 2: autistic versus non-autistic), and viewing (purple: 1st viewing, light blue: 2nd viewing) with response magnitude (y-axis) and timecourse (x-axis). Scenes with significant positive values marked in blue, scenes with significant negative values in orange.

network ROI (for a graph depicting the difference score in each ROI, see Supplementary Fig. 4, and for the average timecourses by ROI and viewing, see Supplementary Fig. 5). Third, we found no evidence for a narrative anticipation effect in the multivariate pattern of response (see Supplementary Fig. 6). Fourth, there was no positive evidence for repetition suppression across the full timecourse, in either experiment (see Supplementary Fig. 7). We also did not observe a narrative anticipation effect in the ToM network specifically during concatenated ToM scenes, as identified in Richardson et al. (2018; for details see Supplementary Material).

4.2.1. *Exploratory reverse correlation analyses in the ToM network*

Finally, we conducted data-driven reverse correlation analyses (Hasson et al., 2004) to discover scenes in which the response magnitude reliably differed across viewings. In non-autistic adults in Experiment 1, we identified two scenes that evoked smaller responses during the second viewing, relative to the first viewing (i.e., timepoints 98–102, and 297–305) and two scenes that evoked larger responses during the second viewing, relative to the first viewing (i.e., timepoints 113–117, and 312–318). We repeated analyses in non-autistic adults in Experiment 2 and replicated the reduced response to one scene during the second viewing (i.e., timepoints 297–304); we also identified a novel scene that evoked a smaller response during the second viewing (i.e., timepoints 259–264) and a novel scene that evoked a larger response during the second viewing (i.e., timepoints 289–292).

In autistic adults in Experiment 2, we identified three scenes that evoked smaller responses during the second viewing, only (i.e., timepoints 21–27, 75–78, and 125–128); autistic adults did not show reduced responses to the one scene we identified in both Experiments with non-autistic adults during the second viewing (see Fig. 3). The range of repetition suppression effects between autistic and non-autistic participants, illustrated by individual subject response difference values, was similar across groups (see Supplementary Fig. 9).

We conducted similar analyses in the pain matrix to understand the specificity of the apparent repetition suppression effects. In Experiment 1, we identified three scenes that evoked smaller responses during the second viewing, relative to the first viewing (i.e., timepoints 97–100, 233–236, and 274–277) and one scene that evoked larger responses during the second viewing, relative to the first viewing (i.e., timepoints 49–52). In Experiment 2, we did not identify any scene in autistic nor non-autistic samples that evoked differential response magnitude across viewings. Taken together, we did not find evidence for robust, replicable repetition suppression effects in the pain matrix.

Interestingly, the scene that evoked reduced responses in the ToM network during the second viewing across both experiments in non-autistic adults was previously identified as a scene that drives responses in ToM brain regions in (non-

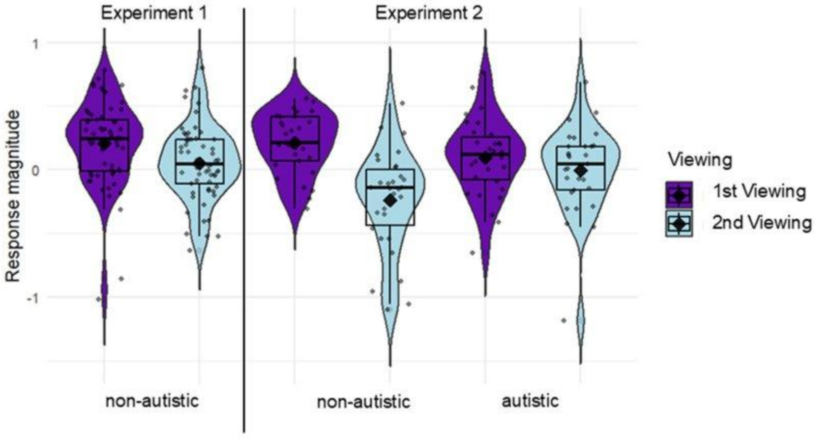


Fig. 4 – **Graph depicting the response magnitude to a key ToM scene (T04; *y*-axis) per group (Experiment 1: non-autistic participants; Experiment 2: non-autistic and autistic participants; *x*-axis) and viewing (purple: 1st; light blue: 2nd viewing) in the ToM network.**

autistic) adults in Richardson et al., 2018 (referred to as scene T04). In this scene, one character initially holds false beliefs about the other's intentions but feels relieved upon discovering the truth. Further, among 3–12-year-old children, ToM network response magnitude during this scene correlated with Theory of Mind behavior, controlling for age. Given these prior results, we tested whether the extent to which responses to T04 were reduced during the second viewing differed significantly between groups in Experiment 2. We extracted the response magnitude to the previously-defined peak timepoint of this scene (timepoint 150 in Richardson et al., 2018; which corresponded to timepoints 300 & 301 in this study) from each individual in each viewing. We found a significant main effect of group ($ß = .44$, $t = 2.31$, $p = .023$), indicating that there was, overall, a reduced response in the ToM network in autistic adults, relative to non-autistic adults, to this scene. There was also a significant group-by-viewing interaction effect ($ß = -.34$, $t = -2.85$, $p = .005$), such that non-autistic adults showed more repetition suppression (i.e., reduced response to T04 during the second viewing) relative to autistic adults (see Fig. 4).

5. Discussion

In this study, we examined predictive coding processes in brain regions associated with Theory of Mind reasoning in non-autistic and autistic adults. We investigated a narrative anticipation effect (i.e., ToM responses shifted earlier in time during the second movie viewing) that was previously described in children by Richardson and Saxe (2019) in three samples of adults (Experiment 1: non-autistic adults; Experiment 2: non-autistic and autistic adults) by applying the exact same paradigm. First, we aimed to find a narrative anticipation effect in non-autistic adults (Experiment 1) and attempted to replicate it in a new sample of non-autistic adults (Experiment 2). Second, we aimed to extend this effect to autistic adults (Experiment 2) to compare predictive coding processes in autistic and non-autistic adults. We expected to find differences between autistic and non-autistic individuals in this effect that could contribute to difficulties in social interaction in autism.

In contrast to our expectations, in our confirmatory analyses, we did not find evidence for narrative anticipation in the ToM network in either non-autistic or autistic adults. Further, we did not find a significant difference in narrative anticipation between non-autistic and autistic adults. Rather, our study showed that the neural responses in the ToM network during a short, naturalistic movie-viewing experiment are highly similar between autistic and non-autistic adults. In all adult samples, response timecourses across viewings were more correlated when temporally aligned than when the second timecourse was shifted earlier in time. In two samples (non-autistic adults in Experiment 1 and autistic adults in Experiment 2), we observed higher correlations in the ToM network, relative to the pain matrix, under the narrative anticipation time scheme - which is consistent with Richardson and Saxe (2019) - but in both cases these higher correlations were still lower than the aligned timecourse correlations. Narrative anticipation effects were also not observed at a faster timescale, in individual ToM regions, or in the multivariate response patterns. These findings of our main analysis may indicate either that a narrative anticipation effect exists, but our task is not sensitive enough to capture it in adults – meaning we cannot draw any conclusions about a group difference (or its absence) – or that adults do not show a narrative anticipation effect when watching this movie. Consequently, our findings show that autistic adults do not differ from non-autistic adults. This would suggest that the core interaction difficulties observed in autism may not lie in an attenuation of prediction processes or may not be captured throughout an anticipation of an entire narrative. Instead, these social difficulties may be measurable in specific situations and/or located in other areas or processes. Moreover, these similar neural responses may suggest that autistic individuals understand narratives similarly to neurotypical individuals but face difficulties more at a level of execution, which is consistent with literature finding no measurable differences between autistic and non-autistic adults in ToM network activation during mentalizing (Dufour et al., 2013; Mangnus et al., 2024; Moessnang et al., 2020). From a theoretical perspective, our confirmatory results would speak against the theoretical explanation of a circumscribed and profound Theory of Mind deficit causing interaction and communication problems in autistic adults (Gernsbacher & Yergeau, 2019).

In exploratory analyses we used data-driven reverse correlation analyses to identify scenes that evoked reliable response differences between the first and second viewing. In both non-autistic samples, we identified a crucial scene at the end of the movie that evoked smaller responses in the ToM network, but not in the pain matrix, during the second viewing, relative to the first viewing. In autistic adults, this key scene evoked a smaller response overall and did not evoke a reduced response during the second viewing (i.e., there was no repetition suppression effect across viewings of this scene); the magnitude of repetition suppression to this scene differed significantly across autistic and non-autistic groups. We did not find global repetition suppression effects - across the whole timecourse - in any sample or networks. This suggests that the observed repetition suppression effect in the ToM network appears to be specific to this key scene.

In comparison to other social scenes in the movie, that also show social interactions involving mental states and emotions, this key scene specifically evokes a more complex reasoning about the false beliefs of the characters: it shows Gus, the grey cloud, revising his beliefs about the intention of his partner Peck, the stork. Because of Peck's absence, Gus became furious, believing that Peck had abandoned him after constantly creating dangerous creatures for Peck to deliver. When Peck returned with protective gear, Gus felt relieved and happy upon realizing Peck's true intentions. In a previous study this same scene drove responses in ToM brain regions in (non-autistic) adults and response magnitude during this scene correlated with ToM behavior in 3- to 12-year-old children, controlling for age (Richardson et al., 2018). Our results might indicate that when complex ToM reasoning is required, different neural processing within the ToM network becomes evident in autistic adults, which is in line with recent findings showing that differences between autistic and non-autistic individuals emerge only when ToM processes become demanding (Schuwerk & Sodian, 2023). Potentially, non-autistic adults benefit more from the initial viewing of this ToM scene, reflected in a more efficient processing (i.e., less response magnitude) during the repeated viewing of the ToM scene. In contrast, autistic adults did not show differences in processing this ToM scene (i.e., the response magnitude was similar) between the first and second viewings, which might indicate that processing demands remained consistent regardless of repetition, suggesting a potential difference in adaptive strategies when processing complex social information. Although this key scene involves a clear false belief and empirical evidence underpins the association between ToM network responses to this scene and ToM reasoning, its emotional content may also contribute to the observed group difference. Future studies could explore the role of empathetic reasoning in processing this scene.

In sum, our confirmatory analyses leave open whether predictive processing is at work in non-autistic and autistic adults when processing social scenes. Additionally, we find no evidence for differential processing in autism. Using a data-driven reverse correlation approach, we identified one scene that evoked differential predictive processes between autistic and non-autistic adults - in line with previous literature. This scene evoked complex ToM reasoning. Thus, we only find exploratory evidence which is not supported by the main analysis, and may either indicate a subtle difference or no difference. But, it is up to future research to confirm this exploratory finding and to better understand this effect.

5.1. *Limitations*

In 3–7-year-old children, the presence of a narrative anticipation effect increased with age (Richardson & Saxe, 2019), which led us to predict that this effect would be evident in neurotypical adults. Speculatively, it is possible that narrative anticipation effects are more present/observable in age-appropriate movie stimuli. That is, movie stimuli that evoke complicated reasoning may be more likely to be processed differently across viewings. As a result, children might benefit more from seeing the complete movie, leading to larger temporal shifts in their responses between the viewings. This is in line with prior evidence for narrative anticipation effects among adults, which tend to use longer movies (Baldassano et al., 2017) designed for adult audiences (Baldassano et al., 2017; Lee et al., 2021). Future research is needed to clarify the extent to which predictive processes differ/depend on stimulus complexity - and how this varies by age and population.

As our study focused on Theory of Mind reasoning from a third-person perspective, our findings cannot be readily generalized to all forms of Theory of Mind reasoning. Early neuroscientific studies on Theory of Mind were limited in ecological validity and explanatory power, as they typically involved processing abstract stimuli from a third-person perspective. In response to this, second-person approaches emerged, focusing on social cognition during real-time interaction—including hyperscanning paradigms in which brain activity from two interaction partners is simultaneously recorded (Misaki et al., 2021; Redcay & Schilbach, 2019). Future neuroscientific studies addressing social interactions that require flexible attunement between partners (Bolis et al., 2023) may be important for understanding the differences in mechanisms underlying differences in social interactions in autistic individuals. This approach could be further extended to the idea that social difficulties are mutual and occur on both sides of social interactions (see Milton, 2012 for double empathy problem), offering a promising direction for future neurocognitive research.

However, in everyday life, humans engage in both types of reasoning: Theory of Mind from a second-person perspective during interactions, and more *offline* Theory of Mind from a third-person perspective—for instance, when observing others or reflecting on past encounters. In fact, a prior experience sampling study found that participants thought more about others' mental states when they were alone, and during interaction, their thoughts were more focused on others' actions rather than mental states (Schuwerk et al., 2019). To investigate this type of offline Theory of Mind reasoning while addressing the limitations of earlier paradigms, researchers have increasingly turned to naturalistic stimuli such as movies (Sonkusare et al., 2019). This is the approach we followed in our study.

While we measured activity in the Theory of Mind network in response to naturalistic interactions depicted on screen—thus increasing ecological validity—we acknowledge that the use of animated stimuli and a fictional narrative remains an artificial scenario that cannot be equated with real–world interactions. Nonetheless, we chose this format because the exaggerated emotional expressions and dense narrative structure were expected to enhance Theory of Mind network activation. We see our study as a step forward in this direction, with future research needed to bridge the remaining gap toward more realistic, ecologically valid settings.

5.2. *Implications*

This study suggests several directions for future research. First, conducting fMRI studies in adults, including non-autistic and autistic samples, using this approach but with stimuli more suited for adults (e.g., the stimuli used in Baldassano et al., 2017 and Lee et al., 2021) could be a next step in testing the predictions of the predictive coding theory of autism. Second, exploring how perceived stimulus complexity and narrative comprehension interact with predictive responses and narrative anticipation is crucial. This research could reveal how these factors influence cognitive processing in both autistic and non-autistic adults. Third, future studies should aim to scale up investigations of scenes that evoke reasoning processes observed in our key ToM scene and systematically examine any potential differences between autistic and non-autistic adults. Finally, to gain further insights into the development of predictive coding processes, it would be beneficial to try to extend Richardson and Saxe's (2019) findings in 3- to 7-year-old children with adolescents (using a movie that is more suitable for this age group). Such an extension could help validate the presence of an anticipation effect in children and clarify the development of this effect across the life span.

6. Conclusion

Confirmatory analyses did not provide evidence for a narrative anticipation effect, as previously defined by Richardson and Saxe (2019), in adults, nor differences in this effect between neurotypical and autistic adults. Both groups showed comparable neural responses within the ToM network during a short, naturalistic movie-viewing experiment. However, exploratory, data-driven analyses revealed a difference in repetition suppression to one particular ToM scene between non-autistic and autistic adults - providing preliminary neural evidence for differences in predictive coding between autistic and non-autistic adults. Yet, this finding should not be readily generalized without further cross-validation in a follow-up study.

CRediT authorship contribution statement

Lucie Zimmer: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Investigation, Formal analysis, Data curation, Conceptualization. **Hilary Richardson:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Formal analysis, Conceptualization. **Carolina Pletti:** Writing – review & editing, Project administration, Investigation. **Markus Paulus:** Writing – review & editing, Resources. **Tobias Schuwerk:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Investigation, Conceptualization.

Funding

The study was done using the NICUM Siemens Prisma scanner supported by German Research Foundation (DFG; Deutsche Forschungsgemeinschaft) [grant number INST 86/1739-1 FUGG].

Declaration of competing interest

The authors declare that they have no competing interests.

Acknowledgements

The authors would like to thank all participants for their contribution to this work, and the student assistants Sarina Drexler, Paula Kirchner, An Thuy Nguyen and Cui Ci Voon for their support in data acquisition. Thanks to Adrian Steffan for technical support.

Scientific transparency statement

DATA: Some raw and processed data supporting this research are publicly available, while some are subject to restrictions: https://osf.io/2uckn/.

CODE: All analysis code supporting this research is publicly available: https://osf.io/2uckn/.

MATERIALS: Some study materials supporting this research are publicly available, while some are subject to restrictions: https://saxelab.mit.edu/theory-mind-and-pain-matrix-localizer-movie-viewing-experiment/, https://osf.io/2uckn/, References for AQ (Baron-Cohen et al., 2001) and BAPQ (Hurley et al., 2007) are contained in the manuscript or supplemental files. Please contact Hogrefe for CFT 20-R and

MWT, and The Walt Disney Company for the Pixar movie *Partly Cloudy*.

DESIGN: This article reports, for all studies, how the author(s) determined all sample sizes, all data exclusions, all data inclusion and exclusion criteria, and whether inclusion and exclusion criteria were established prior to data analysis.

PRE-REGISTRATION: At least part of the study procedures was pre-registered in a time-stamped, institutional registry prior to the research being conducted: https://osf.io/cqnmf At least part of the analysis plans was pre-registered in a time-stamped, institutional registry prior to the research being conducted: https://osf.io/cqnmf The analyses that were undertaken deviated from the preregistered analysis plans. All such deviations are fully disclosed in the manuscript.

For full details, see the *Scientific Transparency Report* in the supplementary data to the online version of this article.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cortex.2025.04.006.

REFERENCES

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5 (5)*. American Psychiatric Publishing.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron, 95*(3), 709–721. https://doi.org/10.1016/j.neuron.2017.06.041

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders, 31*(1), 5–17. https://doi.org/10.1023/A:1005653411471

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage, 37*(1), 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042

Bolis, D., Balsters, J., Wenderoth, N., Becchio, C., & Schilbach, L. (2017). Beyond autism: Introducing the dialectical misattunement hypothesis and a bayesian account of intersubjectivity. *Psychopathology,* 50(6), 355–372. https://doi.org/10.1159/000484353

Bolis, D., Dumas, G., & Schilbach, L. (2023). Interpersonal attunement in social interactions: From collective psychophysiology to inter-personalized psychiatry and beyond. *Philosophical Transactions of the Royal Society B, 378*, Article 20210365. https://doi.org/10.1098/rstb.2021.0365

Cattaneo, L., Fabbri-Destro, M., Boria, S., Pieraccini, C., Monti, A., Cossu, G., & Rizzolatti, G. (2007). Impairment of actions chains in autism and its possible role in intention understanding. *Proceedings of the National Academy of Sciences,* 104(45), 17825–17830. https://doi.org/10.1073/pnas.0706273104

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences, 36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Dufour, N., Redcay, E., Young, L., Mavros, P. L., Moran, J. M., Triantafyllou, C., Gabrieli, J. D., & Saxe, R. (2013). Similar brain activation during false belief tasks in a large sample of adults with and without autism. *Plos One, 8*(9), Article e75468. https://doi.org/10.1371/journal.pone.0075468

Dungan, J. A., Stepanovic, M., & Young, L. (2016). Theory of mind for processing unexpected events across contexts. *Social Cognitive and Affective Neuroscience, 11*(8), 1183–1192. https://doi.org/10.1093/scan/nsw032

Falck-Ytter, T. (2010). Young children with autism spectrum disorder use predictive eye movements in action observation. *Biology Letters, 6*(3), 375–378. https://doi.org/10.1098/rsbl.2009.0897

Frith, U. (2012). The 38th Sir Frederick Bartlett lecture why we need cognitive explanations of autism. *Quarterly Journal of Experimental Psychology, 65*(11), 2073–2092. https://doi.org/10.1080/17470218.2012.697178

Ganglmayer, K., Schuwerk, T., Sodian, B., & Paulus, M. (2020). Do children and adults with autism spectrum condition anticipate others' actions as goal-directed? A predictive coding perspective. *Journal of Autism and Developmental Disorders, 50*, 2077–2089. https://doi.org/10.1007/s10803-019-03964-8

Geng, J. J., & Vossel, S. (2013). Re-evaluating the role of TPJ in attentional control: Contextual updating? *Neuroscience and Biobehavioral Reviews, 37*(10), 2608–2620. https://doi.org/10.1016/j.neubiorev.2013.08.010

Gernsbacher, M. A., & Yergeau, M. (2019). Empirical failures of the claim that autistic people lack a theory of mind. *Archives of Scientific Psychology, 7*(1), 102–118. https://doi.org/10.1037/arc0000067

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science, 303*(5664), 1634–1640. https://doi.org/10.1126/science.1089506

Heil, L., Colizoli, O., Hartstra, E., Kwisthout, J., van Pelt, S., van Rooij, I., & Bekkering, H. (2019). Processing of prediction errors in mentalizing areas. *Journal of Cognitive Neuroscience, 31*(6), 900–912. https://doi.org/10.1162/jocn_a_01381

Hurley, R. S. E., Losh, M., Parlier, M., Reznick, J. S., & Piven, J. (2007). The broad autism phenotype questionnaire. *Journal of Autism and Developmental Disorders, 37*(9), 1679–1690. https://doi.org/10.1007/s10803-006-0299-3

Kana, R. K., Libero, L. E., Hu, C. P., Deshpande, H. D., & Colburn, J. S. (2014). Functional brain networks and white matter underlying theory-of-mind in autism. *Social Cognitive and Affective Neuroscience, 9*(1), 98–105. https://doi.org/10.1093/scan/nss106

Kennedy, D. P., & Adolphs, R. (2012). Perception of emotions from facial expressions in high-functioning adults with autism. *Neuropsychologia,* 50(14), 3313–3319. https://doi.org/10.1016/j.neuropsychologia.2012.09.038

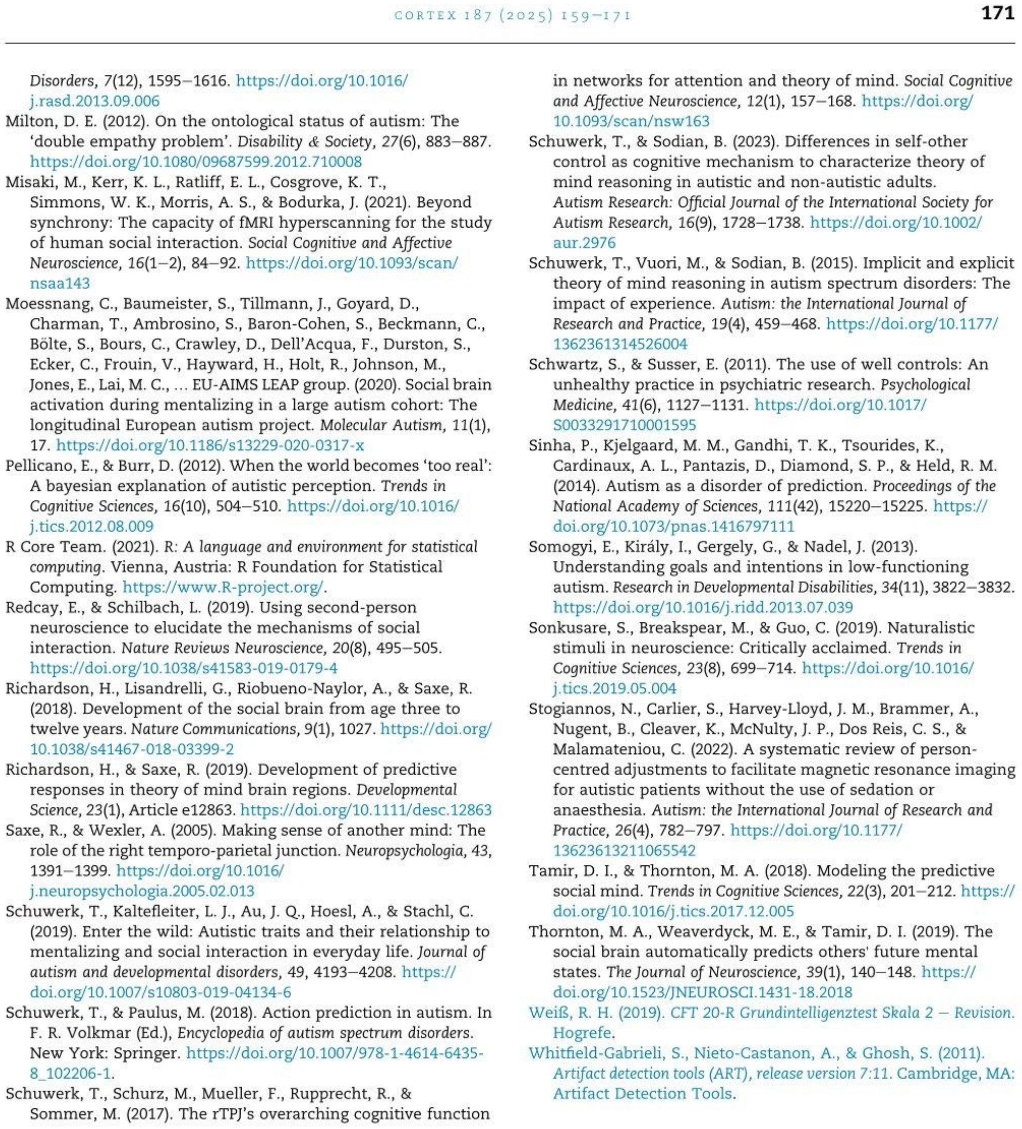
Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron, 79*(5), 836–848. https://doi.org/10.1016/j.neuron.2013.08.020

Lee, A., Aly, M., & Baldassano, C. (2021). Anticipation of temporally structured events in the brain. *eLife,* 10, Article e64972. https://doi.org/10.7554/eLife.64972

Lehrl, S. (2005). *Mehrfachwahl-Wortschatz-Intelligenztest MWT-B*. Spitta.

Mangnus, M., Koch, S. B. J., Cai, K., Greidanus Romaneli, M., Hagoort, P., Bašnáková, J., & Stolk, A. (2024). Preserved spontaneous mentalizing amid reduced intersubject variability in autism during a movie narrative. *bioRxiv*. https://doi.org/10.1101/2024.03.08.583911

Mannion, A., & Leader, G. (2013). Comorbidity in autism spectrum disorder: A literature review. *Research in Autism Spectrum*



Disorders, 7(12), 1595–1616. https://doi.org/10.1016/j.rasd.2013.09.006

Milton, D. E. (2012). On the ontological status of autism: The 'double empathy problem'. *Disability & Society, 27*(6), 883–887. https://doi.org/10.1080/09687599.2012.710008

Misaki, M., Kerr, K. L., Ratliff, E. L., Cosgrove, K. T., Simmons, W. K., Morris, A. S., & Bodurka, J. (2021). Beyond synchrony: The capacity of fMRI hyperscanning for the study of human social interaction. *Social Cognitive and Affective Neuroscience, 16*(1–2), 84–92. https://doi.org/10.1093/scan/nsaa143

Moessnang, C., Baumeister, S., Tillmann, J., Goyard, D., Charman, T., Ambrosino, S., Baron-Cohen, S., Beckmann, C., Bölte, S., Bours, C., Crawley, D., Dell'Acqua, F., Durston, S., Ecker, C., Frouin, V., Hayward, H., Holt, R., Johnson, M., Jones, E., Lai, M. C., … EU-AIMS LEAP group. (2020). Social brain activation during mentalizing in a large autism cohort: The longitudinal European autism project. *Molecular Autism, 11*(1), 17. https://doi.org/10.1186/s13229-020-0317-x

Pellicano, E., & Burr, D. (2012). When the world becomes 'too real': A bayesian explanation of autistic perception. *Trends in Cognitive Sciences, 16*(10), 504–510. https://doi.org/10.1016/j.tics.2012.08.009

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Redcay, E., & Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nature Reviews Neuroscience, 20*(8), 495–505. https://doi.org/10.1038/s41583-019-0179-4

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. https://doi.org/10.1038/s41467-018-03399-2

Richardson, H., & Saxe, R. (2019). Development of predictive responses in theory of mind brain regions. *Developmental Science, 23*(1), Article e12863. https://doi.org/10.1111/desc.12863

Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia, 43*, 1391–1399. https://doi.org/10.1016/j.neuropsychologia.2005.02.013

Schuwerk, T., Kaltefleiter, L. J., Au, J. Q., Hoesl, A., & Stachl, C. (2019). Enter the wild: Autistic traits and their relationship to mentalizing and social interaction in everyday life. *Journal of autism and developmental disorders, 49*, 4193–4208. https://doi.org/10.1007/s10803-019-04134-6

Schuwerk, T., & Paulus, M. (2018). Action prediction in autism. In F. R. Volkmar (Ed.), *Encyclopedia of autism spectrum disorders*. New York: Springer. https://doi.org/10.1007/978-1-4614-6435-8_102206-1.

Schuwerk, T., Schurz, M., Mueller, F., Rupprecht, R., & Sommer, M. (2017). The rTPJ's overarching cognitive function in networks for attention and theory of mind. *Social Cognitive and Affective Neuroscience, 12*(1), 157–168. https://doi.org/10.1093/scan/nsw163

Schuwerk, T., & Sodian, B. (2023). Differences in self-other control as cognitive mechanism to characterize theory of mind reasoning in autistic and non-autistic adults. *Autism Research: Official Journal of the International Society for Autism Research, 16*(9), 1728–1738. https://doi.org/10.1002/aur.2976

Schuwerk, T., Vuori, M., & Sodian, B. (2015). Implicit and explicit theory of mind reasoning in autism spectrum disorders: The impact of experience. *Autism: the International Journal of Research and Practice, 19*(4), 459–468. https://doi.org/10.1177/1362361314526004

Schwartz, S., & Susser, E. (2011). The use of well controls: An unhealthy practice in psychiatric research. *Psychological Medicine, 41*(6), 1127–1131. https://doi.org/10.1017/S0033291710001595

Sinha, P., Kjelgaard, M. M., Gandhi, T. K., Tsourides, K., Cardinaux, A. L., Pantazis, D., Diamond, S. P., & Held, R. M. (2014). Autism as a disorder of prediction. *Proceedings of the National Academy of Sciences, 111*(42), 15220–15225. https://doi.org/10.1073/pnas.1416797111

Somogyi, E., Király, I., Gergely, G., & Nadel, J. (2013). Understanding goals and intentions in low-functioning autism. *Research in Developmental Disabilities, 34*(11), 3822–3832. https://doi.org/10.1016/j.ridd.2013.07.039

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic stimuli in neuroscience: Critically acclaimed. *Trends in Cognitive Sciences, 23*(8), 699–714. https://doi.org/10.1016/j.tics.2019.05.004

Stogiannos, N., Carlier, S., Harvey-Lloyd, J. M., Brammer, A., Nugent, B., Cleaver, K., McNulty, J. P., Dos Reis, C. S., & Malamateniou, C. (2022). A systematic review of person-centred adjustments to facilitate magnetic resonance imaging for autistic patients without the use of sedation or anaesthesia. *Autism: the International Journal of Research and Practice, 26*(4), 782–797. https://doi.org/10.1177/13623613211065542

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences, 22*(3), 201–212. https://doi.org/10.1016/j.tics.2017.12.005

Thornton, M. A., Weaverdyck, M. E., & Tamir, D. I. (2019). The social brain automatically predicts others' future mental states. *The Journal of Neuroscience, 39*(1), 140–148. https://doi.org/10.1523/JNEUROSCI.1431-18.2018

Weiß, R. H. (2019). *CFT 20-R Grundintelligenztest Skala 2 – Revision*. Hogrefe.

Whitfield-Gabrieli, S., Nieto-Castanon, A., & Ghosh, S. (2011). *Artifact detection tools (ART), release version 7:11*. Cambridge, MA: Artifact Detection Tools.

Supporting information: **https://ars.els-cdn.com/content/image/1-s2.0-S0010945225001042-mmc2.docx**