Chapter 4

Engagement Perception and Generation for Social Robots and Virtual Agents

Lee J. Corrigan[1], Christopher Peters[2], Dennis Küster[3] and Ginevra Castellano[4]

Abstract

Technology is the future, woven into every aspect of our lives, but how are we to interact with all this technology and what happens when problems arise? Artificial agents, such as virtual characters and social robots could offer a realistic solution to help facilitate interactions between humans and machines—if only these agents were better equipped and more informed to hold up their end of an interaction. People and machines can interact to do things together, but in order to get the most out of every interaction, the agent must to be able to make reasonable judgements regarding your intent and goals for the interaction. We explore the concept of engagement from the different perspectives of the human and the agent. More specifically, we study how the agent perceives the engagement state of the other interactant, and how it generates its own representation of engaging behaviour. In this chapter, we discuss the different stages and components of engagement that have been suggested in the literature from the applied perspective of a case study of engagement for social robotics, as well as in the context of another study that was focused on gaze-related engagement with virtual characters.

4.1 Introduction

Human interactions are a product of millions of years of evolution, and as such they are typically smooth and effortlessly coordinated, benefiting massively from the fact that both interactants are able to draw upon a multitude of verbal and/or non-verbal cues in ways that help to regulate the interaction. If we are to interact with machines, in the way that the future depicts i.e., with robotic tutors, interactive display points, operator free terminals and the like, then we need to develop machines that can interact with us in a similarly intuitive fashion. However, at present, the way in which we interact with machines is strongly dictated by their design, which is often not optimal in terms of user experience, especially in cases where issues arise. Hence, our interest in building artificial agents, such as virtual characters and social robots with the ability to maintain interactions across a spectrum of task-orientated use cases has a strong applied perspective. Overall, there is growing interest in the engagement concept throughout the human-machine-interaction (HMI) and related fields, but what is engagement and why is it so important? In this chapter we hope to answer this question by unravelling this complex phenomenon, providing both new and existing HMI researchers with an firm underpinning of engagement related theory and concepts.

The remainder of this chapter is organised as follows: In the next section, we provide the reader with some general theory of the various stages and components of engagement. Additionally, we detail related concepts, such as the perception and generation of engagement related behaviours, and novel experimental considerations. In Sect. 4.3, we present two case studies: one of which looks at the perception of engagement

[1] School of Electronic, Electrical and Systems Engineering, University of Birmingham, Birmingham, UK, e-mail: ljc228@bham.ac.uk
[2] Royal Institute of Technology (KTH), Stockholm, Sweden, e-mail: chpeters@kth.se
[3] Jacobs University Bremen, Bremen, Germany, e-mail: d.kuester@jacobs-university.de
[4] Department of Information Technology, Uppsala University, Uppsala, Sweden, e-mail: ginevra.castellano@it.uu.se

for social robotics and another which considers the perception and generation of engagement related behaviours for virtual agents via gaze.

4.2 Theory

4.2.1 Fundamentals

When consulting a dictionary in the English language, the term engagement appears to be used in at least two different ways—as the starting or intention to start, referring to an initiation of contact, and again in the longer term sense, referring to engagement as something that is more involved. In the literature, engagement is defined in a number of ways: as a process; as a stage in a process, or the overall process; as an experience; as a cognitive state of mind; an empathic connection; or as a perceived or theorised indicator describing the overall state of an interaction. Nevertheless, there are two underlying fundamentals that are apparent across most engagement related studies; the existence of various stages and components of engagement. In this section, we discuss each of these in turn.

4.2.1.1 Stages of Engagement

Engagement as a process can be analysed in terms of a number of discrete stages or phases. These may relate to the intensity or degree of involvement of a user with respect to the object or entity of engagement. For example, in a study of engagement with robots, Sidner and Dzikovska [29] refer to engagement as "a process by which individuals in an interaction start, maintain and end their perceived connection to one another". Most often, these stages are considered independently. For example, recognising the desire to *start* an interaction requires the system to detect an intention to engage, e.g., by tracking passers-by to ascertain if there are certain indicators which might suggest an initial interest to become involved with the system [22]. Whereas, to *maintain* an interaction suggests that the intention has been established and that the system must now adapt to the individual user in such a way that it keeps that user engaged for the term of the interaction. Failure to do so may cause the user to *end* the interaction before the system has achieved its purpose, for example to teach, inform or otherwise assist the user. To this extent, the system should be equipped to do both: detect when a user has irrecoverably ended an interaction, e.g., by getting up and walking away, whilst also being able to use appropriate behaviour to end an interaction once either party has achieved their purpose for becoming involved in the interaction.

It is a natural starting point to consider engagement as consisting of at least three broad stages, i.e., intention to engage, engaged and disengaged. However, O'Brien and Toms [17] refer to a fourth possible stage: re-engagement. The concept of reengagement raises the important issue of when an interaction can be considered as complete. If either party is yet to achieve their purpose, but the user is showing signs of becoming disengaged, the system might try to utilise any information that is available, e.g., from current and previous interactions, in order to "understand" the underlying cause of the disengagement and then attempt to re-engage the user with a series of predefined strategies. However, in many cases, disengagement may be difficult to determine with certainty. For example, if the user looks away briefly, it may just mean that she has been temporarily distracted. In certain cases, looking away may in fact even signal engagement, such as during shared attention, when looking at an object under mutual consideration [21].

4.2.1.2 Components of Engagement

While engagement is frequently operationalized by means of measures of visual attention, it is important to distinguish conceptually between engagement and attention. Engagement is a complex phenomenon, a construct consisting of both cognitive (attention, concentration) and affective components (enjoyment) [17, 27].

Attention: As the cognitive component of engagement, attention is often characterised as a global on/off activity, whereas concentration is the ability to pay selective attention to one thing in particular, while ignoring others. For example, a user paying attention to a particular activity or object for a significant amount of time is concentrating. In our work, it is this form of selective attention (relating to concentration) that we are

interested in, and in going forward we refer to this as just attention. This is therefore conceptually distinct from a global measure of wakefulness, or arousal, although the precise focus of selective attention may sometimes be more narrow, and sometimes be wider. Selective attention to a stimulus is a necessary component in most definitions in order for basic forms of engagement to occur. A more sustained form of attention provides a more elaborate requirement for engagement and also allows the possibility of affective involvement [25].

Attention: As the cognitive component of engagement, attention is often characterised as a global on/off activity, whereas concentration is the ability to pay selective attention to one thing in particular, while ignoring others. For example, a user paying attention to a particular activity or object for a significant amount of time is concentrating. In our work, it is this form of selective attention (relating to concentration) that we are interested in, and in going forward we refer to this as just attention. This is therefore conceptually distinct from a global measure of wakefulness, or arousal, although the precise focus of selective attention may sometimes be more narrow, and sometimes be wider. Selective attention to a stimulus is a necessary component in most definitions in order for basic forms of engagement to occur. A more sustained form of attention provides a more elaborate requirement for engagement and also allows the possibility of affective involvement [25].

Another important factor of engagement is considering exactly what it is that a user is engaged with (i.e., the focus of engagement). This can generally only be inferred from the context, particularly for more sophisticated forms of engagement where there may be more than one potential focus of engagement. Gaze can signal attention [26], however, gazing at a particular object is not always indicative of attention. For example, the fact that a user is in the vicinity of a screen or is looking at one does not mean that they are paying attention to it (they may be day-dreaming for example), or that they are paying attention to those aspects that would be the most important ones from the perspective of the experimenter. In fact, even looking away from a screen does not allow the inverse inference that the subject has completely disengaged. Thus, while there is a certain probability that this is the case, looking away from the screen might simply indicate a moment in the interaction during which the user requires additional resources to process what was being said or presented. One way to improve confidence in assessing the attentional component of engagement in this situation is to consider only attention towards currently relevant aspects of the scene, in terms of gaze and other forms of attention related involvement and interaction. For example, in [20] during interaction with a virtual character, three qualities of engagement are defined, relating to the user (1) not looking at the screen at all, (2) looking at irrelevant aspects of the scene, and (3) looking at relevant aspects of the scene with respect to the ongoing interaction.

Enjoyment: As the affective component of engagement, enjoyment also plays a direct role in an interaction. For example, both positive and negative affect has been shown to influence student performance, motivation and effort [4]. More specifically, in terms of object focus, positive emotions such as enjoyment can increase the availability of cognitive resources, having a positive influence on the user's motivation, ability to utilise flexible learning strategies and self-regulation [18]. Positive affect also increases general motivation, which leads people to try harder in tasks, especially where they feel their effort will make a difference [11]. However, these emotions may not always be outwardly expressed, for example, a user is highly unlikely to smile or laugh throughout an interaction, nonetheless, enjoyment is an important component of the engagement construct. Here, enjoyment is most likely to be expressed indirectly, by continuing the interaction with a strong commitment to achieving certain goals. In gaming, for example, players who are immersed in a game tend to make very few facial expressions, but are nonetheless still very much enjoying the interaction. Here the effort afforded to the interaction could be associated with the positive affect (enjoyment), likewise the inverse could also be true, a lack of afforded effort could be associated with negative affect (boredom) [4].

4.2.2 Concepts

So far, in this chapter, we have only discussed engagement in terms of perception. However, if machines are to interact with humans in a natural and intuitive manner, it is not sufficient to focus entirely on one side of an interaction. Rather, we should consider engagement as a communicative process; a sender-receiver loop; to

both perceive and generate engagement-related cues and signals. Here, we discuss the concept of engagement in terms of both perception and generation.

4.2.2.1 Perception

Perception refers to the use of the term engagement as it relates to the decoding of basic cues from another interactant, by a person or by a machine, for example by using computer vision techniques. Of general importance to our sense of engagement with others is our perception of their attention [10], which can be altered by factors such as the effect of distance between interactants on the salience of visual cues and the context of the situation, and by their enjoyment or at least our perception of their interest. Importantly, both cognitive and affective components of engagement can be measured (with a certain probability) on the basis of certain objective indicators and physiological measures. While this is not a one-to-one mapping between indicators and engagement, this allows a certain degree of automatic measurement of engagement that can be expected to become more reliable with the development of new sensors and algorithms. Enjoyment (pleasure) is the affective component of engagement that can be measured on the basis of several potential indicators, such as eyebrow activity (reverse sign) in combination with smiling. Here, eyebrow activity weighs more than smiling, and a moderately negative weight is added for lip-pressing and lip-tightening. Eyebrow movements can also be a predictor for concentration, obstacles and negative valence. Therefore, frowning may indicate effortful processing suggesting high levels of cognitive engagement which are likely to be associated with negative valence (depending on context and intensity). Smiling is expected to be a weak predictor of positive valence, but it might be effective for short-term social responsiveness [11]. Mouth movements, such as lip pressing and tightening can also be associated with task-related attention and concentration.

Furthermore, important non-verbal cues can be obtained based on head direction and gaze [1], blinking, eyebrow movement, posture and posture shifts [14], smiles [3], and engagement gestures [29]. These low-level signals can in some cases be interpreted as direct measures relating to engagement. However, typically, these individual measures become more informative when they are interpreted with respect to specific events in the task, or when they occur in a synchronized fashion with other indicators rather than individually.

4.2.2.2 Generation

The generation of cues and signals (i.e., their “encoding”) requires at least an equal amount of attention. For example, during face-to-face interaction, the face generates a wealth of cues and signals that goes well beyond speech and facial expressions. It may be expected that we naturally pay attention to the face if we are engaged with that person, and may also display feedback such as nods to display our interest and/or show empathy by conducting appropriate facial expressions. In this respect, they might signal engagement, for example, by attending to the other and showing interest in what they say. An important distinction here, is whether such signals are based on a genuine interest or are superficial displays with the implicit or explicit purpose of communicating to the other that one is engaged. One may display signals of interest for a variety of superficial reasons, related to the accomplishment of high-level or abstract goals. Sometimes the display of interest is more important than the actual motivation [8]. In our previous work, an analysis of data extracted from explicit probes [5] and post-experiment questionnaires suggests that one’s own perception of a robot, in terms of helpfulness, friendliness and attentiveness can help to maintain a type of engagement which lasts throughout an interaction [6]. An artificial agent capable of generating, or at least mimicking, certain engagement related behaviours could help to facilitate an intuitive interaction between humans and machines, giving the human the impression that the machine is intelligent enough to warrant further interaction.

4.2.3 Experimental Considerations

Engagement is often reduced to selective visual attention, perhaps as a practical consequence of a limited availability of measures within a given paradigm. However, in order to discuss engagement as a meaningful construct, we argue that the affective components have to be considered as well. Despite this, in practice, this can be difficult to achieve. As part of the work described in Sect. 4.3.1.2, we found the concept of annotating for the entire engagement construct to be extremely complex. In fact, we found ourselves asking “how should

one annotate for both attention and enjoyment at the same time?". For this exact reason, we have started to explore engagement in a de-constructed format, considering engagement-related components individually.

4.2.3.1 Decomposition of the Engagement Construct

Attention and enjoyment are descriptive states in their own right. It is this decomposition of the engagement construct into cognitive and affective components that we believe can account for, and thus allow for, the fact that high engagement can, and often will, represent rather different socio-emotional states during an interaction. For example, a user may show evidence of intense cognitive engagement with a task, while the affective component might be anywhere between highly positive and highly negative. In the immediate situation, the assessment of attention and cognitive engagement with the task may initially be sufficient since it may not appear to matter how much a user is enjoying a task, as long as she/he continues to work hard to solve it. However, in order to anticipate eventual frustration and a high probability of disengagement in one of the subsequent tasks, it is essential that an engagement detector attempts to track also this affective component in order to facilitate appropriate and early interventions by the system.

4.2.3.2 Implicit Probes

During complex interactions, feeding sensor data directly into a computational model might not always be able to provide a accurate measure of the user's state of engagement. For example, in an educational interaction involving a robotic tutor, the system may need to understand why the user is showing signs of is engagement—is it that the task is too difficult, too easy, or is it because the user is simply discounting the advice provided by the robot? In this situation, the system could use a probe to answer some of these questions, i.e., by evaluating certain elements of the interaction.

So, what is a probe? A probe is a non-intrusive, pervasive method of extracting additional supporting features from within the interaction itself, providing highly standardised moments for analysis. Probes are pervasive in the sense that they can be integrated into any stage of an interaction [5]. They do not require the collection of any special additional types of data beyond the measures already stated in the consent forms. Rather, the probes define standardized situations that are naturally embedded into the flow of the task in such a way that they appear to the subject as a completely normal part of the interaction. Their standardization allows the formulation of substantially more meaningful predictions of behaviours within one experiment as well as between experiments and potentially even across different experimental paradigms. In this sense, they could also be described as modular building blocks that can be reused and which remain comparable even when other parts of the interaction require more flexibility. For example, experimenters in laboratory experiments, or doctors with a lot of experience in interviewing patients, will often use a similar approach using highly schematic questions and small talk in order to get a first sense of the participant or patient.
Furthermore, as probes only describe relatively short schematic modules with few degrees of freedom, they do not impede upon the natural flow of the interaction as opposed to, e.g., experimental designs that aim to obtain full control throughout the entirety of an experiment. For this reason, we argue that probes may be an ideal solution when needs for high levels of experimental control and analyses have to be balanced with maintaining a natural flow. Engagement, in this context, is a particularly relevant example because any measurement of engagement has to avoid disrupting the user engagement itself. The features extracted from these probes are embedded within the context of the main task, and are designed to provide the most accurate possible assessment of the user's engagement state. More specifically, we distinguish between two different types of probes: the social probe and the social task probe. This is an additional step beyond the low-level continuous observation already employed elsewhere in HMI and related fields.

The design of a probe causes the user to respond in a certain way and it is that response which is then used to fortify the system's confidence that a user is in a particular state. For example, if the agent is unsure of the user's engagement state because confidence levels are low and social interaction hasn't occurred recently enough to make any inferences, then the agent can trigger a probe by attempting to socially engage the user in a one-to-one interaction. If the user stops what they are doing and responds to the agent's attempt, then we can increase the value associated with social engagement and also increase the overall confidence. Other metrics relating to immediacy, responsiveness and whether or not the user maintains their attention to the social interaction will further affect those values.

Social Probes: The social probe involves a simple, standardised piece of interaction between the agent and the user. Its purpose is to provide a standardised moment in which we can gauge how socially receptive the user is to the agent. To illustrate, in Fig. 4.1, we provide a time-line example of a social probe. The first three seconds are used to attract the attention of the user and the following two segments, lasting five seconds and three and a half seconds respectively, are used for analysis. As an example of this, if the user maintains gaze toward the agent across both maintainer segments, he/she is deemed as showing signs of high attention. We can also use this highly standardised piece of interaction to detect other non-verbal behaviours, such as smiles and facial expressions, including their temporal location within the probe.

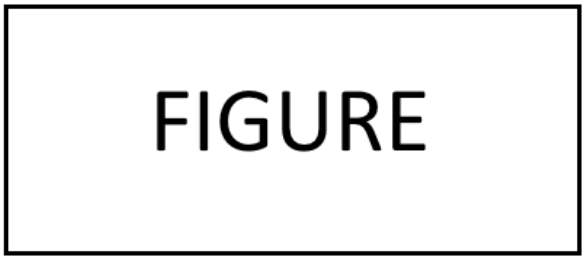

Fig. 4.1 An example of an interaction time-line for a social probe. Timings relate to the amount of interaction allocated to the attractor and two maintainer segments

Actual implemented examples of the content used in this particular type of probe are provided in Sect. 4.3.1.2.
Social-Task Probes: The social-task probe is concerned with the collaborative aspect of an interaction involving an agent and a user, such as the teacher-student or master apprentice relationship. With this type of probe we can measure how receptive the user is to the agent's suggestions and assistance, e.g., by directing gaze toward specific items, or encouraging specific actions. From this we can also measure how reliant or independent the user is on assistance from the agent. In other words, social-task probes are designed to measure aspects of engagement in social-task interactions.

4.3 Practice

4.3.1 Case Study 1: Engagement in Social Robotics

To explore the engagement concept in a task-orientated scenario, we conducted a Wizard-of-Oz (WoZ) style data collection study using a robot. The study was carried out in the classroom environment of an English secondary school. The participants were children aged between 11 and 13; ten boys and ten girls. The demographics survey shows that all of the children had some experience using computers and knowledge of geography, but none had experience with robots. The WoZ-style approach was adopted as it is a common practice in HMI and related fields [24], allowing for a smoother and more believable interaction than what can be achieved on the basis of a fully autonomous robot in the early stages of development. With the help of the wizard, the robot can display realistic behaviours and respond to the child in a timely fashion, without having to implement a fully functioning autonomous system that was not yet available at the time of this case study.

4.3.1.1 Scenario

The children were asked, by their geography teacher, if they wanted to take part in an educational map reading activity with one-to-one support from a robotic tutor. However, the children were not informed, until after the study, that the robot was being controlled by the wizard, i.e., a human. The children were required to employ their existing geography-related knowledge, while also learning how to navigate the map using various combinations of the compass, ruler and map key. The robot provided the child with support that would not only help them to progress further in the task, but also help them to think about how the skills they were learning could be applied to a range of map-related problems. An activity script, using the appropriate level of difficulty, as identified in previous mock-up studies, was written and tested with the help of several teaching experts, ensuring that the content was in-line with the England and Wales National Curriculum for Geography. The robot, a NAO,[1] started each interaction by introducing himself and then asking the child for his/her name, which was then repeated back to the child in a welcome statement.

Next, the robot provided the child with a brief tutorial, including an overview of the activity to help familiarize the child with the interface, tools and the type of support they could expect from the robot. The robot provided support throughout the interaction, and at times when the robot did not need to intervene, it used several idling animations to sustain a certain level of activity, realism, and presence. Additionally, when addressing the child, the robot would attempt to maintain an acceptable level of mutual gaze, looking away occasionally so not to be freaky. It was able to track the child's face and maintain the gaze even as the child moved around in front of the robot. The aim of these activities was to make the robot appear more intelligent and lifelike.

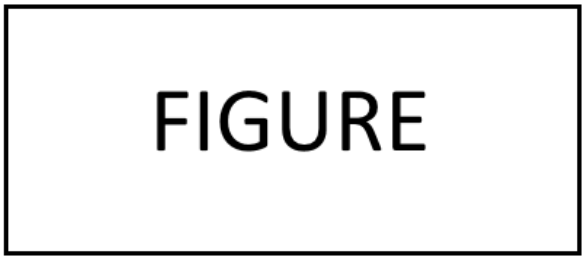

Fig. 4.2 Child interacting with the robot in a social exchange

4.3.1.2 Method

Technical Set-Up

The technical set-up for the WoZ study (*see* Fig. 4.2) comprises of a large touch-screen table that was embedded horizontally into a supporting aluminium structure, forming an interactive table-top surface, a torso-only version of the *NAO* humanoid robot, three video cameras positioned in frontal, lateral and top-down locations, a *Microsoft Kinect*, an *Affectiva Q Sensor* for measuring skin conductivity and OKAO vision software by OMRON for measuring smile intensity and eye-gaze direction.

Implementation of Social Probes

Social Probe 1:
Attractor: *"Nice to meet you Joe"*
[PAUSE]
Maintainer (A): *"I hope that you are doing well today"*
[PAUSE]

Maintainer (B): *"and that you'll have fun hanging out with me for a little while"*
Social Probe 2:
Attractor: *"Joe, Have you ever seen Wallace and Grommit?"*
[PAUSE]
Maintainer (A): *"I think that there is actually a Wallace and Gromit film with robotic trousers."*
[PAUSE]
Maintainer (B): *"I wonder what it would be like to have legs myself"*
Social Probe 3:
Attractor: *"Thank you so much for all you help!"*
[PAUSE]
Maintainer (A): *"I hope you had a good time!"*
[PAUSE]
Maintainer (B): *"I thought you did really good!"*

Level of Automation

All aspects of the interaction, including robot control, helping the child to progress through the activity, implementing the correct teaching strategies, and engaging in social exchanges, were remotely controlled by a qualified teacher using a bespoke interaction control interface (see Fig. 4.3).

Data Collection

The corpus of data collected from *case study 1* included more than seven hours of video material for each of the three viewing angles, interaction data recorded from involvement with the activity and data from the low-level sensors, such as skin conductance, facial action units, smile intensity, posture-related lean information and gaze direction. Data was cleaned of certain artefacts, including sensor noise and incomplete cases, and the raw low-level information was binned into discrete instances of time, more specifically 250 ms, which allowed us to process, analyse and model the data using statistical and machine learning methods.

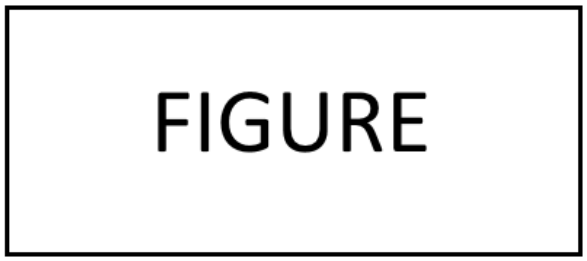

Fig. 4.3 Wizard's view of the interaction control interface

Annotating the Video Material

The simplest method of annotating video material is to use relative ratings and discrete segments of video media. However, in this work we require continuous measures that allow at least a rough estimation of the timing of relevant changes. Such continuous annotation data has the potential to add substantial flexibility to use the final computed 'ground truth' for different purposes, such as statistical analysis and training of machine learning algorithms. However, the trade-off for obtaining continuous data is that the precise moment of changes in subjective states such as engagement can be difficult to pinpoint even for trained raters, resulting in an overall lower reliability compared to a single global Likert scale. Nevertheless, this work follows the more novel approach of adopting the continuous measure which could, potentially, encode far more interesting information.

Annotation Software: Off-the-shelf annotation software could not provide the flexibility to perform Continuous annotations that allow the simultaneous presentation of multiple video streams of data, and the input modality is typically fixed to either mouse or keyboard. We wanted to explore the use of a game-pad, or more specifically the thumb stick of a game-pad. The assumption here is that releasing the thumb stick can be used naturally to indicate a return of the annotated measure back to a neutral state extremely quickly, whereas a mouse or keyboard would cause periods of uncertainty in the output signal, due to the fact that some active effort and time is required to return the rating back to neutral. Furthermore, the latter modalities are unable to offer the same fine grained resolution as an analogue thumb stick. CAT, or Continuous Annotation Tool, is a custom solution designed to facilitate these seemingly "unusual" annotation requirements, i.e., synchronously displaying three different views of the interaction such as, e.g., frontal, lateral and top-down, and providing a simple visual representation of the rating intensity (in real-time) on a vertical slider bar, which transitions from green at the very top to represent a positive or high intensity, orange in the centre to represent a neutral intensity and red at the bottom to represent an extremely low intensity (see Fig. 4.4). In CAT, annotator ratings are automatically logged with two decimal point precision, with maximum and minimum extremes set to 1 and -1 respectively.

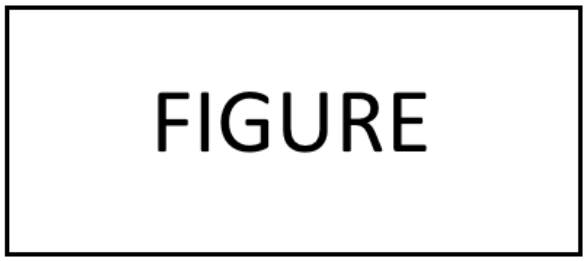

Fig. 4.4 CAT: Continuous Annotation Tool: Bespoke software, developed specifically for the use with continuous ratings

Annotators: For each annotated signal, i.e., social attention and valence, we used the same three annotators. So, for social attention, the annotators were: (1) a pedagogical researcher who could look at attention from a teacher-student perspective, (2) a psychologist to look at attention from a behavioural aspect, and (3) a researcher, specialising in automatic non-verbal behaviour analysis for social robotics.

Annotator Agreement: The issue of reaching agreement is an important part of the annotation process. Our methodology was to reach an acceptable level of agreement in advance, ensuring that the final output signals were the best that we could achieve with the time and resources we had. We adopted a three-step approach, starting with a discussion of the overall objective criteria, in an attempt to pre-emptively list potential indicators, and later in the process the annotators were asked to produce a voice over account and a single continuous rating for the same randomly selected interaction, to visualise the different output signals in a side-by-side analysis. Obviously, there were differences between the signals, but an acceptable level of difference, i.e., less than a second, was achieved.

Ground Truth Extraction

Computing and then extracting a ground truth is an essential process for this type of non-verbal behaviour recognition. The output of the extraction process, which involves aggregating the ratings from multiple annotations into a single signal, represents the final measure for a particular criterion, such as social attention or valence. Producing a ground truth can be a relatively straightforward process when working with discrete labels, but the very nature of our continuous rating process renders many existing methodologies as impractical [13]. In fact, many researchers choose to completely ignore the concept of agreement in favour of simpler methods, such as using the mean from several ratings, or alternatively opting to manually assess the ratings [16]. We wanted to ensure that we were not introducing biases or losing information, so we opted to explore other more suitable methods. An in-depth review of the literature uncovered two potential methods for computing a ground truth, based on annotator agreement. The first method focuses on the use of a correlative threshold, specifically 0.45, meaning that ratings from annotator pairs with correlative coefficients smaller than 0.45 are quite simply omitted from the computation of the ground truth. In contrast to this method, we consider it to be of the utmost importance that the ratings from *all* annotators are included when computing the ground truth, even those who are in disagreement with the others. This provides a more realistic ground truth that takes into account the different backgrounds and perspectives of the annotators. Therefore, we adopted the alternative method of using weighted correlations, similar to the work by Nicolaou et al. [16], which we then further extended for our interaction length non-segmented continuous signals. Here, ratings from all annotators are considered in the computation with the condition that the most highly correlated annotator pairs are given more weight than disagreeing annotators (see Fig. 4.5 for an example of the output ground truth signal). The graph in Fig. 4.6 provides an estimate of inter-rater reliability, in terms of intraclass correlation (ICC), more specifically, we have used ICC (2,3), which denotes the ICC values are calculated for each interaction using Case 2 from the work by Shrout and Fleiss [28], involving the same three annotators for each case.

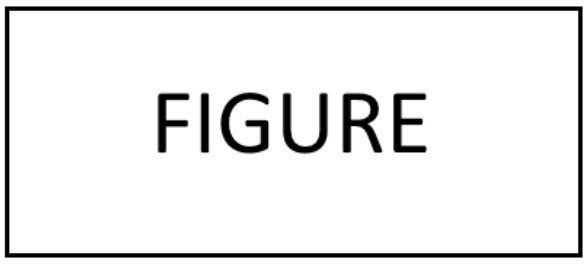

Fig. 4.5 Segment showing the continuous measure of social attention from three annotators (*light dashed lines*) with the final computed 'ground truth' (*dark solid line*)

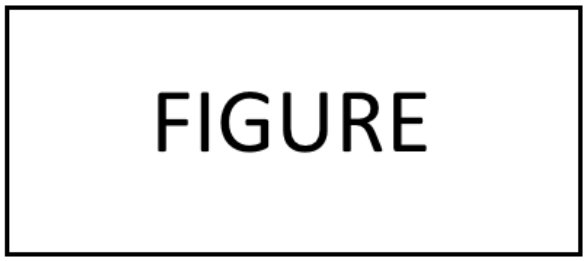

Fig. 4.6 Graph showing an estimate of inter-rater reliability for each of the twenty child-robot interactions, i.e., in terms of intraclass correlation (ICC)

4.3.1.3 Analysis and Results

Analysis of the corpus has, so far, been two-fold: an *interaction-length* analysis of the gaze with the social attention signal, and a *social probe-interval* focussed analysis of the behaviour-related variables with the social attention and valence signals. The motivation for this analysis is to explore what features may be descriptive of engagement in HMI.

Analysis A: Interaction-length

Here, we report the results of a point biserial correlation analysis between the social attention signal (interval scale) and a pre-processed dichotomous nominal scale relating to robot gaze, the two levels are 0 (if the learner was not looking at the robot) and 1 (if the learner is looking at the robot). The results of the analysis, set out in Table 4.1, show that, on average, gazing at the robot is moderately correlated with social attention *($r_{pb} = 0.47$, $p < 0.01$)*.

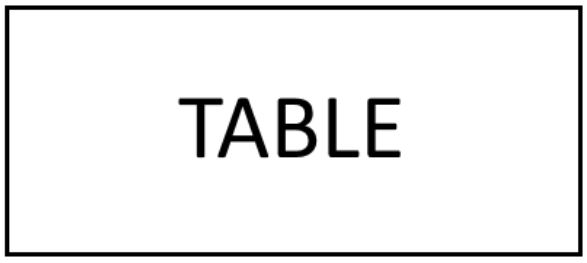

Table 4.1 Results of a point biserial correlation analysis between social-related attention and gazing at the robot. The columns represent interaction ID, point biserial correlation coefficient, statistical significance and number of low-level instances used in the analysis, respectively.

Analysis B: Social Probe-interval

To understand which, if any, behaviour-related variables are an indication of engagement, samples extracted from the *social probes* are compared with samples taken from similar areas of the interaction, +/-30 s, i.e., samples extracted from areas that do not involve *social probes*, in this work we will refer to these as samples taken from *"non-social probes"*. The samples taken from the *social probes* have been extracted in a way that they surround and capture the entire piece of probe-related interaction.

There are three *social probes* embedded into each interaction and the duration of each probe was: 17.75, 13.5, and 6.25 s, respectively, with no overlap between segments (*see* Sect. 4.3.1.2). The *non-social probe* samples were extracted from other random moments outside of the probe-intervals, using an identical process. This process was repeated for each of the 20 interactions, providing a total of 2420 instances of raw low-level data for each case. A comparison of the *social probes* with just a single case of *non-social probes* does not actually tell us anything useful, therefore, we extracted three different cases of *non-social probes* to further support our analysis.

Social Attention

For social attention we consider information from gaze, smile and facial expressions. Pearson product-moment correlation coefficients have been computed to assess relationships between gaze, smile and facial expressions, and the social attention signal, obtained from the ground truth extraction process (Sect. 4.3.1.2). The results of the analysis are set out in Table 4.2.

Table 4.2 Results of a Pearson product-moment correlation between the behavioural indicators and the social attention signal, i.e., for samples taken from the *social probes* and three other *non-social probes* (*NSP*). This table only shows significant correlations, figures marked with ** represent significance at the 0.01 level and others marked with * represent significance at the 0.05 level.

Valence

For valence, the affective component of engagement, we focus on behavioural indicators. Pearson product-moment correlation coefficients have been computed to assess relationships between the smile and facial expressions, and the valence signal, obtained from the ground truth extraction process (Sect. 4.3.1.2). The results of the analysis are set out in Table 4.3.

Table 4.3 Results of a Pearson product-moment correlation between the behavioural indicators and the valence signal, i.e., for samples taken from the *social probes* and three other *non-social probes* (*NSP*). This table only shows significant correlations, figures marked with ** represent significance at the 0.01 level and others marked with * represent significance at the 0.05 level.

4.3.1.4 Discussion

The most obvious finding to emerge from the analysis of the *social probe* versus *non-social probe* samples, is that correlations appear to be stronger, in the majority of cases, in samples taken from the *social probes*. A possible reason for this is that the social probes can be expected to have been particularly engaging in the sense of a simultaneous recruitment of different highly over-learned behavioural response systems. That is, in this case, the existence of well learned norms appears to have led more clearly and consistently communicated social signals—whereas, in the *nonsocial probe* case there are substantially less contextually defined social schemata to help guide the encoding as well as decoding of the behaviours.

4.3.2 Case Study 2: Engagement with Virtual Agents

The development of autonomous virtual agents and animated characters capable of engaging humans in real-time interaction faces many of the same challenges as similar attempts using physical embodiments such as social robots. These include the task of obtaining robust real-time detection of engagement-related behaviour from human users, timely responses from virtual agents, and the generation of appropriate behaviours by agents that are capable of properly expressing their state of engagement and focus of attention. This section describes an example scenario that involved engagement between a human and a virtual agent. The interaction was primarily shaped by gaze behaviour, in particular when it was directed at predefined objects and locations within an artificial scene (for a review, see [26]).

4.3.2.1 Gaze Detection and Representation

There are, in principle, a large number of measures that can be used to detect user engagement. These include, for example, monitoring verbal and non-verbal behaviours, taking physiological measurements and tracking task related actions that are conducted inside an application or virtual environment. However, not all of these measured may be needed at once for a basic analysis of engagement processes. In this scenario, the main method for detecting engagement was based on the gaze behaviours of the user as they engaged in an object identification task with a virtual agent capable of referring to objects non-verbally through its own gaze behaviours.

Gaze Detection

The gaze detection system used facial feature analysis of the image captured from a standard web camera to capture the user's gaze direction (head and eye directions) inside and outside of the screen. The detection process commenced with the eyecenters, which are easily detected, in order to allow the estimation of the eye corners and eyelids and positions on the eyebrows, nostrils, and mouth region. These were subsequently tracked using a Lucas Kanade tracker [12], capable of operating under a wide variety of conditions. Head-pose estimation was calculated based on the displacement of the midpoint of the eye centers from an initial head position in which the user was facing the screen frontally.

Embedded Representations

Since the user's focus of attention is usually highly transient as it shifts around a scene, it can be informative to use predefined objects to track these changes more systematically. Virtual Attention Objects (VAOs) simplify the analysis of what is being looked at in the scene by storing, on a per-object level, when and how much each part of the scene has been looked at. A single VAO is attached to each scene object for which we wish to accumulate user attention information. For example, a single VAO may be defined for each visible object in the scene, including the virtual agent itself. Depending on the requirements of the application, the virtual agent may be represented by a single VAO, or a separate VAO may be defined for each part of the virtual agent for which information about user attention is required. VAOs may also be defined for more abstract objects. For example, a single VAO may be defined to represent the area outside of the screen. This VAO can then be used to record whenever the user gaze wanders outside of the scenario area, which renders it a useful metric for disengagement. Furthermore, VAOs operate in a simple manner: Screen coordinates relating to user gaze are resolved to the specific associated VAO or VAOs. On this basis, the estimated level of attention can be adapted accordingly. The combined VAOs therefore represent a history of how much and when the user has fixated on each object in the scene. Figure 4.7 illustrates a virtual scene from [21], including a virtual agent and a number of objects, and an accompanying VAO representation.

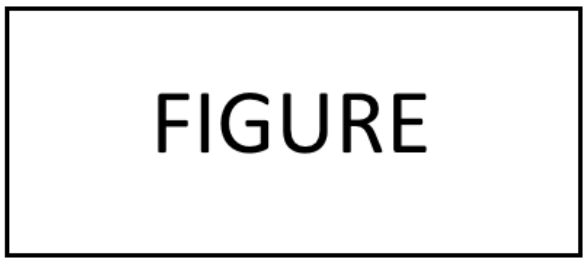

Fig. 4.7 A scenario involving a virtual agent and several objects. The user's gaze behaviour (*left*) is tracked as the agent conducts gaze behaviours towards various objects (*centre*). Gaze and attentive behaviours towards specific elements of the scene are recorded in real-time through Virtual Attention Objects (*right*) as a basis for monitoring user engagement in the scenario [21]

4.3.2.2 Engagement Modelling

In this work, the focus was not only on the different components of engagement, but also the level and quality of those components. Expanding measures beyond what has already been discussed in this chapter, in order to capture varying degrees of engagement and related components. More specifically, we refer to directedness, level of attention, level of engagement and quality of engagement.

Directedness relates to the momentary orienting of the user's body parts with respect to another entity or object from the perspective of that entity or object. The metric is inspired by Baron-Cohen's eye, head and body direction detectors [2] and related work [21]. Directedness as a concept refers to transient and momentary processes that alone do not imply attention or engagement. For example, high directedness was assumed if the eye and head direction was sampled from a user while they are in the process of a gaze change to an alternative location. However, the consideration of directedness over time (and with respect to other aspects of the scene and user) is an important building block towards a model of engagement measurement. Depending on the scenario and, especially, the sensory capabilities of the detection system, the directions of the eyes, head, body and even locomotion trajectories may contribute to directedness measurements.

Level of attention is based on directedness and refers to gaze falling within certain regions over a period of time. It therefore corresponds to the concept of a focus of attention and corresponding dwell time by the eye. An important issue in this respect relates to the clustering of the foci of interest of the user. For scenes that are clearly composed of objects, VAOs may be used as one method for clustering fixation locations on a per-object basis.

Level of interest is based on the stored attention levels over time for each member of a set of VAOs. Each member is categorised according to whether it is a scene object, the agent, the background, or a special object representing the area outside of the screen. It is at this level that specific forms of context can be accounted for: By dynamically defining a set of VAOs containing only those objects relevant to the current interaction, such as recently pointed to or discussed objects, the attention of the user can be compared with this set to obtain a measurement of their level of interest in the interaction itself, referred to here as the level of engagement.

Level of engagement encapsulates how much the user has been looking at the relevant objects in the scene at appropriate times. These will be recently referenced objects in the interaction, e.g. those looked at, pointed to and/or verbally described. These measures are made possible by considering the specific set of VAOs corresponding to currently and recently referenced objects in the interaction. When the agent is talking, but does not refer to anything in the environment, it will be the only VAO in the set, and when it stops talking, this VAO set will be empty.

Quality of engagement accounts for the fact that attention paid to the scene does not necessarily indicate engagement in relation to a particular activity, in this case, interaction with the agent. For example, the user may be looking at the scene for superficial reasons without engaging in the interaction. It provides a slightly more detailed assessment of the type of engagement that the user has entered into. For example, a user who is not engaged in the interaction may not necessarily be looking outside of the scene. Instead, they may be attending to the scene in a superficial manner, looking at objects of interest that are irrelevant to the ongoing interaction. In particular, they may appear to be affectively disengaged from the interaction. We therefore define three broad quality levels: (i) engaged in the interaction (ii) superficially engaged with the scene and action space and (iii) uninterested in the scene/action space. In this way, the behaviour of the user is not being considered in isolation, but in the context of what the agent is doing. If the agent is describing something important for example, a user's disengagement can be considered more serious than if the agent is not doing anything at all.

4.3.2.3 Human Perception

In order to support sustained interactions between humans and virtual agents, models of engagement should not only account for the detection of behaviour from users and the generation of expressive behaviours by agents, but also how humans perceive artificially generated behaviours. This is an important consideration that may involve many factors, such as the embodiment and expressive capabilities of the system, the qualities of the expressive motion, the effects of context on perception and more. One of the fundamental methods of signalling attention and engagement by artificial systems involves gaze [26] and numerous studies have considered the human perception of gaze and other attention related behaviours made by artificial systems.

Gaze and Direction of Attention Perception

Gaze direction perception, which primarily involves eye and head movements and their relationships with

objects in the environment from the perspective of a witness [23], can be extended to a more general concept that relates to the perception of the direction of attention of others. Here, the orientation of the eyes, head, body, and even locomotion trajectories, may contribute to one's impression that they are potentially being attended to by others. For example, at a distance, the eyes of the other may not be clearly visible in which case head and body directions may offer more prominent clues as to the direction that the other is attending to.

Opening Interactions

An example of the use of the concept of direction of attention perception is proposed in [19] for the purposes of opening interactions between interactants in virtual environments and the human perception of the attentive behaviours of virtual characters. Kendon [9] describes a sequence of opening cues for meeting interaction, noting interaction fundamentals, such as the requirements for participants to see each other, while Goffman notes [7] that we generally seek to avoid the social embarrassment of engaging in interaction with an unwilling participant. Therefore, some degree of confidence of interaction reciprocation is necessary, through subtle gaze behaviours for example, before more explicit actions are made, such as verbal greetings. Based on this, a model of engagement opening is described for virtual agents based on a combination of their own gaze behaviours towards others that they wish to interact with and also their interpretations of the gaze and locomotion behaviours of others.

4.3.2.4 Discussion

Important future work is waiting to be done to establish better relationships between the expressive qualities of behaviours of virtual characters, their style and embodiments, and human perception. In humanoid characters, for example, small movements, such as saccadic eye movements and eye-blinks, may play important roles in signalling that the character is concentrating heavily, rather than simply daydreaming or momentarily unresponsive. Embodiment issues related to the use of virtual characters versus their physical counterparts is also an interesting research area. For example, [15] have found that faces that are physically-projected onto face-like surfaces may have a number of advantages over virtual faces constrained to flat screens, especially in relation to the character's ability to engage with individual users and for cueing real objects in the environment. The studies represent some important foundations for constructing artificial entities capable of engaging humans.

Acknowledgments This work was partially supported by the European Commission (EC) and was funded by the EU FP7 ICT-317923 project EMOTE (EMbOdied-perceptive Tutors for Empathy-based learning) and the EU Horizon 2020 ICT-644204 project ProsocialLearn. The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC, and the EC is not responsible for any use that might be made of data appearing therein.

References

1. Asteriadis S, Karpouzis K, Kollias S (2009) Feature extraction and selection for inferring user engagement in an hci environment. In: Human-computer interaction. Springer, New Trends, pp 22–29
2. Baron-Cohen S (1994) How to build a baby that can read minds: cognitive mechanisms in mind reading. Curr Psychol Cogn 13:513–552
3. Castellano G, Pereira A, Leite I, Paiva A, Mcowan PW (2009) Detecting user engagement with a robot companion using task and social interaction-based features interaction scenario. In: Proceedings of the 2009 international conference on multimodal interfaces, pp 119–125
4. Christenson SL, Reschly AL, Wylie C (2012) Handbook of research on student engagement. Springer, Boston
5. Corrigan LJ, Basedow C, Küster D, Kappas A, Peters C, Castellano G (2014) Mixing implicit and explicit probes: finding a ground truth for engagement in social human-robot interactions. In: Proceedings of the 2014 ACM/IEEE international conference on HRI. ACM, pp 140–141
6. Corrigan LJ, Basedow C, Küster D, Kappas A, Peters C, Castellano G (2015) Perception matters! Engagement in task orientated social robotics. In: IEEE RO-MAN 2015. doi:10.1109/ROMAN.2015.7333665
7. Goffman E (2008) Behavior in public places. Simon and Schuster, New York 8. Kappas A, Krämer N (2011) Studies in emotion and social interaction. Face-to-face communication over the internet: emotions in a web of culture, language, and technology. Cambridge University Press, Cambridge
9. Kendon A (1990) Conducting interaction: patterns of behavior in focused encounters, vol 7. CUP Archive, Cambridge
10. Langton SR, Watt RJ, Bruce V (2000) Do the eyes have it? Cues to the direction of social attention. Trends Cogn Sci

4(2):50–59
11. Lewis M, Haviland-Jones JM, Barrett LF (2010) Handbook of emotions. Guilford Press, New York
12. Lucas BD, Kanade T (1981) An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th international joint Conference on Artificial Intelligence, pp 674–679
13. Metallinou A, Narayanan S (2013) Annotation and processing of continuous emotional attributes: challenges and opportunities. In: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), pp 1–8
14. Mota S, Picard RW (2003) Automated posture analysis for detecting learner's interest level. In: Conference on computer vision and pattern recognition workshop, 2003. CVPRW'03, vol 5. IEEE, pp 49–49
15. Moubayed SA, Edlund J, Beskow J (2012) Taming Mona Lisa: communicating gaze faithfully in 2d and 3d facial projections. ACM Trans Interact Intell Syst (TiiS) 1(2):11
16. Nicolaou MA, Gunes H, Pantic M (2010) Automatic segmentation of spontaneous data using dimensional labels from multiple coders. In: Proceedings of LREC int'l workshop on multimodal corpora: advances in capturing, coding and analyzing multimodality, pp 43–48
17. O'Brien HL, Toms EG (2008) What is user engagement? A conceptual framework for defining user engagement with technology. J Am Soc Inf Sci Technol 59(6):938–955
18. Pekrun R, Elliot AJ, Maier MA (2009) Achievement goals and achievement emotions: testing a model of their joint relations with academic performance. J Educ Psychol 101(1):115
19. Peters C (2006) Evaluating perception of interaction initiation in virtual environments using humanoid agents. In: Proceedings of the 2006 conference on ECAI 2006: 17th European conference on artificial intelligence 29 Aug–1 Sept, 2006. IOS Press, Riva del Garda, pp 46–50
20. Peters C, Asteriadis S, Karpouzis K, de Sevin E (2008) Towards a real-time gaze-based shared attention for a virtual agent. In: Workshop on affective interaction in natural environments (AFFINE), ACM international conference on multimodal interfaces (ICMI08)
21. Peters C, Asteriadis S, Karpouzis K (2010) Investigating shared attention with a virtual agent using a gaze-based interface. J Multimodal User Interfaces 3:119–130. doi:10.1007/s12193-009-0029-1
22. Pitsch K, Kuzuoka H, Suzuki Y, Sussenbach L, Luff P, Heath C (2009) The first five seconds: contingent stepwise entry into an interaction as a means to secure sustained engagement in HRI. In: The 18th IEEE international symposium on robot and human interactive communication, 2009. RO-MAN 2009, pp 985–991. doi:10.1109/ROMAN.2009.5326167
23. Qureshi A, Peters C, Apperly I (2013) Interaction and engagement between an agent and participant in an on-line communication paradigm as mediated by gaze direction. In: Proceedings of the 2013 inputs-outputs conference: on engagement in HCI and performance, p 8
24. Riek LD (2012) Wizard of oz studies in hri: a systematic review and new reporting guidelines. J Hum-Robot Interact 1(1):119–136
25. Roseman IJ, Smith CA (2001) Appraisal theory: overview, assumptions, varieties, controversies. In: Appraisal processes in emotion: theory, methods, research. Series in affective science, pp 3–19
26. Ruhland K, Andrist S, Badler J, Peters C, Badler N, Gleicher M, Mutlu B, Mcdonnell R (2014) Look me in the eyes: a survey of eye and gaze animation for virtual agents and artificial systems. In: Eurographics state-of-the-art report. The Eurographics Association, pp 69–91
27. Shernoff DJ, Csikszentmihalyi M, Schneider B, Shernoff ES (2003) Student engagement in high school classrooms from the perspective of flow theory. Sch Psychol Q 18(2):158–176
28. Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86(2):420–428
29. Sidner CL, Dzikovska M (2005) A first experiment in engagement for human-robot interaction in hosting activities. Advances in natural multimodal dialogue systems, pp 55–76