Psychonomic Bulletin & Review
https://doi.org/10.3758/s13423-020-01766-z

THEORETICAL REVIEW

Coordinating attention requires coordinated senses

Lucas Battich [1,2] · Merle Fairhurst [1,3,4] · Ophelia Deroy [1,3,5]

Abstract

From playing basketball to ordering at a food counter, we frequently and effortlessly coordinate our attention with others towards a common focus: we look at the ball, or point at a piece of cake. This non-verbal coordination of attention plays a fundamental role in our social lives: it ensures that we refer to the same object, develop a shared language, understand each other's mental states, and coordinate our actions. Models of joint attention generally attribute this accomplishment to gaze coordination. But are visual attentional mechanisms sufficient to achieve joint attention, in all cases? Besides cases where visual information is missing, we show how combining it with other senses can be helpful, and even necessary to certain uses of joint attention. We explain the two ways in which non-visual cues contribute to joint attention: either as enhancers, when they complement gaze and pointing gestures in order to coordinate joint attention on visible objects, or as modality pointers, when joint attention needs to be shifted away from the whole object to one of its properties, say weight or texture. This multisensory approach to joint attention has important implications for social robotics, clinical diagnostics, pedagogy and theoretical debates on the construction of a shared world.

Keywords Joint attention · Social cognition · Cross-modal attention · Multisensory perception

There is more to joint attention than meets the eye

Infant and caregiver coordinate their attention on a toy while learning its name; jazz musicians jointly attend to the music they play together, and hunters can jointly track the smell or sounds of prey in the forest. The ability to coordinate our perception on a shared object of interest comes to most of us between the ages of 9 and 18 months. In our everyday life, we continue to rely on this non-verbal skill, otherwise known as joint attention, to communicate, share experiences, and coordinate with others.

Joint attention has been proposed as one of the essential ingredients of social skills in humans (Adamson, Bakeman, Suma, & Robins, 2019; Carpenter, Nagell, Tomasello, Butterworth, & Moore, 1998; Eilan, Hoerl, McCormack, & Roessler, 2005; Moore & Dunham, 1995; Seemann, 2011; Tomasello & Farrar, 1986) and, arguably, across other animal species (Ben Mocha, Mundry, & Pika, 2019; Leavens & Racine, 2009). In most of these accounts, joint attention is measured through the capacity to follow gaze and pointing gestures and coordinate on visible targets (Mundy & Newell, 2007). But does coordinating on visible objects only depend on vision? And what happens when we need to coordinate, not on visible targets, but on auditory, tactile, or multisensory ones?

Uncontroversially, shouting or touching someone's shoulder can be useful to make someone pay attention or orient in the right direction. The role of auditory or tactile alerting signals as accessory cues is well established in primate (Liebal, Waller, Burrows, & Slocombe, 2014) and non-primate (Ben Mocha et al., 2019; Bro-Jørgensen, 2010; Rowe, 1999) animal multimodal communication. It is similarly uncontroversial that non-visual senses often act as a *background* or mere

✉ Lucas Battich
lucas.battich@campus.lmu.de

[1] Faculty of Philosophy and Philosophy of Science, Ludwig Maximilian University Munich, Geschwister-Scholl-Platz 1, Munich 80359, Germany

[2] Graduate School of Systemic Neurosciences, Ludwig Maximilian University Munich, Munich, Germany

[3] Munich Center for Neuroscience, Ludwig Maximilian University Munich, Munich, Germany

[4] Institut für Psychologie, Fakultät für Humanwissenschaften, Universität der Bundeswehr München, Munich, Germany

[5] Institute of Philosophy, School of Advanced Study, University of London, London, UK

Published online: 14 July 2020

enabling condition for visual attention (for instance, by using vestibular and proprioceptive cues to determine the spatial orientation of one's body in the world, and orient visual attention accordingly). Existing work in the domain of joint attention would certainly accept that other sensory modalities are involved or that joint attention occurs in multisensory settings. Highlighting that joint attention is fundamentally a multisensory phenomenon, however, stresses that non-visual senses are not merely accessories to what could otherwise be defined as a visual phenomenon. Our goal is to provide a more systematic representation of how non-visual sensory resources contribute to joint attention. More specifically, we argue that non-visual senses play two crucial roles. First, they interact closely with gaze and pointing gestures to prime or *enhance* the coordination of visual attention. Non-visual senses can certainly act as distractors, having a negative impact on joint attention. In most cases, however, and with the exception of rare clinical or artificial cases, which we discuss below, other senses are at least minimally involved in the success of joint attention. Second, they play a *necessary* role when it comes to extending social coordination to non-visual and amodal properties of objects and events in the world.

Consider what would happen if gaze and pointing were indeed all there was to the coordination of attention: without computing information from multiple senses, either serially or in conjunction, our referential intentions would run a much higher risk of remaining ambiguous (see *Non-visual senses enhance visual joint attention*). We could not coordinate on non-visible and more abstract aspects of the world (see *Non-visual senses are necessary to extend joint attention*). The current multisensory account is better than a strictly visual one when it comes to explaining how joint attention establishes a socially shared world, where mind-independent objects can be attended in common (see *Theoretical implications: Sharing more than a visual world*). It also has implications for clinical settings and social robotics which are currently focused on gaze-following: with our new account, deficits in gaze coordination could potentially be compensated for by non-visual modalities, and social robots could coordinate attention with humans even without fine-grained gaze-following capacities (see *Applications: Multisensory strategies for the clinic, the school and social robotics*).

Visual joint attention

When Jerome Bruner and colleagues introduced the term *joint attention* to the research on the ontogeny of communication (Bruner, 1974; Scaife & Bruner, 1975), they referred to infants' developing capacity to share their experiences about objects and events with others, and learn word meanings. Now, the construct is used to explain many aspects of our social activities: joint attention in infancy predicts future social competence (Mundy & Sigman, 2015) and emotion regulation, and may reinforce executive functions (Morales, Mundy, Crowson, Neal, & Delgado, 2005; Swingler, Perry, & Calkins, 2015). For adults, engaging in joint attention modulates multiple cognitive abilities (Shteynberg, 2015), including working memory (Gregory & Jackson, 2017; Kim & Mundy, 2012), mental spatial rotation (Böckler, Knoblich, & Sebanz, 2011), and affective appraisals to objects in the environment (Bayliss, Paul, Cannon, & Tipper, 2006).

Bruner's pioneering work centered on joint *visual* attention (Scaife & Bruner, 1975). By and large, subsequent research has remained exclusively focused on the visual domain. Gaze behavior can be easily measured and controlled in laboratory conditions and is therefore a powerful means to study joint attention. In arguing for a multisensory approach, we do not aim to diminish the important role played by gaze cues. Decades of research on gaze following and gaze alternation have firmly established their importance in development and cognition (Flom, Lee, & Muir, 2017; Frischen, Bayliss, & Tipper, 2007; Schilbach, 2015; Shepherd, 2010), and have provided a solid basis for the study of joint attention.

Research into the early development of joint attention distinguishes between *responding* to joint attention by following the direction of others' attention, and *initiating* joint attention by directing or leading the attention of others to a third object or event (Mundy & Newell, 2007). Responding to joint attention, sometimes considered equivalent to following someone's perceptual cues, is the most studied form of joint attention (Fig. 1a) (Mundy, 2018; but see, e.g., Bayliss et al., 2013; Stephenson, Edwards, Howard, & Bayliss 2018). Whether following social cues for attention differs from following non-social cues like arrows remains a topic of debate and investigation, but uncontroversially engages spatial skills and perceptual gaze processing (Gregory, Hermens, Facey, & Hodgson, 2016; Hermens, 2017; Langton, Watt, & Bruce, 2000; Mundy, 2018; Shepherd, 2010). Senses other than vision can play an instrumental role alongside gaze and pointing gestures to guide spatial attention to visible objects.

Attention following, however, is often not sufficient for joint attention. For example, I can follow your attention without you noticing in any way that I did so, which would not count as joint attention. In addition to gaze following, joint attention requires the ability to engage in a *reciprocal coordination* that guarantees we are looking at the same object together (Mundy, 2018; Siposova & Carpenter, 2019) (Fig. 1b). This triadic coordination exhibits the understanding, even minimally, that both agents are mutually aiming at or aware of the object (Bakeman & Adamson, 1984; Mundy, 2016; Tomasello, 1995). Non-visual senses here may do more than facilitate attention following: they help to strategically select the appropriate target of joint attention between two individuals.

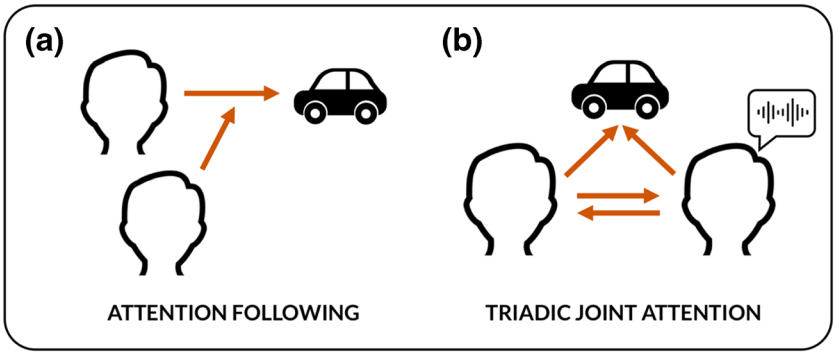

Fig. 1 Following attention is different from coordinating attention. (**a**) Attention following is characterized by the unilateral response of one individual. It can consist of behaviors such as gaze following, or the monitoring of others' bodily posture and gestures, and responding to vocal and haptic cues. Attention following is a pre-condition for full joint attention, and occurs earlier in development. (**b**) Coordination of attention is characterized by the reciprocal interaction between individuals toward a third object. In addition to gaze following, joint attention includes gaze-alternation and directing other's gaze through pointing — but also other senses

Engaging in joint attention requires one to know what one is attending to, as well as what the other is attending to. This in turn requires the combined processing of three types of information: (1) information about one's own attentional state, including interoceptive and proprioceptive information (Mundy & Jarrold, 2010); (2) information about the other's attentional state; (3) information about the target of joint attention (Mundy, 2018; Siposova & Carpenter, 2019). All three types of information and their processing can engage multiple senses, besides vision. Information about my own attention to the object of common reference may include whether I am actively handling the object, or merely looking at it. Information about the other's attentional state will vary depending on whether they have access to the same sensory information I have. The strategies used to establish joint attention will vary when we coordinate on a smell, a sound, the color of an object, or a whole, complex multisensory event.

Non-visual senses enhance visual joint attention

Visual cues provide multisensory expectations

When processing information about the other's attentional state, we can further distinguish between the sense I rely on to monitor the other's attention (e.g., I *gaze* at your hand grasping), and the sense they use, which I monitor to gather information about their attention (e.g., I gaze at your *hand grasping*). This distinction already pleads for the incorporation of richer sensory measures in models of joint attention than mutual eye contact, gaze following or gaze alternation. Observing someone's touching actions, as well as someone being touched, activates similar neural circuits normally involved in the execution of those actions, and the processing of actual touch (Buccino et al., 2001; Keysers et al., 2004), suggesting that tactile expectations regarding the jointly attended object can be gathered vicariously even by sight alone. Studies have here looked at the use of coupled information from eye and hand gestures. When reaching and manipulating objects, gaze and hand movements are systematically coordinated with respect to the target object, with gaze fixation leading the subsequent hand movement (Horstmann & Hoffmann, 2005; Pelz, Hayhoe, & Loeber, 2001). This eye-hand coupling can provide a path for well-coordinated rapid and successful joint attentional interaction: although gaze provides a faster cue to the spatial area where the target is located, the hand trajectory while reaching and grasping provides a slower but more spatially precise and stable cue to the target's location (Yu & Smith, 2013). Additionally, in following a grasping gesture, observers are sensitive to both the direction and the grip aperture size of the reaching hand to facilitate target detection (Tschentscher & Fischer, 2008). Reliance on multiple senses and their interaction may here help provide richer spatial and temporal representations of our environment (Keetels & Vroomen, 2012; Stoep, Postma, & Nijboer, 2017). These multisensory strategies are present during infant–caregiver joint attentional engagement, which reflects the multisensory nature of parent–infant dyadic communication (Gogate, Bahrick, & Watson, 2000; Gogate, Bolzani, & Betancourt, 2006; Hyde, Flom, & Porter, 2016). Multimodal behaviors help sustain joint attention between parents and infants from 12 to 16 months old, in particular when parents express some interest in an object looking at, talking about, and touching the jointly attended object (Suarez-Rivera, Smith, & Yu, 2019). One-year-old infants do not tend to follow the partner's gaze to monitor their attention while playing together with a toy. Instead, they follow their hands (Yu & Smith, 2013). Taken together, this evidence suggests that non-visual senses and multisensory expectations are exploited in joint attention, especially to narrow down the spatial location of the target of joint attention through spatial redundancy.

Recent research on the emergence of pointing gestures reinforces this suggestion. Children interpret pointing gestures

as if they were attempts to touch things (O'Madagain, Kachel, & Strickland, 2019), indicating that understanding visual cues about someone's touch toward a third object are ontogenetically prior to the understanding referential pointing gestures. This recent work suggests new methods to explore whether a similar relation is present in the phylogeny of grasping and pointing cues.

Non-visual cues enhance visual target detection

Joint attention can be established through gaze alone (Flom et al., 2017). In many social contexts, the use of visual cues can be sufficient to coordinate attention, but may not always be the most *efficient*. In information theory, adding redundancy to the initial message so that several portions of the message carry the same information increases the chance that the message is accurately received at the end of a noisy channel (Shannon, 1948). This is also true in perception. For an everyday illustration, consider trying to hit a nail with a hammer. It is possible to push the pointy part of the nail in the wall and then hammer it while relying only on vision, but by holding the nail with one hand, you can gather information about the nail's spatial position both through vision and through your hand position. Studies in multisensory perception demonstrate that redundant information delivered across several sensory modalities increases the reliability of a sensory estimate: it enhances a perceiver's accuracy and response time to detect the presence of a stimulus and to discriminate and identify a sensory feature (e.g., an object's shape or its spatial location), a so-called *redundant-signals effect* (Ernst & Banks, 2002; Miller, 1982). It is safe to assume that redundancy of information across modalities is also usefully exploited when establishing and sustaining joint attention. For example, the caregiver will point to a toy car that the infant can see, and tap on the toy to make a noise. Here, the combination of the visual and auditory information enhances the infant's accuracy and speed in shifting spatial attention (cf. Partan & Marler, 1999) (see Fig. 2a). In this section we review how multisensory information facilitates visual coordination and target detection, focusing on three mechanisms: spatial congruency, temporal synchrony, and cross-modal correspondences.

Redundancy of spatial information is shown to help with the orienting of visual attention in experiments where individual perceivers are presented with a task-irrelevant cue on the same or opposite side of the subsequent visual target. Participants tend to respond more rapidly, and more correctly, to visual targets appearing at the same location as the former task-irrelevant cue, rather than on the opposite side. This works for visual irrelevant cues (Posner, 1980; see Carrasco, 2011; Wright & Ward, 2008, for overviews) and also occurs across modalities: participants are faster and more accurate at detecting target stimuli in one modality when a task-irrelevant cue is presented in the same or similar location (McDonald, Teder-Sälejärvi, & Hillyard, 2000; Spence, McDonald, & Driver 2004a; see Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010, for a review). This evidence suggests that when participants direct their spatial attention to a certain location driven by one modality, their sensitivity to stimuli in that location is also enhanced for other modalities. While these traditional cross-modal attention studies use nonsocial stimuli, there is growing evidence of similar effects with social ones. Gaze-cueing experiments using covert orienting paradigms have shown that cues from another's gaze behavior facilitate the processing of tactile stimuli at the body location corresponding to the other's gaze direction (Soto-Faraco, Sinnett, Alsius, & Kingstone, 2005). Recent work shows that gaze-based cues enhance the processing of tactile (De Jong & Dijkerman, 2019) and auditory (Nuku & Bekkering, 2010) stimuli at what is meant to be the jointly attended location. The current evidence of cross-modal effects in spatial attention gives us reason to think that a wide array of sensory cues, besides someone's gaze or gesture direction, can be exploited to assist spatial coordination between joint attenders.

Temporal synchrony between cross-modal cues, in the absence of spatial congruence, also directs someone's spatial attention. Van der Burg et al. (2008, 2009, 2010) have shown that the presentation of a spatially irrelevant cue in the auditory or tactile modality can facilitate a participant's visual search performance in an environment with color-changing elements, when the non-visual cue is presented at the same time as a color change in the target element. Known as the "pip-and-pop effect," these studies show that even when one sensory cue does not carry relevant spatial information, it can enhance the salience of a spatially relevant cue in a different modality (Ngo & Spence 2010). These cross-modal effects could be exploited in trying to establish joint attention to a target in a changing, dynamic environment. Touching someone's shoulder or vocalizing in synchrony with a certain movement or event (e.g., every time a particular bird jumps from a branch or flutters its wings) may be a better strategy to coordinate attention to it than pointing alone (Fig. 2B).

Finally, *the properties of the non-visual social cues* can also shape congruency effects, besides providing spatial or temporal congruence with visual cues. We are not talking here of semantic congruence (saying "dog" or "woof" while pointing at the visible dog) but of sensory congruence between properties such as pitch or loudness, and visual properties, such as brightness, shape, etc. Humans, like some other animals (Bee, Perrill, & Owen, 2000), exploit the environmental regularities that exist between sensory cues across modalities for communicative purposes. Such regularities show up in cross-modal correspondences, i.e. robust associations between independent features or dimensions across modalities (Spence, 2011; Spence & Deroy, 2013). For example, high-pitched sounds correspond to high spatial positions of a visual stimulus, so that when both features are congruently matched,

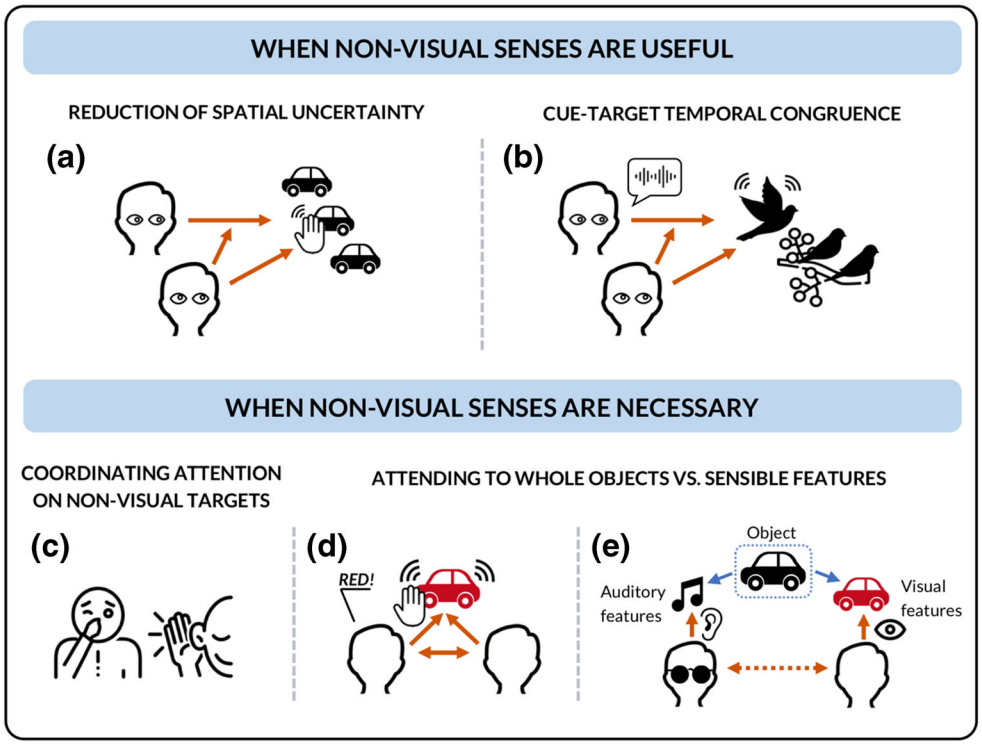

Fig. 2 (**Upper panel**) Non-visual cues can complement visual cues in joint attention. (**a**) Redundant information delivered across modalities can increase accuracy and speed in following spatial cues: by monitoring someone's eye-gaze cues in combination to their hand-grasping actions, the follower's response in localizing the object of joint attention is enhanced. (**b**) Using temporal congruence between a cue and a target in different modalities to facilitate someone's orienting to the correct visual target. (**Lower panel**) Non-visual cues are often necessary for joint attention. (**c**) Establishing joint attention toward a non-visual target by using ostensive visual cues: ostensive pointing at the relevant sensory organ (touching one's ear or one's nose) can provide evidence to another agent of the intention of attending to a non-visual stimulus (a sound, a smell). Such strategies rely on cognitive abilities to infer that the target is non-visual. (**d**) Exploiting temporal synchrony: a parent shakes an object in a temporally synchronous manner congruent with their uttering the word "red." While the visual stimulus and the auditory stimulus have different causal sources (the toy and the parent), the information is conveyed that the word "red" is associated with a visual property of the toy. (**e**) Coordinating on objects we each experience through different modalities: each subject must process information about each other's modal access relative to the target to successfully achieve coordination

attentional orienting to a target visual cue is facilitated (Bernstein & Edelstein, 1971). Other cross-modal correspondences, such as the one that exists between pitch and brightness, work together with temporal synchrony to elicit a "pip-and-pop effect" during visual search: when a visual target changes brightness, a congruent change in pitch of a task-irrelevant auditory cue enhances correct target detection (Klapetek, Ngo, & Spence 2012). The effects of cross-modal correspondence have so far been mostly studied in nonsocial domains. We suggest that they are also relevant in social domains. For example, when trying to direct your attention to an animal hiding in the trees, emitting a high-pitched rather than a low-pitched interjection might help direct attention to the higher part of the scene. To test this suggestion, future work on multisensory joint attention will have to address the role of cross-modal alerting signals, and how the processing of cross-modal social signals compares to nonsocial situations.

Importantly, how much spatial, temporal, and cross-modal congruence facilitate the processing of visual gaze or pointing gestures is ripe for more precise measurements, notably by artificially manipulating the discrepancy between the cues, and measuring the subsequent effects on joint attention.

The interplay between coordinated attention and multisensory processing

Multisensory cues can help the social coordination of attention. Surprisingly, the reverse can also be true. A few innovative studies give evidence that coordinating attention with a partner modulates a participant's multisensory processing. People are better able to ignore task-irrelevant stimuli in a distracting modality when they know that someone else is attending to these distractors (Heed, Habets, Sebanz, & Knoblich, 2010; Wahn, Keshava, Sinnett, Kingstone, & König, 2017).

In the first study (Heed et al., 2010), participants had to judge whether a tactile stimulus was presented on the upper or lower part of a cube, while a distractor visual stimulus was presented synchronously at the same or opposite elevation. In the individual task, participants responded faster and more

accurately when the distractor stimulus was presented at the same elevation as the tactile target, showing a performance difference known as the cross-modal congruency effect (CCE; see Spence, Pavani, Maravita, & Holmes, 2004b, for a review). Interestingly, the CCE was significantly reduced when a partner was instructed to attend to the visual stimuli, indicating that participants could better ignore incongruent distractors when their partner responded on them. This effect was recently replicated in an audiovisual congruency task (Wahn et al., 2017) involving visual flashes and auditory tones originating from the same or opposite spatial vertical location. Knowing that someone else was attending to the incongruent flashes allowed participants to respond faster to the tones, resulting in a reduced CCE.

These studies show that responding jointly reduces the interference of competing stimuli in a multisensory setting (Wahn & König, 2017). The results seem at odds with a recent tradition of research showing that acting jointly increases the interference of irrelevant stimuli, presumably due participants co-representing each other's tasks besides their own (Sebanz, Knoblich, & Prinz, 2003, 2005). For example, performing an object-based visual attention task jointly impairs performance (Böckler, Knoblich, & Sebanz, 2012), and the increase in interference of irrelevant information is well documented in Go/No Go joint Simon tasks (Dolk et al., 2014). The difference between the reduction and the increase of irrelevant interference in different joint attentional tasks may be due to the nature of the tasks studied. An efficient division of labor can be allowed when the different target stimuli of each co-actor's task are presented concurrently, whereas the beneficial effect of filtering irrelevant information disappears when the task involves two competing Go/No Go actions (Dolk & Liepelt, 2018; Sellaro, Treccani, & Cubelli, 2018).

So far, studies have focused on coordinated social attention to separate cross-modal targets. Each participant attends and responds to a different modal stimulus, which facilitates a perceptual division of labor. A multisensory approach to joint attention should encourage us to extend this work to situations where partners attend and respond to the same multisensory stimuli, or try and ignore distractors in the same modality while focusing on another one. For example, when two subjects jointly coordinate their attention toward sounds and flashes presented closely in space and time, the binding of two or more modal features may be further enhanced, compared to conditions where subjects attend to the same sounds and flashes alone. If both are asked to attend jointly to the sounds, and jointly ignore the flashes, they may also be less prone to a ventriloquist effect, where the location of the sounds is displaced toward the location of the flashes (Vroomen & De Gelder, 2004).

Non-visual senses are necessary to extend joint attention

Jointly attending to invisible sounds or smells

The dominance of vision in the study of, and theorizing about, perception and joint attention may reflect the importance of this modality in humans (Colavita, 1974; Emery, 2000; Itier & Batty, 2009; Sinnett, Spence, & Soto-Faraco, 2007), but should not occult the fact that humans also jointly attend and teach words referring to sounds and smells, not to mention musical features.

Establishing joint attention toward a non-visual target requires access to information about both the other's attentional focus and, crucially, the target where the other's attention should be directed. Relative to gaze, a clear limitation of audition and olfaction is that their target of attention is not publicly disclosed to an observer. To establish joint attention coordination on strictly non-visual targets, subjects may be obliged to *indirectly* coordinate on the visual location of these non-visual events and use cognitive strategies to signal and to infer that the target is non-visual. For example, ostensive pointing at the relevant sensory organ (touching one's ear, or one's nose) can provide evidence to another agent of the intention of attending to a non-visual stimulus (Baker & Hacker, 2005) (Fig. 2c).

In addition, ostensive strategies could involve *negative* cues such as standing still, and keeping one's head and eyes motionless to signal that attention should be directed to a non-visual target of joint attention. Here, one prediction would be that such cases would occur only *after* expectations about pointing and gaze have been fully formed – as the strategy rests on using a mismatch between the expectation (that eyes and heads move) and the results (eyes and heads do not move, meaning that the target is non-visible).

Although visual and gestural ostensive cues may be used on some occasions to direct attention to a non-visible target, such behaviors already presuppose that the other agent is capable of understanding that sounds and smells are objects in the world that can be perceived together with others. The developmental onset of the ability to gaze at objects jointly with others is well researched. One outstanding question is when infants start to display an equivalent understanding that others can share with them attention to smells and sounds, and how this understanding is coupled with processing the visual attention of others.

Jointly attending to amodal features

Gaze-based joint attention enhances basic object recognition, even in very young infants (Cleveland & Striano, 2007; Hoehl, Wahl, Michel, & Striano, 2012; Wahl, Marinović, & Träuble, 2019). However, object-recognition development

relies on the ability to perceive global, invariant, and amodal properties like spatial location, tempo, rhythm, and intensity, which can only be conveyed through the combination of different sense modalities (Bahrick & Lickliter, 2014; Hyde et al., 2016). The redundancy introduced by multisensory events can thus be strategically used to establish joint attention on amodal features of objects and events. Bahrick and colleagues suggest that perception of this amodal information is critically important for the development and performance of perceptual object and event recognition (Bahrick, 2010).

One key example of such strategic use is the manner in which the temporal synchrony (when onset, offset, and/or duration of sensory stimuli are the same) between vision and audition can be exploited. For instance, a parent will shake an object in a temporally congruent way with the word they utter, thus enhancing the associating between object and word (Fig. 2d) (Gogate et al., 2000; Gogate & Hollich, 2016; Jesse & Johnson, 2016). The significance of temporal and spatial synchrony across different sensory cues is not only restricted to language learning. Running a toy car over the table or over the infant's arm while saying "vroom" may not directly lead towards word acquisition, as there is no linguistic element to be acquired. But it may help to bind both visual (e.g., shape) and auditory (e.g., vehicle noises) properties to the same object, the toy car.

The use of two cues highlights an important point. Here the target of joint attention is broader than the cues used to attract and coordinate attention: making a sound while moving a toy-car and looking at it ostensibly can be used to draw attention to the whole multisensory object, including its amodal extension, its weight, texture, etc., and not just its auditory or visual properties.

Conversely, the target of joint attention can be narrower than the object of individual attention and even of mutually shared experiences. For example, while musicians may attend to how others move their bow, hands, and heads, their joint attention is focused on the music they produce or, indeed, an element of the music (a particular voice or a particular theme). Moreover, their auditory joint attention will be coordinated through the gestures of a musical conductor, which provide visual cues about particular aspects in the sounds that musicians must follow – the music's tempo, for example. In this sense, the target of coordinated attention is narrower than the visual and auditory cues they use to attract and maintain their attention and narrower than the multisensory production that they know they are mutually experiencing.

Taking into account the role of non-visual senses in coordinating attention highlights that the *target of joint attention* can often be different than *the target of each individual's attention*. Joint attention involves more than merely orienting toward the same target. Perceptual attention can be characterized as the selective information processing of a specific area or features of the sensory world, while ignoring or decreasing processing of other areas and features (Eriksen & James, 1986; Klein & Lawrence, 2012). Joint attention results in a socially mediated enhancement in the processing of sensory information (Mundy, 2018). In other words, joint attention brings about another level of selectivity over an individual's own perceptual attention. Engaging in joint attention allows us to extract from a fundamentally multisensory experience the relevant integrated targets or specific features (visual, auditory, etc.) for further information processing and social coordination.

Sensory deficits: Jointly attending to a multisensory object through different senses

What happens when coordination occurs on objects that the two agents experience through different modalities? This is the case when coordinating attention with blind individuals, or individuals whose vision is temporarily blocked (say, they wear opaque glasses). Here, both or at least one agent knows that the other cannot access the object on which attention needs to be coordinated via the visual modality that they themselves use to access the object.

Cases of sensory deprivation (e.g., deafness, blindness, anosmia, hyposmia) provide methodological tools to study the roles of different senses during joint attention, and how individuals with limited sensory access negotiate coordination. Atypical development highlights the manner in which we share attention with others as a function of information access. In a case study of two congenital blind infants, coordinating attention with their caregivers involved auditory information as well as tactile and kinesthetic information, memory, sound changes, air currents, and echolocation (Bigelow, 2003). Deaf-blind children tend to combine two or more sensory sources for coordinating attention toward an object with their non-deaf-blind parents (Núñez, 2014). A 3-year-old child with profound visual and hearing impairment would first draw on touch to check that she has her caregiver's attention. She would then hold the object of interest towards the caregiver's face with one hand while continuing to monitor their attention with the other hand, vocalizing excitedly and smiling throughout (Núñez, 2014). Social gaze behavior and joint attention through vision alone can also be impacted by auditory deficits (e.g., Corina & Singleton, 2009; Lieberman, Hatrak, & Mayberry, 2014). There is evidence, for example, that auditory deprivation affects the effect of gaze cues and gaze following. Deaf children (aged between 7 and 14 years old) are more susceptible to the influence of task-irrelevant gaze cues than hearing children (Pavani, Venturini, Baruffaldi, Caselli, & van Zoest, 2019). This effect appears to dissipate in deaf adults, suggesting that the salience of social gaze cues changes during development (Heimler et al., 2015)

These studies reinforce the view that our ability to establish the triadic relation characteristic of joint attention can vary

according to the modal pathways used for directing and following the other's attention (Fig. 2e). In multisensory contexts, agents can share across information to which the other person has no access, or is not actively accessing. To illustrate, suppose we are jointly attending to a coffee cup by vision. In addition, I am also touching the object to judge its temperature. Through our coordinated attention to the cup and by monitoring my responses, you can vicariously gather information on my haptic experience and whether the cup is warm.

Theoretical implications: Sharing more than a visual world

Philosophers and psychologists have taken the role of joint attention in our understanding of other minds to argue that joint attention is, in fact, essential to understand the concept of a shared objective world, where mind-independent objects are attended in common (Davidson, 1999; Eilan, 2005; Engelland, 2014; Seemann, 2019; Tomasello, 2014). The ability to coordinate attention to an object together with another individual goes hand in hand with the ability to experience the object as a mind-independent entity separate from oneself (Campbell, 2011). This view has pre-eminent precursors in psychology. Lev Vygotsky (2012), in particular, held the doctrine that all higher cognition in an individual arises from an internationalization process of prior social interactions. Vygotsky's original formulation may seem overly strong, but a Vygotskyan approach has become increasingly influential to account for the social influences observed in the development of cognition and psychiatric disorders (Bolis & Schilbach, 2018; Fernyhough, 2008; Hobson & Hobson, 2011; Tomasello, 2019). Granting that joint attention helps us build a shared objective world, restricting ourselves to gaze and vision alone would make this world incredibly impoverished.

To stress this point, imagine a case where joint attention would *only* occur through gaze-following and looking at pointing gestures: we would only be able to coordinate attention on the visual properties of objects and events. We would certainly be able to learn that most bananas are yellow; we would learn that using color-tinged glasses changes how these properties look; and we would learn that other people may be seeing a drawing upside down when we see it right side up. But how would two people jointly attend to the sound of thunder, or the smell of natural gas? Would they quickly make the difference between pointing at the color of the car, or the car as a whole?

Realizing that we attend to a unitary object or to specific properties cannot occur in a visual-only scenario, or certainly without resorting to more conventional or linguistic means. Using a multisensory combination of cues is necessary to explain that we share an objective world of multisensory objects, sounds, smells, and textures.

Applications: Multisensory strategies for the clinic, the school, and social robotics

A better understanding of the mechanisms through which multisensory and cross-modal processes help and shape the successful coordination of attention on the same object, or on a given aspect of an object, can have direct implications for several sectors and fields.

When gaze coordination is limited

In a caregiver-child pair in which one person has a sensory deficit (deaf-blind, deaf, blind), the information that can be shared will be limited in some way, and compensated for in others. Tactile joint attention is crucial for children with visual impairments and multiple sensory disabilities (Chen & Downing, 2006). A child rolling Play-Doh will lead the adult's hand to share attention to her activity. The adult can follow the child's lead and focus on what the child is doing by keeping non-controlling tactile contact both with the child's hands and with the Play-Doh, establishing a reciprocal relation.

An emphasis on gaze interaction, however, can lead to biased assessments of an individual's ability to coordinate and interact with others. When measured according to vision-based operationalizations, deaf children of hearing parents show a delay in the onset of *visual* joint attentional skills, and symbol-infused joint attention (involving words or symbolic gestures) tends to be less frequent than in typically developing infants (Prezbindowski, Adamson, & Lederberg, 1998). These results have been challenged when factoring the role of other senses: hearing parents do accommodate their deaf children's hearing status by engaging them via multiple modalities, while parents of typically developing children tend to use alternating unimodal (either visual or auditory) cues during a joint attention episode (Depowski, Abaya, Oghalai, & Bortfeld, 2015). Developmental differences are not pronounced in deaf children of deaf parents, who tend to coordinate attention using both visual and tactile signals (Spencer, 2000).

Taken together, these findings suggest that operationalizations of joint attention based on gaze alone may produce unreliable measures of the real ability of infants to coordinate attention with others. They also show that non-visual senses impinge on the development of joint attention, even for non-visually impaired deaf individuals. Finally, the ability to engage in joint attention depends not just on the atypical infant's behavior, but, importantly, on that of their caregivers. Adopting a multisensory perspective on joint attention can provide better measures of the development of atypical children and inspire new complementary strategies to foster the development of joint attention skills.

Multisensory joint attention during learning

The ostensive character of joint attention is central to the acquisition of language (Adamson et al., 2019; Carpenter et al., 1998; Tomasello & Farrar, 1986) and, more generally to the transmission of knowledge and learning (Csibra & Gergely, 2009). In traditional paradigms on the role of joint attention in language development, triadic coordination to a target object is visually established through gaze alternation or pointing, accompanied by the utterance of the linguistic label to be associated with the object (see Akhtar & Gernsbacher, 2007, for a critical overview). As noted above, however, early linguistic development is increasingly recognized as a multisensory process (Gogate & Hollich, 2016; Jesse & Johnson, 2016). Similarly, the importance of multisensory teaching methods is increasingly recognized within pedagogy, both for typically developing children (e.g., Kirkham, Rea, Osborne, White, & Mareschal, 2019; Shams & Seitz, 2008; Volpe & Gori 2019) and for children with learning differences, including dyslexia (e.g., Birsh, 2005) and autistic spectrum disorder (e.g., Mason, Goldstein, & Schwade, 2019).

A better understanding of the interplay of different sense modalities during joint attention, across different ages and neurological conditions, can support the development of multisensory protocols in pedagogical situations. It should also be a reminder of cross-cultural differences when generalizing about teaching: in some cultures, touch, sounds, or smells are more central to social engagement, learning, or communication (Akhtar & Gernsbacher, 2008; Kinard & Watson, 2015). Akhtar and Gernsbacher (2008) review evidence suggesting that in cultures where infants experience continuous physical or vocal contact with their caregivers, and spend less time in face-to-face eye contact, evidence of social engagement will rely on tactile, auditory, and olfactory cues more than mutual gaze cues. Mothers in Kenya, for example, engage in more touching and holding with their infants, and less in eye contact, than mothers in the USA (Richman, Miller, & LeVine, 1992).

Multisensory joint attention with artificial social agents

The field of social robotics strives to bring artificial agents into hospitals, schools, businesses, and homes – complex social environments that require the enactment of naturalistic nonverbal interactions, including joint attention coordination (Clabaugh & Matarić, 2018; Kaplan & Hafner, 2006; Yang et al., 2018). For a robot to help a human partner assemble a piece of furniture, stack blocks with children in the playground, and assist people with disabilities in their daily lives, they need to be sensitive to what the human is attending to, and asking them to attend to.

Whether an artificial agent can successfully engage in joint attention with humans will depend on how well they can meet the behavioral expectations of their human interaction partner. Will they be able to both initiate and follow attentional cues in a naturalistic manner (Pfeiffer-Leßmann, Pfeiffer, & Wachsmuth, 2012)? One current approach is to enable social robots to mimic human gaze behaviors (Admoni & Scassellati, 2017; Kompatsiari, Ciardo, Tikhanoff, Metta, & Wykowska, 2019). However, while human participants do respond to the gaze of artificial agents (Willemse, Marchesi, & Wykowska, 2018), they are also highly sensitive to momentary multimodal behaviors produced by their artificial partner (Yu, Schermerhorn, & Scheutz, 2012). By adopting a multisensory perspective on human-robot joint attention, it is possible to examine non-visual cues emitted by the artificial agent, so that they accord with the expectations of human interaction partners. Being sensitive to the non-visual cues emitted by humans could also improve the spatial and temporal resolution of attention-orienting in robots.

Conclusion

Any episode of visual attention will, de facto, rely on background multisensory processing: we rely on proprioceptive and vestibular cues to visually orient our attention and ourselves in the world. Multisensory interactions, however, play a more substantial role in the coordination of attention across social agents: infants and adults recruit multiple sense modalities to initiate and follow someone's attention to a specific object or location in space. These interactions can be distinguished depending on whether they facilitate the coordination of visual attention, or whether they extend the coordination to non-visual and amodal properties. While non-visual modalities are useful complements for vision in the former case, they are essential in the latter case: some kinds of joint attention are necessarily multisensory, and could not be carried by vision alone.

This multisensory approach has implications for behavioral and developmental models of joint attention. Just as selective attention can be described as a cognitive capacity that both influences and is influenced by perceptual processes across different modalities, models of joint attention must be flexible enough to incorporate how it relies on dynamic information from multiple senses. It also has practical implications to overcome clinical deficits in joint attention, augment its pedagogical role, and address the challenge of coordinating attention between humans and social robots.

Author Note MF was supported in part by funds from LMU Munich's Institutional Strategy LMUexcellent within the framework of the German Excellence Initiative. OD is supported by the NOMIS foundation "Dise" grant.

Funding Information Open Access funding provided by Projekt DEAL.

Compliance with ethical standards

Conflicts of interest The authors have no conflicts of interest to disclose.

References

Adamson, L. B., Bakeman, R., Suma, K., & Robins, D. L. (2019). An expanded view of joint attention: Skill, engagement, and language in typical development and autism. *Child Development*, *90*(1), e1–e18. https://doi.org/10.1111/cdev.12973

Admoni, H., & Scassellati, B. (2017). Social eye gaze in human-robot interaction: A review. *Journal of Human-Robot Interaction*, *6*(1), 25–63. https://doi.org/10.5898/JHRI.6.1.Admoni

Akhtar, N., & Gernsbacher, M. A. (2007). Joint attention and vocabulary development: A critical look. *Language and Linguistics Compass, 1*(3), 195–207. https://doi.org/10.1111/j.1749-818X.2007.00014.x

Akhtar, N., & Gernsbacher, M. A. (2008). On privileging the role of gaze in infant social cognition. *Child Development Perspectives*, *2*(2), 59–65. https://doi.org/10.1111/j.1750-8606.2008.00044.x

Bahrick, L. E. (2010). Intermodal perception and selective attention to intersensory redundancy: Implications for typical social development and autism. In *The Wiley-Blackwell handbook of infant development* (pp. 120–166). https://doi.org/10.1002/9781444327564.ch4

Bahrick, L. E., & Lickliter, R. (2014). Learning to attend selectively: The dual role of intersensory redundancy. *Current Directions in Psychological Science, 23*(6), 414–420. https://doi.org/10.1177/0963721414549187

Bakeman, R., & Adamson, L. B. (1984). Coordinating attention to people and objects in mother-infant and peer-infant interaction. *Child Development*, *55*(4), 1278–1289.

Baker, G. P., & Hacker, P. M. S. (2005). Ostensive definition and its ramifications. In *Wittgenstein: Understanding and meaning. Part i: Essays* (pp. 81–106). https://doi.org/10.1002/9780470752807.ch5

Bayliss, A. P., Paul, M. A., Cannon, P. R., & Tipper, S. P. (2006). Gaze cuing and affective judgments of objects: I like what you look at. *Psychonomic Bulletin & Review*, *13*(6), 1061–1066. https://doi.org/10.3758/BF03213926

Bayliss, A. P., Murphy, E., Naughtin, C. K., Kritikos, A., Schilbach, L., & Becker, S. I. (2013). "Gaze leading": Initiating simulated joint attention influences eye movements and choice behavior. *Journal of Experimental Psychology: General, 142*(1), 76–92. https://doi.org/10.1037/a0029286

Bee, M. A., Perrill, S. A., & Owen, P. C. (2000). Male green frogs lower the pitch of acoustic signals in defense of territories: A possible dishonest signal of size? *Behavioral Ecology*, *11*(2), 169–177. https://doi.org/10.1093/beheco/11.2.169

Ben Mocha, Y., Mundry, R., & Pika, S. (2019). Joint attention skills in wild Arabian babblers (Turdoides squamiceps): A consequence of cooperative breeding? *Proceedings of the Royal Society B: Biological Sciences*, *286*(1900), 20190147. https://doi.org/10.1098/rspb.2019.0147

Bernstein, I. H., & Edelstein, B. A. (1971). Effects of some variations in auditory input upon visual choice reaction time. *Journal of Experimental Psychology*, *87*(2), 241–247. https://doi.org/10.1037/h0030524

Bigelow, A. E. (2003). The development of joint attention in blind infants. *Development and Psychopathology*, *15*(2), 259–275.

Birsh, J. R. (2005). *Multisensory teaching of basic language skills*. Baltimore: Paul Brookes Publishing Co.

Böckler, A., Knoblich, G., & Sebanz, N. (2011). Giving a helping hand: effects of joint attention on mental rotation of body parts. *Experimental Brain Research*, *211*(3-4), 531–545. https://doi.org/10.1007/s00221-011-2625-z

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), 1404–1415. https://doi.org/10.1037/a0027523

Bolis, D., & Schilbach, L. (2018). 'I interact therefore I am': The self as a historical product of dialectical attunement. *Topoi*, 1–14. https://doi.org/10.1007/s11245-018-9574-0

Bro-Jørgensen, J. (2010). Dynamics of multiple signalling systems: animal communication in a world in flux. *Trends in Ecology & Evolution*, *25*(5), 292–300. https://doi.org/10.1016/J.TREE.2009.11.003

Bruner, J. S. (1974). From communication to language: A psychological perspective. *Cognition*, *3*(3), 255–287. https://doi.org/10.1016/0010-0277(74)90012-2

Buccino, G., Binkofski, F., Fink, G.R., Fadiga, L., Fogassi, L., Gallese, V., Seitz, R.J., Zilles, K., Rizzolatti, G. & Freund, H.-J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. *European Journal of Neuroscience, 13*(2), 400–404. https://doi.org/10.1111/j.1460-9568.2001.01385.x

Campbell, J. (2011). An object-dependent perspective on joint attention. In A. Seemann (Ed.), *Joint attention: New developments in psychology, philosophy of mind, and social neuroscience* (pp. 415–430). Cambridge, MA: MIT Press.

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, *63*(4), 1–174. https://doi.org/10.2307/1166214

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484–1525. https://doi.org/10.1016/j.visres.2011.04.012

Chen, D., & Downing, J. E. (2006). *Tactile strategies for children who have visual impairments and multiple disabilities: Promoting communication and learning skills*. New York, NY: AFB Press.

Clabaugh, C., & Matarić, M. (2018). Robots for the people, by the people: Personalizing human-machine interaction. *Science Robotics*, *3*(21), eaat7451. https://doi.org/10.1126/scirobotics.aat7451

Cleveland, A., & Striano, T. (2007). The effects of joint attention on object processing in 4- and 9-month-old infants. *Infant Behavior and Development*, *30*(3), 499–504. https://doi.org/10.1016/J.INFBEH.2006.10.009

Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics*, *16*(2), 409–412. https://doi.org/10.3758/BF03203962

Corina, D., & Singleton, J. (2009). Developmental social cognitive neuroscience: Insights from deafness. *Child Development, 80*(4), 952–967. https://doi.org/10.1111/j.1467-8624.2009.01310.x

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences, 13*(4), 148–153. https://doi.org/10.1016/j.tics.2009.01.005

Davidson, D. (1999). The emergence of thought. *Erkenntnis*, *51*(1), 511–521. https://doi.org/10.1023/A:1005564223855

De Jong, M. C., & Dijkerman, H. C. (2019). The influence of joint attention and partner trustworthiness on cross-modal sensory cueing. *Cortex*, *119*, 1–11. https://doi.org/10.1016/j.cortex.2019.04.005

Depowski, N., Abaya, H., Oghalai, J., & Bortfeld, H. (2015). Modality use in joint attention between hearing parents and deaf children. *Frontiers in Psychology*, *6*, 1556. https://doi.org/10.3389/fpsyg.2015.01556

Dolk, T., & Liepelt, R. (2018). The multimodal go-nogo Simon effect: Signifying the relevance of stimulus features in the go-nogo Simon paradigm impacts event representations and task performance. *Frontiers in Psychology*, *9*, 2011. https://doi.org/10.3389/fpsyg.2018.02011

Dolk, T., Hommel, B., Colzato, L. S., Schütz-Bosbach, S., Prinz, W., & Liepelt, R. (2014). The joint Simon effect: A review and theoretical integration. *Frontiers in Psychology*, *5*, 974. https://doi.org/10.3389/fpsyg.2014.00974

Eilan, N. (2005). Joint attention, communication, and mind. In N. Elian, C. Hoerl, T. McCormack, & J. Roessler (Eds.), *Joint attention: Communication and other minds. Issues in philosophy and psychology* (pp. 1–33). Oxford: Oxford University Press.

Eilan, N., Hoerl, C., McCormack, T., & Roessler, J. (Eds.). (2005). *Joint attention: Communication and other minds. Issues in philosophy and psychology*. Oxford: Oxford University Press.

Emery, N. J. (2000). The eyes have it: The neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews*, *24*(6), 581–604.

Engelland, C. (2014). *Ostension*. https://doi.org/10.7551/mitpress/9780262028097.001.0001

Eriksen, C. W., & James, J. D. S. (1986). Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics*, *40*(4), 225–240. https://doi.org/10.3758/BF03211502

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. https://doi.org/10.1038/415429a

Fernyhough, C. (2008). Getting Vygotskian about theory of mind: Mediation, dialogue, and the development of social understanding. *Developmental Review*, *28*(2), 225–262. https://doi.org/10.1016/j.dr.2007.03.001

Flom, R., Lee, K., & Muir, D. (Eds.). (2017). *Gaze-following: Its development and significance*. New York: Psychology Press.

Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin*, *133*(4), 694–724. https://doi.org/10.1037/0033-2909.133.4.694

Gogate, L. J., & Hollich, G. (2016). Early verb-action and noun-object mapping across sensory modalities: A neuro-developmental view. *Developmental Neuropsychology*, *41*(5-8), 293–307. https://doi.org/10.1080/87565641.2016.1243112

Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development*, *71*(4), 878–894.

Gogate, L. J., Bolzani, L. H., & Betancourt, E. A. (2006). Attention to maternal multimodal naming by 6- to 8-month-old infants and learning of word-object relations. *Infancy*, *9*(3), 259–288. https://doi.org/10.1207/s15327078in0903_1

Gregory, N. J., Hermens, F., Facey, R., & Hodgson, T. L. (2016). The developmental trajectory of attentional orienting to socio-biological cues. *Experimental Brain Research*, *234*(6), 1351–1362. https://doi.org/10.1007/s00221-016-4627-3

Gregory, S. E. A., & Jackson, M. C. (2017). Joint attention enhances visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(2), 237–249. https://doi.org/10.1037/xlm0000294

Heed, T., Habets, B., Sebanz, N., & Knoblich, G. (2010). Others' actions reduce crossmodal integration in peripersonal space. *Current Biology*, *20*(15), 1345–1349. https://doi.org/10.1016/j.cub.2010.05.068

Heimler, B., van Zoest, W., Baruffaldi, F., Rinaldi, P., Caselli, M. C., & Pavani, F. (2015). Attentional orienting to social and nonsocial cues in early deaf adults. *Journal of Experimental Psychology: Human Perception and Performance, 41*(6), 1758–1771. https://doi.org/10.1037/xhp0000099

Hermens, F. (2017). The effects of social and symbolic cues on visual search: Cue shape trumps biological relevance. *Psihologija, 50*(2), 117–140.

Hobson, P., & Hobson, J. (2011). Joint attention or joint engagement? Insights from autism. In A. Seemann (Ed.), *Joint attention: New developments in psychology, philosophy of mind, and social neuroscience* (pp. 115–136). Cambridge, MA: MIT Press.

Hoehl, S., Wahl, S., Michel, C., & Striano, T. (2012). Effects of eye gaze cues provided by the caregiver compared to a stranger on infants' object processing. *Developmental Cognitive Neuroscience*, *2*(1), 81–89. https://doi.org/10.1016/J.DCN.2011.07.015

Horstmann, A., & Hoffmann, K.-P. (2005). Target selection in eye–hand coordination: Do we reach to where we look or do we look to where we reach? *Experimental Brain Research*, *167*(2), 187–195. https://doi.org/10.1007/s00221-005-0038-6

Hyde, D. C., Flom, R., & Porter, C. L. (2016). Behavioral and neural foundations of multisensory face-voice perception in infancy. *Developmental Neuropsychology*, *41*(5-8), 273–292. https://doi.org/10.1080/87565641.2016.1255744

Itier, R. J., & Batty, M. (2009). Neural bases of eye and gaze processing: The core of social cognition. *Neuroscience & Biobehavioral Reviews*, *33*(6), 843–863. https://doi.org/10.1016/j.neubiorev.2009.02.004

Jesse, A., & Johnson, E. K. (2016). Audiovisual alignment of co-speech gestures to speech supports word learning in 2-year-olds. *Journal of Experimental Child Psychology*, *145*, 1–10. https://doi.org/10.1016/j.jecp.2015.12.002

Kaplan, F., & Hafner, V. V. (2006). The challenges of joint attention. *Interaction Studiesinteraction Studies Social Behaviour and Communication in Biological and Artificial Systems*, *7*(2), 135–169.

Keetels, M., & Vroomen, J. (2012). Perception of synchrony between the senses. In M. M. Murray & M. T. Wallace (Eds.), *The neural bases of multisensory processes*. Boca Raton, FL: CRC Press/Taylor & Francis.

Keysers, C., Wicker, B., Gazzola, V., Anton, J.-L., Fogassi, L., & Gallese, V. (2004). A Touching Sight: SII/PV Activation during the Observation and Experience of Touch. *Neuron, 42*(2), 335–346. https://doi.org/10.1016/S0896-6273(04)00156-4

Kim, K., & Mundy, P. (2012). Joint attention, social-cognition, and recognition memory in adults. *Frontiers in Human Neuroscience*, *6*, 172. https://doi.org/10.3389/fnhum.2012.00172

Kinard, J. L., & Watson, L. R. (2015). Joint attention during infancy and early childhood across cultures. In J. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (pp. 844–850). https://doi.org/10.1016/B978-0-08-097086-8.23172-3

Kirkham, N. Z., Rea, M., Osborne, T., White, H., & Mareschal, D. (2019). Do cues from multiple modalities support quicker learning in primary schoolchildren? *Developmental Psychology*, *55*, 2048–2059. https://doi.org/10.1037/dev0000778

Klapetek, A., Ngo, M. K., & Spence, C. (2012). Does crossmodal correspondence modulate the facilitatory effect of auditory cues on visual search? *Attention, Perception, & Psychophysics*, *74*(6), 1154–1167. https://doi.org/10.3758/s13414-012-0317-9

Klein, R. M., & Lawrence, M. A. (2012). On the modes and domains of attention. In M. I. Posner (Ed.), *Cognitive neuroscience of attention,* 2nd (pp. 11–28). New York, NY: Guilford Press.

Kompatsiari, K., Ciardo, F., Tikhanoff, V., Metta, G., & Wykowska, A. (2019). It's in the eyes: The engaging role of eye contact in HRI.

International Journal of Social Robotics, 1–11. https://doi.org/10.1007/s12369-019-00565-4

Langton, S. R. H., 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

Leavens, D., & Racine, T. P. (2009). Joint attention in apes and humans: Are humans unique? *Journal of Conciousness Studies*, *16*(6-8), 240–267.

Liebal, K., Waller, B. M., Burrows, A. M., & Slocombe, K. E. (2014). *Primate communication: A multimodal approach*. https://doi.org/10.1017/CBO9781139018111

Lieberman, A. M., Hatrak, M., & Mayberry, R. I. (2014). Learning to look for language: Development of joint attention in young deaf children. *Language Learning and Development, 10*(1), 19–35. https://doi.org/10.1080/15475441.2012.760381

Mason, G. M., Goldstein, M. H., & Schwade, J. A. (2019). The role of multisensory development in early language learning. *Journal of Experimental Child Psychology, 183*, 48–64. https://doi.org/10.1016/j.jecp.2018.12.011

McDonald, J. J., Teder-Sälejärvi, W. A., & Hillyard, S. A. (2000). Involuntary orienting to sound improves visual perception. *Nature*, *407*(6806), 906–908. https://doi.org/10.1038/35038085

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology*, *14*(2), 247–279. https://doi.org/10.1016/0010-0285(82)90010-X

Moore, C., & Dunham, P. J. (1995). Current Themes in Research of Joint Attention. In C. Moore & P. J. Dunham (Eds.), Joint attention: Its origins and role in development (pp. 15–28). Hillsdale, NJ: Lawrence Erlbaum.

Morales, M., Mundy, P., Crowson, M. M., Neal, A. R., & Delgado, C. E. F. (2005). Individual differences in infant attention skills, joint attention, and emotion regulation behaviour. *International Journal of Behavioral Development*, *29*(3), 259–263. https://doi.org/10.1177/01650250444000432

Mundy, P. (2016). *Autism and joint attention: Development, neuroscience, and clinical fundamentals*. Guilford Publications.

Mundy, P. (2018). A review of joint attention and social-cognitive brain systems in typical development and autism spectrum disorder. *European Journal of Neuroscience*, *47*(6), 497–514. https://doi.org/10.1111/ejn.13720

Mundy, P., & Jarrold, W. (2010). Infant joint attention, neural networks and social cognition. *Neural Networks*, *23*(8), 985–997. https://doi.org/10.1016/j.neunet.2010.08.009

Mundy, P., & Newell, L. (2007). Attention, joint attention, and social cognition. *Current Directions in Psychological Science*, *16*(5), 269–274. https://doi.org/10.1111/j.1467-8721.2007.00518.x

Mundy, P., & Sigman, M. (2015). Joint attention, social competence, and developmental psychopathology. In *Developmental psychopathology* (pp. 293–332). https://doi.org/10.1002/9780470939383.ch9

Ngo, M. K., & Spence, C. (2010). Auditory, tactile, and multisensory cues facilitate search for dynamic visual stimuli. *Attention, Perception, & Psychophysics*, *72*(6), 1654–1665. https://doi.org/10.3758/APP.72.6.1654

Nuku, P., & Bekkering, H. (2010). When one sees what the other hears: Crossmodal attentional modulation for gazed and non-gazed upon auditory targets. *Consciousness and Cognition, 19*(1), 135–143. https://doi.org/10.1016/j.concog.2009.07.012

Núñez, M. (2014). *Joint attention in deafblind children: A multisensory path towards a shared sense of the world*. London: Sense.

O'Madagain, C., Kachel, G., & Strickland, B. (2019). The origin of pointing: Evidence for the touch hypothesis. *Science Advances, 5*(7), eaav2558. https://doi.org/10.1126/sciadv.aav2558

Partan, S., & Marler, P. (1999). Communication goes multimodal. *Science*, *283*(5406), 1272–1273.

Pavani, F., Venturini, M., Baruffaldi, F., Caselli, M. C., & van Zoest, W. (2019). Environmental Learning of Social Cues: Evidence From Enhanced Gaze Cueing in Deaf Children. *Child Development*, *90*(5), 1525–1534. https://doi.org/10.1111/cdev.13284

Pelz, J., Hayhoe, M., & Loeber, R. (2001). The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research*, *139*(3), 266–277.

Pfeiffer-Leßmann, N., Pfeiffer, T., & Wachsmuth, I. (2012). An operational model of joint attention: Timing of gaze patterns in interactions between humans and a virtual human. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 851–856). Austin, TX: Cognitive Science Society.

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology*, *32*(1), 3–25. https://doi.org/10.1080/00335558008248231

Prezbindowski, A. K., Adamson, L. B., & Lederberg, A. R. (1998). Joint attention in deaf and hearing 22 month-old children and their hearing mothers. *Journal of Applied Developmental Psychology*, *19*(3), 377–387. https://doi.org/10.1016/S0193-3973(99)80046-X

Richman, A. L., Miller, P. M., & LeVine, R. A. (1992). Cultural and educational variations in maternal responsiveness. *Developmental Psychology, 28*(4), 614–621. https://doi.org/10.1037/0012-1649.28.4.614

Rowe, C. (1999). Receiver psychology and the evolution of multicomponent signals. *Animal Behaviour*, *58*(5), 921–931. https://doi.org/10.1006/anbe.1999.1242

Scaife, M., & Bruner, J. (1975). The capacity for joint visual attention in the infant. *Nature*, *253*, 265–266.

Schilbach, L. (2015). Eye to eye, face to face and brain to brain: Novel approaches to study the behavioral dynamics and neural mechanisms of social interactions. *Current Opinion in Behavioral Sciences*, *3*, 130–135. https://doi.org/10.1016/J.COBEHA.2015.03.006

Sebanz, N., Knoblich, G., & Prinz, W. (2003). Representing others' actions: Just like one's own? *Cognition*, *88*(3), B11–B21. https://doi.org/10.1016/S0010-0277(03)00043-X

Sebanz, N., Knoblich, G., & Prinz, W. (2005). How two share a task: Corepresenting stimulus-response mappings. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(6), 1234–1246. https://doi.org/10.1037/0096-1523.31.6.1234

Seemann, A. (Ed.). (2011). *Joint attention: New developments in psychology, philosophy of mind, and social neuroscience*. Cambridge, MA: MIT Press.

Seemann, A. (2019). *The Shared World: Perceptual Common Knowledge, Demonstrative Communication, and Social Space*. Cambridge, MA: MIT Press.

Sellaro, R., Treccani, B., & Cubelli, R. (2018). When task sharing reduces interference: Evidence for division-of-labour in Stroop-like tasks. *Psychological Research*, 1–16. https://doi.org/10.1007/s00426-018-1044-1

Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences, 12*(11), 411–417. https://doi.org/10.1016/j.tics.2008.07.006

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shepherd, S. V. (2010). Following gaze: gaze-following behavior as a window into social cognition. *Frontiers in Integrative Neuroscience*, *4*, 5. https://doi.org/10.3389/fnint.2010.00005

Shteynberg, G. (2015). Shared attention. *Perspectives on Psychological Science*, *10*(5), 579–590. https://doi.org/10.1177/1745691615589104

Sinnett, S., Spence, C., & Soto-Faraco, S. (2007). Visual dominance and attention: The Colavita effect revisited. *Perception &*

Psychophysics, *69*(5), 673–686. https://doi.org/10.3758/BF03193770

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

Soto-Faraco, S., Sinnett, S., Alsius, A., & Kingstone, A. (2005). Spatial orienting of tactile attention induced by social cues. *Psychonomic Bulletin & Review*, *12*(6), 1024–1031. https://doi.org/10.3758/bf03206438

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*(4), 971–995. https://doi.org/10.3758/s13414-010-0073-7

Spence, C., & Deroy, O. (2013). How automatic are crossmodal correspondences? *Consciousness and Cognition*, *22*(1), 245–260. https://doi.org/10.1016/j.concog.2012.12.006

Spence, C., McDonald, J., & Driver, J. (2004a). Exogenous spatial-cuing studies of human cross-modal attention and multisensory integration. In C. Spence & J. Driver (Eds.), *Crossmodal space and crossmodal attention* (pp. 276–320). https://doi.org/10.1093/acprof:oso/9780198524861.003.0011

Spence, C., Pavani, F., Maravita, A., & Holmes, N. (2004b). Multisensory contributions to the 3-D representation of visuotactile peripersonal space in humans: Evidence from the crossmodal congruency task. *Journal of Physiology-Paris*, *98*(1), 171–189. https://doi.org/10.1016/j.jphysparis.2004.03.008

Spencer, P. E. (2000). Looking without listening: Is audition a prerequisite for normal development of visual attention during infancy? *Journal of Deaf Studies and Deaf Education*, *5*(4), 291–302. https://doi.org/10.1093/deafed/5.4.291

Stephenson, L. J., Edwards, S. G., Howard, E. E., & Bayliss, A. P. (2018). Eyes that bind us: Gaze leading induces an implicit sense of agency. *Cognition, 172*, 124–133. https://doi.org/10.1016/j.cognition.2017.12.011

Suarez-Rivera, C., Smith, L. B., & Yu, C. (2019). Multimodal parent behaviors within joint attention support sustained attention in infants. *Developmental Psychology*, *55*(1), 96–109. https://doi.org/10.1037/dev0000628

Swingler, M. M., Perry, N. B., & Calkins, S. D. (2015). Neural plasticity and the development of attention: Intrinsic and extrinsic influences. *Development and Psychopathology*, *27*(2), 443–457. https://doi.org/10.1017/S0954579415000085

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences, 14*(9), 400. https://doi.org/10.1016/J.TICS.2010.06.008

Tomasello, M. (1995). Joint attention as social cognition. In C. Moore & P. J. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 103–130). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Tomasello, M. (2014). *A natural history of human thinking*. Cambridge, MA: Harvard University Press.

Tomasello, M. (2019). *Becoming human: A theory of ontogeny*. Cambridge, MA: Harvard University Press.

Tomasello, M., & Farrar, M. J. (1986). Joint attention and early language. *Child Development*, *57*(6), 1454–1463. https://doi.org/10.2307/1130423

Tschentscher, N., & Fischer, M. H. (2008). Grasp cueing and joint attention. *Experimental Brain Research*, *190*(4), 493–498. https://doi.org/10.1007/s00221-008-1538-y

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: Nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(5), 1053–1065. https://doi.org/10.1037/0096-1523.34.5.1053

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2009). Poke and pop: Tactile–visual synchrony increases visual saliency. *Neuroscience Letters*, *450*(1), 60–64. https://doi.org/10.1016/j.neulet.2008.11.002

Van der Burg, E., Cass, J., Olivers, C. N. L., Theeuwes, J., & Alais, D. (2010). Efficient visual search from synchronized auditory signals requires transient audiovisual events. *PLoS ONE*, *5*(5), e10664. https://doi.org/10.1371/journal.pone.0010664

Stoep, N. van der, Postma, A., & Nijboer, T. C. W. (2017). Multisensory perception and the coding of space. In A. Postma & I. van der Ham (Eds.), *Neuropsychology of space: Spatial functions of the human brain* (pp. 123–158). https://doi.org/10.1016/B978-0-12-801638-1.00004-5

Volpe, G., & Gori, M. (2019). Multisensory interactive technologies for primary education: From science to technology. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.01076

Vroomen, J., & De Gelder, B. (2004). Perceptual effects of cross-modal stimulation: Ventriloquism and the freezing phenomenon. In *The handbook of multisensory processes* (Vol. 3, pp. 1–23). The MIT Press, London.

Vygotsky, L. S. (2012). *Thought and language* (E. Hanfmann, G. Vakar, & A. Kozulin, Eds.). Cambridge, MA: MIT Press.

Wahl, S., Marinović, V., & Träuble, B. (2019). Gaze cues of isolated eyes facilitate the encoding and further processing of objects in 4-month-old infants. *Developmental Cognitive Neuroscience*, *36*, 100621. https://doi.org/10.1016/j.dcn.2019.100621

Wahn, B., & König, P. (2017). Can limitations of visuospatial attention be circumvented? A review. *Frontiers in Psychology*, *8*, 1896. https://doi.org/10.3389/fpsyg.2017.01896

Wahn, B., Keshava, A., Sinnett, S., Kingstone, A., & König, P. (2017). Audiovisual integration is affected by performing a task jointly. *Proceedings of the 39th annual conference of the cognitive science society*, 1296–1301.

Willemse, C., Marchesi, S., & Wykowska, A. (2018). Robot faces that follow gaze facilitate attentional engagement and increase their likeability. *Frontiers in Psychology*, *9*, 70. https://doi.org/10.3389/fpsyg.2018.00070

Wright, R. D., & Ward, L. M. (2008). *Orienting of attention*. Oxford: Oxford University Press.

Yang, G.-Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R., … Wood, R. (2018). The grand challenges of Science Robotics. *Science Robotics*, *3*(14), eaar7650. https://doi.org/10.1126/scirobotics.aar7650

Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PLoS ONE*, *8*(11), e79659. https://doi.org/10.1371/journal.pone.0079659

Yu, C., Schermerhorn, P., & Scheutz, M. (2012). Adaptive eye gaze patterns in interactions with human and artificial agents. *ACM Transactions on Interactive Intelligent Systems*, *1*(2), 1–25. https://doi.org/10.1145/2070719.2070726