In: *Proceedings of the 15th International Conference on Intelligent Virtual Agents, (IVA 2015)*, pp. 77–86, Delft, The Netherlands. doi:10.1007/978-3-319-21996-7_8
The final publication is available at link.springer.com



# Modeling a social brain for interactive agents: integrating mirroring and mentalizing

Sebastian Kahl and Stefan Kopp

Social Cognitive Systems Group, Faculty of Technology,
Bielefeld University, Inspiration 1, 33619 Bielefeld, Germany
{skahl,skopp}@uni-bielefeld.de

**Abstract.** Human interaction has a distinct collaborative quality based on the attribution of communicative intentionality. Two networks in the human brain are often described as part of the "social brain": the mirror system for recognizing intentional behavior and the mentalizing (theory of mind) system for processing it. We equip virtual agents with both systems and model their interaction during embodied communication. Results of simulation experiments demonstrate how higher orders of theory of mind lead to more robustness of communication by enabling interactive grounding processes.

**Keywords:** Embodied Virtual Agents, Social Cognition, Mentalizing, Mirroring, Coordination, Gesture

## 1 Introduction

Building artificial agents for natural interaction with humans eventually requires a deep understanding of the mechanisms underlying human social behavior. We are particularly interested in the perceptual and cognitive mechanisms that shape human behavior in social interaction. Those mechanisms have been receiving growing interest in fields such as Cognitive Science and Cognitive Neuroscience in the last decade. Two partially overlapping networks have been identified in the "social brain" [17]: an *action observation system* for perceiving and recognizing others' behaviors, and a *mentalizing system* for understanding others in terms of attributed mental states or theory of mind (ToM). Action observation is widely assumed to rest upon principles of prediction-based processing [2], where predictions about expected sensory stimuli are continuously formed and evaluated against incoming sensory input to inform further processing. A core mechanism to derive such predictions are sensorimotor simulations of the observed behavior, also referred to as *mirroring*. Prediction-based processing has also been argued to underlie language production and comprehension [9] or the social brain more generally [5].

We argue that learning from these mechanisms may help to substantially improve the interactive capabilities of virtual agents. Furthermore, developing models and testing them in human-agent interaction contributes to a more detailed understanding of how the social brain works. For example, it is not clear so far

how the mentalizing system and the mirroring system work together: How does perception change when a behavior is assumed to be "for me"? How are pathways between perception and action modulated by mentalizing in social interaction? And how do these mechanisms participate in coordination processes like feedback, joint attention, or grounding? Indeed, interacting with agents assumed to be "intentional" is fundamentally different from interacting with non-intentional objects [6]. For example, it has been shown that intentionality-attribution and underlying mentalizing influence sensory processing to become "social perception", an altered understanding of each other's actions [19, 14], which is also known as an "intentional stance" [4]. Clearly, these processes play an important role not only in solitary observation events, in which they have been studied mostly so far, but even more so in continuous online interaction [8, 13].

In this paper, we present work towards virtual agents with social brain-like functions by realizing and integrating mirroring and mentalizing abilities in a cognitively inspired fashion. In the following sections, we first review related modeling attempts and then present a model that formalizes the two systems in terms of computational processes, as well as their roles and dynamic interplay in inter-agent communication. Finally, we report results from simulations of embodied communication between two virtual agents, each of which equipped with its own model. We analyze how different abilities for mentalizing enable increasingly complex social coordination, from mere mimicry to eventually shared understanding.

## 2 Related work

Researchers interested in embodied conversational agents (ECAs) have explored many ways that enable agents to respond in interactive settings of verbal and nonverbal communication. The Smart Body animation system [15] enables responsive combination of behavior animations and has been integrated with an improved text and speech analysis system, called Cerebella, to better react by mapping appropriate behavioral responses to derived mental states [7]. The Thalamus framework [10] employs a perceptual loop for continuous interaction with the environment mediated through the agent's body. It has been extended to a generation process shared between "mind" and "body", modeled as a network of behavior models that interface with a body representation [11]. Contradicting with (though referring to) tenets of embodied cognition, this model separates mind and body dualistically. We do not follow this modularized approach but strive for a more consequently embodied and situated account, in which cognitive processes ground in or even arise from sensorimotor layers and bodily shaped interaction with the environment, mediated through perception-action couplings.

As one of the few modeling attempts to combine mentalizing, perception and action control in dynamic social interaction, Wolpert et al. [18] underline in their MOSAIC model that a true communicative model needs to close the communicative loop and must be perceptive to the observer's responses and ul-

timately her understanding. They hypothesize a hierarchy of paired forward and inverse models as a basic mechanism for processing movement as well as beliefs or intentions. Sadeghipour and Kopp [12] proposed an Empirical Bayesian Belief Update model (EBBU), a probabilistic model that implements a mirroring-based account of the perception and production of gestural behavior. They use a hierarchically organized representation of motor knowledge for action perception through forward models that formulate probabilistic expectations about possible continuations of observed gestures. The same representation is used for action generation, with probabilistic interactions between both processes to explain priming and resonance effects. Representations are dynamically augmented by way of inverse models when an unknown action is encountered.

Few attempts have been made to clarify the interaction between mentalizing and mirroring. A meta-analysis of studies on mentalizing found that mirror areas are not recruited unless the task involves inferencing intentionality from action stimuli [17]. Teufel et al. [14] present the "Perceptual Mentalizing Model" which focuses on the influence of the mentalizing system on the mirror system via perceptual processing. They differentiate between explicit and implicit ToM (what we call here mentalizing and mirroring, respectively). Importantly, both kinds of ToM are assumed to be influenced by social sensory processing. Explicit ToM processes are associated with processing of intentionality of a movement and strongly influence perception-action coupling in implicit ToM processing. Wykowska et al. [19] present a model of social attention, the "Intentional Stance Model", in which the mentalizing system either exhibits an intentional stance towards agents (attributing intentionality) or a design stance towards objects. The mentalizing system is assumed to influence sensory processing in a top-down fashion, but also to affect sensory gain control in attention mechanisms. They report the sensory gain manipulation for attentional reorienting mechanisms to be stronger in the intentional stance than in the design stance. A key aspect in triggering this intentional stance seems to be social gaze, which has been found to lead to the attribution of communicative intent [1], which in turn differentially recruits the mirroring and mentalizing system networks in processing the behavior of the interlocutor.

## 3 Towards an integrated model of the social brain

In this paper we present first steps towards a model of how a predictive sensorimotor subsystem and a mentalizing subsystem for attributing mental states interact during situated communication. We thereby not only devise the model but also implement and test it in simulation of social (nonverbal) interactions between two virtual agents, to explore how communicative coordination emerges from the dynamic interplay between the two systems.

We base our modeling approach on a number of assumptions (see Figure 1): First, we define successful communication to be a process that requires *shared* communicative intentionality and establishes perceptual or conceptual common ground between the participants [16]. This state is achieved in a dynamic ground-

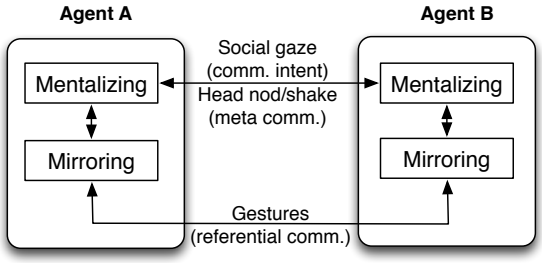


**Fig. 1.** The simulated nonverbal communication between our two virtual agents.

ing process [3], in which communicating agents reciprocally reveal and coordinate their beliefs about each other as well as the state of their interaction. Second, mentalizing plays a pivotal role by facilitating information integration and self-other distinction for coordinated action. It affects and receives information from the mirroring system, which itself processes perceived action in an immediate fashion. Third, we assume that coordinated action in communication highly depends on the degree of ToM realized by the mentalizing system: 2nd order reasoning, i.e. beliefs about other-beliefs, is minimally necessary for any cooperative behavior that goes beyond accidental coordination. Finally, gaze plays a special role in signaling and regulating social attention and is an indicator of communicative intent [1]. Mirror areas were found to be recruited especially when intentional action is expected [17]. We hence assume that gaze triggers mentalizing and thus also mirroring activity. Staying within the confines of the nonverbal domain, we also include head-nods and head-shakes as meta-communicative feedback for signalling agreement or disagreement.

### 3.1 Mentalizing subsystem

The mentalizing subsystem is a model of an agent's subjective ToM, which processes definite information about itself and infers others' mental states from perceptual input. A detailed depiction of the mentalizing subsystem is given in Figure 2, which we will refer to and describe in more detail in section 3.3. In its current version, the system utilizes a simple set of inference heuristics to model how mental state attributions arise and change in social interaction. In detail, this model consists of three sets of mental state attributes for different orders of ToM reasoning: Beliefs held about mental states of myself (*me*) or the interlocutor (*you*) constitute what we call '1st order ToM'. Further, in pursuit of a minimal cognitive model of mentalizing, we assume that only one order of ToM higher is needed for what we want to model. In contrast to the classical recursively nested beliefs, however, we stipulate these beliefs to be held about mental states that both interlocutors have in common (*we*). This is what we call '2nd order ToM'. The functional role that we ascribe to 2nd order ToM is to keep

track of common ground, the desire to agree, and the collaborative state of communication more generally. Generally, mental states consist of beliefs, desires, and intentions.

### 3.2 Mirroring subsystem

The mirroring subsystem employs the abovementioned EBBU model [12] for action observation and production. It implements a probabilistic hierarchical representation of sensorimotor knowledge about hand gestures, along with basic prediction, evaluation and activation processes that are used in both perception and generation of gestures. On the lowest level, *motor commands* are stored that represent segmented movements in time and space. Hand trajectories are given as directed graphs with edges representing motor commands. On the intermediate level, *motor programs* represent paths in the motor command graph and thus stand for meaningful movements. The highest level of abstraction stores *motor schemas* that cluster and represent similar motor programs. When observing a hand trajectory the hierarchical motor knowledge structure is activated and "resonates" to the observed gesture. In each time step the model predicts possible continuations of the observed gesture and compares them to the actual perception. Results lead to updated posterior probability distributions and hence activation of the corresponding motor commands, programs, or schemas. We equipped the mirroring subsystem with knowledge of different trajectories for three iconic ('circle', 'square', 'surface') and one emblematic gesture ('waving'). Those were learned from real human motion data. Single motor programs for a schema can take up to five seconds to produce, with motor commands being activated every tenth of a second. For every new observation of a hand trajectory entering the system the top-most level posterior distributions over motor schemas are taken as a proxy for a gesture's meaning, and are linked to first order mental state attributes in the mentalizig subsystem.

### 3.3 Integration and interplay

Our goal is to integrate mentalizing with mirroring-based action perception to account for how behavior and mental states arise and interact dynamically in a communicative interaction. As a working hypothesis, we consider both systems to be separate but with continuous interactions and projections between each other. In the mentalizing subsystem, any observation of actions can have a direct influence on the mental states held about *you*, where desires, beliefs and intentions are heuristically inferred. For any observed gesture processed by the mirroring subsystem, the most likely motor schema hypothesis is immediately projected into the mentalizing subsystem where it forms a mental state attributed as a *you*-belief, as long as the intention to communicate can be inferred. Correspondingly, a *me*-belief would cause the mirroring subsystem to recruit the intended motor schema for production. The current version of the mirroring subsystem is only capable of processing hand gestures; gaze and head movements are thus directly asserted to the mentalizing subsystem.

Depending on the degree of ToM processed in the agent's mentalizing module, communicative intent can trigger an inferred desire to reach mutual agreement about the understanding of the produced gesture. This is assessed applying a threshold for *good-enough* understanding to the likelihood of beliefs about mental states of *me* and *you* (1st order ToM). Note, however, when this threshold is exceeded the producer agent still cannot be certain about the correct understanding in the recipient unless sufficient feedback is provided. Here we require at least one correct reproduction of the gesture. Further, head-shake and head-nod signals are employed for meta-communication and can either increase or decrease confidence in the respective *you*-belief.

## 4 Simulation results

In order to test the model in online interactions we implemented the model and ran simulations with two virtual agents, each of which equipped with its own integrated model. At the start of the simulation, both agents only have a predefined set of mental states about themselves. They can communicate using four gestures ('circle', 'square', 'surface', and 'waving') that are perceived and generated as 3D hand trajectories, as well as head nods/shakes that are transferred as simple timed key-value pairs. Gestures are produced with a configured amount of white noise, normalized to the maximum movement vectors

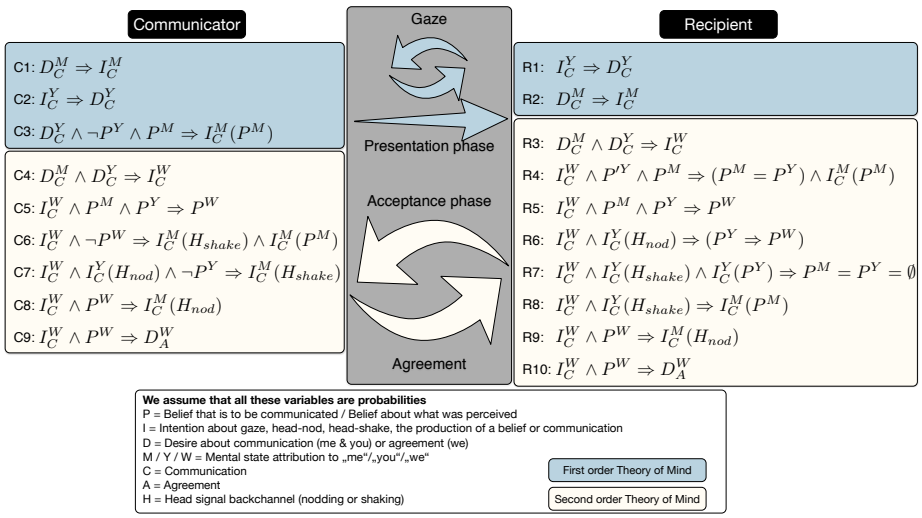


**Fig. 2.** Attributes and inference heuristics in the mentalizing subsystem applied during different phases of the interaction. The basis for complex inference is "Communicative Intention", inferred from social gaze. The "Communicator" agent enters the "Presentation Phase", followed by an "Acceptance Phase" of interactive grounding, where higher order mental attributions are needed for both agents to reach "Agreement".

in the motor schema, so that 10% noise reflect only a small amount of deviation during gesturing. The amount of noise, the ability for 2nd order ToM, and the good-enough threshold for minimal confidence in observing a gesture are the independent variables to parametrize the simulation. We ran six simulation setups: 10%/20%/30% noise with enabled or disabled 2nd order ToM capacity, and a static confidence threshold of 0.8. Each of the setups was run 100 times, always with identically configured agents. Simulations ended either when both agents believed to have reached agreement, or without 2nd order ToM, as soon as the *Communicator* finished its gesture production. As dependent variables we collected the probability distribution of the attributed *you*-belief about a gesture's meaning after every processing of a hand trajectory. We were particularly interested in the effects that different degrees of mentalizing have on the inter-agent coordination dynamics. The complexity of the communication depends on inferred communicative intent, signaled via social gaze. As soon as mutual communicative intent is established, the simulation follows a typical grounding process with presentation and acceptance phases [3], where the *Communicator* always starts with producing a 'circle' gesture.

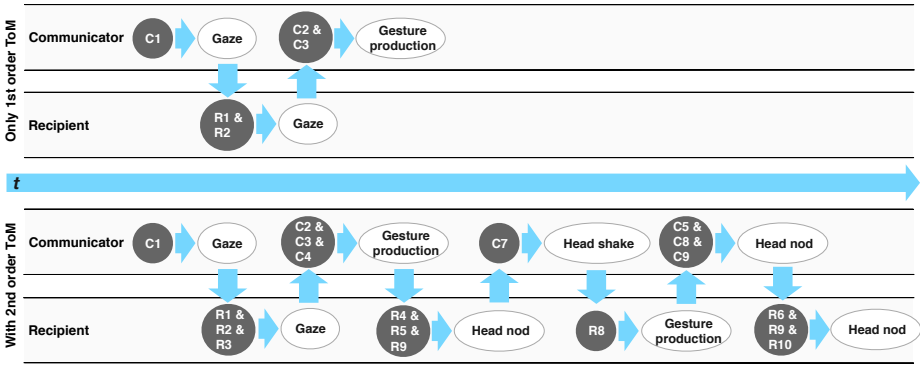


**Fig. 3.** Example interactions from our simulation when both agents have 1st order ToM (top) or 2nd order ToM (bottom). Overt behavior is shown along with the triggered mentalizing inferences (gray circles; indices referring to Figure 2).

To examplify the effect of the mental attributions and inferences possible in 1st and 2nd order ToM, Figure 3 illustrates two typical interaction patterns from our simulation study. The overt behavior of two agents, a *Communicator* and a *Recipient*, are shown along with the inferences drawn after perceiving or producing a certain behavior, with indices referring to the inference rules as shown in Figure 2. The interaction at the top shows a sequence of behavior and inferences typical for 1st order ToM mentalizing. The configured desire to communicate triggers *rule C1*, hence gaze behavior is perceived by the *Recipient* (*rule R1*). Since the *Recipient* is equally configured, its reciprocal gaze behavior

(*rule R2*) triggers an inference about the *Recipient's* desire to communicate in the *Communicator* (*rule C2*), and consequently a gesture is produced(*rule C3*). The interaction at the bottom shows behavior and inferences enabled through 2nd order ToM. While in the beginning there is a similarity to the 1st order ToM interaction, additionally *rules R3 and C4* are triggered and establish the agents' common communicative intent and thus the foundation for meaningful coordination behavior. After the initial gesture production the *Recipient's* mirroring subsystem provides the mentalizing subsystem with the most likely interpretation for the *Communicator's* behavior. That novel behavior triggers *rule R4*, by which the *Recipient* would ideally produce the understood gesture back to the *Communicator*, but in this interaction the gesture was understood with a likelihood above the good-enough threshold. This triggers *rule R5 and R9* as well, leading to a head-nod. Since the *Communicator* has no idea what the *Recipient* has understood the head-nod behavior is answered by a head-shake (*rule C7*), which triggers the *Recipient* to produce its understood gesture back to the *Communicator* (*rule R8*). The *Communicator* understands the gesture, which triggers rules equivalent to those in the *Recipient* (*rule C5, C8, and C9*), leading to a head-nod, which is equivalently answered by the *Recipient* (*rule R6, and R9*) and finalizes the interaction through mutually believed agreement (*rule R10*).

To test the agents' ability to coordinate with and without 2nd order ToM enabled, we analyzed the Kulbach-Leibler Divergence between the probability distributions of the *Recipient's you*-belief and the *Communicator's me*-belief, i.e. the "target belief". Figure 4(a) shows the divergence over interaction time. Without 2nd order ToM only one gesture was produced within 5 seconds. With 2nd order ToM the duration was strongly dependend on the correct understanding of observed gestures. The more mistakes, likely due to noise, the more correction effort emerged and hence longer interactions. Analyzed were interactions with length of at least 10 seconds and 20 seconds, respectively. These plots show

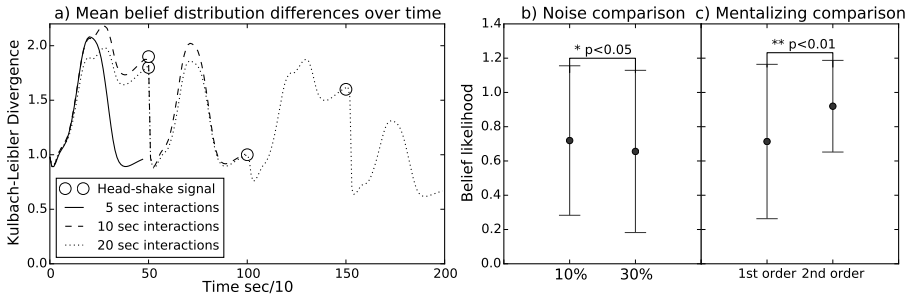


**Fig. 4.** Simulation results show a) KL-divergence between agents' beliefs during interactions of different extent, averaged over noise and ToM conditions, b) mean differences between noise conditions, and c) total mean differences between achieved likelihood about another's belief between ToM conditions.

the average success of coordination, especially in longer interactions. To test the effect of noise we compared the success of both agents reaching the target belief after 5 seconds, averaged over ToM conditions (Figure 4(b)). The comparison shows a significant difference ($t()$=2.4, $p$<0.05) between 10% ($M$=0.6, $SD$=0.4) and 30% ($M$=0.7, $SD$=0.5) noise conditions on gesture understanding. Subsequently, we tested the influence of 2nd order ToM, also by analyzing the success of reaching the target belief (Figure 4(c)). A comparison of the final beliefs averaged over all noise conditions with 2nd order ToM ($M$=0.9, $SD$=0.27) and without ($M$=0.7, $SD$=0.45) showed that 2nd order ToM leads to significantly more likely success in coordination ($t$=6.8, $p$<0.01).

## 5 Conclusions

In this paper we have presented work towards equipping virtual agents with a cognitively inspired model of a "social brain". Our approach is based on the notion that in social interaction, abilities for higher order mentalizing come to interact with predictive action observations in particular ways, and it is this interplay that accounts for the dynamic coordination mechanisms responsible for successful communication. The present step was to implement a 'minimal' mentalizing model that enables distinct mental perspectives, corresponding to beliefs about *me*, *you*, and *we*, and to let it interact with a mirroring-based EBBU model. Actual simulations of dynamically unfolding interaction were run to investigate whether higher order mental state attributions can give virtual agents a distinct advantage in inferring information necessary to successfully act towards a communicative goal. The results we obtained so far demonstrate that endowing IVAs with mechanisms found in the "social brain" enables interactive grounding without scripting it, and thus makes communication significantly more robust and efficient. However, even with higher order mentalizing capacities, a too large perturbation of the communicative signals led to long interaction times due to the inefficient error correcting mechanism emerging from both agents' goal for successful communication. Still, our first prototypical modeling attempt established that mentalizing is crucial for meaningful coordination behavior and success in communication could not be guaranteed without 2nd order ToM. We see the present framework as a good basis for further investigations how an understanding of social cognitive processes can help in the endeavour towards more natural and robust embodied agents, just as the analysis of interactions of humans with agents equipped with our framework will contribute to testing our understanding of how the social brain works. As next steps we will include an account for strategic noise compensation through altered signaling behavior, and want to pursue the question how self-other distinctions can manifest in the sensorimotor system during action observations to help in that attempt.

**Acknowledgements** This research/work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

## References


1. Ciaramidaro, A., Becchio, C., Colle, L., Bara, B.G., Walter, H.: Do you mean me? Communicative intentions recruit the mirror and the mentalizing system. Social Cognitive and Affective Neuroscience 9(7), 909–916 (Jul 2014)
2. Clark, A.: Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences 36(03), 181–204 (Jun 2013)
3. Clark, H.H., Brennan, S.E.: Grounding in communication. In: Perspectives on socially shared cognition, vol. 13, pp. 127–149. American Psychological Association, Washington, DC, US (1991)
4. Dennett, D.C.: The Intentional Stance. The MIT Press, Cambridge, MA (1987)
5. Frith, U., Frith, C.: The social brain: allowing humans to boldly go where no other species has been. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 365(1537), 165–176 (Jan 2010)
6. Gangopadhyay, N., Schilbach, L.: Seeing minds: A neurophilosophical investigation of the role of perception-action coupling in social perception. Social Neuroscience 7(4), 410–423 (Jul 2012)
7. Lhommet, M., Marsella, S.C.: Gesture with meaning. In: Intelligent Virtual Agents. pp. 303–312. Springer (2013)
8. Myllyneva, A., Hietanen, J.K.: There is more to eye contact than meets the eye. Cognition 134, 100–109 (Jan 2015)
9. Pickering, M.J., Garrod, S.: An integrated theory of language production and comprehension. The Behavioral and brain sciences 36(4), 329–47 (Aug 2013)
10. Ribeiro, T., Vala, M., Paiva, A.: Thalamus: Closing the mind-body loop in interactive embodied characters. In: Intelligent virtual agents. pp. 189–195. Springer (2012)
11. Ribeiro, T., Vala, M., Paiva, A.: Censys: A Model for Distributed Embodied Cognition. In: Intelligent Virtual Agents. pp. 58–67. Springer (2013)
12. Sadeghipour, A., Kopp, S.: Embodied Gesture Processing: Motor-Based Integration of Perception and Action in Social Artificial Agents. Cognitive computation 3(3), 419–435 (Sep 2011)
13. Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., Vogeley, K.: Toward a second-person neuroscience. The Behavioral and brain sciences 36(4), 393–414 (Aug 2013)
14. Teufel, C., Fletcher, P.C., Davis, G.: Seeing other minds: Attributed mental states influence perception. Trends in Cognitive Sciences 14(8), 376–382 (2010)
15. Thiebaux, M., Marsella, S., Marshall, A.N., Kallmann, M.: SmartBody: behavior realization for embodied conversational agents. In: Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems. vol. 1, pp. 151–158 (2008)
16. Tomasello, M.: Origins of Human Communication. The MIT Press, Cambridge, MA (2008)
17. Van Overwalle, F.: Social cognition and the brain: a meta-analysis. Human brain mapping 30(3), 829–58 (Mar 2009)
18. Wolpert, D.M., Doya, K., Kawato, M.: A unifying computational framework for motor control and social interaction. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 358(1431), 593–602 (Mar 2003)
19. Wykowska, A., Wiese, E., Prosser, A., Müller, H.J.: Beliefs about the minds of others influence how we process sensory information. PLoS ONE 9(4) (2014)