What if Social Robots Look for Productive Engagement?

Jauwairia Nasir, Barbara Bruno, Mohamed Chetouani, Pierre Dillenbourg

International Journal of Social Robotics
https://doi.org/10.1007/s12369-021-00766-w

What if Social Robots Look for Productive Engagement?

Automated Assessment of Goal-Centric Engagement in Learning Applications

Jauwairia Nasir[1] · Barbara Bruno[1,2] · Mohamed Chetouani[3] · Pierre Dillenbourg[1]

Accepted: 12 February 2021

Abstract
In educational HRI, it is generally believed that a robots behavior has a direct effect on the engagement of a user with the robot, the task at hand and also their partner in case of a collaborative activity. Increasing this engagement is then held responsible for increased learning and productivity. The state of the art usually investigates the relationship between the behaviors of the robot and the engagement state of the user while assuming a linear relationship between engagement and the end goal: learning. However, is it correct to assume that to maximise learning, one needs to maximise engagement? Furthermore, conventional supervised models of engagement require human annotators to get labels. This is not only laborious but also introduces further subjectivity in an already subjective construct of engagement. Can we have machine-learning models for engagement detection where annotations do not rely on human annotators? Looking deeper at the behavioral patterns and the learning outcomes and a performance metric in a multi-modal data set collected in an educational human–human–robot setup with 68 students, we observe a hidden link that we term as Productive Engagement. We theorize a robot incorporating this knowledge will (1) distinguish teams based on engagement that is conducive of learning; and (2) adopt behaviors that eventually lead the users to increased learning by means of being productively engaged. Furthermore, this seminal link paves way for machine-learning models in educational HRI with automatic labelling based on the data.

Keywords Engagement · Human–robot interaction · Learning · Educational robots · Multi-modal learning analytics

1 Introduction

Engagement is a concept widely investigated in human–robot interaction (HRI) and yet still elusive [52]. Commonly adopted definitions include the one of Sidner et al. [68], defining engagement as "the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake", or the one of Poggi et al. [58], defining engagement as "the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and continuing interaction". Castellano et al., investigating predictors and components of engagement, regard engagement as characterised by both an affect and an attention component [17]. Conversely, Salam et al., postulate that "engagement is not restricted to one or two mental or emotional states (enjoyment or attention). During the interaction, as the objective of the current sub-interaction differs, the different concepts

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 765955. Furthermore, this project is supported by he Swiss National Science Foundation through the National Centre of Competence in Research Robotics.

✉ Jauwairia Nasir
jauwairia.nasir@epfl.ch

Barbara Bruno
barbara.bruno@epfl.ch

Mohamed Chetouani
mohamed.chetouani@sorbonne-universite.fr

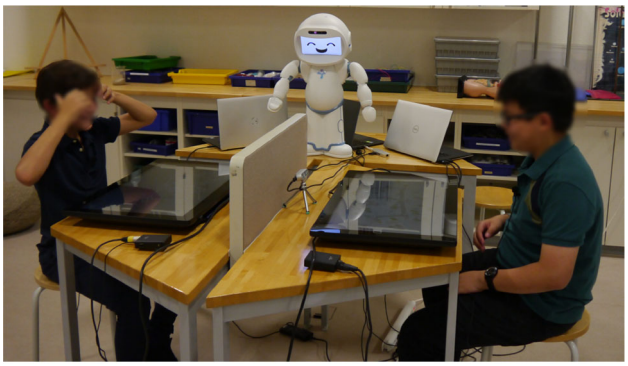
Pierre Dillenbourg
pierre.dillenbourg@epfl.ch

[1] Computer-Human Interaction in Learning and Instruction (CHILI) Lab, Swiss Federal Institute of Technology in Lausanne (EPFL), Lausanne, Switzerland

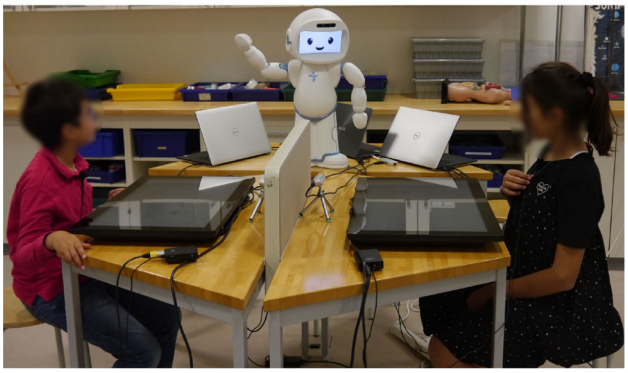
[2] MOBOTS group within the Biorobotics Laboratory (BioRob), Swiss Federal Institute of Technology in Lausanne (EPFL), Lausanne, Switzerland

[3] CNRS UMR 7222, Institute for Intelligent Systems and Robotics, Sorbonne University, Paris, France

Published online: 18 March 2021



(a)



(b)

Fig. 1 Children engaged in the JUSThink educational activity: which of these two teams, apparently similarly performing, will end up actually learning? Can we tell from their behavior? And if so, can we equip a robot with this knowledge, so that it will drive the robots behavior that is helpful for learning?

or cues related to engagement would differ" [63]. Similarly, O'Brien et al. define "user engagement as a multidimensional construct comprising the interaction between cognitive (e.g., attention), affective (e.g., emotion, interest), and behavioural (e.g., propensity to re-engage with a technology) characteristics of users, and system features (e.g., usability)" [48,49].

Studying HRI engagement in educational applications is particularly challenging (and therefore interesting) because of the fact that the robot and the interaction with it is a means to an end, which is learning. In [7], Baxter et al. show "that students who interacted with a robot that simultaneously demonstrated three types of personalization (nonverbal behavior, verbal behavior, and adaptive content progression) showed increased learning gains and sustained engagement when compared with students interacting with a non-personalized robot". Szafir et al. found that "adaptive robotic agent employing behavioral techniques (i.e. the use of verbal and non-verbal cues: increased spoken volume, gaze, head nodding, and gestures) to regain attention during drops in engagement (detected using EEG) improved student recall abilities 43% over the baseline" [69]. In [13], 24 students engage with the robot during a computer-based math test and the results demonstrate increased test performance with various forms of behavioral strategies while combining them with verbal cues result in a slightly better outcome. These studies show how in fact changing the robot's behavior has an impact on learning while making the linear assumption that increasing users engagement leads to increased learning. Hence, the standard approaches in the literature look to *maximize* engagement itself. But, is it enough to assume that *maximizing* engagement, as currently defined, *maximizes* learning?

Inspired by the behaviour and pedagogical principles of human teachers, we propose a paradigm shift for which at a given point in time, an *engaging robot for education* is the one capable of choosing an action that is in line with enhancing the educational goals directly. We postulate that to maximize learning, engagement need not be *maximized*, rather it needs to be *optimized*. This postulation draws some inspiration from the idea of *Productive Failure* proposed by Manu Kapoor [40] where he says "Engaging students in solving complex, ill-structured problems without the provision of support structures can be a productive exercise in failure". We believe that more often that not, there are learners that consecutively fail in a constructivist design, apparently scoring low on perceived engagement that can be biased by performance; however, they end up with higher learning. Same can be true with learners that seem to be succeeding but achieve lower learning. An example of this can be observed in [30] where the authors design a tangible tabletop environment for logistic apprentices for warehouse manipulation. They observe that while the task performance is high compared to the learners using the traditional method of paper and pencil, there is no increase in the learning outcomes. This is due to a phenomenon they termed as *Manipulation Temptation* where there is over-engagement with the task but no high-level reflection. Hence, interventions are incorporated to disengage learners to reflect more and eventually increase learning gains. Going back to the idea of *engaging robot for education*, as pointed out by Belpaeme et al. [9], designing one such robot is then not an easy feat. This is because even experienced human instructors struggle to make the best choice always. We believe that to not be able to distinguish *actual* engagement that potentially will lead to higher learning from *apparent* engagement that has no, or even a detrimental effect on learning plays a role in the struggle to find the appropriate choice.

If optimal engagement does exist, higher learning should then be reflected in certain behavioral patterns of the users. These patterns can then be leveraged to inform the behavior of the robot that is useful for learning. Briefly, this paper makes the following contributions:

– Validate the existence of "a *hidden hypothesis* that links multi-modal behaviors of the users to learning and

performance" that we term as *Productive Engagement* (See Fig. 2).
– The existence of the hidden hypothesis paves way to have machine-learning engagement models for which the labels do not come from human annotators but instead can emerge from the data itself.

Moreover, we define our human–human–robot setting where a learning task is present as a social-task engagement scenario as seen in Fig. 1. The definition by Corrigan et al. in [22] seems to be in line with the *social/task* distinction in the HRI engagement literature with regards to the nature of the HRI scenario/context. They define engagement in terms of three contexts as follows: "task engagement where there is a task and the participant starts to enjoy the task he is doing, social engagement which considers being engaged with another party of which there is no task included and social-task engagement which includes interaction with another (e.g., robot) where both cooperate with each other to perform some task". That said, still in a vast amount of literature, while defining the scenario, the distinction is often blurry since most interactions involve both task as well as social components, intertwined with each other and possibly co-dependent.

Lastly, the choice to have two users in our setting, introducing social engagement with a human, is because we want to grasp all facets of engagement, since we do not know yet which ones will better relate to learning. Social engagement with a human is supported by the idea that collaboration only produces learning if peers engage into rich verbal interactions such as argumentation, explanation, mutual regulation [12,27], or conflict resolution [33,66]. Furthermore, we want the interaction of the user to be as rich as possible and, therefore, the counterpart has to be another human. However, since engagement itself is still rather ambiguous, as explained at the start of the section, having two participants adds the variable of "group engagement", for which, too, multiple definitions exist. Salam et al. define group engagement as,"the joint engagement state of two participants interacting with each other and a humanoid robot" [62]. Oertel et al. define group engagement as "a group variable which is calculated as the average of the degree to which individual people in a group are engaged in spontaneous, non-task-directed conversations" [51] whereas Gatica et al. define group interest as "the perceived degree of interest or involvement of the majority of the group" in [32]. In our human–human–robot setup, we adapt the definition by [51] to our multi-modal data and where the engagement with a robot is dependent on the role of the robot (active, e.g. a team member; or passive, e.g. an instructor). Briefly, for the purpose of analyzing the hidden hypothesis highlighted in the contributions, we want to consider multiple facets of engagement as well as have two human users in the setting to have richer interactions.

In the remainder of the paper, Sect. 2 presents the related work while Productive Engagement (PE) is introduced in Sect. 3. The research questions are highlighted in Sect. 4 followed by the description of the learning activity, and the setup in Sect. 5. Section 6 includes an in-depth analysis, results and discussion. Lastly, concluding remarks follow in Sect. 7.

2 Related Work

The paradigm shift we propose puts us at the crossroad of two fields, social robotics and education. Therefore, this leads us to look at engagement literature from both perspectives of HRI and Multi-modal Learning Analytics (MLA).

It should be noted that in MLA, several studies target "motivation' and its link to learning. This is inspired by the positive relationship established in educational psychology between motivation and success at learning [24,71], For example, in this work by [59], they "demonstrate that motivation in young learners corresponds to observable behaviors when interacting with a robot tutoring system, which, in turn, impact learning outcomes". They observe a correlation between "academic motivation stemming from one's own values or goals as assessed by the Academic Self-Regulation Questionnaire (SRQ-A)" and observable suboptimal help-seeking behavior. The authors then go on to show that an interactive robot that responds intelligently to the observed behaviors positively impacts students learning outcomes. While motivation is not equivalent to engagement, it could rather be the cause of engagement, i.e., if one is motivated to learn intrinsically or extrinsically, one will engage more which is also in line with Maslow's theory of human motivation [44]. These MLA studies are thus sometimes also viewed relevant in the context of understanding engagement in educational settings.

In the literature coming from HRI and MLA, engagement is conventionally described as multi-faceted, meaning that various aspects of the user can be used to model it. Some of the forms found in literature, following the nomenclature proposed by [26], include *affective*, *behavioral*, *cognitive*, *academic*, and *psychological*, etc.

Various methods to *measure* engagement along these facets can then be found in the HRI and MLA literature. In [26], Dewan et al. categorize these methods (for online learning) into *manual*, *semi-automatic*, and *automatic*, and then divide the methods in each category into sub-categories depending upon the modality(ies) of the data used. Adapting the classification mainly from [26], we focus on the *manual* and *automatic* categories:

2.1 Manual

Two of the most popular manual methods found both in HRI and MLA engagement literature include: 1) *Self-Reporting*, where "the learners report their own levels of engagement, attention, distraction, motivation, excitement, etc." [50,70]; 2) *Observational Checklist*, where external observers complete questionnaires on learners engagement or annotate video or speech data [39,55]. While self-reporting is easy to administer and useful for "self-perception and other less observable engagement indicators" [70], there is also the issue of validity that depends on several factors such as learners honesty, willingness, and self-perception accuracy, etc. [29]. On the other hand, disadvantages of the second type of methods include the fact that they require a huge amount of time and effort by the observers, as well as the risk of observational metrics to be affected by confounding factors. For instance, as Whitehill et al. point out in [70], "sitting quietly, good behavior, and no tardy cards appear to measure compliance and willingness to adhere to rules and regulations rather than engagement". Furthermore, while studies with a single observer might suffer from subjectivity, studies with multiple observers might lead to low inter-rater agreement as engagement is a highly subjective construct.

2.2 Automatic

Some of the most widely used methods in MLA and HRI engagement modelling fall under this category. They can further be sub-divided into: 1) *Log-file Analysis*, and 2) *Sensor Data Analysis* methods. In *Log-file Analysis*, interaction traces are analyzed to extract users engagement or even performance (in educational settings) via behavioral indicators like the frequency of doing a particular behavior or the time taken on a particular action, etc. [1,16,20]. Various learning analytics and data mining approaches are used to perform *log-file analysis* in educational settings [4] including prediction methods, structure discovery, relationship mining, etc. While the interaction data is relatively easier to log and, hence, result in considerable amount of data; it lacks information that can be crucial to learning such as where the user is looking at or how the user feels. In the second method, a number of cues are investigated, most commonly through video and audio data: *gaze*, *mutual gaze*, *joint-attention*, *speech*, *posture*, *gestures*, *facial expressions*, *proxemics*, *personality* etc. [3,11,15,31,37,38,41,64,65]. A number of work complement video and audio data with physiological and neurological sensors to provide information such as: *EEG*, *heart rate*, *perspiration rate*, etc. [18,43]. The main advantage of relying on video and audio data only is that the setup can be made relatively unobtrusive and as close to the real settings in a classroom. On the other hand, while physiological and neurological sensors may provide more accurate information about some of the internal states of a learner (namely arousal, alertness, anxiety, etc.), they are specialized sensors that are not very practical in daily classroom settings.

Due to the multi-modality and diversity of the data collected, *Sensor Data Analysis* approaches can differ significantly in terms of the chosen analysis methods. Commonly found solutions include: 1) methods that look to detect the presence of specific engagement cues/events such as directed gaze, back-channels, valence, smile [34,60], 2) supervised classifiers where the labels come from human annotators [15,41,64], and 3) deep-learning [47] and deep reinforcement learning [53,61] approaches for engagement estimation. The deep-learning methods are relatively newer methods in HRI, motivated by the idea that the traditional machine learning methods are not equipped to deal with high-dimensional feature space, require expert engineering, and always rely on data annotation. While the first kind of methods are relatively straight-forward to implement, they are limited to the detectable cues, which are few and possibly affected by confounding factors. Even though supervised classifiers are one of the widely used methods, since engagement is a highly subjective construct, there is the problem of generalization and accuracy of such models since they are modeled in a specific context and the labels are provided by multiple annotators. We must also note that not many studies actually report the annotation protocol. Lastly, the latest deep learning approaches suffer from the lack of interpretability/explainability of results and require an abundance of data.

The state of the art review reported above emphasizes the benefits of multi-modal approaches, which are better suited to capture the nuances of engagement and less severely affected by confounding factors, as well as emphasizes the disadvantages of relying on human observers/annotators, which introduce a hard-to-control-for subjectivity. Hence, in the proposed work, we try to steer away from dependency on human annotators and lack of interpretability (introduced by deep learning approaches) while still making use of multi-modal data as in [57]. We put forward an automatic machine learning method, which relies on both log-files and video/audio data, analysed with clustering techniques. This method can then generate labels for engagement which can then be utilized for training a supervised classifier.

While engagement research in HRI is usually studied as the standalone goal of an experiment and, to the best of our knowledge, no study exists trying to explicitly link it to learning, a large amount of contributions within MLA (and specifically coming from the field of Intelligent Tutoring Systems - ITS) aims at capturing the knowledge state or skill level of the students through the interactions with the system [4,6,21,25,54] in addition to modelling meta-cognitive behaviors, affective states, engagement, and motivation [5,8,14,23,25]. We want to explore the relation between engagement and

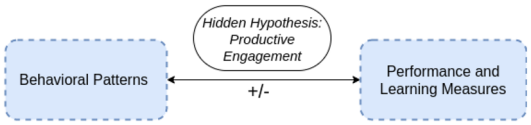


Fig. 2 Overview-productive engagement

learning. The reported MLA literature supports our hypothesis that it is possible to "unveil" learning and performance in the way learners engage with each other and the task at hand. The article investigates this intuition, without forgetting the ultimate goal of turning what we find into something that a robot can use online to drive its behavior to best support learning.

3 Productive Engagement

Our research is motivated by the following conceptions:

1. Maximizing engagement does not necessarily lead to increased learning outcomes, as first noted in Sect. 1, where by here engagement entails the apparent representation through logs, video and audio streams that are annotated by humans.
2. As first discussed in Sect. 2, evaluating engagement in light of domain specific measures like learning outcomes and performance metrics, that are more objective constructs, and relying upon multi-modal data, can be more effective in educational settings than using classifiers with labels from human observers.

We define Productive Engagement as the level of engagement that maximizes learning. Unproductive engagement can occur either due to over engagement (that can happen especially when interacting with gamified educational setups or setups with a robot) or under-engagement, both socially or with the task. We make a distinction between the *social* and *task* aspects of an interaction that happen in an educational setting, adapted from the work of [22]. *Productive Engagement* would then have the following components:

1. *Social Engagement* that we define as the quality and quantity of the verbal and non-verbal social interaction with other entities (learners and robots).
2. *Task Engagement* that we define as the quality and quantity of the interaction with the task.

As seen in Fig. 2, learning and performance can be positive or negatively affected by behavioral patterns pertaining to social and/or task engagement and vice versa. Furthermore, we argue that the other popularly used distinction (*cognitive* and *affective*), as seen in the review by [9], comes under the umbrella of both *task* and *social* engagement aspect of an interaction. To shed more light on the motivation to use this distinction, we include the outcomes classification from the aforementioned review by [9]. They showed that in most of the studies carried out with robots in educational settings, the outcomes (what the robot intervention targets and what the learning activity is designed for) can be classified into *cognitive* and *affective* outcomes [9]."Cognitive outcomes focus on one or more of the following competencies: knowledge, comprehension, application, analysis, synthesis, and evaluation" while the "Affective outcomes refer to qualities that are not learning outcomes per se, for example, the learner being attentive, receptive, responsive, reflective, or inquisitive". Both of these outcomes have been reported to affect learning; however, having a positive affective outcome does not imply positive cognitive outcome or vice versa [9,36]. The use of these two outcomes is also in line with the study of [28] who propose a model to explain the dynamics of affective states that emerge during deep learning that ultimately are also linked with cognitive engagement. Based on the definitions in the engagement literature [19,35,48,49,70], we define them as follows:

1. *Cognitive engagement* refers to the effort that is put into understanding and analyzing the learning concept including meta-cognitive behaviors like reflection.
2. *Affective engagement* encompasses feelings, enjoyment, attitude and the mood of the learners, etc.

The above categorization of engagement facets is presented to ground our definition of productive engagement in the context of existing engagement literature and to illustrate our rationale for selecting engagement-related features. Furthermore, we are aware that separating the cognitive and affective dimensions of interactions is a gross simplification. We nonetheless use this distinction as a convenient way to design the robot behavior as well as to analyse data. Concretely, we propose that a feature can be labelled based on the *type* of engagement (cognitive or affective in task or/and social space) we are using it to measure.

4 Research Questions

We consider our definition of Productive Engagement described above as a *hidden hypothesis* that "links multimodal behaviors of the users to learning and performance". Briefly, this paper investigates the following research questions:

– *RQ1:* Given the behavioral patterns, whether cognitive or affective, social or task, can we reveal a quantitative relationship that links them to learning and performance?

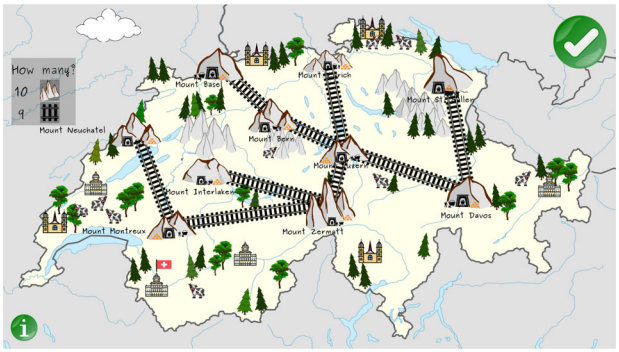
i.e., do people that differ in their behavior also differ in their learning and performance?
- *RQ2:* To feed a machine-learning model of engagement with labelled data, can we replace human annotated labels by measures extracted from learning outcomes?

The link between the stated contributions in the paper, Productive Engagement and the research questions is analogous to a cosco ladder. Previous work on educational HRI and MLA, as aforementioned, agree in suggesting that there is a link between learner engagement and learning. Then, the two fields differ: while the educational HRI side has mostly focused on investigating the relationship between the robot's behavior and learner's engagement, a subset of MLA literature has investigated the relation between learners behaviors (indicative of constructs like engagement, motivation, effortful behavior, that have been used comparably [67]) and learning. In this article, we postulate that it is time to reunite the two sides of the equation: robot behavior to user engagement to user learning. We propose to do so via the concept of Productive Engagement that emerges by investigating such domains in parallel. Productive Engagement is the type of engagement that the robot seeks to raise in the user, because "it is the one that is expected to put the user in conditions likely to trigger learning mechanisms, although there is no guarantee that the expected conditions would occur"[1]. Aforementioned is the first half of the ladder, the one where we climb from the literature to Productive Engagement. Now, on the second half, we descend from Productive Engagement to experiments and implementation. For the full link to work: (1) the robot needs to be able to autonomously infer the user's Productive Engagement in real time (RQ2), and (2) there must exist a link between said engagement and learning (RQ1), so that the robot can verify whether the current user engagement is conducive to learning and plan its actions accordingly.

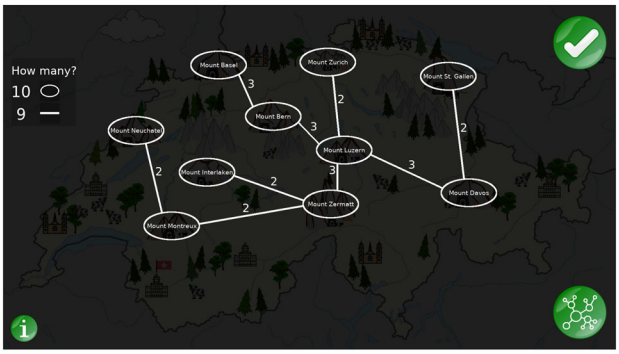
5 User Study

For the evaluation purpose of the hidden hypothesis, we make use of the data from a user study done with a first version of a robot-mediated human–human collaborative learning activity called JUSThink [46]. The JUSThink learning activity aims to (1) improve the computational skills of children by imparting intuitive knowledge about minimum-spanning-tree problems and (2) promote collaboration among the team via its design. As an experimental setup for HRI studies, it also serves as a platform for designing and evaluating robot behaviors that are effective for the pedagogical goals.

[1] This definition is inspired by Dillenbourg's way of defining collaborative learning in [27].



(a)



(b)

Fig. 3 The contents of the screens of the participants during the JUSThink learning activity, where one participant is in the figurative view as seen in (**a**) and the other participant is in the abstract view given by (**b**). The figures show a set of tracks forming a minimum spanning tree for the network of gold mines: finding it and building it collaboratively is the goal of the activity.

The minimum-spanning-tree problem is introduced through a gold mining scenario based on a map of Switzerland, where mountains represent gold mines labelled with Swiss cities names (see Fig. 3).

5.1 Learning Activity

The activity that envisions two children to play as a team consists of several stages spanning approximately 50 minutes. It starts with the robot welcoming the children, then introducing the goal of the task which is then followed by a pre-test. After the pre-test, the robot gives a demo explaining the two game views (see Fig. 3) and their functionalities, which is then followed by the learning task lasting around 25 minutes. After the task, children are asked to fill in a post-test and a self-assessment questionnaire before the robot greets them goodbye. Both the pre-test and post-test are defined in a context other than Swiss gold mines and are based on variants of the graphics in the *muddy city*[2] problem. Both tests are composed of 10 multiple-choice questions, assessing the

[2] https://csunplugged.org/minimal-spanning-trees/.

three concepts: (1) If a spanning tree exists, i.e. if the graph is connected., (2) If the given subgraph spans the graph, and (3) If the given subgraph that spans the graph has a minimum cost.

The learning task lies at the heart of the activity and requires the children to interact with maps such as those shown in Fig. 3 via touch-screens, as shown in Fig. 1. A small humanoid robot, acting as the CEO of a gold-mining company reiterates the problem by asking the participants to help it collect the gold by connecting the gold mines with railway tracks, while spending as little money as possible. The participants collaboratively construct a solution by drawing and erasing tracks that connect pairs of goldmines, and submit it to the robot for evaluation (one of the two optimal solutions is shown in Fig. 3).

The learning task design is scaffolded towards collaboration through precise design choices:

1. The task relies on two different views, respectively called *figurative* and *abstract*, where each gives only *partially observable information* to the user. The nodes and edges of the graph are shown by mountain and railway tracks in the figurative view while in the abstract view, they are denoted by circles and solid lines, respectively. Additionally, in the abstract view, deleted railway tracks are shown with dashed lines and the cost of each edge is indicated as a number.
2. The two views provide *complimentary functionality* and, therefore, in order to make informed decisions, the team members need to communicate. While in the figurative view, one can build and erase tracks, in the abstract view, one can view the cost of every track ever added, access previous solutions and their costs, and bring back a previous solution.
3. Every two edits, the views are swapped between participants, thus allowing each team member to experience the thought process that comes with a view. It also eliminates permanent roles in the game.
4. The cost of each track is initially hidden and only revealed after it is drawn, thus instigating reasoning about an edge in terms of a connection between two entities with an associated cost.
5. The team can submit their solution only if it spans the whole graph and only when both participants press the submit button. This scaffolds for team agreement before submission.

The robot's role in the current activity is two fold: 1) to mediate and automate the entire activity by giving instructions at every stage and moving the activity from one stage to the next as required, and 2) to intervene sparsely during the learning task to provide feedback on the progress, give hints and lend support through minimal verbal and non-verbal behaviors [46].

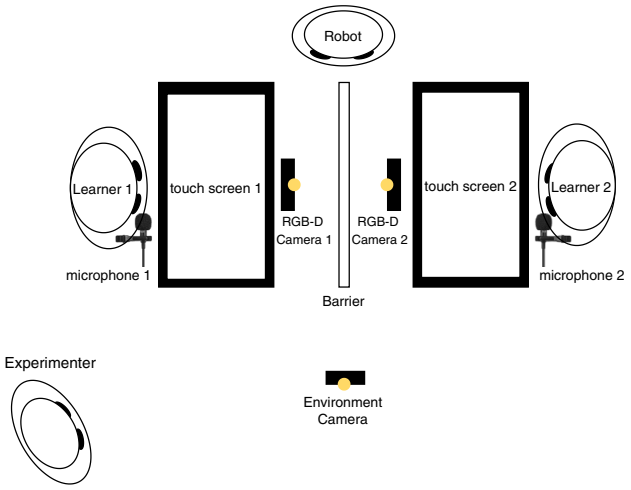


Fig. 4 The layout of the hardware setup for JUSThink

5.2 Setup and Participants

The setup for the experiment is shown in Fig. 1 where the two children in the team sit across each other separated by a barrier. Each of them has a touch screen in front, to interact with the application. The humanoid robot (QTrobot) is placed sideways with respect to the participants, to be visible to both. As depicted in Fig. 4, there are two RGB-D cameras that record the facial streams and one environment camera that films the entire scene. Two lavalier microphones, clipped on the participants, are used to record audio. We use two computers, connected to the screens and the robot, to manage the activity and the synchronous recording of the sensors. On the software side, each participant interacts with an instance of the JUSThink application while a separate robot application is used to manage the robot. All of the applications communicate via Robot Operating System (ROS). Rosbags are used to record all of the participants' actions (the logs) as well as the robot actions. For more details on the hardware and the software setup, see [46].

The study was conducted in two international schools in Switzerland over two weeks[3]. Although the experimenters were always present in the room, the activity was autonomous with little to no intervention required. A total of 96 students participated ranging from 9 to 12 years old; however, to ensure that data used for the study is complete and non-faulty across all sensing modalities (i.e., video, audio and actions logs) as well as homogeneous (e.g., we excluded a team in which participants were speaking French instead of

[3] This study received the approval of the university's ethics committee with reference number HREC No.: 051-2019.

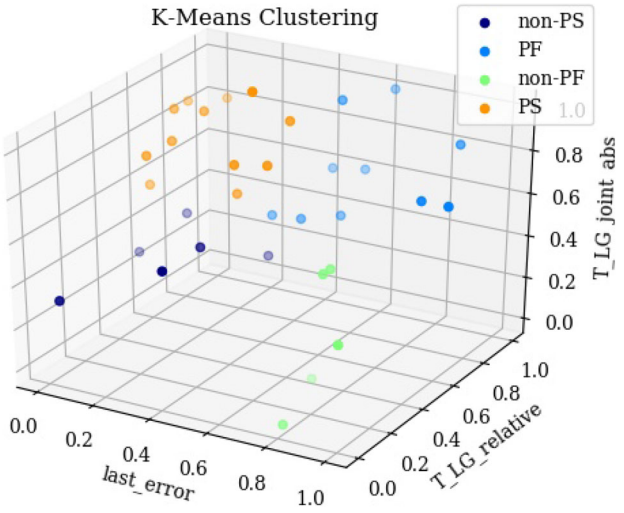


Fig. 5 Clustering of teams based on their learning and performance

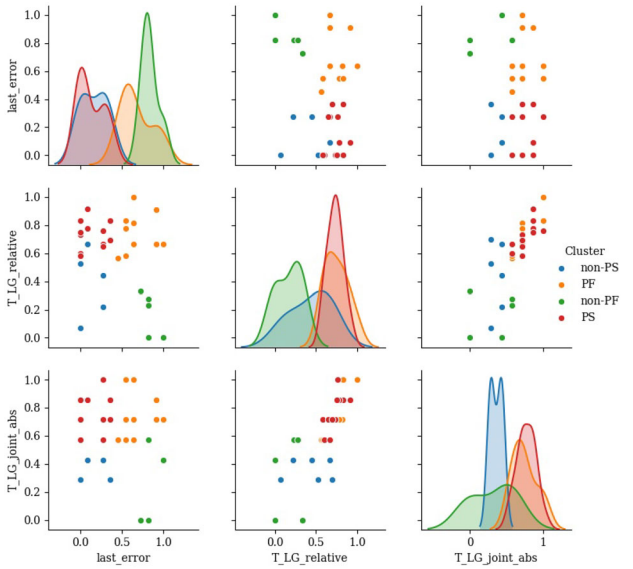


Fig. 6 Pair plots of the clusters obtained through the backward approach. According to their relative placement w.r.t. learning and performance (and in line with terms and concepts used in Education), we can label the clusters as: *non-Productive Success* (cluster non-PS), *Productive Failure* (cluster PF), *non-Productive Failure* (cluster non-PF) and *Productive Success* (cluster PS)

English to communicate with each other), we omitted 28 students, resulting in a corpus of 68 participants (i.e., 34 teams) used for the analysis reported in this article. The dataset that we termed as PE-HRI is made freely and publicly available [45].

6 Evaluating the Hidden Hypothesis

RQ2 assumes that learning and performance data, respectively extracted from the pre- and post-tests and the learning task itself, can provide labels to be used as a reference for the analysis of the engagement features. Concretely, this means that learning and performance data should allow for a separation of teams into different groups, with different learning outcomes and performance. This analysis, which we call "backward" since it allows for moving from learning to engagement (from learning outcomes back to the learning process), is reported in Sect. 6.1. In Sect. 6.2, we first discuss the engagement-related features extracted from video, audio and log data (see Table 1), then investigate the existence of the link between behavior and learning and performance, which we postulate, by verifying whether correspondences exist between the clustering of teams based on their behavior patterns and the learning labels. This is what we call the "forward" approach, since it moves from engagement features to learning outcomes and performance metric. We must point out that by performance, we mean how the teams perform, i.e., fail or succeed at the activity and by learning outcomes, we refer to how the learners score in there pre- and post-tests. For our analysis, we make use of the sklearn machine learning library [56].

6.1 Backward Analysis

We make use of the following learning outcomes and performance metric (which were first outlined in [46]), the definitions of which are outlined as:

- *Last error*: It is a performance metric, denoted by last_error, defined as the error of the last submitted solution by a team. It is computed as the difference between the total cost of the submitted solution and the cost of the optimal solution. Note that if a team has found an optimal solution (last_error = 0) the game stops, therefore making last error = 0.
- *Relative learning gain*: It is a learning outcome, calculated individually and not as a team, defined as the difference between a participant's post-test and pre-test score, divided by the difference between the maximum score that can be achieved and the pre-test score. This grasps how much the participant learned of the knowledge that he/she didn't possess before the activity. At team level, denoted by T_LG_relative, we take the average of the two individual relative learning gains of the team members.
- *Joint learning gain*: It is a learning outcome, denoted by T_LG_joint_abs, defined as the difference between the number of questions that both of the team members answer correctly in the post-test and in the pre-test, which

Table 1 Multi-modal features for the analysis of the participants' engagement in the Forward Approach

Feature	Definition	Feature type
Log features		
Edge Addition	The number of times a team added an edge on the map	Task/Cognitive
Edge Deletion	The number of times a team removed an edge from the map	Task/Cognitive
Ratio of Edge Addition and Deletion	The ratio of addition of edges over deletion of edges by a team	Task/Cognitive
Number of Actions	The total number of actions taken by a team (add, delete, submit, presses on the screen)	Task/Cognitive
History	The number of times a team opened the sub-window with history of their previous solutions	Task/Cognitive
Help	The number of times a team opened the instructions manual	Task/Cognitive
A_A_add	The number of times a team, either member, followed the pattern consecutively: I delete, I add back	Task/Cognitive
A_A_delete	The number of times a team, either member, followed the pattern consecutively: I add, I then delete	Task/Cognitive
A_B_add	The number of times a team, either member, followed the pattern consecutively: I delete, You add back	Task/Social/Cognitive
A_B_delete	The number of times a team, either member, followed the pattern consecutively: I add, You then delete	Task/Social/Cognitive
Redundant Edges	The number of times they had redundant edges in their map	Task/Cognitive
Video Features: Affective states and Gaze		
Positive Valence	The average value of positive valence for the team	Task/Social/Affective
Negative Valence	The average value of negative valence for the team	Task/Social/Affective
Positive Minus Negative Valence	The difference of the average value of positive and negative valence for the team	Task/Social/Affective
Arousal	The average value of arousal for the team	Task/Social/Affective
Smile	The average percentage of time of a team smiling	Task/Social/Affective
Gaze at Partner	The average percentage of time a team has a team member looking at their partner	Social/Cognitive
Gaze at Robot	The average percentage of time a team is looking at the robot	Social/Cognitive
Gaze (Other)	The average percentage of time a team is looking in the direction opposite to the robot	Social/Cognitive
Gaze at Screen_Left	The average percentage of time a team is looking at the left side of the screen	Task/Cognitive
Gaze at Screen_Right	The average percentage of time a team is looking at the right side of the screen	Task/Cognitive
Gaze Ratio of Screen_Right and Screen_Left	The ratio of looking at the right side of the screen over the left side	Task/Cognitive
Audio Features: Speech		
Speech Activity	The average percentage of time a team is speaking over the entire duration of the task	Social/Cognitive
Silence	The average percentage of time a team is silent over the entire duration of the task	Social/Cognitive
Small Pauses	The average percentage of time a team pauses briefly (0.15 sec)	Social/Cognitive
Long Pauses	The average percentage of time a team makes long pauses (1.5 sec)	Social/Cognitive
Speech Overlap	The average percentage of time the speech of the team members overlaps over the entire duration of the task	Social/Cognitive/Affective
Overlap to Speech Activity Ratio	The ratio of the speech overlap over the speech activity	Social/Cognitive/Affective

grasps the amount of knowledge acquired together by the team members during the activity.

We calculate these measures for each team, normalize them to have unit variance, and then perform a K-means clustering on the metrics as observed in Fig. 5. The k is estimated based on the commonly used metric of inertia for analyzing how well the clustering method did. For a better understanding of the resulting clusters, we also generate pair plots for the three metrics in Fig. 6. As the pair plots show, we have four clusters that we can label, in accordance with terminology and concepts commonly adopted in the field of learning and education (more specifically the terms *productive/non-productive* inspired by the terminology of *Productive Failure* [40]), as:

- *Non-Productive Success*, i.e. teams that performed well in the task but did not end up learning; hence, with lower last errors and lower learning gains (BA cluster = non-PS in blue in Fig. 6).
- *Productive Failure*, i.e. teams that did not perform well but did end up learning; hence, with higher last errors and higher learning gains (BA cluster = PF in orange).
- *Non-Productive Failure*, i.e. teams that neither performed well in the task nor did end up learning; hence, with higher last errors and lower learning gains (BA cluster = non-PS in green).
- *Productive Success*, i.e. teams that performed well and also ended up learning; hence, with lower last errors and higher learning gains (BA cluster = PS in red).

In terms of the pedagogical goal as well as the apparent success in the activity, it is quite interesting to see these four types of teams. However, the next question is whether behavioral patterns of teams would cluster in a similar manner or not. In other words, would the different behavioral patterns also indicate such a division among teams?

6.2 Forward Analysis

6.2.1 Joint Analysis of Video, Audio and Log Features

As explained in Sect. 2, in this work we focus on video, audio and log features as some of the most commonly used features for engagement detection, such as speech, affective states, and gaze come from such data. Table 1 lists and details the multi-modal features that we use to analyze participants' behavior in the forward approach. We also mark the feature type as task/social and cognitive/affective, in line with the definitions and rationale outlined in Sect. 3. As a first step, we make sure that the logs, videos, and audios used for generating all the features for a team are aligned and cut for the task duration only and not the entire pipeline given in Sect. 5.

Log features are extracted from the recorded rosbags. The features related to both affective states and gaze are computed through the open source library OpenFace [2]. A common way of calculating affective states, such as *valence* and *arousal*, is via the facial action units generated by OpenFace. For positive and negative valence, we build on action units (AUs) that correspond to positive and negative emotions, respectively, based on the findings from IMotions[4] that uses Affectiva[5] for emotion recognition. These findings are also similar to the ones in *EmotioNet* ([10]). Exponential moving average is applied to smoothen the data for each AU followed by taking an average of the AUs belonging to positive and negative emotions for positive and negative valence, respectively. We calculate arousal by taking average of all the AUs above a certain intensity at a given point in time. Regardless of the valence, the absolute value of arousal is calculated to measure the expressivity of a user. For the smile extraction based on AUs, we base it on the findings from a smile authenticity study conducted by [42]. OpenFace also generates gaze angles that can be used to determine the eye gaze direction in radians in world coordinates. These angles are averaged for both eyes and are converted into more easy to use format than the gaze vectors. Using these gaze angles, it can be approximated if a person is looking straight ahead, left or right. Lastly, voice activity detection (VAD) through audio stream is done by using the python wrapper for the opensource Google WebRTC Voice Activity Detection. All the audio features listed in Table 1 are computed on the output given by the Google WebRTC VAD.

Assessing Forward Clusters: To cluster teams based on their behavior pattern, as captured by the 28 features listed in Table 1, we first apply Principal Component Analysis (PCA) on the normalized features (we use min-max scaler to transform features by scaling each feature between a range of 0 and 1) which return three principal components (PCs). The three principal components identified by the PCA account for 50% of the variance within the features dataset, with the fourth component only contributing for 8%. Then, by applying K-means clustering on the three PCs (with K=4 chosen in accordance with the inertia score), we end up with four clusters as shown in Fig. 7 where each cluster represents a different behavioral pattern.

As outlined in the opening of this section, to investigate RQ1, we compute the average performance metric and learning outcomes for the teams in the clusters obtained from the analysis of behavioral features as shown in Fig. 8. In the rest of the analysis, we disregard cluster F^2_{all} since it is composed of only 2 data points. As the figure shows, while the three clusters F^0_{all}, F^1_{all} and F^3_{all} have similar average

[4] https://imotions.com/blog/facial-action-coding-system/.

[5] https://www.affectiva.com/.

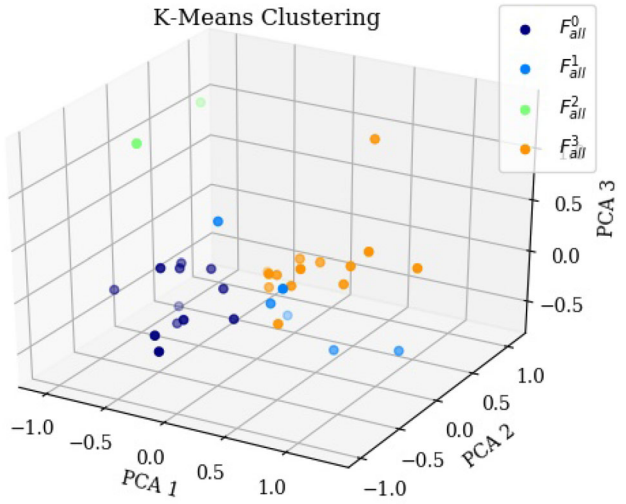


Fig. 7 Clustering of teams based on their behavioural pattern (extracted from video, audio and log features)

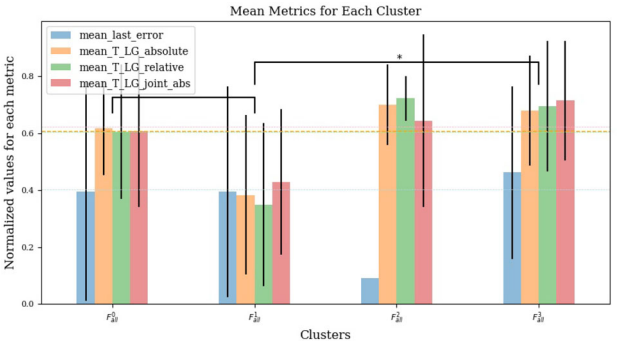


Fig. 8 Learning outcomes and performance metric (averaged within cluster) for the clusters computed with the forward approach. Stars denote statistically significant differences ($p < 0.05$). Dashed horizontal lines indicate the metrics' global averages

performance, they significantly differ in terms of learning outcomes, with clusters F^0_{all} and F^3_{all} having higher averages than cluster F^1_{all} (i.e., F^0_{all} and F^3_{all} including teams that ended up with higher learning, while cluster F^1_{all} includes teams who ended up with low learning). To validate these differences statistically, we perform a Kruskal-Wallis (KW) test on these metrics between each pair. In addition to the learning outcomes first defined in Sect. 6.1, we also include "absolute learning gain" to further validate the results. It is calculated individually and is defined as the difference between a participant's post-test and pre-test score, divided by the maximum score that can be achieved (10), which grasps how much the participant learned of all the knowledge available. At team level, denoted by T_LG_absolute, we take the average of the two individual absolute learning gains of the team members. Coming back to the KW test, for the pair (F^1_{all}, F^3_{all}), there is a significant difference for absolute learning gain, relative learning gain, and joint

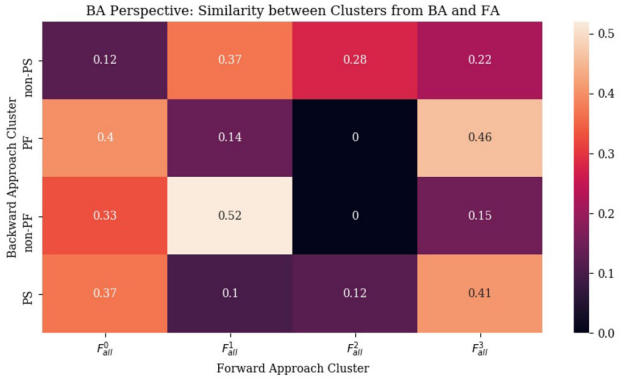


Fig. 9 Similarity matrix between the clusters computed on the learning outcomes and performance metric (backward analysis - rows) and those computed on the engagement features listed in Table 1 (forward analysis-columns)

learning gain respectively as (mean_LG_abs: $p = 0.025$, mean_LG_rel: $p = 0.016$, mean_LG_joint: $p = 0.026$). For the pair (F^0_{all}, F^1_{all}), albeit not statistically significant (for $p < 0.05$), there is a difference in absolute learning gain, and relative learning gain, respectively, as (mean_LG_abs: $p = 0.073$, mean_LG_rel: $p = 0.067$). These results seem to indicate that the teams that end up having significantly higher learning gains behave differently w.r.t. the teams ending up with lower learning gains. In other words, this suggests that participants' behavior is indicative of the separation of teams in high- and low-learners. This, in turn, supports our hypothesis of the existence of a link between engagement and learning (RQ1) and its representability with features that do not require human annotation (RQ2).

Comparing Forward and Backward Clusters: In an effort to further assess our hypothesis, we compare the clusters formed by the backward approach with those obtained in the forward approach. For this, we compute a *similarity score* S^F_B for each backward cluster B with each forward cluster F as:

$$S^F_B = \frac{\textit{common teams in both clusters}}{\textit{total teams in both clusters}} \quad (1)$$

which generates the *Similarity Matrix* shown in Fig. 9. It must be noted here that in Fig. 9, the order of naming of clusters on each axis is unrelated, i.e., we don't expect learners in horizontal cluster non-PS to also be in vertical cluster F^0_{all}, or more specifically we do not expect the diagonal to be filled.

In order to interpret the matrix, let us look at Figs. 6 and 8, along with Fig. 9. Starting from the backward clusters, we can observe that the majority of the teams belonging to low-learning clusters (i.e., cluster non-PS - *non-Productive Success* and cluster non-PF - *non-Productive Failure* in Fig. 6) fall in the forward cluster F^1_{all} ($S^1_{non\text{-}PS} =$

0.37, $S^1_{non\text{-}PF} = 0.52$), which in fact is the one with lowest average learning gain values (see Fig. 8 and Fig. 9). Similarly, the majority of the teams belonging to the high-learning clusters (i.e., cluster PF - *Productive Failure* and cluster PS - *Productive Success* in Fig. 6) fall in the forward clusters F^0_{all} ($S^0_{PF} = 0.40$, $S^0_{PS} = 0.37$) and F^3_{all} ($S^3_{PF} = 0.46$, $S^3_{PS} = 0.41$) that have significantly higher learning gain values (refer to Figs. 8 and 9).

The aforementioned analyses show that there are similarities in the composition of clusters generated by evaluating the teams' learning and performance and those generated by considering their behavior, captured by features extracted from logs, video and audio data. Concretely, in both cases, teams with low learning are grouped together and separated from high-learning teams. This indicates that, irrespective of performance during the task, teams that end up with higher learning exhibit behavioral patterns that can be clearly distinguished from those of teams that do not end up learning. In accordance with the definition put forth in Sect. 3, we deem the teams displaying behavioural patterns conducive to learning as *Productively Engaged*, as opposed to those whose behaviour, albeit possibly appearing engaged and even leading to good performance in the task, is not conducive to learning (*non-Productive Engagement*). We conclude that the reported analysis supports our hypothesis of the existence of a link between behavioral patterns and learning. Moreover, it paves the way for the design of robot behaviours, via the definition of *Productive Engagement*, which aim at putting learners in the best conditions for learning, by optimizing their engagement to that end.

6.2.2 Type-Specific Forward Analysis

The forward analysis presented in the previous section relies on features extracted from action logs, video and audio data. In an effort to verify the robustness of our findings, as well as restrict the feature set, we decided to replicate the forward analysis by first considering only the features extracted from the logs and then only the features extracted from the video and audio data. This separation is based on the idea that log-features are task-specific and, as captured by Table 1, mostly cognitive, while the other two data sources provide mostly social features (both cognitive and affective). Hence, an additional motivation for the analysis is therefore to check whether features of one type contribute more than the other to explaining the results seen in Sect. 6.2.1.

Performing PCA and K-means clustering on the log features (first section of Table 1), returns 3 clusters along 2 significant PCs (accounting for 55% of the variance within the features dataset, with the fourth component only contributing for 10%) as shown in Fig. 10. The similarity matrix given in Fig. 12 between the backward (on learning outcomes and performance metric) and forward (on

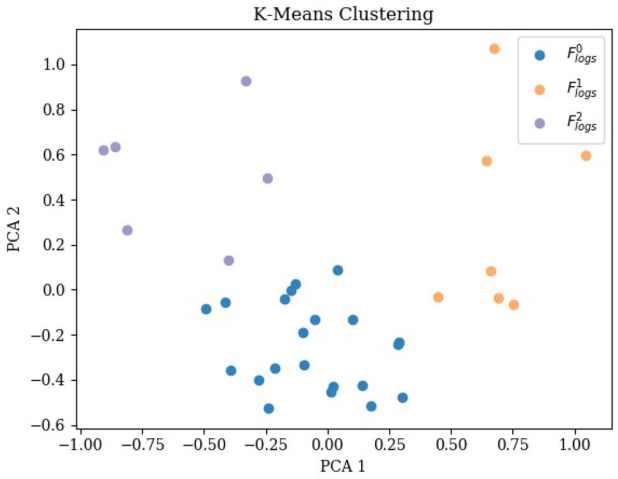


Fig. 10 Clustering of teams based on their behavioural pattern (extracted from log features only)

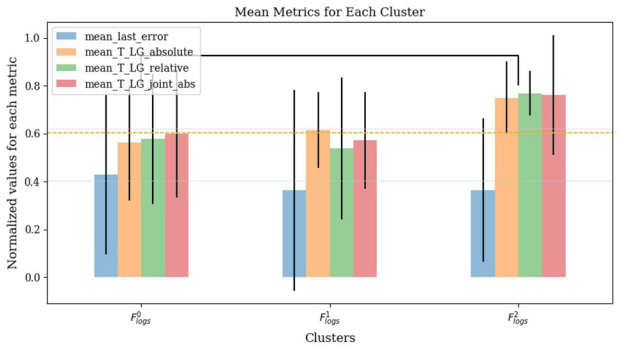


Fig. 11 Learning outcomes and performance metric (averaged within cluster) for the clusters computed with the forward approach using log features only. Dashed horizontal lines indicate the metrics' global averages. No statistically significant difference is found

behavioral features) clusters shows similar results w.r.t. those obtained when considering all features. The low-learning backward clusters (i.e., cluster non-PS - *non-Productive Success* and cluster non-PF - *non-Productive Failure* in Fig. 6) fall more in the forward cluster F^1_{logs} ($S^1_{non\text{-}PS} = 0.44$, $S^1_{non\text{-}PF} = 0.59$) while the high-learning backward clusters (i.e., cluster PF - *Productive Failure* and cluster PS - *Productive Success* in Fig. 6) fall more in the other two forward clusters F^0_{logs} ($S^0_{PF} = 0.68$, $S^0_{PS} = 0.40$) and F^2_{logs} ($S^2_{PS} = 0.41$) (see Figs. 11 and 12). However, a Kruskal-Wallis test run pairwise for the forward clusters over the learning outcomes shown in Fig. 11 reports no statistically significant difference, with only near-significant results we get are for the pair (F^0_{logs}, F^2_{logs}) (mean_LG_abs: $p = 0.060$, mean_LG_rel: $p = 0.065$, mean_LG_joint: $p = 0.096$) .

Similarly, following the backward and forward approach when using only the video and audio features (see Figs. 13, 14, and 15), we see the same conclusion as previously seen. The low-learning backward clusters (i.e., cluster non-PS -

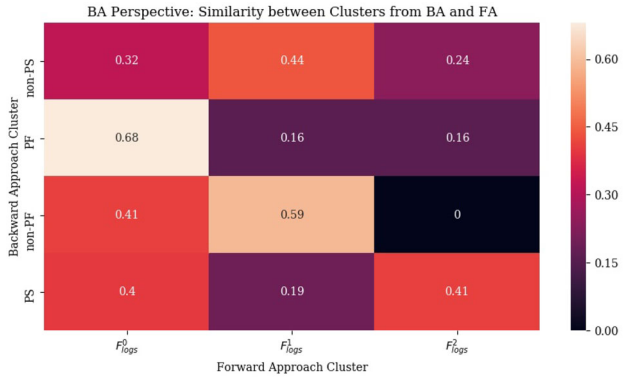


Fig. 12 Similarity Matrix between the clusters computed on the learning outcomes and performance metric (backward analysis - rows) and those computed on the log features listed in the top section of Table 1 (forward analysis - columns)

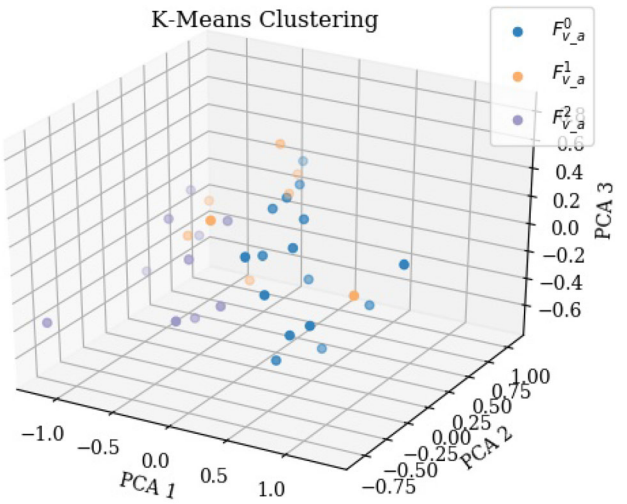


Fig. 13 Clustering of teams based on their behavioural pattern (extracted from video and audio features only)

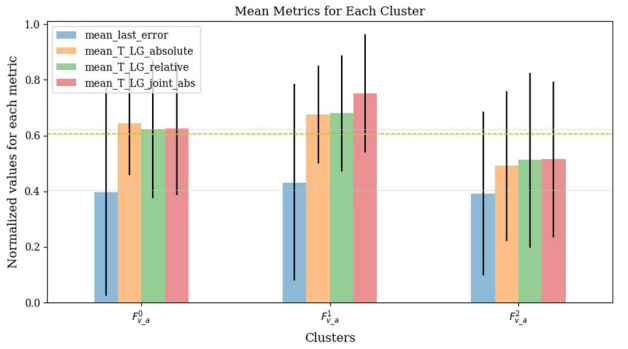


Fig. 14 Learning outcomes and performance metric (averaged within cluster) for the clusters computed with the forward approach using video and audio features only. Dashed horizontal lines indicate the metrics' global averages. No statistically significant difference is found

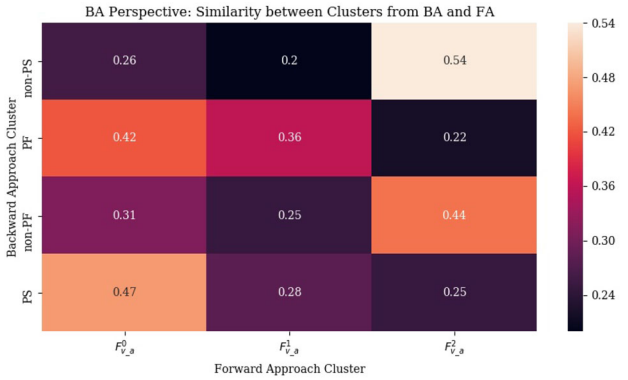


Fig. 15 Similarity Matrix between the clusters computed on the learning outcomes and performance metric (backward analysis - rows) and those computed on the video and audio features listed in the middle and bottom sections of Table 1 (forward analysis - columns)

non-Productive Success and cluster non-PF - *non-Productive Failure* in Fig. 6) fall more in the forward cluster $F^2_{v\_a}$ ($S^2_{non\text{-}PS} = 0.54$, $S^2_{non\text{-}PF} = 0.44$) which in fact is the one with lowest average learning gain values. On the other hand, the high-learning backward clusters (i.e., cluster PF - *Productive Failure* and cluster PS - *Productive Success* in Fig. 6) fall more in the other two forward clusters $F^0_{v\_a}$ ($S^0_{PF} = 0.42$, $S^0_{PS} = 0.47$) and $F^1_{v\_a}$ ($S^1_{PF} = 0.36$) (see Figs. 14 and 15) that have higher learning gain values. However, a Kruskal-Wallis test run pairwise for the forward clusters over the learning outcomes shown in Fig. 14 reports no statistically significant difference.

The results of the type-specific analyses suggest that (1) the results obtained in the global analysis of Sect. 6.2.1 are robust (since type-specific analyses are in line with them, either isolating high-learners or low-learners), and (2) the results obtained in the global analysis are produced by the combined effect of all types of features (since type-specific analyses fail to produce statistically significant results). The latter conclusion is a nice, indirect proof of the multi-dimensional, multi-faceted nature of human engagement, which makes it such a challenging and fascinating research topic.

7 Conclusion and Future Work

As outlined in Sect. 3, our goal is to pave the way for a new way of designing social robots for learning. The behavior of these robots is driven by the effects it will ultimately have on the user's learning, via the effect it has on the user's engagement, inspired by the findings in the fields of Educational HRI and Multi-modal Learning Analytics about the existence of a link between engagement and learning. Fundamental pre-requisites for achieving that goal are that (1) it is possible to compute an approximation of user engagement which is devoid of human intervention, to allow for its automatic

online extraction (RQ2); (2) the operationalization of engagement obtained in step 1 preserves the link with user learning (RQ1). The results we have obtained, reported in Sect. 6, support both hypotheses. Briefly, this paper explores the link between engagement and learning and, thus, proposes the concept of *Productive Engagement*, its validation in an HRI data set, and considerations on its consequences.

Firstly, we conclude that there are behavioral features, pertaining to task or/and social engagement, that predict learning outcomes and that these features are sometimes disconnected from performance in the task. To elaborate on the statement, in light of the results in Sect. 3, we observe that the teams that end up achieving a higher learning gain (i.e., cluster PF - *Productive Failure* and cluster PS - *Productive Success* in Fig. 6) in the JUSThink activity may or may not apparently perform well in the task itself. However, irrespective of their performance, the way those teams interact with the task and express themselves through speech, facial expressions and gaze is distinct from the behavior of the teams who achieve lower learning gains (i.e., cluster non-PS - *non-Productive Success* and cluster non-PF - *non-Productive Failure* in Fig. 6). Hence, these patterns of observable behaviors validate the existence of the hidden hypothesis of *Productive Engagement*.

Secondly, we conclude that the existence of this hidden hypothesis paves way for the design of machine-learning engagement detection models where the labelling for the state of engagement would not need a human annotator but rather come from the data itself. Specifically, the link between the behavioral patterns and the learning outcomes and the performance metric, in the form of statistically significant differences found with KW and the similarity matrix shown in Sect. 6.2.1, allows us to label the teams in forward clusters F^0_{all} and F^3_{all} as *Productively Engaged* and the teams in FA cluster F^1_{all} as *Non-productively Engaged*. At the same time, the results show that the proposed procedure seems better in isolating high-learners than low-learners (see results in Sect. 6.2.1 based on similarity matrix). This finding seems to suggest that while the behavior of people closer to the pedagogical goal of understanding the concept tends to be more distinctive and identifiable, the behavior or people who are (and will end up) not learning is more varied and harder to characterize. Intuitively, this finding reminds of Thomas Edison's famous quote about the many ways in which something can go wrong, and the only (or few) ways in which it can go right.

With this said, while performance is usually a biasing factor for humans when annotating a subjective construct like engagement in such activities; a robot enabled with the aforementioned knowledge around *Productive Engagement* would thus not make its interventions based on whether a team is failing in the task or not, but rather by observing more sophisticated patterns of interaction of a team with the task and with the social environment including the partner and the robot itself.

Furthermore, the analysis presented in this paper considers features computed at global level, i.e., at the end of the interaction. The next logical step along the path that we aim to walk is to transform the features of interest into time-series and verify whether the correlation with learning that we found at a global level still holds in the progression. To further investigate in this direction, as a second step, we plan to design a supervised time-series model with labels adapted through the hidden hypothesis established in the baseline JUSThink scenario, i.e., where the robot's interventions are minimal in order to reduce the confounding effects. The idea is then, as a third step forward, to put the model to test in a real-time scenario where the robot will adapt its behaviors according to the concept of *productive engagement*. The model will, thus, help the robot to answer the question of *when to intervene* effectively. However, to determine *what behavior to induce in the user* while designing for effective robot interventions, the next logical step we envision for this research is the characterization of the forward clusters obtained in Sect. 6.2.1 in terms of the contributions of the single features, and emerging differences between high- and low-learners. The aim is to acquire a deeper understanding of the link between engagement and learning, and therefore reach a refined and more solid definition for *Productive Engagement*.

Funding Open Access funding provided by EPFL Lausanne. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 765955. Furthermore, this project is supported by he Swiss National Science Foundation through the National Centre of Competence in Research Robotics.

Availability of data and material Not applicable

Declarations

Conflict of interest We have no conflict of interest.

Code availability Not applicable

References

1. Alyuz N, Okur E, Oktay E, Genc U, Aslan S, Mete SE, Stanhill D, Arnrich B, Esme AA (2016) Towards an emotional engagement model: can affective states of a learner be automatically detected in a 1:1 learning scenario? CEUR Workshop Proc 1618(1):1–7
2. Amos B, Ludwiczuk B, Satyanarayanan M (2016) Openface: a general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118. CMU School of Computer Science
3. Anzalone SM, Boucenna S, Ivaldi S, Chetouani M (2015) Evaluating the engagement with social robots. Int J Social Robot 7(4):465–478. https://doi.org/10.1007/s12369-015-0298-7
4. Baker R, Siemens G (2012) Educational data mining and learning analytics. In: Sawyer RK (ed) CHLS. Cambridge University Press, Cambridge, pp 253–272. https://doi.org/10.1017/CBO9781139519526.016
5. Baker RS, Corbett AT, Koedinger KR, Wagner AZ (2004) Off-task behavior in the cognitive tutor classroom, pp 383–390. https://doi.org/10.1145/985692.985741
6. Baker RS, Corbett AT, Aleven V (2008) More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 5091 LNCS, pp 406–415. https://doi.org/10.1007/978-3-540-69132-7-44
7. Baxter P, Ashurst E, Read R, Kennedy J, Belpaeme T (2017) Robot education peers in a situated primary school study: personalisation promotes child learning. PLoS ONE. https://doi.org/10.1371/journal.pone.0178126
8. Beal CR, Qu L, Lee H (2004) Basics of feedback control-elements of feedback control | instrumentation and control engineering, pp 151–156
9. Belpaeme T, Kennedy J, Ramachandran A, Scassellati B, Tanaka F (2018) Social robots for education: a review. Sci Robot 3(21):5954. https://doi.org/10.1126/scirobotics.aat5954
10. Benitez-Quiroz CF, Srinivasan R, Martinez AM (2016) Emotionet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 5562–5570. https://doi.org/10.1109/CVPR.2016.600
11. Benkaouar W, Vaufreydaz D (2012) Multi-sensors engagement detection with a robot companion in a home environment multi-sensors engagement detection with a robot companion in a home environment. Workshop on assistance and service robotics in a human environment, pp 45–52
12. Blaye A (1988) Confrontation socio-cognitive et résolution de problèmes. PhD thesis, Centre de Recherche en Psychologie Cognitive, Université de Provence, 13261 Aix-en-Provence, France
13. Brown LV, Kerwin R, Howard AM (2013) Applying behavioral strategies for student engagement using a robotic educational agent. In: Proceedings—2013 IEEE international conference on systems, man, and cybernetics, SMC 2013, pp 4360–4365. https://doi.org/10.1109/SMC.2013.744
14. Conati C, Maclaren H (2009) Empirically building and evaluating a probabilistic model of user affect. User Model User Adap Inter 19:267–303
15. Castellano G, Pereira A, Leite I, Paiva A, Mcowan P (2009) Detecting user engagement with a robot companion using task and social interaction-based features, pp 119–126. https://doi.org/10.1145/1647314.1647336
16. Castellano G, Leite I, Pereira A, Martinho C, Paiva A, McOwan PW (2012) Detecting engagement in hri: an exploration of social and task-based context. In: Proceedings—2012 ASE/IEEE international conference on privacy, security, risk and trust and 2012 ASE/IEEE international conference on social computing, SocialCom/PASSAT 2012, pp 421–428. https://doi.org/10.1109/SocialCom-PASSAT.2012.51
17. 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 4(2):1–25. https://doi.org/10.1145/2622615
18. Chaouachi M, Chalfoun P, Jraidi I, Frasson C (2010) Affect and mental engagement: towards adaptability for intelligent systems. In: Proceedings of the twenty-third international Florida artificial intelligence research society conference (FLAIRS), pp 355–360
19. Chi MT, Wylie R (2014) The ICAP framework: linking cognitive engagement to active learning outcomes. Educ Psychol 49(4):219–243. https://doi.org/10.1080/00461520.2014.965823
20. Cocea M, Weibelzahl S (2009) Log file analysis for disengagement detection in e-Learning environments, vol 19. https://doi.org/10.1007/s11257-009-9065-5
21. Corbett AT, Anderson JR (1995) Knowledge tracing: modeling the acquisition of student knowledge
22. Corrigan LJ, Peters C, Castellano G (2013) Social-task engagement: striking a balance between the robot and the task. Embodied Commun Goals Intentions Work ICSR 13(13):1–7
23. Craig SD, Witherspoon A, D'Mello SK, Graesser A, McDaniel B (2007) Automatic detection of learner's affect from conversational cues. User Model User Adap Inter 18(1–2):45–80. https://doi.org/10.1007/s11257-007-9037-6
24. Deci E (2017) Intrinsic motivation and self-determination. https://doi.org/10.1016/B978-0-12-809324-5.05613-3
25. Desmarais MC, Baker RS (2012) A review of recent advances in learner and skill modeling in intelligent learning environments. User Model User Adap Inter 22(1–2):9–38. https://doi.org/10.1007/s11257-011-9106-8
26. Dewan MAA, Murshed M, Lin F (2019) Engagement detection in online learning: a review. Smart Learn Environ 6(1):1–20. https://doi.org/10.1186/s40561-018-0080-z
27. Dillenbourg P, Baker M, Blaye A, O'Malley C (1996) The evolution of research on collaborative learning. In: Spada H, Reimann P (eds) Learning in humans and machines: towards an interdisciplinary learning science. Elsevier, Oxford, pp 189–211
28. D'Mello S, Graesser A (2012) Dynamics of affective states during complex learning. Learn Instruct 22(2):145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001
29. D'Mello S, Lehman B, Pekrun R, Graesser A (2014) Confusion can be beneficial for learning. Learn Instruct 29:153–170. https://doi.org/10.1016/j.learninstruc.2012.05.003
30. Do-lenh S (2012) Supporting reflection and classroom orchestration with tangible tabletops 5313:241. https://doi.org/10.5075/epfl-thesis-5313
31. Foster ME, Gaschler A, Giuliani M (2017) Automatically classifying user engagement for dynamic multi-party human–robot interaction. Int J Social Robot 9(5):659–674. https://doi.org/10.1007/s12369-017-0414-y
32. Gatica-Perez D, McCowan L, Zhang D, Bengio S (2005) Detecting group interest-level in meetings. In: Proceedings (ICASSP'05). IEEE international conference on acoustics, speech, and signal processing, vol 1. IEEE, pp I–489
33. Glachan M, Light P (1982) Peer interaction and learning: can two wrongs make a right. In: Social cognition: studies of the development of understanding, vol 2 in developing body and mind. Harvester Press, pp 238–262
34. Gordon G, Spaulding S, Westlund JK, Lee JJ, Plummer L, Martinez M, Das M, Breazeal C (2016) Affective personalization of a social robot tutor for children's second language skills. In: Proceedings of the 30th conference on artificial intelligence (AAAI 2016), vol 2011, pp 3951–3957

35. Henrie CR, Halverson LR, Graham CR (2015) Measuring student engagement in technology-mediated learning: a review. Comput Educ 90:36–53. https://doi.org/10.1016/j.compedu.2015.09.005
36. Huang CM, Mutlu B (2014) Learning-based modeling of multimodal behaviors for humanlike robots, pp 57–64. https://doi.org/10.1145/2559636.2559668
37. Ishii R, Nakano YI (2010) An empirical study of eye-gaze behaviors. In: Proceedings of the 2010 workshop on eye gaze in intelligent human machine interaction—EGIHMI '10, pp 33–40. https://doi.org/10.1145/2002333.2002339
38. Ishii R, Shinohara Y, Nakano I, Nishida T (2011) Combining multiple types of eye-gaze information to predict user's conversational engagement. Hum Factors
39. Kapoor A, Picard RW (2006) Multimodal affect recognition in learning environments, p 677. https://doi.org/10.1145/1101149.1101300
40. Kapur M (2008) Productive failure. Cognit Instruct 26(3):379–424. https://doi.org/10.1080/07370000802212669
41. Kim J, Co H, Truong K, Evers V, Truong KP (2016) Automatic detection of children's engagement using non-verbal features and ordinal learning expressive agents for symbiotic education and learning (EASEL) view project squirrel (clearing clutter bit by bit) view project automatic detection of children's engagement using non-verbal features and ordinal learning. https://doi.org/10.21437/WOCCI.2016-5
42. Korb S, With S, Niedenthal P, Kaiser Wehrle S, Grandjean DM (2014) The perception and mimicry of facial movements predict judgments of smile authenticity. PLoS ONE 9(6):99194
43. Kulíc D, Croft E (2007) Affective state estimation for human–robot interaction. IEEE Trans Rob 23(5):991–1000. https://doi.org/10.1109/TRO.2007.904899
44. Maslow A (1943) A theory of human motivation 13:370–396
45. Nasir J, Norman U, Bruno B, Chetouani M, Dillenbourg P (2020a) PE-HRI: a multimodal dataset for the study of productive engagement in a robot mediated collaborative educational setting. https://doi.org/10.5281/zenodo.4288833
46. Nasir J, Norman U, Bruno B, Dillenbourg P (2020b) When positive perception of the robot has no effect on learning. In: 2020 29th IEEE international conference on robot and human interactive communication (RO-MAN), pp 313–320. https://doi.org/10.1109/RO-MAN47096.2020.9223343
47. Nezami OM, Hamey L, Richards D, Dras M (2018) Engagement recognition using deep learning and facial expression 2013
48. O'Brien H, Freund L, Kopak R (2016) Reading environments. In Proceedings of the 2016 ACM on conference on human information interaction and retrieval, pp 71–80. https://doi.org/10.1145/2854946.2854973
49. O'Brien HL, Toms E (2008) What is user engagement? A conceptual framework for defining user engagement with technology. JASIST 59:938–955
50. O'Brien HL, Toms E (2010) The development and evaluation of a survey to measure user engagement. JASIST 61:50–69
51. Oertel C, Scherer S, Campbell N (2011) On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation. In: Twelfth annual conference of the international speech communication association
52. 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. https://doi.org/10.3389/frobt.2020.00092
53. Oggi O, Rudovic, Park HW, Busche J, Schuller B, Breazeal C, Picard RW (2019) Personalized estimation of engagement from videos using active learning with deep reinforcement learning
54. Pardos ZA, Heffernan NT (2010) Modeling individualization in a Bayesian networks implementation of knowledge tracing. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 6075 LNCS, pp 255–266. https://doi.org/10.1007/978-3-642-13470-8_24
55. Parsons J, Leah T (2011) Student engagement: what do we know and what should we do? University of Alberta, Technical report
56. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
57. Perugia G, Boladeras M, Català BE, Rauterberg M (2020) Engagedem: a model of engagement of people with dementia. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2020.2980275
58. Poggi I (2007) Mind, hands, face and body: a goal and belief view of multimodal communication. No. v. 19 = v. 19 in Körper, Zeichen, Kultur ; Body, sign, culture, Weidler, Berlin, oCLC: ocn143609341
59. Ramachandran A, Huang CM, Scassellati B (2019) Toward effective robot–child tutoring: internal motivation, behavioral intervention and learning outcomes. ACM Trans Interact Intell Syst 9(1):1–23. https://doi.org/10.1145/3213768
60. Rich C, Ponsler B, Holroyd A, Sidner CL (2010) Recognizing engagement in human–robot interaction. In: 5th ACM/IEEE International conference on human–robot interaction (HRI), pp 375–382. https://doi.org/10.1109/HRI.2010.5453163
61. Rudovic O, Zhang M, Schuller B, Picard R (2019) Multi-modal active learning from human data: A deep reinforcement learning approach. In: 2019 International conference on multimodal interaction. ACM, New York, pp 6–15
62. Salam H, Chetouani M (2015) Engagement detection based on mutli-party cues for human robot interaction. In: International conference on affective computing and intelligent interaction, ACII 2015, pp 341–347. https://doi.org/10.1109/ACII.2015.7344593
63. Salam H, Chetouani M (2015) A multi-level context-based modeling of engagement in human–robot interaction. In: 11th IEEE international conference and workshops on automatic face and gesture recognition (FG), vol 03, pp 1–6. https://doi.org/10.1109/FG.2015.7284845
64. Salam H, Celiktutan O, Hupont I, Gunes H, Chetouani M (2017) Fully automatic analysis of engagement and its relationship to personality in human–robot interactions. IEEE Access 5:705–721
65. 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-HRI '11, p 305. https://doi.org/10.1145/1957656.1957781
66. Schwarz BB, Neuman Y, Biezuner S (2000) Two wrongs may make a right... if they argue together!. Cognit Instruct 18(4):461–494. https://doi.org/10.1207/S1532690XCI1804_2
67. Sharma K, Papamitsiou Z, Olsen J, Giannakos M (2020) Predicting learners' effortful behaviour in adaptive assessment using multimodal data. https://doi.org/10.1145/3375462.3375498
68. 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. https://doi.org/10.1016/j.artint.2005.03.005
69. Szafir D, Mutlu B (2012) Pay attention! designing adaptive agents that monitor and improve user engagement. In: Conference on human factors in computing systems (CHI). https://doi.org/10.1145/2207676.2207679
70. Whitehill J, Serpell Z, Lin YC, Foster A, Movellan JR (2014) The faces of engagement: automatic recognition of student engagement

from facial expressions. IEEE Trans Affect Comput 5(1):86–98. https://doi.org/10.1109/TAFFC.2014.2316163
71. Wolters CA, Yu SL, Pintrich PR (1996) The relation between goal orientation and students' motivational beliefs and self-regulated learning. Learn Individual Differ 8(3):211–238. https://doi.org/10.1016/S1041-6080(96)90015-1