FINAL REPORT
for the research project ’Perspectives: new vistas in the research of theory of mind’
OTKA 116 779

Ildikó Király
Principle Investigator

Objectives:

The main goal of the project was to unite the classic theories on theory of mind (ToM) and the current empirical findings in a new theoretical framework by resolving the conflicts between them. According to the classic, and re-emerging dogmatic approach the explicit form of theory of mind emerges at the age of 4 years (see Rakoczy, 2012). On the other hand, recent convergent evidence suggests, that even infants (already in their first year of life) can attribute beliefs to others if measured by implicit tasks (Scott and Baillargeon, 2017). However, both the replicability, and the competence underlying the performance of such tasks are questioned (Kulke and Rakoczy, 2018). Relatedly, in light of adult studies only certain contents (level-1 perspectives) can be attributed to others in an automatic, implicit way (Surtees et al, 2012). Based on these results, some researchers theorize that there is a dual system of Theory of Mind (Butterfill and Apperly, 2013), including a quick implicit system that is limited in the contents it can process (this is what operates already in infants but remains available also during adulthood), and also a later developing, explicit, flexible, full-blown system, that can attribute so called “genuine beliefs” as well. In dispute with this proposal, we posit that there is only a unified system of ToM: 1) even the suggested ‘implicit, automatic’ system can attribute *proper, genuine beliefs*, and 2) the salient developmental shift occurring at 4 years does not reflect the emergence of theory of mind itself, rather the integration of other higher order capacities, like episodic memory. Episodic memory is responsible for the enabling of a conscious memory search for information that could define (or refine) the content of the belief. In the current project we worked on these assumptions.

First, we developed a detailed theoretical account on our perspective, namely, that the spontaneous operation of Theory of Mind – which is actually the primary mode of operation of this system - enables quick and real-time behaviour-predictions. (I) This operation mode is supported by an intuitive monitoring interface, a naive model of attention (Elekes and Király, 2020). (II) Furthermore, we argue that the explicit, post-hoc functioning of ToM require the contribution of other higher order cognitive processes, and specifically propose that ToM recruits episodic memory for flexible attributions of beliefs and behaviour (Kampis, Keszei and Király, 2018).

Second, we performed two sets of studies in order to provide supportive empirical evidence for the predictions delineated from the theory formulated.

(I) In the first line of investigations, we searched for evidence for the spontaneous, fast but at the same time full functioning of Theory of Mind. We ran studies with adults and older children that grasped spontaneous level-2 perspective taking. In contrast with previous null-results (REF), we found that spontaneous level-2 perspective taking can be triggered, given that the task and therefore the focus of the other person's attention (whom perspectives is being attributed) is relevant to the participant.

(II) The second pillar of the model we outlined is that the well-known shift in the theory of mind capacity that happens around the age of 4 (success on the Sally-Anne type of false belief tasks) is not generated by a qualitative growth of the theory of mind capacity itself, as on implicit tasks the system operates well before that age. Rather, it is caused by the unfolding of those processes – episodic memory amongst others - that enable the updating of the content of the attributes mental states, and the integration of these processes and theory of mind. This proposal has been addressed by the second line of research, in which children could only give a correct behavioral response in a location change paradigm if they could use episodic recall.

In the following section, the most relevant claims of theory building (Part A) and the supportive empirical evidence (Part B) will be summarized based on the papers published and prepared within the very project.

Part A.

I.

1. Elekes F, Király I. (2021): **Attention in naïve psychology**. *Cognition 206: 104480*
   https://doi.org/10.1016/j.cognition.2020.104480

In everyday life, mentalizing is nested in a rich context of cognitive faculties and background information that potentially contribute to its success. Yet, we know little about these modulating effects. Here we propose that humans develop a naïve psychological model of attention (featured as a goal-dependent, intentional relation to the environment) and use this to fine-tune their mentalizing attempts, presuming that the way people represent their environment is influenced by the cognitive priorities (attention) their current intentions create. The attention model provides an opportunity to tailor mental state inferences to the temporary features of the agent whose mind is in the focus of mentalizing. The ability to trace attention is an exceptionally powerful aid for mindreading. Knowledge about the partner's attention provides background information, however being grounded in his current intentions, attention has direct relevance to the ongoing interaction. Furthermore, due to its causal connection to intentions, the output of the attention model remains valid for a prolonged but predictable amount of time: till the evoking intention is in place. The naïve attention model theory is offered as a novel theory on social attention that both incorporates existing evidence and identifies new directions in research.

We call this theory the naïve attention model theory (AMT). We argue that keeping track of fellow individuals' selective attention provides humans with an opportunity to form more accurate hypotheses about a protagonist's mental representations, while also enabling a relatively effortless computation by presuming long-lasting predispositions in the protagonist to represent certain kinds of information. We build this assumption on that *attention* shapes first person processing, and thus we underline its utmost relevance from the mentalizer's point of view. We review how the cognitive capacity constraints lead to the selective nature of perception, and thus create a similar need for selectivity in mentalizing and point out that tracing others' attentional processes may answer that challenge.

Although theory of mind and social attention have received extensive interest recently, the account we describe offers new insights for both fields by pulling the bond tighter between them. In our attempt to ascertain how humans get to understand their fellow agents' mental states we advocate the notion that Theory of Mind is embedded in (and interacts with) a broader set of cognitive faculties and background knowledge, which contributes to the success of social interactions in a wide range of contexts. We propose that humans have a naïve psychological theory about attention as an intentional relation towards the environment. This model enables them to form predictions on how specific individuals' current cognitive priorities shape their emerging mental representations. It is argued that this includes an integrated understanding of the causal factor that elicits the formation of an attentional set (intentions) as well as behavioral indicators to attention such as gaze.

We have put emphasis on the selective, subjective nature of human perception which notion (as obvious as it may seem) is largely ignored in social cognition and in a sense, even in the narrower field of social attention. We point out that perception, in part, depends on the temporary features of the agent's mind and that this characteristic of perception can be utilized for social purposes. Amongst the different forms of background knowledge that Mentalizer may consider about Agent, attention avails an exceptionally powerful tool for mental state inferences. By representing the agent's attentional set, humans take a step beyond the immediate features of the physical environment, however, the kind of background information they rely on in their mentalization attempt remains to be causally related to the ongoing event. That is, the capacity to trace attention enables ToM to fine-tune its predictions based on the *temporary invariances* of the agent's cognition while also retaining the dynamicity of its predictions.

Nevertheless, there is a lot of hard-to-predict variation in perception and even the mature attention model may fail to grasp that in its full complexity. The characteristic of the human cognitive system to favor attended information only sets priorities or weights in cognitive processing without necessarily cutting off non-attended streams completely. This balance is further modulated by the cognitive demands of the goal that gives rise to A's attentional set in combination with the level of A's motivation, as these factors modulate the amount of attentional resources allocated to the task. Consequently, a loose attentional focus lessens the naïve attention model's power to facilitate mentalization. Also, agents may have several ongoing intentions on various timescales that create multiple attentional sets and channel processing to different kinds of information to varying degrees at all times. Human perception is seldom affected by only *a single* perception goal. Despite these limitations, even a partial model of A's attention will increase the success of mentalization compared to a system which is not sensitive to the agent's top-down attention.

II.

2.

Kampis D., Keszei A., Király, I. (2018): **Encoding third-person epistemic states contributes to episodic reconstruction of memories**. Commentary on Mahr and Csibra: Why do we remember? The communicative function of episodic memory. *Behavioral and Brain Sciences 41*. pp. 29-30. doi:10.1017/S0140525X1700139X

We propose that for successful episodic memory formation, potentially relevant aspects of a situation need to be identified and encoded *online* and retained for prospective interactions. To be maximally convincing, the communicator not only has to encode just *any* contextual detail, but also has to track information *in relation to* social partners.

We suggest that in order to retain the causal history of beliefs, encoding processes need to be sensitive to potential aspects of a situation that can be retrieved when an episodic memory is formed – which also enables avoiding assertions that would lead the social partner not to accept our claims of epistemic authority. This applies to the causal history of first-person, but it is also necessary with regard to third-person beliefs. The latter is especially important because, while it may happen that we have no prior communicative episode with the social partner, assertions of epistemic authority are in fact often preceded by a history of interactions with the addressee.

For successful construction of episodic memories that are used in communication, one often has to encode not just any contextual detail, but track information in relation to a specific social partner. To later recall information that is relevant for that person, online when a specific episode is unfolding the given elements have to be selected, encoded and stored, and crucially, often *indexed* to a specific person. Thus it needs to be taken into account online, what aspects of a certain event may be potential contributors to the later construction of the adequate episodic memory, as required by a communicative episode.

A further challenge is to describe what enables the identification of relevant memory traces at reconstruction. We propose that in order to bridge encoding and retrieval, online theory of mind (belief monitoring) has to support the encoding of information potentially relevant to the basis of belief formation. Episodic memory “hooks” onto these elements (of the causal history of belief formation for the social partner’s belief), and if a later cue refers to these bases of previously formed (attributed) beliefs, this enables the collection of adequate components of episodic memory. Importantly, this process requires the reidentification of the social partner and the reattribution of the social knowledge base and monitoring of potential differences between the self and the partner. Altogether, this mechanism increases the (perceived) veridicality of episodic beliefs reported in a communicative interaction.

The suggested interdependence between episodic memory and theory of mind opens novel perspectives with regard to the developmental trajectory of both domains. Namely, the emergence of episodic memory retrieval would be bootstrapped by communicative situations (e.g., Southgate et al. 2010) especially when mindreading is involved; and relatedly, the mindreading system could learn to update previously attributed beliefs according to relevant new information (Király et al., 2018, see below) through the emergence of episodic memory.

Part B.

I.

3. Elekes, F., Varga, M. Király, I. (2016). **Evidence for spontaneous level-2 perspective taking in adults**. *Consciousness and Cognition 41* . 93–103.

Social interactions are fostered by humans' propensity to compute their partner's perspective online. However, due to the mindreading system's limited capacity perspective taking (PT) was argued to occur spontaneously only for level-1, but not level-2 perspectives. We propose that level-2 perspectives (containing aspectual information) can also be computed spontaneously if participants have reason to assume that the partner is indeed aware of the objects' aspectual properties. Pairs of adult participants took part in the modified version of Surtees, Butterfill, and Apperly's (2012) number verification paradigm. Participants had prior information on their partner's task, which either called for processing aspectual properties or did not. The partner's inconsistent perspective was found to interfere with RT-s providing evidence for spontaneous level-2 PT. However, such interference only occurred when the partner's task involved processing the perspective dependent object feature, suggesting that PT was sensitive to the other's awareness of the to be represented information.

4. Elekes, F., Varga, M., Király, I. (2017). **Level-2 perspectives computed quickly and spontaneously: Evidence from 8 -to 9.5-year-old children.** *British Journal of Developmental Psychology* Vol35, No.4., 609-622. https://doi.org/10.1111/bjdp.12201

It has been widely assumed that computing how a scene looks from another perspective (level-2 perspective taking, PT) is an effortful process, as opposed to the automatic capacity of tracking visual access to objects (level-1 PT). Recently, adults have been found to compute both forms of visual perspectives in a quick but context-sensitive way, indicating that the two functions share more features than previously assumed. However, the developmental literature still shows the dissociation between automatic level-1 and effortful level-2 PT. In the current paper, we report an experiment showing that in a minimally social situation, participating in a number verification task with an adult confederate, eight- to 9.5-year-old children demonstrate similar online level-2 PT capacities as adults. Future studies need to address whether online PT shows selectivity in children as well and develop paradigms that are adequate to test preschoolers' online level-2 PT abilities.

5. Elek, L.P., Király, I., Szücs, R., Oláh, K., Elekes F. **Linguistic but not minimal group membership modulates spontaneous level-2 perspective interference in 8-year-old children.** Paper ready to submit

This paper presents evidence that social categorization affects spontaneous level-2 visual perspective taking (L2PT) differently depending on the type of social category in 8-year-olds. In Experiment 1 (N=46), children were paired with same-age peers, who belonged to the same or a different minimal group. In Experiment 2 (N=42) children participated with an adult confederate, who either shared their cultural group membership or was member of an out-group (inferred from a linguistic cue, accent). Spontaneous L2PT was not affected by the minimal group manipulation. However, accent deteriorated L2PT when it implied that the task partner belonged to an out-group. It is argued that social categories that are indicative of the partner's knowledge states but not ad hoc groups influence spontaneous mentalizing.

6. Elekes, F., Bródy, G., Halász, E., Király, I. (2016). **Enhanced encoding of the co-actor's target stimuli during a shared non-motor task.** *The Quarterly Journal of Experimental Psychology* 69, 2376-2389.

Task co-representation has been proposed to rely on the motor brain areas' capacity to represent others' action plans similarly to one's own. The joint memory (JM) effect suggests that working in parallel with others influences the depth of incidental encoding: Other-relevant items are better encoded than nontask-relevant items. Using this paradigm, we investigated whether task co-representation could also emerge for non-motor tasks. In Experiment 1, we found enhanced recall performance to stimuli relevant to the co-actor also when the participants' task required non-motor responses (counting the target words) instead of key-presses. This suggests that the JM effect did not depend on simulating the co-actor's motor responses. In Experiment 2, direct visual access to the co-actor and his actions was found to be unnecessary to evoke the JM effect in case of the non-motor, but not in case of the motor task. Prior knowledge of the co-actor's target category is sufficient to evoke deeper incidental encoding. Overall, these findings indicate that the capacity of task co-representation extends beyond the realm of motor tasks: Simulating the other's motor actions is not necessary in this process.

II.

7. Peres, K., Kampis, D., Király I. (2021): **The flexibility of early memories: Limited re-evaluation of action steps in 2-year-old infants.** Journal of Experimental Child Psychology 203:105046. doi: 10.1016/j.jecp.2020.105046.

This study investigated the flexibility of 2-year-old infants' retrieval and reenactment processes. In a delayed imitation paradigm, children were exposed to a constraint change (implemented by the distance of a target object) affecting the relevance of using a tool to obtain a goal (reach the object). In Experiment 1, during demonstration in the first session the tool was either relevant or irrelevant for reaching the goal, and 1 week later it either lost or gained its relevance, respectively (Figure 1). We found that when the tool became unnecessary (relevant to irrelevant change), children used it somewhat less than before and used it less compared with when the tool's relevance remained the same (relevant to relevant, no change). When the tool became necessary after a constraint change (irrelevant to relevant change), children used the tool more than before, but not as much as in the Relevant-Relevant control condition. In Experiment 2, the timing of the constraint change (immediate or delayed) was varied in a modified version of the Irrelevant-Relevant condition, where practice before the constraint change was omitted. Children were not significantly more flexible in the immediate condition than in the delayed condition, and comparisons with Experiment 1 showed that performance did not change if we omitted the practice before the change (Figure 2). These results indicate that although 2-year-olds show considerable mnemonic performance, they face difficulties in adapting to constraint changes. We propose that this inflexibility may stem from infants' inability to revise their evaluations formed in previous events due to their immature episodic memory capacities.

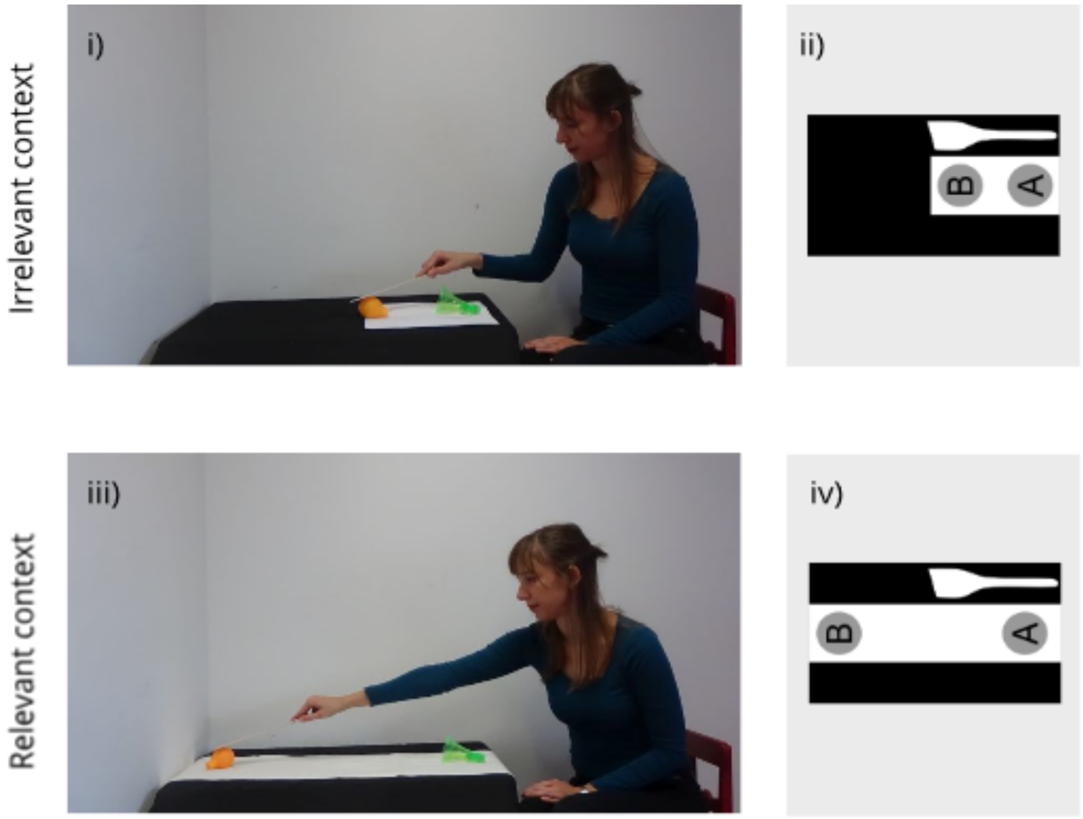


Figure 1.

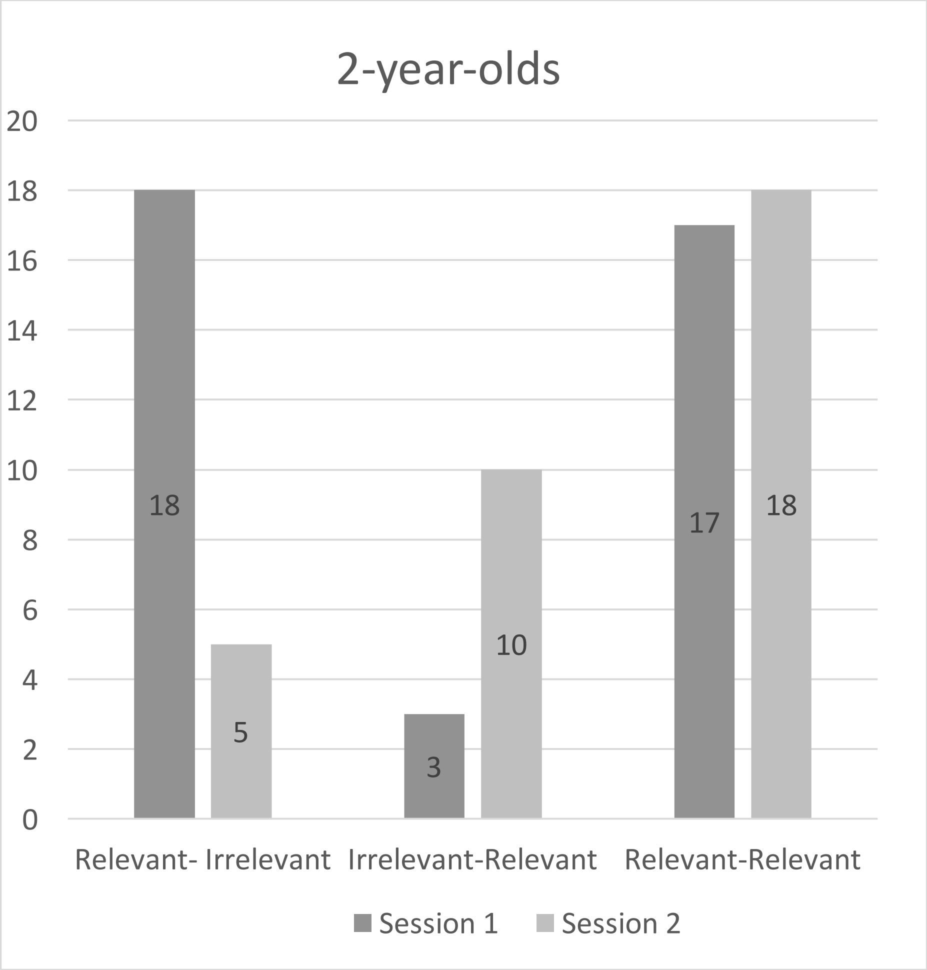
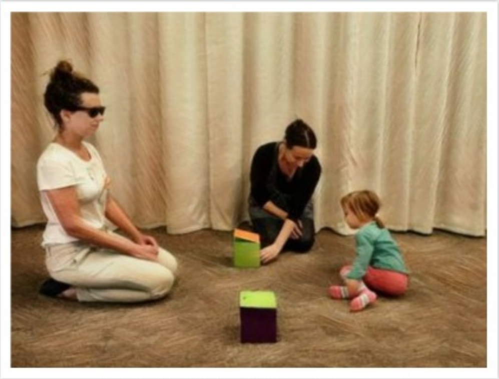


Figure 2.

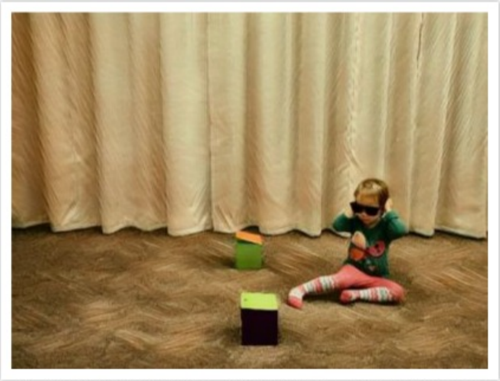
8. Király I, Oláh K, Csibra G, Kovács Á. M. (2018) **Retrospective attribution of false beliefs in 3-year-old children**. *Proceedings of the National Academy of Sciences of the United States of America*. 115 (45) 11477-11482. https://doi.org/10.1073/pnas.1803505115

The continuous flow of social interactions requires humans to monitor others' mental states dynamically, yet this aspect of mind reading remains largely neglected. We tested whether, beyond prospective belief tracking, young children would also attribute beliefs to others retrospectively. We found that 3-year-old children retrospectively inferred the content of someone's beliefs by combining present information with relevant events retrieved from episodic memory. This finding shows that emerging capacities for episodic memory contribute to the development of social cognitive processes, enriching children's ability to monitor others' mental states.

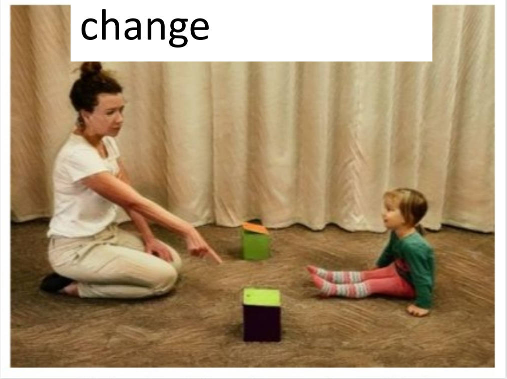
A current debate in psychology and cognitive science concerns the nature of young children's ability to attribute and track others' beliefs. Beliefs can be attributed in at least two different ways: prospectively, during the observation of belief-inducing situations, and in a retrospective manner, based on episodic retrieval of the details of the events that brought about the beliefs. We developed a task in which only retrospective attribution, but not prospective belief tracking, would allow children to correctly infer that someone had a false belief. Eighteen- and 36-month-old children observed a displacement event, which was witnessed by a person wearing sunglasses (Experiment 1, see Figure 3). Having later discovered that the sunglasses were opaque, 36-month-olds correctly inferred that the person must have formed a false belief about the location of the objects and used this inference in resolving her referential expressions. They successfully performed retrospective revision in the opposite direction as well, correcting a mistakenly attributed false belief when this was necessary (Experiment 3, see Figure 4). Thus, children can compute beliefs retrospectively, based on episodic memories, well before they pass explicit false-belief tasks. Eighteen-month-olds failed in such a task, suggesting that they cannot retrospectively attribute beliefs or revise their initial belief attributions. However, an additional experiment provided evidence for prospective tracking of false beliefs in 18-month-olds (Experiment 2). Beyond identifying two different modes for tracking and updating others' mental states early in development, these results also provide clear evidence of episodic memory retrieval in young children.



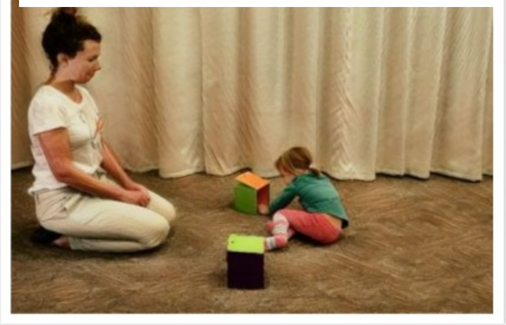
1: Location change



2: Discovering the sunglasses



3: Object request



4: Object selection

Figure 3.

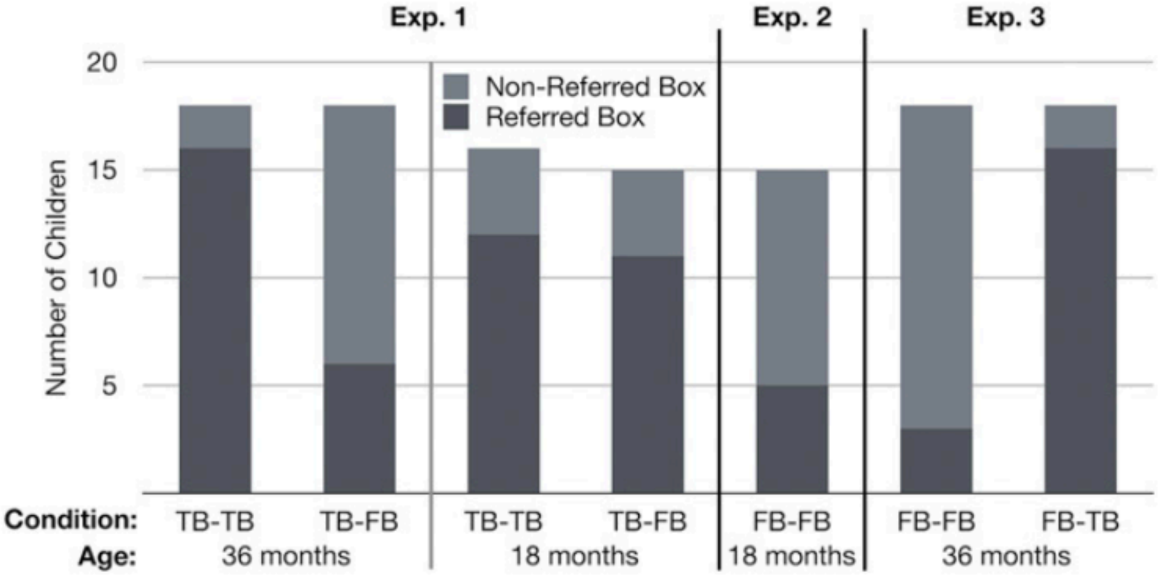


Figure 4.

9. Pomiechowska, B., Téglás, E., Kovács, Á.M., Király, I. **What the eyes tell about beliefs: Increased pupil dilation reflects on-line updating of others' beliefs in adults and 5-year-olds.** Paper ready to submit

Social interactions heavily rely on our ability to continuously keep track of what our partners think and know about the world. However, these attributed beliefs do not always correspond to changes in the state of affairs, or one might realize that a specific attribution may have been incorrect. How does the process of belief monitoring and updating unfold over time?
Across three eye-tracking experiments with adults and 5-year-old children we demonstrate that, from early on, we update others' beliefs on-line, just milliseconds after new information relevant to others' mental states comes in. The belief-updating computations are indexed by an increase in pupil dilation, when an attributed belief has to be revised, both in adults and children (see Figure 5). In addition, our findings suggest that the time-course of spontaneous on-line belief updating can be assessed using eye-tracking technology (Figure 6). This methodological advancement provides a means for future research to better understand how belief tracking emerges in ontogeny and phylogeny.

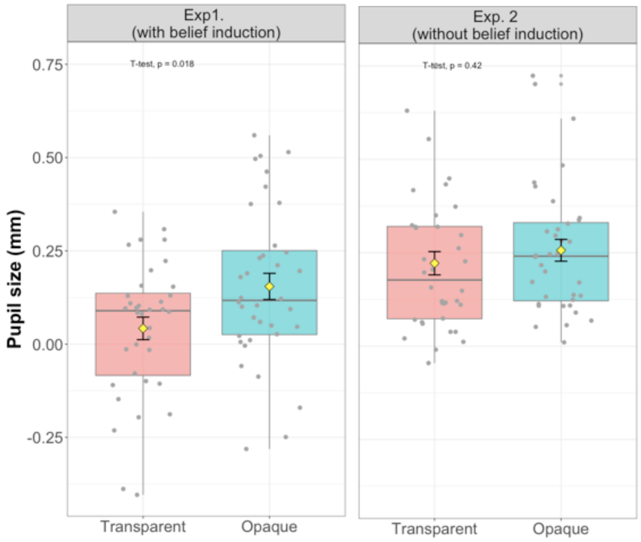


Figure 5.

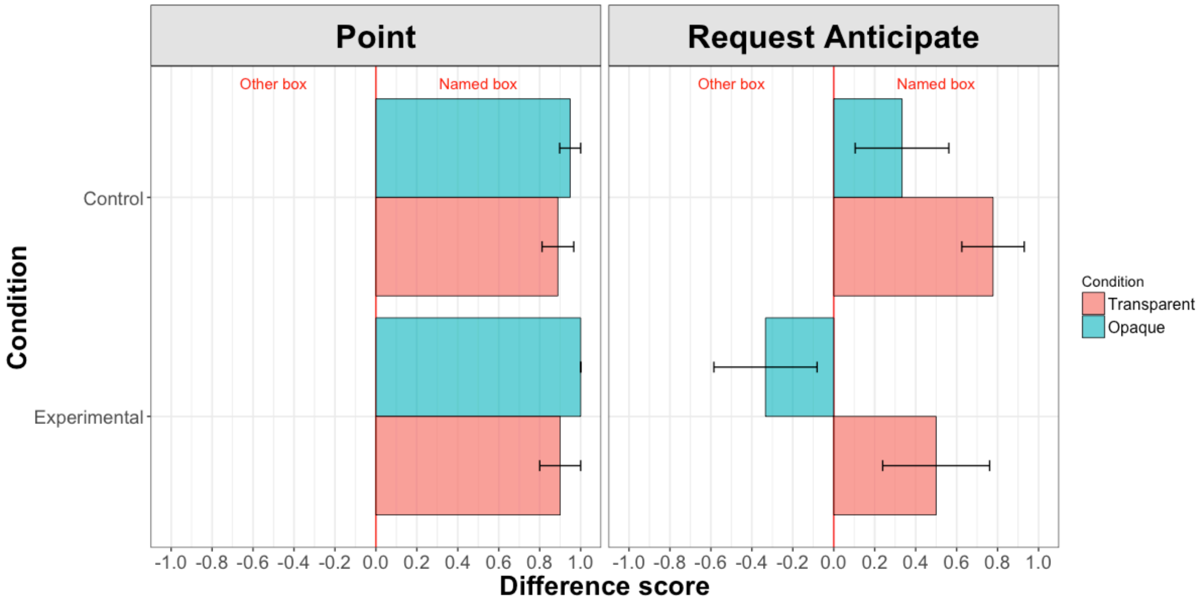
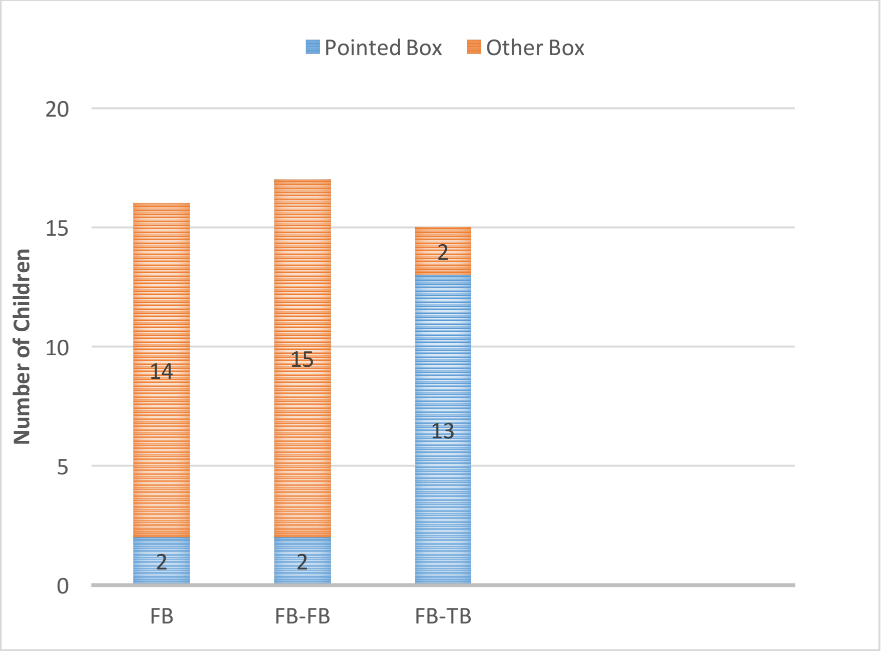


Figure 6.

10. Ildikó Király, Katalin Oláh, Gergely Csibra, Ágnes M. Kovács: **Do 18-month-olds revise attributed beliefs?** Article ready to submit:

Belief attribution can be performed or revised in a retrospective manner by retrieving the details of the events that might have generated the beliefs. Previous studies found that 3-year-olds, but not 18-month-olds could attribute false beliefs retrospectively to other agents after first having the opportunity for attributing a true belief. Here, in contrast, we tested whether 18-month-olds could revise an attributed false belief (FB) into a true belief (TB) when they learned that the person could have witnessed the situation that they initially thought had not been perceived by her. The infants first observed two novel objects hidden by Experimenter 1 (E1) into two boxes. Then E1 left the room, and while she was away, the locations of the objects were swapped. Infants were then asked to accompany Experimenter 2 (E2) to the adjacent room to call E1 back. When they entered the room, infants in the FB-revised-to-TB condition observed E1 peeking into the experimental room through a one-way mirror, while in the FB-stays-FB condition they observed E1 reading a book, while the one-way mirror was covered. Upon return E1 requested an object by pointing to one of the two boxes. Eighty-eight percent of infants chose the non-referred box in the FB-stays-FB condition, but in the FB-revised-to-TB condition, in which they witnessed E1 peeking through the one-way mirror, eighty-seven percent of infants chose the other, the referred box (Figure 7). Thus, 18-month-olds could revise an already attributed false belief after having learnt that this attribution might have been wrong. This points to the flexible use of ToM capacities early on that require the combination of belief-relevant information originating from different sources.



:

Figure 7.

Overall summary:

The studies we designed allowed us to reconcile the classic explanatory theories of Theory of Mind with the more recent findings in cognitive development that challenge those theories. The model we propose re-defines what the primary goal of mindreading is, namely to help to form quick behavioral adjustments in social interactions by mapping others' beliefs online, in a real-time manner.

Theory of mind and perspective taking provide the most fundamental building blocks of human nature. Their significance is indisputable in all situations where people (coming from any age or social group) interact with each other. Theory of mind, the ability to understand someone else's mental states regarding his situation, grounds our social sensitivity, enables us to communicate with confederates, or even to resolve conflicts. With our research we prove that the cognitive capacity necessary for these develops early in childhood. We map others' mental states already during the unfolding of a social event (online) in a quick, efficient and and intentional way. This representation in turn affects our own behaviour and makes it possible to predict others' actions.

Our theoretical proposal has been published in two theoretical papers (Part A), and has been supported by empirical research that has been disseminated in five experimental papers and 3 manuscripts (Part B).

References:

Butterfill, S.A. Apperly, I.A. (2013), How to Construct a Minimal Theory of Mind. *Mind and Language, 28*: 606-637. https://doi.org/10.1111/mila.12036

Rakoczy, H. (2012), Do infants have a theory of mind?. *British Journal of Developmental Psychology*, 30: 59-74. https://doi.org/10.1111/j.2044-835X.2011.02061.x

Scott, R.M., Baillargeon, R. (2017). Early False-Belief Understanding, *Trends in Cognitive Science*s, 21: 4, 237-249. https://doi.org/10.1016/j.tics.2017.01.012.

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.

Surtees, A.D.R. and Apperly, I.A. (2012), Egocentrism and Automatic Perspective Taking in Children and Adults. *Child Development, 83*: 452-460. https://doi.org/10.1111/j.1467-8624.2011.01730.x