Interaction acceptance modelling and estimation for a proactive engagement in the context of human-robot interactions

Timothée Dhaussy
LIA-CERI, Avignon University
timothee.dhaussy@univ-avignon.fr

Bassam Jabaian
LIA-CERI, Avignon University
bassam.jabaian@univ-avignon.fr

Fabrice Lefèvre
LIA-CERI, Avignon University
fabrice.lefevre@univ-avignon.fr

Abstract

Understanding human behavior in social environments provides valuable insights and information. When individuals require interaction with others, they rapidly assess the likelihood of engagement based on social signals and the displayed activity of the potential partner of interaction. We refer to this cognitive process as the Interaction Acceptance Belief (IAB). The concept of IAB finds application in various social robotic scenarios, including service tasks, proactive approaches, and reactive methods. In this paper, we present a comprehensive definition of Interaction Acceptance Belief and propose a methodology for its realistic modeling within real-world scenarios. Our approach aims to enhance the capabilities of social robots to effectively infer and adapt to human preferences, leading to efficient human-robot interactions. By conducting experimental evaluations, we establish the feasibility of developing a model that captures and represents the Interaction Acceptance Belief within a specific social context.

1. Introduction

The research on human-robot social interaction aims at modeling the concept of social intelligence through a robot. Replicating human behaviour in a social environment is a challenging task. A human-robot interaction (HRI) system should be able to naturally interact with a person while being able to respond to stimuli from its environment. Thus, taking action for the agent is performed through the analysis of signals emitted by people within reach, carried out by means of multimodal perceptions. The processes and abilities involved in perceiving, interpreting, and understanding social information are named social recognition [37] and its importance has been emphasized by Sandini et al. [19] in the context of HRI.

When it comes to initiating an interaction with someone, the human brain rapidly assesses the likelihood of its success. A stranger is less likely to accept the interaction, similar to a person already engaged in another interaction or occupied with a task unrelated to the agent. This scan of the person's availability helps to decide to engage or wait for a better opportunity. The person's behavior plays a significant role in this evaluation as we can estimate around 60-65% of all interpersonal communication or interaction is made up of nonverbal behavior [5].

The activity of the user may also give some clues about the person's availability. Furthermore, the activity and nonverbal behaviour can be visually observed. Unlike some other features, such as the individual's mental state or their personal proximity to the agent, these features are too intricate to deduce solely from an image. In this paper, we propose the concept of Interaction Acceptance Belief (IAB) which commonly answers the question ”What chances are my interaction to be accepted by the targeted user?”. This is a measure of uncertainty about the level of acceptance of potential interaction with an agent. Its expression may be manifested through the user's passive or active behaviour towards the agent.

The IAB opens up many possibilities for research on human-robot social interaction with the aim of modeling the concept of social intelligence through a robot. It brings crucial information for robots that may need to proactively [11] initiate an interaction with someone. The need for transparent reasoning [4] for robot actions based on perceptions and beliefs is an important subject in the human-robot interaction, and understanding why a robot would engage one person instead of another in a scenario with proactive interaction is carried out by the IAB.

2. Related Work

Engagement, as referred to by [22], in HRI is described as ”the process by which two participants establish, maintain, and end their perceived connection to one another”. Behavioural engagement is characterised by active participation in the moment of interaction, such as making eye contact, blinking at an abnormal rate, maintaining an appropriate body posture, and using hand gestures appropriately [25]. These indicators have been identified in studies of on-task behaviour and attention. Affective engagement refers to the emotional attachment of a user to the agent [17].

This perceived connection of the engagement shows a lot of similarity with IAB except that it occurs once the interaction has started. However, it can be inferred that the features utilized for the engagement should also be applicable to the IAB.

[26] defines this concept of interaction readiness as ”the extent to which a human prefers to have an interaction. The interest towards an interaction is measured by the interaction expected by a particular human. This interest is evaluated by using the observable cues displayed by that human.”

The level of visual focus of attention has been introduced by Das et al. [8] to help the agent decide to start an interaction. It is modelled through gaze patterns and contextual cues. Webb et al. [35] defined a visual social engagement metric and tested it in a simulation. This concept is derived from two social signals, proxemics (roughly positioning in the social space, management of this position) and mutual gaze, and is assessed in a simulation of group social interactions. The interaction readiness definition and social engagement metric do not represent the case where a person isn’t actively showing any particular interest and instead expresses their readiness or engagement in a passive way through their activity. It can be described as a subtle or latent expression of interest. Therefore, in this study, the primary focus will revolve around incorporating the key elements of social signals and action recognition as the central features.

Social signals processing (SSP) works at modeling, analysis, and synthesis of social signals in human–human, and human–machine interactions [32]. Gaze is a cognitive component of engagement considered the primary cue of attention [21, 12, 18]. Gesture analysis in SSP brings information about the connectedness between people [13]. Facial cues such as blink rate [3], facial action units [36], emotions [29], carry the affective state of participants. Models such as EMONET developed by [30] have increased the ability to continuously estimate variance and arousal. Spatial behaviour, or proxemics, constitutes the dynamic process by which individuals position themselves in social interactions [14]. The social proximity of a person can be indicated by body posture with the orientation of the face and body towards an interlocutor [20, 38].

Recognition of human action has shown growing interest in the last two decades with the emergence of deep learning. It aims at predicting the current activity of a person through a stream of images. Different architectures have been proposed in the last years, two streams CNN-based methods [24, 34] composed of dual spatial flow and temporal flow, RNN-based methods [9, 15] which take advantage of RNN temporal properties, 3D CNN-based methods [31, 6] where video frames with spatial and temporal dimensions are used as input for a 3D CNN model then conveyed to Transformers-like models with attention mechanism [16, 1].

Hence to complement the prior work described above, our study aims to give a definition of IAB, propose a modelization for it, and address the prediction of IAB in real time within the context of proactive HRI scenarios. To achieve this goal, we explore the various modalities. By presenting our approach within a practical use-case scenario, we strive to demonstrate the applicability and effectiveness of the proposed methodology.

3. IAB Model

This section presents a model of the IAB, which is amenable to implementation in a real-life standard scenario.

3.1. IAB Modeling

We aim to perform IAB prediction in a conducive environment within the context of proactive HRI, such as a hospital waiting room. Patients wait for their turn in a rather static manner, they can be standing or sitting on a chair, using their phone, or simply listening to some music. People raising their hands are asking for an interaction; this interaction has almost no chance to be rejected except in some rare cases of wrong calls. A person smiling at the agent may lead to a successful proactive interaction. Both situations yield a similar outcome, where the interaction is accepted. However, the level of uncertainty varies between them.

An observation window is defined as a time interval $[t-\tau, t]$, which captures the last τ seconds of user behaviour leading up to time t. This window serves as the basis for generating a feature vector $[x_t - \tau, ..., x_{t-1}, x_t]$, which includes the frames within the interval. This feature vector is used as input for the classifier. The output of the classifier assigns a label to each observation window, indicating the degree to which the user can be engaged during that time period.

At time step t, we construct a model that classifies the observed user behaviour within the interval $[t - \tau, t]$ as either ”may accept an interaction” or ”may not accept an interaction”. Let $X = [x_1, x_2, ..., x_T]$ represent the sequence of multimodal user behaviour feature vectors, and let $Y = [y_1, y_2, ..., y_T]$ denote the corresponding sequence of binary output labels, with $y_t = C([x_{t-\tau}, ..., x_{t-1}, x_t])$, C being the classifier.

3.2. Pre-processing

To predict IAB levels, a multi-level approach is used, involving the extraction of four different levels of perceptions. These perceptions include:

Head features The first level of perception involves analyzing the eye gaze, head position and action unit of the individual. Action Units (AUs) refer to facial muscle movements or configurations associated with specific facial expressions [28]. This information provides insights into where the person is looking and the orientation of their head. One possible tool to infer gaze orientation and head pose is OpenFace[1], a popular open-source facial behavior analysis toolkit [2].

Body key points The second level of perception investigates the impact of body key points in the image, and how they relate to particular body poses, to enable a deeper understanding of the relationship between body language and IAB levels. These features are extracted per frame with the YoloV7 model [33].

Action Features The third level of perception involves extracting features related to the person's actions. This could include analyzing gestures, body movements, or other behavioral cues that provide information about the individual's ongoing activity, including engagement or interaction. An I3D model [6], pre-trained on the Charades dataset [23], extracts action features through a sliding window of 64 frames on images of users delimited by their bounding boxes. Extracted feature vector has a length of 1024.

Emotions The fourth level of perception focuses on extracting emotional states. This involves analyzing the person's facial expressions to determine their emotional valence and arousal. Additionally, the five primary expressions, namely Neutral, Happy, Sad, Surprise, and Fear, can be detected. The detection of emotions is provided through an implementation of [30] and the extraction is performed per frame.

To achieve synchronized feature vectors, we employ temporal integration (also known as temporal pooling) by applying a common integration window to all feature streams. The integration process involves applying a specific integration function, such as mean and variance, over sliding integration windows of length L seconds. In this study, statistics-based integration functions are used. Specifically, the mean and variance functions are utilized. The integration window length L is set to 500 ms, and there is no overlap between the integration windows.

4. Experiments and Results

In this section, the experiments and results are reported, after the dataset creation process has been described along with the metrics used.

4.1. Dataset

The dataset [2] utilized for this project was collected to create a simulation of patients in a waiting room. To achieve this, we employed non-professional actors who were instructed to play behaviors while being recorded. A dedicated actor, representing the agent's perspective, moved around the room with a camera positioned at the torso level. At times, the actors were requested to switch their behavior scenarios to introduce variety within the scene and be representative of real-world situations. The scenes' durations fall within a range of 30s to 2mn, for a total duration of around one hour. The following enumeration presents a comprehensive list of potential actor behaviors considered in this study:

- Engage in conversations with individuals seated adjacent to them.
- Active use of mobile phones, such as playing games or browsing the Internet.
- Exhibit passive behaviour, where they remain idle and appear to be waiting without any specific engagement or activity.
- Show signs of interest and attentiveness towards the agent.
- Seek the agent's attention or assistance, requesting information, guidance, or support.

A diverse group of actors, comprising 12 individuals of varying genres but all aged in $[20, 30]$, was selected to portray the role of patients in the study. To simplify the process and ensure reproducibility experiences are recorded with the front camera of an Apple iPhone 13 at 30fps. The IAB level of each actor in the video has been meticulously labeled on a scale of 1 to 5 by two annotators. The instruction given to data annotators is for each point, the closer it is to one, the more unlikely it is to accept interaction. Conversely, the closer it is to five, the more likely it is to accept an interaction. An example is also provided to the annotators for each level. After labeling, the Cohen Kappa coefficient score between annotators is 0.88. Examples of behaviors from the datasets are pictured in Figure 1, for which the IAB labels are from top to bottom left to right: 3, 5, 2, 1 for both actors, 4 and 2.

[1] `https://github.com/TadasBaltrusaitis/OpenFace`

[2] The dataset is available for research purposes with a simple request to the authors.

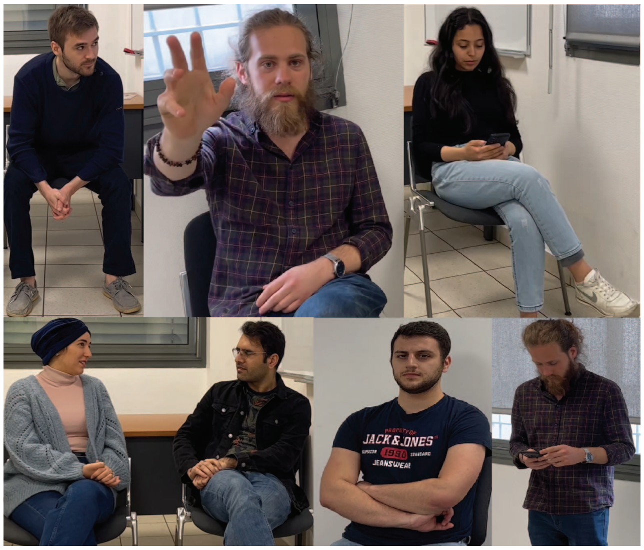

Figure 1. Examples of behaviors from the dataset.

IAB	1	2	3	4	5
#duration (s)	1532	1295	1521	391	322

Table 1. IAB values representation in the dataset (in seconds).

The dataset exhibits a significant class imbalance, with IAB labels 4 and class 5 representing approximately 14% of the overall class distribution. Yet at the same time it aroused naturally from the human playing their scenarios, and choosing their moment to express interest in an interaction.

4.2. Metrics

The initial investigation focuses on evaluating the performance of a classic recurrent neural network (RNN) across different modalities. To evaluate the results, we compute the Area Under the ROC Curve (AUROC) and F1 score. The AUROC serves as an indicator to assess the performance of binary classification models [10]. It provides a comprehensive assessment of the model's ability to differentiate between classes across various threshold values. By considering the balance between recall (proportion of correctly identified positive cases) and specificity (proportion of correctly identified negative cases), the AUROC quantifies the model's overall discriminative power. The F1 score quantifies the balance between precision (proportion of predicted positive cases that are correctly identified as positive) and recall, providing a more robust measure when dealing with imbalanced data sets [27].

4.3. Experiments

The model used in our experiment is based on the many-to-one GRU architecture, as introduced by Cho et al. [7] GRUs are a type of RNN with additional gating mechanisms that help control the flow of information within the network. These gating mechanisms enable GRUs to better manage the flow of information through time and mitigate the vanishing gradient problem, allowing them to capture long-term dependencies more effectively than traditional RNNs. The architecture consists of a single GRU layer with 128 hidden dimensions, which act as memory cells to store and propagate information over time.

In the context of a small-sized dataset, the model can easily become overly sensitive to specific training examples, causing instability during training. Minor variations or outliers in the limited data can have a significant impact on the model's learnt representation. It may also show difficulty in capturing complex patterns: GRUs, are designed to learn intricate patterns and relationships within the data. However, with too little data, the model may not have enough diverse examples to capture the full complexity of the problem, leading to suboptimal performance. To address these challenges, we employ an oversampling technique in the training set to augment the data and achieve a more balanced label representation. A 5-fold cross-validation is employed to leverage the available data, the best model is kept. Models are then evaluated on a test set of 10 minutes.

4.4. Results

We conducted tests involving various features and their combinations. The results obtained for each individual feature and the top 3 combination of features are presented in Table 2.

Features/window length		F1	AUROC
Head	6s	0.59	0.75
Bodypose	3s	0.52	0.72
Action	5s	0.37	0.61
Emotions	4s	0.36	0.5
Head+Bodypose	4s	0.61	**0.8**
Head+Bodypose+Emotion	2s	**0.63**	0.79
Head+Bodypose+Action+Emotion	6s	0.57	0.75

Table 2. F1 and AUROC scores of the various features in the optimal window length.

The action and emotion features overall perform poorly. The body key points provide some clues for the model to evaluate the IAB of a situation. With an AUROC of 0.7, the body pose model has a moderate ability to differentiate between positive and negative cases.

Head and body pose features emerge as common elements among the top three feature combinations. This observation highlights the consistent significance of these perceptual cues in contributing to the overall performance of the predictive model. The presence of head and body pose features in the top-performing combinations reinforces their fundamental role in capturing relevant information and discerning interaction acceptance patterns. Action features do not exhibit a significant influence on the predictive performance or contribute to the improvement of the overall prediction when combined with other features.

A noteworthy observation arising from our experiments is that a significant proportion of the top-performing models demonstrate window lengths that are equal to or exceed 4. This finding suggests that the task of predicting IAB benefits from gathering more observations until a certain threshold is reached. As the temporal context increases through the use of longer window lengths, the predictive models seem to gain an advantage in accurately discerning IAB dynamics.

The combination Head+Bodypose+Emotion yields the best F1 score, achieved with a window length of only 2 seconds. This intriguing result suggests that this particular combination mitigated overfitting up to a certain extent while also facilitating the recognition of specific scenes where the actor's behavior transitioned to exhibit interest in the agent and actively called for interaction. The utilization of emotion cues in conjunction with head and body pose features appears to slightly enhance the model's capacity to capture critical behavioral shifts and expressions of interest, contributing to the successful prediction of interaction acceptance in such instances.

5. Conclusion

This paper introduces and defines Interaction Acceptance Belief (IAB) in the field of Human-Robot Interaction. We introduced a novel modelization approach for IAB prediction and thoroughly validated it using a real-world dataset comprising various scenarios. The obtained results present promising and encouraging outcomes. Notably, our research underscores the critical role played by head and body pose, particularly when combined, in achieving accurate predictions of IAB. The incorporation of these perceptual cues significantly contributes to the model's ability to discern and comprehend the dynamics of interaction acceptance in various real-world contexts.

Moving forward, our future work involves testing our framework with a speech-enabled robotic agent in diverse contexts or increasing complexity. Additionally, we aim to enhance the modelization by considering the distinction between passive and active behaviors to more accurately estimate the IAB. This refined modelization is expected to contribute to a more robust understanding of human-robot interaction dynamics as a whole.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer, 2021.

[2] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2018.

[3] Esube Bekele, Joshua Wade, Dayi Bian, Lian Zhang, Zhi Zheng, Amy Swanson, Medha Sarkar, Zachary Warren, and Nilanjan Sarkar. Multimodal interfaces and sensory fusion in vr for social interactions. volume 8525, 06 2014.

[4] Paul Bremner, Louise A. Dennis, Michael Fisher, and Alan F. Winfield. On proactive, transparent, and verifiable ethical reasoning for robots. *Proceedings of the IEEE*, 107(3):541–561, 2019.

[5] David B. Burgoon, Judee K.and Buller. Interpersonal deception: Iii. effects of deceit on perceived communication and nonverbal behavior dynamics. *Journal of Nonverbal Behavior*, 18(2):155–184, Jun 1994.

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018.

[7] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

[8] Dipankar Das, Md. Golam Rashed, Yoshinori Kobayashi, and Yoshinori Kuno. Supporting human–robot interaction based on the level of visual focus of attention. *IEEE Transactions on Human-Machine Systems*, 45(6):664–675, 2015.

[9] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description, 2016.

[10] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ROC Analysis in Pattern Recognition.

[11] Jasmin Grosinger. On proactive human–ai systems. *International Workshop on Artificial Intelligence and Cognition*, 2022.

[12] Malia F. Mason, Elizabeth P. Tatkow, and C. Neil Macrae. The look of love: Gaze shifts and person perception. *Psychological Science*, 16(3):236–239, 2005.

[13] David Mcneill. Hand and mind: What gestures reveal about thought. *Bibliovault OAI Repository, the University of Chicago Press*, 27, 06 1994.

[14] Ross Mead, Amin Atrash, and Maja J. Matarić. Proxemic feature recognition for interactive robots: Automating metrics from the social sciences. In Bilge Mutlu, Christoph Bartneck, Jaap Ham, Vanessa Evers, and Takayuki Kanda, editors, *Social Robotics*, pages 52–61, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[15] Lili Meng, Bo Zhao, Bo Chang, Gao Huang, Wei Sun, Frederich Tung, and Leonid Sigal. Interpretable spatio-temporal attention for video action recognition, 2019.

[16] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network, 2021.

[17] Reinhard Pekrun and Lisa Linnenbrink-Garcia. *Academic Emotions and Student Engagement*, pages 259–282. Springer US, Boston, MA, 2012.

[18] Christopher Peters, Catherine Pelachaud, Elisabetta Bevacqua, Maurizio Mancini, and Isabella Poggi. Engagement capabilities for ecas. *Autonomous Agents and Multi-agent Systems - AAMAS*, 01 2005.

[19] Giulio Sandini, Vishwanathan Mohan, Alessandra Sciutti, and Pietro Morasso. Social cognition for human-robot symbiosis—challenges and building blocks. *Frontiers in Neurorobotics*, 12, 2018.

[20] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W. McOwan, and Ana Paiva. Automatic analysis of affective postures and body motion to detect engagement with a game companion. HRI '11, page 305–312, New York, NY, USA, 2011. Association for Computing Machinery.

[21] Candace Sidner, Christopher Lee, and Neal Lesh. Engagement when looking: behaviors for robots when collaborating with people. 12 2003.

[22] Candace L. Sidner, Christopher Lee, Cory Kidd, Neal Lesh, and Charles Rich. Explorations in engagement for humans and robots.

[23] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ivan Laptev, Ali Farhadi, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. *ArXiv e-prints*, 2016.

[24] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[25] Gale M. Sinatra, Benjamin C. Heddy, and Doug Lombardi. The challenges of defining and measuring student engagement in science. *Educational Psychologist*, 50(1):1–13, 2015.

[26] Chapa Sirithunge, A. G. Buddhika P. Jayasekara, and D. P. Chandima. Proactive robots with the perception of nonverbal human behavior: A review. *IEEE Access*, 7:77308–77327, 2019.

[27] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing Management*, 45(4):427–437, 2009.

[28] Marwa Mahmoud Tadas Baltrušaitis and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2015.

[29] Michael Thiruthuvanathan, Christ, Balachandran Krishnan, and M. A. Dorai Rangaswamy. Engagement detection through facial emotional recognition using a shallow residual convolutional neural networks. *International Journal of Intelligent Engineering and Systems*, 14:236–247, 2021.

[30] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3, 01 2021.

[31] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks, 2015.

[32] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009. Visual and multimodal analysis of human spontaneous behaviour:.

[33] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022.

[34] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2015.

[35] Nicola Webb, Manuel Giuliani, and Séverin Lemaignan. Measuring visual social engagement from proxemics and gaze. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 757–762, 2022.

[36] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98, 2014.

[37] Robert S. Wyer and Donal E. Carlston. *Social cognition, Inference, and attribution*. HALSTED PRESS, 1979.

[38] Zhou Yu, Xinrui He, Alan Black, and Alexander Rudnicky. User engagement study with virtual agents under different cultural contexts. volume 10011, pages 364–368, 09 2016.