Human-Inspired Socially-Aware Interfaces

Dominik Schiller[1 [0000−0001−7364−5772]], Katharina Weitz[1 [0000−0003−1001−2278]], Kathrin Janowski[1 [0000−0001−5985−4973]], and Elisabeth André[1 [0000−0002−2367−162X]]

Human Centered Multimedia, Augsburg University, Augsburg, Germany
{schiller, weitz, janowski, andre}@hcm-lab.de

Abstract. Social interactions shape our human life and are inherently emotional. Human conversational partners usually try to interpret – consciously or unconsciously – the speaker's or listener's affective cues and respond to them accordingly. With the objective to contribute to more natural and intuitive ways of communicating with machines, an increasing number of research projects has started to investigate how to simulate similar affective behaviors in socially-interactive agents. In this paper we present an overview of the state of the art in social-interactive agents that expose a socially-aware interface including mechanisms to recognize a user's emotional state, to respond to it appropriately and to continuously learn how to adapt to the needs and preferences of a human user. To this end, we focus on three essential properties of socially-aware interfaces: Social Perception, Socially-Aware Behavior Synthesis, and Learning Socially-Aware Behaviors. We also analyze the limitations of current approaches and discuss directions for future development.

Keywords: Socially-Interactive Agents · Social Signal Processing · Affective Computing.

1 Introduction

Rosa [38] sees an essential aspect for a successful life in a resonant world relationship. By resonance, he understands the reaction of humans towards the world around them. In a world that has been strongly dominated by technology in recent decades, and after the great success of technical systems in industry, economy, and our daily life, these systems have become part of this human environment. Initially, the interaction between humans and machines did not seem very "social", but was characterized by the formal processing of tasks. Nowadays, machines are increasingly being used by non-expert users in domestic environments. Often machines are not just employed as tools, but can take on the role of assistants, consultants or even companions. Consequently, there is a need to design interaction technologies that allow humans and machines to interact with each other as naturally as possible. A prominent attempt to create a socially-interactive learning system that can communicate intuitively with the user is the Baby-X-Project [25]. In addition to natural interaction, the developers also modeled the underlying mechanisms based on findings in the field of neuroscience.

People can interact with the system that embodies a virtual baby as they would interact with a human toddler. One important aspect of human communication takes effect here: Interpersonal communication is inherently emotional. Human conversational partners usually try to interpret consciously or unconsciously the speaker's or listener's affective cues and respond to them accordingly. The willingness and ability to empathize with the attitudes and emotions of other people is not only important in interpersonal communication, but should also be considered in the development of socially-interactive agents.

This paper discusses the current state of the art in socially-aware interfaces that dynamically adapt to the affective state of an interlocutor and discusses current limitations and future perspectives of such systems. To this end, we focus on selected capabilities of socially-aware interfaces that fall into the following three categories: (1) Social Perception (2) Socially-Aware Behavior Synthesis and (3) Learning Socially-Aware Behaviors. We conclude this paper by summarizing the presented findings and pointing out potential directions for future developments.

2 Socially-Aware Interfaces

2.1 Social Perception

To incorporate social cues and non-verbal behavior into socially-aware interfaces, robust techniques are required to detect and analyze them. Since humans usually rely on multiple modalities to convey social cues including language, gestures or facial expressions, social cues can be recognized using a large variety of sensory equipment, such as microphones or cameras. Yet the recognition of such social signals is known to be a very hard problem and a real bottleneck on the path to improve human-computer interaction. This section presents a survey of research on automatically sensing and interpreting social signals.

Social Signal Sensing Recent research in the area of social signal processing has focused on a large variety of modalities to determine the affective state of a user. Such modalities include facial expressions [29], gestures and postures [24], speech [46], and physiological measurements [23]. Since modern human-computer interfaces are often offering voice-based interaction, language is an obvious communication channel to explore. Emotions may be determined from the semantic meaning of utterances as well as the paralinguistic acoustic properties, such as jitter or pitch. Also, the recognition of spontaneous displays of emotion, such as sighs, laughs, or moans, have been examined [47].

In the recent past, significant effort has been made to determine an optimal set of such features for emotional speech recognition. As a result, modern frameworks like EmoVoice [45] or OpenSMILE [15] are able to extract thousands of features for this task. In an attempt to establish a generic baseline feature set that is easy to interpret and generalizes well over a magnitude of different tasks, Eyben et al. [14] developed the Geneva Minimalistic Acoustic Parameter Set (GeMAPS). Besides manually engineered features, recent improvements in deep

learning techniques are now providing the foundation for automatically learning a suitable representation of the data [44]. Wagner et al. [48], who compared hand-crafted acoustic features with automatically learned representations, concluded that hand-crafted features are still beneficial at the moment, but especially with recent improvements in deep learning, automatic feature extraction is gaining more and more importance.

Besides the analysis of paralinguistic features, the semantic content of spoken utterances may be exploited to determine the emotional content of an utterance [34]. Traditionally, affective word dictionaries, such as WordNet-Affect [42], are employed to determine the emotional content of a word. However, to obtain acceptable recognition rates, the linguistic context has to be taken into account. Negations represent a particular challenge. They may reverse the polarity of affect conveyed by an utterance as in "I'm not happy", but they may also serve as an amplifier of affect as in "Never have I been so happy." The examples indicate that hand-crafting rules for sentiment analysis is time-consuming. On the analogy of trends in the paralinguistic analysis of emotions, deep learning methods have been proven a promising approach to automatically learn linguistic representations for sentiment analysis [50]. To improve the results of emotion classifiers, the integration of data from multiple sources [27] has been researched. The fusion of multiple modalities usually leads to an improvement of classification reliability compared to a single modality. Though, fusion approaches typically achieve higher gains for acted than for spontaneous behavior [11].

Social Signal Understanding As other areas, social signal sensing benefits from advances in deep neural networks learning. While the robustness of social signal sensing has improved, the use of visualization techniques to enhance the transparency of neural networks revealed that neural networks do not always focus on relevant input components to come up with an interpretation. For example, Weitz et al. [49] analyzed how a deep neural network distinguishes facial expressions of pain and emotions and observed that the deep neural network did not exclusively direct its attention to the face, but also on the (in this case expressionless) background of an image. Even though implausible behavior may lead to correct results, users may loose trust in a system if they find out that the system just attempts to convey the illusion of an understanding system.

To interpret social cues, it is important to equip a system with the ability to understand a user's behavior within the context of an interaction. For example, to determine whether an interaction is enjoyable or not, it does not suffice to look at the laughs of each individual separately. Rather, the temporal dynamics of the laughs within a dialogue has to be considered. In order to take account of the interplay of social cues, Baur et al. [4] developed a probabilistic framework that does not only explicitly model the interlocutors' social cues, but also how they depend on each other.

Furthermore, the interpretation of social cues depends on the situative context in which they occur. For example, a laugh is not always an indicator of joy, but may also allude to negative emotions, such as embarrassment. To interpret

such social cues correctly, the situative context has to be taken into account. This task is a great challenge since it requires not only analyzing, but also reasoning about the interlocutor's situation. As a first step to deal with this task, Gebhard et al. [17] combined a framework for detecting multimodal social cues with a cognitive model of affect (see also section 2.3). The basic idea of the framework is to relate expected emotional appraisal and regulation behaviors to observed social cues.

Overall, it may be said, however, that most approaches to social signal sensing focus on observable indicators of affect as opposed to aiming at deeper understanding of the user's psychological states.

2.2 Socially-Aware Behavior Synthesis

Rosa [38] describes the relationship to the world as fundamentally meaningful for humans, where the intertwining of human beings with their surroundings can be understood as constant mutual interaction. Therefore the environment affects human behavior, just as humans influence the world through their actions. While we have already addressed approaches for social signal perception in the previous section, we will now discuss how an agent should respond to such social signals to demonstrate that it is aware of the user.

Socially-Aware Navigation Socially-aware behaviors include adequate navigation behaviors that follow proxemics conventions, such as maintaining a comfortable distance to nearby people.

Agents that are capable of moving freely in their environment may approximate human proxemic behavior by acting spatially, which in return unlocks a large variety of new possible (interaction) behaviors. Besides maintaining an appropriate distance to the interlocutor, agents may use their body orientation and gaze behavior (e.g., [35,5]). Proxemics is also expressed by conversational behaviors, such as choosing an appropriate topic in small talk that does not appear intrusive to the interlocutor [12].

Typically, the implementation of proxemics behaviors is inspired by studies with human users that interact with a robot in physical environment (see, e.g., [43,13]) or are placed with a synthetic agent in a virtual environment [36]. A particular challenge is to learn appropriate proxemics behaviors in real-time. An example includes the work by Mitsunaga et al. [31] who used comfort and discomfort as input for the reward function, which they calculated based on gaze duration and body movements.

Turn Management Being able to understand and convey each other's turn-taking intentions is an important prerequisite for fluent, natural dialogue. In particular, people involved in a conversation monitor the other person's gaze direction to infer who or what has their attention at any given moment. This belief about their attentional state plays a major part in the constant negotiation of speaker and listener roles. Looking at an interaction partner is believed to signal

that the communication channel is open and one is ready to receive information from the said partner, which is why human speakers establish eye contact when they want to elicit backchannel feedback or a full response from the listener [3]. Conversely, averted gaze is seen as a signal for the opposite, for example, when the speaker wants to take the turn, but is still planning what to say [3].

Bohus and Horvitz [7] used the gaze direction of human quiz players to determine whether a player was yielding the conversational floor to one of their team members or the virtual quiz master agent. Likewise, they directed the agent's gaze towards a certain player whenever the agent expected that player to start speaking, but had the agent avert its gaze while it was waiting for processing results to prevent the humans from taking the turn. Similar gaze aversion signals were applied to social robot behavior by Andrist et al. [2], causing the users to wait more patiently when the robot appeared to be busy thinking. Skantze et al. [41] observed similar effects when a robot was using turn hold signals, such as gaze aversion or filler sounds. They also observed that humans turned their gaze towards the robot when they were waiting for the next piece of information. From these observations, the authors concluded that detecting and correctly interpreting the corresponding behavioral cues from the human would allow social robots to time their responses more appropriately.

Interruptions What timing is appropriate depends on more factors than the belief about the interaction partner's intentions. Deliberately ignoring or being overly sensitive to said intentions can send additional messages about the interaction context, the personality of a participant or their attitude towards the interlocutor. Studies with virtual agents have shown that interrupting and interruption handling behavior associated with specific human personality traits and attitudes also lead human observers to attribute similar characteristics to artificial beings. For instance, ter Maat et al. [28] found that agents who started to speak before the end of another agent's turn were perceived as less agreeable, whereas those who waited for a few seconds appeared less extraverted. Yielding the turn as soon as an overlap was detected led to lower dominance ratings than continuing. Gebhard et al. [19] later varied the interruption response timing for a virtual agent in an interactive dialogue system. Their study confirmed that the agent was perceived as more dominant when it continued talking for a longer time after the user had tried to interrupt it. The opposite was observed for the perceived closeness between user and agent, and the agent was also rated as more friendly when the overlap was minimal.

Turn-taking behavior is shaped by the arbitration between different, possibly conflicting interaction goals, such as being polite while also being assertive. Janowski and André [21] proposed a decision-theoretic model based on psychological theories about how a person's personality, interpersonal stance and different interaction goals relate to each other. One major objective of this research is to mimic human reasoning about timing one's dialogue contributions. This way, the behavior of social agents is intended to become more transparent to interaction designers and end users, as the causal relationships represented in

the model can be used to explain the agent's decision. Furthermore, the same causal relationships can be used to interpret a human's surface behavior in terms of their intentions which, as stated above, the agent needs to understand in order to adapt its behavior. Evaluation results showed that the model could be used to generate interrupting and interruption handling behavior patterns in line with psychological literature and related works.

2.3 Modeling and Simulating Empathy

Various attempts have been made to implement empathic behaviors in computer-based dialogue systems. The simplest form of empathy, *Ideomotoric Empathy*, consists of imitating the emotional cues of the dialogue partner. Mirroring the users' expression is possible without understanding their emotional states. For example, if a user is distressed because he or she was not able to solve a task in a tutoring system, an artificial agent would simply imitate the user's facial expression without knowing why the user is distressed.

This is one of the behaviors realized by Bee et al. [6] with the attentive listener agent Alfred. They used EmoVoice [45] to detect emotional cues in the speaker's voice, from which they calculated the user's current mood tendency via the ALMA model of affect [16]. This mood was then reflected in the agent's facial expression. Another example has been realized by Janowski et al. [22] using the humanoid robot Zeno, developed by Hanson Robotics. The tone of voice as well as the user's facial expression are analyzed to infer the user's emotional state. Zeno then shows the recognized emotion in his face. The semantic content of the user's speech is not taken into account in either of these examples.

Higher forms of empathy can be generally divided into two concepts: *Cognitive* and *Affective Empathy*. While cognitive empathy refers to the ability to understand another person's perspective, affective empathy means the capacity to respond with an appropriate reaction to another's mental state.

Many approaches to simulate emotional processes are based on rules that are inspired by theories from the cognitive sciences. A popular theory is the OCC model by Ortony et al. [33] that includes detailed rules to explain the elicitation of 20 common emotions. On the basis of this model, several computer programs were developed that simulate emotional processes.

An example includes the work by Bee et al. [6] who implemented affective empathy for the virtual Alfred agent by appraising the emotional state inferred from the user's tone of voice. Based on the OCC model [33], the agent perceived negative emotions as "bad event for good other" and positive emotions as "good event for good other". Consequently, observing the user's state elicited the emotions "SorryFor" respectively "HappyFor" in the agent's own affect model, which would then be mapped to facial animations for the Alfred character.

Boukricha et al. [8] presented a computer model for affective empathy that considers the relationship between the single interlocutors when determining an agent's response to the emotional state of others. Their approach is illustrated by a scenario in which empathic behaviors for the agent Emma are automatically generated as a reaction to conversations between the agent Max and the agent

Lisa. For example, Emma would get irritated based on her current mood and her relationship to Max if Lisa should offend Max.

2.4 Learning Socially-Aware Behaviors

In the previous section, we described analytic approaches to simulate socially-aware behaviors based on theories from the cognitive sciences. The question arises of to what extent it is possible to learn sensitive behaviors from recordings of human-human interactions or from life interactions with human interlocutors.

McQuiggan and Lester [30] followed an empirical approach and collected data of empathic interactions between two agents in a virtual environment that were controlled by human trainers. One trainer had to accomplish specific tasks while the other trainer had to observe the behavior and to select appropriate empathic behaviors. Based on these recordings, a computer model for empathic behaviors was created using methods from machine learning (Naïve Bayes and decision trees). The model was tested by having human users interact with virtual agents whose empathic behaviors were based on the learnt model. The study revealed that the human users found the empathic behaviors of the virtual agents appropriate.

While McQuiggan and Lester [30] learnt empathic behaviors from previously recorded computer-mediated interactions between humans, Leite [26] presented an approach to adjust empathic strategies of a robotic cat during the interaction with a child based on Reinforcement Learning (RL). The approach makes use of a reward function that takes into account how the emotional state of a user changes after the application of an empathic strategy. As time progresses, the system learns which strategies have proven promising. Using a sophisticated selection mechanism, the system ensures that successful strategies are selected with greater probability while enabling flexible adaptation to a new situation. For example, the child might require more help when the degree of difficulty of the chess game increases such that the robot needs to adjust in that case. An evaluation showed that the robot managed to keep children interested over a longer period of time, which the authors ascribe to its empathic behavior.

Ritschel et al. [37] presented an RL approach that adapts the linguistic style of a robot based on subliminal feedback provided by a human user: affective signals. RL is used for continuously learning the desired profile over time instead of asking the user explicitly, sticking to a fixed personality. The reward signal required for RL comes directly from the user's level of engagement, which is estimated based on multimodal affective cues. The system's ability to adapt its dialogue style was evaluated using simulation results and an interactive prototype. While an interactive prototype provides more realistic results, a significant amount of effort is required from the human user to provide the system with useful data. Lifelong reinforcement learning [40] represents a promising approach to enable a system to gradually learn appropriate social behaviors over its lifetime by treating novel situations as new tasks in the learning process.

2.5 Socially-Aware Conversational Systems

One form of human-technology interaction that benefits strongly from the integration of the previously presented techniques is natural-language dialogue with embodied conversational agents (see [1,20]). In the following, we will present representative research projects that are focusing on equipping such agents with socially-aware sensing capabilities to dynamically tailor their conversational behaviors to the affective state of a user.

Since listening is an important component to impart appreciation in dialogues, SEMAINE [39] has focused on the implementation of such a behavior. The empathic listeners in the SEMAINE project were characterized by agents with different personality profiles. These agents were able to conduct a conversation with a person and to recognize and respond to a human user's non-verbal behaviors in real-time. The goal of SEMAINE was to create a natural conversational dialogue with the focus on the non-verbal behavior of the human counterpart. To this end, the agent had to be able to produce natural and human-like listener behavior without addressing the challenges of speech recognition and deep natural language understanding.

Cavazza et al. [9] integrated the user's affective state in the dialogue management of an agent to improve the robustness of their speech recognition system. To this end, they inferred the emotion of a user by analyzing the paralinguistic aspects of spoken utterances as well as inferring the general sentiment from the transcription of the spoken statement. If the users employed words to express their emotional state that are unknown to the system, the system would still be able to recognize their emotions from the acoustics of speech. Furthermore, the results of this analysis were used to generate an immediate empathic response when the user stopped speaking.

Morency et al. [32] developed a virtual agent that interacts in a clinical setting with a patient to assess multimodal behaviors related to post-traumatic stress disorder and stress. The automatically analyzed cues by the patient are used for diagnostic purposes and for controlling the dialogue between the agent and the patient. For example, the agent motivates patients who took a lot of pauses to keep talking.

Gebhard et al. [18] implemented a virtual agent called EmmA running on the users mobile phone and helping them cope with stress at work. The agent analyzes the users psychological state based on behavioral data obtained from mobile sensors and a simulation of emotion regulation strategies [17] in order to select appropriate verbal intervention strategies if the stress level is at a critical stage.

Damian et al. [10] focused also on negative emotions, such as stress and nervousness, but in a different setting. They developed a game-like environment for job interview training. In this safe environment, trainees could experiment with different conversational strategies during a job interview with a virtual agent. During the job interview, the verbal and non-verbal signals of the trainee were recorded and analyzed to evaluate the trainee's multimodal behaviors and to regulate the flow of the dialogue.

3 Discussion and Conclusion

In this paper, we presented an overview of current state of the art research towards developing fully socially-aware interfaces. To this end, we identified three essential capabilities of socially-aware interfaces: Social Perception, Socially-Aware Behavior Synthesis and Learning Socially-Aware Behaviors.

A closer look at existing research reveals a clear timeline of progression. Initially, socially-aware interfaces made use of handcrafted features to infer a user's emotional state based on the input of the available sensory equipment, and interactions relied on scripted processes to simulate emotional behavior. Modern approaches are now working towards automatically learned representations of the input data in order to improve recognition results of social signal sensing. Other approaches are even going beyond plain recognition by modeling the dynamics of interpersonal social cues within the situative and conversational context of their occurrence. Such social sensing capacities are often used by the underlying systems to imitate human social behavior during interactions. This way modern agents are capable of displaying human-like interaction behavior, such as mirroring a user's emotion or deciding when to take a turn during a dialog. Finally, attempts are being made to continuously learn appropriate social behaviors from interactions with human users following the paradigm of lifelong learning.

While research endeavors have become increasingly more sophisticated and complex, one key question remains when it comes to determining the direction of future development: Are we creating systems that show true understanding of human social interactions and behaviors, or are we creating systems that pretend to do so by simulating social awareness at a surface level?

While a complete answer to this question is not within the scope of this paper, we argue that current state-of-the-art approaches are focusing rather on the simulation of behaviors as opposed to true understanding. In order to develop systems that are capable of executing not only scripted or isolated individual tasks, but also taking a step towards sensitive behaviors reflecting true understanding, we need to go beyond the pure analysis of observable social cues towards deeper reasoning processes that analyze the context in which they appear.

4 Acknowledgements

This work has been partially funded by the Bundesministerium für Bildung und Forschung (BMBF) within the project VIVA, Grant Number 16SV7960.

References

1. André, E., Pelachaud, C.: Interacting with embodied conversational agents. In: Chen, F., Jokinen, K. (eds.) Speech Technology. pp. 123–149. Springer (2010)
2. Andrist, S., Tan, X.Z., Gleicher, M., Mutlu, B.: Conversational gaze aversion for humanlike robots. In: ACM/IEEE International Conference on Human-Robot Interaction, (HRI), Bielefeld, Germany. pp. 25–32 (2014)

3. Argyle, M., Cook, M.: Gaze and mutual gaze. Cambridge University Press (1976)
4. Baur, T., Schiller, D., André, E.: Modeling users social attitude in a conversational system. In: Tkalcic, M., Carolis, B.D., de Gemmis, M., Odic, A., Kosir, A. (eds.) Emotions and Personality in Personalized Services, pp. 181–199. Springer (2016)
5. Bee, N., André, E., Tober, S.: Breaking the ice in human-agent communication: Eye-gaze based initiation of contact with an embodied conversational agent. In: 9th International Conference on Intelligent Virtual Agents (IVA), Amsterdam. pp. 229–242 (2009)
6. Bee, N., André, E., Vogt, T., Gebhard, P.: The use of affective and attentive cues in an empathic computer-based companion. In: Wilks, Y. (ed.) Natural Language Processing, vol. 8, pp. 131–142. John Benjamins Publishing Company (2010)
7. Bohus, D., Horvitz, E.: Facilitating multiparty dialog with gaze, gesture, and speech. In: International ACM Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (ICML-MLMI), Beijing, China. pp. 5:1–5:8 (2010)
8. Boukricha, H., Wachsmuth, I., Carminati, M.N., Knoeferle, P.: A computational model of empathy: Empirical evaluation. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, (ACII), Geneva, Switzerland. pp. 1–6. IEEE (2013)
9. Cavazza, M., de la Camara, R.S., Turunen, M.: How was your day?: A companion ECA. In: 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1, (AAMAS), Toronto, Canada. pp. 1629–1630. Richland, SC (2010)
10. Damian, I., Baur, T., Lugrin, B., Gebhard, P., Mehlmann, G., André, E.: Games are better than books: In-situ comparison of an interactive job interview game with conventional training. In: International Conference on Artificial Intelligence in Education (AIED), Madrid, Spain. pp. 84–94. Springer (2015)
11. D'Mello, S., Kory, J.: Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In: 14th ACM International Conference on Multimodal Interaction (ICMI), Santa Monica, CA, USA. pp. 31–38. ACM (2012)
12. Endrass, B., Rehm, M., André, E.: Planning small talk behavior with cultural influences for multiagent systems. Computer Speech & Language **25**(2), 158–174 (2011)
13. Eresha, G., Häring, M., Endrass, B., André, E., Obaid, M.: Investigating the influence of culture on proxemic behaviors for humanoid robots. In: IEEE International Symp. on Robot and Human Interactive Communication, (RO-MAN), Gyeongju, South Korea, 2013. pp. 430–435 (2013)
14. Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., et al.: The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. on Affective Computing **7**(2), 190–202 (2015)
15. Eyben, F., Weninger, F., Gross, F., Schuller, B.: Recent developments in openSMILE, the munich open-source multimedia feature extractor. In: ACM Multimedia, Firenze, Italy. pp. 835–838 (2013)
16. Gebhard, P.: ALMA: A layered model of affect. In: 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS). pp. 29–36 (2005)
17. Gebhard, P., Schneeberger, T., Baur, T., André, E.: MARSSI: Model of appraisal, regulation, and social signal interpretation. In: 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS), Stockholm, Sweden. pp. 497–506 (2018)

18. Gebhard, P., Schneeberger, T., Dietz, M., André, E., ul Habib Bajwa, N.: Designing a mobile social and vocational reintegration assistant for burn-out outpatient treatment. In: 19th ACM International Conference on Intelligent Virtual Agents (IVA), Paris, France. pp. 13–15 (2019)
19. Gebhard, P., Schneeberger, T., Mehlmann, G., Baur, T., André, E.: Designing the impression of social agents' real-time interruption handling. In: 19th ACM International Conference on Intelligent Virtual Agents (IVA), Paris, France. pp. 19–21 (2019)
20. Gratch, J., Rickel, J., André, E., Cassell, J., Petajan, E., Badler, N.I.: Creating interactive virtual humans: Some assembly required. IEEE Intelligent Systems **17**(4), 54–63 (2002)
21. Janowski, K., André, E.: What If I Speak Now?: A decision-theoretic approach to personality-based turn-taking. In: 18th International Conference on Autonomous Agents and MultiAgent Systems, (AAMAS). pp. 1051–1059. Richland, SC (2019)
22. Janowski, K., Ritschel, H., Birgit, L., André, E.: Sozial interagierende Roboter in der Pflege. In: Bendel, O. (ed.) Pflegeroboter. pp. 63–87. Springer (2018)
23. Kim, J., André, E.: Emotion recognition based on physiological changes in music listening. IEEE Trans. on Pattern Analysis and Machine Intelligence **30**(12), 2067–2083 (2008)
24. Kleinsmith, A., Bianchi-Berthouze, N.: Affective body expression perception and recognition: A survey. IEEE Trans. on Affective Computing **4**(1), 15–33 (2012)
25. Lawler-Dormer, D.: Baby X: Digital artificial intelligence, computational neuroscience and empathetic interaction. In: ISEA 2013 Conference proceedings. ISEA International (2013)
26. Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., Paiva, A.: The influence of empathy in human–robot relations. International Journal of Human-Computer Studies **71**(3), 250–260 (2013)
27. Lingenfelser, F., Wagner, J., Deng, J., Brueckner, R., Schuller, B., André, E.: Asynchronous and event-based fusion systems for affect recognition on naturalistic data in comparison to conventional approaches. IEEE Trans. on Affective Computing **9**(4), 410–423 (2016)
28. ter Maat, M., Truong, K.P., Heylen, D.K.J.: How Agents' Turn-Taking Strategies Influence Impressions and Response Behaviors. Presence: Teleoperators and Virtual Environments **20**(5), 412–430 (Oct 2011)
29. Martínez, B., Valstar, M.F., Jiang, B., Pantic, M.: Automatic analysis of facial actions: A survey. IEEE Trans. Affective Computing **10**(3), 325–347 (2019)
30. McQuiggan, S.W., Lester, J.C.: Modeling and evaluating empathy in embodied companion agents. International Journal of Human-Computer Studies **65**(4), 348–360 (2007)
31. Mitsunaga, N., Smith, C., Kanda, T., Ishiguro, H., Hagita, N.: Adapting robot behavior for human–robot interaction. IEEE Trans. Robotics **24**(4), 911–916 (2008)
32. Morency, L.P., Stratou, G., DeVault, D., Hartholt, A., Lhommet, M., Lucas, G., Morbini, F., Georgila, K., Scherer, S., Gratch, J., et al.: SimSensei demonstration: a perceptive virtual human interviewer for healthcare applications. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
33. Ortony, A., Clore, G.L., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press (1988)
34. Osherenko, A., André, E.: Lexical affect sensing: Are affect dictionaries necessary to analyze affect? In: International Conference on Affective Computing and Intelligent Interaction, Lisbon, Portugal. pp. 230–241. Springer (2007)

35. Peters, C., Asteriadis, S., Karpouzis, K.: Investigating shared attention with a virtual agent using a gaze-based interface. Journal on Multimodal User Interfaces **3**(1-2), 119–130 (2010)
36. Petrak, B., Weitz, K., Aslan, I., André, E.: Let me show you your new home: Studying the effect of proxemic-awareness of robots on users first impressions. In: 28th IEEE International Conf. on Robot and Human Interactive Communication (RO-MAN), New Delhi, India. IEEE (2019)
37. Ritschel, H., Baur, T., André, E.: Adapting a robot's linguistic style based on socially-aware reinforcement learning. In: 26th IEEE International Symp. on Robot and Human Interactive Communication (RO-MAN), Lisbon, Portugal. pp. 378–384. IEEE (2017)
38. Rosa, H.: Resonanz: Eine Soziologie der Weltbeziehung. Suhrkamp Verlag (2016)
39. Schröder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., ter Maat, M., McKeown, G., Pammi, S., Pantic, M., Pelachaud, C., Schuller, B.W., de Sevin, E., Valstar, M.F., Wöllmer, M.: Building autonomous sensitive artificial listeners. IEEE Trans. on Affective Computing **3**(2), 165–183 (2012)
40. Silver, D.L., Yang, Q., Li, L.: Lifelong machine learning systems: Beyond learning algorithms. In: Lifelong Machine Learning, Papers from the 2013 AAAI Spring Symposium, Palo Alto, California, USA, March 25-27, 2013 (2013)
41. Skantze, G., Hjalmarsson, A., Oertel, C.: Turn-taking, feedback and joint attention in situated human-robot interaction. Speech Communication **65**, 50 – 66 (2014)
42. Strapparava, C., Valitutti, A., et al.: Wordnet affect: an affective extension of wordnet. In: 4th International Conf. on Language Resources and Evaluation, LREC, Lisbon, Portugal. pp. 1083–1086 (2004)
43. Takayama, L., Pantofaru, C.: Influences on proxemic behaviors in human-robot interaction. In: 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, St. Louis, MO, USA. pp. 5495–5502 (2009)
44. Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Schuller, B., Zafeiriou, S.: Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China. pp. 5200–5204 (2016)
45. Vogt, T., André, E., Bee, N.: Emovoice - A framework for online recognition of emotions from voice. In: André, E., Dybkjær, L., Minker, W., Neumann, H., Pieraccini, R., Weber, M. (eds.) Perception in Multimodal Dialogue Systems. pp. 188–199. Springer (2008)
46. Vogt, T., André, E., Wagner, J.: Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation. In: Peter, C., Beale, R. (eds.) Affect and Emotion in Human-Computer Interaction, pp. 75–91. Springer (2008)
47. Wagner, J., Lingenfelser, F., André, E.: Using phonetic patterns for detecting social cues in natural conversations. In: Interspeech, Stockholm. pp. 168–172 (2013)
48. Wagner, J., Schiller, D., Seiderer, A., André, E.: Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant? In: Interspeech, Hyderabad, India. pp. 147–151 (2018)
49. Weitz, K., Hassan, T., Schmid, U., Garbas, J.U.: Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable AI methods. tm-Technisches Messen **86**(7-8), 404–412 (2019)
50. Zhang, L., Wang, S., Liu, B.: Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8**(4), e1253 (2018)