A Transparent Framework towards the Context-Sensitive Recognition of Conversational Engagement

Alexander Heimerl, Tobias Baur, Elisabeth André [1]

Abstract.
Modelling and recognising affective and mental user states is an urging topic in multiple research fields. This work suggests an approach towards adequate recognition of such states by combining state-of-the-art behaviour recognition classifiers in a transparent and explainable modelling framework that also allows to consider contextual aspects in the inference process. More precisely, in this paper we exemplify the idea of our framework with the recognition of conversational engagement in bi-directional conversations. We introduce a multi-modal annotation scheme for conversational engagement. We further introduce our hybrid approach that combines the accuracy of state-of-the art machine learning techniques, such as deep learning, with the capabilities of Bayesian Networks that are inherently interpretable and feature an important aspect that modern approaches are lacking - causal inference. In an evaluation on a large multi-modal corpus of bi-directional conversations, we show that this hybrid approach can even outperform state-of-the-art black-box approaches by considering context information and causal relations.

1 Introduction

Nowadays, machine learning approaches are most often purely data-driven as they use so-called "black-box" approaches that map low-level features or decisions of previous classifiers onto abstract labels following statistical methods. Here we usually have no transparent concept of how the model is internally represented, e.g. how and why weights on the nodes of artificial neural networks are related.

In most research areas (e.g., in psychology, behaviour analysis, but also physics), the goal of creating a model is to reason about observations in the world, while creating and validating theories that aim to find causation and explanations. Then, such models are often validated in simulations, or collated with real-world observations. That means on the one hand, we have data-driven models in machine learning that do a decent job in creating predictions for a huge amount of recognition problems, but deliver no transparent way to understand their decisions and don't necessarily have a theory behind them. On the other hand, we have models that aim to explain interrelations of observations of the world and/or of their inner states. Such models are also called "white-box" approaches.

In this paper, we suggest a hybrid approach that combines state-of-the-art "black-box" recognition models with a transparent causal inference model. Lately, the focus of research tends towards deep end-to-end learning with artificial neural networks. While such approaches deliver promising results on audio-visual data, they only give little insight on how and why they predict behaviours the way they do. In this work, we investigate the recognition of "conversational engagement". Especially in scenarios where it is essential to know *why* a person's behaviour is interpreted as, e.g., "strongly disengaged", the idea is often to identify cues that led to this interpretation, providing an additional abstraction layer. Here, the relevance of a comprehensible model becomes very clear. Imagine a system that gives feedback on how engaged a person appeared in a social coaching scenario. A model should be able to give feedback on *why* it decided a person appeared to be strongly engaged or disengaged, so that a human can learn from the feedback. In order to infer complex social signals with a transparent model, we combine predictions of multiple high-precision classifiers with dynamic Bayesian networks (DBN) [32]. DBNs are probabilistic models that allow expressing causal relationships between nodes in a network, while at the same time considering previous observations. Even tough the parameters for such nodes and even the overall network structure may be learned with machine learning techniques, DBNs allow retracing the decisions they are making for each node or layer of nodes visually and are therefore inherently interpretable. While the structure of a DBN may be modelled based on a theory and grounded in social sciences, our framework allows to consider parallel observations, so it can learn correlations between concurrent behaviours, context and the complex phenomena of interest.

2 Related work

2.1 Engagement in psychology

Engagement is a complex social attitude. This becomes apparent when being confronted by the mass of available definitions. In fact Glas et. al [17] gave an overview of many different engagement definitions, with some of them being very context specific. The definition of Poggi coincides best with a general understanding of engagement. She describes it 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." [36]. As complex as it is to acquire a fitting definition, equally complex is the manifestation of engagement in conversations. There are multiple behaviours that are strongly connected to it.

In general, body language is an elemental part in expressing conversational engagement. To be more precise, the alignment of the body and the limbs play an important role on broadcasting the state of engagement [31]. Interlocutors, that are engaged during a conversation, align their bodies to each other, as described in [22], "to create a frame of engagement".

However not only the body position and body movement relative to each other is an important criteria, also the individual body behaviour is of great interest. Lots of body movement may indicate some kind

[1] Augsburg University, Germany, Universitätstr. 6a, email: heimerl@hcm-lab.de

of restlessness. This was found to be connected to boredom, which is a manifestation of low engagement [14]. Also depending on the level of engagement the body reacts with more subtle signals. Heart rate, blood pressure, EEG and galvanic skin response are all potential candidates to draw conclusions about engagement [51].
Moreover specific gestures may allow to draw conclusions about the level of engagement. Lausberg [25] investigated, among other things, the origin of self-touch gestures. She describes, that self-touch gestures occur when people are emotionally engaged. Alongside self-touch gestures there are also more complex gestures, that reflect different affective states [28].
Another crucial part in human interaction are "Feedback / Backchannels". It describes a high-level behaviour that is related to engagement. Backchannels are a kind of feedback. They occur between interlocutors and are typically in the form of non-intrusive acoustic or visual signals, e.g. a simple "Yes" or a headnod. Backchannels are a tool, to not only signal the success of communication, but also provide information about the level of engagement [17].
A strong form of engagement manifestation is mirroring of behaviours, be it acoustic or visual, from one interlocutor by the other. Those go by the terms "Synchrony", "Mimicry" or "Alignment". All of those represent a connection or bonding between interlocutors [17].

2.2 Recognition of engagement

Engagement has been investigated from various research angles, e.g. how to define engagement, how to annotate engagement or how to automatically predict engagement. Therefore it is no surprise that there are many different systems available to automatically predict engagement.
Rich et al. [39] introduced a reusable module for the recognition of engagement in human-robot interaction. They identified four connection events that they found to be tools for the maintenance of engagement. The four events were, directed gaze, mutual facial gaze, adjacency pairs, verbal and non-verbal backchannels. Those concepts built the theoretical foundation for their engagement recognition module.
Sanghvi et al. [45] predicted engagement based on body posture features. All their features have been extracted from video signals. They identified following important posture features: "Body lean angle", "Slouch factor", "Quantity of motion" and "Contraction index". For the classification they used Weka [16] and evaluated 63 different classifiers. The best ones achieved a prediction accuracy of 82% on the two classes "engaged" and "not engaged".
Roman Bednarik et al. [7] focused on recognising conversational engagement with gaze data. Further, they introduced an annotation scheme for the different levels of conversational engagement. They defined a total of six levels. In ascending order, the first being the lowest level of engagement and the last being the highest level of engagement: "No interest", "Following", "Responding", "Conversing", "Influencing discussion discourse/topic" and "Governing/managing discussion". To ease down the classification task the authors decided to reduce the six classes of engagement to a two-classes problem - low and high engagement. For the automatic estimation they computed a total of 26 features from the raw eye gaze data, e.g. number of fixations, number of saccades, minimal and maximal fixation duration, minimal and maximal saccade amplitude, quantity of fixation at the speakers' face. Those features have been used to train a SVM. Following this approach they achieved a prediction accuracy of 74%.
Yun et al. [56] proposed a convolutional neural network(CNN) to automatically predict engagement of children. For training their CNN they relied solely on facial images. However due to limited training data they used CNNs that have been pre-trained on face recognition tasks. Their network architecture includes a new layer combination to model temporal dynamics in order to extract high-level features from low-level features. For predicting engagement they distinguished between four levels of engagement, high engagement, low engagement, low disengagement and high disengagement. On the given task their network architecture achieved a balanced accuracy of 0.7807.

There is already plenty of research available that targets recognising engagement. However most of the systems focus solely on finding feasible features, either handcrafted or extracted from convolutional layers to optimise prediction accuracy. Little attention is payed to context, which is important when it comes to recognising engagement in everyday scenarios. Depending on the environment individuals are in it can affect how people behave and also what kind of cues they are using during a conversation. Imagine a student talking to his friend during a break in comparison to a student attending an oral exam. However not only external factors can influence the broadcasting of engagement. Also the very unique psychological traits every person has can influence their behaviour. An extrovert person in comparison to an introvert person can appear totally different during a conversation. Those examples illustrate potential context information that should be considered when recognising engagement.

2.3 Bayesian networks

Bayesian networks have been successfully applied in earlier work in the area of high-level interpretation of social signals. One of the pioneer studies is the work by Conati et al. [11]. They have incorporated bio-feedback sensors into a complex emotion model, that was based on a subset of the emotions proposed by OCC theory [34]. They employed a dynamic decision network (a generalisation of a dynamic Bayesian network) to capture many of the complex phenomena associated with appraisal theories. In particular, their model estimated student goals based on personality traits and events which represent changes in the environment (e.g., progress in the system) as well as evidence from physical feedback channels to support the model's prediction.
Sabourin et al. [43] focused, similar to Conati et al., on learners' emotions, and employed multiple variations of Bayesian networks. More specifically, they investigated the benefits of using cognitive models of learner emotions, to guide the development of Bayesian networks for prediction of student affect. Predictive models were empirically trained on data, acquired from 260 students interacting with a game-based learning environment. As a dynamic Bayesian network turned out to be the most successful model, they emphasised the importance of temporal information in predicting learner emotions. They concluded that predictive models may be used to validate theoretical models of emotion.
Wöllmer et al. [55] combined a hierarchical dynamic Bayesian network to detect linguistic keyword features together with long short-term memory (LSTM) neural networks [19] which model phoneme context and emotional history to predict the affective state of the user. This way, they are combining acoustic, linguistic, and long-term context information to continuously predict the current valence and activation in a two-dimensional emotion space.
Lugrin et al. [26] used Bayesian networks to incorporate culture into intelligent systems by combining theory-based and data-driven approaches. Their network aims to generate non-verbal culture-

dependent behaviours. While the model is structured based on cultural theories and theoretical knowledge of their influence on prototypical behaviour, the parameters of the model are learned from a multi-modal corpus recorded in the German and Japanese cultures. In their work, they aim to generate adequate behaviours for an agent to show, based on its simulated culture.
Finally, one could conclude that (dynamic) Bayesian networks have been successfully employed for some predefined contexts and applications. Especially when considering context, as it is essential in e.g. appraisal emotion models, or in specific applications, DBNs turn out to be a promising approach. In contrast to most other fusion mechanisms their structure may be actively modelled, based on existing theories, so that the structure contains valuable information implicitly, allowing to include existing knowledge in the model. This is especially useful when it is required to make assumptions *why* the model predicted one outcome and not another. It is worth mentioning that context information has only rarely been taken into account - or in most cases, limited to aspects like temporal context in previous research. Yet, in human communication multiple aspects of context [6] continuously influence our behaviours.

2.4 Explainable AI Approaches

The current trend in machine learning tends towards deep learning and neural network architectures that in contrast to Bayesian networks aren't inherently interpretable. Therefore efforts are made to provide explanations for such "black-box" approaches. In general we can distinguish between two kinds of systems providing explanations: model-agnostic or model-specific. Model-agnostic systems are capable of generating explanations independent of the underlying model. Ribeiro et al. introduce in [38] LIME, a model-agnostic approach for the generation of explanations. LIME is able to provide explanations for any given model by approximating an interpretable model around the passed model.
Alber et al. [3] introduced a library named iNNvestigate that provides implementations of common analysis methods for neural networks, e.g. PatternNet and LRP. The generated explanations come in the form of highlighted regions, that have been important for the classification. The supported methods are in contrast to Lime model-specific.
Same goes for SHAP developed by Lundberg et al. [27]. Their framework generates explanations by assigning each feature a value, that describes its importance in regard to the prediction.

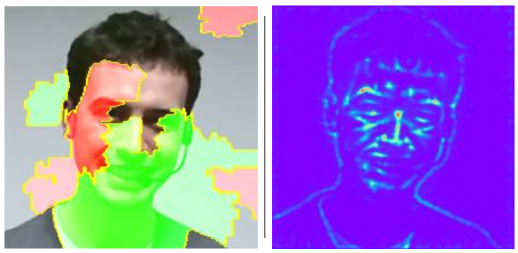

Figure 1. The left image shows an explanation generated with LIME. The right image displays an explanation generated with the iNNvestigate Library using Guided Backpropagation. The neural network to be explained was trained on raw image data from the NoXi corpus (see section 4) to predict different emotions, in this particular case the network predicted happiness as the subject's emotional state.

Figure 1 displays what visual explanations generated by LIME and iNNvestigate could possibly look like. The images have been generated within the scope of the presented work. While such visual explanation systems are of great value in helping to better understand which part of the input data was relevant for a decision, they don't provide causal explanations. The explanation generated by LIME highlights areas that are important for predicting a specific class in green colour, whereas the red coloured shapes describe areas that speak against the predicted class. In the example provided in Figure 1 it is evident that a large part of the face including the smile of the person is important for classifying happiness. However the other half of the face is coloured red and even some areas in the background are coloured green. With this information alone it is not easily comprehensible what the exact reasoning to predict a particular class has been. The explanations generated with iNNvestigate are even harder to correctly interpret. In the provided examples several edges outlining the facial features of the subject are marked being relevant for predicting. Those explanations often leave the user guessing and applying self made causal coherencies to further explain the prediction. Rather these approaches help to get better insight on the decisions of a network on a feature level. A big advantage of Bayesian networks is that the structure of a network can be modelled to have intrinsic meaning. Those causal coherencies might be used as a foundation for generating human-interpretable textual explanations.

3 The Role of Context

In current systems for recognising human behaviours only little attention is given to *context* (e.g. context that is represented by surrounding frames when training a model). Yet there are behaviours that are difficult to analyse and interpret correctly without further information about the *context* of a situation. *Context* is a wide-ranging term that has different meanings depending on the paradigm of research, application and scenario. Duranti et al. [15] noted that it seems impossible to present a single, precise and technical definition of context. Context information might appear as a single impact factor on the interaction or as a combination of multiple types of information. In addition to that, various challenges occur when it comes to context in multimodal communication [50]. In this section we approach different aspects of context:

Temporal context: In classical linguistics, context is "a frame that surrounds the event and provides resources for its appropriate interpretation" [15]. Wöllmer et al. [54] considered context as the temporal surroundings of an observation. In their work they successfully applied bidirectional long-short-term memory (BLSTM) neural networks to consider contextual long-range observations for the prediction of emotions. They further investigated algorithms such as multidimensional dynamic time wrapping (DTW) and asynchronous hidden-markov models to fuse mutual information from multiple modalities, while considering their temporal alignment [53]. An overview on algorithmic approaches, such as dynamic and canonical time wrapping in the context of facial expression analysis is given in [35].
When analysing complex social signals and emotions, the temporal order of behaviours is of vast importance. As an example, Keltner [21] describes a typical time series of behaviours in multiple modalities, that represent a typical instance for the complex emotion "embarrassment" in a social situation - a similar times series of events as we consider here for recognising engagement. Typically, the gaze shifts towards the bottom, the lips make slight

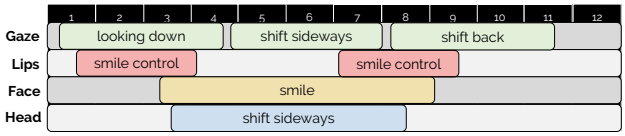

Figure 2. A typical time series of social cues that are performed when a person is feeling "embarrassed"

movements that often turn into a smile followed by the gaze and head shifting to the side and back. Considering such sequences of social signals adds valuable information to the interpretation, compared to the analysis of isolated single cues.

Interaction dynamics context Analysing the dynamics in human communication includes being able to investigate both, the individual multi-modal dynamics (see temporal context) as well as the interpersonal dynamics. Researchers consider interpersonal dynamics on multiple abstractions. For example, Delaherche et al. and Varni et al. [13, 46] consider the synchronicity of people in dyadic interactions on a signal level. Therefore, they developed a set of synchronicity measurements. Rich et al. [40] defined state machines to automatically recognise the four interpersonal cues "mutual gaze", "directed gaze", "adjacency pairs" and "backchannels". In their work they counted the appearance of such bi-directional cues and considered their appearance as an indicator of a person's engagement. Another aspect is the current role in a conversation. Depending on whether the user is in the role of a listener or a speaker, the same kind of behaviour might be interpreted in a completely different way. The influence of the interaction role is illustrated by the following example. Let us assume we observe a person showing a high amount of gestural activity. If the person is in the role of a listener, the observed activity could be interpreted as restlessness. On the opposite, if the person is in the role of a speaker, we might conclude that the person is actively engaged in the conversation. Salam et al. [44] classify multiple aspects of context as parts of the relationship of a social robot and a human during an interaction. More precisely, the interaction context in their definition describes how a scenario relates multiple interlocutors.

Semantic context: The interpretation of detected social cues can be entirely altered through the semantics of accompanying verbal utterances. For example, a laughter in combination with an utterance commenting a negative event would no longer be interpreted as a sign of happiness, but rather be taken as sarcasm. By considering the semantics of accompanying spoken content, detected social cues could be interpreted more accurately. Studies further indicate that humans use semantic context for the interpretation of facial expressions [8, 37, 48].

Environmental context: The location and environmental surroundings may also influence the way we behave during an interaction. As an example, Zimmermann et al. [57] argues that the environmental surroundings directly influence our behaviours e.g. in the way we breathe or speak. In human-computer interaction and especially in ubiquitous computing, a system is called context-aware when it understands the circumstances and conditions surrounding the user. Abowd et al. [1], define context as "any information that can be used to characterise the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves". They further state that context is highly dependable on the current perspective.

Social context: Another aspect of context is the so called "social context". Riek et al. [41] stress the importance of considering social context when creating automated behaviour analysis systems. In their definition, social context is the "environment where a particular person is situated with four factors that may influence (their) behaviour: situational context, cultural context, the person's social role context, and the environmental social norms". Such aspects may be addressed by the following questions: In what kind of situation does the conversation happen? What is the setting of the interaction? (situational context), How well do the interlocutors know each other? Do they share common knowledge? What culture or gender do they have? What is their personality like? (cultural context). How is their relationship? How is their social status? (the person's social role). What are the social norms in the location of the interaction? What are the social norms in the community of the interlocutors? (environmental social norms). Questions like these play an important role, especially when interpreting non-verbal behaviour. Some of these aspects might be difficult to retrieve in an automated manner during the interaction between multiple interlocutors. However, if it is not possible to automatically gather such context information, it could be collected upfront.

When humans interpret behaviours of other people, they consciously or unconsciously include these and similar considerations in their reasoning process. Machines that aim to correctly interpret human behaviours should therefore consider contextual aspects in their interpretation models as well. Yet, besides temporal context (e.g. [54]), only little attention has been put to contextual aspects in current social signal processing research.

4 NoXi Database

The data for the upcoming evaluation tasks has been gathered from the NoXi Database [9]. NoXi provides dyadic novice-expert conversations. One participant took the role of the expert and the other one the role of the novice. Experts were free to chose the topic they wanted to talk about. Furthermore, the novices were evenly free in choosing what to listen to. This resulted in conversations covering a broad scope of different topics ranging from photography to dementia. Both participants were placed in separate rooms during the recording. They interacted remotely through TV screens and microphones. An example for the setup can be seen in Figure 3.

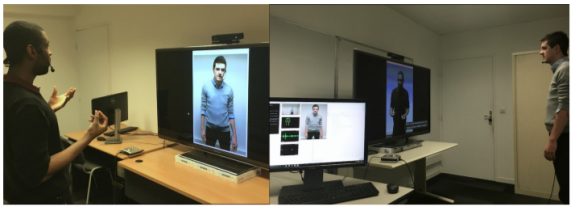

Figure 3. Recording of a novice-expert conversation in the NoXi database [9].

The database covers multiple languages and ethnicities, e.g. English, French, German, Indonesian, Arabic, Spanish, Italian. However, English, German and French have been the languages that occurred the most. A total of 84 sessions have been recorded, providing

25 hours and 18 minutes of conversational data. Additionally, demographic information of the participants have been collected, which include gender, cultural identity, age and level of education. The range of age has been from 21 to 50 years. We decided for the NoXi corpus due to the fact that it contains multi-modal multi-person interaction data and its transferability to social coaching scenarios. Moreover the setup of the corpus allowed for both, engaging, as well as non-engaging interactions.
A total of 19 sessions of the NoXi corpus have been annotated regarding conversational engagement. The annotators followed the engagement definition of Poggi, which we introduced in subsection 2.1. For most of the sessions novice and expert annotations have been created. Of the 19 sessions twelve are associated with French, four with English and three with German. For annotating, a continuous scheme has been chosen. The engagement annotations were created on the ratings of 4-7 different annotators. To measure the quality of the created annotations, from every annotator, they are validated against each other using the Pearson Correlation Coefficient (PCC). Based on the PCC, a gold standard for the annotations has been created. Whenever different annotators have scored a PCC value greater than 0.5 they have been merged to a gold standard annotation. Depending on the definition a value greater than 0.5 is considered a strong uphill (positive) linear relationship. However, at least two annotators have to score higher than 0.5, otherwise no gold standard has been created for the specific session and the session has been discarded. The gold standard itself is calculated by averaging the corresponding annotations.

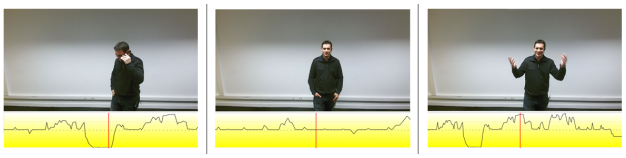

Figure 4. Examples for very low (left), medium (middle) and very high (right) engagement. In addition, the corresponding gold standard annotation is provided.

Figure 4 displays examples for very low, medium and very high engagement, with the corresponding gold standard annotation. The first image has been interpreted by the annotators as very low engagement. This scene occurred, as the novice decided to answer his phone, during the conversation (there were planned interruptions in the NoXi corpus, e.g. by calls from the experimenters or walk-ins). Answering the phone can be considered as a strong signal of the individual not willing to maintain the interaction. The alignment of head and body, away from the interlocutor, go along with a very low level of engagement. The next picture displays a neutral body position of the novice. This behaviour has been associated with a medium level of engagement. He aligned his body towards the other participant and is focusing the TV-screen. The last image represents very high engagement. The novice is smiling and shows a very open body posture, with the arms wide spread using a large gesture space. Again his body is aligned towards his interlocutor.
Engagement comes in various facets and sometimes the determination of its degree is distinct, like the just presented examples for very low and very high engagement. However, sometimes things are less obvious and leave room for a different interpretation. During the continuous annotation of conversational engagement we faced similar problems, as the ones mentioned by Whitehill et al. in [51]. They faced the issue, that an annotator tends to classify the level of engagement in the context of the currently annotated individual. Furthermore, they argue this could lead to annotations that are not comparable between different sessions. In fact, during the process of annotating, we often caught ourselves with statements like, “For their type of character, this should be considered as low/medium/high engagement”. However, we figured out that this causal chain is not wrong. It shows, that the way the level of engagement of an individual is perceived, also depends on the psychological traits the annotator attributes to the individual. Those traits can be considered as context information, which could be modelled inside the Bayesian network.

5 Engagement Model

Based on the evidences presented in subsection 2.1 we developed an annotation scheme that has been used to train our Bayesian networks. We considered different modalities besides context information.

Audio: First of all we considered the general voice activity of the interlocutors as valuable information. Even though it is very basic in its nature it allows to draw a conclusion about the overall involvement of the individuals regarding the conversation. An overall low voice activity may imply a conversation with low engaged interlocutors. On top of that we distinguished between different types of voice activity. We considered speech, filler and silence. The fillers are a particularly interesting type of voice activity as they also cover audio backchannels. In subsection 2.1 we mentioned that backchannels are a very common tool during conversation and provide information about the level of engagement [17].
Further Knapp et al. [23] argue that emotions are reliably transported by the voice. Therefore we trained a support vector machine (SVM) to predict the arousal of the voice [5]. The output of both SVM models (arousal, speech/filler/silence) is used to train the Bayesian network.

Face/Head: During conversations the face usually occupies most of the interlocutors attention. A lot of important information regarding the level of engagement can be extracted from the face respectively the head. Therefore we aimed in our annotation scheme to cover a general impression of the region, as well as looking for specific behaviour that is strongly connected to engagement. We defined features that represent the overall movement of the head in regard to X,Y and Z-Axis. Those features were mainly inspired by the research of Ryota Ooko et al. [33]. They found that a moderate positive correlation of head movement regarding the level of conversational engagement is present. Further we considered the individual gaze behaviour of the participants. There are multiple studies present about the recognition of engagement solely based on gaze data, with good recognition scores [20] [7]. Finally we trained a neural network on the facial action units (FACS) extracted with Openface [4] to predict valence [5]. We used the output of the neural network to train our Bayesian network.

Body: We mentioned earlier in subsection 2.1 that the alignment and movement of the body play an important role in the recognition of engagement. We followed an approach that has been similar to the head features. We tried to cover the general behaviour of the body, as well as specific gestures or poses that are connected to engagement. Therefore we defined a group of features, called body properties. They are mainly inspired by the coding system introduced in [12]. It contains values for the distance between the arms and the hips for X and Z-Axis. Moreover, the alignment of the arms is covered, by calculating the rotation of the elbow joints. Those values are supposed to describe a general level of openness.

Also the distances of each arm to the hip allow interpretation of the symmetry of the arms. In addition to that, the standard deviation of the distance travelled by the head during a frame and the rotation of the head is calculated. Those values have been chosen based on [29] [12].
In subsection 2.1 we identified restlessness to be connected to low levels of engagement. This is the reason we decided to calculate the continuous movement of the interlocutors. Continuous movement is a cumulative value, which describes the overall body movement. Lots of movement may indicate restlessness. In addition to that we wanted to cover the amount of gesticulation an individual performs. Gesticulation is mentioned in [29] and [12] as a crucial nonverbal queue in communication. Therefore we mapped the amount of movement done by both hands onto a real number value, which represents a numeric value for gesticulation.
Furthermore we considered the crossed arms and head touch gestures. The crossing of the arms is a common and often observed gesture. In research it is often interpreted as the expression of a negative emotional attitude by individuals [18] [49]. Based on this we argue that a negative emotional state is bonded to low engagement. In subsection 2.1 we mentioned self touches as a possible signal of being emotionally engaged. Moreover, Gunes et al. [18] were able to achieve good recognition rates for emotions, based on face and body features. Their system associated the emotions of fear, sadness and surprise mostly with gestures of the hands touching the head.

We believe that context plays an important role when it comes to correctly identifying social behaviour. The same applies to recognising conversational engagement. Depending on the context a specific gesture or behaviour may have a different meaning. Recall the example of the very actively moving engaged expert. His continuous movement is not a sign of restlessness. Given the fact that he is talking and gesticulating he should be considered as actively engaged in the conversation. Based on the different types of context we defined in section 3 we considered following context to predict conversational engagement.

Turn hold: During a conversation the interlocutors usually alternate their speaking turns. Therefore we determine the interlocutor that is currently holding the turn. Turn taking and vocal cues play an important part during conversations [23]. This kind of information can be considered as interaction dynamics context.

Role: In the used corpus two roles have been present: novice and expert. The novice has been the one with little to no knowledge about the topic presented by the expert. Accordingly, the expert has been the one introducing and providing information about the topic to the novice. Furthermore, it is in the nature of the expert to be more talkative than the novice, therefore a rather silent expert tends to be in a state of lower engagement, when compared to a similar silent novice, who might be just interestedly listening. In terms of context the information about the role covers multiple aspects. As we just elaborated, most of the time novices and experts operate differently during conversations. Therefore this can be seen as interaction dynamics context. Besides that, the role also covers social context. This is due to the fact, that specific expectations are raised towards the expert. By putting themselves in the role of an expert they signal the novice that they have sophisticated knowledge about their topic. This may result in novices being rather reserved regarding their interactions and comments. Moreover, it is common for the expert to take the lead during the conversation, which automatically results in more speaking time.

Gender: There are differences in the behaviour during conversations depending on the gender of the interlocutors [29]. For example, in same-gender conversation pairs females tend to have more eye contact with each other then males do. Also, males are more prone to decrease eye contact over time, while females have a tendency to increase it [29]. That is only one of many examples where the different genders behave differently. Due to that we think that not only gender itself, but also the constellation of interlocutor pairs, e.g. male-male, male-female, female-female, will be beneficial to the recognition of engagement. By considering the gender we aim to cover another aspect of social context.

Temporal context: In section 3 we argued that the temporal order of behaviours is important when it comes to analysing complex social signals, such as engagement. That means, time series and patterns of behaviours have different meaning when performed differently.

Coming up with a suitable architecture for the Bayesian network has been an incremental approach. This process included systematically adding, removing and exchanging classifiers, because even though specific characteristics for engagement are suggested in the literature, it does not necessarily mean they will work for any given context.

To provide more insight about the actual architecture Figure 5 displays an excerpt of the multi person dynamic Bayesian network. Basically the network is a graphical representation of the just presented annotation scheme. However, a big advantage of Bayesian networks is that the structure has intrinsic meaning compared to other models (e.g. artificial neural networks). This way, we were able to take knowledge about causal coherencies into account. Context nodes such as the gender or role are represented by conditional nodes, so that engagement is predicted "given" the context information, while social cues are "symptoms" shown by the observed person. In other words, social cues can be observed, given that a person has a certain level of engagement. Most of the context information we considered important is focused on a single interlocutor. However we also identified interaction dynamics context as a key element in correctly interpreting conversational engagement. Therefore we chose to model a multi person Bayesian network that also takes the interaction context and the interaction dynamics of the different interlocutors into account when estimating conversational engagement. For the NoXi Database this resulted in a network considering two persons - expert and novice. Moreover, we modelled our network as a dynamic Bayesian network. This way we were able to take temporal context into account.

6 Transparency

Bayesian networks not only allow us to easily model context and other causal coherencies, but also provide transparency by default [52]. In subsection 2.4 we mentioned that machine learning models, in the context of explainable AI, can be distinguished between inherently interpretable models and black-box models. Bayesian networks are inherently interpretable. This is due to the fact that for a given set of variables a Bayesian network is a representation of the joint probability distribution [30]. Usually we want a trained Bayesian network - given a set of observation - to predict what the most likely class of our target node is. In our use case we want to know how engaged one of the interlocutors is. However in a Bayesian network we are not only able to find out how engaged a person is but also what are the most important features for a specific class and what characteristics

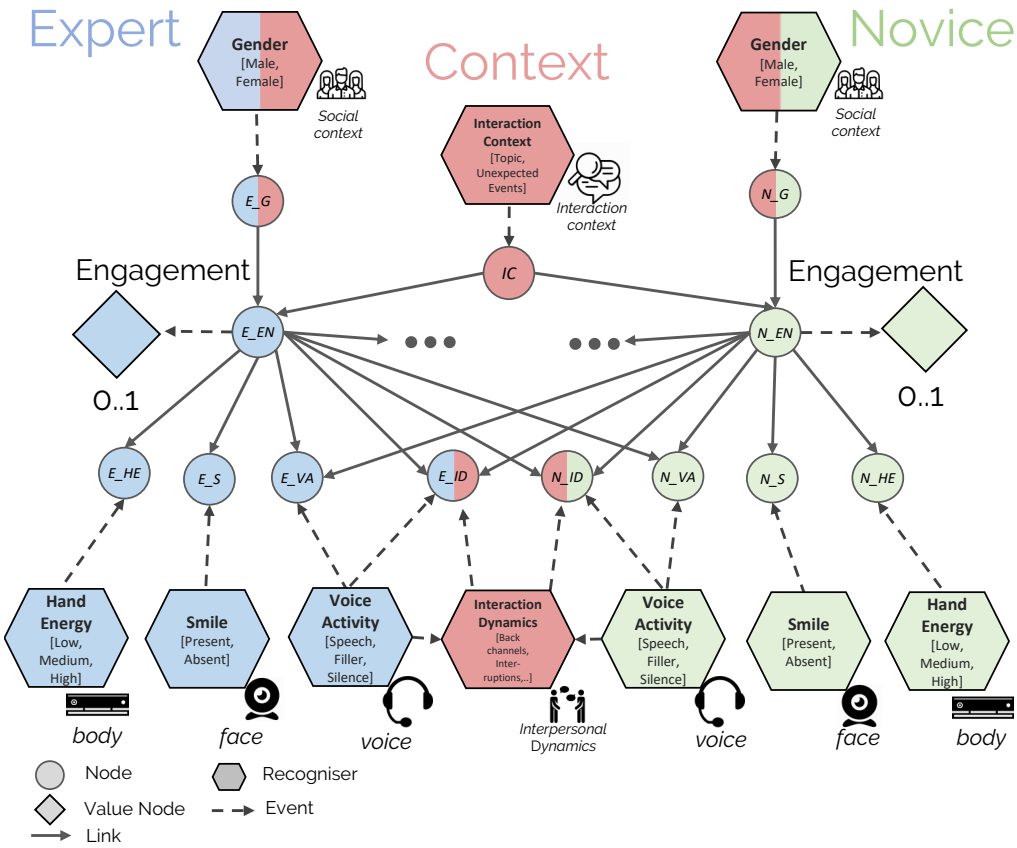

Figure 5. Schematic of a single time slide in a dynamic Bayesian network for two persons.

does the feature have. In Figure 6 a schematic of a reduced Bayesian network for the recognition of engagement is presented. The network contains the features Hand Energy and Voice Activity, which can take the characteristics low, medium and high. Moreover we have our target node Engagement, which also can be low, medium and high. Finally we considered some social context by adding the Role of the interlocutors. The schematic displays the probability distribution of the nodes given the person is highly engaged. This information tells us that when a person is highly engaged they are most likely in the role of the expert (70%) and show most likely high levels of Hand Energy and Voice Activity. We could now apply the same approach to find out more about low and medium engagement and get extensive insight about the learnt representations of our network.

7 Evaluation

Even though transparency is important in the context of machine learning, there is little use for a transparent model that isn't able to accurately predict the task at hand. That is why we investigate in the following the performance of the introduced architectures compared to other state-of-the-art machine learning approaches.

We split the acquired data into dedicated sets for training and evaluation. The training set included 13 sessions and had a size of 616374 samples. The evaluation set consisted out of six sessions, with a total of 328385 samples. So we ended up with the evaluation set having roughly half the samples of the training set.

To evaluate the different models, the Pearson correlation coefficient has been calculated between the model's prediction and the gold standard annotation.

[ht]

Table 1. Average PCCs on multimodal inputs

Method	Modalities	PCC
LSVM	Face, Body, Voice	.6253
Keras RNN	Face, Body, Voice	.6034
BN	Face, Body, Voice, Context	.7373
DBN (10 timesteps)	Face, Body, Voice, Context	.7443
MDBN (10 timesteps)	Face, Body, Voice, Context	.7680

As described earlier, developing a suitable Bayesian network has been an incremental approach by adjusting the classifier composition. An early Bayesian network (BN) based on multiple modalities including some context information achieved promising results with a PCC of 0.7373. By extending this network with temporal context for selected nodes that are related to body and face movement as well as voice activity we were able to further improve the correlation score to 0.7443. During our tests the network that performed best has been a multi-person dynamic Bayesian network (MDBN). It incorporates interpersonal dynamics, like mutual gaze and turn transitions between the novice and expert. The network achieved a PCC of 0.768 which is significantly better ($p < 0.001$) than the best single-user DBN (0.7443).

The (D)BNs we applied are created using a hybrid approach where classification results for sub-recognition tasks, as well as threshold

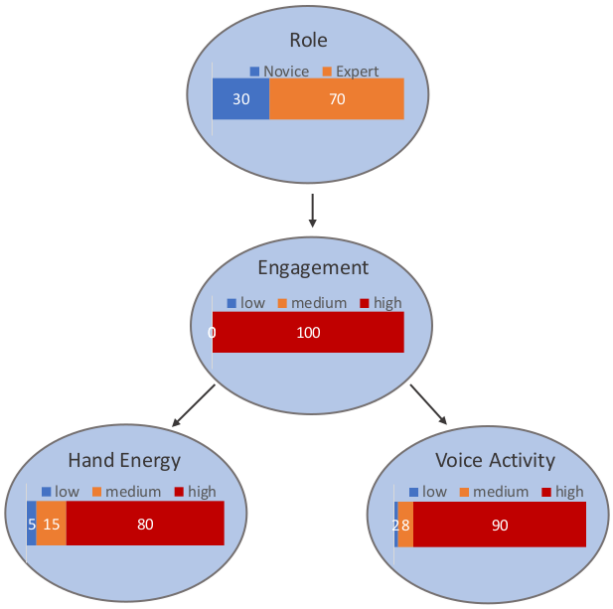

Figure 6. Schematic of a simplified Bayesian network displaying the probability distribution given the observation of high engagement.

based features are used to update the evidences in the network. This makes it difficult to compare the multi-modal model with other classification models that rely on low level features. In order to have a baseline to evaluate our approach, we created an engagement feature set that is heavily influenced by the previously introduced engagement annotation scheme. It contains features on body movement, body posture, head movement, facial expression and audio. We trained a linear support vector machine (LSVM) on this feature set and achieved a PCC of 0.6253. Moreover, we tested several neural networks implemented in Keras. The best one has been a fully connected deep recurrent neural network (RNN) and was able to score a PCC of 0.6034 on the engagement feature set. Those results are significantly ($p < 0.001$) worse than our introduced hybrid model.

8 Discussion

We were able to show that our hybrid approach using a theory-modelled DBN can deliver comparable results to purely statistical black-box approaches. This is in compliance with the research of Rudin [42]. On our corpus it even slightly outperformed the other classification methods. With the introduction of a multi person dynamic Bayesian network architecture we were able to further increase the prediction accuracy. We explain this with several aspects: by employing the transparent DBN we could intuitively refine our first assumptions on what influences engagement, which allowed us to incrementally add classifiers, until the network achieved satisfying correlations with our gold standard annotation. Further, through the update mechanism on annotation/event abstraction we aimed to simulate a decision making and reasoning process that's similar to the one of humans. To our understanding, humans will consciously or unconsciously map abstractions of behaviours (e.g. smiles) on their perception of the other person (e.g. happiness). Further, we conclude that for our particular use-case of recognising conversational engagement, considering different types of context information leads to improvements in terms of the correct and adequate interpretation. In fact the more context information we added the better our model performed.

9 Conclusion

Deep learning can be considered as the current gold standard in machine learning. Deep neural networks proved themselves on various problem domains by performing exceptionally well [24] [10] [2]. However their biggest weakness is their lack of interpretability. That is why efforts are made to provide additional insight to otherwise "black-boxes" (see subsection 2.4). Even though there are approaches present that help in gaining additional insight on the decision-making of neural network architectures, they rather provide additional information on a feature-level basis. In contrast to that there are models, like Bayesian networks that are inherently interpreteable and can be modelled to have intrinsic meaning. This enables a user to gather causal coherencies on why a model made a specific prediction. Often this seems to come down to a trade-off between prediction performance and transparency. However, we showed for the use case of multi-modal engagement recognition that by applying a hybrid approach that fuses abstractions of multiple social cues in a causal recognition model, accuracy and transparency do not necessarily need to exclude each other. Moreover we were able to improve the recognition rates of our model by incorporating social, temporal and interaction dynamics context. The significant impact of context on recognition scores stresses the importance of context in correctly and adequately interpreting conversational engagement. The proposed system has been implemented within the SSI Framework [47], so that all social cue classification models, as well as the overall BN inference step can be performed in a real-time system. This allows to apply this approach in a variety of applications, such as human-agent or human-robot scenarios.

ACKNOWLEDGEMENTS

This work has received funding from the DFG under project number 392401413, DEEP.
Further this work presents and discusses results in the context of the research project ForDigitHealth. The project is part of the Bavarian Research Association on Healthy Use of Digital Technologies and Media (ForDigitHealth), funded by the Bavarian Ministry of Science and Arts.

REFERENCES

[1] Gregory D. Abowd, Anind K. Dey, Peter J. Brown, Nigel Davies, Mark Smith, and Pete Steggles, 'Towards a better understanding of context and context-awareness', in *Handheld and Ubiquitous Computing, First International Symposium, HUC'99, Karlsruhe, Germany, September 27-29, 1999, Proceedings*, ed., Hans-Werner Gellersen, volume 1707 of *Lecture Notes in Computer Science*, pp. 304–307. Springer, (1999).

[2] Igor Aizenberg and Gonzalez Alexander, 'Image recognition using mlmvn and frequency domain features', *International Joint Conference on Neural Networks*, (2018).

[3] Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans, 'innvestigate neural networks!', *CoRR*, **abs/1808.04260**, (2018).

[4] T. Baltrušaitis, P. Robinson, and L. P. Morency, 'Openface: An open source facial behavior analysis toolkit', in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, (March 2016).

[5] S. Basu, N. Jana, A. Bag, Mahadevappa M, J. Mukherjee, S. Kumar, and R. Guha, 'Emotion recognition based on physiological signals using valence-arousal model', in *2015 Third International Conference on Image Information Processing (ICIIP)*, pp. 50–55, (2015).

[6] Tobias Baur, Dominik Schiller, and Elisabeth André, 'Modeling user's social attitude in a conversational system', in *Emotions and Personality in Personalized Services*, 181–199, Springer, (2016).

[7] Roman Bednarik, Shahram Eivazi, and Michal Hradis, 'Gaze and conversational engagement in multiparty video conversation: An annotation scheme and classification of high and low levels of engagement', in *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*, Gaze-In '12, pp. 10:1–10:6, New York, NY, USA, (2012). ACM.

[8] Vicki Bruce and Andy Young, *In the eye of the beholder: the science of face perception.*, Oxford University Press, 1998.

[9] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar, 'The noxi database: Multimodal recordings of mediated novice-expert interactions', *ICMI'17*, (November 2017).

[10] Dan Cireşan, Ueli Meier, and Jürgen Schmidhuber, 'Multi-column deep neural networks for image classification', *arXiv preprint arXiv:1202.2745*, (2012).

[11] Cristina Conati and Heather Maclaren, 'Modeling user affect from causes and effects', in *User Modeling, Adaptation, and Personalization, 17th International Conference, UMAP 2009, formerly UM and AH, Trento, Italy, June 22-26, 2009. Proceedings*, pp. 4–15, (2009).

[12] Nele Dael, Marcello Mortillaro, and Klaus R. Scherer, 'The body action and posture coding system (bap): Development and reliability', *Journal of Nonverbal Behavior*, **36**(2), 97–121, (Jun 2012).

[13] Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen, 'Interpersonal synchrony: A survey of evaluation methods across disciplines', *IEEE Trans. Affective Computing*, **3**(3), 349–365, (2012).

[14] Sidney S D'Mello, Patrick Chipman, and Art Graesser, 'Posture as a predictor of learner's affective engagement', in *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29, (2007).

[15] Alessandro Duranti and Charles Goodwin, *Rethinking context: Language as an interactive phenomenon*, number 11 in Studies in the Social and Cultural Foundations of Language, Cambridge University Press, 1992.

[16] Stephen R Garner et al., 'Weka: The waikato environment for knowledge analysis', in *Proceedings of the New Zealand computer science research students conference*, pp. 57–64, (1995).

[17] N. Glas and C. Pelachaud, 'Definitions of engagement in human-agent interaction', in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 944–949, (Sept 2015).

[18] Hatice Gunes and Massimo Piccardi, 'Affect recognition from face and body: early fusion vs. late fusion', in *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, volume 4, pp. 3437–3443. IEEE, (2005).

[19] Sepp Hochreiter and Jürgen Schmidhuber, 'Long short-term memory', *Neural Computation*, **9**(8), 1735–1780, (1997).

[20] Ryo Ishii and Yukiko I. Nakano, 'An empirical study of eye-gaze behaviors: Towards the estimation of conversational engagement in human-agent communication', in *Proceedings of the 2010 Workshop on Eye Gaze in Intelligent Human Machine Interaction*, EGIHMI '10, pp. 33–40, New York, NY, USA, (2010). ACM.

[21] Dacher Keltner, 'Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame', *Journal of personality and social psychology*, **68**(3), 441, (1995).

[22] Mardi Kidwell, 'Framing, grounding, and coordinating conversational interaction: Posture, gaze, facial expression, and movement in space', in *Body - Language - Communication. An International Handbook on Multimodality in Human Interaction*, 100 – 113, De Gruyter Mouton, (2013).

[23] Mark Knapp, L. and Judith Hall, A., *Nonverbal Communication in Human Interaction*, Harcourt Brace, 1997.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, 'Imagenet classification with deep convolutional neural networks', in *Advances in Neural Information Processing Systems 25*, eds., F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, 1097–1105, Curran Associates, Inc., (2012).

[25] Hedda Lausberg, 'Neuropsychology of gesture production', in *Body - Language - Communication. An International Handbook on Multimodality in Human Interaction*, 168 – 182, De Gruyter Mouton, (2013).

[26] Birgit Lugrin, Julian Frommel, and Elisabeth André, 'Combining a data-driven and a theory-based approach to generate culture-dependent behaviours for virtual characters', in *Advances in Culturally-Aware Intelligent Systems and in Cross-Cultural Psychological Studies*, 111–142, Springer, (2018).

[27] Scott M Lundberg and Su-In Lee, 'A unified approach to interpreting model predictions', in *Advances in Neural Information Processing Systems 30*, eds., I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4765–4774, Curran Associates, Inc., (2017).

[28] Marwa Mahmoud and Peter Robinson, 'Interpreting hand-over-face gestures', in *Affective Computing and Intelligent Interaction*, 248–255, Springer, (2011).

[29] Albert Mehrabian, *Nonverbal Communication*, AldineTransaction, 2007.

[30] Michael Mitchell, Tom, *Machine Learning*, 177–197, MacGraw-Hill, 1997.

[31] Cornelia Müller, Alan Cienki, Ellen Fricke, Silva Ladewig, David McNeill, and Sedinha Tessendorf, *Body - Language - Communication. An International Handbook on Multimodality in Human Interaction*, De Gruyter Mouton, 2013.

[32] Kevin Patrick Murphy and Stuart Russell, 'Dynamic bayesian networks: representation, inference and learning', *Ph.D Thesis*, (2002).

[33] Ryota Ooko, Ryo Ishii, and Yukiko I. Nakano, 'Estimating a user's conversational engagement based on head pose information', in *Intelligent Virtual Agents*, eds., Hannes Högni Vilhjálmsson, Stefan Kopp, Stacy Marsella, and Kristinn R. Thórisson, pp. 262–268, Berlin, Heidelberg, (2011). Springer Berlin Heidelberg.

[34] Andrew Ortony, Gerald L Clore, and Allan Collins, *The cognitive structure of emotions*, Cambridge university press, 1990.

[35] Yiannis Panagakis, Ognjen Rudovic, and Maja Pantic, 'Learning for multi-modal and context-sensitive interfaces', *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*, **2**, in press, (2018).

[36] Isabella Poggi, *Mind, hands, face and body: a goal and belief view of multimodal communication*, Weidler, 2007.

[37] Carl Ratner, 'Back to dr. ratner's home page journal of mind and behavior, 1989, 10, 211-230 a social constructionist critique of naturalistic theories of emotion', *Journal of Mind and Behavior*, **10**, 211–230, (1989).

[38] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, '"why should I trust you?": Explaining the predictions of any classifier', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144, (2016).

[39] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, 'Recognizing engagement in human-robot interaction', in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 375–382, (March 2010).

[40] Charles Rich, Brett Ponsleur, Aaron Holroyd, and Candace L. Sidner, 'Recognizing engagement in human-robot interaction', in *Proceedings of the 5th ACM/IEEE International Conference on Human Robot Interaction, HRI 2010, Osaka, Japan, March 2-5, 2010*, eds., Pamela J. Hinds, Hiroshi Ishiguro, Takayuki Kanda, and Peter H. Kahn Jr., pp. 375–382. ACM, (2010).

[41] Laurel D. Riek and Peter Robinson, 'Challenges and opportunities in building socially intelligent machines [social sciences]', *IEEE Signal Process. Mag.*, **28**(3), 146–149, (2011).

[42] Cynthia Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence*, **1**(5), 206–215, (2019).

[43] Jennifer Sabourin, Bradford W. Mott, and James C. Lester, 'Modeling learner affect with theoretically grounded dynamic bayesian networks', in *Affective Computing and Intelligent Interaction - 4th International Conference, ACII 2011, Memphis, TN, USA, October 9-12, 2011, Proceedings, Part I*, pp. 286–295, (2011).

[44] Hanan Salam and Mohamed Chetouani, 'A multi-level context-based modeling of engagement in human-robot interaction', in *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015, Ljubljana, Slovenia, May 4-8, 2015*, pp. 1–6. IEEE Computer Society, (2015).

[45] 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', in *Proceedings of the 6th International Conference on Human-robot Interaction*, HRI '11, pp. 305–312, New York, NY, USA, (2011). ACM.

[46] Giovanna Varni, Marie Avril, Adem Usta, and Mohamed Chetouani, 'Syncpy: a unified open-source analytic library for synchrony', in *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence, INTERPERSONAL@ICMI 2015, Seattle, Washington, USA, November 13, 2015*, pp. 41–47, (2015).

[47] Johannes Wagner, Florian Lingenfelser, Tobias Baur, Ionut Damian, Felix Kistler, and Elisabeth André, 'The social signal interpretation (ssi) framework: Multimodal signal processing and recognition in real-time', in *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, p. 831–834, New York, NY, USA, (2013). Association for Computing Machinery.

[48] Harald G Wallbott, 'In and out of context: Influences of facial expression and context information on emotion attributions', *British Journal of Social Psychology*, **27**(4), 357–369, (1988).

[49] Harald G Wallbott, 'Bodily expression of emotion', *European journal of social psychology*, **28**(6), 879–896, (1998).

[50] Rebekah Wegener, *Studying Language in Society and Society through Language: Context and Multimodal Communication*, 227–248, Palgrave Macmillan UK, London, 2016.

[51] 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).

[52] Wim Wiegerinck, Willem Burgers, and Bert Kappen, *Bayesian Networks, Introduction and Practical Applications*, 401–431, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[53] Martin Wöllmer, Marc Al-Hames, Florian Eyben, Björn W. Schuller, and Gerhard Rigoll, 'A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams', *Neurocomputing*, **73**(1-3), 366–380, (2009).

[54] Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn W. Schuller, and Shrikanth S. Narayanan, 'Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling', in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pp. 2362–2365, (2010).

[55] Martin Wöllmer, Björn W. Schuller, Florian Eyben, and Gerhard Rigoll, 'Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening', *J. Sel. Topics Signal Processing*, **4**(5), 867–881, (2010).

[56] W. Yun, D. Lee, C. Park, J. Kim, and J. Kim, 'Automatic recognition of children engagement from facial video using convolutional neural networks', *IEEE Transactions on Affective Computing*, 1–1, (2018).

[57] Heinz Zimmermann, *Speaking, listening, understanding*, SteinerBooks, 1996.