International Journal of Social Robotics (2024) 16:1641–1663
https://doi.org/10.1007/s12369-024-01146-w

REVIEW

From the Definition to the Automatic Assessment of Engagement in Human–Robot Interaction: A Systematic Review

Alessandra Sorrentino[1] · Laura Fiorini[1] · Filippo Cavallo[1]

Accepted: 8 May 2024 / Published online: 4 June 2024

Abstract

The concept of engagement is widely adopted in the human–robot interaction (HRI) field, as a core social phenomenon in the interaction. Despite the wide usage of the term, the meaning of this concept is still characterized by great vagueness. A common approach is to evaluate it through self-reports and observational grids. While the former solution suffers from a time-discrepancy problem, since the perceived engagement is evaluated at the end of the interaction, the latter solution may be affected by the subjectivity of the observers. From the perspective of developing socially intelligent robots that autonomously adapt their behaviors during the interaction, replicating the ability to properly detect engagement represents a challenge in the social robotics community. This systematic review investigates the conceptualization of engagement, starting with the works that attempted to automatically detect it in interactions involving robots and real users (i.e., online surveys are excluded). The goal is to describe the most worthwhile research efforts and to outline the commonly adopted definitions (which define the authors' perspective on the topic) and their connection with the methodology used for the assessment (if any). The research was conducted within two databases (Web of Science and Scopus) between November 2009 and January 2023. A total of 590 articles were found in the initial search. Thanks to an accurate definition of the exclusion criteria, the most relevant papers on automatic engagement detection and assessment in HRI were identified. Finally, 28 papers were fully evaluated and included in this review. The analysis illustrates that the engagement detection task is mostly addressed as a binary or multi-class classification problem, considering user behavioral cues and context-based features extracted from recorded data. One outcome of this review is the identification of current research barriers and future challenges on the topic, which could be clustered in the following fields: engagement components, annotation procedures, engagement features, prediction techniques, and experimental sessions.

Keywords Engagement definition · Engagement detection · Human–robot interaction · Social robotics

1 Introduction

To develop socially intelligent robots, one key issue is to provide them with the capability to evaluate several aspects of the interaction [1] and of the user [2]. The user profile is multifaceted, and the robot should take into account any aspect to shape the appropriate behavior [2]. Among the social phenomena that characterize the user profile, engagement is one of the aspects that the robot should be aware of to personalize the interaction [3]. In most of the studies where the users interact with technology (i.e. computers, virtual agents, and robots), the term engagement is commonly used without any explicit definition or interpretation [4, 5]. The concept seems characterized by vagueness and great variability [6], leaving the reader to fill the void [6, 7]. When dealing with this vagueness, several studies analyzed the definitions and the role of engagement with technological devices in specific contexts. In human–computer interaction (HCI), [5] describes the role of user engagement across computer science studies, mostly focusing on the conception, theories, and measurement of engagement. In the human–agent interaction context, [6] proposes an overview of the different factors considered when dealing with engagement, distinguishing especially between different types of engagement, the environment, and the involved participants. Similarly, [4] exploited the interpretation of engagement in human–agent interaction by highlighting a subset of concepts that

✉ Laura Fiorini
laura.fiorini@unifi.it

[1] Department of Industrial Engineering, University of Florence, Via Santa Marta 3, 50139 Florence, Italy

are interchangeably used, namely: attention [8], involvement, interest, immersion, rapport [9], empathy, and stance. Some of these concepts also emerge in the work related to the HCI [5]. These overlapping factors in different disciplines highlight the "multi-faceted" nature of the phenomenon [6] as well as the difficulty of proving a complete definition of the term, even when just considering a specific interaction scenario. The challenge of providing a proper definition of the concept increases if additional variables are included, like the duration of interaction (i.e., short-term vs. long-term engagement [10]), and the number of involved participants in the scene (i.e., individual vs group engagement [11]).

A common methodology for assessing the engagement of the participant(s) during a human–robot interaction (HRI) is through self-reports and questionnaires (e.g., User Engagement Scale [12]). When the administration of self-reports is unfeasible (e.g., young or older people with cognitive disabilities are involved), one common strategy relies on the usage of observational methods, such as observational rating scales (e.g., Observational Measurement of Engagement [13], the Menorah Park Engagement Scale [14], and Observed Emotion Rating Scale), ethograms (e.g., Video-Coding Incorporating Observed Emotions [15], Ethological Coding System for Interviews [16]), and coding schemes (e.g., Ethnographic and Laban-Inspired Coding System of Engagement [7]). To reduce the effort of manually assessing engagement, several researchers started developing strategies for automatically detecting it [17]. Most of the recent works adopt a "cue-centric" approach [18], identifying the social cues that may characterize the behavior of an "engaged" user. The recent advancements in machine and deep learning strategies led to new possibilities for improving automatic engagement detection in terms of accuracy and computational time. As computer vision-based techniques are usually adopted for assessing student engagement in online learning [19], the engagement state of the user can be used by the robot to adapt its behavior during the ongoing interaction [20]. Even if the concept itself is not well-defined and the multiple aspects composing it are still vague, engagement is considered a core aspect of human–robot interaction. It is part of the broad spectrum of factors that influence the quality of interaction, thus the acceptability and the perception of the robotic platform.

There have been previous efforts in investigating the engagement topic and its assessment in human–machine interaction (HMI) [21], human–agent interaction (HAI) [4] and human–computer interaction (HCI) [5]. In those studies, robots are indented as technology [5] and physical embodied agents [4, 6, 21]. We believe that social robots are more than just a technology tool, due to their capabilities of undertaking a large variety of complex human-like tasks such as navigation, object manipulation, and social interactions [22], thus expressing human-like social behaviors. Similarly, we support the idea that the embodied interaction with a robot may provoke different social phenomena in the interaction with respect to virtual agents or other machine interfaces [21], due to its physical presence. This is why we decided to address the engagement topic by focusing on human–robot interaction only. In HRI settings, [23] analyzed the socially aware engagement concept that characterizes human–robot first encounters, while [22] presented the latest works addressing engagement in children during child-robot interactions in educational and therapeutic settings. The current systematic review aims to provide a broad overview of the engagement concept, covering every phase of interaction and enlarging the spectrum of interaction contexts and settings. In detail, we are interested in investigating the definition of engagement used in HRI studies when a physical robot is present, identifying the components and any relationship with the interaction domains. Additionally, we would like to investigate the influence of the recent advancements in machine and deep learning solutions on the methodology used for the automatic engagement assessment. In this manuscript, we reviewed the literature with the intention of answering the following research questions:

1. Which are the most common definitions of engagement in the HRI research field? Which aspects of engagements are considered? Does the interpretation of engagement and its components change based on the application scenario?
2. Which methods and features are commonly used to automatically detect and assess user engagement?

2 Methods

2.1 Search Strategy

An electronic database search was performed in January 2023 using Scopus and Web of Science databases to identify articles concerning the automatic assessment of engagement in HRI. Specifically, the terms and the keywords used for the literature research were *(autom* OR continuous) AND (assess* OR detect* OR recogni* OR estimate* OR evalua*) AND engage* AND robot** located within the title and/or abstract. Only original, full-text articles published in English that reported automatic techniques for engagement estimation were included in this review. According to the research, there were 376 references from Scopus and 214 references from Web of Science. During the screening phase, two independent reviewers were involved. In cases of disagreements, meetings and discussions were organized to solve them.

2.2 Selection Criteria

First, duplicated documents were eliminated. Thereafter, the abstracts of the papers, retrieved by the electronic search, were examined to identify which deserved a full evaluation. During the screening procedure, the papers were excluded if (i) they belonged to a different research field (e.g., UAV, agriculture support, autonomous driving, aerospace applications); (ii) they were an abstract, a short communication, a review article or chapter published in a non-scientific book; (iii) they addressed research problems out of the scope of this review (i.e., speech recognition, localization, human activity recognition, affect recognition with no robot, automatic speaker verification). Among the 123 selected for the evaluation procedure, several papers were excluded if (i) they did not involve a robot; (ii) they did not focus on the engagement estimation with automatic techniques; (iii) they did not appear appropriate for this review after the reading of title and abstract; and (iv) they were not full access. Additionally, if multiple papers with similar content were published by the same authors, the ones published in journals were selected instead of papers presented at conferences. In cases where similar studies by the same authors were presented at conferences, the most recent paper was selected. Finally, the reference lists of included papers were examined to identify relevant studies that the electronic search might have missed (total number of papers: 8). At the end of the screening and evaluation phase, 28 papers were included in this review (see Fig. 1).

3 Results

3.1 Application Overview

The idea of automatically assessing engagement in HRI is quite new. As shown in Fig. 2a, the first work on this topic dates to 14 years ago. Of the reviewed papers, 22 papers (75.86%) were published in the last six years (2017–2022). It suggests that the interest in automating the engagement assessment gained more popularity in recent years.

Analyzing the context of interaction, it is possible to distinguish five application contexts in which engagement has been investigated (see Fig. 2b): game activity, conversation, cognitive therapy, working roles, and education. The main areas in which automatic engagement has been investigated are the game activity and the conversation categories (25% each). In the game category, we included the scenarios in which the individual is playing under the supervision of the robot (e.g., the child is playing chess and the robot monitors his/her moves [24–27]) and when the robot is an active participant in the game, e.g., poses the question of a quiz [28], plays a pointing game [29] or performs a handover task [30]. In the second category, we clustered the studies in which humans and robots are involved in a conversation, e.g., the robot is a storyteller [31], the robot explains some paintings [11], the robot speaks with the users [32–35], fostering the interaction between them [36]. Another scenario in which the detection of user engagement is quite relevant regards cognitive therapy, where a robotic platform is used to elicit certain behaviors in children with Autism Spectrum Disorder [20, 37–40]. Despite the recent trend of assessing students' engagement in online learning [19], automatic engagement detection in the education context with the robot is still not very popular (18% of the reviewed works). On the other hand, the engagement concept is present in a consistent way, when the robot is assigned to perform human working roles, like bartender [18, 41], museum guide [42] and salesperson [43].

3.1.1 Experimental Sessions

Overall, there is a balanced number of works involving young users (children: 12 studies; children with ASD: 5 studies) and adult individuals (13 works). The work of [42] is the only one that addressed both children and adults as target users. In most of the works, the interaction occurs between one participant and one robot (dyadic interaction:60.71%). To a lesser extent, the robot deals with a group of participants simultaneously (multi-party interaction: 28.57%). The works of [41–43] investigate engagement in dyadic and multi-party interactions (i.e.,, 10.71%). All the reviewed studies analyzed engagement in short-term interactions. It means that most of the time the users are requested to interact with the robot once, namely in a short time frame. Since the interaction relies on constructed protocols, user engagement is usually analyzed in 10–15 min, according to the reviewed works. The duration varies according to the context and the task. The maximum duration reported is 25 min in [38, 39], when the NAO robot is used in autism therapy. The minimum duration is 3 and a half minutes of interaction in the bartending scenario reported in [18]. It is worth noticing, that the duration is not always reported in the study (see Table 2 of the Supplementary Material). From the reviewed works, it emerged that, in some cases, the users were requested to interact over multiple sessions over a longer period (i.e., long-term interaction). One example is the work of [20], where the robot is installed at home for one month. Analogously, in the work of [44], the users were involved in six experimental sessions (three with the robot, three with the tablet) daily over 2 weeks. The work of [45] presents the results of children interacting with the TEGA robot in 6–8 experimental sessions over 3 months. Considering the overall duration of the experimental sessions, the longest one lasted for 278 days [42]. Excluding the works that did not specify it (i.e., 9 out of 28), in one case the robot was tested at home [20], in two cases the robot was tested in public places (e.g., museum [42] and shops [43]),

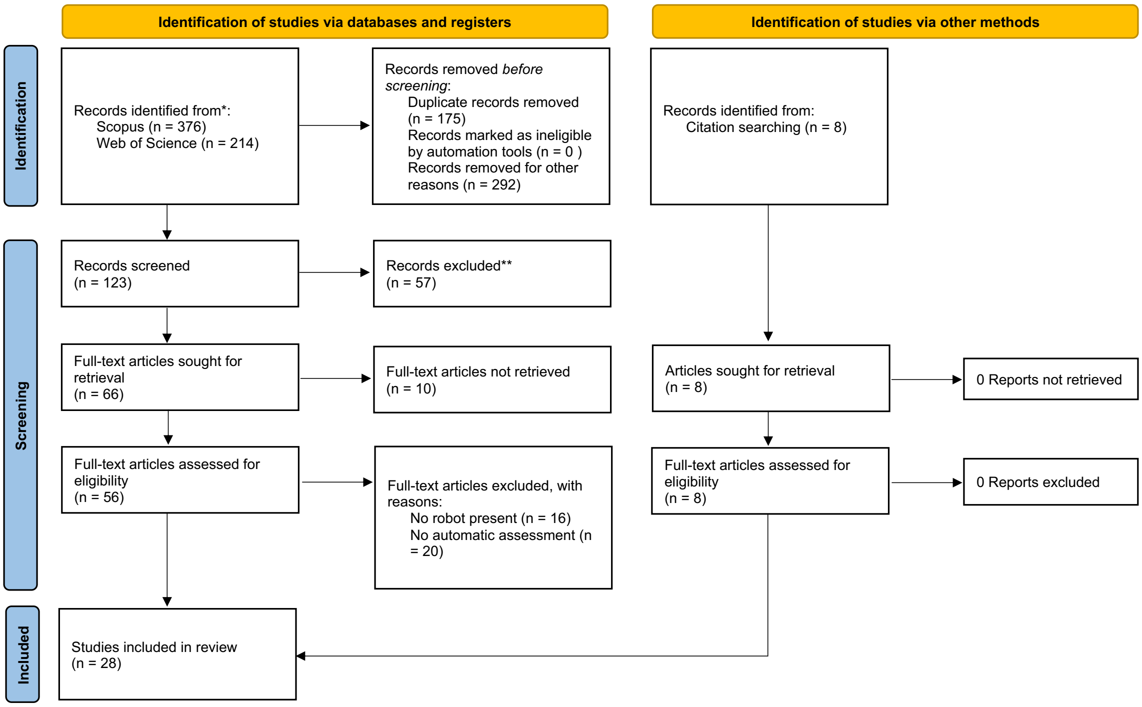
Identification of studies via databases and registers
Identification of studies via other methods
Identification
Records identified from*:
Scopus (n = 376)
Web of Science (n = 214)
Records removed *before screening*:
Duplicate records removed (n = 175)
Records marked as ineligible by automation tools (n = 0)
Records removed for other reasons (n = 292)
Records identified from:
Citation searching (n = 8)
Screening
Records screened (n = 123)
Records excluded** (n = 57)
Full-text articles sought for retrieval (n = 66)
Full-text articles not retrieved (n = 10)
Articles sought for retrieval (n = 8)
0 Reports not retrieved
Full-text articles assessed for eligibility (n = 56)
Full-text articles excluded, with reasons:
No robot present (n = 16)
No automatic assessment (n = 20)
Full-text articles assessed for eligibility (n = 8)
0 Reports excluded
Included
Studies included in review (n = 28)

Fig. 1 PRISMA flow chart of the study selection process

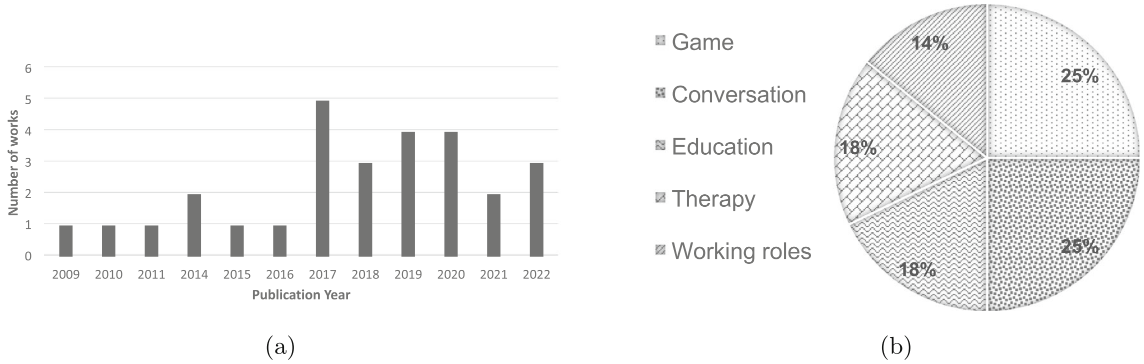

Fig. 2 Statistics of the application overview: **a** number of the reviewed papers, clustered by the year of publication; **b** cake graph of the application scenarios

and in 6 cases the interaction took place at school (i.e., elementary school [24–27, 44, 46, 47], kindergarten [45]). In the remaining work, the interaction occurred in a laboratory setting (i.e., office [33, 48] and university environments [29, 35, 36, 41, 49]). Based on the duration and the settings chosen for the interactions, the number of involved participants largely varied among the works (maximum: 227 participants; minimum: 2 participants). A complete overview is reported in Table 2 of the Supplementary Material.

3.2 Definition of Engagement

Our analysis reports that the definition of engagement is missing in 6 out of 28 studies. It means that in 6 studies, there are no references or explicit statements on the engagement concept. In the remaining cases, two main interpretations of engagement emerge (reported in Table 1). The first definition is proposed by [50], which states that engagement is "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 the interaction, evaluating whether to stay involved, and deciding when to end the connection" [50]. The proposed interpretation of engagement identifies engagement as a continuous and synchronous process that has a clear beginning and an end [5].[1] Additionally,

[1] Similarly, [51] referred to it as a process "subsuming the joint, coordinated activities by which participants initiate, maintain, join,

the interpretation of [50] assumes a dynamic nature (i.e.,, changing over time and between interactions), which can be framed by user actions [5]. While the first definition describes engagement as a process, the second most common definition is theorized by [52] and it defines engagement as "the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction". It interprets engagement as a quality metric of the interaction.

In other studies, the authors mentioned different definitions of the concept. The work of [30] refers to the definitions of [52] and of [53], giving the reader some insights on engagement without specifying the connection between the two theories. The definition proposed by [53] identifies four main discrete events in the dynamics of engagement: point of engagement, period of sustained engagement, disengagement, and re-engagement. Even if [53] depicts engagement as a dynamic process, in their formal definition they refer to it as a "quality of user experience characterized by attributes of challenge, positive affect, endurability, aesthetic and sensory appeal, attention, feedback, variety/novelty, interactivity, and perceived user control". In their work, [30] recognized the complexity of the engagement concept, and they focused on the "with-me-ness" concept proposed by [54], which is the extent to which the human is with the robot during an interactive task. On the contrary, the studies conducted by [26, 27] associate the definitions of engagement provided by [50] and [52] with the description of the concept of social engagement. Following the taxonomy theorized by [55], social engagement represents the involvement of the person with a robot capable of sociable and friendly interaction. They stated that social engagement differs from task engagement and social-task engagement, due to the different conscious focus during the interaction. In the first case, i.e.,, task engagement, the human finds himself immersed in the task (enjoying and concentrating on the inclusion in the task),[2] ignoring the robot's presence. In the latter case, i.e., social-task engagement, the individual finds himself with a socially capable robot, where both the individual and the robot work together to perform an explicit task. Similarly, [42] focused on the social dimension of engagement, mentioning the definition of [52], to underline the view of interpreting engagement as a quality of interaction and mentioning the definition provided by [11], which describes engagement as the "measure of the intention-to and the quality-of interaction as perceived by the user".

Some of the remaining studies rely on definitions of engagement, which are strongly related to the application setting and the number of involved participants. In the educational setting, [47] recalls the concepts of social- and task- engagement as components of productive engagement, defined "as the level of engagement that maximizes learning" performances. In the same context, [44] refers to the definition of collaborative engagement provided by [57], which states that "engagement refers to a student's participation in the learning process, and it is considered an expression of internal state, such as commitment, motivation, or interest". Considering the hosting application scenario, [43] defines the visitors' engagement as the probability that the visitor will reply to the robot's utterances. In a conversation scenario, the authors of [36] adopted the definition of group engagement reported in [11], expressed as "the joint engagement state of two participants interacting with each other and a humanoid robot". In the multi-party educational scenario proposed in [48], the engagement concept recalls the definitions of social/task engagement, but it also refers to "how interested the learner is in taking part in the interaction or being an active listener". It highlights the need to evaluate the engagement state of the group as well as the engagement state of each individual, separately.

3.2.1 Engagement's Components

Besides providing a proper definition of engagement, several works share the idea that engagement is a complex phenomenon, composed of multiple constructs that are strongly related to each other while being individually identified through specific behavioral indicators [44]. The components we are referring to are affective, cognitive, and behavioral constructs (see Table 1). The affective component of engagement is usually reflected by emotions and reactions between the parts (human and robot) involved in the interaction [18]. It encompasses the feelings, enjoyment, attitudes, and moods of the involved users [47]. In most of the works, the affective component of engagement is embodied by the enjoyment [36] and enthusiastic feelings [26]. The predominant connection between positive emotions and engagement is enforced by the theory that "positive emotions give a signal of purpose and excitement to the brain, accelerating learning and enhancing motivation" [35]. The cognitive facet of engagement includes some conscious components such as the effort [44, 47], investment [18], and attention [36] of the parts in the task and in the interaction. To a lesser extent, some works also include a behavioral aspect in the definition of engagement. Taking the definition from the Pediatric Assessment of Rehabilitation Engagement scale, [18] defines behavioral engagement as a proactive tendency to adapt to the changes and experiences of the interaction, as well as sharing intentions and desire to improve or change the interaction. It can be

Footnote 1 continued
abandon, suspend, resume or terminate an interaction". This interpretation enriches the one theorized by [50], by including the concepts of abandon, suspension, and resuming the interaction experience.

[2] This concept recalls the flow theory proposed by [56], and the concept of rapport theorized by [9].

Table 1 Overview of the definitions and the components of engagement, explicitly stated by the authors

	Definition			Components		
Work	Sidner et al. [50]	Poggi et al. [52]	Other	Affective	Cognitive	Behavioral
[46, 47]			x	x	x	
[48]			x	x		
[18]	x			x	x	x
[20]			x	x	x	x
[42]		x	x			
[44]			x	x	x	x
[37]			x			
[34]	x					
[43]			x		*	
[30]		x			*	
[45]	x					
[33]		x				x
[35]			x	x		
[27]	x	x				
[36]			x	x	*	
[11]	x					
[28]	x					
[26]	x	x		x	*	
[25]		x		x	*	
[29]	x					
[24]		x		x	*	

*Indicates that only attention is considered

simplified as the motivation [58], which mostly encourages action and participation in the task [5].

In [20, 44], the three constructs are conjoined. Similarly, [18] investigated each construct independently and conjointly with the others. Even if separating each dimension is a gross simplification [47], most of the works just consider one of the aspects when dealing with the definition of engagement. As shown in Table 1, attention is the component most often related to engagement. In 4 of the six works in which attention is mentioned, this aspect is investigated in tandem with the affective component. Similarly, [47] includes affective and cognitive components with social/task engagement for defining productive engagement.

3.3 Automatic Assessment

Engagement detection and recognition task is intended as the perceptual capability of correctly identifying the user state during the task and the interaction. It is commonly treated as a prediction problem, where the performances are evaluated by comparing the predicted engagement label/value with a ground truth one. Based on the categories used for the assessment, user engagement has been considered a discrete (i.e.,, binary or multi-class) or continuous state. From this review, a common pattern of automatic assessment emerged, composed of three main steps, namely:

- Data annotation: it refers to ground truth assessment, thus to the process of generating the engagement labels. The ground truth value could be associated with a self-report score, or it could be performed manually (i.e, performed by a group of experts that label the data recorded during the interaction, based on a common annotation scheme), or automatically (i.e., generating labels without a human expert in the loop).
- Features extraction: it relies on automatically extracting the features that describe the engagement concept. Before feeding the extracted features to the prediction framework, some studies include a correlation analysis to evaluate the significance of the extracted features.
- Automatic Prediction: application of ruled-based and/or machine learning algorithms on the extracted features.

Each phase is detailed in the following subsections.

3.3.1 Engagement Annotation

The annotation procedure is detailed in 21 works out of 28. The work of [27] is the only one that associated as ground truth the final engagement score of a self-report questionnaire, based on the following dimensions: quality of the interaction (adapted by the social engagement domain of

[59]), friendship (i.e., help and self-validation domains of [60]), and Perceived affective interdependence (as social presence measure [61]). In most of the remaining works (see Table 2), the annotation procedure was manually performed by third-party observers, usually expert coders, who separately assign the engagement label based on a common annotation scheme. In five works, the annotation is performed automatically. Namely, the work of [35] associated engagement values with a cluster of emotions. Similarly, the work of [47] applied unsupervised learning methods for generating labels. The work of [45] proposed a personalized active learning approach to get the majority of engagement labels, requiring a small sample set of annotated videos.

The annotation process is based on the common trend of considering engagement detection as a binary classification problem. In 10 works, data were labeled as "engage" or "not engage", based on whether certain behaviors were (not) present. In the work presented by [36], the label associated with group engagement is given based on the similarity of the degree of engagement of the participants. If both participants were engaged in the interaction with the robot, then the group engagement is similar, otherwise not.

To a lesser extent, automatic engagement detection is treated as a multi-class classification problem (7 works out of 28). Besides the two labels reported in the binary classification, some of these works include a third discrete state representing "a partial degree of engagement" [30], a "mid-engagement" [45] or a "neutral" state [35, 40, 48]. Three labels are also used in [41] to also distinguish a user that is "not seeking engagement". Similarly, [34] used four different labels related to the different phases of engagement: "approaching", "interacting", "leaving" and "uninterested". A continuous score was associated with the engagement state by [39, 42], ranging [0,1] and [-1, +1], respectively. In [46, 47], the labels used belong to the concept of Productive Engagement (as shown in Table 2). In the remaining cases, the authors did not specify the labels used for the classification task.

Annotated Datasets in HRI Even if some works included the adopted annotation scheme, the annotated datasets are usually not publicly available or vaguely described. A complete overview is reported in Table 2 and in the Supplementary Material. A brief description of the publicly available datasets is reported below:

- *Engagement datasets* In the work of [41], several engagement datasets have been created, involving young adults interacting with the James Robot Bartender. The multimodal corpora are composed of annotated video recordings and system logs of several participants playing the role of customers in a drink-ordering scenario, as in [62]. The anonymized and annotated corpora can be downloaded as reported in [41].
- *PE-HRI dataset* The dataset consists of team-level data collected from 34 teams of two (68 children), where the children, aged between 9 and 12, are involved in a learning activity using the JUSThink platform. The JUSThink platform consists of two screens and a QTrobot acting as a guide and a mediator. The dataset contains the team-level multi-modal behavioral data (i.g. log files with speech behavior, setup, gaze patterns, and affective states), team-level performance, and learning metrics. More details on the dataset and download procedure are reported in [63].
- *PE-HRI-temporal* The dataset is composed of the same information reported in PE-HRI dataset, with the addition that the features were computed in windows of 10 s. More information is reported in [64].
- *UE-HRI dataset* The User Engagement-HRI (UE-HRI) dataset consists of 195 recordings of humans freely interacting with the robot Pepper, standing in a fixed position. In the proposed setup, the participants were free to join the interaction if they wished, free to leave when they wanted, and were expected to behave in an unconstrained way [65]. The recordings belong to a wide range of heterogeneous sensors, namely: a microphone array, cameras, depth sensors, sonars, and lasers, along with user feedback captured through Pepper's touch screen. All data streams available on Pepper are packaged in the open-source Robot Operating System and indexed using the robot timestamps (to avoid synchronization issues). A subset of 54 interactions (each one lasting between 4 and 15 min) is freely available for download and use.[3]
- *MHHRI* The Multimodal Human–Human–Robot Interactions (MHHRI) Dataset [66] was introduced for studying the relationship between engagement and personality simultaneously in human–human interactions (HHI) and human–robot interactions (HRI). It is composed of a set of multi-modal data recorded during 48 interactions, in which participants asked personal questions to each other. The recorded data belongs to Kinect depth sensors, egoview cameras (worn by the participants), and biosensors. The engagement state of the users was assessed with a post-study questionnaire asking the participants about their perceived enjoyment of the interaction, and by an external annotator (as reported in [36]). Further information is detailed in [66].
- *Month-length intervention dataset* Despite not having a proper name, the work of [20] attaches, as supplementary material, the dataset built and used in their study. It is a multimodal dataset containing the annotated engagement value, as well as the visual, audio, and game performance features extracted from the recordings of each

[3] https://adasp.telecom-paris.fr/resources/2017-05-18-ue-hri/.

Table 2 Overview of the annotation procedures adopted in the reviewed works

Work	Procedure	Label type	Label list/meaning	Collected data	Dataset
[48]	Manual	Discrete (audio data) Discrete (visual data)	Very engaged, engaged, neutral disengaged, very disengaged High- and low- engagement	Dialog log, survey audio and image stream	N.P.A.
[47]	Automatic	Discrete	Productively engaged Non-productively engaged[a]	Log files, image and audio streams	PE-HRI
[18]	Manual	Continuous	High- and low- engagement	Image stream	N.P.A
[20]	Manual	Discrete	Engaged, disengaged	Log file, image and audio streams	Available here
[42]	Manual	Continuous	High engagement (1) Low engagement (0)	Image stream	TOGURO (N.P.A.) UE-HRI [65]
[44]	Manual	Continuous	All behaviors detected (45) No behaviors detected (0)	Image and audio streams	N.P.A.
[37]	Manual	Continuous	All behaviors detected (1) No behaviors detected (0)	Image and audio streams	N.P.A.
[34]	Manual	Discrete	Approaching, interacting Leaving, uninterested	Image stream	N.P.A.
[43]	Manual	Discrete	engaged, partially engaged, disengaged	image stream and laser data	N.P.A.
[30]	Manual	Discrete	Engaged, partially engaged, disengaged	Image stream	N.P.A.
[45]	Automatic	Discrete	Low, med, high Engagement	Image stream	N.P.A.
[33]	Manual	Discrete	Engaged, disengaged	Image and audio streams	N.P.A.
[38, 39]	Manual	Continuous	Completely engaged (+1) Completely disengaged (-1)	Image and audio stream, physiological data	N.P.A.
[35]	Automatic	Continuous	Engaged (1), Neutral (0.5) Disengaged (0)	Image stream	N.P.A.
[41]	Manual	Discrete	Not Seeking/Seeking engagement, and Engaged	Image and audio streams	Engagement datasets
[27]	Self-assessment	Discrete	High- and Low- Engagement	Log files	N.P.A
[36]	Manual	Discrete	Engaged, Not Engaged (individual) Similar/Dissimilar (group)	Image stream	MHHRI [66]
[11]	Manual	Discrete	Engagement Disengagement	log files, image and audio streams	Vernissage dataset [67]
[28]	Manual	Discrete	Engaged, not engaged	Image stream	
[26]	Manual	Discrete	Medium-to-high and Medium-to-low engagement	Log files Image stream	Inter-ACT (N.P.A.)
[24, 25]	Manual	Discrete	Engaged, not engaged	Image stream	N.P.A

N.P.A.: Not publicly available

[a] Learner profiles associated with the same concept of productive engagement

child (seven in total) with a clinical diagnosis of ASD from mild to moderate ranges, that interacted with Kiwi robot in several sessions over one month. Further details are reported in [20].

- *Vernissage dataset* The Vernissage dataset [67] contains 13 sessions of NAO interacting with two persons. The robot serves as an art guide, explaining the paintings to the users and then quizzing them in art and culture [67]. The dataset comprises synchronized recordings from multiple auditory, visual, and robotic system information channels. These are multi-party interactions that were manually annotated with several nonverbal cues, such as speech utterances, 2D head-location, nodding, visual focus of attention (VFOA), and addressees [67]. The dataset is publicity available upon request.[4]

3.3.2 Features Extraction

The features used for the automatic assessment of user engagement can be clustered into two categories: behavioral features and context-based features. Under the umbrella term of behavioral features, we grouped the features related to user behavior and user emotional state. As reported in Table 3, they mostly rely on the data recorded by the sensors mounted over the robotic platform and surrounding the environment. On the other hand, the context-based features refer to the information stored in log files (see Table 3), which keep track of the robot's behavior and/or of the task performed by the user(s) with the robot.

Behavioral Features From visual data, the descriptors used to assess engagement are body posture, head pose, eye gazing, and facial expressions. Of the reviewed works, six studies adopted only visual features for detecting engagement. The work of [25] represents the first attempt at detecting engagement considering the body postures and body motions of the individual. In this work, expressive postural features (i.e., body lean angle, slouch factor, quantity of motion, and contraction index) of children playing chess with iCat robot were extracted from videos recorded from the lateral view. Since head pose has been proven to be highly correlated with human engagement in face-to-face interactive scenarios [1], several works combined the features related to body activity and the head pose of the users as descriptors of engagement. One case study is described in [36], which used the Kinect RGB recordings to extract individual features based on body posture and quantity of motion. In the same work, as descriptors of group engagement, interpersonal features (i.e., the global quantity of movement, relative orientation and distance of the participants, and relative orientation to the robot) are obtained geometrically from the individual features in tandem with head orientation (i.e., Visual Focus of Attention), a geometrical approximation of the user's eye gazing. Similarly, [30] extracted 3D body poses of children freely interacting with the NAO robot to derive high-level features of body motion and head orientation (i.e., the angle between the child's gaze and the robot, the angle between the child's body and the robot, and the distance of the hands from the respective shoulders). In this work, multiple cameras were installed in the environment and the authors included a preliminary step for fusing and interpolating the child's pose detection from multiple views. In [30], the head pose was computed geometrically from the facial key points of interest. Analogously, [34] extracted 55 descriptors of engagement, including an affective component (i.e., the user is smiling) into the categories listed so far (i.e., body posture and head pose). The same authors distinguished between features associated with head pose and eye gazing of the user, extracting pitch, roll, and yaw angles in the first case, and examining the focus of the gaze in the second case (i.e., the user is looking at the Kinect, the user is looking away from the Kinect, left eye is closed, right eye is closed). The work of [37] included three features derived from the Laban Movement Analysis (i.e., space, weight, and time) to represent the dynamics of human movement (i.e., the effort), as well as 2-dimensional positions of the facial landmarks to describe the affective state of the participants. In recent work, [18] used features belonging to facial expressions (e.g., smile, inner brow raise, brow raise, brow furrow, mouth open), emotions (e.g., joy, anger, fear, disgust, contempt, sadness, and surprise), head pose (i.e., roll, pitch, and yaw angles), eye gazing (e.g., spatial coordinates), body postures (i.e., 2D body pose of 25 main joints), and additional behavioral indices (i.e., attention, disappointment, relax).

From this review, it emerged that another common strategy is to extract descriptors of engagement from visual and audio data, simultaneously. In a long-term experimental scenario, [20] investigated the engagement of children interacting with the Kiwi robot, considering the voice quality (i.e., harmonicity, intensity, pitch frequency, and periodicity) in tandem with eye gazing, head position, and facial expressions. Similarly, [44] selected voice quality features (i.e., intensity and pitch frequency) and alignment features (i.e., responding to the robot's question, extending or elaborating talks by the peer or the robot, initiating a talk) in addition to eye gazing, body posture and smiling activity as social descriptors of engagement. Another frequent approach is to consider the quantity of speech occurring in the interaction, referred to as speech activity. In [47], visual behavioral features (e.g., smile, emotions in terms of positive/negative valence and arousal, eye gazing) were combined with some features derived by the quantity of speech of the participants (e.g., speech activity, silence, small pauses, speech overlaps). Without considering visual data, [32] detected the less engaged individual by monitoring the total speak-

[4] http://vernissage.humavips.eu/.

Table 3 Overview of the extracted features, i.e., behavioral (on the left) and contextual (on the right)

Work	Facial expressions	Emotion	Head pose	Eye gaze	Body posture	Speech activity	Voice quality	Text content	Others	Robot's speech	Robot gaze	Robot's expressions	Task difficulty	Task duration	Performed actions	Progress in the game	Task repetition
[48]	●	○	○	●	○	○	●	●	○	○	○	○	○	○	○	○	○
[46, 47]	◑	●	○	●	○	●	○	○	○	○	○	○	○	○	●	○	○
[18]	●	●	●	●	●	○	○	○	○	○	○	○	○	○	○	○	○
[32]	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○
[44]	◑	○	○	●	●	●	●	○	○	●	○	○	○	○	○	○	○
[20]	●	○	●	●	○	○	●	○	○	○	○	○	●	●	○	●	●
[42]	○	○	○	○	○	○	○	○	⊖	○	○	○	○	○	○	○	○
[37]	●	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○
[34]	◑	○	●	●	●	○	○	○	○	○	○	○	○	○	○	○	○
[43]	◑	○	○	●	○	○	○	○	■	○	○	○	○	○	○	○	○
[30]	○	○	●	○	●	○	○	○	○	○	○	○	○	○	○	○	○
[45]	○	○	○	○	○	○	○	○	⊖	○	○	○	○	○	○	○	○
[33]	○	○	◑	●	○	○	○	◑	○	○	○	○	○	○	○	○	○
[38]	●	○	●	●	●	○	●	○	⬢	○	○	○	○	○	○	○	○
[39]	○	○	○	○	○	○	○	○	⊖	○	○	○	○	○	○	○	○
[31]	○	○	○	○	○	○	○	○	●	○	○	○	●	○	○	○	○
[31]	○	○	●	○	●	○	○	◑	○	○	○	○	○	○	○	○	○
[35]	○	○	○	○	○	○	○	○	⊖	○	○	○	○	○	○	○	○
[41]	○	○	●	○	●	●	○	○	○	○	○	○	○	○	○	○	○
[40]	○	○	●	○	○	○	●	○	○	○	○	○	○	○	○	○	○
[36]	○	○	●	○	●	○	○	○	○	○	○	○	○	○	○	○	○
[11]	○	○	●	●	○	●	○	○	○	●	○	○	○	○	●	○	○
[28]	◕	○	○	●	●	●	○	●	○	●	○	○	○	○	●	○	○
[26]	◑	○	●	●	○	○	○	○	○	○	○	●	●	○	●	●	○
[25]	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○
[29]	○	○	○	●	○	○	○	◑	○	○	●	○	○	○	○	○	○
[24]	◑	○	○	●	○	○	○	○	○	○	○	●	●	○	○	○	○

While black bullets mean that the mentioned features are extracted, half-colored bullets highlight that a reduced set of the mentioned features are extracted. Namely, in the first column, smiles and laughers are associated with vertically and horizontally half-colored circles, respectively. A 3/4-colored circle is used when both smile and laughter are considered. Otherwise, half-colored circles represent head nodding in the third column, and back-channeling utterances in the eighth column. In the ninth column, the proxemics features are represented by squared bullets, physiological parameters are represented by hexagons, and filled circles with a bar refer to deep features ing time. In the work of [41], the speech activity and the visual data are used to detect the location of who is willing to interact with the bartender robot James. This information is adopted in tandem with additional behaviors of the individuals (e.g., head pose and quantity of motion). Features related to head pose and eye gazing are merged with features related to speech activity in the work of [11]. Specifically, the authors distinguished between the moments in which everyone was speaking, silence, and laughing sounds. The works reported so far extract behavioral features from speech quality and speech quantity. Another common approach is to detect specific utterances in the spoken text. The type of utterances pronounced by the participants were used to define the type of social signals expressed by the child (e.g., question, answer, greeting, suggestion introduction, request), descriptors of engagement together with facial expressions, body posture, eye gazing and the presence of other annotated behaviors (e.g., headshake, hand-writing touching-hair, touching-face). Similarly, [29] identified as descriptors of sustained engagement, the observable behaviors occurring in the so-called *connection events*, namely: direct gaze, mutual facial gaze, adjacency pairs, and backchannels. Each event is modeled as a finite state machine, where the transition from one state to another depends on the occurrence (and detection) of specific human (and robot) behaviors. While the direct gaze event occurs when the responders look at the object gazed at or pointed by the initiator, mutual face gaze happens when both responder and initiator look at each other faces. The adjacency pair consists of two utterances by two speakers, with minimal overlap or gap between them (e.g., turn-taking), and the backchannels are events in which one individual (e.g., the responder) communicates to the other party (e.g., the initiator) comprehension and/or desire to keep listening (e.g., expressions of backchanneling are “hmm”, “yeah”, “uh-huh”) [29]. Behavioral features related to backchanneling are also considered in [31], in tandem with body and head pose analysis, and in the work of [33], together with eye gazing, head nodding features, and the presence of laughing events. In two different contexts (i.e., [40, 43]), proxemics is also included as a descriptor of the interpersonal distance between the user and the robot. As shown in Table 3, [38] included heart rate, electrodermal activity, and body temperature recorded by wearable devices as additional features to detect engagement in children with ASD. One of the latest trends is to use deep learning features as behavioral features. Another approach is to feed the image frames of the human–robot interaction through a pre-trained CNN (i.e., ResNet-50) to extract automatically the facial features [39, 45] as well as surrounding events [42] related to engagement, without deriving them geometrically or analytically. One advantage of deep learning approaches is that the features of interest do not have to be explicitly defined a priori, but they need annotated data [42].

Context-Based Features From the analysis of log files, it is possible to assess engagement keeping track of the user task performances, and the robot behaviors. The works of [24, 26] represent the first attempts to introduce context-based features in engagement detection, considering children playing chess with iCat robot. In their first work, the authors modeled children’s engagement by integrating features describing user behavior (i.e., the user is smiling, the user is looking at the robot) and two different levels of contextual information: game state and the presence of the robot’s facial expressions [24]. In [26], the authors included additional features related to the user task’s performance, namely: game evolution (i.e., the difference between the current and the previous value of the game state), captured pieces (i.e., to specify if the child or the robot or both captured a chess’ piece), user emotivector (i.e., the result of the mismatch between expectation and actual outcome of the user’s progress in the game) and user anticipation (i.e., whether the user looks at the robot immediately after making a move and before the robot generates a reaction). In the latest work, [27] proposed a new set of features for detecting engagement, which did not include any behavioral feature of the user. The authors reduced the set of context-based features related to the game, selecting the game state, game evolution, and emotivector as in [26], and adding game result, number of moves, and duration of the game to the list. In the same work, the authors increased the number of robot behaviors (that could affect the engagement of the participant), including four different empathetic behaviors: encouraging comments, scaffolding (i.e., providing feedback on the user’s last move), offering help (i.e., suggesting a good move for the user to play), and intentionally playing a bad move to favor the child.

Reconsidering the features related to the task performance, [20] included the duration of the game, the challenge level of the task, the number of task repetitions, and the number of incorrect mistakes of the user. Similarly, [47] defined some context-based features related to productive engagement, computing the number of times the children were performing certain actions (e.g., a team opened the instructions manual, a team added or removed an edge on the map) and the total number of actions performed in each session.

Focusing on the robot’s behavior, [29] considered some human-like robot’s behavioral cues, namely robot gazing (e.g., if the robot is looking at the user or the object), and speaking. Similarly, [28] considered the speech capabilities of the robot (e.g., asking questions, answering, greeting), which are identical to the speech features extracted from the behavior of the human partner. The speech activity, as well as the topic of the speech, and the person the robot is referring to in the speech, are the robot’s features that are considered by [11].

3.3.3 Automatic Prediction

Three methodologies characterize engagement assessment models: rule-based, machine-learning-based, and deep learning-based.

Rule-Based Methods As shown in Table 4, three works adopt a rule-based model, selecting different rules among them. The work of [29] relies on the presence of four main social signals (i.e., direct gaze, mutual facial gaze, adjacency pair, backchannelling), each one detected by a dedicated state machine, that sends to a general integrator the final engagement score. In their work, [44] adopted a formula to compute engagement at each timestamp. Namely, engagement is obtained by summing the values of behavioral,[5] cognitive, and emotional engagement, each one described by specific behavioral features (reported in Table 1). On the other hand, [32] adopted a threshold-based rule for detecting the passive subject in the interaction, by monitoring the number of turn-takings of each individual. A rule-based method is also present in the works of [25, 41], in comparison with other machine-learning approaches. In [25], four different algorithms are compared to the rule-based approach, namely: an alternating decision tree, an additive logistic regression, a metaclassifier for handling multi-class datasets, and a multinomial logistic regression model. The best performance in terms of average accuracy was obtained by the rule-based approach and the alternating decision tree (average accuracy=82%). Similarly, [41] compared the rule-based approach with several training and classification procedures, reported in Table 4. In detail, the authors underlined the need to treat engagement detection more like a sequence labeling problem, than a frame-level classification task since the evaluation of engagement should consider current and previous estimates of engagement. In general, the work of [41] concludes that the performances of the rule-based and the machine learning methods are comparable in terms of accuracy in both offline and online classification tasks, even if the rule-based method tended to be less precise in detecting the changes of engagement state.

Machine-Learning Methods The Support Vector Machine (SVM) algorithm with Radial Basis Function (RBF) is the most used machine-learning algorithm for engagement detection. When used alone, the authors reported different recognition rates, obtained by combining different feature sets. In the chess game scenario, the SVM recognition rates achieved the best performance when a reduced set of affective features (i.e., valence and interest) in tandem with one contextual feature (i.e., user anticipation) were selected (accuracy=93.75%) [26], and when a subset composed by game-based and turn-based features were considered (F-measure=0.80) [27]. In the individual engagement assessement in a group scenario, [11] compared the classification performances of SVM fed with individual features (e.g., features of the primary user), interpersonal features (e.g., features of another person present), and of robot's features. The results show that the best accuracy was obtained by testing the algorithm with only the individual features (accuracy=75.91%). Regarding group engagement, [36] compared the performances of two supervised algorithms: SVM with linear kernel and the Random Forest (RF). In this case, the best performances of detecting individual and group engagement were achieved by using the RF algorithm. The best classification result for detecting individual engagement was achieved using individual features with personality labels (F-measure=0.81). Similarly, the best classification result of group engagement was achieved adopting only individual features, and in conjunction with interpersonal features and personality labels, respectively (F-measure=0.60). In learning a second language context, a combined SVM classification was proposed to assess emotional engagement from video, obtaining the best detection performance for low engagement level(accuracy=79%) [48].

Another popular technique for detecting engagement is the Bayesian network. The outcomes of [24] confirm that a multimodal feature set (behavioral and context-based features) improves the recognition of the engagement of children playing with iCat robot (ROC Area=0.96). Similarly, [40] designed a dynamic Bayesian network for recognizing engagement in children affected by ASDs interacting with the NAO robot, by transforming the qualitative evaluation from professional caregivers as parameters of the model by fuzzy logic. The outcomes of the proposed model coincided with the expert's ratings, with an accuracy of 93.60%. A Bayesian model is also adopted by [31] for assessing engagement in users listening to the story told by the Reeti robot, and by [33], for detecting user engagement during the conversation with ERICA robot. In the latter case, the authors included, in the Bayesian model, a latent character representing the perception of engagement of the annotator. The hypothesis behind this strategy is that the annotation process is influenced by the subjectivity of the perception of engagement from the annotator's point of view. Comparing the proposed hierarchical Bayesian model with other machine learning and deep learning models (reported in Table 4), the results confirmed that the inclusion of the latent character improved the overall engagement recognition performance (accuracy= 70%) [33].

For estimating the engagement of visitors entering a shop, [43] adopted a logistic regression model, obtaining an 88.9% accuracy rate in online fashion. The logistic regression model

[5] The author mentioned behavioral engagement as *bodily engagement* as expressed by the following children's behaviors: eye contact, gaze orientation (looking at the robot), body orientation (facing peer or robot), posture (e.g., leaning forward), gestures or enactments of ideas (e.g., representing a concept), and facial expressions (e.g., smile).

also appears in the list of algorithms tested by [20] for detecting engagement in long-term scenarios. Among the seven algorithms used for the classification task, the best performance was obtained by the gradient boost decision tree algorithm (AUROC=88%). Interestingly, [20] compared different classification algorithms, fed with different sets of features (i.e., visual features, audio features, game performance features, and all features together), proving the visual features outperformed the classification task. As shown in Table 4, the work of [18] also compared different classification strategies, applied to different feature sets. In [18], the group of features was obtained by applying different feature selection techniques (e.g., Best First, Correlation Attribute Evaluation, and Random Search). The Random Forest got the best classification performances for almost every considered dataset. On the other hand, testing the model on a specific set of features, the results of [28] confirmed that the decision tree C4.5 algorithm outperformed the other tested approaches (recall= 84.83%). Only the works of [47] adopted an unsupervised algorithm (i.e., K-means) for clustering the behavioral and context-based features associated with Productive Engagement. Interestingly, the results suggest that individual and merged features could be used for automatic labeling and, thus for automatic assessment.

Deep-Learning Methods The third cluster of classification models regards deep learning techniques, detailed in Table 5. In some works mentioned above, deep learning models were included in the list of classification methods, as in [18, 20, 28, 33], without specifying the details of the network. On the contrary, [37] proposed a multi-channel and multi-layer convolutional neural network (CNN) for their temporal multi-label classification problem, and then compared the performance of the proposed model with some standard machine learning algorithms. The proposed network is composed of two convolutional layers, to identify temporal data patterns, and three dense layers for the classification [37]. The evaluation results returned that the proposed model achieved an accuracy comparable to the machine learning algorithms, without outperforming. The best result in detecting children engagement while freely playing with the robot was the RF algorithm (accuracy= 81%). Similarly, [35] proposed a multi-layered convolutional neural network, composed of five convolutional layers and three fully connected layers, which obtained a classification accuracy equal to 82%. A convolutional module is also present in the deep learning architecture[6] proposed by [42] for detecting the engagement of the museum's visitors. In the proposed architecture, the role of the convolutional module (i.e., a pre-trained ResNetXt-50) is to extract the frame features from the video. These features are then passed to a recurrent module, composed of a single layer of Long Short Term Memory (LSTM) with 2048 units, which extracts the temporal behavior of humans within the considered time window. In the end, a 2048 × 1 fully connected layer returns the predicted engagement value. The model was trained and tested on the TOGURO dataset. Additionally, the same model was also tested on the UE-HRI dataset [65], achieving high accuracy (AUC = 0.89). It is worth mentioning, that the work of [42] is the only work in HRI in which the same engagement detection model is tested on different datasets, belonging to different interaction contexts. A convolutional module is also included in the architecture proposed by [45] for extracting the engagement values of children interacting with TEGA robot. Namely, the authors chose a pre-trained CNN for extracting the video features. These features are then fed into the Temporally Consistent Deep Q-Learning (TC-DQL) model, composed of LSTM cells followed by linear fully connected layers. Deep Q-learning is here used to select the most appropriate action to perform, namely: store the video for further labeling or estimate the engagement level of the child. The same authors proposed a novel personalized deep learning architecture, i.e., CultureNet, for estimating the engagement of children with ASD by using their faces and information about their culture in [39]. This architecture is composed of a CNN layer for extracting the most discriminative (deep) facial features (i.e., Faster R-CNN [68]) and five fully connected layers, that exploit the cultural label information in learning the engagement levels. Namely, the culturalization step is performed using culture-specific data to fine-tune the last fully connected layer of the network [39]. In the same context, the same authors propose an additional personalized deep learning architecture, i.e., PPA-net (Personalized Perception of Affect network), to automatically perceive children's affective states and engagement during robot-assisted autism therapy [38]. In this work, the personalized approach relies not only on cultural information but also on other contextual information (e.g., demographic and behavioral info), specific to each individual. At features layer, the PPA-net handles missing and noisy data by adopting supervised auto-encoders, which convert signals into hidden representations. At the second level (i.e., context layer), the feature representation is augmented by the expert's input, which represents the complete assessment of the child. In the last layer, a multitasking learning phase is included to predict emotional valence, emotional arousal, and engagement. An auto-encoder method is also adopted by [34] to estimate user engagement. Single- and multi-task learning personalization based on Efficient Neural Architecture Search is proposed by [46] to personalize productive engagement models. The results of [46] highlight that personalized models performed better than the non-personalized ones and that the speech modality was the most informative feature for predicting productive engagement. In the child-robot interaction con-

[6] The trained model and the software is publicity available at https://github.com/LCAS/engagement_detector.

Table 4 Overview of the prediction models for engagement assessment adopted in the reviewed studies, i.e., rule-based (italic), machine-learning-based (bold), and deep-learning-based (bolditalic)

Work	*Rule-based*	**Naïve Bayes**	**Nearest-neighbors**	**Random tree**	**Random forest**	**Linear regression**	**Logistic regression (LR)**	**Additive LR**	**Multinomial LR**	**SVM-linear**	**SVM-RBF**	**Decision tree**	**Bayesian network**	**Meta-classifier**	**Conditional random fields**	**K-means**	***DL methods***
[48]										x	x						x
[47]																x	
[2]				x	x	x											x
[32]	x																
[20]		x	x	x			x				x	x					x
[42]																	x
[44]	x																
[37]			x	x							x	x					x
[34]																	x
[43]							x										
[30]				x							x						x
[45]																	x
[33]						x					x		x				x
[38, 39]																	x
[31]												x					
[35]																	x
[41]	x	x	x			x			x		x				x		
[40]													x				
[26, 27]											x						
[36]				x						x							
[11]											x						
[28]		x		x								x					x
[25]	x							x	x			x		x			
[29]	x																
[24]													x				

Table 5 Overview of deep-learning models adopted for engagement classification

Work	MLP Regressor	Multilayer Perception	Neural Network	CNN	Autoencoder	Recurrent NNs	Multi-Task Learning	Other/ Custom
[46]							▒	▒
[48]							■	
[18]	■	■						
[20]			■			■		
[42]				▒		▒		
[37]				■				
[34]					■			
[30]			▒			▒		
[45]								■
[33]		■						
[39]				■				
[38]							■	
[35]				■				
[28]		■						

While the black cells mean that the specific model is used alone, the gray cells mean that the highlighted models are combined

text, [30] proposed an alternative approach based on deep learning for estimating engagement. The proposed neural network is composed of three fully connected layers, a single LSTM layer in the middle, and a fully connected layer coupled with a softmax function on the output. The classification performance of the proposed network outperformed the other popular classifiers, achieving an accuracy of 77%. A bidirectional LSTM is also adopted by [48] for assessing the arousal level of the speech data, only.

4 Discussion

One of the outcomes of this review is to identify challenges and opportunities for future research works, by considering the limitations mentioned in the reviewed works. In Table 6, we clustered the current barriers, reporting some suggestions for future improvements.

4.1 Engagement Concept

From the analysis of the current literature on engagement in human–robot interaction, we can conclude that the definition of engagement is still not clear. The main definitions of engagement that emerged in this review were also reported in [33, 42]. While [33] specifies that the great difference between the definitions relies on the included components (i.e., the definition provided by [50] is mostly related to the concepts of attention and involvement, while the interpretation of engagement as metrics of the quality of interaction recalls also concepts like interest and rapport), [42] suggests that the definition influences the classification task (e.g., adopting the definition of [50], the engagement detection aims to detect the different phases that compose engagement process). In this review, works ranging over different application scenarios and classification strategies mention the same engagement definition, suggesting that the context of interaction and the annotation procedure have no link with the adopted definition. This is why future research works should provide a better and more precise definition of engagement, which will guide the researchers in designing appropriate prediction strategy to detect it automatically. More emphasis should be also invested in specifying the concept of group engagement, moving the focus from the individual to the overall group. As reported in [36, 49], deriving the engagement of a group member from the analysis of individual engagement and the interpersonal interaction with the other members is not effective. By formulating a precise definition of group engagement, an alternative approach could be designed. The same approach should be used to clarify the role of engagement in long-term interactions since the current definitions mostly refer to short-term interactions.

This lack of clarity makes also it difficult to identify the components and the aspects that we should consider when trying to assess engagement automatically, as well as their relationship. Merging the data reported in Table 1 (i.e., component definition) and Table 3 (i.e., list of the extracted features), it is hard to find a clear and strong correspondence between the two, highlighting that the lack of clarity in the engagement definition is reflected also in the feature selection process. Additionally as reported in Table 1, most of the works consider attention and the affective component of engagement as two separate aspects, without investigating their relationship (i.e., how they are related, when they are related, etc.). It may suggest that engagement is present when the values associated with both categories are high. Addition-

Table 6 Challenges and opportunities on engagement definition and assessment

Keyword	Barrier/limitation	Challenge/opportunity	Research topics
Engagement concept	Vagueness of the definition [30, 44]	Propose a new definition of engagement, which considers the application scenario, the robot's task, the target users, and the type of interaction	Revise the definition of engagement based on the involved participants, the context, and the robot's capabilities
			Highlight the relationship between the meaning of the concept and its automatic assessment
			Clarify the concept in multi-party and long-term interactions
	Limited interpretation of the components [26]	Propose a precise taxonomy of the engagement components	Clearly uncover the potential relationships between engagement and its components
			Investigate the relationships between components and the context of interaction
Annotation	Manual and offline procedure [20, 27]	Labelling throughout ongoing interaction	Define a standard approach or shared policy for assessing ground truth values during the ongoing interaction
			Introduce novel labelling procedures to select the most informative instances that need labelling, and reduce annotator workload
			Increase variance in data, deploying the frameworks in real-world settings
	Quality of datasets [37, 38]	Design secure data-sharing framework for improving the quality of the annotated data	Obtain balanced datasets, integrating more annotated data
			Conceptualize a secure data-sharing framework to host databases belonging to similar interaction context
Engagement features	Limited number of features [25, 30, 35, 37, 40]	Fusion of multimodal features for improving the quality of the assessment, by selecting appropriate cues	Include multiple data modalities
			Define the trade-off between quantity and quality of features
			Correlate features of interest with the definition of engagement as well as with its components
	Low quality of features extraction tools [11, 28, 32, 35, 38, 40, 44]	Improve hardware and software technology to obtaining reliable engagement features, that can be tested in real-world settings	Design more robust, not invasive, ecological hardware solutions
			Improve software tools accuracy for behavioral analysis, integrating pre-processing techniques
			Design not invasive alerts that make the user aware of any technical issue that may happen during the interaction
	Context-dependent features [25, 26]	Contextual features could not be used, exactly as they stand,in a scenario that is substantially different	Investigate the relationship between engagement and additional users' characteristics (i.e., age, culture, nationality)
			Spawn the possibility of adapting existing engagement model on different contextual factors

Table 6 continued

Keyword	Barrier/limitation	Challenge/opportunity	Research topics
Automatic Prediction	Static framework [37, 38]	Design engagement estimator as a dynamic framework	Include temporal parameters
			Learning and adaptation over time
	Off-line deployment of engagement models [18, 20, 30, 33, 34, 40–42, 47]	Deploy prediction framework online, in real-time interactions improving the learning capacity of the model	Transform the features of interest into time-series
			Usage of engagement prediction for rewarding the robot's action
	Parameter tuning not performed [26, 41]	Design an optimized framework able to handle previously unseen individuals	Fine-tuning the parameters of the engagement estimator
			Determine the complementarity of the classification result and the classification confidence
	Lack of interpretability of machine-learning and deep-learning models [47]	Identify novel techniques to monitor the learning performances of the models	Adopt eXplainable AI tools for highlighting hidden patterns
			Introduce and define objective measures to validate the obtained output
Experimental sessions	Limited sample set [18, 25, 27, 29, 30, 39, 40, 43]	Design experimental sessions involving a larger set of participants, characterized by different age, and social-cultural background	Enlarge the number of participants
			Investigate engagement expressions of different target users performing the same task
	Single experimental episode [31, 38]	Include multiple interaction episodes with the same participant on a longer timespan	Accessing multiple sessions
			Investigate more engagement in long-term interaction cases
			Analysis on the relationship between short-term and long-term engagement
	Laboratory setting [32, 35, 36, 38]	Move the experimental settings in real-world settings to improve the reliability of the proposed model	Design less constrained interaction settings, for a more naturalistic user behavior expression
			Improve the positions of the sensors in the scene for a more ego-centric view of the user
			Enlarge the possible interaction scenarios

The list of works reported in the second column refers to the works in which the corresponding limitation is pointed out

ally, [26] highlights that each component, especially interest and affective state, does not fully explain engagement when it is taken alone. The relationship between engagement and its components should be investigated in relation with the context [26]. From a deep analysis of the relationship among the components, it may be possible that additional dimensions of engagement emerge, which may be more appropriate to the context (e.g., imitation, behavioral contingency, and synchrony).

4.2 Annotation Procedure

The main limitation of the annotation procedure is associated to the manual annotation. Besides being the most adopted strategy,the drawbacks of this technique are multiple. On one side, it is a laborious and a subjective procedure. To overcome the former aspect, future works should foster the adoption of automatic annotation techniques (as in [35, 45, 47]), reducing the burden oh human annotators, while guaranteeing the quality of the annotations. Additionally, the manual annotation is strongly affected by the subjectivity of the annotators, as highlighted in [33, 48]. Even if the annotators involved in this task share the same annotation procedure, it requires the additional step of verifying the agreement between annotators each time to obtain reliable labels. To overcome this limitation, objective measurements of engagement should be defined and shared among the different works. Shared and standardized annotation tools (e.g., based on well-known inventories, such as the Pediatric Assessment of Rehabilitation Engagement introduced in [18] or the Temple Presence Inventory used in [36]) could may reduce this bias. This strategy may open other opportunities, like the improvement of the quality of the datasets, which are often unbalanced and not available for further studies. Due to privacy issues related to the type of data (i.e., images), a secure data-sharing framework should be conceptualized to simplify the comparison and evaluation of the same prediction model over different datasets, as highlighted in [37, 38].

Besides being manual or automatic, most of the reviewed annotation procedures were performed offline, namely at the end of the interaction. Thus, one possible improvement is to identify a way to associate ground truth values with engagement dimension(s) during the ongoing interaction [20, 27]. One attempt in this direction has been made by [45], which designed a deep reinforcement learning framework for active learning. Future research works should exploit this technique in other application scenarios. Additionally, the adoption of categorical labels may prevent the identification of engagement variations over the interaction [48]. Thus, whether choosing between categorical and continuous labels for engagement requires further discussion.

4.3 Engagement Features

The predominant presence of behavioral features extracted by the camera suggests that engagement is mostly described by nonverbal behaviors. Namely, comparing the performances of the same engagement detector with multiple sets of features, it often emerged that the non-verbal behaviors extracted by the visual sensors are more discriminating than the other alternatives [20, 38]. Similarly, previous studies showed that engagement detection performances improves when more than one feature is considered [25, 30, 35, 37, 40]. This is the reason why future works should define the trade-off on the quality and the number of features. Besides adopting correlation analysis and features selection strategies to identify the most representative features, future "cue-centric" strategies should focus more on investigating the relationship between the features of interest and the components or the engagement concept itself they are interested in. Recent trends tend to accept the quality of the features extracted in a black-box manner by deep learning networks, which could be pre-trained with different purposes. This approach may be affected by lack of interpretability, and from the evaluation results, it emerged that deep learning strategies mostly failed when compared with traditional machine learning algorithms [18, 20, 28, 33, 37]. We believe that by clearly identifying the relevant features, as well as clarifying the theoretical background behind them, the quality of the deep learning strategies could improve as well.

In addition, several works list in their limitation the low performances of the features' detectors (see Table 6). As an example, when working frame-wise, the skeleton tracker may not easily differentiate the skeleton of the participant of interest (e.g., children) and the skeleton of an additional individual present in the scene (e.g., members of the research team and/or parent), causing unreliable geometrical features as results [36, 37]. Erroneous speech analysis performances are caused by internet malfunctioning problems since most of the speech recognition tools require it [34], as well as the hardware's quality [20] and the presence of background noise in real-world settings [38]. In future works, it is advisable to include preliminary audio pre-processing, based on background noise reduction, speaker diarisation, and a better selection of audio descriptors [38], as well as clear alerts on the robot for making the user aware of the problem [34]. The latest reported limitation of engagement features is that they strictly depend on the context and the application scenario. As reported in [27], the list of context-based features needs to be defined, validated, and proved in every context, reducing the portability and the generalization ability of the proposed model. Additionally, the influence of user characteristics on the expression of engagement should be further explored [44]. In this direction, the works of [31, 36] investigate the role of personality in the detection of user engagement, obtaining promising results. Similarly, [38, 39] included the users' cultural background in their engagement detection model. The combination of social aspects with user characteristics may lead to a different interpretation of engagement, and thus, to a new category of engagement features.

4.4 Engagement Prediction

The main drawback of the proposed frameworks is that they are static, which means that temporal information is not included. Few works, i.e., [41, 42, 46] consider the temporal information, thus investigating the evolution of the parameters along the interaction. The introduction of the temporal domain, in the engagement detection is important to validate that the offline results (i.e., global) are also reflected by the pattern over time (i.e., local) [47]. One improvement of the engagement detection models is to become dynamic, learning over time and adapting to the user. For properly adapting the behavior of the robot to user engagement, future studies should focus not only on the prediction accuracy of the proposed methods but also on the interpretability of the results. Deep-learning models have the advantage of not explicitly defining the features of interest a priori, since they only require the phenomenon to be annotated [42]. However, deep-learning methods may lack interpretability [47], especially when used as black-box tools, which complicate the design of appropriate and effective robot interventions. In this direction, techniques of eXplainable Artificial Intelligence (XAI) could used to enforce the usage of deep-learning models, while identifying the cues and/or information of interest that could be used in the adaptation process. Similarly, XAI methods could be also exploited to properly validate the selected features of interest in machine-learning models.

Another limitation is related to the fact that most of the prediction models are trained and tested offline. The prediction models investigated in this review highlight that there is a general tendency to train and test the models at the end of the interaction, thus investigating their performances on recorded data. The only exceptions are represented by some works that adopted rule-based approaches (e.g., [31, 32]), in which the engagement value is computed based on the presence of certain behaviors of interest (or based on other simple logics) in real-time. Similarly, the works of [40–43] include an online deployment of the (trained) model. The shift from the offline to the online model deployment is important not only to foster the integration of detection module as part of the behavioral model of the robot, but also to check the validity of the proposed approach. In the work of [41], the authors claim that online (i.e., run-time) evaluation is fundamental for properly rating any classifier performance in the engagement detection task. In their work, the performances of the offline validation stage and frame-by-frame evaluation testing were not indicative or representative of the online performances. This is a core point for developing robust engagement detectors, together with fine-tuning the inference models.

4.5 Experimental Sessions

Most of the reviewed works highlight the conduction of the experimental session as the main limitation. This is mostly due to the low number of recruited participants (as reported in Table 6), which may not be a good representation of the overall population. The largest number of participants (≥ 60) is present when the interaction involves a team (i.e., in [47]) and when the robot is tested for a month-length time in public spaces, like in museums [42] or shops [43]. Excluding those cases and the works in which the number of participants is not specified (i.e., 2), 42.30% of the reviewed works recruited less than 10 participants. Aside from the small number of participants, most of the works highlight as an additional limitation the fact that the involved participants belong to a certain category of users, which also restricts the generalization of the proposed procedure. One of the proposals is to improve the detection capabilities of the model by leveraging data from users of different ages and different levels of cognitive abilities [40] and from multiple cultures [39].

According to some authors, additional limitations of data gathering rely on the fact that the experimentation scenario is performed in a single-interaction episode and a controlled setting. Regarding the first point, some authors propose to plan several sessions with the same participant, to detect engagement over a longer period, like in [37, 44, 45]. This approach could also provide elements for investigating the dependency between short-term and long-term engagement natures. Regarding the controlled settings, which constrained the participant to sit on a chair facing the robot, recent works started moving towards a free scenario, especially in the child-robot interaction, allowing the user to freely move in the environment and interact with the robot (see [30, 37]). The free scenario allows the detection of more naturalistic behaviors and expressions. One main concern about the experimental setting is the sensors' position, which are frequently installed in the environment or on the tablet, based on the best perspective for recording the interaction. As reported by [36, 38], a more naturalistic and ego-centric view could be obtained by posing a camera on the robotic platform. With this strategy, the same experimental setup could be used to evaluate multiple application scenarios, which can extend the trial dimensions, as suggested by [32]. In this direction, future works should also validate the reliability of the collected dataset, as highlighted in [69].

5 Conclusion

This review study aimed to investigate the concept of engagement and its connection with the automated frameworks currently developed for assessing it. Considering the research questions reported in Section 1, this analysis reported that, despite the engagement definition is not always pointed out, there are two main interpretations of the concept, sometimes interconnected. The first interpretation intends engagement as a process with a clear begin and an end, orchestrated by the connectedness between interactors [50], conversely the second definition depicts engagement as a quality metrics (i.e. being together and continuing the interaction). Both interpretations refer to a connection between two agents, e.g. user(s) and robot, in an interactive scenario, without specifying the context of interaction (i.e. target users, robot task, duration of the interaction). As result, works addressing different interaction contexts adopt the same definition, suggesting that the vagueness related to the concept could be attributed to the usage of the term, more than to the term itself. Since the modality of interaction could be influenced by the user profile and context of interaction, the concept of

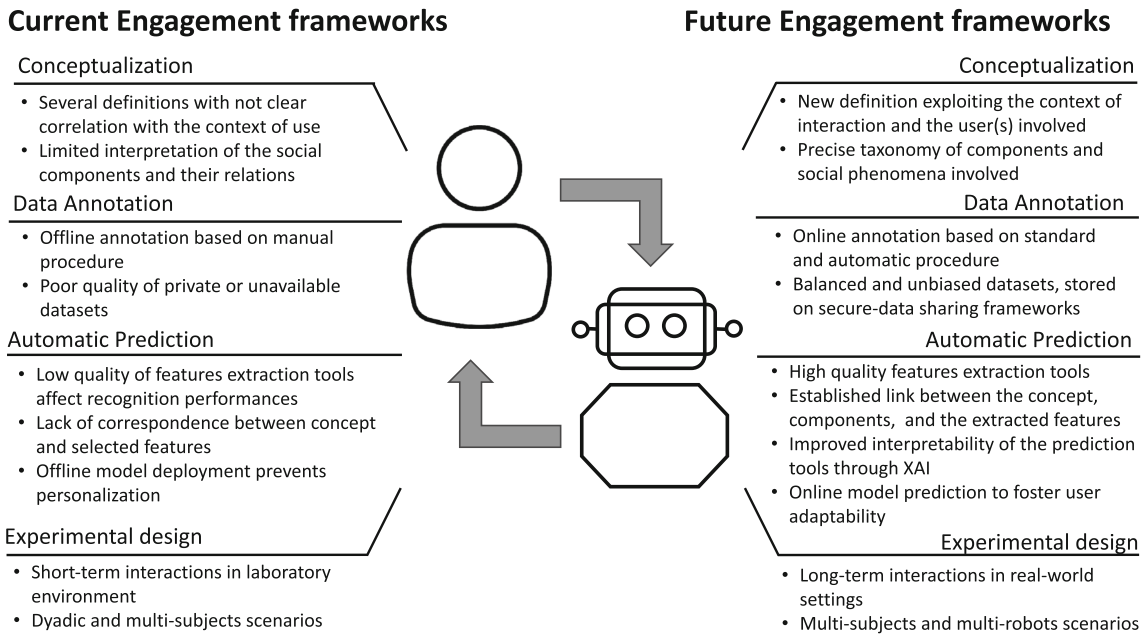
Current Engagement frameworks
Conceptualization
• Several definitions with not clear correlation with the context of use
• Limited interpretation of the social components and their relations
Data Annotation
• Offline annotation based on manual procedure
• Poor quality of private or unavailable datasets
Automatic Prediction
• Low quality of features extraction tools affect recognition performances
• Lack of correspondence between concept and selected features
• Offline model deployment prevents personalization
Experimental design
• Short-term interactions in laboratory environment
• Dyadic and multi-subjects scenarios
Future Engagement frameworks
Conceptualization
• New definition exploiting the context of interaction and the user(s) involved
• Precise taxonomy of components and social phenomena involved
Data Annotation
• Online annotation based on standard and automatic procedure
• Balanced and unbiased datasets, stored on secure-data sharing frameworks
Automatic Prediction
• High quality features extraction tools
• Established link between the concept, components, and the extracted features
• Improved interpretability of the prediction tools through XAI
• Online model prediction to foster user adaptability
Experimental design
• Long-term interactions in real-world settings
• Multi-subjects and multi-robots scenarios

Fig. 3 Current state of the art and expected improvements in the engagement concept and assessment

engagement should consider these details to provide a more clear and precise interpretation. To overcome the vagueness barriers, some of the reviewed works propose alternative definitions (e.g., Social/Task engagement [55], and Productive engagement [47]) that remark the aspects of this concept that should be taken into account according to a specific context of interaction (i.e., involved users, robot task, duration of the interaction).

In our view, the engagement phenomenon in HRI reflects the user intent to establish and maintain a connection with the robotic agent for the duration of the task as well as with the task, which depends on the achievement of a personal or a shared goal (if any). It may not be limited on the user positive feelings and attention during the interactive task (i.e., affective and cognitive components), as long as there is a commitment in achieving the task with the other, sharing intentions and desire to improve the interaction (i.e., behavioral component). When the "other" is a robot, the presence of the robot should return an advantage to the user experience, so that the user is prone to engage again or to keep the interaction, driven by an intrinsic interest and a personal reward. Performances and experience's improvements, related to the presence of the robot, should be considered as additional components of the engagement phenomenon, since they could remark the user intent and interest to connect with the robot. This aspect is also strictly related with the context of use and the role/strategy associated to the robot. In educational settings, the role of the robot is to support the learning performances of users as well as to motivate and "optimize" the learning outcome. In the cognitive therapy, the robot is perceived as a tool for fostering and improving the treatment adherence in patients. In game scenarios, the antagonist role of the robot could be used to incentivize and encourage the users to win.

In a more general overview, our idea is that the intentions of the users reflect the intentions of the robot, and vice versa, in a continuous loop that stops (i.e. disengagement) when the robot is not perceived as useful to reach the goal, or when the goal is reached. In this direction, more exhaustive definitions of engagement related to the context of the interaction should be provided.

Considering the methodology for engagement estimation, there is a growing trend of adopting advanced deep-learning solutions for ad-hoc scenarios. At the current time, annotated datasets for general purposes are not available. Since it is missing a clear interpretation of the concept, also the features used for the automatic assessment are several and they have no relation with the definition as well as its components (most of the time). As in human–human interaction there is a continuous exchange of messages between the involved partners, the perception of user engagement must rely on a multimodal and not-invasive approach (see Fig. 3). The descriptors of the engagement should be selected based on the interaction and expected improvements of the user in the task, which may vary among the users. Thus, the engagement assessment should stress more on the continuous aspect of engagement, which could better cover any shade of the user's intention. An additional step is to integrate the information on the user engagement in the decision module of the robotic platform to customize and adapt the robot's behaviors accordingly.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12369-024-01146-w.

Author Contributions AS was responsible for paper structure and writing, synthesizing the information from the papers into text and tables. LF contributed to the introduction, methodology definition, and search strategies. FC was the scientific supervisor, and guarantor for the review and contributed to the methodology definition, paper writing, discussion, and conclusion. All authors were involved in defining the keywords for the literature analysis, in paper screening and selection. All authors read, provided feedback, and approved the final manuscript.

Funding Open access funding provided by Universitá degli Studi di Firenze within the CRUI-CARE Agreement. We acknowledge funding from Next Generation EU, in the context of the National Recovery and Resilience Plan, Investment PE8 - Project Age-It: “Ageing Well in an Ageing Society” [DM 1557 11.10.2022]. The views and opinions expressed are only those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Data availability Not applicable.

Code Availability Not applicable.

Declarations

Conflict of interest The authors declare they have no financial interests.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Materials availability Not applicable.

References

1. Anzalone SM, Boucenna S, Ivaldi S, Chetouani M (2015) Evaluating the engagement with social robots. Int J Soc Robot 7(4):465–478
2. Rossi S, Ferland F, Tapus A (2017) User profiling and behavioral adaptation for hri: A survey. Pattern Recogn Lett 99:3–12
3. Sorrentino A, Mancioppi G, Coviello L, Cavallo F, Fiorini L (2021) Feasibility study on the role of personality, emotion, and engagement in socially assistive robotics: a cognitive assessment scenario. In: Informatics, vol 8, pp 23, MDPI
4. Glas N, Pelachaud C (2015) Definitions of engagement in human-agent interaction. In: 2015 international conference on affective computing and intelligent interaction (ACII), pp 944–949, IEEE
5. Doherty K, Doherty G (2018) Engagement in hci: conception, theory and measurement. ACM Comput Surv (CSUR) 51(5):1–39
6. Oertel C, Castellano G, Chetouani M, Nasir J, Obaid M, Pelachaud C, Peters C (2020) Engagement in human-agent interaction: an overview. Front Robot AI 7:92
7. Perugia G, Van Berkel R, Díaz-Boladeras M, Català-Mallofré A, Rauterberg M, Barakova E (2018) Understanding engagement in dementia through behavior. the ethographic and laban-inspired coding system of engagement (elicse) and the evidence-based model of engagement-related behavior (emodeb). Front Psychol 9:690
8. Yu C, Aoki PM, Woodruff A (2004) Detecting user engagement in everyday conversations. arXiv preprint arXiv:cs/0410027
9. Tickle-Degnen L, Rosenthal R (1990) The nature of rapport and its nonverbal correlates. Psychol Inq 1(4):285–293
10. Bickmore T, Schulman D, Yin L (2010) Maintaining engagement in long-term interventions with relational agents. Appl Artif Intell 24(6):648–666
11. Salam H, Chetouani M (2015) Engagement detection based on mutli-party cues for human robot interaction. In: 2015 international conference on affective computing and intelligent interaction (ACII), pp 341–347, IEEE
12. O’Brien HL, Cairns P, Hall M (2018) A practical approach to measuring user engagement with the refined user engagement scale (ues) and new ues short form. Int J Hum Comput Stud 112:28–39
13. Cohen-Mansfield J, Dakheel-Ali M, Marx MS (2009) Engagement in persons with dementia: the concept and its measurement. Am J Geriatr Psychiat 17(4):299–307
14. Orsulic-Jeras S, Judge KS, Camp CJ (2000) Montessori-based activities for long-term care residents with advanced dementia: effects on engagement and affect. Gerontologist 40(1):107–111
15. Jones C, Sung B, Moyle W (2015) Assessing engagement in people with dementia: a new approach to assessment using video analysis. Arch Psychiatr Nurs 29(6):377–382
16. Troisi A (1999) Ethological research in clinical psychiatry: the study of nonverbal behavior during interviews. Neurosci Biobehav Rev 23(7):905–913
17. Nasir J, Bruno B, Dillenbourg P (2020) Is there’ one way’ of learning? A data-driven approach. In: Companion publication of the 2020 international conference on multimodal interaction, pp 388–391
18. Rossi A, Raiano M, Rossi S (2021) Affective, cognitive and behavioural engagement detection for human-robot interaction in a bartending scenario. In: 2021 30th IEEE international conference on robot & human interactive communication (RO-MAN), pp 208–213, IEEE
19. Dewan M, Murshed M, Lin F (2019) Engagement detection in online learning: a review. Smart Learn Environ 6(1):1–20
20. Jain S, Thiagarajan B, Shi Z, Clabaugh C, Matarić MJ (2020) Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders. Sci Robot 5(39):eaaz3791
21. Salam H, Celiktutan O, unes H, Chetouani M (2023) Automatic context-aware inference of engagement in hmi: a survey. IEEE Trans Affect Comput
22. Lytridis C, Bazinas C, Papakostas GA, Kaburlasos V (2020) On measuring engagement level during child-robot interaction in education. Robot Educ Curr Res Innov 10:3–13
23. Avelino J, Garcia-Marques L, Ventura R, Bernardino A (2021) Break the ice: a survey on socially aware engagement for human-robot first encounters. Int J Soc Robot 13(8):1851–1877
24. Castellano G, Pereira A, Leite I, Paiva A, McOwan PW (2009) Detecting user engagement with a robot companion using task and

social interaction-based features. In: Proceedings of the 2009 international conference on multimodal interfaces, pp 119–126

25. Sanghvi J, Castellano G, Leite I, Pereira A, McOwan PW, Paiva A (2011) Automatic analysis of affective postures and body motion to detect engagement with a game companion. In: Proceedings of the 6th international conference on human-robot interaction, pp 305–312
26. Castellano G, Leite I, Pereira A, Martinho C, Paiva A, Mcowan PW (2014) Context-sensitive affect recognition for a robotic game companion. ACM Trans Interact Intell Syst (TiiS) 4(2):1–25
27. Castellano G, Leite I, Paiva A (2017) Detecting perceived quality of interaction with a robot using contextual features. Auton Robot 41(5):1245–1261
28. Jang M, Park C, Yang H-S, Kim J-H, Cho Y-J, Lee D-W, Cho H-K, Kim Y-A, Chae K, Ahn B-K (2014) Building an automated engagement recognizer based on video analysis. In: Proceedings of the 2014 ACM/IEEE international conference on human–robot interaction, pp 182–183
29. Rich C, Ponsler B, Holroyd A, Sidner CL (2010) Recognizing engagement in human-robot interaction. In: 2010 5th ACM/IEEE international conference on human-robot interaction (HRI), pp 375–382, IEEE
30. Hadfield J, Chalvatzaki G, Koutras P, Khamassi M, Tzafestas CS, Maragos P (2019) 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 (IROS), pp 1251–1256, IEEE
31. Ritschel H, Baur T, André E (2017) Adapting a robot's linguistic style based on socially-aware reinforcement learning. In: 2017 26th IEEE international symposium on robot and human interactive communication (roman), pp 378–384, IEEE
32. Ayllon D, Chou T-S, King A, Shen Y (2021) Identification and engagement of passive subjects in multiparty conversations by a humanoid robot. In: Companion of the 2021 ACM/IEEE international conference on human–robot interaction, pp 535–539
33. Inoue K, Lala D, Takanashi K, Kawahara T (2018) Engagement recognition by a latent character model based on multimodal listener behaviors in spoken dialogue. APSIPA Trans Signal Inf Process, 7
34. Pattar SP, Coronado E, Ardila LR, Venture G (2019) Intention and engagement recognition for personalized human-robot interaction, an integrated and deep learning approach. In: 2019 IEEE 4th international conference on advanced robotics and mechatronics (ICARM), pp 93–98, IEEE
35. Poltorak N, Drimus A (2017) Human-robot interaction assessment using dynamic engagement profiles. In: 2017 IEEE-RAS 17th international conference on humanoid robotics (humanoids), pp 649–654, IEEE
36. Salam H, Celiktutan O, Hupont I, Gunes H, Chetouani M (2016) Fully automatic analysis of engagement and its relationship to personality in human-robot interactions. IEEE Access 5:705–721
37. Javed H, Lee W, Park CH (2020) Toward an automated measure of social engagement for children with autism spectrum disorder-a personalized computational modeling approach. Front Robot AI, pp 43
38. Rudovic O, Lee J, Dai M, Schuller B, Picard RW (2018) Personalized machine learning for robot perception of affect and engagement in autism therapy. Sci Robot 3(19):eaao6760
39. Rudovic O, Utsumi Y, Lee J, Hernandez J, Ferrer EC, Schuller B, Picard RW (2018) Culturenet: a deep learning approach for engagement intensity estimation from face images of children with autism. In: 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 339–346, IEEE
40. Feng Y, Jia Q, Chu M, Wei W (2017) Engagement evaluation for autism intervention by robots based on dynamic Bayesian network and expert elicitation. IEEE Access 5:19494–19504
41. Foster ME, Gaschler A, Giuliani M (2017) Automatically classifying user engagement for dynamic multi-party human-robot interaction. Int J Soc Robot 9(5):659–674
42. Del Duchetto F, Baxter P, Hanheide M (2020) Are you still with me? Continuous engagement assessment from a robot's point of view. Front Robot AI 7:116
43. Iwasaki M, Zhou J, Ikeda M, Onishi Y, Kawamura T, Nakanishi H (2019) Acting as if being aware of visitors' attention strengthens a robotic salesperson's social presence. In: Proceedings of the 7th international conference on human-agent interaction, pp 19–27
44. Kim Y, Butail S, Tscholl M, Liu L, Wang Y (2020) An exploratory approach to measuring collaborative engagement in child robot interaction. In: Proceedings of the tenth international conference on learning analytics & knowledge, pp 209–217
45. Rudovic O, Park HW, Busche J, Schuller B, Breazeal C, Picard RW (2019) Personalized estimation of engagement from videos using active learning with deep reinforcement learning. In: 2019 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 217–226, IEEE
46. Chithrra R, Vetha V, Salam H, Nasir J, Bruno B, Celiktutan O (2022) Personalized productive engagement recognition in robot-mediated collaborative learning. In: Proceedings of the 2022 international conference on multimodal interaction, pp 632–641
47. Nasir J, Bruno B, Chetouani M, Dillenbourg P (2022) What if social robots look for productive engagement? Int J Soc Robot 14(1):55–71
48. Engwall O, Cumbal R, Lopes J, Ljung M, Månsson L (2022) Identification of low-engaged learners in robot-led second language conversations with adults. ACM Trans Hum Robot Interact (THRI) 11(2):1–33
49. Salam H, Chetouani M (2015) A multi-level context-based modeling of engagement in human-robot interaction. In: 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG), vol 3, pp 1–6, IEEE
50. Sidner CL, Lee C, Kidd CD, Lesh N, Rich C (2005) Explorations in engagement for humans and robots. Artif Intell 166(1–2):140–164
51. Bohus D, Horvitz E (2009) Models for multiparty engagement in open-world dialog. In: Proceedings of the SIGDIAL 2009 conference, the 10th annual meeting of the special interest group on discourse and dialogue, pp 10
52. Poggi I (2007) Mind, hands, face and body: a goal and belief view of multimodal communication. Weidler
53. O'Brien HL, Toms EG (2008) What is user engagement? A conceptual framework for defining user engagement with technology. J Am Soc Inform Sci Technol 59(6):938–955
54. Lemaignan S, Garcia F, Jacq A, Dillenbourg P (2016) From real-time attention assessment to "with-me-ness" in human-robot interaction. In: 2016 11th ACM/IEEE international conference on human-robot interaction (HRI), pp 157–164, IEEE
55. Corrigan LJ, Peters C, Castellano G, Papadopoulos F, Jones A, Bhargava S, Janarthanam S, Hastie H, Deshmukh A, Aylett R (2013) Social-task engagement: striking a balance between the robot and the task. Embodied Commun Goals Intentions Workshop ICSR 13:1–7
56. Nakamura J, Csikszentmihalyi M (2014) The concept of flow. In: Flow and the foundations of positive psychology, pp 239–263, Springer
57. Skinner EA, Pitzer JR (2012) Developmental dynamics of student engagement, coping, and everyday resilience. In: Handbook of research on student engagement, pp 21–44, Springer
58. Brown L, Kerwin R, Howard AM (2013) Applying behavioral strategies for student engagement using a robotic educational agent. In: 2013 IEEE international conference on systems, man, and cybernetics, pp 4360–4365, IEEE

59. Sidner CL, Kidd CD, Lee C, Lesh N (2004) Where to look: a study of human-robot engagement. In: Proceedings of the 9th international conference on Intelligent user interfaces, pp 78–84
60. Mendelson MJ, Aboud FE (1999) Measuring friendship quality in late adolescents and young adults: Mcgill friendship questionnaires. Can J Behav Sci 31(2):130
61. Biocca F (1997) The cyborg's dilemma: Progressive embodiment in virtual environments. J Comput Med Commun 3(2):JCMC324
62. Foster ME, Gaschler A, Giuliani M, Isard A, Pateraki M, Petrick RP (2012) Two people walk into a bar: dynamic multi-party social interaction with a robot agent. In: Proceedings of the 14th ACM international conference on multimodal interaction, pp 3–10
63. Nasir J, Norman U, Bruno B, Chetouani M, Dillenbourg P (2020) PE-HRI: a multimodal dataset for the study of productive engagement in a robot mediated collaborative educational setting.'
64. Nasir J, Bruno B, Dillenbourg P (2021) PE-HRI-temporal: a multimodal temporal dataset in a robot mediated collaborative educational setting
65. Ben-Youssef A, Clavel C, Essid S, Bilac M, Chamoux M, Lim A (2017) Ue-hri: a new dataset for the study of user engagement in spontaneous human-robot interactions. In: Proceedings of the 19th ACM international conference on multimodal interaction, pp 464–472
66. Celiktutan O, Skordos E, Gunes H (2017) Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. IEEE Trans Affect Comput 10(4):484–497
67. Jayagopi DB, Sheiki S, Klotz D, Wienke J, Odobez J-M, Wrede S, Khalidov V, Nyugen L, Wrede B, Gatica-Perez D (2013) The vernissage corpus: a conversational human-robot-interaction dataset. In: 2013 8th ACM/IEEE international conference on human-robot interaction (HRI), pp 149–150, IEEE
68. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst, vol 28
69. Vigni F, Andriella A, Rossi S (2024) A rosbag tool to improve dataset reliability. In: Companion of the 2024 ACM/IEEE international conference on human-robot interaction, pp 1085–1089

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Alessandra Sorrentino is a PostDoc Research Fellow at the University of Florence (Florence). She received a master's degree in Artificial Intelligence and Robotics from the University of Rome "La Sapienza", Italy, in 2018, and a Ph.D. (cum laude) degree in BioRobotics from the BioRobotics Institute, Scuola Superiore Sant'Anna, in 2022. Her research focuses on developing socially intelligent robots, aiming at adapting their behaviors according to their users for establishing trustworthy interactions, and to end-users needs for providing personalized assistance. She is interested in socially assistive robotics, human-robot interaction, affective computing, and AI-based control. Over the years, she collaborated on national and international projects, such as ACCRA, CloudIA, Pharaon. Currently, she is Task leader with AGE-IT project.

Laura Fiorini is an assistant professor at the University of Florence, Department of Industrial Engineering, Florence, Italy. She received the M.Sc. Degree in Biomedical Engineering at the University of Pisa in 2012 (full marks, cum laude). She obtained a Ph.D. in Biorobotics (full marks, cum laude) at the BioRobotics Institute of Scuola Superiore Sant'Anna, in 2016. In 2015 she visited the Bristol Robotics Laboratory at the University of West England (Bristol, UK). From 2016 to 2020, she was a postdoc researcher at the BioRobotics Institute and she collaborated on different EU and national projects such as Robot-Era, ACCRA, CloudIA, and SI-ROBOTICS. Currently, she is the coordinator of the Italian pilot site of the Pharaon Project.

Filippo Cavallo is Associate Professor in Biomedical Robotics and Biomechatronics at the University of Florence, Department of Industrial Engineering, Florence, Italy. He received the Master Degree in Electronics Engineering, Curriculum Bioengineering, from the University of Pisa, Italy, and the Ph.D. degree in Bioengineering at the BioRobotics Institute of Scuola Superiore Sant'Anna, Pisa, Italy. From 2007 to 2013, he was postdoc researcher and, from 2013 to 2019, he was assistant professor and head and scientific responsible of the Assistive Robotics Lab at the BioRobotics Institute, Scuola Superiore Sant'Anna. Since 2020, he is an associate professor with the University of Florence in Biomedical Robotics and Bio-Mechatronics and an affiliate professor at the BioRobotics Institute, Scuola Superiore Sant'Anna. The objectives of his research activities are to promote and evaluate novel service robotics for active and healthy ageing, and to identify and validate disruptive healthcare paradigms for neurodegenerative and chronic diseases, focusing on prevention and support for physical and cognitive declines. The main scientific and technological challenges concern social robotics, human-robot interaction, wearable sensors, Internet of Things, and artificial intelligence for robot companion and healthcare applications. He has participated in various national and European projects and is the author of 180+ papers at conferences and ISI journals.