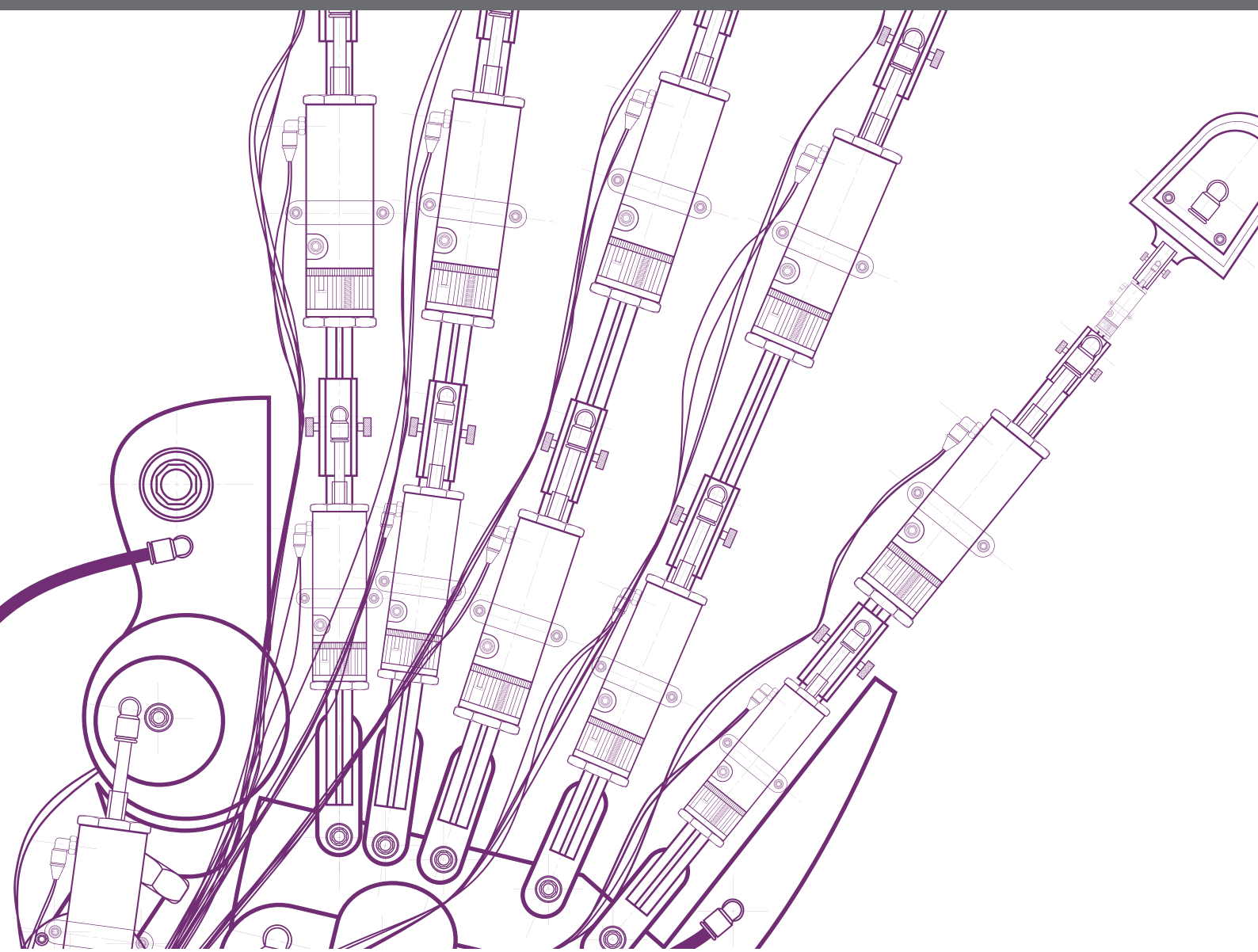
# COMPUTATIONAL APPROACHES FOR HUMAN-HUMAN AND HUMAN-ROBOT SOCIAL INTERACTIONS

EDITED BY: Vittorio Murino, Cigdem Beyan, Gentiane Venture and Agnieszka Wykowska

PUBLISHED IN: Frontiers in Robotics and AI



frontiers Research Topics

# frontiers



ISSN 1664-8714
ISBN 978-2-88963-807-9
DOI 10.3389/978-2-88963-807-9

## About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

# COMPUTATIONAL APPROACHES FOR HUMAN-HUMAN AND HUMAN-ROBOT SOCIAL INTERACTIONS

Topic Editors:
**Vittorio Murino,** Italian Institute of Technology (IIT), Italy
**Cigdem Beyan,** Italian Institute of Technology (IIT), Italy
**Gentiane Venture,** Tokyo University of Agriculture and Technology, Japan
**Agnieszka Wykowska,** Italian Institute of Technology (IIT), Italy

**Citation:** Murino, V., Beyan, C., Venture, G., Wykowska, A., eds. (2020). Computational Approaches for Human-Human and Human-Robot Social Interactions. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88963-807-9

# Table of Contents

frontiers in Robotics and AI

EDITORIAL
published: 30 April 2020
doi: 10.3389/frobt.2020.00055

# Editorial: Computational Approaches for Human-Human and Human-Robot Social Interactions

*Cigdem Beyan*[1]*, *Vittorio Murino*[1,2,3], *Gentiane Venture*[4] *and Agnieszka Wykowska*[5]

[1] *Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia, Genoa, Italy,* [2] *Department of Computer Science, University of Verona, Verona, Italy,* [3] *Huawei Technologies Ltd., Ireland Research Center, Dublin, Ireland,* [4] *GV Lab., Department of Mechanical Systems Engineering, Tokyo University of Agriculture and Technology, Tokyo, Japan,* [5] *Social Cognition in Human-Robot Interaction, Istituto Italiano di Tecnologia, Genoa, Italy*

**Keywords: human-human interaction, human robot interaction (HRI), machine learning, computational approaches, social interactions**

**Editorial on the Research Topic**

**Computational Approaches for Human-Human and Human-Robot Social Interactions**

## 1. INTRODUCTION

Non-verbal behaviors such as gaze, facial expressions, gestures, and vocal behavior carry significant information regarding human personality, emotions, engagement, intentions, action goals, and focus of attention. A large part of human communication takes place non-verbally (and often implicitly) during an explicit exchange of thoughts, attitudes, concerns, and feelings. Analyzing the basic principles of human communication through non-verbal signals is a long-standing research focus in cognitive and social psychology. However, the automatic realization of such analyses, especially by using machine learning (ML), or, in general, computational techniques, is a relatively unexplored avenue, although these techniques can be very efficient and effective.

Automatized detection and analysis of non-verbal social signals can be of particular relevance not only to human-human interaction (HHI) but also in human-robot interaction (HRI). Over the last decade, much research effort has been dedicated to improving robots' capabilities regarding perceiving, interacting, and cooperating with humans. Indeed, social HRI requires augmentation of robots' standard functionality with the ability to recognize and interpret human social signals in order to be able to engage naturally and intuitively with a human. Simultaneously, research efforts are being directed toward examining the human side of HRI, namely, the human mechanisms of social cognition in interactions with artificial agents (embodied robots specifically). This is crucial in order to understand how the human brain processes social signals coming from non-human agents and whether such agents can evoke mechanisms of social cognition in humans. ML techniques have also proved to be useful in this case to explore the patterns of neural and behavioral activity of the human counterparts.

This Research Topic is dedicated to exploring computational techniques for the analysis of non-verbal social signals in HHI as well as HRI. Specifically, we focus on ML methodologies, as well as other computational approaches for understanding non-verbal behavior and analyzing multi-modal data. It brings together ten selected papers that reflect some of the current computational approaches applied to HHI and HRI.

Bartlett et al. focus on movement analysis based on internal state identification. Video clips of social interactions, either the original scene or in the form of 2D body pose data, were shown to participants whose internal state perception was later assessed. These data were analyzed to

OPEN ACCESS

***Edited and reviewed by:***
*Alexandre Bernardino, Instituto Superior Técnico, Portugal*

***Correspondence:**
*Cigdem Beyan*
*cigdem.beyan@iit.it*

***Specialty section:***
*This article was submitted to Humanoid Robotics, a section of the journal Frontiers in Robotics and AI*

***Received:*** *27 February 2020*
***Accepted:*** *30 March 2020*
***Published:*** *30 April 2020*

***Citation:***
*Beyan C, Murino V, Venture G and Wykowska A (2020) Editorial: Computational Approaches for Human-Human and Human-Robot Social Interactions. Front. Robot. AI 7:55. doi: 10.3389/frobt.2020.00055*

determine whether the full scene clips were more informative than the 2D body pose. The results showed that participants were able to identify interaction imbalance, valence, and engagement independent of the types of videos. ML achieved similar performances as well, which can be interpreted as indicating that it can successfully decode and classify internal states using low-dimensional data.

Kory-Westlund and Breazeal investigate whether a social robot can increase children's rapport, positive emotion, acceptance, engagement, closeness, and learning. The robot entrained its speech and behavior to individual children and provided an appropriate backstory about its abilities. The data analysis performed showed that the robot's entrainment led children to show more positive emotions; it affected children's emulation of the robot's words in their own stories. Additionally, children who heard the robot's backstory were more accepting of it, find it more human-like, and agreed more to its requests.

Bloch et al. study the relevance of interpersonal synchrony (IS) for Autism Spectrum Disorder (ASD). IS is related to empathy and rapport and thus enables successful HHI, while individuals with ASD have difficulties with IS. The authors present a comprehensive review of IS in ASD and then propose a theoretical concept based on temporal processing of sensory input of interactions. Georgescu et al. present an ML method to study IS difficulties in ASD. IS between the head and upper body was quantified using Motion Energy Analysis, the results of which were used to train a Support Vector Machine to classify individuals with ASD and typically developed individuals.

Biancardi et al. propose a computational model that allows changes in the impression of warmth and competences of an embodied conversational agent that can interact with a human. The impressions of warmth and competence are changed in real-time to adapt to the human in order to maximize engagement. The system is tested as a museum guide, and it is shown that the hypothesis of warmth primacy may be valid.

Niewiadomski et al. focus on the analysis of social activities related to food and eating, as well as computational and technological approaches addressing such activities. The paper describes the approach of treating food-related activities as a social phenomenon that requires psychological and sociological analyses. It also presents problems that need to be tackled from the computational perspective, such as detection and recognition of food-related or eating activities.

Amiriparian et al. address interpersonal synchronization of acoustic signals during speech communication. They present an auto-encoder-based method trained on a large set of dyads across six different cultures. The results show that in all six cultures, partners tended to synchronize their speech, but inter-cultural differences were also observed.

Lammers et al. present a dataset of everyday actions expressing various emotions. This dataset was created based on motion capture data collected from human volunteers and then rendered into video files with a standardized, unified virtual character performing the actions. The stimulus material was then homogenized in terms of low-level physical features and tested for sufficiently high recognition rates.

Iwasaki et al. conducted an in-the-wild experiment, where a Pepper robot was in the role of a salesperson. The robot responded to various social situations and tried to attract customers' attention. Many customers ignored the robot's presence. However, if it managed to create a first impression of being capable of recognizing and appropriately responding to human behavior, it had higher chances of engaging customers. In a lab-environment experiment, the robot's "looking back behavior" was manipulated such that participants subjectively felt that they were being observed. The paper points out that for attracting the attention of users and maintaining their engagement, it is important to create an impression that a robot is aware of and reactive to the situational context, environment, and current state of the interaction.

Dinh et al. describe a framework for legible and safe robot behavior for HRI based on reinforcement learning. In a collaborative scenario, where both human and robot need to reach the same objects, the robot learns how to be legible to the human and how to avoid dynamic obstacles, thereby improving the safety of the human. This was tested in a virtual reality setup and in a physical HRI with a KUKA robot arm. The results showed that over the course of the experiment, participants efficiently learned how to predict robot movements and rated the robot's legibility increasingly higher. That improvement was better compared to a non-adaptive condition. The important advantage of this approach is that it is generalizable to other tasks.

## AUTHOR CONTRIBUTIONS

All co-authors drafted, revised the manuscript for important intellectual content, and approved the final version to be published.

**Conflict of Interest:** VM is employed by the company Huawei Technologies.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

PERSPECTIVE
published: 20 August 2019
doi: 10.3389/frobt.2019.00073

# INTRApersonal Synchrony as Constituent of INTERpersonal Synchrony and Its Relevance for Autism Spectrum Disorder

*Carola Bloch*[1,2]*, *Kai Vogeley*[1,3], *Alexandra L. Georgescu*[4] and *Christine M. Falter-Wagner*[2,5]*

[1] Department of Psychiatry and Psychotherapy, Medical Faculty, University of Cologne, Cologne, Germany, [2] Department of Psychiatry and Psychotherapy, Medical Faculty, Ludwig-Maximilians-University, Munich, Germany, [3] Institute of Neuroscience and Medicine (INM3), Research Center Jülich, Jülich, Germany, [4] Department of Psychology, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom, [5] Department of Psychology, Faculty of Human Science, University of Cologne, Cologne, Germany

INTERpersonal synchrony leads to increased empathy, rapport and understanding, enabling successful human-human interactions and reciprocal bonding. Research shows that individuals with Autism Spectrum Disorder (ASD) exhibit difficulties to INTERpersonally synchronize but underlying causes are yet unknown. In order to successfully synchronize with others, INTRApersonal synchronization of communicative signals appears to be a necessary prerequisite. We understand INTRApersonal synchrony as an implicit factor of INTERpersonal synchrony and therefore hypothesize that atypicalities of INTRApersonal synchrony may add to INTERpersonal synchrony problems in ASD and their interaction partners. In this perspective article, we first review evidence for INTERpersonal dissynchrony in ASD, with respect to different approaches and assessment methods. Second, we draft a theoretical conceptualization of INTRApersonal dissynchrony in ASD based on a temporal model of human interaction. We will outline literature indicating INTRApersonal dissynchrony in ASD, therefore highlighting findings of atypical timing functions and findings from clinical and behavioral studies that indicate peculiar motion patterns and communicative signal production in ASD. Third, we hypothesize that findings from these domains suggest an assessment and investigation of temporal parameters of social behavior in individuals with ASD. We will further propose specific goals of empirical approaches on INTRApersonal dissynchrony. Finally we present implications of research on INTRApersonal timing in ASD for diagnostic and therapeutic purposes, what in our opinion warrants the increase of research efforts in this domain.

**Keywords: human-human interaction, INTERpersonal synchrony, INTRApersonal synchrony, timing, non-verbal behavior, autism spectrum disorder**

OPEN ACCESS

***Edited by:***
*Cigdem Beyan,*
*Istituto Italiano di Tecnologia, Italy*

***Reviewed by:***
*Chung Hyuk Park,*
*George Washington University, United States*
*Nicola Vanello,*
*University of Pisa, Italy*

***Correspondence:**
*Carola Bloch*
*carola.bloch@med.uni-muenchen.de*
*Christine M. Falter-Wagner*
*christine.falter@med.uni-muenchen.de*

***Specialty section:***
*This article was submitted to Humanoid Robotics, a section of the journal Frontiers in Robotics and AI*

***Received:*** *22 March 2019*
***Accepted:*** *30 July 2019*
***Published:*** *20 August 2019*

***Citation:***
*Bloch C, Vogeley K, Georgescu AL and Falter-Wagner CM (2019) INTRApersonal Synchrony as Constituent of INTERpersonal Synchrony and Its Relevance for Autism Spectrum Disorder. Front. Robot. AI 6:73. doi: 10.3389/frobt.2019.00073*

## 1. INTERPERSONAL DISSYNCHRONY

> "*Terms such as interactional synchrony, non-verbal mirroring, shared rhythmicity, motor mimicry or chameleon effect embrace the underlying dimension of coordination between two or more individuals in the domain of nonverbal action*" (Ramseyer and Tschacher, 2008, p.332).

Across different terminology INTERpersonal synchrony describes the phenomenon that people automatically align behavior while interacting. This is thought to strengthen their social bond

by means of increased rapport (LaFrance, 1979; Tickle-Degnen and Rosenthal, 1990; Lakin and Chartrand, 2003; Vacharkulksemsuk and Fredrickson, 2012), mutual affiliation (Hove and Risen, 2009), enhanced mentalizing (Baimel et al., 2018), successful joint action (Valdesolo et al., 2010; Lorenz et al., 2014), as well as empathy (Behrends et al., 2012; Koehne et al., 2016). Autism Spectrum Disorder (ASD) is defined as a neurodevelopmental disorder that entails difficulties in social communication and interaction together with repetitive behaviors and restricted interests (American Psychiatric Association, 2013) and there exists evidence for INTERpersonal dissynchrony of individuals with ASD with interaction partners.

INTERpersonal synchrony as a dependent variable in groups of healthy control persons was measured in reference to parameters in a dynamical model of human movements dynamical model of human movements (Haken et al., 1985). Those studies measured coordinated movements between two individuals in terms of reduced changes in relative phase angles between reference points in two oscillating systems (Richardson et al., 2007; Schmidt and Richardson, 2008; Romero et al., 2015). With respect to individuals with ASD one study found less coordination of movements measured by the alignment of phase angles between two rocking chairs (Marsh et al., 2013). Similarly, individuals with ASD synchronized pendulum swings less with their parents (Fitzpatrick et al., 2016).

Fitzpatrick et al. (2017a) separately investigated performance in intentional vs. spontaneous synchrony tasks and found lower INTERpersonal coherence scores in both domains for children with ASD. Additionally, the authors found distinct cognitive mechanisms underlying both kinds of alignment problems (Fitzpatrick et al., 2017b). In a naturalistic setting, spontaneous INTERpersonal synchrony was measured by coherence of body motion of two interaction partners in predefined regions of interest, thereby not focusing on external oscillators or specific limbs, rather on general body motion (Ramseyer and Tschacher, 2011; Romero et al., 2015). So-called Motion Energy Analysis (MEA) (Ramseyer and Tschacher, 2011) calculates cross-correlation time series of pixel changes from video-recorded interactions as an indicator for coordinated movements. Noel et al. (2018) used MEA in their study and showed that individuals with ASD exhibited less INTERpersonal synchrony and less complex movements in an interview setting.

Relevant for understanding INTERpersonal synchrony are also joint action paradigms, in which participants have to conduct actions that require consideration of another person's perspective and movement affordances in the course of motion planning. When assessing the motor anticipation of a partner's grip comfort when passing objects, participants with ASD showed more variable grip positions indicating atypical social motor planning (Gonzalez et al., 2013). Moreover, with increasing severity of ASD traits, participants modulated grip movements less in adaption to a partner's movements, but performed well in a non-social replication task indicating deficits only for the social domain (Curioni et al., 2017). Other studies found less grip-to-reach positions that enhanced end-state comfort for the partner (Scharoun and Bryden, 2016; Studenka et al., 2017) and more variable reaction times, slower movements and more movement dissynchrony with an interaction partner (Fulceri et al., 2018).

Besides investigations of alignments of whole body or limb movements, mutual gaze and the establishment of joint attention are of particular interest for INTERpersonal coupling processes (Emery, 2000; Senju and Johnson, 2009). Gaze behavior is the first non-verbal source for the coordination of behavior between newborn and parent and therefore a driving force for the development of non-verbal reciprocity (Feldman, 2007). Empirical evidence for atypical gaze behavior and atypical processing of gaze cues in individuals with ASD is now overwhelming, in particular early aversion of social gaze (Jones and Klin, 2013), altered attention preferences for social cues in form of gaze avoidance (Madipakkam et al., 2017) and less contact or involvement evoked by direct gaze (Schwartz et al., 2010). Gaze idiosyncrasies were already found in children with ASD (Jones and Klin, 2013) and are still present in adults (Schwartz et al., 2010; Georgescu et al., 2013; Madipakkam et al., 2017; Caruana et al., 2018) and are not caused by oculomotor disfunctions (Caruana et al., 2018). In conclusion, empirical evidence from several domains indicate reduced body motion alignment, less anticipation of other persons' kinematics in motor planning as well as atypical social gaze as features of individuals with ASD that contribute to INTERpersonal dissynchrony (see **Table 1** for an overview).

## 2. INTRAPERSONAL DISSYNCHRONY IN ASD

In their social entrainment model, McGrath and Kelly (1986) consider social interaction in terms of temporal patterns or rhythms in behavior. This model states that endogenous (i.e. individual) rhythms in behavior become temporally aligned in phase and period in the course of interaction. This implies the emergence of systematic temporal patterns of verbal and non-verbal turn-taking during INTERpersonal encounters. Based on this, one can assume that there exist temporal windows of signal production that are critical for communication efficiency and INTERpersonal alignment. From an individual perspective, communication signals are composed of various non-verbal sources (e.g., gaze and gestures). These need to be coordinated with each other and with verbal output to achieve the intended communicative effects. We define INTRApersonal synchrony as the temporal coordination of communication signals in a socially informative manner. In the following, we will review evidence of atypical temporal processing and movement patterns in ASD. We will then introduce the idea that those peculiarities may be related to individuals with ASD missing the assumed temporal windows for producing socially effective communication signals.

### 2.1. Temporal Processing in ASD

Temporal processing of sensory input seems to be altered in individuals with ASD. For instance, in a perceptual simultaneity task individuals with ASD judged the presentation of two visual stimuli to be temporally asynchronous for smaller stimulus onset asynchronies compared to typically developed (TD) control participants (Falter et al., 2012a). Further empirical evidence

**TABLE 1 |** Studies on INTERpersonal synchrony in ASD.

| Study | *N* (m;f) | Age *M*(*SD*) | Paradigm |
|---|---|---|---|
| **OSZILLATION** | | | |
| Fitzpatrick et al. (2016) | 9 (8;1) | 13.7 (1.3) | Pendulum task |
| Marsh et al. (2013) | 8 (6;2) | 6.2 (1.2) | Rocking chair task |
| **BODY ALIGNMENT** | | | |
| Fitzpatrick et al. (2017a,b) | 45 (39;6) | 8.6 (4.8) | Social motor synchronization tasks and cognitive measures |
| Noel et al. (2018) | 12 (8;4) | 12.2 (3.8) | Multisensory temporal binding task and MEA |
| **JOINT ACTION** | | | |
| Curioni et al. (2017) | 16 (13;3) | 26.1 (/) | Grasping objects in social vs. non-social condition |
| Fulceri et al. (2018) | 11 (10;1) | 7.8 (1.3) | Joint action task with clear and unclear end point |
| Gonzalez et al. (2013) | 10 (9;1) | 32.7 (10.8) | Helping partner by passing objects |
| Scharoun and Bryden (2016) | 14 (9;5) | 8.6 (/) | Grasp-to-reach task with experimenter |
| Studenka et al. (2017) | 5 (3;2) | 9.8 (/) | Narrative task and motor perspective taking |
| **SOCIAL GAZE** | | | |
| Caruana et al. (2018) | 17 (11;6) | 26.5 (11.9) | Initiating and responding to joint attention |
| Jones and Klin (2013) | 11 (11;0) | 0.2–0.4 | Gaze preferences in longitudinal study design |
| Madipakkam et al. (2017) | 14 (8;4) | 35.4 (2.3) | Unconscious reactions to direct and averted gaze |
| Schwartz et al. (2010) | 20 (11;9) | 39.3 (9.2) | Socioaffective effects of direct gaze |

*All studies recruited age-matched control groups.*

shows an enhanced temporal parsing of auditory (Jones et al., 2009) and visual events (Falter et al. 2013; but see Isaksson et al. 2018), lower hit rates for the detection of differences in temporal intervals between auditory signals (Falter et al., 2012b), atypical judgment and reproduction of durations (Szelag et al., 2004) and wider multisensory temporal binding windows for simultaneity judgments (Noel et al., 2018). All of these findings support the notion of atypical temporal processing in ASD, possibly associated with a detail-focused, less holistic cognitive style as postulated in the 'Weak Central Coherence Theory' of autism (Happé and Frith, 2006). Atypical temporal processing also manifests in higher level processes, such as the subjective experience of time. Allman et al. (2014) in this context argue that stereotypical behavior patterns and behavioral routines serve the structuring of subjective time experience in ASD which compensates for atypical internal timing functions. In line with that are results of a high tendency in ASD to rely on self-structured routines and repetitive behavior to control bottom-up perceptual input, thereby generating experiences of timelessness or "flow" (Vogel et al., 2018a,b).

Atypical temporal processing in ASD most likely influences behavior as well, given that sensorimotor frameworks propose feedback loops of sensory and motor systems (Wolpert et al., 2003; Torres et al., 2013). In this line, Gowen and Miall (2005) found atypical motor timing in an ASD sample, namely faster and more variable responses in a finger tapping task and results were replicated by Isaksson et al. (2018). In addition to this evidence for altered motor timing, behavioral research underpins the assumption that individuals with ASD exhibit atypical movement patterns, as shown in the following literature.

## 2.2. Motor Production in ASD

Clinically, "clumsiness" in motor production is a major feature of autism (Asperger, 1944). Although still a secondary criterion for diagnosis, Parma and de Marchena (2015) argue that atypical motor patterns in ASD need to be further investigated as they occur across the spectrum and may constitute a possible diagnostic marker. In this context a study by Anzulewicz et al. (2016) successfully discriminated children with ASD from TD children by a machine learning algorithm that deployed motor variables from a gaming task with a touch screen. Children with ASD exhibited significantly faster movements with peculiar pressure patterns. Other studies further highlight jerky limb movements (Cook et al., 2013), atypical gait (Barrow et al., 2011; Kindregan et al., 2015; Dufek et al., 2017; Eggleston et al., 2017), enhanced postural sway (Gowen and Miall, 2005; Doumas et al., 2016) and enhanced variability in motor output (Brincker and Torres, 2013; Gowen and Hamilton, 2013; Parma and de Marchena, 2015; Kaur et al., 2018). A meta-analysis by Fournier et al. (2010) included 41 studies on motor coordination, motor impairment, arm movement, gait, or postural stability. They found a significant effect indicating weaker motor performance in ASD individuals, independent from symptom severity. A review by Gowen and Hamilton (2013) systematically inspected approaches on motor abilities in ASD on the background of a computational model that postulates intermediate cognitive steps of motor processing. The authors suggest poorer integration of sensory input for motor planning as well as increased variability in motor output or "motor noise" as integral characteristics in ASD.

Taken together, there is cumulative evidence for atypical movement patterns in terms of reduced coordination and greater variability in motor production. Together with evidence for peculiarities in motor timing (Gowen and Miall, 2005; Isaksson et al., 2018), movement aberrations may influence INTERpersonal communication because communicative signals as motor acts dissociate from typical signal production with respect to temporal emergence.

### 2.3. INTRApersonal Dissynchrony in Interactions

The Autism Diagnostic Observation Schedule (ADOS) as a standard diagnostoc tool targets the coordination of communication channels as a symptom of ASD (Lord et al., 2000). Regarding social contexts there exists research in atypical gesture production in individuals with ASD. In a study by de Marchena and Eigsti (2010), the authors counted gesture usage and coded types of gestures in a narrative task with ASD and TD adolescents. They found no differences in frequency and the kind of gesture used but atypical timing of gestures related to co-occurring speech led to reduced ratings of communication quality in naive observers. In another study on gesture usage in infants with ASD, Colgan et al. (2006) also found no differences in the frequency of gestures, but infants with ASD showed a reduced variety of gestures compared to TD control participants. That is in line with findings of less complexity in non-verbal behavior found by Noel et al. (2018). These results indicate that it is not the quantity of communicative signals that leads to the known communication difficulties but the quality of signals and how they fit in the interactional flow. In the social context it is noteworthy to mention, that the temporal thresholds of perceptual simultaneity (that indicated enhanced temporal parsing of sensory events) in ASD were significantly correlated with difficulties in the communications domain, especially when difficulties were assessed with items encompassing the use of communicative gestures and social imitation (Falter et al., 2012b). Isaksson et al. (2018) likewise found an association of enhanced temporal parsing and symptom severity in communication and social interaction. Noel et al. (2018) further demonstrated that multisensory temporal binding windows correlated with INTERpersonal synchrony in TD participants but not in participants with ASD, indicating distinctive associations in the multisensory temporal domain.

Those findings imply that temporal processing in ASD may be associated with the reduced INTERpersonal alignments. If this association is mediated by atypical social signal timing, is targeted by our proposed perspective on INTRApersonal dissynchrony in ASD.

## 3. PERSPECTIVE ON FUTURE RESEARCH

The temporal model of social interactions by McGrath and Kelly (1986) implies that synchronous alignments require mutual responsiveness and coordinated signal production. Individuals with ASD exhibit atypical temporal processing and motor patterns, what most likely disrupts the emergence or maintenance of systematic INTERpersonal coupling. In line with that, we argue that future research needs to extend findings of deviant motor timing (Gowen and Miall, 2005; Barakova and Chonnaparamutt, 2009; Isaksson et al., 2018) to the domain of socially expressive behavior and investigate the impact of INTRApersonal dissynchrony on interactions.

Therefore we suggest approaches on INTRApersonal dissynchrony should pursue two consecutive goals. First, the aim is to quantify temporal deviations in communication behavior in ASD and to find critical temporal windows of INTRApersonal synchronous signal production. State-of-the-art techniques, such as motion capture, eye tracking and video tracking should be used to assess time series of communication behavior. This allows the investigation of peculiarities of signal timing in multiple communication contexts. Combining such techniques makes it possible to assess the temporal coordination of separate signal sources (e.g., gaze, gestures, facial expressions, speech) in terms of relational signal onsets, durations and end points. Thereby one may gain insights into the temporal composition of individual signal streams. On the background of findings of enhanced motor variability (Brincker and Torres, 2013; Gowen and Hamilton, 2013; Kaur et al., 2018) the investigation of measures of dispersion in ASD samples will be of particular interest. Furthermore, implementing perceptual timing tasks may lead to insights in functional relations of temporal processing and social motor timing in ASD. Thus, the questions if timing of communicative channels is affected by a general sensory timing deficit or by social contexts or both can be addressed by comparing task performance in social and non-social tasks of varying sensory complexity. With regard to the neurophysiological framework of predictive coding, it would be highly interesting to investigate, if INTRApersonal dissynchrony also manifests in EEG patterns with social cues produced by individuals with ASD possibly missing predictive timing windows (Arnal and Giraud, 2012). As physical arousal and stress may be enhanced in social tasks in ASD participants, further assessment of heart rate and skin conductance constitute important covariates.

A subsequent goal would be to analyze the perception of idiosyncratic communication patterns, here targeting causal effects of INTRApersonal dissynchrony on INTERpersonal outcomes. Therefore, motion capture data should be used to animate virtual characters in order to create ecologically valid and standardized stimulus material for perception studies (Bente and Krämer, 2002; Georgescu et al., 2014; Pan and Hamilton, 2018). By assessing impression, evaluation and recognition of altered signal production in ASD, one may draw causal conclusions for deficits in social interactions. Dependent variables should be included in such perception studies that are critical for the quality of the produced signal, e.g., communication efficiency and measures of INTERpersonal bonding, e.g., likeability. Creating an "autistic avatar" would allow experimental manipulation of movement parameters under high experimental control. It is of great relevance to illuminate the perspective of the interaction partner to fully understand developmental pathways and resulting communication deficits. There is evidence that TD participants show poorer performance in decoding expressive movements generated by individuals with ASD, indicating reciprocal lack of mentalization (Edey et al., 2016). On presentation of short video clips or still frames of individuals with ASD, independent raters judged individuals with ASD less favorably and reported less motivation to socially approach them (Sasson et al., 2017). An avatar that exhibits specific autistic movement patterns could therefore be employed for research into reciprocal effects of INTERpersonal dissynchrony as well as for training of staff to improve interaction with patients.

This approach on INTRApersonal dissynchrony in ASD potentially has further implications for diagnosis and therapy. In

this context, measures of INTRApersonal communication signal coordination could serve as implicit measures that can be used for diagnostic purposes. Implicit diagnostic tools are strongly needed to account for symptomatical heterogeneity in ASD. Subjective observational tools are based on clinical observations or self-report with limited objectivity, especially in adults given behavioral adjustment throughout their lives. Time series data of motion patterns may be used for diagnostic purposes, e.g. supported by machine learning (Georgescu et al., 2019). Our recent work shows that automatized classification of ASD from non-ASD is possible on the mere basis of motion energy assessed using video analysis (ibid.). Specifically motion capture data of INTRApersonal movement parameters is likely to further increase classification power due to richer data retrieval.

An INTRApersonal approach has conceivable implications for the field of robotics in autism research as temporal parameters of signal production may inform models of interactive robotic behavior.

Recent research of human-robot interaction (HRI) with children with ASD revealed positive effects, as robots attract attention and elicit novel behavior while social complexity can be controlled for (Duquette et al., 2008; Scassellati et al., 2012; Srinivasan et al., 2016). The AURORA project (AUtonomous RObotic platform as a Remedial tool for children with Autism) used the robot "Robota," which resembles a human doll and is able to exhibit interactive movements via video-, speech-, and motion-tracking (Dautenhahn and Billard, 2002). The project showed that "Robota" could serve a mediating role for eliciting joint attention in triadic human-human-robot interactions and elicited spontaneous imitation behavior (Dautenhahn and Werry, 2004; Robins et al., 2004, 2005), thereby potentially reinforcing social skills.

Amplifying joint attention via HRI is of great potential for endorsing social engagement and reciprocity in children with ASD, but the literature is not yet fully convincing. In their study, Anzalone et al. (2014) found that children with ASD were less responsive to joint attention initiatives by the social robot "Nao" and both groups responded less to the robot compared to a human therapist. Another approach investigated interactions of four children with "Nao" and again found mixed results, including facilitated joint attention only for one child (Tapus et al., 2012). Possibly, the design of the social robot with respect to its anthropomorphism may be highly important for eliciting and reinforcing social interactive behavior in children with ASD, for example a realistic eye design in joint attention paradigms (Admoni and Scassellati, 2017; Luria et al., 2018). However, the target of the intervention is not yet properly defined.

Our perspective suggests that individuals with ASD exhibit social interaction in different ways (e.g. peculiar temporal parameters of communicative signal production). Socially interactive robots generally need to be able to recognize communicative signals and exhibit appropriate reactions (Breazeal et al., 2016). Thus, models of robotic behavior could be adjusted to temporal parameters of signal production in ASD in order to reinforce reciprocity, similar to computational approaches in Admoni and Scassellati (2014) or Barakova and Chonnaparamutt (2009). Such an adjustment may enhance compliance and responsiveness of individuals with ASD toward the robotic interaction partner. Furthermore, given that Gowen and Hamilton (2013) suggest intact motor learning in ASD, parameters of typical signal timing may be used for robotic interventions that aim to train proper timing in communicative signal coordination, thereby providing a possible quantitative outcome measure of treatment success. Building upon findings of the AURORA project, the creation of HRI scenarios in which human-like robots serve as interactive tutors for training specific communicative skills (e.g. joint attention) are promising. Creating game-based robot interactions that prompt spontaneous imitation of properly coordinated signals could be a great opportunity to support children in their development of non-verbal skills.

There exist a number of aspects that need to be considered when planning approaches on INTRApersonal dissycnhrony in ASD. One potentially confounding factor when measuring INTRApersonal synchrony lies in the distinction between spontaneously and voluntarily produced behavior, as different cognitive processes are thought to underlie these processes (Frith and Frith, 2008; Torres et al., 2013). Thus, future studies should investigate how INTRApersonal dissynchrony differs under the impact of explicit instructions or implicit and natural task conditions. Furthermore, highly standardized study designs that strictly control sensory surroundings are crucial for studying INTRApersonal synchronization, given deviant sensory processing may contribute to behavioral variability in ASD.

Further research should broaden this approach to other psychiatric disorders that entail INTRApersonal coordination peculiarities like schizophrenia (Walther et al., 2015) or depression (Schrijvers et al., 2008). But especially for ASD, we think that a perspective on INTRApersonal dissynchrony is fundamentally relevant for understanding INTERpersonal difficulties. A quantification of temporally atypical coordination of communication signals in ASD is an important explanatory approach that potentially informs diagnosis as well as intervention programs.

## AUTHOR CONTRIBUTIONS

In accordance with theoretical discussions with CF-W and KV, CB wrote the first manuscript version. AG contributed literature and theoretical ideas. All authors read and modified the manuscript several times. All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

## FUNDING

CB was supported by a German Research Foundation grant (granted to CF-W; FA876/3-1). CF-W was supported by a Bavarian Gender Equality Grant.

## REFERENCES


Admoni, H., and Scassellati, B. (2014). "Toward a data-driven generative behavior model for human-robot interaction," in *Proceedings of the 2014 Workshop on Mobile Augmented Reality and Robotic Technology-Based Systems* (Bretton Woods, NH: ACM), 19–20.

Admoni, H., and Scassellati, B. (2017). Social eye gaze in human-robot interaction: a review. *J. Hum. Robot Interact.* 6, 25–63. doi: 10.5898/JHRI.6.1.Admoni

Allman, M. J., Yin, B., and Meck, W. H. (2014). "Time in the psychopathological mind," in *Subjective Time: The Philosophy, Psychology, and Neuroscience of Temporality*, eds V. Arstila and D. Lloyd (Cambridge, MA: MIT Press), 637–654.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. Washington, DC: American Psychiatric Pub.

Anzalone, S. M., Tilmont, E., Boucenna, S., Xavier, J., Jouen, A.-L., Bodeau, N., et al. (2014). How children with autism spectrum disorder behave and explore the 4-dimensional (spatial 3d+ time) environment during a joint attention induction task with a robot. *Res. Autism Spectr. Disord.* 8, 814–826. doi: 10.1016/j.rasd.2014.03.002

Anzulewicz, A., Sobota, K., and Delafield-Butt, J. T. (2016). Toward the autism motor signature: gesture patterns during smart tablet gameplay identify children with autism. *Sci. Rep.* 6:31107. doi: 10.1038/srep31107

Arnal, L. H., and Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends Cognit. Sci.* 16, 390–398. doi: 10.1016/j.tics.2012.05.003

Asperger, H. (1944). Die "Autistischen Psychopathen" im Kindesalter. *Arch. Psychiatr. Nervenkr.* 117, 76–136. doi: 10.1007/BF01837709

Baimel, A., Birch, S. A., and Norenzayan, A. (2018). Coordinating bodies and minds: behavioral synchrony fosters mentalizing. *J. Exp. Soc. Psychol.* 74, 281–290. doi: 10.1016/j.jesp.2017.10.008

Barakova, E. I., and Chonnaparamutt, W. (2009). Timing sensory integration. *IEEE Robot. Autom. Mag.* 16, 51–58. doi: 10.1109/MRA.2009.933626

Barrow, W. J., Jaworski, M., and Accardo, P. J. (2011). Persistent toe walking in autism. *J. Child Neurol.* 26, 619–621. doi: 10.1177/0883073810385344

Behrends, A., Müller, S., and Dziobek, I. (2012). Moving in and out of synchrony: a concept for a new intervention fostering empathy through interactional movement and dance. *Arts Psychother.* 39, 107–116. doi: 10.1016/j.aip.2012.02.003

Bente, G., and Krämer, N. C. (2002). "Virtuelle Gesten: VR-Einsatz in der nonverbalen kommunikationsforschung" in *Virtuelle Realitäten*, *Vol. 5*, eds G. Bente, N. C. Krämer, and A. Petersen (Göttingen: Hogrefe), 81–107.

Breazeal, C., Dautenhahn, K., and Kands, T. (2016). "Social robotics." in *Springer Handbook of Robotics*, eds B. Siciliano and O. Khatib (New York, NY: Springer International Publishing), 1935–1972.

Brincker, M., and Torres, E. B. (2013). Noise from the periphery in autism. *Front. Integr. Neurosci.* 7:34. doi: 10.3389/fnint.2013.00034

Caruana, N., Stieglitz Ham, H., Brock, J., Woolgar, A., Kloth, N., Palermo, R., et al. (2018). Joint attention difficulties in autistic adults: an interactive eye-tracking study. *Autism* 22, 502–512. doi: 10.1177/1362361316676204

Colgan, S. E., Lanter, E., McComish, C., Watson, L. R., Crais, E. R., and Baranek, G. T. (2006). Analysis of social interaction gestures in infants with autism. *Child Neuropsychol.* 12, 307–319. doi: 10.1080/09297040600701360

Cook, J. L., Blakemore, S.-J., and Press, C. (2013). Atypical basic movement kinematics in autism spectrum conditions. *Brain* 136, 2816–2824. doi: 10.1093/brain/awt208

Curioni, A., Minio-Paluello, I., Sacheli, L. M., Candidi, M., and Aglioti, S. M. (2017). Autistic traits affect interpersonal motor coordination by modulating strategic use of role-based behavior. *Mol. Autism* 8:23. doi: 10.1186/s13229-017-0141-0

Dautenhahn, K., and Billard, A. (2002). "Games children with autism can play with Robota, a humanoid robotic doll," in *Universal Access and Assistive Technology*, eds S. Keates, P. M. Langdon, P. J. Clarkson, and P. Robinson (London, UK: Springer), 179–190.

Dautenhahn, K., and Werry, I. (2004). Towards interactive robots in autism therapy: background, motivation and challenges. *Pragmat. Cognit.* 12, 1–35. doi: 10.1075/pc.12.1.03dau

de Marchena, A., and Eigsti, I.-M. (2010). Conversational gestures in autism spectrum disorders: asynchrony but not decreased frequency. *Autism Res.* 3, 311–322. doi: 10.1002/aur.159

Doumas, M., McKenna, R., and Murphy, B. (2016). Postural control deficits in autism spectrum disorder: the role of sensory integration. *J. Autism Dev. Disord.* 46, 853–861. doi: 10.1007/s10803-015-2621-4

Dufek, J. S., Eggleston, J. D., Harry, J. R., and Hickman, R. A. (2017). A comparative evaluation of gait between children with autism and typically developing matched controls. *Med. Sci.* 5:1. doi: 10.3390/medsci5010001

Duquette, A., Michaud, F., and Mercier, H. (2008). Exploring the use of a mobile robot as an imitation agent with children with low-functioning autism. *Autonom. Robots* 24, 147–157. doi: 10.1007/s10514-007-9056-5

Edey, R., Cook, J., Brewer, R., Johnson, M. H., Bird, G., and Press, C. (2016). Interaction takes two: typical adults exhibit mind-blindness towards those with autism spectrum disorder. *J. Abn. Psychol.* 125:879. doi: 10.1037/abn0000199

Eggleston, J. D., Harry, J. R., Hickman, R. A., and Dufek, J. S. (2017). Analysis of gait symmetry during over-ground walking in children with autism spectrum disorder. *Gait Posture* 55, 162–166. doi: 10.1016/j.gaitpost.2017.04.026

Emery, N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neurosci. Biobehav. Rev.* 24, 581–604. doi: 10.1016/S0149-7634(00)00025-7

Falter, C. M., Braeutigam, S., Nathan, R., Carrington, S., and Bailey, A. J. (2013). Enhanced access to early visual processing of perceptual simultaneity in autism spectrum disorders. *J. Autism Dev. Disord.* 43, 1857–1866. doi: 10.1007/s10803-012-1735-1

Falter, C. M., Elliott, M. A., and Bailey, A. J. (2012a). Enhanced visual temporal resolution in autism spectrum disorders. *PLoS ONE* 7:e32774. doi: 10.1371/journal.pone.0032774

Falter, C. M., Noreika, V., Wearden, J. H., and Bailey, A. J. (2012b). More consistent, yet less sensitive: interval timing in autism spectrum disorders. *Q. J. Exp. Psychol.* 65, 2093–2107. doi: 10.1080/17470218.2012.690770

Feldman, R. (2007). Parent–infant synchrony: biological foundations and developmental outcomes. *Curr. Direct. Psychol. Sci.* 16, 340–345. doi: 10.1111/j.1467-8721.2007.00532.x

Fitzpatrick, P., Frazier, J. A., Cochran, D. M., Mitchell, T., Coleman, C., and Schmidt, R. (2016). Impairments of social motor synchrony evident in autism spectrum disorder. *Front. Psychol.* 7:1323. doi: 10.3389/fpsyg.2016.01323

Fitzpatrick, P., Romero, V., Amaral, J. L., Duncan, A., Barnard, H., Richardson, M. J., et al. (2017a). Evaluating the importance of social motor synchronization and motor skill for understanding autism. *Autism Res.* 10, 1687–1699. doi: 10.1002/aur.1808

Fitzpatrick, P., Romero, V., Amaral, J. L., Duncan, A., Barnard, H., Richardson, M. J., et al. (2017b). Social motor synchronization: insights for understanding social behavior in autism. *J. Autism Dev. Disord.* 47, 2092–2107. doi: 10.1007/s10803-017-3124-2

Fournier, K. A., Hass, C. J., Naik, S. K., Lodha, N., and Cauraugh, J. H. (2010). Motor coordination in autism spectrum disorders: a synthesis and meta-analysis. *J. Autism Dev. Disord.* 40, 1227–1240. doi: 10.1007/s10803-010-0981-3

Frith, C. D., and Frith, U. (2008). Implicit and explicit processes in social cognition. *Neuron* 60, 503–510. doi: 10.1016/j.neuron.2008.10.032

Fulceri, F., Tonacci, A., Lucaferro, A., Apicella, F., Narzisi, A., Vincenti, G., et al. (2018). Interpersonal motor coordination during joint actions in children with and without autism spectrum disorder: the role of motor information. *Res. Dev. Disabil.* 80, 13–23. doi: 10.1016/j.ridd.2018.05.018

Georgescu, A. L., Koehler, J. C., Weiske, J., Vogeley, K., Koutsouleris, N., and Falter-Wagner, C. (2019). Machine learning approaches to study social interaction difficulties in ASD. *Front. Robot. AI*.

Georgescu, A. L., Kuzmanovic, B., Roth, D., Bente, G., and Vogeley, K. (2014). The use of virtual characters to assess and train non-verbal communication in high-functioning autism. *Front. Hum. Neurosci.* 8:807. doi: 10.3389/fnhum.2014.00807

Georgescu, A. L., Kuzmanovic, B., Schilbach, L., Tepest, R., Kulbida, R., Bente, G., et al. (2013). Neural correlates of "social gaze" processing in high-functioning autism under systematic variation of gaze duration. *Neuroimage Clin.* 3, 340–351. doi: 10.1016/j.nicl.2013.08.014

Gonzalez, D. A., Glazebrook, C. M., Studenka, B. E., and Lyons, J. (2013). Motor interactions with another person: do individuals with autism spectrum disorder plan ahead? *Front. Integr. Neurosci.* 7:23. doi: 10.3389/fnint.2013.00023

Gowen, E., and Hamilton, A. (2013). Motor abilities in autism: a review using a computational context. *J. Autism Dev. Disord.* 43, 323–344. doi: 10.1007/s10803-012-1574-0

Gowen, E., and Miall, R. C. (2005). Behavioural aspects of cerebellar function in adults with asperger syndrome. *Cerebellum* 4, 279–289. doi: 10.1080/14734220500355332

Haken, H., Kelso, J. A., and Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biol. Cybernet.* 51, 347–356. doi: 10.1007/BF00336922

Happé, F., and Frith, U. (2006). The weak coherence account: detail-focused cognitive style in autism spectrum disorders. *J. Autism Dev. Disord.* 36, 5–25. doi: 10.1007/s10803-005-0039-0

Hove, M. J., and Risen, J. L. (2009). It's all in the timing: interpersonal synchrony increases affiliation. *Soc. Cognit.* 27, 949–960. doi: 10.1521/soco.2009.27.6.949

Isaksson, S., Salomäki, S., Tuominen, J., Arstila, V., Falter-Wagner, C. M., and Noreika, V. (2018). Is there a generalized timing impairment in autism spectrum disorders across time scales and paradigms? *J. Psychiatr. Res.* 99, 111–121. doi: 10.1016/j.jpsychires.2018.01.017

Jones, C. R., Happé, F., Baird, G., Simonoff, E., Marsden, A. J., Tregay, J., et al. (2009). Auditory discrimination and auditory sensory behaviours in autism spectrum disorders. *Neuropsychologia* 47, 2850–2858. doi: 10.1016/j.neuropsychologia.2009.06.015

Jones, W., and Klin, A. (2013). Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. *Nature* 504:427. doi: 10.1038/nature12715

Kaur, M., Srinivasan, S. M., and Bhat, A. N. (2018). Comparing motor performance, praxis, coordination, and interpersonal synchrony between children with and without autism spectrum disorder (ASD). *Res. Dev. Disabil.* 72, 79–95. doi: 10.1016/j.ridd.2017.10.025

Kindregan, D., Gallagher, L., and Gormley, J. (2015). Gait deviations in children with autism spectrum disorders: a review. *Autism Res. Treat.* 2015:741480. doi: 10.1155/2015/741480

Koehne, S., Hatri, A., Cacioppo, J. T., and Dziobek, I. (2016). Perceived interpersonal synchrony increases empathy: insights from autism spectrum disorder. *Cognition* 146, 8–15. doi: 10.1016/j.cognition.2015.09.007

LaFrance, M. (1979). Nonverbal synchrony and rapport: analysis by the cross-lag panel technique. *Soc. Psychol. Q.* 42, 66–70. doi: 10.2307/3033875

Lakin, J. L., and Chartrand, T. L. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychol. Sci.* 14, 334–339. doi: 10.1111/1467-9280.14481

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., et al. (2000). The autism diagnostic observation schedule—generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J. Autism Dev. Disord.* 30, 205–223. doi: 10.1023/A:1005592401947

Lorenz, T., Vlaskamp, B. N., Kasparbauer, A.-M., Mörtl, A., and Hirche, S. (2014). Dyadic movement synchronization while performing incongruent trajectories requires mutual adaptation. *Front. Hum. Neurosci.* 8:461. doi: 10.3389/fnhum.2014.00461

Luria, M., Forlizzi, J., and Hodgins, J. (2018). "The effects of eye design on the perception of social robots," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (Nanjing: IEEE), 1032–1037.

Madipakkam, A. R., Rothkirch, M., Dziobek, I., and Sterzer, P. (2017). Unconscious avoidance of eye contact in autism spectrum disorder. *Sci. Rep.* 7:13378. doi: 10.1038/s41598-017-13945-5

Marsh, K. L., Isenhower, R. W., Richardson, M. J., Helt, M., Verbalis, A. D., Schmidt, R., et al. (2013). Autism and social disconnection in interpersonal rocking. *Front. Integr. Neurosci.* 7:4. doi: 10.3389/fnint.2013.00004

McGrath, J. E., and Kelly, J. R. (1986). *Time and Human Interaction: Toward a Social Psychology of Time.* New York, NY: Guilford Press.

Noel, J.-P., De Niear, M. A., Lazzara, N. S., and Wallace, M. T. (2018). Uncoupling between multisensory temporal function and nonverbal turn-taking in autism spectrum disorder. *IEEE Trans. Cognit. Dev. Syst.* 10, 973–982. doi: 10.1109/TCDS.2017.2778141

Pan, X., and Hamilton, A. F. C. (2018). Why and how to use virtual reality to study human social interaction: the challenges of exploring a new research landscape. *Br. J. Psychol.* 109, 395–417. doi: 10.1111/bjop.12290

Parma, V., and de Marchena, A. B. (2015). Motor signatures in autism spectrum disorder: the importance of variability. *J. Neurophysiol.* 115, 1081–1084. doi: 10.1152/jn.00647.2015

Ramseyer, F., and Tschacher, W. (2008). "Synchrony in dyadic psychotherapy sessions," in *Simultaneity: Temporal Structures and Observer Perspectives*, eds S. Vrobel, O. E. Rössler, and T. Marks-Tarlow (Singapore: World Scientific), 329–347.

Ramseyer, F., and Tschacher, W. (2011). Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. *J. Consult. Clin. Psychol.* 79:284. doi: 10.1037/a0023419

Richardson, M. J., Marsh, K. L., Isenhower, R. W., Goodman, J. R., and Schmidt, R. C. (2007). Rocking together: dynamics of intentional and unintentional interpersonal coordination. *Hum. Mov. Sci.* 26, 867–891. doi: 10.1016/j.humov.2007.07.002

Robins, B., Dautenhahn, K., Te Boekhorst, R., and Billard, A. (2005). Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills? *Univ. Access Inform. Soc.* 4, 105–120. doi: 10.1007/s10209-005-0116-3

Robins, B., Dickerson, P., Stribling, P., and Dautenhahn, K. (2004). Robot-mediated joint attention in children with autism: a case study in robot-human interaction. *Interact. Stud.* 5, 161–198. doi: 10.1075/is.5.2.02rob

Romero, V., Amaral, J., Fitzpatrick, P. A., Schmidt, R. C., and Richardson, M. (2015). "Capturing social motor coordination: a comparison of the microsoft kinect, video-motion analysis and the polhemus latus motion tracking system," in *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (Pasadena, CA).

Sasson, N. J., Faso, D. J., Nugent, J., Lovell, S., Kennedy, D. P., and Grossman, R. B. (2017). Neurotypical peers are less willing to interact with those with autism based on thin slice judgments. *Sci. Rep.* 7:40700. doi: 10.1038/srep40700

Scassellati, B., Admoni, H., and Matarić, M. (2012). Robots for use in autism research. *Annu. Rev. Biomed. Eng.* 14, 275–94. doi: 10.1146/annurev-bioeng-071811-150036

Scharoun, S. M., and Bryden, P. J. (2016). Anticipatory planning in children with autism spectrum disorder: an assessment of independent and joint action tasks. *Front. Integr. Neurosci.* 10:29. doi: 10.3389/fnint.2016.00029

Schmidt, R. C., and Richardson, M. J. (2008). "Dynamics of interpersonal coordination," in *Coordination: Neural, Behavioral and Social Dynamics*, eds A. Fuchs and V. Jirsa (Heidelberg: Springer), 281–308.

Schrijvers, D., Hulstijn, W., and Sabbe, B. G. (2008). Psychomotor symptoms in depression: a diagnostic, pathophysiological and therapeutic tool. *J. Affect. Disord.* 109, 1–20. doi: 10.1016/j.jad.2007.10.019

Schwartz, C., Bente, G., Gawronski, A., Schilbach, L., and Vogeley, K. (2010). Responses to nonverbal behaviour of dynamic virtual characters in high-functioning autism. *J. Autism Dev. Disord.* 40, 100–111. doi: 10.1007/s10803-009-0843-z

Senju, A., and Johnson, M. H. (2009). The eye contact effect: mechanisms and development. *Trends Cognit. Sci.* 13, 127–134. doi: 10.1016/j.tics.2008.11.009

Srinivasan, S. M., Eigsti, I.-M., Gifford, T., and Bhat, A. N. (2016). The effects of embodied rhythm and robotic interventions on the spontaneous and responsive verbal communication skills of children with autism spectrum disorder (ASD): a further outcome of a pilot randomized controlled trial. *Res. Autism Spectr. Disord.* 27, 73–87. doi: 10.1016/j.rasd.2016.04.001

Studenka, B. E., Gillam, S. L., Hartzheim, D., and Gillam, R. B. (2017). Motor and verbal perspective taking in children with autism spectrum disorder: changes in social interaction with people and tools. *Res. Dev. Disabil.* 66, 64–79. doi: 10.1016/j.ridd.2017.02.017

Szelag, E., Kowalska, J., Galkowski, T., and Pöppel, E. (2004). Temporal processing deficits in high-functioning children with autism. *Br. J. Psychol.* 95, 269–282. doi: 10.1348/0007126041528167

Tapus, A., Peca, A., Aly, A., Pop, C., Jisa, L., Pintea, S., et al. (2012). Children with autism social engagement in interaction with nao, an imitative robot: a series of single case experiments. *Interact. Stud.* 13, 315–347. doi: 10.1075/is.13.3.01tap

Tickle-Degnen, L., and Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychol. Inq.* 1, 285–293. doi: 10.1207/s15327965pli0104_1

Torres, E. B., Brincker, M., Isenhower III, R. W., Yanovich, P., Stigler, K. A., Nurnberger Jr., J. I., et al. (2013). Autism: the micro-movement perspective. *Front. Integr. Neurosci.* 7:32. doi: 10.3389/fnint.2013.00032

Vacharkulksemsuk, T., and Fredrickson, B. L. (2012). Strangers in sync: achieving embodied rapport through shared movements. *J. Exp. Soc. Psychol.* 48, 399–402. doi: 10.1016/j.jesp.2011.07.015

Valdesolo, P., Ouyang, J., and DeSteno, D. (2010). The rhythm of joint action: synchrony promotes cooperative ability. *J. Exp. Soc. Psychol.* 46, 693–695. doi: 10.1016/j.jesp.2010.03.004

Vogel, D., Falter-Wagner, C. M., Schoofs, T., Krämer, K., Kupke, C., and Vogeley, K. (2018a). Interrupted time experience in autism spectrum disorder: empirical evidence from content analysis. *J. Autism Dev. Disord.* 49, 22–33. doi: 10.1007/s10803-018-3771-y

Vogel, D. H., Falter-Wagner, C. M., Schoofs, T., Krämer, K., Kupke, C., and Vogeley, K. (2018b). Flow and structure of time experience–concept, empirical validation and implications for psychopathology. *Phenomenol. Cognit. Sci.* 18, 1–24. doi: 10.1007/s11097-018-9573-z

Walther, S., Stegmayer, K., Sulzbacher, J., Vanbellingen, T., Müri, R., Strik, W., et al. (2015). Nonverbal social communication and gesture control in schizophrenia. *Schizophr. Bull.* 41, 338–345. doi: 10.1093/schbul/sbu222

Wolpert, D. M., Doya, K., and Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 358, 593–602. doi: 10.1098/rstb.2002.1238

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

PERSPECTIVE
published: 29 November 2019
doi: 10.3389/frobt.2019.00132

# Machine Learning to Study Social Interaction Difficulties in ASD

*Alexandra Livia Georgescu*[1,2*†], *Jana Christina Koehler*[3*†], *Johanna Weiske*[3], *Kai Vogeley*[2,4], *Nikolaos Koutsouleris*[3] *and Christine Falter-Wagner*[3,5]

[1] Department of Psychology, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom, [2] Department of Psychiatry and Psychotherapy, University Hospital of Cologne, Cologne, Germany, [3] Department of Psychiatry and Psychotherapy, Medical Faculty, LMU Munich, Munich, Germany, [4] Institute of Neuroscience and Medicine, Cognitive Neuroscience (INM-3), Research Center Juelich, Jülich, Germany, [5] Institute of Medical Psychology, Medical Faculty, LMU Munich, Munich, Germany

OPEN ACCESS

**Edited by:**
Cigdem Beyan,
Italian Institute of Technology (IIT), Italy

**Reviewed by:**
Lori-Ann Rosalind Sacrey,
University of Alberta, Canada
Concetto Spampinato,
University of Catania, Italy

***Correspondence:**
Alexandra Livia Georgescu
alexandra.georgescu@kcl.ac.uk
Jana Christina Koehler
jana.koehler@med.uni-muenchen.de

[†] These authors have contributed equally to this work

**Specialty section:**
This article was submitted to Human-Robot Interaction, a section of the journal Frontiers in Robotics and AI

**Received:** 03 March 2019
**Accepted:** 13 November 2019
**Published:** 29 November 2019

**Citation:**
Georgescu AL, Koehler JC, Weiske J, Vogeley K, Koutsouleris N and Falter-Wagner C (2019) Machine Learning to Study Social Interaction Difficulties in ASD. Front. Robot. AI 6:132. doi: 10.3389/frobt.2019.00132

Autism Spectrum Disorder (ASD) is a spectrum of neurodevelopmental conditions characterized by difficulties in social communication and social interaction as well as repetitive behaviors and restricted interests. Prevalence rates have been rising, and existing diagnostic methods are both extremely time and labor consuming. There is an urgent need for more economic and objective automatized diagnostic tools that are independent of language and experience of the diagnostician and that can help deal with the complexity of the autistic phenotype. Technological advancements in machine learning are offering a potential solution, and several studies have employed computational approaches to classify ASD based on phenomenological, behavioral or neuroimaging data. Despite of being at the core of ASD diagnosis and having the potential to be used as a behavioral marker for machine learning algorithms, only recently have movement parameters been used as features in machine learning classification approaches. In a proof-of-principle analysis of data from a social interaction study we trained a classification algorithm on intrapersonal synchrony as an automatically and objectively measured phenotypic feature from 29 autistic and 29 typically developed individuals to differentiate those individuals with ASD from those without ASD. Parameters included nonverbal motion energy values from 116 videos of social interactions. As opposed to previous studies to date, our classification approach has been applied to non-verbal behavior objectively captured during naturalistic and complex interactions with a real human interaction partner assuring high external validity. A machine learning approach lends itself particularly for capturing heterogeneous and complex behavior in real social interactions and will be essential in developing automatized and objective classification methods in ASD.

**Keywords: autism spectrum disorder, machine learning, nonverbal synchrony, support vector machine, motion energy analysis, classification, intrapersonal synchrony, nested cross-validation**

## INTRODUCTION

Autism spectrum disorder (ASD) is an umbrella term for neurodevelopmental conditions characterized by severe difficulties in social interaction and communication, as well as by repetitive behaviors and restricted interests (American Psychiatric Association, 2013). The prevalence rates of ASD are on the rise (Elsabbagh et al., 2012) and diagnostic services are experiencing an increased demand, in particular in adults seeking diagnostic advice (Murphy et al., 2011). Diagnostics according to medical guidelines are time-consuming, the clinical assessment is complicated by the phenotypic heterogeneity and the language-dependency of assessment with verbal skills being affected by the ASD.

Recently, computational methods of classification have been employed to increase diagnostic reliability and efficiency (Thabtah, 2018). In particular, machine learning (ML) employs algorithms to uncover patterns in complex datasets, which are utilized to improve decision making. ASD diagnostics come down to a decision-making problem that can be supported by automated models (classifiers) using ML to decide whether a newly assessed patient has ASD or not. This works by splitting available data into a training set, on which an algorithm is trained, which is then applied to a test set, resulting in a measure of accuracy of the resulting model. Without making assumptions ML finds classification solutions in a data-driven, bottom-up approach that can be applied to individual prediction making (Dwyer et al., 2018). The primary purposes of using ML are (1) to reduce assessment time to reach a diagnostic decision in order to provide quicker access to health care services, (2) to improve diagnostic reliability, and (3) diagnostic validity by reducing dimensionality of input data so as to identify those features that have the most diagnostic value in ASD (Thabtah, 2018). However, first applications of ML in studies on autism diagnostics have been inconsistent in terms of methodology and outcome, with inconsistent classification accuracy and specificity.

The aim of the present paper is twofold: First, we aim to give an overview of previous research that has attempted to apply ML methods to the classification of ASD, while suggesting guidelines for future research in terms of setup and algorithm design. Second, in a proof-of-principle analysis of data from a social interaction study we aim to establish the potential of using full-body non-verbal behavior data extracted from video recordings of naturalistic social interactions to classify autistic adults.

## MACHINE LEARNING APPLICATIONS IN THE CLASSIFICATION OF ASD

First ML attempts in ASD have been used with the aim of shortening ADOS [Autism Diagnostic Observation Schedule, (Lord et al., 2000)] and ADI-R [Autism Diagnostic Interview, (Lord et al., 1994)] administration time by item-reduction yielding a classification accuracy of autism vs. typically-developing (TD) individuals of up to 99.9% (Wall et al., 2012a,b; Bone et al., 2016). In a similar attempt to predict case status words and expressions contained in 8 year old children's developmental evaluations across a network of multiple clinical sites were used for algorithm development (Maenner et al., 2016) with 86.5% prediction accuracy and high concordance with the respective clinician. Home videos of children have been rated by naïve and/or expert raters in terms of ASD-typical behavior and ratings fed into a predictive model along with other features of the diagnostic process (Glover et al., 2018; Tariq et al., 2018). However, while all these first studies using ML in ASD yield fairly high accuracies, the features utilized for classification are still highly subjective and not independent of the respective clinician who bases the diagnostic decision on just those features (circularity). Importantly, when using subjectively influenced data, resulting classification algorithms must be validated in an independent sample in order to prevent circularity.

An increasing number of studies are also using ML to separate individuals with ASD from TD individuals based on neuroimaging data. For example, Ecker et al. (2010) used regional gray and white matter volume measures from whole-brain structural MRI scans of individuals with ASD to investigate their diagnostic value. They used a common variant of ML, the support vector machine (SVM). This is an algorithm aiming at finding a boundary (the so-called "hyperplane") that can be used to optimally classify groups while being able to generalize to new cases (Dwyer et al., 2018). In their sample, the SVM correctly classified individuals with ASD and controls on the basis of their neuroanatomy with about 80% accuracy (Ecker et al., 2010). These original observations are supported by findings from several other neuroimaging studies with similar levels of classification accuracy in younger age groups (Wee et al., 2014), females with ASD (Calderoni et al., 2012) and with various anatomical and functional measurements (Coutanche et al., 2011). These results based on objective data are very promising, although not widely applicable due to high costs.

## WHOLE-BODY MOVEMENTS AS A FEATURE IN ML ALGORITHMS IN ASD

Another source of objective data with high potential for diagnostics can be found in the motor domain. Approximately 80% of children with ASD are suspected to exhibit pronounced motor difficulties (Green et al., 2009). Difficulties with balance, gait, movement speed and timed movements have demonstrated to hold a high level of discrimination between children with ASD and TD children (Jansiewicz et al., 2006) and correlate strongly with measures of social and communicative functioning (Parma and de Marchena, 2016). Hence, movement parameters of social interactions in ASD should be investigated for their potential as a diagnostic marker.

Particularly relevant for ASD motor symptomology are gestures and non-verbal communicative behaviors (Georgescu et al., 2014). Accordingly, atypical non-verbal behavior has been included in the DSM-5 criteria for ASD. Yet, the assessment is not straightforward or standardized so far and is hampered by the fact that non-verbal behavior is not necessarily reduced in ASD, but abnormal in the *quality* of its temporal coordination with

own verbal output (de Marchena and Eigsti, 2010) and that of an interaction partner. Literature provides evidence for aberrations in temporal processing (Allman and Falter, 2015) and time experience in ASD (Vogel et al., 2019), potentially affecting non-verbal communication. In fact, findings have shown that ASD can be characterized by increased temporal resolution associated with the severity of (non-verbal) communication impairments in ASD (Falter et al., 2012, 2013; Menassa et al., 2018; but see Isaksson et al., 2018).

Recently, movement in ASD has taken up increasing interest (for a review see Bo et al., 2016). In a proof-of-concept study to explore whether low-functioning children with ASD could be identified by means of a kinematic analysis of a simple motor task, 15 children with ASD and 15 TD children (2–4 years) were asked to pick up a ball and drop it into a hole while their movements were recorded using a motion tracker (Crippa et al., 2015). Seventeen kinematic parameters were extracted from the upper-limb movement and seven of these were found significant for discrimination. The classifier distinguished ASD from non-ASD with a classification accuracy of 96.7%, suggesting the validity of assuming a motor signature of ASD. Reach and throw movements of 10 ASD and 10 TD children were analyzed for "peculiar features" using ML and fed into a classification algorithm yielding an accuracy of 92.5% (Perego et al., 2009). Furthermore, Li et al. (2017) extracted 40 kinematic parameters of imitative movements and identified 9 of them that best describe variance of participant groups, resulting in a classification accuracy of 93%.

These studies demonstrate the potential of using kinematic biomarkers in diagnostics of ASD. However, the movements under investigation were staged, thus, highly unnatural. Yet, it has been established that individuals with ASD have particular difficulties with spontaneous "on-line" social interaction requiring intuitive decisions and behavior (Redcay et al., 2013) constituting an urgent need to move this type of research to more external validity and investigate movement in a more naturalistic context.

## CLASSIFICATION USING INTRAPERSONAL SYNCHRONY: A PROOF-OF-CONCEPT STUDY

Whole-body movements in more naturalistic conversations were tested for their classification potential in 29 high functioning adults with ASD and 29 TD individuals. The data for this investigation came from a study on interpersonal coordination in dyadic interactions (Georgescu et al., under revision). The autistic participants were diagnosed and recruited at the Autism Outpatient Clinic of the Department of Psychiatry, University Hospital Cologne, Germany. The sample included only patients with the diagnoses high-functioning autism (ICD-10: F84.0) or Asperger syndrome (ICD-10: F84.5). Two medical specialists confirmed the diagnosis independently in clinical interviews, according to the criteria of the International Classification of Diseases (ICD-10) and supplemented by extensive neuropsychological examination. The TD sample was recruited online from the student and staff population at the University of Cologne and the University Hospital of Cologne, Germany. The study was conducted with the approval of the local ethics committee of the Medical Faculty of the University of Cologne. Participants were paired to conduct five 5 min social interaction tasks. Conversational dyads consisted of either two TD individuals, two individuals with ASD or a TD individual with an individual with ASD. An ice-breaker task, two debating tasks, a meal-planning task and a roleplay were included resulting in a total of 145 videos of social interactions (for more information, see Georgescu et al., under revision). All conversations were recorded in a room with standardized artificial lighting and using a high-definition video camera (Panasonic DV C Pro HD P2), mounted on a tripod 320 cm away from the chairs which were 60 cm apart from each other. Since one of the MIXED dyads did not understand instructions on the ice-breaker task, for the purpose of this analysis the whole task was abandoned, resulting in a total of 116 videos submitted for final analysis. Intrapersonal Synchrony between the head and upper body was quantified using Motion Energy Analysis, a widely used semi-automated frame-differencing method that continuously monitors the amount of movement occurring in manually pre-defined regions of interest and the method of lagged cross-correlations (Nagaoka and Komori, 2008; MEA; Altmann, 2011; Ramseyer and Tschacher, 2011). MEA offers the advantage of a constraint-free, objective analysis tool for non-verbal behavior (e.g., Ramseyer and Tschacher, 2011; Schmidt et al., 2012; Paxton and Dale, 2013). This method has been used to capture body movement in different contexts (e.g., Grammer et al., 1999; Ramseyer and Tschacher, 2011, 2014; Schmidt et al., 2012, 2014; Paxton and Dale, 2013). MEA and other frame-differencing methods have been successfully used in clinical research before (e.g., Kupper et al., 2015) and in particular in autism (Noel et al., 2017; Romero et al., 2017, 2018). We followed the MEA pipeline described in Ramseyer and Tschacher (2014). We manually selected two regions of interest (ROI) for each participant, covering (1) the head and (2) the rest of the body including the legs. Changes in grayscale values in these ROIs were detected and separately recorded as two continuous time series measuring the amount of movement in the head and the body region of each person. Data were submitted for quantification of Intrapersonal Synchrony (for more information on the MEA procedure in general, please see Ramseyer and Tschacher, 2014 and on this sample, Georgescu et al., under revision). Input time series were smoothed and scaled to account for different-sized ROIs using custom software in R (package rMEA, Kleinbub and Ramseyer, 2019) and cross-correlated in windows of 60 s with a time lag of $\pm 5$ s (step size 0.04 s). Windows were not allowed to overlap. The resulting 1,004 lagged cross-correlations were then z-standardized and aggregated over the four conditions for every participant, yielding 4,016 features per participant which were implemented in the open-source machine learning tool NeuroMiner (https://www.pronia.eu/neurominer/). A support vector machine with linear kernel was chosen as a classification algorithm, a multivariate supervised learning technique widely

**TABLE 1 |** Performance metrics of the ASD vs. TD SVM classifier.

| | |
|---|---|
| True positives/true negatives | 28/16 |
| False positives/false negatives | 13/1 |
| Accuracy [%] | 75.9 |
| Sensitivity [%] | 96.6 |
| Specificity [%] | 55.2 |
| Area under the curve | 0.71 |

*For detailed explanation of performance metrics please refer to Dwyer et al. (2018).*

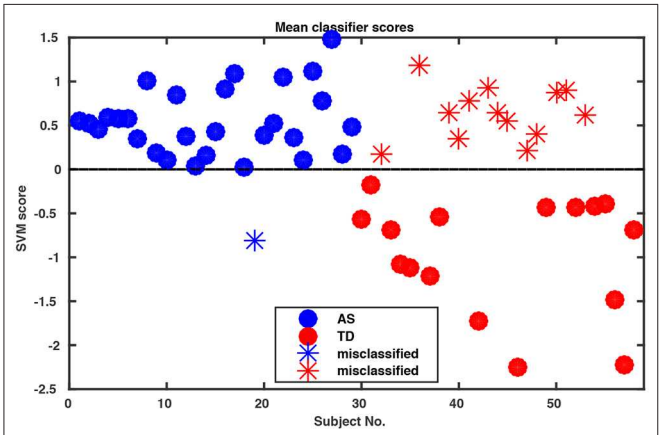



**FIGURE 1 |** Decision scores of SVM classification performance. The algorithm assigns a score to each participant indicating the probability of this participant as belonging to Group 1 or 2 (in our case ASD vs. TD) where the decision boundary between the two groups is zero. Notably, our algorithm misclassified only one of the ASD participants.

used in psychiatric research (Bone et al., 2016; Duda et al., 2016). Our repeated nested *k*-fold cross-validation (CV) structure consisted of 10-folds and five permutations for the outer cross-validation cycle ($CV_2$) and repeated 5-by-5-fold inner cross-validation cycle ($CV_1$), with participants being shuffled prior to each definition of folds. This way, the data available for training was maximized while ensuring enough heterogeneity within the inner test sample to avoid overfitting and create stable models. Parameter optimization was performed in $CV_1$, while model performance was evaluated in $CV_2$. Prior to analysis, data was preprocessed using principal component analysis (PCA) for dimensionality reduction, retaining the principal components that cumulatively explained 80% of the variance in each $CV_1$ fold, and subsequently, scaled feature-wise from 0 to 1. The slack parameter C was estimated in the inner CV cycle using eight parameters ranging from 0.015625 to 16. Overall classification performance resulted in 75.9% accuracy (**Table 1**). Remarkably, sensitivity was 96.6%, correctly classifying all but one individual with ASD (**Figure 1**).

Thus, with a portable and inexpensive video-setup in a naturalistic setting and a semi-automated analysis pipeline, we reached a good diagnostic classification of ASD within four 5 min interaction excerpts on the mere basis of objective motion data. Feeding further clinical and interaction variables into the algorithm promises a high potential for classification (see Future Perspectives section).

**BOX 1** | Minimum requirements for reliable clinical application of ML in ASD research (adapted from Dwyer et al., 2018)

- Combination of objective variables and standard diagnostic measures as input features to classify ASD.
- Use of nested CV as a standard procedure.
- Prevent unstable model outcomes through *k*-fold CV.

## METHODOLOGICAL ISSUES IN MACHINE LEARNING APPROACHES TO CLASSIFYING ASD

Unlike e.g., Bone et al. (2016) or Li et al. (2017), most ML studies in ASD research have relied on simple cross-validation (CV) methods. This increases the likelihood of choosing an overly optimistic model (Cawley and Talbot, 2010). We therefore suggest the application of a second layer of CV to allow for parameter selection and model performance evaluation to not be performed on the same data and to prevent overfitting. The test fold is completely held out until parameter optimization within the inner CV cycle is achieved by splitting the training data once more into an (inner) test and (inner) training set. The optimized models can then be tested for generalizability on the outer test fold. This so-called nested CV maximizes generalizability and has now been established as a gold standard procedure in psychiatric research (Dwyer et al., 2018). In order to account for the small sample sizes in ASD research, often predictions are made in a leave-one-out approach whereby only one individual's data is held out in the test set while parameters are optimized on the others (Crippa et al., 2015; Li et al., 2017). Especially, for ASD with its highly heterogeneous phenotype, leave-one-out creates overly variable test sets, rendering model outcomes unstable (Varoquaux et al., 2017). This can be prevented through k-fold nested CV and simultaneous permutation of individual data sets within the inner cross-validation cycle (Dwyer et al., 2018). An overview of best-practice standards is outlined below.

## FUTURE PERSPECTIVES

Impairments of non-verbal communication are seen across the entire spectrum of ASD warranting the use as a behavioral biomarker. Yet, its intricacy requires multivariate analysis methods to capture complex interdependencies across domains. Machine learning offers the potential to incorporate high-dimensional data for the detection of underlying mechanisms and classification if certain minimum practice requirements are fulfilled (see **Box 1**).

In our proof-of-principle study, we were able to classify high-functioning adults with ASD from TD adults on the mere basis of non-verbal intrapersonal motion synchrony in

social interactions with an accuracy of 75.9%, which can be regarded a conservative estimate on the basis of a state-of-the art ML approach. Due to relatively small sample sizes available with high phenomenological heterogeneity in ASD, it is of utmost importance to choose adequate methods of cross-validation in order to maximize generalizability. The use of repeated nested cross-validation prevents overfitting and should be incorporated as a standard procedure in ML applications. However, given our rather limited sample size, the next steps for future research will be to apply the resulting algorithm to a completely new and larger data set and to investigate its transdiagnostic specificity across different psychiatric disturbances.

Future research should furthermore consider combining multiple non-verbal communication parameters and clinical data (e.g., questionnaires) in order to improve prediction and classification accuracy further and to possibly detect potential associations across domains. For instance, peculiarities in eye-gaze (Merin et al., 2007; Georgescu et al., 2013) and facial expression (McIntosh et al., 2006) in ASD demonstrate feasible approaches.

One future avenue would be to explore methods to quantify non-verbal behavior in a fully-automated fashion. In the present proof-of-principle study, a dataset was used that was analyzed using MEA, a classic frame-differencing approach. It has been shown that MEA is able to capture movements and even complex coordinative patterns to a similar extent as more expensive motion capture equipment such as the Polhemus (Romero et al., 2017). A main advantage for autism research of this method of extracting whole-body motor movement is that it does not involve any wearable technology. Given the hypersensitivity exhibited by many individuals with ASD, not having to add any attachable piece of equipment or body suit to their bodies is helpful. However, while MEA automatically detects pixel changes, corresponding regions of interest are drawn in manually. Although resulting values are standardized, there remains a subjective component. Computer vision tools that can estimate the coordinates of limb positions and even extract gaze location and body poses would offer similar benefits while balancing out subjective biases in the motion extraction process (Marín-Jiménez et al., 2014; Mehta et al., 2017; Tome et al., 2017; Cao et al., 2018). In addition, they offer even more flexibility, given it could be possible to include less strict and standardized experimental setups (no requirement for standardized camera or lighting conditions). However, the validity for movement extraction compared to other standard motion capture methods has not been demonstrated yet. Moreover, such tools vary greatly with respect to their susceptibility to tracking failures, or the type of videos they can support (single vs. multiple agent, indoor vs. outdoor etc.). Overall, with the current methodology that is available for motion extraction, the present semi-automated method offers a realistically applicable diagnostic value. Nevertheless, incredible advances are being made (Li et al., 2018; Tran et al., 2018) such that they are very promising tools for future non-verbal behavior in autism research and beyond.

Taken together, given the recent advances in predictive psychiatry, adequately applied ML offers the potential to fully capture the autistic phenotype in all its complexity with sufficient specificity across psychiatric disorders with a special focus on the spontaneous non-verbal behavior during social encounters with others and irrespective of clinician or site.

## DATA AVAILABILITY STATEMENT

The video datasets generated and analysed during the current study are not publicly available due this being identifiable patient data from a sample that did not consent to their data being shared in any form.

## ETHICS STATEMENT

Written informed consent was obtained from all participants in accordance with the Declaration of Helsinki (1964). All participants received a monetary compensation for their participation of 50 Euros and were debriefed at the end. The study was conducted with the approval of the local ethics committee of the Medical Faculty of the University of Cologne.

## AUTHOR CONTRIBUTIONS

AG, JK, and CF-W contributed equally to the drafting of this manuscript. AG provided the data. JW and JK performed the statistical analysis. All authors contributed to the manuscript revision, read, and approved the submitted version.

## FUNDING

JK was supported by Stiftung Irene, LMU excellent seed funding (granted to CF-W; 867900-3) and a German Research Foundation grant (granted to CF-W; FA876/3-1). CF-W was supported by a Bavarian Gender Equality Grant. The data was collected under a postdoctoral grant awarded to AG under the Institutional Strategy of the University of Cologne within the German Excellence Initiative.

## ACKNOWLEDGMENTS

Marius Kuschefski, Sevim Koeroglu, and Helen Fischer deserve much appreciation for their assistance with the data collection. We also thank the UK Media Team of the University Hospital of Cologne for their help and support with the video recordings. We are grateful to Wolfgang Tschacher and Fabian Ramseyer for help with Motion Energy Analysis.

## REFERENCES


Allman, M., and Falter, C. (2015). "Abnormal timing and time perception in autism spectrum disorder?: a review of the evidence," in *Time Distortions in Mind: Temporal Processing in Clinical Populations,* eds A. Vatakis and M. Allman (Leiden; Boston, MA: Brill), 37–56.

Altmann, U. (2011). "Investigation of movement synchrony using windowed cross-lagged regression," in *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues.* Lecture Notes in Computer Science, Vol. 6800, eds A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, and A. Nijholt (Berlin; Heidelberg: Springer).

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®).* American Psychiatric Publishing. doi: 10.1176/appi.books.9780890425596

Bo, J., Lee, C.-M., Colbert, A., and Shen, B. (2016). Do children with autism spectrum disorders have motor learning difficulties? *Res. Autism Spectr. Disord.* 23, 50–62. doi: 10.1016/j.rasd.2015.12.001

Bone, D., Bishop, S. L., Black, M. P., Goodwin, M. S., Lord, C., and Narayanan, S. S. (2016). Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion. *J. Child Psychol. Psychiatr.* 57, 927–937. doi: 10.1111/jcpp.12559

Calderoni, S., Retico, A., Biagi, L., Tancredi, R., Muratori, F., and Tosetti, M. (2012). Female children with autism spectrum disorder: an insight from mass-univariate and pattern classification analyses. *NeuroImage* 59, 1013–1022. doi: 10.1016/j.neuroimage.2011.08.070

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2018). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *ArXiv:1812.08008*. doi: 10.1109/CVPR.2017.143

Cawley, G. C., and Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Machine Learn. Res.* 11, 2079–2107.

Coutanche, M. N., Thompson-Schill, S. L., and Schultz, R. T. (2011). Multi-voxel pattern analysis of fMRI data predicts clinical symptom severity. *NeuroImage* 57, 113–123. doi: 10.1016/j.neuroimage.2011.04.016

Crippa, A., Salvatore, C., Perego, P., Forti, S., Nobile, M., Molteni, M., et al. (2015). Use of machine learning to identify children with autism and their motor abnormalities. *J. Autism Dev. Disord.* 45, 2146–2156. doi: 10.1007/s10803-015-2379-8

de Marchena, A., and Eigsti, I. M. (2010). Conversational gestures in autism spectrum disorders: asynchrony but not decreased frequency. *Autism Res.* 3, 311–322. doi: 10.1002/aur.159

Duda, M., Ma, R., Haber, N., and Wall, D. P. (2016). Use of machine learning for behavioral distinction of autism and ADHD. *Transl. Psychiatr.* 6:e732. doi: 10.1038/tp.2015.221

Dwyer, D. B., Falkai, P., and Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annu. Rev. Clin. Psychol.* 14, 91–118. doi: 10.1146/annurev-clinpsy-032816-045037

Ecker, C., Rocha-Rego, V., Johnston, P., Mourao-Miranda, J., Marquand, A., Daly, E. M., et al. (2010). Investigating the predictive value of whole-brain structural MR scans in autism: a pattern classification approach. *NeuroImage* 49, 44–56. doi: 10.1016/j.neuroimage.2009.08.024

Elsabbagh, M., Divan, G., Koh, Y.-J., Kim, Y. S., Kauchali, S., Marcín, C., et al. (2012). Global prevalence of autism and other pervasive developmental disorders. *Autism Res.* 5, 160–179. doi: 10.1002/aur.239

Falter, C. M., Braeutigam, S., Nathan, R., Carrington, S., and Bailey, A. J. (2013). Enhanced access to early visual processing of perceptual simultaneity in autism spectrum disorders. *J. Autism Dev. Disord.* 43, 1857–1866. doi: 10.1007/s10803-012-1735-1

Falter, C. M., Elliott, M. A., and Bailey, A. J. (2012). Enhanced visual temporal resolution in autism spectrum disorders. *PLoS ONE* 7:e32774. doi: 10.1371/journal.pone.0032774

Georgescu, A. L., Kuzmanovic, B., Roth, D., Bente, G., and Vogeley, K. (2014). The use of virtual characters to assess and train non-verbal communication in high-functioning autism. *Front. Human Neurosci.* 8:807. doi: 10.3389/fnhum.2014.00807

Georgescu, A. L., Kuzmanovic, B., Schilbach, L., Tepest, R., Kulbida, R., Bente, G., et al. (2013). Neural correlates of "social gaze" processing in high-functioning autism under systematic variation of gaze duration. *NeuroImage* 3, 340–351. doi: 10.1016/j.nicl.2013.08.014

Glover, E., Garberson, F., Abbas, H., and Wall, D. P. (2018). Machine learning approach for early detection of autism by combining questionnaire and home video screening. *J. Am. Med. Inform. Assoc.* 25, 1000–1007. doi: 10.1093/jamia/ocy039

Grammer, K., Honda, M., Juette, A., and Schmitt, A. (1999). Fuzziness of nonverbal courtship communication unblurred by motion energy detection. *J. Pers. Soc. Psychol.* 77:487. doi: 10.1037//0022-3514.77.3.487

Green, D., Charman, T., Pickles, A., Chandler, S., Loucas, T., Simonoff, E., et al. (2009). Impairment in movement skills of children with autistic spectrum disorders. *Dev. Med. Child Neurol.* 51, 311–316. doi: 10.1111/j.1469-8749.2008.03242.x

Isaksson, S., Salomäki, S., Tuominen, J., Arstila, V., Falter-Wagner, C. M., and Noreika, V. (2018). Is there a generalized timing impairment in Autism Spectrum Disorders across time scales and paradigms? *J. Psychiatr. Res.* 99, 111–121.

Jansiewicz, E. M., Goldberg, M. C., Newschaffer, C. J., Denckla, M. B., Landa, R., and Mostofsky, S. H. (2006). Motor signs distinguish children with high functioning autism and Asperger's syndrome from controls. *J. Autism Dev. Disord.* 36, 613–621. doi: 10.1007/s10803-006-0109-y

Kleinbub, J. R., and Ramseyer, F. (2019). *rMEA Synchrony in Motion Energy Analysis (MEA) Time-Series.* R package version 1.1.0. Available online at: https://CRAN.R-project.org/package=rMEA

Kupper, Z., Ramseyer, F., Hoffmann, H., and Tschacher, W. (2015). Nonverbal synchrony in social interactions of patients with schizophrenia indicates socio-communicative deficits. *PLoS ONE* 10:e0145882. doi: 10.1371/journal.pone.0145882

Li, B., Sharma, A., Meng, J., Purushwalkam, S., and Gowen, E. (2017). Applying machine learning to identify autistic adults using imitation: an exploratory study. *PLoS ONE* 12:e0182652. doi: 10.1371/journal.pone.0182652

Li, Z., Gavrilyuk, K., Gavves, E., Jain, M., and Snoek, C. G. M. (2018). Videolstm convolves, attends and flows for action recognition. *Comp. Vis. Image Underst.* 166, 41–50. doi: 10.1016/j.cviu.2017.10.011

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., et al. (2000). The autism diagnostic observation schedule—generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J. Autism Dev. Disord.* 30, 205–223. doi: 10.1023/A:1005592401947

Lord, C., Rutter, M., and Le Couteur, A. (1994). Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J. Autism Develop. Disord.* 24, 659–685. doi: 10.1007/BF02172145

Maenner, M. J., Yeargin-Allsopp, M., Van Naarden Braun, K., Christensen, D. L., and Schieve, L. A. (2016). Development of a machine learning algorithm for the surveillance of autism spectrum disorder. *PLoS ONE* 11:e0168224. doi: 10.1371/journal.pone.0168224

Marín-Jiménez, M. J., Zisserman, A., Eichner, M., and Ferrari, V. (2014). Detecting people looking at each other in videos. *Int. J. Comput. Vis.* 106, 282–296. doi: 10.1007/s11263-013-0655-7

McIntosh, D. N., Reichmann-Decker, A., Winkielman, P., and Wilbarger, J. L. (2006). When the social mirror breaks: deficits in automatic, but not voluntary, mimicry of emotional facial expressions in autism. *Dev. Sci.* 9, 295–302. doi: 10.1111/j.1467-7687.2006.00492.x

Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.-P., et al. (2017). Vnect: real-time 3d human pose estimation with a single rgb camera. *ACM Trans. Graph.* 36:44. doi: 10.1145/3072959.3073596

Menassa, D. A., Braeutigam, S., Bailey, A., and Falter-Wagner, C. M. (2018). Frontal evoked γ activity modulates behavioural performance in Autism Spectrum Disorders in a perceptual simultaneity task. *Neurosci. Lett.* 665, 86–91. doi: 10.1016/j.neulet.2017.11.045

Merin, N., Young, G. S., Ozonoff, S., and Rogers, S. J. (2007). Visual fixation patterns during reciprocal social interaction distinguish a subgroup of 6-month-old infants at-risk for autism from comparison infants. *J. Autism Dev. Disord.* 37, 108–121. doi: 10.1007/s10803-006-0342-4

Murphy, D. G. M., Beecham, J., Craig, M., and Ecker, C. (2011). Autism in adults. New biologicial findings and their translational implications to the cost of clinical services. *Brain Res.* 1380, 22–33. doi: 10.1016/j.brainres.2010.10.042

Nagaoka, C., and Komori, M. (2008). Body movement synchrony in psychotherapeutic counseling: A study using the video-based quantification method. *IEICE Trans. Inf. Syst.* 91, 1634–1640. doi: 10.1093/ietisy/e91-d.6.1634

Noel, J.-P., De Niear, M. A., Lazzara, N. S., and Wallace, M. T. (2017). Uncoupling between multisensory temporal function and nonverbal turn-taking in autism spectrum disorder. *IEEE Trans. Cognit. Develop. Syst.* 10, 973–982. doi: 10.1109/TCDS.2017.2778141

Parma, V., and de Marchena, A. B. (2016). Motor signatures in autism spectrum disorder: the importance of variability. *J. Neurophysiol.* 115, 1081–1084. doi: 10.1152/jn.00647.2015

Paxton, A., and Dale, R. (2013). Frame-differencing methods for measuring bodily synchrony in conversation. *Behav. Res. Methods* 45, 329–343. doi: 10.3758/s13428-012-0249-2

Perego, P., Forti, S., Crippa, A., Valli, A., and Reni, G. (2009). "Reach and throw movement analysis with support vector machines in early diagnosis of autism," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Minneapolis, MN), 2555–2558. doi: 10.1109/IEMBS.2009.5335096

Ramseyer, F., and Tschacher, W. (2011). Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. *J. Consult. Clin. Psychol.* 79, 284–295. doi: 10.1037/a0023419

Ramseyer, F., and Tschacher, W. (2014). Nonverbal synchrony of head- and body-movement in psychotherapy: different signals have different associations with outcome. *Front. Psychol.* 5:979. doi: 10.3389/fpsyg.2014.00979

Redcay, E., Dodell-Feder, D., Mavros, P. L., Kleiner, M., Pearrow, M. J., Triantafyllou, C., et al. (2013). Atypical brain activation patterns during a face-to-face joint attention game in adults with autism spectrum disorder. *Hum. Brain Mapp.* 34, 2511–2523. doi: 10.1002/hbm.22086

Romero, V., Amaral, J., Fitzpatrick, P., Schmidt, R. C., Duncan, A. W., and Richardson, M. J. (2017). Can low-cost motion-tracking systems substitute a Polhemus system when researching social motor coordination in children? *Behav. Res. Methods* 49, 588–601. doi: 10.3758/s13428-016-0733-1

Romero, V., Fitzpatrick, P., Roulier, S., Duncan, A., Richardson, M. J., and Schmidt, R. C. (2018). Evidence of embodied social competence during conversation in high functioning children with autism spectrum disorder. *PLoS ONE* 13:e0193906. doi: 10.1371/journal.pone.0193906

Schmidt, R. C., Morr, S., Fitzpatrick, P., and Richardson, M. J. (2012). Measuring the dynamics of interactional synchrony. *J. Nonverbal. Behav.* 36, 263–279. doi: 10.1007/s10919-012-0138-5

Schmidt, R. C., Nie, L., Franco, A., and Richardson, M. J. (2014). Bodily synchronization underlying joke telling. *Front. Hum. Neurosci.* 8:633. doi: 10.3389/fnhum.2014.00633

Tariq, Q., Daniels, J., Schwartz, J. N., Washington, P., Kalantarian, H., and Wall, D. P. (2018). Mobile detection of autism through machine learning on home video: a development and prospective validation study. *PLoS Med.* 15:e1002705. doi: 10.1371/journal.pmed.1002705

Thabtah, F. (2018). Machine learning in autistic spectrum disorder behavioral research: a review and ways forward. *Inform. Health Soc. Care* 44, 278–297. doi: 10.1080/17538157.2017.1399132

Tome, D., Russell, C., and Agapito, L. (2017). "Lifting from the deep: convolutional 3d pose estimation from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 2500–2509. doi: 10.1109/CVPR.2017.603

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. (2018). "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 6450–6459. doi: 10.1109/CVPR.2018.00675

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B. (2017). Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage* 145, 166–179. doi: 10.1016/j.neuroimage.2016.10.038

Vogel, D., Falter-Wagner, C. M., Schoofs, T., Krämer, K., Kupke, C., and Vogeley, K. (2019). Interrupted time experience in autism spectrum disorder: empirical evidence from content analysis. *J. Autism Dev. Disord.* 49, 22–33. doi: 10.1007/s10803-018-3771-y

Wall, D. P., Dally, R., Luyster, R., Jung, J.-Y., and DeLuca, T. F. (2012a). Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PLoS ONE* 7:e43855. doi: 10.1371/journal.pone.0043855

Wall, D. P., Kosmicki, J., DeLuca, T. F., Harstad, E., and Fusaro, V. A. (2012b). Use of machine learning to shorten observation-based screening and diagnosis of autism. *Transl. Psychiatr.* 2:e100. doi: 10.1038/tp.2012.10

Wee, C.-Y., Wang, L., Shi, F., Yap, P.-T., and Shen, D. (2014). Diagnosis of autism spectrum disorders using regional and interregional morphological features. *Hum. Brain Mapp.* 35, 3414–3430. doi: 10.1002/hbm.22411

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

REVIEW
published: 05 December 2019
doi: 10.3389/frobt.2019.00119

# Computational Commensality: From Theories to Computational Models for Social Food Preparation and Consumption in HCI

*Radoslaw Niewiadomski*[1]*, *Eleonora Ceccaldi*[2], *Gijs Huisman*[3], *Gualtiero Volpe*[2] and *Maurizio Mancini*[4]

[1] CONTACT Unit, Istituto Italiano di Tecnologia, Genoa, Italy, [2] InfoMus Lab, DIBRIS, University of Genoa, Genoa, Italy, [3] Digital Society School, Amsterdam University of Applied Sciences, Amsterdam, Netherlands, [4] School of Computer Science and Information Technology, University College Cork, Cork, Ireland

Food and eating are inherently social activities taking place, for example, around the dining table at home, in restaurants, or in public spaces. Enjoying eating with others, often referred to as "commensality," positively affects mealtime in terms of, among other factors, food intake, food choice, and food satisfaction. In this paper we discuss the concept of "Computational Commensality," that is, technology which computationally addresses various social aspects of food and eating. In the past few years, Human-Computer Interaction started to address how interactive technologies can improve mealtimes. However, the main focus has been made so far on improving the individual's experience, rather than considering the inherently social nature of food consumption. In this survey, we first present research from the field of social psychology on the social relevance of Food- and Eating-related Activities (F&EA). Then, we review existing computational models and technologies that can contribute, in the near future, to achieving Computational Commensality. We also discuss the related research challenges and indicate future applications of such new technology that can potentially improve F&EA from the commensality perspective.

**Keywords: commensality, food, food recognition, HCI, social signal processing, embodied interfaces, social robots, augmented experience**

OPEN ACCESS

***Edited by:***
*Agnieszka Wykowska,*
*Italian Institute of Technology (IIT), Italy*

***Reviewed by:***
*Oya Aran,*
*Idiap Research Institute, Switzerland*
*Rob Comber,*
*Royal Institute of Technology, Sweden*

***Correspondence:**
*Radoslaw Niewiadomski*
*radoslaw.niewiadomski@iit.it*

***Specialty section:***
*This article was submitted to Humanoid Robotics, a section of the journal Frontiers in Robotics and AI*

***Received:*** *02 March 2019*
***Accepted:*** *28 October 2019*
***Published:*** *05 December 2019*

***Citation:***
*Niewiadomski R, Ceccaldi E, Huisman G, Volpe G and Mancini M (2019) Computational Commensality: From Theories to Computational Models for Social Food Preparation and Consumption in HCI. Front. Robot. AI 6:119. doi: 10.3389/frobt.2019.00119*

## 1. INTRODUCTION

Food and drink consumption is a vital human activity aimed at providing the body with nutrients that are necessary for survival. What is more, eating and drinking are also highly social activities that take place, for example, around the dining table at home, in restaurants, or in public spaces. People use food to regulate their own and others' emotions, for example, by offering food to cheer others up or by eating some particular food they associate with positive memories. Humans learn that food can have a social and emotional meaning from a very young age, for example, by associating food offering with soothing (Hamburg et al., 2014). Food-related interaction, often referred to as "commensality," is very important for personal health and well-being (e.g., Grevet et al., 2012).

Given the importance of food consumption, researchers in human-computer interaction (HCI) and artificial intelligence (AI) have recently started to address how interactive technologies can

improve mealtimes. For example, devices like sensor networks or connected appliances offering multi-sensory eating experiences (Kortum, 2008) are increasingly entering the processes of food preparation and consumption, while virtual agents (Gardiner et al., 2017) and robot companions (Baroni et al., 2014) are used to motivate children to eat more healthily. The variety of the topics related to *Food- and Eating-related Activities* (F&EA) has attracted researchers' interest from several AI-related disciplines: from computer vision to multimodal interaction and from positive to social computing, as demonstrated by the recently born series of workshops titled "Multi-sensory Approaches to Human-Food Interaction" and the "ACM Future of Computing & Food Manifesto[1]".

However, research in AI and HCI and technologies dedicated to F&EA often focus on food (or eating) itself (e.g., food recognition and sensory augmentation) rather than on its social dimension. In this work, we introduce the concept of *Computational Commensality (CC)*[2] to gather different attempts to computationally address various social aspects of food and eating. CC extends commensality in humans (see **Figure 1**), which is *the practice of sharing food and eating together in a social group* (Ochs and Shohet, 2006) by introducing technology as a "social glue" for food-related interaction. CC will focus on creation of rich physical or mediated multimodal interaction between two or more agents (being humans, or humans and machines) which may enable or enhance outcomes of the "traditional" commensality (i.e., in the sense of Ochs and Shohet's definition) studied so far mainly by psychologists and sociologists. CC needs, for example, F&EA recognition modules as building blocks to create food-related interaction. However, it goes beyond these topics already extensively studied in HCI and AI. It must also be distinguished from the other food-related concepts recently proposed, such as *gastroludology* (Chisik et al., 2018) and *human–food interaction* (Comber et al., 2014; Altarriba Bertran et al., 2018). The first one focuses on experiences involving playing with food (e.g., games). Indeed, such experiences can sometimes be social, e.g., when two or more persons use the technology to feed each other (Mehta et al., 2018) (see section 6.1 for other examples), but still the main focus is on sensorial, playful experience with food and the technological innovation enabling it. In this sense, Mueller et al. (2018) proposes considering the eating activity "as something not serious, with neither a clear goal nor real-world consequences." The second one mainly investigates the individual experience, and rarely considers social context (although we present some recent interesting works in this field in section 5).

In our view, CC may appear in two main scenarios. First of all, two (or more) humans can use technology to enable or enhance human-human interactions during meals. Examples can be: using technology to enhance co-located social dining (e.g., Ganesh et al., 2014), or using tele-dining technology to enable social interaction between people who do not share the same physical space (e.g., Nawahdah and Inoue, 2013). In the second CC scenario, human(s) interact with an artificial companion, such as a social robot (e.g., Khot et al., 2019) during meal time. The companion uses sensors and computational models of commensality to guide its behavior toward the human interlocutor. In both cases, computational models can also be used to analyze and quantify the interaction during meals, for example, by detecting quantity of food consumed together or identifying the social roles at the table.

The main goal of this article is 2-fold: (1) we discuss psychological and sociological studies on the social aspects of food and eating activities, showing how they can be exploited to create CC; and (2) we present computational models, devices, and applications focusing on their social dimension, illustrating how they could be used in CC scenarios.

In the next section we will review contributions from social psychology dealing with food related interaction and social influence on food behavior, and we will illustrate recent HCI and AI works that deal with food preparation and consumption. In section 3, we will start our survey by illustrating existing works on food and eating recognition, which is probably the food-related topic most explored in computer science, with applications ranging from food production to virtual dining experiences. Existing solutions are usually based on computer vision and machine learning techniques, although other modalities such as audio have been sometimes explored. The most recent trends include the application of deep learning techniques for life-logging. We will present works dealing with human movement tracking and monitoring in food-related activities—for example, the recognition of drinking and swallowing actions from multimodal data coming from wearable sensors, audio or visual devices. In section 4, we will turn our attention to systems applying these technologies to provide physical or psychological support in eating activities. For example, systems offering physical assistance (e.g., for physically impaired people), mainly using robots, cooking assistants (e.g., in augmented reality), or serious games aiming to change bad eating habits. In section 5, we will discuss systems that use similar techniques with the aims to manipulate, augment, or enhance eating and drinking experiences through multimodal technologies. For example, several devices have been designed and tested to detect and simulate odors to be presented during food consumption as an additional sensory cue, while in other cases dining tables enhanced with projection mapping visualizations have been developed. These efforts provide insights on how technology can be introduced in dining activities to enhance the sensory experience of food and drink. In section 6 we will outline systems that use technology to, in a broad sense, enable, or stimulate interaction during food preparation and consumption. This includes multi-user games, robot-mediated physical interaction, as well as tele-dining systems. Most of these works already contain some aspects of CC, as they provide the technology and computational models

[1] https://acm-fca.org/2018/07/01/future-of-computing-food-manifesto

[2] The term "commensality" was used previously in a narrower sense, among others, in works by Ferdous et al. (2016a) on the influence of existing technology (e.g., tablets, smart phones) in familial interaction, or by Grevet et al. (2012) on promoting social awareness around mealtimes through communication of the "eating"-related statuses to remote confederates. In this paper, we propose a broader perspective, which includes, for instance, social interaction not only with another human but also the technology itself (e.g., social robots).

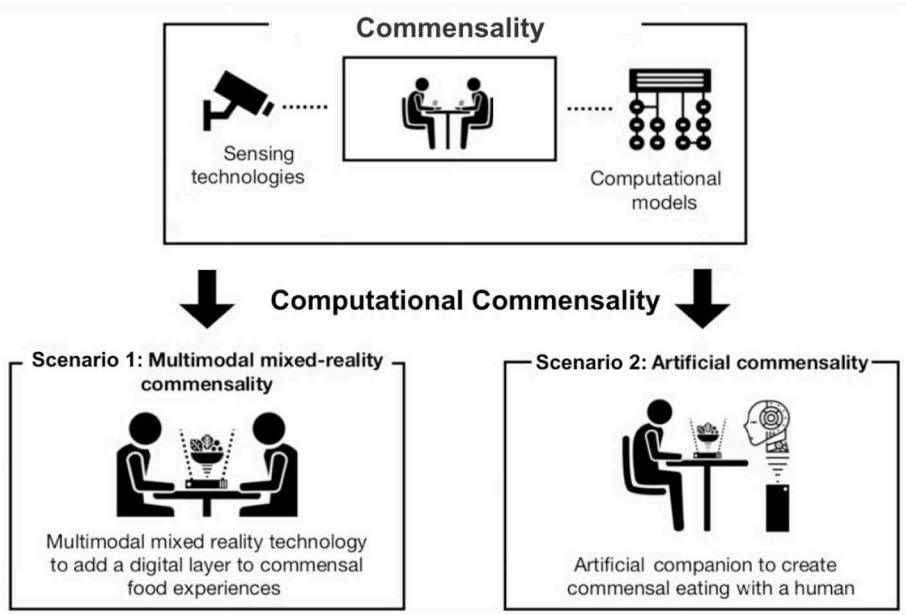



**FIGURE 1 |** From commensality to computational commensality. We introduce research on "traditional" commensality in section 2, the papers related to the CC Scenario 1 in sections 4 and 6, and to the CC Scenario 2—in sections 3, 5, and 6.

to enhance or extend the human-human interaction around the food.

Despite the long list of topics related to eating and AI we present in our survey (see also **Figure 2**), we believe that the investigation of the link between social aspects of F&EA and technology has just started. The variety of possible applications in this field is enormous, pushing this discipline to grow up quickly in the near future. We will conclude the paper by discussing some possible future research directions in section 7.

## 1.1. Selection of Sources

Given the novelty of the topic and its inherently multidisciplinary nature, we drew on many different sources from different disciplines and kept an open perspective in the selection of literature relevant to the current survey. Nevertheless, our main focus was on computational technologies that either had a direct relation with social dining practices (e.g., tele-presence dining), or computational technologies that would be part of or would enable such practices (e.g., computer vision, food recognition). For this reason, in preparing this survey, we mainly focused on "technology-oriented" online libraries, such as the IEEE Xplore Digital Library and the ACM Digital Library. Our initial search focused on work published in the past 5 years, to come up with the works we surveyed in sections 3–6. The following search terms were used: *commensality*, *eating*, *dining*, and *food* (see **Table 1** for details). Our initial search resulted in an initial selection of 2174 (i.e., the total number of ACM and IEEE references from 2014 to 2019, see **Table 1**). To get a more complete overview of the field, in the next step, we searched for relevant sources cited in the papers from the initial selection. This process resulted in an additional selection of 3040 sources (**Table 1**). It is important to notice that not all papers in this pool deal with food-related technology. For instance, several papers contain the keyword "eating" in used as a verb in the title, e.g., "how to have the cake and eat it too," without addressing eating-related research questions.

In addition to this, we also relied on sources from psychology and social sciences (i.e., Frontiers in Psychology and Psychological bulletin) and from Appetite. In doing so, we leveraged the aforementioned keywords, often combined with: *social facilitation*, *social comparison*, and *social context*. The final number of sources used in this survey is indicated in parentheses in **Table 1**. It is important to stress that our intention, when preparing this survey, was rather to show the variety of topics relevant to CC than performing a systematic survey of one research field. For example, we do not aim to enumerate all papers on serious games for changing eating habits published in the last 5 years, but we show a broad spectrum of works we deem relevant to CC systems and applications. Consequently, this survey is different from previous attempts of reviewing the existing works (see next section 1.2 for more detail), which usually focus on one aspect of food-related technology only.

## 1.2. Related Work

Min et al. (2019) proposed a survey on Food Computing, defined as an interdisciplinary field addressing food-related studies via computer science. In their view, Food Computing applies computational approaches for acquiring and analyzing heterogeneous food data from various sources for perception, recognition, retrieval, recommendation and monitoring of food to address food related issues in health, biology, gastronomy, and agronomy. Furthermore, Food Computing is conceptualized

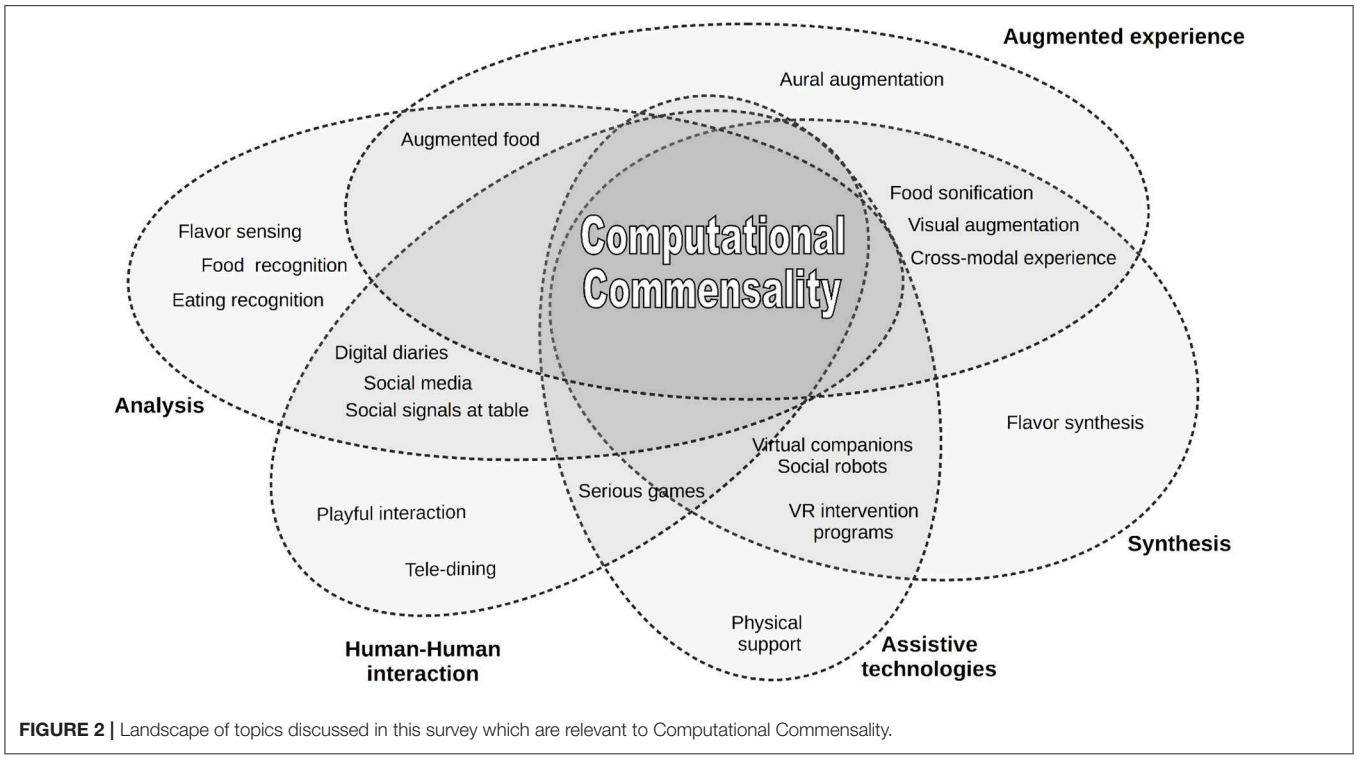



**FIGURE 2 |** Landscape of topics discussed in this survey which are relevant to Computational Commensality.

**TABLE 1 |** The total number of the papers per keyword found in the online libraries (in parenthesis the number of the papers cited in this survey corresponding to given period and the source).

| Keyword | ACM digital library | | IEEE Xplore digital library | |
|---|---|---|---|---|
| | 2014–2019 | 2010–2019 | 2014–2019 | 2010-2019 |
| Commensality | 13 (5) | 17 (6) | 1 (0) | 1 (0) |
| Eating | 322 (14) | 466 (19) | 670 (6) | 926 (10) |
| Food | 971 (10) | 1299 (17) | — | — |
| Dining | 66 (0) | 106 (3) | 131 (1) | 225 (3) |

*The research was performed on 23rd of August, 2019.*

as a collection of new methodologies and technologies for food science. According to the authors, Food Computing involves several steps. It requires data collection coming from different sources (e.g., social media, leveraging pictures and videos posted by users) and analysis carried out, for example, through machine learning or data mining techniques. At the same time, Food Computing has several applications, from food perception to recognition and from food recommendation to intake monitoring. The authors conclude by illustrating future directions and challenges of Food Computing. Although they mention that Food Computing might be involved in human behavior understanding especially in terms of the interaction of humans with food, Food Computing does not explicitly deal with the social dimension of food.

Shifting to Human-Computer Interaction (HCI), Grimes and Harper (2008) examined the literature on food and technology, pointing out that most of the existing works fall into two main categories: (1) technologies to solve food-related issues (for instance by helping inexperienced cooks to prepare a dish), and (2) technologies to modify the user's bad eating habits, namely "corrective technologies." In their view, HCI should also focus on the pleasurable and socio-cultural aspects of eating, and they introduce the concept of *celebratory technologies* "that celebrate the way that people interact with foods" (Grimes and Harper, 2008). Within this aim, according to the authors, several concepts related to eating should be explored, such as creativity, pleasure, nostalgia, gifting, family connectedness, trend-seeking, and relaxation. In the last part of their work, they describe challenges and provide a possible research direction toward this new type of technology. Interestingly, they point out different "social" aspects of food, such as food *gifting*, which has a symbolic social function. Comber et al. (2014) illustrate how food practices have gained importance in HCI, building what is called Human Food Interaction (HFI). The main areas of interest in HFI are health and wellbeing, sustainability, food experiences, and alternative food cultures. In this, HFI differentiates from commensality as the commensal experience seems not to be the core of HFI. HFI is in fact more focused on food practices as socio-cultural artifacts, examining cultural environmental and political aspects of human-food interaction. Furthermore, a new body of research on the experience of food, known by the term of *gastrophysics* (Spence, 2017b), has grown, describing the many factors driving food perception and enjoyment. The *scientific study of dining* illustrates how all sensory modalities drive food experience, along with the context in which the meal is consumed. In this sense, the social dimension can

influence the food experience as much as the edible elements on the plate.

## 2. FOOD AS A SOCIAL PHENOMENON

Food itself has an inherently social and emotional meaning. As such, it has been the subject of psychological and socio-psychological studies aiming at investigating it as a social phenomenon. Studies on traditional commensality investigated several interaction-related phenomena, such as conversational patterns in families (Laurier and Wiggins, 2011). Research on conversation analysis has drawn its attention to F&EA, for example, preparing (Paay et al., 2015) and sharing (Sterponi, 2003; Goodwin, 2007; Mondada, 2009) a meal. Research on workspace interaction has demonstrated how food and beverages, for istance, coffee, can have important communicative functions, with sipping becoming a cue to turn taking during conversations (Laurier, 2008). Furthermore, meal time is a key moment for addressing gestural interaction, as shown by Nansen et al. (2014). Other researchers focused on the cultural aspects of joint food practices (Fischler, 2011). They showed how the importance of the social dimension of food changes across cultures, being more relevant for French and Italian than for U.S. eaters. Ferdous et al. (2016b) illustrated how commensality and technology are often blended together, with technologies such as smartphones and tablets often being included in family dinners. The social dimension of eating is further explored by works on *social comparison*. In psychology, the term *social comparison* indicates how people spontaneously compare themselves to others as a means of self-evaluation (Festinger, 1954). Researchers have shown that mealtime is often a chance for social comparison (Polivy, 2017), with food perception and intake being influenced by others' presence (Polivy and Pliner, 2015). Social psychology has demonstrated how being with others can affect mealtime in terms of food intake (Bell and Pliner, 2003; Herman et al., 2003; Paquet et al., 2008; Hermans et al., 2009, 2012; Howland et al., 2012), food choice (Stroebele and De Castro, 2004; McFerran et al., 2009; Prinsen et al., 2013), calory consumption (Young et al., 2009), and taste perception (King et al., 2004; Poor et al., 2013) and how this influence is modulated by group membership (Cruwys et al., 2012), relationship status (Salvy et al., 2007), and, interestingly, that such influence still holds in virtual eating scenarios (Fox et al., 2009).

According to Simmel (1997), the meal has an *immeasurable sociological significance*. In fact, it is well known that, from a very young age, humans learn that eating is more than just introducing food into the body. Hamburg et al. (2014) illustrated how infants learn the soothing effect food can have and, more interestingly, according to the scope of this survey, that food offering can be a way to show empathy for others' distress. The term *comfort food* refers to those foods whose consumption provides consolation or a feeling of well-being (Spence, 2017a). In this view, food has the possibility of fostering positive emotions. Troisi et al. (2015) stated how food can have the capability of making people feel socially connected and call this property the *social utility* of food. Furthermore, they stress the idea of food as a social surrogate, demonstrating how (comfort) food choice is often associated with social isolation. As Connor and Armitage (2002) claimed, social psychologists contribute to food research by addressing topics as factors in food choice (McFerran et al., 2009), dietary change (Lange et al., 2018), weight control and well-being (Utter et al., 2018), snacking (Schüz et al., 2018), and food and self-presentation (Herman et al., 2003). In this context, the most thoroughly explored topic is what social psychology defines as the *social facilitation of eating* (e.g., Herman, 2015, 2017). This term describes the influence of presence of others on food intake. Diary studies have shown how meal size increases with the degree of intimacy with meal companions (De Castro, 1994) and how the presence of others models food intake, by acting as a guide on what and how to eat (Cruwys et al., 2015).

### 2.1. Summary

In our view, CC is a multidimensional phenomenon. So, the social and psychological aspects of F&EA should be taken into consideration when proposing computational models aimed at augmenting, analyzing or simply recognizing commensality. On one hand, several results mentioned in this section can relatively easily be replicated to check whether the technology-enhanced interaction can foster similar (positive) outcomes on interactants as it was observed in "traditional" commensality. This can include, for example, experiments on the quantity of food intake, or the degree of intimacy experienced by interactants. Work on socio-affective values of F&EA might contribute to new computational techniques for analyzing social interaction by relying on social, commensality-related cues. For instance, as humans can infer relationship statuses from observing people sharing food (Erwin et al., 2002), we could envisage technologies able to exploit F&EA-related cues in a similar way.

On the other hand, some existing computational models can be useful to quantify the social interaction around the table, and, thus, to address important research questions in the field. An example of how CC could benefit, e.g., social psychology, could consist in using a quantitative approach to detect the person at the table who is marginally participating in the conversation while eating (e.g., from their gaze behavior and amount of food intake (see Otsuka and Inoue, 2012).

## 3. TECHNOLOGY FOR FOOD AND EATING RECOGNITION

In order to account for the role of food and food related behaviors as non-verbal social signals, technologies must be able to afford their automated recognition. Technologies able to detect food related activities (for instance swallowing or drinking) contribute to CC because interaction in ecological settings often revolves around food (e.g., parties, dates, meetings). In particular, in the CC scenario 2, social robots and virtual agents would benefit from being able to detect when human interactants are involved in a F&EA, such as taking a bite of food or sip of a drink. Such a F&EA during interaction could influence, for example, turn-taking, which should be taken into account by the social robot

or virtual agent when it serves as a companion during food consumption. In addition, the artificial companion might be able to detect the type of food the human interactants are consuming and can comment on the qualities of the food in conversation.

This section reviews existing studies on food and eating detection: going from computer vision algorithms for food recognition (both on food image datasets and on pictures or video captures in the wild) to automatic trackers of eating activities (e.g., automatically detecting food intake quantity and speed).

## 3.1. Food Detection and Recognition

Several algorithms and techniques for food detection and recognition rely on the huge amount of pictures people share on social media every day. It must be noted that such pictures are mostly egocentric pictures, meaning that they are taken from the point of view of the user. They can be automatically captured by wearable cameras (e.g., for medical purposes) or taken by users of mobile devices (e.g., smartphones). They usually have low quality and poor framing, so the objects to be detected (e.g., plates) are far from the center of the picture, have scarce illumination, are deformed by the camera lens, and so on. These pictures are the most commonly shared by users on social media or on instant messaging apps, making them particularly interesting for automated analysis. To deal with food pictures, authors can exploit several existing datasets on food: *ILSVRC 2013* by Russakovsky et al. (2015), *Food101* by Bossard et al. (2014), *UECFood256* by Kawano and Yanai (2014), or *Egocentric Food*[3]. The resulting recognition models showed high accuracy in locating food, both in traditional and egocentric pictures, when there is no overlap between objects. Such approaches and resources, although not directly aimed at investigating commensality, are required steps toward analyzing human behavior during mealtime.

Bolaños et al. (2013) implemented a technique, based on machine learning, for the labeling of huge amounts of images. The algorithm they presented, based on a Hierarchical Sampling (HS) method, determines whether or not a plate is present in an image. According to the test on about 90k images, the algorithm can label all images in about 40 min in a totally unsupervised setting. Similarly, Ciocca et al. (2017) created the UNIMIB2016 dataset consisting of 3616 food instances belonging to 73 food classes (e.g., "pasta," "pizza," "yogurt"). The dataset is manually annotated to separate the food from the background. The authors also performed the automated recognition of food types using K-Nearest Neighbors (k-NN) and Support Vector Machine (SVM).

Aguilar et al. (2017) aims to build an application for automatic food habits tracking. It is a multi-labeling task, that is, a machine learning problem in which there are multiple output labels (instead of a single one), and it is solved using a Convolutional Neural Network (CNN). Results have shown good recognition rates also for recognizing ingredients of recipes that were not present in the training set.

Herranz et al. (2018) propose to take into account context and external knowledge in automated food detection. Context is, for example, the location, date, and time a food picture was taken. External knowledge includes food recipes, nutrition information, restaurant information, and food images and videos. In this framework, the authors review existing works on multimodal cuisine analysis, focusing on food recognition in restaurants. As mentioned before, studies exploiting egocentric pictures often have to deal with poor image quality; despite this, egocentric pictures (due to their great availability) are still leveraged, as in Jia et al. (2018). This work illustrates the development of the *eButton*, a small box containing a camera and a motion sensor. Using it, the authors showed that, even if the quality of egocentric images is lower than that of a smartphone, still it allows for food detection. To do that, they chose to exploit an existing CNN, the Clarifai CNN (Zeiler and Fergus, 2014), and compared food recognition on images captured by the eButton vs. images belonging to the Food-5K dataset (Singla et al., 2016). Results showed that the performances of the CNNs are comparable.

It is worth noticing that, as a consequence of the works we have described above, there already exist solutions for food recognition that take into consideration the context (e.g., location, like in Herranz et al., 2018). However, it seems that the social context is not considered (yet). The information about the group (e.g., the number of people involved) and the group bonds (e.g., their level of intimacy) can be contextual information helpful to recognizing some type of food. For example, some types of food are eaten usually in close company, such as, birthday cakes, raclette, fondue, or Korean BBQ, while the others are more often eaten alone, for istance, fast food. This example demonstrates the need to introduce models of CC in food recognition.

## 3.2. Eating Activity Detection and Recognition

As far as eating activity recognition is concerned it must be noted that activities linked to food preparation present a high intra-class variability, as highlighted in Stein and McKenna (2013b). They observed that recognition would be possible if large datasets were available, but this is not the case with food preparation activities. For this reason, they present work in which activity recognition is carried out by performing a training on a limited amount of data, collected in the publicly available *50 Salads dataset* presented in Stein and McKenna (2013a) and Chen et al. (2017). They compared two approaches: classifying (e.g., SVM, k-NN) activity of single users and then combining the results vs. performing a combined classification. They argued that the first one gives the most accurate results as it takes into account intra-user variability. Features were extracted from accelerometers attached to objects and from environmental video data (e.g., a camera framing the cooking area from top).

Wearable sensors can be used for eating recognition and detection, but they are typically intrusive. For instance, Bi et al. (2014) exploited a necklace-like device and a smartphone to capture throat sounds, and applied machine learning techniques (e.g., kNN, SVM) to determine the eating-related user activity (e.g., chewing, swallowing, breathing). The device could be applied to monitoring what and how people eat during the day to

[3]https://github.com/MarcBS/keras

better address food-related health problems like dysphagia and indigestion. Their system was based on a microphone and on the extraction of acoustic features to be later used for training and classification of eating-related activities, which reached over 95% accuracy. Amft and Tröster (2008), similarly to Bi et al. (2014), developed an on-body sensing approach to detect three key activities during food intake: arm movements, chewing, and swallowing. They applied Hidden Markov Models (HMM) on inertial sensors data to recognize arm movements. Chewing was recognized by analyzing the produced sounds. Swallowing was detected from the fusion data captured by two sensors: a surface EMG sensor and a stethoscope microphone. Moreover, Mendi et al. (2013) propose an application for eating activity recognition based on an accelerometer placed on the user's wrist, providing the user with information on the total number of bites, bites-taken rate and eating speed. The application is based on acceleration peaks detection and sends real-time warnings to the user when the eating speed is over a given threshold.

Rahman et al. (2015) highlight that eating is difficult to be accurately and unobtrusively recognized and analyzed. Worn sensors, for example, are deemed as uncomfortable and, in their opinion, they should be avoided. As an alternative, they propose to use Google Glass to track head movements, and they demonstrate that the captured inertial data (i.e., accelerometer, gyroscope, and magnetometer) from this device are informative enough to automatically recognize users' eating activity with traditional machine learning techniques, such as k-NN and RF. Interestingly, similarly to Stein and McKenna (2013b), Rahman et al. (2015) also see as the primary application of their work the possibility to better monitor and cure chronic diseases like obesity and diabetes. Other approaches to overcoming the intrusiveness of wearable sensors for eating recognition have been proposed. An interesting approach was illustrated by Chang et al. (2006), who designed a "diet-aware" dining table that used weight sensors and radio frequency identification (RFID) readers in order to measure food intake of diners at the table. The combination of weight sensors and RFID tags embedded in food containers enables the detection of food being moved from a central container to an individual's plate, allowing measures to be taken during a multi-party dinner. A first small-scale evaluation of the system showed recognition accuracy of food transfer from a central container to an individual plate and eating events to be around 80%. Another example of a device for the monitoring of food intake in ecological settings was proposed by Fontana et al. (2014), who developed a wearable system composed of a jaw motion sensor, a hand gesture sensor, and an accelerometer. The system is integrated with a smartphone equipped with a food intake recognition module which uses dedicated sensor fusion and pattern recognition techniques. The device was validated in real-life conditions over a one-day period by 12 subjects. Results showed that the system was able to detect food intake with an average accuracy of 89.8%.

An interesting contribution to CC was proposed by Kiriu et al. (2017) who, using the data from a smartwatch and a smartphone, are able to recognize whether a person is eating alone or in company, with an accuracy of 96%. The data consist of kinematics data (e.g., from 3D accelerometer) and several metrics of the smartphone. At the moment, this approach was tested only on a small dataset of 20 participants, but it showed a very interesting direction of research to be more deeply investigated in the framework of CC.

With the aim of addressing eating activities ecologically, while preserving their social dimension, it might be necessary to discriminate eating from speech. The goal of the work illustrated by Hantke et al. (2016) and Hantke et al. (2018), which was part of the EU iHEARu Project[4], was in fact to demonstrate that Automatic Speech Recognition can be improved by introducing the automatic recognition of eating conditions. To do that, they collected the iHEARu-EAT audio/video database, featuring 1.6k utterances of 30 subjects, 6 food types, and read/spontaneous speech. The authors performed a number of experiments in different conditions to discriminate between normal speech and eating speech, and to detect the type of food that was eaten while speaking. Results were positive, though the authors highlighted that the accuracy of detection, based on SVM, was strictly linked to the training that was carried out on some specific food (apple, nectarine, banana, Haribo Smurfs, biscuit, and crisps).

### 3.3. Future Developments Toward Computational Commensality

Overall, the existing technologies for food and eating recognition are not yet ready to be exploited in real-life applications. In a recent study, Alharbi et al. (2017) addressed the challenges of wearing devices (video camera, neck-worn sensor, and a wrist-worn sensor) for food activity monitoring to support weight management, mainly in terms of comfort of wearing a camera, and privacy. Results showed that participants had many concerns about privacy and had the feeling of a *social stigma* of wearing electronic devices that could worry other people around them.

Eating recognition might benefit from introducing the social context to the recognition models. One can imagine the recognition systems, in which subject data coming, for example, from an accelerometer placed on their wrist (e.g., similar to Mendi et al., 2013), is compensated by the data coming from similar devices placed on wrists of their eating companions. Research has in fact demonstrated that people eating together tend to mimic their companions' food intake, for instance in terms of bite rhythm (Hermans et al., 2012). Indeed, in the commensal scenarios, models for the recognition of eating activities should take into account social dynamics between the interaction partners, for istance, conversation turn-taking, social relations between the eaters (e.g., leadership, level of intimacy) but also other contextual data, such as the place of eating (e.g., fast food or an exclusive restaurant).

## 4. ASSISTIVE TECHNOLOGY

Much attention has been oriented toward the development of technologies to provide support in eating activities or during food preparation (as in Mennicken et al., 2010; Angara et al., 2017). Such technologies include systems offering physical support and assistance (e.g., for physically impaired people), mainly through

[4]http://www.tangsoo.de

the use of robots. A separate category consists of mobile apps that monitor and help to change the eating habits and increase overall well-being. Here we can distinguish two subcategories: some of the systems can supplement therapy related to some concrete health problem such as diabetes, while others can be used to improve the general habits, for example, by promoting a balanced diet, and, consequently, increasing the well-being of users. With a similar goal, several virtual and robotic assistants and serious games were developed. In particular, the gamification approach is a very popular method used specially in systems dedicated to young end users. All these applications focusing, at least at the moment, mainly on health and well-being related goals, are relevant to CC as they may relatively easily become CC use cases, in which one or more humans enter into interaction with a socially intelligent system (tutor, coach, assistant). One can, for example, easily imagine a virtual character which would not only assist the user by explicitly instructing her about healthy eating, but create a rich and fruitful social interaction, which can, indirectly, influence the well-being of the user and consequently her eating habits, too. To reach this aim, overall social skills of machines need to be improved.

## 4.1. Artificial Companions

Several systems have been designed to assist humans in changing their eating habits by leveraging the communicative (and sometimes affective) skills of humanoid assistants, be they virtual (e.g., embodied agents) or physical (e.g., robots). They usually address some very specific populations, such as hospital patients, children, or the elderly, as those groups often benefit from healthier life style, including healthier eating habits. Angara et al. (2017), proposed an interactive kitchen assistant giving health recommendations. Interestingly, the virtual agent's food-related interaction with the user was enriched by taking into account the user's food habits and cultural food preferences. Gardiner et al. (2017) evaluated the use of such technology to promote healthier eating behaviors. A virtual assistant is able to interact face-to-face with a user providing personalized dietary suggestions and health information (e.g., food recipes) and asking food-related questions to the user. A study on 61 female participants using the system during a 30-day span demonstrated a decrease of negative eating habits (e.g., drinking alcohol) and an increase of positives ones (e.g., fruit consumption). Such assistants do not need to have a human-like appearance, as in the case of the work by Pollak et al. (2010), who developed *Time to Eat*, a mobile virtual pet game designed to enhance healthy eating habits in teenagers. The pet sends healthy eating daily reminders to the user. In response, children take photos of their meals and snacks, of which the "healthiness" is evaluated by the app, which in turn influences the pet's emotional state (e.g., junk food corresponds to sadness, healthy food corresponds to happiness). Again, an evaluation involving 53 children showed that the app had positive effects on their eating habits. Parra et al. (2018) proposed an interesting combination between a human-like virtual assistant and crowd-sourcing. They developed an app with an e-assistant able to discuss the preparation of a meal with a human user. Then, the user can upload a photo of the meal and receive an evaluation provided by another user of the same app. The final system was evaluated on 59 patients, who found it useful, easy to use, and helpful in maintaining tasks "related to their diet."

With the aim of solving issues caused by solitary eating, Takahashi et al. (2017) proposed a virtual co-eating system allowing enjoyable conversations related to the meal, as well as typical daily conversation to be maintained. A virtual character is displayed on a mesh fabric, and the character has an embedded facial expression recognition module. The results of a preliminary evaluation of the system are particularly interesting in the CC context, as 4 out of 5 participants reported improvement when comparing the eating alone condition with the one in which they ate together with the virtual character.

Assistants promoting healthier eating styles can have physical bodies, as in the case of robots. Baroni et al. (2014) evaluated the effect of a humanoid robot on children's dietary choices. The robot can successfully persuade children to eat more fruit and vegetables by communicating verbally (modulating the voice, and using encouragements) and non-verbally (through gestures, proxemics, gaze). Eating habits in young children are also addressed by Randall et al. (2018) with their Health-e-Eater, a sensor-equipped plate and a simple robotic companion which motivates and educates children during meals. Health-e-Eater is a low-cost robot architecture based on a Raspberry Pi 3, equipped with LEDs, a vibration motor, a servomotor, and a speaker. LED lights and verbal messages are supposed to focus the attention of the child on the food, encouraging and rewarding them when a healthy eating style is detected. McColl and Nejat (2013) proposed an assistive robot designed to cognitively stimulate and engage the elderly during eating. Starting from existing studies on the role of the interaction during meals in the improvement of dietary intake (e.g., Schell and Kayser-Jones, 1999), they designed an autonomous robot able to detect the amount of food intake while interacting with the user, both verbally and non-verbally. Some of the robot utterances are directly related to the eating itself (e.g., encouragements), while the others are aimed at enhancing the interaction (e.g., greetings, telling jokes, laughing). An exploratory study was conducted on a group of elderly, and results showed that participants felt engaged, enjoyed themselves, and cooperated with the robot in response to its prompting behaviors.

In general, we believe that social robots are particularly appropriate to become commensal partners, but very little research has been presented on this topic so far. Within the aim to exploit the positive outcomes of commensality, Khot et al. (2019) recently proposed a robotic dining companion called FoBo (see **Figure 3**). The role of this robot is to create playful and entertaining interactions around a meal with no clear "real-world" goal. So, it does not instruct or correct human's behavior but, instead, for istance, it "consumes" batteries, performs sounds related to eating (e.g., burping and purring), as well as mimics some human behaviors.

## 4.2. Virtual and Augmented Reality

Virtual and augmented reality are also used to create situations aimed to change the human eating habits. For instance, Celikcan et al. (2018) proposed the Virtual Cafeteria–VR immersive environment designed for nutrition education of adolescents.

**FIGURE 3 |** FoBo—a robotic dining companion. Reproduced from Khot et al. (2019) with permission from the authors who hold the copyrights.

The virtual buffet offers a large selection of foods and drinks covering the three meals. The users create their own meals and can pick any portion for any available food. At the same time, they are given age appropriate recommendations on healthy eating and recommended portion sizes of each food group. The data related to the user activity are collected, so as soon as the session ends, the detailed nutritional information of the assembled meal is immediately available. In a similar vein, the Virtual Food Court (Nordbo et al., 2015) is a VR environment for studying humans' food choices in the context of policy-based interventions. It was successfully used to analyze the effects of introducing taxes for unhealthy food on food choices. Narumi et al. (2012) propose a system based on Augmented Reality (AR) and a Head Mounted Display (HMD) to influence food consumption and the perception of satiety by exploring the phenomenon of cross-modality. The system allows for the augmentation of food volume using the shape deformation techniques to give the user the impression of consuming more than she does in reality. The evaluation shows a significant effect of size (enlarged vs. shrunken) on the quality of food consumed. Participants ate significantly more food when the size was virtually decreased as compared to the condition of virtually increased food. According to the authors, such a system might be useful in treating obesity. Additional evidence for positive effects of VR on helping people in acquiring more healthy eating habits is provided by Tuanquin et al. (2018). The authors used VR to change the visual and olfactory appearance of food items in order to gradually change a person's eating preferences toward more healthy food choices. The goal was to help individuals with eating disorders through VR cue-exposure therapy (CET). Results of a first study, in which participants were presented with actual and virtual chocolate chip cookies, showed that the VR setup was able to successfully induce food craving and the urge to eat the cookie. According to the authors, the VR setup shows potential to aid in CET by presenting virtual food cues during therapy. Indeed, VR is increasingly recognized as a potentially useful tool to study human behavior regarding food choice (Nordbo et al., 2015; Ung et al., 2018), and food cravings (Ledoux et al., 2013), as well as research into the sensory aspects of food selection and consumption in general (Stelick and Dando, 2018).

## 4.3. Dedicated Sensing Devices

As for other devices that can positively intervene in human eating habits, Hermsen et al. (2016) carried out an experiment involving a smart fork (i.e., a fork-shaped device augmented with sensors and actuators). The fork can provide real-time haptic and visual feedback to the user (Kadomura et al., 2013), for example, producing alerts if the user eats too quickly. Eleven participants who perceive themselves as "fast eaters" were asked to use the fork during 3 days. Most of them reported an increased awareness of their eating rate, and decrease of the eating speed. Drink-O-Mender by Ritschel et al. (2018) is able to sense the type and amount of drinks consumed by an adult, providing verbal advice depending on calories and nutritional values. For example, it may try to attract the person's attention toward the drinks with the lowest amount of calories.

## 4.4. Future Developments for Creation of Computational Commensality With Artificial Companions

Assistive technologies, at least at the moment, rarely explore the social bonds with their users. An interesting exception we mentioned above are the works by McColl and Nejat (2013) and Khot et al. (2019), where the robot builds an interaction for which the aim is not only functional (i.e., assistive) but also social. Future solutions may include, for example, tools for reciprocal assistance. Moreover, even if social aspects of eating are considered, it is usually limited to dyadic interactions. A robot bartender by Foster et al. (2012) is a rare example of an artificial companion able to deal with multiple humans in a dynamic social setting. Their robot is able to engage in multiple socially appropriate interactions at the same time when performing a task-oriented activity (i.e., serving drinks).

Similarly, immersive VR/AR systems currently focus on the individual experience. Systems such as the Virtual Cafeteria mentioned in this section can easily become multi-user social systems, where different users can interact and exchange their experiences regarding the food, similarly to how they do now using dedicated forums (see, e.g., Parra et al., 2018).

At the same time, it is important to stress that examples of artificial (eating) companions that do not have either an assistive or coaching role, are even more rare. Liu and Inoue (2014) propose a virtual eating companion that aims to be an active listener in order to support the generation of new ideas. According to the authors, the person who has a meal is likely to become an attentive listener, while the other interactant more likely becomes a speaker. Based on this assumption, the authors created a virtual character whose eating behavior is modeled on the quantitative analysis of actual dining behavior. For this purpose, recordings of multiple students eating together were used. Performing such analysis of human-human interaction is a good first step to create CC applications. Unfortunately,

such works are still rare, especially when we consider papers that use technology to automatically quantify social interactions during meals.

## 5. TECHNOLOGY FOR AUGMENTED FLAVOR EXPERIENCES

In the literature dealing with food and technology there is a large body of work on the use of technology to alter flavor experiences (Spence and Piqueras-Fiszman, 2013; Bruijnes et al., 2016). These works are grounded in research into the multi-sensory nature of flavor experiences (Velasco et al., 2018). The central notion is that flavor is a multi-sensory construct of which the percept results from a combination of information from several sensory channels (Auvray and Spence, 2008). A change to one sensory channel (e.g., the color of the food) can potentially influence the flavor experience of the food consumed. Several techniques exist that can be used to digitally alter the flavor experience of food. This is of interest to CC because such alterations could be used to create new social dining experiences, new ways of socially sharing food experiences, and can give robotic or virtual dining companions some form of control over actual food being consumed by human co-diners.

### 5.1. Visual Flavor Augmentation

Considering flavor as a multi-sensory construct, changing the visual appearance of food has been demonstrated to have an impact on flavor experiences (Zampini et al., 2007). A potentially interesting method to digitally alter the visual appearance of actual food is the use of projection mapping (Kita and Rekimoto, 2013). In one study, projection mapping was used to alter the visual appearance of yogurt. Colors, shapes, and animations were found to have the potential to change flavor experiences (Huisman et al., 2016).

### 5.2. Auditory Flavor Augmentation

Auditory feedback can be used to change the perceived texture of food, altering the overall experience of eating, say, crisps (Zampini and Spence, 2004; Koizumi et al., 2011). Wang et al. (2018) propose a five-keys framework for augmentation of the eating experience with sounds. According to them, (1) new sounds can be generated (that are different from the natural sounds of consumed food), (2) the natural sounds can be amplified, (3) removed, or (4) blended with other (food related) sounds. Finally, the sounds can also be (5) distorted. Within this framework, the authors propose the Singing Carrot, a platform for the exploration of food sonifications, which generates sounds when the user eats a carrot. The system detects food consumption through capacitive touch sensing, and the value of sensed capacitance is mapped to the frequency of a sinewave, resulting in eating sound that dynamically changes. A similar concept is used in the iScream! system (Wang et al., 2019, see **Figure 4**), which allows the use of a novel "gustosonic" experience of digital sounds which are automatically created as a result of eating an ice cream. It also uses capacitive sensing to detect eating actions, and based on these actions, it plays different sounds to create a playful eating experience.

**FIGURE 4 |** The user is playing with iScream! Reproduced from Wang et al. (2019) with permission from the authors who hold the copyrights.

### 5.3. Haptic Flavor Augmentation

Augmentation of haptic sensations, for example through electrical muscle stimulation, can create augmented experiences of food texture (Niijima and Ogawa, 2016). Iwata et al. (2004) in their research focused specifically on creating a simulation of mastication, using haptic technology. The biting force used to chew on real food items (e.g., a cracker) was recorded and used as data in the system to modulate the physical resistance provided by the haptic device in order to produce sensations of biting into different food items.

### 5.4. Chemical Flavor Augmentation

Ranasinghe et al. (2011) and Ranasinghe and Do (2017) highlight how taste and smell are the senses allowing us to remember emotions and feelings, as they directly influence our mood, stress, retention, and recall functions (Drewnowski, 1997). With this in mind, Ranasinghe and Do (2017) created the Digital Lollipop, a device that synthesizes taste (e.g., sweet, salty, sour, bitter, and umami) by applying a small electrical and thermal (i.e., applying heat vs. cool) stimulation to tongue. The Digital Lollipop consists of two silver electrodes, a sphere and a plate, and can generate square wave pulses with a current ranging from 20 to 200 $\mu A$ with frequencies in the range 50–1,200 Hz. The tongue must be placed between the electrodes. In an experiment presented in Ranasinghe and Do (2017) the authors observed that, by placing the electrodes on the tongue tip and sides, 90% of participants perceived sourness, 70% saltiness, 50% bitterness, and 5% sweetness (corresponding to the case in which the current was inverted). Some participants perceived a tingling, pineapple-like sensation when current increased. Tip and side stimulations exhibited slight variations in the observations, mainly related to the intensity needed to elicit the sensations, which was lower. Note, however, that there are specific challenges to using technology to address the chemical senses, such as the notion that people generally pay less attention to smells when they are engaged in another task as well as the need for using capsules to produce artificial scents (Spence et al., 2017).

### 5.5. Multimodal Flavor Augmentation

There are also multimodal approaches where digital augmentation is used to address multiple sensory channels at the same time. For example, Narumi et al. (2011) used a HMD to alter the visual appearance of cookies. An olfactory display was synchronized to the visuals to create the illusion of, for example, chocolate flavors. In another example, researchers used electrical stimulation of the taste buds on the tongue, in addition to lights, and an olfactory display to augment the flavor of an actual drink (Ranasinghe et al., 2017, see also Narumi et al., 2010; Ranasinghe et al., 2014, 2016).

One solution could be to create VR experiences that more carefully integrate with "real" experiences, by having actual objects (e.g., food or drinks) also be virtually represented in VR (Harley et al., 2018). Automatic computer vision-based food recognition solutions could be used in such an approach way to create compelling mixed reality experiences (Kanak et al., 2018).

### 5.6. Future Developments for Multi-Sensory Commensality

There is a large body of research on multi-sensory flavor experiences (for an overview see Auvray and Spence, 2008), and this research has since found its way into high-end restaurants such as Sublimotion[5] and Ultra Violet[6] (see Spence and Piqueras-Fiszman, 2014 for additional examples), where diners not only experience haute cuisine but also high-tech.

While the research discussed above does not directly bear on commensality, it can be argued that flavor augmentation and flavor synthesis have a potential future role to play in social communication and CC. For example, Ranasinghe et al. (2011) discuss how their flavor synthesizing technology could be used for flavor communication between remotely located individuals, creating a new kind of remote communication. Similarly, flavor augmentation technology could be used to share experiences around actual food items. Remotely located diners could potentially adapt the flavor experience of their companion's food to their own flavor experience (see also section 6.2). The potential future application of these technologies could be envisioned to be in the realm of social media communication and the sharing of experiences through social media, similar to how applications such as Instagram are now used to share visual aspects of food. Connecting such social food sharing to food printing technologies would create interesting forms of social "food messaging" (Wei et al., 2014).

Other applications of food augmenting technologies in commensal scenarios might include using food augmentation to communicate emotional and social states of interaction partners, for exmaple, by dynamically changing their food properties such as a color or by adding the relevant sonification. These alterations could be part of CC models that also drive social behaviors of robotic or virtual dining companions so that they can interact with humans through the food on the table.

[5]https://www.sublimotionibiza.com
[6]https://uvbypp.cc

## 6. TECHNOLOGY FOR FOSTERING HUMAN-HUMAN INTERACTION

Several technologies have been proposed that can be considered digital extensions of social activities. For example, job meetings can take place through video-conference (Jo et al., 2016) or artistic performances through Networked Music Performance (Rottondi et al., 2016). Similarly, technologies such as tele-dining platforms were proposed as a digital counterpart of eating activities. In this section, we describe technologies that are supposed to be able to deal with the social aspects of food and eating-related activities and which are designed to make eating more social as well as more enjoyable.

### 6.1. Serious Games and Playful Interaction

Particularly popular are solutions which combine educational purposes related to food intake with entertainment, often in the form of a serious game. Such games often introduce elements of competition or cooperation between two or more players, and thus, a social dimension to the activity. We4Fit by Pereira et al. (2014) is an example of such a mobile app that uses a gamification approach to modify the motivation of users to change eating habits and promote a healthy lifestyle. Interestingly, it is a rare example of a collaborative food-related game: in the game, the user (or a team) posts pictures of the consumed food. Other participants rank the photos indicating how healthy the photographed food is. At the end of each seven-day round, the sum of the ranks is used to establish the winning team. Playing in teams, according to the authors, should enhance the motivational effect, as the members of the team can influence each other to obtain a better final score. Another application that uses similar mechanisms of interaction between the users is called Foodie Moodie (ElSayed et al., 2018). This app promotes the awareness about the relation between the type of food consumed and the mood. It allows the users to keep track of what they eat and understand how it may affect their mood, as well as provides guidelines to other users about the possible interrelation between mood and food. A gamification techniques were also included in the app: first, the users can collaborate by adding and (re-)viewing the others' hints related to the topic. They may also compete with each other trying to obtain the highest total score on their tips. Other serious games use immersive virtual environments and AR. Ganesh et al. (2014) used interactive projection mapping to introduce game elements on children's plates during eating, with the goal of addressing children's reluctance to eat certain types of, predominantly healthy, food. The system is composed of two applications: one which changed the color of food items and one that awarded points and virtual badges for eating healthy items. The system was evaluated with children and their families at home. Observational data indicated that children ate food items they were otherwise reluctant to eat, and showed a playful attitude to food and the system. In addition, the system served to stimulate interactions between parents and children regarding healthy food intake. The authors considered the system's role in enhancing parent-child interaction and interactions between siblings to be particularly important. In this sense, it can be considered an example of commensal technology.

*You Better Eat to Survive!* by Arnold (2017) is a VR game in which eating real food becomes an input to control the narrative of the VR game. The players work in teams: when one player tries to realize some tasks in a virtual world, the other needs to feed him with real food to keep the first one "alive." In the background, the system is able to detect the eating events via a microphone placed near to the first player's mouth. The VR game objective is to get rescued from an island. During its exploration, the player must keep himself alive by eating regularly. Otherwise, due to hunger, he loses consciousness, which is the end of the game. The team players can succeed in the game only if they collaborate, thus the game becomes a social experience. *Feed the Food Monsters!* by Arza et al. (2018) is a two-player AR game that uses chewing real food as an input to control the flow of the game. In order to achieve the goal (i.e., feed the virtual monsters that live in the stomach) the participants need to chew slowly. The participants can monitor each other's chewing behaviors through an interface that is displayed on their torso. Thus, they can also interact and guide each other to chew properly. AR is used to visualize the process of digestion though the means of using playful animations rather than showing the actual human anatomy.

The *Restaurant Game* (Orkin and Roy, 2010) is an example of a virtual commensality platform for human virtual agent interaction. The aim is 2-fold: it is designed to collect the data of human interactions, for example, when playing the a role of a customer or waiter in a 3D virtual environment, as well as to generate plausible behaviors of virtual agents. In the system, humans control characters from a first-person perspective using the mouse and chat. Agents are also able to interact and build a conversation both with humans and other agents. The behavior patterns of the agent are learned automatically from logs of the previous game sessions. Although it is not clear whether the previous game sessions include also the logs of human-human interactions in the VR system, such an extension would definitely be valuable in the context of CC.

Other examples of playful approaches to enhance interaction during dining using technology involve physical installations. One playful approach is presented by Mehta et al. (2018). The *Arm-a-dine* system involves two users both wearing a robot arm attached to their chests. A mobile phone camera is used to detect the facial expressions of the diners. For example, if a negative facial expression is detected the wearers own robotic arm will pick up a food item and present it to the wearer. If a positive facial expression, such as a smile, is detected the robotic arm of the partner will offer a food item to the person smiling. The central concept of the system revolves around taking away some bodily control in order to stimulate new kinds of social interactions around food consumption. One finding from a first exploratory study indicated that feeding another person using the robotic arm was an enjoyable social experience, and that it could potentially serve as an ice-breaker between strangers. In a similar vein, Mitchell et al. (2015) designed an actuated dining table where two people can eat together. The table lowers the plate of the person who is eating too fast, and raises the plate of the other so that diners' eating speeds become aligned. According to the authors, misalignment in eating pace between co-diners can create social friction and discomfort, something their system aims to address. The concept of the interactive table is also explored by Kado et al. (2010), who present a more abstract approach of agency with their *sociable dining table* (SDT). Users interact with the SDT by knocking on it which, according to the authors, serves as a minimal social cue to interact with "creatures"; actuated tableware such as a pot and a dish that can move around the table. An exploratory study indicated that users were able to guide the creatures around the table through the knocking interaction. While the social dimension of the installation deserves further study it is interesting to consider agency in a broader scope through interactions with robotic table wear. Li et al. (2018) explore ingestible sensors, i.e., microsystems that perform sensing inside the body. As an example they propose HeatCraft–two user interactive system which measures the internal body temperature of the one player and communicates it to the other player though thermal stimuli.

Humans often interact not only when eating but also when preparing food. Foodie by Wei and Nakatsu (2012) is an example of system that allows for joint design and creation of real food. The system is composed of the Food Creation Interface—a mobile app to design the food (e.g., to define its shape, color), and the Food Crafting Mechanism—a robot which crafts the designed food. In the use case scenario, multiple persons, by using their mobile devices, design new food together and send the project to another user whose robot generates edible food (e.g., a sandwich).

## 6.2. Tele-Dining

As demonstrated in Ferdous (2015), technology can enhance commensal experiences. During family mealtimes technology can scaffold and shape social interaction (Ferdous et al., 2016a). In Ferdous et al. (2016b), they present *TableTalk*, affording diners the possibility to bring together their photos, videos, audio, and other digital media to create a shared commensal technological experience. Nevertheless, the positive role of technology in these instances is focused on the traditional setting of the family dinner. Nawahdah and Inoue (2013) highlight that family dining is becoming very difficult today. Young people and the elderly are increasingly living independently, and working people tend to either travel frequently or work remotely, as also observed by Sellaeg and Chapman (2008). Hence, technology to enhance commensality can no longer only focus on the shared family dining table, but has to take into account distal interactions. Commensality can be re-introduced by exploring the possibility of what is called *remote commensality* (Foley-Fisher et al., 2010; Wei et al., 2011a; Grevet et al., 2012; Komaromi Haque, 2016). An example of a system aimed at creating a sense of remote commensality is the *KIZUNA* system (Nawahdah and Inoue, 2013; Inoue and Nawahdah, 2014). The system enables asynchronous dining interaction between people living in different time zones. The idea is that a person can experience remote co-dining with another person by watching a pre-recorded video of the other person's dining. The system works by separately recording the dining actions of two persons dining at different times and plays back these recordings by modulating

the playback time to ensure the synchronization between the real and the recorded person. As the authors highlight in their work, it is not enough that people can merely watch another person dining, as it also happens, for example, with the *Culater* (Tsujita et al., 2010), to have the illusion of co-dining. It is the synchronization between their actions that contributes to achieving this illusion. The *KIZUNA* system was validated through questionnaires asking participants to rate the sense of presence they perceived from the remote (pre-recorded) person and the overall satisfaction of communication. The test had two conditions, one in which participants had dinner while watching a pre-recorded video of someone else eating and another one using a Wizard-of-Oz approach to simulate the *KIZUNA system*. Results showed that the system was preferred both in terms of presence perception and overall satisfaction of communication.

Similarly, Heidrich et al. (2012) presented the *Room XT* concept which consists of a wall-sized projection, head-tracking, and 3D rendering to create the illusion of sitting across from another person at a table. In this setup, head-tracking would allow for the projection to be adjusted to the point-of-view of the person looking at the projection to create the illusions of depth. The concept was implemented in a scaled-down setup using a computer monitor and Kinect sensor to demonstrate the potential of the depth illusion in a shared dining experience.

The importance of synchronized multimodal signals during remote co-dining is underlined by the work of Wei et al. (2011b) in the design of the *CoDine* system. The system consists of a large video screen, Kinect sensor, augmented table and tablecloth, and food printer. Remotely located diners can see each other through the screen and use the combination of the screen and Kinect sensor to engage in gesture-based interactions with the system through icons displayed on the screen. The icons can be selected to share messages through the tablecloth which, through thermochromic ink and Peltier elements, can change color to display simple shapes. Similarly, on-screen icons can be selected to create printed food shapes on the remotely located other's food items using the food printer. Finally, the augmented table is embedded with a movable magnet that allows a diner to remotely move another person's tableware, the idea being that this enables a form of sharing that is typically only possible during co-located dining. Aspects of the *CoDine* system were later implemented in *Foodie*, aimed at social interactions through printable food. Foodie by Wei and Cheok (2012) is an integrated system that allows for joint design and creation of real food (see section 6.1).

Where *CoDine* and *Foodie* enable interactions between two remotely located diners, the telematic dinner party system by Barden et al. (2012) allows remotely located groups of people to engage in interactions during dinner. After observational studies of dinner parties and an initial prototype design, the final design consisted of a set of round tables where three diners would gather around. Webcams were used to capture diners at one table and a projector was used to project visuals of remotely located diners. A projection area showed a visualization of a remotely located diner from the other table. In the center of each table a rotating platform was used to present food. Diners could physically rotate the platform on their table, which would result in an identical rotation in the platform of the remotely located table. The setup was evaluated during several dining scenarios (e.g., murder mystery dinner party), showing that communication was not as fluent as during a co-located dinner party. Conversely, participants did engage in playful behavior during the scenarios predominantly by manipulating the rotating platform, for example, while a remotely located diner was just about to reach for an item of food. The authors suggest that the element of playfulness helped overcome technical limitations while at the same time resulting in more of a performance than an actual dinner party.

### 6.3. Future Developments for Enhance Computational Commensality Between Humans

To conclude this section it is worth noticing that the systems mentioned in section 6.1 allow for some interaction between multiple humans. It can be as simple as trying to perform a better score in a game than all other competitors, or very complex scenarios requiring cooperation and which revolve around the topics of eating and food. The latter also show that the technology can change *eating* into a play and create an experience, which is enjoyable not because of the (consumed) food, but mainly due to connecting people by tasks that require joint actions in the physical space (Altarriba Bertran et al., 2018; Chisik et al., 2018). In this sense, for istance, works by Mehta et al. (2018) or Mitchell et al. (2015) are examples of CC, which would not be possible without using the technology (i.e., in traditional human-human setting).

The technology to enable remote commensality is becoming increasingly more sophisticated, and researchers have made headway in creating systems that allow individuals some form of visual communication and in some cases shared interaction with actual food items. However, it remains to be investigated whether or not these systems provide the same benefits of actual commensality—for example, the ones mentioned in section 2. In addition, these systems do not necessarily provide solutions for individuals experiencing (chronic) loneliness due to a lack of sufficiently satisfying social connections. Nevertheless, one may wonder whether the use of these systems could also be sought in shared dining with strangers, which could be seen as a potential approach to create new, hopefully in the end, satisfying, social relationships. The technology implemented in tele-dining systems could then also be used as conversation support technology to stimulate strangers to engage with each other socially (Otsuka and Inoue, 2012).

## 7. FINAL DISCUSSION

Eating is a highly social activity, and so are everyday eating-related actions. For this reason, we believe computational models and techniques aimed at reading and understanding human

non-verbal social interaction should pay attention to eating-related behaviors. In this survey, we hope to have provided an overview of existing psychological studies, approaches and technologies aimed at addressing, creating or augmenting commensality. The body of literature discussed shows that CC draws on many different fields. It is a complex, multi-disciplinary field of research still in its early stages. Therefore, a number of hiatuses remain that deserve to be addressed in future research.

Like current smart phone use, adding technology to the dining table will change social situations and rituals around the food consumed. Therefore, CC should take into account the impact of the technology once it is introduced into the dining sphere. For example, consider using a VR headset to visually augment food experiences. In such a situation, it becomes very difficult to share food experiences between co-diners due to the fact that the headset will make regular face-to-face interactions very difficult. Similarly, gamification, and augmentation technology in general, can also serve as potential distractors from food and food consumption. As an educational approach, one could question whether creating distractions is beneficial to long-term food enjoyment and healthy eating habits of children, for example. More generally, the argument can be put forth that any kind of technology that distracts from the actual food or genuine interaction during dinner can have potentially detrimental effects on food enjoyment, healthy eating, and conducive social eating habits.

At the same time, potential opportunities for sharing food and flavor experiences across distances (e.g., while connected through the internet) can be enabled by the same technologies, potentially providing commensal experiences where none were possible before. Future work on CC should carefully consider how the addition of technology to commensality will impact already existing social eating practices.

One way to have technology more seamlessly integrated into current dining practices is to move it into the background, and to adapt it to different eating situations as needed. However, technology to track and recognize food items is not fully implemented yet in many of the systems that have been discussed in this review. Therefore the manipulations, for example those aimed at guiding diners' behavior using AR, are typically hand-built to match the food items presented to participants. Considerable effort should be put into automatic recognition of food items in order to create seamless, automatic augmentations in an interaction loop where the food is recognized and the digital augmentation is automatically generated based on the recognition (e.g., to match or contrast some of the qualities of the food). Only through such integrations can these systems have a real place at the dining table, especially as far as commensal dining is concerned (see also previous point).

Works discussed in this survey that provide augmented food experiences often do so in controlled lab settings. One can question how strongly lab-based manipulations affect (commensal) food experiences in a real-life dining settings. Some restaurants do experiment with CC (e.g., Sublimotion, Ultra Violet), but there is currently a lack of research showing effects of technological augmentations on food experiences in ecologically valid settings. In relation to ecologically valid research, it is important to stress cultural aspects of commensality. There are strong differences between various cultures in commensal eating (Kittler et al., 2011; Counihan and Van Esterik, 2012; Anderson, 2014). However, little research in CC, be it related to food recognition, changing flavor experiences, or providing support through artificial social agents, takes into account cultural differences in a structural way. Therefore, it can be recommended for research to move away from focusing on WEIRD (Henrich et al., 2010) samples, and include more culturally diverse samples. At the very least, research in CC should be mindful of the fact that results may be limited to specific socio-cultural settings, and be difficult to generalize beyond that setting. As an example, in this context, nearly all commensal technologies listed in this paper (e.g., Kado et al., 2010; Mitchell et al., 2015) assume that eating is organized around the table (whether real, augmented or virtual). At the same time, it is well-known that people in several cultures eat and interact when eating without using such furniture. Consequently, it might be important also to develop culture-specific CC.

From a more technological perspective, it is important to carefully consider the validity of computational models of human-human behaviors at the table. Existing models dedicated to social signal processing, for example, leadership (Beyan et al., 2018; Niewiadomski et al., 2018), cohesion (Hung and Gatica-Perez, 2010), and turn-taking (de Kok and Heylen, 2009) might not necessarily be appropriate for analyzing behaviors in commensal scenarios. Indeed, when eating together, humans perform at least two different activities in parallel: eating and socializing. Both of these activities could be considered to interfere with each other. For example, when chewing or focusing on the food on the plate social behaviors that are typical in non-commensal social settings can be disrupted (e.g., turn-taking). In addition, the non-verbal behaviors and the communication with eating partners are limited by the position at the table and the distance to the interlocutors. These unique aspects of social interactions that occur while consuming food highlight the need to build new, multimodal corpora of commensality. Here, it is important to take into account spontaneous behavioral aspects related to the food specifically (e.g., food recognition, mastication) as well as the social behaviors that occur between co-located humans.

Indeed, such models may be a requirement to create truly social interactions with artificial social entities such as assistive robots or virtual coaches. These social interactions should not only be focused on food, diet, and eating behaviors, to name but a few application areas covered in this review, but should include interactions around other topics, from small talk about the weather to discussing the day at work. Through such more complete social interactions the bond between the user and the artificial social entity can potentially be strengthened. Only when embodied computational systems, such as social robots, can participate on some level in all the complexities of social interactions during meal time can we move toward true CC.

To conclude, we have seen in this paper that even if technology is often integrated in eating practices already,

there is still the need for technologies capable of reading and generating social signals that are associated with such practices. This should motivate HCI and AI researchers to give more attention to different social aspects of food related interactions. Our hope is that this work will contribute to kick off new research and strengthen existing research initiatives in diverse fields toward the creation of novel computational models dealing with commensal food preparation and consumption.

## AUTHOR CONTRIBUTIONS

All authors contributed to the introduction and revised the manuscript. EC: took main charge of shaping food as a social phenomenon. GH: technology for augmented flavor experiences and final discussion. MM: technology for food and eating recognition and technology for fostering human-human interaction. RN: assistive technology and technology for fostering human-human interaction.

## REFERENCES


Aguilar, E., Bolaños, M., and Radeva, P. (2017). "Food recognition using fusion of classifiers based on CNNs," in *International Conference on Image Analysis and Processing* (Catania: Springer), 213–224.

Alharbi, R., Pfammatter, A., Spring, B., and Alshurafa, N. (2017). "Willsense: adherence barriers for passive sensing systems that track eating behavior," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, CO: ACM), 2329–2336.

Altarriba Bertran, F., Jhaveri, S., Lutz, R., Isbister, K., and Wilde, D. (2018). "Visualising the landscape of human-food interaction research," in *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems*, DIS '18 Companion (New York, NY: ACM), 243–248.

Amft, O., and Tröster, G. (2008). Recognition of dietary activity events using on-body sensors. *Artif. Intell. Med.* 42, 121–136. doi: 10.1016/j.artmed.2007.11.007

Anderson, E. N. (2014). *Everyone Eats: Understanding Food and Culture*. New York, NY: NYU Press.

Angara, P., Jiménez, M., Agarwal, K., Jain, H., Jain, R., Stege, U., et al. (2017). "Foodie fooderson a conversational agent for the smart kitchen," in *Proceedings of the 27th Annual International Conference on Computer Science and Software Engineering* (Toronto, ON: IBM Corp), 247–253.

Arnold, P. (2017). "You better eat to survive! exploring edible interactions in a virtual reality game," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17 (New York, NY: ACM), 206–209.

Arza, E. S., Kurra, H., Khot, R. A., and Mueller, F. F. (2018). "Feed the food monsters!: helping co-diners chew their food better with augmented reality," in *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, CHI PLAY '18 Extended Abstracts (New York, NY: ACM), 391–397.

Auvray, M., and Spence, C. (2008). The multisensory perception of flavor. *Conscious. Cogn.* 17, 1016–1031. doi: 10.1016/j.concog.2007.06.005

Barden, P., Comber, R., Green, D., Jackson, D., Ladha, C., Bartindale, T., et al. (2012). "Telematic dinner party: designing for togetherness through play and performance," in *Proceedings of the Designing Interactive Systems Conference* (Newcastle upon Tyne, UK: ACM), 38–47.

Baroni, I., Nalin, M., Zelati, M. C., Oleari, E., and Sanna, A. (2014). "Designing motivational robot: how robots might motivate children to eat fruits and vegetables," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication* (Edinburgh), 796–801.

Bell, R., and Pliner, P. L. (2003). Time to eat: the relationship between the number of people eating and meal duration in three lunch settings. *Appetite* 41, 215–218. doi: 10.1016/S0195-6663(03)00109-0

Beyan, C., Capozzi, F., Becchio, C., and Murino, V. (2018). Prediction of the leadership style of an emergent leader using audio and visual nonverbal features. *IEEE Trans. Multimedia* 20, 441–456. doi: 10.1109/TMM.2017.2740062

Bi, Y., Xu, W., Guan, N., Wei, Y., and Yi, W. (2014). "Pervasive eating habits monitoring and recognition through a wearable acoustic sensor," in *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '14 (Brussels:Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, ICST), 174–177.

Bolaños, M., Garolera, M., and Radeva, P. (2013). "Active labeling application applied to food-related object recognition," in *Proceedings of the 5th International Workshop on Multimedia for Cooking & Eating Activities* (Barcelona: ACM), 45–50.

Bossard, L., Guillaumin, M., and Van Gool, L. (2014). "Food-101–mining discriminative components with random forests," in *European Conference on Computer Vision* (Zurich: Springer), 446–461.

Bruijnes, M., Huisman, G., and Heylen, D. (2016). "Tasty tech: human-food interaction and multimodal interfaces," in *Proceedings of the 1st Workshop on Multi-sensorial Approaches to Human-Food Interaction* (Tokyo: ACM), 4.

Celikcan, U., Bulbul, A. S., Aslan, C., Buyuktuncer, Z., Isgin, K., Ede, G., et al. (2018). "The virtual cafeteria: an immersive environment for interactive food portion-size education," in *Proceedings of the 3rd International Workshop on Multisensory Approaches to Human-Food Interaction*, MHFI'18 (New York, NY: ACM), 5:1–5:5.

Chang, K.-H., Liu, S.-Y., Chu, H.-H., Hsu, J. Y.-J., Chen, C., Lin, T.-Y., et al. (2006). "The diet-aware dining table: observing dietary behaviors over a tabletop surface," in *International Conference on Pervasive Computing* (Pisa: Springer), 366–382.

Chen, C., Jafari, R., and Kehtarnavaz, N. (2017). A survey of depth and inertial sensor fusion for human action recognition. *Multimedia Tools Appl.* 76, 4405–4425. doi: 10.1007/s11042-015-3177-1

Chisik, Y., Pons, P., and Jaen, J. (2018). "Gastronomy meets ludology: towards a definition of what it means to play with your (digital) food," in *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, CHI PLAY '18 Extended Abstracts (Melbourne, VIC: ACM), 155–168.

Ciocca, G., Napoletano, P., and Schettini, R. (2017). Food recognition: a new dataset, experiments, and results. *IEEE J. Biomed. Health Informatics* 21, 588–598. doi: 10.1109/JBHI.2016.2636441

Comber, R., Choi, J. H.-J., Hoonhout, J., and O'hara, K. (2014). Editorial: designing for human-food interaction: an introduction to the special issue on 'food and interaction design'. *Int. J. Hum. Comput. Stud.* 72, 181–184. doi: 10.1016/j.ijhcs.2013.09.001

Connor, M., and Armitage, C. J. (2002). *The Social Psychology of Food.* Philadelphia, PA: Open University Press.

Counihan, C., and Van Esterik, P. (2012). *Food and Culture: A Reader*. New York, NY: Routledge.

Cruwys, T., Bevelander, K. E., and Hermans, R. C. (2015). Social modeling of eating: a review of when and why social influence affects food intake and choice. *Appetite* 86, 3–18. doi: 10.1016/j.appet.2014.08.035

Cruwys, T., Platow, M. J., Angullia, S. A., Chang, J. M., Diler, S. E., Kirchner, J. L., et al. (2012). Modeling of food intake is moderated by salient psychological group membership. *Appetite* 58, 754–757. doi: 10.1016/j.appet.2011.12.002

De Castro, J. M. (1994). Family and friends produce greater social facilitation of food intake than other companions. *Physiol. Behav.* 56, 445–455. doi: 10.1016/0031-9384(94)90286-0

de Kok, I., and Heylen, D. (2009). "Multimodal end-of-turn prediction in multi-party meetings," in *Proceedings of the 2009 International Conference on Multimodal Interfaces*, ICMI-MLMI '09 (New York, NY: ACM), 91–98.

Drewnowski, A. (1997). Taste preferences and food intake. *Annu. Rev. Nutr.* 17, 237–253. doi: 10.1146/annurev.nutr.17.1.237

ElSayed, M. N., Abdennadher, S., and Gabr, F. M. (2018). "Foodie moodie: a crowdsourcing platform for interrelating food with mood," in *2018 IEEE 6th International Conference on Serious Games and Applications for Health (SeGAH)* (Wien), 1–5.

Erwin, P. G., Burke, A., and Purves, D. G. (2002). Food sharing and perceptions of the status of a relationship. *Percept. Motor Skills* 94, 506–508. doi: 10.2466/pms.2002.94.2.506

Ferdous, H. S. (2015). "Technology at mealtime: beyond the'ordinary,'" in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (Seoul: ACM), 195–198.

Ferdous, H. S., Ploderer, B., Davis, H., Vetere, F., and O'hara, K. (2016a). Commensality and the social use of technology during family mealtime. *ACM Trans. Comput. Hum. Interact.* 23:37. doi: 10.1145/2994146

Ferdous, H. S., Ploderer, B., Davis, H., Vetere, F., O'Hara, K., Farr-Wharton, G., et al. (2016b). "Tabletalk: integrating personal devices and content for commensal experiences at the family dinner table," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg: ACM), 132–143.

Festinger, L. (1954). A theory of social comparison processes. *Hum. Relat.* 7, 117–140. doi: 10.1177/001872675400700202

Fischler, C. (2011). Commensality, society and culture. *Soc. Sci. Inform.* 50, 528–548. doi: 10.1177/0539018411413963

Foley-Fisher, Z., Tsao, V., Wang, J., and Fels, S. (2010). "Netpot: easy meal enjoyment for distant diners," in *International Conference on Entertainment Computing* (Seoul: Springer), 446–448.

Fontana, J. M., Farooq, M., and Sazonov, E. (2014). Automatic ingestion monitor: a novel wearable device for monitoring of ingestive behavior. *IEEE Trans. Biomed. Eng.* 61, 1772–1779. doi: 10.1109/TBME.2014.2306773

Foster, M. E., Gaschler, A., Giuliani, M., Isard, A., Pateraki, M., and Petrick, R. P. (2012). "Two people walk into a bar: dynamic multi-party social interaction with a robot agent," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12 (New York, NY: ACM), 3–10.

Fox, J., Bailenson, J., and Binney, J. (2009). Virtual experiences, physical behaviors: the effect of presence on imitation of an eating avatar. *Presence* 18, 294–303. doi: 10.1162/pres.18.4.294

Ganesh, S., Marshall, P., Rogers, Y., and O'Hara, K. (2014). "Foodworks: tackling fussy eating by digitally augmenting children's meals," in *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational* (Helsinki: ACM), 147–156.

Gardiner, P. M., McCue, K. D., Negash, L. M., Cheng, T., White, L. F., Yinusa-Nyahkoon, L., et al. (2017). Engaging women with an embodied conversational agent to deliver mindfulness and lifestyle recommendations: a feasibility randomized control trial. *Patient Educ. Counsel.* 100, 1720–1729. doi: 10.1016/j.pec.2017.04.015

Goodwin, M. H. (2007). Occasioned knowledge exploration in family interaction. *Discourse Soc.* 18, 93–110. doi: 10.1177/0957926507069459

Grevet, C., Tang, A., and Mynatt, E. (2012). "Eating alone, together: new forms of commensality," in *Proceedings of the 17th ACM International Conference on Supporting Group Work* (Sanibel Island, FL: ACM), 103–106.

Grimes, A., and Harper, R. (2008). "Celebratory technology: new directions for food research in HCI," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence: ACM), 467–476.

Hamburg, M. E., Finkenauer, C., and Schuengel, C. (2014). Food for love: the role of food offering in empathic emotion regulation. *Front. Psychol.* 5:32. doi: 10.3389/fpsyg.2014.00032

Hantke, S., Schmitt, M., Tzirakis, P., and Schuller, B. (2018). "Eat-: the icmi 2018 eating analysis and tracking challenge," in *Proceedings of the 2018 on International Conference on Multimodal Interaction* (Boulder: ACM), 559–563.

Hantke, S., Weninger, F., Kurle, R., Ringeval, F., Batliner, A., Mousa, A. E.-D., et al. (2016). I hear you eat and speak: automatic recognition of eating condition and food type, use-cases, and impact on ASR performance. *PLoS ONE* 11:e0154486. doi: 10.1371/journal.pone.0154486

Harley, D., Verni, A., Willis, M., Ng, A., Bozzo, L., and Mazalek, A. (2018). "Sensory VR: Smelling, touching, and eating virtual reality," in *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction* (Stockholm: ACM), 386–397.

Heidrich, F., Kasugai, K., Röcker, C., Russell, P., and Ziefle, M. (2012). "RoomXT: advanced video communication for joint dining over a distance," in *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2012 6th International Conference on* (San Diego, CA: IEEE), 211–214.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). Most people are not weird. *Nature* 466:29. doi: 10.1038/466029a

Herman, C. P. (2015). The social facilitation of eating. A review. *Appetite* 86, 61–73. doi: 10.1016/j.appet.2014.09.016

Herman, C. P. (2017). The social facilitation of eating or the facilitation of social eating? *J. Eat. Disord.* 5:16. doi: 10.1186/s40337-017-0146-2

Herman, C. P., Roth, D. A., and Polivy, J. (2003). Effects of the presence of others on food intake: a normative interpretation. *Psychol. Bull.* 129:873. doi: 10.1037/0033-2909.129.6.873

Hermans, R. C., Engels, R. C., Larsen, J. K., and Herman, C. P. (2009). Modeling of palatable food intake. The influence of quality of social interaction. *Appetite* 52, 801–804. doi: 10.1016/j.appet.2009.03.008

Hermans, R. C., Lichtwarck-Aschoff, A., Bevelander, K. E., Herman, C. P., Larsen, J. K., and Engels, R. C. (2012). Mimicry of food intake: the dynamic interplay between eating companions. *PLoS ONE* 7:e31027. doi: 10.1371/journal.pone.0031027

Hermsen, S., Frost, J. H., Robinson, E., Higgs, S., Mars, M., and Hermans, R. C. (2016). Evaluation of a smart fork to decelerate eating rate. *J. Acad. Nutr. Dietetics* 116, 1066–1067. doi: 10.1016/j.jand.2015.11.004

Herranz, L., Min, W., and Jiang, S. (2018). Food recognition and recipe analysis: integrating visual content, context and external knowledge. *arXiv preprint* arXiv:1801.07239.

Howland, M., Hunger, J. M., and Mann, T. (2012). Friends don't let friends eat cookies: effects of restrictive eating norms on consumption among friends. *Appetite* 59, 505–509. doi: 10.1016/j.appet.2012.06.020

Huisman, G., Bruijnes, M., and Heylen, D. K. (2016). "A moving feast: effects of color, shape and animation on taste associations and taste perceptions," in *Proceedings of the 13th International Conference on Advances in Computer Entertainment Technology* (Osaka: ACM), 13.

Hung, H., and Gatica-Perez, D. (2010). Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Trans. Multimedia* 12, 563–575. doi: 10.1109/TMM.2010.2055233

Inoue, T., and Nawahdah, M. (2014). "Influence of dining-progress synchrony in time-shifted tele-dining," in *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14 (New York, NY: ACM), 2089–2094.

Iwata, H., Yano, H., Uemura, T., and Moriya, T. (2004). "Food simulator: a haptic interface for biting," in *Virtual Reality, 2004. Proceedings. IEEE* (Chicago, IL: IEEE), 51–57.

Jia, W., Li, Y., Qu, R., Baranowski, T., Burke, L. E., Zhang, H., Bai, Y., et al. (2018). Automatic food detection in egocentric images using artificial intelligence technology. *Public Health Nutr.* 22, 1168–1179. doi: 10.1017/S1368980018000538

Jo, D., Kim, K. H., and Kim, J. (2016). "Effects of avatar and background representation forms to co-presence in mixed reality (MR) tele-conference systems," in *SA 2016 - SIGGRAPH ASIA 2016 Virtual Reality Meets Physical Reality: Modelling and Simulating Virtual Humans and Environments* (Anaheim, CA: Association for Computing Machinery, Inc).

Kado, Y., Kamoda, T., Yoshiike, Y., De Silva, P. R., and Okada, M. (2010). "Sociable dining table: The effectiveness of a "konkon" interface for reciprocal adaptation," in *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on* (Osaka: IEEE), 105–106.

Kadomura, A., Li, C.-Y., Chen, Y.-C., Tsukada, K., Siio, I., and Chu, H.-H. (2013). "Sensing fork: eating behavior detection utensil and mobile persuasive game," in *CHI'13 Extended Abstracts on Human Factors in Computing Systems* (Paris: ACM), 1551–1556.

Kanak, A., Özlü, A., Polat, S. O., and Ergün, Ö. Ö. (2018). "An intelligent dining scene experience," in *2018 26th Signal Processing and Communications Applications Conference (SIU)* (Izmir: IEEE), 1–3.

Kawano, Y., and Yanai, K. (2014). "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *European Conference on Computer Vision* (Zurich: Springer), 3–17.

Khot, R. A., Arza, E. S., Kurra, H., and Wang, Y. (2019). "Fobo: Towards designing a robotic companion for solo dining," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19 (New York, NY: ACM).

King, S. C., Weber, A. J., Meiselman, H. L., and Lv, N. (2004). The effect of meal situation, social interaction, physical environment and choice on food acceptability. *Food Qual Prefer.* 15, 645–653. doi: 10.1016/j.foodqual.2004.04.010

Kiriu, K., Ochiai, K., Inagaki, A., Yamamoto, N., Fukazawa, Y., Kimoto, M., et al. (2017). "Recognizing whether a person is eating alone or has company by using wearable devices," in *2017 Tenth International Conference on Mobile Computing and Ubiquitous Network (ICMU)* (Toyama), 1–2.

Kita, Y., and Rekimoto, J. (2013). "Spot-light: multimodal projection mapping on food," in *International Conference on Human-Computer Interaction* (Las Vegas, NV: Springer), 652–655.

Kittler, P. G., Sucher, K. P., and Nelms, M. (2011). *Food and Culture*. Belmont, CA: Cengage Learning.

Koizumi, N., Tanaka, H., Uema, Y., and Inami, M. (2011). "Chewing jockey: augmented food texture by using sound based on the cross-modal effect," in *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology* (Lisbon: ACM), 21.

Komaromi Haque, J. (2016). Synchronized dining tangible mediated communication for remote commensality. *Master Thesis*. Malmo: Malmo University.

Kortum, P. (2008). *HCI Beyond the GUI: Design for Haptic, Speech, Olfactory, and Other Nontraditional Interfaces*. San Francisco, CA: Elsevier.

Lange, D., Corbett, J., Knoll, N., Schwarzer, R., and Lippke, S. (2018). Fruit and vegetable intake: the interplay of planning, social support, and sex. *Int. J. Behav. Med.* 25, 421–430. doi: 10.1007/s12529-018-9718-z

Laurier, E. (2008). Drinking up endings: conversational resources of the café. *Lang. Commun.* 28, 165–181. doi: 10.1016/j.langcom.2008.01.011

Laurier, E., and Wiggins, S. (2011). Finishing the family meal. The interactional organisation of satiety. *Appetite* 56, 53–64. doi: 10.1016/j.appet.2010.11.138

Ledoux, T., Nguyen, A. S., Bakos-Block, C., and Bordnick, P. (2013). Using virtual reality to study food cravings. *Appetite* 71, 396–402. doi: 10.1016/j.appet.2013.09.006

Li, Z., Chen, W., Wang, Y., Hoang, T., Wang, W., Boot, M., et al. (2018). "Heatcraft: playing with ingestible sensors via localised sensations," in *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, CHI PLAY '18 Extended Abstracts (New York, NY: ACM), 521–530.

Liu, R., and Inoue, T. (2014). "Application of an anthropomorphic dining agent to idea generation," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct (New York, NY; Seattle, WA: ACM), 607–612.

McColl, D., and Nejat, G. (2013). Meal-time with a socially assistive robot and older adults at a long-term care facility. *J. Hum.Robot Interact.* 2, 152–171. doi: 10.5898/JHRI.2.1.McColl

McFerran, B., Dahl, D. W., Fitzsimons, G. J., and Morales, A. C. (2009). I'll have what she's having: effects of social influence and body type on the food choices of others. *J. Consum. Res.* 36, 915–929. doi: 10.1086/644611

Mehta, Y. D., Khot, R. A., Patibanda, R., and Mueller, F. (2018). "Arm-a-Dine: towards understanding the design of playful embodied eating experiences," in *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play* (Melbourne: ACM), 299–313.

Mendi, E., Ozyavuz, O., Pekesen, E., and Bayrak, C. (2013). "Food intake monitoring system for mobile devices," in *Advances in Sensors and Interfaces (IWASI), 2013 5th IEEE International Workshop on* (Bari: IEEE), 31–33.

Mennicken, S., Karrer, T., Russell, P., and Borchers, J. (2010). "First-person cooking: a dual-perspective interactive kitchen counter," in *CHI'10 Extended Abstracts on Human Factors in Computing Systems* (Atlanta, GA: ACM), 3403–3408.

Min, W., Jiang, S., Liu, L., Rui, Y., and Jain, R. (2019). A survey on food computing. *ACM Comput. Surv.* 52:92. doi: 10.1145/3329168

Mitchell, R., Papadimitriou, A., You, Y., and Boer, L. (2015). "Really eating together: a kinetic table to synchronise social dining experiences," in *Proceedings of the 6th Augmented Human International Conference* (Singapore: ACM), 173–174.

Mondada, L. (2009). The methodical organization of talking and eating: assessments in dinner conversations. *Food Qual. Prefer.* 20, 558–571. doi: 10.1016/j.foodqual.2009.03.006

Mueller, F. F., Kari, T., Khot, R., Li, Z., Wang, Y., Mehta, Y., et al. (2018). "Towards experiencing eating as a form of play," in *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, CHI PLAY '18 Extended Abstracts (New York, NY: ACM), 559–567.

Nansen, B., Davis, H., Vetere, F., Skov, M., Paay, J., and Kjeldskov, J. (2014). "Kitchen kinesics: situating gestural interaction within the social contexts of family cooking," in *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: the Future of Design* (Sydney: ACM), 149–158.

Narumi, T., Ban, Y., Kajinami, T., Tanikawa, T., and Hirose, M. (2012). "Augmented perception of satiety: controlling food consumption by changing apparent size of food with augmented reality," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12 (New York, NY: ACM), 109–118.

Narumi, T., Nishizaka, S., Kajinami, T., Tanikawa, T., and Hirose, M. (2011). "Augmented reality flavors: gustatory display based on edible marker and cross-modal interaction," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC: ACM), 93–102.

Narumi, T., Sato, M., Tanikawa, T., and Hirose, M. (2010). "Evaluating cross-sensory perception of superimposing virtual color onto real drink: toward realization of pseudo-gustatory displays," in *Proceedings of the 1st Augmented Human International Conference* (Mègeve: ACM), 18.

Nawahdah, M., and Inoue, T. (2013). "Virtually dining together in time-shifted environment: Kizuna design," in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, TX: ACM), 779–788.

Niewiadomski, R., Chauvigne, L., Mancini, M., and Camurri, A. (2018). "Towards a model of nonverbal leadership in unstructured joint physical activity," in *Proceedings of the 5th International Conference on Movement and Computing*, MOCO '18 (New York, NY: ACM), 20:1–20:8.

Niijima, A., and Ogawa, T. (2016). "A proposal of virtual food texture by electric muscle stimulation," in *Multimedia & Expo Workshops (ICMEW), 2016 IEEE International Conference on* (Seattle; Washington: IEEE), 1–6.

Nordbo, K., Milne, D., Calvo, R. A., and Allman-Farinelli, M. (2015). "Virtual food court: a VR environment to assess people's food choices," in *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*, OzCHI '15 (New York, NY: ACM), 69–72.

Ochs, E., and Shohet, M. (2006). The cultural structuring of mealtime socialization. *New Dir. Child Adolesc. Dev.* 2006, 35–49. doi: 10.1002/cd.154

Orkin, J., and Roy, D. K. (2010). "Capturing and generating social behavior with the restaurant game," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1*, AAMAS '10 (Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems), 1765–1766.

Otsuka, Y., and Inoue, T. (2012). "Designing a conversation support system in dining together based on the investigation of actual party," in *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on* (Manchester: IEEE), 1467–1472.

Paay, J., Kjeldskov, J., and Skov, M. B. (2015). "Connecting in the kitchen: an empirical study of physical interactions while cooking together at home," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC: ACM), 276–287.

Paquet, C., St-Arnaud-McKenzie, D., Ma, Z., Kergoat, M.-J., Ferland, G., and Dube, L. (2008). More than just not being alone: the number, nature, and complementarity of meal-time social interactions influence food intake in hospitalized elderly patients. *Gerontologist* 48, 603–611. doi: 10.1093/geront/48.5.603

Parra, M. O., Favela, J., Castro, L. A., and Morales, A. (2018). Monitoring eating behaviors for a nutritionist e-assistant using crowdsourcing. *Computer* 51, 43–51. doi: 10.1109/MC.2018.1731078

Pereira, C. V., Figueiredo, G., Esteves, M. G. P., and de Souza, J. M. (2014). "We4Fit: A game with a purpose for behavior change" Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD)," in *Hsinchu*, 83–88. doi: 10.1109/CSCWD.2014.6846821

Polivy, J. (2017). What's that you're eating? social comparison and eating behavior. *J. Eat. Disord.* 5:18. doi: 10.1186/s40337-017-0148-0

Polivy, J., and Pliner, P. (2015). "She got more than me". Social comparison and the social context of eating. *Appetite* 86, 88–95. doi: 10.1016/j.appet.2014.08.007

Pollak, J., Gay, G., Byrne, S., Wagner, E., Retelny, D., and Humphreys, L. (2010). It's time to eat! Using mobile games to promote healthy eating. *IEEE Pervas. Comput.* 9, 21–27. doi: 10.1109/MPRV.2010.41

Poor, M., Duhachek, A., and Krishnan, H. S. (2013). How images of other consumers influence subsequent taste perceptions. *J. Market.* 77, 124–139. doi: 10.1509/jm.12.0021

Prinsen, S., de Ridder, D. T., and de Vet, E. (2013). Eating by example. Effects of environmental cues on dietary decisions. *Appetite* 70, 1–5. doi: 10.1016/j.appet.2013.05.023

Rahman, S. A., Merck, C., Huang, Y., and Kleinberg, S. (2015). "Unintrusive eating recognition using Google Glass," in *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare* (Graz: ICST, Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 108–111.

Ranasinghe, N., and Do, E. Y.-L. (2017). Digital lollipop: studying electrical stimulation on the human tongue to simulate taste sensations. *ACM Trans. Multimedia Comput. Commun. Appl.* 13:5. doi: 10.1145/2996462

Ranasinghe, N., Karunanayaka, K., Cheok, A. D., Fernando, O. N. N., Nii, H., and Gopalakrishnakone, P. (2011). "Digital taste and smell communication," in *Proceedings of the 6th International Conference on Body Area Networks* (Berlin: ICST, Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 78–84.

Ranasinghe, N., Lee, K.-Y., and Do, E. Y.-L. (2014). "Funrasa: an interactive drinking platform," in *Proceedings of the 8th International Conference on Tangible, Embedded and Embodied Interaction* (New York, NY: ACM), 133–136.

Ranasinghe, N., Lee, K.-Y., Suthokumar, G., and Do, E. Y.-L. (2016). Virtual ingredients for food and beverages to create immersive taste experiences. *Multimedia Tools Appl.* 75, 12291–12309. doi: 10.1007/s11042-015-3162-8

Ranasinghe, N., Nguyen, T. N. T., Liangkun, Y., Lin, L.-Y., Tolley, D., and Do, E. Y.-L. (2017). "Vocktail: a virtual cocktail for pairing digital taste, smell, and color sensations," in *Proceedings of the 2017 ACM on Multimedia Conference* (Mountain View, CA: ACM), 1139–1147.

Randall, N., Joshi, S., and Liu, X. (2018). "Health-e-Eater: Dinnertime companion robot and magic plate for improving eating habits in children from low-income families," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18 (New York, NY: ACM), 361–362.

Ritschel, H., Seiderer, A., Janowski, K., Aslan, I., and André, E. (2018). "Drink-o-mender: an adaptive robotic drink adviser," in *Proceedings of the 3rd International Workshop on Multisensory Approaches to Human-Food Interaction*, MHFI'18 (New York, NY: ACM), 3:1–3:8.

Rottondi, C., Chafe, C., Allocchio, C., and Sarti, A. (2016). An overview on networked music performance technologies. *IEEE Access* 4, 8823–8843. doi: 10.1109/ACCESS.2016.2628440

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Salvy, S.-J., Jarrin, D., Paluch, R., Irfan, N., and Pliner, P. (2007). Effects of social influence on eating in couples, friends and strangers. *Appetite* 49, 92–99. doi: 10.1016/j.appet.2006.12.004

Schell, E. S., and Kayser-Jones, J. (1999). The effect of role-taking ability on caregiver-resident mealtime interaction. *Appl. Nurs. Res.* 12, 38–44. doi: 10.1016/S0897-1897(99)80167-0

Schüz, B., Papadakis, T., and Ferguson, S. G. (2018). Situation-specific social norms as mediators of social influence on snacking. *Health Psychol.* 37:153. doi: 10.1037/hea0000568

Sellaeg, K., and Chapman, G. E. (2008). Masculinity and food ideals of men who live alone. *Appetite* 51, 120–128. doi: 10.1016/j.appet.2008.01.003

Simmel, G. (1997). *Sociology of the Meal*. London: Sage.

Singla, A., Yuan, L., and Ebrahimi, T. (2016). "Food/non-food image classification and food categorization using pre-trained googlenet model," in *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management* (Amsterdam: ACM), 3–11.

Spence, C. (2017a). Comfort food: a review. *Int. J. Gastron. Food Sci.* 9, 105–109. doi: 10.1016/j.ijgfs.2017.07.001

Spence, C. (2017b). *Gastrophysics: The New Science of Eating*. Oxford:Penguin.

Spence, C., Obrist, M., Velasco, C., and Ranasinghe, N. (2017). Digitizing the chemical senses: possibilities & pitfalls. *Int. J. Hum. Comput. Stud.* 107, 62–74. doi: 10.1016/j.ijhcs.2017.06.003

Spence, C., and Piqueras-Fiszman, B. (2013). Technology at the dining table. *Flavour* 2:16. doi: 10.1186/2044-7248-2-16

Spence, C., and Piqueras-Fiszman, B. (2014). *The Perfect Meal: The Multisensory Science of Food and Dining*. New York, NY: John Wiley & Sons.

Stein, S., and McKenna, S. J. (2013a). "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Zurich: ACM), 729–738.

Stein, S., and McKenna, S. J. (2013b). "User-adaptive models for recognizing food preparation activities," in *Proceedings of the 5th International Workshop on Multimedia for Cooking & Eating Activities* (Barcelona: ACM), 39–44.

Stelick, A., and Dando, R. (2018). Thinking outside the booth-the eating environment, context and ecological validity in sensory and consumer research. *Curr. Opin. Food Sci.* 21, 26–31. doi: 10.1016/j.cofs.2018.05.005

Sterponi, L. A. (2003). Account episodes in family discourse: the making of morality in everyday interaction. *Discourse Stud.* 5, 79–100. doi: 10.1177/1461445603005001840

Stroebele, N., and De Castro, J. M. (2004). Effect of ambience on food intake and food choice. *Nutrition* 20, 821–838. doi: 10.1016/j.nut.2004.05.012

Takahashi, M., Tanaka, H., Yamana, H., and Nakajima, T. (2017). "Virtual co-eating: making solitary eating experience more enjoyable," in *Entertainment Computing – ICEC 2017*, eds N. Munekata, I. Kunita, and J. Hoshino (Cham: Springer International Publishing), 460–464.

Troisi, J. D., Gabriel, S., Derrick, J. L., and Geisler, A. (2015). Threatened belonging and preference for comfort food among the securely attached. *Appetite* 90, 58–64. doi: 10.1016/j.appet.2015.02.029

Tsujita, H., Yarosh, S., and Abowd, G. D. (2010). "Cu-later: a communication system considering time difference," in *Proceedings of the 12th ACM International Conference Adjunct Papers on Ubiquitous Computing-Adjunct* (Copenhagen: ACM), 435–436.

Tuanquin, N. M. B., Hoermann, S., Petersen, C. J., and Lindeman, R. W. (2018). "The effects of olfactory stimulation and active participation on food cravings in virtual reality," in *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (Reutlingen: IEEE), 709–710.

Ung, C.-Y., Menozzi, M., Hartmann, C., and Siegrist, M. (2018). Innovations in consumer research: the virtual food buffet. *Food Qual. Prefer.* 63, 12–17. doi: 10.1016/j.foodqual.2017.07.007

Utter, J., Larson, N., Berge, J. M., Eisenberg, M. E., Fulkerson, J. A., and Neumark-Sztainer, D. (2018). Family meals among parents: associations with nutritional, social and emotional wellbeing. *Prev. Med.* 113, 7–12. doi: 10.1016/j.ypmed.2018.05.006

Velasco, C., Obrist, M., Petit, O., and Spence, C. (2018). Multisensory technology for flavor augmentation: a mini review. *Front. Psychol.* 9:26. doi: 10.3389/fpsyg.2018.00026

Wang, Y., Li, Z., Jarvis, R., Khot, R. A., and Mueller, F. F. (2018). "The singing carrot: designing playful experiences with food sounds," in *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, CHI PLAY '18 Extended Abstracts (New York, NY: ACM), 669–676.

Wang, Y., Li, Z., Jarvis, R., Khot, R. A., and Mueller, F. F. (2019). "iscream!: towards the design of playful gustosonic experiences with ice cream," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19 (New York, NY: ACM), INT047:1–INT047:4.

Wei, J., and Cheok, A. D. (2012). Foodie: play with your food promote interaction and fun with edible interface. *IEEE Trans. Consum. Electron.* 58, 178–183. doi: 10.1109/TCE.2012.6227410

Wei, J., Ma, X., and Zhao, S. (2014). "Food messaging: using edible medium for social messaging," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, ON: ACM), 2873–2882.

Wei, J. and Nakatsu, R. (2012). "Leisure food: derive social and cultural entertainment through physical interaction with food," in *International Conference on Entertainment Computing* (Bremen: Springer), 256–269.

Wei, J., Peiris, R. L., Koh, J. T. K. V., Wang, X., Choi, Y., Martinez, X. R., et al. (2011a). "Food media: exploring interactive entertainment over telepresent

dinner," in *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology* (Lisbon: ACM), 26.

Wei, J., Wang, X., Peiris, R. L., Choi, Y., Martinez, X. R., Tache, R., et al. (2011b). "Codine: an interactive multi-sensory system for remote dining," in *Proceedings of the 13th International Conference on Ubiquitous Computing* (Beijing: ACM), 21–30.

Young, M. E., Mizzau, M., Mai, N. T., Sirisegaram, A., and Wilson, M. (2009). Food for thought. what you eat depends on your sex and eating companions. *Appetite* 53, 268–271. doi: 10.1016/j.appet.2009.07.021

Zampini, M., Sanabria, D., Phillips, N., and Spence, C. (2007). The multisensory perception of flavor: assessing the influence of color cues on flavor discrimination responses. *Food Qual. Prefer.* 18, 975–984. doi: 10.1016/j.foodqual.2007.04.001

Zampini, M., and Spence, C. (2004). The role of auditory cues in modulating the perceived crispness and staleness of potato chips. *J. Sens. Stud.* 19, 347–363. doi: 10.1111/j.1745-459x.2004.080403.x

Zeiler, M. D., and Fergus, R. (2014). "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision* (Zurich: Springer), 818–833.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 08 November 2019
doi: 10.3389/frobt.2019.00116

# Synchronization in Interpersonal Speech

*Shahin Amiriparian[1]*, Jing Han[1], Maximilian Schmitt[1], Alice Baird[1], Adria Mallol-Ragolta[1], Manuel Milling[1], Maurice Gerczuk[1] and Björn Schuller[1,2]*

*[1] ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany, [2] Group on Language, Audio & Music, Imperial College London, London, United Kingdom*

During both positive and negative dyadic exchanges, individuals will often unconsciously imitate their partner. A substantial amount of research has been made on this phenomenon, and such studies have shown that synchronization between communication partners can improve interpersonal relationships. Automatic computational approaches for recognizing synchrony are still in their infancy. In this study, we extend on previous work in which we applied a novel method utilizing hand-crafted low-level acoustic descriptors and autoencoders (AEs) to analyse synchrony in the speech domain. For this purpose, a database consisting of 394 in-the-wild speakers from six different cultures, is used. For each speaker in the dyadic exchange, two AEs are implemented. Post the training phase, the acoustic features for one of the speakers is tested using the AE trained on their dyadic partner. In this same way, we also explore the benefits that deep representations from audio may have, implementing the state-of-the-art Deep Spectrum toolkit. For all speakers at varied time-points during their interaction, the calculation of reconstruction error from the AE trained on their respective dyadic partner is made. The results obtained from this acoustic analysis are then compared with the linguistic experiments based on word counts and word embeddings generated by our *word2vec* approach. The results demonstrate that there is a degree of synchrony during all interactions. We also find that, this degree varies across the 6 cultures found in the investigated database. These findings are further substantiated through the use of 4,096 dimensional Deep Spectrum features.

**Keywords: speech synchronization, human-human interaction, computational paralinguistics, machine learning, speech processing, autoencoders**

OPEN ACCESS

***Edited by:***
*Agnieszka Wykowska, Italian Institute of Technology Istituto Italiano di Tecnologia (IIT), Italy*

***Reviewed by:***
*Sean Andrist, Microsoft Research, United States*
*Kerstin Fischer, University of Southern Denmark, Denmark*

****Correspondence:***
*Shahin Amiriparian*
*shahin.amiriparian@informatik.uni-augsburg.de*

***Specialty section:***
*This article was submitted to Human-Robot Interaction, a section of the journal Frontiers in Robotics and AI*

***Received:*** *01 March 2019*
***Accepted:*** *22 October 2019*
***Published:*** *08 November 2019*

***Citation:***
*Amiriparian S, Han J, Schmitt M, Baird A, Mallol-Ragolta A, Milling M, Gerczuk M and Schuller B (2019) Synchronization in Interpersonal Speech. Front. Robot. AI 6:116. doi: 10.3389/frobt.2019.00116*

## 1. INTRODUCTION

It has been shown that during a dyadic human-human interaction, companions will often synchronize their communication style with their partner. This synchrony happens not only on a linguistic level, e.g., syntactic alignment (Gries, 2005; Dale and Spivey, 2006; Branigan et al., 2010), but also occurs across modes, with partners shifting their posture (Scheflen, 1964), facial expression (Blairy et al., 1999), as well as verbal cues (Chartrand and Bargh, 1999)—a topic which has been an area of interest across fields, including psychology (Likowski et al., 2012) and neuroscience (Seibt et al., 2015; Rymarczyk et al., 2018).

An alteration in the rapport between partners is one outcome in relation to synchronous behaviors, and can be described as an interpersonal aspect of a given dyadic exchange in which both

partners are experiencing positivity (Tickle-Degnen and Rosenthal, 1990). From early-research in the field of psychology an increase in rapport was found from interactions in which body posture synchrony had occurred (LaFrance, 1979). However, due to the intrinsic complexity of human behavior, the measurement of interaction synchrony as an indicator of rapport has posed a substantial challenge for researchers (Bernieri et al., 1994). Nevertheless, in social psychological research a non-invasive measurement of interpersonal synchrony, which can be performed without the knowledge of participants, shows great potential for the analysis of human interaction (Bernieri et al., 1994).

Pickering and Garrod presented a mechanistic model of language processing during a dialogue (Pickering and Garrod, 2004). Their interactive alignment account describes how interlocutors automatically synchronize their linguistic representations on multiple levels, from syntax to semantics and phonetics. They argue that alignment on one level also increases alignment on other levels through mechanisms like *routinization* (i.e., the establishment of semi-fixed expressions encoding specific meanings). In recent years, approaches testing mimicry (synchrony) as a tool to enhance rapport have been popular in the field of Human Robot Interaction (HRI) (Riek et al., 2010; Li and Hashimoto, 2011). Valdesolo et al. analyzed the influence of synchrony on individuals who pursue joint goals (Valdesolo et al., 2010). The authors demonstrated that synchrony in body motions can enhance individuals' perceptual sensitivity to the movements of other persons and therefore can increase their success in a following cooperative task which requires the ability to respond appropriately to a partner's movement (Valdesolo et al., 2010). Furthermore, it was discussed that success in achieving common goals is motivated by the enhanced sense of collective spirit, and that synchrony could also predict cooperative ability (Valdesolo et al., 2010).

Previously studies in the area of automatic synchrony detection, have come largely from the vision domain (Michelet et al., 2012), some of which evaluating behaviors such as rate of head nods, and smiling (Sun et al., 2011a; Bilakhia et al., 2013). For this study, we focus on the acoustic signal, as it has been shown that aside from body-language, partners will additionally shift their speech style toward that of their partner (Giles, 1973; Giles et al., 1987).

Although there are similar previous works on this topic (Brdiczka et al., 2005; Burgoon and Hubbard, 2005), we have first proposed an acoustic-based approach to evaluate individual communication styles for the phenomenon of dyadic synchrony across a broad group of cultures (Han et al., 2018). First, we attempt a brute-force conventional approach in which we extract low-level descriptors (LLDs) such as log-energy, and pitch, to measure similarities in the speech turns, resulting in limited success (Han et al., 2018). To explore a state-of-the-art machine learning approach for this task, an autoencoder-based framework is implemented. The framework consists of two autoencoders (AEs), in which each is trained on the speech of one of the communication partners, subject A and B, respectively. On training completion, the data subsets are then switched, and fed to the opposing AE. In choosing this approach, we hypothesize that when a subject is behaving in a more synchronous manner, the reconstruction error of the features from the AE trained on their communication partner should decrease over time. Compared to other state-of-the-art computational approaches for unsupervised learning, e.g., Generative Adversarial Networks, AEs are relatively easy to train and chose hyperparameters for.

In the following section, the related work is summarized both from a sociological and a technical perspective. We then describe our multicultural dataset and the extracted acoustic and DEEP SPECTRUM features used in our research. In section 4, we analyse the behavioral similarities of dyads and explain the experimental settings and discuss about our findings. Afterwards, in section 5, we analyse the linguistic behavior and compare the results to the ones obtained from our acoustic approach, before concluding the paper in section 6.

## 2. RELATED WORK

Synchronous behavior (often referred to as mimicry), can play an important role as a mechanism of *emotional contagion* (Hatfield et al., 1993) i.e., the phenomenon an individual's emotional response to activate a similar emotion in their partner., and is either emotion- or motor-based (Hess and Fischer, 2013). Emotional synchrony is the change in affective states such as *happiness* or *anger*, and the motor-based synchrony would refer to physical changes, e.g., facial expression or position of the hands, although there is also literature indicating that vocal expression is often an unconscious motor act (McGettigan, 2015). Of the two, motor-based synchrony is a more effectively tracked aspect, as there is an object component which can be classified by a human observer, subsequently showing improved accuracy for automatic approaches such as body posture recognition (Hu et al., 2016).

Toward the end of the 1970s, the Facial Action Coding System' (Ekman and Friesen, 1978) based on so-called *facial action units* (FAUs), descriptors of 44 facial activations, was first proposed. Since this time FAUs have been utilized for an array of computational tasks (Kaiser and Wehrle, 1992; Tian et al., 2001; Jaiswal and Valstar, 2016). When combining active FAUs various facial expressions can be constructed, with a strong relationship between typical FAU combinations, e.g., frowning, or smiling, and an individual's affective state (Ekman and Friesen, 2003). These combinations have shown to be independent from culture (Ekman and Friesen, 2003), and can be robustly extracted utilizing state-of-the-art toolkits such as the well-known OPENFACE (Baltrušaitis et al., 2016).

In general partners will likely show synchrony of traits such as gestures and posture, from their partner, nearer to the end of a conversation (Chartrand and Bargh, 1999; Delaherche et al., 2012). Motor-based synchrony can be applied as a persuasive tool during human-to-human exchange, specifically when including the mimicry of the partners spoken opinion (Hess and Fischer, 2013). From both the auditory and visual channels, humans are vulnerable to this behavior (Parrill and Kimbara, 2006). To this end, although there has been evidence of communication

partners synchronizing when they do not agree, there is more prevalent factors of synchrony when partners discuss a common topic of which they hold a similar opinion (Sun et al., 2011a).

From a computational point of view, automatic detection approaches for motor-based synchronous behavior are varied. A time-based regression model which utilized long short-term memory (LSTM) recurrent neural networks (RNNs) was proposed as a prediction method for audio-visual features of *chat* partners (Bilakhia et al., 2013). In Bilakhia et al. (2013), the authors utilized *Mel-frequency cepstral coefficients* (MFCCs) as acoustic features and *facial landmarks* as visual features. They then trained an ensemble of models to predict the features of one chat partner based on the features of their dyadic partner in order to solve the binary classification task of *mimicry* or *non-mimicry*. The model in which the lowest reconstruction error was provided gave the class. In contrast to their work, our approach is unsupervised, i.e., the models are not trained to predict a ground truth occurrence of mimicry.

In general, emotion-based synchrony has not been extensively researched, and has shown to be highly dependent on social context, with individuals not synchronizing at all if they are not in favor with one another (Hess and Fischer, 2014). As well as having a positive outcome on negotiations (Swaab et al., 2011), a similar observation for the favored partner was found within linguistic information (Scissors et al., 2008). In a text-based interaction individuals were found to repeat the style of their partner over time, particularly in scenarios where trust was already established. In this same way, rapport during interactions was found to develop more highly between partners over time when repeating the counterpart's behaviors (LaFrance, 1979).

## 3. DATASET AND FEATURES

To validate the proposed approaches, we use the SEWA corpus of audio-visual interaction in-the-wild (Kossaifi et al., 2019)[1]. A database which has in the past been used as the official benchmark database for the 2017 and 2018 Audio-Visual Emotion Challenges (AVEC) (Ringeval et al., 2017, 2018). Extracting both hand-crafted acoustic features and deep representations of the audio signal on the frame-level of all sessions. We decided to extract both acoustic and DEEP SPECTRUM features, due to their previous performance and proven ability in capturing characteristics of speech (Schuller et al., 2013; Amiriparian et al., 2016, 2018; Eyben, 2016). Both feature sets are different in their nature; COMPARE is a hand-crafted, expert-designed feature set which can cover time-dependent frame-level information for the input signals, and DEEP SPECTRUM is based on the spectrograms of audio signals, focusing mostly on the time-frequency properties of the speech.

### 3.1. The SEWA Video Chat Dataset

The SEWA database includes audio-visual recordings of 197 dyadic conversations (including 201 male and 197 female subjects), from individuals of six differing cultures (Chinese, Hungarian, German, British, Serbian, and Greek). A summary of the SEWA database is given in **Table 1**, including number and total duration of conversation for each culture. An example conversation is shown in **Figure 1** and during such conversations, subjects discuss with each other their view of a 90 s advertisement of a (water) tap that they have just been shown via the web platform.

**TABLE 1 |** SEWA corpus: Quantity of conversations and subjects, as well as total duration given in minutes for each culture.

| Index | Culture | # Conversations | # Subjects | Total duration |
|---|---|---|---|---|
| C1 | Chinese | 35 | 70 | 101 |
| C2 | Hungarian | 33 | 66 | 67 |
| C3 | German | 32 | 64 | 89 |
| C4 | British | 33 | 66 | 94 |
| C5 | Serbian | 36 | 72 | 98 |
| C6 | Greek | 28 | 56 | 81 |
| Sum | | 197 | 394 | 530 |

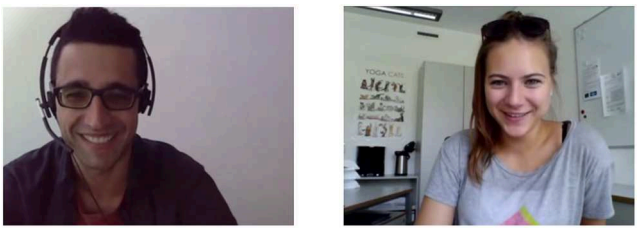


**FIGURE 1 |** Screenshots taken from sample video chats in the SEWA corpus (German culture).

The subjects were "in-the-wild" and using a personal computer, with recordings captured from either their home or office. The chat partners were already acquainted with one another before the chat (either family, friends, or colleagues), and included varied gender pairings (female-male, female-female, male-male), which were balanced across all sessions. Subject were aged between 18 and 60, and communication was held in the native language of the partners, with no specified limitation on what to discuss about the advertisement. From post analysis, it was found that the conversations in the SEWA Dataset contain a variety of emotional states, as well as instances of both agreement/disagreement, and additionally positive/negative rapport (Ringeval et al., 2017, 2018; Kossaifi et al., 2019).

### 3.2. Acoustic Features

The COMPARE feature set of acoustic features (Eyben, 2016) is used for our first approach. For each audio recording, acoustic low-level descriptors are extracted using the OPENSMILE toolkit (Eyben et al., 2013) at a step size of 10 ms. COMPARE LLDs are extracted at frame-level. *Functionals* defined in the feature set are not applied in this work, as the time-dependent frame-level information is of most interest. Extracted with a window size of 20 to 60 ms length, there are 65 LLDs in the COMPARE feature set and these have been summarized in **Table 2**. Feature vectors

[1]https://sewaproject.eu/

**TABLE 2 |** Interspeech 2013 Computational Paralinguistics Challenge feature set.

| **4 energy related LLD** | **Group** |
|---|---|
| Loudness | Prosodic |
| Modulation loudness | Prosodic |
| RMS energy, zero-crossing rate | Prosodic |
| **55 spectral related LLD** | **Group** |
| RASTA auditory bands 1–26 | Spectral |
| MFCC 1–14 | Cepstral |
| Spectral energy 250-650 Hz, 1–4 kHz | Spectral |
| Spectral roll-off pt. .25, .50, .75, .90 | Spectral |
| Spectral flux, entropy, variance | Spectral |
| Spectral skewness and kurtosis | Spectral |
| Spectral slope | Spectral |
| Spectral harmonicity | Spectral |
| Spectral sharpness (auditory) | Spectral |
| Spectral centroid (linear) | Spectral |
| **6 voicing related LLD** | **Group** |
| $F_0$ via SHS | Prosodic |
| Probability of voicing | Voice quality |
| Jitter (local and delta) | Voice quality |
| Shimmer | Voice quality |
| Log harmonics-to-noise ratio | Voice quality |

*An overview of the 65 acoustic low-level descriptors (LLDs). SHS, Sub-Harmonic Summation.*

of size 130 for each 10 ms step are given by calculating the first order derivative (deltas).

## 3.3. Deep Spectrum Features

In addition to the acoustic features (cf. section 3.2), we apply the feature extraction DEEP SPECTRUM toolkit[2] to extract deep representations from the audio signals using pre-trained convolutional neural networks (CNNs) (Amiriparian et al., 2017c). First, audio signals are transformed into Mel-spectrogram plots using a Hanning window of width 500 ms and an overlap 10 ms. From these, 128 Mel-frequency bands are then computed. Afterwards, the generated spectrograms are forwarded through VGG16 (Simonyan and Zisserman, 2014), a pre-trained CNN, and the activations of the penultimate fully connected layer (*fc7*) of the network are extracted, resulting in a 4,096 dimensional DEEP SPECTRUM feature vector. These features can be considered as being a high-level representation of the Mel-spectrograms (Amiriparian et al., 2017c), and have shown to be highly effective in various speech and audio analysis tasks (Amiriparian et al., 2017a,c, 2018, 2019; Baird et al., 2017; Ringeval et al., 2018).

# 4. BEHAVIOR SIMILARITY TENDENCY ANALYSIS WITH AUTOENCODER

In order to investigate the temporal-based patterns, as well as interpersonal sentiment which may occur in speech, we first need to get machine readable representations from the speech signals of each individual (cf. section 3.2 and 3.3) and then use these features for our machine learning experiments (cf.section 4.1). Based on the experimental results (cf. section 4.2), we then analyse the behavior similarities in various cultures.

To minimize the variance between recording environments the acoustic features (130 frame level) are first standardized (zero mean and unit standard deviation) across the same recordings. We have neither standardized nor normalized the DEEP SPECTRUM features, since we found during our preliminary evaluation that this negatively impacts autoencoder performance. Before beginning to train the AE (cf.section 4.1), the feature sequences are first segmented based on the transcriptions which are also included in the SEWA database. The feature sequences of each recording are then split in two sub-sequences, with each having the features of only one of the subjects.

We then use a machine learning framework based on autoencoders for investigating the effect of synchrony found in the feature sequences. Autoencoders are a special type of neural network architecture trained in an unsupervised manner to find a compact, information rich representation of the input data from which this input can be reconstructed (Vincent et al., 2008). Further, the reconstruction error that is made by a trained autoencoder on unseen test data can give an indication on how similar this data is to the training domain: In the context of audio analysis, this has for example been used for automatic acoustic novelty detection (Marchi et al., 2015), the intuition being that audio events that are foreign to the training data will be harder to accurately reconstruct for the autoencoder. For our experiments, the training domain of each autoencoder are the feature sequences of one speaker while the sequences of the speaker's partner are used for evaluation. In our approach, AEs use the features extracted at each frame as independent instances, without considering the evolution of features over time. For each individual dyadic interaction in the dataset, we proceed as follows: Features of one subject are applied frame-wise to train the first AE, with the features of the other used frame-wise for testing. Although training the AEs and reconstructing the features using each frame as an independent instance, we preserve the order of the test frames in order to generate the reconstructed sequence of features. Then, the root-mean-squared errors (RMSEs) are calculated between the reconstructed and actual features as a means of evaluating the extent to which the RMSE varies over time. For each conversation, we end with two AEs trained on the two subjects involved, with two one-dimensional RMSE sequences, whose slopes can be measured by computing their first derivatives and later averaged for further analysis.

## 4.1. Experimental Settings

For the AEs, we made use of a common bottleneck architecture: The input layer of the encoder and the output layer of the decoder match the size of the feature vectors whilst the size of neurons on the hidden layers is halved (doubled) for each layer in the encoder (decoder). As shown in **Figure 2**, the AE framework that has been constructed consists of a 3-layer encoder with a 3-layer decoder. During the initial experiments, nodes in each layer were selected as follows: 130–64–32–12–32–64–130, with

[2]https://github.com/DeepSpectrum/DeepSpectrum

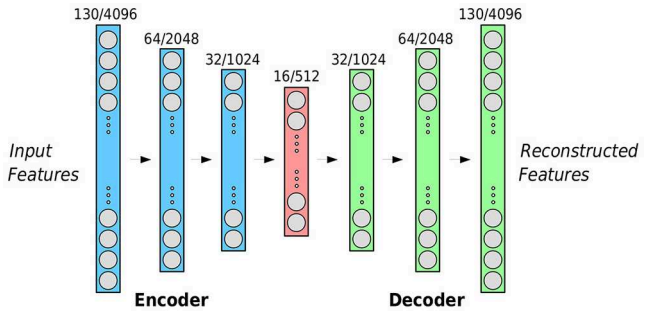



**FIGURE 2 |** The autoencoder framework implemented for our study. The error between the input features (left) and the reconstructed features (right) is minimized for each subject using the RMSE. The given number of neurons in each layer (indicated above the neurons) refers to the COMPARE/DEEP SPECTRUM features, respectively.

**TABLE 3 |** Average slope of RMSE sequences of all subjects and the Pearson correlation coefficients of pairs in each culture (C1: Chinese, C2: Hungarian, C3: German, C4: British, C5: Serbian, and C6: Greek).

| Feature set | C1 | C2 | C3 | C4 | C5 | C6 |
|---|---|---|---|---|---|---|
| **Acoustic features** | | | | | | |
| *average slope* | −0.07 | −0.11 | −0.10 | −0.07 | −0.08 | −0.12 |
| *pcc of pairs* | −0.03 | 0.34 | 0.15 | 0.39 | 0.39 | −0.26 |
| **DEEP SPECTRUM features** | | | | | | |
| *average slope* | −0.03 | −0.05 | −0.03 | −0.02 | −0.05 | −0.07 |
| *pcc of pairs* | 0.03 | 0.16 | 0.18 | 0.09 | 0.13 | −0.15 |

*The autoencoders were trained on both acoustic and DEEP SPECTRUM features. For all cultures the average slope shallower when using DEEP SPECTRUM features.*

the dimensions of the output matching that of the input low-level audio descriptors. For the DEEP SPECTRUM features, we use a larger number of neurons on each layer: 4,096–2,048–1,024–512–1,024–2,048–4,096. We train all AEs with a batch size of 256 for 512 epochs minimizing the mean squared reconstruction error using the Adagrad (Duchi et al., 2011) optimizer with a learning rate of 0.01.

When the temporal reconstruction errors had been generated for each of the test subjects, the sequence is then utilized for a linear regression task, assuming that the learnt slope will indicate a behavior pattern change. In other words, when there is a negative slope, this may indicate that the dyadic partners are becoming more similar. Counter to this if there is a positive slope, it would indicate that the partners are less synchronized. As well as this, we make the additional assumption that the overall amplitude of the slope will denote the level of synchrony as well.

Our approach for using the slope for synchrony analysis between dyads is mainly motivated by the works introduced in Sun et al. (2011b), Delaherche et al. (2012), and Bilakhia et al. (2013). In Delaherche et al. (2012), the authors state that the interactive alignment/synchrony can be observed in conversation from a variety of features such as intonation, intensity, and rhythm in speech. In addition, in Bilakhia et al. (2013), the authors applied MSE to measure the reconstruction error of an unseen example with a trained model to detect non-verbal vocal mimicry vs. non-mimicry categories. In particular, 6 MFCCs were adopted as audio features instead of pitch or energy, whilst in the present work, more hand-crafted features, as well as deep representations, are investigated. Moreover, in Sun et al. (2011b), the results have shown that a long-term increasing correlation is consistently obtained between two speakers in a discussion. Thus, though the term "slope" was not well-supported in any of previous work, these previous findings motivate this work to adopt the RMSE slope overall interaction to indicate progressive synchronization. Furthermore, in **Table 3**, it has been demonstrated that the slope tendencies have a negative correlation with the answer to the question if an individual feels of holding the same opinion with the partner, demonstrating that the detected synchronization tendency has a high correlation with their self-reported labels.

## 4.2. Results and Discussion

The first culture from the SEWA dataset; C1 (Chinese) will be where we begin our discussion. This culture consists of 35 sessions, and the average RMSE sequence slope for all 70 subjects is −0.07, and −0.03 when using acoustic and DEEP SPECTRUM features, respectively. Using both feature sets, which differ in their nature, we show that very low average RMSE can be achieved for the Chinese culture. This finding indicates a relatively high synchrony between Chinese dyadic partners.

From the analysis shown in **Figure 3** it can be seen that most subject slopes for both feature sets (54 /70 for the acoustic features and 47 /70 for the DEEP SPECTRUM features) are negative, with less being positive. With our previous assumption in mind, these results indicate that the acoustic LLD features and the DEEP SPECTRUM features of these subjects have a smaller reconstruction error over time. As the AE is trained with the opposing subject from the same session a smaller reconstruction error should indicate higher synchrony between the communication partners. We also see a similar trend across other cultures in the dataset, however the ratios for negative / positive slope vary across cultures. **Figures 4**, **5** show the slope of RMSE for all subjects and all cultures obtained from both feature sets.

With these results in mind, the average slopes *s* were calculated for all cultures, as well as the Pearson correlation coefficients (PCCs). This was made with the intention of investigating cultural-based variation across the spontaneous in the wild conversations. For this analysis, results are summarized in **Table 3**. As mentioned a negative slope indicates a more synchronous speech-based relationship. The *average slope* is computed to demonstrate the overall tendency throughout all subjects in one specific culture, whilst the *pcc of pairs* is applied to indicate the tendency between two conversation partners given that specific culture.

From the correlation analysis shown in **Table 3**, it can be noticed that generally when observed as group pairings

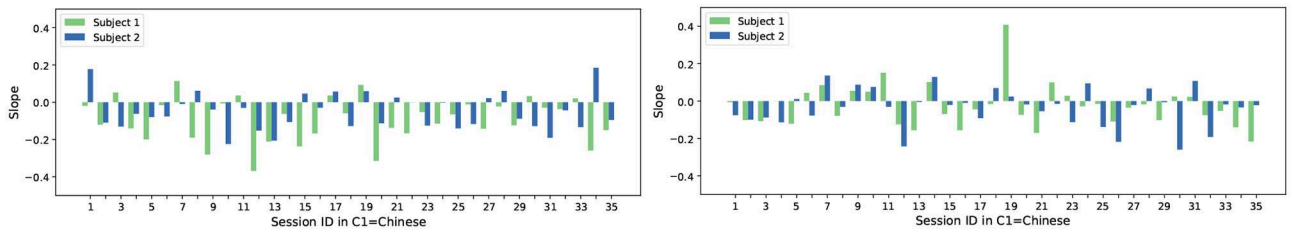



**FIGURE 3 |** Slope of RMSE sequences of 70 Chinese subjects from 35 recordings. In each recording, there are two subjects as denoted with green and blue bars. The diagrams (from left to right) are generated based on the acoustic and DEEP SPECTRUM features, respectively.

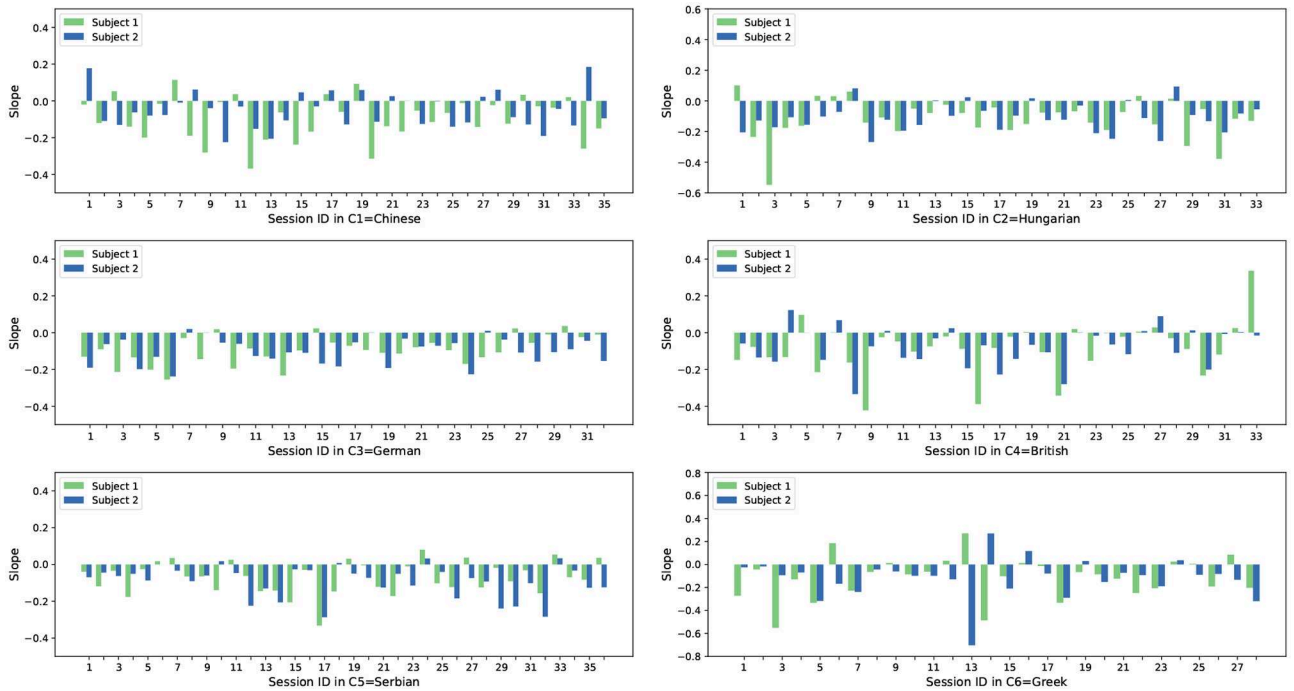



**FIGURE 4 |** Slope of RMSE sequences of paired subjects from all recordings in all six cultures. The results are calculated based on the acoustic features.

A/B, individuals across the six cultures show a tendency to synchronize. Given that *s* for each culture is always negative. The Greek culture (C6) shows the largest slope, i.e., lower synchrony between the Greek dyads, and the smallest slope is observed for both Chinese and British cultures.

As well as this, when looking only at the PCC, we can see an alternative culture variance. In the case of PCC, positive values indicate that the subjects of a culture converge to a similar place, either both behaving in synchrony or out of synchrony with one another. Conversely, a negative PCC would indicate that one subject is dominating the other. No correlation is seen in the C1 (Chinese) pairs for example, with a PCC of −0.03 and 0.03 when using acoustic and DEEP SPECTRUM features which is close to 0. On the other side, a linear correlation is shown as either positive for the Hungarian (C2), German (C3), British (C4), and Serbian (C5) or negative for the Greek (C6) culture. Although out of the scope of our study, it would be of benefit to verify these findings based on literature across other fields, such as the anthropological linguistics domain and the field of conversation analysis (Stivers et al., 2009). We should also note that variances such as educational background, occupation, and health status of the individuals in the SEWA dataset may have some effect on the result, however, although the dataset providers did implement a control of aspects such as age and gender, variation between complex characteristics such as these would be difficult to avoid.

## 5. LINGUISTIC BEHAVIOR ANALYSIS AND SIMILARITY PATTERNS

Motor-based synchrony, e.g., raising an eyebrow, can be detected from visual mid-level features such as Facial Action Units (FAUs) (Surakka and Hietanen, 1998). Nonetheless, the detection of similarity in speech from raw features is challenging due

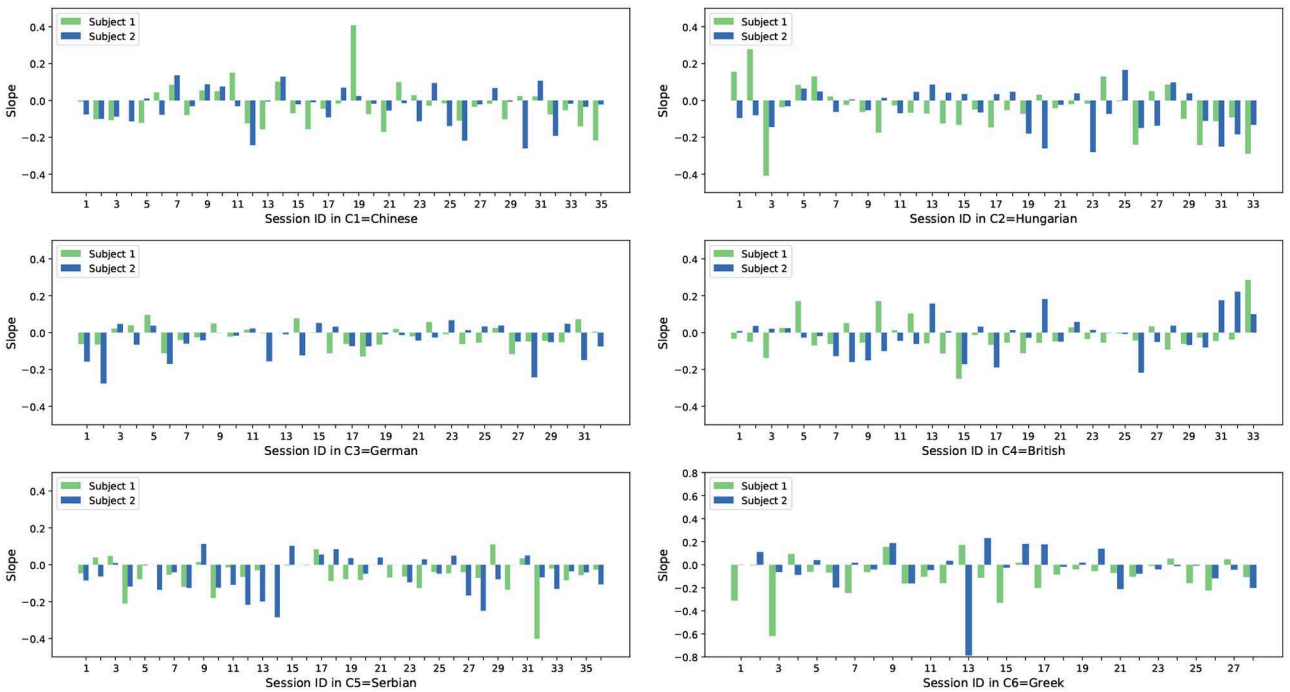



**FIGURE 5 |** Slope of RMSE sequences of paired subjects from all recordings in all six cultures. The results are calculated based on the DEEP SPECTRUM features.

to the variability of speech descriptors. To name a few, these descriptors are sensitive to the environment and the voice of the subject, which is influenced by factors such as age and gender, amongst others.

Besides the acoustic similarities, we should also investigate the behavioral synchronization shown in video chats from other modalities, including linguistic information. In this regard, rather than the conventional bag-of-words (BoW) approach, which represents a text as a sparse histogram vector, word embeddings are the current state-of-the-art (Kusner et al., 2015; Liu et al., 2015; Amiriparian et al., 2017b; Chung and Glass, 2018). With this technique, the sparse histogram vectors, with a dimensionality higher than $\mathbb{R}^{1\times5000}$, are transformed into a lower dimensionality vector, typically $\mathbb{R}^{1\times300}$, where each component in the vector space represents a concept. As a relevant property of word embeddings, the distance between this concept and words with similar meanings is lower than the distance between this concept and words with completely different meanings. The architecture of neural networks for word embeddings usually includes a single layer, which converts the BoW into the embedding vector. Currently, *word2vec*, introduced by Mikolov et al. (2013), is a popular technique to generate word embeddings, as it is trained on large text corpora, such as Wikipedia. This technique employs a specialized objective function, called"negative sampling." One of the benefits of using such word embedding technique is that the representations generated from the words quantitatively capture several properties of the object they describe (Mikolov et al., 2013).

We base our analysis on the manual transcriptions of the video chats from the six different cultures included in the SEWA database (cf. section 3 for details). Word embeddings are extracted using pre-trained *word2vec* models available on the internet. While a word embedding model for the British culture trained on a Google News corpus is employed[3], word embedding models for the Hungarian and German cultures trained on Wikipedia dumps are used[4]. For the other cultures, suitable word embedding pre-trained models are not currently available and, as a consequence, we exclude these cultures from our experiments with the *word2vec* approach. Furthermore, training our own word embedding models on the transcriptions of the SEWA database is discarded due to limitations on the available data. Word embedding models require large amounts of data to be trained, usually requiring more than a million running words.

In order to analyse the linguistic synchronization as the interaction progresses, we decide to split the chat sessions in two halves, the first and second half of each conversation. The measurement of similarities on a smaller scale, e.g., on utterance or speaker turn level, is not possible, as some particular speaker turns are quite long (more than 30 s). For every half of the interaction *word2vec* embeddings are extracted from both the speaker and their partner, and the cosine similarity between the word embeddings is computed. In addition to word embeddings, a simple evaluation of word usage is also made by counting how often the same words were used by the two subjects in each segment and normalizing the result by the number of words per segment. The averaged similarities of both scenarios in both halves of the interactions for all participants

[3]https://github.com/3Top/word2vec-api
[4]https://github.com/Kyubyong/wordvectors

**TABLE 4 |** Evaluation of linguistic similarities between dyadic companions in the two halves of the video chat.

| Culture | Word usage similarity | | *word2vec* similarity | |
|---|---|---|---|---|
| | 1st half | 2nd half | 1st half | 2nd half |
| C1 (Chinese) | 0.710 | 0.880 | — | — |
| C2 (Hungarian) | 0.738 | 0.902 | 0.809 | 0.794 |
| C3 (German) | 1.063 | 1.128 | 0.301 | 0.327 |
| C4 (British) | 1.714 | 1.787 | 0.364 | 0.383 |
| C5 (Serbian) | 1.241 | 1.353 | — | — |
| C6 (Greek) | 0.849 | 1.125 | — | — |

*The linguistic information is analyzed using two different approaches: by computing word usage and by extracting word2vec embeddings from the transcripts included in the SEWA database.*

belonging to the same culture are calculated and summarized in **Table 4**.

The results reported in **Table 4** show that for all cultures the linguistic similarity increases during the video chat in regards to the word usage. For *word2vec* embeddings the increase is very subtle and in particular, for the Hungarian culture, we observe that the similarity slightly decreases. The very weak or even non-existent linguistic synchronization we measured with the *word2vec* approach could be explained by the nature of the rather complex features. It seems possible that a synchronization on such a high linguistic level takes even more time than the acoustic synchronization or the linguistic synchronization on the word level and could therefore not be measured in short conversations. This result leads us to assume that rapport and synchrony in the linguistic domain is manifested in the direct synchrony of terminology, rather than in synchrony of concepts and topics.

The differences of linguistic similarity across cultures is quite noticeable as the values of word usage similarity in the first half of the conversations range from 0.710 in the Chinese culture up to 1.714 in the British culture. In the *word2vec* approach the similarity values for the first half of the conversations range from 0.301 in the German culture up to 0.809 in the Hungarian culture. Reasons for this, as for the different changes of the similarity through the conversations, might lie in the respective languages of the different cultures or culture-specific behaviors during conversation.

## 6. CONCLUSION AND OUTLOOK

In this work, we have demonstrated that, an autoencoder-based framework has great potential to recognize the spontaneous and unconscious synchronization which occur during social interactions. We can see this evidence through the observation of the reconstruction error, when using the acoustic and DEEP SPECTRUM features extracted from the speech of each dyadic companion.

From this work, we have also explored culturally dependent synchronization of vocal behavior in dyadic conversations. In section 4, we have analyzed the behavior similarities and ability of interpersonal chats to synchronize. It was found that both feature sets are suitable for this task. Most subjects slopes are negative when observing the feature sets (54 /70 for the acoustic features and 47 /70 for the DEEP SPECTRUM features). From additional correlation analysis, it was found that individuals do tend to synchronize, however from this analysis, the cultural differences were more noticeable, e.g., C6 (Greek) and C1 (Chinese) show quite opposing average slopes (−0.07 and −0.03, respectively with DEEP SPECTRUM features).

Furthermore, the results provided in **Table 4** demonstrated that for all six cultures the linguistic similarity increases during the video chat.

Future work will focus on utilizing further unsupervised representation learning techniques, such as unsupervised feature learning with deep neural networks using the AUDEEP toolkit (Amiriparian et al., 2017b; Freitag et al., 2018), and feature quantization methods, such as *bag-of-audio-words* (Schmitt et al., 2016). Moreover, we are planing to exploit the linguistic domain through state-of-the-art *word2vec* embeddings (Mikolov et al., 2013). Given the findings in relation to cultures from the utilized dataset, it would also be of value to further explore this, possibly through a deeper analysis of non-verbal synchrony and the known occurrence of this during dyadic interactions (Tschacher et al., 2014). It is also of big interest to analyse the amount of alignment between speakers across different dyads. Finally, in addition to the slope of the reconstruction errors, we want to explore further evaluation strategies to measure the degree of synchrony between subjects (Delaherche et al., 2012).

## DATA AVAILABILITY STATEMENT

The dataset analyzed for this study, SEWA, is a public dataset and can be found under the following link: https://db.sewaproject.eu/.

## ETHICS STATEMENT

For recording the SEWA dataset the local ethics board, the Imperial College Research Ethics Committee (ICREC), has approved the recording of the audio-visual database and the study of audio-visual behavior in the collected data. All subjects analyzed for the study described in this article have given their written informed consent to participate prior to recording. The two participants shown in **Figure 1** have given their written informed consent to publish excerpts from their recordings in academic documents, articles, and presentations.

## AUTHOR CONTRIBUTIONS

SA, JH, and MS conceptualized the study and ran the machine learning experiments. AB, AM-R, MM, and BS did literature analysis, manuscript preparation and editing. MG helped with running the experiments and testing the codes. All authors revised, developed, read, and approved the final manuscript.

## FUNDING

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 688835 (RIA DE-ENIGMA) and No. 826506 (sustAGE), and from the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B).

## REFERENCES

Amiriparian, S., Cummins, N., Gerczuk, M., Pugachevskiy, S., Ottl, S., and Schuller, B. (2019). "Are you playing a shooter again?!" deep representation learning for audio-based video game genre recognition. *IEEE Trans. Games* 11 doi: 10.1109/TG.2019.2894532

Amiriparian, S., Cummins, N., Ottl, S., Gerczuk, M., and Schuller, B. (2017a). "Sentiment analysis using image-based deep spectrum features," in *Proceedings of the 7th Biannual Conference on Affective Computing and Intelligent Interaction (ACII 2017)* (San Antonio, TX), 26–29.

Amiriparian, S., Freitag, M., Cummins, N., and Schuller, B. (2017b). "Sequence to sequence autoencoders for unsupervised representation learning from audio," in *Proceedings of the DCASE 2017 Workshop* (Munich), 17–21.

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., et al. (2017c). "Snore sound classification using image-based deep spectrum features," in *Proceedings of INTERSPEECH 18th Annual Conference of the International Speech Communication Association* (Stockholm: ISCA), 3512–3516.

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Pugachevskiy, S., and Schuller, B. (2018). "Bag-of-deep-features: Noise-robust deep feature representations for audio analysis," in *Proceedings of the 31st International Joint Conference on Neural Networks (IJCNN)* (Rio de Janeiro: IEEE), 2419–2425.

Amiriparian, S., Pohjalainen, J., Marchi, E., Pugachevskiy, S., and Schuller, B. (2016). "Is deception emotional? An emotion-driven predictive approach," in *Proceedings INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association* (San Francisco, CA: ISCA), 2011–2015.

Baird, A., Amiriparian, S., Cummins, N., Alcorn, A. M., Batliner, A., Pugachevskiy, S., et al. (2017). "Automatic classification of autistic child vocalisations: A novel database and results," in *Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association* (Stockholm: ISCA), 849–853.

Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). "OpenFace: an open source facial behavior analysis toolkit," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)* (Lake Placid, NY), 1–10.

Bernieri, F. J., Davis, J. M., Rosenthal, R., and Knee, C. R. (1994). Interactional synchrony and rapport: measuring synchrony in displays devoid of sound and facial affect. *Pers. Soc. Psychol. Bull.* 20, 303–311. doi: 10.1177/0146167294203008

Bilakhia, S., Petridis, S., and Pantic, M. (2013). "Audiovisual detection of behavioural mimicry," in *Proceedings Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)* (Geneva), 123–128.

Blairy, S., Herrera, P., and Hess, U. (1999). Mimicry and the judgement of emotional facial expressions. *J. Nonverbal Behav.* 23, 5–41. doi: 10.1023/A:1021370825283

Branigan, H. P., Pickering, M. J., Pearson, J., and McLean, J. F. (2010). Linguistic alignment between people and computers. *J. Pragmatics* 42, 2355–2368. doi: 10.1016/j.pragma.2009.12.012

Brdiczka, O., Maisonnasse, J., and Reignier, P. (2005). "Automatic detection of interaction groups," in *Proceedings of the 7th International Conference on Multimodal Interfaces*, ICMI '05 (Trento), 32–36.

Burgoon, J. K., and Hubbard, A. E. (2005). "Cross-cultural and intercultural applications of expectancy violations theory and interaction adaptation theory," in *Theorizing About Intercultural Communication*, ed W. B. Gudykunst (Thousand Oaks, CA: Sage) 149–171.

Chartrand, T. L., and Bargh, J. A. (1999). The chameleon effect: the perception–behavior link and social interaction. *J. Pers. Soc. Psychol.* 76, 893–910. doi: 10.1037//0022-3514.76.6.893

Chung, Y.-A. and Glass, J. (2018). Speech2vec: a sequence-to-sequence framework for learning word embeddings from speech. *arXiv preprint* arXiv:1803.08976.

Dale, R., and Spivey, M. J. (2006). Unraveling the dyad: using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Lang. Learn.* 56, 391–430. doi: 10.1111/j.1467-9922.2006.00372.x

Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., and Cohen, D. (2012). Interpersonal synchrony: a survey of evaluation methods across disciplines. *IEEE Trans. Affect. Comput.* 3, 349–365. doi: 10.1109/T-AFFC.2012.12

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* 12, 2121–2159.

Ekman, P., and Friesen, W. V. (1978). *Facial Action Coding System.* Consulting Psychologists Press. Available online at: https://books.google.fr/books?id=08l6wgEACAAJ

Ekman, P., and Friesen, W. V. (2003). *Unmasking the Face: A Guide to Recognizing Emotions From Facial Clues, 1 Edn*. Los Altos, CA: Ishk.

Eyben, F. (2016). *Real-Time Speech and Music Classification by Large Audio Feature Space Extraction, 1 Edn*. Basel: Springer.

Eyben, F., Weninger, F., Groß, F., and Schuller, B. (2013). "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proceedings the 21st ACM International Conference on Multimedia (ACMM)* (Barcelona), 835–838.

Freitag, M., Amiriparian, S., Pugachevskiy, S., Cummins, N., and Schuller, B. (2018). audeep: Unsupervised learning of representations from audio with deep recurrent neural networks. *J. Mach. Learn. Res.* 18, 1–5.

Giles, H. (1973). Accent mobility: a model and some data. *Anthropol. Linguist.* 15, 87–105.

Giles, H., Mulac, A., Bradac, J. J., and Johnson, P. (1987). Speech accommodation theory: the first decade and beyond. *Ann. Int. Commun. Assoc.* 10, 13–48. doi: 10.1080/23808985.1987.11678638

Gries, S. T. (2005). Syntactic priming: a corpus-based approach. *J. Psycholinguist. Res.* 34, 365–399. doi: 10.1007/s10936-005-6139-3

Han, J., Schmitt, M., and Schuller, B. W. (2018). "You sound like your counterpart: Interpersonal speech analysis," in *Proceedings of Speech and Computer - 20th International Conference, SPECOM* (Leipzig), 188–197.

Hatfield, E., Cacioppo, J. T., and Rapson, R. L. (1993). Emotional contagion. *Curr. Dir. Psychol. Sci.* 2, 96–100. doi: 10.1017/CBO9781139174138

Hess, U., and Fischer, A. (2013). Emotional mimicry as social regulation. *Pers. Soc. Psychol. Rev.* 17, 142–157. doi: 10.1177/1088868312472607

Hess, U., and Fischer, A. (2014). Emotional mimicry: why and when we mimic emotions. *Soc. Pers. Psychol. Compass* 8, 45–57. doi: 10.1111/spc3.12083

Hu, F., Wang, L., Wang, S., Liu, X., and He, G. (2016). A human body posture recognition algorithm based on bp neural network for wireless body area networks. *China Commun.* 13, 198–208. doi: 10.1109/CC.2016.7563723

Jaiswal, S., and Valstar, M. (2016). 'Deep learning the dynamic appearance and shape of facial action units," in *Proceedings of 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)* (New York, NY: IEEE), 1–8.

Kaiser, S., and Wehrle, T. (1992). Automated coding of facial behavior in human-computer interactions with facs. *J. Nonverbal Behav.* 16, 67–84. doi: 10.1007/BF00990323

Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., et al. (2019). SEWA DB: a rich database for audio-visual emotion and sentiment research in the wild. *CoRR*, abs/1901.02839.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). "From word embeddings to document distances," in *International Conference on Machine Learning* (Lille), 957–966.

LaFrance, M. (1979). Nonverbal synchrony and rapport: Analysis by the cross-lag panel technique. *Soc. Psychol. Q.* 42, 66–70. doi: 10.2307/3033875

Li, Y., and Hashimoto, M. (2011). "Effect of emotional synchronization using facial expression recognition in human-robot communication," in *Proceedings of 2011 IEEE International Conference on Robotics and Biomimetics (ROBIO)* (Phuket), 2872–2877.

Likowski, K., Muehlberger, A., Gerdes, A., Wieser, M., Pauli, P., and Weyers, P. (2012). Facial mimicry and the mirror neuron system: simultaneous acquisition of facial electromyography and functional magnetic resonance imaging. *Front. Hum. Neurosci.* 6:214. doi: 10.3389/fnhum.2012.00214

Liu, Y., Liu, Z., Chua, T.-S., and Sun, M. (2015). "Topical word embeddings," in *Proceedings of Conference on Artificial Intelligence (AAAI)*.

Marchi, E., Vesperini, F., Eyben, F., Squartini, S., and Schuller, B. (2015). "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional lstm neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Brisbane: IEEE), 1996–2000.

McGettigan, C. (2015). The social life of voices: studying the neural bases for the expression and perception of the self and others during spoken communication. *Front. Hum. Neurosci.* 9:129. doi: 10.3389/fnhum.2015.00129

Michelet, S., Karp, K., Delaherche, E., Achard, C., and Chetouani, M. (2012). "Automatic imitation assessment in interaction," in *Human Behavior Understanding*, eds A. A. Salah, J. Ruiz-del Solar, Ç. Meriçli, and P.-Y. Oudeyer (Berlin; Heidelberg: Springer), 161–173.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality," in *Proceedings of NIPS* (Lake Tahoe, NV), 3111–3119.

Parrill, F., and Kimbara, I. (2006). Seeing and hearing double: the influence of mimicry in speech and gesture on observers. *J. Nonverbal Behav.* 30:157. doi: 10.1007/s10919-006-0014-2

Pickering, M. J., and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* 27, 169–190. doi: 10.1017/S0140525X04000056

Riek, L. D., Paul, P. C., and Robinson, P. (2010). When my robot smiles at me:enabling human-robot rapport via real-time head gesture mimicry. *J. Multimodal User Interfaces* 3, 99–108. doi: 10.1007/s12193-009-0028-2

Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Kaya, H., Schmitt, M., et al. (2018). "Avec 2018 workshop and challenge: bipolar disorder and cross-cultural affect recognition," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop (AVEC)* (Seoul: ACM), 3–13.

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., et al. (2017). "AVEC 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of n Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop (AVEC)* (Mountain View, CA), 3–9.

Rymarczyk, K., Zurawski, L., Jankowiak-Siuda, K., and Szatkowska, I. (2018). Neural correlates of facial mimicry: Simultaneous measurements of emg and bold responses during perception of dynamic compared to static facial expressions. *Front. Psychol.* 9:52. doi: 10.3389/fpsyg.2018.00052

Scheflen, A. E. (1964). The significance of posture in communication systems. *Psychiatry* 27, 316–331. doi: 10.1080/00332747.1964.11023403

Schmitt, M., Ringeval, F., and Schuller, B. (2016). "At the border of acoustics and linguistics: Bag-of-Audio-Words for the recognition of emotions in speech," in *Proceedings INTERSPEECH 2017, 17th Annual Conference of the International Speech Communication Association* (San Francisco, CA), 495–499.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., et al. (2013). "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings of INTERSPEECH* (Lyon), 148–152.

Scissors, L. E., Gill, A. J., and Gergle, D. (2008). "Linguistic mimicry and trust in text-based cmc," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (San Diego, CA), 277–280.

Seibt, B., Mühlberger, A., Likowski, K., and Weyers, P. (2015). Facial mimicry in its social setting. *Front. Psychol.* 6:1122. doi: 10.3389/fpsyg.2015.01122

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., et al. (2009). Universals and cultural variation in turn-taking in conversation. *Proc. Natl. Acad. Sci. U.S.A.* 106, 10587–10592. doi: 10.1073/pnas.0903616106

Sun, X., Nijholt, A., Truong, K. P., and Pantic, M. (2011a). "Automatic visual mimicry expression analysis in interpersonal interaction," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Colorado Springs, CO), 40–46.

Sun, X., Truong, K. P., Pantic, M., and Nijholt, A. (2011b). "Towards visual and vocal mimicry recognition in human-human interactions," in *2011 IEEE International Conference on Systems, Man, and Cybernetics* (Anchorage, AK: IEEE), 367–373.

Surakka, V., and Hietanen, J. K. (1998). Facial and emotional reactions to duchenne and non-duchenne smiles. *Int. J. Psychophysiol.* 29, 23–33. doi: 10.1016/S0167-8760(97)00088-3

Swaab, R. I., Maddux, W. W., and Sinaceur, M. (2011). Early words that work: when and how virtual linguistic mimicry facilitates negotiation outcomes. *J. Exp. Soc. Psychol.* 47, 616–621. doi: 10.1016/j.jesp.2011.01.005

Tian, Y.-I., Kanade, T., and Cohn, J. F. (2001). Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Anal. Mach. Intellig.* 23, 97–115. doi: 10.1109/34.908962

Tickle-Degnen, L., and Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychol. Inquiry* 1, 285–293. doi: 10.1207/s15327965pli0104_1

Tschacher, W., Rees, G. M., and Ramseyer, F. (2014). Nonverbal synchrony and affect in dyadic interactions. *Front. Psychol.* 5:1323. doi: 10.3389/fpsyg.2014.01323

Valdesolo, P., Ouyang, J., and DeSteno, D. (2010). The rhythm of joint action: synchrony promotes cooperative ability. *J. Exp. Soc. Psychol.* 46, 693–695. doi: 10.1016/j.jesp.2010.03.004

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine Learning* (Helsinki: ACM), 1096–1103.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 31 July 2019
doi: 10.3389/frobt.2019.00069

# Adaptation and Transfer of Robot Motion Policies for Close Proximity Human-Robot Interaction

*Khoi Hoang Dinh*, Ozgur S. Oguz, Mariam Elsayed and Dirk Wollherr*

*Chair of Automatic Control Engineering, Department of Electrical and Computer Engineering, Technical University of Munich, Munich, Germany*

In the context of human-robot collaboration in close proximity, safety and comfort are the two important aspects to achieve joint tasks efficiently. For safety, the robot must be able to avoid dynamic obstacles such as a human arm with high reliability. For comfort, the trajectories and avoidance behavior of the robot need to be predictable to the humans. Moreover, these two aspects might be different from person to person or from one task to another. This work presents a framework to generate predictable motions with dynamic obstacle avoidance for the robot interacting with the human by using policy improvement method. The trajectories are generated using Dynamic Motion Primitives with an additional potential field term that penalizes trajectories that may lead to collisions with obstacles. Furthermore, human movements are predicted using a data-driven approach for proactive avoidance. A cost function is defined which measures different aspects that affect the comfort and predictability of human co-workers (e.g., human response time, joint jerk). This cost function is then minimized during human-robot interaction by the means of policy improvement through black-box optimization to generate robot trajectories that adapt to human preferences and avoid obstacles. User studies are performed to evaluate the trust and comfort of human co-workers when working with the robot. In addition, the studies are also extended to various scenarios and different users to analyze the task transferability. This improves the learning performance when switching to a new task or the robot has to adapt to a different co-worker.

**Keywords: human robot interaction, motion generation, black-box optimization, dynamic motion primitives, policy improvement, close proximity**

OPEN ACCESS

***Edited by:***
*Agnieszka Wykowska,*
*Istituto Italiano di Tecnologia, Italy*

***Reviewed by:***
*Alessandro Roncone,*
*University of Colorado Boulder, United States*
*Marc Hanheide,*
*University of Lincoln, United Kingdom*

****Correspondence:***
*Khoi Hoang Dinh*
*khoi@lsr.ei.tum.de*

***Specialty section:***
*This article was submitted to Human-Robot Interaction, a section of the journal Frontiers in Robotics and AI*

***Received:*** *28 February 2019*
***Accepted:*** *17 July 2019*
***Published:*** *31 July 2019*

***Citation:***
*Hoang Dinh K, Oguz OS, Elsayed M and Wollherr D (2019) Adaptation and Transfer of Robot Motion Policies for Close Proximity Human-Robot Interaction. Front. Robot. AI 6:69. doi: 10.3389/frobt.2019.00069*

## 1. INTRODUCTION

Nowadays, robots are no longer only industrial machines behind fences. Instead, they are being integrated more in our daily lives as well as in collaborative manufacturing scenarios. The new generation of robots is expected to assist elderly people in daily tasks, to support customers in markets, to work as a partner with humans in factories, etc. For all of these tasks, the robots are required to interact with the human. Especially in collaborative scenarios, where robots work with humans as co-partners in joint tasks, they need to interact more efficiently since it will increase the overall performance. Looking at the case when two humans perform a joint task as an example, the humans can anticipate each others' movements and perform a complementary action without the need of verbal communication. This facilitates teamwork and increases the efficiency of joint

tasks (Erlhagen et al., 2007). Similarly, robots are expected to move in a natural way, similar to human-human interaction.To achieve such an interaction between humans and robots, the first requirement is the robot's motion must be readable to the human (Kirsch et al., 2010), which means the human partner is able to understand its intentions and the motion/behavior of the robot has to meet the expectations of the human partner. In the work of Lichtenthäler and Kirsch (2016), this is defined as legible robot behavior. Another requirement is that the robot has to be aware of its surroundings to provide a safe environment, while still being efficient in performing its task. Legibility and safety are therefore the two important criteria that increase the efficiency of joint collaboration between human and robot.

In order for humans to feel comfortable working with robots, especially in close proximity, they have to understand the robot's behavior and be able to infer their actions or in other words, the robot's behavior must be legible to the human partner. Identifying the factors that contribute to these natural movements is not trivial. According to a study conducted by Dautenhahn et al. (2005), participants want robots assisting at home to be predictable, controllable and have human-like communication. Another study (Koay et al., 2007) that investigated the subjective effects of direction of approach and distance of robots when handing an object over to humans, came to the conclusion that the frontal approach is subjectively preferred most by the participants since it is the most predictable. In addition, Bortot et al. (2013) discovered that understanding and predicting the behavior of the robot increases the well-being of humans.

The question arises how such legible robot motion can be generated. Dragan and Srinivasa (2013) tried to find one mathematical metric for legibility. However, this is insufficient as robot motion gets perceived differently by individual humans and depends on several factors including the configuration of tasks, robot positions and human positions. It is therefore necessary to have a framework in which the robot is able to learn legible motions by interacting directly with the human. In this way, all possible influencing factors will indirectly be included.

In addition, to fulfill the requirements mentioned above, legibility alone is not sufficient. In order to ensure the safety of humans in close proximity scenarios and allow joint collaborations, the robot has to know the position of the human and possibly predict their motion (Oguz et al., 2017) to modify its trajectories in real-time and reliably avoid collision with the human. Combining this safe behavior with legibility increases human comfort.

It is also worth mentioning that the main drawback of many learning approaches is the training time. The learning process usually requires several iterations of training and is time consuming to repeat for each new task and each human partner. In a lot of scenarios, it might be useful to have a flexible algorithm that still works if any parameter changes i.e., robot position, task configuration, human perspective, etc without the need of retraining. Therefore, the algorithm must be capable of extending to different scenarios and different tasks.

In this work, we develop a framework to generate legible robot motion that is transferable to different tasks and that is safe to allow collaboration in close proximity through a reinforcement learning approach. The interdependency between legibility, safety and efficiency is tackled for achieving natural human-robot interaction. Both human and robot collaborate in a joint scenario, i.e., in our case they have to reach similar objects, and the robot will adapt its motions over time corresponding to the reaction/prediction of the human partner. After training, the robot will be able to perform its tasks more efficiently and more predictable. This helps increase human comfort and the effectiveness of the collaboration. Our framework is also generalizable to similar tasks using learned policies in order to save training time.

## 2. RELATED WORK

Safety and legibility of robot motion in close proximity have always been investigated independently. Several methods were proposed that produce real-time obstacle avoiding trajectories, while others developed optimization based algorithms for legible robot motions.

Legible (or predictable) robot motion was first introduced in Dautenhahn et al. (2005). The result from their survey confirms the necessity of predictable behaviors in future robot companions. However, the paper does not focus on how to generate predictable behaviors for the robot. In the works from the Robotics and Artificial Intelligence Group at LAAS/CNRS (Alami et al., 2005; Sisbot et al., 2007, 2008; Sisbot et al., 2010; Sisbot and Alami, 2012), they developed a human aware motion and manipulation framework which is able to generate safe, comfortable and socially acceptable motions. The framework is verified on a mobile robot manipulator in simulated environment and in a hand-over scenario on real setup. The safety criterion introduced in their works, however, is based on the distance between the robot and the human, i.e., the robot should keep its distance from the human when performing tasks. While the framework is able to generate safe and legible motion, it is not applicable for joint tasks in close proximity since it does not allow the interaction between human and robot. As shown in the results of their papers, only the robot performs its tasks and there is no collaboration between them.

The work from Dragan et al. (2013) focuses explicitly on generating predictable and legible robot motion. In their work, the authors differentiate between legibility and predictability and provide a mathematical model to produce and evaluate such motions. They assume that humans expect robots to be efficient in their movements and compare all possible goals in the scene to determine the most probable one. This probability is formulated mathematically and is being maximized for the targeted goal. This approach has some limitations. The algorithm was tested only with two goals for the robot, which the human had to predict when pausing a video which showed the robot moving to one of the two (see **Supplementary Video**). This setup was very simple as the probability of selecting a goal (randomly) is already 50%. Another limitation is that the subjective evaluation of robot efficiency differs from one individual to another and the algorithm does not allow to adjust the robot's movements to individual preferences of each participant.

In the work of Stulp et al. (2015), the team generates robot motions that learn from the observation of a human participant and iteratively reduce the human's reaction time. Here, Dynamic Motion Primitives (DMPs) are used for motion planning. Policy Improvement through Black Box Optimization (PIBBO) (Stulp and Sigaud, 2012) is applied to improve the robot's legibility to the human iteratively. This is done by only optimizing human guessing time about the action of the robot and the correctness of the prediction without defining formal criteria about legibility. This approach provides flexibility in choosing the relevant parameters to be optimized to obtain legible motion. Recently, Busch et al. (2017) showed that transferring the learned policy to other individuals leads to better prediction in the beginning and can thus lead to shorter adaptation times for new subjects. However, in this work no close interaction scenarios were considered as no necessary collision avoidance methods were integrated and only the policy transfer to other subjects was investigated, not the policy transfer to new tasks.

Safety for humans during interaction with the robot, in general, involves several aspects and criteria (Robla-Gómez et al., 2017). There are also different categories of methods to ensure safety for the human partner (Lasota et al., 2017) i.e., safety through control, motion planning, consideration of psychological factors, etc. Within this work, we limit the safety aspect to the obstacle avoidance behavior of the robot and therefore only mention about methods that are able to provide this functionality to the robot. In this aspect, potential field (Khatib, 1985) is a very popular and widely used approach due to its simplicity and real-time capability. Flacco et al. (2012) and Dinh et al. (2015) utilize the potential field idea in their works to provide obstacle avoidance behavior on the end-effector of an articulated robot. In the work of Park et al. (2008), the authors introduce the dynamic potential field to adapt robot trajectories while avoiding obstacles in mid-motion. This dynamic potential field is used with the inverse kinematics with null-space constraints to further ensure collision avoidance between the human and robot's links. However, the aim of these approaches was not to enable the robot to interact with humans, but rather to perform desired movements in the presence of obstacles. In a recent study by Oguz et al. (2017), a stochastic motion planning algorithm is introduced that predicts human motions and adjusts the robot's trajectories on-line to avoid the predicted region. For the prediction of the human movement, Probabilistic Movement Primitives (ProMPs) were used, which were first introduced by Paraschos et al. (2013). This method learns the distribution of the motion during training and allows prediction of human motion in the online phase. This allows close interaction between humans and robots, but does not examine predictable or legible motion.

Inspired by the work of Stulp et al. (2015) and considering the requirements of joint human-robot collaboration in close proximity, in this paper we extend the learning approach in Stulp et al. (2015) with the potential field method. Our contribution is therefore a learning framework incorporating real-time obstacle avoidance to allow humans and robots working together in close proximity and therefore both legibility and safety aspects are tackled within our framework. This means that the human partner no longer stays outside of the robot workspace as a silent observer, but really cooperates with the robot in joint tasks in the same workspace. Apart from that, we also develop a task generalization method to generate policies for new tasks from previously learned tasks. With our task generalization approach, the robot is able to adapt to new tasks faster and hence the training time is reduced. We evaluate our approach on an articulated KUKA IIWA robot in virtual reality (VR) as well as in a real robot and complete the evaluation with a human study.

In the following, we first introduce our legible motion framework in section 3 then present our idea on the task generalization method in section 4. The improvement of our framework and task generalization approach is evaluated through experiments in section 5. Sections 6 and 7 provides further discussion and concludes our work.

## 3. LEGIBLE MOTION FRAMEWORK IN HUMAN ROBOT INTERACTION IN CLOSE PROXIMITY

A general overview of our framework is shown in **Figure 1**. The goal of the framework is to generate legible motion for the robot directly through interaction between the human and robot. Both of them collaborate in a joint scenario, i.e., in our case they have to reach similar objects, and the robot will adapt its motions over time corresponding to the reaction/prediction of the human partner. After training, the robot will be able to perform its tasks more efficiently and more predictable. This helps increase human comfort and the effectiveness of the collaboration. The framework therefore can be described in three steps as follow:

1. Firstly, Dynamic Movement Primitives (DMPs) are used to generate smooth trajectories with modifiable parameters. These trajectories are generated in Cartesian space and converted into the joint space of the robot using inverse kinematics.
2. DMP trajectories are then executed by the robot in the online phase where the robot collaborates with the human in a joint task. During execution, a potential field force is added to modify the DMP trajectories to ensure safety of the human.
3. A cost function which evaluates how the human partner perceives each trajectory is computed. These costs are then used to update the policy, which comprises the parameters of the DMPs in our framework. In the next iteration, new trajectories are sampled based on the updated policy and the procedure repeats until it converges to an optimal predictable trajectory or the maximum number of iterations is reached.

DMP trajectories are the trials/samples that the robot performs to understand how his human partner perceives a legible motion. By changing the parameters of the DMPs, the robot is able to exploit the working area and approach the goal from different angles. The human reacts to the robot by moving to his corresponding task. Each trajectory performed by the robot is then evaluated based on the human reaction formulated in a predefined cost function. This cost function reflects the perception of the human on how legible this trajectory is. Base on the evaluation of the cost

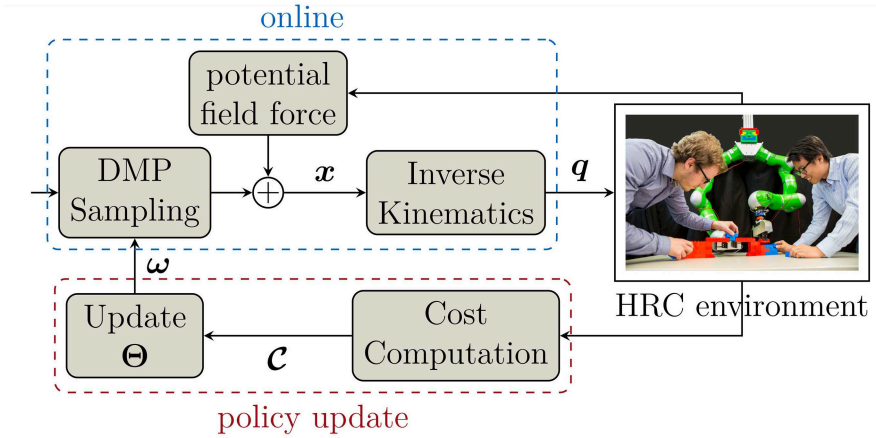



**FIGURE 1 |** Overview of the human-guided policy improvement framework. In each iteration, DMPs generate Cartesian trajectories $\boldsymbol{x}$, which are converted to joint angles $\boldsymbol{q}$ using inverse kinematics. During interaction, potential field force is added to modify the trajectories online to provide safety when the human gets close to the robot. After execution, a cost function $\boldsymbol{C}$ is evaluated to update the policy $\boldsymbol{\Theta}$ of the DMPs using PIBBO for the next iteration.

function of each trajectory, the DMP parameters will be modified in favor of the ones that are more predictable to the human (smaller costs). This is done by the policy update method called Policy Improvement through Black Box Optimization (PIBBO). After the DMP parameters (policies) are updated, the robot rolls out new samples from these parameters for the next iteration. The procedure is then repeated until the trajectories converge or the maximum number of iterations is reached. Note that all of these computations are done at the beginning of each iteration.

To prevent collision between human and robot during execution (online phase), the DMP trajectories are modified using a potential field force. This potential field force is proportional to the relative distance between the human and robot and returns an error vector that is added into the current DMP trajectory. As a result, the robot will move away when the human comes close, and recovers his task when the area is free. Additionally, in order to increase safety in close proximity, human motion is predicted using Probabilistic Movement Primitives (ProMPs) (Paraschos et al., 2013) and serves as **Supplementary Information** added into the potential field force. ProMPs is a recent approach that is able to generate/represent movement from a given trajectory distribution. After training with a set of human motion observations, we used ProMPs in the online phase to predict the movement of the human hand and incorporate this information into the potential field. This helps the robot react faster and can avoid the human more actively.

In section 3.1, we first briefly introduce DMPs and describe how they are used to generate smooth trajectories. The policy update method PIBBO is introduced and explained in section 3.2. This is followed by the explanation of how safety for the human partner is ensured through potential field force with the assistance of ProMPs in section 3.3. Finally, the cost function that evaluates the performance of each trajectory especially with a focus on collaboration effectiveness, is explained in detail in section 3.4.

## 3.1. Dynamic Movement Primitives

DMPs provide a method for trajectory control and planning that is able to represent complex motions and is flexible to be adjusted without manual parameter tuning or having to worry about instability (Ijspeert et al., 2002). DMPs comprise two parts, a dynamical system, and a nonlinear forcing term. In our work, the dynamical system is defined as a closed loop spring-damper system

$$\tau\ddot{y} = \alpha(\beta(y_g - y) - \dot{y}) \tag{1}$$

that converges to the defined attractor state $y_g$ where $\tau$ is the time constant, $\alpha$ and $\beta$ are positive constants. By setting $\beta$ to $\alpha/4$ we get a critically damped system. The variables $y$, $\dot{y}$ and $\ddot{y}$ are the position, velocity and acceleration, respectively.

The forcing term, which forms the second part of the DMPs, deforms the trajectory to match a desired shape. Thus, the spring-damper system is modulated to

$$\tau\ddot{y} = \alpha(\beta(y_g - y) - \dot{y}) + f(x), \tag{2}$$

where $f(x)$ is the forcing term consisting of a weighted sum of Gaussian basis functions multiplied by a canonical dynamical system, denoted as $x$. The canonical system $x$ is obtained by

$$\dot{x} = -\alpha_x x, \tag{3}$$

where $\alpha_x$ is a constant. The canonical system state $x$ in (3) starts at some arbitrary value and goes to 0 as time goes to infinity. This ensures convergence to the goal while keeping the forcing term not directly dependent on time. The forcing function $f(x)$ hence has the form

$$f(x) = \frac{\sum_{i=1}^{N} \psi_i(x)\omega_i}{\sum_{i=1}^{N} \psi_i(x)} x, \tag{4}$$

where

$$\psi_i(x) = \exp\left(-\frac{1}{2\sigma_i^2}(x - c_i)^2\right) \tag{5}$$

defines the Gaussian basis functions with means $c_i$ and variances $\sigma_i$. In (4), $N$ is the number of basis functions and $\omega_i$ are modifiable weights, which are adjusted to match the desired trajectory. They are optimized by the policy improvement method explained in section 3.2.

Since the mass spring-damper system leads to high initial accelerations, which is not desirable for robots, we use a goal system, which moves the attractor state of the system from the initial state $y_0$ to the goal state $y_g$ during the movement. This delayed goal attractor $y_{gd}$ itself is given as an exponential dynamical system that starts at $y_0$ and converges to $y_g$.

$$\dot{y}_{gd} = -\alpha_g(y_g - y_{gd}) \tag{6}$$

Thus the equation for the DMPs resolves to

$$\tau\ddot{y} = \alpha(\beta(y_{gd} - y) - \dot{y}) + f(x) \tag{7}$$

The DMPs has several advantages, which make it suitable for our framework:

- It is guaranteed to converge to the goal, since the canonical system is 0 at the end of every movement.
- The weights $\omega_i$ can be adapted to generate any desired trajectory. In our case this is especially relevant, since we want to learn the optimal trajectory and adjust the weights online with each interaction.
- As there is no time-dependency, the duration of the movement can simply be altered by adjusting $\tau$.

## 3.2. Policy Improvement Through Black-Box Optimization

Policy improvement methods seek to optimize the parameters of a policy w.r.t. a utility function. In our work, we use a policy improvement method to iteratively update the weights of the DMP to obtain a desired trajectory. Policy improvement methods have two basic steps:

1. Exploration by perturbation: The exploration noise $\boldsymbol{\epsilon_t}$ can be either added to the actions, i.e., the output of the policy ($\pi_{\boldsymbol{\theta}}(x) + \boldsymbol{\epsilon_t}$), or directly to the input parameters of the policy ($\pi_{\boldsymbol{\theta}+\boldsymbol{\epsilon}_t}(x)$).
2. Policy update: Here, the parameters of the policy are updated in order to minimize a predefined cost metric $C$. Usually, gradient descent is applied to iteratively converge to a local minimum. Another method is the reward-weighted averaging, which is used in our application.

Reward-weighted averaging does not require differentiability of the cost function, which makes it more stable than gradient descent if the cost function is not continuous.

Specifically for this work, we choose Policy Improvement through Black-box Optimization (PIBBO) as our policy improvement method (Stulp and Sigaud, 2012). PIBBO treats the whole control trajectory as a black-box, i.e., no assumptions are made about the search space or the cost function. An important property of PIBBO is that the search is done in the space of policy parameters, thus it is a parameter perturbing approach. The output $u_t$ of the policy is computed as:

$$u_k = \pi_{\boldsymbol{\theta}+\boldsymbol{\epsilon}_k}(x), \quad \text{with} \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \Sigma) \tag{8}$$

In our case the policy $\pi_{\boldsymbol{\theta}}$, is the DMP and $\boldsymbol{\theta}$ are the corresponding weights for the Gaussians.

The parameter update is done using reward-weighted averaging. First, the cost $C_k$ for each trajectory roll-out is computed. Then we assign higher probabilities $P_k$ to trajectories with a lower cost and vice versa.

$$P_k = \frac{e^{-1/\lambda C_k}}{\sum_{k=1}^{K} e^{-1/\lambda C_k}} \tag{9}$$

$k$ is the number of roll-outs and $\lambda$ is a constant between 0 and 1.

The parameter update is then given as

$$\delta\boldsymbol{\theta} = \sum_{k=1}^{K} P_k \boldsymbol{\epsilon}_k \tag{10}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \delta\boldsymbol{\theta}. \tag{11}$$

After taking the weighted average of all roll-outs, the new DMP with updated parameters $\boldsymbol{\theta}$ follows the trend of trajectories with high probabilities (i.e., low costs). This process of perturbing and updating is repeated until the desired cost value is achieved or the maximum number of updates is reached.

The exploration is done by rolling out different trajectories and evaluating them using the cost values resulting from the interaction with the human. Before outlining the cost function in detail, we discuss the safety aspect of the human partner in close proximity.

## 3.3. Safety Aspect in Close Proximity

As the human works together with the robot in close proximity, safety of the human needs to be considered. In essence, the robot should be able to physically avoid the human to prevent any collision. In this section, we describe our approach to provide a safety aspect for the robot. The main idea is to create an artificial repulsive force to push the robot away whenever the human comes close (Khatib, 1990; Park et al., 2008). Furthermore, to improve the robot reactivity, the human motion is also considered. In our approach, we use Probabilistic Movement Primitives (ProMPs) to predict the human motion and incorporate its effect into the repulsive force. Our idea about generating repulsive force for obstacle avoidance will be introduced in section 3.3.1, after that, an introduction about ProMPs and how human motion prediction extracted from ProMPs is incorporated will be given in section 3.3.2.

### 3.3.1. Repulsive Force With Artificial Potential Field

The robot trajectory is generated by the DMP at the beginning of every update. We want to modify this trajectory to avoid the human partner while still generating smooth motions and following the original DMP trajectory when the human is out of reach. As the DMP trajectory is already smooth based on its

formulation (see section 3.1), the artificial repulsive force also has to generate a smooth transition on the robot. This is important for the human partner to feel comfortable when working with the robot. A simple solution is to make the robot behave like a virtual mass-spring-damper system regarding to external forces (Hogan, 1984)

$$\boldsymbol{F}_{\text{ext}} = \boldsymbol{M}\ddot{\boldsymbol{e}} + \boldsymbol{D}\dot{\boldsymbol{e}} + \boldsymbol{K}\boldsymbol{e}, \tag{12}$$

where $\boldsymbol{F}_{\text{ext}} \in \mathbb{R}^3$represents an external virtual force, which is excited whenever the human enters the safety area around the end-effector of the robot. This virtual mass-spring-damper system results in a smooth transition in the vector $\boldsymbol{e} \in \mathbb{R}^3$ regardless $\boldsymbol{F}_{\text{ext}}$. This vector indicates the modification length and direction to be added to the DMP. $\boldsymbol{M}, \boldsymbol{D}, \boldsymbol{K} \in \mathbb{R}^{3\times3}$are positive definite matrices that represent the mass, damping and stiffness of the virtual system. In our proposed setup, $\boldsymbol{M}$ is chosen as the identity matrix, $\boldsymbol{K}$ and $\boldsymbol{D}$ are diagonal matrices chosen to adapt the desired reaction to virtual forces. Increasing the damping results in a slower reaction but smoother movement of the robot. The external virtual force $\boldsymbol{F}_{\text{ext}}$ is computed based on potential fields w.r.t the distance between the end-effector of the robot and obstacles.

The idea of potential fields was first introduced in the work of Khatib (1990). Whenever an obstacle is inside a threshold region of the end-effector, a repulsive force vector $\boldsymbol{F}_{\text{ext}}$ according to (12) is generated. Here, we use the same idea of repulsive vectors (Flacco et al., 2012; Dinh et al., 2015) to generate a smooth reaction force

$$\boldsymbol{F}_{\text{ext}} = \frac{\boldsymbol{F}_{\max}}{1 + \exp\left(\left(\left\|\boldsymbol{d}(E,O)\right\|(2/\rho) - 1\right)\gamma\right)}, \tag{13}$$

where $\mathbf{F}_{\max}$ is the maximum force applied, $\left\|\boldsymbol{d}(E,O)\right\|$is the distance between obstacle $O$ and end-effector $E$, $\rho$is the threshold distance that defines the collision region around the end-effector and $\gamma$is a shape factor. The force reaches its maximum if the distance equals zero, and zero if the obstacle is outside the region, respectively. The steepness of the force profile within the threshold region regarding the distance can be adjusted by the shape factor $\gamma$. With $\boldsymbol{F}_{\text{ext}}$, the error vector $\boldsymbol{e}$ is obtained from (12) which return in the deviation needs to be added into the DMP to avoid the obstacle.

### 3.3.2. Human Motion Prediction With ProMP

Although the robot is able to avoid the human with the repulsive force generated from the potential field, its reaction time is an important factor that needs to be considered. In a confined workspace where the human usually interferes with the robot, the robot might not have enough time to react and fail to avoid the human partner. Increasing the safety region around the robot can improve the reaction time but results in a smaller workspace. Thus, in our framework, we estimate and predict the human motion and add this additional information into the repulsive force to increase the responsiveness of the robot.

In general, human motion estimation requires a specialized prediction method due to the inter- and intra-personal movement variations (Todorov, 2004). To imitate such behavior online, we use ProMPs and learn a distribution of a motion behavior by training with multiple trajectories performed for a specific task (Paraschos et al., 2013). ProMPs represent a discrete trajectory $X = \{x_n\}, n = 0 \ldots N$ defined by states $x_n$ over time $N$ with the formulation

$$\boldsymbol{y}_n = [x_n, \dot{x}_n]^\top = \boldsymbol{\Phi}_n^\top \boldsymbol{\omega} + \boldsymbol{\epsilon}_y, \tag{14}$$

where $\boldsymbol{\omega} \in \mathbb{R}^{k\times2}$ is the weighting matrix over the $k \times 2$ dimensional time-dependent basis matrix $\boldsymbol{\Phi}_n = [\phi_n, \dot{\phi}_n]$ with $k$ being the number of basis functions and $\boldsymbol{\epsilon}_y \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_y)$ is zero-mean independent Gaussian noise, while $\boldsymbol{\Phi}_n^\top \boldsymbol{\omega}$ gives the mean of the trajectory. Introducing a Gaussian distribution to also represent variance $p(\boldsymbol{\omega}; \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\omega}|\boldsymbol{\mu}_{\boldsymbol{\omega}}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}})$ over the weighting vector $\boldsymbol{\omega}$ results in the following distribution for the trajectory:

$$\begin{aligned} p(\boldsymbol{y}_n; \boldsymbol{\theta}) &= \int \mathcal{N}(\boldsymbol{y}_n|\boldsymbol{\Phi}_n^\top \boldsymbol{\mu}_{\boldsymbol{\omega}}, \boldsymbol{\Sigma}_y)\mathcal{N}(\boldsymbol{\omega}|\boldsymbol{\mu}_{\boldsymbol{\omega}}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}})d\boldsymbol{\omega} \\ &= \mathcal{N}(\boldsymbol{y}_n|\boldsymbol{\Phi}_n^\top \boldsymbol{\mu}_{\boldsymbol{\omega}}, \boldsymbol{\Phi}_n^\top \boldsymbol{\Sigma}_{\boldsymbol{\omega}} \boldsymbol{\Phi}_n + \boldsymbol{\Sigma}_y). \end{aligned} \tag{15}$$

Using a set of motion observations, the parameters $\boldsymbol{\mu}_{\boldsymbol{\omega}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}$ can be computed by maximum likelihood estimation (Lazaric and Ghavamzadeh, 2010).

By this formulation, an online human motion prediction, where a trajectory along with the variance for each discretized time point is generated. This predicted trajectory can be used in different ways within our framework. An intuitive way is to select some predictions at different time points along the trajectory. These predictions represent the points in space where the human *might* occlude in the future and thus are treated as *incoming* obstacles that the robot has to avoid. This triggers the reaction of the robot even if the human is not currently within the safety region, which in turn increases the responsiveness of the robot. In case the human does not move toward the robot, these *incoming* obstacles do not create any disturbance, thus do not alter the robot desired position.

## 3.4. Cost Computation

In this section, we will explain how the cost function in our framework (**Figure 3**) is defined. There are different aspects that we want to evaluate through the cost function:

- First is the legibility of the robot trajectories. There are different methods to measure this aspect. In the works of Dehais et al. (2011) and Lichtenthäler et al. (2011), they show the participants robot motions and afterwards ask them to rate how legible the motions were perceived. In a quantitative level, Dragan and Srinivasa (2013) and Busch et al. (2017) show the participants robot motions through videos/experiments and ask them to indicate immediately or press a button when they feel certain about the robot's intention. Time and correctness of the prediction are used as the indicators for legibility in their works. Using the same approach as in Busch et al. (2017), we also use the human prediction time and accuracy to form the cost of legibility.
- Second is the smoothness of the trajectories. This helps the human partners feel comfortable when working with

the robot and be more confident approaching their goals. Smoothness also contributes in the legibility aspect since a jerky motion does not meet the expectation of the human. In our framework, we use the third derivative of the trajectories to form the cost of smoothness.

From the two aspects that we want to evaluate, several components are identified and also mixed up depending on the experimental setup. Here, we list all the costs used in this work:

- End-effector jerk $V_{\text{ej}}$: the sum of the third derivative of the end-effector position of the robot at each sample along trajectory.
- Angular jerk $V_\theta$: the sum of the third derivative of the angular positions of the controlled joints of the robot at each sample along trajectory.
- Human prediction time $V_{\text{pred}}$: the time taken by the human to make a prediction about the robot's target. It starts when the robot starts moving and ends when the human reaches one of the targets.
- Accuracy $V_{\text{task}}$: whether the human prediction was correct, translating to 0 cost ($V_{\text{task}} = 0$), or if the prediction was wrong which results to a cost of 1 ($V_{\text{task}} = 1$).
- Human duration $V_{\text{dur}}$: the duration of the human movement between when the human starts moving and reaches the goal. It is a measurement of human's confidence in the robot's presence.
- The weighted distance between the robot trajectories, $V_\delta$, which measures how distinct the trajectory to the targeted goal is in comparison to the trajectories to the other goals. This cost is calculated using the following equation:

$$V_\delta = \left( \sum_{g=1}^{G} \sum_{t=0}^{T} \frac{1}{t} \left\| \boldsymbol{p}_t, \boldsymbol{q}_t \right\|_2 \right)^{-1} \tag{16}$$

where $G$ is the number of the goals excluding the targeted goal, $g$ is the other goal whose trajectory is compared to the targeted goal trajectory, $t$ is the time step at which we calculate the distance, $T$ is the total time of the trajectory, $\boldsymbol{p}_t$ is the point at $t$ in the trajectory to the targeted goal, $\boldsymbol{q}_t$ is the position at $t$ in the trajectory to the goal $g$ and $\left\| \boldsymbol{p}_t, \boldsymbol{q}_t \right\|_2$ is the Euclidean distance between $\boldsymbol{p}_t$ and $\boldsymbol{q}_t$.

In summary, the cost function has the form

$$V = \lambda_{\text{ej}} V_{\text{ej}} + \lambda_\theta V_\theta + \lambda_{\text{pred}} V_{\text{pred}} + \lambda_{\text{task}} V_{\text{task}} + \lambda_{\text{dur}} V_{\text{dur}} + \lambda_\delta V_\delta \tag{17}$$

where each cost component is weighted differently. In general, $\lambda_{\text{pred}}, \lambda_{\text{task}} > \lambda_{\text{ej}}, \lambda_\theta, \lambda_{\text{dur}}, \lambda_\delta$ as we want to have a high reward for trajectories that are more predictable to the human partner.

## 4. TASK GENERALIZATION

Even though our framework generates predictable policies, the learning procedure requires a considerable amount of data and thus time until a convergent behavior is achieved. Furthermore, the trained polices directly depend on the specific setup. When the environment changes, e.g., the start/goal positions of the robot or the relative position of the human w.r.t. the robotic partner, the robot needs to adapt to this new configuration.

Given a fixed number of policies that have been learned on specific settings, the existing knowledge can be exploited, such that the adjustment to variations of similar tasks can be achieved given limited data. In other words, since the prior policies learned already encode some preference of human perception, they can be used to improve the learning convergence rate for the cases that the robot has not been trained for. We propose an approach to realize such a generalization capability for the policy improvement framework within HRI settings.

Suppose that the set of tasks for the robot is defined as

$$\boldsymbol{\Phi} = \left\{ \boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_M \mid M \in \mathbb{N} \right\}, \tag{18}$$

where $M$ is the number of available tasks. Within the scope of this work, a task is defined as a reaching motion, where the starting position is the same for all of the tasks and $\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_M$ are $M$ different goal positions. Learning via PIBBO is done by selecting a subset $\boldsymbol{\mathcal{T}_i}$ out of $\boldsymbol{\Phi}$ and training trajectories for each goal in $\boldsymbol{\mathcal{T}}_i$, where

$$\boldsymbol{\mathcal{T}}_i = \left\{ \boldsymbol{g}_{i1}, \boldsymbol{g}_{i2}, \cdots, \boldsymbol{g}_{iS} \right\} \subset \boldsymbol{\Phi}, S \in \mathbb{N}, \boldsymbol{g}_{ij} \neq \boldsymbol{g}_{ik}, \forall j \neq k \tag{19}$$

with a predefined $S < M$. The result of PIBBO is $S$ policies that generate predictable trajectories for each $\boldsymbol{g}_{ij}$ over $\boldsymbol{\mathcal{T}_i}$. Each policy is parameterized by $\boldsymbol{\Theta}^{ij}_{\boldsymbol{\mathcal{T}}_i}$, e.g., in our case given as the weighted basis functions of the DMP. Note that the policy of $\boldsymbol{g}_{ij}$ depends on the remaining goals in $\boldsymbol{\mathcal{T}}_i$, which means a similar task will have different policies if it belongs to a different subset. We then denote the generated policy for a goal $\boldsymbol{g}_{ij}$ from $\boldsymbol{\mathcal{T}_i}$ as

$$\boldsymbol{\pi}_{\boldsymbol{\mathcal{T}}_i}(\boldsymbol{g}_{ij} \mid \boldsymbol{g}_{i\setminus j}) = \boldsymbol{\pi}_{\boldsymbol{\mathcal{T}}_i}(\boldsymbol{g}_{ij}) = \boldsymbol{\pi}(\boldsymbol{\Theta}^{ij}_{\boldsymbol{\mathcal{T}}_i}) \tag{20}$$

where $\boldsymbol{g}_{i\setminus j}$ is an abbreviation of all tasks in $\boldsymbol{\mathcal{T}_i}$ except $j$. This can be interpreted as the policy that generates the most predictable motion for goal $\boldsymbol{g}_{ij}$ given the remaining tasks in $\boldsymbol{\mathcal{T}}_i$.

Given a training set $\mathbb{T} = \{\boldsymbol{\mathcal{T}}_1, \boldsymbol{\mathcal{T}}_2, \cdots, \boldsymbol{\mathcal{T}}_k\}$ consisting of $k$ batches of $S$ elements from $\boldsymbol{\Phi}$ each, a new $\widetilde{\boldsymbol{\mathcal{T}}} \notin \mathbb{T}$ is drawn from $\boldsymbol{\Phi}$. The objective here is to find a new policy for a goal $\boldsymbol{g}_m \subset \widetilde{\boldsymbol{\mathcal{T}}}$ such that the DMPs initialized using this policy improve the convergence rate of the learning procedure of $\boldsymbol{g}_m$ afterwards. This requires finding a mapping

$$\boldsymbol{\pi}_{\widetilde{\boldsymbol{\mathcal{T}}}}(\boldsymbol{g}_m) = \boldsymbol{h}\left( \boldsymbol{\pi}_{\boldsymbol{\mathcal{T}}_1}(\boldsymbol{g}_{11}), \ldots, \boldsymbol{\pi}_{\boldsymbol{\mathcal{T}}_1}(\boldsymbol{g}_{1S}), \ldots, \boldsymbol{\pi}_{\boldsymbol{\mathcal{T}}_k}(\boldsymbol{g}_{kS}) \right) \tag{21}$$

with $\boldsymbol{h}(\cdot)$ is a function of all policies obtained from the training set $\mathbb{T}$. In fact, solving (21) is equivalent to finding the parameterized vector $\boldsymbol{\Theta}^m_{\widetilde{\boldsymbol{\mathcal{T}}}}$ in Equation (20) for goal $\boldsymbol{g}_m$ in the new subset $\widetilde{\boldsymbol{\mathcal{T}}}$.

We claim that a predictable trajectory for each goal in $\boldsymbol{\mathcal{T}_i}$ depends on a set of features $\boldsymbol{\chi}$. These features characterize the interrelation between $\boldsymbol{g}_{ij}$ and $\boldsymbol{g}_{i\setminus j}$ in the subset $\boldsymbol{\mathcal{T}}_i$. They can be relative distances, angles, etc, depending on how the set of tasks $\boldsymbol{\Phi}$ is defined. These features vary for each goal in each subset. Given a predefined set of $p$ features for goal $\boldsymbol{g}_{ij}$ in $\boldsymbol{\mathcal{T}}_i$, we denote the resulting feature vector for each goal as $\boldsymbol{\chi}_{\boldsymbol{\mathcal{T}}_i}(\boldsymbol{g}_{ij}) \in \mathbb{R}^p$. We now want to establish a relation between $\boldsymbol{\chi}_{\boldsymbol{\mathcal{T}}_i}(\boldsymbol{g}_{ij})$ and vector

$\boldsymbol{\Theta}^{ij}_{\boldsymbol{\mathcal{T}}_i}$, which is the policy of $\boldsymbol{g}_{ij}$ in $\boldsymbol{\mathcal{T}}_i$. Furthermore, the weighted basis functions of the DMP in $\boldsymbol{\Theta}^{ij}_{\boldsymbol{\mathcal{T}}_i}$ are independent from each other, hence can be evaluated individually. Therefore, we propose an approximation to initialize each individual weight $\Theta \in \boldsymbol{\Theta}^{ij}_{\boldsymbol{\mathcal{T}}_i}$ as follow

$$\Theta = \beta_0 + \sum_{k=1}^{p} \beta_k \chi_k(\boldsymbol{g}_{ij}), \tag{22}$$

where $\chi_k \in \boldsymbol{\chi}_{\boldsymbol{\mathcal{T}}_i}$ represents a single feature in the set of $p$ features for goal $\boldsymbol{g}_{ij}$. Given the trained policies from $\mathbb{T}$ and a predefined set of features $\boldsymbol{\chi}$, $\boldsymbol{\beta} = \{\beta_0, \boldsymbol{\beta}_k\}$ in (22) is obtained by solving the linear regression problem. Assuming $\boldsymbol{\Theta}^{ij}_{\boldsymbol{\mathcal{T}}_i}$ has $N$ basis functions, then $N$ linear regression problems of (22) are solved individually to obtain $N$ sets of $\boldsymbol{\beta}$, denoted as $\boldsymbol{\beta}_i$ where $i$ denotes the according basis function index.

From there, given the new subset $\widetilde{\boldsymbol{\mathcal{T}}}$, the generalized policy for a goal $\boldsymbol{g}_m$ in $\widetilde{\boldsymbol{\mathcal{T}}}$ is initialized by the approximate value as

$$\boldsymbol{\Theta}^{m}_{\widetilde{\boldsymbol{\mathcal{T}}}} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_N \end{bmatrix} \boldsymbol{\chi}_{\widetilde{\boldsymbol{\mathcal{T}}}}(\boldsymbol{g}_m), \tag{23}$$

where $\boldsymbol{\chi}_{\widetilde{\boldsymbol{\mathcal{T}}}}(\boldsymbol{g}_m)$ is the features of $\boldsymbol{g}_m$ in $\widetilde{\boldsymbol{\mathcal{T}}}$. We then use this policy as an initialization for the DMP when learning predictable motion for the new subset $\widetilde{\boldsymbol{\mathcal{T}}}$. Details about our implementation and results are outlined in section 5.3.

## 5. RESULTS

In this section, we present different experiments to evaluate our framework and task generalization method. In section 5.1, we first describe the experimental setup in virtual reality (VR). The legibility results are shown in section 5.2 while in section 5.3, we present the results of our task generalization approach. Finally, to verify the safety aspect of our approach we conducted an experiment on a real KUKA LWR 4+ robot and present its results in section 5.4.

### 5.1. Experimental Setup in Virtual Reality

We conduct our main experiments in a VR environment as shown in **Figure 2**. There are advantages of VR that facilitate our work: First, it is easier to change the environment or switch to different robots and second, VR provides a first person point of view that is similar to how humans would perceive their environment, which makes it suitable for our work.

In the experiment, the participant wears a VIVE pro headset and stands in front of a table with the robot mounted on it in VR. We added a real table at the exact location as in VR, which both acts as a physical support and improves the realism of the interaction. The position of the robot is different depending on the experiments. In our case, we use two configurations: (*i*) the robot is mounted on the same side of the participant, and (*ii*) on the opposite side of the participant relative to the collaborative task area. The first case emphasizes the side-by-side perspective of the human toward the robot motions and the second case highlights the direct point of view when the human observes the robot motions from the opposite side. Here we want to investigate if this perspective also affects the predictable motion of the robot. To facilitate the collaboration between the human and the robot, we design the tasks for both as reaching designated goals. The goals of the robot are visualized as cylinders and the goals of the human are visualized as spheres. Each goal of the robot has a corresponding goal of the human with the same color. They are positioned near each other to evaluate obstacle avoidance behavior (**Figure 2**).

For each experiment, there are three different goals for the robot and three corresponding goals for the human. The robot starts first by moving to one randomly chosen goal and the participant has to predict which one the robot is aiming at and moves the VR controller to the corresponding goal with same color when they feel confident about the target of the robot. After that, both the participant and robot move back to their starting positions and the procedure repeats. The participant is informed that this is a collaborative task, therefore they are expected to find a balance between making a correct prediction or being fast and reacting early. For example, making many wrong predictions results in failing the tasks, whereas having long prediction time increases the total amount of time for both to finish their tasks. Both cases reduce the efficiency of the collaboration.

Each experiment consists of a habituation phase and an evaluation phase. The purpose of the habituation phase is to get the participants acquainted to the VR environment and the used equipment as well as familiarized to the robot motions and their own task. This habituation phase reduces the learning effect during the main evaluation phase. During the evaluation phase, the participants are asked to answer a questionnaire. The answers are scaled onto 5 different levels: *strongly disagree*, *disagree*, *neutral*, *agree*, *strongly agree*. There are 11 questions in total, that are classified into 5 categories:

- How does the participant feel about the smoothness of the trajectories?
- Does the participant feel safe when working with the robot?
- Are the robot trajectories predictable?
- How natural and comfortable the participant feel about the robot trajectories?
- How does the participant like and want to work with the robot again?

The user study was approved by the ethics committee of the TUM School of Medicine. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

For all experiments, if not mentioned specifically, we use configurations and parameters described as follow:

- For DMP, we use three goal systems for the three Cartesian goal positions of the end-effector. These goal systems are first initialized with straight lines. The DMP has 5 equally spaced Gaussian radial basis functions and there are 5 samples per update for each goal. In the sampling phase, we add perturbations with the covariance size as 200 to the DMP parameters and run the policy for each sample. With each

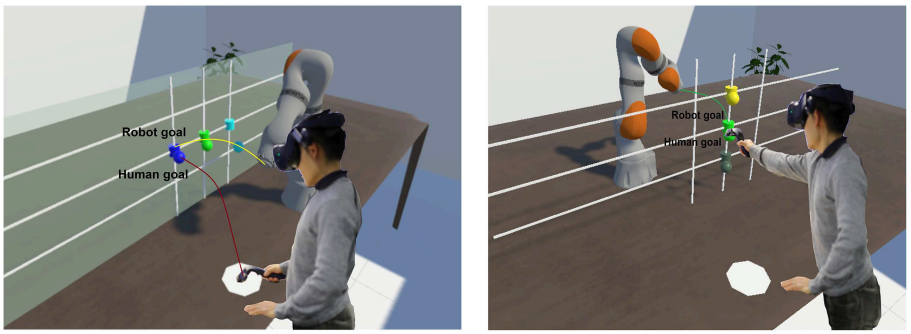



**FIGURE 2 |** Experiment setup with different configurations: **(A)** Human and robot are on the same side. **(B)** Robot is on the opposite side of the human.

iteration, we let the variance factor for the perturbations decay as it helps reducing the search space for the parameters over time.
- For obstacle avoidance, we use a motion capture system to detect the position of the human (and the velocity), which are then used to compute the repulsive force. The maximum force $\boldsymbol{F}_{\max}$ is set to $300N$ and the obstacle threshold is 20 cm around the end-effector of the robot.
- The weights of cost components used in the experiment are: $\lambda_{\text{ej}} = 1, \lambda_{\theta} = 2, \lambda_{\text{pred}} = 8, \lambda_{\text{task}} = 10, \lambda_{\text{dur}} = 1, \lambda_{\delta} = 3$. The weights of the human prediction time and accuracy costs are relatively higher than the others.

Over time, the policy of the robot is updated to adapt to the preferences of the human and produces more predictable movements to the human partner. The results of this adaptation are presented in the following section.

In order to convert the Cartesian trajectory produced by the DMP into joint positions, we use traditional inverse differential kinematics:

$$\dot{\boldsymbol{\theta}} = \boldsymbol{J}^{+}\dot{\boldsymbol{x}} \tag{24}$$

with $\boldsymbol{J}^{+}$ being the pseudo inverse of the Jacobian $\boldsymbol{J}$ of the end-effector (Penrose, 1955), $\boldsymbol{\theta} \in \mathbb{R}^7$ is joint configuration and $\boldsymbol{x} \in \mathbb{R}^3$ is Cartesian position. The pseudo inverse gives the least square approximation to the real inverse. In our case only the pseudo inverse is applicable, as we map three Cartesian values to seven joint positions, which makes the Jacobian not quadratic and thus not regular. We constrain the covariance size of the DMPs to avoid generating trajectories out of the robot's reach. In addition, the joint configuration corresponding to the starting position is fixed for all trajectories. In this way, the elbow position of the robot resulting from joint redundancy does not change significantly during the experiment. Hence, the adaptation effect is mainly visible on the end-effector movement. The motion of the end-effector is formed based on the DMPs trajectories and the potential field force applied to it and it is the major factor for the human partner to differentiate between different robot motions. Our detailed implementation is provided in https://github.com/khoilsr/hrc_legible_motion_generation.

## 5.2. Predictable Robot Motion for a Specific Setup

Given a specific setup, which in our case comprises the goals of the human and the robot in addition to the robot mounting position (either in the same side or opposite side of the human), the predictable trajectories are obtained through the learning framework. We conduct experiments with different participants on different configurations to evaluate overall performance. To quantify the performance of our framework, we look at the following criteria:

- The total cost $V$ and human prediction time cost $V_{\text{pred}}$ (section 3.4) for each update. $V_{\text{pred}}$ is used to quantify the legibility of the robot motion while $V$ shows the overall efficiency of the learning framework.
- The opinion of the subject about how legible robot motions are after each phase.
- The converged trajectory for each goal after learning w.r.t each subject.

The first two criteria will be discussed in sections 5.2.1 and 5.2.2 while the last one will be analyzed in section 5.2.3.

### 5.2.1. Evaluation of the Learning Framework

Fifteen participants took part in this study. As mentioned, each experiment consists of a habituation phase and an evaluation phase. In the habituation phase, 30 trials are executed using invariable DMP trajectories. After its completion, the evaluation phase starts, which consists of 10 updates with 5 trials per update for each of the three goals, resulting in 150 trials in total. This number is comparable to Stulp et al. (2015) and Busch et al. (2017). To evaluate the participants' perception during the experiment, this phase is divided in three blocks with two breaks after the $4^{th}$ and $7^{th}$ update, respectively, in which the participants are asked to fill a short questionnaire (see results in **Figure 4**).

The prediction time and accuracy from all participants is collected using a motion capturing system after each trial to update the cost function and evaluate the framework over time. The human prediction time is calculated by measuring the time between the start of the robot's motion until the participant reaches their goal. Since each human being has a different

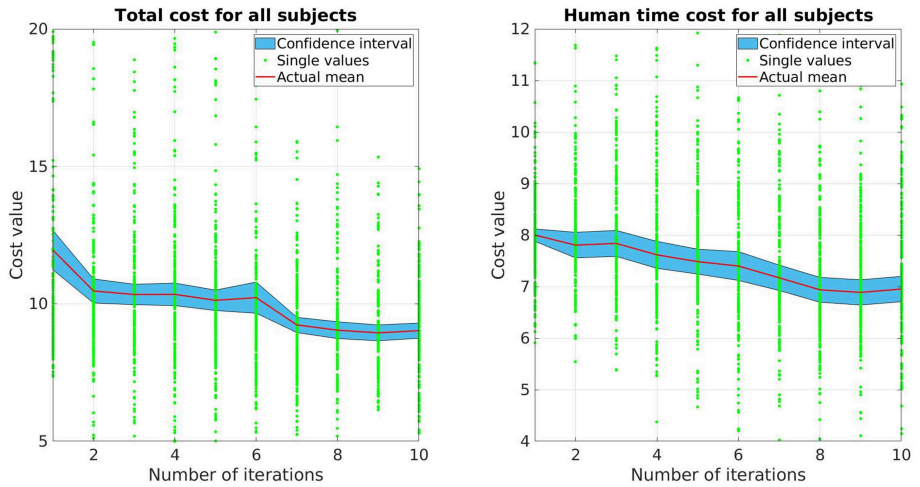



**FIGURE 3 |** The mean and confidence interval of the total cost and human prediction time cost for all subjects.

inherent reaction speed, we normalize the measurement of the human prediction time of each participant by their responses on the first update, which is computed as the average of 15 values of the human prediction time.

Both total cost $V$ and human prediction time cost $V_{\text{pred}}$ for all subjects are presented in **Figure 3**. For each update, there are 225 data points (15 trials per update for 15 participants), each data point represents the measurement of one single movement of the participants. The red line depicts the mean while the blue area illustrates the 95% confidence interval. As shown, both cost values decrease over time. The human prediction time cost $V_{\text{pred}}$ drops around 13%, while the total cost $V$ drop is around 24%. Comparing the data between the first and the last update, a pair-sampled $t$-test indicates that they are both significantly different from each other ($t = 7.142, p < 0.001$ for $V$ and $t = 7.437, p < 0.001$ for $V_{\text{pred}}$). The decrease of human prediction time indicates that the subjects are able to predict and react faster to the robot motions while the reduction of total cost also implies that the subjects predict more accurately over time (accuracy cost has the highest weight).

The subjective legibility of robot trajectories is measured by the questionnaire during the breaks and after the last update. Here, we asked the participants' opinion on two statements: *the robot's intention was clear* and *it was easy to predict which goal the robot is targeting*. We get the average of the two answers as the measurement of legibility aspect from the human perspective. The trajectories become more predictable as the median increases over time (**Figure 4A**). An interesting result that can be observed here is that the interquartile range is reduced from phase 1 to phase 2, however it slightly increases from phase 2 to phase 3. This means the improvement from phase 2 to phase 3 is not very clear as the mean increase but the data spread is also larger. One reason for this is due to the trajectories of the robot start to get close to the converged one after a few updates and the updated trajectories of phase 2 and phase 3 are quite close together. An example of this behavior is shown in **Figure 4B**, where the trajectories start as a straight line toward the goals and after a few updates, get close to the converged trajectories depicted as the bold and dark curves for each goal.

Overall, it can be concluded that, given a specific setup, human prediction time and subjective legibility can be improved through our framework and therefore can boost the efficiency of the collaboration between human and robot. However, the question arises here whether the learning effect of the participants plays a significant role in the improvement of the results, since the experiment is designed as a repetitive task. This will be discussed further in the next section.

### 5.2.2. Comparison With Non-adaptive Robot

In this section, we compare our method with a non-adaptive baseline. Even though we reduce the learning effect from the participants through the habituation phase, there is still probability that the human adapts to the motions of the robot over time. Therefore, the goal of this section is to investigate if the prediction of the human is improved due to the legible motions of the robot or because of human adaption. We design two experiments with the same environment setup, i.e., the tasks and the positioning of human and robot are the same. We use the counterbalanced ABBA design and define the following two groups:

- Group I: Subjects within this control group first interact with the non-adaptive robot, then with the adaptive robot subsequently.
- Group II: Subjects within this control group first interact with the adaptive robot, then with the non-adaptive robot subsequently.

In the case of non-adaptive robot, we also use our framework, but the policies (the parameters of the DMP) will not be updated. Therefore, the non-adaptive robot will always follow a straight line from the start toward the goal in every motion. As there is no adaption from the robot, the results from

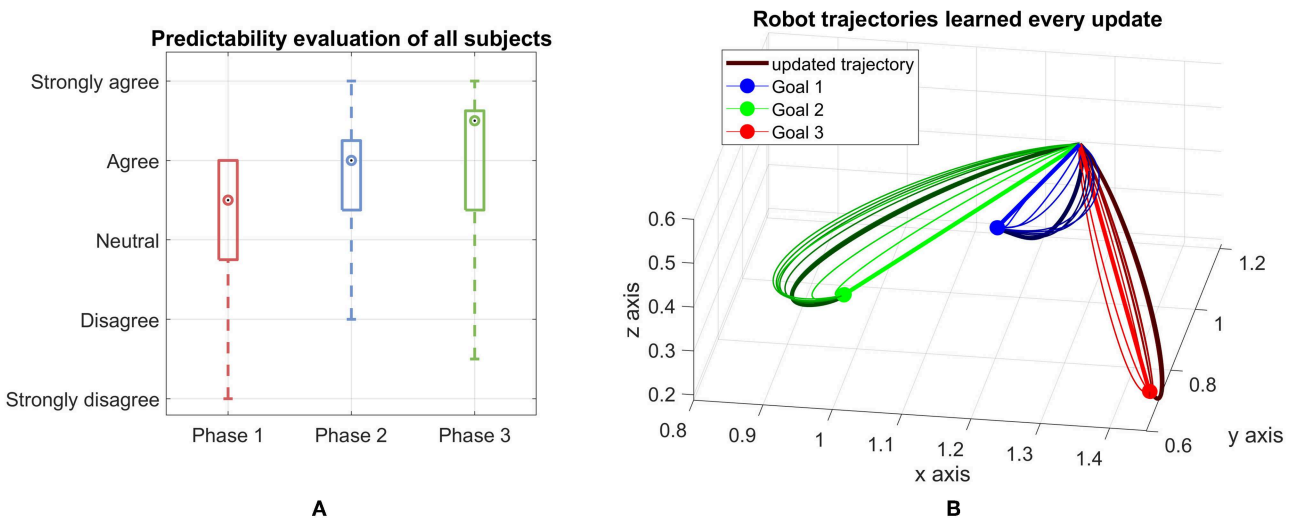



**FIGURE 4 | (A)** Predictability evaluation from all subjects for each phase. The evaluation of participants for the first 4 updates, then the next 3 updates and the last 3 ones are shown in phase 1, phase 2 and phase 3, respectively. The mark shows the median of each group. The box contains 50% of the middle half of all given answers thus representing the interquartile range of the data. The whiskers mark the most extreme answers. **(B)** Robot trajectories learned for each goal every update. The first updates (straight lines) and last updates are marked thicker than the others.

the non-adaptive robot solely reflect the learning capability of the human over time. This configuration also guarantees that the trajectory of the robot is smooth based on the DMP formulation (section 3.1) and the avoidance behavior is identical to the adaptive robot. The only difference between the two robots is the method to generate their motions which can be evaluated by comparing the results from the two experiments.

The experiments are then conducted on 14 new subjects, divided into 2 groups of 7 participants each. The procedure for each experiment is identical to the experiment described in section 5.2.1.

The total cost and human prediction time for both cases, adaptive and non-adaptive robot, are shown in **Figure 5**. The error bar represents the mean value and standard deviation for each update. For the adaptive robot, there is a clear tendency for decreasing in both total cost and human time cost over the course of iterative updates. On average, the total cost deceases around 22% and human time cost decreases around 10%. In the case of non-adaptive robot, these values are 3.8% and 4.5%, respectively. It can also be seen that for the first few updates, the subjects collaborate better with the non-adaptive robot as both of the costs are lower. This is due to the fact that the adaptive robot uses a trial and error method to understand how the human perceives legibility by exploiting different motions. Motions that are harder to predict result in a higher cost, as shown in the slightly increasing in the human time cost on the second and third updates of the adaptive robot. But overtime, its motions become more predictable and easier for the subjects to predict compared to the non-adaptive robot, as indicated by the better performance in both cost values from the sixth update and after.

We also perform pair-sampled $t$-test to evaluate how significantly different is the performance between the adaptive and non-adaptive robot. On the first update, the performance between both robots is not significantly different ($t = -2.464, p > 0.001$ for the total cost $V$ and $t = -0.266, p > 0.001$ for the human prediction time cost $V_{\text{pred}}$). In contrast, on the last update, the $t$-test results in $t = 4.139, p < 0.001$ for $V$ and $t = 3.185, p < 0.001$ for $V_{\text{pred}}$, which indicates that the difference is significant. Overall, our conclusion drawn from this section is that the improvement in the human prediction time and the overall performance is mainly from the legible behavior of the robot. The learning effect from the human partner, while also reducing the human time and cost, does not have a significant contribution within our framework.

### 5.2.3. Predictable Trajectory Evaluation

To analyze the converged trajectories from the policy improvement framework, we first pick three different configurations: 3 goals in a horizontal line, 3 goals in a vertical line and 3 goals in a diagonal line. These configurations are illustrated in **Figure 7**. Combined with two different mounting positions of the robot (same or opposite to the human), we have 6 cases in total. The experiments are conducted with several participants for each case. In **Figure 6** we representatively show 3 converged robot trajectories for each goal configuration.

For the horizontal configuration, **Figures 6A,B** are with the robot on the same side and **Figure 6C** is with the robot on the opposite side. The robot tends to bend more on the left or right for the blue or red goal, respectively, while for the green goal, the robot tries to keep the trajectory in the middle with a small variance, i.e., the green line diverges slightly to the left side in

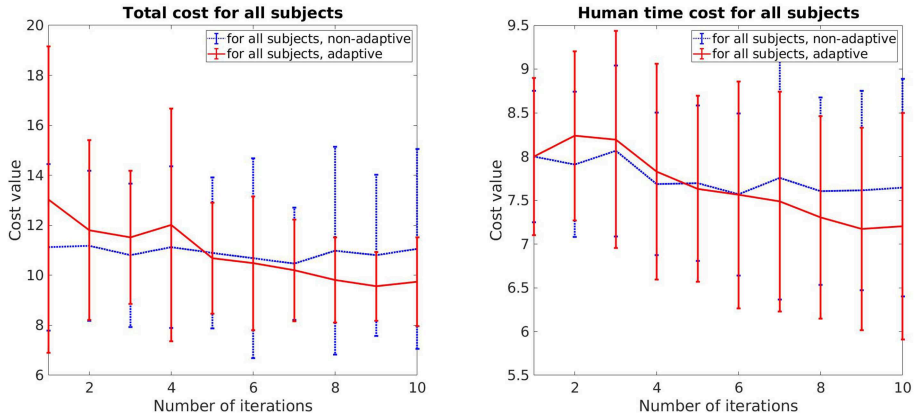



**FIGURE 5 |** Comparison of the total cost and human prediction time between adaptive robot and non-adaptive robot.

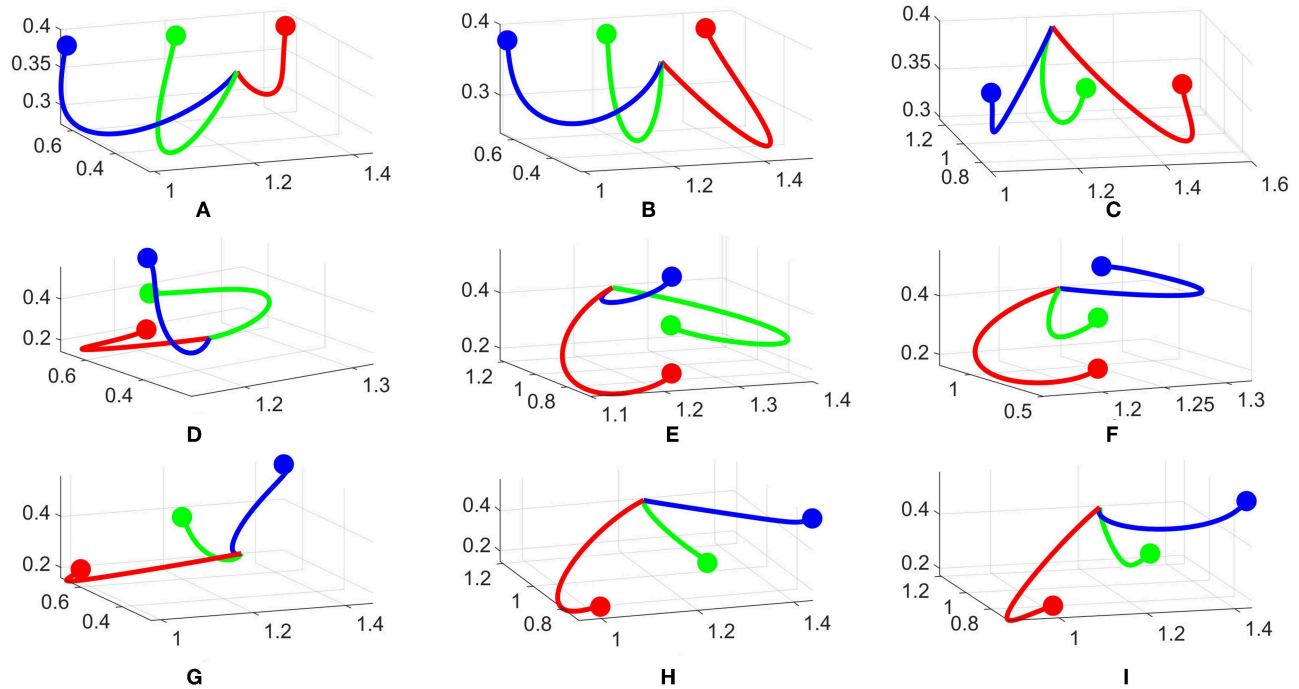



**FIGURE 6 |** Converged trajectories from different subjects and configurations. **(A–C)** Horizontal, **(D–F)** Vertical, **(G–I)** Diagonal configurations.

**Figure 6A**. Another variance is the length of the trajectories, e.g., the red line is the shortest in **Figure 6A** and longest in **Figure 6B**. All trajectories tend to go downward for all three results.

The vertical configuration is one of the most interesting case as the trajectories converge quite differently. For example, the green line curves to the left in **Figures 6D,E** but keeps in the middle-left in **Figure 6F**. The red line is the only one bending to the left in all three results. However, we observe the same pattern for all three results. For each case, one trajectory bends to the left side, one to the right side and one stays in the middle. This creates a divergence between the three trajectories and makes it easier to predict. The difference in trajectory shape toward each goal comes from the random sampling of DMPs during the rollout phase. For example, if there are more rollouts for the green goal to the left side and being predicted correctly by the human, these rollouts will be rewarded more and push the next update to the left. Another reason is the personal preference of each participant, i.e., for the blue goal, it is easier for one participant to predict if it bends to the left side, but for another the right side is favorable. Hence, these trajectories are rewarded differently.

For the diagonal configuration, we observe similar behaviors as in the horizontal one. The green line stays in the middle while the two others diverge to the corresponding directions. Also in this configuration, the distance between two goals is larger than previous cases, therefore it is easier for the human to predict in this configuration. The blue line is one example as it tends to go straight toward the goal in **Figure 6H**.

For all configurations, we observed slightly different trajectories w.r.t the mounting position of the robot. It seems the perspective affects the shape, but it's not always significant. This is probably because from the human point of view, the shape of trajectories does not change a lot, therefore it does not affect the predictability too much.

In summary, there are differences between trajectories w.r.t different subjects and configurations i.e., length, bending angle, etc. However, we also observed several similarities and patterns in the robot trajectories that make them become more predictable to the human. This motivates us to learn these patterns such that they can be generalized to other cases.

## 5.3. Task Generalization Evaluation

As learning a policy for each task and each configuration requires considerable amount of time, it is preferable to take advantage of the knowledge of the prior polices as it already encodes some preference of human perception. In this section, we evaluate our task generalization presented in section 4. To generalize the policy for task $\boldsymbol{g}_m$ in a new set $\widetilde{\mathcal{T}}$, we have to find a set of features $\boldsymbol{\chi}_{\widetilde{\mathcal{T}}}(\boldsymbol{g}_m)$ (see section 4). From our observation and from the results in section 5.2.3, we identified some critical features that a predictable trajectory depends on:

- The relative distances from the target goal $\boldsymbol{g}_m$ to other goals in $\widetilde{\mathcal{T}}$.
- The angles between the target goal $\boldsymbol{g}_m$ to other goals in $\widetilde{\mathcal{T}}$ w.r.t the horizontal line.
- The relative angle between the human and the robot.

Without loss of generality, we illustrate our idea for the case $\widetilde{\mathcal{T}}$ consisting of 3 goals as depicted in **Figure 7**. The workspace of the robot is divided into a $3 \times 3$ lattice where robot goals can be located in 9 different positions. For the sake of simplicity, the height of the workspace is normalized as 1. **Figure 7** depicts some possible configurations and how $\boldsymbol{\chi}_{\widetilde{\mathcal{T}}}(\boldsymbol{g}_m)$ is calculated. For example, for *G1* in **Figure 7A**, the relative distances to *G2* and *G3* are 0.5 and 1, respectively, the angles to *G2* and *G3* are both $0^\circ$. For *G2* in **Figure 7B**, the angles are $90^\circ$ and $-90^\circ$ while for the same *G2* in **Figure 7C**, these values are $45^\circ$ and $-135^\circ$. The relative angles between the human and robot is set $0^\circ$ if the robot in mounted on the same side with the human and $180^\circ$ if the robot is mounted on the opposite side of the human. Within the scope of this work, we only investigate these two mounted positions of the robot, but it can be extended to other cases, e.g. the robot is positioned on one side of the table such that the perspectives of the human and robot are orthogonal.

To verify our task generalization approach, three configurations **Figures 7A–C** combined with two different robot positions are used for the training phase (6 different cases in total). The training phase consists of 18 subjects, equally distributed for all cases. For each experiment, we obtain the policy w.r.t each subject for each case. The weights of the converged trajectories are extracted to construct a regression model. Then, we use a new setup depicted in **Figure 7D** with the robot mounted on the same side with the human as a testing sample. Using the corresponding features for the new setup as the input, we initialize the DMP with the output of the regression function in Equation (23).

The robot trajectories in the first and final update are depicted in **Figure 8**. The trajectories are initialized as curves toward the three goals in the first update instead of straight lines in the non-trained case. For *G1*, the curve bends upward while for *G2* and *G3*, the curves deviate downward, more to the left and right from the human point of view, respectively. These behaviors match the expectation that we observed in section 5.2.3. During the updates, the robot continues exploring new motions around the initial ones. The covariance size of the DMP perturbation is set to half of the value of the non-trained case so that the rollout trajectories are sampled in a smaller area. The converged trajectories for each goal are shown in the final update in **Figure 8**. Compared to the first update, the shape of the trajectories does not change a lot, which indicates that the learning algorithm stays close to the minimum from the beginning.

Next, we analyze the outcome of the total cost and the human prediction time. Our goal here is to compare the performance of the learning method to the non-trained case. Therefore, we establish two groups with 6 new participants each:

- Group A: Subjects within this control group interact with the untrained robot on a specific experimental setup different from the ones used for training the data.
- Group B: Subjects interact with the robot, whose trajectories are initialized by the regression model. The experimental setup is identical to the one of Group A.

The experiment procedure is the same as described in section 5.2.1. The human prediction time cost of each subject is also normalized for cross comparison. The means and standard deviations of the total cost and human prediction time cost from both groups are plotted together for comparison (**Figure 9**). A clear improvement of the trained robot can be observed directly from the result as both the total cost and the human prediction time cost are lower than the untrained robot. In addition, the cost values of the trained robot start decreasing from the start while in the case of untrained robot, they start increasing at first then decrease due to high exploration in the beginning. As the experiment is designed exactly the same between both groups, the improvement of the trained robot comes from the initialized trajectories derived from our task generation approach. Instead of exploring the whole area, the trained robot only needs to search around the given trajectories, which inherit the properties of legibility from training data. As a result, the human predicts easier and faster over time, i.e., the human time cost drops substantially 20% in the case of the trained robot compare to 10% of the untrained robot after 10 updates.

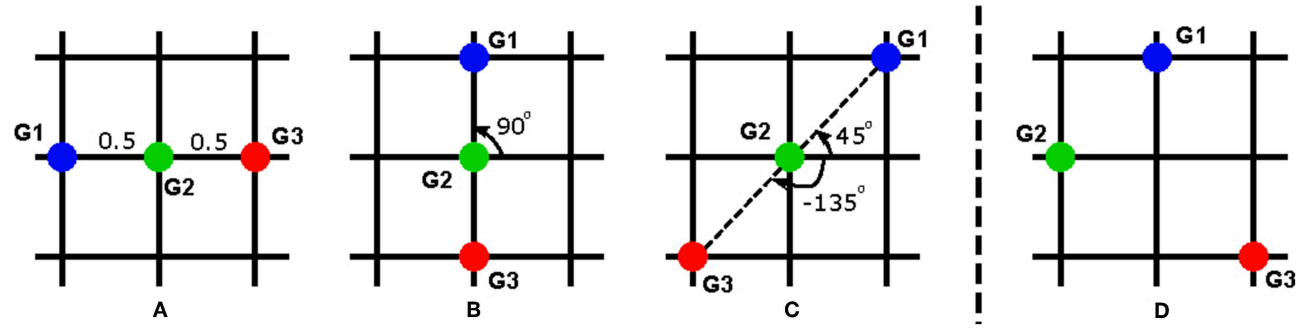



**FIGURE 7 |** Different configurations of the robot goals: **(A–C)** are used for training, **(D)** is used for testing the task generalization approach.

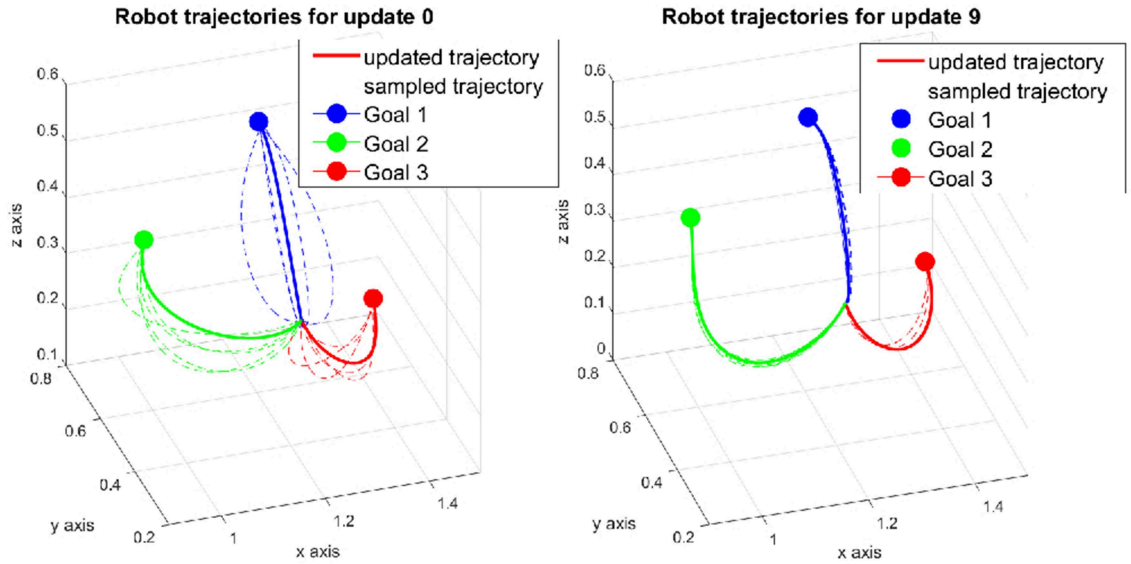



**FIGURE 8 |** Robot trajectories in the task generalization experiment. The trajectories were initialized by the weights generated from the regression model.

As a conclusion, the task generalization approach that we proposed increases the efficiency of the learning framework. Starting from an initial trajectory generated from the approach, the robot trajectory converges quickly to the predictable one, which is also close to the initial trajectory. This helps to reduce the number of updates and the number of sampled trajectories per update, which in turn reduces the amount of time needed for training.

## 5.4. Experimental Results on a Real Robot

As shown in previous sections, our approach is efficient in learning predictable motions for the robot through interaction in VR. We take one step further and bring our framework into a real robot. While performing the experiments in VR allows us to evaluate our hypotheses in different setups and configurations without the need to account for the system limits, safety, etc. in the performance, it is difficult to judge the safety aspect from the human perspective since there is no real collision possibility during the experiment. Therefore, the safety aspect is additionally evaluated in this section. For this purpose, we design the experiment as illustrated in **Figure 10** with the robot on the opposite side of the human. The robot used in this experiment is the KUKA LWR 4+ which has 7 degrees of freedom. The same inverse kinematics introduced in section 5.1 are applied to convert the Cartesian position to joint configuration for the robot. The trajectories generated from our framework are sent to the robot via ROS (Robot Operating System) at the frequency of 100$Hz$. The KUKA robot uses the joint position control internally to keep track of the sent trajectories. Slightly different from the setup in VR, here the goals of the human and robot are chosen to be the same and are constructed in the form of three LEGO blocks (red, blue, and yellow). With this configuration, the human needs to enter the robot workspace to reach the goals and therefore triggers the possibility of collision at every movement. The human hand and robot end effector are equipped with passive retroreflective markers which are tracked by a Qualisys tracking system. This information is then used by the robot to avoid the human and provide safety during the experiment.

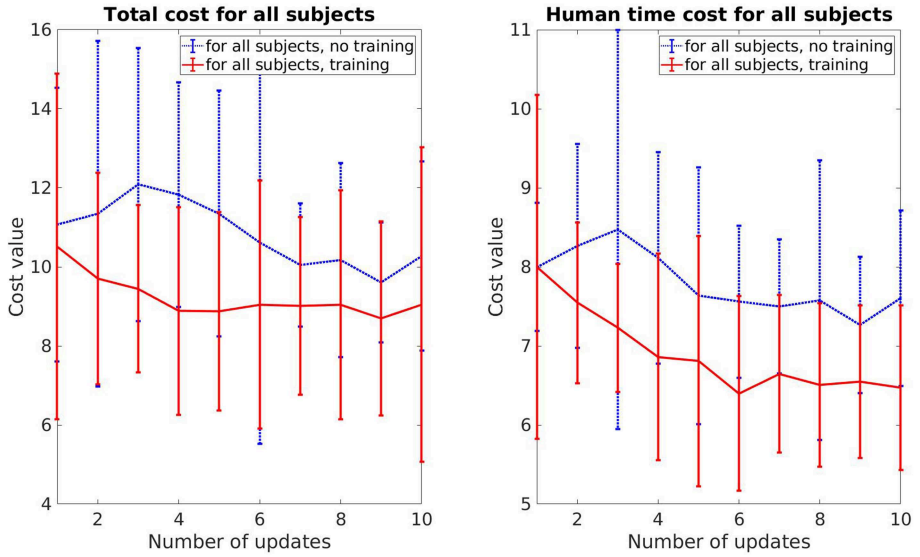



**FIGURE 9 |** Cost plots that show the difference between the control group that interacted with the untrained robot and the results for the interaction with the trained robot.

The experiment procedure is then designed identically to previous sections with a habituation phase and three main blocks in the evaluation phase. The first block contains 4 updates while the second and third one contain 3 updates each. After each block, there is a short break for the subject to answer a questionnaire. The questionnaire is designed similarly to section 5.2.1 with the same questions about the legibility of robot motions. Additionally, new questions are added to evaluate the safety aspect and comfort of the participants. For safety, we asked the participants' opinions about three statements: *The robot is responsive to my movement*, *The robot does not hit me while moving* and *I feel safe working with the robot*. The first two statements focus on the avoidance behavior of the robot since this is the key feature to provide safety for the human. The last statement is a direct question to the participants if they feel safe when working with the robot. The average of three answers is used as the measurement for safety aspect. Similarly, for comfort, two statements were asked: *The motion of the robot is natural to me* and *I feel comfortable working with the robot*. Here we want to evaluate if our framework also provides comfort to the human partner. The experiment lasts around 30 min in total. During the experiment, the participants are asked to wear a headphone with concentration music so that they do not get distracted by the surrounding environment.

We collect data from 10 new participants who have not participated in or known about the VR experiments. The results therefore only reflect the performance of the real robot. Regarding the total cost $V$ and the human prediction time cost $V_{\text{pred}}$, we observe similar patterns as in VR experiments. Both cost values decreases over time (**Figure 11**). In this case, the total cost $V$ drops around 20% and for the human prediction time cost $V_{\text{pred}}$, the drop is around 19%. The improvement in the cost

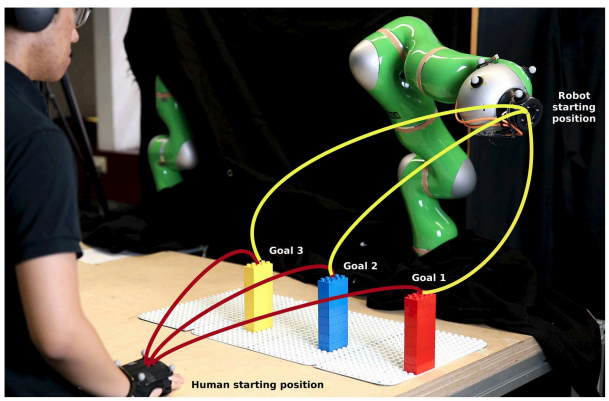



**FIGURE 10 |** Real experiment setup on a KUKA LWR 4+ robot.

values indicates that the trajectories of the KUKA robot is more predictable over time.

The bottom side of **Figure 11** shows the evaluation of safety and comfort aspects in box-plot. In case of safety, there is almost no negative answer from the participants as the data spreads only from *neutral* to *strongly agree* in all three phases of the experiment. The boxes, which contains 50% of the answers spread around *agree* level in phase 1 and phase 2. In phase 3, there is a larger variation since the box spreads from above *neutral* to *strongly agree*. In general, the data shows positive feedback which means the participants are confident that the robot will not hit them while moving and therefore they feel safe when working with the robot. Some participants, that we observed that during the experiment, even show their interest in the behavior of the robot by repetitively interacting with the robot after finishing their task (they keep moving their hand toward the robot to

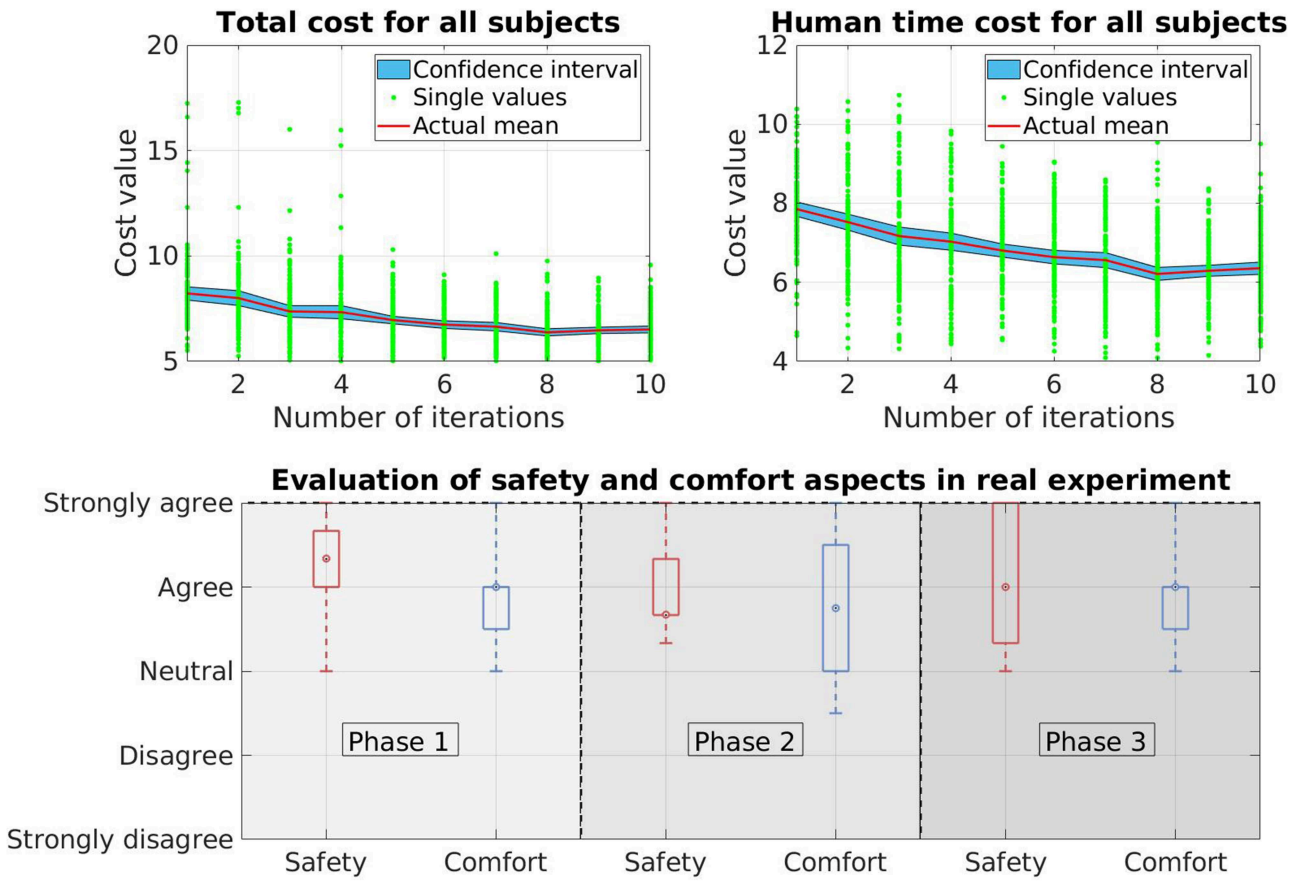



**FIGURE 11 |** Results of the real experiment on a KUKA LWR 4+ robot. The top side of the figure shows the total cost and human prediction time cost while the bottom side shows the human evaluation in the safety and comfort aspects when working with the real robot.

see how the robot reacts to their movement). For comfort, the participants also give positive feedback as most of the answers are above *neutral* level. Only in phase 2, one of the whiskers stays below *neutral* level. However this is an extreme case (1 out of 10 subjects) which also reflects the difference in subject's personality. Overall, we can conclude that our framework is able to provide a safe and comfortable environment for the interaction between human and robot during the learning process.

## 6. DISCUSSION

Our learning framework is a framework that combines learning and interaction into one. By ensuring safety for the human partner, we are able to change from "learning from observation" to "learning through interaction." The results in sections 5.2.1 and 5.2.2 show that our framework is able to generate motions that are legible to the human partner during interaction. A substantial improvement compared to the non-adaptive baseline also points out that the robot motion is more legible over time due to its own adaption and the learning effect from human does not play a significant role during the learning process. We also present some preliminary results in our task generalization approach. We first learn the policies of three sampled tasks and use our approach to generate the policy for a new one. Results presented in section 5.3 indicate that the robot initialized with this policy achieves a better performance. This confirms our hypothesis that legibility can also be transferred to similar tasks and our framework therefore is generalizable using our task generalization method. We also verify our framework in a real experiment setup and show that it is able to provide a safe environment for the human partner. Even though the results that we presented show the effectiveness of our framework, there are some other aspects that we want to discuss in detail.

In our study, we evaluate and verify different hypotheses as presented in section 5. Beside that, there are also other case studies that are worth investigating in further experiments. One case study that is interesting to further investigate is how the predictable trajectories learned from the framework are affected by the relative perspective of the human and robot. The motivation of this study comes from the fact that the human partner usually does not stay at a fixed position, but rather goes around when working with the robot. Therefore the robot trajectories also change from the human point of view. In our work, two mounting positions of the robot were evaluated and we obtained some preliminary results. However, further positions need to be investigated to justify this proposition.

Another case study is about the variation in perception of different types of participants, e.g., participants who have robotics background behave and react differently when working with the robot compared to others who do not have robotics background. Comparing the outcomes of the learning framework from these types of participants requires further inspection but might lead to interesting results.

Task generalization is a concept to estimate the policy for a new task from the existing policies of the prior trained tasks by exploiting the relation between human perception in term of predicting robot trajectories and task specifications. As a result, for a new task, the robot starts from a trajectory that is more predictable to the human and therefore the convergence rate of the learning framework is improved. We demonstrated our idea in a $3 \times 3$ lattice environment with 3 tasks for the robot per configuration and showed the effectiveness of the approach. The advantages of our method are: First, it does not require the exact positions of tasks but only the relative positions between them as we only estimate the basis functions of the DMP; second, it can be extended to an $n \times n$ case with larger number of tasks per configuration without lots of modifications. However, since the features that specify the differences of tasks are defined from the start and do not change during the learning phase, the variation of new tasks whose policies can be estimated by our approach are limited. The reason is these new tasks need to be described using the same features. For example, in our work, all tasks or the robot are reaching a goal on a vertical plane.

With the promising outcome of the task generalization method, there are some consequent open questions that are worth investigating further. The first question is how to identify features and how to qualify the influence of each feature to the trajectories of the robot. In this work, we did it mainly by observing from a certain number of participants and identifying some critical features. However, more data is required to properly justify these features. Another interesting question is how many cases are needed for the training phase of the task generalization approach and how to select these cases such that it comprises enough information about the interrelation between tasks. Too many training cases requires lots of training time, thus reduces the efficiency of the approach. But too few training cases might not contain enough variation, therefore affect the outcome of the generalization method.

Finally, the experiment on the KUKA LWR 4+ robot is our first step to bring our learning framework to reality. The avoidance behavior of the robot is reliable such that the human feels safe and confident to cooperate with the robot. Here, we want to emphasize the importance of this avoidance behavior and its contribution on the success of the learning process since it allows a smooth and consistent behavior from the human partner in term of prediction and hand movement. One example is that in case of collision during the experiment, the human would feel uncomfortable and hesitant to do the next movements, which may lead to inaccurate measurement of the human prediction time. Beside that, one limitation in our setup on the KUKA LWR 4+ is the working area of the robot. Due to the joint limits of the robot (especially the elbow), the mounting position and our configuration to avoid singularity, the workspace of the robot is quite small as we can only setup 3 goals with the distance between them being around 20 cm. As a result, it is difficult to extend the framework to different tasks and evaluate the task generalization method in a real setup. A solution for this is to change the mounting position i.e., mount the robot on the ceiling to have a larger range on the elbow or use a different robot with larger working space.

## 7. CONCLUSION

In this work, a framework is developed to generate predictable robot motion that can adapt to human preferences and can avoid dynamic obstacles, which in our case is the human hand during interaction. The experiments that were conducted show that robots are able to adapt their behavior to human preferences. They can learn to become more predictable while still giving humans the freedom to move safely in the same work space. The humans became faster and more confident in their predictions. Furthermore, a task generalization approach is also developed and tested. In our experiment, the learned policy produces better results in the new task than the control group without a pre-learned policy. This confirms our hypothesis that the policy learned by this framework is indeed transferable to other tasks.

## DATA AVAILABILITY

No datasets were generated or analyzed for this study.

## ETHICS STATEMENT

The user study was approved by the ethics committee of the TUM School of Medicine. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

## AUTHOR CONTRIBUTIONS

KH and OO conceived of the presented idea and developed the framework and formulated the formulations. ME provided inputs for the framework. KH and ME performed the experiments and collected data. OO and DW verified the approach and supervised the experiment. All authors discussed the results and contributed to the final manuscript.

## FUNDING

The research leading to these results has received funding from the Horizon 2020 research and innovation programme under grant agreement no. 820742 of the project HR-Recycler - Hybrid Human-Robot RECYcling plant for electriCal and eLEctRonic equipment.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2019.00069/full#supplementary-material

## REFERENCES


Alami, R., Clodic, A., Montreuil, V., Sisbot, E. A., and Chatila, R. (2005). "Task planning for human-robot interaction," in *Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-aware Services: Usages and Technologies, sOc-EUSAI '05* (New York, NY: ACM), 81–85.

Bortot, D., Born, M., and Bengler, K. (2013). "Directly or on detours? How should industrial robots approximate humans?," in *International Conference on Human-Robot Interaction* (Tokyo).

Busch, B., Grizou, J., Lopes, M., and Stulp, F. (2017). Learning legible motion from human-robot interactions. *Int. J. Soc. Robot.* 9, 765–779. doi: 10.1007/s12369-017-0400-4

Dautenhahn, K., Woods, S., Kaouri, C., Walters, M., Koay, K., and Werry, I. (2005). "What is a robot companion - friend, assistant or butler?," in *International Conference on Intelligent Robots and Systems* (Edmonton, AB).

Dautenhahn, K., Woods, S., Kaouri, C., Walters, M. L., Koay, K. L., and Werry, I. (2005). "What is a robot companion - friend, assistant or butler?," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1192–1197.

Dehais, F., Sisbot, E. A., Alami, R., and Causse, M. (2011). Physiological and subjective evaluation of a human robot object hand-over task. *Appl. Ergon.* 42, 785–791. doi: 10.1016/j.apergo.2010.12.005

Dinh, K. H., Oguz, O., Huber, G., Gabler, V., and Wollherr, D. (2015). "An approach to integrate human motion prediction into local obstacle avoidance in close human-robot collaboration," in *2015 IEEE International Workshop on Advanced Robotics and its Social Impacts (ARSO)* (Lyon), 1–6.

Dragan, A., Lee, K., and Srinivasa, S. (2013). "Legibility and predictability of robot motion" in *IEEE International Conference on HRI* (Tokyo).

Dragan, A. D., and Srinivasa, S. S. (2013). *Generating Legible Motion.* Berlin: Robotics: Science and Systems.

Erlhagen, W., Mukovskiy, A., Chersi, F., and Bico, E. (2007). "On the developement of intention understanding for joint action tasks," in *International Conference on Developement and Learning* (London, UK).

Flacco, F., Krger, T., Luca, A. D., and Khatib, O. (2012). "A depth space approach to human-robot collision avoidance" in *2012 IEEE International Conference on Robotics and Automation* (Saint Paul, MN), 338–345.

Hogan, N. (1984). "Impedance control: an approach to manipulation," in *1984 American Control Conference* (San Diego, CA), 304–313.

Ijspeert, A., Nakanishi, J., and Schaal, S. (2002). "Movement imitation with nonlinear dynamical systems in humanoid robots," in *Proceedings of the IEEE International Conference on Robotics and Automation, 2002. ICRA02*, Vol. 2 (Washington, DC), 1398–1403.

Khatib, O. (1985). "Real-time obstacle avoidance for manipulators and mobile robots," in *Proceedings. 1985 IEEE International Conference on Robotics and Automation*, Vol. 2 (St. Louis, MO), 500–505.

Khatib, O. (1990). *Real-Time Obstacle Avoidance for Manipulators and Mobile Robots*. New York, NY: Springer.

Kirsch, A., Thibault, K., Sisbot, A., Alami, R., Lawitzky, M., Hirche, S., et al. (2010). *Plan-Based Control of Joint Human-Robot Activities,* Vol. 24. KI - Künstliche Intelligenz (Springer), 223–231.

Koay, K., Sisbot, E., Syrdal, D., Walters, M., Dautenhahn, K., and Alami, R. (2007). "Exploratory study of a robot approaching a person in the context of handing over an object," n *AAAI Spring Symposium: Multidisciplinary Collaboration for Socially Assistive Robotics* (Stanford, CA), 18–24.

Lasota, P. A., Fong, T., and Shah, J. A. (2017). A survey of methods for safe human-robot interaction. *Found. Trends Robot.* 5, 261–349. doi: 10.1561/2300000052

Lazaric, A., and Ghavamzadeh, M. (2010). "Bayesian multi-task reinforcement learning," in *ICML - 27th International Conference on Machine Learning* (Haifa: Omnipress), 599–606.

Lichtenthäler, C., and Kirsch, A. (2016). *Legibility of Robot Behavior: A Literature Review.*

Lichtenthäler, C., Lorenz, T., and Kirsch, A. (2011). "Towards a legibility metric: how to measure the perceived value of a robot," in *International Conference on Social Robotics, ICSR 2011* (Amsterdam).

Oguz, O., Sari, O., Dinh, K., and Wollherr, D. (2017). "Progressive stochastic motion planning for human-robot interaction," in *IEEE International Symposium on Robot and Human Interactive Communication* (Lisbon).

Paraschos, A., Daniel, C., Peters, J., and Neumann, G. (2013). "Probabilistic movement primitives," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13* (Nevada: Curran Associates Inc.), 2616–2624.

Park, D., Hoffmann, H., Pastor, P., and Schaal, S. (2008). "Movement reproduction and obstacle avoidance with dynamic movement primitives and potential fields," in *IEEE-RAS International Conference on Humanoid Robots* (Daejeon).

Penrose, R. (1955). A generalized inverse for matrices. *Math. Proc. Camb. Philos. Soc.* 51, 406–413. doi: 10.1017/S03050041000 30401

Robla-Gómez, S., Becerra, V. M., Llata, J. R., Gonzlez-Sarabia, E., Torre-Ferrero, C., and Prez-Oria, J. (2017). Working together: a review on safe human-robot collaboration in industrial environments. *IEEE Access* 5, 26754–26773. doi: 10.1109/ACCESS.2017.2773127

Sisbot, E. A., and Alami, R. (2012). A human-aware manipulation planner. *IEEE Trans. Robot.* 28, 1045–1057. doi: 10.1109/TRO.2012.21 96303

Sisbot, E. A., Clodic, A., Alami, R., and Ransan, M. (2008). "Supervision and motion planning for a mobile manipulator interacting with humans," in *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Amsterdam), 327–334.

Sisbot, E. A., Marin-Urias, L. F., Alami, R., and Simeon, T. (2007). A human aware mobile robot motion planner. *IEEE Trans. Robot.* 23, 874–883. doi: 10.1109/TRO.2007.904911

Sisbot, E. A., Marin-Urias, L. F., Broquère, X., Sidobre, D., and Alami, R. (2010). Synthesizing robot motions adapted to human presence. *Int. J. Soc. Robot.* 2, 329–343. doi: 10.1007/s12369-010-0059-6

Stulp, F., Grizou, J., Busch, B., and Lopes, M. (2015). "Facilitating intention prediction for humans by optimizing robot motions," in *Intelligent Robots and Systems* (Hamburg).

Stulp, F., and Sigaud, O. (2012). Policy improvement methods: between black-box optimization and episodic reinforcement learning. *hal-00738463.*

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nat. Neurosci.* 7, 907–915. doi: 10.1038/nn1309


**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

METHODS
published: 27 September 2019
doi: 10.3389/frobt.2019.00094

# Introducing ACASS: An Annotated Character Animation Stimulus Set for Controlled (e)Motion Perception Studies

*Sebastian Lammers*[1,2]*, *Gary Bente*[3], *Ralf Tepest*[1], *Mathis Jording*[2], *Daniel Roth*[4] and *Kai Vogeley*[1,2]

[1] Department of Psychiatry, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany, [2] Cognitive Neuroscience (INM-3), Institute of Neuroscience and Medicine, Research Center Jülich, Jülich, Germany, [3] Department of Communication, Michigan State University, East Lansing, MI, United States, [4] Human-Computer Interaction, Institute for Computer Science, University of Würzburg, Würzburg, Germany

OPEN ACCESS

**Edited by:**
Agnieszka Wykowska,
Italian Institute of Technology, Italy

**Reviewed by:**
Ulysses Bernardet,
Aston University, United Kingdom
Ruud Hortensius,
University of Glasgow,
United Kingdom

***Correspondence:**
Sebastian Lammers
sebastian.lammers@uk-koeln.de

**Specialty section:**
This article was submitted to Humanoid Robotics, a section of the journal Frontiers in Robotics and AI

**Received:** 23 February 2019
**Accepted:** 13 September 2019
**Published:** 27 September 2019

**Citation:**
Lammers S, Bente G, Tepest R, Jording M, Roth D and Vogeley K (2019) Introducing ACASS: An Annotated Character Animation Stimulus Set for Controlled (e)Motion Perception Studies. Front. Robot. AI 6:94. doi: 10.3389/frobt.2019.00094

Others' movements inform us about their current activities as well as their intentions and emotions. Research on the distinct mechanisms underlying action recognition and emotion inferences has been limited due to a lack of suitable comparative stimulus material. Problematic confounds can derive from low-level physical features (e.g., luminance), as well as from higher-level psychological features (e.g., stimulus difficulty). Here we present a standardized stimulus dataset, which allows to address both action and emotion recognition with identical stimuli. The stimulus set consists of 792 computer animations with a neutral avatar based on full body motion capture protocols. Motion capture was performed on 22 human volunteers, instructed to perform six everyday activities (mopping, sweeping, painting with a roller, painting with a brush, wiping, sanding) in three different moods (angry, happy, sad). Five-second clips of each motion protocol were rendered into AVI-files using two virtual camera perspectives for each clip. In contrast to video stimuli, the computer animations allowed to standardize the physical appearance of the avatar and to control lighting and coloring conditions, thus reducing the stimulus variation to mere movement. To control for low level optical features of the stimuli, we developed and applied a set of MATLAB routines extracting basic physical features of the stimuli, including average background-foreground proportion and frame-by-frame pixel change dynamics. This information was used to identify outliers and to homogenize the stimuli across action and emotion categories. This led to a smaller stimulus subset ($n = 83$ animations within the 792 clip database) which only contained two different actions (mopping, sweeping) and two different moods (angry, happy). To further homogenize this stimulus subset with regard to psychological criteria we conducted an online observer study ($N = 112$ participants) to assess the recognition rates for actions and moods, which led to a final sub-selection of 32 clips (eight per

category) within the database. The ACASS database and its subsets provide unique opportunities for research applications in social psychology, social neuroscience, and applied clinical studies on communication disorders. All 792 AVI-files, selected subsets, MATLAB code, annotations, and motion capture data (FBX-files) are available online.

**Keywords: body motion, experimental paradigms, human interaction, motion capture, non-verbal behavior, social cognition, visual stimuli**

## INTRODUCTION

Observations of others' movements provide important information about our social environment. Not only do movements tell us what people are doing or what they intend to do (Dittrich, 1993; Thompson and Parasuraman, 2012; Cavallo et al., 2016), they also build the basis for far-reaching inferences about others' motivational states, moods, and emotions (Atkinson et al., 2004; Loula et al., 2005; Chouchourelou et al., 2006; Gross et al., 2012; Barliya et al., 2013). The cognitive mechanisms and the putatively distinct neural mechanisms underlying action recognition on the one hand and emotion inferences on the other hand are not yet fully understood (Vogeley, 2017). A limiting factor in previous studies has been the lack of naturalistic movement stimuli that are free of confounds and allow for high levels of experimental control (cf. Bente, 2019). This is a general requirement in motion perception studies, but particularly crucial for studies in the field of cognitive neuroscience, where distinct stimulus features that are not subject to the experimental variation, can contaminate the observed effects and aggravate their interpretation. Problematic confounds can derive from low-level physical features, such as differences in luminance or pixel changes, as well as from higher-level psychological features, such as differences in the stimulus difficulty and recognition base rates. The demand for internal validity, stands opposite to the quest for ecologically valid social stimuli, which has led to the use of more complex, real-life samples of human behavior, as captured in video documents (Bartels and Zeki, 2004; Hasson et al., 2004; Nishimoto et al., 2011; Lahnakoski et al., 2012; de Borst and de Gelder, 2015). Beyond the mentioned threats to internal validity, the disadvantage of video stimuli, in particular those collected in naturalistic settings, is evident: video documents usually disclose person variables such as age, ethnicity, gender, or attractiveness relevant to stereotypes that might interfere with inferences based on movement (Meadors and Murray, 2014). Further confounds concern the visibility of context, which has been shown to massively influence the recognition of bodily expressions (Kret and de Gelder, 2010). Last but not least, when falling back on existing media content, such as samples from TV shows or movies (Hasson et al., 2004; Spunt and Lieberman, 2012; Schmälzle et al., 2015) there is no way to control any of the visual features and no access to behavioral information of the actors, except through time consuming coding.

Different methods for stimulus production have been proposed to preserve the natural movement dynamics while avoiding the typical issues of video stimuli (cf. Bernieri et al., 1994) such as the use of point light displays (Johansson, 1973, 1976) or video quantization techniques (Berry et al., 1991, 1992). However, both methods come along with specific limitations. Although point-light displays have been shown to carry relevant information for the recognition of intentions (Manera et al., 2010) and emotions (Atkinson et al., 2004; Chouchourelou et al., 2006; Gross et al., 2012; Barliya et al., 2013; von der Lühe et al., 2016) they can only portray movements but not postural patterns (see Cutting and Proffitt, 1981), which also convey relevant emotional information (cf. Aviezer et al., 2012). Quantization techniques used to degrade video images to rougher mosaic patterns are restricted as they cannot completely obscure person characteristics, such as gender and ethnicity (see stimulus examples in Bernieri et al., 1994). These limitations can be overcome by using motion capture technologies and hereon based character animations (cf. Kret and de Gelder, 2010). Such procedures for stimulus production not only allow to systemically vary or obscure aspects of physical appearance (Bente et al., 2008, 2010) but also provide rich datasets to analyze the behavioral variations in the stimuli (Poppe et al., 2014). Importantly, we could show that character animations (lacking several visible features) produce similar impressions as videos of the original human movement they are based on (Bente et al., 2001a,b).

A setback of motion capture and character animation applications can be seen in the time consuming production process including marker application and calibration and particularly the labor intense post-production to clear the motion data from measurement artifacts and jitter before rendering. To protect these considerable investments it is reasonable to produce and publish larger stimulus data sets for multiple (re-)use. Ideally, these stimulus sets should contain annotations of low-level and high-level stimulus features, which allow other researchers to select stimulus subsets tailored to their specific research questions and methodological requirements. This is particularly true for brain imaging studies that might require the control of physical stimulus features such as brightness, contrast or pixel change dynamics in order to avoid contaminations of low-level sensory effects and high-level inferential processes. We here introduce such an annotated stimulus database suitable for the study of action recognition and emotion inferences in social perception research and social neuroscience.

Motion capture was performed on 22 human volunteers, instructed to perform six everyday activities (mopping, sweeping, painting with a roller, painting with a brush, wiping, sanding) in three different moods (angry, happy, sad; see **Table 1**). The six activities were chosen to be recognizable for the majority of viewers without specific expertise in contrast to movements

**TABLE 1 |** Activities and moods recorded in the motion capture setup.

| Activities | | | Moods |
|---|---|---|---|
| 1. Mopping | 2. Sweeping | | 1. Happy |
| 3. Wiping a table with a rag | 4. Sanding a piece of wood on a table | X | 2. Angry |
| 5. Painting a wall with a brush | 6. Painting a wall with a roller | | 3. Sad |

*All six activities were performed in three designated recording blocks for each mood.*

requiring expert knowledge (e.g., particular dancing styles). Five-second clips of each motion protocol were rendered into AVI-files using two virtual camera perspectives for each clip, yielding a set of 792 stimuli. Based on this, we identified an exemplary subset of clips controlled for low- and high-level confounds: By applying a MATLAB routine for feature extraction we identified a subset of 83 clips free of outliers and characterized by maximal similarity of low-level physical stimulus features across actions and moods (see **Figure 1** for an overview). In the next step we conducted an online observer study to obtain recognition rates for action and emotion which could serve as high-level psychological selection criteria for stimulus sets. Applying this data to further homogenize the stimulus set we ended with a fully balanced subset of 32 animation clips (eight variations of each of four possible combinations: two actions × two moods). This specific subset was prepared for a particular fMRI study that focused on the differential activation of the action observation network and the mentalizing system (also called theory of mind system) as related to action and emotion recognition (Geiger et al., 2019).

The current article introduces the ACASS database (**A**nnotated **C**haracter **A**nimation **S**timulus **S**et) and reports the details of stimulus generation, the algorithm used for feature extraction, as well as the exemplary stepwise stimulus selection procedure leading to the subset(s). The publication includes the complete database including all animations ($N = 792$) annotated with low-level features along with two subsets: (a) with additional recognition rate annotation ($n = 83$ animations) and (b) selected for maximum homogenous and balanced properties ($n = 32$ animations). Additionally, we provide the 3D data (FBX-files, $N = 396$). Readers interested in existing motion capture databases can refer to **Table 2** and the respective publications mentioned therein.

# STIMULUS DATABASE

## Performers

We recruited 31 volunteers (17 females, mean age = 25.55, $SD$ = 6.01) via (a) mailing lists of the study programs Psychology and Neuroscience of the University of Cologne, (b) word of mouth or (c) publicly visible notices. The volunteers which participated in the study to produce motion capture data will in the following be called "performers." Four performers were excluded due to technical issues. Five other performers were excluded because they stated that they did not empathize sufficiently with the demanded moods during the procedure (see section Instructions and Recording-Procedures for details), resulting in a total sample of $n = 22$ (12 females, mean age = 24.73, $SD = 4.84$).

All performers were informed about the scientific background of the envisaged use of their motion capture recordings as stimulus material and gave informed consent prior to participation. All performers were either compensated monetarily (15€) or with credits for participation. Procedures were approved by the ethics committee of the Medical Faculty of the University of Cologne.

## Instructions and Recording-Procedures

All performers filled out a questionnaire via computer which included basic demographic information, as well as the following psychological traits (see **Supplementary Data Sheet 1**): a short version of the *Big Five Inventory* (Rammstedt and John, 2007), the *Toronto Empathy Questionnaire* (Spreng et al., 2009), and the *Emotional Intelligence Scale* (Schutte et al., 1998). Correlations between these traits and the subsequent recognition rates (see section Homogenizing for Recognition Rates) showed that the personality traits of the performers have no significant influence on the subsequent recognition rates when presenting the animations to naïve volunteers (Lammers, 2017).

We selected six everyday household activities (mopping, sweeping, painting with a roller, painting with a brush, wiping a table, sanding a piece of wood) in combination with three moods (angry, happy, sad; see **Table 1**) to yield animations that contain information about a specific activity (*What is the person doing?*) and at the same time about the underlying mood that the person was in (*How is the person doing it?*). The six activities can be separated in three domains (floor, table, wall) with two pairs of actions each. For instance, sweeping and mopping (floor) are not too easily differentiated when shown as wooden mannequin without the used tool (see **Figure 2A**).

Each volunteer performed all activities in combination with the different moods resulting in 18 recordings per performer (see **Table 1**). To ensure that the performers execute the different movements naturally while displaying the different moods, we used the following *mood induction procedure*. Specific instructions were presented as audio recordings to which the volunteers listened before each of the 18 recordings. Mood induction was achieved by an *Imagination Mood Induction Procedure*, which is considered to be one of the most effective ways to induce different moods (for a meta-analysis on mood induction procedures, see Westermann et al., 1996; a transcript of the instructions is provided in **Supplementary Table 2**).

The recordings were organized in three recording blocks according to the moods: angry, happy, and sad. The order of the three moods was randomized for each performer, while the order of activities remained the same in all three blocks. To control for immersion of the volunteers into the different moods, the performers' level of immersion into the demanded mood was assessed after each recording block via a Likert scale (*How well were you able to empathize with the required feeling?; German: Wie gut konnten Sie das von Ihnen geforderte Gefühl nachempfinden?*) ranging from 1 (*not at all*) to 11 (*very well*). The mean level of immersion was 9.197 (*SD* 1.184). Performers' data as a whole were excluded from further processing if they responded with a

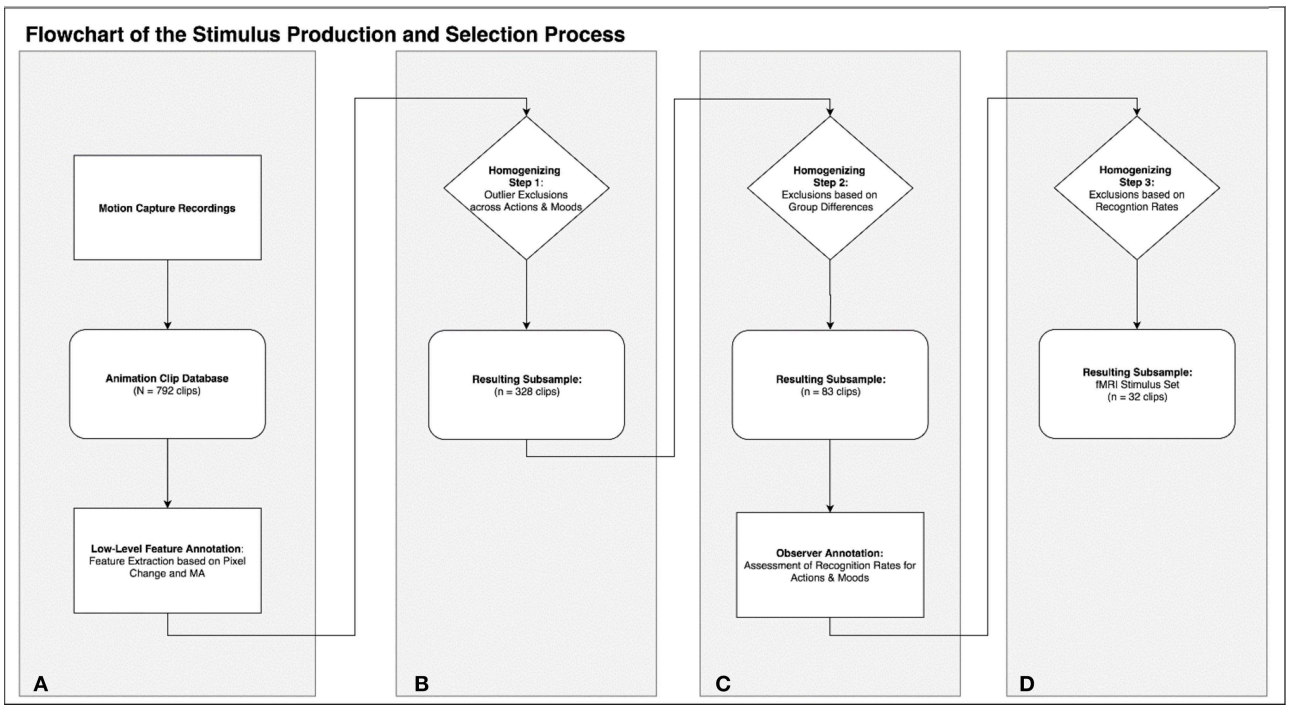



**FIGURE 1 |** This flowchart summarizes the rationale and process of stimulus production **(A)**, annotation **(A,C)**, and selection of the stimulus subsets **(B–D)**.

value equal to or smaller than five for any of the recording blocks to ensure sufficiently mood-influenced movements. Additionally, performers were asked to briefly describe the situation(s), which they imagined in order to immerse into the different moods. Directly before the next recording block they were presented with a 90 s relaxation-video (showing a tree with relaxing background music) to neutralize the mood.

## Technical Setup and Processing

The movements were recorded using an optical motion capture system with 16 infrared cameras (frame rate = 100 Hz) and the Motive Software (Optitrack™, NaturalPoint, Inc., Oregon, USA). After recordings, the 3D-data were processed and rendered using MotionBuilder® and Maya® (Autodesk Inc., California, USA) to retarget the human movements onto a virtual character in a virtual scene. We used a virtual character on a black background that looked like a wooden mannequin without a face, with detectable gross hand movements but without visibility of the fingers and the used tools (see **Figure 2A**).

Light sources and virtual cameras were added to all recordings in an identical fashion to ensure uniform brightness conditions. The virtual cameras defined the perspective (position, orientation, field of view) from which the resulting animation showed the mannequin. We placed two virtual cameras in each virtual scene to render the material from both the left-hand 45 degree angle and the right-hand 45 degree angle from the frontal axis. We chose this angle, because in pretests it achieved the best tradeoff between ecological validity and recognizability compared to other orientations.

From the total recording length of ∼30 s only the first 5 s of the respective action were batch-rendered as PNG-files with the mental ray Plugin for Maya. We decided to use the first 5 s, because we expect the mood to be performed at peak intensity at the beginning of the recording sequence. Using a custom MATLAB script, these image-files were subsequently converted to high definition AVI-files (1280 × 720 pixels) with a frame rate of 25 frames per second.

The rendering resulted in 792 animation clips featuring 22 volunteers performing six everyday household activities in combination with three moods (see **Table 1**).

Additionally we provide the 396 FBX-files that allow the use in virtual reality and to further change camera angles, choose different appearances of the avatar or computations based on the 3D data.

## Low-Level Physical Feature Extraction and Stimulus Annotation

Our aim is to provide solid animation stimuli for research paradigms. As such, we deem it most important to be able to characterize the stimuli that are shown to (future) participants. While the analysis of the motion capture data would yield additional insight about the individual movements, we aimed at specifying details about the stimulus material that is presented to volunteers of future studies. This means that the analysis of the visual features of the AVI-files gains the best insight into what future participants will perceive when confronted with the stimuli.

**TABLE 2 |** Existing motion capture databases.

| Name | Publication | Availability |
|---|---|---|
| The Korea University Gesture Database | Hwang, B. W., Kim, S., and Lee, S. W. (2006). A full-body gesture database for automatic gesture recognition. 7th International Conference on Automatic Face and Gesture Recognition (FGR06), 243–248. https://doi.org/10.1109/FGR.2006.8 | Upon request: gesturedb@image.korea.ac.kr |
| The Biological Motion Library | Ma, Y., Paterson, H. M., and Pollick, F. E. (2006). A motion capture library for the study of identity, gender, and emotion perception from biological motion. Behavior Research Methods, 38(1), 134–141. https://doi.org/10.3758/BF03192758 | http://paco.psy.gla.ac.uk/index.php/res/download-data |
| CMU Mocap Database | Not available | http://mocap.cs.cmu.edu |
| HDM05 | Müller, M., Röder, T., Clausen, M., Eberhardt, B., Krüger, B., and Weber, A. (2007). Documentation Mocap Database HDM05 (No. CG-2007-2). Universität Bonn. | http://resources.mpi-inf.mpg.de/HDM05 |
| HMDB | Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). HMDB: A large video database for human motion recognition. 2011 International Conference on Computer Vision, 2556–2563. https://doi.org/10.1109/ICCV.2011.6126543 | http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database |
| ICS Action Database | Not available | Upon request: tmori@ics.t.u-tokyo.ac.jp Overview: http://www.miubiq.cs.titech.ac.jp/action/index.html |
| IEMOCAP | Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Narayanan, S. S. (2008). IEMOCAP: interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4), 335. https://doi.org/10.1007/s10579-008-9076-6 | Upon request: https://sail.usc.edu/iemocap/release_form.php |
| GEMEP Corpus | Bänziger, T., Mortillaro, M., and Scherer, K. R. (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. Emotion, 12(5), 1161–1179. https://doi.org/10.1037/a0025827 | Upon request: https://www.unige.ch/cisa/gemep |
| The KIT whole-body human motion database | Mandery, C., Terlemez, O., Do, M., Vahrenkamp, N., and Asfour, T. (2015). The KIT whole-body human motion database. 2015 International Conference on Advanced Robotics (ICAR), 329–336. https://doi.org/10.1109/ICAR.2015.7251476 | https://motion-database.humanoids.kit.edu/ |

*Only databases that were available to the authors are listed here. Databases that have an accompanying article but can no longer be accessed are not listed.*

**TABLE 3 |** Overview of Value Categories Computed by Matlab Algorithm.

| No. | Value category | Description |
|---|---|---|
| 1. | pixelamount | Number of non-black pixels in current frame |
| 2. | intensitydiff | Changes of gray-scale values across time |
| 3. | rel_intensitydiff | Amount of pixels in avatar ÷ sum of intensity differences (1 ÷ 2) |
| 4. | MA_X | The horizontal extension of the motion area |
| 5. | MA_Y | The vertical extension of the motion area |
| 6. | MA_size | MA-X-Dimension × MA-Y-Dimension (4 × 5) |

*To help understand the variable-names in the supplementary spreadsheets, the value categories are named accordingly here. One of the six categories always builds the first part of the variable-name. For each of these six categories, ten values (see **Supplementary Table 1**) were computed, resulting in a total of 60 variables. Example for the variable-name for the mean amount of pixels of a clip: pixelamount_mean. MA, motion area.*

To this end, we developed a special algorithm, which accepts most common video file formats (e.g., AVI, MPEG-1, MPEG-4). The algorithm is implemented and executed in MATLAB (R2017a, The MathWorks, Inc., Natick, USA). The routine performs a frame-by-frame comparison based on 8-bit gray-scale converted images with a black threshold of 30. The resulting signal is filtered with a moving average filter (window size = 5). The algorithm extracts two main features: (a) the size of a "*motion area*" (MA) and (b) differences in pixel intensity (i.e., pixel change). The MA is automatically defined by the 2D-area that the avatar occupies per frame and can be thought of as the smallest possible rectangle encompassing the whole body including the most distal parts (minimum bounding box). Usually these are head and feet, as well as hands, elbows or shoulders (see **Figure 2B** for illustration). The MA gives an impression of the extension of movements (e.g., stretched arms) and the frequency of occurring motion patterns (e.g., back and forth movements). On a more abstract level, the MA measures the size of the area in a given frame that is occupied by non-black pixels (proportion of foreground to background).

Pixel change is computed by comparing the absolute differences of gray values of each pixel frame-by-frame. This allows to infer motion parameters in general, but is particularly interesting for cases when the changes in MA are subtle (e.g., small movements in front of the body). These concepts are based on common approaches, namely *motion energy analysis* (Ramseyer and Tschacher, 2011) and *motion energy detection* (Grammer et al., 1999). The output of the *low-level feature annotation* is structured in 60 variables, with six main categories (**Table 3**) and 10

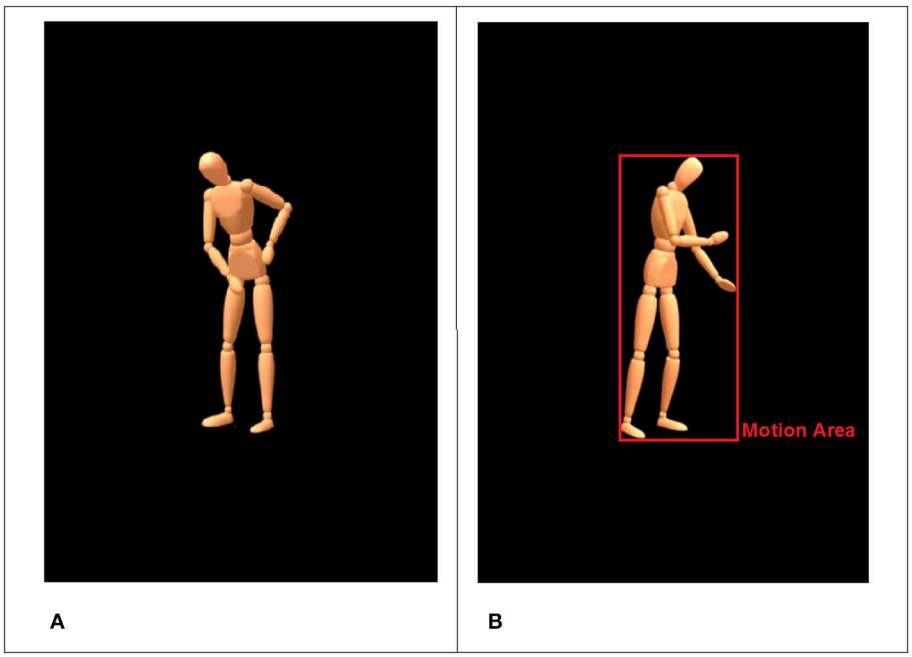



**FIGURE 2 |** Standardized virtual character with blank face used in the animations **(A)**. The red rectangle illustrates the detected motion area for the current frame **(B)**.

values each (see **Supplementary Table 1**). Three of the six categories are centered on pixel change computations (categories 1–3), while the other three reflect characteristics of the MA (categories 4–6). Automated curve sketching is implemented to compare the progression of motion features within and between animation clips (see **Figure 3** for an example). One core element of this procedure is the translation of visible motion features into quantitative properties (e.g., number of maxima; see **Supplementary Table 1**, Values 3–10).

Based on these values we defined *motion frequency* as the number of maxima of the *MA-size*-curve (e.g., how often does the avatar stretch its arms) and *motion expansiveness* as the amplitude of the *MA-size*-curve (e.g., how far does the avatar stretch its arms).

Most of the 60 parameters show weak correlations, however some are inherently connected and thus show strong correlations (e.g., the number of maxima and the mean distance between those maxima; for a graphical representation of correlations between all parameters, see **Figure 4**).

## Resulting Database

The 60 variables resulting from the low-level feature extraction were computed for all 792 animation clips and included in the database metafile (see **Supplementary Data Sheet 2**; see also **Figures 5**, **6** for an overview of all animations across actions and moods).

We used R (R Core Team, 2019), RStudio (RStudio Team, 2018) and the *lme4* package (Bates et al., 2015) to fit generalized linear mixed effects models of the relationship between motion frequency and action, as well as mood. Likelihood ratio tests were used to assess the general influence of predictors, comparing how well models including different predictors fit a given data set while taking into account the models' complexity. The significance of the effect of each predictor was tested by comparing a model including the predictor with the same model without the predictor against a significance level of 0.05.

*Post hoc* tests were computed for the comparison between factor levels (correcting for multiple comparisons) with the *glht()* function from the *multcomp* package (Hothorn et al., 2008). To analyze motion frequency, a model including action and mood (without interaction term) as fixed effects with random intercepts for motion capture performers was fitted and performed significantly better than the null model including only the intercept or models with only one of the fixed effects [$\chi^2_{(2)} = 176.31$, $p < 0.001$].

In *post hoc* tests we found significant differences in the mean motion frequency for sanding vs. wiping ($M = -0.14$, $SE = 0.04$, $z = -3.16$, $p < 0.01$; see also **Figure 7**), but not between the two other pairs of activities. The tests further revealed significant differences in the mean motion frequency between happy and sad movements, $M = -0.19$, $SE = 0.04$, $z = -5.28$, $p < 0.001$, angry and sad movements, $M = -0.44$, $SE = 0.03$, $z = -12.99$, $p < 0.001$ and notably also between happy and angry movements, $M = 0.25$, $SE = 0.03$, $z = 7.84$, $p < 0.001$ (see also **Figure 8**).

# DEFINING STIMULUS SUBSETS

In the following we exemplarily demonstrate a stimulus selection procedure which results in an optimal set to compare neural correlates of action and emotion recognition. This selection is based on the low level video features described above, as well as on an additional annotation based on observer recognition

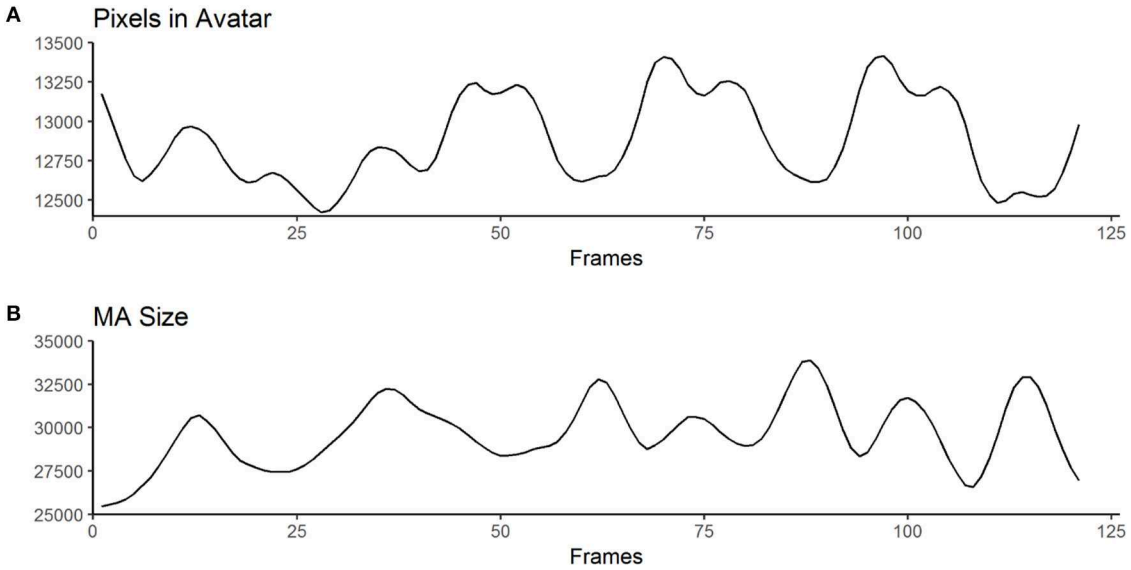



**FIGURE 3 |** Exemplary curves computed from the raw output of the MATLAB algorithm for one animation clip. **(A)** shows an example for "Pixels in Avatar" (categories 1 – 3 in **Table 2**), while **(B)** displays an example for "MA-Size" (categories 4 – 6 in **Table 2**). The trajectories of the curves are used to derive variables such as the number of maxima or the mean amplitude (for a full list of computed variables, see **Table 2** and **Supplementary Table 1**).

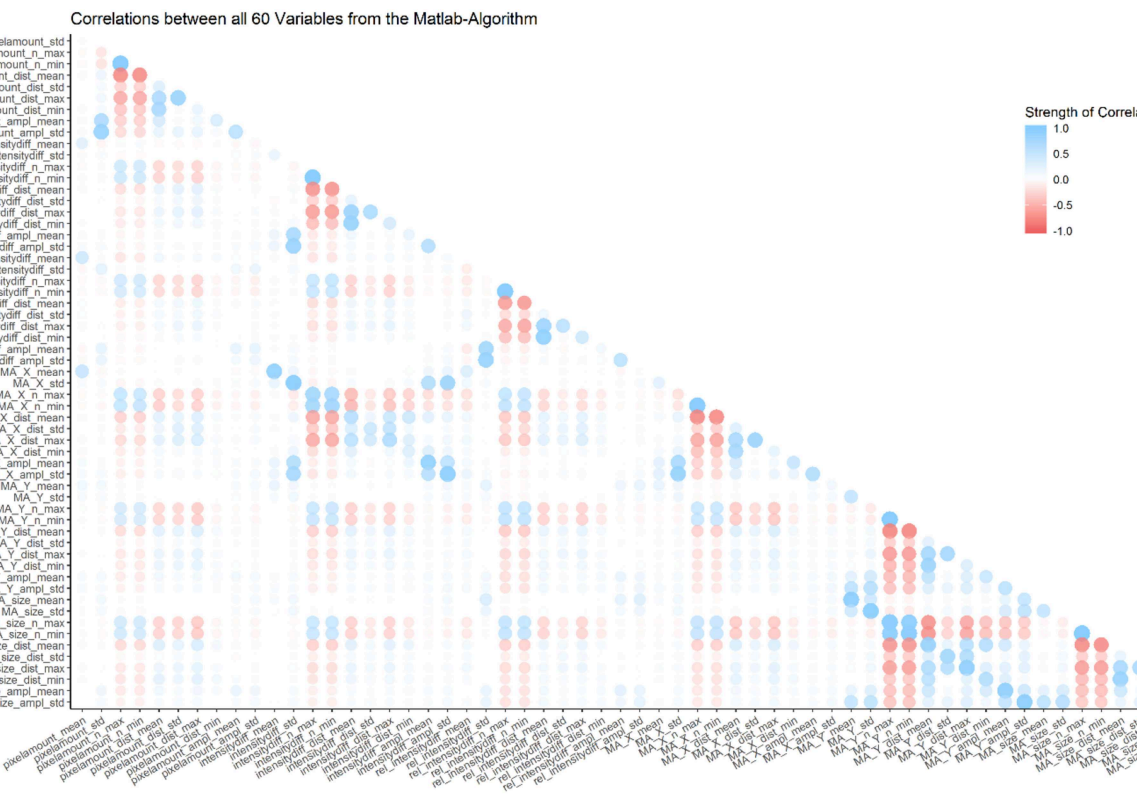



**FIGURE 4 |** Overview of correlations between all 60 variables from the MATLAB algorithm. Most correlations are weak, but some parameters are inherently connected, and thus show strong correlations (e.g., number of minima and number of maxima in the same category).

rates for actions and emotions (see section Homogenizing for Recognition Rates). The procedure comprises three selection steps, which lead to a highly homogenous set of 32 stimuli with eight clips for each of the four different possible combinations (two actions × two emotions; see **Figures 1B–D** for an overview of the selection procedure).

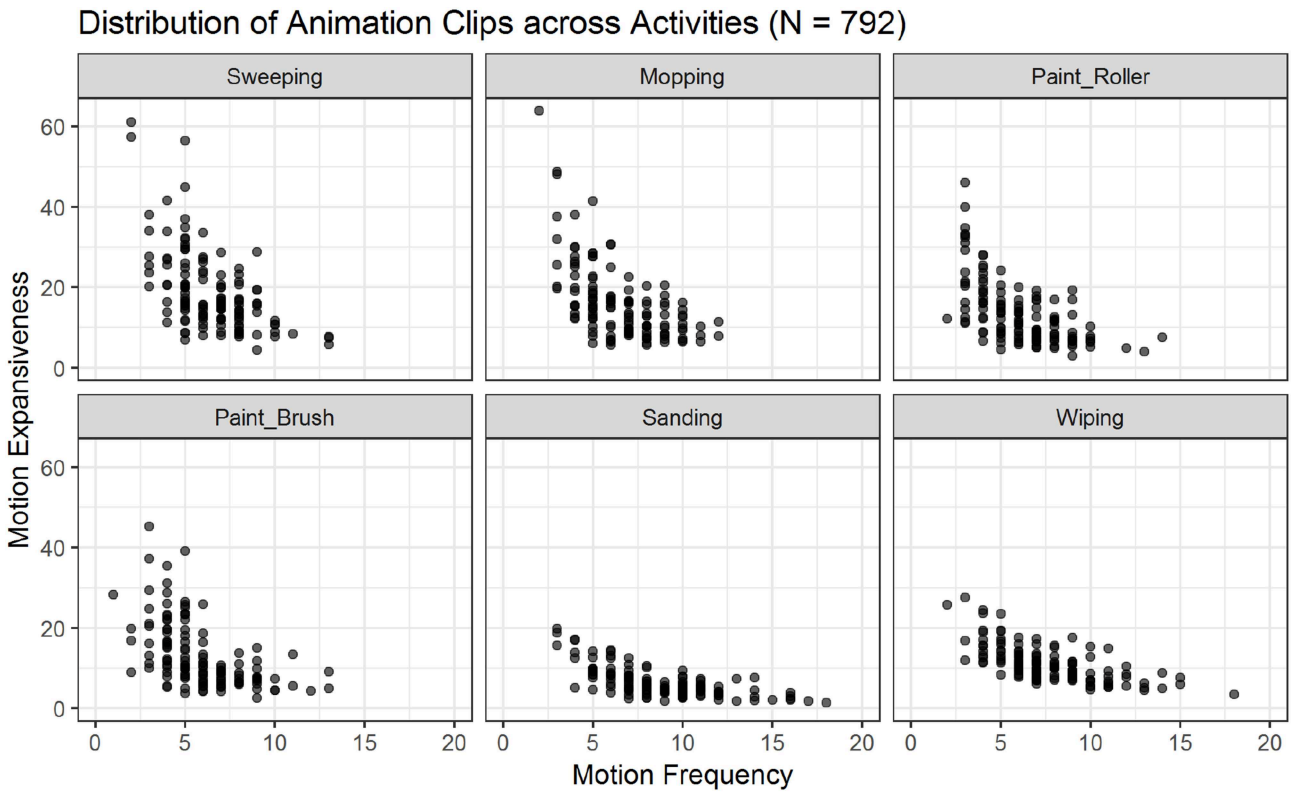



**FIGURE 5 |** Scatter plots showing the relations of motion frequency and motion expansiveness for all six activities. Darker areas indicate the overlap of multiple animation-files. Sanding and wiping show the highest values for motion frequency, while sweeping, and mopping show the highest values for motion expansiveness.

## Homogenizing for Low-Level Physical Features

### Procedure

First, we excluded single animation clips with outliers in any of the 60 variables (outlier defined as a value outside the range of $M \pm 2 \times SD$) to ensure comparability across action and emotion categories. To this end a z-score for each variable was computed. After excluding clips with outlier data in any of the 60 variables, 328 of the initial 792 animations remained (see **Figure 1B**). The distribution of the remaining clips across conditions (actions, moods) is illustrated in **Figure 9**. In a second step, the remaining 328 videos were subsequently analyzed with R (R Core Team, 2019) and RStudio (RStudio Team, 2018) in (generalized) linear mixed effects models, followed by *post hoc* tests as described above.

The goal was to remove groups that show significant differences in their motion frequency and to identify the subset of clips with the highest possible homogeneity (see **Figure 1C**). Since motion frequency is reported to be the most characteristic parameter of movements under varying emotional conditions (Paterson et al., 2001; Sawada et al., 2003), we decided to focus on this variable in the selection process. The results for motion expansiveness are reported as an additional descriptive parameter.

## RESULTS

Motion frequency was analyzed in generalized linear mixed effects models with action and mood as fixed effects and random intercepts for motion capture performers. A model including action and mood (without interaction term) as fixed effects fitted the data significantly better than the null model including only the intercept or models with only one of the fixed effects [$\chi^2_{(2)} = 16.67, p < 0.001$].

Even after filtering outliers there were still significant differences between sad and happy activities, $M = -0.13, SE = 0.05, z = -2.52, p < 0.05$, as well as sad and angry actions, $M = -0.22, SE = 0.05, z = -4.05, p < 0.001$. No significant difference was found between happy and angry actions, $M = 0.09, SE = 0.05, z = 1.82, p = 0.16$. Hence animations containing sad actions were excluded, to homogenize the stimulus set with respect to motion frequency.

In contrast to the analysis prior to the exclusion of outliers, the *post hoc* tests now did not show any significant differences between the motion frequency of either of the three pairs of activities (floor, table, wall). Painting activities were excluded more often by the procedure of outlier removals (see **Figure 9**). In the four other actions (domains: floor, table) there was an uneven distribution among sanding and wiping across moods (see **Figure 9**). Thus, we decided to exclude table- and wall-activities.

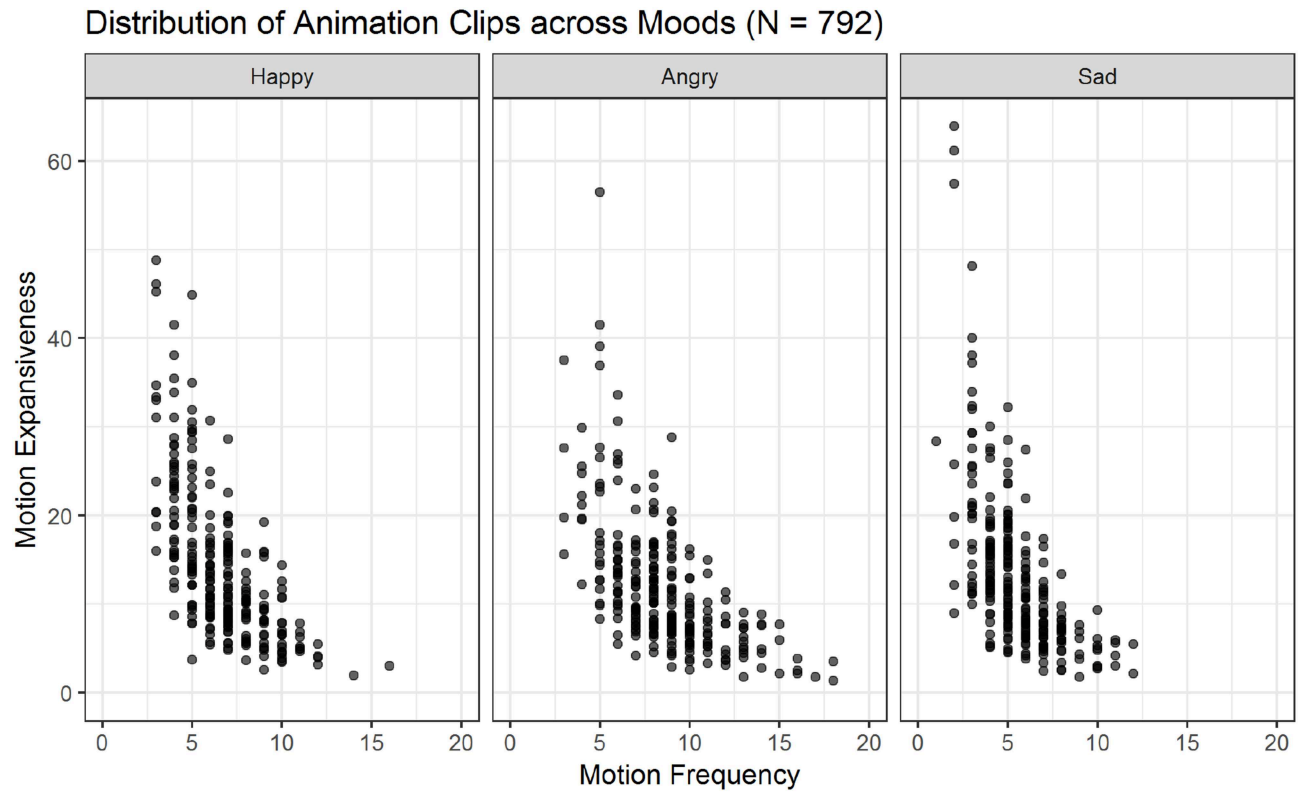



**FIGURE 6 |** Scatter plots showing the relations of motion frequency and motion expansiveness for all three moods. Darker areas indicate the overlap of multiple animation-files. Sad movements show the highest values for motion expansiveness, while angry movements show the highest motion frequency values.

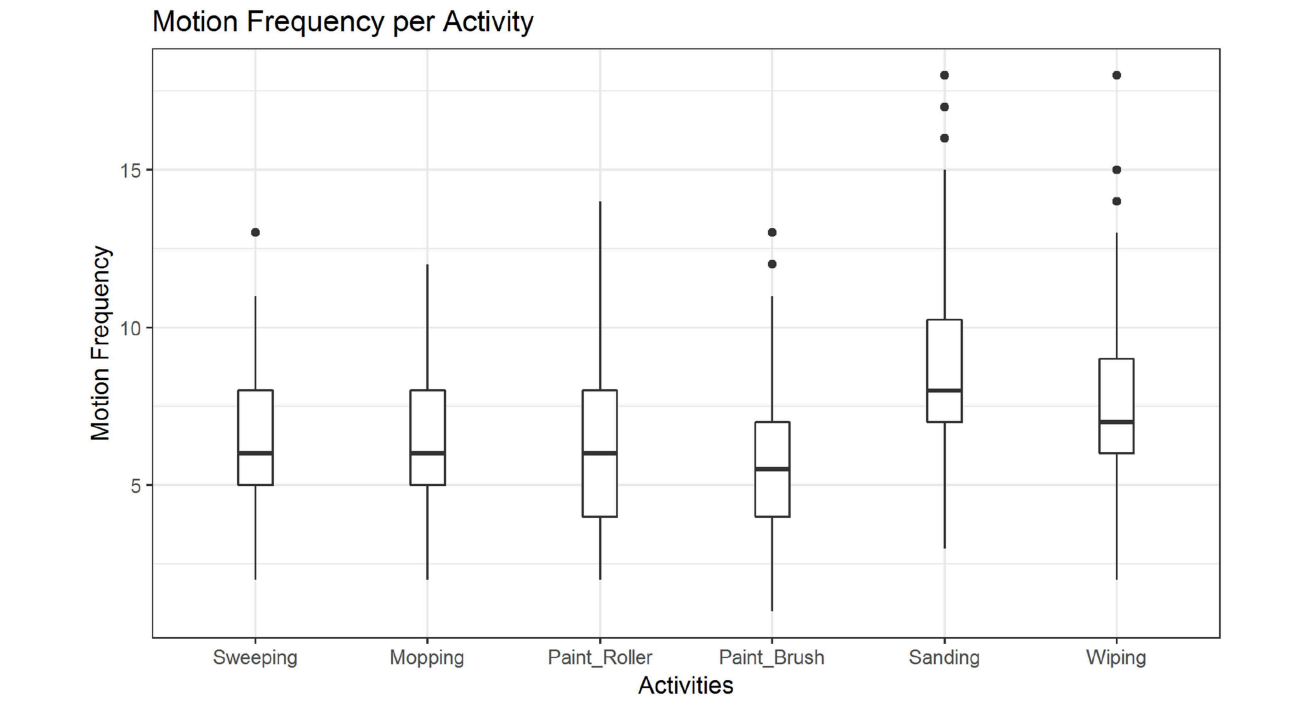



**FIGURE 7 |** Distribution of motion frequency across activities. A significant difference is found between the mean motion frequency of sanding vs. wiping ($p < 0.01$).

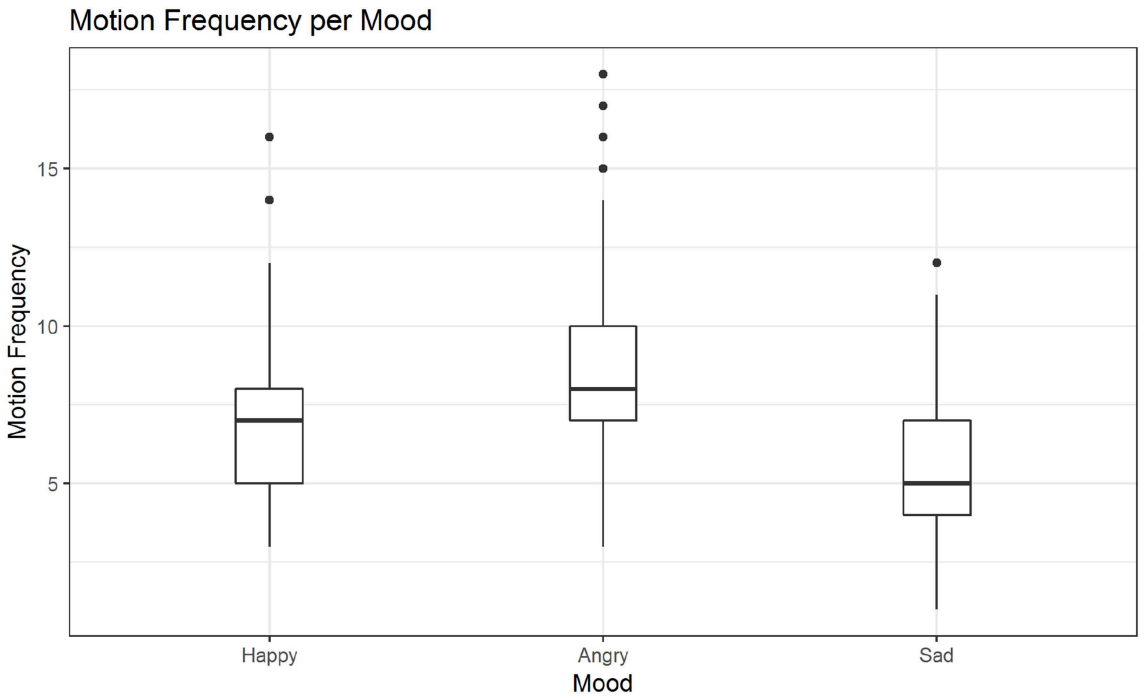



**FIGURE 8 |** Distribution of motion frequency across moods. Significant differences are found in the mean motion frequency between happy and sad movements, angry and sad movements and notably also between happy and angry movements (in all mentioned contrasts: *ps* < *0.001*).

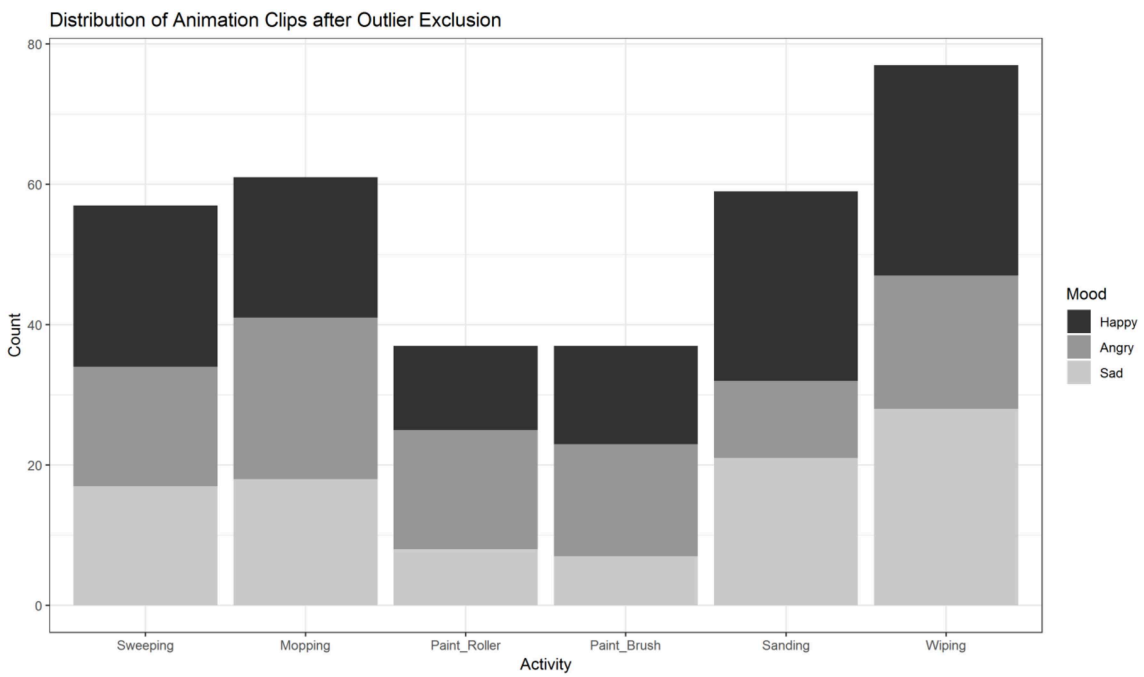



**FIGURE 9 |** Distribution of animation clips ($n = 328$) across activities and moods after exclusion of outliers based on low-level feature extraction. Painting-activities were excluded significantly more often than the four other activities. The distribution of sanding and wiping across moods is unbalanced compared to sweeping and mopping.

Motion expansiveness was investigated by comparing the fit of linear mixed effects models with random intercepts for motion capture performers. A model including action as fixed effect fitted the data significantly better than the null model including only the intercept [$\chi^2_{(5)} = 123.90$, $p < 0.001$]. Adding mood as fixed effect (without interaction term) did not significantly improve the model fit [$\chi^2_{(2)} = 1.94$, $p = 0.38$] and was thus not included in the model.

*Post hoc* tests revealed significant differences between mopping and sweeping, $M = -3.38$, $SE = 0.84$, $z = -4.04$, $p < 0.001$, as well as between sanding and wiping, $M = 4.23$, $SE = 0.78$, $z = 5.40$, $p < 0.001$, but no significant difference between the two painting-activities, $M = -0.19$, $SE = 1.05$, $z = -0.18$, $p = 0.99$.

On the basis of these arguments we decided to focus the following steps on a 2 × 2 design with the actions being mopping vs. sweeping, and the moods being happy vs. angry ($n = 83$ remaining clips).

## Homogenizing for Recognition Rates

This particular selection was intended for a functional neuroimaging study where task difficulty across conditions was ideally balanced between both tasks (Geiger et al., 2019). We therefore conducted an online survey using the remaining 83 clips to receive an additional annotation for these animations. In this survey we showed each animation to volunteers to compute recognition rates for actions and moods. Taking recognition rates as estimate of task difficulty, we further selected clips to homogenize for this high-level feature (see **Figure 1D**). This is especially important in cognitive neuroscience studies to avoid confounding effects of task difficulty on observed brain activity.

### Participants (Observers)

We recruited 112 volunteers (73 females, mean age = 31.66, $SD$ = 11.71) independently from the group of performers (see section Performers) via (a) mailing lists of the study programs Biology, Neuroscience, Philosophy and Psychology of the University of Cologne, (b) word of mouth or (c) a designated mailing list of volunteers of the Research Center Jülich.

Three participants who's answering behavior differed significantly (deviations > 2 × $SD$) from the rest of volunteers were excluded. Additionally, six participants were excluded because they were presented with too many incomplete animations (> 2 × $SD$). The number of incomplete animation playbacks was dependent on the computer hardware and internet connection of each participant. To ensure that the majority of ratings are based on the viewing of complete animations, we excluded participants' ratings with many incomplete animation playbacks. Four participants were excluded, because of technical difficulties, resulting in a total remaining sample of $n = 99$ (64 females, mean age = 31.52, $SD$ = 12.03).

### Procedure

At the beginning of the survey, all participants received structured instructions. It was pointed out that all data were collected and analyzed anonymously. It was further emphasized that the task was either to focus on (a) the action or (b) the mood displayed. Tasks were always indicated before the start of the video and were additionally displayed above the video during its presentation. After the presentation, participants were prompted with an explicit forced-choice format [for the activity: (a) mopping or (b) sweeping; for the mood: (a) happy or (b) angry]. The animations were divided into four subgroups, containing either 20 or 21 clips with approximately equal amounts of clips per mood and activity. Each volunteer was randomly assigned to one of four subgroups and rated each animation of that subgroup for activity and mood. The order of the clips was randomized within the subgroups. After completing the video ratings, basic information (age, gender, handedness, sportiness, years of education) was assessed. The experiment was finished with a short debriefing that informed the participants about the general purpose of the survey and the overarching project. The recognition rates were computed by dividing the amount of correct answers by the total amount of given answers for each animation (for both activities and moods). The survey was conducted via Unipark (Questback GmbH, EFS Survey, Version 10.9, http://www.unipark.com). Results were analyzed in SPSS (Version 24). For the purpose of data cleansing, z-scores were computed for (a) responses, (b) the amount of incomplete clips (see section Participants (Observers) for details).

### Results

The majority of animations were rated above chance level within a range from 55 to 100% correctness in at least one condition (see **Figure 10** and **Supplementary Data Sheet 3**). Thirty-six clips were rated both for action as well as mood at a rate of ≥55%, with a maximum accuracy difference of 30 percentage points between the two scores. For the selection of the fMRI stimulus set, we controlled for two parameters: (a) difference between the two recognition rates (<30 percentage points), (b) equal amount of combinations between activities and moods [angry mopping ($n = 8$), angry sweeping ($n = 8$), happy mopping ($n = 8$), happy sweeping ($n = 8$); see **Supplementary Data Sheet 4**].

# DISCUSSION AND FUTURE PROSPECTS

We herewith present the ACASS database including 792 animations with their respective annotations about basic motion features and emotional expressions inscribed therein. The outstanding features of this newly generated database are (a) the uniform presentation across actors after transferring all human movements onto the same avatar and (b) the motion feature annotation of all animations. The low-level physical feature annotation allows to define various subsets, for instance selecting maximum heterogeneous or homogenous subsets. Furthermore, additional annotations, for instance regarding psychological evaluations as provided by neutral observers can enrich the database and extend its usefulness even beyond the possible applications sketched here.

As a show case, we have demonstrated here as one example how to extract a homogeneous stimulus subset with respect to perceived difficulty of action and mood recognition for the purpose of a particular functional neuroimaging study in the field of social cognitive neuroscience that aimed at identifying

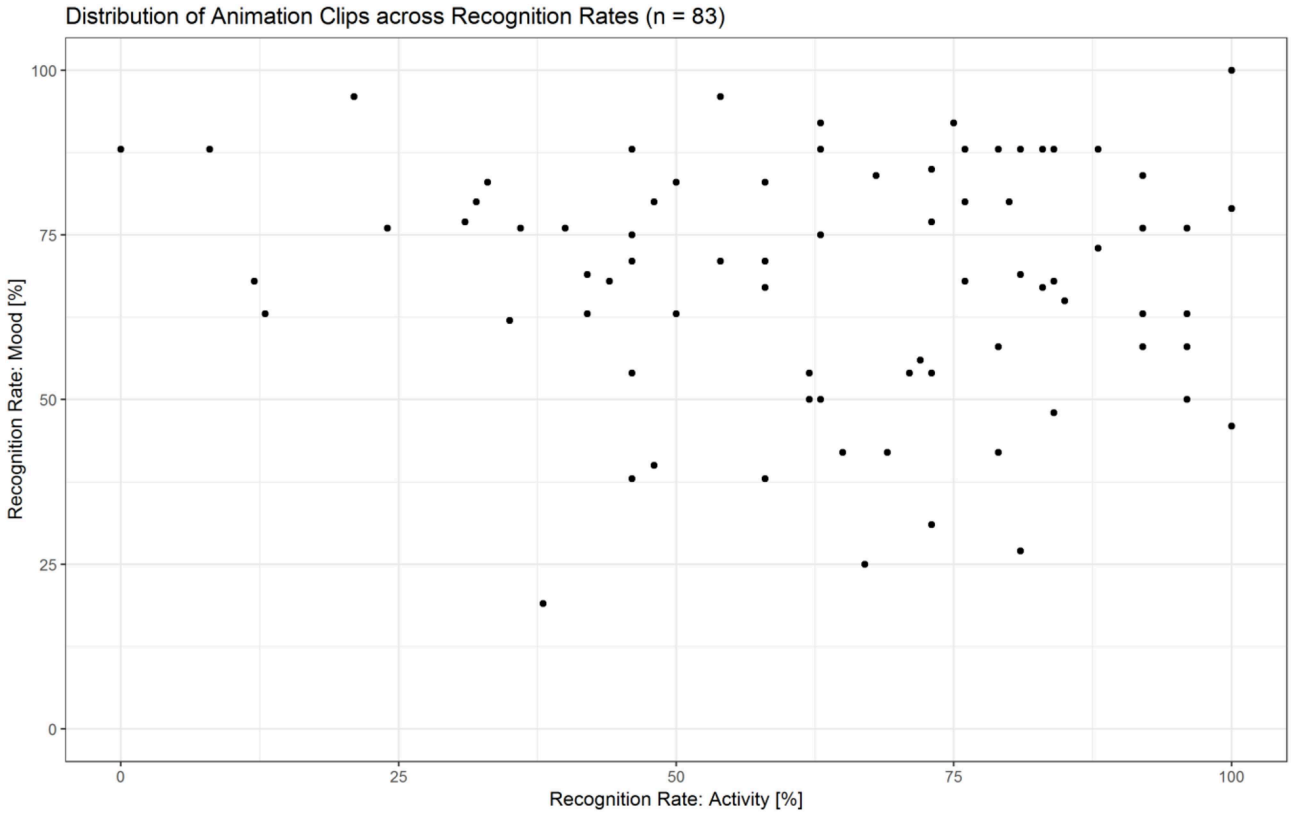



**FIGURE 10 |** Distribution of animation clips ($n = 83$) across recognition rates for activities and moods. The majority of animations were recognized above chance level within a range from 55–100% correctness in at least one condition.

the neural correlates of action recognition and mood recognition (Geiger et al., 2019).

For this subset of the database, different types of application within social neuroscience come to mind: it would be very interesting and timely to investigate the temporal relations of the involved brain systems with more suitable technology like magnetoencephalography. Another obvious question is that of functional connectivity of the involved brain regions. This leads to questions about changes in psychopathological conditions. Abnormalities have been reported for mentalizing abilities in conditions such as schizophrenia and autism spectrum disorders (Frith, 2004). Functional connectivity has been shown to be altered between and within the mentalizing system and the action observation network in autism spectrum disorders (Fishman et al., 2014). With our novel stimulus subset the neural correlates of the involved systems can be investigated in more detail.

Aside from possible applications in the field of social cognitive neuroscience, the stimulus subset, as well as other individually chosen subsets from the database can serve in behavioral studies that use the annotational information to systematically vary e.g., task difficulty (recognition rates). For instance, this could be interesting to contrast ambiguous animations with recognition rates close to guessing rate with other animations that are mostly correctly recognized according to the observer annotation. A further interesting study could be to examine animations that are easily recognized for only one category (e.g., action but not mood). A free viewing task could be conducted to see what the spontaneous attributions of observers are, when no specific instructions and answering options are given. The stimuli could be further enhanced to use in studies about perspective taking and embodiment, e.g., by use in virtual reality or systematically varying the camera angle. Another interesting line of investigation could be to ask participants to rate animations for valence and arousal.

The ACASS database, including the subsets, as well as the source code of the algorithm are hosted at FigShare (doi.org/10.6084/m9.figshare.c.4443014) (preview during review-process). Annotational information are provided in designated CSV-files to enable the selection of individual sets of animations.

## LIMITATIONS

The ACASS database contains recordings of six different household activities that we expect the vast majority of viewers to recognize. All activities were performed stand-alone. Thus, the recordings do not cover interactive situations like dyadic activities or those that address the viewer as an interaction partner. Our main field of application is aimed to be person perception as a well-established domain in social psychology, which includes the processing of social information derived from mere observation beyond true interactions.

## DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the manuscript and the **Supplementary Files**.

## ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the ethics committee of the Medical Faculty of the University of Cologne with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the ethics committee of the Medical Faculty of the University of Cologne.

## AUTHOR CONTRIBUTIONS

GB and KV conceived the project. DR and SL prepared the motion capture recordings. SL conducted the motion capture recordings and prepared the first draft of the manuscript. RT developed the MATLAB algorithm. SL and MJ analyzed the data. All authors reviewed and edited the manuscript.

## ACKNOWLEDGMENTS

We thank Alexander Geiger for his contributions throughout the stimulus production, annotation, and selection. We would also like to show our gratitude to Carola Bloch, Arvid Hofmann, Felix Stetter, and Kristoffer Waldow for their support during the motion capture recordings and processing procedures. Parts of the data have been used in the bachelor's thesis by SL.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2019.00094/full#supplementary-material

## REFERENCES

Atkinson, A. P., Dittrich, W. H., Gemmell, A. J., and Young, A. W. (2004). Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception* 33, 717–746. doi: 10.1068/p5096

Aviezer, H., Trope, Y., and Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science* 338, 1225–1229. doi: 10.1126/science.1224313

Barliya, A., Omlor, L., Giese, M. A., Berthoz, A., and Flash, T. (2013). Expression of emotion in the kinematics of locomotion. *Exp. Brain Res.* 225, 159–176. doi: 10.1007/s00221-012-3357-4

Bartels, A., and Zeki, S. (2004). Functional brain mapping during free viewing of natural scenes. *Hum. Brain Mapp.* 21, 75–85. doi: 10.1002/hbm.10153

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Bente, G. (2019). "New tools - new insights: using emergent technologies in nonverbal communication research," in *Reflections on Interpersonal Communication*, eds S. R. Wilson and S. W. Smith (San Diego, CA: Cognella), 161–188. Available online at: https://titles.cognella.com/reflections-on-interpersonal-communication-research-9781516530427

Bente, G., Krämer, N. C., Petersen, A., and de Ruiter, J. P. (2001a). Computer animated movement and person perception: methodological advances in nonverbal behavior research. *J. Nonverbal Behav.* 25, 151–166. doi: 10.1023/A:1010690525717

Bente, G., Leuschner, H., Issa, A. A., and Blascovich, J. J. (2010). The others: universals and cultural specificities in the perception of status and dominance from nonverbal behavior. *Conscious. Cogn.* 19, 762–777. doi: 10.1016/j.concog.2010.06.006

Bente, G., Petersen, A., Krämer, N. C., and de Ruiter, J. P. (2001b). Transcript-based computer animation of movement: evaluating a new tool for nonverbal behavior research. *Behav. Res. Methods Instrum. Comput.* 33, 303–310. doi: 10.3758/BF03195383

Bente, G., Senokozlieva, M., Pennig, S., Al-Issa, A., and Fischer, O. (2008). Deciphering the secret code: a new methodology for the cross-cultural analysis of nonverbal behavior. *Behav. Res. Methods* 40, 269–277. doi: 10.3758/BRM.40.1.269

Bernieri, F. J., Davis, J. M., Rosenthal, R., and Knee, C. R. (1994). Interactional synchrony and rapport: measuring synchrony in displays devoid of sound and facial affect. *Pers. Soc. Psychol. Bull.* 20, 303–311. doi: 10.1177/0146167294203008

Berry, D. S., Kean, K. J., Misovich, S. J., and Baron, R. M. (1991). Quantized displays of human movement: a methodological alternative to the point-light display. *J. Nonverbal Behav.* 15, 81–97. doi: 10.1007/BF00998264

Berry, D. S., Misovich, S. J., Kean, K. J., and Baron, R. M. (1992). Effects of disruption of structure and motion on perceptions of social causality. *Pers. Soc. Psychol. Bull.* 18, 237–244. doi: 10.1177/0146167292182016

Cavallo, A., Koul, A., Ansuini, C., Capozzi, F., and Becchio, C. (2016). Decoding intentions from movement kinematics. *Sci. Rep.* 6:37036. doi: 10.1038/srep37036

Chouchourelou, A., Matsuka, T., Harber, K., and Shiffrar, M. (2006). The visual analysis of emotional actions. *Soc. Neurosci.* 1, 63–74. doi: 10.1080/17470910600630599

Cutting, J. E., and Proffitt, D. R. (1981). "Gait perception as an example of how we may perceive events," in *Intersensory Perception and Sensory Integration Perception and Perceptual Development*, eds R. D. Walk and H. L. Pick (Boston, MA: Springer US), 249–273. doi: 10.1007/978-1-4615-9197-9_8

de Borst, A. W., and de Gelder, B. (2015). Is it the real deal? Perception of virtual characters versus humans: an affective cognitive neuroscience perspective. *Front. Psychol.* 6:576. doi: 10.3389/fpsyg.2015.00576

Dittrich, W. H. (1993). Action categories and the perception of biological motion. *Perception* 22, 15–22. doi: 10.1068/p220015

Fishman, I., Keown, C. L., Lincoln, A. J., Pineda, J. A., and Müller, R. A. (2014). Atypical cross talk between mentalizing and mirror neuron networks in autism spectrum disorder. *JAMA Psychiatry* 71, 751–760. doi: 10.1001/jamapsychiatry.2014.83

Frith, C. D. (2004). Schizophrenia and theory of mind. *Psychol. Med.* 34, 385–389. doi: 10.1017/S0033291703001326

Geiger, A., Bente, G., Lammers, S., Tepest, R., Roth, D., Bzdok, D., et al. (2019). Distinct functional roles of the mirror neuron system and the mentalizing system. *Neuroimage* 202:116102. doi: 10.1016/j.neuroimage.2019.116102

Grammer, K., Honda, M., Juette, A., and Schmitt, A. (1999). Fuzziness of nonverbal courtship communication unblurred by motion energy detection. *J. Pers. Soc. Psychol.* 77, 487–508. doi: 10.1037/0022-3514.77.3.487

Gross, M. M., Crane, E. A., and Fredrickson, B. L. (2012). Effort-Shape and kinematic assessment of bodily expression of emotion during gait. *Hum. Mov. Sci.* 31, 202–221. doi: 10.1016/j.humov.2011.05.001

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., and Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science* 303, 1634–1640. doi: 10.1126/science.1089506

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biom. J.* 50, 346–363. doi: 10.1002/bimj.200810425

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.* 14, 201–211. doi: 10.3758/BF03212378

Johansson, G. (1976). Spatio-temporal differentiation and integration in visual motion perception. *Psychol. Res.* 38, 379–393. doi: 10.1007/BF00309043

Kret, M. E., and de Gelder, B. (2010). Social context influences recognition of bodily expressions. *Exp. Brain Res.* 203, 169–180. doi: 10.1007/s00221-010-2220-8

Lahnakoski, J. M., Glerean, E., Salmi, J., Jääskeläinen, I. P., Sams, M., Hari, R., et al. (2012). Naturalistic FMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. *Front. Hum. Neurosci.* 6:233. doi: 10.3389/fnhum.2012.00233

Lammers, S. (2017). *Production and perception of whole-body movements to probe the social brain* (bachelor's thesis). University of Cologne, Cologne, Germany. doi: 10.31237/osf.io/swucv

Loula, F., Prasad, S., Harber, K., and Shiffrar, M. (2005). Recognizing people from their movement. *J. Exp. Psychol. Hum. Percept. Perform.* 31, 210–220. doi: 10.1037/0096-1523.31.1.210

Manera, V., Schouten, B., Becchio, C., Bara, B. G., and Verfaillie, K. (2010). Inferring intentions from biological motion: a stimulus set of point-light communicative interactions. *Behav. Res. Methods* 42, 168–178. doi: 10.3758/BRM.42.1.168

Meadors, J. D., and Murray, C. B. (2014). Measuring nonverbal bias through body language responses to stereotypes. *J. Nonverbal Behav.* 38, 209–229. doi: 10.1007/s10919-013-0172-y

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol.* 21, 1641–1646. doi: 10.1016/j.cub.2011.08.031

Paterson, H., Pollick, F., and Sanford, A. (2001). The role of velocity in affect discrimination. *Proc. 23rd Annu. Conf. Cogn. Sci. Soc.* 23, 756–761. Available online at: https://escholarship.org/uc/item/3191m9bh

Poppe, R., Van Der Zee, S., Heylen, D. K. J., and Taylor, P. J. (2014). AMAB: automated measurement and analysis of body motion. *Behav. Res. Methods* 46, 625–633. doi: 10.3758/s13428-013-0398-y

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna. Available online at: https://www.R-project.org/

Rammstedt, B., and John, O. P. (2007). Measuring personality in one minute or less: a 10-item short version of the big five inventory in English and German. *J. Res. Personal.* 41, 203–212. doi: 10.1016/j.jrp.2006.02.001

Ramseyer, F., and Tschacher, W. (2011). Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. *J. Consult. Clin. Psychol.* 79, 284–295. doi: 10.1037/a0023419

RStudio Team (2018). *RStudio: Integrated Development Environment for R*. Boston, MA. Available online at: http://www.rstudio.com/

Sawada, M., Suda, K., and Ishii, M. (2003). Expression of emotions in dance: relation between arm movement characteristics and emotion. *Percept. Mot. Skills* 97, 697–708. doi: 10.2466/pms.2003.97.3.697

Schmälzle, R., Häcker, F. E. K., Honey, C. J., and Hasson, U. (2015). Engaged listeners: shared neural processing of powerful political speeches. *Soc. Cogn. Affect. Neurosci.* 10, 1137–1143. doi: 10.1093/scan/nsu168

Schutte, N. S., Malouff, J. M., Hall, L. E., Haggerty, D. J., Cooper, J. T., Golden, C. J., et al. (1998). Development and validation of a measure of emotional intelligence. *Personal. Individ. Differ.* 25, 167–177. doi: 10.1016/S0191-8869(98)00001-4

Spreng, R. N., McKinnon, M. C., Mar, R. A., and Levine, B. (2009). The Toronto empathy questionnaire. *J. Pers. Assess.* 91, 62–71. doi: 10.1080/00223890802484381

Spunt, R. P., and Lieberman, M. D. (2012). An integrative model of the neural systems supporting the comprehension of observed emotional behavior. *Neuroimage* 59, 3050–3059. doi: 10.1016/j.neuroimage.2011.10.005

Thompson, J., and Parasuraman, R. (2012). Attention, biological motion, and action recognition. *Neuroimage* 59, 4–13. doi: 10.1016/j.neuroimage.2011.05.044

Vogeley, K. (2017). Two social brains: neural mechanisms of intersubjectivity. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 372:20160245. doi: 10.1098/rstb.2016.0245

von der Lühe, T., Manera, V., Barisic, I., Becchio, C., Vogeley, K., and Schilbach, L. (2016). Interpersonal predictive coding, not action perception, is impaired in autism. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 371:20150373. doi: 10.1098/rstb.2015.0373

Westermann, R., Spies, K., Stahl, G., and Hesse, F. W. (1996). Relative effectiveness and validity of mood induction procedures: a meta-analysis. *Eur. J. Soc. Psychol.* 26, 557–580. doi: 10.1002/(SICI)1099-0992(199607)26:4<557::AID-EJSP769>3.0.CO;2-4

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 24 September 2019
doi: 10.3389/frobt.2019.00093

# Managing an Agent's Self-Presentational Strategies During an Interaction

*Beatrice Biancardi[1]*, Maurizio Mancini[2], Paul Lerner[1,3] and Catherine Pelachaud[1]*

[1] CNRS-ISIR, Sorbonne University, Paris, France, [2] School of Computer Science and Information Technology, University College Cork, Cork, Ireland, [3] CNRS-LIMSI, Université Paris-Saclay, Paris, France

In this paper we present a computational model for managing the impressions of warmth and competence (the two fundamental dimensions of social cognition) of an Embodied Conversational Agent (ECA) while interacting with a human. The ECA can choose among four different self-presentational strategies eliciting different impressions of warmth and/or competence in the user, through its verbal and non-verbal behavior. The choice of the non-verbal behaviors displayed by the ECA relies on our previous studies. In our first study, we annotated videos of human-human natural interactions of an expert on a given topic talking to a novice, in order to find associations between the warmth and competence elicited by the expert's non-verbal behaviors (such as type of gestures, arms rest poses, smiling). In a second study, we investigated whether the most relevant non-verbal cues found in the previous study were perceived in the same way when displayed by an ECA. The computational learning model presented in this paper aims to learn in real-time the best strategy (i.e., the degree of warmth and/or competence to display) for the ECA, that is, the one which maximizes user's engagement during the interaction. We also present an evaluation study, aiming to investigate our model in a real context. In the experimental scenario, the ECA plays the role of a museum guide introducing an exposition about video games. We collected data from 75 visitors of a science museum. The ECA was displayed in human dimension on a big screen in front of the participant, with a Kinect on the top. During the interaction, the ECA could adopt one of 4 self-presentational strategies during the whole interaction, or it could select one strategy randomly for each speaking turn, or it could use a reinforcement learning algorithm to choose the strategy having the highest reward (i.e., user's engagement) after each speaking turn.

**Keywords: embodied conversational agents, warmth, competence, human-agent interaction, impression management, non-verbal behavior**

OPEN ACCESS

***Edited by:***
*Gentiane Venture, Tokyo University of Agriculture and Technology, Japan*

***Reviewed by:***
*Soheil Keshmiri, Advanced Telecommunications Research Institute International (ATR), Japan*
*Bipin Indurkhya, Jagiellonian University, Poland*

****Correspondence:***
*Beatrice Biancardi*
*beatrice.biancardi@gmail.com*

***Specialty section:***
*This article was submitted to Human-Robot Interaction, a section of the journal Frontiers in Robotics and AI*

***Received:*** *11 February 2019*
***Accepted:*** *10 September 2019*
***Published:*** *24 September 2019*

***Citation:***
*Biancardi B, Mancini M, Lerner P and Pelachaud C (2019) Managing an Agent's Self-Presentational Strategies During an Interaction. Front. Robot. AI 6:93. doi: 10.3389/frobt.2019.00093*

## 1. INTRODUCTION AND MOTIVATION

During the last decades, anthropomorphic interfaces, such as humanoid robots and virtual characters, have been increasingly deployed in several roles, such as pedagogical assistants, companion, trainers. When conceiving Embodied Conversational Agents (ECAs), which are anthropomorphic virtual characters capable of interacting with users using verbal and non-verbal behavior (for more details, see Cassell, 2000), it is very important to take into account how users

perceive them during the course of the interaction. Virtual agents ought to be endowed with the capability of maintaining engaging interactions with users (Sidner and Dzikovska, 2005). This would make it easier for a virtual guide to transmit information, would ensure change behavior for a virtual coach, would create rapport with a virtual companion. Like in human-human interactions, the first moments of an interaction with a virtual character are critical since users form impressions about them, that can affect the rest of the interaction, in terms of engagement and willingness to continue it (Cafaro et al., 2016).

During the first moments of a new encounter, people automatically collect information to infer the intentions of the others (also called "warmth" dimension Fiske et al., 2007), that is, how the others seem friendly, social, moral, as well as the consequent ability to enact those intentions (called "competence" dimension Fiske et al., 2007), that is, how the others seem intelligent, competent, skillful. People are quite accurate at forming this kind of impressions, by collecting and integrating information from others' appearance and behaviors. This process, defined by Goffman and his colleagues as *impression formation*, is naturally coupled with *impression management*, that is, the attempt to control the impressions that one gives to the others (Goffman et al., 1978). Impression management concerns, among other, dressing and hairstyle, the choice of the moment when smiling, as well as behaviors such as body orientation, posture, etc. People adopt verbal and non-verbal self-presentational strategies in order to elicit in the other a specific impression. According to the context and the goal, one can choose a strategy to convince a target other that he is likable or competent for example (Jones and Pittman, 1982).

Non-verbal behaviors play an important role in these processes (Goffman et al., 1978; Judd et al., 2005). If we want to investigate the effects of these behaviors on the interaction, this could be difficult since we cannot have full control of them in a spontaneous interaction between humans. We can exploit ECAs, which allow us to fully manage their behaviors, to investigate the effect of non-verbal behaviors on the interaction.

In the work presented in this paper, we manage agent's behaviors. To choose the set of possible behaviors for the agent to display, we previously started from the analysis of human-human interaction, in order to identify non-behavioral cues eliciting different impressions of warmth and competence (Biancardi et al., 2017). We then implemented them into an ECA in order to investigate how these cues are perceived when displayed by a virtual character instead of a human (Biancardi et al., 2018). Starting from these findings, we now focus on two main questions:

- What is the impact of these behaviors on a real interaction between an ECA and a human?
- How can an ECA manage its behaviors in order to engage the user, and so to improve the quality of the interaction?

To address them, we have developed a model to manage the impressions generated by an ECA on the user, by endowing it with the capability of adapting its behaviors, and the strategies that drive them, according to user's reactions. The goal of the agent is to maximize user's engagement during the interaction. If the user is engaged, it is more probable for her to have a longer interaction and to appreciate it.

In the following sections, we will describe the dimensions studied in this work in section 2 and the related work in section 3, we will present the architecture of our system in section 4 and the evaluation study of the system in section 5. We will finally discuss the results in section 6 and the limitations and possible improvements of our system in section 7.

## 2. BACKGROUND

In this section we provide definitions and related theories about the psychological dimensions that are investigated in our research: the two fundamental dimensions of social cognition, that is, Warmth and Competence (W&C), and Engagement.

### 2.1. Warmth and Competence

Several authors investigated the fundamental dimensions of social cognition, that is, those characteristics of the others that are processed from the initial moments of an interaction.

These authors converged, even if adopting different terminology, to two main dimensions (Abele and Wojciszke, 2013). The first includes traits like friendliness, morality, sociability, trustworthiness, and it is commonly labeled as warmth. The second one includes traits like agency, efficacy, intelligence, and it is commonly labeled as competence. In the current work we refer to competence as cognitive competence (knowledge, abstract intelligence and experience).

We can already find W&C in Asch's research (Asch, 1946). He was the first who intuited the centrality of W&C in impression formation. Later, Rosenberg et al. distinguished intellectual good/bad traits (such as intelligent, skillful, determined, foolish, unintelligent, irresponsible) and social good/bad traits (such as sociable, honest, warm, unsociable, cold, unhappy) as the main dimensions of person's judgements (Rosenberg et al., 1968). Wojciszke et al. showed that W&C account for almost 82% of the variance in global impressions of well-known others: when people interpret behaviors or their impressions of others, W&C form basic dimensions that almost entirely account for how people characterize others (Wojciszke et al., 1998).

According to the evolutionary explanation given by Fiske et al. warmth is judged before competence, as others' intentions matter more to survival whether the other can act on those goals (Fiske et al., 2007). Primacy of warmth is supported by a large evidence (Willis and Todorov, 2006; Wojciszke and Abele, 2008). In Wojciszke and Abele (2008) participants were asked to list the most important personality traits: they listed significantly more warmth traits than competence traits, and the five most frequently listed traits were warmth-related. Moreover, evaluations based on warmth information were strong and stable, while those based on competence information were weak and dependent on accompanying warmth information. Finally, cognitive performance is better for warmth than for competence. For example, in rapidly judging faces at 100 ms exposure times, social perceivers judged trustworthiness (as a warmth trait Fiske et al., 2007) most reliably, followed by competence (Willis and Todorov, 2006).

Whether the previous authors investigated W&C at a person-perception level, Fiske et al. with their Stereotype Content Model (Fiske et al., 2002), showed the role of W&C in group stereotypes. Groups' warmth is judged according to their level of competition with the in-group, while competence depends on the group status. Different levels of W&C elicit unique emotional (admiration, contempt, envy, and pity; Fiske et al., 2002) and behavioral responses (active and passive, facilitative and harmful; Cuddy et al., 2008).

Another topic of interest concerning W&C is the relationship between the judgements about them. According to Rosenberg et al. they are positively correlated, that is, a *halo effect* occurs (Rosenberg et al., 1968). This effect led people who were given information about only one dimension (warmth or competence), to make judgements about the other (non-described) dimension toward the same direction of the described one.

Yzerbyt et al. showed evidence for an opposite effect instead, called *compensation effect* (Yzerbyt et al., 2008). This effect also occurred in Judd et al. experiments, where they asked to compare two targets. Some participants received information about the competence of the two targets (high in one target and low in the other one), while other participants received information about the warmth of the two targets (again, high in one target and low in the other one). Judgements about the manipulated dimension (competence for some participants, warmth for the others) corresponded to the given information, while for the non-manipulated dimension they went toward the opposite direction of those about the manipulated dimension (Judd et al., 2005).

More recent studies showed the occurrence of compensation effect also in absence of any explicit comparative context, that is without evoking any explicitly comparison to another target. Kervyn et al. called it *amplification effect* (Kervyn et al., 2016).

#### 2.1.1. Behavioral Cues of Warmth and Competence

While most of the studies described above used written descriptions of traits and situations as cues of W&C (e.g., "X helped a blind woman to cross the street," "X wrote a little computer program that solved a tough calculus integration problem"), other works focused on non-verbal cues conveying these dimensions.

Previous research in human-human interaction showed an important effect of smiling on warmth (Bayes, 1972; Cuddy et al., 2008), as well as the presence of immediacy cues that indicate positive interest or engagement (e.g., leaning forward, nodding, orienting the body toward the other), touching and postural openness, and mirroring (i.e., copying the non-verbal behaviors of the interaction partner). Leaning backwards, orientating the body away from the other, tense and intrusive hand gestures (e.g., pointing) are related to impressions of low warmth (Cuddy et al., 2008).

Non-verbal behaviors eliciting competence are more related to dominance and power, such as expansive (i.e., taking up more space) and open (i.e., keeping limbs open and not touching the torso) postures. People who express high-power or assertive non-verbal behaviors are perceived as more skillful, capable, and competent than people expressing low-power or passive non-verbal behaviors (Cuddy et al., 2008). Hand gestures have been found to influence competence perception too, in particular, ideationals (i.e., gestures related to the semantic content of the speech) and object-adaptors resulted in higher judgements of competence, while self-adaptors resulted in lower ones (Maricchiolo et al., 2009).

#### 2.1.2. Self-Presentational Strategies

Jones and Pittman argued that people can use different verbal and non-verbal behavioral techniques to create the impressions they desire in their interlocutor (Jones and Pittman, 1982). The authors proposed a taxonomy of these techniques, that they called self-presentational strategies. We illustrate here 4 of their strategies that can be associated to different levels of W&C. We did not consider the 5th strategy of the taxonomy, called *Exemplification*. This strategy is used when people want to be perceived as self-sacrificing and to gain the attribution of dedication from others, thus it is not related neither to warmth nor to competence. Concerning the other 4 strategies, two of them focus on one dimension at a time, the other two focus on both dimensions by giving them opposite values:

- *Ingratiation*: its goal is to get the other person to like you and attribute positive interpersonal qualities (e.g., warmth and kindness). The person selecting this strategy has the goal to elicit impressions of high warmth, without considering its level of competence.
- *Supplication*: it occurs when individuals present their weaknesses or deficiencies to receive compassion and assistance from others. The person selecting this strategy has the goal to elicit impressions of high warmth and low competence.
- *Self-promotion*: it occurs when individuals call attention to their accomplishments to be perceived as capable by observers. The person selecting this strategy has the goal to elicit impressions of high competence, without considering its level of warmth.
- *Intimidation*: it is defined as the attempt to project its own power or ability to punish to be viewed as dangerous and powerful. In the context of our research, we interpret this strategy in a smoother way, as the goal to elicit impressions of low warmth and high competence.

### 2.2. Engagement in Human-Agent Interaction

An important aspect of human-agent interaction is engagement which ensures the interaction to move forwards. Despite of being a major theme of research and a universal goal in Human-Computer Interaction (HCI), engagement is a difficult concept to define (102 different definitions of engagement exist according to Doherty and Doherty review Doherty and Doherty, 2018), due to its multidimensional nature and the difficulty to measure it.

A detailed summary of engagement definitions in human-agent interaction is provided in Glas and Pelachaud (2015a). Among others, it can be defined as "the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction" (Poggi,

2007), and as "the process by which participants involved in an interaction start, maintain and terminate an interaction" (Sidner and Dzikovska, 2005; Corrigan et al., 2016).

Engagement is not measured from single cues, but rather from several cues that arise over a certain time window (Peters et al., 2005). Engagement can be defined by high-level behavior like, synchrony—which is the temporal coordination during social interactions; mimicry—which is the automatic tendency to imitate others; feedback—which can indicate whether the communication is successful or not. Similarly, engagement can also be defined by low-level behavior like eye gaze - providing feedback and showing interest; head movements - nods (in agreement, disagreement, in between); gestures—to greet, to take turns; postures—body orientation, lean; facial expressions. Clavel et al. provided a review on methodologies for assessing user engagement in human-agent interaction (Clavel et al., 2016).

In the work presented in this Chapter we used low-level signals, such as facial Action Units activation, trunk and head rotation, to measure engagement. The engagement detection model is described in section 4.1.

# 3. RELATED WORK

Some works already exist that included W&C dimensions in ECAs. Nguyen et al. analyzed gestures, use of space and gaze behaviors in videos of actors performing different degrees of W&C (Nguyen et al., 2015). They applied an iterative methodology which included theory from theater, animation and psychology, expert reviews, user testing and feedback, in order to extract a set of rules to be encoded in a virtual agent. They then asked participants to rate W&C of an agent behaving by following these rules. Bergmann et al. found that human-like vs. robot-like appearance positively affects impressions of warmth, while the presence of co-speech gestures increases competence judgements (Bergmann et al., 2012).

The goal of our current work is to model W&C dimensions in order to obtain an engaging ECA, by following the idea that a more engaging agent is likely to form a positive impression and be accepted by the user, thus promoting further interactions (Bergmann et al., 2012; Cafaro et al., 2017). Several authors attempted to design engaging virtual agents, by focusing on the use of feedback and backchannels (Truong et al., 2010), by adopting politeness strategies (Glas and Pelachaud, 2015b), or by investigating the role of verbal alignment for improving user's engagement (Campano et al., 2015). Other studies focused on how to improve user's engagement by adapting social agents (mainly robots) behaviors, using reinforcement learning (RL) methods. These works incorporate user's social signals to measure user's engagement and exploit it as the reward of the RL algorithm. For example, Ritschel et al. computed user's engagement as a reward, with the goal to adapt robot's personality expressed by linguistic style (Ritschel et al., 2017). Gordon et al. exploited facial expressions to measure child's engagement in order to adapt a robot's behaviors (Gordon et al., 2016), while Liu et al. exploited user's physiological signals (Liu et al., 2008).

## 3.1. Our Previous Work

In our previous research, we investigated the associations between non-verbal cues and W&C impressions in human-human interaction (Biancardi et al., 2017). To do that, we annotated videos form NoXi dataset (Cafaro et al., 2017), a corpus of spontaneous interactions involving an expert and a novice discussing about a given topic (e.g., sports, videogames, travels, music, etc.). We annotated the type of gesture, the type of arms rest poses, head movements and smiling, as well as the perceived W&C of the expert. We found a negative association with warmth and competence for some arms rest poses like arms crossed. We also found that the presence of gestures was positively associated with both W&C, in particular the presence of beat gestures (rhythmic gestures not related to the speech content) for both W&C and ideationals for warmth. In addition, when gestures were performed with a smile, warmth judgements increased. A *compensation effect* was found for smiling: warmth judgements were positively related to the presence of smiles, while competence judgements were negatively related to it.

With respect to the works cited at the beginning of the section, we considered more behaviors than only co-speech gestures, in particular the position of the arms when not performing gestures. In addition, we analyzed W&C elicited by non-verbal behaviors performed during natural interactions, instead of behaviors performed by actors.

We then continued our research by questioning how these cues are perceived when displayed by an ECA (Biancardi et al., 2018). To do that, we manipulated in an ECA the most interesting findings from the previous study and asked people to rate videos of the agent displaying different combinations of these manipulations. We found an effect of type of gesture on W&C judgements. In particular, W&C ratings were higher when the agent displayed ideationals than compared to when it displayed beats. In addition, this effect occurred for warmth judgements only when the frequency of gestures was high rather than low.

Our previous works did not investigate W&C impressions in an interaction, where participants are no more passive observers but active agents. The work presented in this paper aims to improve the previous ones, by starting from their findings and focusing on two main questions:

- What is the impact of these behaviors on a real interaction between an ECA and a human?
- How can an ECA manage its behaviors in order to engage the user, and so to improve the quality of the interaction?

We conceived an interaction scenario where the agent manages the impressions of W&C it gives by adopting one of the 4 self-presentational strategies described in section 2.1.2. We exploited the results of our previous works in order to define the non-verbal behaviors associated to each strategy, while we relied on literature to select the verbal behavior for each strategy (see section 5).

In order to make the agent learn how to manage its impressions, that is, to adapt its behavior in real-time to user's engagement level, we adopt a reinforcement learning (RL) approach rather than supervised learning techniques. Since the ECA's behavioral adaptation has the goal to maximize user's engagement, we use this variable as reward in the RL algorithm.

The action space, that is, the set of possible choices of the agent, concerns different behavioral strategies, eliciting impressions of different levels of W&C.

Differently from the existing works described above, the system presented in this paper is the first one using behaviors eliciting different W&C impressions as variables in a RL algorithm for ECAs.

To do this aim, we implemented a system architecture that is described in more details in the following section.

# 4. SYSTEM ARCHITECTURE

We conceived a system architecture to enable the interaction between an ECA and a user. To do that, we implemented software modules to capture user's behavior (speech, facial expressions, head and torso orientation), analyse/interpret it (e.g., detect the user's level of engagement) and decide what the ECA should say and how (i.e., the non-verbal behaviors accompanying speech). The ECA's speech and behavior are decided not only based on the detected user's level of engagement but also by taking into account the ECA's self-presentational intention. That is, the ECA has the goal of communicating a given level of W&C that will influence the choice of the verbal and non-verbal signals to be produced.

**Figure 1** illustrates the system we designed and implemented. We can distinguish 2 main parts:

1. *User analysis*–We exploit the EyesWeb platform (Camurri et al., 2004), that extracts in real-time: (1) user's non-verbal signals (i.e., torso and head orientation), starting from the Kinect depth camera skeleton data; (2) user's face Action Units (AUs), by running the OpenFace framework (Baltrušaitis et al., 2016); (3) user's speech, by executing the Microsoft Speech Platform[1]. After that, as illustrated in section 4.1, EyesWeb computes the user's overall engagement.
2. *ECA generation*–Agent's behavior generation is performed by VIB/Greta, a software platform supporting the creation of socio-emotional embodied conversational agents (Pecune et al., 2014). For the presented work, we implemented a self-presentational intention manager using Flipper (van Waterschoot et al., 2018) to process the detected user's overall engagement and speech and to choose the verbal and non-verbal signals the ECA has to perform in the next speaking turn, according to a reinforcement learning algorithm. The self-presentational intention manager also includes a Natural Language Processing (NLP) module for user's speech interpretation. As explained in section 4.2, Flipper selects the proper communicative intention of the ECA while VIB/Greta generates the ECA animation consisting of gestures, facial expressions and gaze, in sync with speech.

## 4.1. Overall Engagement Detection

As mentioned earlier in the paper, in this work we aim at endowing ECAs with the capability of adapting their behavior according to the user's reactions. In particular, we focus on the user's level of engagement. So, we now present our computational model of user's engagement based on the works of Corrigan et al. (2016) and Sidner and Dzikovska (2005). In our model, user's engagement can be expressed at three different levels, corresponding to different types of non-verbal signals:

- *Attention engagement*–Engagement can be expressed by continuously gazing at relevant objects/persons during the interaction. The more a person continuously focuses her attention on a relevant object/person, the more engaged she is (Sidner and Dzikovska, 2005).
- *Cognitive engagement*–(Corrigan et al., 2016) claims that "frowning may indicate effortful processing suggesting high levels of cognitive engagement." The same work also refers to signals such as "looking for a brief interval outside the scene" as indicators of cognitive engagement.
- *Affective engagement*–Smiling could indicate that a person is enjoying the interaction, while some postures (e.g., crossed arms, hands in pockets) or posture shifts can indicate a lack of engagement.

The *Affective and Cognitive Engagement Detection* module is based on a Long Short-Term Memory (LSTM) prediction model using Recurrent Neural Networks implemented with the Keras toolkit and TensorFlow. More details about this model can be found in Dermouche and Pelachaud (2018). The prediction model takes as input the user's face AUs during the last second, and predicts the user's affective and cognitive engagement: for example, when non-verbal signals like frowning or smiling are extracted, the affective and cognitive engagement increases.

The *Attention Engagement Computation* module is implemented in EyesWeb as a set of rules. It takes as input the user's head and torso orientation and computes the user's attention engagement: for example, if the user is facing the ECA (with both her head and torso) then the attention engagement increases.

Finally, affective, cognitive and attention engagement are summed up by the *Overall User Engagement Computation* module.

Overall user's engagement is computed continuously at 10 Hz during every speaking turn, starting when the agent starts to pronounce its question for the user and ending when the user stops replying to the agent (or, if the user does not respond, until a 1,500 ms of continuous silence is detected). After the end of the speaking turn, the overall mean engagement is sent from EyesWeb to the Self-presentational Intention Manager, described in the following section, that will plan the verbal and non-verbal behavior the ECA will produce in the next speaking turn.

**Figure 2** depicts the user analysis interface, developed in EyesWeb.

## 4.2. Self-Presentational Intention Manager

User's speech and overall engagement are sent to the Self-presentational Intention Manager implemented in the Dialog Manager Flipper, an open-source engine for pragmatic yet robust interaction management for ECAs (van Waterschoot et al., 2018).

The Dialog Manager Flipper is based on two main components described in XML: the *information state* and the

[1]https://www.microsoft.com/en-us/download/details.aspx?id=27225

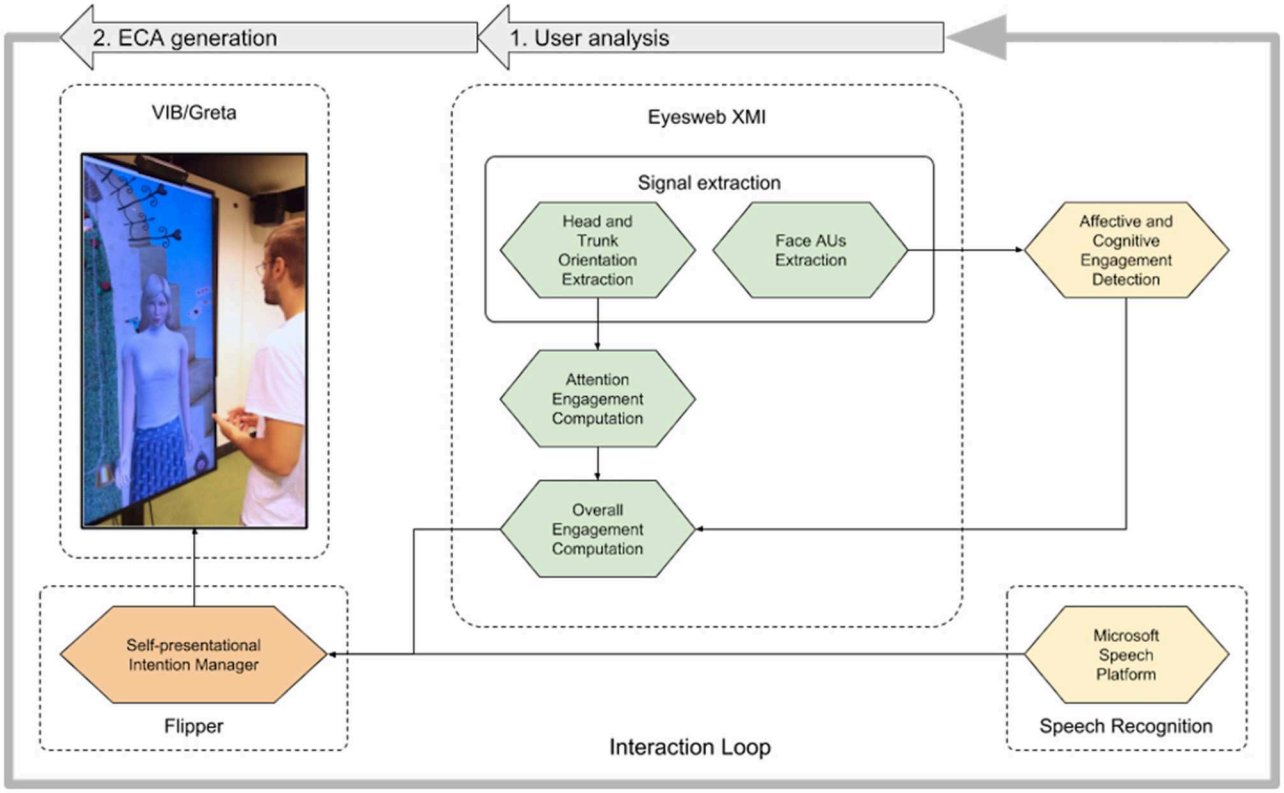



**FIGURE 1 |** System architecture: user non-verbal and verbal signals are extracted by EyesWeb and the Microsoft Speech Platform, respectively; user's overall engagement, computed by EyesWeb, is provided to the Self-presentational Intention Manager that decides the verbal and non-verbal signals to be produced by VIB/Greta. (The person in this image agrees for publication).

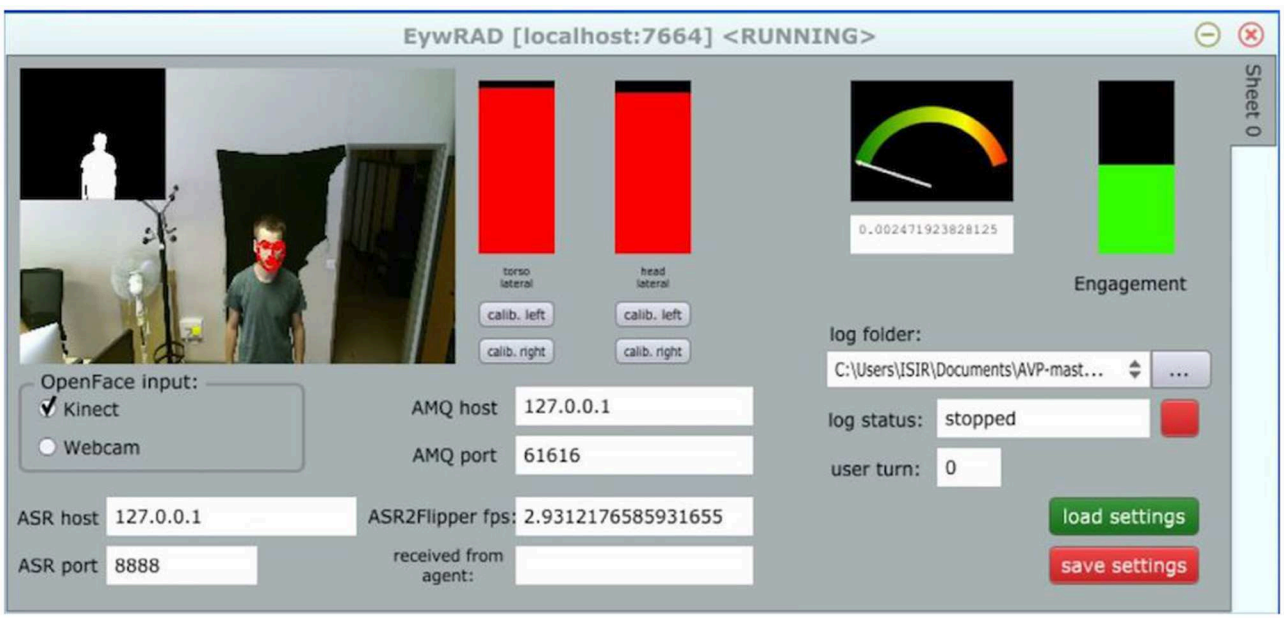



**FIGURE 2 |** The user analysis interface implemented in EyesWeb. On the left, user's silhouette is extracted from Kinect's depth data. The two red bars in the middle indicate that the user is looking at the screen, with both her trunk (left bar) and head (right bar). Audio intensity is low (volume meter on the right), that is, the user is not speaking. Overall engagement level is represented by the green bar on the right (The person in this image agrees for publication).

*declarative templates*. The information state stores interaction-related information and data in a hierarchical tree-based structure. Declarative templates can be grouped and organized in different files according to their related functionality (van Waterschoot et al., 2018). Each template consists of:

- *preconditions*: sets of rules that describe when a template should be executed;
- *effects*: associated updates to the information state.

So, for example, we defined a template whose precondition is that the user's overall engagement value has been computed by EyesWeb (see section 4.1) and the effect is that the expected reward of the current self-presentational intention has to be updated depending on the engagement value (see section 4.2.1).

Flipper has been also exploited to implement a dialogue manager based on NLP, aiming at interpreting user's speech to select the ECA's next self-presentational intention. Since the generation of a realistic and complex dialogue is not the main focus of our work, the agent takes into account only the polarity of user's answers rather than the semantic content of user's speech. For example, the agent can ask whether or not the user wants a more detailed explanation about a topic: if the user's answer is positive, then the agent will talk about it in more detail, or will move to another topic in case of a negative answer (see section 5.5).

#### 4.2.1. Self-Presentational Intention Selection

During its interaction with the user, the ECA has the goal of selecting its self-presentational intention (e.g., to communicate verbally and non-verbally a given utterance with high warmth and low competence). The ECA will choose its intention among a given set of possible utterances depending on the user's overall engagement value: for example, if the last self-presentational intention had the effect of decreasing the detected user's engagement, then the ECA will select a different intention for the next speaking turn, that is it will select an utterance associated with a different value of warmth and of competence; conversely, if the last intention increased user's engagement, that intention will be maintained.

This problem can be seen as a *multi-armed bandit problem* (Katehakis and Veinott Jr, 1987), which models agents evolving in an environment where they can perform several actions, each action being more or less rewarding for them.

In our case, the actions that the ECA can perform are the verbal and non-verbal behaviors corresponding to the self-presentational intention the ECA aims to communicate, and they are selected by the Formula 1. The environment is the interaction with the user, while the state space is the set of the topics discussed at each speaking turn, and it is defined by the Dialog Manager. That is, the choice of the action does not change the state (i.e., the topic discussed during the actual speaking turn), but rather it acts on how this topic is realized by verbal and non-verbal behavior.

In order to maximize user's engagement during the interaction, the ECA will, at the beginning, explore the environment (i.e., by randomly choosing an initial self-presentational intention) and then exploit its knowledge (i.e., user's engagement) to find the most rewarding self-presentational intention.

To do that, we choose to exploit the $\epsilon$-decreasing learning approach: the exploration rate $\epsilon$ continuously decreases in time. In this way, the ECA starts the interaction with the user by exploring the environment without taking into account knowledge (i.e., user's engagement) and finishes it by exploiting its knowledge only (i.e., without performing any further environment exploration). That is, the ECA explores with probability $\epsilon$, and exploits knowledge with probability $1 - \epsilon$.

The ECA updates its knowledge through a table where it iteratively approximates the expected reward $Q(int)$ of a self-presentation intention $int$. This is done using the formula below:

$$Q(int)_{t+1} \leftarrow (1 - \alpha) \times Q(int)_t + \alpha \times e_t \quad (1)$$

where:

- $Q(int)$ is the expected value of the intention, $int \in$ [*ingratiation, supplication, self-promotion, intimidation*];
- $\alpha$ is the learning rate, set at 0.5, a very high number compared to other works (e.g., in Burda et al., 2018 it was set to 0.0001). This is because the ECA needs to learn quickly (i.e., in few dialogue steps) the self-presentational intention to use;
- $e$ is the overall engagement score, that is the reward for the ECA.

## 5. EVALUATION STUDY

We now present the evaluation study we conceived to investigate whether or not an ECA endowed with the architecture described in the previous section, that is, able to manage its impressions of W&C according to user's engagement, could affect user-agent interaction. In the study, we compared different conditions where the ECA could interact with the user by adapting or not its behaviors.

We created a scenario where the virtual agent, called Alice, plays the role of a virtual guide of a museum. The experiment took place in the Carrefour Numerique, an area of the Cité des sciences et de l'industrie in Paris, one of the largest sciences museums in Europe.

### 5.1. Independent Variables

The independent variable manipulated in this study concerns agent's **Strategy**, that is, how the agent manages its behaviors to influence user's perception of its W&C.

For each speaking turn, the agent plays one out of 4 self-presentational techniques presented in section 2.1.2, inspired from Jones & Pittman's taxonomy (Jones and Pittman, 1982), in order to appear more or less warmth and/or competent. According to the different **Strategy** conditions, the agent can select one of the 4 self-presentational techniques at the beginning and display it during the whole interaction, or select one of the 4 at each speaking turn, either randomly or by using our self-presentational intention model based on user's overall engagement detection.

These 4 self-presentational techniques are realized by the agent through its non-verbal and verbal behavior. The choice

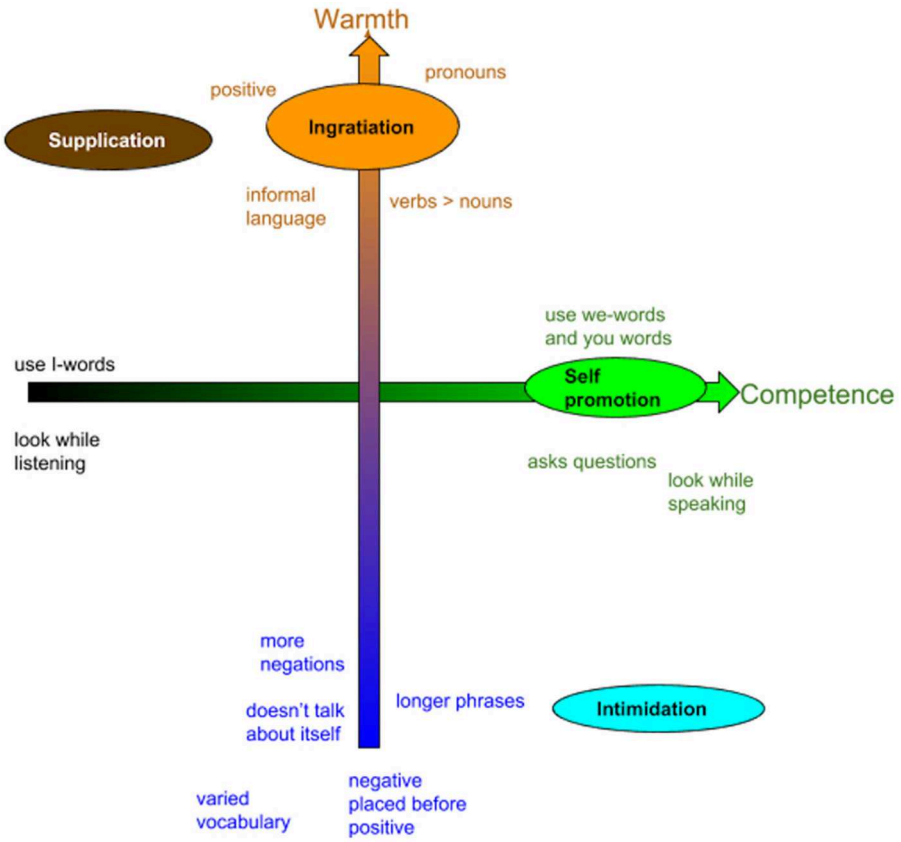



**FIGURE 3 |** Use of pronouns, verbs, type of language, and other verbal behaviors associated to each self-presentational technique, inspired from Pennebaker (2011) and Callejas et al. (2014).

of its non-verbal behavior is based on our previous studies described in section 3.1. The verbal behavior characterizing the different strategies is inspired by the works of Pennebaker (2011) and Callejas et al. (2014). According to their findings, we manipulated the use of *you-* and *we-* pronouns, the level of formality of the language, the length of the sentences. For example, sentences aiming at eliciting high warmth contain more pronouns, less synonyms, more informal language, so that the phrases are more casual and give the impression to be less meditated; more verbs rather than nouns, and positive contents are predominant. Sentences aiming at eliciting low warmth contain more negations, longer phrases, formal language, and do not refer to the speaker. Sentences aiming at eliciting high competence contain high rates of we- and you-words, and I-words at low rates. **Figure 3** shows the use of verbal behavior according to each self-presentational technique, while **Table 1** shows an example of a speaking turn for each of the 4 techniques.

The independent variable **Strategy** has 6 levels: the first 4 levels are static conditions, where one self-presentational technique is chosen at the beginning of the interaction and does not change; in the last 2 levels the self-presentational technique is chosen at each speaking turn. They are:

- **INGR:** when the agent selects the Ingratiation self-presentational technique from the beginning to the end of the interaction, without considering user's reactions;
- **SUPP:** when the agent selects the Supplication self-presentational technique from the beginning to the end of the interaction, without considering user's reactions;
- **SELF:** when the agent selects the Self-promotion self-presentational technique from the beginning to the end of the interaction, without considering user's reactions;
- **INTIM:** when the agent selects the Intimidation self-presentational technique from the beginning to the end of the interaction, without considering user's reactions;
- **RAND:** it consists in selecting one of the 4 self-presentational techniques, randomly, at each speaking turn, **without considering user's reactions**;
- **IMPR:** it consists in selecting one of the 4 self-presentational techniques, at each speaking turn, **by using our self-presentational intention model** based on user's overall engagement detection (see section 4.1).

According to the **Strategy** level, the self-presentational intention selection module of the Dialog Manager Flipper (see section 4.2.1) will apply (or not) the reinforcement learning formula 1 to update the action (i.e., the following self-presentational intention) of the agent.

**TABLE 1 |** An example of 4 different sentences for the same speaking turn (the agent introduces the videogames exhibition), according to the 4 different self-presentational techniques.

| Strategy | Translated sentence | Original sentence |
|---|---|---|
| **INGR** | "You can test some games, if you wanna." | *Tu vas pouvoir tester des jeux si tu veux.* |
| **SUPP** | "I dunno about the other exhibits of the museum, but here you can test some games, it's cool!" | *J'connais pas les autres expo du musée, mais ici on peut tester des jeux, c'est trop bien !* |
| **SELF** | "In this exhibition, you can test some videogames." | *Dans cette expo tu va pouvoir tester des jeux-vidéos.* |
| **INTIM** | "In this exhibition, you can try out some games on different platforms." | *Dans cette exposition tu peux essayer des jeux sur différents supports.* |

*The original sentences in French are provided.*

**TABLE 2 |** Items of the NARS questionnaire, adapted from Nomura et al. (2006).

| Items |
|---|
| 1. I would feel uneasy if virtual characters had emotions. |
| 2. I would feel relaxed talking with virtual characters. |
| 3. I feel comforted being with virtual characters that have emotions. |
| 4. The word "virtual character" means nothing to me. |
| 5. I would hate the idea that virtual characters were making judgements about things. |
| 6. I would feel very nervous just standing in front of a virtual character. |
| 7. I would feel paranoid talking with a virtual character. |
| 8. I am concerned that virtual characters would be a bad influence on children. |

**TABLE 3 |** Items of the questionnaire about user's perception of the interaction, adapted from Bickmore et al. (2011).

| Measure | Question |
|---|---|
| **Satisfaction** | *I am satisfied with my interaction with Alice.* |
| **Continue** | *I would like to talk with Alice again.* |
| **Like** | *I liked Alice.* |
| **Learnfrom** | *I have learned something from Alice.* |
| **Exhib** | *Alice gave me want to visit the exhibition (if you haven't yet)* |
| **Rship** | *I would describe Alice as a complete stranger vs. a close friend.* |
| **Likeperson** | *I would describe Alice just as a computer vs. like a person.* |

*Alice is the name of the virtual character.*

## 5.2. NARS

Before the interaction, we collected information about users' attitudes and prejudices toward virtual characters. We used a slightly adapted version of the Negative Attitudes toward Robots Scale (Nomura et al., 2006). This questionnaire measures people's negative attitudes toward situations and interactions with robots, toward the social influence of robots, and toward emotions in interaction with robots. We selected the most relevant questions according to our context and adapted the questions by referring to virtual characters instead of robots. Participants gave their rating on a 5-points Likert scale, from 1 = "I completely disagree" to 5 = "I completely agree." The items of the questionnaires (translated in English) are available in **Table 2**.

## 5.3. Dependent Variables

The dependent variables were measured during and after the interaction with the virtual character. During the interaction, if the participant agreed in the consent form, we recorded the user's speech audio, in order to measure user's cues of engagement from his verbal behavior. After the interaction we asked the participants to rate the agent's W&C, and their overall satisfaction of the interaction.

### 5.3.1. Verbal Cues of Engagement

For people who agreed with audio recording of the experiment, we collected quantitative information about their answers, in particular:

- The polarity of the answer to Topic1_question (see section 5.5);
- The polarity of the answer to Topic2_question (see section 5.5);
- The number of any verbal feedback produced by the user during a speaking turn.

### 5.3.2. Self-Report Assessment

After the interaction, the participants filled in a final questionnaire, divided in several parts. In particular we measured:

- User's perception of agent's warmth (**w**) and competence (**c**): we presented a list of adjectives referring to W&C and asked participants to indicate their agreement on a 5-points Likert scale (1 = "I completely disagree," 5 = "I completely agree") about how precisely each adjective described the character. The items were taken from Aragonés et al. (2015) scale, and were: *kind, pleasant, friendly, warm* for warmth, and *competent, effective, skilled, intelligent* for competence.
- User's perception of the interaction (**perception**): the second part of the questionnaire concerned a list of items adapted from those already used by Bickmore et al. (2011). They are shown in **Table 3**.

## 5.4. Hypotheses

The first experiment's goal was to demonstrate that the ECA's 4 self-presentational techniques during all the interaction are correctly perceived by users, for example, if users rate the agent in **INGR** condition as warm, and the agent in **INTIM** as cold and competent.

In particular, we hypothesize that:

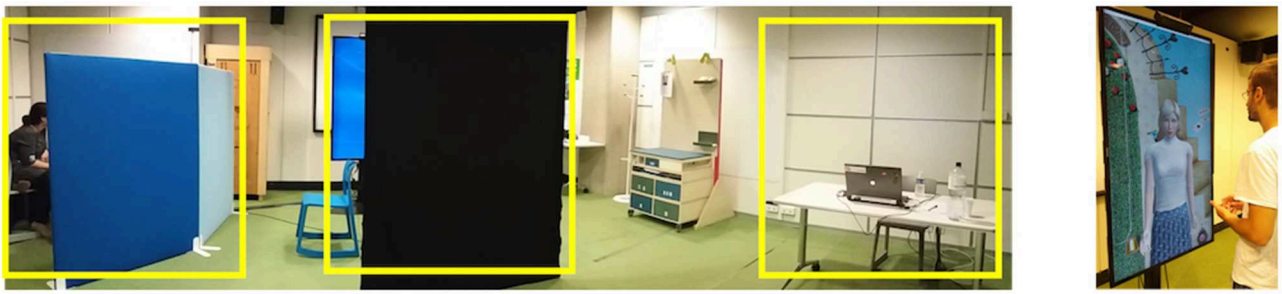


**FIGURE 4 |** The experimenter room and an example of an interaction (the person in this image agrees for publication). In the yellow squares, on the left, the control place, in the middle the interaction place, and on the right the questionnaires space.

- **H1ingr**: The agent in **INGR** condition will be perceived as **warm** by users;
- **H1supp**: The agent in **SUPP** condition will be perceived as **warm** and **not competent** by users;
- **H1self**: The agent in **SELF** condition will be perceived as **competent** by users;
- **H1intim**: The agent in **INTIM** condition will be perceived as **competent** and **not warm** by users.

Then, our main hypothesis is that the use of the self-presentational intention model based on user's overall engagement detection (i.e., when the virtual character adapts its behaviors) positively affects user's perception of the interaction. Thus, we hypothesize that:

- **H2a**: The scores of **perception** items are higher in **IMPR** condition compared to all the other conditions;
- **H2b**: The agent in **IMPR** condition influences how it is perceived in terms of W&C.

## 5.5. Protocol

The experiment took place in a room of the Carrefour Numérique. As shown in **Figure 4**, the room was divided in three areas:

- The questionnaires place, including a desk with a laptop, and a chair;
- The interaction place, with a big screen displaying the virtual character, a Kinect 2 on the top of the screen and a black tent in front of the screen;
- The control station, separated by the rest of the room by 2 screens. This place included a desk with the computer controlling the system.

The experiment was completed in three phases:

1. Before the interaction begun, the participant sat at the questionnaires place, read and signed the consent form, and filled in a first questionnaire (see section 5.2), then moved to the interaction place, where the experimenter gave the last instructions (5 min);
2. During the interaction phase, the participant stayed right in front of the screen, between it and the black tent. He\she wore a headset and was free to interact with the virtual character as he \she wanted. During this phase, the experimenter stayed in the control place, behind the screens (3 min);
3. After the interaction, the participant came back to the questionnaires place and filled in the last questionnaires (see section 5.3.2). After that, the experimenter proceeded with the debriefing (5 min).

The interaction with the virtual character lasted about 3 min. It included 25–36 steps, according to user's answers. A step includes one or few sentences played by the virtual character and user's answer. If user did not reply in a certain interval of time, the agent started the following step. After each step, user's engagement was computed through our overall engagement detection model (see section 4.1).

The dialogue is divided into 4 main parts that were always played by the agent, no matter what answers the users gave:

1. Start interaction (8 steps);
2. Topic 1 (3 steps);
3. Topic 2 (4 steps);
4. End of the interaction (4 steps).

At the end of parts 1, 2, and 3, the agent asked a question to the user. After parts 2 and 3, if the user gave a positive answer, the agent continued to talk about the same topic (6 steps for Topic 1, 5 steps for Topic 2), otherwise it skipped to the next part. The dialogue flowchart is shown in **Figure 5**.

## 5.6. Analysis and Results

We analyzed data from 75 participants, of which were 30 females and 2 preferred not to specify their gender. The majority of the participants were in the 18–25 or 36–45 age range, 5 of them were not native French speakers, and 72% of them had at least a Bachelor. Participants were almost equally distributed across the levels of the independent variable **Strategy** ($12.5 \pm 1$ participants per each strategy).

Before conducting our analyses, we computed Cronbach's alphas and explored the distribution of data. Good reliability for **w** and **c** items was found ($\alpha = 0.9$ and $\alpha = 0.8$, respectively). We then used the mean of these items for our analyses. Since the distributions of this data satisfy assumptions for ANOVA, we run this type of test on them.

Nars scores got an acceptable score of reliability ($\alpha = 0.66$), we therefore computed the means of these items in order to

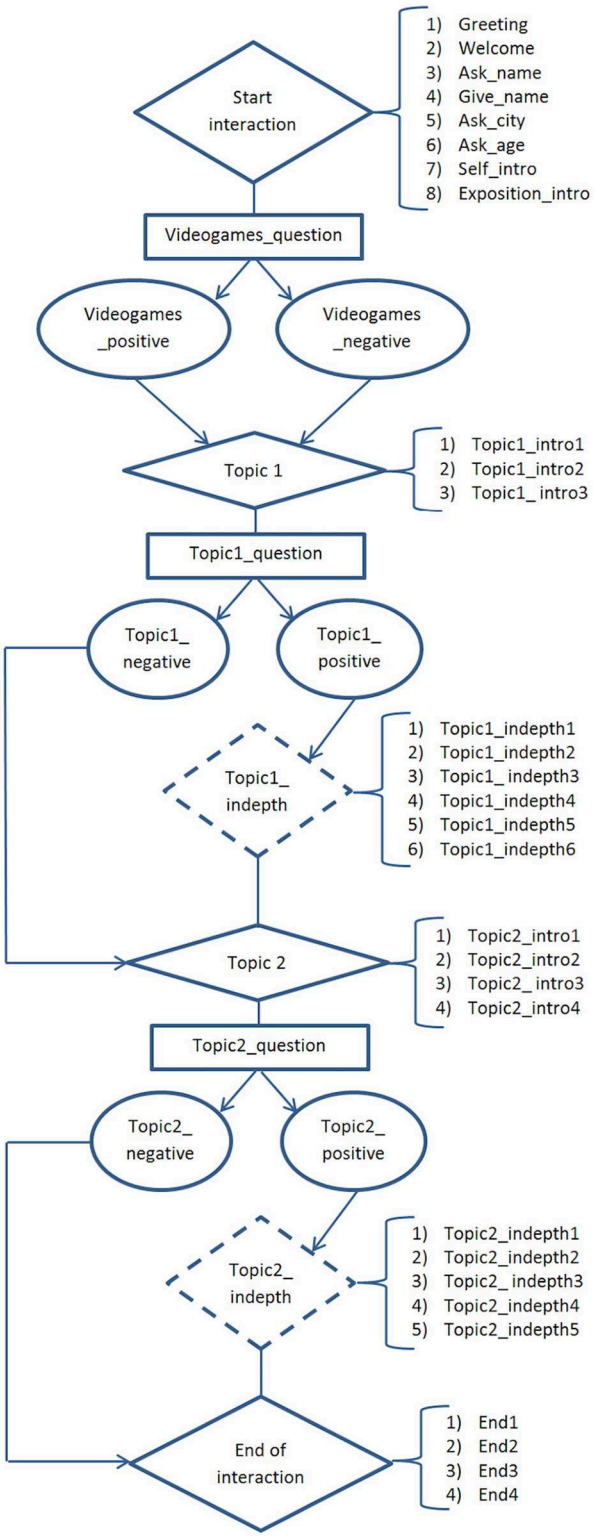



**FIGURE 5 |** The dialogue flowchart. The diamond shapes represent the main parts that always occur during the dialogue, the rectangles represent questions, the rounds represent agent's reply to user's answer and the dotted shapes the optional parts. Where not specified, each shape represents one step of the dialogue.

obtain one overall mean for each participant. We then divided participants into 2 groups, "high" and "low," according to whether they obtained a score higher than the overall mean or not, respectively. Participants were almost equally distributed into the two groups (39 in the "high" group, 36 in the "low" group, almost equally distributed across the other variables, too).

### 5.6.1. Warmth

A 4-way between-subjects ANOVA, including age, sex and Nars scores as factors, was first run in order to check for any effect of these variables. A main effect being found for Nars scores, we then conducted a 4 × 2 between-subjects ANOVA with **Strategy** and Nars as factors. The analysis revealed a main effect of **Strategy** [$F_{(5,62)} = 4.75$, $p = 0.000974$, $\eta^2 = 0.26$] and Nars [$F_{(1,62)} = 5.74$, $p = 0.02$, $\eta^2 = 0.06$]. Warmth ratings were higher from participants with a high Nars score ($M = 3.74$, $SD = 0.77$) than from those with a low Nars score ($M = 3.33$, $SD = 0.92$).

In **Table 4** are showed mean and SD of **w** scores for each level of **Strategy**. Multiple comparisons $t$-test using Holm's correction shows that the **w** mean for **INTIM** is significantly lower than all the others (see **Figure 6**). As consequence, the others conditions are rated as warmer than **INTIM**. **H1ingr**, **H1supp** are thus validated, and **H1intim** and **H2b** are validated for the warmth component.

### 5.6.2. Competence

A 4-way between-subjects ANOVA, including age, sex and Nars scores as factors, was first run in order to check for any effect of these variables. No effects were found for any factor, even when considering only **Strategy** as factor. When looking at the means of **c** for each condition (see **Table 5**), **SUPP** is the one with lower score, even if its difference with the other scores does not reach statistically significance (all $p$-values $> 0.1$). **H1supp** and **H1intim** (for the competence component) are not validated.

### 5.6.3. User's Perception of the Interaction

We analyzed each item of **perception** separately, by applying non-parametric tests since data were not normally distributed.

Concerning **satisfaction** scores, a Kruskal-Wallis rank test showed a statistically significant difference according to **Strategy** [$H_{(5)} = 11.99$, $p = 0.03$]. In particular, Dunn's test for multiple comparisons found that **INGR** scores were significantly higher than **SUPP** ($z = 2.88$, p-adj $= 0.03$) and **INTIM** ($z = 2.56$, p-adj $= 0.04$) (see **Figure 7A**). No differences were found between **IMPR** scores and the other conditions. In addition, a statistically significant difference between scores was found according to Nars scores ($U = 910.5$, $p = 0.02$): participants who got high scores in the Nars questionnaire were more satisfied by the interaction ($M = 3.62$, $SD = 0.94$) than people who got low scores in the Nars questionnaire ($M = 3.00$, $SD = 1.07$). Another interesting results concerns the effect of age on **satisfaction** [$H(4) = 15.05$, $p = 0.005$]: people in the age range 55+ were more satisfied than people of any other age range (see **Figure 7B**, all p-adj $\leq 0.03$).

Concerning **continue** scores, no effect of **Strategy** was found. In general, mean scores were not very high, with only scores

**TABLE 4 |** Mean and standard deviation of warmth scores for each level of **Strategy**.

| Condition | Warmth mean ± SD |
|---|---|
| **INGR** | 3.77 ± 0.57 |
| **SUPP** | 3.54 ± 0.999 |
| **SELF** | 3.81 ± 0.70 |
| **INTIM** | 2.63 ± 0.93 |
| **RAND** | 3.71 ± 0.80 |
| **IMPR** | 3.89 ± 0.38 |

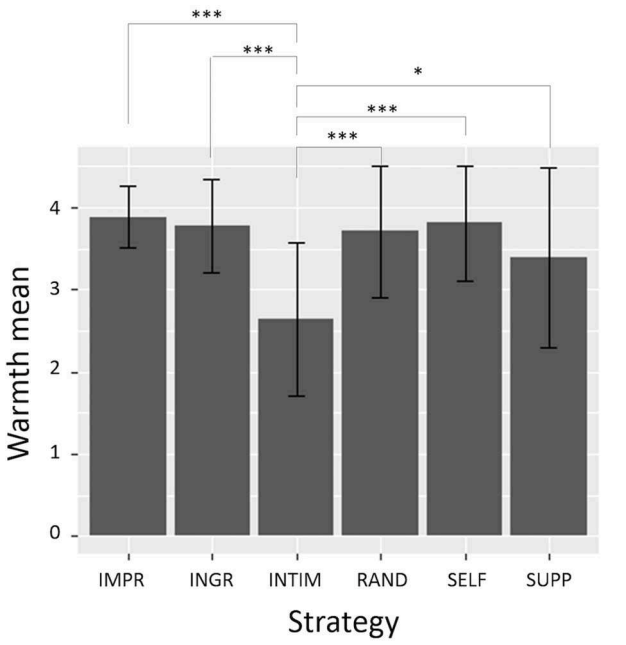



**FIGURE 6 |** Mean and SD values of warmth ratings for each level of **Strategy**. **INITM** scores are significantly lower than any other condition. Significance levels: $^{*}p < 0.05$, $^{***}p < 0.005$.

in **INGR** and **SELF** conditions being higher than 3. A Mann-Whitney *U*-Test showed a statistically significant difference according to Nars scores ($U = 998$, $p = 0.001$): participants who got high scores in the Nars questionnaire were more motivated to continue the interaction ($M = 3.28$, $SD = 1.12$) than people who got low scores in the Nars questionnaire ($M = 2.36$, $SD = 1.13$).

Concerning **like** scores, a Kruskal-Wallis rank test showed a very near to significance difference according to **Strategy** [$H_{(5)} = 10.99$, $p = 0.05$]. In particular, Dunn's test for multiple comparisons found that **INGR** scores were significantly higher ($M = 3.75$, $SD = 0.62$) than **INTIM** ($M = 2.62$, $SD = 0.96$; $z = 2.87$, p-adj $= 0.03$) (see **Figure 7C**). No differences were found between **IMPR** scores and the other conditions. In addition, a statistically significant difference between scores was found according to Nars scores ($U = 970$, $p = 0.003$): participants who got high scores in the Nars questionnaire liked Alice more ($M = 3.62$, $SD = 0.91$) than people who got low scores in the Nars questionnaire ($M = 2.92$, $SD = 0.99$).

**TABLE 5 |** Mean and standard deviation of competence scores for each level of **Strategy**.

| Condition | Competence mean ± SD |
|---|---|
| **INGR** | 3.6 ± 0.62 |
| **SUPP** | 2.98 ± 0.77 |
| **SELF** | 3.75 ± 0.63 |
| **INTIM** | 3.65 ± 0.79 |
| **RAND** | 3.5 ± 0.70 |
| **IMPR** | 3.43 ± 0.76 |

*No significant differences among the conditions were found.*

Concerning **learnfrom, exhib**, and **rship**, no significant differences in scores were found according to any variable. Participants' scores about **learnfrom** and **exhib** were all over the mean value, while for **rship** the mean scores for each condition were quite low (all means $\leq 2.75$), suggesting that participants considered Alice as very distant from them.

Concerning **likeperson** scores, no significant differences were found according to **Strategy**. Mean scores for each condition were quite low (all means $\leq 2.25$), suggesting that in general Alice was perceived more similar to a computer than a person. A Mann-Whitney *U*-Test showed a statistically significant difference according to Nars scores ($U = 1028$, $p = 0.0003$): participants who got high scores in the Nars questionnaire perceived Alice less closed to a computer ($M = 2.49$, $SD = 1.12$) than people who got low scores in the Nars questionnaire ($M = 1.58$, $SD = 0.69$).

On the whole, these results do not allow us to validate **H2a**, but agent's adaptation was found to have at least an effect on its level of warmth (**H2b**, see section 5.6.1).

### 5.6.4. Verbal Cues of Engagement

Only one person gave a negative answer to Topic1_question, while people gave different responses to Topic2_question. In general, participants which did not give much verbal feedback (i.e., <13 reactions over all the speaking turns) gave a positive answer to this question ($OR = 4.27$, $p = 0.04$). In addition, we found that ratings about **likeperson** item were significantly lower for people giving much verbal feedback ($M = 1$, $SD = 0$) compared to those of people who did not talk a lot ($M = 2.16$, $SD = 1.07$; $U = 36.5$, $p = 0.02$). This means that, even than in general users found the agent closer to a computer than to a real person, all the people who gave much verbal feedback during the interaction perceived totally agreed with this definition. No differences in any of the dependent variables were found according to **Strategy**.

## 6. DISCUSSION

In this section we discuss the details of the results of our evaluation study.

First of all, regarding **H1**, the only statistically significant results concern the perception of agent's warmth. Alice was rated as colder when she adopted **INTIM** strategy, compared to the other conditions. This supports the thesis of the primacy

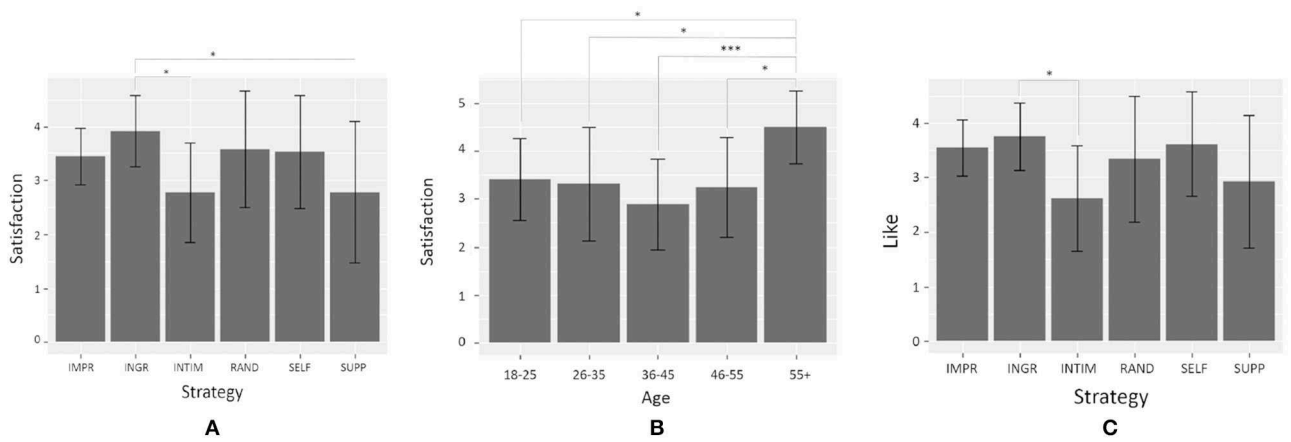



**FIGURE 7 |** Mean values with sd for the different items of **perception** where an effect of **Strategy** and age was found. Significant results of Dunn's test for multiple comparisons are reported, with the following significance levels: $^{*}p < 0.05$, $^{***}p < 0.001$: **(A)** mean values of **satisfaction** for each level of **Strategy**; **(B)** mean values of **satisfaction** for each age range; **(C)** mean values of **like** for each level of **Strategy**.

of warmth dimension (Wojciszke and Abele, 2008, see section 2), and it is in line with the positive-negative asymmetry effect described by Peeters and Czapinski (1990), who argues that negative information has generally a higher impact in person perception than positive information. In our case, when the agent displays cold (i.e., low warmth) behaviors (i.e., in **INTIM** condition), it is judged by participants with statistically significant lower ratings of warmth. Regarding the other conditions (**INGR**, **SUPP**, **SELF**, **IMPR**, and **RAND**), they elicited warmer impressions in the user, but there is not one strategy better than the others in this regard. The fact that also the **SELF** elicited the same level of warmth than the others could reflect an halo effect: the behaviors displayed to appear competent influenced its warmth perception in the same direction.

Regarding **H2**, the results do not validate our hypothesis **H2a** that the interaction is improved when the virtual agent manages its impressions by adapting its strategy according to user's engagement. When analyzing scores for **perception** items, we found that participants were more satisfied by the interaction and they liked Alice more when the agent wanted to be perceived as warm (i.e., in **INGR** condition), compared to when it wanted to be perceived cold and competent (i.e., in **INTIM** condition). An hypothesis is that since the agent was perceived warmer in **INGR** condition, it could have positively influenced the ratings of the other items, like **satisfaction**. Concerning **H2b** about a possible effect of agent's adaptation on user's perception of its W&C, it is interesting to see that when the agent adapts its self-presentational strategy according to user's overall engagement, it is perceived as warm. This highlights a link between agent's adaptation, user's engagement and warm impression: the more the agent adapts its behaviors, the more the user is engaged and the more s/he perceives the agent as warm.

When looking at participants' verbal cues of engagement (see section 5.6.4), we could divide people into two groups: those who gave much verbal feedback during the speaking turns, and those who mainly answered to agent's questions and did not talk during the rest of the interaction. Participants talking a lot may ask questions to the agent, give their opinion on a game, etc. Since the agent is not endowed with natural language understanding capacities, it could not answer participant's request, nor could it argument on user's opinion. Even though we did not explain agent's limitation to participants before starting the experiment, users who gave many feedback at the beginning of the interaction often became aware that the agent could not react to their speech, since it did not consider what they said, interrupt them, continue talking on its topic as if the participants had not talked. This could had a negative effect on their experience and had led them to choose not to continue to discuss with the agent. When looking at the interaction with this group of people, we notice that they stop proving feedback after the virtual agent missed answering them properly. There is a clear distinction in their verbal behaviors before and after the agent missed their input. In our quantitative analyses we found that the majority of people replying a lot to the agent often gave a negative answer to the question Topic2_question asked by the agent about continuing the discussions. On the other hand, people who did not talk a lot had less probability to experience weird situations such as asking a question to the agent and not being heard. These people were less disappointed than the others and more likely to accept to continue the interaction. Indeed, according to our results, the majority of people who did not give much verbal feedback gave a positive answer to the question Topic2_question. This hypothesis that participants giving much feedback at the beginning of the interaction discovered the limits of the agent seems in line with the lower scores found for **likeperson** item given by people talking a lot compared to the others. The fact that the agent did not behave in the appropriate way and that the agent did not stand up to their expectancies could have highlighted even more the fact that they were in front of a system that simulates a "mock" of interaction. Another possible explanation to this result could concern the fact that people who did not talk a lot were intimidated and so they did not dare to give a negative answer

to the agent. This could be in line too with the results about **likeperson** item: considering the agent closer to a person, they could have answered "yes" as not to offend, somehow, the agent.

In this discussion we should take into account how participants' expectancies may affect their perception of the interaction. People expectancies about others' behaviors have already been demonstrated to affect human-human interaction (Burgoon, 1993), as well as when people are in front of an ECA (Burgoon et al., 2016; Biancardi et al., 2018). In this study we found some effects of people's a priori about virtual character: people who got higher scores in the Nars questionnaire generally perceived the agent warmer, compared to people who got lower scores in the Nars questionnaire. In addition, it should not be forgotten that the fact of being in a Sciences museum, combined with people exposition to films and TV shows about artificial intelligence could have had a strong impact on participants' expectancies. People could have difficulties in distinguishing between what is shown in science-fiction films and the current state of the technology of interactive ECAs. Thus, people could have exaggerated expectancies about our virtual agent's capabilities. These expectancies, and the related disappointment showed by some participants when interacting with a less sophisticated virtual character, could have become an uncontrollable variable preventing any other effect of the independent variables of our experiment. Nevertheless, it has to be remembered that in this experiment we mainly focused on the non-verbal behaviors rather than on the dialogical dimension, limiting therefore the dialogue complexity to better control the other variables. The agent had the floor during the majority of the interaction; our system took into account the polarity of user's answers only at 2 specific moments, Topic1_question and Topic2_question (see section 5.5, thus the variability of the agent's dialogue was very limited.

## 7. CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

In this paper, we presented a computational model for an Embodied Computational Agent, aimed at managing its self-presentational intentions eliciting different impressions of warmth and competence, in order to maximize user's engagement during the interaction. We built an architecture which takes as input participants facial Action Units, torso and head rotation, use them to compute user's overall engagement and sends it to the dialog manager of the agent. Through a reinforcement learning algorithm which takes user's engagement as reward, the agent can select the self-presentational intention which maximizes user's engagement. In order to evaluate the system, we conceived an interaction scenario where the agent played a role of museum guide. In the experiment we manipulated how the agent selected its self-presentational intention at each speaking turn. It could adapt its behavior by using the reinforcement learning algorithm, or choose it randomly, or use the same self-presentational intention during the whole interaction. The agent which adapted its behavior to maximize user's engagement was perceived as warm by participants, but we did not find any effect of agent's adaptation on users' evaluation of the interaction.

We are aware of some limitations of our system: we discuss them in the following paragraph, and suggest some future improvements to deal with these limitations. First of all, many participants did not like the virtual character, as we can see from their answers to the questionnaires, as well as from their comments during the debriefing. They reported their disappointment about the quality of the animation and of the voice of the agent. They described the experience as "disturbing," "creepy." So probably their very first impression about the appearance and the voice of the agent was too strong and affected the rest of the experience. During the interaction, participants did not show many non-verbal behaviors. This could be due to the setup of the experiment, where participants stood in front of the screen and the virtual agent was displayed at human size. According to their comments, many people were a bit frightened by the dimension of the agent and for almost all of them it was their first interaction with an ECA. Many of them stared at the ECA without moving much. They did not vary their facial expression, move their head or gesture. Since our overall engagement detection module relies on the interpretation of non-verbal behaviors, the lack of behavioral change impacts directly the output values it returns.

In our work, we have done qualitative analyses and some quantitative ones. In the future, it would be interesting to conduct further quantitative measures, such as analyzing facial expressions, gaze direction and posture of the participants to measure phenomena like synchronization and alignment. This will allow us to have a complementary measure to their subjective evaluation.

One of the main limits of the interaction was that agent's strategies did not focus on building a rapport with the participant: it just managed its impressions of warmth and competence without considering the social relation with the user. Rapport, meant as the feeling of harmony and connection with another, is an important aspect of human interaction, as well as of human-agent interaction (Gratch et al., 2007; Zhao et al., 2016). Agent's self-presentational intentions should take into account this dimension, at both verbal and non-verbal level. For example, we could include some conversational strategies such as self-disclosure, enhance the gaze behavior of the agent to improve mutual attentiveness, and provide agent's non-verbal listening feedback, such as postural mimicry and synchronization of its movements with the user's ones.

## DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

## ETHICS STATEMENT

This study was exempt from the above requirements, as for the French law, ethical approval is required only for experiments involving invasive interventions on human subjects, which was

not the case of this study. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol followed the ethical guidelines of the Institut des Systémes Intelligents et de Robotique.

## AUTHOR CONTRIBUTIONS

BB wrote the manuscript with support and feedback from all the other authors. More specifically, MM collaborated to the implementation of the system and was in charge of the system architecture section. CP globally supervised the manuscript. PL was the main contributor of the design, implementation of the system, and collaborated to run the experiment.

## FUNDING

This work was supported by ANR IMPRESSIONS project number ANR-15-CE23-0023.

## ACKNOWLEDGMENTS

Authors want to thanks the Carrefour Numérique of Cité des sciences et de l'industrie for hosting the evaluation study and proving technical material for the experiment. MM carried out the work described in this paper while being at the Casa Paganini–InfoMus Research Centre of the University of Genoa (Italy). He would like to thank Prof. Antonio Camurri for his support.

## REFERENCES

Abele, A. E., and Wojciszke, B. (2013). The big two in social judgment and behavior. *Soc. Psychol.* 44, 61–62. doi: 10.1027/1864-9335/a000137

Aragonés, J. I., Poggio, L., Sevillano, V., Pérez-López, R., and Sánchez-Bernardos, M.-L. (2015). Measuring warmth and competence at inter-group, interpersonal and individual levels/medición de la cordialidad y la competencia en los niveles intergrupal, interindividual e individual. *Revista de Psicología Soc.* 30, 407–438. doi: 10.1080/02134748.2015.1065084

Asch, S. E. (1946). Forming impressions of personality. *J. Abnorm. Soc. Psychol.* 41:258.

Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)* (Lake Placid, NY: IEEE), 1–10.

Bayes, M. A. (1972). Behavioral cues of interpersonal warmth. *J. Consult. Clin. Psychol.* 39:333.

Bergmann, K., Eyssel, F., and Kopp, S. (2012). "A second chance to make a first impression? how appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time," in *International Conference on Intelligent Virtual Agents* (Santa Cruz, CA: Springer), 126–138.

Biancardi, B., Cafaro, A., and Pelachaud, C. (2017). "Analyzing first impressions of warmth and competence from observable nonverbal cues in expert-novice interactions," in *19th ACM International Conference on Multimodal Interaction* (Glasgow).

Biancardi, B., Cafaro, A., and Pelachaud, C. (2018). "Étude des effets de différents types de comportements non-verbaux sur la perception d'un agent virtuel," in *Workshop Affect, Compagnon Artificiel, Interaction (WACAI)* (Porquerolles).

Bickmore, T., Pfeifer, L., and Schulman, D. (2011). "Relational agents improve engagement and learning in science museum visitors," in *International Workshop on Intelligent Virtual Agents* (Berlin: Springer), 55–67.

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., and Efros, A. A. (2018). Large-scale study of curiosity-driven learning. *arXiv:1808.04355*.

Burgoon, J. K. (1993). Interpersonal expectations, expectancy violations, and emotional communication. *J. Lang. Soc. Psychol.* 12, 30–48.

Burgoon, J. K., Bonito, J. A., Lowry, P. B., Humpherys, S. L., Moody, G. D., Gaskin, J. E., et al. (2016). Application of expectancy violations theory to communication with and judgments about embodied agents during a decision-making task. *Int. J. Hum. Comput. Stud.* 91, 24–36. doi: 10.1016/j.ijhcs.2016.02.002

Cafaro, A., Vilhjálmsson, H. H., and Bickmore, T. (2016). First impressions in human–agent virtual encounters. *ACM Trans. Comput. Hum. Int.* 23:24. doi: 10.1145/2940325

Cafaro, A., Wagner, J., Baur, T., Dermouche, S., Torres Torres, M., Pelachaud, C., et al. (2017). "The noxi database: multimodal recordings of mediated novice-expert interactions," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (Glasgow), 350–359.

Callejas, Z., Ravenet, B., Ochs, M., and Pelachaud, C. (2014). "A computational model of social attitudes for a virtual recruiter," *13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2014* (Paris).

Campano, S., Langlet, C., Glas, N., Clavel, C., and Pelachaud, C. (2015). "An eca expressing appreciations," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (Xian: IEEE), 962–967.

Camurri, A., Coletta, P., Massari, A., Mazzarino, B., Peri, M., Ricchetti, M., et al. (2004). "Toward real-time multimodal processing: eyesweb 4.0," in *2004 Convention: Motion. Emotion and Cognition Proceedings of the Artificial Intelligence and the Simulation of Behaviour (AISB)* (Leeds: Citeseer), 22–26.

Cassell, J. (2000). *Embodied Conversational Agents*. Boston, MA: MIT press.

Clavel, C., Cafaro, A., Campano, S., and Pelachaud, C. (2016). "Fostering user engagement in face-to-face human-agent interactions: a survey," in *Toward Robotic Socially Believable Behaving Systems-Volume II* (Cham: Springer), 93–120.

Corrigan, L. J., Peters, C., Küster, D., and Castellano, G. (2016). "Engagement perception and generation for social robots and virtual agents," in *Toward Robotic Socially Believable Behaving Systems-Volume I* (Cham: Springer), 29–51.

Cuddy, A. J., Fiske, S. T., and Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Adv. Exp. Soc. Psychol.* 40, 61–149. doi: 10.1016/S0065-2601(07)00002-0

Dermouche, S., and Pelachaud, C. (2018). "From analysis to modeling of engagement as sequences of multimodal behaviors," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)* (Miyazaki).

Doherty, K., and Doherty, G. (2018). Engagement in hci: conception, theory and measurement. *ACM Comput. Surv.* 51, 99:1–99:39. doi: 10.1145/3234149

Fiske, S. T., Cuddy, A. J., and Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends Cogn. Sci.* 11, 77–83. doi: 10.1016/j.tics.2006.11.005

Fiske, S. T., Cuddy, A. J., Glick, P., and Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *J. Personal. Soc. Psychol.* 82:878. doi: 10.1037//0022-3514.82.6.878

Glas, N., and Pelachaud, C. (2015a). "Definitions of engagement in human-agent interaction," in *International Workshop on Engagment in Human Computer Interaction (ENHANCE)* (Xian), 944–949.

Glas, N., and Pelachaud, C. (2015b). Politeness versus perceived engagement: an experimental study. *Nat. Lang. Proc. Cogn. Scie. Proc.* 2014:135. doi: 10.1515/9781501501289.135

Goffman, E. et al. (1978). *The Presentation of Self in Everyday Life*. London: Harmondsworth.

Gordon, G., Spaulding, S., Westlund, J. K., Lee, J. J., Plummer, L., Martinez, M., et al. (2016). "Affective personalization of a social robot tutor for children's second language skills," in *Thirtieth AAAI Conference on Artificial Intelligence* (Phoenix, AZ), 3951–3957.

Gratch, J., Wang, N., Gerten, J., Fast, E., and Duffy, R. (2007). "Creating rapport with virtual agents," in *International Workshop on Intelligent Virtual Agents* (Paris: Springer), 125–138.

Jones, E. E., and Pittman, T. S. (1982). Toward a general theory of strategic self-presentation. *Psychol. Perspect. Self* 1, 231–262.

Judd, C. M., James-Hawkins, L., Yzerbyt, V., and Kashima, Y. (2005). Fundamental dimensions of social judgment: understanding the relations between judgments of competence and warmth. *J. Personal. Soc. Psychol.* 89:899. doi: 10.1037/0022-3514.89.6.899

Katehakis, M. N., and Veinott, A. F. Jr. (1987). The multi-armed bandit problem: decomposition and computation. *Math. Operat. Res.* 12, 262–268.

Kervyn, N., Bergsieker, H. B., Grignard, F., and Yzerbyt, V. Y. (2016). An advantage of appearing mean or lazy: amplified impressions of competence or warmth after mixed descriptions. *J. Exp. Soc. Psychol.* 62, 17–23. doi: 10.1016/j.jesp.2015.09.004

Liu, C., Conn, K., Sarkar, N., and Stone, W. (2008). Online affect detection and robot behavior adaptation for intervention of children with autism. *IEEE Trans. Robot.* 24, 883–896. doi: 10.1109/TRO.2008.2001362

Maricchiolo, F., Gnisci, A., Bonaiuto, M., and Ficca, G. (2009). Effects of different types of hand gestures in persuasive speech on receivers' evaluations. *Lang. Cogn. Process.* 24, 239–266. doi: 10.1080/01690960802159929

Nguyen, T.-H. D., Carstensdottir, E., Ngo, N., El-Nasr, M. S., Gray, M., Isaacowitz, D., et al. (2015). "Modeling warmth and competence in virtual characters," in *International Conference on Intelligent Virtual Agents* (Delft: Springer), 167–180.

Nomura, T., Kanda, T., and Suzuki, T. (2006). Experimental investigation into influence of negative attitudes toward robots on human–robot interaction. *Ai Soc.* 20, 138–150. doi: 10.1007/s00146-005-0012-7

Pecune, F., Cafaro, A., Chollet, M., Philippe, P., and Pelachaud, C. (2014). "Suggestions for extending saiba with the vib platform," in *W'shop Architectures and Standards for IVAs, Int'l Conf. Intelligent Virtual Agents* (Boston, MA: Citeseer), 16–20.

Peeters, G., and Czapinski, J. (1990). Positive-negative asymmetry in evaluations: the distinction between affective and informational negativity effects. *Eur. Rev. Soc. Psychol.* 1, 33–60.

Pennebaker, J. W. (2011). The secret life of pronouns. *New Sci.* 211, 42–45. doi: 10.1016/S0262-4079(11)62167-2

Peters, C., Pelachaud, C., Bevacqua, E., Mancini, M., Poggi, I., and Tre, U. R. (2005). "Engagement capabilities for ecas," in *AAMAS'05 Workshop Creating Bonds With ECAs* (Utrecht).

Poggi, I. (2007). *Mind, Hands, Face and Body: A Goal and Belief View of Multimodal Communication*. Berlin: Weidler.

Ritschel, H., Baur, T., and André, E. (2017). "Adapting a robot's linguistic style based on socially-aware reinforcement learning," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (Lisbon: IEEE), 378–384.

Rosenberg, S., Nelson, C., and Vivekananthan, P. (1968). A multidimensional approach to the structure of personality impressions. *J. Personal. Soc. Psychol.* 9:283.

Sidner, C. L., and Dzikovska, M. (2005). "A first experiment in engagement for human-robot interaction in hosting activities," in *Advances in Natural Multimodal Dialogue Systems*, eds J. Van Kuppevelt, L. Dybkjaer, and N. Ole Bernsen (Dordrecht: Springer), 55–76.

Truong, K. P., Poppe, R., and Heylen, D. (2010). "A rule-based backchannel prediction model using pitch and pause information," in *Eleventh Annual Conference of the International Speech Communication Association*.

van Waterschoot, J., Bruijnes, M., Flokstra, J., Reidsma, D., Davison, D., Theune, M., et al. (2018). "Flipper 2.0: a pragmatic dialogue engine for embodied conversational agents," in *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (Sydney: ACM), 43–50.

Willis, J., and Todorov, A. (2006). First impressions making up your mind after a 100-ms exposure to a face. *Psychol. Sci.* 17, 592–598. doi: 10.1111/j.1467-9280.2006.01750.x

Wojciszke, B., and Abele, A. E. (2008). The primacy of communion over agency and its reversals in evaluations. *Eur. J. Soc. Psychol.* 38, 1139–1147. doi: 10.1002/ejsp.549

Wojciszke, B., Bazinska, R., and Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personal. Soc. Psychol. Bull.* 24, 1251–1263.

Yzerbyt, V. Y., Kervyn, N., and Judd, C. M. (2008). Compensation versus halo: the unique relations between the fundamental dimensions of social judgment. *Personal. Soc. Psychol. Bull.* 34, 1110–1123. doi: 10.1177/01461672083 18602

Zhao, R., Sinha, T., Black, A., and Cassell, J. (2016). "Automatic recognition of conversational strategies in the service of a socially-aware dialog system," in *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (Los Angeles, CA), 381–392.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 26 June 2019
doi: 10.3389/frobt.2019.00049

# What Can You See? Identifying Cues on Internal States From the Movements of Natural Social Interactions

*Madeleine E. Bartlett*[1]*, *Charlotte E. R. Edmunds*[2], *Tony Belpaeme*[1,3], *Serge Thill*[4,5] and *Séverin Lemaignan*[6]

[1] Centre for Robotics and Neural Systems (CRNS), University of Plymouth, Plymouth, United Kingdom, [2] Warwick Business School, University of Warwick, Coventry, United Kingdom, [3] ID Lab—imec, University of Ghent, Ghent, Belgium, [4] Interaction Lab, School of Informatics, University of Skövde, Skövde, Sweden, [5] Donders Institute for Brain, Cognition, and Behavior, Radboud University, Nijmegen, Netherlands, [6] Bristol Robotics Lab, University of the West of England, Bristol, United Kingdom

OPEN ACCESS

**Edited by:**
*Cigdem Beyan, Istituto Italiano di Tecnologia, Italy*

**Reviewed by:**
*Radoslaw Niewiadomski, University of Genoa, Italy*
*Atesh Koul, Istituto Italiano di Tecnologia, Italy*
*Giulia Perugia, Uppsala University, Sweden*
*Jordi Vallverdu, Autonomous University of Barcelona, Spain*

***Correspondence:**
*Madeleine E. Bartlett*
*madeleine.bartlett@plymouth.ac.uk*

**Specialty section:**
*This article was submitted to Human-Robot Interaction, a section of the journal Frontiers in Robotics and AI*

**Received:** *12 January 2019*
**Accepted:** *06 June 2019*
**Published:** *26 June 2019*

**Citation:**
*Bartlett ME, Edmunds CER, Belpaeme T, Thill S and Lemaignan S (2019) What Can You See? Identifying Cues on Internal States From the Movements of Natural Social Interactions. Front. Robot. AI 6:49. doi: 10.3389/frobt.2019.00049*

In recent years, the field of Human-Robot Interaction (HRI) has seen an increasing demand for technologies that can recognize and adapt to human behaviors and internal states (e.g., emotions and intentions). Psychological research suggests that human movements are important for inferring internal states. There is, however, a need to better understand what kind of information can be extracted from movement data, particularly in unconstrained, natural interactions. The present study examines which internal states and social constructs humans identify from movement in naturalistic social interactions. Participants either viewed clips of the full scene or processed versions of it displaying 2D positional data. Then, they were asked to fill out questionnaires assessing their social perception of the viewed material. We analyzed whether the full scene clips were more informative than the 2D positional data clips. First, we calculated the inter-rater agreement between participants in both conditions. Then, we employed machine learning classifiers to predict the internal states of the individuals in the videos based on the ratings obtained. Although we found a higher inter-rater agreement for full scenes compared to positional data, the level of agreement in the latter case was still above chance, thus demonstrating that the internal states and social constructs under study were identifiable in both conditions. A factor analysis run on participants' responses showed that participants identified the constructs *interaction imbalance*, *interaction valence* and *engagement* regardless of video condition. The machine learning classifiers achieved a similar performance in both conditions, again supporting the idea that movement alone carries relevant information. Overall, our results suggest it is reasonable to expect a machine learning algorithm, and consequently a robot, to successfully decode and classify a range of internal states and social constructs using low-dimensional data (such as the movements and poses of observed individuals) as input.

**Keywords: social psychology, human-robot interaction, machine learning, social interaction, recognition**

# 1. INTRODUCTION

One of the main goals in the field of Human-Robot Interaction (HRI) is to create robots capable of recognizing and adapting to human interaction partners in an appropriate manner (Dautenhahn and Saunders, 2011). In human-human interactions, the appropriateness of our responses to others is often a result of our ability to recognize the internal states (e.g., intentions, dispositions) of our interaction partner (Domes et al., 2007). Here we focus on internal states and social constructs relevant to task engagement and social relations between interaction partners. For example, we consider states that can be thought of as dispositional judgments (e.g., friendliness), states which can be considered emotional and are embedded within a social context (e.g., aggression), and states relevant to task performance (e.g., boredom). These states are communicated through both verbal and non-verbal cues (Pollick et al., 2001; Manera et al., 2011). Endowing robots and behavior classification systems with a similar ability to recognize internal states based on non-verbal behaviors would allow for more appropriate, autonomous human-robot interactions (Breazeal et al., 2009; Vernon et al., 2016), and for classification systems to provide more detailed insights into human behavior, e.g., for security purposes (Gowsikhaa et al., 2014).

## 1.1. Internal State Recognition

HRI research exploring approaches to achieving on-line recognition of human internal states/behavior draws on our understanding of how humans themselves infer internal states and social constructs. For example, a rich history of research has led to the assumption that humans are able to infer the internal states of others by observing their actions and movements (Gallese and Goldman, 1998; Manera et al., 2011; Quesque et al., 2013) and facial expressions (Ekman and Friesen, 1971; Haidt and Keltner, 1999; Tracy and Robins, 2008). In their paper, Manera et al. (2011) claim that *"in some circumstances, the movement of a human body... is sufficient to make judgments... in relation to the actor's intention"* [p. 548]. The idea here is that our intentions or emotions influence differences in the movements we make and, as observers, we are able to pick up on these differences and use them to infer the internal state of the person performing the action (Pollick et al., 2001; Ansuini et al., 2014; Becchio et al., 2017). To examine this researchers have used point-light displays and other methods to isolate movement information from other sources of information. Point-light displays denote the position and movements of an actor's joints on an otherwise blank display. Studies using this type of stimulus have shown that humans are able to use observed movement to infer an actor's gender (Kozlowski and Cutting, 1977; Mather and Murdoch, 1994; Hufschmidt et al., 2015), intention (Manera et al., 2010; Quesque et al., 2013) and emotional state (Pollick et al., 2001; Alaerts et al., 2011).

Available evidence also suggests that internal states and social constructs which fall under our definition of being socially relevant, dispositional or related to task engagement/performance are recognizable from observable movement. Okada et al. (2015) found that observable movements and non-verbal audio information produced during spontaneous, naturalistic interactions were sufficient for classifying dispositions and social behaviors such as dominance and leadership. Similarly, Sanghvi et al. (2011) demonstrated that postural behaviors could be used to classify a child's engagement with a robotic opponent, with which the children are playing a game. Beyan et al. (2016) asked four unacquainted individuals to complete a group decision task. They found that a classifier, when fed the 3D positional data of the interaction, was able to identify leaders within the group based on head pose and gaze direction information. Sanchez-Cortes et al. (2011) applied a computational framework to the inference of leadership and related concepts (e.g., dominance, competence) from non-verbal behaviors in a group interaction. Interactions in this study took place between four previously unacquainted individuals whose interactions were spontaneous and minimally structured. Sanchez-Cortes and colleagues were able to identify which behaviors were most informative for the recognition of the different leadership concepts. For example, conversational turn-taking and body movement behaviors were found to be the most informative for inferring leadership, whereas head activity and vocal pitch were the most informative for inferring competence.

States which are socially relevant, dispositional or task related, (such as friendliness, dominance or engagement) are particularly relevant for HRI research where the aim is to provide a socially interactive agent. In such scenarios it is preferable to have an agent which can provide appropriate social behaviors and responses (Dautenhahn and Saunders, 2011). Whilst emotion and intention recognition are definitely important for generating appropriate autonomous social behaviors from a robot, some HRI scenarios would also benefit from an ability to recognize internal states as we have defined them here. For instance, a teaching robot, such as those developed by the L2TOR project (Belpaeme et al., 2015), would be better able to provide appropriately timed encouragements or prompts if able to recognize when a student is bored or not engaged with the learning task.

As a result, HRI researchers have begun exploring ways in which observed movement can be utilized by robots and artificial systems to enable automated interpretation of, and responding to, the internal states of humans (Schrempf and Hanebeck, 2005; Han and Kim, 2010). Whilst humans also use other cues such as tone of voice (Walker-Andrews, 1997), findings such as those described above suggest that movement information may be sufficient for recognizing some, if not all, human internal states.

## 1.2. Current Study

### 1.2.1. Motivation and Approach

To take advantage of this information for the purposes of internal state recognition it is important to first identify what internal state information is available in movements and body postures. This knowledge is particularly useful for streamlining the design process for a robot or classifier able to interpret such data. For example, if we want to design a system able to recognize when a human is bored, we first need to know what data is sufficient, if not optimal, for recognizing this state. Would the system need

to take multiple behaviors into account, e.g., movements and prosodic features, or would movement alone be enough? In the case of internal states such as emotions and intentions, previous research suggests that movement information is sufficient for gaining insight (e.g., Tracy and Robins, 2008; Manera et al., 2011; Quesque et al., 2013). Given that the aim of HRI research is to create systems and robots which can be deployed in the real world, it is also important to consider that a classifier must be able to deal with natural, spontaneous human behaviors. Consequently, it is important to explore whether (and which) internal states can be recognized from the movements produced in natural human interactions. A a growing pool of studies have examined this (e.g., Sanchez-Cortes et al., 2011; Sanghvi et al., 2011; Shaker and Shaker, 2014; Okada et al., 2015; Beyan et al., 2016; Okur et al., 2017; Kawamura et al., 2019). However, further research is needed to provide a better understanding of which internal states can be inferred from such movements.

We therefore propose that an exploration into how readily different types of internal states can be identified from naturalistic human behavior would be beneficial for the streamlining of future HRI research. That is, by identifying which internal states are best recognized from a particular behavioral modality (e.g., biological motion), future research can identify which data sources are most useful for a given recognition task.

This study takes the first steps in this direction by developing a method for determining which internal state information is reported as identifiable by humans when they observe people in natural interactions. Given the strength of evidence suggesting that movement information is useful for identifying emotional and other internal states or social constructs (e.g., Pollick et al., 2001; Gross et al., 2012; Quesque et al., 2013; Beyan et al., 2016), this modality is likely to be a rich source of internal state information. Further, by extending this work to naturalistic interactions, we will find which internal states are likely to be identified in more ecologically valid settings. The usefulness of these states to HRI, indicate that an exploration of which internal states, from a selection of several, are recognizable from human movements would be helpful in guiding future research and development. To address this, we aim to examine and compare how reliably humans report identifying a number of different internal states and social constructs from observable movements.

To summarize, the main aim of this study is to demonstrate a method for identifying: (1) whether the data source of choice (in this case observable movements) can be used by humans to infer internal states and social constructs, and (2) what internal states and social constructs are readable from the movements within the data set. To do so, we will present short video clips of social interactions (exhibiting seven different internal states and social constructs) to participants. These clips come from the PInSoRo (Lemaignan et al., 2017) data set made openly available by our group[1]. This data set consists of videos of child-child or child-robot interactions. Children were asked to play for as long as they wanted on a touch-screen table-top device. For this study, we will solely use the child-child interactions as these are more likely to involve spontaneous behaviors throughout the children's interactions with one another. Some participants will view short clips including the full visual scene (full-scene condition) and others clips containing only movement and body posture information (movement-alone condition). These clips will contain at least one noticeable internal state (for details of the selection process see the Method section). Following each clip, participants respond to a series of questions where they can describe the internal states (e.g., boredom, friendliness) or social constructs (e.g., cooperation, dominance) they identified in the children's behaviors. By comparing responses in each condition we expect to be able to identify constructs which are likely to be recognizable from movement information alone.

#### 1.2.2. Hypotheses and Predictions

Based on previous findings that humans are able to recognize internal states such as emotions (Gross et al., 2012) and group dynamics such as leadership (Beyan et al., 2016) from human motion information, we expect the following:

1. Participants will report being able to draw internal state information from the movement-alone videos (Hypothesis 1). Specifically, we predict that even in the impoverished movement-alone condition, the provided ratings will be sufficient to describe the internal states and social constructs identified in the observed interaction. This can be tested by training a classifier on the full-scene ratings, and assessing its performance when tested on the movement-alone ratings.
2. However, given that participants in this condition are provided with fewer visual cues than those viewing the full-scene videos (e.g., lack of resolution for facial expressions) we expect a higher recognition error rate in the movement-alone condition compared to the full-scene condition (Hypothesis 2). If this is the case, we predict that inter-rater agreement levels amongst participants will be above chance in both conditions (i.e. the same constructs are robustly identified in the clips by the participants), but with higher levels of agreement in the full-scene condition.

## 2. METHOD

### 2.1. Design and Participants

This study examined the effect of video type (full-scene vs. movement-alone) on responses to questions about the nature of the interaction depicted in the videos. We used a between-subject design: participants saw either *full-scene* clips (**Figure 1**, left) or *movement-alone* clips (**Figure 1**, right). 284 participants were recruited from Amazon's Mechanical Turk (MTurk). A total of 85 participants were excluded from analysis due to incorrect answers to an attention check (discussed in Procedure), leaving 199 participants (see **Table 1** for demographics). All participants were remunerated $1 (USD) upon completion of the experiment.

### 2.2. Materials

The stimuli used for this experiment were extracted from the PInSoRo data set. This data set contains videos (up to 40 min long) of pairs of children interacting whilst playing on a touch-screen table-top. For the present study we extracted twenty 30 s clips from these videos. We wanted to provide participants with

[1]https://freeplay-sandbox.github.io

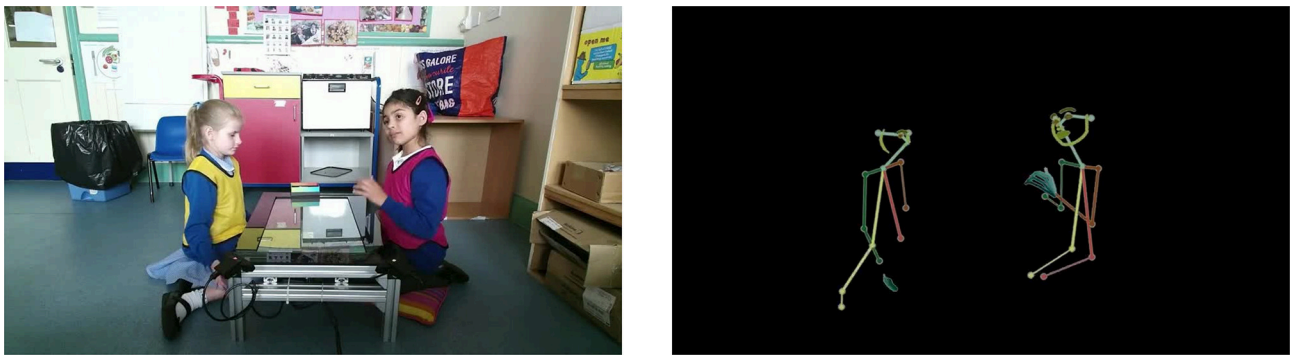


**FIGURE 1 |** Captures of one of the twenty video-clips, *full-scene* condition on the left, *movement-alone* condition on the right. Written consent for these images to be shared was obtained during collection.

**TABLE 1 |** Demographics of participants included in the analyses.

| Condition | *N* | Mean Age (Range) | Gender (%M, %F) | % American | % English First Language |
|---|---|---|---|---|---|
| Movement-Alone | 100 | 34.52 (22–70) | 55%, 44% | 75% | 80% |
| Full-Scene | 99 | 33.54 (19–72) | 65%, 34% | 69% | 73% |
| Both | 199 | 34.03 (19–72) | 60%, 39% | 72% | 76% |

clips which showed both children in the frame at the same time. We therefore selected our stimuli from videos filmed using a camera which had been positioned roughly 1.4m away from the touch-screen table-top, with the table-top in the center of the camera's view, thus allowing for each child to be viewed on either side of the frame (see **Figure 1**, left).

Two versions of the same clips were extracted: the *full-scene* clips were the raw video footage of the children playing, recorded from a static camera (**Figure 1**, left); the *movement-alone* clips were based on the exact same clips, but post-processed to extract skeletal and facial landmarks (using the OpenPose library[2]; Cao et al., 2017). Resulting landmarks were rendered on a black background, and connected to each other using colored lines, so that each child was depicted as a stick-man-style figure (**Figure 1**, right).

Clip selection was made based on whether a notable "event" or social dynamic occurred, defined as the labels listed in **Table 2**. This was done by watching the full-scene clips and working out what internal states and social constructs might be inferred from the children's movements. Specifically, two experimenters selected and labeled clips (by first independently extracting and annotating clips from the PInSoRo dataset, and second discussing to reach consensus) wherein at least one of the following seven concepts described the children's behavior or their interaction in the full-scene clips (see **Table 2**):

1. Boredom - at least one child was bored or not engaging with the task on the touch-screen (e.g., resting head in hand, interacting with touch-screen in slow/lazy manner).
2. Aggression - at least one child exhibited a physical aggressive action either toward the touch-screen or the other child (e.g., hitting the screen, pushing the other child's hand away).
3. Cooperation - the children were working together and/or communicating about how to perform a task [e.g., talking, joint attention (looking at the same object together), nodding].
4. Dominance - one child was bossy, performing most of the actions on the touch-screen or clearly in charge (e.g., pointing to touch-screen and talking at the other child, stopping the other child from using the touch-screen, being the only child to use the touch-screen).
5. Aimless play - at least one child was interacting with the touch-screen in a non-goal-directed manner or without being very engaged in their task (e.g., sitting slightly away from touch-screen whilst still using it, slow/lazy movements on touch-screen, not always looking at what they're doing).
6. Fun - at least one child was having fun (e.g., laughing, smiling).
7. Excitement - at least one child behaved excitedly (e.g., more dynamic than just "having fun," hearty laughter, open smiling mouth, fast movements).

It was decided that multiple labels could be applied to each clip for two reasons. First, the two children in each clip could have behaved in very different ways. Thus, if one child was bored and the other excited, the clip would be assigned both the Boredom and Excitement labels (see **Table 2**). Second, we recognized that a lot can happen in 30 s (the duration of the clips) resulting in changes in the internal states or social constructs which could be inferred from the children's behaviors. For example, an interaction might involve an excited child pushing the other away so they didn't have to share the touch-screen, causing the second child to sit and watch in a manner denoting boredom, this clip could be labeled with Excitement, Aggression and Bored. These labels were selected based on two considerations: (a) the events and internal states which appear available the dataset, and (b) events and internal states which would be useful to a robot which might observe or mediate

[2]https://github.com/CMU-Perceptual-Computing-Lab/openpose/

**TABLE 2 |** Labels that experimenters assigned to each clip during clip selection.

| Clip | Label 1 | Label 2 | Label 3 |
|---|---|---|---|
| 01 | Aggressive | | |
| 02 | Aggressive | Excited | Aimless |
| 03 | Excited | Fun | |
| 04 | Cooperative | | |
| 05 | Bored | Aimless | |
| 06 | Cooperative | | |
| 07 | Dominance | | |
| 08 | Bored | | |
| 09 | Cooperative | | |
| 10 | Cooperative | Dominance | |
| 11 | Cooperative | Dominance | |
| 12 | Aggressive | Aimless | |
| 13 | Excited | Aggressive | Aimless |
| 14 | Aggressive | Fun | |
| 15 | Dominance | | |
| 16 | Cooperative | Dominance | |
| 17 | Excited | Aggressive | |
| 18 | Aggressive | Dominance | |
| 19 | Dominance | | |
| 20 | Excited | | |

such an interaction. Recognizing boredom and aimless behavior would allow a robot to appropriately encourage a child to take part in a task. Recognizing when a child is being dominant or aggressive could provide a robot with cues to mediate and balance the interaction, or request assistance from a human adult (e.g., in the case of aggressive behavior). Recognizing excitement, fun and cooperation could be used to cue positive feedback from the robot, or to signal that the robot need not interject. The selection was made independently by two of the authors, using a consensus method to reach agreement. It is important to note that interactions in this data set were minimally controlled - pairs of children from the same school class were asked to play on a touch-screen table-top for as long as they wanted. Whilst structured play options were provided, they were not enforced. The selected clips were stored on a private server for the duration of the experiment.

Similarly to the selection of clip labels, the questions were constructed by the experimenters based on the types of internal states and social constructs we might want an artificial system to recognize within a scene. The open question was a single item which asked participants *"What did you notice about the interaction?."* The closed questions were a series of 4 unique questions concerning group dynamics, and 13 2-part questions wherein participants were asked the same question twice, once regarding the child on the left and once regarding the child on the right. Each of these 13 pairs were displayed one after the other. Otherwise, the order in which the questions were presented was random (see **Appendix A** for the questions and response options).

It is important to note that the ground-truth of what internal states the children were experiencing during their interactions is not available. As such, neither the labels used for clip selection and labeling, nor the inferences participants provide in their questionnaire responses can be truly validated. The labels were, therefore, also an attempt to work out what naive observers would infer from the videos.

## 2.3. Apparatus

The experiment was designed using the jsPsych library[3], and remotely hosted from a private server (**Figure 2** shows a screenshot of the experiment). The experiment was accessible via Amazon Mechanical Turk (MTurk) to MTurk Workers. An advert was posted on MTurk containing a link to the experiment. The remote/online nature of this study means that we had no control over the physical set-up experienced by the participants.

## 2.4. Procedure

The two video conditions were posted as separate experiments. To ensure that participants did not complete both conditions, the experiments were posted one at a time. Upon opening the experiment participants were asked to provide their MTurk ID and then shown a welcome screen. This was followed by a consent form where participants were asked to provide consent by selecting one of two response options ("I do not consent," or "I do consent"). If participants selected "I do not consent," the experiment would close. If they selected "I do consent" participants were able to press a "Continue" button and proceed to an instruction screen. This was followed by a series of 4 demographic questions (age, nationality, first language and gender). An instruction screen was then presented for a minimum of 3,500 ms, containing the following text:

> *"During this experiment you will be shown 4 30-second clips of children interacting. The children are sat either side of a touch-screen table-top on which they can play a game. Pay particular attention to the way the children interact. After each video you will be asked some questions about what you have watched."*

Participants could then press any button to continue on to the experimental trials.

All participants were asked to complete 4 trials and were presented with the same series of events within each trial. Each trial started with a 30 s clip selected randomly from the list of 20, which was immediately followed by the questions. Upon completion of the fourth trial, participants were shown an additional 2 questions which acted as an attention check (see **Figure 3**). Responses to these questions were used to assess how attentive participants were and how diligently they completed the experiment. Participants who responded incorrectly were excluded from analysis.

Participants then viewed a debrief page which thanked them, explained the purpose of the study and attention-check questions, and provided participants with contact information if they had further questions or desired to withdraw their data. Participants were then provided with a "survey code" which was randomly generated and were instructed that they had completed

[3]https://www.jspsych.org/

Page 1 of 4.

How much do you agree with the following statements?

The children were competing with one another.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The child on the left was sad.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The child on the right was sad.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The child on the left was aggressive.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The child on the right was aggressive.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The children were cooperating with one another.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The child on the left was excited.

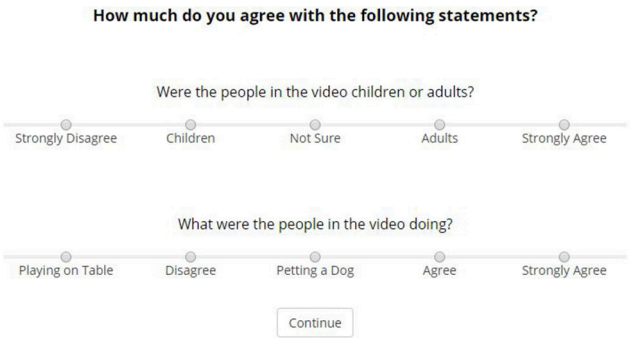
Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The child on the right was excited.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Continue

**FIGURE 2 |** Screenshot of the online experimental setup showing the questionnaire, just after watching the video clip (here in the full-scene condition). The poster image displayed at the top is a static snapshot of the clip. Written consent for these images to be shared was obtained during collection.



How much do you agree with the following statements?

Were the people in the video children or adults?

Strongly Disagree | Children | Not Sure | Adults | Strongly Agree

What were the people in the video doing?

Playing on Table | Disagree | Petting a Dog | Agree | Strongly Agree

Continue

**FIGURE 3 |** Capture of attention check questions presented at the end of the questionnaire. Single correct answer provided. Questions and responses are presented in the same format as the rest of the questions in order to test whether participants read the questions.

the experiment and should now return to the MTurk page in order to submit their survey code. The survey codes participants submitted were later compared to those generated to validate participation and payment was authorized via the MTurk system. The experiment took between 20 and 30 min to complete.

The resulting data set is fully anonymous, and made publicly available at https://github.com/severin-lemaignan/pinsoro-kinematics-study/blob/master/fulldata.csv.

## 3. RESULTS

All data analyses were performed with the Python `pandas` and `sklearn` toolkits. The notebook used for this article, allowing for the replication of our results, is available online, see section 5.

The responses to the open questions revealed no insights beyond those addressed in the specific questions. Therefore, the analyses of these responses are not included in this report.

### 3.1. Inter-rater Agreement

To determine inter-rater agreement and reliability, we calculated agreement scores across all 30 questions for each clip in each condition separately. This analysis was performed to examine whether participants in each condition gave similar ratings across all questions when they had viewed the same clip. High agreement would indicate that participants had interpreted similar things from a given clip, e.g., participants might all have felt that the children in a clip were being friendly and cooperative, or aggressive and competitive. Whilst this analysis does not reveal exactly what participants interpreted from the videos, it does indicate whether they gave similar ratings, and therefore reported recognizing similar states/behaviors. Given that each clip was rated by a varying subset of participants, Krippendorff's alpha (Hayes and Krippendorff, 2007) was the most appropriate metric of rater agreement (see **Table 3** for number of raters and agreement per clip). The alpha scores ranged from 0.058 to 0.463 i.e., from "slight" to "moderate" agreement (Landis and Koch, 1977).

**TABLE 3 |** Table of inter-rater agreement scores for responses to each clip in each condition.

| Clip | Krippendorff's Alpha (3 d.p.) | |
|---|---|---|
| | Full-Scene (N) | Movement Alone (N) |
| 1 | 0.446 (16) | 0.186 (26) |
| 2 | 0.181 (24) | 0.270 (20) |
| 3 | 0.393 (22) | 0.369 (18) |
| 4 | 0.444 (22) | 0.262 (23) |
| 5 | 0.328 (23) | 0.283 (20) |
| 6 | 0.463 (19) | 0.359 (19) |
| 7 | 0.091 (19) | 0.236 (23) |
| 8 | 0.339 (19) | 0.312 (17) |
| 9 | 0.097 (20) | 0.058 (18) |
| 10 | 0.396 (18) | 0.086 (13) |
| 11 | 0.280 (17) | 0.234 (23) |
| 12 | 0.368 (25) | 0.298 (16) |
| 13 | 0.334 (20) | 0.189 (21) |
| 14 | 0.310 (17) | 0.309 (21) |
| 15 | 0.422 (26) | 0.242 (14) |
| 16 | 0.192 (16) | 0.272 (21) |
| 17 | 0.273 (17) | 0.183 (21) |
| 18 | 0.334 (16) | 0.331 (24) |
| 19 | 0.415 (22) | 0.304 (19) |
| 20 | 0.451 (18) | 0.250 (23) |

A $t$-test was conducted to assess whether the two conditions differed in their agreement scores across all 20 clips. This analysis revealed that participants in the full-scene condition showed significantly higher agreement ($M = 0.328$, $SD = 0.110$) than participants in the movement-alone condition ($M = 0.252$, $SD = 0.079$) (Paired Samples $T$-Test: $t_{(39)} = 2.95$, $p = 0.008$, $d = 0.78$). These analyses show that participants viewing the full-scene clips demonstrated higher levels of agreement in their ratings than those viewing the movement-alone clips. However, participants in the latter condition still showed some agreement compared to chance (chance level Krippendorff's Alpha = 0.0; One Sample $T$-Test: $t_{(19)} = 13.95$, $p =< 0.001$, $d = 3.12$), suggesting that some internal states and social constructs were recognizable within the movement information in both conditions.

## 3.2. Automatic Labeling of Internal States

The following analysis explored the question of whether the internal states and social constructs which were available to/inferred by humans when viewing the full visual scene was also available in the movement-alone condition.

We investigated this question using supervised machine learning: would a classifier, trained to label internal states and social constructs from the full-scene ratings, then label the social situations equally well from the movement-alone ratings? If so, this would suggest that the same interaction information was recognized by, and therefore available to, participants in each video condition.

**Pre-processing** Participants' ratings were coded from 0 (*strongly disagree*) to 4 (*strongly agree*), each construct being recorded as $left_{construct}$ and $right_{construct}$ (see **Appendix A**). Before the following analyses were run, the data from the right-left paired questions was transformed so that results could be more easily interpreted in terms of what behaviors were evident in the interactions, ignoring whether it was the child on the right or the left who was exhibiting this behavior. First, for each question we calculated the absolute difference $diff_{construct} = abs(left_{construct} - right_{construct})$ between the score for the left child and the right child. This score was calculated so that we could more easily see if the children were rated as behaving in the same way, or experiencing similar internal states. Examining the individual scores for each child would have meant that in order to see the dynamics between the children, each clip would have needed to be analyzed separately. Second, for each question we calculated the sum (shifted to the range $[-2, 2]$) $sum_{construct} = left_{construct} + right_{construct} - 4$ of the scores for both children. This score was calculated because the difference score does not contain information about the strength of the rater's belief that the behavior or internal state was evident in the clip. For example, we might have the same difference score for clips where raters believed that both children behaved aggressively and that neither child behaved aggressively. The sum score tells us the degree to which a state was identifiable in the clip.

**Multi-label classification** To test whether the same interaction information was reported in each video condition we examined whether the ratings from each condition were sufficient to identify the types of internal states or social constructs which were depicted in the videos.

The classifier was trained in a supervised manner, using the 30 ratings provided by the participants (questions from **Appendix A**, pre-processed as indicated above) as input, and the seven labels assigned to each clip during selection (**Table 2**) as the target classification classes. Because the clips could be assigned multiple labels (e.g., a given interaction can be *fun* and *cooperative* at the same time), we used a multi-label classifier (Pieters and Wiering, 2017), using 7-dimensional binary vectors (wherein a zero value denoted that a label was not present in the clip, and a value of one denoted that it was).

We compared the performances of four of classifier (random forest classifier, extra-tree classifier, multi-layer perceptron classifier and a k-Nearest Neighbor classifier, using implementations from the Python `sklearn` toolkit; hyper-parameters were optimized using a grid search where applicable), and eventually selected a k-Nearest Neighbor (with $k = 3$) classifier as providing the best overall classification performance.

Accuracy, precision, recall and F1 score were calculated to assess the performance of the classifier (following recommendations in Sorower (2010) and using the *weighted* implementations of the metrics available in the Python `sklearn` toolkit). Specifically, in the following, *Accuracy* reports the percentage of instances where the predicted labels match exactly with the actual labels; *Precision* is calculated as the ratio $\frac{tp}{tp+fp}$ of true positives divided by the total number of predicted labels (true positives + false positives); *Recall* is calculated as $\frac{tp}{tp+fn}$, i.e. the ratio true positives over the total number of labels that *should* have been found (true positives +

false negatives). Finally, the *F1 score* is the harmonic average of the precision and recall, calculated as $\frac{2precision \cdot recall}{precision+recall}$.

To see how well the classifier performed, we compared performance against chance. Chance levels for these metrics were calculated by training the classifier with randomly generated labels (using the same distribution of labels as found in the real data set), and then measuring the classifier's performance on the actual testing data set.

Results are shown in **Table 4**. In both testing conditions, performance is poor to moderate (for instance 15.8% accuracy for the exact predictions of correct labels in the movement-alone clips), but remain markedly above chance levels (following Ojala and Garriga (2010) permutation-based $p$-value for classification significance, we found $p = 0.02$ for the full-scene classification, and $p = 0.01$ for the movement-alone classification, ruling out with high probability the null hypothesis that the classification results are due to chance).

Importantly, we found that prediction scores are very similar when testing the classifier on the full-scene ratings or when testing on the movement-alone ratings. This indicates that, from the perspective of automatic data classification, participants who viewed the movement-alone videos were able to report similar details as participants in the full-scene condition. This suggests that the movement-alone videos contain sufficient information to identify different internal states and social constructs.

To identify whether there were particular internal states or social constructs which were easier to recognize than others, the F1 score for each label was calculated. These results are reported in **Table 5** and **Figure 4**. We can see that in both conditions the labels "Bored" and "Aggressive" have higher F1 scores than the other labels. Additionally, the F1 scores for these labels when classifying the full-scene ratings (Bored: 60.0%, Aggressive: 39.0%) are similar to the F1 scores when testing was done on the movement-alone ratings (Bored: 58.5%, Aggressive: 43.7%). This suggests that these constructs are as readily recognized when viewing the full visual scene as when viewing only body movements. In contrast, the F1 score for "Aimless" when testing on full-scene ratings is similar to the scores for most of the rest of the labels (30.3%) but drops to be much lower than any other label when testing was done on the movement-alone ratings (19.4%). This could be interpreted as showing that aimless play, whilst fairly well recognized from the ratings of full visual scene videos, is much harder to recognize from ratings produced when participants viewed only movement information.

This analysis relied on the labels assigned by some of the authors during clip selection. However, participants may have been able to recognize other internal states or social constructs not covered by these labels. In order to investigate possible latent constructs that participants in both conditions may have relied on, we next performed a factor analysis on the dataset.

## 3.3. Factor Analysis

An Exploratory Factor Analysis (EFA) was performed to explore what types of information participants reported recognizing from the videos. If similar latent constructs are found to underlie participants responses in each condition, this would support the conclusion that participants reported identifying the same types of information in each type of video. Additionally, exploring what factors load into each construct would provide an indication of what these types of information are.

**EFA** Preliminary assessments revealed a Kaiser-Meyer-Olkin (KMO) statistic of 0.89 and the Bartlett's Test of Sphericity was significant, indicating that the data was suitable for performing an EFA. EFA was performed on the ratings data from each video condition separately to examine what types of interaction information participants were able to draw from the full visual scene compared to movement information alone. We used the

**TABLE 4 |** Classification results. *Full-scene* results are obtained by training the classifier on 80% of the full-scene ratings, and testing on the remaining 20%; *Movement-alone* results are obtained by training the classifier on 100% of the full-scene data, and testing on the movement-only ratings.

| | Accuracy | Precision | Recall | F1-measure |
|---|---|---|---|---|
| Full-scene | 15.1 | 44.5 | 32.0 | 36.1 |
| Chance | 3.7 | 27.3 | 14.0 | 17.4 |
| Movement-alone | 15.8 | 41.6 | 32.7 | 36.3 |
| Chance | 3.9 | 28.2 | 14.2 | 17.9 |

*Results are averaged over a 300-fold cross-validation. Values are given as percentages.*

**TABLE 5 |** F1 scores for each independent label.

| | Aggressive | Aimless | Bored | Cooperative | Dominant | Excited | Fun |
|---|---|---|---|---|---|---|---|
| Full-scene | 42.2 | 29.5 | 56.6 | 30.7 | 37.9 | 32.2 | 25.1 |
| Chance | 18.8 | 17.3 | 11.7 | 18.2 | 20.0 | 18.6 | 11.4 |
| Movement Alone | 43.7 | 19.4 | 58.5 | 29.6 | 43.4 | 31.2 | 27.5 |
| Chance | 20.1 | 16.1 | 10.7 | 18.7 | 19.9 | 17.3 | 10.4 |

*See **Table 4** for the meaning of each row. Values are given as percentages.*

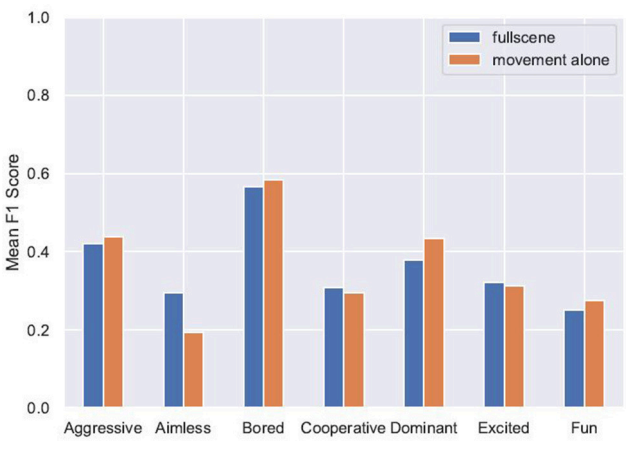



**FIGURE 4 |** F1 scores of individual label predictions in both conditions.

**TABLE 6 |** Factor loadings for the three-factor solution using EFA, with factor loadings > 0.35.

| | Factor 1: imbalance | | Factor 2: valence | | Factor 3: engagement | |
|---|---|---|---|---|---|---|
| | *Full-scene* | *Mov.-alone* | *Full-scene* | *Mov.-alone* | *Full-scene* | *Mov.-alone* |
| Diff sad | 0.41 | 0.52 | | | | |
| Sum sad | | | 0.72 | 0.53 | | 0.49 |
| Diff happy | 0.49 | 0.53 | | | | |
| Sum happy | | | | –0.51 | –0.55 | |
| Diff angry | 0.40 | 0.62 | | | | |
| Sum angry | | | 0.81 | 0.85 | | |
| Diff excited | 0.53 | 0.63 | | | | |
| Sum excited | | | | | –0.71 | |
| Diff calm | 0.45 | 0.63 | | | | |
| Sum calm | | | | –0.45 | | |
| Diff friendly | 0.69 | 0.56 | | | | |
| Sum friendly | | | | –0.60 | –0.43 | |
| Diff aggressive | 0.78 | 0.79 | | | | |
| Sum aggressive | | | 0.80 | 0.72 | –0.36 | |
| Diff engaged | | 0.39 | | | 0.65 | 0.52 |
| Sum engaged | | | | | –0.64 | –0.64 |
| Diff distracted | | | | | 0.65 | 0.63 |
| Sum distracted | | | 0.63 | | | 0.82 |
| Diff bored | | 0.44 | | | 0.61 | 0.54 |
| Sum bored | | | 0.58 | | 0.48 | 0.83 |
| Diff frustrated | 0.53 | 0.61 | | | | |
| Sum frustrated | | | 0.70 | 0.69 | | |
| Diff dominant | 0.75 | 0.81 | | | | |
| Sum dominant | | | 0.53 | 0.52 | | |
| Diff submissive | 0.68 | 0.72 | | | | |
| Sum submissive | | | 0.54 | | | |

`factor_analyzer` Python module[4] to perform the EFA, additionally using a *promax* rotation. Three factors were found to explain 44% of the variance in the full-scene ratings, and 46% in the movement-alone ratings. The factor loadings for each component can be seen in **Table 6**.

A Pearson correlation was conducted to examine the similarity of components found in the full-scene and movement-alone ratings. A strong positive correlation was found between each pair of components: for Factor 1: $r = 0.94, p < 0.001$; for Factor 2: $r = 0.84, p < 0.001$; for Factor 3: $r = 0.81, p < 0.001$. This supports the hypothesis that the same latent constructs are relied upon by the participants to rate social interactions, be it based on raw video footage (full-scene) or on a simplified, movement-only, stick-man-style representation (movement-alone).

By inspecting the distribution of factors loadings in **Table 6**, the latent constructs can be further interpreted. It appears that the first component is describing how different the children's behaviors and emotional states are, i.e. this factor describes an *imbalance* in the children,s social, behavioral, and emotional states. For instance, a high value on this scale would show that the children were reported as behaving very differently, e.g., if one child was highly engaged, the other was not very engaged at all.

The second component describes the overall *valence* of the interaction. A high value on this factor would indicate a negative, adversarial interaction where the children were rated as being sad, aggressive etc. Alternatively, a (lower) positive valence value might result from an interaction where one child was rated as being more sad or aggressive than the other child was happy. For both conditions this component has positive correlations with the *Sum* items for negative emotions and behaviors (e.g., Anger, Aggression). For the movement-alone condition, this component also has negative correlations with *Sum* items for positive emotions and behaviors (e.g., Happiness, Friendliness).

The third component is mostly describing the children's *engagement* with their task. In comparison to the other two components it contains more of a mix of *Sum* and *Difference* items, and therefore describes both how similar the children were in how engaged they were, and the overall level of engagement within the interaction. A high value on this third factor would show that the children were rated as showing different levels of engagement, but a strong indication of boredom within the interaction as a whole.

**Social Expressiveness of the EFA-Space Embedding** One may wonder whether these three factors alone would allow by themselves for an effective assessment of a social interaction, i.e. is the social "expressiveness" of our EFA factors as good as the original 26 factors? This can be investigated by re-applying the same classification methodology as used in section 3.2 to the EFA embedding of the participants' ratings.

To this end, the 26-dimensional participant ratings were projected onto the smaller, 3-dimensional, space spanned by the EFA factors (the *EFA-space*):

$$M^{EFA}_{fullscene} = M_{fullscene} \cdot \Lambda^{EFA}_{fullscene}$$
$$M^{EFA}_{movementalone} = M_{movementalone} \cdot \Lambda^{EFA}_{fullscene}$$

with $M_{fullscene}$ the $396 \times 26$ matrix of the participants' ratings, $M^{EFA}_{fullscene}$ the $396 \times 3$ matrix of the participants' ratings projected onto the EFA space, and $\Lambda^{EFA}_{fullscene}$ the $26 \times 3$ matrix of the EFA factor loadings (**Table 6**). Both the full-scene clips and the movement-alone clips where projected into the same space (spanned by the factors found during the full-scene EFA).

Then, we retrained the same classifier (a kNN with $k = 3$) as in section 3.2, and tried to predict social labels from EFA-projected ratings unseen at training time. **Tables 7**, **8** show the results. We observe a drop of about 4–6% in performance, but still above chance.

## 4. DISCUSSION AND CONCLUSION

Psychology literature has long established the importance of observing physical group behaviors to provide us with a unique window onto the agents' internal states, as well as the current state of the social interaction. Specifically, we have previous

[4]https://github.com/EducationalTestingService/factor_analyzer

**TABLE 7 |** Classification results, including classification in EFA-space. *EFA-space* means that the dimensionality of the training and testing data is reduced to 3 by projecting the ratings onto the 3-dimensional space spanned by the EFA factors; non-EFA values copied from **Table 4** for comparison.

| | Accuracy | Precision | Recall | F1-measure |
|---|---|---|---|---|
| Full-scene, EFA | 11.2 | 38.3 | 26.2 | 30.0 |
| Full-scene | 15.1 | 44.5 | 32.0 | 36.1 |
| Chance | 3.8 | 28.1 | 14.2 | 17.8 |
| Movement-alone, EFA | 11.7 | 35.1 | 27.0 | 30.3 |
| Movement-alone | 15.7 | 41.6 | 32.7 | 36.3 |
| Chance | 3.9 | 28.3 | 14.2 | 17.9 |

*Values are given as percentages.*

**TABLE 8 |** F1 scores for each independent label, including after classification in the EFA-space.

| | Aggressive | Aimless | Bored | Cooperative | Dominant | Excited | Fun |
|---|---|---|---|---|---|---|---|
| Fullscene, EFA | 37.8 | 16.2 | 53.9 | 29.4 | 29.7 | 25.9 | 20.6 |
| Fullscene | 42.2 | 29.5 | 56.6 | 30.7 | 37.9 | 32.2 | 25.1 |
| Chance | 19.1 | 16.5 | 11.7 | 19.0 | 19.6 | 17.4 | 11.0 |
| Movement alone, EFA | 36.5 | 24.0 | 49.2 | 24.6 | 33.7 | 27.4 | 12.2 |
| Movement alone | 43.7 | 19.4 | 58.5 | 29.6 | 43.4 | 31.2 | 27.5 |
| Chance | 19.8 | 16.4 | 10.7 | 18.9 | 19.9 | 17.9 | 10.5 |

*Non-EFA values copied from **Table 5** for comparison. Values are given as percentages.*

evidence of the role of *movements/actions* as an important social signal (Gallese and Goldman, 1998; Alaerts et al., 2011). The main contribution of this paper is to investigate the question of what different states are identified by observers of naturalistic interactions, looking at the (rather messy) social interactions occurring between children while playing together.

This study aimed to examine the kinds of information humans report recognizing from the movements of such naturalistic social interactions. We investigated the following question: is movement information alone (in our case, the moving skeletons of two children playing together, pictured on a uniform black background) sufficient for humans to successfully infer the internal states and social constructs experienced and present within a social interaction? Our methodology involved a between-subject, on-line study, where participants were asked to rate children's behaviors along 17 dimensions, having either watched the raw footage of short interaction videos, or only the skeletons and facial landmarks extracted from the same video clips. This resulted in about 800 unique human ratings, covering both conditions, across 20 different clips, selected for displaying a range of different internal states and social constructs.

We explored the ratings data set (which is publicly available, see the details in the following section) using two main data mining techniques. We first trained a classifier on the full-scene ratings with hand-crafted social labels to then attempt to automatically identify these social labels on the movement-alone ratings. Our results show that training our best performing classifier (a 3-kNN) on 80% of the full-scene ratings and testing on the remaining 20% results in a (cross-validated) precision of 46.2% and recall of 33.6%. We found very similar levels of precision and recall (respectively 41.6 and 32.7%) when testing on the movement-alone ratings: the assessment of the social interaction taking place between two children, made by naive observers watching a low-dimensional, movement-alone video-clip of the interaction, carries similar informational content regarding the internal states and social constructs as the original raw video footage. Based on this finding, we can tentatively conclude that whilst the movement alone videos contain fewer pieces of information, the pieces of information available are as meaningful as those in the full scene videos. Furthermore, we can assess that these pieces of information can be interpreted by human observers in a similar way as those in the full scene videos.

To better make sense of these results, we employed a second data mining technique (Exploratory Factor Analysis, EFA) to attempt to uncover underlying latent factors that would in effect embody stronger cognitive constructs, implicitly relied upon by the humans when assessing a social interaction. We ran independent EFAs on the ratings provided for the full-scene videos and those provided for the movement-alone clips.

To our surprise, the latent factors found by the EFA were strongly correlated between both conditions. In both condition, one factor was measuring the **behavioral imbalance** between the two children (i.e. how similar or dissimilar their behaviors were); a second factor reflected the **valence of the interaction**, from adversarial behaviors and negative emotions, to pro-social and positive behaviors and emotions; finally a third factor embodied **the level of engagement** of the children. These constructs may be indicative of the constructs humans use to interpret social interactions in general. Further research is needed to confirm whether or not this is the case. However, if it is it would provide further insights into how humans approach the interpretation and understanding of social interactions. That is, these three factors may represent the basic cognitive constructs humans use to understand social interactions. Consequently, HRI research could use these constructs as a basic framework for exploring human behavior for classification purposes.

Using the 3-dimensional subspace spanned by these three EFA factors, we have furthermore shown that 'summarizing' the internal states and social constructs inferred by the participants into the 3 latent constructs—imbalance, valence, engagement—only slightly degrades the ability of the classifier to predict the social labels associated with the interaction. This reinforces the hypothesis that these three constructs might play a foundational role in the human understanding of social interactions.

The results of both the classification analysis and EFA demonstrate that it is reasonable to expect a machine learning algorithm, and in consequence, a robot, to successfully decode and classify a range of internal states and social constructs using a low-dimensional data source (such as the movements and poses of observed individuals) as input. Specifically, whilst this study does not examine the ability to identify the correct internal states or social constructs, we have shown that, in a robust way, people agree in their reports of what they have seen both within and between conditions. As such, our study

shows that, even though assessing social interactions is difficult even for humans, using skeletons and facial landmarks only does not significantly degrade the assessment. Future studies aiming to train a robotic system would ideally utilize a training dataset where the internal states and social constructs have been verified (and therefore a ground-truth is available). This study provides the evidence to guide this type of work, for example by demonstrating that training a robot to recognize aggression from movement information is likely to be more successful than recognizing aimlessness.

## 4.1. Opportunities for Future Work

Given that this work is exploratory in nature, it presents a number of opportunities for future work. First, while above chance, the accuracy of the classifier is relatively low. This may reflect the inherent difficulty of rating internal states and social constructs for an external, naive observer (such as the raters recruited for this study). The literature on emotion recognition does show that humans are able to recognize emotional states from impoverished stimuli with a high level of accuracy [e.g., 44–59% in Alaerts et al. (2011), 59–88% in Gross et al. (2012)]. Similarly, research regarding the recognition of dispositions and social behaviors indicate that computational techniques can achieve a higher recognition accuracy than the current study. For example, Okada et al. (2015) achieved around 57% accuracy in classifying dominance. However, there is some evidence to suggest that humans may not be as accurate as computational classifiers in identifying internal states as we define them here. To demonstrate, Sanghvi et al. (2011) found that whilst human observers were able to recognize engagement to an average of 56% accuracy, their best classifier achieved an 82% level of accuracy. Whilst the accuracy scores presented here are much lower, the existing literature suggests that this may be a result of the fact that humans do seem to demonstrate some difficulty in recognizing these types of states. Additionally, it is important to remember that the classifier in this study labeled the clips using the ratings of all the left/right child questionnaire items, whereas previous research has tended to use the raw visual and/or audio information for classification by both computational systems (Okada et al., 2015) and human observers (Sanghvi et al., 2011). This high dimensional input may have had the effect of diluting the specificity and causing the classifier to use irrelevant or unhelpful inputs when making classification decisions. Additionally, the low classification accuracy may result from the fact that the questionnaire used in this study might not have been good enough. As such, future research would benefit from developing and optimizing the questionnaire.

Additionally, the present study does not explore precisely which movement characteristics were useful for participants in making inferences about the internal states of the children in the videos. In this study we employed a supervised classification technique to demonstrate that social interaction assessments based on full-scene or movement-only stimuli were of similar quality–most notably, our input were ratings of social interactions by human observers. This technique is *not* practically transferable to a robot, as robots would have to directly classify the raw stimuli (a video stream or skeletons), without having access to intermediate ratings of the agents' states. Creating such a classifier is an important next step in deciphering how humans recognize internal states, and therefore in deciding how a robot or classifier can be endowed with a similar skill, for which our present results provide a solid foundation.

The fact that the internal states experienced by the children in the videos could not be validated does present a further limitation for this study. A number of datasets demonstrating one or a subset of the internal states we are interested in are available. For example, the Tower Game Dataset consists of human-human pairs collaborating on a task, and has been annotated for joint attention and entrainment behaviors reflecting cooperation and collaboration (Salter et al., 2015). Similarly, the DAiSEE dataset contains videos of individuals watching videos in an e-learning setting and is annotated for the internal states of boredom, confusion, engagement, and frustration (Gupta et al., 2016). Other datasets include: the UE-HRI annotated for engagement (Ben-Youssef et al., 2017), the ELEA annotated for perceived leadership and dominance (Sanchez-Cortes et al., 2011) among others. Replicating this experiment using a validated dataset may provide stronger classification and inter-rater agreement results. However, few ecologically-valid datasets present the range and variety of internal states as are available in the PInSoRo dataset. As such, this present research represents an important first step in framing the research methodology for analysis of complex, real-life social interactions.

## 4.2. Conclusion

The aim of this study was to identify social constructs or human internal states which a socially interactive robot could be made to recognize. Analyzing the weighted precision scores for each classification label revealed that "Aggressive" and "Bored" were classified correctly more often in both conditions, whilst "Aimless" was classified correctly much less from the movement-alone ratings. This suggests that training a robot to recognize aimlessness based on movement information might not be as successful as training recognition of boredom. Practically speaking, this finding suggests that designing a tutor robot, such as those used by L2TOR (Belpaeme et al., 2015), to recognize when a child is bored by their task based on movement information would be more successful than having the robot recognize when a child is performing the task in an "aimless" or "non-goal-directed" manner. Such a robot could then appropriately offer encouragement or an alternative task.

Additionally, these findings suggest that exploring other data sources for recognizing human internal states may reveal that certain behavioral modalities may be more useful for recognizing different states. In this way, the method we have demonstrated here can be used to streamline research aimed at teaching robots [and other classification technologies, e.g., automatic classification of security footage (Gowsikhaa et al., 2014)] to recognize human internal states. By applying this method to different types of input data, research can identify the optimal behavioral modality for recognizing a particular human internal state.

These findings have significant impact for both social psychology and artificial intelligence. For social psychology,

it consolidates our understanding of implicit social communication, and confirms previous findings that humans are able to recognize socially relevant information from observed movements (Iacoboni et al., 2005; Alaerts et al., 2011; Quesque et al., 2013). For artificial intelligence, and in particular, for social robotics and human-robot interaction, it provides support for the intuition that low-dimensional (about 100 skeletal and facial points per agent vs. full video frames comprising of hundred of thousands of pixels), yet structured observations of social interactions might effectively encode complex internal states and social constructs. This provides promising support for fast and effective classification of social interactions, a critical requirement for developing socially-aware artificial agents and robots.

## 5. RESOURCES FOR REPLICATION

Following recommendations by Baxter et al. (2016), we briefly outline hereafter the details required to replicate our findings.

### 5.1. Study

The protocol and all questionnaires have been provided in the text. The code of the experiment is available at https://github.com/severin-lemaignan/pinsoro-kinematics-study/. Note that, due to data protection regulations, the children' video clips are not available publicly. However, upon signature of an ethical agreement, we can provide them to the interested researcher.

### 5.2. Data Analysis

The full recorded experimental dataset, as well as the complete data analysis script allowing for reproduction of the results and plots presented in the paper (using the Python *pandas* library) are open and available online, in the same Git repository. In particular, a iPython notebook with all the steps followed for our data analysis is available here: https://github.com/severin-lemaignan/pinsoro-kinematics-study/blob/master/analysis/analyses_notebook.ipynb.

## ETHICS STATEMENT

This study was carried out in accordance with the recommendations of Ethics guidelines of the University of Plymouth Ethics Committee, with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Ethics Committee of the University of Plymouth.

## AUTHOR CONTRIBUTIONS

MB, CE, and SL contributed to the design, running, analysis, and write-up of this study. ST and TB contributed to the supervision and funding of this study.

## FUNDING

This work is part of the EU FP7 project DREAM project (www.dream2020.eu), funded by the European Commission (grant no. 611391). It has received additional funding by the EU H2020 Marie Sklodowska-Curie Actions project DoRoThy (grant no. 657227).

## REFERENCES

Alaerts, K., Nackaerts, E., Meyns, P., Swinnen, S. P., and Wenderoth, N. (2011). Action and emotion recognition from point light displays: an investigation of gender differences. *PLoS ONE* 6:e20989. doi: 10.1371/journal.pone.0020989

Ansuini, C., Cavallo, A., Bertone, C., and Becchio, C. (2014). The visible face of intention: why kinematics matters. *Front. Psychol.* 5:815. doi: 10.3389/fpsyg.2014.00815

Baxter, P., Kennedy, J., Senft, E., Lemaignan, S., and Belpaeme, T. (2016). "From characterising three years of HRI to methodology and reporting recommendations," in *The Eleventh ACM/IEEE International Conference on Human Robot Interation* (Christchurch: IEEE Press), 391–398.

Becchio, C., Koul, A., Ansuini, C., Bertone, C., and Cavallo, A. (2017). Seeing mental states: an experimental strategy for measuring the observability of other minds. *Phys. Life Rev.* 24, 67–80. doi: 10.1016/j.plrev.2017.10.002

Belpaeme, T., Kennedy, J., Baxter, P., Vogt, P., Krahmer, E. E. J., Kopp, S., et al. (2015). "L2TOR-second language tutoring using social robots," in *Proceedings of the ICSR 2015 WONDER Workshop* (Paris).

Ben-Youssef, A., Clavel, C., Essid, S., Bilac, M., Chamoux, M., and Lim, A. (2017). "Ue-hri: a new dataset for the study of user engagement in spontaneous human-robot interactions," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (New York, NY: ACM), 464–472. doi: 10.1145/3136755.3136814

Beyan, C., Carissimi, N., Capozzi, F., Vascon, S., Bustreo, M., Pierro, A., et al. (2016). "Detecting emergent leader in a meeting environment using nonverbal visual features only," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (New York, NY: ACM), 317–324. doi: 10.1145/2993148.2993175

Breazeal, C., Gray, J., and Berlin, M. (2009). An embodied cognition approach to mindreading skills for socially intelligent robots. *Int. J. Robot. Res.* 28, 656–680. doi: 10.1177/0278364909102796

Cao, Z., Simon, T., Wei, S. E., and Sheikh, Y. (2017). "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR* (Honolulu, HI).

Dautenhahn, K., and Saunders, J. (2011). *New Frontiers in Human Robot Interaction*, vol 2. Amsterdam: John Benjamins Publishing.

Domes, G., Heinrichs, M., Michel, A., Berger, C., and Herpertz, S. C. (2007). Oxytocin improves mind-reading in humans. *Biol. Psychiat.* 61, 731–733. doi: 10.1016/j.biopsych.2006.07.015

Ekman, P., and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* 17:124.

Gallese, V., and Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends Cogn. Sci.* 2, 493–501.

Gowsikhaa, D., Abirami, S., and Baskaran, R. (2014). Automated human behavior analysis from surveillance videos: a survey. *Artif. Intell. Rev.* 42, 747–765. doi: 10.1007/s10462-012-9341-3

Gross, M. M., Crane, E. A., and Fredrickson, B. L. (2012). Effort-shape and kinematic assessment of bodily expression of emotion during gait. *Hum. Move. Sci.* 31, 202–221. doi: 10.1016/j.humov.2011.05.001

Gupta, A., D'Cunha, A., Awasthi, K., and Balasubramanian, V. (2016). Daisee: Towards user engagement recognition in the wild. *arXiv* arXiv:1609.01885.

Haidt, J., and Keltner, D. (1999). Culture and facial expression: Open-ended methods find more expressions and a gradient of recognition. *Cogn. Emot.* 13, 225–266.

Han, J.-H., and Kim, J.-H. (2010). "Human-robot interaction by reading human intention based on mirror-neuron system," in *2010 IEEE International Conference on Robotics and Biomimetics (ROBIO)* (Tianjin: IEEE), 561–566.

Hayes, A. F., and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Commun. Methods Meas.* 1, 77–89. doi: 10.1080/19312450709336664

Hufschmidt, C., Weege, B., Rder, S., Pisanski, K., Neave, N., and Fink, B. (2015). Physical strength and gender identification from dance movements. *Pers. Individ. Diff.* 76, 13–17. doi: 10.1016/j.paid.2014.11.045

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., and Mazziotta, J. C. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biol.* 3, 0529–0535. doi: 10.1371/journal.pbio.0030079

Kawamura, R., Toyoda, Y., and Niinuma, K. (2019). "Engagement estimation based on synchrony of head movements: application to actual e-learning scenarios," in *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion* (New York, NY: ACM), 25–26.

Kozlowski, L. T., and Cutting, J. E. (1977). Recognizing the sex of a walker from a dynamic point-light display. *Percept. Psychophys.* 21, 575–580.

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174.

Lemaignan, S., Edmunds, C., Senft, E., and Belpaeme, T. (2017). The Free-play Sandbox: a Methodology for the Evaluation of Social Robotics and a Dataset of Social Interactions. *arXiv* arXiv:1712.02421.

Manera, V., Becchio, C., Cavallo, A., Sartori, L., and Castiello, U. (2011). Cooperation or competition? Discriminating between social intentions by observing prehensile movements. *Exp. Brain Res.* 211, 547–556. doi: 10.1007/s00221-011-2649-4

Manera, V., Schouten, B., Becchio, C., Bara, B. G., and Verfaillie, K. (2010). Inferring intentions from biological motion: a stimulus set of point-light communicative interactions. *Behav. Res. Methods* 42, 168–178. doi: 10.3758/BRM.42.1.168

Mather, G., and Murdoch, L. (1994). Gender discrimination in biological motion displays based on dynamic cues. *Proc. R. Soc. Lond. B* 258, 273–279.

Ojala, M., and Garriga, G. C. (2010). Permutation tests for studying classifier performance. *J. Mach. Learn. Res.* 11, 1833–1863. doi: 10.1109/ICDM.2009.108

Okada, S., Aran, O., and Gatica-Perez, D. (2015). "Personality trait classification via co-occurrent multiparty multimodal event discovery," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (New York, NY: ACM), 15–22.

Okur, E., Alyuz, N., Aslan, S., Genc, U., Tanriover, C., and Esme, A. A. (2017). "Behavioral engagement detection of students in the wild," in *International Conference on Artificial Intelligence in Education* (Wuhan: Springer), 250–261.

Pieters, M., and Wiering, M. (2017). "Comparison of machine learning techniques for multi-label genre classification," in *Benelux Conference on Artificial Intelligence* (Groningen: Springer), 131–144.

Pollick, F. E., Paterson, H. M., Bruderlin, A., and Sanford, A. J. (2001). Perceiving affect from arm movement. *Cognition* 82, B51–B61. doi: 10.1016/S0010-0277(01)00147-0

Quesque, F., Lewkowicz, D., Delevoye-Turrell, Y. N., and Coello, Y. (2013). Effects of social intention on movement kinematics in cooperative actions. *Front. Neurorobot.* 7:14. doi: 10.3389/fnbot.2013.00014

Salter, D. A., Tamrakar, A., Siddiquie, B., Amer, M. R., Divakaran, A., Lande, B., and Mehri, D. (2015). "The tower game dataset: a multimodal dataset for analyzing social interaction predicates," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (Xi'an: IEEE), 656–662.

Sanchez-Cortes, D., Aran, O., Mast, M. S., and Gatica-Perez, D. (2011). A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Trans. Multi.* 14, 816–832. doi: 10.1109/TMM.2011.2181941

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., and Paiva, A. (2011). "Automatic analysis of affective postures and body motion to detect engagement with a game companion," in *Proceedings of the 6th international conference on Human-robot interaction* (New York, NY: ACM), 305–312.

Schrempf, O. C., and Hanebeck, U. D. (2005). "A generic model for estimating user intentions in human-robot cooperation," in *ICINCO* (Barcelona), 251–256.

Shaker, N., and Shaker, M. (2014). "Towards understanding the nonverbal signatures of engagement in super mario bros," in *International Conference on User Modeling, Adaptation, and Personalization* (Aalborg: Springer), 423–434.

Sorower, M. S. (2010). *A Literature Survey on Algorithms for Multi-Label Learning*. Corvallis: Oregon State University.

Tracy, J. L., and Robins, R. W. (2008). The nonverbal expression of pride: evidence for cross-cultural recognition. *J. Personal. Soc. Psychol.* 94:516. doi: 10.1037/0022-3514.94.3.516

Vernon, D., Thill, S., and Ziemke, T. (2016). "The role of intention in cognitive robotics," in *Toward Robotic Socially Believable Behaving Systems-Volume I* (Springer), 15–27.

Walker-Andrews, A. S. (1997). Infants' perception of expressive behaviors: differentiation of multimodal information. *Psychol. Bull.* 121:437.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

# A. APPENDIX

## A.1. Questions

Open Question: "What did you notice about the interaction?"

Specific Questions: For all of the following questions participants were asked to report how much they agreed with each statement. Answers : *Strongly Disagree / Disagree / Not Sure / Agree / Strongly Agree*

1. "The children were competing with one another."
2. "The children were cooperating with one another."
3. "The children were playing separately."
4. "The children were playing together."

6-7 "The character on the left/right was sad."
8-9 "The character on the left/right was happy."
10-11 "The character on the left/right was angry."
12-13 "The character on the left/right was excited."
14-15 "The character on the left/right was calm."
16-17 "The character on the left/right was friendly."
17-18 "The character on the left/right was aggressive."
19-20 "The character on the left/right was engaged with what they were doing on the table."
21-22 "The character on the left/right was distracted from the table."
23-24 "The character on the left/right was bored."
25-26 "The character on the left/right was frustrated."
27-28 "The character on the left/right was dominant."
29-30 "The character on the left/right was submissive."

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 06 September 2019
doi: 10.3389/frobt.2019.00085

# "That Robot Stared Back at Me!": Demonstrating Perceptual Ability Is Key to Successful Human–Robot Interactions

*Masaya Iwasaki*[1], *Jian Zhou*[1], *Mizuki Ikeda*[1], *Yuki Koike*[1], *Yuya Onishi*[1], *Tatsuyuki Kawamura*[2] *and Hideyuki Nakanishi*[1]*

[1] Department of Adaptive Machine Systems, Osaka University, Osaka, Japan, [2] Kyoto Innovation, Inc., Kyoto, Japan

Communication robots, such as robotic salespeople and guide robots, are increasingly becoming involved in various aspects of people's everyday lives. However, it is still unclear what types of robot behavior are most effective for such purposes. In this research, we focused on a robotic salesperson. We believe that people often ignore what such robots have to say owing to their weak social presence. Thus, these robots must behave in ways that attract attention encouraging people to nod or reply when the robots speak. In order to identify suitable behaviors, we conducted two experiments. First, we conducted a field experiment in a shop in a traditional Kyoto shopping street to observe customers' real-world interactions with a robotic salesperson. Here, we found that the first impression given by the robot had a crucial influence on its subsequent conversations with most customer groups and that it was important for the robot to indicate it could understand how much attention customers were paying to the robot in the early stages of its interactions if it was to persuade customers to respond to what it said. Although the field experiment enabled us to observe natural interactions, it also included many external factors. In order to validate some of our findings without the involving these factors, we further conducted a laboratory experiment to investigate whether having the robot look back at the participants when they looked at it increased their perception that the robot was aware of their actions. These results supported the findings of the field experiment. Thus, we can conclude that demonstrating that a robot can recognize and respond to human behavior is important if it is to engage with people and persuade them to nod and reply to its comments.

**Keywords: robotic salesperson, field trial, multimodal conversation analysis, social presence, situation awareness**

OPEN ACCESS

***Edited by:***
*Agnieszka Wykowska,*
*Istituto Italiano di Tecnologia, Italy*

***Reviewed by:***
*Silvia Rossi,*
*University of Naples Federico II, Italy*
*Cesco Willemse,*
*Istituto Italiano di Tecnologia, Italy*

***Correspondence:**
*Hideyuki Nakanishi*
*nakanishi@ams.eng.osaka-u.ac.jp*

***Specialty section:***
*This article was submitted to Human-Robot Interaction, a section of the journal Frontiers in Robotics and AI*

***Received:*** *11 January 2019*
***Accepted:*** *22 August 2019*
***Published:*** *06 September 2019*

***Citation:***
*Iwasaki M, Zhou J, Ikeda M, Koike Y, Onishi Y, Kawamura T and Nakanishi H (2019) "That Robot Stared Back at Me!": Demonstrating Perceptual Ability Is Key to Successful Human–Robot Interactions. Front. Robot. AI 6:85. doi: 10.3389/frobt.2019.00085*

## INTRODUCTION

In recent years, several attempts have been made to integrate robots that can communicate with people into different aspects of daily life (Shiomi et al., 2006; Yamazaki et al., 2008; Gehle et al., 2014) because robots are seen as more engaging than animated characters and are perceived as more credible and informative as well as more enjoyable to interact with (Kidd and Breazeal, 2004).

Many studies have considered ways to utilize these types of robots. For example, experiments have been conducted on the use of guidance robots in museums (Shiomi et al., 2006; Lee et al., 2010; Tanaka et al., 2015); further, robots have been adopted for educational purposes (Gehle et al., 2014). Additionally, much research has focused on employing robotic salespeople in real-world shops. For example, several studies have shown that specific robot motions have a large influence on people's impressions of the robots (Kanda et al., 2001; Sidner et al., 2004, 2005; Ham et al., 2011), while other researchers have attempted to find particular robot behaviors that attract customers' attention (Yamazaki et al., 2008, 2009). In addition, several researchers have developed robots that can recognize human social behaviors and take advantage of these to attract attention (Gaschler et al., 2012; Das et al., 2015; Fischer et al., 2015). However, these types of behaviors are not always effective in different aspects of daily life. Some researchers developed a robot that can recognize social behavior recognition of human and attract the attention depending on typical social behaviors of human (Gaschler et al., 2012; Das et al., 2015; Fischer et al., 2015). However, these kind of behaviors are not efficient in every aspects of daily life.

In this research, we focus on the behaviors of a robotic salesperson. When there are foreign travelers in a shopping mall, the salespeople in the mall may not be able to speak their language. In such cases, robotic salespeople could help to serve customers, but they are easily ignored by customers due to their lack of social presence, making it difficult for them to work as salespeople. The robots' behavior should draw human attention to them and encourage customers to listen carefully to what they have to say.

In this paper, our goal is to investigate these types of behaviors of robotic salespeople. First, we conducted a field experiment in a shop in a traditional Kyoto shopping street in order to identify behaviors that could draw people's attention to the robot. In this experiment, although we observed natural, real-world interactions between the robotic salesperson and the customers, there were also many external factors. In order to validate some of our experimental findings without involving these extraneous factors, we also conducted a laboratory experiment to examine whether demonstrating the robot's ability to perceive how much attention people were paying the robot could encourage them to respond to its comments.

# RELATED WORK

## Field Trials

Many real-world experiments have already attempted to study natural interactions between robots and humans. For example, experiments have been conducted in museums (Bennewitz et al., 2005; Kuno et al., 2007; Yamazaki et al., 2008; Gehle et al., 2014) and a classroom (Tanaka et al., 2015). In addition, several experiments have employed robots as salespeople for different purposes (Lee et al., 2012; Nakagawa et al., 2013; Niemelä et al., 2017a,b). Two of these experiments were conducted in a shopping mall (Kanda et al., 2010; Shiomi et al., 2013). The first one aimed to use a robot to build customer relationships, while in the second one a robotn offered customers product coupons to improve product sales. However, the robots in these experiments did not introducepresent products to the customers directly. By contrast, in this research we would develop a robotic salesperson that can introduce customers to products in a real shop.

## Attracting Customers to Robots

Many experiments have also been conducted into attracting customers' attention to robots. For example, it was found that tracking customers' faces and head movements could attract their attention in a museum (Yamazaki et al., 2008, 2009). However, that robot was automated and could not communicate naturally with customers. In another study, they placed a robot in an information kiosk to encourage customers to communicate with the robot (Lee et al., 2010), but did not generate a large dataset. In our research, we used remotely controlled robots and conducted two long-term experiments to investigate how to attract and communicate with customers.

# FIELD EXPERIMENT

Robotic salespeople's comments tend to be easily ignored due to their weak social presence, meaning that they may not be effective. The robots' actions must therefore attract human attention and encourage people to listen to what robots have to say. In this section, we conduct a field experiment in order to identify robot behaviors that can draw people's attention to it by observing its natural, real-world interactions with customers. The fact that the robot is not-ignored means that the customer responds continuously to the robot's speech. That is, two-way conversation is established. In this section, we focus on the robot's initial utterances, drawing on previous research suggesting that first impressions are important in human–human interactions (Kelley, 1950), then investigate how to establish two-way conversations.

## Method

### Experimental Setup

We conducted the experiment in a shichimi (seven-spice blend) shop located in a traditional Kyoto shopping mall. **Figure 1A** shows a photograph of the shop, where we installed a Pepper robot as a salesperson. We used Pepper because it has many sensors, enabling us to easily obtain real-time data from the customers, as well as a robot-mounted tablet that we could use to show them pictures of the products. We developed a remote controler and installed a predetermined set of actions in the robot before the experiment began.

Here, we used the Wizard of Oz (WOZ) method (Saerbeck et al., 2010) to control the robot remotely, with an experimenter selecting appropriate reactions for Pepper based on the current situation. With the WOZ method, it takes time for the operator to determine the robot's next behavior and implement it, but this is not a serious problem when the robot is talking with visitors. Pepper's behaviors were

divided into two types: ordinary conversation and product introduction. Its ordinary conversation behaviors included greetings (such as "Hello"), handshake requests, and self-introduction, while its product introduction behaviors included offering customers a sample to try, trying to promote sales of shichimi and soft-serve ice cream, and asking a salesperson for help. When offering customers a sample, Pepper would point to the sample's location with its left hand and say, "Would you like to try a sample? You can taste here."

We placed a camera behind the robot to enable the experimenters to observe the situation in the shop and choose its next action. The robot could also turn its head automatically to focus on people's faces using a camera on its head. To record data, we set up three obscure recording cameras in the shop, as shown in **Figure 1B**. Here, the customers' faces have been obscured. We also set up one clear recording camera in the shop. When customers approached the robot, they were shown a consent form on the tablet. Only when they had given their consent did we begin recording with the camera. We also placed some handouts on the robot's leg that gave further information about the whole experiment. This experiment was approved by Osaka University's Research Ethics Committee.

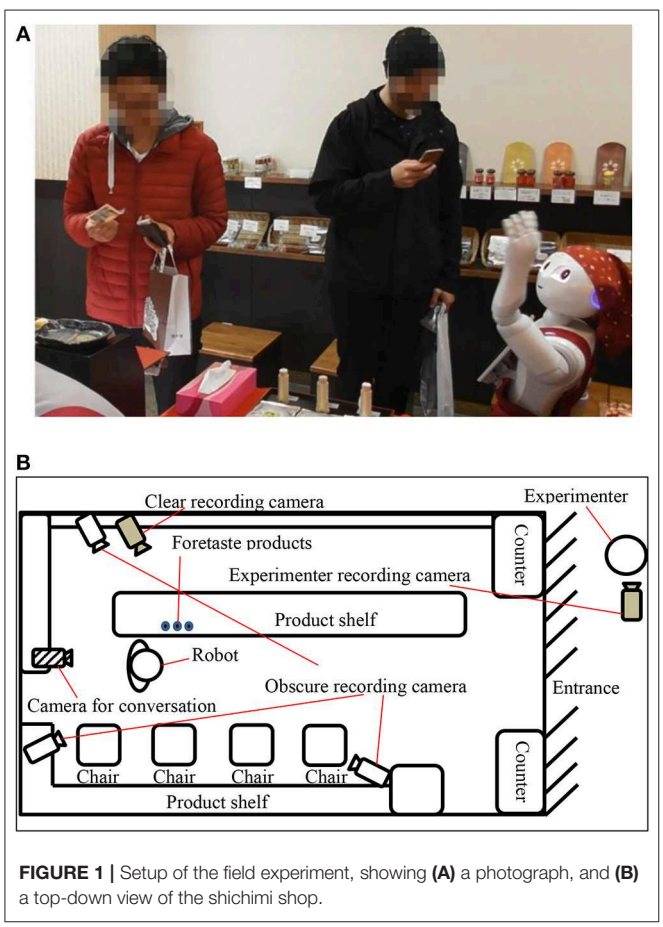



**FIGURE 1 |** Setup of the field experiment, showing **(A)** a photograph, and **(B)** a top-down view of the shichimi shop.

### Analysis Method

In order to observe and analyze the structure and patterns in the robot's interactions with customers, we conducted a multimodal conversation analysis. First, we transcribed the conversations with each group in detail based on the acquired video footage. In addition to the words spoken, the transcripts also described the timing of the customers' remarks, as well as their body movements, gaze direction, and so on. Then, we used these transcripts and videos to analyze the interactions, taking into account both verbal and non-verbal information. Here, we defined a customer group as a group of people who knew each other and entered the shop at the same time, determining this by using the video to confirm that they entered the shop together and talked to each other.

## Results and Discussion

This experiment was carried out over 10 days in 2017. During this time, around 360 customers visited the shop, divided into 164 groups with an average of 2.2 people per group.

In order for a robotic salesperson to offer services to customers and encourage them to make purchases, it needs to attract their attention to what it has to say. Thus, it was vital to investigate which types of action the robot could use to attract the customers' attention. When Pepper received two or more consecutive replies from the same customer, we defined it as a two-way conversation. However, if the customer either did not respond or only replied once, we defined it as a one-way interaction. When we looked for these two types of conversation in our experimental data, we found that 45 groups engaged in two-way conversations, compared with 119 groups whose interactions were one-way. These results suggest that the robot was ignored by most customers.

### Customers' First Impressions of the Robot Strongly Influenced Their Conversations

In society, robots are generally perceived as mechanical beings that are merely tasked with executing human orders accurately. However, unlike industrial robots, some robots now coexist with people in everyday society. Thus, the relationships between humans and robots should not only involve humans giving commands to robots, but also robots being able to communicate interactively with humans on an equal footing. In this section, we investigate which of the robotic salesperson's behaviors persuaded customers to respond to its comments.

A previous study of human–human interactions found that people's behavior toward others is shaped by their first impressions, with people who have favorable first impressions of someone tending to interact more with them than others who have formed unfavorable impressions (Kelley, 1950). Although that research focused on human–human interactions, this finding may also be applicable to human–robot interactions, so we focused on the robot's initial utterances and examined how best to establish two-way conversations.

First, we compare the group that had a two-way conversation with the group that robot spoke one-way utterances.

**Transcript 1.** The group that had a two-way conversation with Pepper (December 5th 16:17:17-16:18:26)

| | | |
|---|---|---|
| 1 | | ((C1 looks at Pepper)) |
| 2 | P | Hello!= |
| 3 | C1 | Hi! |
| 4 | C2 | Hello |
| 5 | P | My name is Pepper. |
| 6 | C1 | Hi, Pepper! |
| 7 | P | Nice to meet you. |
| 8 | C2 | Nice to meet you too. |
| 9 | C1 | Nice to meet you too. (1.0) |
| 10 | C1 | Hi, Pepper::. |
| 11 | P | May I shake hands with you? |
| 12 | C1 | Sure! Hi! Hello! ((C1 is shaking hands with Pepper)) |
| | | (…) |
| 19 | P | Would you like to try a sample? You can taste here. |
| 20 | C1 | OK! ((C1 looks at the tasting sample)) |

(()) : Supplementary explanation · Speaker's behavior
! : Lively tone
= : Speech and utterance are connected without interruption
(number) : The length of silence
":" : Stretched sound
(…) : Omission
(P = Pepper, C1 = young woman1, C2 = young woman2)

Transcript 1 shows an example of customers having a two-way conversation with Pepper. At the beginning, when they had just entered the shop (**Figure 2A**), Pepper said "Hello!" (Line 2), to which the customer replied "Hi!" (Line 3). After that, Pepper made some brief comments, to which the customers replied "Nice to meet you," (Line 7) and "May I shake hands with you?" (Line 11). We can therefore say that they engaged in a two-way conversation. Once the conversation had begun, even though the robot made slightly longer comments, such as "Would you like to try a sample? You can taste here,"(Line 19) the customer answered "OK!" and looked at the samples (Line 20). **Figure 2B** shows C1 looking at the samples.

**Transcript 2.** The group that robot spoke one-way utterances (August15th 15:59:33-16:00:50)

| | | |
|---|---|---|
| 1 | | ((Looking at the products)) |
| 2 | P | Medium hot shichimi is standard spicy for normal use. |
| 3 | P | Very hot shichimi is characterized by a numbing and exciting spicy taste. |
| 4 | | ((C1, C2 and C3 are looking at the products)) |
| 5 | P | Hello. |
| 6 | P | My name is Pepper. |
| 7 | | ((C1, C2 and C3 get away from Pepper)) |
| 8 | P | Wait, wait. Come on! Let's talk together. |

(P = Pepper, C1 = man1, C2 = man2, C3 = man3, S = salesperson)

However, Transcript 2 gives an example of a one-way interaction. When the customers entered the shop, Pepper gave a lengthy description of the shop's products, saying "Medium hot shichimi is standard spicy for normal use." (Line 2, **Figure 3A**), but they did not respond. Here, we can see that once the one-way interaction had begun, even when the robot made short and easy-to-answer comments, such as "Hello!" (Line 5) and "My name is Pepper" (Line 6), it was simply ignored (**Figure 3B**).

Comparing these two examples, we see that once the customers had responded to Pepper's comments, the subsequent conversation became two-way. By contrast, when they did not respond to Pepper's comments, the subsequent interaction was one-way. These differences are particularly noticeable at the beginning of the conversation, so the initial impression the robot gives to customers appears to be extremely important, and possibly determines the customers' subsequent attitude to it.

For all the customers who entered the shop, we looked at the robot's first utterance and the customer's initial response. For 31 out of the 35 customer groups that replied to the robot's first utterance (88.6% of cases), this resulted in a two-way conversation. By contrast, only 14 out of the 129 customer groups who did not respond to the robot's first utterance (10.8% of cases) went on to have a two-way conversation. We also validated these results using chi-squared tests, finding that the difference between the two conditions was significant ($x^2 = 83.5$, $p = 0.0063 \times 10^{-17} < 0.05$). Consequently, we believe that the initial impression given by the robot had a crucial influence on the subsequent conversation for most customer groups.

Among the 14 groups that did not initially respond to Pepper but then went on to have interactive conversations, this was mostly due to the robot using the wrong language or the customers not paying attention to its first comment. In these cases, when the robot said something later, most of the customers were surprised and responded willingly. In addition, four groups replied to the robot's first utterance but then let the interaction become one-way. However, in these cases, the customers included words that seemed to be spontaneous like "I was surprised."

So far, it is unclear whether comment length is all that matters, or whether the content is also important. We therefore compared the interactions in cases where the robot's first utterance was the same, namely "Hello," which was its most frequent initial comment. The results are shown in **Figure 4**. For 23 out of the 29 customer groups that replied to the robot's initial greeting, this resulted in a two-way conversation (79.3%). By contrast, only 5 out of the 39 customer groups who did not respond to the robot's greeting went on to have a two-way conversation (12.8%). The results of our chi-squared tests showed that the difference between the two conditions was significant ($x^2 = 30.36$, $p = 0.03 \times 10–6 < 0.05$).

Given the above, it is reasonable to assume that the impression given by the robot at the beginning of the interaction had a decisive influence on the subsequent conversation for most customer groups. Essentially, the customers' impressions of the robot were determined at the start of the interaction. If they initially perceived the robot as being similar to a voice guidance machine, its subsequent actions tended to be ignored, resulting

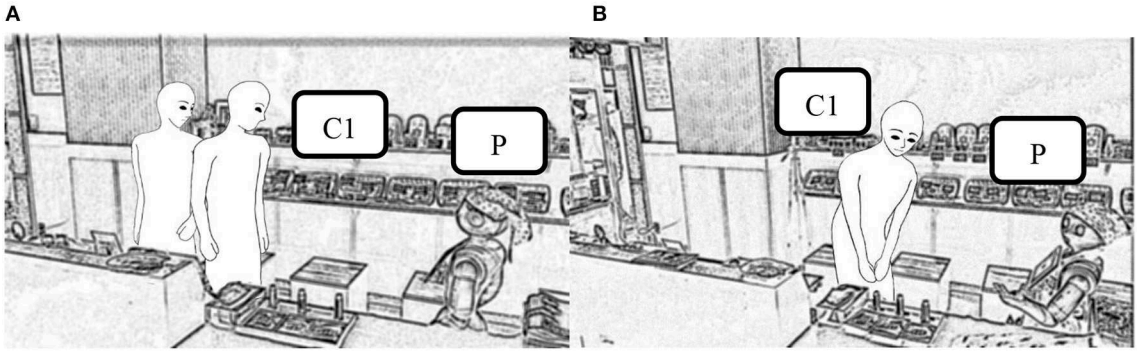



**FIGURE 2 |** Scenes from Transcript 1, showing **(A)** the robot saying "Hello!" (Line 2), and **(B)** the customers looking at the samples (Line 20).

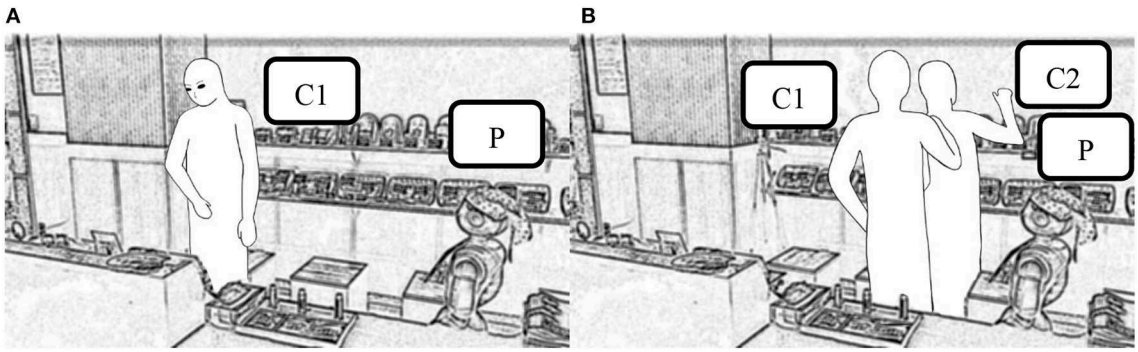



**FIGURE 3 |** Scenes from Transcript 2, showing **(A)** the robot saying a lengthy explanation (Line 2), and **(B)** the robot being ignored (Line 6).

in a one-way interaction. However, if the customers initially saw the robot as being capable of two-way dialogue, they were much more likely not to ignore its subsequent actions, resulting in a two-way conversation.

## Establishing the Two-Way Conversation

Having found that it was important for the robot to persuade customers to reply to its first utterance if it was to establish a two-way conversation, we investigated how to encourage customers to reply to the robot. Here, we focused on its initial interactions with customers and compared two customer groups, one where the robot was unable to start a conversation and another where it could.

**Transcript 3.** The robot did not start a conversation with the customers(16th August 14:31:07-14:33:54).

| | | |
|---|---|---|
| 1 | | ((Entering the shop)) |
| 2 | | ((Looking at the products)) |
| 3 | P | Would you like to try a sample? You can taste here. |
| 4 | | (3.3) |
| 5 | P | Welcome. Please feel free to watch the products. |
| 6 | P | Are you troubled to select? |
| 7 | P | May I shake hand with you? |

(P = robot, C1 = old man, C2 = old woman)

**Transcript 4.** The robot started a conversation with the customers(10th April 15:52:45-15:57:30).

| | | |
|---|---|---|
| 1 | | ((Entering the shop)) |
| 2 | | ((Looking at the products)) |
| 3 | P | May I help you? |
| 4 | | ((C1 turns his head to look at the robot)) (0.5) |
| 5 | P | Hello! |
| 6 | C2 | Hello↑! |
| 7 | C1 | Hi::! |
| 8 | P | My name is Pepper. |

(P = robot, C1 = man, C2 = woman)

In Transcript 3, the robot suggested that the customers try a sample (Line 3), but they were looking at the products and did not reply. We believe this was because they did not know whom the robot was speaking to. In Transcript 4, the robot said "Hello!" (Line 5) while the customers were looking at it (Line 4, **Figure 5**). In that case, the customers replied to the robot (Line 6), and we believe this was because the robot greeted them while they were looking at it. Thus, they realized that the robot was talking to them, establishing a state of mutual perception.

We also wanted to discover whether the robot had to greet customers quickly when they turned their heads to look at it. In Transcript 5, the customer turned her head to look at the robot, but it was slow to greet them: for 3.6 s, she was looking

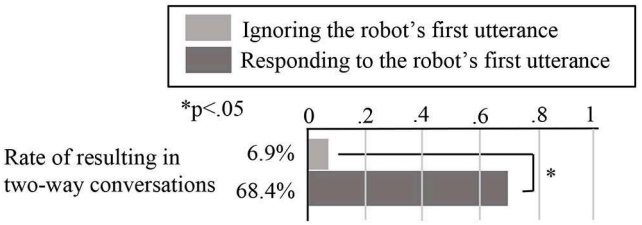



**FIGURE 4 |** Influence of an initial interaction on the following interaction.

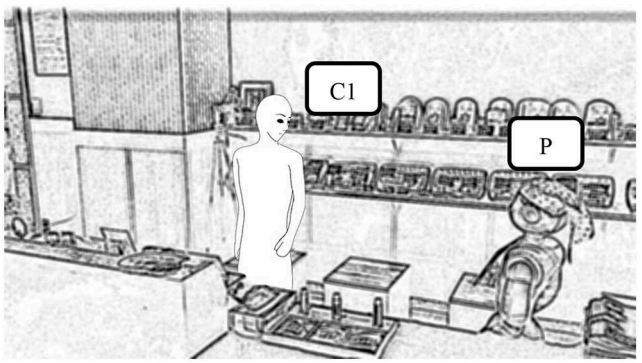



**FIGURE 5 |** The customer turns his head to look at the robot (Line 4 of Transcript 4).

at the robot but it took no action (Lines 3 and 4), and then she turned her head away to look at the products (Line 5). Thus, we believe the robot must greet customers quickly when they turn their heads to look at it, otherwise they will rapidly lose interest.

**Transcript 5.** The customer turns his head to look at the robot(16th August 15:47:27-15:50:49).

| | | |
|---|---|---|
| 1 | | ((Entering the shop)) |
| 2 | P | May I help you? |
| 3 | | ((C1 turns her head to look at the robot)) |
| 4 | | (3.6) |
| 5 | P | (C1 turns her head to look at the products) |
| 6 | P | Nice to meet you! |

(P = robot, C1 = old woman)

From the above, when customers turn to look at the robot, that is a good time for it to greet them. Engaging with them at such moments helps them to believe that the robot is aware they have turned their heads. We therefore investigated all the customer groups to see whether they responded to the robot's utterances. Of the 98 customer groups who were looking at the robot when it greeted them, 78 responded (79.6%). By contrast, only 8 out of the 66 customer groups who were looking elsewhere responded (12.1%). Thus, we can see that most of the customers who responded to the robot were looking at it when it greeted them.

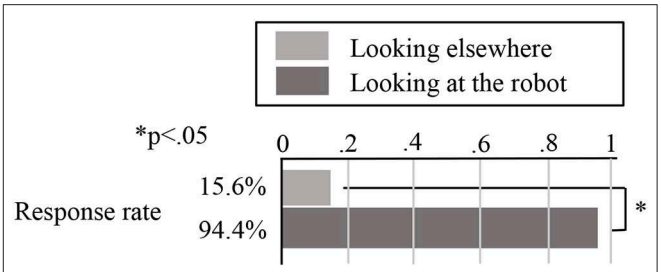



**FIGURE 6 |** Result of the customers' response.

Our chi-squared test results show that the difference between the two conditions was significant ($x^2 = 71.9$, $p = 0.021 \times 10^{-15} < 0.05$). However, this does not account for differences in the content of the robot's first utterance, so we conducted another chi-squared test for just the groups where the robot's first utterance was "Hello!" The results, shown in **Figure 6**, indicate that 34 out of the 36 customer groups who were looking at the robot when it greeted them responded to it (94.4%), compared with only 5 out of the 27 customer groups who were looking elsewhere (15.6%). Again, we can see that most of the customers who responded to the robot were looking at it when it greeted them, and our chi-squared test results show that the difference between the two conditions was significant ($x^2 = 43.02$, $p = 0.054 \times 10^{-13} < 0.05$).

If the robot responds to customers the moment they see it, this suggests that it is able to perceive the customers' behavior and degree of attention. As a result, customers are more likely to respond to the robot. Essentially, when it shows its perceptual ability to customers, its conversations with them are more likely to be interactive.

## LABORATORY EXPERIMENT

In the field experiment (Field Experiment), we found that giving the impression that the robot could recognize human behavior encouraged customers to reply. However, since this result was derived from a field experiment, there were many external factors. In order to validate some of our experimental findings without involving these extraneous factors, we also conducted a laboratory experiment to examine whether demonstrating the robot's ability to perceive how much attention people were paying it could encourage them to respond to its comments.

### Hypothesis

In this experiment, we investigated which robot's behaviors persuaded people to respond to its utterances. We believed that it needed to give the participants the impression that it could understand its surroundings, including how much attention they were paying to it, by responding to their non-verbal information. To test this, we adopted a looking-back behavior, where the robot would look back at the participants as soon as they turned their heads to look at it. We then examined whether invoking this looking-back behavior before the conversations began could

capture the participants' subsequent attention and encourage them to respond to the robot. Here, we considered the following two hypotheses.

**Hypothesis 1:** The robot's looking-back behavior increases the participants' perception that it is looking at them.

**Hypothesis 2:** The robot's looking-back behavior encourages the participants to respond to it.

# Method

## Experimental Setting

For this experiment, we adopted the simply designed robot shown in **Figure 7**. We did not add features such as eyes, a nose, or a mouth to the robot's face because we suspected that its expression might influence the participants' impressions of it. However, we did make the robot wear glasses to show the direction of its line of sight. **Figure 8** shows the experimental setup. We placed the robot behind the participant's chair because we assumed that robots would talk to customers from different directions in real-world shops. This meant that, in order to see the robot, the participants had to turn their heads first.

As a task for the participants to complete, we chose sudoku, an easy logic-based number-placement puzzle, because there was plenty that the robot could say about sudoku puzzles. This experiment was based on an experiment plan that was approved by Osaka University's Research Ethics Committee.

## Robot Design

We placed a motor in the robot's shoulder (**Figure 9A**), enabling it to move its left arm with 2° of freedom, swinging it back and forth and rotating it in and out. In addition, we added a motor to control the neck with two strings (**Figure 9B**), enabling it to move its head with 1° of freedom, namely left and right. To control the robot remotely, we developed PC-based controler software in advance. During the experiment, we observed the participants and controlled the robot with a camera that we placed beside it. The robot's utterances came from a speaker that we placed behind it, so the participants could locate it based on the direction of its voice, which was synthesized.

## Procedure

When each participant first entered the experimental room, the robot had already started talking about sudoku. The participant then stood in a waiting area and listened to one of the experimenters explain the following three points about the experiment: the participant was to solve a sudoku puzzle, the robot would signal them when to begin, and the experiment would end when they solved the puzzle. After that, the participant sat down on the chair and waited for the robot to signal them to begin the puzzle. After the robot had talked about sudoku for around 4 min, it gave a signal for the participant to begin. When the participant completed the puzzle, they rang a bell to call the experimenter. After the experiment was over, the participant answered a questionnaire about their impressions of the robot and we discussed their reasons for awarding particular scores and taking the actions they did during the experiment. Finally, the participants were debriefed after the interview.

## Conditions

To validate the hypotheses (Hypothesis), we focused on one factor and two experimental conditions.

**Factor:** the robot's looking-back behavior

**No looking-back behavior condition:** After the participant entered the laboratory, the robot kept speaking until it signaled them to begin the puzzle.

**Looking-back behavior condition:** After the participant entered the laboratory, the robot kept speaking. However, when they sat down, the robot stopped speaking to show them that it suspected they were not listening to it. After that, when the participant turned their head to look at it, it also turned its head to look at them and resumed speaking until it signaled them to begin the puzzle.

The robot spoke for the same amount of time under both conditions (around 4 min). However, it simply talked about sudoku puzzles in general, and did not include tips or ways to solve the current puzzle, so that the content of its comments did not attract the participants' attention. In addition, the robot's utterances were decided before the experiment. While speaking, its head swung from side to side every few seconds so that it turned toward each participant several times. Its left arm also moved up and down so the participants could see it was a robot when they looked at it. **Figure 7** shows the experimental procedures under both conditions. In this example, the participant turned his head to look at the robot even without the looking-back behavior, but not all participants did this. In order to provide a clear understanding of the different conditions in the laboratory experiment, a video is available (**Supplementary Material**).

## Participants

Twenty participants (10 females and 10 males) were involved in the experiment. They were all 18–24-year-old university students living in Japan, recruited for the experiment and paid for their contributions. None of them were known to the experimenters. In addition, we used a between-subjects design, because their impressions of the robot under one condition could influence their responses under the other condition.

## Behavior Evaluation

In this experiment, we first counted the number times each participant responded to the robot's utterances. A participant was seen as responding to the robot if they either made an utterance of their own or nodded without saying anything. Under the looking-back behavior condition, the experimenter observed the participants and made the robot say "Hello" and "Nice to meet you" to them when they turned their heads to look at it. By contrast, under the no looking-back behavior condition, the robot uttered each sentence at predetermined intervals.

We suspected that, if the robot left a wider interval between utterances, it was more likely that the participant would respond, so we analyzed the participants' behavior while keeping the robot's utterances exactly the same under both conditions, during the period when the robot was talking bout sudoku after potentially having looked back. Specifically, we counted the

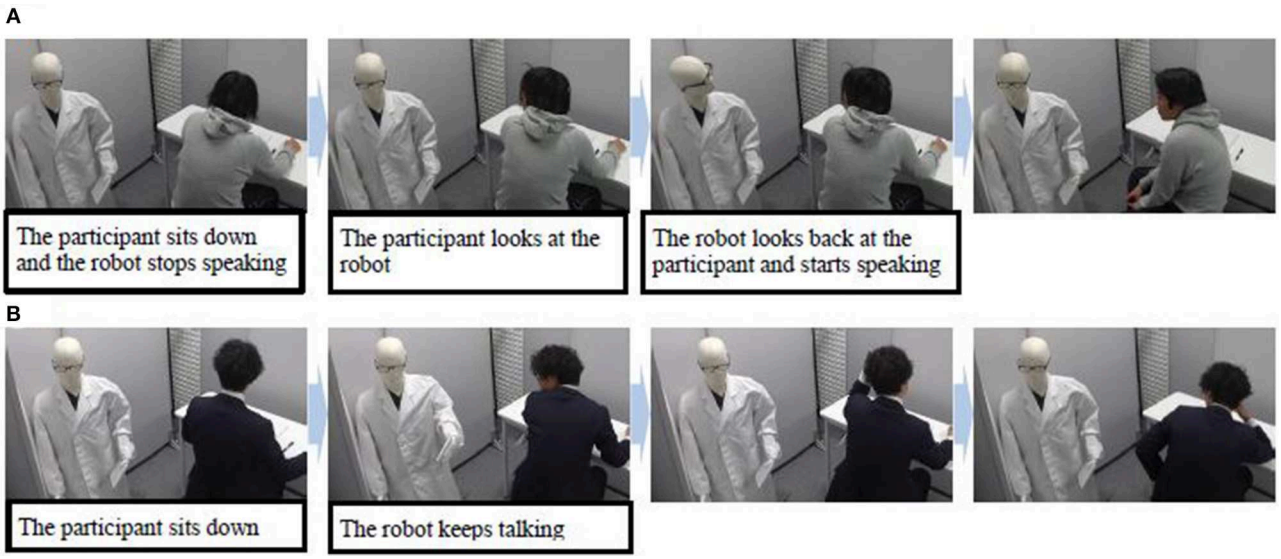



**FIGURE 7 |** Experimental procedures under both the **(A)** looking-back and **(B)** no looking-back behavior conditions.

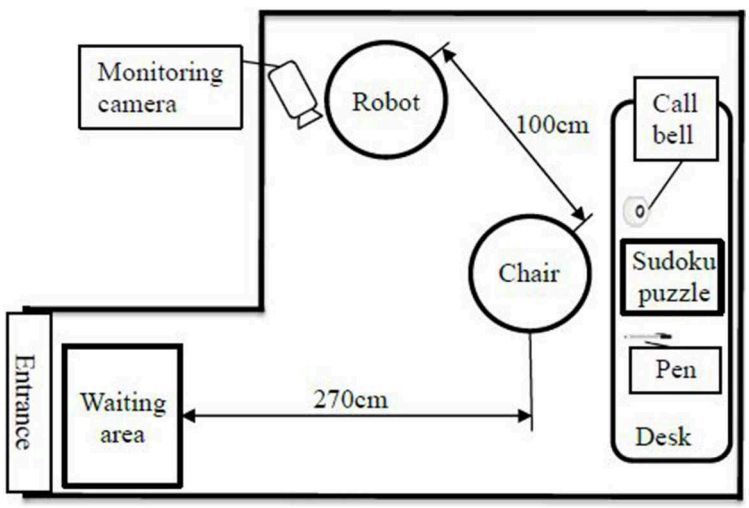



**FIGURE 8 |** Overview of the laboratory experiment.

number of responses and measured how long the participants watched the robot when it left equal intervals following each utterance under both conditions. We measured this time based on video recordings taken from the camera shown in **Figure 8**. In order to examine the changes in the participants' responses over time, we divided the robot's utterances into four parts based on time, splitting its 4 min of speech into four 1-min parts.

Under the looking-back behavior condition, all the participants had to look at the robot so if, during the experiment, the participants did not turn to look at it, we would make the robot say "Please look at me." However, we felt that this utterance ("Please look at me") led the participants to look at the robot on purpose so, when we analyzed the experimental results, we also analyzed the data with these cases excluded. In this study, two coders collected the behavioral data and we adopted Cohen's kappa statistic to validate its inter-rater reliability. The results showed that κ values for the participants' responses ($\kappa = 0.79$), time spent looking at the robot ($\kappa = 0.80$), and number of spoken replies ($\kappa = 1.0$) were all above 0.75.

## Questionnaire Evaluation

After the experiment, the participants filled out questionnaires regarding their impressions of the robot in order to evaluate Hypothesis 1. They responded using a 7-point Likert scale going from 1 (strongly disagree) through 4 (neutral) to 7 (strongly agree), and we also included a free description section.

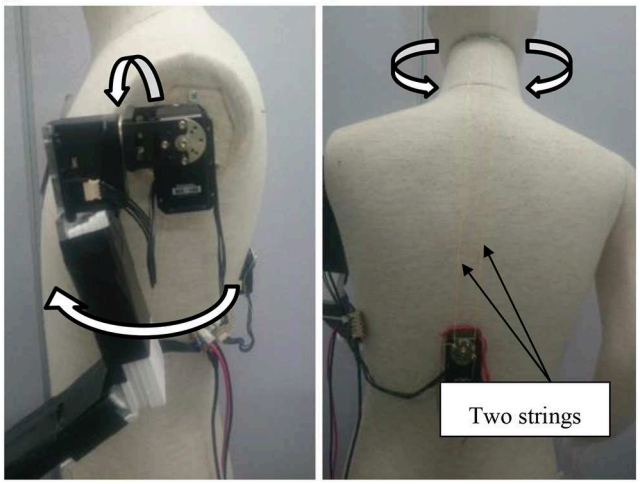



**FIGURE 9 |** Structure of the robot, showing the **(A)** left arm, and **(B)** neck.

Afterwards, we interviewed the participants about their reasons for awarding particular scores and acting as they did during the experiment. The questionnaire included the following seven questions. Here, Q1 assessed the quality of the robot's speech, Q2 checked for manipulation, and the remaining questions were related to the participants' impressions of the robot.

Q1. The robot's voice was sufficiently clear.
Q2. I felt I was being observed by the robot.
Q3. I felt the robot was waiting for my reply.
Q4. I felt the robot's behaviors were similar to human ones.
Q5. I felt I was being forced to listen to the robot.
Q6. I felt I was being forced to respond to the robot.
Q7. I felt the robot was reacting to my behavior.

## Results and Discussion

### Results

**Figure 10** shows the behavior evaluation results, and **Figure 11** shows the questionnaire results. For these analyses, we carried out two-tailed independent $t$-tests. For the speech quality question (Q1 in **Figure 11**), we found no significant difference between the two conditions, so we believe that the robot's speech quality did not influence the participants' behavior or their impressions of the robot. In addition, the results for Q2 (whether the participants felt the robot was observing them) showed a significantly higher score under the looking-back behavior condition than under the no looking-back behavior condition ($t(18) = 2.11$, $p = 0.049 < 0.05$, Cohen's $d = 0.95$), supporting Hypothesis 1. There were no significant differences between the two conditions for any of the other questions.

In the behavior evaluation results shown in **Figure 10**, the difference in the total number of responses across all four parts of the experiment between the two conditions was not significant ($t_{(18)} = 1.86$, $p = 0.079 < 0.1$, Cohen's $d = 0.83$). From **Figure 10**, the more the robot talked, the fewer responses the participants made, under both conditions. We therefore analyzed whether

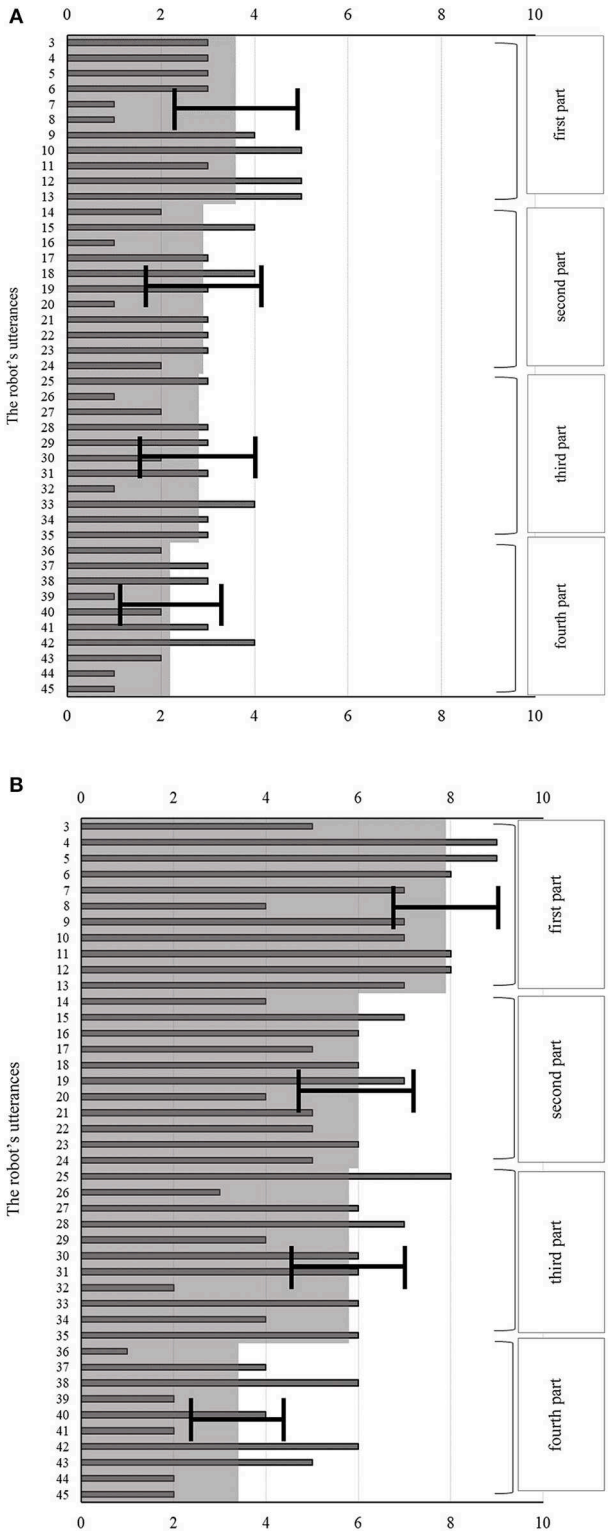



**FIGURE 10 |** Histograms representing the number of responses to each of the robot's utterances, under the **(A)** no looking-back, and **(B)** looking-back behavior conditions. Here, the bars represent the average numbers of responses, while the light gray shaded areas represent the overall average numbers of responses for each part and the error bars represent the standard errors.

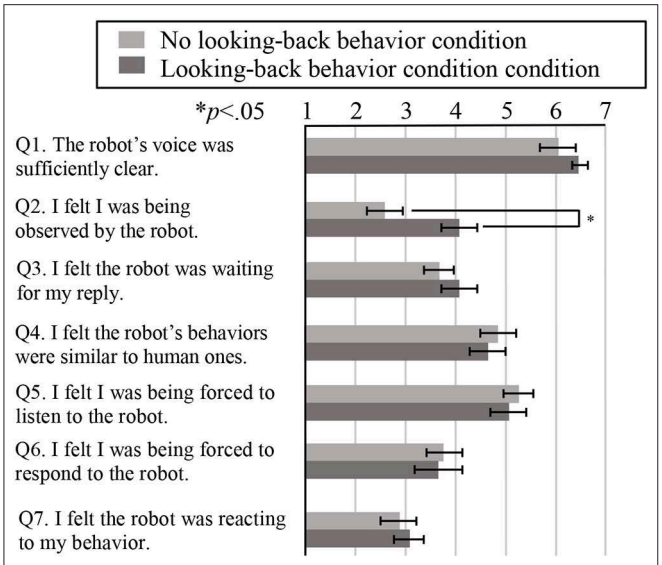



**FIGURE 11 |** Questionnaire evaluation results. Here, the bars represent the average scores for each question, while the error bars represent the standard errors.

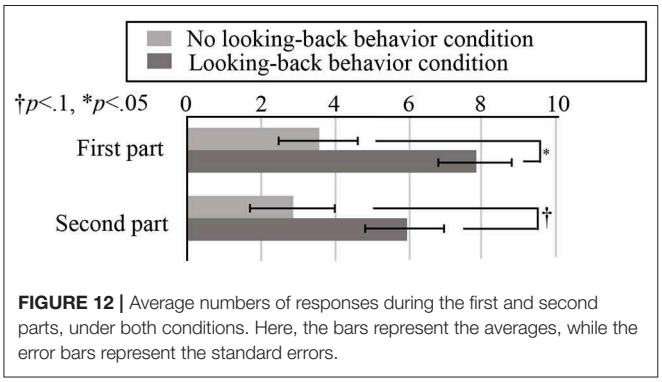



**FIGURE 12 |** Average numbers of responses during the first and second parts, under both conditions. Here, the bars represent the averages, while the error bars represent the standard errors.

the number of responses during each part differed between the two conditions.

**Figure 12** shows the average numbers of responses during the first and second parts. For the first part, we found that the difference was significant ($t_{(18)} = 2.47$, $p = 0.024 < 0.05$, Cohen's $d = 1.10$). By contrast, the difference in the numbers of responses during the second part was a non-significant tendency ($t(18) = 1.77$, $p = 0.093 < 0.1$, Cohen's $d = 0.79$). Finally, we found no significant differences in the numbers of responses during the third and fourth parts. Thus, Hypothesis 2 was only supported during the early stages of the robot's comments.

## Discussion

Regarding the questionnaire, the results for Q2 showed that the feeling of being observed by the robot was significantly stronger under the looking-back behavior condition. During the interviews, the participants made comments such as, "When I turned around, I made eye contact with the robot," and "When I turned my head to the robot, it also looked at me." This confirms that the participants had the impression that the robot was looking back in response to them turning their heads and looking at it.

From the behavior evaluation results, we see that although there was a difference in the total number of responses to the robot's utterances, it was not significant. We believe this was because the robot's looking-back behavior only occurred at the start of each experiment, so the participants' impression of the robot faded away over time and they stopped feeling that it could understand their behavior. In addition, there was no significant difference between the two conditions in the responses to questions Q5–Q7 on the questionnaire. We believe this was because the participants had much stronger impressions of the latter half of the experiment because they answered the questionnaire afterward. Moreover, when interviewed, one of the participants said that "I felt that my actions were being observed by the robot when it looked back at the beginning, but as time went on, this faded away."

Given these results, we divided the robot's utterances into four 1-min parts to investigate how the number of responses changed over time (**Figure 10**). This showed that the more the robot talked, the fewer responses the participants made, under both conditions. When we focused only on the first part of the robot's utterances, there was a significant difference in the number of responses between the two conditions, which we believe is because the robot's looking-back behavior made a strong impression on the participants during this time.

It is also possible that another reason for this was that we did not consider the concept of turn-taking. In a previous study, a robot only looked at people when it was asking them to respond (Chao and Thomaz, 2010). Thus, we might have been able to maintain the number of responses by repeating the robot's looking-back behavior while it was talking. During their post-experiment interviews, most of the participants said that "I wanted to show that I was listening to the robot." Under the looking-back behavior condition, not only their faces but also often their bodies were turned toward the robot when they were listening to it. Under the looking-back behavior condition, 9 out of the 10 participants turned their bodies toward the robot. By contrast, only 4 out of the 10 participants did the same under the no looking-back behavior condition.

From the above, we concluded that indicating the robot can understand the participants' behavior and mental state is important for increasing its social presence. The robot's initial behavior enhanced their perception of being looked at by it. After that, they would have felt guilty if they had ignored the robot, so they tried to suggest that they were listening to it and, consequently, were more willing to respond to it.

In this experiment, although we investigated the effect of the robot's looking back, we did not study which behaviors would enable it to show that it was aware of how much attention the participants were paying to it. Moreover, the looking-back behavior was performed before the conversations began, so it is possible that the participants simply become bored when the impression created by this behavior faded away. It is probable that the robot could maintain a strong social presence by performing such behaviors several times during the

conversation or adopting the previously mentioned approaches considered in related studies (Shiomi et al., 2006; Yamazaki et al., 2008, 2009). We plan to investigate these points in future work.

## GENERAL DISCUSSION

From the field experiment, we found that most of the customer groups fell into one of two categories: either the group replied to the robot's first utterance, resulting in a two-way conversation, or it did not, resulting in a one-way interaction. Therefore, we believe that the impression made by the robot at the beginning of the interaction had a crucial influence on the subsequent conversation for most customer groups. The key was to persuade the customers to reply to the robot's first utterance.

We also clarified that the best time for the robot to first talk to a customer is the moment when they turn their head to look at it. Greeting them at this time potentially makes them believe that the robot is aware they have turned their head. Thus, we believe that giving the impression of recognizing human behavior encourages customers to reply. A previous study found that gaze-based feedback can be used to signal the robot's perception, understanding, and attitude toward the communicated content (Allwood et al., 1992), which also supports our conclusion. The laboratory experiment also supported this conclusion. In addition, the robot established eye contact with the customers when they looked at it, and some studies have shown that eye contact has an impact on various cognitive processes (Senju and Hasegawa, 2005; Dalmaso et al., 2017; Xu et al., 2018). We therefore believe that establishing eye contact was also a factor in our results.

Taken together, our qualitative and statistical results lead us to conclude that indicating the robot can understand people's behavior and mental state is important for attracting their attention and makes it easier to persuade people to listen to it.

In this paper, we conducted both a field experiment and a laboratory experiment. The field experiment did not consider customer differences, such as the number of people in the group or their age, gender, or nationality, even though these could have affected their interactions with the robot. We also did not consider the effect of different utterances. We plan to investigate these issues in future work.

## CONCLUSION

This paper has focused on encouraging customers to respond to a robotic salesperson's initial utterances. With this in mind, we conducted two experiments to investigate the initial stages of human–robot interactions, namely a field experiment and a laboratory experiment, in order to investigate what types of behaviors the robot should adopt and when it should perform them. First, we conducted a field trial to observe natural interactions between a robot and customers in a real shop. Then, we conducted a laboratory experiment to investigate whether having a robot look back at the participant when they looked at it increased their perception that the robot was aware of their actions.

Based on the results, we found that suggesting the robot could recognize human behavior in the initial stages of its interactions with customers made them feel as if it was looking at them and encouraged them to respond to its utterances. Our most important finding is that, in conversations between people and robots, it is important to suggest that the robots are aware of their behavior and state of mind. Such behavior makes people feel that the robot can understand their behavior and respond accordingly, so they are more likely to show they are listening to it. We hope that this research will promote human–robot conversation and enable us to use robots more effectively.

## ETHICS STATEMENT

Our experiments have approved from Research Ethics Committee of Osaka University.

## AUTHOR CONTRIBUTIONS

MIw contributed to design, conduction, data collection, analysis of the both experiments, and writing the paper. JZ contributed to design, conduction, data collection, analysis of the field experiment and writing the paper. MIk contributed to design, conduction, and data collection of the field experiment. YK contributed to design, conduction, data collection, and analysis of the laboratory experiment. YO contributed to design and analysis of the laboratory experiment. TK contributed to design and conduction of the field experiment. HN critically reviewed the paper and gave the final approval of the version submitted.

## FUNDING

This study was supported by HAYAO NAKAYAMA Foundation for Science & Technology and Culture, ROIS NII Open Collaborative Research 19FC01, and JSPS KAKENHI Grant Numbers JP19H00605, JP19K21718.

## ACKNOWLEDGMENTS

We thank Dintora for providing us an environment for conducting a field experiment. And we also thank Yusuke Inoue for helping us conduct the field experiment.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2019.00085/full#supplementary-material

## REFERENCES


Allwood, J., Nivre, J., and Ahlsen, E. (1992). On the semantics and pragmatics of linguistic feedback. *J. Semantics* 9, 1–26. doi: 10.1093/jos/9.1.1

Bennewitz, M., Faber, F., Joho, D., Schreiber, M., and Behnke, S. (2005). "Towards a humanoid museum guide robot that interacts with multiple persons," in *Humanoid Robots, 2005 5th IEEE-RAS International Conference on* (Tsukuba: IEEE), 418–423. doi: 10.1109/ICHR.2005.1573603

Chao, C., and Thomaz, A. L. (2010). "Turn taking for human-robot interaction," in *2010 AAAI Fall Symposium Series* (Arlington, VA).

Dalmaso, M., Castelli, L., and Galfano, G. (2017). Attention holding elicited by direct-gaze faces is reflected in saccadic peak velocity. *Exp. Brain Res.* 235, 3319–3332. doi: 10.1007/s00221-017-5059-4

Das, D., Rashed, M. G., Kobayashi, Y., and Kuno, Y. (2015). Supporting human robot interaction based on the level of visual focus of attention. *IEEE Trans. Hum. Mach. Syst.* 45, 664–675. doi: 10.1109/THMS.2015.2445856

Fischer, K., Yang, S., Mok, B., Maheshwari, R., Sirkin, D., and Ju, W. (2015). "Initiating interactions and negotiating approach: a robotic trash can in the field," in *2015 AAAI Spring Symposium Series* (Palo Alto, CA).

Gaschler, A., Jentzsch, S., Giuliani, M., Huth, K., de Ruiter, J., and Knoll, A. (2012). "Social behavior recognition using body posture and head pose for human-robot interaction," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems* (Vilamoura: IEEE), 2128–2133. doi: 10.1109/IROS.2012.6385460

Gehle, R., Pitsch, K., and Wrede, S. (2014). "Signaling trouble in robot-togroup interaction. emerging visitor dynamics with a museum guide robot," in *Proceedings of the Second International Conference on Human-Agent Interaction* (Tsukuba: ACM), 361–368. doi: 10.1145/2658861.2658887

Ham, J., Bokhorst, R., Cuijpers, R., van der Pol, D., and Cabibihan, J. J. (2011). "Making robots persuasive: the influence of combining persuasive strategies (gazing and gestures) by a storytelling robot on its persuasive power," in *Proceedings of the Third International Conference on Social Robotics, ICSR'11* (Berlin: Springer), 71–83. doi: 10.1007/978-3-642-25504-5_8

Kanda, T., Ishiguro, H., and Ishida, T. (2001). "Psychological analysis on human-robot interaction," in *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, vol. 4. (Seoul: IEEE), 4166–4173. doi: 10.1109/ROBOT.2001.933269

Kanda, T., Shiomi, M., Miyashita, Z., Ishiguro, H., and Hagita, N. (2010). A communication robot in a shopping mall. *IEEE Trans. Robot.* 26, 897–913. doi: 10.1109/TRO.2010.2062550

Kelley, H. (1950). The warm-cold variable in first impressions of persons. *J. Personal.* 18, 431–439. doi: 10.1111/j.1467-6494.1950.tb01260.x

Kidd, C., and Breazeal, C. (2004). "Effect of a robot on user perceptions," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (Sendai: IROS), 3559–3564. doi: 10.1109/IROS.2004.1389967

Kuno, Y., Sadazuka, K., Kawashima, M., Yamazaki, K., Yamazaki, A., and Kuzuoka, H. (2007). "Museum guide robot based on sociological interaction analysis," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose: ACM), 1191–1194. doi: 10.1145/1240624.1240804

Lee, M. K., Forlizzi, J., Kiesler, S., Rybski, P., Antanitis, J., and Savetsila, S. (2012). "Personalization in HRI: a longitudinal field experiment," in *Proceedings of the Seventh Annual ACMIEEE International Conference on Human-Robot Interaction* (Boston, MA), 319–326. doi: 10.1145/2157689.2157804

Lee, M. K., Kiesler, S., and Forlizzi, J. (2010). "Receptionist or information kiosk: how do people talk with a robot?" in *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (Savannah: ACM), 31–40. doi: 10.1145/1718918.1718927

Nakagawa, K., Shiomi, M., Shinozawa, K., Matsumura, R., Ishiguro, H., and Hagita, N. (2013). Effect of robots whispering behavior on peoples' motivation. *Int. J. Soc. Robot.* 5, 5–16. doi: 10.1007/s12369-012-0141-3

Niemelä, M., Arvola, A., and Aaltonen, I. (2017a). "Monitoring the accep-¨ tance of a social service robot in a shopping mall: first results," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna: ACM), 225–226. doi: 10.1145/3029798.3038333

Niemelä, M., Heikkilä, P., and Lammi, H. (2017b). "A social service robot¨ in a shopping mall: expectations of the management, retailers and consumers," in *The Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna: ACM), 227–228. doi: 10.1145/3029798.3038301

Saerbeck, M., Schut, T., Bartneck, C., and Janse, M. D. (2010). "Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta: ACM), 1613–1622. doi: 10.1145/1753326.1753567

Senju, A., and Hasegawa, T. (2005). Direct gaze captures visuospatial attention. *Visual Cognit.* 12, 127–144. doi: 10.1080/13506280444000157

Shiomi, M., Kanda, T., Ishiguro, H., and Hagita, N. (2006). "Interactive humanoid robots for a science museum," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction* (Salt Lake City: ACM), 305–312. doi: 10.1145/1121241.1121293

Shiomi, M., Shinozawa, K., Nakagawa, Y., Miyashita, T., Sakamoto, T., Terakubo, T., et al. (2013). Recommendation effects of a social robot for advertisement-use context in a shopping mall. *Int. J. Soc. Robot.* 5, 251–262. doi: 10.1007/s12369-013-0180-4

Sidner, C. L., Kidd, C. D., Lee, C., and Lesh, N. (2004). "Where to look: a study of human-robot engagement," in *Proceedings of the 9th International Conference on Intelligent User Interfaces* (Funchal, Madeira: ACM), 78–84. doi: 10.1145/964442.964458

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., and Rich, C. (2005). Explorations in engagement for humans and robots. *Art. Intel.* 166, 140–164. doi: 10.1016/j.artint.2005.03.005

Tanaka, F., Isshiki, K., Takahashi, F., Uekusa, M., Sei, R., and Hayashi, K. (2015). "Pepper learns together with children: development of an educational application," in *Humanoid Robots (Humanoids), 2015 IEEE-RAS 15th International Conference* on (Seoul: IEEE), 270–275. doi: 10.1109/HUMANOIDS.2015.7363546

Xu, S., Zhang, S., and Geng, H. (2018). The effect of eye contact is contingent on visual awareness. *Front. Psychol.* 9:93. doi: 10.3389/fpsyg.2018.00093

Yamazaki, A., Yamazaki, K., Kuno, Y., Burdelski, M., Kawashima, M., and Kuzuoka, H. (2008). "Precision timing in human-robot interaction: coordination of head movement and utterance," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence: ACM), 131–140. doi: 10.1145/1357054.1357077

Yamazaki, K., Yamazaki, A., Okada, M., Kuno, Y., Kobayashi, Y., Hoshi, Y., et al. (2009). "Revealing gauguin: engaging visitors in robot guide's explanation in an art museum," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ACM), 1437–1446. doi: 10.1145/1518701.1518919


**Conflict of Interest Statement:** TK is the CEO of Kyoto Innovation, Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 09 July 2019
doi: 10.3389/frobt.2019.00054

# Exploring the Effects of a Social Robot's Speech Entrainment and Backstory on Young Children's Emotion, Rapport, Relationship, and Learning

*Jacqueline M. Kory-Westlund* and Cynthia Breazeal*

*MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, United States*

OPEN ACCESS

***Edited by:***
*Cigdem Beyan,*
*Istituto Italiano di Tecnologia, Italy*

***Reviewed by:***
*Vicky Charisi,*
*Joint Research Centre, European Commission, Belgium*
*Sofia Serholt,*
*University of Gothenburg, Sweden*
*Takamasa Iio,*
*University of Tsukuba, Japan*

****Correspondence:***
*Jacqueline M. Kory-Westlund*
*jakory@alum.mit.edu*

***Specialty section:***
*This article was submitted to Human-Robot Interaction, a section of the journal Frontiers in Robotics and AI*

***Received:*** *06 January 2019*
***Accepted:*** *24 June 2019*
***Published:*** *09 July 2019*

***Citation:***
*Kory-Westlund JM and Breazeal C (2019) Exploring the Effects of a Social Robot's Speech Entrainment and Backstory on Young Children's Emotion, Rapport, Relationship, and Learning. Front. Robot. AI 6:54. doi: 10.3389/frobt.2019.00054*

In positive human-human relationships, people frequently mirror or mimic each other's behavior. This mimicry, also called entrainment, is associated with rapport and smoother social interaction. Because rapport in learning scenarios has been shown to lead to improved learning outcomes, we examined whether enabling a social robotic learning companion to perform rapport-building behaviors could improve children's learning and engagement during a storytelling activity. We enabled the social robot to perform two specific rapport and relationship-building behaviors: speech entrainment and self-disclosure (shared personal information in the form of a backstory about the robot's poor speech and hearing abilities). We recruited 86 children aged 3–8 years to interact with the robot in a 2 × 2 between-subjects experimental study testing the effects of robot entrainment *Entrainment* vs. *No entrainment* and backstory about abilities *Backstory* vs. *No Backstory*. The robot engaged the children one-on-one in conversation, told a story embedded with key vocabulary words, and asked children to retell the story. We measured children's recall of the key words and their emotions during the interaction, examined their story retellings, and asked children questions about their relationship with the robot. We found that the robot's entrainment led children to show more positive emotions and fewer negative emotions. Children who heard the robot's backstory were more likely to accept the robot's poor hearing abilities. Entrainment paired with backstory led children to use more of the key words and match more of the robot's phrases in their story retells. Furthermore, these children were more likely to consider the robot more human-like and were more likely to comply with one of the robot's requests. These results suggest that the robot's speech entrainment and backstory increased children's engagement and enjoyment in the interaction, improved their perception of the relationship, and contributed to children's success at retelling the story.

**Keywords: children, entrainment, language development, peer modeling, rapport, relationship, robotics, storytelling**

# 1. INTRODUCTION

Social robots have been designed as peers, tutors, and teachers to help children learn a variety of subjects (Belpaeme et al., 2018), including math (Clabaugh et al., 2015; Kennedy et al., 2015), language (Movellan et al., 2009; Kory and Breazeal, 2014; Gordon et al., 2016; Kory Westlund et al., 2017a,b; Vogt et al., 2017; Rintjema et al., 2018), reading (Gordon and Breazeal, 2015), handwriting (Hood et al., 2015), social skills (Robins et al., 2005; Scassellati et al., 2018), curiosity (Gordon et al., 2015), and a growth mindset (Park et al., 2017b). Prior work has explored how social robots can best engage children in learning activities and improve learning outcomes, using, e.g., personalization of behavior or curriculum (Gordon and Breazeal, 2015; Hood et al., 2015; Gordon et al., 2016; Baxter et al., 2017; Scassellati et al., 2018), appealing appearance and personality (Kory and Breazeal, 2014), and appropriate nonverbal behaviors (Kennedy et al., 2015; Kory Westlund et al., 2017a,b). One aspect of human-human interpersonal interaction that has been linked to improved learning outcomes in peer tutoring situations is rapport and positive relationships (Sinha and Cassell, 2015a,b). Because of this link, we hypothesize that improving a social robot's capabilities for building rapport and positive relationships with children may similarly lead to improved learning outcomes.

Some prior work with adults provides evidence in support of this hypothesis (Kidd and Breazeal, 2008; Lubold et al., 2016, 2018; Lubold, 2017); however, there is little work yet exploring a social robot's rapport and relationship with young children. Thus, in this paper, we explored whether enabling a social robot to perform rapport-building behaviors, including speech and behavior entrainment, and giving the robot an appropriate backstory regarding its abilities, could help establish rapport and generate positive interactions with children, which we hypothesized could improve children's learning and engagement.

# 2. BACKGROUND

## 2.1. Relationships, Rapport, and Learning

We have strong evidence that children's peer relationships provide bountiful opportunities for learning via observing peers, being in conflict with peers, and cooperating with peers (Piaget, 1932; Bandura and Walters, 1963; Bandura, 1971; Vygotsky, 1978; Tudge and Rogoff, 1989; Rubin et al., 1998; De Lisi and Golbeck, 1999; Whitebread et al., 2007). The research so far on children's peer learning discusses how children might learn from other, but does not yet thoroughly address what precisely modulates peer learning. That is: Are all peers approximately equivalent as sources to promote learning, or is there something about some peers that makes them "better inputs" than others? In the context of social robots, what is it about a social robot that could lead children to learn more, or less?

Two possible modulating factors are rapport and a positive relationship. Some recent work has linked rapport to improved learning outcomes in older children's human-human peer tutoring situations (Sinha and Cassell, 2015a,b). In addition, the social bonds between children and teachers can predict learner performance (Wentzel, 1997). Other research has shown that children may learn math concepts from media characters more effectively when they have stronger parasocial relationships with those characters (Gola et al., 2013; Richards and Calvert, 2017).

Many different social and relational factors can increase rapport, trust, and engagement with virtual agents and robots. For example, using appropriate social cues (Desteno et al., 2012; Lee et al., 2013; Breazeal et al., 2016b), contingent backchanneling (Park et al., 2017a), nonverbal mirroring (Bailenson et al., 2005; Burleson and Picard, 2007; Lubold et al., 2018), responsiveness and proactivity (Kim et al., 2006), increased social presence (Lester et al., 1997), and matching ethnic communication styles (Cassell et al., 2009) all have had positive effects.

We chose to implement two rapport- and relationship-building behaviors in a social robot to explore their effects on young children's engagement and learning: speech entrainment and self-disclosure (shared personal information).

## 2.2. Speech Entrainment

In positive human-human interpersonal interactions, people frequently mimic each other's behavior—such as posture, affect, speech patterns, gestures, facial expressions, and more—unconsciously, without awareness or intent (Davis, 1982; Grammer et al., 1998; Philippot et al., 1999; Provine, 2001; Lakin et al., 2003; Semin and Cacioppo, 2008; Reitter et al., 2011; Borrie and Liss, 2014). This mimicry, also called entrainment, is considered a signal of rapport and has been observed in a variety of human relationships (Tickle-Degnen and Rosenthal, 1990; Dijksterhuis and Bargh, 2001; Rotenberg et al., 2003; Dijksterhuis, 2005; Chartrand and van Baaren, 2009; Wiltermuth and Heath, 2009; Lubold, 2017), as well as with robots and virtual agents (Breazeal, 2002; Bell et al., 2003; Suzuki and Katagiri, 2007; Levitan et al., 2016). While there is less work exploring mimicry and rapport in children, there is some showing that infants and children mimic emotions with humans (Haviland and Lelwica, 1987; Chisholm and Strayer, 1995; Rotenberg et al., 2003) and with robots (Gordon et al., 2016). Thus, enabling a robot to perform entrainment could significantly increase children's rapport with it. We chose speech entrainment because language learning is often a dialogue-heavy activity, and thus, would perhaps be more noticeable and relevant than entraining other behaviors. In addition, given the morphology and technical limitations of the robot platform we had available for this study (the Tega robot, described below), speech entrainment was one of the most feasible behaviors to study, though other behaviors could also be examined in the future (such as posture or affect).

Speech entrainment involves matching the vocal features such as speaking rate, intensity, pitch, volume, and prosody of one's interlocutor. This mimicry tends to happen unconsciously, and more often when rapport has been established—i.e., when one feels closer to or more positively about one's interlocutor (Porzel et al., 2006; Reitter et al., 2011; Borrie and Liss, 2014). Some recent work has explored increasing prosodic synchrony in a speech-controlled child-robot game in order to promote cooperation and improve enjoyment (Chaspari and Lehman, 2016; Sadoughi et al., 2017). In addition, Lubold and colleagues developed several social voice-adaptive robots that adjust the

pitch of the robot's text-to-speech voice to match that of its human interlocutor (Lubold et al., 2015, 2016, 2018; Lubold, 2017). This vocal entrainment contributed to increased learning with undergraduate students as well as middle school students during math tasks, but did not increase self-reported rapport. However, our work differs in several ways. We are investigating the impact of entrainment with younger children in a more social task—language learning—that may be more affected by social relationships. Second, these prior studies compared a robot with a text-to-speech voice to one that had a more expressive (albeit contingently adapted) voice. They did not control for the expressivity of the voice. Other recent work found that a robot with a more expressive voice was more effective as a learning companion, leading to greater engagement and learning, than a robot that used a flat voice, similar to a classic text-to-speech voice (Kory Westlund et al., 2017b). This work raises the question of whether the effects seen in Lubold et al.'s studies are strictly a result of the entrainment or a result of the robot's voice being more expressive. In the work presented here, we control for the robot's expressivity.

## 2.3. Backstory (Personal Self-Disclosure)

Backstory is the story told by or about an agent, including personal story (e.g., origin, family, hobbies), capabilities, limitations, and any other personal information that might be disclosed. With young children in particular, we expect that sharing information about an agent in a story context could make it easier for children to understand.

Prior work has shown that the story told about a robot prior to interaction can change how people perceive the robot and interact with it. Telling participants that a robot is a machine vs. a human-like, animate agent (Stenzel et al., 2012; Klapper et al., 2014; Kory Westlund et al., 2016b) or giving the robot a name and a story involving greater agency and experience (Darling et al., 2015) can manipulate people's perceptions of the robot as an animate, social agent as well as their empathy for the agent. These studies build on extensive work in social cognition and social psychology literature regarding the idea that framing or priming can influence subsequent behavior and perception (Dijksterhuis and Bargh, 2001; Biernat, 2004). However, it is not only stories told before an interaction, but also the content *of* an interaction that affects people's perceptions of their interlocutor. For example, one aspect of children's friendships and positive relationships is self-disclosure. Children disclose more information, and more personal information, in closer relationships (Rotenberg and Mann, 1986; Rotenberg, 1995). The amount of disclosure during conversation reflects how close two children feel to one another. A robot that discloses personal information may impact not only relationship formation and perception, but the story it tells could also impact how a child perceives how social an agent the robot is.

Backstory can also increase engagement with an agent. For example, in one study, giving a robot receptionist a scripted backstory during a long-term deployment increased engagement, since the story added interesting variation and history to the interactions people had with it (Gockley et al., 2005). However, no research as yet has examined the impact a backstory can have on young children's learning.

Part of our goal in giving the robot a backstory was to promote a more positive relationship. Thus, we examined specific interventions regarding the acceptance of peers and how these interventions might play into the story told about the robot. Favazza and colleagues explored how to promote the acceptance of peers with disabilities in children's kindergarten classrooms, as well as how to measure that acceptance (Favazza and Odom, 1996; Favazza et al., 2000). One component of the intervention they used involved telling stories with guided discussion about children with disabilities; a second component involved structured play with the peers who had disabilities. We combined the idea of telling a story about one of the robot's relevant difficulties that could be perceived as a disability—namely, its hearing and listening abilities—with the idea of self-disclosure as a component of children's friendships; and followed this story/disclosure with several structured activities with the robot.

There are ethical concerns regarding deception when giving robots stories that may elicit empathy, trust, or acceptance. In this study, the backstory we chose to use was fairly reflective of the actual limitations and capabilities of social robots. It pertained to the robot's difficulties with hearing and listening and was thus fairly realistic and not particularly deceptive, given general difficulties in social robotics with automatic speech recognition and natural language understanding. The remainder of the backstory discussed the robot's interest in storytelling and conversation, which was deceptive in that robots do not really have interests, but served to present the robot as a character with interests in these subjects in order to promote engagement in learning activities.

# 3. METHODOLOGY

## 3.1. Research Questions

We wanted to explore whether a social robot that entrained its speech and behavior to individual children and provided an appropriate backstory about its abilities could increase children's rapport, positive relationship, acceptance, engagement, and learning with the robot during a single session.

## 3.2. Design

The experiment included two between-subjects conditions: Robot entrainment (*Entrainment* vs. *No entrainment*) and Backstory about abilities (*Backstory* vs. *No Backstory*). We abbreviate the four conditions as *E-B, E-NB, NE-B,* and *NE-NB*. In the *Entrainment (E)* condition, the robot's speech was entrained based on each child's speaking rate, pitch, and volume, and exuberance. In the *Backstory (B)* condition, the experimenter explained that the robot was not so good at hearing and needed practice; this backstory was reinforced by the robot later.

## 3.3. Participants

We recruited 95 children aged 3–8 years (47 female, 48 male) from the general Boston area to participate in the study. We recruited a wide age range in order to recruit a sufficient number of participants and also because we were interested in seeing

**TABLE 1 |** Demographic information about the participants by condition.

| Condition | Mean age (SD) | Girls | Boys | Monolingual | Bilingual |
|---|---|---|---|---|---|
| E-B | 5.40 (1.54) | 11 | 9 | 12 | 8 |
| E-NB | 5.21 (1.34) | 7 | 9 | 9 | 7 |
| NE-B | 5.44 (1.67) | 13 | 15 | 18 | 10 |
| NE-NB | 5.27 (1.35) | 13 | 9 | 11 | 11 |

whether older children (e.g., 6–8 years) or younger children (e.g., 3–5 years) might relate differently to the robot's relational behavior, since children may develop relationships differently as they grow older (Hartup et al., 1988; Rubin et al., 1998).

Nine children were removed from analysis because they did not complete the study[1]. The children in the final sample included 86 children aged 3–8 (44 female, 42 male), with a mean age of 5.31 years ($SD = 1.43$). Of these, 3 were 3-year-olds, 30 were 4-year-olds, 19 were 5-year-olds, 15 were 6-year-olds, and 9 were 7-year-olds, and 10 were 8-year-olds. Forty-nine children spoke English only; 37 children were bilingual.

We used random counterbalanced assignment to assign children to conditions. There were 20 in the *E-B* condition, 16 in the *E-NB* condition; 28 children in the *NE-B* condition; and 22 in the *NE-NB* condition. The imbalance was a result of the children who did not complete the study. **Table 1** lists age, gender, and bilingual status by condition. Age did not significantly differ by condition. We asked parents to rate their children's social behavior on a variety of dimensions; these ratings also did not significantly differ by condition.

Children's parents gave written informed consent prior to the start of the study, and all children assented to participate. The protocol was approved by the MIT Committee on the Use of Humans as Experimental Subjects.

## 3.4. Hypotheses

We expected that the robot's entrainment and backstory might affect both children's rapport and social behavior, as well as learning and retention, during a single session with the robot. Accordingly, we used a variety of measures to explore the effects of the robot's entrainment and backstory. We tentatively expected the following results:

**Learning**

- **H1:** In all conditions, children would learn the target vocabulary words presented in the robot's story. In prior studies, we have seen children learn new words from stories told by robots (Kory, 2014; Kory Westlund et al., 2017b; Park et al., 2019). However, we expected that children would learn more as a result of the robot's entrainment or from an increased relationship, i.e., the most in the *E-B* condition, followed by the *E-NB* and *NE-B* conditions, and the least in the *NE-NB* condition.
- **H2:** Children who learned the target vocabulary words would also use them in their story retells. We have previously seen children mirror a robot's vocabulary words in their own stories (Brennan, 1996; Iio et al., 2015; Kory Westlund et al., 2017b).
- **H3:** Because of the expected connection between the robot's entrainment and backstory to children's rapport and relationship, as well as prior work showing that the story told about a computer's limitations influenced participants' lexical entrainment (Pearson et al., 2006), we expected the entrainment and backstory would lead to differences in children's mirroring of the robot's story in their retells. Children in the *E-B* condition would produce more vocabulary, longer stories, and phrase mirroring because of more rapport and a closer relationship.

**Rapport, Relationship, and Social Behavior**

- **H4:** A robot with an appropriate backstory about its abilities (*E-B* and *NE-B* conditions) would lead to greater acceptance by children of the robot and more helping behaviors.
- **H5:** Both entrainment and backstory would lead children to treat the robot as a greater social other, such as laughing and smiling more (Provine, 2001; Smidl, 2006), and affording the robot courtesies such as saying goodbye or considering its preferences (Reeves and Nass, 1996). We expected to see this more in the *E-B* than the other conditions; and least in the *NE-NB* condition.
- **H6:** Children would show greater rapport, entrainment, mirroring, and helping behaviors with a robot that entrained to them (*E-B* and *E-NB* conditions). We also expected that a robot with both an appropriate backstory and entrainment (*E-B*) would promote a stronger relationship, and as a result, greater attention, engagement, rapport, and mirroring than in the *E-NB* condition. Furthermore, children's attention, engagement, and positive emotions would increase—or at least decrease less—over the course of the session than in the other conditions.
- **H7:** Children who reported a closer relationship to the robot would also show more mirroring behaviors, more helping behaviors, greater rapport, greater engagement, and more learning. We expected a connection between children's relationship and their learning because of prior work showing that rapport can facilitate learning in peer tutoring scenarios (Sinha and Cassell, 2015a,b).

## 3.5. Procedure

Five different experimenters (three female adults and two male adults) ran the study in pairs in a quiet room in the lab. The study setup is shown in **Figure 1**. One experimenter interacted with the child. The second experimenter was present in the room, but sat back behind a laptop and did not interact directly with

[1]The children who failed to complete the study were primarily younger children (one 3-year-old, five 4-year-olds, one 5-year-old, and two six-year-olds). Most were very distracted during the session and did not want to play with the robot for the full duration of the session. One 4-year-old and the 3-year-old appeared scared of the robot and did not want to interact at all, even with parental prompting. One of the 6-year-olds had accidentally signed up for the study twice, and this was not noticed until after we began the session.

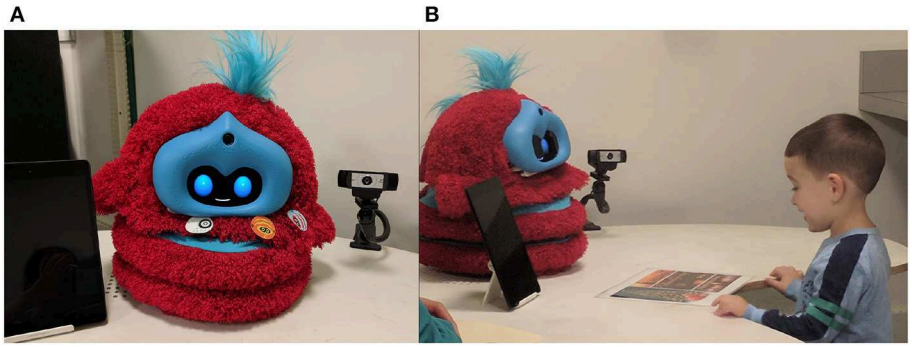



**FIGURE 1 | (A)** The robot was placed on a table. The tablet was set upright to the left (when facing the robot), and the camera behind the robot and to the right. **(B)** A child discusses holidays with the robot in the picture conversation task. Written informed consent was obtained to use this image.

the child; their role was to teleoperate the robot and manage the other equipment. Some children wished their parents to stay with them (e.g., if they were particularly shy); in these cases children's parents were instructed to watch only and let their children do as much as possible by themselves.

For each child, the interaction with the robot lasted about 20 min, followed by 5–10 min for the posttests. The interaction script, full interaction procedure, and other study materials are available for download from figshare at: https://doi.org/10.6084/m9.figshare.7175273; they are available for download as **Supplementary Materials**.

The experimenter introduced the sleeping robot, Tega, to the child and explained that it liked looking at pictures and telling stories. If the child was in the Backstory condition, the experimenter also explained that Tega sometimes had trouble hearing: "Do you see Tega's ears? Tega's ears are hiding under all the fur, so sometimes Tega's ears don't work very well. Tega sometimes has a lot of trouble hearing. You should talk to Tega in a loud and clear voice so Tega can hear you. Try to be understanding if Tega needs to hear something again." Then, in all conditions, the experimenter invited the child to help wake up the robot.

The robot interaction had four main sections: A brief introductory conversation (providing context for sharing the backstory, 2–3 min), a conversation about pictures (providing opportunities for speech entrainment and a helping/compliance request, 5–6 min), a sticker task (a sharing/compliance request, 1 min), a storytelling activity (providing opportunities to learn words and mirror the robot's speech, 10–12 min), and a brief closing conversation (1–2 min).

In the introductory conversation, the robot introduced itself, shared personal information about its favorite color and an activity it liked doing, and prompted the child for disclosure in return. Then, in the Backstory condition, the robot reinforced the backstory provided by the experimenter earlier, telling the child, "Sometimes I have trouble hearing and I can't always understand what people tell me. I try really hard, but sometimes I just don't hear things right. I need help and practice to get better!"

The picture conversation took approximately 5 min and was designed to provide many conversation turns for the child, and thus provide the robot with opportunities to entrain its speech to the child's. The experimenter placed photos one at a time in front of the robot and child (e.g., a collage of holidays or pictures from children's movies). For each picture, the robot introduced the picture content, expressed something it liked about the picture, asked the child a question, responded with generic listening responses (e.g., "Can you tell me more?," "Oh, cool!," "Keep going!"), shared another fact relevant to the picture, and asked another question. At two points during this activity, there were scripted moments where the robot had difficulty hearing (saying, e.g., "I didn't hear that, can you say it again?"), to reinforce its backstory. The experimenter explained that the robot and child had to do at least three pictures, but they could do one more if they wanted—this set up a later compliance/helping task after the third picture, in which the robot asked if the child would do a fourth picture with it to help it practice extra. If the child declined the fourth picture, the experimenter moved on.

The sticker task was used to see how likely the child was to agree to a request by the robot to share a favorite object. The child was allowed to pick out a sticker from a small selection. The robot stated that it wanted the child's sticker and asked for it. The child could spontaneously speak or give their sticker to the robot, or decline. If the child gave their sticker, the experimenter would conveniently find a duplicate sticker in their pocket to replace it, so that the child would not have to forgo their favorite sticker.

The storytelling activity was modeled after the story retelling task used in Kory Westlund et al. (2017b). The robot told a story consisting of a 22-page subset of the wordless picture book "Frog, Where Are you?" by Mercer Mayer. The pages of the book were shown one at a time on the tablet screen. On each page, the robot said 1–2 sentences of the story. Every few pages, the robot asked a dialogic reading comprehension question about the events in the story, e.g., "Where is the deer taking the boy?," 'and "How do you think the boy feels now?" (3 questions total, decreased from the 11 questions in the prior study to decrease the length of the story activity). As in the prior study, the robot responded to children's answers with encouraging, non-committal phrases such as "Mmhm," "Good thought," and "You may be right."

We embedded six target vocabulary words (all nouns) into the story. As in the prior study, we did not test children on their knowledge of these words prior to the storytelling activity because we did not want to prime children to pay attention to

these words, since that could bias our results regarding whether or not children would learn or use the words after hearing them in the context of the robot's story. We used the six key nouns identified in the original story in Kory Westlund et al. (2017b), which were replaced with the target words "gopher"(original word: animal), "crag" (rock),"lilypad" (log), "hollow" (hole), "antlers" (deer), and "cliff" (hill).

After the robot told the story, the robot prompted children to retell the story. Children could use the tablet while retelling the story to go through the story pages, so they could see the pictures to help them remember the story. Twice during the retell, the robot had difficulty hearing ("What? Can you say that again?"), which reinforced the backstory. Children's retellings were used as a measure of their story recall, mirroring of the robot's speech, and expressive use of the vocabulary words.

As part of the closing conversation, we included a goodbye gift task. The experimenter brought out a tray with several objects on it: a small toy frog (because the frog was present in the robot's story), a small book (because the robot expressed great interest in stories), a sticker of the robot's favorite color (blue), and an orange sticker. The child could pick an object to give to the robot, and the experimenter followed up by asking why the child had picked that gift.

After the robot interaction, the experimenter administered a receptive vocabulary test of the six target words in the story. For each word, four pictures taken from the story's illustrations were shown to the child. The child was asked to point to the picture matching the target word. We examined both children's receptive knowledge of the words as well as children's expressive or productive abilities during the story retelling, since children who can recognize a word may or may not be able to produce it themselves.

This was followed by the Inclusion of Other in Self task, adapted for children as described in Kory-Westlund et al. (2018). In this task, children are shown seven pairs of circles that proceed from not overlapping at all to overlapping almost entirely. They are asked to point to the circles showing how close they feel to five different entities: their best friend, their parent, a bad guy they saw in a movie, their pet (or if they have no pet, their favorite toy), and the robot. These five entities were included because we were curious how children might rate the robot compared to other people and things they might feel close to.

Then the experimenter asked several questions taken from the Social Acceptance Scale for Kindergarten Children (Favazza and Odom, 1996; Favazza et al., 2000) regarding how accepting children might be of the robot and its hearing difficulties, as well as of other children who might have hearing difficulties, as described in Kory-Westlund and Breazeal (2019). Finally, children performed a Picture Sorting Task (Kory-Westlund and Breazeal, 2019), in which they were asked to arrange a set of eight entities along a line. The entities included a baby, a frog, a cat, a teddy bear, a computer, a mechanical robot arm, a robot from a movie (e.g., Baymax, WALL-e, or R2D2, depending on which the child was familiar with), and Tega. The line was anchored at one end with a picture of an adult human female and at the other with a picture of a table. We wanted to see where children placed the robot in relation to the other entities.

## 3.6. Materials

We used the Tega robot, a colorful, fluffy squash and stretch robot designed for interactions with young children (Kory Westlund et al., 2016a) (see **Figure 1**). The robot is covered in red fur with blue stripes and uses an Android phone to display an animated face and run control software. The face has blue oval eyes and a white mouth, both of which can change shape to display different facial expressions and mouth movements (visemes) during speech. The robot can move up and down, tilt sideways, rotate from side to side, and lean forward and backward. The experimenters referred to the robot by name (not with pronouns) in a non-gendered way throughout the study.

Speech was recorded by a human adult female and shifted to a higher pitch to sound more child-like. All robot speech was sent through the automated audio entrainment module and streamed to the robot. For the *Entrainment* conditions, all speech was entrained; for the *No Entrainment* conditions, processing still occurred, but the speech simply passed through and was not changed. The reason for this was to incur the same delay (generally a latency of less than 1–2 s) that results from entraining and streaming speech in both conditions. More details regarding entrainment are provided below.

We used a Google Nexus 9 8.9-inch tablet to display the story. Touchscreen tablets have effectively engaged children and social robots in shared tasks (Park et al., 2014), including storytelling activities (Kory and Breazeal, 2014; Kory Westlund et al., 2017b). We used the same custom software on the tablet to display the story pages as in Kory Westlund et al. (2017b), which allowed the teleoperator to turn the pages at appropriate times. This software is open-source and available online under the MIT License at https://github.com/mitmedialab/SAR-opal-base/.

## 3.7. Teleoperation

As in the prior study (Kory Westlund et al., 2017b), we used custom teleoperation software to control the robot and digital storybook. The teleoperation software is open-source and available online under the MIT License at https://github.com/mitmedialab/tega_teleop/. The experimenters were all trained to control the robot by an expert teleoperator.

Using teleoperation allowed the robot to appear autonomous while removing technical barriers, primarily natural language understanding, since the teleoperator could be in the loop to parse language. The teleoperator triggered when the robot began each sequence of actions (speech, physical motions, and gaze), and when the storybook should turn the page. Thus, the teleoperator had to attend to timing in order to trigger action sequences at the right times. The timing of actions within sequences was automatic and thus consistent across children. There were also several occasions when the teleoperator had to listen to children's speech and choose the most appropriate of a small set of different action sequence options to trigger, namely during the picture conversation task.

The teleoperator performed one of two actions if the child asked an unexpected question or said something unusual. During the conversation portion of the interaction, the teleoperator could trigger one of the generic responses (e.g., "Mmhm!," "Hm, I don't know!") in reply. During the remainder of the

interaction, the teleoperator had to continue in accordance with the interaction script, which essentially ignored unexpected behaviors. While this is not ideal from an interaction standpoint, it was necessary to ensure reasonably consistent behavior on the part of the robot across children.

## 3.8. Entrainment

In the *Entrainment* condition, the speaking rate and pitch of the robot's voice were automatically adjusted to be more similar to the child. In addition, the robot's volume and exuberance were manually adapted by the teleoperator.

For speaking rate and pitch entrainment, the child's speech was automatically collected via the robot's microphone when it was the child's turn to speak in the conversation. Using automatic software scripts with Praat (audio analysis software), various features of the children's speech were extracted and used to modify the robot's recorded speech files. These modified audio files were then streamed to the robot for playback.

For speaking rate, the robot's speech was sped up or slowed down to match the child's speaking rate. Thus, if a child spoke slowly, the robot slowed down its speech as well. We included ceiling and floor values such that the robot's speech would only ever be sped up or slowed down by a maximum amount, ensuring that the speech stayed within a reasonable set of speeds. We used the Praat script for speaking rate detection from de Jong and Wempe (2009). The code for our entrainment module is open-source and available online under a GNU General Public License v3.0 at https://github.com/mitmedialab/rr_audio_entrainer/.

The mean pitch of the robot's speech was shifted up or down. In doing this, the robot matches two features: (1) the child's age, (2) the child's current mean pitch. In general, people speak at a particular fundamental frequency, but there is variation within an individual (pitch sigma). Thus, we provided a table of mean fundamental frequencies for different age children based on the values computed in prior work (Weinberg and Zlatin, 1970; Bennett, 1983; Sorenson, 1989; Hacki and Heitmüller, 1999; Baker et al., 2008; Gelfer and Denor, 2014). For a given child, all of the robot's speech was first shifted to have the mean pitch for children of that age. Then, since an individual may vary their pitch in each utterance, the pitch of each utterance was also shifted up or down slightly based on whether the child's most recent utterance was higher or lower. Unlike Lubold and colleagues (Lubold et al., 2016, 2018), we did not adapt the pitch contour of the robot's speech. Because the base sounds for the robot's speech were recorded by a human (not flat text-to-speech as in Lubold et al.'s work), the sounds had their own pitch contours. Pilot tests showed that morphing or replacing this contour led to speech that sounded unnatural (e.g., placing emphasis on the wrong syllables).

We also manually adapted the robot's volume and exuberance. During the introduction and first picture in the picture task, the teleoperator observed the child's behavior and personality: were they shy, passive, reserved, or quiet (less exuberant/quiet children)? Or were they loud, extroverted, active, smiley, or expressive (more exuberant/loud children)? Based on this binary division, the teleoperator adjusted the robot's audio playback volume twice, at two specific points during the interaction, to either be slightly quieter (for less exuberant/quiet children) or slightly louder (for more exuberant/louder children). Furthermore, the teleoperator triggered different animations to be played on the robot at six different points during the interaction—more excited and bigger animations for more exuberant/louder children; quieter, slower, animations for less exuberant/quieter children.

## 3.9. Data

We recorded audio and video of each interaction session using a camera set up on a tripod behind the robot, facing the child. All audio was transcribed by human transcriptionists for later language analyses. Children's responses to the posttest assessments were recorded on paper and later transferred to a spreadsheet.

## 3.10. Data Analysis

For the analysis of children's story retellings, we excluded the three 3-year-olds because one did not retell the story, and the other two needed extra prompting by the experimenter and were very brief in their responses. Of the remaining 83 children, one child's transcript could not be obtained due to missing audio data. Fifteen children did not retell the story (the number from each condition who did not retell the story was not significantly different). Thus, in total, we obtained story retell transcripts for 67 children (15 *E-B*; 9 *E-NB*; 22 *NE-B*; 21 *NE-NB*).

We analyzed children's transcribed story retells in terms of story length (word count), overall word usage, usage of target vocabulary words, and similarity of each child's story to the robot's original story. We created an automatic tool to obtain similarity scores for each child's story as compared to the robot's story, using a phrase and word matching algorithm. The algorithm proceeded as follows: First, take both stories (the original story and the child's story) and remove stopwords (i.e., words with no significant information such as "the," "uh," and "an"). Second, stem words—i.e., convert words to their original form. For example, "jumping" would be converted to "jump." Third, find all N-grams in each story, where an N-gram is a continuous sequence of N words from both texts. Fourth, remove duplicate N-grams from one text. Fifth, count how many N-grams are the same in both texts. The number of matches is the similarity score. This algorithm produces a score reflecting the number of exact matching phrases in both stories—i.e., words used in the same order by both the child and robot. It also produces a higher match score for texts that have both more matching phrases and longer matching phrases. We also implemented an algorithm for counting similar matches that are close to each other, but not exactly the same. This algorithm was the same as the above, where the fifth step (counting matching N-grams) used a fuzzy string matching algorithm to determine if the N-grams matched.

When running the algorithm to match stories, we used $N = 3$ for computing exact match scores because a smaller $N$ may not retain enough information to be considered actual phrase matching, while a larger $N$ may encompass more information than would constitute a single phrase. For determining similar match scores, we used $N = 4$, so that when phrases differed by

one word, or used a different word in the middle of a similar phrase, they might still match, as would be expected for similar phrases. We combined the exact and similar match scores to get a single overall similarity score for each child's story that reflected the child's overall use of exact and similar matching phrases.

For example, the robot's story included the sentences, "The baby frog liked the boy and wanted to be his new pet. The boy and the dog were happy to have a new pet frog to take home." After stopword removal and stemming, this was converted to: "baby frog like boy want be new pet boy dog happy new pet frog take home." One child's story included the similar section, "Then he hopped on his hand and he wanted to be his pet. And then the dog and the boy was happy to have a new pet," which was converted to: "hop hand want be pet dog boy happy new pet." There were several exactly matching phrases, e.g., *"happy new pet."* There were also several similar matching phrases, e.g., (robot) *"be pet boy dog"*/(child) *"be pet dog boy."*

We obtained children's facial expressions from the recorded videos using Affdex, emotion measurement software from Affectiva, Inc., Boston, MA, USA (McDuff et al., 2016). Affdex can detect 15 facial expressions, which are used to detect whether the face is displaying nine different affective states. Affdex only recognizes outward expressions of affect (i.e., facial configuration patterns), which does not imply detecting any underlying feelings or inferring deep internal states (though they are believed to be correlated). For each frame of a video, Affdex attempts to detect a face. If a face is detected, Affdex scores each affective state as well as the presence of each expression in the range 0 (no expression/affective state detected) to 100 (expression or state fully present); middle values represent an expression or state that is partially present. However, these values are relative and Affdex does not specify what the exact difference between scores means. For more detail on the algorithms used for facial affect classification, see Senechal et al. (2015). We analyzed affect data for 74 children (16 *E-B*; 11 *E-NB*; 26 *NE-B*; 21 *NE-NB*). For the remaining 12 children, little or no affect data were collected as a result of system failures, such as children's faces not being recognized by Affdex.

We focused our analysis on the following affective states and facial expressions: joy, fear, sadness, surprise, concentration, disappointment, relaxation, engagement, valence, attention, laughter, and smiles. We included valence in addition to specific emotions such as joy because Affdex uses different sets of facial expressions to detect the likelihood that a face is showing each affective state. Thus, valence is not detected from, e.g., the emotions joy or sadness; instead, it is calculated from a set of facial expressions that is somewhat different than, though overlapping with, the set of expressions used to calculate other emotions. The expression "concentration" was called "contempt" by Affectiva. Affectiva has no label for concentration or thinking expressions. Affectiva uses brow furrows and smirks to classify contempt; prior work has found that brow furrowing and various lip movements present in smirks such as mouth dimpling and lip tightens are also associated with concentration (Oster, 1978; Rozin and Cohen, 2003; Littlewort et al., 2011). Furthermore, contempt is generally defined as "the feeling that a person or thing is worthless or beneath consideration," which, as in Kory Westlund et al. (2017b), did not make sense in this context; children's expressions were more indicative of concentration.

We coded children's responses to the Social Acceptance Scale questions on a 3-point scale, with *"no"* as 0, *"maybe"* as 1, and *"yes"* as 2. We labeled children's placement of the entities in the Picture Sorting Task, with the anchor on one end (the human) at position 1 and the anchor at the other (the table) at position 10. Thus, a lower rank indicated that children placed the entity closer to the adult woman. We counted positions to determine what rank was held by each picture. We also computed scores for Tega's rank relative to the other entities. For example, we subtracted the human baby's rank from Tega's rank to get Tega's rank relative to the human baby and human adult. Because Tega's position among the entities was dependent on where children placed the other entities in the task, we examined where children placed all the different entities.

We coded whether children agreed to do the fourth picture and whether they gave the robot their sticker with *"no"* as 0 and *"yes"* as 1. We coded children's selections in the goodbye gift task as follows: *frog* as 4, *book* as 3, *blue sticker* as 2, and *orange sticker* as 1. We also coded the comments children made regarding why they selected a particular gift with the following rubric: 2 if they referenced the robot or the robot's feelings (e.g., "Tega would like it because frog jumped out in story," "Tega likes books," "Because he wanted a sticker"); 1 for a somewhat relevant comment, mentioning the interaction (e.g., "It was in the story"); 0 for no explanation, reference to themselves, or an irrelevant comment (e.g., "It is swamp week at camp," "I don't know").

# 4. RESULTS

Our results are divided below into two parts, each reflecting one of our hypothesis areas: (1) *Learning*: We asked whether the robot's entrainment and backstory would increase children's learning with the robot and emulation of the robot's story; and (2) *Rapport, relationship, and social behavior:* We asked whether children would show greater rapport, acceptance, positive emotion, engagement, and closeness to the robot as a result of its entrainment and backstory.

## 4.1. Learning (H1, H2, H3)

For all learning-related analyses of variance, we included Age as a covariate because we expected that children's age would be related to their language ability and thus to their vocabulary scores and the complexity and/or length of their stories.

### 4.1.1. Target Vocabulary Word Identification (H1)

We performed 2×2 between-subjects analyses of variance with Entrainment (*E* vs. *NE*) and Backstory (B vs. NB) with Age as a covariate. We found a significant effect of Age on the total vocabulary words identified correctly, $F_{(5,\ 77)} = 2.76$, $p = 0.024$, $\eta_p^2 = 0.15$. Eight-year-olds correctly identified the most words, while 3-year-olds correctly identified the least (**Figure 2A**). We also found a significant effect of Entrainment on children's identification of the target words, $F_{(1,\ 77)} = 5.47$, $p = 0.022$, $\eta_p^2$

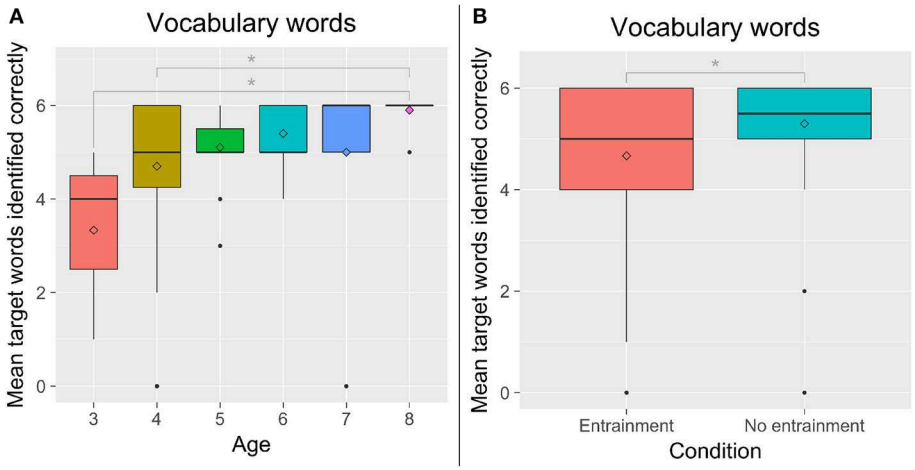



**FIGURE 2 | (A)** The number of words correctly identified by children of each age group. **(B)** The number of words correctly identified by entrainment condition. $^{*}p < 0.05$.

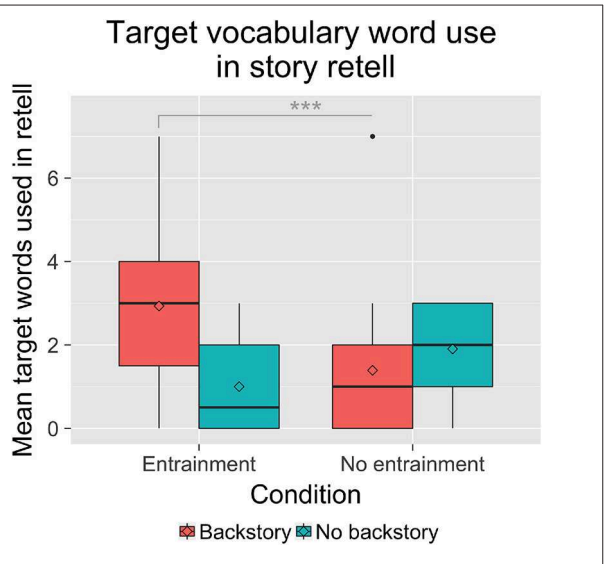



**FIGURE 3 |** Children in the *E,B* condition used more target words in their story retells than children in the other conditions. $^{***}p < 0.001$.

= 0.07. Contrary to our hypotheses, children in the *NE* condition correctly identified more words than children in the *E* condition; however, in both conditions, there appeared to be a ceiling effect (**Figure 2B**). Older children were more likely to correctly identify words than younger children, $r_{s(85)} = 0.367$, $p < 0.001$.

### 4.1.2. Target Vocabulary Word Use (H2, H3)

A 2×2 between-subjects analyses of variance with Entrainment (*E* vs. *NE*) and Backstory (B vs. NB) with Age as a covariate revealed a significant interaction between Entrainment and Backstory regarding children's use of the target vocabulary words in the story, $F_{(1,59)} = 9.45$, $p = 0.003$, $\eta_p^2 = 0.14$. Children in the *E,B* condition used significantly more of the target words than children in all three other conditions (**Figure 3**).

Overall, we saw no correlation between children's recognition of words on the vocabulary test and their subsequent use of those words in their retells, $r_{s(67)} = 0.047$. However, there were trends showing that this did vary by condition, though none of the correlations were significant. If the robot entrained, children were more likely to use the words themselves if they had identified the words correct on the test, *E-B* $r_{s(15)} = 0.253$; *E-NB* $r_{s(10)} = 0.254$; children who did not receive entrainment were less likely to do so, *NE-B* $r_{s(23)} = -0.077$; *NE-NB* $r_{s(21)} = 0.024$.

In summary, given that children's scores on the vocabulary identification test were not significantly different by condition, these results suggest that the robot's entrainment and backstory did not impact children's initial encoding of the words, but did affect children's expressive use of the words in their retelling.

### 4.1.3. Story Length (H3)

The robot's story was 435 words long, including the dialogic questions. The mean length of children's retells was 304 words ($SD = 110.9$). After stopword removal, the robot's story was 185 words, of which 99 were unique, non-overlapping words. The mean length of children's stories after stopword removal was 113 ($SD = 41.7$), with a mean of 63.1 unique words ($SD = 19.0$).

We performed 2×2 between-subjects analyses of variance with Entrainment (*E* vs. *NE*) and Backstory (*B* vs. *NB*) with Age as a covariate, which revealed a significant effect of Age on the length of children's stories after stopword removal, $F_{(4, 59)} = 3.77$, $p = 0.008$, $\eta_p^2 = 0.20$, and on the number of unique words children used, $F_{(4, 59)} = 3.19$, $p = 0.019$, $\eta_p^2 = 0.17$. *Post-hoc* tests revealed that 6- and 7-year-old children told longer stories than 4-year-old children, and 7-year-old children used more unique words than 4-year-old children (**Figures 4A,B**). The length of children's stories before stopword removal followed the same pattern, but was not statistically significant. This suggests that the primary difference between older (6–7 years) and younger (4–5 years) children's stories was their use of significant content words vs. stopwords.

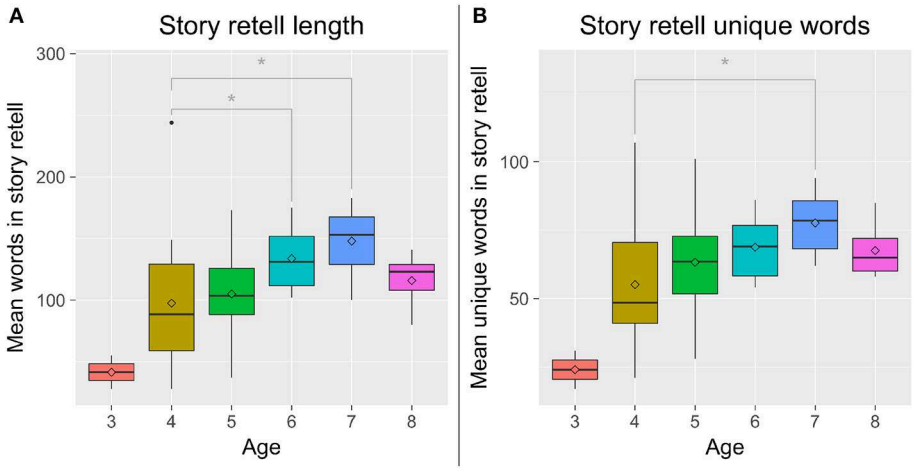



**FIGURE 4 | (A)** Older children told longer stories than younger children. **(B)** Older children used more unique words than younger children. *$p < 0.05$.

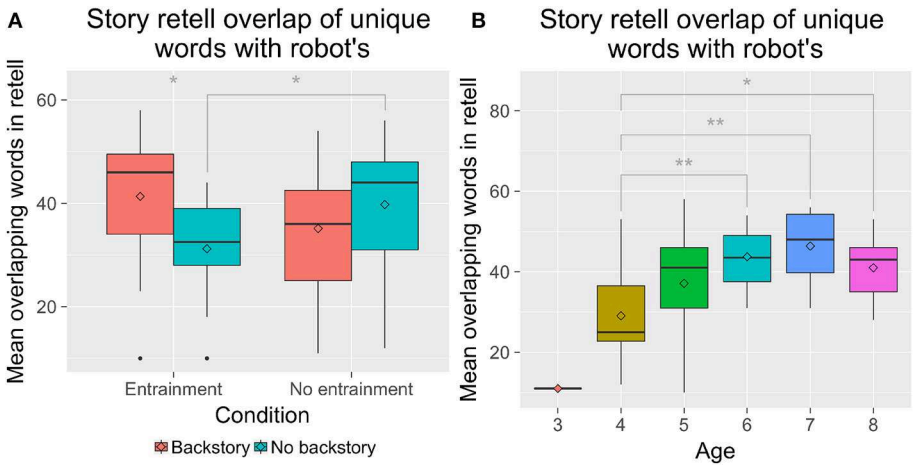



**FIGURE 5 |** The number of overlapping words children used by entrainment condition **(A)** and by age **(B)**. *$p < 0.05$; **$p < 0.01$.

#### 4.1.4. Mirroring the Robot's Story (H2, H3)

Children used a mean of 37.7 unique words ($SD = 12.3$) in their retells of the 99 unique words that the robot had used in its story. A 2×2 between-subjects analyses of variance with Entrainment (*E* vs. *NE*) and Backstory (*B* vs. *NB*) with Age as a covariate revealed that the number of overlapping unique words used was significantly different by Age, $F_{(4, 60)} = 6.12$, $p < 0.001$, $\eta_p^2 = 0.29$. We also observed a significant interaction of Entrainment with Backstory, $F_{(1, 60)} = 6.42$, $p = 0.013$, $\eta_p^2 = 0.10$. *Post-hoc* tests showed that older children overlapped more than younger children (**Figure 5A**). Children in the *E-NB* condition ($M = 31.2$, $SD = 10.9$) overlapped less than children in the *E-B* and *NE-NB* conditions (*E-B:* $M = 41.3$, $SD = 13.2$; *NE-B:* $M = 36.2$, $SD = 10.6$; *NE-NB:* $M = 39.8$, $SD = 13.3$) (**Figure 5B**).

Children's stories received mean scores of 41.3 ($SD = 36.2$) for their use of exact and similar phrases that mirrored the robot's phrases. However, we observed no significant differences between conditions in children's use of exact and similar matching phrases.

### 4.2. Rapport, Relationship, and Social Behavior (H4, H5, H6, H7)

#### 4.2.1. Acceptance of the Robot (H4)

We performed 2×2 between-subjects analyses of variance with Entrainment (*E* vs. *NE*) and Backstory (*B* vs. *NB*) for the questions asked about children's social acceptance of the robot and of other children. We found a significant main effect of Backstory of children's responses to the question "Would you like to be good friends with a robot who can't hear well," $F_{(1, 82)} = 7.55$, $p = 0.007$, $\eta_p^2 = 0.08$. Children who heard the robot's backstory were more likely to respond positively than children who did not hear the robot's backstory. Children who heard the backstory were also somewhat more likely to respond positively to the question, "Would you like to be good friends with a

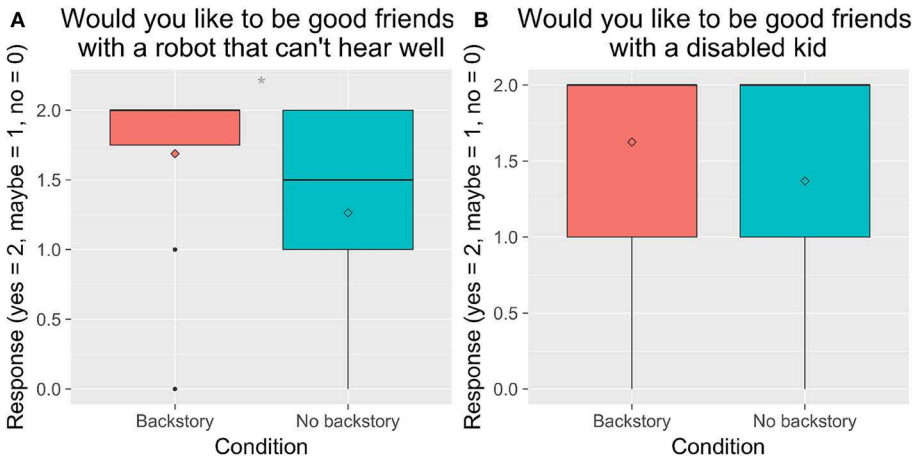



**FIGURE 6 |** Children's responses to the question, "Would you like to be good friends with a robot who can't hear well?" and the question, "Would you like to be good friends with a handicapped or disabled kid?" by condition. *$p < 0.05$.

**TABLE 2 |** Analysis of facial expressions during the interaction by condition.

| Expression | Overall | E-B | E-NB | NE-B | NE-NB |
|---|---|---|---|---|---|
| Engagement | 30.8 (11.7) | 33.3 (13.3) | 30.5 (12.0) | 29.6 (11.2) | 30.5 (11.4) |
| Attention | 68.9 (13.4) | 62.2 (21.1) | 67.8 (15.2) | 71.9 (5.56) | 72.0 (9.51) |
| Valence | −0.738 (9.11) | 3.51 (8.81) | 5.75 (13.72) | −4.13 (5.20) | −2.72 (8.47) |
| Joy | 7.13 (8.04) | 9.13 (8.81) | 12.1 (12.5) | 5.48 (5.02) | 5.61 (7.26) |
| Smiles | 8.98 (8.82) | 10.9 (9.35) | 14.6 (13.4) | 7.16 (5.65) | 7.52 (8.31) |
| Laughter | 0.13 (0.22) | 0.23 (0.31) | 0.28 (0.36) | 0.08 (0.09) | 0.07 (0.11) |
| Relaxation | 3.53 (5.31) | 4.13 (5.38) | 6.63 (9.61) | 2.49 (2.42) | 3.06 (5.03) |
| Surprise | 7.21 (6.96) | 8.47 (9.22) | 4.53 (4.63) | 7.40 (5.32) | 7.43 (7.84) |
| Disappointment | 4.98 (3.98) | 2.58 (2.01) | 3.58 (3.03) | 6.58 (4.37) | 5.72 (4.05) |
| Fear | 1.48 (2.06) | 1.00 (1.40) | 0.38 (0.66) | 1.87 (2.04) | 1.93 (2.72) |
| Concentration | 2.92 (2.48) | 2.02 (1.79) | 2.11 (1.87) | 3.20 (2.45) | 3.72 (3.03) |
| Sadness | 0.27 (0.46) | 0.22 (0.34) | 0.49 (0.54) | 0.32 (0.59) | 0.17 (0.24) |

*Values can range from 0 (no expression present) to 100 (expression fully present), except Valence, which can range from −100 to 100. Each column lists mean and standard deviation.*

handicapped or disabled kid," though it was not statistically significant (**Figure 6**).

### 4.2.2. Children's Expressivity and Positive Emotion (H5, H6)

Overall, children were highly attentive and engaged, and displayed surprise and other emotions during the story (see **Table 2**). To evaluate whether children showed greater engagement or positive emotion with the robot that entrained, we performed 2×2 between-subjects analyses of variance with Entrainment (*E* vs. *NE*) and Backstory (*B* vs. *NB*).

We found a significant main effect of Entrainment on children's expressions of joy, $F_{(1, 69)} = 6.25$, $p = 0.015$, $\eta_p^2 = 0.070$; fear, $F_{(1, 69)} = 5.31$, $p = 0.024$, $\eta_p^2 = 0.074$; concentration, $F_{(1, 69)} = 5.09$, $p = 0.027$, $\eta_p^2 = 0.074$; disappointment, $F_{(1, 69)} = 12.7$, $p < 0.001$, $\eta_p^2 = 0.17$; attention, $F_{(1, 69)} = 5.66$, $p = 0.02$, $\eta_p^2 = 0.091$; laughter, $F_{(1, 69)} = 12.02$, $p < 0.001$, $\eta_p^2 = 0.13$; smiles, $F_{(1, 69)} = 5.82$, $p = 0.019$, $\eta_p^2 = 0.064$; and valence, $F_{(1, 69)} = 14.7$, $p = < 0.001$, $\eta_p^2 = 0.16$. *Post-hoc* tests showed that children expressed less fear, concentration, disappointment, and attention in the *E* condition than in the *NE* condition (**Figure 7**). Children showed higher mean joy, laughter, valence (i.e., showed more affect with a positive valence), and more smiles in the *E* condition than in the *NE* condition (**Figure 8**). There were no significant differences in sadness, surprise, relaxation, or engagement; however, there was a trend for children in the *E* condition to show more relaxation than in the *NE* condition, which could have contributed to the higher valence seen in the *E* condition.

Next, we asked whether children's affect changed during the session. We split the affect data into the first half of the session and the second half of the session, using the data timestamps to determine the halfway point. We ran a 2×2×2 mixed ANOVA with time (within: first half vs. second half) × Entrainment (between: *E* vs. *NE*) × Backstory (between: *B* vs. *NB*). Although we hypothesized several changes in children's affect over time as a result of condition, we corrected for multiple comparisons here and only considered results significant when $p < 0.004$.

Like before, we found a significant main effect of Entrainment on disappointment, $F_{(1, 70)} = 14.7$, $p < 0.001$; laughter, $F_{(1, 70)} = 8.94$, $p = 0.004$; and valence, $F_{(1, 70)} = 14.6$, $p < 0.001$. There were trends for a main effect of Entrainment on joy, $F_{(1, 70)} = 4.25$, $p = 0.043$; fear, $F_{(1, 70)} = 5.88$, $p = 0.018$; attention, $F_{(1, 70)} = 4.37$, $p = 0.040$; and smiles, $F_{(1, 70)} = 3.99$, $p = 0.0497$. Children showed fewer expressions of fear and disappointment in the *E* than in the *NE* condition (**Figure 9**). Children showed more joy, more smiles, and higher valence in the *E* than the *NE* condition.

We found a significant main effect of time on joy, $F_{(1, 67)} = 34.6$, $p < 0.001$; valence, $F_{(1, 67)} = 17.7$, $p < 0.001$; engagement, $F_{(1, 67)} = 10.3$, $p = 0.002$; smiles, $F_{(1, 67)} = 40.5$, $p < 0.001$; relaxation, $F_{(1, 67)} = 27.2$, $p < 0.001$; laughter, $F_{(1, 67)} = 11.9$, $p = 0.001$. All of these decreased from the first half to the second half of the session.

We saw trends for interactions of Entrainment with time: concentration, $F_{(1, 67)} = 6.79$, $p = 0.011$; attention, $F_{(1, 67)} =$

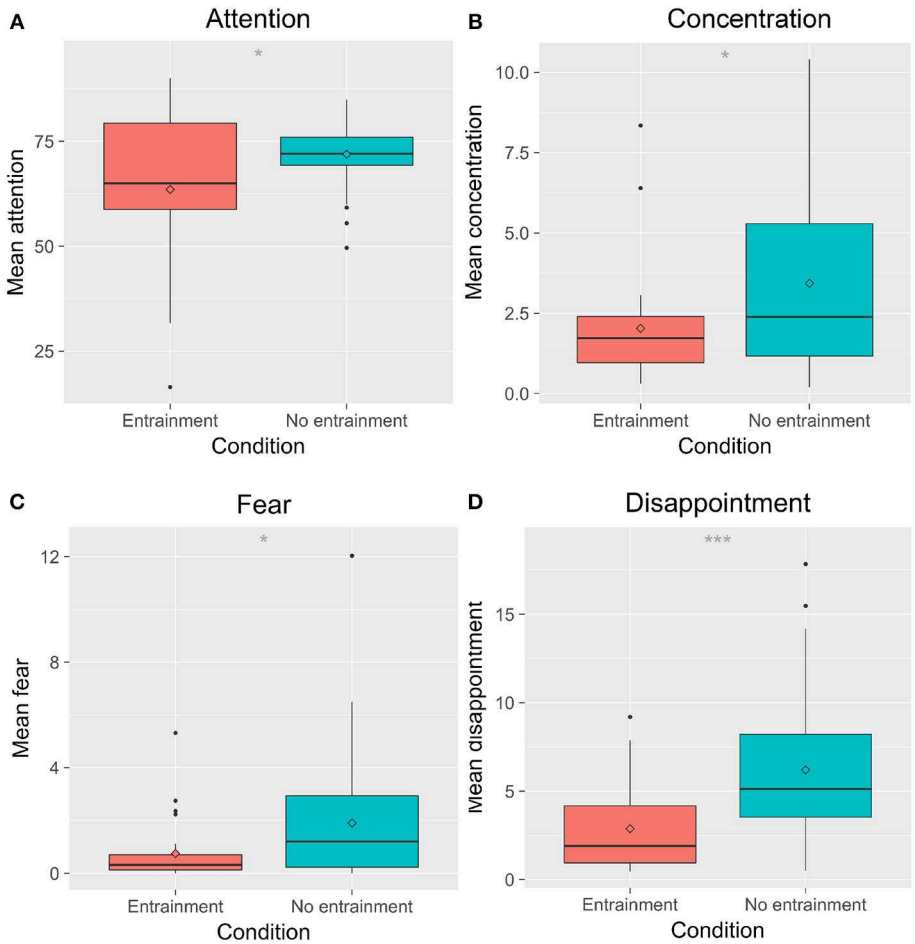



**FIGURE 7 |** Children's overall negative affect varied by entrainment condition. **(A)** shows attention; **(B)** shows concentration; **(C)** shows fear; **(D)** shows disappointment. $^{*}p < 0.05$; $^{***}p < 0.001$.

5.47, $p = 0.022$; and laughter, $F_{(1, 67)} = 7.82$, $p = 0.007$. Children showed more concentration during the first half in the *NE* than in the *E* condition. Children showed more attention during the first half for *NE* vs. *E,* but they did not differ during the second half. Children laughed more in the first half in the *E* condition than in the *NE* condition, and decreased to the second half, while in the *NE* condition the amount of laughter did not change over time.

We also saw trends for interactions of time with Backstory for fear, $F_{(1, 67)} = 8.55$, $p = 0.005$; sadness, $F_{(1, 67)} = 7.01$, $p = 0.010$; disappointment, $F_{(1, 67)} = 7.70$, $p = 0.007$; attention, $F_{(1, 67)} = 4.88$, $p = 0.031$; and valence, $F_{(1, 67)} = 8.12$, $p = 0.006$ (**Figure 10**). Children expressed less fear in the second half of the session when they did not hear the backstory, but expressed somewhat more fear in the second half if they had heard the backstory. They expressed less sadness in the second half in *NB* condition, but did not change in *B* condition. Children's expressions of disappointment increased slightly in the *B* condition from first to second half, but not for the *NB* condition. Children's attention was higher initially in the *NB* condition and decreased slightly, while children's attention started lower in the *B* condition and increased slightly. Children showed decreased valence in the *B* condition from first half to second half, but not in the *NB* condition.

### 4.2.3. Closeness to the Robot (H5, H6)

We performed a 2×2×5 mixed ANOVA with Entrainment (*E* vs. *NE*) × Backstory (*B* vs. *NB*) × IOS agent (within: Friend, Parent, Tega, Pet/Toy, Bad guy). We found a significant effect of agent, $F_{(4, 302)} = 61.9$, $p < 0.001$. *Post-hoc* Tukey's HSD tests showed that the bad guy was rated significantly lower than all other agents. In addition, the robot was rated significantly lower than the friend, but was not significantly different from the parent or pet/toy (**Figure 11A**). Older children were more likely to rate Tega as closer, $r_{s(86)} = 0.410$, $p < 0.001$ (**Figure 13A**).

Regarding the Picture Sorting Task, overall, Tega was placed at a mean position of 4.78 ($SD = 1.80$) (**Figure 11B**). **Figure 12A** shows results by condition for Tega's distance to the human, and **Figure 12B** shows the relative distance of each entity from the Tega robot by condition.

We performed a mixed ANOVA with Entrainment (between: *E* vs. *NE*) × Backstory (between: *B* vs. *NB*) × Entity (within:

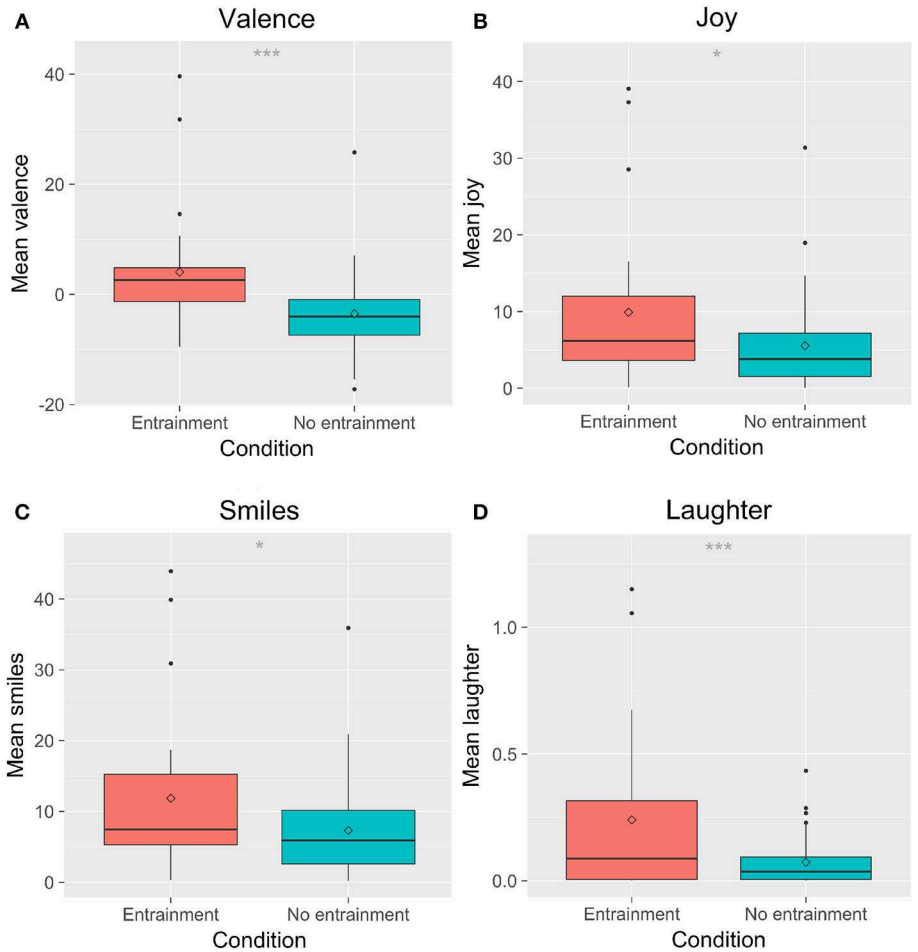



**FIGURE 8 |** Children's overall postive affect varied by entrainment condition. **(A)** shows valence; **(B)** shows joy; **(C)** shows smiles; **(D)** shows laughter. $^{*}p < 0.05$; $^{***}p < 0.001$.

Tega robot, baby, cat, frog, teddy bear, movie robot, robot arm, computer) for the entity positions, as well as for the entity positions relative to the Tega robot. For entity positions, we observed a significant main effect of Entity, $F_{(7,\ 574)} = 71.7$, $p < 0.001$. We also observed a significant interaction of Entity with Entrainment, $F_{(7,\ 574)} = 2.15$, $p = 0.037$; and a significant interaction of Entity with Backstory, $F_{(7,\ 574)} = 2.35$, $p = 0.022$.

*Post-hoc* tests revealed that the baby was placed significantly closer to the human adult than all other entities. The cat was placed significantly closer to the human adult than all entities except for the Tega robot in the *E* condition, and closer to the human than all entities except Tega and the frog in the *NB* condition. In both the *NE* and *B* conditions, the cat was not placed significantly differently from Tega, the frog, movie robot, or teddy bear.

In the *E* condition, the Tega robot was significantly closer to the human adult than the robot arm, computer, movie robot, and teddy bear. It was farther from the human adult than the baby and was not placed in a significantly different position from the cat or frog. In the *NE* condition, Tega was only placed significantly closer to the human adult than the robot arm and computer; it was not placed significantly differently from the cat, frog, movie robot, or teddy bear. Tega was not placed in a significantly different position from the movie robot in the *B* condition, but was placed significantly farther from it (closer to the human) in the *NB* condition.

The frog was placed significantly closer to the human adult than the robot arm and computer, and significantly farther from the human adult than the baby, but otherwise its position did not differ significantly from any other entities, except in the *NB* condition, where it was placed closer than the movie robot.

In the *NE* condition, the robot arm was placed closer to the table than the frog and movie robot, but in the *E* condition, the robot arm was not placed significantly differently from the frog or movie robot. By Backstory, children in the *B* condition placed the robot arm closer to the table than all other entities except the computer and teddy bear, while in the *NB* condition the robot arm's position was also not signficantly different from the movie robot's. Finally, in the *NE* and *B* conditions, the computer was placed closer to the table than all entities except the robot arm, while in the *E* and *NB* conditions, the computer was also not significantly different from the movie robot.

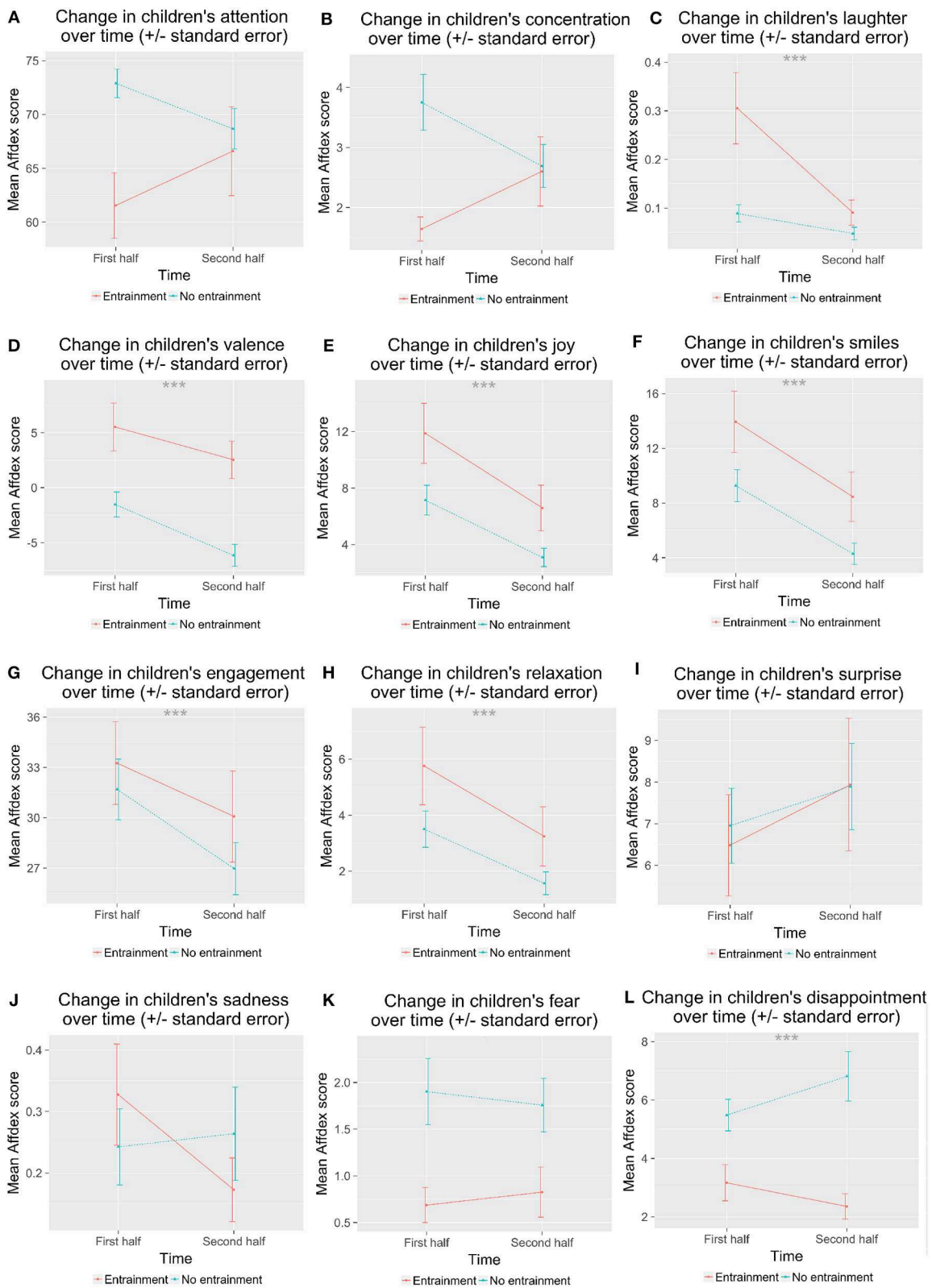



**FIGURE 9 |** Children's affect during the first half and the second half of the interaction varied by entrainment condition. **(A)** shows attention; **(B)** shows concentration; **(C)** shows laughter; **(D)** shows valence; **(E)** shows joy; **(F)** shows smiles; **(G)** shows engagement; **(H)** shows relaxation; **(I)** shows surprise; **(J)** shows sadness; **(K)** shows fear; **(L)** shows disappointment. ***$p < 0.001$.

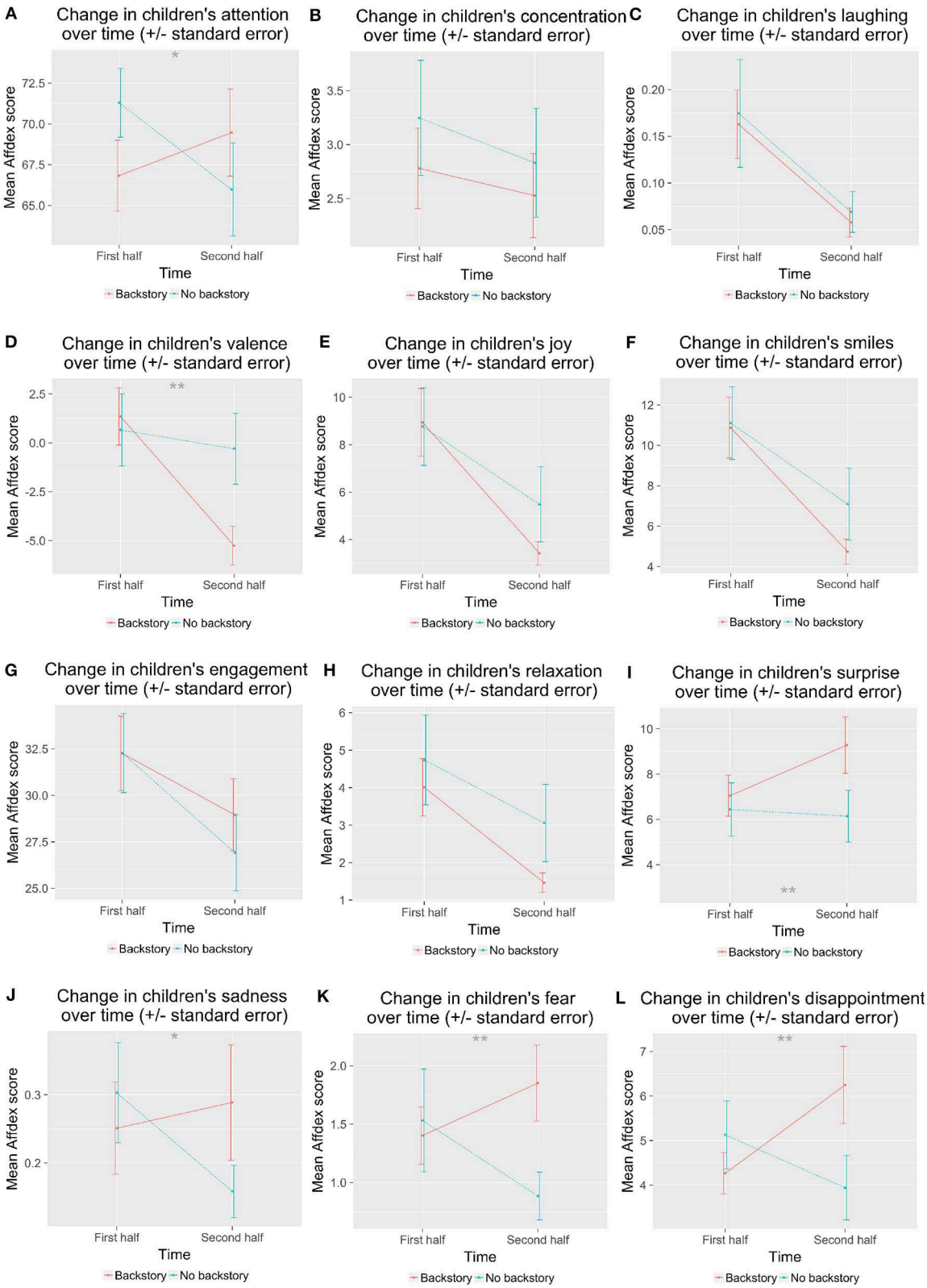



**FIGURE 10 |** Children's affect during the first half and the second half of the interaction varied by backstory. **(A)** shows attention; **(B)** shows concentration; **(C)** shows laughter; **(D)** shows valence; **(E)** shows joy; **(F)** shows smiles; **(G)** shows engagement; **(H)** shows relaxation; **(I)** shows surprise; **(J)** shows sadness; **(K)** shows fear; **(L)** shows disappointment. $^{*}p < 0.05$; $^{**}p < 0.01$.

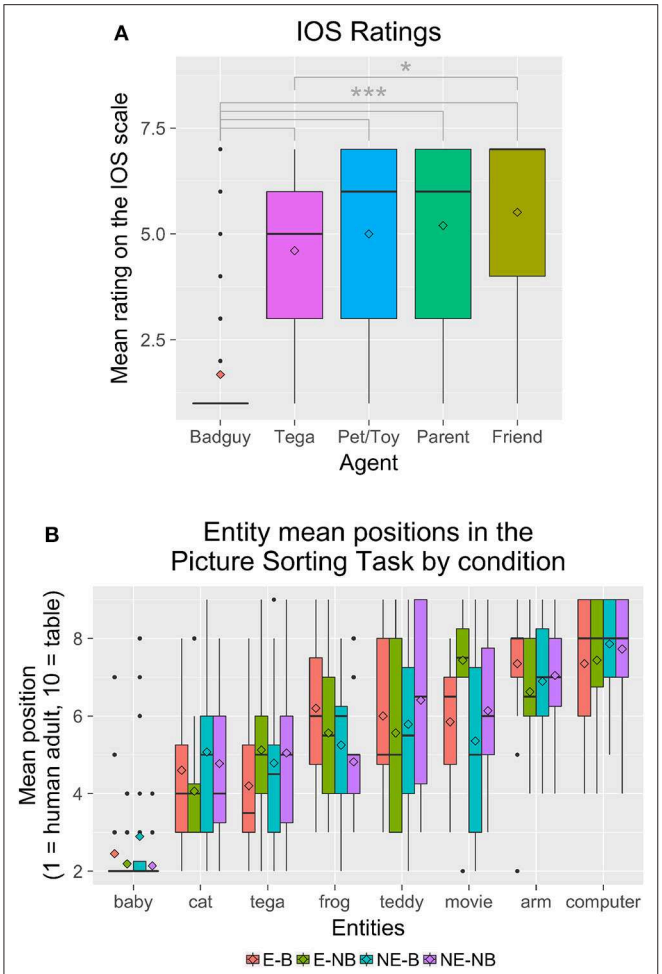



**FIGURE 11 | (A)** Children's IOS ratings for each agent. **(B)** The mean position where children placed each entity in the Picture Sorting Task by condition. $^{*}p < 0.05$; $^{***}p < 0.001$.

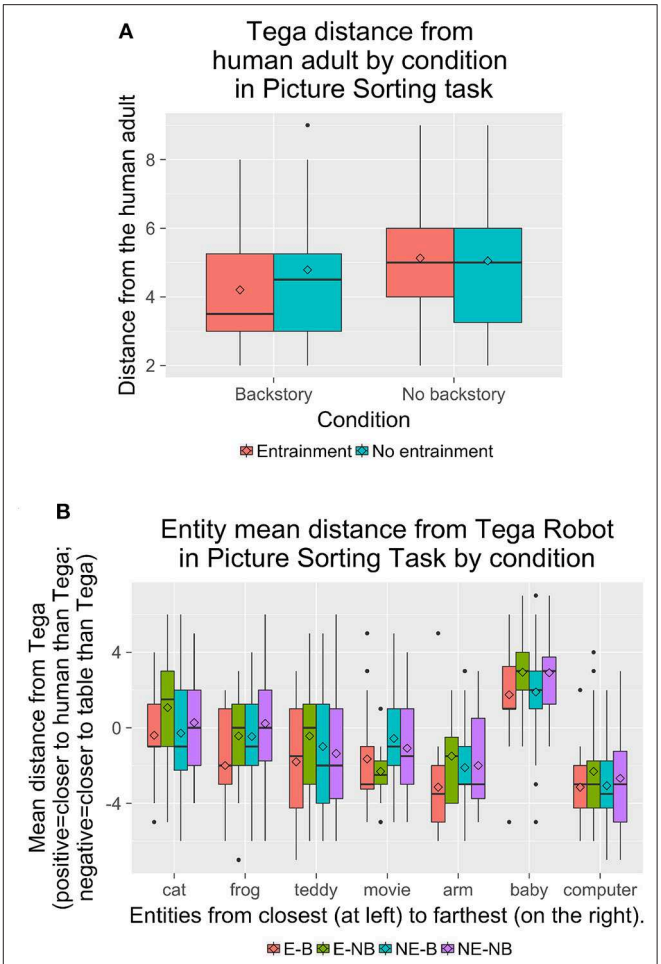



**FIGURE 12 | (A)** Tega's mean distance from the human adult in the Picture Sorting Task by condition. **(B)** The distance of each entity from the Tega robot in the Picture Sorting Task by condition. There were trends for the Tega robot to be placed closer to the baby in the *B* condition than in the *NB* condition, closer to the movie robot in the *E* condition than in the *NE* condition, and closer to the frog in the *E-B* condition than in the other conditions.

Regarding the distance of each entity relative to the Tega robot, we observed a significant main effect of Entity, $F_{(6, 492)} = 71.8$, $p < 0.001$. We also observed a significant interaction of Entity with Entrainment, $F_{(6, 492)} = 2.13$, $p = 0.049$; and a trend toward an interaction of Entity with Backstory, $F_{(6, 492)} = 2.11$, $p = 0.051$. *Post-hoc* tests revealed that the baby was placed farther from Tega, and closer to the human adult than Tega was, than all other entities. There was a trend for children to place the Tega robot closer to the baby (and the baby closer to the human adult than Tega) in the *B* condition (mean difference = 1.83, $SD = 2.55$) than in the *NB* condition ($M = 2.92$, $SD = 2.01$).

The cat was placed closer to Tega than most other entities. It was not placed significantly differently than the teddy bear in the *E* condition; from the frog, movie robot, or teddy bear in the *NE* and *B* conditions; and from the frog in the *NB* condition.

The computer was placed farther from Tega than all entities except the robot arm and, in the *E* and *NB* conditions, the movie robot. The robot arm, in turn, was placed farther from Tega than all entities except the computer and teddy bear. In the *NB* and *NE* conditions, the robot arm was also not different than the movie robot; and in the *E* condition, the robot arm was also not different from the movie robot or frog. There was a trend for children to place Tega farther from the movie robot, and closer to the human than the movie robot was, in the *E* condition ($M = -1.94$, $SD = 2.40$) than in the *NE* condition ($M = -0.80$, $SD = 2.69$).

Finally, we also observed trends for Tega to be placed farther from the frog, and also closer to the human adult than the frog was, in the *E* (*E*: $M = -1.31$, $SD = 2.77$, *NE*: $M = -0.16$, $SD = 2.62$) and *B* conditions (*B*: $M = -1.11$, $SD = 2.76$, *NB*: $M = -0.05$, $SD = 2.60$).

We observed no significant differences between conditions regarding whether children were more likely to agree to do the fourth picture with the robot, give the robot their sticker in the sticker task, or give the robot a bigger goodbye gift (in terms of how meaningful the robot might think it to be). About half the children in each condition chose to do the fourth picture; we did not see any effects of the number of picture conversations (i.e., the three required vs. the optional fourth one) on the results. If

we looked at children's likelihood to perform all three activities (adding up the fourth picture, the sticker, and the goodbye gift, rather than any one individually), we saw a trend for children in the *E-B* condition to be slightly more likely to do all three activities, though this was not statistically significant.

#### 4.2.4. Children's Mirroring, Learning, and Relationship (H7)

We found that children who gave Tega a closer score on the IOS task were also more likely to use the target words in their stories, $r_{s(67)} = 0.359$, $p = 0.003$ (**Figure 13C**). They were also more likely to emulate the robot's stories as reflected by the number of exact and similar phrases used in their retells, $r_{s(67)} = 0.273$, $p = 0.025$ (**Figure 13B**). Given that age also correlated with children's ratings of Tega on the IOS task, we might suspect that age is more relevant than how close children felt to the robot. However, age did not correlate with children's use of exact and similar phrases, which suggests a deeper story.

In addition, children who placed Tega closer to the human in the Picture Sorting Task were also more likely to use phrases similar to the robot's, $r_{s(67)} = -0.299$, $p = 0.014$ (**Figure 13D**). There was a trend for children who placed Tega closer to the human to also rate Tega more closely on the IOS task, $r_{s(86)} = -0.197$, $p = 0.069$.

We did not observe any significant correlations of children's vocabulary scores with their phrase mirroring or any of the relationship assessments.

## 5. DISCUSSION

We asked whether a social robot that entrained its speech and behavior to individual children and provided an appropriate backstory about its abilities could increase children's rapport, positive relationship, acceptance, engagement, and learning with the robot. Below, we discuss the main findings and then discuss the implications of these findings.

### 5.1. Learning

Children learned the target vocabulary words in the robot's story and were generally attentive and engaged with the robot regardless of the experimental condition. They showed a variety of emotional expressions throughout the interaction. Children remembered the robot's story as evidenced by their ability to retell the story and their identification of target words on the vocabulary test. These results are in line with the prior study using this story activity (Kory Westlund et al., 2017b), which found significant learning gains.

We did see differences in children's learning by condition. Contrary to our hypotheses (H1), children in the *No Entrainment* condition correctly identified more target words than children in the *Entrainment* condition (**Figure 2B**). This could be for several reasons. A prior study found that a robot tutor that employed social adaptive behaviors led to lower learning gains than a robot that did not act as socially (Kennedy et al., 2015). Thus, perhaps the entraining robot was perceived more socially, which was detrimental in learning. This is contrary to our hypotheses regarding the importance of social behavior, rapport, and relationship in language learning with peers. However, in the prior study, children performed a math task with the robot tutor. The authors hypothesized that perhaps children were paying attention to the robot's social behavior as opposed to the lessons it was providing, or, alternatively, that the social behavior placed greater cognitive load on children thus inhibiting their ability to perform in the math task. Performance on a math task in a tutoring format may indeed benefit less from a robot's social behaviors than performance in a language-based story activity in a peer-learning format.

A second explanation pertains to the learning results we observed. There was a ceiling effect and little variance in children's responses, with 43% of children correctly identifying all six target words, and 41% correctly identifying 5 of the target words. If a significant number of children were already familiar with the target words, then the vocabulary tests would not reflect their learning during the task with the robot; the difference between conditions may not reflect children's learning in the task. Furthermore, given that children's receptive language abilities may precede their expressive abilities (Bloom, 1974; Ingram, 1974; Sénéchal, 1997), we would expect that children who correctly identified more words to also use more of them in their stories (H2), reflecting greater understanding and deeper encoding of the words (this was also seen in the prior study, Kory Westlund et al., 2017b). However, we did not see this correlation: children's use of the target words was not significantly correlated with correct identification of the words. In fact, children's use of the target words was significantly greater in the *E-B* condition than all others, in line with our hypotheses (H3) (**Figure 3**). Additionally, while the patterns were not significant, children were moderately more likely to use the words if they had identified them correctly in the *Entrainment* condition than in the *No Entrainment* condition. These results suggest that the robot's rapport- and relationship-building behaviors affected either or both of (a) children's learning and deeper understanding of the words such that they were more able to expressively use the words, or (b) children's mirroring of the robot's speech such that they used more of these target words, both of which would be in line with prior work linking rapport to learning (Sinha and Cassell, 2015a,b). This was also a short-term encounter. Given the positive aspects we see here regarding word use and mirroring, we expect that over multiple sessions, we would see greater differences in word learning.

When we examined children's mirroring of the robot's speech, we saw that children did mirror the robot (H2, **Figures 3**, **5**), in line with past work suggesting that children may mirror adults' syntax and speech (Huttenlocher et al., 2004) and earlier work in human-computer interaction showing that adults will entrain to computers and robots (e.g., Pearson et al., 2006; Lubold et al., 2018). However, we saw no significant differences in children's emulation of the robot's phrases, and in fact, less overlap in the number of unique words used by children that mirrored the words the robot used in the *E-NB* condition, and little difference among the other conditions (contrary to H3). This suggests that perhaps entrainment did not affect children's mirroring of the words the robot used so much as their expressive ability to use the key words present in the story. Prior work has shown that social robots can be successful at prompting children to demonstrate expressive vocabulary skills in both

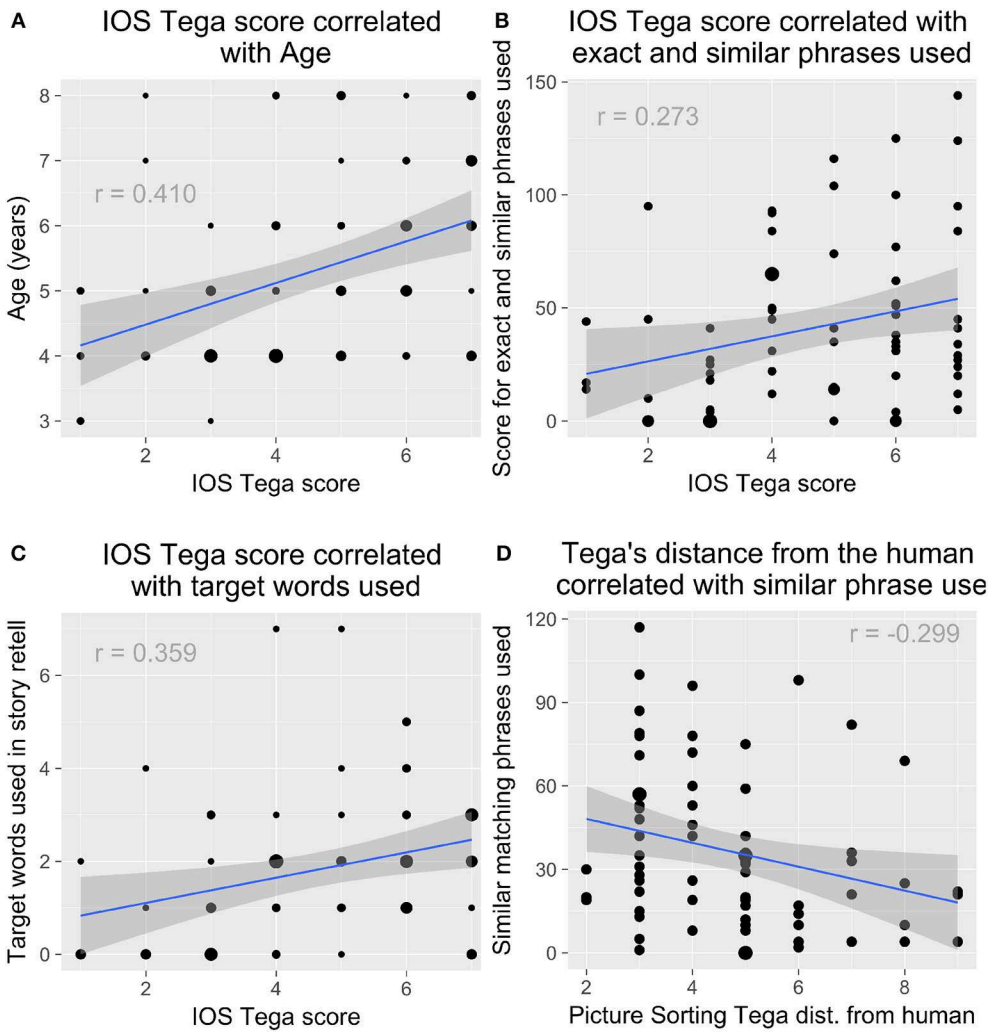



**FIGURE 13 | (A)** Older children rated the robot as closer in the IOS task. Children who rated the robot as closer were more likely to **(B)** use the target words in their stories and **(C)** emulate the robot's phrases. **(D)** Children who placed the robot closer to the human in the Picture Sorting Task were also more likely to emulate the robot.

vocabulary test and storytelling contexts (e.g., Kory and Breazeal, 2014; Kory Westlund et al., 2017b; Wallbridge et al., 2018). The present study suggests that the robot's entrainment may influence expressive ability.

The lack of difference in phrase mirroring was counter to our hypotheses (H3). Perhaps children did not feel sufficiently more rapport with the entraining robot for this to affect their storytelling. Indeed, in all conditions, the robot was a friendly, expressive character, which children generally said they felt close to—as close as to pet or parent, though less close than to a best friend. The entrainment only affected the robot's speech and some animations (which were played primarily in accompaniment with speech). In particular, if a child was very shy and rarely spoke, then the robot had fewer opportunities to adapt and entrain to that child. Perhaps greater difference would be seen if the robot also entrained other behaviors, such as posture, gesture, or word use. Another explanation is that perhaps language mirroring is not as closely linked to rapport as we expected; there is limited research so far suggesting this link, and more is needed.

## 5.2. Rapport, Relationship, and Social Behavior

The robot's entrainment and backstory also affected children's displays of positive emotions during the interaction. All children were engaged, but children in the *E-B* condition showed more positive emotions (e.g., joy, laughter, smiles, and positive valence), as well as fewer negative emotions (e.g., disappointment, fear) (supporting H5 and H6; see **Figures 7–10**). Laughter and smiling are social behaviors (Provine, 2001; Smidl, 2006; Manson et al., 2013). We also saw trends for children to be more helpful and accommodating in the *E-B* condition, as one might expect with a more social agent (Reeves and Nass, 1996), as evidenced by their behavior with fourth picture, the sticker task, and the goodbye gift. This is evidence that the robot's entrainment and backstory improved children's enjoyment of the

interaction and may have perceived it as more of a social agent, perhaps a result of increased rapport (supporting H5 and H6).

Children in the *E-B* condition also showed fewer attentive expressions, though only during the first half of the interaction (they did not differ later on). This could mean that these children were in fact less attentive initially, or it could mean that they were showing more positive attentive expressions that were coded by the affect recognition software as engagement and joy. If they were less attentive, we might expect this to be reflected in their vocabulary scores and story retellings—perhaps this is why these children did not identify as many words correctly. However, children in the *E-B* condition showed just as many expressions of engagement as children in the other conditions, were just as likely to retell the story, and as noted earlier, there were few significant differences by condition in children's story retellings beyond *more* use of the target words by children in the *E-B* condition. An alternative explanation is that perhaps children's attentive looks were related to how much cognitive effort was involved in performing the task. The robot's entrainment and backstory could have improved rapport and made the interaction more fluent, easier, and smoother, thus requiring less intense attention by children. This would be especially apparent earlier in the interaction, immediately following the robot's backstory disclosure and during the picture conversation task, when the robot was entraining more frequently due to the increased number of conversational turns during that task.

Related to this, we saw that children's attention increased over time in the *B* condition, but decreased in the *NB* condition, while multiple negative emotions (fear, disappointment, sadness) were displayed more frequently over time in the *B* condition than in the *NB* condition. For all other affective states measured, the change over time was not significant, though there were patterns for decreases in positive affect (e.g., joy, smiles, etc.) over time for all children. If children's attentive expressions were related to cognitive effort, this could indicate that in the *B* condition, children felt that over time, they had to attend more carefully to the robot (putting in more effort) in order to help it and deal with its hearing limitations. This could, perhaps, have led to increased feelings of difficulty interacting with the robot over time, which could have led to the increased displays of negative emotions that we observed in the *B* condition.

Regarding the decrease in attention in the *NB* condition, it may be that these children became less attentive because they were growing bored or were not as invested in the interaction. Indeed, while not statistically significant, children's engagement did decrease slightly more over time in the *NB* condition than in the *B* condition. There were also no affective states for which children in the *NB* condition increased their expression over time, suggesting that they became less expressive overall, which may be indicative of boredom or less emotional investment in the interaction.

We observed that children showed greater acceptance of the robot when they had heard the robot's backstory, as we expected (H4; **Figure 6**). Children's increased negative affect seen in the *B* condition may also reflect increased sympathy for the robot. Regardless, it seems that the robot's story influenced children's perceptions of it, in line with prior work showing that a robot's story does influence how people understand and react to it (Stenzel et al., 2012; Klapper et al., 2014; Darling et al., 2015; Kory Westlund et al., 2016b). Interestingly, this effect seemed to carry over to children's ideas about being friends with other children. While only a trend, it suggests room for future interventions using robots to help children understand and accept others different from themselves.

As noted above, children generally felt as close to the robot as they did to a pet, favorite toy, or parent, though not quite so close as to their best friend (**Figure 11A**). They generally placed Tega closer to the human adult than the table in the Picture Sorting Task, and frequently close to the human baby and to the cat (**Figures 11B**, **12**). These results present an intriguing picture regarding children's perceptions of the robot as a peer- or friend-like, non-human, animate entity. Children did not confuse the robot with a human; they knew it was different. Children seemed to clearly find companionship in the robot and to place it in a category between friend, pet, and authority figure. It was not merely a machine or computer; it was seen as more animate and alive—but not in the same category as a human. This jibes with prior work suggesting that children may categorize robots as in-between entities, with attributes of both living beings and mechanical artifacts (Kahn et al., 2002, 2012; Severson and Carlson, 2010). Perhaps children observed that some of the things that are messy about human relationships, such as the kinds of conflict that arise and the emotions that others display, are not the same in robot relationships—perhaps they are more like pet relationships. In this case, the robot did not get overly upset when it did not receive the sticker it wanted in the sticker task; it was generally cheerful throughout the interaction, which perhaps would not have been the case with another child. It is also likely that the robot's morphology influenced children's perceptions, since the robot we used was fluffy, colorful, and moved more like an animated character or sidekick than a humanoid being.

In support of our hypotheses regarding the connection between children's feelings of closeness, rapport, and relationship with learning and mirroring the robot (H7), we observed that children who rated the robot as closer to themselves also used the target words more often and emulated the robot's story more (**Figure 13**). This is in line with earlier work linking rapport to learning (Sinha and Cassell, 2015a,b). However, we also saw that age correlated with children's ratings of Tega on the IOS task. Older children rated the robot as closer; younger children as less closer. Perhaps younger children were less sure of the robot and needed more time to become comfortable with it. Given these correlations, we might suspect that age was more relevant to children's use of the target words and emulation of the robot's story than children's closeness ratings. However, children's age did not correlate with children's emulation of the robot's phrases at all, which suggests that this emulation was in fact related to children's feelings of closeness.

Finally, we also observed a few age differences. The length of children's story retellings differed with respect to their age, but did not vary by condition (**Figure 4**). Notably, the stories told by 6- and 7-year-old children were longest. The stories of 8-year-old children were not quite so long, which may have been because they were less interested in the story, rather than less capable. The

story and activity were designed with 4–7-year-olds in mind. The story may have been a little on the difficult side for the younger children, and on the easy side (and thus perhaps a little boring) for the oldest. However, even the children outside the target age range for the activity were receptive to the social robot, showing engagement, learning, and emulation.

Taken together, these results show that the robot's rapport and relationship-building behaviors do matter in interactions with young children. A robot that deliberately emulates a child's speech in a way similar to how people mirror each other can elicit more positive emotion and greater emulation of key words in a language learning activity. Children's feelings of closeness are related to their emulation of the robot's words in their stories.

## 5.3. Relation to Related Work

Our results also mirror, to an extent, the results in the prior study that explored a robot's use of expressive vs. flat speech (Kory Westlund et al., 2017b). In both studies, the robot's entrainment, backstory, and expressivity reflected the sensitivity the robot showed to the interaction. This sensitivity influenced children's engagement and learning. This is in line with work examining nonverbal behaviors in human-human learning interactions, in particular, nonverbal immediacy. *Nonverbal immediacy* refers to the *perceptual availability* of one's interaction partner, i.e., the use of nonverbal behaviors including gaze, gesture, posture, facial expressions, and vocal qualities such as prosody to signal general responsiveness and attentiveness. In human-human learning interactions, nonverbal immediacy has been linked to increased learning gains (Mehrabian, 1968; Christophel, 1990; Witt et al., 2004). When we examine prior child-robot interaction studies, we see that they have found a similar pattern of results to these human-human studies: The use of nonverbal immediacy behaviors including socially contingent behavior, appropriate gaze and posture, and vocal expressivity increased children's learning, engagement, and trust in a learning companion (Breazeal et al., 2016a; Kennedy et al., 2017; Kory Westlund et al., 2017a,b). Thus, it may be that the entrainment behaviors used by the robot increased its perceived immediacy and perceived sensitivity to the interaction.

However, in other work on language learning with social robots, the robot's social interactive capabilities have been found to influence children's relationships and social acceptance of the robot, but not their learning (e.g., Kanda et al., 2004, 2007, 2012). Indeed, some work has shown no significant differences in children's word learning from a social robot (with numerous embodied social capabilities) than from a tablet (e.g., Kory Westlund et al., 2015; Vogt et al., 2019). Arguably, these studies suggest a contrary story in which the robot's social capabilities may not affect children's learning that much.

These studies, however, have generally included learning tasks that did not require a robot or much social behavior for learning to proceed. For example, the second language learning activities used by Vogt et al. (2019) involved educational games presented on a tablet, for which the robot provided instructions, feedback, and support, but in which—as the authors acknowledge—the robot appeared to be non-critical for the learning interaction. The robot's social behavior may matter more for conversation and storytelling-based activities than for tablet games or simpler word learning tasks. Thus, we suspect that the robot's social capabilities (such as nonverbal immediacy) can influence children's learning—as we have seen here and in multiple other studies discussed earlier—but that the influence of social behavior is moderated by other factors, such as the extent to which the robot's sociality is necessary for the learning activity to proceed smoothly (as in the case of conversation and storytelling-based activities), and the extent to which the robot's social behavior helps build rapport.

This hypothesis is supported by Lubold and colleagues' recent work with middle school children and adults, in which a social robot with vocal entrainment contributed to increased learning on math tasks, though not increases in self-reported rapport (Lubold et al., 2016, 2018; Lubold, 2017). Because the vocal entrainment served not only to match pitch and other vocal features, but also made the robot's text-to-speech voice much more expressive, these studies could not disentangle the effects of expressivity from entrainment—however, both expressivity and entrain increase the robot's social capabilities. Our results here are similar to Lubold et al.'s, in that we also found that the robot's vocal entrainment was related to learning, but unlike Lubold's work, we also found connections between the robot's entrainment and aspects of children's relationship and rapport, including increased positive emotion and language emulation. This difference could be for numerous reasons, including the different age groups studied, the different learning matter (math vs. language), and the additional social and expressive capabilities of our robot.

Our results also extend prior work showing that children learn through storytelling with peer-like robot companions in ways that are significantly different from how children learn and engage with other technologies. We are seeing a peer learning dynamic similar to that seen in child-child interactions. Children socially model and emulate the behavior of the robots, like they do with other children. For example, children are more emotionally expressive when the robot is more expressive (Spaulding et al., 2016), show more curiosity in response to a robot's increased curiosity (Gordon et al., 2015), teach new tasks to robot peers (Park and Howard, 2015), and emulate linguistic phrases and vocabulary (Kory Westlund et al., 2017b). This study extends these previous works to explore not only *whether* children will learn with and emulate a robot peer, but the *mechanisms* by which robots can influence peer learning. Rapport and relationship appear to be two such mechanisms.

## 5.4. Limitations

This study had several limitations. First, we did not control for children's individual differences, particularly with regards to learning ability, language ability, or socio-economic status, all of which may affect individual children's social interactions and learning with the robot. Furthermore, we did not obtain an equal number of children at each age group to participate in the study. Future work should examine a more homogeneous sample as well as explore the stability of results across individual differences and across ages as children grow older.

We also lacked complete story retelling data and affect data for all children. Some children did not retell the story and in

a few cases, we had issues regarding the audio quality of the recorded stories. Some children's faces were not recognized by the Affdex software, and a few videos were missing or insufficiently captured a full frontal view of the children's faces, which was necessary for affect recognition. As a result, the analyses reported are underpowered. Future work should take greater effort to obtain quality audio and video recordings for all children during the study.

As mentioned in Kory Westlund et al. (2017b), the target vocabulary words were uncommon, but some children still may have known them. In particular, older children may have been familiar with some of the words, given the correlation we observed between children's age and the number of words identified correctly. The words' uncommonness may have cued children to pay attention to them; as such, future work should consider using nonce words or include a vocabulary pretest. Including a vocabulary pretest would also help ensure that children's language abilites did not differ by condition.

The robot's automated entrainment was limited to its speaking rate and pitch, so if a child was very quiet or spoke rarely, the robot would not have been able to entrain to that child. Because volume and exuberance were teleoperated, these occurred for all children. Future work could explore ways of encouraging shy children to speak up, or explore other modalities for entrainment, such as posture, gesture, facial expressions, and word use.

It is also unclear how generalizable the results are to robots with different embodiments or morphologies. The Tega robot that we used appears much like a fluffy stuffed animal, and thus is morphology could be seen as more familiar to children than a robot such as the Aldebaran NAO, which is humanoid. Children may feel a different level of comfort or uncanniness with a humanoid robot than with the Tega robot.

Finally, this study explored only a single one-on-one interaction with the robot. As such, any overall effects could be related to the novelty of the robot. However, children had the same amount of exposure to the robot in all conditions, so novelty cannot explain the differences we observed between conditions regarding the effects of entrainment and backstory.

Because learning tends to happen over time, as does the development of relationships, future work should explore longitudinal interactions to help us better understand the relationship between learning and rapport. Furthermore, children are frequently accompanied by friends and siblings in educational contexts. We do not know how multiple encounters with the robot or how interacting in groups might affect children's development of a relationship and rapport with the robot. Exploring group interactions that include multiple children, or children in concert with parents and teachers, could help us learn how to integrate robots into broader educational contexts and connect learning with peers to learning in school and at home.

## 6. CONCLUSION

In this work, we explored the impact of a robot's entrainment and backstory on children's engagement, rapport, relationship, and learning during a conversation and story activity. We found that the robot's rapport- and relationship-building behaviors affected children's emulation of the robot's words in their own stories, their displays of positive emotion, and their acceptance of the robot, and their perception of the robot as a social agent. This study adds to a growing body of work suggesting that the robot's social design impacts children's behavior and learning. The robot's story, use of relationship behaviors, nonverbal immediacy and rapport behaviors, social contingency, and expressivity are all important factors in a robot's social design.

## ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the MIT Committee on the Use of Humans as Experimental Subjects with written informed consent from all child subjects' parents and verbal assent from all child subjects. All child subjects' parents gave written informed consent and all child subjects gave verbal assent in accordance with the Declaration of Helsinki. The protocol was approved by the MIT Committee on the Use of Humans as Experimental Subjects.

## AUTHOR CONTRIBUTIONS

The study was conceived and designed by JK-W and CB. Data analysis was performed by JK-W. The paper was drafted, written, revised, and approved by JK-W and CB.

## FUNDING

This research was supported by an MIT Media Lab Learning Innovation Fellowship. The MIT Libraries Open Access Fund provided support for open access publication fees.

## ACKNOWLEDGMENTS

We would like to thank Paul Harris for his advice regarding the assessments, Kika Arias and Adam Gumbardo for help creating study materials and performing data collection, and Farida Virani, Branden Morioka, Anastasia Ostrowski, and David Cruz for additional help with data collection.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2019.00054/full#supplementary-material

## REFERENCES

Bailenson, J., Swinth, K., Hoyt, C., Persky, S., Dimov, A., and Blascovich, J. (2005). The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments. *Presence* 14, 379–393. doi: 10.1162/105474605774785235

Baker, S., Weinrich, B., Bevington, M., Schroth, K., and Schroeder, E. (2008). The effect of task type on fundamental frequency in children. *Int. J. Pediatr. Otorhinolaryngol.* 72, 885–889. doi: 10.1016/j.ijporl.2008.02.019

Bandura, A. (1971). *Social Learning Theory*. Morristown, NJ: General Learning Press.

Bandura, A., and Walters, H. (1963). *Social Learning and Personality Development*. New York, NY: Holt Rinehart and Winston.

Baxter, P., Ashurst, E., Read, R., Kennedy, J., and Belpaeme, T. (2017). Robot education peers in a situated primary school study: personalisation promotes child learning. *PLoS ONE* 12:e0178126. doi: 10.1371/journal.pone.0178126

Bell, L., Gustafson, J., and Heldner, M. (2003). "Prosodic adaptation in human-computer interaction," in *Proceedings of ICPHS*, Vol. 3 (Barcelona), 833–836.

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., and Tanaka, F. (2018). Social robots for education: a review. *Sci. Robot.* 3:eaat5954. doi: 10.1126/scirobotics.aat5954

Bennett, S. (1983). A 3-year longitudinal study of school-aged children's fundamental frequencies. *J. Speech Lang. Hear. Res.* 26, 137–141. doi: 10.1044/jshr.2601.137

Biernat, M. (2004). *Standards and Expectancies: Contrast and Assimilation in Judgments of Self and Others*. London: Psychology Press.

Bloom, L. (1974). "Talking, understanding, and thinking: developmental relationship between receptive and expressive language," IN *Language Perspectives, Acquisition, Retardation and Intervention*, eds R. L. Schiefelbusch and L. L. Lloyd (Baltimore, MD: University Park Press), 285–311.

Borrie, S. A., and Liss, J. M. (2014). Rhythm as a Coordinating Device: entrainment With Disordered Speech. *J. Speech Lang. Hear. Res.* 57, 815–824. doi: 10.1044/2014_JSLHR-S-13-0149

Breazeal, C. (2002). Regulation and entrainment in human-robot interaction. *Int. J. Robot. Res.* 21, 883–902. doi: 10.1177/027836490202101009 6

Breazeal, C., Dautenhahn, K., and Kanda, T. (2016a). "Social robotics," in *Springer Handbook of Robotics*, eds B. Siciliano and O. Khatib (New York, NY: Springer International Publishing), 1935–1972.

Breazeal, C., Harris, P. L., DeSteno, D., Kory Westlund, J. M., Dickens, L., and Jeong, S. (2016b). Young children treat robots as informants. *Topics Cogn. Sci.* 8, 481–491. doi: 10.1111/tops.12192

Brennan, S. E. (1996). Lexical entrainment in spontaneous dialog. *Proc. ISSD* 96, 41–44.

Burleson, W., and Picard, R. W. (2007). Gender-specific approaches to developing emotionally intelligent learning companions. *Intell. Syst. IEEE* 22, 62–69. doi: 10.1109/MIS.2007.69

Cassell, J., Geraghty, K., Gonzalez, B., and Borland, J. (2009). "Modeling culturally authentic style shifting with virtual peers," in *Proceedings of the 2009 International Conference on Multimodal Interfaces, ICMI-MLMI '09* (New York, NY: ACM), 135–142.

Chartrand, T. L., and van Baaren, R. (2009). Human mimicry. *Adv. Exp. Soc. Psychol.* 41, 219–274. doi: 10.1016/S0065-2601(08)00405-X

Chaspari, T., and Lehman, J. F. (2016). "An acoustic analysis of child-child and child-robot interactions for understanding engagement during speech-controlled computer games," in *INTERSPEECH* (San Francisco, CA), 595–599.

Chisholm, K., and Strayer, J. (1995). Verbal and facial measures of children′s emotion and empathy. *J. Exp. Child Psychol.* 59, 299–316. doi: 10.1006/jecp.1995.1013

Christophel, D. M. (1990). The relationships among teacher immediacy behaviors, student motivation, and learning. *Commun. Educ.* 39, 323–340. doi: 10.1080/03634529009378813

Clabaugh, C., Sha, F., Ragusa, G., and Mataric, M. (2015). "Towards a personalized model of number concepts learning in preschool children," in *Proceedings of the ICRA Workshop on Machine Learning for Social Robotics* (Seattle, WA), 26–30.

Darling, K., Nandy, P., and Breazeal, C. (2015). "Empathic concern and the effect of stories in human-robot interaction," in *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (Kobe), 770–775.

Davis, M. (ed.). (1982). *Interaction Rhythms: Periodicity in Communicative Behavior*. New York, NY: Human Sciences Press.

de Jong, N. H., and Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behav. Res. Methods* 41, 385–390. doi: 10.3758/BRM.41.2.385

De Lisi, R., and Golbeck, S. L. (1999). "Implications of piagetian theory for peer learning," in *Cognitive Perspectives on Peer Learning*, The Rutgers Invitational Symposium On Education Series, eds A. M. O'Donnell and A. King (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 3–37.

Desteno, D., Breazeal, C., Frank, R. H., Pizarro, D., Baumann, J., Dickens, L., et al. (2012). Detecting the trustworthiness of novel partners in economic exchange. *Psychol. Sci.* 23, 1549–1556. doi: 10.1177/0956797612448793

Dijksterhuis, A. (2005). Why we are social animals: the high road to imitation as social glue. *Perspect. Imit.* 2, 207–220.

Dijksterhuis, A., and Bargh, J. A. (2001). The perception-behavior expressway: automatic effects of social perception on social behavior. *Adv. Exp. Soc. Psychol.* 33, 1–40. doi: 10.1016/S0065-2601(01)80003-4

Favazza, P. C., and Odom, S. L. (1996). Use of the acceptance scale to measure attitudes of kindergarten-age children. *J. Early Intervent.* 20, 232–248. doi: 10.1177/105381519602000307

Favazza, P. C., Phillipsen, L., and Kumar, P. (2000). Measuring and promoting acceptance of young children with disabilities. *Except. Child.* 66, 491–508. doi: 10.1177/001440290006600404

Gelfer, M. P., and Denor, S. L. (2014). Speaking fundamental frequency and individual variability in caucasian and African American school-age children. *Am. J. Speech Lang. Pathol.* 23, 395–406. doi: 10.1044/2014_AJSLP-13-0016

Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S., et al. (2005). "Designing robots for long-term social interaction," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005 (IROS 2005)* (Edmonton, AB), 1338–1343.

Gola, A. A. H., Richards, M. N., Lauricella, A. R., and Calvert, S. L. (2013). Building meaningful parasocial relationships between toddlers and media characters to teach early mathematical skills. *Media Psychol.* 16, 390–411. doi: 10.1080/15213269.2013.783774

Gordon, G., and Breazeal, C. (2015). "Bayesian active learning-based robot tutor for children's word-reading skills," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence* (Austin, TX).

Gordon, G., Breazeal, C., and Engel, S. (2015). "Can children catch curiosity from a social robot?," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland: ACM), 91–98.

Gordon, G., Spaulding, S., Kory Westlund, J., Lee, J. J., Plummer, L., Martinez, M., et al. (2016). "Affective personalization of a social robot tutor for children's second language skill," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence* (Palo Alto, CA).

Grammer, K., Kruck, K. B., and Magnusson, M. S. (1998). The courtship dance: patterns of nonverbal synchronization in opposite-sex encounters. *J. Nonverb. Behav.* 22, 3–29. doi: 10.1023/A:1022986608835

Hacki, T., and Heitmüller, S. (1999). Development of the child's voice: premutation, mutation. *Int. J. Pediatr. Otorhinolaryngol.* 49(Suppl. 1), S141–S144. doi: 10.1016/S0165-5876(99)00150-0

Hartup, W. W., Laursen, B., Stewart, M. I., and Eastenson, A. (1988). Conflict and the friendship relations of young children. *Child Dev.* 59, 1590–1600. doi: 10.2307/1130673

Haviland, J. M., and Lelwica, M. (1987). The induced affect response: 10-week-old infants' responses to three emotion expressions. *Dev. Psychol.* 23, 97–104. doi: 10.1037/0012-1649.23.1.97

Hood, D., Lemaignan, S., and Dillenbourg, P. (2015). "When children teach a robot to write: an autonomous teachable humanoid which uses simulated handwriting," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI '15* (New York, NY: ACM), 83–90.

Huttenlocher, J., Vasilyeva, M., and Shimpi, P. (2004). Syntactic priming in young children. *J. Mem. Lang.* 50, 182–195. doi: 10.1016/j.jml.2003.09.003

Iio, T., Shiomi, M., Shinozawa, K., Shimohara, K., Miki, M., and Hagita, N. (2015). Lexical entrainment in human robot interaction. *Int. J. Soc. Robot.* 7, 253–263. doi: 10.1007/s12369-014-0255-x

Ingram, D. (1974). "The relationship between comprehension and production," in *Language Perspectives: Acquisition, Retardation, and Intervention* (Baltimore, MD: University Park Press), 670.

Kahn, P. H., Friedman, B., and Hagman, J. (2002). "I care about him as a pal: conceptions of robotic pets in online aibo discussion forums," in *CHI'02 Extended Abstracts on Human Factors in Computing Systems* (Minneapolis, MN: ACM), 632–633.

Kahn, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., et al. (2012). "Robovie, you'll have to go into the closet now": children's social and moral relationships with a humanoid robot. *Dev. Psychol.* 48:303. doi: 10.1037/a0027033

Kanda, T., Hirano, T., Eaton, D., and Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: a field trial. *Hum. Comput. Interact.* 19, 61–84.

Kanda, T., Sato, R., Saiwaki, N., and Ishiguro, H. (2007). A two-month field trial in an elementary school for long-term human #x2013;robot interaction. *IEEE Trans. Robot.* 23, 962–971. doi: 10.1109/TRO.2007.904904

Kanda, T., Shimada, M., and Koizumi, S. (2012). "Children learning with a social robot," in *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (London), 351–358.

Kennedy, J., Baxter, P., and Belpaeme, T. (2015). "The robot who tried too hard: social behaviour of a robot tutor can negatively affect child learning," in *Proceedings of HRI*, Vol. 15 (Portland, OR).

Kennedy, J., Baxter, P., and Belpaeme, T. (2017). Nonverbal immediacy as a characterisation of social behaviour for human–robot interaction. *Int. J. Soc. Robot.* 9, 109–128. doi: 10.1007/s12369-016-0378-3

Kidd, C. D., and Breazeal, C. (2008). "Robots at home: understanding long-term human-robot interaction," in *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference On* (Nice: IEEE), 3230–3235.

Kim, Y., Baylor, A. L., and PALS Group (2006). Pedagogical agents as learning companions: the role of agent competency and type of interaction. *Educ. Technol. Res. Dev.* 54, 223–243. doi: 10.1007/s11423-006-8805-z

Klapper, A., Ramsey, R., Wigboldus, D., and Cross, E. S. (2014). The control of automatic imitation based on bottom–up and top–down cues to animacy: insights from brain and behavior. *J. Cogn. Neurosci.* 26, 2503–2513. doi: 10.1162/jocn_a_00651

Kory Westlund, J. M., Dickens, L., Jeong, S., Harris, P. L., DeSteno, D., and Breazeal, C. L. (2017a). Children use non-verbal cues to learn new words from robots as well as people. *Int. J. Child Comput. Interact.* 13, 1–9. doi: 10.1016/j.ijcci.2017.04.001

Kory Westlund, J. M., Dickens, L., Sooyeon, J., Paul, H., DeSteno, D., and Breazeal, C. (2015). "A comparison of children learning new words from robots, tablets, and people," in *New Friends: The 1st International Conference on Social Robots in Therapy and Education* (Almere).

Kory Westlund, J. M., Jeong, S., Park, H. W., Ronfard, S., Adhikari, A., Harris, P. L., et al. (2017b). Flat versus expressive storytelling: young children's learning and retention of a social robot's narrative. *Front. Hum. Neurosci.* 11:295. doi: 10.3389/fnhum.2017.00295

Kory Westlund, J. M., Lee, J. J., Plummer, L., Faridi, F., Gray, J., Berlin, M., et al. (2016a). "Tega: a social robot," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Christchurch).

Kory Westlund, J. M., Martinez, M., Archie, M., Das, M., and Breazeal, C. (2016b). "Effects of framing a robot as a social agent or as a machine on children's social behavior," in *The 25rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (New York, NY: IEEE Press), 688–693.

Kory, J. (2014). *Storytelling with robots: effects of robot language level on children's language learning*. (Master's Thesis). Massachusetts Institute of Technology, Cambridge, MA.

Kory, J., and Breazeal, C. (2014). "Storytelling with robots: learning companions for preschool children's language development," in *2014 RO-MAN: The 23rd IEEE International Symposium on Robot and Human Interactive Communication* (Edinburgh), 643–648.

Kory-Westlund, J. M., and Breazeal, C. (2019). "Assessing children's perceptions and acceptance of a social robot," in *Proceedings of the 18th ACM International Conference on Interaction Design and Children, IDC '19* (New York, NY: ACM), 38–50.

Kory-Westlund, J. M., Park, H. W., Williams, R., and Breazeal, C. (2018). "Measuring young children's long-term relationships with social robots," in *Proceedings of the 17th ACM Conference on Interaction Design and Children* (Trondheim: ACM), 207–218.

Lakin, J. L., Jefferis, V. E., Cheng, C. M., and Chartrand, T. L. (2003). The chameleon effect as social glue: evidence for the evolutionary significance of nonconscious mimicry. *J. Nonverb. Behav.* 27, 145–162. doi: 10.1023/A:1025389814290

Lee, J. J., Knox, W. B., Wormwood, J. B., Breazeal, C., and DeSteno, D. (2013). Computationally modeling interpersonal trust. *Front. Psychol.* 4:893. doi: 10.3389/fpsyg.2013.00893

Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., and Bhogal, R. S. (1997). "The persona effect: affective impact of animated pedagogical agents," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, CHI '97* (New York, NY: ACM), 359–366.

Levitan, R., Benus, S., Gálvez, R. H., Gravano, A., Savoretti, F., Trnka, M., et al. (2016). "Implementing acoustic-prosodic entrainment in a conversational avatar," in *INTERSPEECH* (San Francisco, CA), 1166–1170.

Littlewort, G. C., Bartlett, M. S., Salamanca, L. P., and Reilly, J. (2011). "Automated measurement of children's facial expressions during problem solving tasks," in *Face and Gesture 2011* (Santa Barbara, CA), 30–35.

Lubold, N. (2017). "Building rapport through dynamic models of acoustic-prosodic entrainment," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA '17* (New York, NY: ACM), 297–300.

Lubold, N., Pon-Barry, H., and Walker, E. (2015). "Naturalness and rapport in a pitch adaptive learning companion," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (Scottsdale), 103–110.

Lubold, N., Walker, E., and Pon-Barry, H. (2016). "Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion," in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction, HRI '16* (Piscataway, NJ: IEEE Press), 255–262.

Lubold, N., Walker, E., Pon-Barry, H., and Ogan, A. (2018). "Automated pitch convergence improves learning in a social, teachable robot for middle school mathematics," in *Artificial Intelligence in Education*, Lecture Notes in Computer Science, eds C. Penstein Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, and B. du Boulay (New York, NY: Springer International Publishing), 282–296.

Manson, J. H., Bryant, G. A., Gervais, M. M., and Kline, M. A. (2013). Convergence of speech rate in conversation predicts cooperation. *Evol. Hum. Behav.* 34, 419–426. doi: 10.1016/j.evolhumbehav.2013.08.001

McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., and el Kaliouby, R. (2016). "AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA '16* (New York, NY: ACM), 3723–3726.

Mehrabian, A. (1968). Some referents and measures of nonverbal behavior. *Behav. Res. Methods Instrum.* 1, 203–207. doi: 10.3758/BF03208096

Movellan, J., Eckhardt, M., Virnes, M., and Rodriguez, A. (2009). "Sociable robot improves toddler vocabulary skills," in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* (San Diego, CA: ACM), 307–308.

Oster, H. (1978). "Facial expression and affect development," in *The Development of Affect*, Genesis of Behavior, eds M. Lewis and L. A. Rosenblum (Boston, MA: Springer US), 43–75.

Park, H. W., Coogle, R. A., and Howard, A. (2014). "Using a shared tablet workspace for interactive demonstrations during human-robot learning scenarios," in *2014 IEEE International Conference on Robotics and Automation (ICRA)* (Hong Kong), 2713–2719.

Park, H. W., Gelsomini, M., Lee, J. J., and Breazeal, C. (2017a). "Telling stories to robots: the effect of backchanneling on a child's storytelling," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI '17* (New York, NY: ACM), 100–108.

Park, H. W., Grover, I., Spaulding, S., Gomez, L., and Breazeal, C. (2019). "A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education," in *AAAI* (Honolulu).

Park, H. W., and Howard, A. M. (2015). "Retrieving experience: interactive instance-based learning methods for building robot companions," in *2015 IEEE International Conference on Robotics and Automation (ICRA)* (Seattle, WA), 6140–6145.

Park, H. W., Rosenberg-Kima, R., Rosenberg, M., Gordon, G., and Breazeal, C. (2017b). "Growing growth mindset with a social robot peer," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI '17* (New York, NY: ACM), 137–145.

Pearson, J., Hu, J., Branigan, H. P., Pickering, M. J., and Nass, C. I. (2006). "Adaptive language behavior in HCI: how expectations and beliefs about a system affect users' word choice," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '06* (New York, NY: ACM), 1177–1180.

Philippot, P., Feldman, R. S., and Coats, E. J. (1999). *The Social Context of Nonverbal Behavior*. Cambridge, UK: Cambridge University Press.

Piaget, J. (1932). *The Moral Development of the Child*. London: Kegan Paul.

Porzel, R., Scheffler, A., and Malaka, R. (2006). "How entrainment increases dialogical effectiveness," in *Proceedings of the IUI*, Vol. 6 (Sydney, NSW: Citeseer), 35–42.

Provine, R. R. (2001). *Laughter: A Scientific Investigation*. London: Penguin.

Reeves, B., and Nass, C. (1996). *How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge, UK: CSLI Publications and Cambridge University Press.

Reitter, D., Keller, F., and Moore, J. D. (2011). A computational cognitive model of syntactic priming. *Cogn. Sci.* 35, 587–637. doi: 10.1111/j.1551-6709.2010.01165.x

Richards, M. N., and Calvert, S. L. (2017). "Media characters, parasocial relationships, and the social aspects of children's learning across media platforms," in *Media Exposure During Infancy and Early Childhood: The Effects of Content and Context on Learning and Development*, eds R. Barr and D. N. Linebarger (Cham: Springer International Publishing), 141–163.

Rintjema, E., van den Berghe, R., Kessels, A., de Wit, J., and Vogt, P. (2018). "A robot teaching young children a second language: the effect of multiple interactions on engagement and performance," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, HRI '18* (New York, NY: ACM), 219–220.

Robins, B., Dautenhahn, K., Boekhorst, R. T., and Billard, A. (2005). Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills? *Univ. Access Inform. Soc.* 4, 105–120. doi: 10.1007/s10209-005-0116-3

Rotenberg, K. J. (1995). Development of children's restrictive disclosure to friends. *J. Genet. Psychol.* 156, 279–292. doi: 10.1080/00221325.1995.9914823

Rotenberg, K. J., Eisenberg, N., Cumming, C., Smith, A., Singh, M., and Terlicher, E. (2003). The contribution of adults' nonverbal cues and children's shyness to the development of rapport between adults and preschool children. *Int. J. Behav. Dev.* 27, 21–30. doi: 10.1080/01650250143000571

Rotenberg, K. J., and Mann, L. (1986). The development of the norm of the reciprocity of self-disclosure and its function in children's attraction to peers. *Child Dev.* 57, 1349–1357. doi: 10.2307/1130414

Rozin, P., and Cohen, A. B. (2003). High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans. *Emotion* 3, 68–75. doi: 10.1037/1528-3542.3.1.68

Rubin, K. H., Bukowski, W., and Parker, J. G. (1998). "Peer interactions, relationships, and groups," in *Handbook of Child Psychology: Social, Emotional, and Personality Development*, eds W. Damon and N. Eisenberg (Hoboken, NJ: John Wiley & Sons Inc), 619–700.

Sadoughi, N., Pereira, A., Jain, R., Leite, I., and Lehman, J. F. (2017). "Creating prosodic synchrony for a robot co-player in a speech-controlled game for children," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna: ACM), 91–99.

Scassellati, B., Boccanfuso, L., Huang, C.-M., Mademtzi, M., Qin, M., Salomons, N., et al. (2018). Improving social skills in children with ASD using a long-term, in-home social robot. *Sci. Robot.* 3:eaat7544. doi: 10.1126/scirobotics.aat7544

Semin, G. R., and Cacioppo, J. T. (2008). "Grounding social cognition: synchronization, coordination, and co-regulation," in *Embodied Grounding: Social, Cognitive, Affective, and Neuroscientific Approaches*, eds G. R. Semin and E. R. Smith (New York, NY: Cambridge University Press), 119–147. doi: 10.1017/CBO9780511805837.006

Sénéchal, M. (1997). The differential effect of storybook reading on preschoolers' acquisition of expressive and receptive vocabulary. *J. Child Lang.* 24, 123–138. doi: 10.1017/S0305000996003005

Senechal, T., McDuff, D., and el Kaliouby, R. (2015). "Facial action unit detection using active learning and an efficient non-linear kernel approximation," in *Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), ICCVW '15* (Washington, DC: IEEE Computer Society), 10–18.

Severson, R. L., and Carlson, S. M. (2010). Behaving as or behaving as if? Children's conceptions of personified robots and the emergence of a new ontological category. *Neural Netw.* 23, 1099–1103. doi: 10.1016/j.neunet.2010.08.014

Sinha, T., and Cassell, J. (2015a). "Fine-grained analyses of interpersonal processes and their effect on learning," in *Artificial Intelligence in Education* (Cham: Springer), 781–785.

Sinha, T., and Cassell, J. (2015b). "We click, we align, we learn: impact of influence and convergence processes on student learning and rapport building," in *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence, INTERPERSONAL '15* (New York, NY: ACM), 13–20.

Smidl, S. L. (2006). *Portraits of laughter in "kid" ergarten children: the giggles and guffaws that support teaching, learning, and relationships* (Doctoral dissertation). Virginia Tech.

Sorenson, D. N. (1989). A fundamental frequency investigation of children ages 6–10 years old. *J. Commun. Disord.* 22, 115–123. doi: 10.1016/0021-9924(89)90028-2

Spaulding, S., Gordon, G., and Breazeal, C. (2016). "Affect-aware student models for robot tutors," in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, AAMAS '16* (Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems), 864–872.

Stenzel, A., Chinellato, E., Bou, M. A. T., del Pobil, A. P., Lappe, M., and Liepelt, R. (2012). When humanoid robots become human-like interaction partners: corepresentation of robotic actions. *J. Exp. Psychol.* 38:1073. doi: 10.1037/a0029493

Suzuki, N., and Katagiri, Y. (2007). Prosodic alignment in human-computer interaction. *Conn. Sci.* 19, 131–141. doi: 10.1080/09540090701369125

Tickle-Degnen, L., and Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychol. Inq.* 1, 285–293. doi: 10.1207/s15327965pli0104_1

Tudge, J., and Rogoff, B. (1989). "Peer influences on cognitive development: piagetian and Vygotskian perspectives," in *Interaction in Human Development*, Crosscurrents in contemporary psychology, eds M. H. Bornstein and J. S. Bruner (Hillsdale, NJ: Lawrence Erlbaum Associates, Inc), 17–40.

Vogt, P., de Haas, M., de Jong, C., Baxter, P., and Krahmer, E. (2017). Child-robot interactions for second language tutoring to preschool children. *Front. Hum. Neurosci.* 11:73. doi: 10.3389/fnhum.2017.00073

Vogt, P., van den Berghe, R., de Haas, M., Hoffmann, L., Kanero, J., Mamus, E., et al. (2019). "Second language tutoring using social robots. A large-scale study," in *IEEE/ACM Int. Conf. on Human-Robot Interaction (HRI 2019)* (Daegu).

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Wallbridge, C. D., van den Berghe, R., Hernández Garcia, D., Kanero, J., Lemaignan, S., Edmunds, C., et al. (2018). "Using a robot peer to encourage the production of spatial concepts in a second language," in *Proceedings of the 6th International Conference on Human-Agent Interaction, HAI '18* (New York, NY: ACM), 54–60.

Weinberg, B., and Zlatin, M. (1970). Speaking fundamental frequency characteristics of five- and six-year-old children with mongolism. *J. Speech Lang. Hear. Res.* 13, 418–425. doi: 10.1044/jshr.1302.418

Wentzel, K. R. (1997). Student motivation in middle school: the role of perceived pedagogical caring. *J. Educ. Psychol.* 89, 411–419. doi: 10.1037/0022-0663.89.3.411

Whitebread, D., Bingham, S., Grau, V., Pasternak, D. P., and Sangster, C. (2007). Development of metacognition and self-regulated learning in young children: role of collaborative and peer-assisted learning. *J. Cogn. Educ. Psychol.* 6, 433–455. doi: 10.1891/194589507787382043

Wiltermuth, S. S., and Heath, C. (2009). Synchrony and cooperation. *Psychol. Sci.* 20, 1–5. doi: 10.1111/j.1467-9280.2008.02253.x

Witt, P. L., Wheeless, L. R., and Allen, M. (2004). A meta-analytical review of the relationship between teacher immediacy and student learning. *Commun. Monogr.* 71, 184–207. doi: 10.1080/036452042000228054

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

# Advantages of publishing in Frontiers

**OPEN ACCESS**
Articles are free to read for greatest visibility and readership

**FAST PUBLICATION**
Around 90 days from submission to decision

**HIGH QUALITY PEER-REVIEW**
Rigorous, collaborative, and constructive peer-review

**TRANSPARENT PEER-REVIEW**
Editors and reviewers acknowledged by name on published articles

**Frontiers**
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

**Visit us:** www.frontiersin.org
**Contact us:** info@frontiersin.org | +41 21 510 17 00

**REPRODUCIBILITY OF RESEARCH**
Support open data and methods to enhance research reproducibility

**DIGITAL PUBLISHING**
Articles designed for optimal readership across devices

**FOLLOW US**
@frontiersin

**IMPACT METRICS**
Advanced article metrics track visibility across digital media

**EXTENSIVE PROMOTION**
Marketing and promotion of impactful research

**LOOP RESEARCH NETWORK**
Our network increases your article's readership