Automatic Context-Aware Inference of Engagement in HMI: A Survey

Hanan Salam, Oya Celiktutan, Hatice Gunes, Mohamed Chetouani

Automatic Context-Driven Inference of Engagement in HMI: A Survey

Hanan Salam, *Member, IEEE,* Oya Celiktutan, *Member, IEEE,* Hatice Gunes, *Senior Member, IEEE,* and Mohamed Chetouani, *Member, IEEE,*

Abstract—An integral part of seamless human-human communication is engagement, the process by which two or more participants establish, maintain, and end their perceived connection. Therefore, to develop successful human-centered human-machine interaction applications, automatic engagement inference is one of the tasks required to achieve engaging interactions between humans and machines, and to make machines attuned to their users, hence enhancing user satisfaction and technology acceptance. Several factors contribute to engagement state inference, which include the interaction context and interactants' behaviours and identity. Indeed, engagement is a multi-faceted and multi-modal construct that requires high accuracy in the analysis and interpretation of contextual, verbal and non-verbal cues. Thus, the development of an automated and intelligent system that accomplishes this task has been proven to be challenging so far. This paper presents a comprehensive survey on previous work in engagement inference for human-machine interaction, entailing interdisciplinary definition, engagement components and factors, publicly available datasets, ground truth assessment, and most commonly used features and methods, serving as a guide for the development of future human-machine interaction interfaces with reliable context-aware engagement inference capability. An in-depth review across embodied and disembodied interaction modes, and an emphasis on the interaction context of which engagement perception modules are integrated sets apart the presented survey from existing surveys.

Index Terms—Engagement Detection, Human-Machine Interaction, Socially Intelligent Systems.

1 Introduction

The field of human-machine interaction (HMI) is rapidly developing to address various societal challenges. Human interactions with machines can take different forms, depending on the scenario, machine (dis)embodiment (referred to as interaction mode hereafter), and interaction goal. Examples include delivering remote education [1], enhancing mental well-being [2], and supporting elderly individuals [3]. The success of such applications highly depends on users' satisfaction, trust and technology acceptance; therefore, it is becoming increasingly desirable that human-machine interaction systems develop social intelligence and become attuned to their users through the effective use of multimodal communication channels, ultimately leading to the maximization of the targeted interaction's outcomes [4].

One key component of social intelligence is unarguably engagement [5], [6], [7]. Engagement is a complex multimodal and multi-faceted social phenomenon that requires the perception and recognition of social signals and their interpretation at a higher level for social behaviour regulation. In the past decade, there has been a significant body of work that aims to develop engagement inference models and machine behaviour adaption mechanisms in various human-machine interaction contexts [8], [9]. From the perspective of disembodied interaction (machine interface without embodiment, e.g. human-computer interaction), for example in the context of HMI for learning, it is important to design engaging learning systems that have the capability of detecting the user's engagement state and adapting to it, allowing the user to acquire the learning outcomes objectives [10], [11]. Within the context of game entertainment, designing engaging games is essential for making the user's experience enjoyable and preventing withdrawal [12]. From the perspective of embodied interaction (machine interface with physical or virtual embodiment, e.g. human-agent interaction and human-robot interaction), engagement is an essential rubric, which allows a smooth and natural interaction between the user and the robot/agent, and can contribute to achieving effective long-term interactions that go beyond the novelty effect. Across different HMI contexts, achieving engaging human-machine interactions requires that the machine is able to 1) interpret human's engagement from the observation of their multimodal cues [13] and 2) express its engagement in an appropriate manner beyond on-off interactions.

Lately there has been an increasing trend towards integrating contextual information in social signal processing and affective computing research [14], [15]. During an interaction, any information that allows the characterization of an entity's situation can be considered as context, provided that the entity is an individual, a location, or any object relevant to the human-machine interaction, including the user and the machine [16]. Context is used intuitively by humans in social interactions to act and react properly, as

- *H. Salam is with Center of AI & Robotics (CAIR), SMART Lab, New York University, Abu Dhabi, E-mail: hanan.salam@nyu.edu*
- *O. Celiktutan is with the Department of Engineering, King's College London, United Kingdom, WC2R 2LS E-mail: oya.celiktutan@kcl.ac.uk*
- *H. Gunes is with the Department of Computer Science and Technology, University of Cambridge, United Kingdom, CB3 0FD, E-mail: hatice.gunes@cl.cam.ac.uk*
- *M. Chetouani is with the Institute of Intelligent Systems and Robotics, CNRS UMR7222, Sorbonne University, Paris, France, E-mail: mohamed.chetouani@sorbonne-universite.fr*

well as to correctly infer the others' state of mind [16]. In particular, humans might manifest their engagement state in different ways that largely depends on context [17]. The user's mental, emotional, and behavioural states associated with their engagement state was also found to vary with the interaction context [18]. Consequently, improving the machine's access to context information would increase its social intelligence skills and promote more accurate, adaptive, and engaging user experience [16].

The importance of integrating context in the design of engagement inference systems has been loosely underlined in the literature through the use of various contextual cues as input to automatic engagement inference models [19]. Despite the global recognition of the importance of context-aware engagement modeling and inference by the community, however, the literature lacks a systematic context-driven overview on the topic. In this paper, we present an in-depth overview of the engagement modelling, detection, and recognition approaches across different interaction modes (i.e. disembodied and embodied interaction). We put a special emphasis on the importance of context for modelling engagement and for the development of context-aware, accurate and adaptive engagement perception algorithms. The current engagement survey is context-driven in the sense that it discusses engagement definitions across different interaction modes, outlines different contextual factors that have an effect on human-machine engagement (e.g. interaction mode and scenario, and personal factors), and reviews contextual features used in engagement inference models in the literature.

There has been a couple of previous efforts that reviewed the definition of engagement [20] and its implications in human-agent interaction [6]. There is also a recent survey by Oertel *et al*. [21], which reviewed the definition of engagement and how it differs across different interaction settings (i.e., real-world versus laboratory, short-term versus long-term, social versus task oriented) and user profiles (e.g., adults versus children). However, in their survey, the emphasis is more on the behaviour adaptation strategies and the review of the engagement perception methods does not go beyond the deep learning approaches. In [22], a survey on engagement level recognition in child-robot interaction was also presented. However, the survey's scope was limited to education and therapeutic settings. To the best of our knowledge, this paper is the first comprehensive context-driven survey of automatic engagement inference, starting from the definition of engagement to the design of the full detection pipeline including data acquisition, feature extraction, and inference.

This review will serve as a guide for researchers interested in the topic of engagement to acquire a holistic understanding of the concept. Specifically, the emphasis on the contextual factors of engagement, and how engagement was defined and detected across different contexts in the literature will inform context-aware artificially intelligent systems. Consequently, this will allow the design of human-machine interactions with increased usability, accuracy, and efficiency in real-time settings, leading to improved user experience.

Fig. 1. Examples of disembodied [34] and embodied (HAI [35], HRI [15]) interaction scenarios.

2 Context-Driven Engagement Definition

In order to build effective systems for engagement inference, it is essential to establish a clear and precise definition of the notion. Engagement is a complex construct composed of various components or factors, which were covered across various interaction modes, namely, human-human interaction (HHI), human-computer interaction (HCI), human-agent interaction (HAI) and human-robot interaction (HRI). Different concepts (e.g. attention, involvement, interest, immersion, rapport) were related to engagement and sometimes even used interchangeably in the literature [20].

From the perspective of the user, engagement is often seen to be composed of three factors: emotional, cognitive, and behavioural. Depending on the context, some factors can be predominant with respect to the others [18]. For instance, cognitive factors such as concentration might be more predominant in a learning context, compared to a purely social context. It is crucial to understand how the notion of engagement changes based on the different context categories as they appeared in the different studies. We distinguish between three context types that influence the user's engagement state, and process in their interaction with intelligent systems, namely, (1) the interaction mode (i.e., embodied vs. disembodied), (2) the interaction scenario (e.g., competitive vs. collaborative), and (3) personal factors (e.g., personality and gender). In this section, we summarise the most commonly used definitions of engagement and associated attributes. We investigate how engagement was related to different concepts across different contexts. We underline how these change based on the context. Table 2 summarizes the commonly used definitions across the different modes of interaction.

2.1 Interaction Mode: Embodied vs. Disembodied

In the sequel, we review widely used definitions of engagement for human-human interaction, disembodied human-machine interaction (e.g., mobile devices and web applications), and embodied human-machine interaction (e.g., agents, and robots) settings.

2.1.1 Engagement in Human-Human Interaction

The notion of engagement in human-human interaction (HHI) was addressed in social sciences by Goffman [5] who differentiates between *unfocused interaction* and *focused interaction*. Unfocused interaction, a pre-requisite of engagement, is concerned with what can be communicated between people due to their co-presence in the same social situation (e.g.

TABLE 1
Summary of the most commonly used definitions across the different modes of interaction: HHI, HCI, HAI, and HRI. Interaction Mode (IM).

IM	Definition	Paper
HHI	Engagement occurs when people gather closely together and openly cooperate to sustain a single focus of attention, typically by taking turns at talking.	[5]
HCI	Engagement with technology is a measure of the quality of user experience.	[6], [23]
	A connection that exists at any point of time and possibly over time between a user and a resource. The cognitive, affective, and behavioural state of interaction that makes the user want to be there.	[24]
	Engagement in online learning is a construct that encompasses student's behaviours and involvement in consistent engagement with resources or activities within the online environment, with the end-goal of achieving learning.	[25]
	Engagement in computer supported collaborative learning is with-me-ness which measures how much are the students with the instructor.	[26]
HAI	An emotional state linked to the participant's goal of receiving and elaborating new and potentially useful knowledge.	[27]
	Empathic engagement is fostering of emotional involvement intending to create a coherent cognitive and emotional experience which results in empathic relations.	[28]
	The value that a participant in an interaction attributes to the goal of being together with the other participant(s) and continuing interaction.	[29]
HRI	The process by which two (or more) participants establish, maintain and end their perceived connection. This process includes initial contact, negotiating a collaboration, checking that the other is still taking part in the interaction, evaluating, staying involved, and deciding when to end connection.	[7]
	The process of subsuming the joint, coordinated activities by which participants initiate, maintain, join, abandon, suspend, resume, or terminate an interaction.	[30]
	Magnitude of an intrinsically motivated behaviour that is initiated by an organism to reach a specific goal	[31]
	Task engagement where there is a task and the participant starts to enjoy the task they are doing, *social engagement* which considers being engaged with another party when there is no task included, and *social-task engagement* which includes interaction with another (e.g., robot) where both cooperate with each other to perform a task.	[32]
	Productive Engagement is defined as the level of engagement that maximizes learning.	[33]

glancing to glean information about the other). On the other hand, focused interaction (face engagements/encounter) takes place when there is *cooperation* between individuals gathering closely together to *sustain a single focus of attention*, commonly via turn-taking and conversing.

In HHI, several researchers investigated engagement in the context of learning. Learner's engagement was linked to emotional (valence, value, interest), cognitive (motivation, effort, strategy), and behavioural (participation, persistence) components [36] [37] [38] [39] [40]. Another view of learner's engagement in school setting relates engagement to behavioural, academic, cognitive, and psychological dimensions [41]. For instance, behavioural engagement in a class refers to class attendance, concentrating on tasks, listening and following the teacher's directions. A cognitive component (concentration) is apparent in what is described as behavioural in [41]. *Emotional engagement* is the emotional attitude concerning the learning task. A student may have high behavioural engagement (a student obtaining high grades on exams) and low emotional engagement (but is bored). *Cognitive engagement* is related to learning with cognitive abilities such as memory, attention, or strategy. Another work considered student's engagement as the linear sum of the student's perceived focused attention, percieved involvement in the task, and the endurability (perception of the experience as worthwhile) [42]. Few studies discussed that a learner's engagement state can be modelled by the learner's affective state [43]. Specifically, engagement was directly related to the state of flow [44], which can be reached during periods of full engagement (e.g. when improving or enjoying the learning activity), whereas disengagement can be depicted during period of lack of enjoyment or non-advancement in the learning activity. Negative affective states such as boredom, frustration, and confusion are connected to disengagement [44]. In addition to enjoyment, factors of challenge and being in a zone of proximal development (distance between what a learner can do with support and without support) were considered for student's engagement [45]. The study of Pekrun *et al*. [46] proposed that learning and cognition are highly affected by affect, which has an impact on motivation, attention, and strategy use.

The definitions of engagement in HHI focus on the co-presence of interaction partners, cooperation on mutual activities, sustaining a single focus of attention, and establishing and maintaining a connection. The interaction between humans is an embodied one, and the embodiment is similar. Humans as interaction partners are of similar nature, and the interaction is governed by various factors such as the (in)existence of a history between the individuals, their social roles, the goal of the encounter, their knowledge of each others, their rapport, their similarities and differences, their characteristics and backgrounds, etc.

2.1.2 Engagement in Disembodied HMI

Disembodied HMI (also referred to as HCI) can take several forms, such as interacting with a computer application, web searching, online shopping, webcasting, learning, and gaming among others. In such interaction, the focus is primarily on the task. While some form of social interaction might exist, the absence of embodiment, limits the machine's expression of some forms of social intelligence (e.g. through non-verbal signals like gestures and emotions), and consequently, the expectation of the user from the system in this regard. Certain social channels (e.g interactive pop-up

messages in learning contexts) can be used in such contexts to increase the human's engagement, but this depends on the machine's design and social expressivity capabilities. However, the design of engaging human-computer interactions, even with limited social intelligence expressivity, is of utmost importance, and serves in attaining the system's end-goal, ensuring its usability and long-term usage. The factors pertaining to engagement in HCI depend on the system's characteristics, social expressivity capabilities, and the interaction's end goal, among others.

Engagement with technology is regarded as a measure of the quality of user experience [6], [23], a connection between a user and a resource that can exist at any instant and even in the long term [47]. Engagement with a computer application was referred to by Obrien *et al*. [24] as the cognitive, affective, and behavioural state of interaction that "makes the user want to be there". It is a process comprised of four distinct stages including point of engagement, period of sustained engagement, disengagement, and re-engagement. It is characterized by attributes that concern the user, system, and their interaction. These include challenge, positive affect, endurability, aesthetic and sensory appeal, attention, feedback, variety/novelty, interactivity, and perceived user control [6]. Reputation, trust and expectation, and user context were also underlined as characteristics of engagement in the context of web-applications [47]. In practice, engagement characteristics were narrowed down in some HCI contexts to cognitive components, e.g. attention [48], [49], or affective components, e.g. frustration [50].

Engagement in the context of online learning has been widely studied. For instance, the concept of with-me-ness for computer supported collaborative learning was introduced in [26] in an attempt to assess "how much are the students with the instructor?". Two components of with-me-ness are distinguished: perceptual - learner engages with what the instructor is referring to via deictic gestures (e.g. pointing), and conceptual - learner engages with what the instructor is referring to verbally. In [25], engagement in online learning is regarded as a construct that encompasses student's behaviours and involvement in consistent engagement with resources or activities within the online environment, with the end-goal of achieving learning. Concerning factors pertaining to engagement in online learning were referred to as social, cognitive, behavioural, collaborative and emotional elements [51]. The student's access to learning material was underlined as a factor of student's engagement in [52]. A longitudinal study aiming at understanding online student engagement over a semester [53] demonstrated a dynamic and fluctuating nature of student engagement, which is affected by factors of assessment, course units workload, course units content, lecturer presence and behaviour, and work/life commitments.

The definitions of engagement in disembodied HMI focus on the user-system interaction end-goal. While some factors are common to HHI, such as the user's cognitive, emotional, and behavioral factors, other factors pertaining to the system (e.g. aesthetics, reputation, perceived user control) or to the interaction context (e.g. course units content and workload in online learning) are more relevant to HCI.

2.1.3 Engagement in Embodied HMI

Embodiment in HMI usually takes two forms: virtual (embodied conversational agents), and physical (robots). Embodiment results in higher social intelligence expressivity capabilities in the system and interaction as compared to disembodied interaction. Interaction with virtual agents is often conversational, which adds more weight to the social aspect of engagement. On the other hand, physical embodiment increases the domain of applications of the machine. Activities that require physical capacities (e.g. waitressing or care-giving) as well as physical collaborative activities (e.g robot-mediated collaborative learning) become possible with physical embodiment.

HAI. In the literature of virtual embodied agents, some definitions of engagement were associated with the goal of acquiring knowledge, with an emphasis on the emotional dimension and empathy. For instance, in the context of a conversational scenario, Peters *et al*. [27] defined engagement as "an emotional state linked to the participant's goal of receiving and elaborating new and potentially useful knowledge". They presented engagement as the direct consequence of interest and and attention. The notion of empathic engagement was referred to by [28] as "fostering of emotional involvement intending to create a coherent cognitive and emotional experience which results in empathic relations". According to Glas and Pelachaud [20], the definition of Poggi [29] of engagement as "the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and continuing interaction" is the most suitable definition for engagement in embodied HMI as it resonates with making interactions with virtual agents as believable and human-like as possible.

HRI. The history of engagement in the field of HRI dates back to the year 2002. In [7], Sidner and Dzikovska presented a robot endowed with basic engaging capabilities including the capacity to initiate, maintain, and end a conversation with a human. According to Sidner and Dzikovska, engagement is defined as "the process by which two (or more) participants establish, maintain and end their perceived connection. This process includes initial contact, negotiating a collaboration, checking that other is still taking part in interaction, evaluating, staying involved, and deciding when to end connection." In their work [7], engagement was studied in the context of hosting activities, namely, a class of *collaborative activities* where an agent provides services to the human in a certain context (e.g. information, entertainment, education), such that the human may be requested by the agent to perform some actions, necessary for the fulfilment of such services. They underlined attention as a direct correlate of engagement by demonstrating that, during an interaction, a robot performing "engagement gestures" (e.g. following the user's face) would lead to an increase of user's attention, and consequently, their engagement.

The definition of Sidner and Dzikovska is among the most widely used definitions in the area of HRI. For example, [54], [55], and [19] adopted this definition of engagement, particularly, for describing emotional interaction level and social bonding established in child-robot interaction.

Other definitions widely used in the HRI literature were proposed by Bohus *et al*. [30] and Peters [56]. According to Bohus *et al*. [30], engagement is "the process of subsuming the joint, coordinated activities by which participants initiate, maintain, join, abandon, suspend, resume, or terminate an interaction." Various HRI studies employed the conceptualisation of Bohus *et al*.'s engagement definition in their studies [57], [58], [59]. The definition of Poggi [29] was adopted by Peters [56] and used in the context of child-robot interactions by Castellano *et al*. [60] and Sanghvi *et al*. [61] among others. Another definition that builds upon motivation theory instead of cognitive and emotional constructs is that of Drejing *et al*. [31] who suggests that engagement can be defined as the magnitude of an intrinsically motivated behaviour initiated by an individual to reach a specific goal.

With physical embodiment, the interaction components can vary between social, task, and social-task. Consequently, an interesting view on engagement, in this regard, is that of Corrigan *et al*. [32] who introduced a context-dependent engagement definition, in terms of task, social and social-task contexts. *Task engagement* corresponds to scenarios where a participant is performing a task, and enjoys it. *Social engagement* involves engagement with an interaction partner in a social interaction, with an absence of a task. *Social-task engagement* involves a cooperation with another to perform a task. Some studies such as that of [62] focused on social-task engagement, by concentrating on the degree of engagement of the user with the robot during a collaborative task. Consequently, a metric of engagement was defined as the normalized fraction of time an interaction party directs their attention to the attention target the robot expects for the current task (or sub-task). In the context of collaborative learning, the concept of Productive Engagement (PE) was introduced in [33] with the aim to conceptualize *engagement that is conducive to learning*. PE is defined as the level of engagement that maximizes learning and is composed of social and task engagement. In contrast to existing work in engagement conceptualization, the authors argue that being overly engaged can result in decreased learning outcomes.

Similar to disembodied HMI, cognitive and emotional elements of attention and valence of feeling were emphasised as components of engagement [14]. According to Corrigan [32], engagement is characterized by elements of participation, commitment, concentration, involution and immersion. Causation elements constitute internal states or desires like intrigue, curiosity, amazement, interest, concern, or wonder. Furthermore, engagement may evoke more emotional aspects of awareness, states of pleasure or arousal, thus justifying the initial investment in engagement.

Engagement and its underlying factors in embodied HMI can be seen as a middle-way construct, borrowing aspects from engagement in HHI and disembodied HMI. On the one hand, embodiment provides presence, and humans are less likely to ignore the system, and consequently the expectations in terms of social interaction naturalness and human-likeness. On the other hand, the focus on the task and its differentiation from the social aspect of the interaction is apparent in the embodied HMI engagement literature.

2.2 Interaction Scenario

Interaction scenario describes the interaction between the machine (embodied or disembodied) and the human. Different interaction scenarios might trigger different cognitive, emotional, and behavioural user states, indicative of their engagement state [18]. Based on existing literature in HMI, we differentiate between 7 interaction scenarios: (1) *Purely social*, (2) *Informative*, (3) *Educative*, (4) *Competitive*, (5) *Collaborative*, (6) *Negotiation*, and (7) *Guide-and-follow*. Some scenarios are possible for all interaction modes (e.g. competitive, informative), while others are restricted to embodied interaction (e.g. collaboration, negotiation, social).

Purely social [63] context is a social context that does not involve performing a task. In such context, social interaction may include greetings, self-introductions, or informal talking, etc. In conversational HCI context, temporal characteristics [64] composed of the user's past engagement state (temporal continuity), their current emotional state, and the other participants' engagement states were considered. Cognitive factors including attention and interest in the conversation were related to engagement.

Informative [63], [65] context entails transmitting general information that does not fall in the category of educating, e.g. giving navigation directions to reach a certain location.

Educative [66] context describes a form of learning which entails the transfer of knowledge or skills from an educator to a learner. Student engagement within the context of technology supported learning has been widely studied in the literature [67], [68], [69] from physiological signals or videos recorded during Massive Open Online Courses (MOOCs), or data collected in the class [70], [71], [72].

Competitive [19], [73], [74] context is characterized by elements of skills or knowledge testing or a form of competition over a certain profit (e.g. quiz, non-collaborative game, etc.). In such context, states of concentration and reflection might be triggered.

Collaborative [75] context involves a form of collaboration to achieve a pre-defined task. Such scenario is more prevalent in HRI, since physical embodiment allows a wider range of collaborative activities.

Negotiation [76] context involves the adoption of several strategies to achieve goals. Various parties confer and reach an agreement.

Guide-and-follow [77] context is concerned with cases involving a form of guidance to accomplish a specific task, where one party leads while the other follows the directions.

2.3 Personal Factors

Personal factors such as the user's gender, culture, age, ethnicity, personality, or if the user has a certain pathology were underlined in the literature as intrinsic factors affecting the engagement state of an individual [78]. From the system's perspective, specifically in embodied HMI, the system's personal factors (e.g. gender, personality) were also found to directly affect the user's engagement state. Other personal factors such as age and ethnicity are not explored in the literature of engagement inference. This might be related to the fact that most studies in engagement restrict their studies to specific age groups (e.g. adults or children), and

it is rare to find studies with datasets that include different age groups or different ethnicities.

Gender. The effect of gender on user's engagement was discussed in some studies [77], [79]. For example, the HRI study of [79] found that most participants preferred interacting with robots of the opposite sex, with a stronger gender effect when the participant was a male, and the robot is a female. This suggests that designing human-machine interactions that adapts the gender of the machine to that of the user would result in higher user engagement. Similarly, also in an HRI context, it was found that female and male users engage differently with robots [77]. Personalized models such as gender-specific models improved the accuracy of inference compared to general models in HCI scenarios [39].

Culture. Differences in social behaviour among different societies and cultures has been thoroughly underlined in the literature [80]. For instance, studies on emotion recognition have reported higher accuracy when cultural factors were taken into consideration [81], encouraging the adoption of a culture-sensitive approach in the assessment of emotions [82], and consequently emotion-dependent constructs such as engagement. Similarly, in HRI, an analysis of children engagement have revealed differences in engagement displays across different cultural backgrounds [83], which have been taken into account in the computational models of engagement [84]. A study on proxemics (a relevant cue of engagement) in HRI, particularly on robot approaching groups of people, have also shown different preferences in proxemics behaviour among different cultures [85]. In HCI, it was found that between- and within-country cultural differences have an impact on digital consumer engagement and engagement with online marketing material [85]. In the context of learners engagement, cultural factors were found to be correlated with organisational, technological, and pedagogical components of online learning [86].

Personality. The effect of personality on the engagement state is also evident in the literature. For instance, in contrast to gender, previous studies found that human interactions with a robot having the similar personality traits were perceived as more comfortable compared to interactions with a robot having different personality traits [78]. In a triadic HRI study, results showed a significant correlation between the perceived enjoyment with an extroverted robot and the participants' agreeableness and extroversion traits [87]. The effect of the user's personality regardless of the robot's personality on the user's engagement state was also investigated. Findings indicate that higher extroversion scores were correlated with longer interactions with robots [88]. High conscientiousness scores were associated with higher expression of attentiveness and responsiveness in interaction [89]. On the other hand, individuals scoring high on the agreeableness dimension reported higher enjoyment in interactions compared to others. In learning contexts, it was found that extroverted and introverted students exhibit different behaviours to indicate the same cognitive and affective states [42].

Pathology. The presence of certain pathologies can alter the way an individual behaves. For example, pathologies that have an effect on social behaviour include Autism Spectrum Disorder (ASD), Attention Deficit Hyperactivity Disorder (ADHD), Major Depression Disorder (MDD), Bipolar Disorder (BP), among others. For instance, individuals with ASD are characterized by deficiencies in demonstrating proper social cues in social interaction contexts [90]. Individuals with ADHD may exhibit an increased quantity of movement, in addition to an impaired capacity of sustaining attention on tasks. The characteristic social behaviours of individuals with such pathologies compromises a generic engagement model ability to accurately infer the human's engagement state. The fact that individuals attained with such pathologies exhibit unusually diverse styles in the expression of their affective-cognitive states makes the inference task even more challenging [91]. Consequently, integrating the pathology information or clinical assessments in engagement inference models might better inform the decisions of such models. Except for few approaches, the use of such information is scarce in the literature. For instance, clinical assessment in addition to culture, gender and individual traits information were proposed in [91] to condition autoencoders for the task of inferring child's engagement level continuously in time, in the context of robot-assisted autism therapy.

3 Automatic Inference of Engagement

In this section, we present a detailed summary of the existing approaches to automatic engagement recognition, including an overview of data acquisition for engagement inference, the multimodal behavioural cues commonly used as features for engagement in the literature and the employed machine learning approaches for the task, both traditional and modern solutions (e.g., deep learning).

3.1 Data Acquisition

Usually data is collected using unimodal or multimodal sensors such as microphones, cameras, 3D sensors (e.g. Kinect©), and physiological sensors. However, the choice of modalities depends on the context of application. Collected samples are then given ground truth labels reflecting the perceived or reported engagement state of the user. Engagement ground truth assessment largely depends on the context in which the engagement is being measured. Compared to emotion data annotation where categorical and dimensional scales are commonly used to assess emotions ground truth [92], there are no common scales used to collect engagement ground truth. Whether to treat engagement as a process, a discrete or a continuous value, or to concentrate on a specific component of the construct (e.g. behavioural vs. cognitive engagement) is highly dependent on the interaction context and end-goal. Moreover, there is no agreement on the optimal time scale for engagement annotation, e.g. frame-level or segment-level (cf. Section 4.2). Indeed, in the literature, there is no unified strategy for user engagement state annotation. In the following, we give an overview of the publicly available engagement datasets (Section 3.1.1) as well as common annotation strategies to obtain ground truth data (Section 3.1.2). We finalise with reviewing the problem formulation for engagement and the

definition of categories that commonly appear in the state-of-the-art (Section 3.1.3).

3.1.1 Publicly Available Datasets

Since engagement is a context-dependent construct and a relatively recent subject in human-machine interaction, there are only a few publicly available datasets in the literature, which provide engagement annotations. These datasets are given in Table 2 and are introduced based on the interaction mode below.

TABLE 2
Overview of publicly available datasets in engagement inference. Modalities: Audio (A), Video (V), Physiological (P), Log – data that captures interaction with the setup (L).

Mode	Dataset	Modality	# S	Context	Papers
HHI	RECOLA [93] (2013)	V, A, P	46	Collaborative	−−
HCI	DAiSEE [94] (2016)	V	112	Educative	[95]
	HBCU [40] (2014)	V	34	Educative	−−
	in-the-wild [69] (2018)	V	78	Educative	[96]
HRI	MHHRI [97] (2017)	V, A, P	18	Social	[98]
	PE-HRI [99] (2020)	V, A, L	68	Educative	[33]
	PE-HRI-temporal [100] (2021)	V, A, L	68	Educative	[11]
	UE-HRI [101] (2017)	V, A	54	Social	[102], [103]

HHI. The Remote Collaborative and Affective Interactions (RECOLA) database [93] provides engagement labels in addition to set of affective and social behaviour annotations including arousal, valence, agreement, dominance, performance, and rapport in a mediated interaction context. The participants in this corpus were recorded remotely in dyads during a video conference while completing a collaborative task (the survival task). In addition to video, the corpus includes audio and physiological data (ECG and EDA). Engagement annotations in this corpus are performed for each interaction session with a discrete Likert scale of $(1-7)$.

HCI. Existing datasets in HCI are mostly recorded in the context of online learning. For instance, the DAiSEE dataset [94] includes learner's videos captured while they were watching a video tutorial, with a webcam mounted on a computer. It was collected in unconstrained conditions at different locations with varying illumination settings. Learner's engagement level as well as relevant emotions (bored, confused, and frustrated) annotations on a scale of $(0-3)$ are provided for each video. Similarly, "in-the-wild" dataset [69] for engagement assessment includes student's videos collected via Skype in an unconstrained environment. Engagement levels were annotated using crowdsourcing in terms of 4 classes, including disengaged, barely, normally, and highly engaged. Finally, the HBCU dataset [40] includes student's engagement level annotations assessed via crowdsourcing on data captured as the participants were engaged in a cognitive skill training study.

HRI. The Multimodal Human-Human-Robot Interactions (MHHRI) Dataset [97] was introduced for studying the relationship between personality and engagement simultaneously in dyadic HHI and triadic HRI. The context of the dataset is purely social revolving around personal questions asked by the interaction entities to each other. The dataset was recorded using biosensors, Kinect depth sensors in addition to first-person vision cameras attached to the participants heads. The engagement state of the users was assessed with a post-study questionnaire asking the participants about their perceived enjoyment of the interaction. In a later study [98], labels from external annotators were obtained for this dataset using crowdsourcing. Another HRI dataset, the User Engagement in spontaneous HRI (UE-HRI) dataset [101], was presented to study spontaneous social interactions between humans and a robot. The dataset includes 54 dyadic HRI interactions with the robot Pepper situated in a public space, collected over a period of 56 days. The dataset was recorded using two 2D cameras, a 3D depth sensor, 4 directional microphones, sonar, and laser sensors. Recorded streams include face, speech, gesture, and dialog features. Engagement labels were obtained by external annotators.

In an educative HRI context, the Productive Engagement in HRI (PE-HRI) dataset [99] is a multimodal dataset that allows studying engagement in collaborative robot-mediated educational contexts. The dataset includes productive engagement scores which are computed via a linear combination of the most discriminatory features [104]. Additionally, the dataset consists of multimodal team level behaviours and learning outcomes (34 teams of two children). In a later version [100], the PE-HRI-temporal dataset was introduced where temporal features were computed in windows of 10 seconds for each team.

3.1.2 Engagement Annotation

An essential step for building a reliable engagement inference system is acquiring the ground truth data. Engagement ground truth labels are usually assessed via validated or self-designed questionnaires such as the Temple Presence Inventory (TPI) [105] which was adapted and employed in HRI & HAI contexts [98], [106]. Approaches to the collection of engagement ground truth labels can be divided in three categories: (1) *self-report* labels, (2) *external* measures, and (3) combination of self-report and external annotations.

Self-report Labels. These constitute pre- or/and during- or/and post-interaction self reports that gather information from the user about their experience with the technology. Using self-report evaluation of engagement can be considered as reliable since in theory one can truly know how they really felt during an experience. However, asking people to evaluate their experience after the experiment may be prone to error, as it relies on memory recall and on their attention to and communication of what made their sense of engagement to be perceived as powerful, or weak [107]. Moreover, another issue that arises with post-experience questionnaires is that they do not take into account the interaction dynamics. Interactions are dynamic and change over time, making engagement to fluctuate [6]. To account for these issues, some HRI works explored strategies to obtain the ground truth data from the users during the

interaction by introducing implicit and explicit probes for collecting self-reported engagement ground truth labels at different stages of the interaction [108]. Also in the context of interaction with an online learning platform, emotional engagement self-labels were collected during the interaction via two modes: (1) voluntary where the students can provide their engagement label at any time via a window that appeared during the entire interaction on the interface and (2) mandatory via a pop-up window that appeared at random time intervals [109].

External Annotations. These constitute recruiting a number of external annotators that assess an individual's level or state of engagement based on audio-visual recordings of the individual's interaction during the experience [110]. External evaluation by external observers might be prone to errors due to the difficulty of perceiving the true engagement state of the user. In addition, it is often difficult to reach an agreement on the perceived engagement due to annotators' subjectivity. This is a general problem involving any social phenomenon, and there is already a line of work focusing on obtaining reliable ground truth labels from multiple annotations. For instance, in HRI, independent models trained with different annotator's labels and then aggregated to obtain one integrated label [111]. Similarly, the subjectivity of the annotators was considered by modeling each annotator's latent character affecting their engagement perception using hierarchical Bayesian model in [112]. The annotator's character and engagement level are estimated as latent variables. Experimental results show such modeling outperforms baseline models which do not take into consideration the differences in annotations and annotator's characters.

Combining Self-report and External Labels. As discussed above, both self-report and external annotations used to assess engagement ground truth suffer from certain disadvantages. To account for this, few approaches have attempted to combine self-report and external annotations to obtain engagement ground truth data. The hypothesis is that engagement prediction accuracy would be increased with a combined label reflecting a ground truth closer to reality, which is a combination between the perceived and reported engagement states. In the context of conversational HAI, past attempts for obtaining a combined self-report and external engagement label was to sum the user's and observer's judgments (e.g., user is labelled as disengaged when assessed as such by both parties) [110], [113], [114].

3.1.3 Categorical View of Engagement

Most of the existing approaches to automatic engagement recognition formalise this problem as a classification task. It is crucial to identify what classes are relevant to the interaction and application context. For instance, one might be interested in detecting the user's intention to engage with one or all of the interaction parties, disengagement or the user's level of engagement. For this reason, we identify five categories of engagement in the state-of-the-art, summarised in Table 3.

Intention To Engage. Various engagement inference approaches concentrate on detecting of user's intention to engage in the interaction [57], [65]. Particularly, in HAI and HRI scenarios that include a social context, detecting the user's intention to initiate an interaction is of utmost importance for the agent to show an intelligent behaviour, and engage in a successful interaction. For instance, [57] focus on detecting user's intention to engage in an interaction with a robot using two classes: not seeking engagement vs. seeking engagement. Similarly, in [65] the occurrence of engagement intentions was modeled with two classes: E-intention (user wants to start a conversation with the agent or intends to speak, provided that the speaking floor was being held by others) and D-intention (user intends to disengage).

Engagement/Disengagement. Different approaches focus on binary classification of engagement, aiming to detect the presence or absence of engagement vs. disengagement. For instance, in [115], disengagement detection in individual and group interactions was investigated. On the other hand, linking engagement to attention, in [59] the focus was on the user's attention and lack of attention states.

Engagement Process. Several approaches focused on the classification of the phases of the engagement process, which constitutes mainly user's intention to engage, engagement, and disengagement. For example, in [116], [117], a model for recognizing 5 engagement classes was trained: no one, will interact, interact, leave interact, and someone around.

Engagement Level. Another line of work deals with the recognition of user's engagement level or degree defining the engagement state on a spectrum, e.g. low to high. Examples of defined engagement level classes from the literature are summarized in Table 3 together with the respective context they where defined in.

Behavioural Engagement. These approaches consider the behavioural aspect of the user during their engagement in an interaction. For instance, [118] distinguish different states of conversational engagement: no interest, following, responding, conversing, influencing and managing. Similarly, [15], [18] decomposed the engagement state into several mental, behavioural, and emotional states. In [119] two classes in addition to high and low engagement levels were explored, namely, "lead" referring to when the game leader was directing the conversation and "org" corresponding to when the group was forming itself. In the context of team interaction [120], the level of participation in a meeting was annotated with respect to six engagement states: disengagement (no participation, distraction and no attention to the meeting), relaxed engagement (attention to the meeting, listening, observing, but no participation), involved engagement (attention and non-verbal feedback), intention to act (preparation for active participation indicated through an increase in activity), action (speaking and/ or interacting with participants or content on displays), and involved action (intense gesture and voice).

3.2 Engagement Modalities and Features

In the literature, a rich set of features were explored, extracted from various data modalities including audio, video, text, and physiological data, for engagement inference in different interaction contexts. While different categorisations can be found such as static vs. dynamic features [125], or low-level vs. high-level features [126], [127], we divide the engagement features into five categories: (1) contextual

TABLE 3
Overview of the engagement categories that are widely used in the literature. IM: Interaction Mode

Category	Paper	Classes	Context	IM
Intention To Engage	[57]	Not seeking engagement, Seeking engagement		HRI
	[65]	Engagement intentions: E-intention , D-intention , Attention saliency		HAI
Engagement/Disengagement	[115]	Disengagement		HRI
Engagement Process	[116]	Intention to engage, Engaged, Disengaged		HRI
	[63]	User is present, Interacting, Engaged, Just attending	Robotic Receptionist	HRI
	[117]	Will interact, Interact, Leave interact, No one, Someone around		HRI
Attention Level	[121]	Distracted, Tired/Sleepy, Not paying attention, Attentive, Full of interest, Curious	Educative	HCI
Attention/Frustration	[48]	Attention/Non attention, Frustration/Non Frustration		HCI
Attention	[59]	Attention, In-attention		HRI
Engagement Level	[19]	Medium-high to high engagement, Medium-high to low engagement	Competitive	HRI
	[119]	Group involvement: High, Low, Lead, Group formation	Collaborative game	HHI
	[122]	Engaged in the interaction, Superficially engaged with the scene and action space, Uninterested in the scene or action space	Agent salesperson	HAI
	[123]	High and low engagement	Educative	HRI
Interest Level	[124]	High interest, Low interest, Refreshing, Bored, neutral, Other	Competitive	HCI
Engagement Behaviour	[118]	Conversational engagement: No interest, Following, Responding, Conversing,Influencing, Managing		HHI
	[18]	Intends to Engage, Listening, Concentrating, Responding, Positive reaction, Negative reaction, Waiting feedback, Thinking, Disengaged	Social, Informative Competitive	HRI

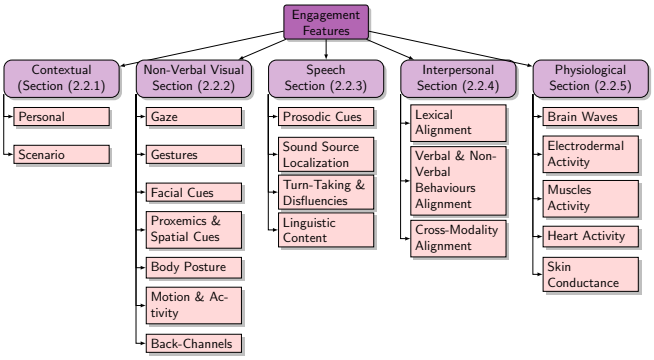

Fig. 2. Engagement modalities and features categories.

cues, (2) non-verbal visual cues, (3) speech cues, (4) interpersonal cues, and (5) physiological cues. These categories are depicted in a feature tree in section 2.

Our literature analysis of engagement modalities and features has shown that there is no observable trend of using specific features in specific contexts, as this depended on the available modalities and data, and engagement categories/scores, which differed from one study to another. Nevertheless, in our review of contextual cues (Section 3.2.1), we have found an important difference in the used contextual features among different interaction contexts (disembodied vs. embodied HMI). Consequently, in this section, we provide an overview of the features that are most commonly used in the literature of engagement inference. We include the context in which the different features are used, whenever it is relevant. We also discuss the difference in the used contextual features among different disembodied vs. embodied HMI interaction contexts.

3.2.1 Contextual cues

We have discussed in Section 2 the importance of context for engagement. In this section, we review how context was used in the engagement inference literature, and what contextual features were extracted for this purpose across the different interaction modes (disembodied and embodied HMI) and scenarios.

Disembodied HMI. Contextual features have been extracted depending on different **interaction scenarios**. In an HCI *competitive* game context, sequence mining was applied to combine physiological signals and game-context information for the prediction of computer game player affective states [128]. For detecting a user's attention level for tasks that commonly occur in a workplace setting, machine-specific contextual features were fused with person-specific features. The contextual features were collected from the keyboard, the mouse, and the active window size [50]. In an *educative* HCI scenario involving children engaged in solving a puzzle on a computer, the game state was used as contextual features, and combined with facial and posture features for engagement inference [66]. In the context of interaction with a learning platform, contextual information were extracted from the URL logs and combined with appearance features to predict whether the user is engaged with the platform by analysing on-task vs. off-task behaviours [129].

Embodied HMI. In terms of **interaction scenario**, in an HRI *competitive* game scenario (robot vs. child), task and social-based contextual features in the form of robot's social behaviour, game events, and game progress were explored for the inference of engagement [19] [130]. In a multiparty HRI scenario with varying contexts including *informative*, *competitive*, and *social*, [15], the robot's and other participant's behavioural cues were used as contextual information to detect the user's engagement state. The study has shown that behavioural cues from the other parties (robot and other participant) can be used as predictors of the engagement state of the user in question. This might happen when in multi-party interactions, there is a significant cohesion and synchrony within the group, which might signal high engagement. Similarly, in [131], to integrate context into the proposed framework of engagement inference, the current user's engagement is modeled as a function of

previous user's engagement state. The approach is validated in the context of *negotiation* scenario (i.e., multi-agent job interview) and a *social* robot scenario.

In terms of **personal factors**, a line of work investigated the impact of personal cues such as individual's *personality* as well as their interaction partner's personality on the inference of engagement [98] or rapport [132]. Salam *et al*. [98] focused on a triadic HRI interaction scenario (i.e., between two humans and a robot) and showed that using individual's personality traits only was sufficiently informative to detect their engagement, and combining those with interpersonal features resulted in further improvement in the accuracy of engagement inference. Cerekovic *et al*. [132] analysed the impact of personality on the perception of rapport in an HAI scenario. Participants interacted with two different agents, displaying different characteristics (i.e., cheerful vs. gloomy), and they found that extroverted and agreeable people tend to report higher levels of rapport with both agents. *Culture* and *gender* were used implicitly for personalizing engagement recognition deep architectures to specific population of differing gender or culture [84], [133], [134]. We touch into this more in detail in the Section 3.3.2.2, reviewing personalized deep learning models for engagement inference.

Despite evidence on the predictive power of contextual information for engagement inference, however, methods using contextual features remain scarce in the literature. Devillers *et al*. [135] highlight the importance of taking into account context in the assessment of engagement in HRI, namely the behaviour of the robot, the interaction dynamics, and the dialogue participants communicative behaviour variations. They identified linguistic, paralinguistic, interactional, non-verbal, and specific emotional and mental states features to be of utmost importance for engagement prediction.

3.2.2 Non-verbal Visual Cues

Non-verbal cues comprise social signals exchanged during social interactions along with the uttered verbal words. This type of behaviour is the most informative, when individuals can easily verbally pretend that they are engaged in a task or in a conversation. However, it is much harder to fake non-verbal behaviour which is indicative of engagement. Some researcher even claim that initiating and maintaining engagement is fully possible without verbal conversation [7]. Features in this category include visual cues (eye gaze, gestures, facial expressions, proxemics and spatial features, body posture, and motion and activity features), and audio cues (prosodic, sound source localization, turn-taking, and back-channels). Commonly used non-verbal visual cues include: (1) gaze, (2) gestures, (3) facial cues, (4) proxemics and spatial cues, (5) body Posture, (5) motion and activity, and (6) non-verbal back-channels.

Gaze. Eye gaze is the mirror to the mind and is strongly coupled with cognitive and emotional processes. It is considered as the primary cue of attention [136], a cognitive component of engagement. For instance, engagement with a speaking conversational partner can be signaled by gaze cues such as looking at them. On the other hand, looking away from them for long periods indicates disinterest and consequently a disengagement intention [137]. Gaze is among the most commonly used features for inferring engagement across various interaction contexts. Employed gaze features include gaze direction [19], [43], [59], [138], visual focus of attention [58], [98], gaze transition patterns [110], [113], [114], the amount of time the user's gaze was directed at the interaction partner [13], [60], gaze following of the subject of speech [26], and mutual gaze [54], and gaze fixation duration [139]. Eye gaze and head orientation are closely linked and often one accompany the other. Pointing gestures can also contribute to the computation of the other's direction of attention [140]. In some scenarios, it might be difficult to detect the eye gaze accurately (e.g., due to low resolution images, adversarial head poses, or occlusions). Several researchers, opted to approximate the user's gaze by the head pose [13], [30], [43], [57], [116], [141], especially in HRI where there is a big distance between the robot and the interactant.

Gestures. Gestural behaviour in interactions conveys a lot information on the engagement state and on the connectedness between participants [142]. Moreover, in HAI and HRI the generation of appropriate gestures via the agent/robot and the successful interpretation of human gestures have a significant effect on the user's engagement and consequently the interaction success [143]. Relevant gestures include head gestures such as head nods and shakes (convey (dis)agreement and attention) [66], [141], and hand gestures and speed such as hand raising (convey intention to engage) [120], [144], [145].

Facial Cues. As discussed in previous sections, engagement has an affective component. Consequently, facial expressions and emotions are important features for detecting user's engagement state across various interaction modes and scenarios. Some approaches focused on a small class of expressions, such as the smile [19], [120] or the probability of smile and fidget [66]. Other approaches detected multiple emotion states [146] and facial expressions such as smiling, surprised, neutral, and angry states [59]. Fine-grained facial movements such as the displacement and velocity of the mouth and eye landmarks [147], or facial action units (AUs) were also used as engagement features [40], [42], [148]. [39] concluded that eye's region is more informative than mouth's region. Eye physiological indices such as pupil diameter and blink rate and eye behavioural indices are also relevant to engagement inference [139].

Proxemics and Spatial Cues. Spatial behaviour (proxemics) constitutes the dynamic process by which individuals position themselves in social interactions [149]. Proxemics carry significant attentional and interpersonal factors [150], which are indicative of an individual's engagement state. For instance, being at a relatively small distance indicates higher probability of paying attention to a robot [63]. Holthaus *et al*. [151] proposed a spatial model for a receptionist robot to infer the user's intention. Other features to consider are the relative distance between the interaction partners [98], or between the user and the robot or the machine [42], [59], [152], as well as the trajectory or speed of a person such as walking towards a robot. The size of the face detected by the camera can also be used to approximate the distance between a person and a robot [43], [57], [116]. Detecting if the face is currently invisible, and for how long was also used to indicate a person's spatial

situation [55].

Body Posture. Previous research has shown that the face's and body's orientation towards an interlocutor is a signal of engagement [56], [153]. The body orientation and relaxation were also found to be significant indicators of the communicator's liking for their addressee [154]. The dynamics of postural shifts might even signal shifts or changes in the conversation topic (boundaries of information units) during an interaction [155]. Moreover, as the posture is often displayed unintentionally, this makes it an effective indicator of the user's real affective state [156], [157], [158], as well as their attitude and alertness [159]. Various approaches employed body posture features in the literature of engagement inference [124]. Among such features, the upper body orientation, the back's curvature, and the upper body degree of contraction and expansion (contraction index) were used in the context of human-child game interaction [61]. Naturally occurring seated postures were also used as features for child engagement during a computer learning task [160]. Body lean angle, slouch factor, body direction, hand vertical position, and posture were also explored to detect engagement in team meetings [120] and student engagement with MOOCs [161]. Other features include hand pose and body orientation [57], upper body posture and body openness [162], the position and orientation of the feet, hips, and torso and the shoulders positions and orientation relative to each other [116], the relative orientation between the participants and the robot [98].

Motion and Activity. Previous studies found correlations between the user's quantity of movement and their degree of involvement in an interaction [17], [163]. It is actually possible to judge users' level of engagement by measuring their movements as they use a computer [17]. For instance, increased user movement can be an indicator of an increased involvement, consequently an increased engagement state [163]. On the other hand, the absence of certain movements can signal cognitive engagement in seated situations [17]. In [17], the authors distinguished between instrumental movements referring to physical movements serving a direct goal in a given situation, and non-instrumental movements that are involuntary tiny movements that people usually exhibit to reflect the person's internal states. If someone is absorbed in what they are watching or doing, referred to as "rapt engagement" by [17], there is a decrease in these non-instrumental movements. Motion features used for engagement inference include the Quantity of Motion (QoM), which is a measure of the amount of detected motion. QoM was used by [61], [98] to infer engagement with a robot companion. The approach of [98] also detected the global QoM of a multi-party interaction for group and individual engagement inference. The body activity computed from skeleton joints of upper body bounding boxes was also used in this context [98].

Back-Channels. Back-channels correspond to events where an interaction party responds back to a primary communication initiator with a brief verbal or gestural communication [54]. Example non-verbal back-channels include head nods and shakes signaling to the initiator that the responder understands, listens or desires to continue the conversation. In HHI, head nods were combined with other non-verbal features to detect team engagement in meetings [120]. In HRI, [13] a robot was endowed with the capability of interpreting nods as back-channels and agreements in conversation in order to recognize the user's state of engagement. Back-channels together with laughing and nodding were also found to be related to the level of engagement in social HRI scenario [15].

3.2.3 *Speech Cues*

Speech cues are used in the context of HMI for engagement recognition, since some spoken linguistic behaviour might convey engagement [13]. Speech cues can be extracted from verbal speech, or written text. Other than the linguistic content, various non-verbal signals are encoded in speech. The analysis of written or spoken language is interesting for the context in question because it permits a straightforward way to give an input to the engagement system: greetings can be considered as a cue for intention to engage [30], [137], while closing comments may signify the disengagement of the user [13]. Using verbal behaviour, the user has an active role during the interaction in a way that would not be possible if only non-verbal visual behaviours were considered. Embodied HMI (HRI and HAI), for instance, can be more natural if humans can speak directly with robots/agents. Several works consider acoustic and text-based data to obtain information about the user's engagement [164]. These can be categorized into: (1) prosodic cues, (2) sound source localization, (3) turn-taking and disfluencies (4) and linguistic content.

Prosodic Cues. Previous research [163] observed a relationship between the voice's prosodic characteristics (level, span, intensity) and involvement. The authors of [64], for example, used sound and prosodic features such as the speech rhythm, stress, and intonation, to estimate conversational engagement level between a voice communication system users. Another work [120] used the speech volume in combination with other non-verbal features to predict team engagement in meetings.

Sound Source Localization (SSL). Detecting the sound source can be an indicator of user's intention to engage, engagement or disengagement. For example, the robot's ability of localising someone next to it from the sound, identifying speech activity or even prosody allows it to recognise the engagement state of a user. SSL was used in the literature for locating the user's voice or footsteps, allowing to detect the direction in which the user approaches a robot, which indicates their intention to engage [116]. SSL in an HRI setting [165] was also used to estimate user's intention to engage and engagement level in [164].

Turn-Taking. The notion of turn-taking also plays an important role in the context of conversational HAI and HRI. As Sidner *et al*. [137] observed, failure of an interaction party to take an expected turn is an indication of disengagement. Thus, related to this notion, adjacency pairs which consist of "two utterances by two speakers, with minimal overlap or gap between them, such that the first utterance provokes the second utterance" were used as an engagement cue in [54]. Similarly, engagement annotation of conversational data was studied in the context of HRI, where it has been demonstrated that engagement level is correlated with turn-taking behaviours such as the duration of the next turn [166]. Turn-taking features employed in

the literature include speech and pause duration statistics, speaker change with and without overlap, successful and unsuccessful interruption, and speech overlap [167].

Disfluencies, such as filled pauses (also referred to as fillers), short and long speech pauses between words, or hesitations were also considered in the literature of engagement detection and management. For instance, combined with gaze, filled pauses (e.g. "uh" or "umm") were used to detect whether a user wishes to disengage from the interaction with a robot [168]. In [55], linguistic hesitations (e.g. filled and non-filled pauses) were used for managing conversational disengagement in HRI when uncertainty about whether the user wishes to stay engaged in the interaction arises. In a robot-mediated collaborative learning context involving teams of two children, speech behaviors including the team members speech overlap or the amount of their long pauses were found to be the most discriminating between teams exhibiting high learning engagement and teams exhibiting low learning engagement. An engagement score was generated using a linear combination of such behaviours [104].

Linguistic Content. Certain spoken words can signal engagement and related affective states [169]. The use of linguistic content in engagement inference is scarce in the literature. Example works include [57] who presented a method for speech recognition using grammar implementation. The proposed method aimed to extract syntactic and semantic information from the user's speech to detect engagement. In [42] a tutorial scenario was proposed to predict engagement and learning, where the user was able to send textual messages to a virtual agent.

3.2.4 Interpersonal cues

Interpersonal features represent the interpersonal behaviours of the interaction parties relative to each other. The cues discussed in the previous sections can be considered as individual cues that are extracted by isolating the interaction partners or action-reaction processes where interaction partners exhibit a behaviour in response to each other (e.g., turn-taking, back-channels). In this section, we look at more elaborate cues such as synchrony or alignment. Synchrony is defined by [170] as "the dynamic and reciprocal adaptation of the temporal structure of behaviours between interacting partners". Unlike mirroring or mimicry, synchrony is dynamic in the sense that the important element is the timing, rather than the nature of the behaviours [170]. Several studies in HMI have referred to the importance of synchrony for engagement inference. Detecting synchrony between the interaction parties involves not only the detection of their reactivity, but also their agency and their engagement within the ongoing interaction [171]. For instance, Prepin and Gaussier [172] showed that synchrony is a viable indicator of the user's satisfaction and level of engagement during an interaction with a robot. The better the interaction is, the more the human is synchronous with the robot. They designed a robotic architecture that can detect temporal synchrony between the user and agent's actions and use this parameter to adapt the robot's behaviour using reinforcement learning.

Three categories of approaches that use participant's synchrony for engagement recognition can be identified in the literature: (1) lexical, (2) verbal and non-verbal behaviour, and (3) cross-modality alignment.

Lexical alignment refers to the adoption of one's interlocutor's lexical items [173]. State-of-the-art approaches investigated lexical alignment by detecting the use of shared vocabulary and verbal repetitions [174] as a cue of engagement in an attempt to develop a virtual agent capable of employing alignment strategies to maintain user's engagement.

Verbal and non-verbal behaviour alignment refers to the alignment or synchrony between participants verbal or non-verbal cues during a social or collaborative interaction. Studies in HHI found similar inphase/antiphase dynamic among interaction parties engageing in an interpersonal task [175]. In the context of multi-party HRI, Salam *et al*. [15] showed that it is possible to detect the engagement of an interaction party using as input the behavioural cues of the other interaction parties (robot and other participant) with an acceptable accuracy compared to solely using the interaction party individual behavioural cues. Such approach can be considered an indirect way of using synchrony features since the high correlation of others' cues with the engagement of the participant in question means that they were in synchrony with her. They [15] also extracted a set of features describing the synchrony and alignment between robot's and participants' behaviours. These include, among others, mutual gaze and laughter. Other features included events where a participant speaks with the other during the speech of the robot. This may signal a disengagement behaviour as it means that the participant is not listening to what the robot is saying.

Cross-modality alignment refers to the alignment or synchrony between different modality cues such as speech and gaze, or speech and gesture. Findings suggest that gazing towards objects relevant to the conversation is an indicator of engagement [137]. For instance, in the context of a multi-party HRI scenario [15], events where one participant looks at the other who is speaking to the robot or events where a participant looks at objects corresponding to the topic of robot's speech (i.e., gaze-speech alignment) were used as predictors of the user's engagement state. In [131], a study was performed in the context of two scenarios: a social robot scenario and a multi-agent job interview scenario, proposing to model the interpersonal cues dynamics that reflect the social attitude of the user with the context. The user's engagement was considered as a combination of the user's individual behaviour patterns and their interpersonal behaviour patterns as well as their temporal alignment.

3.2.5 Physiological cues

Previous research provided evidence on the relationship of physiological signals with affective and cognitive states. Composed of both affective and cognitive components, engagement can be successfully predicted from physiological responses. Various state-of-the-art methods have employed physiological signal analysis for the inference or analysis of engagement or any of its related constructs. For instance, in [176] physiological features were extracted from Electrocardiogram (ECG), electrodermal activity (EDA), and Photoplethysmography (PPG) signals to differentiate between boredom (associated with disengagement), pain and

surprise. In [138], EEG signals were used with the goal to capture brain activity for understanding the link between intention and attention. In [49], EEG, ECG, and Heat Flux (HF) signals were investigated for the identification and evaluation of engagement level variations during cognitive tasks. In the context of tele-operated HRI, human engagement (attention) on a specific task was also detected from physiological signals [177]. The extracted features included cardiac activity, heart sound, Bioimpedance, EDA, muscle activity and Skin Temperature (SKT). Results showed that the correlations between engagement ground truth labels and features were not similar among all participants indicating that people do not express engagement in the same manner. In [39], video-based heart rate (HR) measurement combined with appearance and geometric features was explored for engagement inference. In [178], it was found that a notable percentage of the variance for task engagement can be predicted from ECG, electro-oculogram (EOG), skin conductance (SC), and respiratory rate (RR) signals. In [179], it was shown that blood volume pulse (BPV) amplitude varies significantly among task and resting periods, which signals variations in sympathetic activity when engaging with a task.

The advantage of using physiological signals for predicting the affective and cognitive components of engagement is that they are able to provide precise measurements, since they give an objective insight on the true state experienced by the user. However, their intrusive nature requiring to attach various sensors to the user's body limits their use in natural human-machine interaction settings and their scalability to a wide range of applications. Another important issue with physiological signals comes from the inter-intra individual variability.

3.3 Engagement State Inference

Engagement state inference methods can be classified into three different categories: (1) rule-based, (2) supervised learning and (3) unsupervised learning. The employment of traditional machine learning techniques was more prevalent prior to the emergence of deep learning methods and their success in various pattern recognition problems. On the other hand, unsupervised learning approaches to engagement inference were also explored in the literature. However, these are limited to educative HCI contexts.

3.3.1 Rule-based Approaches

Rule-based approaches for engagement inference infer a decision on the user's engagement state based on a set of custom-defined rules (IF some condition THEN some action). Most of the rule-based engagement inference methods can be found in HRI contexts. This is mainly because rule-based methods are less expensive in terms of resources requirements, making them more suitable for integration is real-time settings. Sidner *et al*. [143] implemented one of the first robotic architectures that is endowed with engagement rules to generate appropriate engagement behaviour for the robot, and to detect the human's engagement state enabling a successful collaborative conversation between the human and the robot. Similarly [13] endowed their robot with rule-based engagement inference capabilities. The robot judges the user's engagement based on the head's position which indicates if the user is looking at the robot, at objects necessary for the collaboration or to other objects or to empty space. Similarly, [180] implemented a multimodal rule-based real-time robotic architecture that detects user engagement based on movement detection (head and body tracking), face recognition, gaze direction, proxemic distance, and audio cues (sound direction localization, audio signal power). Other works that employed rule-based engagement inference methods include [54], [58], [57], [55], [152]. Comparing rule-based user engagement classification to machine learning methods, [57] found that trained classifiers were faster and more accurate at detecting the user's intention to engage, while the rule-based approach resulted in more stability.

An important advantage of rule-based engagement inference approaches is explainability, an important aspect to take into account to avoid potential unwanted bias towards certain protected social groups (e.g. gender or race). Other advantages include the simplicity and rapidity of implementation, and the non-necessity of large training datasets. On the other hand, such advantages come at the cost of ignoring important complex patterns that can be automatically discovered from data by machine learning models, compromising the accuracy of rule-based methods.

3.3.2 Supervised Learning Approaches

Supervised learning methods were widely used for engagement inference in the literature. We categorize the employed supervised learning methods into two categories: (1) traditional machine learning, and (2) deep learning approaches.

3.3.2.1 *Traditional Machine Learning Approaches*: A range of classification techniques have been used to classify multi-modal observations into one of the pre-defined engagement classes. The performance depends on the general framework and used modalities. Examples include Support Vector Machine (SVM) [19] [42] [138] [116] [40] [39], GentleBoost, AdaBoost [40], Multinomial logistic regression [55] [40], General Regression Models (GRM), C4.5 (decision tree) [39], and Random forests [49], Dtree and OneR [61], boosted decision trees models [55], Fuzzy min-max neural networks (FMMNN) [59]. Comparing various classifiers, [57] concluded that trained classifiers are faster and more efficient compared to rule-based methods. However rule-based methods are more stable (prediction variables vary less frequently during an interaction). Moreover, in order to obtain better results, it's possible to add some temporal features to the states using Conditional Random Field or HMM. For instance, employing a multilevel structure with coupled HMM led to suitable performance for engagement inference in a usual daily conversations [64]. In [160], learner's engagement level was estimated using a combination of neural networks and Hidden Markov Models. Other probabilistic methods that were employed in the literature include Bayesian inference methods [60], [162] and Sugeno-type fuzzy inference system [48].

3.3.2.2 *Deep Learning Approaches*: The rise of deep learning and its track record in achieving high performance for various affective computing problematics has led researchers to propose various deep learning approaches for

engagement recognition. A clear added value of deep learning is their ability to learn new featureres presentations. This is also a nice way to learn mixed representations using contextual information, which is harder with other approaches. The most recent deep learning approaches concentrate on students engagement inference in online learning. This is due to the recent shift to online learning due to the Covid-19 pandemic. Deep learning approaches for engagement inference either use raw data, or behavioural cues as input to deep architectures.

In HRI, various architectures composed of Convolutional Neural Networks (CNNs) and LSTM were proposed for engagement classification or regression. For instance [181] proposed to classify engagement using body pose estimated from multiple RGB depth cameras as input to an LSTM layer. Other architectures composed of CNNs followed by LSTM [8] were also employed to predict a continuous engagement score from robot-view video streams.

In online learning context, an approach for student engagement prediction has been proposed by fusing facial and body features into a single long short-term memory (LSTM) model [182]. Dilated Temporal Convolutional Networks (TCN) has also been proposed for predicting student engagement intensity [183]. It has been demonstrated that TCN capture long term dependencies and it outperforms other sequential models like LSTMs.

A recent trend, is the development of personalized engagement models. ResNet-50 architecture [84] was proposed to train culture-wise personalized engagement models (CultureNet) from face images for engagement inference in the context of robot-assisted therapy for autistic children. [91] proposed a personalized multitask learning framework to simultneously predict engagement, valence and arousal. The network learns behavioural multimodal (visual, audio, physiological) features representations using autoencoders. Personalization with respect to culture, gender and each individual is performed using different fine-tuning strategies.

3.3.3 Unsupervised Learning Approaches

There has been an interest in developing unsupervised learning approaches to engagement inference, although mainly these approaches are based clustering techniques and are limited to HCI and education context. These approaches aim to discover learners' engagement patterns including engagement state, level or style from data, e.g., system log files.

Engagement State. A line of work has focused on identifying patterns corresponding to a binary engagement state of the learner, namely, engagement vs. disengagement. For instance, in the context of MOOC, Coffrin *et al*. [184] were able to differentiate between engagement and disengagement in the course based on student's learning analytics including histogram of student's performance and weekly student's participation.

Engagement Level. Some approaches looked into detecting the level of user's engagement with the learning environment. For instance, in the context of exploratory learning environments [185] unsupervised clustering (k-means) were applied to identify interaction patterns corresponding to different levels of learner engagement from gaze and context-based features reflecting students actions during their learning experience. Two engagement levels were identified, namely, high and low learners.

Engagement Style. These approaches aim to identify learners engagement style or behaviour based on interaction patterns with the learning environment. For example, in [186], [187], learners were clustered based on their degree of lectures and assignment completion. In the context of robot-mediated learning [11], approaches were designed using forward and backward clustering from the multimodal behavioural features to the learners learning outcome metrics and vice-and-versa, allowing the identification of learner profiles (gainers vs. non-gainers). An engagement score was then generated using a linear combination of the most discriminating behaviors between the identified gainers and non-gainers profiles [104].

4 Discussion and Open Questions

In this section we discuss the main points presented in this survey paper as well as open questions in automatic engagement perception and modeling which remain underexplored and deserve attention.

4.1 Context-aware Engagement Modeling

One important question, related to context-aware engagement modeling, is how to incorporate contextual information in the automatic engagement inference systems. For instance, in HRI, the work of [135] highlighted the importance of accounting for high-level and low-level communication processes when measuring engagement. In addition to communicative behaviour cues (e.g., visual, linguistic), and interpersonal cues (communicative behaviour of interaction parties w.r.t each other), it is necessary to consider the interaction dynamic (i.e. variations in interaction parties communicative behaviour), as well as contexts (e.g., human-robot relationship, social, situational, human profile).

As discussed in Section 3.2.1, most of the existing approaches incorporated contextual information in the form of features extracted from the task or from the interaction parties' behaviours. However, research has shown that the context of the interaction may evolve over time, impacting the interaction parties' behaviours and consequently, the engagement models' accuracies [18]. Regardless of the ultimate interaction goal, various interpersonal sub-contexts (e.g. social, informative, etc.) [18] might emerge during the same interaction, which might trigger different cognitive, emotional and behavioural user states, indicative of their engagement state. While previous studies relate some mental states to engagement, the literature lacks a clear indication of when a certain state (e.g emotional, cognitive) is a significant indicator of engagement in a certain context. It is thus necessary to understand how the engagement cues emerge and fluctuate during the same interaction in relation to the sub-context, and to consider such variations when developing automatic engagement inference models.

4.2 Temporal Dynamics of Engagement

Previous works have indicated that engagement is a dynamic process that changes over time and that is dependent on the participants of a continuously evolving interaction [20], [135]. This dynamic aspect was not thoroughly

studied in the literature of automatic engagement inference. As a matter of fact, we find different ways by which researchers segment their videos and address the problem of timescale without giving any justification on why the specific time scale was chosen. For instance, [117] detected engagement in 80 milliseconds video segments, [65] used 0.5 second fragments, and [124] segment their videos into a maximum of 8 seconds fragments. One study [40], has addressed this matter in the context of student engagement and performed a comparison between 1 frame, 10 second segments and 1 minute segments engagement labeling by external annotators. They found that the labeling task was easier and more reliable (high inter-rater agreement) when annotators labeled 10 second video clips, and that a reliable prediction of engagement labels of 10 second video segments can be obtained from the average of their constituting frames labels. Consequently, an important aspect to take into consideration in engagement inference is the optimal observation window in which engagement can be detected. Relevant research questions that are still open for investigation include the following. Is it sufficient to perform a *static* (frame-level) inference? If yes, is this achieved by re-using clip-level engagement label as the labels for the constituent frames? What are the advantages and disadvantages of this? Or a *dynamic* (segment-level) inference is more relevant? In the case of dynamic inference, what is the optimal time window, and is this context-dependent? Is engagement in specific time segments or frames affected by the users' past behaviors, and if yes, to what extent? Such questions merit further investigation in future studies.

4.3 Personalised Models and Bias

As discussed in Section 3.3.2.2, a recent trend in engagement prediction systems is the development of personalized models. In such frameworks, user profiling w.r.t personal factors (cf. Section 2.3) can be performed prior to training the models. Such information can then be used at the level of the data, or within the machine learning process to adapt the models to the specific profiles. Compared to one-fits-all approaches which are simpler to train but can compromise the models accuracy and adaptability, personalized models seem to be promising. However, they remain underexplored. One concern that arises in this context is the problem of bias and potential unfairness of personalized models decisions to certain social groups (e.g. gender, age). Generic models of affect were shown to present certain biases to such groups [188], if not properly tackled [189]. However, bias investigation and mitigation in generic engagement models is still not explored in the literature. Moreover, the question of whether personalization of automatic engagement inference systems increase or mitigate bias and fairness remains open, and merits further investigation.

5 Conclusion

In this paper, we presented a context-driven survey on engagement in human-machine interaction across various modes of interaction (HHI, HCI, HRI, and HAI). We reviewed more than 200 papers and we introduced widely used engagement definitions, available datasets, widely used features, and machine learning approaches. Engagement is a key component of social intelligence. We believe our survey will be a helpful guide for researchers working or planning to work on the problem of engagement inference, and aiming to equip machines with social intelligence. We finalised our survey with discussions and open questions to present our insights into how to advance this area of research further.

References

[1] A. Ikedinachi, S. Misra, P. A. Assibong, E. F. Olu-Owolabi, R. Maskeliūnas, and R. Damasevicius, "Artificial intelligence, smart classrooms and online education in the 21st century: Implications for human development," *Journal of Cases on Information Technology*, vol. 21, no. 3, pp. 66–79, 2019.

[2] A. Thieme, D. Belgrave, and G. Doherty, "Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems," *ACM Transactions on Computer-Human Interaction*, vol. 27, no. 5, pp. 1–53, 2020.

[3] J. Broekens, M. Heerink, H. Rosendal *et al.*, "Assistive social robots in elderly care: a review," *Gerontechnology*, vol. 8, no. 2, pp. 94–103, 2009.

[4] K. Tatarian, R. Stower, D. Rudaz, M. Chamoux, A. Kappas, and M. Chetouani, "How does modality matter? investigating the synthesis and effects of multi-modal robot behavior on social intelligence," *International Journal of Social Robotics*, pp. 1–19, 2021.

[5] E. Goffman, *Behavior in public places: Notes on the social organization of gatherings*. Free Press New York, 1963.

[6] H. L. O'Brien and E. G. Toms, "What is user engagement? a conceptual framework for defining user engagement with technology," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 6, pp. 938–955, 2008.

[7] C. L. Sidner and M. Dzikovska, "Human-Robot Interaction: Engagement between humans and robots for hosting activities," in *4th IEEE International Conference on Multimodal Interfaces*, 2002, pp. 123–137.

[8] F. Del Duchetto, P. Baxter, and M. Hanheide, "Are you still with me? continuous engagement assessment from a robot's point of view," *Frontiers in Robotics and AI*, vol. 7, 2020.

[9] L. El Hamamsy, W. Johal, T. Asselborn, J. Nasir, and P. Dillenbourg, "Learning by collaborative teaching: An engaging multiparty cowriter activity," in *28th IEEE International Conference on Robot and Human Interactive Communication*, 2019, pp. 1–8.

[10] M. Dewan, M. Murshed, and F. Lin, "Engagement detection in online learning: a review," *Smart Learning Environments*, vol. 6, no. 1, pp. 1–20, 2019.

[11] J. Nasir, A. Kothiyal, B. Bruno, and P. Dillenbourg, "Many are the ways to learn identifying multi-modal behavioral profiles of collaborative learning in constructivist activities," *International Journal of Computer-Supported Collaborative Learning*, pp. 1–39, 2022.

[12] A. Rapp, "Time, engagement and video games: How game design elements shape the temporalities of play in massively multiplayer online role-playing games," *Information Systems Journal*, vol. 32, no. 1, pp. 5–32, 2022.

[13] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, no. 1, pp. 140–164, 2005.

[14] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. Mcowan, "Context-sensitive affect recognition for a robotic game companion," *ACM Transactions on Interactive Intelligent Systems*, vol. 4, no. 2, pp. 1–25, 2014.

[15] H. Salam and M. Chetouani, "Engagement detection based on mutli-party cues for human robot interaction," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 341–347.

[16] G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles, "Towards a better understanding of context and context-awareness," in *Handheld and ubiquitous computing*, 1999, pp. 304–307.

[17] H. J. Witchel, C. E. Westling, J. Tee, A. Healy, R. Needham, and N. Chockalingam, "What does not happen: Quantifying embodied engagement using nimi and self-adaptors," *Participations*, vol. 11, no. 1, pp. 304–331, 2014.

[18] H. Salam and M. Chetouani, "A multi-level context-based modeling of engagement in human-robot interaction," in *11th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, vol. 3, 2015, pp. 1–6.

[19] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. McOwan, "Detecting engagement in hri: An exploration of social and task-based context," in *International Conference on Social Computing*, 2012, pp. 421–428.

[20] N. Glas and C. Pelachaud, "Definitions of engagement in human-agent interaction," in *International Workshop on Engagment in Human Computer Interaction (ENHANCE)*, 2015, pp. 944–949.

[21] C. Oertel, G. Castellano, M. Chetouani, J. Nasir, M. Obaid, C. Pelachaud, and C. Peters, "Engagement in human-agent interaction: An overview," *Frontiers in Robotics and AI*, vol. 7, p. 92, 2020.

[22] C. Lytridis, C. Bazinas, G. A. Papakostas, and V. Kaburlasos, "On measuring engagement level during child-robot interaction in education," in *International Conference on Robotics in Education (RiE)*, 2019, pp. 3–13.

[23] J. Lehmann, M. Lalmas, E. Yom-Tov, and G. Dupret, "Models of user engagement," in *User Modeling, Adaptation, and Personalization*, 2012, pp. 164–175.

[24] H. L. O'Brien and E. G. Toms, "The development and evaluation of a survey to measure user engagement," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 1, pp. 50–69, 2010.

[25] A. Brown, J. Lawrence, M. Basson, and P. Redmond, "A conceptual framework to enhance student online learning and engagement in higher education," *Higher Education Research & Development*, vol. 41, no. 2, pp. 284–299, 2022.

[26] K. Sharma, P. Jermann, and P. Dillenbourg, ""with-me-ness": A gaze-measure for students' attention in moocs," in *International conference of the learning sciences*, no. EPFL-CONF-201918, 2014.

[27] C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi, "A model of attention and interest using gaze behavior," in *Intelligent Virtual Agents*, 2005, pp. 229–240.

[28] L. Hall, S. Woods, R. Aylett, L. Newall, and A. Paiva, "Achieving empathic engagement through affective interaction with synthetic characters," in *Affective computing and intelligent interaction*, 2005, pp. 731–738.

[29] P. I., *Mind, hands, face and body. A goal and belief view of multimodal communication*. Weidler, 2007.

[30] D. Bohus and E. Horvitz, "Models for multiparty engagement in open-world dialog," in *SIGDIAL Conference*, 2009, p. 10.

[31] K. Drejing, S. Thill, and P. Hemeren, "Engagement: A traceable motivational concept in human-robot interaction," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 956–961.

[32] L. J. Corrigan, C. Peters, G. Castellano, F. Papadopoulos, A. Jones, S. Bhargava, S. Janarthanam, H. Hastie, A. Deshmukh, and R. Aylett, "Social-task engagement: Striking a balance between the robot and the task," in *Embodied Communication of Goals and Intentions Workshop at International Conference on Social Robotics*, 2013, pp. 1–6.

[33] J. Nasir, B. Bruno, M. Chetouani, and P. Dillenbourg, "What if social robots look for productive engagement?" *International Journal of Social Robotics*, pp. 1–17, 2021.

[34] K. Delgado, J. M. Origgi, T. Hasanpoor, H. Yu, D. Allessio, I. Arroyo, W. Lee, M. Betke, B. Woolf, and S. A. Bargal, "Student engagement dataset," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3628–3636.

[35] N. Glas and C. Pelachaud, "User engagement and preferences in information-giving chat with virtual agents," in *Workshop on Engagement in Social Intelligent Virtual Agents (ESIVA)*, 2015, pp. 33–40.

[36] D. Zyngier, "(re)conceptualising student engagement: Doing education not doing time," *Teaching and Teacher Education*, vol. 24, no. 7, pp. 1765–1776, 2008.

[37] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School engagement: Potential of the concept, state of the evidence," *Review of educational research*, vol. 74, no. 1, pp. 59–109, 2004.

[38] C. Peters, G. Castellano, and S. d. Freitas, "An exploration of user engagement in hci," *International Workshop on Affective-Aware Virtual Agents and Social Robots AFFINE'09*, 2009.

[39] H. Monkaresi, "Recognizing complex mental states from naturalistic human-computer interactions," Ph.D. dissertation, University of Sydney, 2014.

[40] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagement from facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, 2014.

[41] A. R. Anderson, S. L. Christenson, M. F. Sinclair, and C. A. Lehr, "Check & connect: The importance of relationships for promoting engagement with school," *Journal of School Psychology*, vol. 42, no. 2, pp. 95–113, 2004.

[42] A. K. Vail, J. F. Grafsgaard, J. B. Wiggins, J. C. Lester, and K. E. Boyer, "Predicting learning and engagement in tutorial dialogue: A personality-based model," in *16th International Conference on Multimodal Interaction*, 2014, pp. 255–262.

[43] F. Papadopoulos, L. J. Corrigan, A. Jones, and G. Castellano, "Learner modelling and automatic engagement recognition with robotic tutors," *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pp. 740–744, 2013.

[44] A. Graesser, B. McDaniel, P. Chipman, A. Witherspoon, S. D'Mello, and B. Gholson, "Detection of emotions during learning with autotutor," in *28th annual meetings of the cognitive science society*, 2006, pp. 285–290.

[45] A. L. Brown, S. Ellery, and J. C. Campione, *Creating zones of proximal development electronically.* Lawrence Erlbaum Associates publishers, 1998.

[46] R. Pekrun, "The impact of emotions on learning and achievement: Towards a theory of cognitive/motivational mediators," *Applied psychology*, vol. 41, no. 4, pp. 359–376, 1992.

[47] S. Attfield, G. Kazai, M. Lalmas, and B. Piwowarski, "Towards a science of user engagement (position paper)," in *WSDM workshop on user modelling for Web applications*, 2011.

[48] S. Asteriadis, K. Karpouzis, and S. Kollias, "Feature extraction and selection for inferring user engagement in an hci environment," in *International Conference on Human-Computer Interaction*, 2009, pp. 22–29.

[49] A. Belle, R. H. Hargraves, and K. Najarian, "An automated optimal engagement and attention detection system using electrocardiogram." *Computational and mathematical methods in medicine*, pp. 1–12, 2012.

[50] H. J. Sun, M. X. Huang, G. Ngai, and S. C. F. Chan, "Nonintrusive multimodal attention detection," in *7th International Conference on Advances in Computer-Human Interactions, ACHI 2014*, 2014, pp. 192–199.

[51] P. Redmond, L.-A. Abawi, A. Brown, R. Henderson, and A. Heffernan, "An online engagement framework for higher education." *Online learning*, vol. 22, no. 1, pp. 183–204, 2018.

[52] T. Soffer and A. Cohen, "Students' engagement characteristics predict success and completion of online courses," *Journal of computer assisted learning*, vol. 35, no. 3, pp. 378–389, 2019.

[53] T. Muir, N. Milthorpe, C. Stone, J. Dyment, E. Freeman, and B. Hopwood, "Chronicling engagement: Students' experience of online learning over time," *Distance Education*, vol. 40, no. 2, pp. 262–277, 2019.

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

[55] D. Bohus and E. Horvitz, "Managing human-robot engagement with forecasts and ...um ...hesitations," in *16th International Conference on Multimodal Interaction*, 2014, pp. 2–9.

[56] C. Peters, "Direction of attention perception for conversation initiation in virtual environments," in *Virtual Agents*, 2005, pp. 215–228.

[57] M. E. Foster, A. Gaschler, and M. Giuliani, "How can i help you? comparing engagement classification strategies for a robot bartender," in *15th ACM on International conference on multimodal interaction*, 2013, pp. 255–262.

[58] D. Klotz, J. Wienke, J. Peltason, B. Wrede, and S. Wrede, "Engagement-based multi-party dialog with a humanoid robot," in *SIGDIAL 2011 Conference*, 2011, pp. 341–343.

[59] S.-S. Yun, M.-T. Choi, M. Kim, and J.-B. Song, "Intention reading from a fuzzy-based human engagement model and behavioural features," *International Journal of Advanced Robotic Systems*, 2012.

[60] G. Castellano, I. Leite, A. Pereira, A. Paiva, and P. W. McOwan, "Detecting user engagement with a robot companion using task and social interaction-based features," *ICMI-MLMI'09*, pp. 119–126, 2009.

[61] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva, "Automatic analysis of affective postures and body motion to detect engagement with a game companion," in *6th ACM/IEEE International Conference on Human-Robot Interaction*, 2011, pp. 305–311.

[62] S. Lemaignan, F. Garcia, A. Jacq, and P. Dillenbourg, "From real-time attention assessment to "with-me-ness" in human-robot interaction," in *TO FIND*, 2016.

[63] M. P. Michalowski, S. Sabanovic, and R. Simmons, "A spatial model of engagement for a social robot," in *9th IEEE International Workshop on Advanced Motion Control*, 2006, pp. 762–767.

[64] C. Yu, P. M. Aoki, and A. Woodruff, "Detecting user engagement in everyday conversations," *arXiv preprint cs/0410027*, 2004.

[65] Q. Xu, L. Li, and G. Wang, "Designing engagement-aware agents for multiparty conversations," in *SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 2233–2242.

[66] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *13th annual ACM international conference on Multimedia*, 2005, pp. 677–682.

[67] A. Kamath, A. Biswas, and V. Balasubramanian, "A crowdsourced approach to student engagement recognition in e-learning environments," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–9.

[68] A. Gupta, R. Jaiswal, S. Adhikari, and V. Balasubramanian, "DAISEE: dataset for affective states in e-learning environments," *CoRR*, vol. abs/1609.01885, 2016. [Online]. Available: http://arxiv.org/abs/1609.01885

[69] A. Kaur, A. Mustafa, L. Mehta, and A. Dhall, "Prediction and localization of student engagement in the wild," in *2018 Digital Image Computing: Techniques and Applications (DICTA)*, 2018, pp. 1–8.

[70] N. Alyuz, E. Okur, E. Oktay, U. Genc, S. Aslan, S. E. Mete, D. Stanhill, B. Arnrich, and A. A. Esme, "Towards an emotional engagement model: Can affective states of a learner be automatically detected in a 1:1 learning scenario," in *6th Workshop on Personalization Approaches in Learning Environments (PALE)*, 2016.

[71] M. Cukurova, Q. Zhou, D. Spikol, and L. Landolfi, "Modelling collaborative problem-solving competence with transparent learning analytics: is video data enough?" in *10th International Conference on Learning Analytics & Knowledge*, 2020, pp. 270–275.

[72] A. Kasparova, O. Celiktutan, and M. Cukurova, "Inferring student engagement in collaborative problem solving from visual cues," in *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion)*, 2020.

[73] D. B. Jayagopi and J. Odobez, "Given that, should i respond? contextual addressee estimation in multi-party human-robot interactions," in *8th ACM/IEEE International Conference on Human-Robot Interaction*, 2013, pp. 147–148.

[74] S. Sheikhi, D. Babu Jayagopi, V. Khalidov, and J.-M. Odobez, "Context aware addressee estimation for human robot interaction," in *6th workshop on Eye gaze in intelligent human machine interaction: gaze in multimodal interaction*, 2013, pp. 1–6.

[75] C. L. Sidner, C. Lee, and N. Lesh, "Engagement rules for human-robot collaborative interactions," in *IEEE International Conference on Systems, Man and Cybernetics*, 2003, pp. 3957–3962.

[76] E. Nouri, S. Park, S. Scherer, J. Gratch, P. Carnevale, L.-P. Morency, and D. R. Traum, "Prediction of strategy and outcome as negotiation unfolds by using basic verbal and behavioral features," in *INTERSPEECH*, 2013, pp. 1458–1461.

[77] C. L. Sidner, C. D. Kidd, C. Lee, and N. Lesh, "Where to look: a study of human-robot engagement," in *9th international conference on Intelligent user interfaces*, 2004, pp. 78–84.

[78] E. Park, D. Jin, and A. P. del Pobil, "The law of attraction in human-robot interaction," *International Journal of Advanced Robotic Systems*, vol. 9, 2012.

[79] E. Park, K. J. Kim, and A. P. del Pobil, "The effects of robot's body gesture and gender in human-robot interaction," *Human-Computer Interaction*, vol. 6, pp. 91–96, 2011.

[80] H. C. Triandis *et al.*, *Culture and social behavior*. McGraw-Hill New York, 1994.

[81] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: a meta-analysis." *Psychological bulletin*, vol. 128, no. 2, p. 203, 2002.

[82] B. Mesquita and R. Walker, "Cultural differences in emotions: A context for interpreting emotional experiences," *Behaviour research and therapy*, vol. 41, no. 7, pp. 777–793, 2003.

[83] O. Rudovic, J. Lee, L. Mascarell-Maricic, B. W. Schuller, and R. W. Picard, "Measuring engagement in robot-assisted autism therapy: a cross-cultural study," *Frontiers in Robotics and AI*, vol. 4, p. 36, 2017.

[84] O. Rudovic, Y. Utsumi, J. Lee, J. Hernandez, E. C. Ferrer, B. Schuller, and R. W. Picard, "Culturenet: A deep learning approach for engagement intensity estimation from face images of children with autism," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018, pp. 339–346.

[85] M. P. Joosse, R. W. Poppe, M. Lohse, and V. Evers, "Cultural differences in how an engagement-seeking robot should approach a group of people," in *5th ACM international conference on Collaboration across boundaries: culture, distance & technology*, 2014, pp. 121–130.

[86] J. Hannon and B. D'Netto, "Cultural diversity online: Student engagement with learning technologies," *International journal of educational management*, 2007.

[87] O. Celiktutan and H. Gunes, "Computational analysis of human-robot interactions through first-person vision: Personality and interaction experience," in *24th IEEE International Symposium on Robot and Human Interactive Communication*, 2015, pp. 815–820.

[88] S. Ivaldi, S. Lefort, J. Peters, M. Chetouani, J. Provasi, and E. Zibetti, "Towards engagement models that consider individual factors in hri: On the relation of extroversion and negative attitude towards robots to gaze and speech during a human–robot assembly task," *International Journal of Social Robotics*, vol. 9, no. 1, pp. 63–86, 2017.

[89] R. Cuperman and W. Ickes, "Big five predictors of behavior and perceptions in initial dyadic interactions: Personality similarity helps extraverts and introverts, but hurts "disagreeables"." *Journal of personality and social psychology*, vol. 97, no. 4, p. 667, 2009.

[90] *Diagnostic and statistical manual of mental disorders: DSM-5*. American Psychiatric Association, 2013.

[91] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Science Robotics*, vol. 3, no. 19, 2018.

[92] L. Constantine and H. Hajj, "A survey of ground-truth in emotion data annotation," in *IEEE International Conference on Pervasive Computing and Communications Workshops*, 2012, pp. 697–702.

[93] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *10th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–8.

[94] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "Daisee: Towards user engagement recognition in the wild," *arXiv preprint arXiv:1609.01885*, 2016.

[95] A. Abedi and S. Khan, "Affect-driven engagement measurement from videos," *arXiv preprint arXiv:2106.10882*, 2021.

[96] X. Niu, H. Han, J. Zeng, X. Sun, S. Shan, Y. Huang, S. Yang, and X. Chen, "Automatic engagement prediction with gap feature," in *20th ACM International Conference on Multimodal Interaction*, 2018, pp. 599–603.

[97] O. Celiktutan, E. Skordos, and H. Gunes, "Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement," *IEEE Transactions on Affective Computing*, 2017.

[98] H. Salam, O. Celiktutan, I. Hupont, H. Gunes, and M. Chetouani, "Fully automatic analysis of engagement and its relationship to personality in human-robot interactions," *IEEE Access*, vol. 5, pp. 705–721, 2016.

[99] J. Nasir, U. Norman, B. Bruno, M. Chetouani, and P. Dillenbourg, "Pe-hri: A multimodal dataset for the study of productive engagement in a robot mediated collaborative educational setting [dataset]," https://doi.org/10.5281/zenodo.4288833, 2020.

[100] J. Nasir, B. Bruno, and P. Dillenbourg, "PE-HRI-temporal: A multimodal temporal dataset in a robot mediated collaborative educational setting [dataset]," https://doi.org/10.5281/zenodo.5576058, 2021, zenodo.

[101] A. Ben-Youssef, C. Clavel, S. Essid, M. Bilac, M. Chamoux, and A. Lim, "Ue-hri: A new dataset for the study of user engagement in spontaneous human-robot interactions," in *19th ACM international conference on multimodal interaction*, 2017, pp. 464–472.

[102] T. Liu and A. Kappas, "Predicting engagement breakdown in hri using thin-slices of facial expressions," in *Workshops at the thirty-second AAAI conference on artificial intelligence*, 2018.

[103] A. Ben-Youssef, G. Varni, S. Essid, and C. Clavel, "On-the-fly detection of user engagement decrease in spontaneous human–robot interaction using recurrent and deep neural networks," *International Journal of Social Robotics*, vol. 11, no. 5, pp. 815–828, 2019.

[104] J. Nasir, B. Bruno, M. Chetouani, and P. Dillenbourg, "A speech-based productive engagement metric for real-time human-robot interaction in collaborative educational contexts," *Preprint*, 2022.

[105] M. Lombard, L. Weinstein, and T. Ditton, "Measuring telepresence: the validity of the temple presence inventory (tpi) in a gaming context," in *ISPR 2011: The International Society for Presence Research Annual Conference*, 2011.

[106] S. Campano, C. Langlet, N. Glas, C. Clavel, and C. Pelachaud, "An eca expressing appreciations," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 962–967.

[107] E. Champion, B. Dave, and I. Bishop, "Interaction, agency and artefacts," in *Digital Design: Research and Practice, 10th International Conference on CAAD Futures' 2003*, 2003, pp. 249–258.

[108] L. J. Corrigan, C. Basedow, D. Küster, A. Kappas, C. Peters, and G. Castellano, "Mixing implicit and explicit probes: finding a ground truth for engagement in social human-robot interactions," in *ACM/IEEE international conference on Human-robot interaction*, 2014, pp. 140–141.

[109] S. Aslan, E. Okur, N. Alyuz, S. E. Mete, E. Oktay, U. Genc, and A. A. Esme, "Students' emotional self-labels for personalized models," in *7th International Learning Analytics & Knowledge Conference*, 2017, pp. 550–551.

[110] Y. I. Nakano and R. Ishii, "Estimating user's engagement from eye-gaze behaviors in human-agent conversations," in *15th international conference on Intelligent user interfaces*, 2010, pp. 139–148.

[111] K. Inoue, D. Lala, K. Takanashi, and T. Kawahara, "Engagement recognition in spoken dialogue via neural network by aggregating different annotators' models." in *Interspeech*, 2018, pp. 616–620.

[112] ——, "Latent character model for engagement recognition based on multimodal behaviors," in *9th International Workshop on Spoken Dialogue System Technology*, 2019, pp. 119–130.

[113] R. Ishii and Y. I. Nakano, "Estimating user's conversational engagement based on gaze behaviors," in *Intelligent Virtual Agents*, 2008, pp. 200–207.

[114] R. Ishii, Y. Shinohara, I. Nakano, and T. Nishida, "Combining multiple types of eye-gaze information to predict user's conversational engagement," in *2nd workshop on eye gaze on intelligent human machine interaction*, 2011.

[115] I. Leite, M. McCoy, D. Ullman, N. Salomons, and B. Scassellati, "Comparing models of disengagement in individual and group interactions," in *Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2015, pp. 99–105.

[116] W. Benkaouar and D. Vaufreydaz, "Multi-sensors engagement detection with a robot companion in a home environment," in *Workshop on Assistance and Service robotics in a human environment at IROS 2012*, 2012, pp. 45–52.

[117] D. Vaufreydaz, W. Johal, and C. Combe, "Starting engagement detection towards a companion robot using multimodal features," *Robotics and Autonomous Systems*, 2015.

[118] R. Bednarik, S. Eivazi, and M. Hradis, "Gaze and conversational engagement in multiparty video conversation: An annotation scheme and classification of high and low levels of engagement," in *4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*, 2012, pp. 1–6.

[119] C. Oertel and G. Salvi, "A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue," in *15th ACM on International conference on multimodal interaction*, 2013, pp. 99–106.

[120] M. Frank, G. Tofighi, H. Gu, and R. Fruchter, "Engagement detection in meetings," *arXiv preprint arXiv:1608.08711*, 2016.

[121] S. Asteriadis, P. Tzouveli, K. Karpouzis, and S. Kollias, "Non-verbal feedback on user interest based on gaze direction and head pose," in *2nd International Workshop on Semantic Media Adaptation and Personalization*, 2007, pp. 171–178.

[122] C. Peters, S. Asteriadis, and K. Karpouzis, "Investigating shared attention with a virtual agent using a gaze-based interface," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, pp. 119–130, 2010.

[123] L. J. Corrigan, C. Peters, and G. Castellano, "Identifying task engagement: Towards personalised interactions with educational robots," in *Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 655–658.

[124] A. Kapoor, R. W. Picard, and Y. Ivanov, "Probabilistic combination of multiple modalities to detect interest," in *Pattern Recognition, 2004. ICPR 2004. 17th International Conference on*, vol. 3, 2004, pp. 969–972.

[125] S. M. Anzalone, S. Boucenna, S. Ivaldi, and M. Chetouani, "Evaluating the engagement with social robots," *International Journal of Social Robotics*, pp. 1–14, 2015.

[126] N. Glas, K. Prepin, and C. Pelachaud, "Engagement and virtual agents," Master's thesis, Télécom ParisTech, 2013.

[127] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[128] H. P. Martínez and G. N. Yannakakis, "Mining multimodal sequential patterns: A case study on affect detection," in *13th international conference on multimodal interfaces*, 2011, pp. 3–10.

[129] E. Okur, N. Alyuz, S. Aslan, U. Genc, C. Tanriover, and A. A. Esme, "Behavioral engagement detection of students in the wild," in *International Conference on Artificial Intelligence in Education*, 2017, pp. 250–261.

[130] G. Castellano, I. Leite, and A. Paiva, "Detecting perceived quality of interaction with a robot using contextual features," *Autonomous Robots*, vol. 41, no. 5, pp. 1245–1261, 2017.

[131] T. Baur, D. Schiller, and E. André, "Modeling user's social attitude in a conversational system," in *Emotions and Personality in Personalized Services*, 2016, pp. 181–199.

[132] A. Cerekovic, O. Aran, and D. Gatica-Perez, "Rapport with virtual agents: What do human social cues and personality explain?" *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 382–395, 2017.

[133] O. Rudovic, M. Zhang, B. Schuller, and R. Picard, "Multi-modal active learning from human data: A deep reinforcement learning approach," in *International Conference on Multimodal Interaction*, 2019, pp. 6–15.

[134] O. Rudovic, H. W. Park, J. Busche, B. Schuller, C. Breazeal, and R. W. Picard, "Personalized estimation of engagement from videos using active learning with deep reinforcement learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 217–226.

[135] L. Devillers and G. D. Duplessis, "Toward a context-based approach to assess engagement in human-robot social interaction," in *Dialogues with Social Robots*, 2017, pp. 293–301.

[136] M. F. Mason, E. P. Tatkow, and C. N. Macrae, "The look of love gaze shifts and person perception," *Psychological Science*, vol. 16, no. 3, pp. 236–239, 2005.

[137] C. L. Sidner, C. Lee, and N. Lesh, "Engagement when looking: behaviors for robots when collaborating with people," in *7th workshop on the Semantic and Pragmatics of Dialogue*, 2003, pp. 123–130.

[138] S. Ehrlich, A. Wykowska, K. Ramirez-Amaro, and G. Cheng, "When to engage in interaction and how? eeg-based enhancement of robot's ability to sense social signals in hri," in *Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on*, 2014, pp. 1104–1109.

[139] E. Bekele, J. W. Wade, D. Bian, L. Zhang, Z. Zheng, A. Swanson, M. Sarkar, Z. Warren, and N. Sarkar, "Multimodal interfaces and sensory fusion in vr for social interactions," in *International Conference on Virtual, Augmented and Mixed Reality*, 2014, pp. 14–24.

[140] S. R. Langton, R. J. Watt, and B. Vicki, "Do the eyes have it? cues to the direction of social attention," *Trends in Cognitive Sciences*, vol. 4, no. 2, 2000.

[141] C. L. Sidner, C. Lee, L. P. Morency, and C. Forlines, "The Effect of head-nod recognition in human-robot conversation," *ACM Conference on Human-Robot Interaction*, pp. 290–296, 2006.

[142] D. McNeill, *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

[143] C. L. Sidner and C. Lee, "An architecture for engagement in collaborative conversations between a robot and humans," *Mitsubishi Electric Research Labs TR2003-13, Cambridge, MA*, 2003.

[144] G. Tofighi, H. Gu, and K. Raahemifar, "Vision-based engagement detection in virtual reality," in *2016 Digital Media Industry & Academic Forum (DMIAF)*, 2016, pp. 202–206.

[145] A. TS and R. M. R. Guddeti, "Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks," *Education and Information Technologies*, vol. 25, no. 2, pp. 1387–1415, 2020.

[146] M. M. Thiruthuvanathan, B. Krishnan, and M. Rangaswamy, "Engagement detection through facial emotional recognition using a shallow residual convolutional neural networks," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 2, 2021.

[147] D. Silva and P. North, "Audiovisual recognition," in *wrong*, 2004, pp. 584–592.

[148] J. Grafsgaard, J. Wiggins, a.K. Vail, K. Boyer, E. Wiebe, and J. Lester, "The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring," *International Conference on Multimodal Interaction*, pp. 42–49, 2014.

[149] E. T. Hall *et al.*, *The silent language*. Doubleday New York, 1959, vol. 3.

[150] R. Mead, A. Atrash, and M. J. Matarić, "Proxemic feature recognition for interactive robots: automating metrics from the social sciences," in *Social Robotics*, 2011, pp. 52–61.

[151] P. Holthaus, K. Pitsch, and S. Wachsmuth, "How can i help? spatial attention strategies for a receptionist robot," *International Journal of Social Robotics*, vol. 3, no. 4, pp. 383–393, 2011.

[152] D. Bohus and E. Horvitz, "Learning to predict engagement with a spoken dialog system in open-world settings," in *SIGDIAL Conference*, 2009, pp. 244–252.

[153] J. Fast, *Body language*. Simon and Schuster, 1970, vol. 82348.

[154] A. Mehrabian, "Relationship of attitude to seated posture, orientation, and distance," *Journal of Personality and Social Psychology*, vol. 10, no. 1, pp. 26–30, 1968.

[155] J. Cassell, Y. I. Nakano, T. W. Bickmore, C. L. Sidner, and C. Rich, "Non-verbal cues for discourse structure," in *39th Annual Meeting of the Association for Computational Linguistics*, 2001, pp. 114–123.

[156] N. K. Person, S. D'Mello, and A. Olney, "Toward socially intelligent interviewing systems," *Envisioning the survey interview of the future*, pp. 195–214, 2008.

[157] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *Affective Computing, IEEE Transactions on*, vol. 4, no. 1, pp. 15–33, 2013.

[158] H. J. Witchel, C. E. Westling, J. Tee, R. Needham, A. Healy, and N. Chockalingam, "A time series feature of variability to detect two types of boredom from motion capture of the head and shoulders," in *2014 European Conference on Cognitive Ergonomics*, 2014, p. 37.

[159] P. Bull, *Body movement and interpersonal communication*. John Wiley & Sons Inc, 1983.

[160] S. Mota and R. W. Picard, "Automated posture analysis for detecting learner's interest level," in *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, vol. 5, 2003, pp. 49–49.

[161] A. Kaur, B. Ghosh, N. D. Singh, and A. Dhall, "Domain adaptation based topic modeling techniques for engagement estimation in the wild," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019, pp. 1–6.

[162] T. Baur, I. Damian, F. Lingenfelser, J. Wagner, and E. André, "Nova: Automated analysis of nonverbal signals in social interactions," in *Human Behavior Understanding*, 2013, pp. 160–171.

[163] C. Oertel, C. De Looze, S. Scherer, A. Windmann, P. Wagner, and N. Campbell, "Towards the automatic detection of involvement in conversation," in *Analysis of verbal and nonverbal communication and enactment. The processing issues*, ser. Lecture Notes in Computer Science, A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, and A. Nijholt, Eds., 2011, vol. 6800, pp. 163–170.

[164] J. Maisonnasse, "Estimation des relations attentionnelles dans un environnement intelligent," Ph.D. dissertation, wrong, 2007, thèse de doctorat dirigée par Crowley, James L Sciences cognitives Grenoble 1 2007. [Online]. Available: http://www.theses.fr/2007GRE10300

[165] H.-D. Kim, J.-S. Choi, M. Kim, and C.-H. Lee, "Reliable detection of sound's direction for human robot interaction," in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, vol. 3, 2004, pp. 2411–2416.

[166] K. Inoue, D. Lala, S. Nakamura, K. Takanashi, and T. Kawahara, "Annotation and analysis of listener's engagement based on multi-modal behaviors," in *Workshop on Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction*, 2016, pp. 25–32.

[167] J. Kim, K. P. Truong, V. Charisi, C. Zaga, V. Evers, and M. Chetouani, "Multimodal detection of engagement in groups of children using rank learning," in *International Workshop on Human Behavior Understanding*, 2016, pp. 35–48.

[168] M. Bilac, M. Chamoux, and A. Lim, "Gaze and filled pause detection for smooth human-robot conversations," in *IEEE-RAS 17th International Conference on Humanoid Robotics*, 2017, pp. 297–304.

[169] C. Strapparava, A. Valitutti *et al.*, "Wordnet affect: an affective extension of wordnet." in *Lrec*, vol. 4, no. 1083-1086, 2004, p. 40.

[170] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen, "Interpersonal synchrony: A survey of evaluation methods across disciplines," *Affective Computing, IEEE Transactions on*, vol. 3, no. 3, pp. 349–365, 2012.

[171] J. Nadel, K. Prepin, and M. Okanda, "Experiencing contingency and agency: First step toward self-understanding in making a mind?" *Interaction studies*, vol. 6, no. 3, pp. 447–462, 2005.

[172] K. Prepin and P. Gaussier, "How an agent can detect and use synchrony parameter of its own interaction with a human?" in *Development of Multimodal Interfaces: Active Listening and Synchrony*, 2010, pp. 50–65.

[173] A. Foltz, J. Gaspers, K. Thiele, P. Stenneken, and P. Cimiano, "Lexical alignment in triadic communication," *Frontiers in psychology*, vol. 6, p. 127, 2015.

[174] S. Campano, J. Durand, and C. Clavel, "Comparative analysis of verbal alignment in human-human and human-agent interactions." in *LREC*, 2014, pp. 4415–4422.

[175] M. J. Richardson, K. L. Marsh, and R. Schmidt, "Effects of visual and verbal interaction on unintentional interpersonal coordination." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 31, no. 1, p. 62, 2005.

[176] E.-H. Jang, B.-J. Park, M.-S. Park, S.-H. Kim, and J.-H. Sohn, "Analysis of physiological signals for recognition of boredom, pain, and surprise emotions," *Journal of physiological anthropology*, vol. 34, no. 1, p. 1, 2015.

[177] P. Rani and N. Sarkar, "Operator engagement detection for robot behaviour adaptation," *International Journal of Advanced Robotic Systems*, vol. 4, no. 1, pp. 1–12, 2007.

[178] S. H. Fairclough and L. Venables, "Psychophysiological predictors of task engagement and distress," *Human Factors in Design, Safety, and Management In D. de Waard, KA Brookhuis, R. van Egmond, and Th. Boersema (Eds.)*, pp. 349–362, 2005.

[179] C. Iani, D. Gopher, and P. Lavie, "Effects of task difficulty and invested mental effort on peripheral vasoconstriction," *Psychophysiology*, vol. 41, no. 5, pp. 789–798, 2004.

[180] U. Maniscalco, P. Storniolo, and A. Messina, "Bidirectional multimodal signs of checking human-robot engagement and interaction," *International Journal of Social Robotics*, pp. 1–15, 2022.

[181] J. Hadfield, G. Chalvatzaki, P. Koutras, M. Khamassi, C. S. Tzafestas, and P. Maragos, "A deep learning approach for multi-view engagement estimation of children in a child-robot joint attention task," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019, pp. 1251–1256.

[182] Y.-Y. Li and Y.-P. Hung, "Feature fusion of face and body for engagement intensity detection," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 3312–3316.

[183] C. Thomas, N. Nair, and D. B. Jayagopi, "Predicting engagement intensity in the wild using temporal convolutional network," in *20th ACM International Conference on Multimodal Interaction*, 2018, pp. 604–610.

[184] C. Coffrin, L. Corrin, P. de Barba, and G. Kennedy, "Visualizing patterns of student engagement and performance in moocs," in *fourth international conference on learning analytics and knowledge*, 2014, pp. 83–92.

[185] S. Amershi, C. Conati *et al.*, "Combining unsupervised and supervised classification to build user models for exploratory learning environments," *Journal of educational data mining*, vol. 1, no. 1, pp. 18–71, 2009.

[186] R. F. Kizilcec, C. Piech, and E. Schneider, "Deconstructing disengagement: analyzing learner subpopulations in massive open online courses," in *third international conference on learning analytics and knowledge*, 2013, pp. 170–179.

[187] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Engaging with massive online courses," in *23rd international conference on World wide web*, 2014, pp. 687–698.

[188] T. Xu, J. White, S. Kalkan, and H. Gunes, "Investigating bias and fairness in facial expression recognition," in *Computer Vision - ECCV 2020 Workshops - Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*, 2020, pp. 506–523.
[189] J. Cheong, S. Kalkan, and H. Gunes, "The hitchhiker's guide to bias and fairness in facial affective signal processing: Overview and techniques," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 39–49, 2021.

Hanan Salam is Assistant Professor in Computer Science at New York University Abu Dhabi. She is also the director of Social Machines & Robotics Lab (SMART) & a member of the Center of AI & Robotics (CAIR). She is the co-founder of Women in AI, an international non-profit whose mission is to close the gender gap in the domain of Artificial Intelligence through education, research, and events. Her scientific interests include Artificial Intelligence for mental healthcare, Human-Machine Interaction, social robotics, computer vision, machine learning and affective computing.

Oya Celiktutan received the Ph.D. degree in electrical and electronic engineering from Bogazici University, Turkey, in collaboration with the National Institute of Applied Sciences of Lyon, France, in 2013. She spent several years as a postdoctoral researcher with the Queen Mary University of London; the University of Cambridge; and Imperial College London, United Kingdom. Since 2018, she has been an Assistant Professor (Lecturer) with the Centre for Robotics Research, Department of Engineering, King's College London, United Kingdom, where she is the Head of Social AI & Robotics Laboratory. Her research focuses on computer vision and machine learning, applied to human behaviour understanding and generation, social signal processing, and human–robot interaction.

Hatice Gunes (Senior Member, IEEE) is a Professor of Affective Intelligence and Robotics (AFAR) at the Department of Computer Science and Technology, University of Cambridge, United Kingdom, leading the AFAR Lab. Her expertise is in the areas of affective computing and social signal processing cross-fertilizing research in human behaviour understanding, computer vision, signal processing, machine learning, and human–robot interaction. She has published over 125 papers in the above areas. Dr Gunes is the former President (2017-2019) of the Association for the Advancement of Affective Computing, was the General Co-Chair of ACII 2019, and Program Co-Chair of ACM/IEEE HRI 2020 and IEEE FG 2017. Her research has been supported by various competitive grants, with funding from the Engineering and Physical Sciences Research Council, UK (EPSRC), Innovate UK, British Council, Alan Turing Institute and EU Horizon 2020. She is currently a Fellow of the EPSRC and was previously a Faculty Fellow of the Alan Turing Institute.

Mohamed Chetouani is a Professor of signal processing and machine learning for human-machine interaction. He is affiliated to the PIRoS research team at the Institute for Intelligent Systems and Robotics (CNRS UMR 7222), Sorbonne University. His activities cover social signal processing, social robotics and interactive machine learning with applications in psychiatry, psychology, social neuroscience and education. Since 2018, he is the coordinator of the ANIMATAS H2020 Marie Sklodowska Curie European Training Network. Since 2019, he is the President of the Sorbonne University Ethical Committee. He is member of the management board of the International AI Doctoral Academy initiated by European networks of AI excellence centers. He is member of the EU Network of Human-Centered AI. He was Program co-chair of ACM ICMI 2020. He is General Chair of VIHAR 2021 and ACM ICMI 2023.