Attention in naïve psychology

Fruzsina Elekes^{1,2} & Ildikó Király^{1,3,4}

1 MTA-ELTE Social Minds Research Group, Budapest, Hungary

2 Social Mind and Body Group, Central European University, Budapest, Hungary

3 Cognitive Development Center, Central European University, Budapest, Hungary

4 Institute for Advanced Studies, Paris, France

Corresponding author: Fruzsina Elekes, elekes.fr@gmail.com

Revised manuscript submitted to *Cognition* on 24.07.2020.

Abstract

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.

Keywords: theory of mind; top-down attention; selective processing; naïve attention model

1. Mentalizing in its everyday context

A multitude of factors influence how the world is represented in one's mind, how it is evaluated and how it is stored in memory. Information does not simply flow through the mind leading to the formation of objective marks of the environment, like light on photo paper. The process through which the physical environment becomes accessible to humans has an active, constructive nature (Gregory, 1997; Helmholtz, 1866; Norman, 2002). The perceiver's mind (its current features and state) has a great role in determining whether and how information will be represented in it (Gibson & Rader, 1979), rendering perception largely subjective and personal. The particular way in which something is represented constitutes the agent's perspective on that information. Perspective differences may take various forms. For instance, previous experiences may influence how current events are interpreted: a frightening past encounter with a dog may render any subsequent exemplars scary to one person, even if due to their size and demeanor others see them as adorable. Current goals may also shape information processing: an urge to drink a coffee before giving a lecture will make a scholar's mind tuned to coffee-related details, potentially missing other salient information in plain sight (like a colleague waving from across the street).

The idiosyncrasy of this process poses a particular challenge for human social life. Humans' social success depends in great part on their ability to trace how others represent their physical and social surroundings. That is, the representation that an agent (A) forms is often the target of implicit and explicit inferences performed by observing individuals for social purposes. This process is called *mentalization* and the capacity itself, Theory of Mind (Premack & Woodruff, 1978). The lack of one to one correspondence between the physical world and the reflection it creates in one's mind imposes great challenges for the mindreader (M) who is essentially interested in the latter. The mere presence of an intentional agent says very little about which aspects of the environment the agent will process, and even less on the

subjective manner in which it will be processed. Consequently, for M to reach accurate enough predictions about A's representations, she must have some degree of understanding about the characteristics *of A or A's mind* that gives rise to his mental states.

In everyday life, mentalizing is embedded in a wide context of background knowledge that may contribute to its success. Humans enter any social situation with a broad range of expectations, based on, for instance, prior knowledge about the target agent's personality traits, behavioral tendencies and cognitive abilities (Conway, Catmur, & Bird, 2018; Tamir & Thornton, 2018) or social group memberships (Oláh, Elekes, & Király, 2019; Spaulding, 2018; Westra, 2017). Such characteristics of the agent may *inform* the mentalizing process in important ways. Conway and colleagues (2018) argue that mental state inferences are made based not only on how adequate a certain mental state is under the current environmental constraints, but also on how likely A's mind (due to the above features) is to hold that mental state. In a similar vein, knowledge about the social group memberships of A may help ascertain the partner's epistemic states, as categories carry information regarding the circle of knowledge group members would likely possess (Oláh et al, 2019).

The above accounts point out that achieving a deep understanding of A's mental contents entails integrating information from two different time-scales: the everchanging features of the current interaction on the one hand, coupled with *a priori* expectations on stable, enduring social characteristics on the other (a similar notion is reflected in recent models of common ground as well, see Kecskes & Zhang, 2009). Here we will propose that humans can use another layer of inference as well, that falls in-between the temporal scope of the above two. This enables M to form temporarily extended expectations about A (like personality or group memberships), while better answering the requirements of flexible, dynamic, real-time interactions.

As illustrated before, A's *existing* mental states (and intentions specifically) exert a modulatory effect on how he forms *subsequent* mental contents through creating a cognitive context for processing. This modulatory effect is in great part realized by channeling his attention selectively to goal-relevant information (Serences & Yantis, 2006). We argue that knowledge about A's goal dependent selective attention, i.e. attentional set, is a uniquely powerful source of social inference. Rather than being linked to either the past (detached from the specifics of the current interaction) or the immediate present (without considering the broader context of the interaction), attention provides a form of background knowledge that originates in and is valid throughout *the length of the current event*. Importantly, being grounded in the ongoing event, selective attention has direct inferential power regarding the epistemic states that A obtains within that event.

In the following paragraphs, we will introduce a theory that we call naïve attention model theory (AMT).¹ We will 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. The argumentation will unfold in the following structure. In section 2, we will give a summary of how attention shapes first person processing, pointing out its utmost relevance from the mentalizer's point of view and we will delimit which aspects of attention

¹ The "naïve attention model theory" (AMT) is the theory described here on the possibility that humans develop a naïve model of attention. The expression "naïve model of attention" will be used to refer to a set of general, rule-based presumptions, which presumptions likely differ in the extent to which they are available for symbolic manipulation. It is not intended to mark a single, explicit theory – as spelled out in detail in section 4.

our theory will target. In section 3, we review how the cognitive capacity constraints that lead to the selective nature of perception create a similar need for selectivity in mentalizing and point out that tracing others' attentional processes may answer that challenge. Section 4 gives a detailed description of the naïve attention model, discussing how knowledge about an agent's top-down attentional focus interacts with the informational complexity of the surroundings in determining the accuracy and subjective confidence of mental state predictions. We outline how the naïve attention model may contribute to online predictions, give a tentative account on the model's ontogenetic development and outline some of the empirical predictions of the naïve attention model theory. Finally, in chapter 5, we position our theory among the existing theoretical accounts on social attention.

2. Attention, please!

Let us first consider in more detail the role that attention plays in cognition. Attention is the mechanism that controls orienting to, filtering of, and searching for information both in the external and in the internal environment (such as one's memories, thought or internal semantic structures, Posner, 1980). By orchestrating these processes, it enables the formation of a coherent experience against the blurring fuse of information the cognizer is continuously exposed to (Cohen, Cavanagh, Chun, & Nakayama, 2012; Fritz, Elhilali, David, & Shamma, 2007). As only a limited amount of information can be kept in consciousness at any given time, attention, as a gatekeeper, protects the cognitive system from being flooded by detail (Serences & Yantis, 2006). Attended information receives more processing (Anderson, Hillyard, & Müller, 2008) and is retained in memory more successfully, while those bits of information that do not get prioritized by attention will likely fall outside of conscious perception (e.g. Bentin, Kutas, & Hillyard, 1995; Mack & Rock, 1998).

Information may receive attention for different reasons: while bottom-up attention capture is a saliency based, involuntary process, top-down attention can be characterized as

being allocated voluntarily based on the cognizer's mental states, goals, desires, expectations, or emotions (Serences & Yantis, 2006).² The interaction of these competing processes determines the priorities of A's cognitive processing (Awh et al., 2012), or, from a phenomenological point of view, A's conscious visual experience (Egeth & Yantis, 1997; Serences & Yantis, 2006).

Taken together, a large portion of the idiosyncrasy in the way people represent their surroundings comes from the way attention orients cognition and creates priorities in processing. From the mentalizer's point of view, monitoring another person's attention appears to have the power to promote identifying his epistemic states and interpreting or predicting his behaviors based on these. We argue that humans are capable of utilizing this relation through obtaining a *naïve psychological understanding of attention* over ontogenetic development.

We will limit our discussion to how top-down, goal-driven attention may be adopted for the sake of social understanding, while we acknowledge that bottom-up attention capture can also be used to predict or influence the partner's behavior (Philips, 2019). Top down attention can be conceptualized as the *intention to see* (Gibson & Rader, 1979) or more generally the intention to perceive, and is an excellent candidate for social purposes for two reasons. On the one hand, being causally linked to mental states, top-down attentional sets should be accessible to social cognition through basic forms of mentalizing, like goal

² More recently, it has been put forward that top-down attention includes involuntary forms of selection as well, in addition to the goal-driven, voluntary selection processes. Such involuntary processes are nevertheless considered "top-down" as they depend on the agent's prior experiences, i.e. his history of attention allocation (Awh, Belopolsky, & Theeuwes, 2012).

attribution or preference-desire reasoning. On the other hand, acknowledging the relation between a goal and the resulting attentional set not only allows M to identify what kinds of information A is *currently* prioritizing, but also when he will stop prioritizing them. When a goal is obtained, the consequent attentional set ceases to exist as well, providing M with a criterion for both maintaining and discarding an expectation about A's attention.

Humans' mental states are shaped by information coming from multiple modalities at the same time. Accordingly, attention plays a similar role in filtering and prioritizing information received through different perceptual modalities (Hansen & Hillyard, 1980; Spence, 2002) and may even exert its effect across modalities (Fritz, et al., 2007). We presume that humans' understanding of selective attention entails attention as a domain general process and therefore can be applied to refine predictions about mental states acquired through any perceptual modality. Following an existing trend in the mindreading literature we will center this paper around how humans understand the role of selective attention in the formation of visually informed epistemic states. However, we will also discuss how the presumptions of the naïve attention model extend to other modalities, such as hearing.

We do not anchor our model to the wide-spread overt (fixated) / covert (non-fixated) distinction, as by distinguishing attentional processes along gaze, it unnecessarily constrains discussion to visual perception while also fails to grasp the entirety of how gazing and attention relate in vision (see chapter 4.1.).

3. The selectivity problem in mindreading

In their attempt to map out the mental realm of their fellow social agents, humans often form mental state predictions based on sparse evidence. Although a good enough prediction is essential for the success of our everyday social interactions, the mentalizer faces (at the very least) two challenges in this process. First, the strongly filtered nature of

conscious processing means that A's mental states can only be partially predicted based on the environment itself (and A's physical position in it). Secondly, although the computational problem at hand is quite complex, M faces similar capacity constraints to A. This means that just as the agent is forced to be selective in what he forms a mental experience of, the mentalizer will also be unable to compute A's perspective including *all* the information that is visually accessible to A. M will necessarily be selective even in cases of automatic or spontaneous instances of mentalizing (Carruthers, 2017). On what basis this obligatory selectivity takes place has ample consequences for the success of M's mentalization attempts. In Carruthers' view (2017), selectivity is determined by the objects' saliency to M, where saliency may be endowed to an object for its *relation to A*, that is, due to being perceived as target of A's behavior, like his gaze, or communicative utterances. This view on saliency as the basis of selectivity in mentalizing corresponds to Wilson and Sperber's (2002) argument that humans have the ability to predict what is relevant to a partner, which information they use to infer what the partner will be most likely to derive genuine knowledge from.

Building on and extending the above line of work, here we propose that humans can take a more fine-grained approach that allows them to fine tune their mentalizing to the characteristics of its current target, leading to more accurate and confident predictions. We argue that a naïve grasp of the nature of the challenge itself allows humans to overcome the obstacle. Specifically, that, the selectivity of cognition is shaped by regularities: the way people prioritize information for mental representation reflects their current goal-driven attentional sets. Others have also argued for the theoretical relevance of tracking fellow agents' attention to facilitate ToM computations (Cole & Millett, 2019). These authors emphasize that perspective taking is necessarily a knowledge-based, inferential process as someone else's raw visual experience is impossible to be represented. However, regarding the role of attention in perspective taking, Cole and Millett (2019) conclude that "no other

observer can know where our attention is” (p. 1794), rejecting the possibility that such mechanism would exist. Here, we aim to provide a detailed account on the tools that, in our view, humans may be in possession of serving just that purpose.

We propose that humans develop a naïve psychological understanding of attention and consider it a factor while making inferences about the representations that A will form about his environment in a given situation. That is, agent A is regarded not simply as *an* intentional agent but one that is engaged in *a specific intentional relation* with his surroundings; whose idiosyncratic way of representing the world can be understood and predicted by considering the wider set of representations and mental states in which it is embedded. Assuming that, M’s primary goal is to understand A and that she has a general naïve understanding of how attention mediates between the outside world and its internal representation, we expect M to adjust the way she selects information for mentalizing to *the way A is expected to filter his experiences*. Therefore, although selectivity in mentalizing is a cognitive necessity, it may promote the accuracy of mentalizing provided that M can direct her constrained amount of resources in a way that mirrors A’s current perceptual goals. In the following paragraphs, we summarize the principles we assume the naïve attention model may contain, and how the naïve model may operate.

4. The naïve attention model

We propose that humans construct a naïve model that contains presumptions about attention as a cognitive skill. This model works like a generative framework: it contains a set of predefined features the model is applicable to and includes inherent conditions of use that guide information selection and facilitate the interpretation of an observed event or behavior in a format that will allow predictions. The general presumptions of the model are used by M to form expectations about the perception of particular individuals under specific environmental conditions, that is, the naïve attention model is applied to A, yielding an

implemented model, a set of expectations, of A's cognitive priorities. In the following, we describe our construal of the presumptions of the naïve attention model in detail:

Firstly, the attention model assumes that cognition, in conscious, awake states at least, is always oriented towards *something*. We propose that the default expectation of the naïve attention model is that this attentional focus is external to the agent. In other words, humans expect other humans to have a general interest in, and an intentional relation towards their surroundings even if the specific form of intentionality set by the agent's goal is unknown. This assumption draws on Dennett's (1981) theory of the intentional stance, which proposes that humans are inclined to see rational agents engaged in an intentional relation with their environment, knowing all relevant truths their sensory experience avails them. The default assumption of external focus can be revisited if contradictory evidence arises (Benedek, Daxberger, Annerer-Walcher, & Smallwood, 2018).

Secondly, the attentional model expects humans to be selective processors, acknowledging that exposure to information does not necessarily lead to conscious experience. As perception is embedded in the current cognitive-motivational context of the agent's mind, based on exposure only, it cannot be determined with absolute certainty whether someone has perceived something. An external attentional focus does not yield experience of all information out there. Rather, attention is selectively allocated to specific kinds of information, determining what the agent forms an epistemic state of.

Thirdly, the naïve model presumes that the focus of an agent's attention is determined by the *goals* they are currently pursuing so that the agent's attentional set promotes processing information that is relevant with respect to the goal in question. Furthermore, the status of the agent's goal, be it ongoing or fulfilled, is used to predict whether the corresponding attentional set should be maintained further or discarded.

The naïve model of attention is modality independent. The model is centered around the notion that attention originates in the intentions of an individual, enhancing their perception for attended information (yielding knowledge of attended information more probably than unattended information) and this presumption is not restrained to any perceptual modality. Note however, that there are modality dependent factors, in addition to the modality independent workings of attention, that the naïve attention model needs to entail. In visual perception such factor is mere visual access to information, defined by the agent's body orientation and gaze direction. Only information within the visual field can be brought into consciousness through attentional selection. That is, the epistemic state that an agent is most likely to acquire is at the intersection of their current attentional set *and* their visual field. Disregarding one or the other would yield specific mentalization errors that we will explore later. Importantly, the constraint of visual access provides a powerful, though rudimentary, behavioral cue to perception: gaze (see 4.1.). In another leading modality, hearing, perception is not constrained in such manner. Auditory information can be perceived regardless of the perceiver's orientation in space, and so the naïve attention model's presumptions can be implemented in a more direct way.

Finally, we assume that the mature attention model entails an understanding of the probabilistic relation between attention and consciously accessible epistemic states. In its richest form, the naïve model of attention may acknowledge that the intention to seek out something (voluntary attention) means that *priority* is given to its processing in the cognitive system without an absolute certainty of conscious experience. And vice versa, that a lack of top-down attention means less probable processing, even though it does not render it impossible.

When recruited, the above presumptions of the naïve attention model receive input about the specifics of the current event and generate a set of expectations, an implemented

model about A's (a particular individual's) processing priorities. These expectations then shape the way subsequent mental state inferences are made. The output of the naïve attention model enables ToM to determine which is the *most viable* epistemic state of all those that are physically possible in the given environmental context, by considering the probability with which A would obtain that epistemic state under the temporary "attentional settings" of his mind. We argue that the implemented model is created and enriched incrementally, depending on the amount of cognitive resources allocated to its computation (see 4.2.).

Note that the presumptions of the naïve model may conflict with the scientifically established state of affairs regarding attention. For instance, the exact role attention plays in creating conscious experience is still disputed in cognitive science (Cohen et al., 2012; Koch & Tsuchiya, 2006; Lamme, 2003; Posner, 1980). Also, the tendency to select certain kinds of information for processing does not vanish instantly and without leaving a trace when the intention it serves is fulfilled (Awh et al., 2012) and there might be many other factors influencing perception in any given moment. Certainly, knowledge of A's attention is not sufficient to unequivocally predict the content of his mind. However, mapping out even such probabilistic regularities of the environment (the workings of the human mind in this case) can improve the quality of prediction if it is detected by the agent. The naïve model does not have to be scientifically accurate to be useful as long as it enables more efficient predictions than the organism would be capable of otherwise. In the following paragraphs, we describe the operation of the model in more detail focusing on the visual modality.

4.1. Intentions and gaze direction as indicators of attention

Intentions and gaze are both profoundly interconnected with visual attention, but their roles with respect to shaping attention, and, therefore the inferential power they carry for social life, differ. While intentions *create* attentional sets, gaze determines the range of information that can potentially be heightened and brought to consciousness by top-down

attention. We will now analyze the nature of these cues in more detail and the inferences they allow the mindreader to make. We present evidence on how humans can use this information as indicators of attention and point out that the role of intentions is largely unexplored so far.

There is overwhelming evidence that fellow agents' eyes constitute a special stimulus for the human brain. The human eye's unique morphology makes gaze direction easily perceptible to others (Kobayashi & Kohshima, 1997), and this enabled gaze to gather an evolutionarily novel function in humans, serving cooperative social interactions (Tomasello, Hare, Lehman, & Call, 2007). A's gaze direction impacts *M's attention* in a speeded and seemingly uncontrolled fashion, indicating that there may be a dedicated mechanism for gaze processing (Baron-Cohen, 1991; Langton, Watt, & Bruce, 2000). Humans allocate attention to gaze-cued locations, detecting stimuli more easily at the cued location and this facilitation occurs at the early phases of visual processing (Schuller & Rossion, 2001). Another's gaze is also known to impact subsequent memory of gaze cued information (Gregory & Jackson, 2017; Nie, Ding, Chen, & Conci, 2018). The pervasive nature of gaze cueing is further demonstrated by its resistance to information about the gaze's lack of informational validity (e.g. Friesen & Kingstone, 1998; Downing, Dodds, Bray, 2004; Driver, Davis, Ricciardelli, Kidd, Maxwell, & Baron-Cohen 1999; Farroni, Johnson, Brockbank, & Simion, 2000; Ristic & Kingstone, 2005).

Some argue that humans use gaze as the primary cue to others' attention (Baron-Cohen, 1991). There is in fact accumulating evidence that the cuing effect of gaze is susceptible to information that is social in nature, suggesting that A's gaze orients M's attention on a fundamentally social basis (e.g. Bayliss, Schuch & Tipper, 2010 – but see: Hietanen & Leppanen, 2003 –; Deaner, Shepherd, & Platt, 2007; Gobel, Tufft, & Richardson, 2017; Gregory & Jackson, 2019; Marotta, Lupiáñez, Martella, & Casagrande, 2012; Pavan, Dalmaso, Castelli, & Galfano, 2011). The literature suggests that gaze cuing is more likely to

occur when the gazer has the necessary means to hold mental states (is intentional, Wiese, Wykowska, Zwickel, & Müller, 2012; Gobel et al., 2017; Terrizzi & Beier, 2016) and when the gaze cue is interpreted as part of an intentional gaze-shift (Hietanen, 2002). Although when two masks are worn by the same individual on front and back of the head, the gaze cuing effect fails to emerge even though participants can verbally report which of the two mask directions hid intentionality (Kingstone, Kachkovski, Vasilyev, Kuk, & Welsh, 2019). Finally, there is evidence that the gazer's assumed visual access to the target modulates the effect of the gaze itself (Király, Oláh, Csibra, & Kovács, 2018; Nuku & Bekkering, 2007; Teufel, Alexis, Clayton, & Davis, 2010, but see: Cole, Smith, & Atkinson, 2015 or Wilson, Soranzo, & Bertamini, 2017). Based on these findings some conclude that participants interpret another agent's gaze as reflecting the gazer's attention, and consequently, they allocate their attention to objects that they assume another person is also *attending to*. In fact, some go as far as to suggest that the gaze cuing paradigm measures joint attention under laboratory conditions (Bayliss et al., 2010, for an analysis of what psychological processes "jointness" may entail with respect to attention, see Siposova & Carpenter, 2019).

However, the acuity of predictions regarding A's attention based on gaze direction (and consequent inferences about epistemic states) encounters limitations for multiple reasons. While it is true that people often look where they attend, this relation is not mandatory in either its components. On the one hand, information that is not fixated may still be attended covertly and therefore processed preferentially (Cameron, Tai, & Carrasco, 2002; Posner, 1980): attention can shift to peripheral locations evidenced by RT benefits for targets presented at the cued versus the un-cued location – while maintaining central fixation. On the other hand, looking at something does not necessarily mean that the target is attended to or represented consciously (Mack & Rock, 1998). Voluntary attention may be directed to certain features of a scene so efficiently that other, fully visible, salient and sometimes relevant

features go unnoticed (e.g. Neisser & Becklen, 1975). As a matter of fact, humans have a striking tendency not to perceive events they did not expect to happen, or which are irrelevant for their ongoing goals and intentions. This phenomenon, called inattentional blindness (Mack & Rock, 1998), is a pervasive feature of the human mind and it even extends to objects or events that are *directly fixated by A* (Mack, 2003). These two phenomena, covert attention and inattentional blindness demonstrate that attention can fully dissociate from the spatial location of fixation.

Furthermore, humans can efficiently broaden and constrain their spatial focus of attention changing the total size of the attended spatial area depending on task demands (Goodhew & Plummer, 2019), and this process is independent of gaze. Also, attention can be directed at information on a feature level, such as color or orientation, in which case it spreads over the *entire visual field*, highlighting the target feature (Corbetta, Miezin, Dobmeyer, Shulman, & Petersen, 1990; Liu, Stevens, & Carrasco, 2007; Maunsell & Treue, 2006; Störmer & Alvarez, 2014). The accompanying gaze pattern is thus often not descriptive of its target, rendering gaze direction inapt for social use. In light of these phenomena, the existing trend in empirical research to operationalize attention as targeting spatial locations or objects (e.g. Fink, Dolan, Halligan, Marshall, & Frith, 1997; O'Craven, Downing, Kanwisher, 1999; Serences, Schwarzbach, Courtney, Golay, & Yantis, 2004) potentially leads this field to overestimate the social informativeness of gaze cues.

To summarize, due to the nature of its relation to attention, gaze has an important but constrained use in mentalizing. Gaze direction determines which part of the environment will be projected to A's retina, thereby determining the agent's visual field. Only the information present in the visual field can be brought into consciousness at any given time with the help of attention, therefore, gaze defines the *range of information* that A may acquire knowledge of. At the same time, A's goal-dependent attentional set defines the circle of information that he

intends to process in depth, causing a temporary sensitivity of A's cognition to that information. Therefore, a fuller understanding of the mental states that A acquires calls for an integrated understanding of both of these things.

4.1.1. Empirical evidence for the use of intentions. In the following paragraphs we review empirical evidence suggesting that humans can go beyond the specificity of gaze direction, allocating more processing capacities to the subset of information that their task partner *intends to* process selectively. This has been found to impact memory performance, visual perspective taking, as well as stimulus detection.

In an array of visual stimuli, participants' attention is drawn to the information that their task partner is instructed to keep in working memory, and this results in better retrieval of these items compared to baselines (He, Lever, & Humphreys, 2011). In a similar experiment (Eskenazi, Doerrfeld, Logan, Knoblich, & Sebanz, 2013), participants were instructed to observe a stream of words jointly with a partner and react selectively to a subset of these (e.g. P1 to animals, P2 to objects, no one to plants). Later they were faced with a surprise recall test of all items. Participants showed enhanced memory performance to the subset of words that their partner attended to compared with the non-attended subset, called the joint memory effect (JME) (Elekes, Bródy, Halász, & Király, 2016; Elekes & Sebanz, 2020; Eskenazi, et al., 2013; Wagner, Giesen, Knausenberger, & Echterhoff, 2017). In the joint memory paradigm gaze direction does not differentiate between attended and non-attended information as all items are presented centrally, looked at and processed semantically by both participants. Rather, attention is operationalized as each individual's task induced intention to detect their own subset.

A subsequent study in the same paradigm has provided further evidence for the importance of the partner's intentions in the JME (Elekes & Sebanz, 2020). The authors found no memory benefit for the partner's targets when the partner's task was to detect words of a

specific presentation color instead of semantic category. This finding implies that others' attention is not treated as a spatial spotlight which highlights an object as a whole (a word in this case), rather a task partner's attention is represented on a more fine-grained scale, on the level of object features. As a consequence of this selectivity, encoding is only enhanced for those words that the partner is likely to obtain epistemic states of (processes semantically). Elekes and Sebanz also found that participants could *link* partner-relevant items to the partner, suggesting that the JME extends to the social source of the learnt content.

The partner's task induced attentional set has been found to impact perspective taking as well. Ambiguous visual stimuli, like the character 6/9, carry different meanings from opposing perspectives. Computing how such an object appears, and what it means to someone occupying the opposing viewing angle, is referred to as level-2 perspective taking (Flavell, Everett, Croft, & Flavell, 1981). Elekes, Varga and Király (2016) found that participants inferred the alternative perspective spontaneously when their partner was known to attend to the characters' meaning. However, the number's identity was not inferred from the partner's perspective when the partner's task called for attending to the presentation color of the same characters. That is, perspective taking emerged selectively, depending on the specific type of intentional relation the partner was engaged in with his environment. In a social context including repeated eye-contact and gaze shifting between partner and stimuli, others found no such goal-dependent modulation of level-2 PT (Surtees, Apperly & Samson, 2016). This indicates that the effect of intention based selective attention can be outweighed by coarser, more robust cues.

Furthermore, Frischen, Loach and Tipper (2009) report evidence that participants spontaneously simulate their partner's selective attentional processes in a spatial context. A spatial location that contained a distractor during a previous prime suffers from inhibition (negative priming) when it contains a target in the subsequent probe even when the prime was

responded to by the partner instead of the participant him/herself. In a similar vein, there is also evidence that participants spontaneously consider the breadth of their task partner's spatial attention when performing a Navon task in parallel (Böckler, Knoblich, & Sebanz, 2012). Specifically, when making identity judgments about the stimuli (a large letter composed of small letters), participants' RT-s were longer when their breadth of attention differed (being globally or locally defined) than when it was aligned. Others found indication that misaligned attentional foci have particularly devastating effect on response accuracy when the letters in the global and local foci differ, that is, not only the partner's breadth of attention but also the content of the partner's attention is in conflict with one's own (Fabbri, Frisoni, Martoni, Tonetti, & Natale, 2018).

Finally, it has been posited that the perception of gaze triggers certain cognitive effects, e.g. in a crowd of people with averted gaze, a single gazer whose eyes are directed at the perceiver pops out and is much easier to find than an averted gaze in a crowd of directed gaze (Conty, Tijus, Hugueville, Coelho, & George, 2006). Also, perceiving a gaze directed at oneself is known to disrupt one's performance in a working memory task (Wang & Apperly, 2017). The work of Colombatto, van Buren, and Scholl (2019; in press) suggest that instead of gaze per se, these effects might rely on perceiving intentionality, i.e. the mind behind the gaze. When clearly inanimate objects with no gaze-like features are arranged in a way that indicates coordination or design, the above effects can be elicited. The authors argue that directed attention can be inferred even in the absence of gaze cues.

The above findings provide the first evidence that humans' naïve conception about attention surpasses a mere sensitivity to gaze. In fact, under various conditions, prior information about the partner's intentions were found to outweigh the effect of a gaze cue. Through acknowledging that attention *originates* in intentions (and can thus be approximated from them), the mentalizer obtains an inferential tool that allows her to reach mental state

inferences of heightened accuracy even when behavioral cues, such as gaze, are not available or are uninformative regarding attention.

Note, however that representing a partner's task dependent attentional focus does not induce the same pattern of processing as one's own attention (Constable, Pratt, & Welsh, 2018). When two targets (T1, T2) follow each other closely in a stream of stimuli, participants often miss the second target if it falls within a critical time window, called an attentional blink (Raymond, Shapiro, & Arnell, 1992). According to Constable and colleagues (2018), when participants only attended to one of the targets and their partners attended to the other, detection accuracy for T2 remained high and independent from the lag between T1 and T2 even when participants were directly *asked* to monitor their partner's task.

4.2. The naïve attention model in online predictions

For real-time interactions to be successful, people need to be able to perform quite rich inferences under serious time constraints and need to update those inferences continuously as the interaction unfolds. The empirical work discussed above indicates that it may be possible to recruit the inferential principles of the attention model efficiently enough to contribute to such interactions.

There is a growing body of evidence that humans (both infants and adults) are susceptible to others' mental states, which is demonstrated even in cases of passive observation, or when participants are engaged in a task that does not call for social interaction and is evidenced in a very fast time-scale. These include studies on visual perspective taking (e.g. Baker, Levin, & Saylor, 2016; Beck, Rossion, & Samson, 2018; Elekes, et al., 2016; Freundlieb, Kovács, Sebanz, 2018; Luo & Beck, 2010) and false beliefs (e.g. Grosse Wiesmann, Friederici, Singer, & Steinbeis, 2017; Schneider, Slaughter, Becker, & Dux, 2014). However, it has also been suggested that the understanding that M can form under

these conditions lacks a truly mentalistic nature (Apperly & Butterfill, 2009; Heyes, 2014; Low, Apperly, Butterfil, & Rakoczy, 2016; Todd, Simpson, & Cameron, 2019). Others (Carruthers, 2017) argue that the representation formed about A's mind is gradually enriched as processing unfolds and the elaboration of the mental state inference depends on the amount of resources allocated to it. That is, although opinions on the nature of spontaneous mindreading(like) effects differ, there seems to be a consensus that the initial processing stages of mentalizing can only generate an incomplete output.

Consequently, it may appear that our theory proposes computations too demanding to be achieved in real-time interactions. We believe this is not necessarily the case. Firstly, humans are extremely skilled at detecting agents or gaze direction and inferring goals already from infancy, and are known to do so spontaneously (Farroni, et al., 2000; Gergely & Csibra, 2003; Gergely, Nádasdy, Csibra, & Bíró, 1995; Johnson, 2003; Kuhlmeier, Bloom, & Wynn, 2004; Movellan & Watson, 1987; Saxe, Tenenbaum, & Carey, 2005). This set of skills provides the foundation for applying the naïve attention model to specific interactions. Secondly, corresponding to the 1-system account's view (Carruthers, 2016, 2017) we argue that the implemented model that M forms about A's attentional priorities can be elaborated gradually over time and depending on the amount of allocated cognitive resources, enabling only implicit modulation of mental state inferences at the initial stages of computation. That is, the principles of the naïve model that guide the inferences and predictions could remain implicit while successfully leading to conclusions on what A thinks and intends to do. Finally, we propose that ToM draws on the capacity of event segmentation to lessen the demand of online computation by identifying the scope of validity of certain computed contents. The latter two arguments will be elaborated further in the following paragraphs.

4.2.1. Gradual enrichment. The level of elaboration that M's naïve attention model has reached through ontogenetic development (see 4.3.) sets the upper limit of how advanced

the implemented attention model centering on A can get *under optimal conditions*. Optimal conditions refer to social contexts where M has both the time and cognitive resources necessary for building the implemented model as well as the motivation to allocate those resources to computation. However, this is often not the case in real-time interactions. We argue that a rudimentary model of A's attention can be created as soon as the presence of an agent is detected and this implemented model is maintained and can be gradually enriched if further resources are allocated to it (Carruthers, 2017). The same naïve attention model can, thus output implemented models of different representational complexity, ranging from primary models that guide M's expectations about A in an implicit manner, to full-fledged models that can be subjected to symbolic manipulation and possibly even verbal report.

In what way might primary models differ from the elaborated versions? It can be that at the initial stages of processing, expectations about A's processing priorities (and the mental states which computation is adjusted to those priorities) are not directly linked to the agent whose presence and intentions set the model in action. There is empirical indication that at the earliest stages of mentalizing information from the self-perspective and information from others' perspectives is computed without indexing which belongs to whom (Kampis, Parise, Csibra, & Kovács, 2015), which requires an additional computational step (Qureshi, Apperly, & Samson, 2010). The belief of another agent may be just as easy to access during a self-perspective decision as one's own (Kovács, Téglás, & Endress, 2010) and people may even arrive to an explicit understanding about A's perspective quicker than about their own (Ramsey, Hansen, Apperly, & Samson, 2013; Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010). However, even such rudimentary outcome may serve behavior prediction under right conditions. For instance, it has been shown that when infants infer preference based on an agent's behavior, they do not necessarily link the preference to the specific agent but will still use the preference to form expectations about future actions

(Kampis, Somogyi, Itakura, & Király, 2013). That is, the inferential power of the computed mental state is retained even when it is not linked to a specific agent. This indicates that computed mental states can form the basis of expectations about their behavioral (and possibly mental) effects. The computation of mental states is triggered by a specific agent, but further inferences are not applied to that specific individual. Such a rudimentary understanding may lead to correct predictions if the identity of the agent, whose actions elicited intention and attention computation doesn't change suddenly (as in most everyday contexts).

In fact, M may even form a mental state prediction adjusted to A's attentional set without the mediation of a symbolic model of attention. As we will argue in section 4.3. in more depth, it is assumed that at the early stages of development, the presumptions of the naïve attention model are built into the organism's special purpose procedures in the form of inherent inferential principles and do not take a form that is accessible for symbolic manipulation. Some presumptions may remain to be implicit in adulthood as well. If the presumptions are implicit in the domain-specific inference mechanisms that are used to conclude what A thinks, then M can get from predicting what A intends, to what he most likely sees, without explicitly representing that he *attends* to goal-relevant information.

4.2.2. Event segmentation. Online predictions are most often embedded in an interaction that unfolds over time, which may allow M to use certain shortcuts. Specifically, it seems reasonable to assume that in the process of mental state computation, humans use a “compute once and update/discard if necessary” type of strategy, instead of disregarding their previously computed expectations and re-computing them each moment. Although information accessible in the physical world might be rapidly changing therefore creating the need to update A's inferred mental states quite often, there are nevertheless some aspects of the social context that can be expected to be *locally invariant*. If M can identify these locally

invariant features, she can save resources by simply keeping her expectations regarding them in place. The consequent questions are then: which expectations should be maintained over time, and on what basis should M determine when they eventually lose validity? We argue that humans' tendency to parse experience into meaningful events might provide the answer.

In naturalistic settings information flows in a continuous manner, and it is the perceiver herself that creates smaller units for perception. These units are parsed in a *meaningful* way, depending on the information present in the environment coupled with those available to the observer as background knowledge and is called event segmentation (Zacks & Swallow, 2007). Segmentation occurs spontaneously even during passive observation and results in the formation of hierarchically structured events, where the smallest meaningful units are considered parts of broader, coarser ones (Zacks, Tversky, & Iyer, 2001).

Some types of information that people are predisposed to use for segmenting social events have great relevance for ToM too. For instance, the identity of the agent rarely changes within what is perceived as an event (Boggia & Ristic, 2015), and Theory of Mind may build on such invariance. Drawing on a similar notion, Kovács (2016) proposes that the basic representational structure in ToM is the so-called belief file in which the agent and the content of the belief are stored in separate slots. This format enables the mindreader to update only the belief content without re-representing the invariant agent information, thereby gaining speed and flexibility in computation.

Goals, similarly to agents, denote the borders or broader, extended events (Zacks & Swallow, 2007). That is, one of the most important cues that humans rely on when parsing events, corresponds to the factor that establishes one's attentional set. As goals signal event borders, any inference made based on A's goal can be expected to remain valid throughout the event and lose effectiveness when the event comes to an end. We assume that ToM exploits this regularity and uses perceived event borders to determine the informational

validity of its expectations regarding A's attention. The naïve attention model generates an implemented model regarding A *for the length of the current event* which needs not be reconstructed continuously for it to continuously modulate subsequent mental state predictions.³

By using event borders to determine the applicability of her expectations about A, M can achieve fast predictions about A's epistemic states which are at the same time tailored to A's idiosyncratic ways, thereby leading to highly accurate and personalized inferences. We assume that event segmentation thus facilitates mentalizing as long as A's goals parse information to large enough segments. If A's goals change rapidly (or if there is discontinuity in A's presence), the event and the assumed invariance within is short lived. This creates the need to recruit the attention model repeatedly, and so the process potentially loses efficiency.

4.3. The acquisition of the naïve model

The naïve model of attention that the adult mindreader is proposed to be equipped with includes a wide range of presumably *learnt* regularities, and thus, we may expect it to be gradually refined through development, reaching its adult complexity and abstraction at a developmentally advanced stage. It is possible, for instance, that children first use the rules that they acquire in a deterministic manner, and only later acknowledge the probabilistic nature of the relation between attention and awareness, enabling them to have different levels of subjective confidence in their predictions.

³ In contrast, in most computerized experimental paradigms participants are presented with social events (trials) lasting a few seconds. Participants are expected to process every trial independently of the previous one, preventing them from utilizing event segmentation to ease the demands of social understanding.

Despite its prolonged development, the origins of the model may already be present in infancy. Carruthers (2016) for instance argues that humans may be born with a rudimentary concept of awareness. Children in their first year of life produce and understand behaviors that attempt to influence the direction of others' attention. Infants guide the attention of a partner by pointing at an object in order to obtain new information about it (Kovács, Tauzin, Téglás, Gergely, & Csibra, 2014). They also selectively address such information requests to individuals whom they see as knowledgeable sources in relation to the referent about which they want to learn (Begus & Southgate, 2012).

Although the role of visual cues (like pointing, or gazing) during development is substantial, infants can also use pragmatic inference to pinpoint the target of an interaction partner's attention. Reference disambiguation studies show that based on their knowledge about a model's prior experience with certain (but not other) objects, infants in their first and second year can infer which object the model must be attending (Akhtar, Carpenter, & Tomasello, 1996; Moll, Koring, Carpenter, & Tomasello, 2006; O'Neill, 1996). In fact, an understanding of others in terms of their attention might be essential for language development and has been proposed to be a precursor of Theory of Mind (Baron-Cohen, 1991; 1993). Importantly, Baron-Cohen proposes (1993) that this early understanding is based on infants' propensity to infer attention and an action's target object (goal) *from gaze direction*. Crucially, the AMT proposes that at later stages of development humans become capable of making inferences in the opposite direction as well: using *goals or intentions* to better predict attention, modulating their expectations formed based on gaze direction.

The ontogenetic development of the naïve attention model is, thus, envisioned as a gradual refinement and differentiation of concepts, and a continuous learning of presumptions that over time results in the adult-like capacity (Carruthers, 2016). In addition to the enrichment of the model's content, we speculate that the model gradually becomes integrated

with executive systems and linguistic faculties and accessible to genuine thought (i.e. explicit evaluation and usage) as development progresses (Carruthers, 2018). The presumptions of the naïve model start off as implicit regularities that, being built into the computational processes, inferential principles themselves, guide behavior without the mediation of a symbolic representational structure. Over time, some of the model's presumptions may evolve into knowledge that is available in a symbolic form for other processes, such as explicit mentalizing or common-sense reasoning. It is, however, conceivable that some presumptions remain to be implicit even in adulthood (evidenced, for instance, by the surprise that the phenomenon of inattentional blindness evoked amongst scientists).

To summarize, we take a one-system view on the naïve attention model, where presumptions are enriched in terms of content and representational complexity. Importantly, the naïve attention model is envisaged as a collection of interconnected presumptions rather than a single, uniform theory, and accordingly, the process of enrichment does not have to unfold at the same pace for each presumption of the model. We do not, however, rule out the possibility that humans' understanding of attention can be explained by positing multiple systems.⁴

⁴ Apperly and Butterfil's (2009) two-systems view posits that social skills are fragmented insomuch that the cognitively efficient, spontaneous forms of mindreading are encapsulated in a distinct system (system-1), including its dedicated processes and types of input. In ontogenetically and phylogenetically more advanced states, system-2 is also available, which is slower, requires deliberation but allows for representing representational relations. The two systems are proposed to function in parallel in adulthood without the possibility to interact with each other.

A subsequent question regarding the model's development concerns what types of experiences lead M to develop the attention model in the first place. Does the necessary input for acquiring the model come from observing others or from the child's experience about her *own* attentional processes? Although we do not wish to make a strong stand in this respect, it seems plausible to assume that children acquire the naïve model of attention through observing others, instead of themselves (Carruthers, 2009; Southgate, 2018). Southgate (2018) for instance argues that in the absence of a self-concept, infants develop an understanding of others' perspective developmentally earlier compared to when they obtain a self-perspective. To develop the naïve model based on their own attentional processes, children would have to treat their own perception as the object of their thought which is a metarepresentational skill that develops *after* infants start to demonstrate a primary understanding of others' attention (Gopnik & Wellman, 1992). Furthermore, discovering, for instance, the principle that gazing does not necessarily lead to seeing calls for noticing the *discrepancy* between the information present and the representation formed of it. Intuitively, this task seems easier as a third-party observer, not being influenced by the same perception goals as the agent (A) who is the target of naïve investigation.

In this paper, M is only discussed as an agent who is motivated to understand *others*, which, as outlined above, may be the developmentally earlier emerging function of the naïve model. However, we also assume that the same model should be applied to the self in those rarer, developmentally later occurring occasions when M focuses her mentalizing capacities towards herself.

4.4. Research guidelines and predictions of the naïve attention model theory

The AMT calls attention to the fact that certain research traditions themselves limit our understanding of humans' grasp of attention. Research often treats gaze information interchangeably with attention which assumption then constrains how complex humans'

understanding of attention can appear to be. In order to fully comprehend humans' understanding of the social significance of selective attention (potentially represented as an intentional relation towards the environment that influences the formation of epistemic states) we need to go beyond testing the effects of certain physical correlates, like gaze or head orientation. As selective attention is the result of *a particular intention*, empirical research needs to target mental states as indicators of attention.

Importantly, the visual complexity of the environment that A is part of (or the experimental stimuli mimicking that) will modulate the extent to which knowledge about A's intentions can enhance the acuity of mental state inferences on top of the information provided by his gaze. Top-down attention has traditionally been characterized as a process of selection amongst information that are in constant *competition* for processing (Desimone & Duncan, 1995; Lavie, 1995). When the complexity of the environment is extremely low such that there is practically no competition between information and the scene can be represented in its full informational complexity, the importance of M's access to A's perception goals decreases. Under such conditions, relying on the presumption that A's focus is external (see section 4.), M can reach relatively accurate predictions based on gaze direction only, although knowledge about intentions may still increase M's *confidence* in the inference. To summarize, due to the very nature of selective attention, simplistic experimental contexts will underestimate the significance of intentions as indicators of attention.

There is a large amount of evidence for humans' sensitivity to gaze. Experimental work is yet to disentangle whether these effects draw on an internal model of *attention*, as an intentional relation, or some limited scope theory about gazing. A testcase of this question is provided by covert attention, i.e. attention to non-fixated locations. Those models that assume that humans are informed about others' attention by their gaze direction only, rule out the possibility that covert attention could be grasped by human social understanding. On the other

hand, the naïve attention model proposed here provides the necessary cognitive basis for keeping track of covert attention by grasping attention at its origin (an intention) instead of its behavioral manifestation (gaze). Empirical support for this can be gathered by systematically manipulating the location of task-relevant information relative to the interactional partner's fixation.

Gaze and intention-based selective attention can be pitted against each other in other ways as well originating from the constraint that gaze-dependent visual field imposes on the perceiver. Specifically, some information may be relevant with respect to A's attentional set but still be outside of A's visual field (e.g. A collects blue dots amongst a set of black ones, some of which are in front of him, others behind). Although M should only attribute knowledge to A about those that he has visual access to, A's intention to perceive all blue dots may, however, endow social relevance and consequent processing on the ones he fails to see as well. How the hierarchical relation of gaze and intentions are represented, and the conditions under which one or the other may outweigh each other is yet to be explored.

A further prediction of the theory is that the naïve model of attention features attention as a domain general capacity. This is expected to allow humans to apply their assumptions about the role of attention in modulating the acquisition of mental states to perceptual modalities beyond vision, like hearing. Note also, that while the phenomenon of joint triadic attention is extensively researched, especially regarding ontogenetic development (e.g. Striano & Stahl, 2005) and deviations from its typical path (e.g. Leekam & Ramsden, 2006), the possibility of jointly attending auditory information is largely ignored.

Finally, the AMT assumes that ToM recruits the naïve attention model in order to enhance the accuracy of its mental state predictions, tailoring them to the temporary features of A's cognition. This will to some degree guide M's own attentional processes, cognitive priorities to align with A's – as she will be using her own limited amount of capacities to map

out A's epistemic states. However, for the sake of social success, A's attentional set should not be treated as M's own, neither should it be merged with M's. We predict that M does not only process more deeply the information selectively attended by A, she can also link that information to A (for evidence in this direction see Elekes & Sebanz, 2020). Relatedly, the theory also predicts that people presume that selective attention leads to knowledge about the subset of information that A had attended previously. Therefore, attention creates "selective knowledgeability", which effect should influence all behaviors sensitive to the social partner's knowledgeability, e.g. social learning in children (Sabbagh & Baldwin, 2001), or conformity in adults (Allan, Midjord, Martin, & Gabbert, 2012).

5. Contemporary approaches to social attention

Attention and human social cognition interweave through multiple routes. This multifaceted relationship has been vigorously explored during the last decades of social and cognitive psychology (Tomasello, 1995) including an accentuated interest as of late (e.g. Capozzi & Ristic, 2018; Milward & Carpenter, 2018; Siposova & Carpenter, 2019; Steinmetz & Pfattheicher, 2017). Importantly, attention can be social in several ways. Thus far, we have discussed research on the role of observed gaze in directing M's attention and the extent to which gaze can potentially contribute to understanding A' attention. We now turn to existing theoretical accounts on humans' ability to share attention with each other and point out how our theory relates to these.

Substantial theoretical interest has been devoted to how humans represent attention in others. Importantly, these accounts emphasize the advantages of experiencing that attention is *shared with someone else*. For instance, Clark and Marshall (1981) propose that for the success of linguistic discourses, dyads aim to achieve a common ground, the mutual acknowledgment that they share some knowledge. What is, in fact, part of the common ground can never be identified with certainty, but, they argue, if the partners have reason to

believe that both agents are attending to an object, and they assume that both infer knowledge from attendance, the assumption that there is shared knowledge can (relatively) safely be made. Others' attentive state is thus taken as an indication for her establishing an epistemic state on the attended information.

A similar phenomenon is reflected in the term joint attention (JA, Tomasello, 1995; Mundy & Newell, 2007). Joint attention refers to the ability to coordinate attention on some external object with a conspecific, accompanied by the mutual realization that they attend jointly. This entails the human specific ability to share an intentional relation with others towards the environment (Tomasello, 1995), and provides a cornerstone of human social learning and language acquisition specifically (Baldwin, 1995; Bruner, 1981; Carpenter & Tomasello, 2000; Morales, Mundy, Delgado, Yale, Messinger, Neal, & Schwartz, 2000). Our naïve attention model theory (AMT) shares important assumptions with Tomasello's theory of joint attention. First, on the grounds laid by Gibson and Rader (1979), both theories posit that attention (at least in its top-down manifestation) can be defined as the intentional act of perceiving, where the goals of the perceiver will influence the representation that arises through perception. Second, both theories emphasize that attention does not equal physically orienting to something; that fixating on certain stimuli is neither necessary nor sufficient to attend to it. Although the premises are similar, the scope of our theory, both in terms of the input and the output of the mechanisms, differs from that of JA in a way that addresses current trends and challenges in ToM research.

Specifically, the JA theory can only be applied to situations that involve mutuality between interaction partners. The cognitive and social benefits of joint attention require the phenomenological experience of not only aligning attention with others' but also sharing that intentional relation. The AMT claims that humans are generally predisposed to infer what others know, how they represent their environment, which aim is served by their ability to

monitor how a fellow individual's attentional set modulates their mental experiences. This is a more elementary process than engaging in joint attention, covering a wider range of social contexts (social at least in a unidirectional manner that one social agent represents something about another's mind). For the attention model to be recruited, M only needs to infer the current intentions of A. A does not even have to be aware of M's presence, let alone M's interest in his mental states, or to have a motivation to reciprocate that interest. In fact, a great chunk of our everyday lives composes of scenes that are social only in this unidirectional manner. The benefit of using the attention model is purely epistemic; it does not induce affiliative responses as experiencing JA does (Wolf, Launay, & Dunbar, 2016). The naïve attention model potentially contributes to the emergence of shared intentionality but is not sufficient in itself to establish it.

Finally, we will consider how our theory relates to and differs from the shared attention theory (SAT, see Shteynberg, 2010, 2015, 2018). The aim of the SAT is to give an account on what individual level cognitive mechanisms enable cultures to emerge and endure. The SAT proposes that the form of group-living that humans are characterized by rendered the existence of coordinated group knowledge adaptive. To cultivate such knowledge, humans learnt to prioritize information co-experienced, shared by group members. In a shared attention state, people experience attending to something from a first-person plural perspective (we), where the self and other are experienced as a unified agent with a single target of attention. This could be the process that enables the formation of a shared reality without engaging in any form of mentalization about the other's inner states.

The SAT account takes a very different standpoint from ours in its most fundamental claims. First, the SAT hypothesizes that people experience a sense of "we", and the information is cognitively prioritized because it is in the focus of "our" group-level attention. In our view, such a mechanism may serve the motivation to create the highest possible

amount of shared knowledge with the group but does not serve the epistemic goal of understanding and predicting other *individuals*. We argue that for their everyday social success humans need to keep track of *why* they prioritize certain bits of information: knowledge that is directly relevant to themselves needs to be kept separate from knowledge that gains relevance through being relevant to someone else. Thus, the mature mentalizing process (aided by the attention model) is headed towards attributing mental representations to *specific individuals*. As we have noted before, at the earliest stages of processing, self and other associated mental contents may not be linked to individuals yet. However, this does not mean that these contents are experienced in a unified perspective, as *ours*. Secondly, the SAT posits that the formation of a shared attention state does not require mentalizing and its effect on cognitive processing is independent of ToM. The AMT, on the other hand, argues that the mechanism is originated in basic forms of mentalizing (inferring goal based attentional sets) and may feed information forward to more complex forms of social reasoning. Finally, the SAT presumes that people only enter a shared attention state with similar others, with whom they share a group – which selectivity is needed to create and maintain stable group level knowledge. The AMT does not posit such a difference. As long as M is personally motivated to understand A, A's group membership or general similarity/dissimilarity to M does not make a difference. In fact, if M discovers that A belongs to another cultural group, and so they cannot assume to share much of their background knowledge, M's motivation to interpret A could even lead to *enhanced processing* regarding A's current knowledge states, including those modulated by his attentional set.

6. Concluding remarks

Although theory of mind and social attention have received extensive interest recently, the account we present here 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 harvested for social purposes. Amongst the different forms of background knowledge that M may consider about A, 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 (Cohen et al., 2012). 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.

In the current paper we outline the theory in its basics with the aim of launching a scientific dialogue and facilitating research. In our view, the significance of the AMT lies in its ability to generate novel, testable predictions that do not deduce from other models of social attention, and at the same time in its ability to accommodate existing empirical evidence as well.

Acknowledgements

We thank Krisztina Andrási, Bálint Forgács, Mikolaj Hernik, Katalin Oláh, Sunnae Kim and Tibor Tauzin for their valuable comments on the manuscript. Preparation of the manuscript was financially supported by Momentum (grant number: LP-2017-17/2017) and OTKA (grant number: K-116779) grants. This article benefited from a fellowship at the Paris Institute for Advanced Study (France), with the financial support of the French State, programme “Investissements d’avenir” managed by the Agence Nationale de la Recherche (ANR-11-LABX-0027-01 Labex RFIEA+).

References

- Akhtar, N., Carpenter, M., & Tomasello, M. (1996). The role of discourse novelty in early word learning. *Child Development*, 67(2), 635–645. doi:10.1111/j.1467-8624.1996.tb01756.x
- Allan, K., Midjord, J. P., Martin, D., & Gabbert, F. (2012). Memory conformity and the perceived accuracy of self versus other. *Memory & Cognition*, 40(2), 280-286.
<https://doi.org/10.3758/s13421-011-0141-9>
- Andersen, S. K., Hillyard, S. A., & Müller, M. M. (2008). Attention facilitates multiple stimulus features in parallel in human visual cortex. *Current Biology*, 18(13), 1006-1009.
<https://doi.org/10.1016/j.cub.2008.06.030>
- Apperly, I. A. & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological review*, 116(4), 953. <http://dx.doi.org/10.1037/a0016923>
- Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in cognitive sciences*, 16(8), 437-443.
<https://doi.org/10.1016/j.tics.2012.06.010>
- Baker, L. J., Levin, D. T., & Saylor, M. M. (2016). The extent of default visual perspective taking in complex layouts. *Journal of Experimental Psychology: Human Perception and Performance*, 42(4), 508. <http://dx.doi.org/10.1037/xhp0000164>
- Baldwin, D. A. (1995). Understanding the link between joint attention and language. In (Eds: Moore, C., Dunham, P. J.) *Joint attention: Its origins and role in development*, 131-158.
- Baron-Cohen, S. (1993). From attention-goal psychology to belief-desire psychology: the development of a theory of mind, and its dysfunction. In S. B. Cohen, H. T. Flusberg (Eds.), *Understanding Other Minds: Perspectives From Autism*. New York, NY, US: Oxford University Press.

- Baron-Cohen, S. (1991). Precursors to a theory of mind: Understanding attention in others. *Natural theories of mind: Evolution, development and simulation of everyday mindreading*, (233-251). Cambridge, MA, US: Basil Blackwell.
- Bayliss, A. P., Schuch, S., & Tipper, S. P. (2010). Gaze cueing elicited by emotional faces is influenced by affective context. *Visual Cognition*, 18(8), 1214-1232.
<https://doi.org/10.1080/13506285.2010.484657>
- Beck, A. A., Rossion, B., & Samson, D. (2017). An objective neural signature of rapid perspective taking. *Social cognitive and affective neuroscience*, 13(1), 72-79.
<https://doi.org/10.1093/scan/nsx135>
- Begus, K., & Southgate, V. (2012). Infant pointing serves an interrogative function. *Developmental science*, 15(5), 611-617. <https://doi.org/10.1111/j.1467-7687.2012.01160.x>
- Benedek, M., Daxberger, D., Annerer-Walcher, S., & Smallwood, J. (2018). Are you with me? Probing the human capacity to recognize external/internal attention in others' faces. *Visual Cognition*, 26(7), 511-517. <https://doi.org/10.1080/13506285.2018.1504845>
- Bentin, S., Kutas, M., & Hillyard, S. A. (1995). Semantic processing and memory for attended and unattended words in dichotic listening: Behavioral and electrophysiological evidence. *Journal of Experimental Psychology: Human Perception and Performance*, 21(1), 54. <http://dx.doi.org/10.1037/0096-1523.21.1.54>
- Boggia, J., & Ristic, J. (2015). Social event segmentation. *The Quarterly Journal of Experimental Psychology*, 68(4), 731-744. <https://doi.org/10.1080/17470218.2014.964738>
- Böckler, A., Knoblich, G., & Sebanz, N. (2012). Effects of a coactor's focus of attention on task performance. *Journal of Experimental Psychology: Human Perception and Performance*, 38(6), 1404. <http://dx.doi.org/10.1037/a0027523>

- Bruner, J. S. (1981). The social context of language acquisition. *Language & Communication*, 1(2-3), 155-178. [http://dx.doi.org/10.1016/0271-5309\(81\)90010-0](http://dx.doi.org/10.1016/0271-5309(81)90010-0)
- Cameron, E. L., Tai, J. C., & Carrasco, M. (2002). Covert attention affects the psychometric function of contrast sensitivity. *Vision research*, 42(8), 949-967. [https://doi.org/10.1016/S0042-6989\(02\)00039-1](https://doi.org/10.1016/S0042-6989(02)00039-1)
- Capozzi, F., & Ristic, J. (2018). How attention gates social interactions. *Annals of the New York Academy of Sciences*. <https://doi.org/10.1111/nyas.13854>
- Carpenter, M., & Tomasello, M. (2000). Joint attention, cultural learning, and language acquisition: Implications for children with autism. In A. M. Wetherby & B. M. Prizant (Eds.), *Communication and language intervention series; Vol. 9. Autism spectrum disorders: A transactional developmental perspective* (pp. 31-54). Baltimore, MD, US: Paul H Brookes Publishing.
- Carruthers, P. (2018). Young children flexibly attribute mental states to others. *Proceedings of the National Academy of Sciences*, 115(45), 11351-11353. <https://doi.org/10.1073/pnas.1816255115>
- Carruthers, P. (2017). Mindreading in adults: evaluating two-systems views. *Synthese*, 194(3), 673-688. <https://doi.org/10.1007/s11229-015-0792-3>
- Carruthers, P. (2016). Two systems for mindreading?. *Review of Philosophy and Psychology*, 7(1), 141-162. <https://doi.org/10.1007/s13164-015-0259-y>
- Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and brain sciences*, 32(02), 121-138. <https://doi.org/10.1017/S0140525X09000545>

- Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In, A.K. Joshi, B. L. Webber, and I.A. Sag (Eds.), *Elements of Discourse Understanding* (10-66). Cambridge: Cambridge University Press, 1981.
- Cohen, M. A., Cavanagh, P., Chun, M. M., & Nakayama, K. (2012). The attentional requirements of consciousness. *Trends in cognitive sciences*, 16(8), 411-417.
<http://dx.doi.org/10.1016/j.tics.2012.06.013>
- Cole, G. G., & Millett, A. C. (2019). The closing of the theory of mind: A critique of perspective-taking. *Psychonomic bulletin & review*, 26(6), 1787-1802.
<https://doi.org/10.3758/s13423-019-01657-y>
- Cole, G. G., Smith, D. T., & Atkinson, M. A. (2015). Mental state attribution and the gaze cueing effect. *Attention, Perception, & Psychophysics*, 77(4), 1105-1115.
10.3758/s13414-014-0780-6
- Colombatto, C., van Buren, B., & Scholl, B. J. (in press). Gazing without eyes: A ‘stare-in-the-crowd’ effect induced by simple geometric shapes. *Perception*.
- Colombatto, C., van Buren, B., & Scholl, B. J. (2019). Intentionally distracting: Working memory is disrupted by the perception of other agents attending to you—even without eye-gaze cues. *Psychonomic bulletin & review*, 26(3), 951-957.
<https://doi.org/10.3758/s13423-018-1530-x>
- Constable, M. D., Pratt, J., & Welsh, T. N. (2018). “Two Minds Don’t Blink Alike”: The Attentional Blink Does Not Occur in a Joint Context. *Frontiers in psychology*, 9.
10.3389/fpsyg.2018.01714
- Conty, L., Tijus, C., Hugueville, L., Coelho, E., & George, N. (2006). Searching for asymmetries in the detection of gaze contact versus averted gaze under different head

views: A behavioural study. *Spatial Vision*, 19, 529-545.

<https://doi.org/10.1163/156856806779194026>

Conway J., Catmur C., Bird G. (2018). Understanding individual differences in theory of mind via representation of minds, not mental states; Conway, Catmur and Bird. PsyArXiv. 10.17605/OSF.IO/DXPVQ

Corbetta, M., Miezin, F. M., Dobmeyer, S., Shulman, G. L., & Petersen, S. E. (1990). Attentional modulation of neural processing of shape, color, and velocity in humans. *Science*, 248(4962), 1556

Deaner, R. O., Shepherd, S. V., & Platt, M. L. (2007). Familiarity accentuates gaze cuing in women but not men. *Biology Letters*, 3(1), 65-68. 10.1098/rsbl.2006.0564

Dennett, D. C. (1981). True Believers: The Intentional Strategy and Why it Works. In A. F. Heath (Ed), *Scientific Explanation: Papers based on Herbert Spencer Lectures given in the University of Oxford* (pp. 53-75), Oxford, University Press.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1), 193-222. <https://doi.org/10.1146/annurev.ne.18.030195.001205>

Downing, P., Dodds, C., & Bray, D. (2004). Why does the gaze of others direct visual attention? *Visual Cognition*, 11(1), 71-79. <http://dx.doi.org/10.1080/13506280344000220>

Driver IV, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual cognition*, 6(5), 509-540. <http://dx.doi.org/10.1080/135062899394920>

Egeth, H. E., & Yantis, S. (1997). Visual attention: Control, representation, and time course. *Annual review of psychology*, 48(1), 269-297. 10.1146/annurev.psych.48.1.269

- 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(12), 2376-2389. <http://dx.doi.org/10.1080/17470218.2015.1120332>
- Elekes, F. & Sebanz, N. (2020). Effects of a partner's task on memory for content and source. *Cognition*, 198, 104221. <https://doi.org/10.1016/j.cognition.2020.104221>
- Elekes, F., Varga, M., Király, I. (2016). Evidence for spontaneous level-2 perspective taking in adults. *Consciousness and Cognition*, 41, 93-103. doi:10.1016/j.concog.2016.02.010
- Eskenazi, T., Doerrfeld, A., Logan, G. D., Knoblich, G., & Sebanz, N. (2013). Your words are my words: Effects of acting together on encoding. *The Quarterly Journal of Experimental Psychology*, 66(5), 1026-1034. <http://dx.doi.org/10.1080/17470218.2012.725058>
- Fabbri, M., Frisoni, M., Martoni, M., Tonetti, L., & Natale, V. (2018). Influence of time-of-day on joint Navon effect. *Cognitive processing*, 19(1), 27-40. <https://doi.org/10.1007/s10339-017-0849-y>
- Farroni, T., Johnson, M. H., Brockbank, M., & Simion, F. (2000). Infants' use of gaze direction to cue attention: The importance of perceived motion. *Visual cognition*, 7(6), 705-718. <https://doi.org/10.1080/13506280050144399>
- Fink, G. R., Dolan, R. J., Halligan, P. W., Marshall, J. C., & Frith, C. D. (1997). Space-based and object-based visual attention: shared and specific neural domains. *Brain: a journal of neurology*, 120(11), 2013-2028. <https://doi.org/10.1093/brain/120.11.2013>
- 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. <http://dx.doi.org/10.1037/0012-1649.17.1.99>

- Freundlieb, M., Kovács, Á. M., & Sebanz, N. (2018). Reading Your Mind While You Are Reading—Evidence for Spontaneous Visuospatial Perspective Taking During a Semantic Categorization Task. *Psychological science*, 29(4), 614-622.
<https://doi.org/10.1177/0956797617740973>
- Friesen, C. K., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic bulletin & review*, 5(3), 490-495.
<https://doi.org/10.3758/BF03208827>
- Frischen, A., Loach, D., & Tipper, S. P. (2009). Seeing the world through another person's eyes: Simulating selective attention via action observation. *Cognition*, 111(2), 212-218.
<https://doi.org/10.1016/j.cognition.2009.02.003>
- Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007). Auditory attention—focusing the searchlight on sound. *Current opinion in neurobiology*, 17(4), 437-455.
<https://doi.org/10.1016/j.conb.2007.07.011>
- Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in cognitive sciences*, 7(7), 287-292.
[https://doi.org/10.1016/S1364-6613\(03\)00128-1](https://doi.org/10.1016/S1364-6613(03)00128-1)
- Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, 56(2), 165-193. [https://doi.org/10.1016/0010-0277\(95\)00661-H](https://doi.org/10.1016/0010-0277(95)00661-H)
- Gibson, E., & Rader, N. (1979). Attention. The perceiver as performer. In G. A. Hale, M. Lewis (Eds.), *Attention and cognitive development* (pp. 1-21). Plenum Press, New York.
- Gobel, M. S., Tufft, M. R., & Richardson, D. C. (2018). Social beliefs and visual attention: how the social relevance of a cue influences spatial orienting. *Cognitive science*, 42, 161-185. <https://doi.org/10.1111/cogs.12529>

- Goodhew, S. C., & Plummer, A. S. (2019). Author accepted manuscript: Flexibility in resizing attentional breadth: Asymmetrical versus symmetrical attentional contraction and expansion costs depends on context. *Quarterly Journal of Experimental Psychology*, <https://doi.org/10.1177/1747021819846831>.
- Gopnik, A., Wellman, H.M. (1992). Why the child's theory of mind is really a theory? *Mind and Language* 7 (1-2):145-71. <https://doi.org/10.1111/j.1468-0017.1992.tb00202.x>
- Gregory, S. E. A., & Jackson, M. C. (2019). Barriers block the effect of joint attention on working memory: Perspective taking matters. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(5), 795-806. <http://dx.doi.org/10.1037/xlm0000622>
- Gregory, S. E., & Jackson, M. C. (2017). Joint attention enhances visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(2), 237. <http://dx.doi.org/10.1037/xlm0000294>
- Gregory, R. L. (1997). Knowledge in perception and illusion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 352(1358), 1121-1127. 10.1098/rstb.1997.0095
- 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. <https://doi.org/10.1111/desc.12445>
- Hansen, J. C., & Hillyard, S. A. (1980). Endogeneous brain potentials associated with selective auditory attention. *Electroencephalography and clinical neurophysiology*, 49(3-4), 277-290. [https://doi.org/10.1016/0013-4694\(80\)90222-9](https://doi.org/10.1016/0013-4694(80)90222-9)
- He, X., Lever, A. G., & Humphreys, G. W. (2011). Interpersonal memory-based guidance of attention is reduced for ingroup members. *Experimental brain research*, 211(3-4), 429-438. 10.1007/s00221-011-2698-8

- Helmholtz, H. von (1866). Concerning the perceptions in general. In *Treatise on physiological optics*, vol. III, 3rd edn (translated by J. P. C. Southall 1925 *Opt. Soc. Am.* Section 26, reprinted New York: Dover, 1962).
- Heyes, C. (2014). Submentalizing I am not really reading your mind. *Perspectives on Psychological Science*, 9(2), 131-143. 10.1177/1745691613518076
- Hietanen, J. K. (2002). Social attention orienting integrates visual information from head and body orientation. *Psychological Research*, 66(3), 174-179.
<https://doi.org/10.1007/s00426-002-0091-8>
- Hietanen, J. K., & Leppänen, J. M. (2003). Does facial expression affect attention orienting by gaze direction cues? *Journal of Experimental Psychology: Human Perception and Performance*, 29(6), 1228. 10.1037/0096-1523.29.6.1228
- Johnson, S. C. (2003). Detecting agents. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 358(1431), 549-559. 10.1098/rstb.2002.1237
- Kampis, D., Parise, E., Csibra, G., & Kovács, Á. M. (2015). Neural signatures for sustaining object representations attributed to others in preverbal human infants. *Proceedings of the Royal Society B*, 282(1819), 20151683. 10.1098/rspb.2015.1683
- Kampis, D., Somogyi, E., Itakura, S., & Király, I. (2013). Do infants bind mental states to agents? *Cognition*, 129(2), 232-240. <https://doi.org/10.1016/j.cognition.2013.07.004>
- Kecskes, I., & Zhang, F. (2009). Activating, seeking, and creating common ground: A socio-cognitive approach. *Pragmatics & Cognition*, 17(2), 331-355.
<https://doi.org/10.1075/pc.17.2.06kec>
- Kingstone, A., Kachkovski, G., Vasilyev, D., Kuk, M., & Welsh, T. N. (2019). Mental attribution is not sufficient or necessary to trigger attentional orienting to gaze. *Cognition*, 189, 35-40. <https://doi.org/10.1016/j.cognition.2019.03.010>

- 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. <https://doi.org/10.1073/pnas.1803505115>
- Kobayashi, H., & Kohshima, S. (1997). Unique morphology of the human eye. *Nature*, 387(6635), 767. <https://doi.org/10.1038/42842>
- Koch, C., & Tsuchiya, N. (2007). Attention and consciousness: two distinct brain processes. *Trends in cognitive sciences*, 11(1), 16-22. <http://dx.doi.org/10.1016/j.tics.2006.10.012>
- Kovács, Á. M. (2016). Belief files in theory of mind reasoning. *Review of Philosophy and Psychology*, 7(2), 509-527. <https://doi.org/10.1007/s13164-015-0236-5>
- 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. [10.1126/science.1190792](https://doi.org/10.1126/science.1190792)
- Kovács, Á. M., Tauzin, T., Téglás, E., Gergely, G., & Csibra, G. (2014). Pointing as epistemic request: 12-month-olds point to receive new information. *Infancy*, 19(6), 543-557. <https://doi.org/10.1111/infa.12060>
- Kuhlmeier, V. A., Bloom, P., & Wynn, K. (2004). Do 5-month-old infants see humans as material objects? *Cognition*, 94(1), 95-103. <https://doi.org/10.1016/j.cognition.2004.02.007>
- Lamme, V. A. (2003). Why visual attention and awareness are different. *Trends in cognitive sciences*, 7(1), 12-18. [http://dx.doi.org/10.1016/S1364-6613\(02\)00013-X](http://dx.doi.org/10.1016/S1364-6613(02)00013-X)
- Langton, S. R., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in cognitive sciences*, 4(2), 50-59. [https://doi.org/10.1016/S1364-6613\(99\)01436-9](https://doi.org/10.1016/S1364-6613(99)01436-9)

- Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human perception and performance*, 21(3), 451.
<https://doi.org/10.1037/0096-1523.21.3.451>
- Leekam, S. R., & Ramsden, C. A. (2006). Dyadic orienting and joint attention in preschool children with autism. *Journal of autism and developmental disorders*, 36(2), 185.
<https://doi.org/10.1007/s10803-005-0054-1>
- Liu, T., Stevens, S. T., & Carrasco, M. (2007). Comparing the time course and efficacy of spatial and feature-based attention. *Vision research*, 47(1), 108-113.
<https://doi.org/10.1016/j.visres.2006.09.017>
- Low, J., Apperly, I. A., Butterfill, S. A., & Rakoczy, H. (2016). Cognitive architecture of belief reasoning in children and adults: A primer on the two-systems account. *Child Development Perspectives*, 10(3), 184-189. <https://doi.org/10.1111/cdep.12183>
- Luo, Y., & Beck, W. (2010). Do you see what I see? Infants' reasoning about others' incomplete perceptions. *Developmental Science*, 13(1), 134-142.
<https://doi.org/10.1111/j.1467-7687.2009.00863.x>
- Mack, A. (2003). Inattention blindness looking without seeing. *Current Directions in Psychological Science*, 12(5), 180-184. [10.1016/j.visres.2010.02.024](https://doi.org/10.1016/j.visres.2010.02.024)
- Mack, A., & Rock, I. (1998). Inattention blindness: Perception without attention. In (Wright, R.D) *Visual attention*, 55-76.
- Marotta, A., Lupiáñez, J., Martella, D., & Casagrande, M. (2012). Eye gaze versus arrows as spatial cues: Two qualitatively different modes of attentional selection. *Journal of experimental psychology: human perception and performance*, 38(2), 326.
<http://dx.doi.org/10.1037/a0023959>

- Maunsell, J. H., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in neurosciences*, 29(6), 317-322. <https://doi.org/10.1016/j.tins.2006.04.001>
- Milward, S. J., & Carpenter, M. (2018). Joint action and joint attention: Drawing parallels between the literatures. *Social and Personality Psychology Compass*, 12(4), e12377. <https://doi.org/10.1111/spc3.12377>
- Moll, H., Koring, C., Carpenter, M., & Tomasello, M. (2006). Infants determine others' focus of attention by pragmatics and exclusion. *Journal of Cognition and Development*, 7(3), 411-430. http://dx.doi.org/10.1207/s15327647jcd0703_9
- Morales, M., Mundy, P., Delgado, C. E., Yale, M., Messinger, D., Neal, R., & Schwartz, H. K. (2000). Responding to joint attention across the 6-through 24-month age period and early language acquisition. *Journal of applied developmental psychology*, 21(3), 283-298. [https://doi.org/10.1016/S0193-3973\(99\)00040-4](https://doi.org/10.1016/S0193-3973(99)00040-4)
- Movellan, J. R., & Watson, J. S. (1987). Perception of directional attention. In *Infant behavior and development: Abstracts of the sixth international conference of infant studies*.
- Mundy, P., & Newell, L. (2007). Attention, joint attention, and social cognition. *Current directions in psychological science*, 16(5), 269-274. [10.1111/j.1467-8721.2007.00518.x](https://doi.org/10.1111/j.1467-8721.2007.00518.x)
- Neisser, U., & Becklen, R. (1975). Selective looking: Attending to visually specified events. *Cognitive psychology*, 7(4), 480-494. [https://doi.org/10.1016/0010-0285\(75\)90019-5](https://doi.org/10.1016/0010-0285(75)90019-5)
- Nie, Q. Y., Ding, X., Chen, J., & Conci, M. (2018). Social attention directs working memory maintenance. *Cognition*, 171, 85-94. <https://doi.org/10.1016/j.cognition.2017.10.025>
- Norman, J. (2002). Two visual systems and two theories of perception: An attempt to reconcile the constructivist and ecological approaches. *Behavioral and brain sciences*, 25(1), 73-96. <https://doi.org/10.1017/S0140525X0200002X>

- Nuku, P., & Bekkering, H. (2008). Joint attention: Inferring what others perceive (and don't perceive). *Consciousness and Cognition*, 17(1), 339-349.
<http://dx.doi.org/10.1016/j.concog.2007.06.014>
- O'Craven, K. M., Downing, P. E., & Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature*, 401(6753), 584. <https://doi.org/10.1038/44134>
- Oláh, K., Elekes, F., & Király, I. (2019). Creating a shared representational space. PsyArxiv. 10.31234/osf.io/g2z74
- 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/10.1111/j.1467-8624.1996.tb01758.x>
- Pavan, G., Dalmasso, M., Galfano, G., & Castelli, L. (2011). Racial group membership is associated to gaze-mediated orienting in Italy. *PLoS One*, 6(10), e25608.
<https://doi.org/10.1371/journal.pone.0025608>
- Phillips, B. (2019). The evolution and development of visual perspective taking. *Mind & Language*, 34(2), 183-204. <https://doi.org/10.1111/mila.12214>.
- Posner, M. I. (1980). Orienting of attention. *Quarterly journal of experimental psychology*, 32(1), 3-25. <http://dx.doi.org/10.1080/00335558008248231>
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(04), 515-526.
<https://doi.org/10.1017/S0140525X00076512>
- Qureshi, A. W., Apperly, I. A., & Samson, D. (2010). Executive function is necessary for perspective selection, not Level-1 visual perspective calculation: Evidence from a dual-task study of adults. *Cognition*, 117(2), 230-236.
<http://dx.doi.org/10.1016/j.cognition.2010.08.003>

- Ramsey, R., Hansen, P., Apperly, I., & Samson, D. (2013). Seeing it my way or your way: frontoparietal brain areas sustain viewpoint-independent perspective selection processes. *Journal of Cognitive Neuroscience*, 25(5), 670-684.
https://doi.org/10.1162/jocn_a_00345
- Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink?. *Journal of experimental psychology: Human perception and performance*, 18(3), 849.
- Ristic, J., & Kingstone, A. (2005). Taking control of reflexive social attention. *Cognition*, 94(3), B55-B65. <https://doi.org/10.1016/j.cognition.2004.04.005>
- Sabbagh, M. A., & Baldwin, D. A. (2001). Learning words from knowledgeable versus ignorant speakers: Links between preschoolers' theory of mind and semantic development. *Child development*, 72(4), 1054-1070. <https://doi.org/10.1111/1467-8624.00334>
- Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, 36(5), 1255. <http://dx.doi.org/10.1037/a0018729>
- Saxe, R., Tenenbaum, J. B., & Carey, S. (2005). Secret agents: Inferences about hidden causes by 10-and 12-month-old infants. *Psychological Science*, 16(12), 995-1001.
<https://doi.org/10.1111/j.1467-9280.2005.01649.x>
- Schuller, A. M., & Rossion, B. (2001). Spatial attention triggered by eye gaze increases and speeds up early visual activity. *Neuroreport*, 12(11), 2381-2386.

- Schneider, D., Slaughter, V. P., Becker, S. I., & Dux, P. E. (2014). Implicit false-belief processing in the human brain. *NeuroImage*, *101*, 268-275.
<https://doi.org/10.1016/j.neuroimage.2014.07.014>
- Serences, J. T., Schwarzbach, J., Courtney, S. M., Golay, X., & Yantis, S. (2004). Control of object-based attention in human cortex. *Cerebral Cortex*, *14*(12), 1346-1357.
<https://doi.org/10.1093/cercor/bhh095>
- Serences, J. T., & Yantis, S. (2006). Selective visual attention and perceptual coherence. *Trends in cognitive sciences*, *10*(1), 38-45. <http://dx.doi.org/10.1016/j.tics.2005.11.008>
- Shteynberg, G. (2018). A collective perspective: shared attention and the mind. *Current Opinion in Psychology*, *23*, 93-97. <https://doi.org/10.1016/j.copsyc.2017.12.007>
- Shteynberg, G. (2015). Shared attention. *Perspectives on Psychological Science*, *10*(5), 579-590. <https://doi.org/10.1177/1745691615589104>
- Shteynberg, G. (2010). A silent emergence of culture: the social tuning effect. *Journal of personality and social psychology*, *99*(4), 683. <http://dx.doi.org/10.1037/a0019573>
- Siposova, B., & Carpenter, M. (2019). A new look at joint attention and common knowledge. *Cognition*, *189*, 260-274. <https://doi.org/10.1016/j.cognition.2019.03.019>
- Southgate, V. (2018). The puzzle of early mentalizing: A unitary theory of infants' successes and preschoolers' failures. PsyArXiv. 10.31234/osf.io/y2ghq
- Spaulding, S. (2018). Do you see what I see? How social differences influence mindreading. *Synthese*, *195*, 4009-4030. <https://doi.org/10.1007/s11229-017-1404-1>
- Spence, C. (2002). Multisensory attention and tactile information-processing. *Behavioural brain research*, *135*(1-2), 57-64. [https://doi.org/10.1016/S0166-4328\(02\)00155-9](https://doi.org/10.1016/S0166-4328(02)00155-9)

- Steinmetz, J., & Pfattheicher, S. (2017). Beyond social facilitation: A review of the far-reaching effects of social attention. *Social Cognition*, 35(5), 585-599.
<https://doi.org/10.1521/soco.2017.35.5.585>
- Störmer, V. S., & Alvarez, G. A. (2014). Feature-based attention elicits surround suppression in feature space. *Current Biology*, 24(17), 1985-1988.
<https://doi.org/10.1016/j.cub.2014.07.030>
- Striano, T., & Stahl, D. (2005). Sensitivity to triadic attention in early infancy. *Developmental Science*, 8(4), 333-343. <https://doi.org/10.1111/j.1467-7687.2005.00421.x>
- Surtees, A., Apperly, I., & Samson, D. (2016). I've got your number: Spontaneous perspective-taking in an interactive task. *Cognition*, 150, 43-52.
<http://dx.doi.org/10.1016/j.cognition.2016.01.014>
- 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>
- Terrizzi, B. F., & Beier, J. S. (2016). Automatic cueing of covert spatial attention by a novel agent in preschoolers and adults. *Cognitive Development*, 40, 111-119.
<https://doi.org/10.1016/j.cogdev.2016.08.001>
- Teufel, C., Alexis, D. M., Clayton, N. S., & Davis, G. (2010). Mental-state attribution drives rapid, reflexive gaze following. *Attention, Perception, & Psychophysics*, 72(3), 695-705.
10.3758/APP.72.3.695
- Todd, A. R., Simpson, A. J., & Cameron, C. D. (2019). Time pressure disrupts level-2, but not level-1, visual perspective calculation: A process-dissociation analysis. *Cognition*, 189, 41-54. <https://doi.org/10.1016/j.cognition.2019.03.002>

- Tomasello, M. (1995). Joint attention as social cognition. In C. Moore, P. J. Dunham (Eds.), *Joint attention: Its origins and role in development* (103-130), Psychology Press, Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.
- Tomasello, M., Hare, B., Lehmann, H., & Call, J. (2007). Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. *Journal of human evolution*, 52(3), 314-320. <https://doi.org/10.1016/j.jhevol.2006.10.001>
- Wagner, U., Giesen, A., Knausenberger, J., & Echterhoff, G. (2017). The Joint Action Effect on Memory as a Social Phenomenon: The Role of Cued Attention and Psychological Distance. *Frontiers in psychology*, 8, 1697. <https://doi.org/10.3389/fpsyg.2017.01697>
- Wang, J. J., & Apperly, I. A. (2017). Just one look: Direct gaze briefly disrupts visual working memory. *Psychonomic bulletin & review*, 24(2), 393-399. <https://doi.org/10.3758/s13423-016-1097-3>
- Westra, E. (2017). Stereotypes, theory of mind, and the action–prediction hierarchy. *Synthese*, 1-26. <https://doi.org/10.1007/s11229-017-1575-9>
- Wiese, E., Wykowska, A., Zwickel, J., & Müller, H. J. (2012). I see what you mean: how attentional selection is shaped by ascribing intentions to others. *PloS one*, 7(9), e45391. <https://doi.org/10.1371/journal.pone.0045391>
- Wilson, C. J., Soranzo, A., & Bertamini, M. (2017). Attentional interference is modulated by salience not sentience. *Acta psychologica*, 178, 56-65. <https://doi.org/10.1016/j.actpsy.2017.05.010>
- Wilson, D., & Sperber, D. (2002). Truthfulness and relevance. *Mind*, 111(443), 583-632. <https://doi.org/10.1093/mind/111.443.583>

Wolf, W., Launay, J., & Dunbar, R. I. (2016). Joint attention, shared goals, and social bonding. *British Journal of Psychology*, 107(2), 322-337.

<https://doi.org/10.1111/bjop.12144>

Zacks, J. M., & Swallow, K. M. (2007). Event segmentation. *Current directions in psychological science*, 16(2), 80-84. <https://doi.org/10.1111/j.1467-8721.2007.00480.x>

Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of experimental psychology: General*, 130(1), 29.

<http://dx.doi.org/10.1037/0096-3445.130.1.29>