EPFL

Thèse n° 9781

Introducing Productive Engagement for Social Robots Supporting Learning

Présentée le 31 octobre 2022

Faculté informatique et communications
Laboratoire d'ergonomie éducative
Programme doctoral en robotique, contrôle et systèmes intelligents

pour l'obtention du grade de Docteur ès Sciences

par

Jauwairia NASIR

Acceptée sur proposition du jury

Dr D. Gillet, président du jury
Prof. P. Dillenbourg, Dr B. Bruno, directeurs de thèse
Prof. S. Rossi , rapporteuse
Prof. E. André, rapporteuse
Prof. T. Käser, rapporteuse

École polytechnique fédérale de Lausanne

2022

If I cannot do great things,
I can do small things in a great way.
— Martin Luther King Jr.

To my dearest husband,
my beloved parents, and
the One and Only Allah.

Acknowledgements

Brilliant minds *accompanied* with humanity is a rare commodity in today's world but I was lucky to have found that in my supervisor **Pierre Dillenbourg**. He is the best supervisor and mentor one can ask for, who encourages, motivates, and constructively criticizes all in a way that it only lifts you up to strive to be better and better at what you do. He also has the unique ability to have a special connection with each and every student of his. I am eternally grateful for all the opportunities he gave me to explore various research directions, question what is taken for granted in educational HRI research, think out of the box, and in the process allowing me to grow as a researcher and as a person. It would only be fair to say that I aspire to create a similar nurturing and fun environment as a leader wherever life takes me and I look forward to continuing our relationship.

Getting one great PhD supervisor is considered lucky but what if you get two? I would like to express my deepest gratitude towards **Barbara Bruno**, my co-supervisor, another amazing supervisor that I had the chance to be mentored by during my PhD. She makes sure that not only she is always there for us especially in moments of doubts but also she leads from the front when it comes to 'cracking up' the lab. I am thankful to you for all your brilliant feedback, for your amazing mentorship and discussions inside and outside CHILI, for using me as your reminder app, and for all the jokes and your attempts to tease me, and much more. I wholeheartedly credit you to play a major role in making my PhD an amazing experience for me. You deserve the best. See you in the land of beers, and Lederhosen.

Bureaucracy takes up half of our lives if not more. I am eternally grateful to **Florence Colomb**, the *real* boss of our lab, for taking care of a lot of this mess for us and for the amazing scavenger hunt she prepared for us to discover *Eburodunum*.

I would like to thank my esteemed jury committee: **Denis Gillet**, **Tanja Käser**, **Elisabeth André**, and **Silvia Rossi** for taking the time to evaluate my work and give their valuable feedback.

When I started my PhD, the superhuman **Wafa Johal**, was the first postdoc to help me navigate through the foggy first year to its successful completion. Apart from your feedback and inputs,

Acknowledgements

I have always admired your ability to be so productive and to be able to multitask successfully with a 1000 things on your plate. Next, I am very grateful to one of my most amazing collaborators, and *not* my postdoc, **Aditi Kothiyal**. I had so much fun working with you and successfully dealing with *Reviewer 1s*; learning so much from you about research, learning sciences; and sharing common interests such as dislike for word templates over tex, desi food, culture and bringing Pakistan and India on the same page.

This research work would not have been possible without the support of wonderful International schools in Switzerland that enabled all of my studies during my PhD. Special thanks to **Kearon Mcnicol**, **Paul Magnuson**, **Jessica O'Neill Casas**, **Adrian Hirst**, **Felicia de Lucia**, **Jon Snell**, **Danielle Allard** and **Marie-France Labelle** for all their effort in organizing the studies.

I would also like to extend my thanks to my collaborators: **Mohamed Chetouani** for his valuable feedback, discussions, for always being available for his students, as well as being an awesome coordinator of the ANIMATAS project; **Catharine Oertal** for giving me the opportunity to work with her on the review paper and in the process learning a lot; **Hanan Salam**, **Vetha Vikashini**, and **Oya Celiktutan** for our all girl collaboration.

I would like to thank the European Union's Horizon 2020 research and innovation programme that funded this research under the grant agreement No 765955. A special thanks to **Lebet Corinne**, the coordinator of the Doctoral Program of Robotics, Control, and Intelligent Systems (EDRS), for helping to navigate through the PhD process.

I am thankful to all my masters and bachelors semester project students as well as research scholars: **Mortadha Abderrahim**, a brilliant student of mine who broke the record for missing back-up trains during experiments; **Pierre Oppliger**, from whom I learned a lot (about our research ideas and degrees of murders :P); **William Ouensanga**, **Haoyu Sheng** who was not only great at her work, and hardworking, but also very sweet, **Malek El Mekki**, **Leandro Graziano**, **Anna Donnet** and **Laura Mathex**.

I am also very grateful for my Bachelor's final year project supervisor **Yasar Ayaz** and Master thesis supervisor **Jong-Hwan Kim** for giving me the wonderful opportunities, environment, and resources to explore various robot research ideas with them leading to some great outcomes and growing my passion further.

I cannot be thankful enough for being a part of this amazing CHILI team that became more like a family and gave me the best working environment ever. In alphabetical order of last name, I would like to thank everyone for supporting me, each in their own way, in this journey: the multi-talented **Thibault Asselborn**, aka Tyibo, for his support and for being a kind *french* friend to me throughout, my best cracking partner, the inventor of *the Jauwairia dance*, and who always has a war to fight back at me ;), **Sruti Bhattacharjee** for her child-like fun spirit, **Laurent Boatto** for his great humor and considerate nature, **Victor Borja** for being a very kind

and supportive friend to me and for always bringing good positive spirit when he decides to eat with us :D, **Lucas Burget** for the several nice talks, laughter and joy he brings, and for being an awesome first-ever male happiness manager (of course he learned from the best ;P), **Zhenyu Cai** for his ever smiling personality and warmth, **Melike Cezayirlioglu**, **Jules Cortois**, **Richard Davis** for always being a very wise, positive and considerate person to be around, **Louis Faucon** for his positivity and most intriguing discussions laden with bayesian probabilities and inferences, **Jie Gao** for all her love, her 'urgent' talks in the kitchen :D, for her believing that I am a calm person always which is so not true, and for her deep questions on life, **Thomas Gargot** for how he cracked up the lab during his stay here, **Kevin Gonyop** for being such a nice and wise person to have around, **Arzu Guneyzu** for officially welcoming me into CHILI, for being a dearest and kindest friend to me, for giving her beautiful heart to everyone around her, and for being the best neighbor even though she took half of my desk space :P, **Stian Haklev**, **Soheil Kianzad**, **Corinne Lebourgeoi** for being an amazing person and friend to me, **Franziska Margrith**, **Utku Norman** for not only being the kindest person to all, running to open every door possible for everyone, but for being my biggest work companion, collaborator, and support from day one in the lab - we literally started on the same day - and this PhD journey would not have been the same without you :), **Jennifer Olsen** for her valuable feedback and positive energy, **Ayberk Ozgur** for being a friend imparting knowledge and wisdom around him, and for giving Cellulos to the lab :D, **Anthony Peguet** for his kind personality, **Dorsa Safaei** for her smile and positive spirit, **Sina Shahmoradi** for being a friend in this journey from the very beginning, for the many ANIMATAS adventures together, for his friendly nature, and for inspiring me with his spirit to embark many new challenges and adventures, **Xu Tianyang** for her sweetest personality, **Daniel Tozadore**, aka Dani, for very quickly becoming an amazing and caring friend and a big source of support and advice for me, for his incredible efforts and ability to tease me, for being a good listener, for sharing many discussions, laughter, and Tiramisus with me (and sorry but not sorry for spoiling Harry Potter for you ;)), **Sven Viquerat** for being a good smiling friend always and for bringing cheerfulness and joy around me and everyone, **Killian Viquerat** for his undying energy, **Chenyang Wang** for his friendliness, and for being a very nice neighbor to sit next to in the last year of my PhD, **Yi-Shiun Wu** for being a genuine helping friend who taught us all a lot about the art of BBQ-ing the right way, and how to find cheap business class tickets, **Su Xiaotian** for her friendly personality, **Ramtin Yazadanian** for his many many jokes, **Teresa Yeo** for her kindness always.

I would also like to thank other colleagues and friends at EPFL: **Jade Cock** for her smiling, friendly and very caring nature :), **Laila El-Hamamsy** - there from my first PhD year with whom I can have never ending conversations :D, **Alexandra Niculescu** for her support for me but love for QTrobot :D, and **Melissa Skewers** for her utmost help throughout to help navigate with schools and her incredible energy.

Being a part of an EU Horizon 2020 Marie Skłodowska-Curie Innovative Training Network, ANIMATAS, gave me a chance to experience some amazing memories all over Europe with

incredible fellow PhDs in the program from all over the world. I would like to thank my amazing friends **Rebecca Stower**, **Sooraj Krishna**, **Natalia Calvo**, **Karen Tatarian**, **Sahba Zojaji**, **Ramona Merhej**, **Silvia Tulli** for giving me wonderful moments to always cherish; **Sera Buyukgoz** for being the sweetest friend and the biggest supporter and cheerleader one can ask for, **Tanvi Dinkar** for being a great friend, for the many many long voice notes and conversations about life and for all the support, **Manuel Bied** for being a wonderful lab mate when I was at Paris and for the friendship that followed after, **Maha El Garf** for being a big support and a friend even while being hundreds of miles apart, **Sebastian Wallkötter** for our very interesting discussions in Paris to begin with and all after that, for being the ESR representative so wonderfully, and for helping me write the German abstract of my thesis better than the English version.

I would like to thank my entire Women in AI Switzerland team, with whom I had several adventures in Switzerland to help promote awareness about fairer and unbiased AI, including many brilliant women. Special thanks to **Marisa Tschopp** for her amazing leadership, passion and for always showing support, **Kristina Jonkuviene** for her wonderful energy and friendship along the way, and **Gisela Andrade** for her warmth.

Friends like family are one of the biggest blessings and I was lucky to have a multicultural family during my stay in Switzerland. I would like to especially thank: **Hala Khodr**, aka Halalala, for giving me the best view in the lab, for her companionship not just inside the lab during my PhD journey but outside the lab too in the journey of life, for being a true sister to me, for literally coming to my rescue, for twinning with me, for giving me the honor to be her roommate on Skiminaires and conferences and still her maintaining friendship with me :P, and for having spent so many other amazing moments together; **Mayssa Bouaouina** for becoming one of my closest friends and a sister over the years with whom I feel like I could talk about everything and anything, for sharing the toughest time of my PhD by being next to me always, for *understanding* the *Jauwairia-ness* in me :D, for her positive and smiling personality, and for sharing Italian adventures with me among many other amazing things. I am also extremely grateful to **Zeinab Shmeis**, **Alaa Rushdy**, **Caroline Savio**, **Zahraa Ghanem**, **Kaouthar Najim**, **Firdaous Najim**, and **Ghewa Alsabeh** for being a great source of comfort for me in my Swiss life and for sharing many unforgettable moments and trips. **Arooj Akbar** - I still do not know how we missed knowing each other for almost two years at EPFL but after we finally met, it was an instant connection - thank you for all our amazing conversations, and your support; **Saqib Javed** for being a true Pakistani friend who came a little too late to EPFL :P.

I would like to remember all the wonderful people, who became lifelong friends, that I met along the path my academic career took me in various countries (whom I miss now for being far away geographically): starting from the most amazing set of sisters I found during my bachelors **Aleena Hassan**, **Nida Khalil**, **Samreen Siddique**, **Aisha Zia**; my best *junior* friend **Saad Butt**; my childhood friends **Rameesha Khalid** for always checking up on me even being so far apart physically, **Sumaira Shamim**, and **Hajira Shahid**; and then the several group of friends like family in the Republic of Korea and Germany including **Reem Abdelhamid**, **Asma**

Achek, **Mubarak Buser**, **Chahrazad Esalim**, **Sajid Iqbal**, **Abdul Saboor**; a friend like sister with whom I have travelled the most number of countries from Korea to Europe **Sadaf Gulshad**; **Gehan Fatima**, **Fatima Malik**; **Fatima Rizwan**, **Anam Hammad**, **Muhammad Rizwan**, **Faizan Naeem**, **Hammad Raza**; **Hamza Zulfiqar**, and **Farwa Ishtiaq**.

I have found a second family in my in-laws who have given me so much love and support during this journey: my father-in-law **Munir Naz**, who left us a little too soon, from whom I have experienced utmost support always and wish he was still here; my mother-in-law **Naseem Akhtar**, an incredible woman who has treated me like a daughter always, from whom I am learning the art of being selflessly giving; my sister-in-law **Maryam Munir**, in whom I found another sister of my own, who is the best at giving surprises; my brother-in-laws Arsalan and Sheharyar; Arsala Aapi, and then the cutest Abeeha and Ali for their most adorable messages.

I would especially like to thank my family, starting from my relatives who cheered on me at every achievement and milestone. My siblings have been a great source of support in all the ups and downs of life. I am grateful to my brothers: **Murtaza Nasir** for inspiring me with his hard work to reach where he is now - you deserve every bit of happiness and success, **Jauwad Nasir** for his strength to face life - I know you to be one of the most talented people ever who deserves the best. I have deepest gratitude for having my sisters as greatest blessings in my life: **Memoona Nasir** for being a great role model for me in many aspects of life - your strength, and self-confidence has been the most inspiring and I can never forget how you have been as a big sister looking over me starting from the very early years till now, **Ayesha Nasir** for being a true companion in life - with every passing day, I am seeing you, my lil baby sister, grow into this amazing, creative and talented human being I am so proud of. However the circumstances are, you are always there with your ever so fun stickers and best gifs. Now we are all in 5 different countries, it means we do not get to see much of each other (I miss us all together a lot) but you all are always very close to my heart.

My parents are the most precious blessing for me in life, they are the reason of where I am today. I can never thank you enough for everything you both have done for your children, for the life you have given us in so many different countries of the world: my father **Nasir-ud-din Gohar** for teaching me to be a person of principles, values, and kindness above everything else while achieving great success and pursuing all your dreams, for showing me that you can be an honest person with integrity and still rise to the top; my mother **Yasmin Nasir** for teaching me how to navigate through the ups and downs of life with a beautiful giving heart that is so big that there is room for everyone in it, for giving us unconditional warmth and love in all seasons of life. You both are my inspiration in many ways. No words here will ever do justice to what you both mean to me. I am proud to be your daughter. :')

At the end, I would like to thank *my person*, a gem, coolest companion, biggest cheerleader and best friend, my handsome husband **Abdul Hannan** with whom I am lucky to share my life. Despite whatever the situation is, you have this incredible ability to make my life calmer,

peaceful and exciting all at the same time. You have always supported me in my ambitions so ardently, sometimes even more than myself and for that I am eternally grateful. This “long-distance PhD" of *ours* would not have been possible without all that you did and in the process, I am glad in addition to me getting my PhD, you also became a gold member of Deutsche Bahn with many perks :D. Thank you for your love, patience, understanding and most importantly *for existing*. You are one to keep :). I can’t wait for our next adventure together in life. :’)

Lausanne, 26 August 2022 Jauwairia Nasir

Abstract

We have all been one such student or seen such students who can maintain the 'good student' image while playing a video game under the table or those loyal backbenchers, seemingly always distracted, who then ace their exams. These intricacies of human behaviors are just a few examples of what makes it non-trivial and challenging even for expert teachers to know how students' visible behaviors relate with learning. As research investigates ways in which robots and AI can support teachers and students, it is faced with the same challenge of inferring students' engagement; thus, making the investigation of this topic increasingly popular in educational HRI. The state of the art usually explores the relationship between the robot behaviors and the engagement state of the learner while assuming a linear relationship between engagement and learning. However, is it correct to assume that to maximize learning, one needs to maximize engagement? Furthermore, conventional supervised engagement models require human annotators to get labels. This not only is laborious but can also introduce subjectivity. Can we have machine-learning engagement models where annotations do not rely on human annotators? Additionally, with the increase in open-ended learning activities which by design employ the 'learning by failing' paradigm, in-task performance can not be the best measure for learning. Can we instead rely on multi-modal behaviors?

In an effort to cater for these challenges, this thesis dives deep to identify and quantify the relationship between learning and engagement, which we term as *Productive Engagement* (PE). In order to develop, design, and evaluate our PE framework, (1) we first designed and developed an open-ended collaborative learning activity that served as a platform for evaluating different robot variants over time. With 98 children interacting with the baseline version from 2 international Swiss schools, we showed that in-task performance and learning are indeed not correlated. Thus, this showed the importance of not being limited to robot interventions that affect only superficial measures of students' learning. (2) Then, with learner's multi-modal behaviors, we showed that indeed there is a hidden link between learner's behaviors and learning that can be quantified, i.e., validating the proposed concept of *Productive Engagement*. (3) This quantifiable link surfaced three collaborative multi-modal learner profiles, by using a *forward and backward clustering and classification technique*, two of which are linked to higher learning. This technique gave a possibility to surface data driven labels for engagement; thus,

evading the process of human annotations. We then identified similarities and differences between these learner profiles both at an *aggregate* and at the *temporal* level. (4) Based on (3), we constructed a PE score that can either be directly used as an assessment metric by a social robot in real-time or as data driven labels for building more sophisticated regression models. (5) With the learner profiles and the PE score, we designed and evaluated more advanced robot variants for the final studies with ~160 students from 7 international Swiss schools. With the design of different robot variants that employ knowledge about the learner's skills conducive to learning, rather than domain knowledge, in order to provide interventions; we provided a complementary perspective on the role of social robots in educational settings.

Keywords: Engagement, Human-Robot Interaction, Social Robotics, Educational Robotics, Multi-modal Learning Analytics, Collaborative Learning, Time-series Analysis

Résumé

Nous avons tous été ou vu de tels étudiants qui peuvent maintenir l'image de "bon étudiant" tout en jouant à un jeu vidéo sous la table ou ces fidèles élèves d'arrière-ban, apparemment toujours distraits, qui réussissent ensuite leurs examens. Ces subtilités du comportement humain ne sont que quelques exemples de ce qui rend non trivial et difficile, même pour les enseignants experts, de savoir comment les comportements visibles des élèves sont liés à l'apprentissage. Alors que la recherche étudie les moyens par lesquels les robots et l'IA peuvent aider les enseignants et les élèves, elle est confrontée au même défi de déduire l'engagement des élèves, ce qui rend l'étude de ce sujet de plus en plus populaire dans le domaine des Interactions Human-Robot. L'état de l'art explore généralement la relation entre les comportements du robot et l'état d'engagement de l'apprenant en supposant une relation linéaire entre l'engagement et l'apprentissage. Cependant, est-il correct de supposer que pour maximiser l'apprentissage, il faut maximiser l'engagement ? En outre, les modèles d'engagement supervisés classiques nécessitent des annotateurs humains pour obtenir des labels. Cela est non seulement laborieux, mais peut également introduire de la subjectivité. Est-il possible d'avoir des modèles d'engagement par apprentissage automatique où les annotations ne dépendent pas des annotateurs humains ? De plus, avec l'augmentation des activités d'apprentissage ouvertes qui, de par leur conception, utilisent le paradigme "apprendre en échouant", la performance en cours de tâche ne peut pas être la meilleure mesure de l'apprentissage. Peut-on alors se baser sur les comportements multimodaux ?

Dans un effort pour répondre à ces défis, dans cette thèse, nous nous efforçons d'identifier et de quantifier la relation entre l'apprentissage et l'engagement, que nous appelons l'engagement productif (EP). Afin de développer, de concevoir et d'évaluer notre cadre d'engagement productif, (1) nous avons d'abord conçu et développé une activité d'apprentissage collaborative ouverte qui a servi de plateforme pour évaluer différentes variantes de robots au fil du temps. Avec 98 enfants interagissant avec la version de base provenant de 2 écoles internationales suisses, nous avons montré que la performance en tâche et l'apprentissage ne sont effectivement pas corrélés. Ainsi, cela a montré l'importance de ne pas se limiter à des interventions robotiques qui n'affectent que des mesures superficielles de l'apprentissage des élèves. (2) Ensuite, avec les comportements multimodaux de l'apprenant, nous avons montré qu'il existe en effet un lien caché entre les comportements de l'apprenant et l'apprentissage qui peut être quantifié, validant ainsi le concept proposé d'engagement productif. (3) Ce lien

quantifiable a fait apparaître trois profils d'apprenants multimodaux collaboratifs, en utilisant une technique de classification et de regroupement progressive et retrogressive, dont deux sont liés à un apprentissage supérieur. Cette technique a permis de faire apparaître des labels d'engagement basées sur des données, évitant ainsi le processus d'annotation humaine. Nous avons ensuite identifié les similitudes et les différences entre ces profils d'apprenants, tant au niveau agrégé qu'au niveau temporel. (4) Sur la base de (3), nous avons construit un score EP qui peut être soit directement utilisé comme une métrique d'évaluation par un robot social en temps réel, soit comme des labels orientés donnés pour construire des modèles de régression plus sophistiqués. (5) Avec les profils des apprenants et le score EP, nous avons conçu et évalué des variantes de robots plus avancées pour les études finales avec 160 étudiants de 7 écoles internationales suisses. Avec la conception de différentes variantes de robots qui utilisent des connaissances sur les compétences de l'apprenant propices à l'apprentissage, plutôt que des connaissances du domaine, afin de fournir des interventions ; nous avons fourni une perspective complémentaire sur le rôle des robots sociaux dans les environnements éducatifs.

Mots-clés : Engagement, interaction homme-robot, robotique sociale, robotique éducative, analyse de l'apprentissage multimodal, apprentissage collaboratif, analyse des séries temporelles.

Zusammenfassung

Wir alle waren schon einmal ein solcher Schüler oder haben einen solchen Schüler gesehen, der das Image des "guten Schülersäufrechterhalten konntewährend er unter dem Tisch ein Videospiel spielte oder jene loyalen Hinterbänkler, die scheinbar immer abgelenkt sind und dann ihre Prüfungen mit Bravour bestehen. Diese Beispiele des menschlichen Verhaltens sind nur einige Beispiele dafür, dass es selbst für erfahrene Lehrer nicht trivial ist zu wissen, wie das sichtbare Verhalten der Schüler mit dem Lernerfolg zusammenhängt. Die Forschung untersucht wie Roboter und KI Lehrer und Schüler unterstützen können und steht hier ebenfalls der Herausforderung gegenüber das Engagement der Schüler erkennen zu müssen. Momentan fokussieren sich Forscher in der Regel auf die Beziehung zwischen dem Verhalten des Roboters und dem Engagement des Lernenden, wobei gewöhnlich eine lineare Beziehung zwischen Engagement und Lernerfolg angenommen wird. Ist es jedoch richtig anzunehmen, dass eine Maximierung des Lernerfolges eine Maximierung des Engagements voraussetzt? Zudem erfordern herkömmliche Engagement-Modelle die auf überwachtem Lernen basieren menschliche Aufwand, um Labels zu erhalten. Dies ist nicht nur mühsam, sondern auch subjektiv. Gibt es Engagement-Modelle die maschinelles Lernen benutzen und bei denen die Labels nicht von menschlichen Annotatoren abhängen? Mit der Zunahme von Lernaktivitäten die ein offenes Ende haben und die das Paradigma Lernerfolg durch Scheitern"verwenden ist die Leistung wärend der Aktivität oft nicht das beste Maß des Lernerfolgs. Können wir stattdessen multimodale Verhaltensweisen verwenden?

Um diesen Herausforderungen zu meistern gehen wir in dieser Arbeit in die Tiefe um die Beziehung zwischen Lernerfolg und Engagement, welche wir als Productive Engagement (PE) bezeichnen, zu identifizieren und zu quantifizieren. Um unsere PE-Methode zu entwickeln, zu gestalten und zu evaluieren, (1) haben wir zunächst eine offene kollaborative Lernaktivität konzipiert und entwickelt, die als Plattform für die Evaluierung verschiedener Robotervarianten diente. Mit 98 Kindern aus zwei internationalen schweizer Schulen die mit der Basisversion interagierten konnten wir zeigen, dass die der Erfolg bei der Aufgabe tatsächlich nicht mit dem Lernerfolg zusammenhängt. Dies zeigte wie wichtig es ist sich nicht auf Roboterinterventionen zu beschränken die sich nur auf oberflächliche Messungen des Lernens von Schülern auswirken. (2) Anschließend haben wir anhand des multimodalen Verhaltens der Lernenden gezeigt, dass es tatsächlich einen versteckten Zusammenhang zwischen dem Verhalten der Lernenden und ihrem Lernerfolggibt und dass dieser quantifiziert

werden kann, d. h. wir konnten das vorgeschlagene Konzept des produktiven Engagements bestätigen. (3) Diese quantifizierbare Verbindung hat es uns ermöglicht mittels einer eine Vorwärts- und Rückwärts-Clustering- und Klassifizierungstechnik drei kollaborative multimodale Lernerprofile aufzudecken von denen zwei mit höherem Lernerfolg verbunden sind. Diese Technik ermöglichte es, datengesteuerte Labels für das Engagement zu erstellen und so die Verwendung von menschlichen Labels zu umgehen. Anschließend haben wir Ähnlichkeiten und Unterschiede zwischen diesen Lernerprofilen sowohl auf aggregiertem als auch auf zeitlichem Niveau ermittelt. (4) Auf der Grundlage von (3) haben wir einen PE-Score konstruiert, der entweder direkt als Bewertungsmaßstab für einen sozialen Roboter in Echtzeit oder als datengesteuerte Kennzeichnung für das Erstellen komplexerer Regressionsmodelle verwendet werden kann. (5) Mit den Lernerprofilen und dem PE-Score haben wir fortgeschrittenere Robotervarianten für die abschließenden Studien entworfen und mit 160 Schülern aus 7 internationalen schweizer Schulen evaluiert. Damit eröffnen wir durch die Entwicklung verschiedener Robotervarianten, die, anstatt Fachwissen, das Wissen über lernfördernde Fähigkeiten der Lernenden nutzen um Interventionen anzubieten, eine ergänzende Perspektive auf die Rolle sozialer Roboter in Bildungsumgebungen.

Schlüsselwörter: Engagement, Mensch-Roboter-Interaktion, soziale Robotik, Bildungsrobotik, Multimodale Lernanalyse, kollaboratives Lernerfolg, Zeitreihenanalyse

Astratto

Tutti noi siamo stati uno di questi studenti o abbiamo visto studenti che riescono a mantenere l'immagine di "bravo studente" mentre giocano a un videogioco sotto il tavolo o quegli irriducibili dell'ultimo banco, apparentemente sempre distratti, che poi superano brillantemente tutti gli esami. Questi sono solo alcuni esempi di ciò che rende non banale e impegnativo anche per gli insegnanti esperti sapere come i comportamenti visibili degli studenti siano in relazione con il loro apprendimento. Man mano che la ricerca indaga sui modi in cui i robot e l'IA possono supportare insegnanti e studenti, si trova ad affrontare la stessa sfida di dedurre l'apprendimento degli studenti dal loro comportamento, rendendo così l'indagine di questo argomento sempre più popolare nell'HRI educativa. Lo stato dell'arte sull'argomento esplora la relazione tra il comportamento del robot e il coinvolgimento dell'allievo nell'attività didattica, assumendo una relazione lineare tra coinvolgimento e apprendimento. Tuttavia, è corretto assumere che per massimizzare l'apprendimento sia necessario massimizzare il coinvolgimento? Inoltre, i modelli usualmente adottati per l'analisi del coinvolgimento richiedono l'intervento di esperti che annotino i dati. Questo non solo è laborioso, ma può anche introdurre soggettività. Possiamo avere modelli automatici per l'analisi del coinvolgimento e dell'apprendimento, in cui le annotazioni non richiedono l'intervento di esperti? Inoltre, con l'aumento delle attività di apprendimento esplorative, basate sul paradigma dell'"imparare fallendo", il successo nell'attività non può essere considerato come misura per l'apprendimento. Possiamo quindi affidarci, a questo scopo, all'analisi multimodale del comportamento?

Nel tentativo di rispondere a queste sfide, in questa tesi esploriamo come identificare e quantificare la relazione tra apprendimento e coinvolgimento, che definiamo Coinvolgimento Produttivo (PE dall'inglese "Productive Engagement"). Per sviluppare, progettare e valutare il nostro sistema per l'analisi del PE, (1) abbiamo prima progettato e sviluppato un'attività di apprendimento collaborativo ed esplorativo che è servita come piattaforma per valutare diverse varianti di comportamento del robot. Con 98 bambini di 2 scuole internazionali svizzere che hanno interagito con la versione di base, abbiamo dimostrato che il successo nell'attività e l'apprendimento non sono correlati. Ciò ha dimostrato l'importanza di non limitarsi a interventi che influenzano solo misure superficiali dell'apprendimento degli studenti. (2) In seguito, con l'analisi multimodale del comportamento degli studenti, abbiamo dimostrato che esiste un legame nascosto tra il comportamento degli studenti e l'apprendimento, che può essere quantificato, convalidando così il concetto proposto di Coinvolgimento Produttivo.

(3) Questo legame ha portato all'individuazione di tre profili di studenti, due dei quali collegati ad un migliore apprendimento, emersi utilizzando una tecnica di clustering e classificazione. Da ciò abbiamo potuto estrarre annotazioni per il coinvolgimento, evitando così il passaggio attraverso esperti. Abbiamo poi analizzato le somiglianze e le differenze tra questi profili, sia a livello aggregato che temporale. (4) Sulla base di (3), abbiamo costruito un punteggio PE che può essere utilizzato direttamente e in tempo reale come metrica di valutazione da un robot sociale o come annotazione per la costruzione di modelli di regressione più sofisticati. (5) Con i profili degli studenti e il punteggio PE, abbiamo progettato e valutato varianti di robot più avanzate per gli esperimenti finali con 160 studenti di 7 scuole internazionali svizzere. Con la progettazione di diverse varianti di robot equipaggiati con la conoscenza delle abilità che favoriscono l'apprendimento, piuttosto che la conoscenza dei concetti da apprendere, questa tesi fornisce una prospettiva complementare sul ruolo dei robot sociali in contesti educativi.

Parole chiave: Coinvolgimento, interazione uomo-robot, robotica sociale, robotica educativa, Analisi multimodale dell'apprendimento, Apprendimento collaborativo, Analisi delle serie temporali

Contents

Acknowledgements i

Abstract (English/Français/Deutsch/Italian) vii

List of figures xix

List of tables xxv

1 Engagement in HRI: Deceptively Simple, Endlessly Complicated 1
- 1.1 Concept of Engagement in Human Robot Interaction . . . 2
- 1.2 Challenges . . . 3
  - 1.2.1 Relationship between Engagement and Learning . . . 3
  - 1.2.2 Human Subjectivity when Modelling Engagement . . . 4
  - 1.2.3 In-Task Performance as a Measure of Learning . . . 4
  - 1.2.4 Real-time Constraints . . . 5
- 1.3 Research Goals . . . 5
- 1.4 Organization of the Thesis . . . 6

2 Designing JUSThink platform for building our Engagement Framework 9
- 2.1 Introduction . . . 10
- 2.2 Background . . . 12
- 2.3 Activity Design . . . 15
  - 2.3.1 Learning Task Design . . . 15
  - 2.3.2 The Robot's Role . . . 18
  - 2.3.3 Setup Design . . . 18
- 2.4 User Study . . . 20
  - 2.4.1 Evaluation Metrics . . . 20
  - 2.4.2 Participants . . . 23
- 2.5 Analysis and Discussion . . . 24
  - 2.5.1 RQ1: On Participants' Self-assessment . . . 24
  - 2.5.2 RQ2: On the Relation Between Performance and Learning Gain . . . 24
  - 2.5.3 RQ3: On the Impact of Performance and Learning Gain on Participants' Self- and Robot assessment . . . 25
- 2.6 Key Take-Aways . . . 27

3 Productive Engagement **29**
3.1 Introduction . . . 30
3.2 Background . . . 31
3.2.1 Manual . . . 31
3.2.2 Automatic . . . 32
3.3 Productive Engagement . . . 34
3.4 Research Questions . . . 35
3.5 Generating an Open-Source Dataset: PE-HRI . . . 36
3.6 Evaluating the Hidden Hypothesis . . . 37
3.6.1 Backward Analysis . . . 37
3.6.2 Forward Analysis . . . 41
3.7 Conclusion . . . 50

4 Identifying multi-modal behavioral profiles of collaborative learning in constructivist activities **53**
4.1 Introduction . . . 54
4.2 Related Work . . . 55
4.2.1 Indicators of collaborative learning . . . 56
4.2.2 Building multi-modal models of collaborative learning . . . 58
4.3 Methods . . . 59
4.3.1 Dataset and preprocessing . . . 59
4.3.2 Analysis Approach . . . 62
4.4 Results . . . 66
4.4.1 Pairwise Significantly Distinct Behaviors . . . 66
4.4.2 Interaction Analysis of Multi-Modal Cases . . . 70
4.5 Discussion . . . 76
4.5.1 Speech Behaviors . . . 77
4.5.2 Log Actions . . . 77
4.5.3 Affective Behaviors . . . 78
4.5.4 Gaze Behaviors . . . 79
4.5.5 Tying it All Together: How the Different Modalities Interplay? . . . 79

5 Temporal Pathways to Learning **83**
5.1 Introduction . . . 84
5.2 Literature Review . . . 85
5.2.1 Performance Based Systems . . . 85
5.2.2 Behavior Based Systems . . . 86
5.3 Methods . . . 89
5.3.1 Dataset . . . 89
5.3.2 Analysis Methodology . . . 91
5.4 Results . . . 94
5.5 Discussion . . . 96
5.5.1 Temporal Multi-modal behavioral Profiles . . . 96

5.5.2 Interplay between PS Strategies and other behaviors 103
5.5.3 Connections to Computer-supported Collaborative Learning Literature 105
5.5.4 Implications for Design of Adaptive Learning Interventions 106

6 A Speech-based Productive Engagement Metric for Real-time Human-Robot Interaction in Collaborative Educational Contexts 109
6.1 Introduction 109
6.2 Revisiting Productive Engagement 111
6.3 Problem Statement 113
6.3.1 Treatment of Learning 113
6.3.2 Treatment of Behavioral Patterns 115
6.4 Methods 119
6.4.1 Dataset 119
6.4.2 Analysis 119
6.5 Results 123
6.5.1 Clustering 123
6.5.2 Classification Models 123
6.5.3 PE score 126
6.6 Discussion 129
6.7 Conclusion 129

7 Designing and Evaluating Autonomous Social Robots using the Productive Engagement Framework 131
7.1 Theoretical Description of the Robots 131
7.2 Designing Harry and Hermione 133
7.2.1 Designing Pool of Robot Behaviors 133
7.2.2 Generation of the PE Score in Real-Time 135
7.2.3 Profile Comparison in Real-Time 137
7.2.4 Robot Control Architecture 139
7.2.5 Validation of robot behaviors with Harry in a small online study 144
7.3 Hypotheses 147
7.4 User Study 147
7.4.1 Participants 147
7.4.2 Real-time setup 148
7.4.3 Evaluation Metrics 149
7.4.4 Validation of the thresholds 149
7.5 Results 150
7.5.1 Comparing Harry and Hermione on the Evaluation Metrics 151
7.5.2 Comparing Harry and Hermione on Robot Interventions 154
7.6 Discussion 155

8 Broadening The Horizon **161**
8.1 An Alternate Design for the robot Hermione . . . 161
8.1.1 Construction of the nPE score . . . 162
8.1.2 nPE based real-time Control Architecture for Snape . . . 163
8.1.3 Pilot Study . . . 164
8.1.4 Preliminary Results . . . 165
8.2 Personalization models for Productive Engagement . . . 166
8.2.1 Methodology . . . 166
8.2.2 Initial Results . . . 167
8.3 Incorporating Personality in an Educational Robot . . . 168
8.3.1 Methodology . . . 170
8.3.2 User study with Adults . . . 171
8.3.3 Results . . . 171

9 Synthesis **173**
9.1 Overview . . . 173
9.2 Contributions . . . 177
9.3 Take-aways . . . 180
9.4 Limitations . . . 181

A Appendix A **183**

B Appendix B **185**

C Appendix C **193**

Bibliography **218**

Curriculum Vitae **219**

List of Figures

1.1 Theoretical description of the three robot versions 7

2.1 QTrobot welcomes children to the JUSThink activity. 11

2.2 (a) The introduction sheet to familiarize children with the Cellulo robot, specifically how to move it and to distinguish various types of haptic feedback given by the robot. (b, c) Map 1 and Map 2 used in the learning activity, where the goal on a map is to find the optimal path from *home* represented with a clip-art of a house to (b) *gym* and (c) *cinema*. The source and destination nodes, and the optimal paths are highlighted here with dashed circles and dashed lines respectively. 13

2.3 Photos of (a) a single team while answering the collaborative quiz and (b) several teams participating concurrently in the learning activity in our pilot study . . . 14

2.4 The contents of the screens of the participants, where one participant is in the figurative view and the other participant is in the abstract view. The shown set of tracks forms a minimum spanning tree for the network of gold mines to be constructed together by the participants. Participants swap view after every 2 moves. 16

2.6 The layout of the hardware setup for JUSThink. 21

2.7 Box plots showing the distribution of the ratings given in the questionnaire ($N = 39$ teams) for each question. The questions are listed in Table 2.2. 21

2.8 Distribution of learning and performance metrics. 25

2.9 Relative learning gain vs. last error plot for the teams ($N = 39$ teams). We denote the teams that felt stressed (with team average rating ≥ 4) by a circle 'O', those that said the robot was distracting (rating of 4 or above) by a cross 'X', and those that believed the robot should give more useful feedback (rating of 4 or above) by a plus '+', as rated for questions 11, 18 and 19, respectively. The line represents the linear regression line with a 95% confidence interval. 26

3.1 Overview - Productive Engagement 30

3.2 Clustering of teams in the PE-HRI dataset based on their learning and performance. 39

3.3 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* (non-PS), *Productive Failure* (PF), *non-Productive Failure* (non-PF) and *Productive Success* (PS). 39
3.4 Percentage of variance explained by each individual PC. 43
3.5 Clustering of teams based on their behavioural pattern (extracted from video, audio and log features). 43
3.6 Learning outcomes and performance metric (averaged within cluster) for the clusters computed with the forward approach. Stars denote statistically significant differences ($p < 0.05$) which exist for the pair (F^1_{all}, F^3_{all}). For the pair (F^0_{all}, F^1_{all}), the differences are only marginally significant. Dashed horizontal lines indicate the metrics' global averages. 44
3.7 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 3.1 (forward analysis - columns). 45
3.8 Clustering of teams based on their behavioural pattern (extracted from log features only). 47
3.9 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 between clusters is found. 48
3.10 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 3.1 (forward analysis - columns). . . 48
3.11 Clustering of teams based on their behavioural pattern (extracted from video and audio features only). 49
3.12 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 between clusters is found. 49
3.13 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 3.1 (forward analysis - columns). 50

4.1 Learning gains vs performance. All values here are non-normalized. 62
4.2 Overview of our technique in Nasir, Bruno, and Dillenbourg, 2020. 63
4.3 Features with highest variance between all three behavioral clusters. For the ease of comprehension, each modality is represented by a unique pattern and each behavior within a modality by several shades of the same color. 68

4.4 Significantly distinctive features between the *Expressive Explorers* and the *Silent Wanderers*. . . . 69
4.5 Significantly distinctive features between the *Calm Tinkerers* and the *Silent Wanderers*. . . . 70
4.6 Significantly distinctive features between the two type of gainers. . . . 71
4.7 The two views of the JUSThink game, namely *figurative* and *abstract*, as shown on the screens of the participants when they are empty. . . . 72
4.8 The dialogue for an *Expressive Explorers* team where the blue and red rectangles indicate the duration in which learner A and B are speaking, respectively. Speech overlap is indicated by the overlapping rectangles. Other relevant log and affective features are also shown in a parallel table. . . . 73
4.9 The two views of the JUSThink game, namely *figurative* and *abstract*, as shown on the screens of the participants from a team belonging to the group of *Calm Tinkerers*. . . . 74
4.10 The dialogue for a *Calm Tinkerers* team. . . . 75
4.11 The dialogue for a non-gainer team of *Silent Wanderers*. . . . 75
4.12 The interplay between the problem-solving strategies and the emotional expressivity for the gainer teams. . . . 81

5.1 Behaviors Clustering step . . . 92
5.2 The HMM step . . . 93
5.3 The Analysis Methodology . . . 95
5.4 HMM State diagram for the Expressive Explorers . . . 96
5.5 HMM State diagram for the Calm Tinkerers . . . 97
5.6 HMM State diagram for The Silent Wanderers . . . 97
5.7 Temporal profile for Expressive Explorers . . . 99
5.8 Temporal profile for Calm Tinkerers . . . 100
5.9 Temporal profile for Silent Wanderers . . . 100

6.1 Representation of evolution of learning: in the left hand side figures, it is assumed that learning evolves linearly while in the right hand figures, the assumption is that learning evolves non-linearly where t, L, g, and ng represent time, learning, gainers and non-gainers. The thicker/green lines correspond to the gainers. . . . 114
6.2 Representation of a direct mapping between behavioral patterns and learning. 114
6.3 Representation of an indirect mapping between behavioral patterns and learning. 114
6.4 Behavior of the proposed *PE_Score* when keeping speech level at 0, 0.5 and 1, respectively . . . 122
6.5 Behavior of the proposed *PE_Score* when keeping the Overlap_to_Speech_Ratio level at 0, 0.5 and 1, respectively . . . 122
6.6 Behavior of the proposed *PE_Score* when keeping the Long_Pauses level at 0, 0.5 and 1, respectively . . . 122
6.7 k-means clustering on the principle components generated from the behavioural windows. . . . 124

6.8 Raw PE scores for two random gainer (top) and non-gainer teams (bottom) . . 126
6.9 Two clusters returned by agglomerative clustering where the numbers on the x-axis represent the team index. The smaller cluster on the left consists all of the non-gainer teams 128

7.1 Theoretical description of the three robot versions 132
7.2 The placement of our robots *Ron*, *Harry* and *Hermione* on the space of domain knowledge and behavior knowledge 133
7.3 Facial expressions of QTrobot in horizontal order from top left corner: neutral, smiling, happy, sad, confused, surprised, bored/yawning, puffing cheeks/being cute, and winking. 134
7.4 Pipeline for the generation of the PE score 137
7.5 Pipeline for the generations of profiles 139
7.6 Robot Control Architecture for both *Harry* and *Hermione* 141
7.7 A simplified visualization of the action selection technique for *Hermione* 145
7.8 Setup adapted for an online study with *Harry* 146
7.9 Children interacting with *JUSThink-Pro* at the six schools that participated in the *Harry* and *Hermione* study 148
7.10 A zoomed in view of the *JUSThink-Pro* setup at one of the schools 149
7.11 Validation of the thresholds for profile classification 150
7.12 Linear regression between the the PE scores and the learning gains of the teams that interacted with *Harry* (on the left) and *Hermione* (on the right). For the former, the *PE score* significantly predicts the learning gain with a β of 0.39 and a p-value of 0.01 while for the latter, we do not find a significant result. 151
7.13 Comparison of Harry and Hermione in terms of our evaluation metrics where the asterisk on the graph represents a significant difference on the Kruskal Wallis test. There is a significant difference between the two robots in terms of the *PE score* (p-value: 0.03) and the *suggestion_usefulness* score (p-value: 0.06). 152
7.14 Comparison between the low learning teams in the two conditions. None of the metrics differ with statistical significance. 153
7.15 Comparison between the high learning teams in the two conditions where the asterisk on the graph represents a significant difference on the Kruskal Wallis test in terms of the *PE score* (p-value: 0.004) and the *suggestion_usefulness* score (p-value: 0.05). 153
7.16 Comparison of the intervention types received by the high learning teams in the two conditions where *Harry* received significantly more *exploration inducing* (p-value: 0.03) and *reflection inducing* (p-value: 0.0019) interventions while *Hermione* received significantly more *communication inducing* (p-value: $1.64e^{-05}$) interventions 154
7.17 Linear regression between the three intervention types and the *PE score* for the teams that interacted with *Harry*. None of the intervention types are statistically significant predictors of the *PE score* 156

7.18 Linear regression between the three intervention types and the PE score for the teams that interacted with *Hermione*. *communication inducing* and are statistically significant predictors of the *PE score* with p-values of 0.02 and 0.02, respectively. 156
7.19 Percentage of effective interventions for the two robots: 42% and 33% of the intervention type *Exploration inducing*, 6% and 10% of the intervention type *Reflection inducing*, and 53% and 48% of the intervention type *Communication inducing* are effective for *Harry* and *Hermione*, respectively. 157

8.1 The analysis methodology that led to the construction of the nPE score 164
8.2 Children interacting with *JUSThink-Pro* along with the *Snape* robot at a Swiss school 165
8.3 The general architecture for the personalization models for Productive Engagement starting with learner's profile, feature extraction, and then efficient neural architecture system 167
8.4 Comparison matrix between the 3 personalities' suggestions 169
8.5 Description of the personalities in terms of the OCEAN traits 169
8.6 Examples of the suggestions for each robot personalities 170
8.8 Adults interacting with *JUSThink-Pro* at EPFL 171
8.7 Confusion matrix for the recognition of Kauri and Zuri (N = 290) 171
8.9 Matrix showing p-values obtained with Kruskal Wallis test for each parameter . 172

A.1 Comparison between the clusters of the two approaches in terms of the teams they consist of. 184

List of Tables

2.1 Pipeline of the JUSThink activity. 13
2.2 Categorisation of the questions in the questionnaire. 19

3.1 Multi-modal features for the analysis of the participants' engagement in the Forward Approach 40
3.2 Actions units employed for the calculation of positive and negative valence . . 42

4.1 Multi-modal features that represent behaviors and constructs 61
4.2 The three clusters in approach B with mean values for learning gains (LG) as well as the last error. The significantly different learning gains are represented in bold. 66
4.3 p-values for the Kruskal-Wallis analysis on each pair with significance level of 0.05 67

5.1 Log features from our PE-HRI-Temporal dataset 90
5.2 Video based features from our PE-HRI-Temporal dataset 90
5.3 Audio based features from our PE-HRI-Temporal dataset 91
5.4 Interplay between stages of problem solving strategies and behaviors of speech, gaze, and affect 101

6.1 Characterization of Productive Engagement 116
6.2 Factors of our problem statement and their associated Characteristics 116
6.3 Log features from our PE-HRI-Temporal dataset 117
6.4 Video features from our PE-HRI-Temporal dataset 118
6.5 Audio features from our PE-HRI-Temporal dataset 118
6.6 Classification Results I for when we consider each sequence as a data point. Please note that there is one multi-variate sequence per team 124
6.7 Classification Results II for when we consider each window (non-incremental) in a sequence as a data point 125
6.8 Classification Results III for when we consider each window (incremental) in a sequence as a data point 125
6.9 Validation test 1: Kruskal Wallis tests for the averages (test 1a) as well as for all the points (test 1b) in PE score sequences of the gainers (G) and non-gainers (NG) 127
6.10 Validation test 3: Kruskal Wallis tests between the DTW distances of gainers (G) with the two groups (test 3a) as well as the non-gainers (NG) with the two groups (test 3b) 128

7.1 Examples of Robot Interventions 136
7.2 Threshold values for *Hermione* where *EE* and *CT* stand for *Expressive Explorers* and *Calm Tinkerers* respectively and the numbers represent the minutes into the game. 139
7.3 Robot tasks for both *Harry* and *Hermione* 141
7.4 Regression tests for both *Harry* and *Hermione* 155

A.1 Classification Results 184

B.1 Features' Mean values in each of the Expressive Explorers' states 186
B.2 Features' Mean values in each of the Calm Tinkerers' states 187
B.3 Features' Mean values in each of the Silent Wanderers' states 188
B.4 p-values from Kruskal-Wallis test on the Expressive Explorers' states 189
B.5 p-values from Kruskal-Wallis test on the Calm Tinkerers' states 190
B.6 p-values from Kruskal-Wallis test on the Silent Wanderers' states 191

1 Engagement in HRI: Deceptively Simple, Endlessly Complicated

"*If our brains were simple enough for us to understand them, we'd be so simple that we couldn't.*"
— Ian Stewart, The Collapse of Chaos: Discovering Simplicity in a Complex World

Did you ever talk with someone who was profusely nodding to your profound statements with an attentive expression and a focused gaze only to realize many minutes after that they didn't hear a word you said? or with someone who looks bored to death when you speak but then surprises you with an insightful comment? Being *engaged* in a certain situation/interaction seems to be more of a *hidden* state that does not appear to manifest in the same way in all humans given the complex species that we are. To say the very least, engagement and the way it is surfaced in observable cues can depend on many varied factors such as the context, the personality of the people involved and then their personal circumstances that day or hour, the weather in the moment, or even political climate such as Trump becoming the president of United States, etc.

Particularly in educational contexts, the engagement of a learner, in addition to the factors mentioned above, can also be influenced by the use of technology, such as tablets, robots, virtual agents, etc. that are now increasingly being incorporated in learning scenarios (Belpaeme et al., 2018; Elgarf et al., 2022; Johal, 2020; Krishna & Pelachaud, 2022; Stower & Kappas, 2021; Tatarian et al., 2020; Tulli et al., 2020). Since in such settings, the learner's engagement towards their learning environment is a *means* to an *end*, which is learning; it becomes particularly important for the human or the robot tutor or mediator to perceive it effectively and intervene when required. While it is not possible to gauge all the aforementioned factors on which engagement can depend, this thesis aims at proposing a new framework to conceptualize, model, validate and utilize engagement based on *multiple* observable cues in a learning context. Precisely, the context used in the thesis is that of an open-ended collaborative human-human-robot learning activity in which the robot mediates the activity.

With regards to an extensive overview of engagement in Human-Agent interaction, the author of this thesis contributed in:

C. Oertel, G. Castellano, M. Chetouani, **J. Nasir**, M. Obaid, C. Pelachaud, and C. E. Peters, "Engagement in Human-Agent Interaction: An Overview," in *Frontiers in Robotics and AI (2020), 7:92* Oertel et al., 2020.

1.1 Concept of Engagement in Human Robot Interaction

Engagement is a concept widely investigated in Human-Robot Interaction (HRI) and yet still elusive (Oertel et al., 2020). While some researchers see it as a process, others view it as a state. Commonly adopted definitions include the one of Sidner et al. (Sidner et al., 2005) where engagement is considered 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. (Poggi, 2007) where engagement is considered as "the **value** that a participant in an interaction attributes to the goal of being together with the other participant(s) and continuing interaction".

While there is a general consensus on the idea that engagement is a **multi-dimensional construct**, opinions differ concerning the dimensions composing it. Castellano et al., investigating predictors and components of engagement, regard engagement as characterised by both an affect and an attention component (Castellano et al., 2014). 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 or cues related to engagement would differ" (Salam & Chetouani, 2015a). 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)" (H. O'Brien et al., 2016; H. L. O'Brien & Toms, 2008).

Then with regards to the nature of the HRI scenario/context, there seems to be a **social/task** distinction in the HRI engagement literature that is covered in the definition by Corrigan et al. in Corrigan et al., 2013. 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, there is a possibility for having a multi-party scenario, i.e., when there are two or more people involved in the interaction. Since engagement itself is still rather ambiguous, as explained so far, 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" (Salam

& Chetouani, 2015b). 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” (Oertel et al., 2011) whereas Gatica et al. define group interest as “the perceived degree of interest or involvement of the majority of the group” in (Gatica-Perez et al., 2005).

1.2 Challenges

As with every evolving line of research, there are many open questions, challenges, and limitations in the field of HRI when it comes to how engagement should be modelled (Oertel et al., 2020) and especially how its understanding can be incorporated in a robot for providing effective interventions for advancing learning. We have identified four challenges that serve as the motivation for this thesis as well as what we tackle in this thesis.

1.2.1 Relationship between Engagement and Learning

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. A long-term study (Park et al., 2019) in a story telling context with a robot found that an affective policy trained using reinforcement learning approach successfully personalized to each child and led to a boost in their learning outcomes and engagement. Baxter et al., 2017 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 and Mutlu, 2012 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". In Brown et al., 2013, 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 changing the robot’s behavior has an impact on learning, while carrying a linear assumption that increasing users engagement leads to increased learning, i.e., they manipulate engagement and see an improvement in learning. Hence, the standard approaches in the literature look to *maximize* engagement itself.

But, is it correct to assume that maximizing engagement, as currently defined and modelled, maximizes learning? We believe this relationship has not been explicitly or extensively investigated in HRI. With the engagement framework that we propose in this thesis, we aim to critically assess this relationship.

1.2.2 Human Subjectivity when Modelling Engagement

For a robot to assist students, automatic detection of engagement would be a necessity so that the robot could give immediate feedback to the learners. Currently in HRI, for building such automatic models, one of the most popular methods is to employ several human experts to annotate the data corpora where chunks of videos are annotated on different scales of engagements. These scales can be nominal, ordinal, interval, or ratio. Studies by Rossi et al., 2021; Salam, Çeliktutan, et al., 2017; Sanghvi et al., 2011a outline the process for their use case. Once the annotation process is completed, inter-rater reliability is calculated using metrics like Cohen's Kappa or Krippendorf's Kappa among others. However, keeping into consideration the various ways in which engagement is defined and understood, there remains a huge risk of subjectivity that can lead to low inter-rater reliability (Oertel et al., 2020).

In addition to the challenge of low inter-rate reliability, this process is also very time and effort intensive. In our work, we aim at building a data driven pipeline for modelling engagement that could surface labels without having a human expert in the loop; hence, moving away from a method prone to subjectivity as well as that is time and effort intensive.

1.2.3 In-Task Performance as a Measure of Learning

Taking inspiration from Intelligent Tutoring Systems (ITS) or more generally educational software, which provide a customized feedback to learners, a robot meant to provide interventions in an educational HRI setting should be equipped with a *student model* and a *pedagogical model* (Akkila et al., 2019; Nwana, 1990). The pedagogical model is responsible for making appropriate intervention[s] in the activity (i.e., interventions that have a positive effect on the student's learning), knowing the details of the learning activity and being informed by the student model about the student's status. Bayesian Knowledge Tracing (BKT) (Corbett & Anderson, 1995) is one of the most widely used approaches to model student knowledge (Desmarais & Baker, 2012; Sabourin et al., 2016; Siemens & Baker, 2012). One of the assumptions in BKT is that at each step, the "student can either succeed or fail the task", i.e. there is a straightforward, binary mapping between performance in the task and learning, which makes the approach most "relevant for tutors that use exercises and scaffolding as the main vehicle for learning" (Desmarais & Baker, 2012). However, there is an increasing emphasis towards incorporating more open-ended/constructivist learning activities that encourage the awareness of the knowledge construction process, e.g. by promoting, among other things, *Problem Based Learning* that requires the learners to devise a solution to a real world problem together, and/or *Cooperative Learning* in which interdependence among group members is needed to solve a problem; thus, violating the requirement of a chain of binary right/wrong steps towards the goal (Brooks & Brooks, 1993; Schulte, 1996). The learners rather "become engaged by applying their existing knowledge and real-world experience, learning to hypothesize, testing their theories, and ultimately drawing conclusions from their findings" (Olusegun, 2015).

As a consequence, in-task performance can no longer act as the sole indicator of learning in scenarios that, by design, require the learners to fail and make mistakes along the way as they explore and exploit their environment. Keeping this in mind, in this thesis, we aim to utilize measures, other than in-task performance, to model students knowledge.

Our motivation to tilt towards an open-ended learning activity is exactly the idea that it is a more complex learning environment and being able to asses learner's engagement in such a setting will allow us to build a more robust technique.

1.2.4 Real-time Constraints

With various sources of information on the behavioral data of the students, highly heterogeneous data is generated that requires synchronization before it can be processed or evaluated (Crescenzi-Lanna, 2020; Sharma & Giannakos, 2020; Wagner et al., 2013). Further, some sensor techniques are more intrusive than others such as eye-tracking, EEG, physiological data and currently are not practical in classrooms due to being more expensive and needing high expertise (Sharma & Giannakos, 2020), as well as sensors such as eye-tracking seem to not be compatible for younger children (Crescenzi-Lanna, 2020). In settings where the *timing* of feedback matters such as a robot in an educational setting or an Intelligent Tutoring System (ITS), every sensor that is added to understand the situation comes at a computational cost. However, learning happens in real-time and cannot be paused because the perception-to-inference-to-action loop of the robot needs more time. To have a fast efficient system, which is a challenge on its own, one needs to make a choice of what sensors to focus.

This choice is not very straightforward and cannot be made without prior knowledge on what modality or modalities could be most useful in capturing learning best in a scenario. In our framework, while we start off with a broader range of modalities to try to better understand learning; for the purpose of evaluation in real-time, we plan to converge to a minimalist setup to tackle the aforementioned constraints.

1.3 Research Goals

This thesis aims to critically investigate the relationship between engagement and learning in educational settings. We envision a educational social robot that:

1. can detect what *being engaged in the learning process* looks like
2. can provide feedback to improve such engagement *if* and *when* required
3. can do so in soft real-time
4. can personalize/adapt its interventions

Tying our vision of an educational social robot to the challenges highlighted in the last section, we outline four broader research goals for this thesis:

1. **Research Question 1**: Given the learners behavioral patterns, can we reveal a quantitative relationship that links them to learning?
2. **Research Question 2**: Which learner behaviors are predictive of learning and how?
3. **Research Question 3**: Can we build representations of engagement using the behaviors identified in RQ2 that can then be used for its detection in real-time?
4. **Research Question 4**: How can a robot make use of these representations to induce the relevant behaviors, found as a result of RQ2, in the learners?

In this thesis, we iteratively design robots from *Ron* to *Harry* to *Hermione* with the goal to not only endow a robot with useful knowledge through our proposed engagement framework but also to use it only in an *if*-and-*when*-needed fashion. We refer to the robot with the aforementioned names throughout the thesis to help referring to the different versions of the robot at the different stages of development. Briefly and theoretically, as elaborated in Figure 1.1, what we envision is that *Ron* helps to automate the entire interaction, provides basic motivational feedback to the learner, and it does so while being least aware of its surroundings, i.e., the sensory information coming from learners and the activity. *Harry* has all the capabilities that *Ron* has and additionally it has an idea of *what* behaviors could be useful for learning in the context of the learning activity. Hence, it suggests randomly one among those behaviors at fixed times. *Hermione* too has all the capabilities of *Ron* and additionally it not only has the knowledge of *what* behaviors could be useful for learning, like *Harry*, but it also has an idea of *when* to suggest a particular behavior and *why* to suggest that specific behavior.

With the outlined RQs, we think this thesis lies at the intersection of the fields of educational Human-Robot Interaction and Learning Analytics, particularly multi-modal and collaborative learning analytics, and therefore could be of interest to researchers in both domains.

1.4 Organization of the Thesis

The rest of the thesis is organized in the following way:

Chapter 2: This chapter introduces the learning context that will be used throughout the thesis. More precisely, it outlines the open-ended robot mediated collaborative learning platform *JUSThink* that has been designed and implemented during this thesis together with my colleague Utku Norman. The chapter also shows initial results from a data collection study, also referred to in this thesis as the *Ron* study, conducted with the platform and the robot *Ron*.

Chapter 3: This chapter introduces the concept of *Productive Engagement*, ground it in literature, and validate it in the context of the data collected with the *JUSThink* platform in the

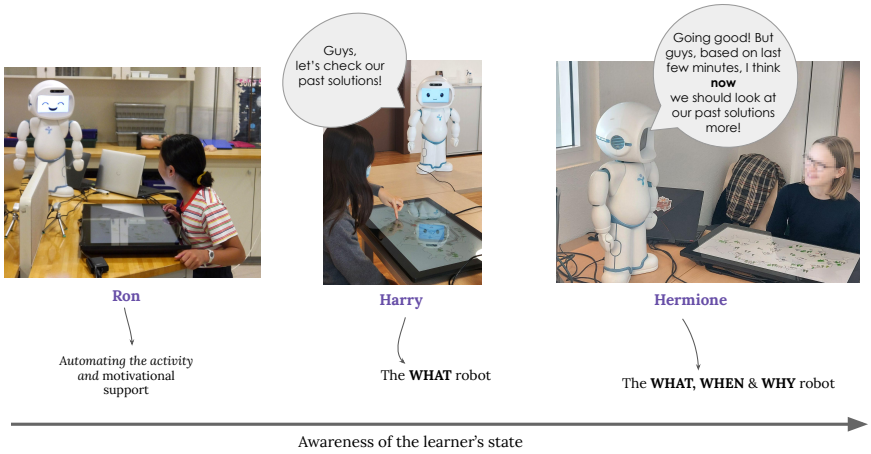


Figure 1.1: Theoretical description of the three robot versions

study described in Chapter 2. The validation results in surfacing productively engaged groups as well as a non-productive group of learners. Precisely, this chapter targets RQ1.

Chapters 4 and 5: In chapters 4 and 5, we investigate what the visible behavioral profiles of these groups reveal about learning in a collaborative open-ended learning activity. In chapter 4, we focus on looking at learner groups at an aggregate level (behaviors averaged over the entire learning activity) whereas in chapter 5, we look at the evolution of behaviors to understand the changes over time within the groups. The outcomes from chapters 4 and 5 identify behaviors that might be more conducive to learning, i.e., indicative of *Productive Engagement*, in such a collaborative learning context and thus constitute a fundamental reference for the robot interventions. Hence, these chapters target RQ2.

Chapter 6: Building on chapters 3, 4 and 5, in this chapter, we investigate method(s) for computing *Productive Engagement* reliably and online. Therefore, this chapter targets RQ3.

Chapter 7: Now that we have obtained some answers for our RQs 1, 2, and 3 through the previous chapters, this chapter focuses on research question 4. We design and implement action selection strategies for the robots *Harry* and *Hermione* that incorporate varying levels of knowledge acquired via the *Productive Engagement* framework developed in chapters 3, 4, 5, and 6. This chapter also presents an extensive study with the robots endowed with the concept of *Productive Engagement* to evaluate the effectiveness of their interventions. This study is also referred to as the *Harry* and *Hermione* study. Precisely, this chapter targets RQ4.

Chapters 8 and 9 and : Chapter 8 discusses the ongoing work particularly some expansions related to the thesis, and possible directions for future research while Chapter 9 synthesizes the findings and contributions of this thesis along with the take-aways and limitations.

2 Designing JUSThink platform for building our Engagement Framework

In order to build our engagement framework and iteratively design the robot *Hermione*, we first need to build an educational HRI activity with a rich context. In this chapter, we present our novel robot-mediated, collaborative problem solving activity for school-children, called JUSThink, aiming at improving their computational thinking skills. JUSThink will serve as a reference for investigating how the robot's behaviour can influence the engagement of the children with the learning process, as well as their collaboration while working on it.

To this end, the *JUSThink* version serving as a baseline with a minimalist supportive robot, presented in this chapter, aims at investigating (i) participants' engagement with the activity (Intrinsic Motivation Inventory—IMI), their mutual understanding (IMI-like) and perception of the robot (Godspeed Questionnaire); (ii) participants' in-task performance and learning metrics. We carried out an extensive user study in two international schools in Switzerland, in which 98 children participated in pairs in one-hour long interactions with the activity. We observe that in-task performance is not correlated with learning. Furthermore, surprisingly, while a teams' in-task performance significantly affects how team members evaluate their own competence, mutual understanding and task engagement, it does not affect their perception of the robot and its helpfulness, a fact which highlights the need for baseline studies and multi-dimensional evaluation metrics when assessing the impact of robots in educational activities.

This work corresponds to the following publications:

J. Nasir*, **U. Norman***, B. Bruno, and P. Dillenbourg, "When Positive Perception of the Robot Has No Effect on Learning," in *29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Naples, Italy, 2020* (Nasir, Norman, Bruno, & Dillenbourg, 2020a).

J. Nasir*, **U. Norman***, W. Johal, J. K. Olsen, S. Shahmoradi and P. Dillenbourg, "Robot Analytics: What Do Human-Robot Interaction Traces Tell Us About Learning?," *28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), New Delhi, India,*

2019, pp. 1-7 (Nasir et al., 2019).

*equal contribution of work.

2.1 Introduction

"Computational thinking (CT) is going to be needed everywhere. And doing it well is going to be a key to success in almost all future careers." The words of Stephen Wolfram[1] capture the urgency seen in the efforts to introduce CT in educational curricula before high school (Menon et al., 2019). At the same time, the potential of robots is increasingly being explored in educational settings across the globe, under the intuition that robots could be an effective tool for advancing CT skills (Chalmers, 2018), as well as for increasing participants' engagement with the educational activity (Belpaeme et al., 2018) and collaboration (Hamamsy et al., 2019; Ioannou & Makridou, 2018). However, crafting pedagogical designs and robot interventions that truly succeed in achieving such objectives is a challenging and to-date open question.

Inspired by this challenge, the JUSThink platform[2] (see Fig. 2.1) aims to:

1. improve the computational thinking skills of children by exercising their abstract reasoning with and through graphs (posed as a way to represent, reason about and solve a problem),
2. promote collaboration between participants, by providing team members with different, complementary information at all times during the activity,
3. serve as a platform for the design and evaluation of robot behaviours aiming to ultimately improve learning, by improving participants' engagement with the task as well as collaboration and mutual understanding between them (Nasir, Norman, Bruno, & Dillenbourg, 2020b).

From a research perspective, and in line with the objectives outlined above, the designed robot-mediated activity is also aiming to surface cues relevant to (i) participants' engagement with the task at hand, their partner and the robot, (ii) mutual understanding and misunderstandings between the participants.

The contribution of the first version of the JUSThink platform is twofold:

1. Provide a baseline for a robot-mediated human-human collaborative learning activity in which the robot automates the entire interaction moving the activity from one phase to another, gives instructions, and provides basic motivational feedback to the learner,

[1] https://blog.stephenwolfram.com/2016/09/how-to-teach-computational-thinking/
[2] https://www.epfl.ch/labs/chili/index-html/research/animatas/justhink/

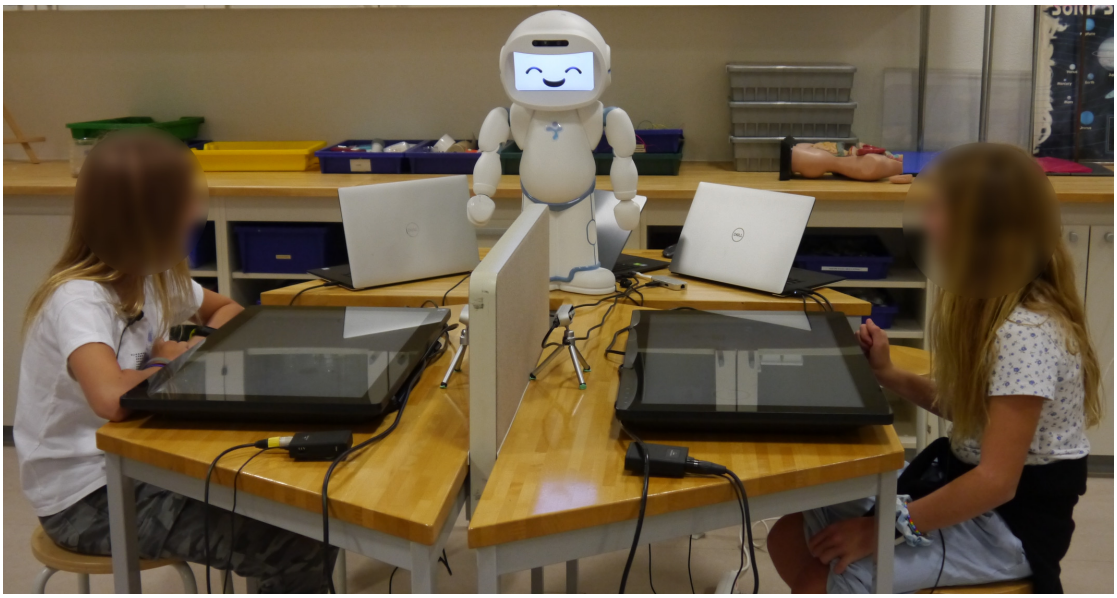


Figure 2.1: QTrobot welcomes children to the JUSThink activity.

without causing unnecessary distractions (we will be referring to this version of the robot as *Ron*);

2. Enable an analysis of the participants' self-assessment of engagement, mutual understanding, and perception of the robot, both independently and in connection with performance and learning in the collaborative activity.

The second contribution serves a double purpose. On the one hand, it is meant as a baseline reference for future studies on the impact that robot behaviours have on participants' learning, performance, engagement, collaboration and mutual understanding. This is the reason why the robot's behaviour in this version is purposefully designed to be minimal and detached from the participants' situation. On the other hand, participants' assessment of a "useless robot", especially if they are struggling with the task at hand, is an interesting insight into the appropriateness of commonly adopted tools for robot evaluation in educational settings. For this reason, in our analysis we complement standard HRI questionnaires with learning and performance metrics.

Moreover, linking back to the distinction in the literature regarding the nature of the HRI context, introduced in Chapter 1, we define our human-human-robot setting where a learning task is present as a *social-task engagement scenario*.

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 (Blaye, 1988; Dillenbourg et al., 1996), or

conflict resolution (Glachan & Light, 1982; Schwarz et al., 2000). To ensure that the interactions of the user are as rich as possible, the activity has to envision another human as a counterpart.

Concretely, in this chapter, with the proposed JUSThink platform and a data collection study, we address the following research questions:

1. **RQ1:** How do participants assess their engagement, mutual understanding and perception of the robot *Ron*, for the proposed JUSThink activity?
2. **RQ2**: Is the first version of the JUSThink activity effective in its pedagogical objective?
3. **RQ3**: Is there a correlation between the performance in the task, or the learning gain, and participants' self-assessment of engagement, mutual understanding, competence, stress and, above all, the robot's behaviour and its helpfulness?

From the above research questions, we derive the corresponding, following hypotheses:

1. **H1**: H1(a): Participants' self-assessment of engagement and mutual understanding is positive.

   H1(b): Their self-assessment of the robot is negative because of its few and limited interventions.
2. **H2**: Performance in the learning task positively correlates with learning gain.
3. **H3**: H3(a): Teams with high performance will rate their engagement, mutual understanding, self-competence higher than teams with low performance, and will have a more positive perception of the robot.

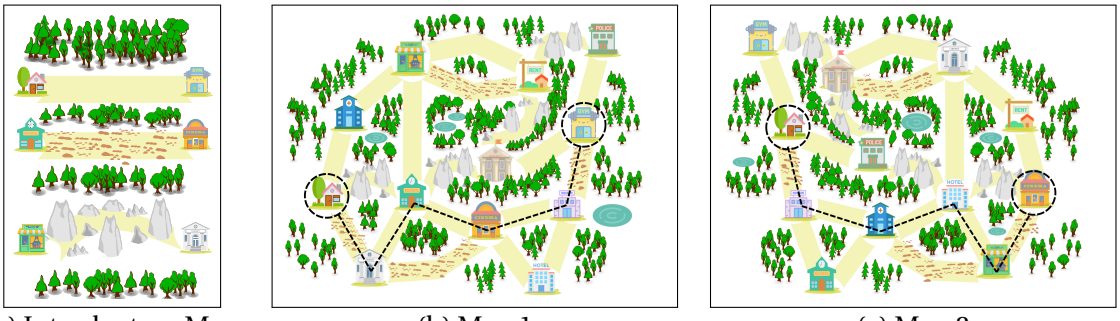
   H3(b): Teams with low performance will rate their stress higher than teams with high performance, and will have a more negative perception of the robot and its helpfulness.

2.2 Background

Robots have been incorporated in collaborative learning activities to support the interaction in various ways. For instance, a robot equipped with emphatic competencies was used to support the interactions of a collaborative learning activity about sustainable development through constructing a sustainable city in a group setting. The robot provided support by considering the affective states of the participants (Alves-Oliveira et al., 2019). Within a learning-by-teaching paradigm (Chase et al., 2009), robots were used: to promote children's responsibility in a collaborative learning activity in which children write on a tablet and the robot gives corrective feedback (Chandra et al., 2015), to aid the reading of children where a child and a robot collaboratively read stories (Yadollahi et al., 2018), and to be collaboratively tutored by children in order to improve handwriting (Hamamsy et al., 2019).

Table 2.1: Pipeline of the JUSThink activity.

Stage	What are the participants supposed to do?	What does the robot do?	Level	Duration
Welcome	Enter their name, age and gender on the screen	Welcome the participants, ask them for personal details	individual	2 min
Introduction	Listen to the robot	Introduce the task goal: connecting the gold mines by spending as little money as possible	team	2 min
Pre-test	Answer a list of multiple-choice questions on the screen	Ask the participants to answer the pre-test questions	individual	≤ 10 min
Demo	Listen to the robot and follow the illustrations on the screen	Explain the two game views and their functionalities	team	3 min
Learning Task	Find a cheapest railway network (a minimum spanning tree) connecting all gold mines by: i) drawing or erasing tracks that connect pairs of gold mines ii) submitting any agreed-upon solution to the robot for evaluation and feedback	At the submission of a solution: If the submitted solution is optimal, congratulate the participants and move to the post-test stage. Otherwise, reveal the cost difference between the submitted solution and an optimal one and motivate the participants to try harder. Point out the availability of the history of submitted solutions if the participants are not successful after several attempts.	team	≤ 25 min
Post-test	Answer a list of multiple-choice questions on the screen	Ask the participants to answer the post-test questions	individual	≤ 10 min
Questionnaire	Rate on a 5-point Likert scale a set of items about engagement, mutual understanding and the robot	Ask the participants to answer the questionnaire	individual	≤ 5 min
Goodbye	See the robot wave goodbye	Thank the participants for their help, say goodbye	team	≤ 1 min



(a) Introductory Map (b) Map 1 (c) Map 2

Figure 2.2: (a) The introduction sheet to familiarize children with the Cellulo robot, specifically how to move it and to distinguish various types of haptic feedback given by the robot. (b, c) Map 1 and Map 2 used in the learning activity, where the goal on a map is to find the optimal path from *home* represented with a clip-art of a house to (b) *gym* and (c) *cinema*. The source and destination nodes, and the optimal paths are highlighted here with dashed circles and dashed lines respectively.

(a) A team of two children

(b) Several teams

Figure 2.3: Photos of (a) a single team while answering the collaborative quiz and (b) several teams participating concurrently in the learning activity in our pilot study

Here, we would like to note that the implications that the *design* of an activity has on enforcing collaboration cannot be ignored: one cannot merely put two students together and expect them to collaborate. This is something we practically experienced via the design of another activity called *Cellulo City* (Nasir et al., 2019), as shown in Figure 2.2 and Figure 2.3. *Cellulo City* is an open-ended collaborative learning activity using tangible haptic-enabled Cellulo robots in a classroom-level setting where the idea is to highlight some of the core concepts involved in path planning by exploratory behavior; hence, serving as an advance organizer (Ausubel, 1960) to a conventional lesson or even as a stand alone session with some modifications. The pilot study, spanning over approximately an hour, was conducted with 25 children aged between 11-12, playing in teams of two. While the study itself contributes to highlighting the potential of the use of learning analytics in educational robotics, we noticed that the design did not enforce collaboration. A more pro-active student in the team could just be doing all the work while the other student passively observes, or one student could dominate the manipulation of the robot, etc.

A careful activity design is thus needed to maximise the chances for the learning mechanisms to occur. Therefore, in the second learning activity design while moving to a very different kind of platform, our design enforces, through specific design choices (elaborated on in the next section), collaboration between the team members while also leaving space for exploration: thus, the participants are expected to have productive interactions (Dillenbourg, 1999) while contributing to a solution together.

2.3 Activity Design

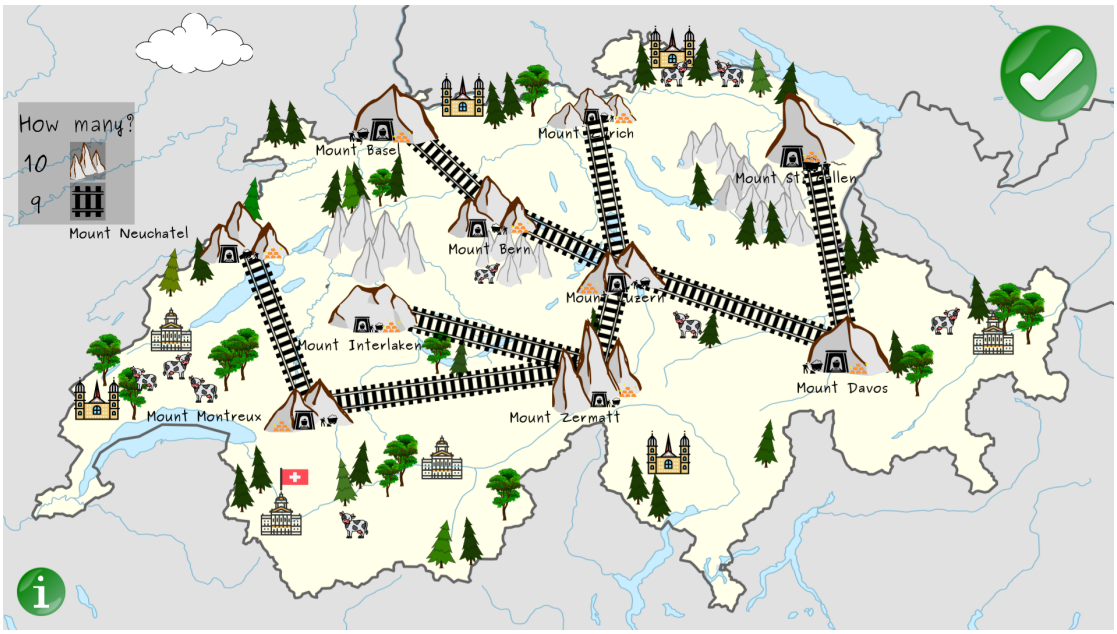
The JUSThink activity is organised in a sequence of stages as described in Table 2.1, the core of which is the learning task.

2.3.1 Learning Task Design

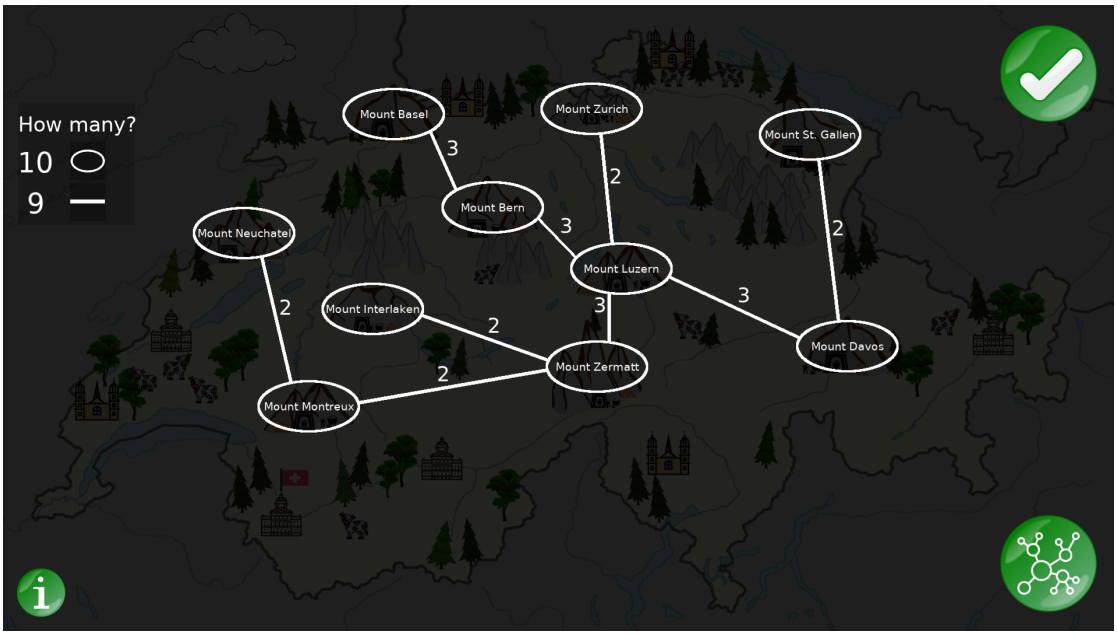
Swiss Gold Mines Scenario

The objective of the JUSThink activity is to give participants an intuitive knowledge about minimum-spanning-tree problems[3] and how to solve them. To introduce the minimum-spanning-tree problem to the participants as a game and with minimal terminology, we created a scenario based on a map of Switzerland. On the map, gold mines are depicted with mountains, animated with glittering gold on them, and labelled with names of Swiss cities (e.g. “Mount Zermatt" and “Mount Zurich"): these make up the nodes V of the graph $G = (V, E)$.

[3]Let $G = (V, E)$ denote a connected, undirected, edge-weighted graph. V is the set of nodes, $E \subseteq V \times V$ is the set of edges that connects node pairs, and $c : E \to \mathbb{R}$ is the edge cost function for G. A subgraph of G is said to “span" the graph G if it connects all nodes of G, i.e. each node is reachable from every other. The problem is to find a subgraph T of G that spans G and minimises $o_T(T) = \sum_{e \in E_T} c(e)$. An optimal solution T is called a minimum spanning tree for G.



(a) Figurative view



(b) Abstract view

Figure 2.4: The contents of the screens of the participants, where one participant is in the figurative view and the other participant is in the abstract view. The shown set of tracks forms a minimum spanning tree for the network of gold mines to be constructed together by the participants. Participants swap view after every 2 moves.

At the start of the Learning Task stage, the robot *Ron*, 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, spending as little money as possible on the tracks. Then , the participants collaboratively construct a solution by drawing and erasing tracks that connect pairs of gold mines, and submit it to the robot for evaluation (one of the two optimal solutions is shown in Fig. 2.4). The cost function and the graph layout draw inspiration from the *muddy city* problem[4]. Note that the cost function is strictly positive.

Scaffolding for Collaboration and Abstract Reasoning

We chose to have an open-ended collaborative activity where learners collaborate to solve an open-ended problem without receiving direct guidance, and this is inspired by the inherent characteristic of such problem-solving followed by instruction (PS-I) activities that encourage the awareness of knowledge gaps, stimulate knowledge construction processes and lead to increased learning gains (Loibl et al., 2017; T. Sinha & Kapur, 2021). Additionally, it is known that collaborative activities need to be scripted for better collaboration and learning (Kollar et al., 2006; Vogel et al., 2017). Therefore, we designed a script based on partial information, role switching and complementarity. A number of design choices have been made in that regard.

Firstly, the screens display two different views that present only *partially observable information* to the participants, with a barrier preventing each participant from seeing the other's screen (see Fig. 2.1). At every point in time within the task, one of the participants is shown the *figurative view* and the other is shown the *abstract view* (see Fig. 2.4) In the figurative view, nodes are shown as mountains and edges as railway tracks connecting two gold mines. Edges' costs are not visible. In the abstract view, nodes are shown as labelled circles, drawn edges as solid lines, while edges drawn and then deleted appear as dashed lines, superimposed over the figurative drawing as a semitransparent overlay. The costs of edges (solid or dashed) are indicated as a number near their center point.

Secondly, the views offer *complementary functionalities*, allowing different actions for constructing a solution. In the figurative view, one can edit the graph by drawing a track or erasing an existing track. In the abstract view, one can see the cost of the tracks, access the previous solutions and their costs, and bring back a previous solution after discarding the current selection. A track is explored after drawing it for the first time, and its cost is displayed in the abstract view until the constructed solution is submitted. Hence, in order to make an informed decision on which action to take (add/delete edges, submit), the participants need to communicate their understanding of what the best move would be based on the information available to them.

Thirdly, every two edits, the views of the participants are swapped, i.e. the participant in the figurative view is then in the abstract view and vice versa. This is so that there are no

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

permanent roles, and that the participants could participate equally in the thought process associated with each view.

Fourthly, the participants can submit only if their solution spans the whole graph. The participants can submit as many times as they want, until they find an optimal solution or the allocated time is over. This allows the participants to experiment with different solutions. The participants are informed of the remaining time only a few minutes before the allocated time is over.

Fifthly, the cost of each track is initially hidden and revealed only after it is drawn. This could promote reasoning about an edge in terms of a connection between two entities with an associated cost.

Lastly, in order to submit a solution, both participants have to select submit (for the same solution) by clicking the submit button on their respective screens. A selection for submission is revoked by an edit on the solution. Thus, the participants need to agree on a solution.

2.3.2 The Robot's Role

The robot's role in the first version of JUSThink, i.e., *Ron's* role is two fold: (i) mediate and automate the entire interaction (see Table 2.1), pausing the participants' applications, giving instructions, and moving from a stage to the next upon its completion; as well as (ii) intervene when a solution is submitted and support through minimal expressive behaviours (as mentioned in Table 2.1). The expressive behaviours include verbal support, using participants' names, and the display of emotions and supporting gestures. Some of these behaviours can be seen in Fig. 2.3.2.

2.3.3 Setup Design

Hardware Setup

The hardware layout required for the JUSThink activity is shown in Fig. 2.6. Two children sit across each other, separated by a barrier. In front of each child, a touch screen is placed horizontally. The humanoid robot (QTrobot[5]) is placed on the side, visible by both children. The children can see each other but not their partner's screen. The experimenter is at all times in the room, ready to intervene. The interaction is recorded by three cameras: one environment camera filming the whole scene and two RGB-D cameras each focused on a child's face. Audio is recorded with two lavalier microphones, clipped on the children's shirts. Two computers, connected to the two touchscreens and to the robot's local network, manage the activity and the synchronous recording of the cameras and microphones. The face cameras are connected to a third computer to alleviate the burden of bandwidth in the local network.

[5] https://luxai.com/humanoid-social-robot-for-research-and-teaching/

Table 2.2: Categorisation of the questions in the questionnaire.

No	Question	Group	Category
1	I was trying very hard to find the best solution.	Cognitive at Task Level (IMI)	Task Engagement
2	It was important to me to do well at this task.		
3	I thought this activity was quite enjoyable.	Affective at Task Level (IMI)	
4	I enjoyed trying to find the best solution.		
5	I was trying very hard while discussing with my friend about the activity.	Cognitive at Social Level (IMI)	Social Engagement
6	It was important for me to discuss with my friend while finding the best solution.		
7	Discussions with my friend were quite interesting.	Affective at Social Level (IMI)	
8	I enjoyed discussing with my friend about the activity.		
9	I think I did pretty well at this activity.	Perceived Competence (IMI)	Own Competence
10	I am satisfied with my performance at this task.		
11	I felt tense while doing this activity.	Pressure/Tension (IMI)	Stress
12	I think my friend understood my instructions very well.	Cognitive (IMI-like)	Mutual Understanding
13	I think my friend understood my emotions very well.	Affective (IMI-like)	
14	I think the robot is competent (capable).	Robot (Godspeed)	Robot
15	I think the robot is intelligent.		
16	I think the robot is friendly.		
17	I think the robot is likeable.		
18	I think the robot is distracting.	Robot (Godspeed-like)	Robot Behaviour
19	I think the robot should give more useful feedback.		
20	I liked the robot.		
21	I would like to play the same game with the same friend.	Game and Friend	
22	I would like to play the same game with another friend.		
23	I knew my friend well.	Known Friend	
24	How many minutes do you think you spent on the part where you played with your friend to find the best solution?	Perception of Time	

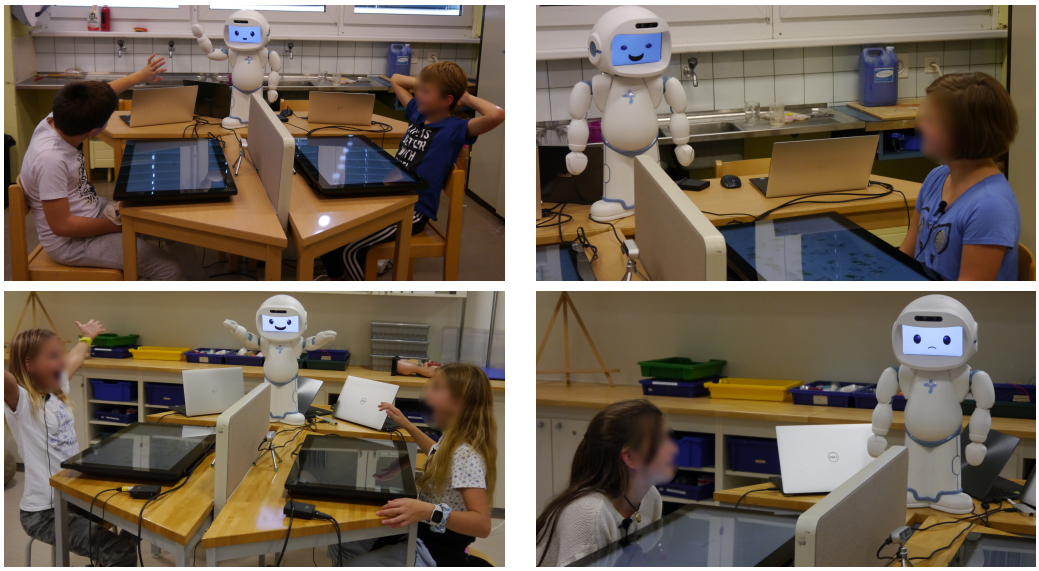


Figure 2.5: Various robot behaviours during the activity. On the top left, *Ron* is waving while welcoming a team to the activity and on the top right, *Ron* is smiling after explaining the rules of the goldmine scenario and all the gold the students will be collecting. On the bottom left is a moment captured right after a team submits an optimal solution and *Ron* is excitedly congratulating the team while the bottom right shows *Ron* exhibiting sadness after saying goodbye to a team at the end of the activity.

Software Setup

Each participant interacts with an instance of the JUSThink participant application that is written in Python and uses pyglet as the windowing and multimedia library. Hence, a separate instance of the application is run for each participant in a team. The JUSThink robot behaviour application is also developed in Python and governs what the robot does and when. The applications communicate with each other via the Robot Operating System (ROS).

2.4 User Study

2.4.1 Evaluation Metrics

Learning Metrics

We generate our learning metrics from the scores of the pre-test and post-test, which are defined in a context other than Swiss gold mines and based on variants of the graphics in the *muddy city*[4] problem.

Specifically, pre-test and post-test are composed of 10 multiple-choice questions, assessing the following concepts:

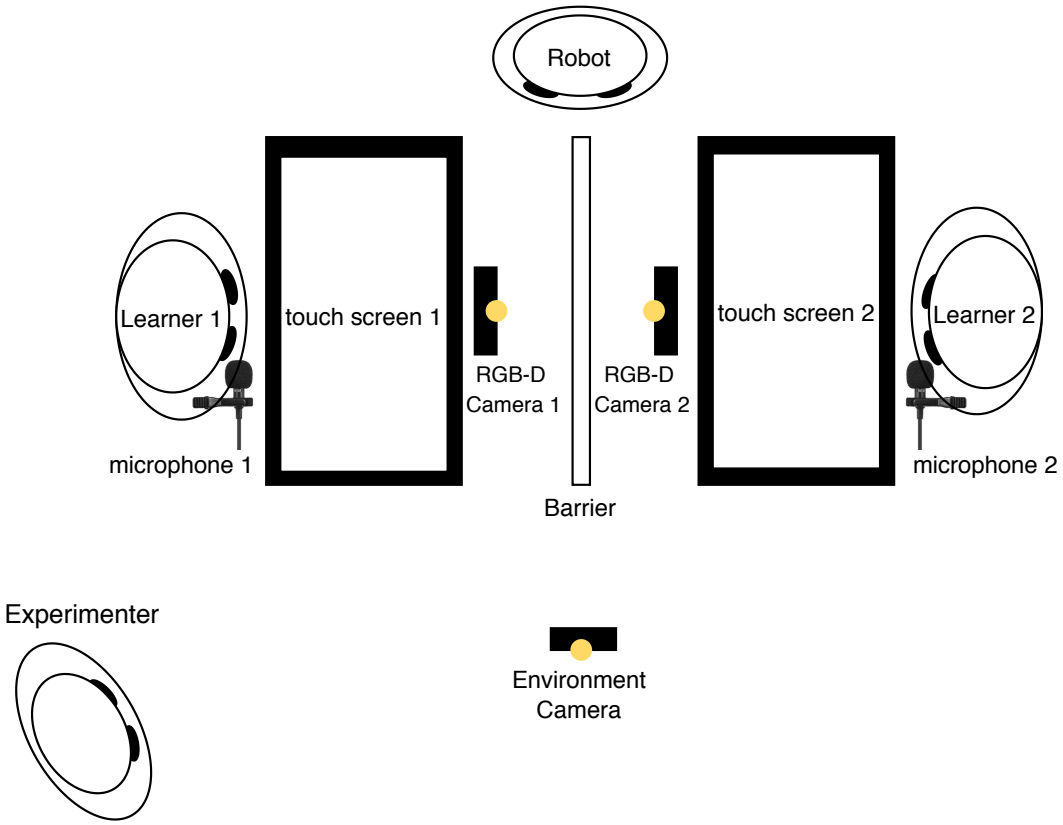


Figure 2.6: The layout of the hardware setup for JUSThink.

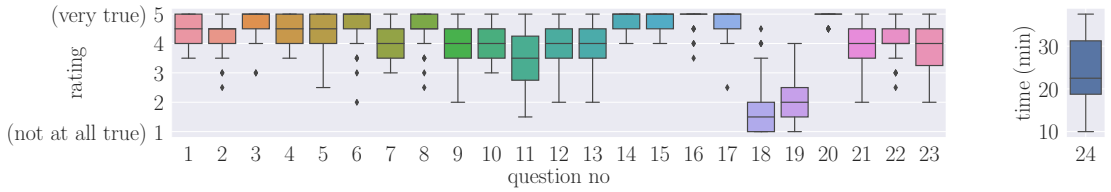


Figure 2.7: Box plots showing the distribution of the ratings given in the questionnaire ($N = 39$ teams) for each question. The questions are listed in Table 2.2.

- **C1:** (exists-or-not, 3 questions). If a spanning tree exists, i.e. if the graph is connected. Example question: “In which map can a postman visit *all the houses* using only the roads?"
- **C2:**(spans-or-not, 3 questions). If the given subgraph spans the graph. Example question: “In which map can a postman visit *all the houses* using *only the black roads*?"
- **C3:**(minimum-or-not, 4 questions). If the given subgraph that spans the graph has a minimum cost. Example question: “In which map can the city build another path with *fewer* stones than the *black path shown* to visit all the houses?".

In C2 and C3, the black path illustrates the given subgraph. The questions are given here in verbatim, where the emphases (here in *italics*) are presented to the participants in uppercase.

The post-test is obtained by randomly shuffling the questions and the response choices within and across the questions related to the same concept, as well as vertically mirroring the images given in the response choices.

On the basis of the pre- and post-test scores, we define two learning metrics:

- **absolute learning gain**, i.e. 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,
- **relative learning gain**, i.e. 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, which grasps how much the participant learned of the knowledge that he/she didn’t possess before the activity.

In the analysis, the absolute learning gains of two team members are averaged, to provide a measure of the **team’s absolute learning gain**. The same procedure is used to obtain a **team’s relative learning gain**.

Performance Metrics

Let error be the difference between the cost of a submitted solution and the cost of an optimal/correct solution (optimal cost), normalised by the optimal cost. Then, we define two metrics to measure the task performance as follows:

- **last error**, i.e. error of the last submitted solution. Note that if a team has found an optimal solution (error = 0) the game stops, therefore making last error = 0.
- **minimum error**, i.e. the minimum of the error values, considering all submitted solutions. This metric is interesting since the last submission does not necessarily correspond to the best solution of a team, in case they have not found an optimal solution.

Questionnaire

The questionnaire consists of 24 questions as reported in Table 2.2. Among them, 11 belong to the Intrinsic Motivation Inventory (IMI) (Ryan & Deci, 2000), which "is a multidimensional measurement device intended to assess participants' subjective experience related to a target activity in laboratory experiments" and relate to engagement, own competence and stress, 5 are ad-hoc questions exploring mutual understanding and the relationship with the team partner (Items 12, 13, 21-23), 4 belong to the Godspeed questionnaire (Bartneck et al., 2009), a widely used instrument in HRI to assess the perception of the robot, which we complement with 3 additional ad-hoc questions on the robot's behaviour and its helpfulness. Question 24 is on the perception of time elapsed.

The Godspeed items concerning the perception of the robot refer to its competence, intelligence, friendliness and likeability and are complemented by behavioural items on being distracting and giving useful feedback. Engagement here entails the effort put in for solving the task (cognitive engagement at task level - Items 1-2) as well as for discussions with the partner to solve the given problem (cognitive engagement at social level - Items 5-6). It also includes the enjoyment that the participants had with regards to the task (affective engagement at task level - Items 3-4) as well as when discussing with their partner (affective engagement at social level - Items 7-8). More on this division of engagement will be elaborated in the upcoming chapter. Similarly, mutual understanding was also measured both in terms of their understanding of each other's instructions for solving the task (cognitive - Item 12) and their understanding of each others' emotions (affective - Item 13).

With respect to the Research Questions driving this study, the questionnaire by itself is meant to investigate RQ1, the learning and performance metrics allow for investigating RQ2, while all metrics together are used to investigate RQ3.

2.4.2 Participants

The *Ron* study was conducted with 96 children aged 9 to 12 years[6]. Due to technical issues during the experiment, 18 participants are omitted from the analysis, resulting in a dataset of 78 children (41 females: $M = 10.3, SD = 0.75$; 37 males: $M = 10.4, SD = 0.60$). The experiment took place over the span of two weeks in two international schools in Switzerland and the participants participated in teams of two, in sessions lasting approx. 50 minutes. The activity pipeline is summarised in Table 2.1. There were always two experimenters available in the room but the system was fully automated to require the least intervention by the experimenters. While the participants were generally familiar with robots as a part of their curriculum and STEM activities, they did not have a prior experience with the robot platform used in this study which could introduce some novelty effect; however, that is a well-known HRI problem.

[6]Ethical approval for this study was obtained from EPFL Human Research Ethics Committee (051-2019/05.09.2019).

2.5 Analysis and Discussion

It is to be noted that while the questionnaire and tests were done individually ($N = 78$ participants); for the purposes of our analyses, we report values as a team average ($N = 39$ teams).

2.5.1 RQ1: On Participants' Self-assessment

In Fig. 2.7, we see the distribution of the team-averaged ratings for all questions in the questionnaire.

Engagement and Mutual Understanding

Participants rated themselves to be engaged highly at both task and social level (mean(1-8)= 4.43). Similar to engagement, the participants rated the understanding of their instructions and emotions by their partners as very high (12, 13). These results support H1(a).

Perception of the Robot

Despite the robot having a basic role in the current setup, the participants rated it very highly with regards to competence, intelligence, friendliness, and likeability (mean(14-17)= 4.78)—see 14-17 in Fig. 2.7. It must be noted that the interaction lasted 45 to 50 minutes; hence giving ample time for the participants to form an opinion on the characteristics of the robot (and their limitations). Also, despite a high number of teams not being successful in finding an optimal solution, we see that the majority of them think that the robot does not need to give more useful feedback (19). Furthermore, very few participants found the robot's behaviour as distracting (18). Hence, contrary to our expectations, H1(b) is rejected.

2.5.2 RQ2: On the Relation Between Performance and Learning Gain

We observe a spectrum of gains from negative to positive for the two learning gains described in Sec. 2.4.1.

Fig. 2.8 shows the distribution of the learning and performance metrics. The error of a team in their last submission is $M = 20.2\%$ ($SD = 16.0\%$), where a team that has found an optimal solution has last error = 0. Specifically, 8 out of 39 teams have found an optimal solution. The minimum error achieved by a team has $M = 11.2\%$ ($SD = 9.6\%$), absolute learning gain has $M = 1.0\%$ ($SD = 11.1\%$), and relative learning gain has $M = -7.1\%$ ($SD = 38.4\%$).

To have an in depth view, we calculated Spearman's correlation between each pair of performance and learning metrics; however we did not find any significant correlations.

In Fig. 2.9, we plot all the teams to see how they are scattered in the 2D space spanned

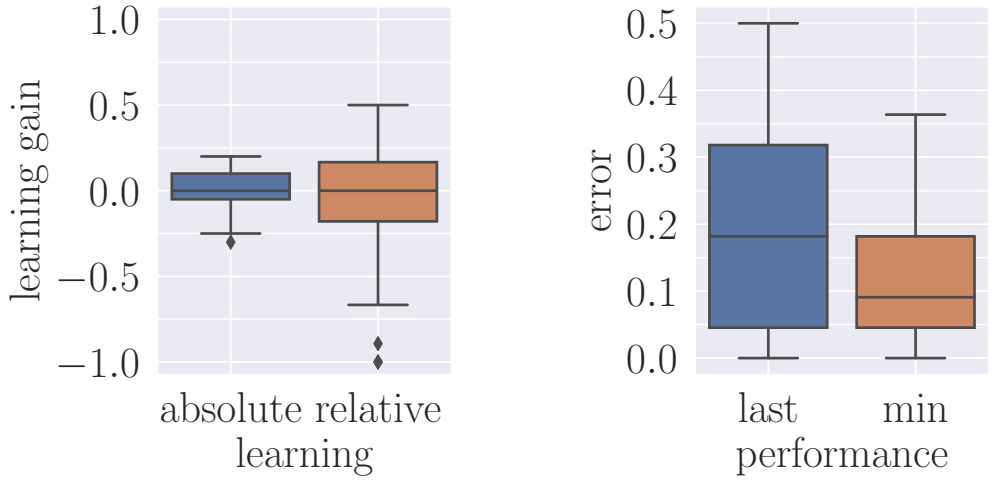


Figure 2.8: Distribution of learning and performance metrics.

by the last error and the relative learning gain. In line with Spearman's correlation results ($r_s = -0.08, p = 0.627$), we observe that the relative gain that a participant achieves is not proportional to their success in the game, which is an important observation for the design of the robot's interventions. In conclusion, participants who appear to be performing well (left side of Fig. 2.9) may not necessarily be developing an understanding of the task, a finding which does not support hypothesis H2.

Lastly, a possible explanation for the observed low learning gains, as well as the lack of a relation between performance and learning gain, is that our pre- and post-tests rely on a high transfer between the task and the test, which is not spontaneous. To elaborate, as mentioned in section 2.4.1, the pre- and post-tests are in a different context where the questions are posed differently than how the problem is presented in the Swiss goldmine task while testing the same underlying concepts.

2.5.3 RQ3: On the Impact of Performance and Learning Gain on Participants' Self- and Robot assessment

In this section, we investigate whether performance or learning gains are related with participants' self-assessment on engagement, mutual understanding, perception of the robot, self-competence, stress, and especially the need for the robot to give more feedback or its assessment as a distraction. Spearman's correlation reports three medium correlations that are significant between last error and competence ($r_s = -0.369, p = .02$), minimum error and competence ($r_s = -0.417, p = .008$), and minimum error and mutual understanding ($r_s = -0.336, p = .03$). This indicates that 1) the higher the last error or the minimum error, the lower would the participants rate their self competence, and 2) the higher the minimum error, the lower mutual understanding is rated.

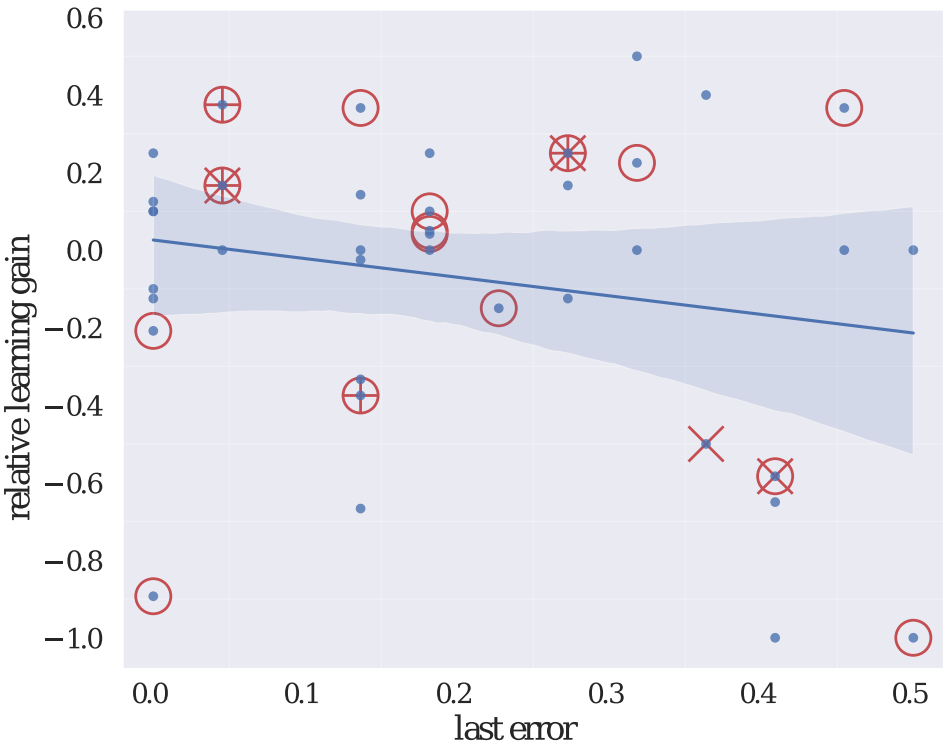


Figure 2.9: Relative learning gain vs. last error plot for the teams ($N = 39$ teams). We denote the teams that felt stressed (with team average rating ≥ 4) by a circle 'O', those that said the robot was distracting (rating of 4 or above) by a cross 'X', and those that believed the robot should give more useful feedback (rating of 4 or above) by a plus '+', as rated for questions 11, 18 and 19, respectively. The line represents the linear regression line with a 95% confidence interval.

It is important to note here that there were no significant correlations found between self-assessment metrics and the two learning gains: participants seem to have based their assessment of self-competence on apparent representations of learning and achievement, e.g. success-failure in the game, rather than the tests which are used to measure learning. A similar result was reported in (Fry, 1976) where the authors observed that “subjects who experienced success made significantly greater gains in positive self-assessments, and failure subjects made significantly greater gains in negative self-assessments". Note that the participants did not receive feedback on their scores in the tests.

We then performed a Kruskal-Wallis test to inquire if teams belonging to high and low in-task performance groups report differently on the aforementioned questions. In line with our hypothesis H3(a), high performing teams (in terms of last error) rated their task engagement significantly higher than those who did not perform as well ($H = 5.669$, $p = .017$, Cohen's $d = 1.11$). Conversely, their perception of the robot is higher than that of low performing teams, but the result is not significant ($H = 2.785$, $p = .095$, Cohen's $d = 0.68$). For this reason, we deem H3(a) to be only partially supported by our findings, and specifically to be rejected concerning the perception of the robot.

Concerning H3(b), we see that there is no significant result neither with Spearman's correlation nor with Kruskal-Wallis test. Indeed, as shown in Fig. 2.9, teams that reported high levels of stress, the robot being distracting, or wished for more useful feedback are dispersed throughout the plot, regardless of their performance. As the figure shows, actually most of the teams that perceived the robot to be distracting or wished for more useful feedback (marked in the figure by a cross and a plus sign, respectively) lie more on the top-left area of the plot, which denotes high learning and high performance (low error). This means that low performance does not make the participants rate their stress higher, have a more negative opinion of the robot or, interestingly, wish the robot could give more useful feedback.

2.6 Key Take-Aways

In this chapter, we presented a novel robot-mediated collaborative educational activity that is evaluated in a user study involving 78 children aged 9-12. The user study aims at assessing various performance and learning metrics, alongside task and social engagement, mutual understanding between partners, self-perception of competence, stress, robot and robot behaviour. We report three key findings: 1) in-task performance and learning are not correlated, and also do not correlate similarly with other metrics; 2) while affecting how a participant perceives their own competence, task engagement, and mutual understanding with their partner, performance has no significant effect on the perception of the robot. Moreover, low performance has no correlation with wishing the robot to give more useful feedback; 3) despite *Ron's* rudimentary behaviour, participants perceive it as highly competent, intelligent, friendly, likeable, not distracting, and report not feeling a need for more feedback from the robot.

Such findings allow for drawing conclusions which, albeit far from definitive, provide insights

for robot-mediated pedagogical activity design. Specifically:

1. The lack of correlation between learning and performance metrics highlights the importance of not being limited to robot interventions that affect (and refer to) only superficial measures of students' learning, such as in-task performance and rather also focus on behavioural patterns that more solidly indicate whether participants would end up learning or not. This links with one of the four challenges discussed in Chapter 1 section 1.2.3.
2. The fact that the performance, low or high, did not have any effect on the perceived usefulness of the robot by the participants highlights the need for well-crafted domain specific metrics to truly assess the effectiveness of the robot and complement the general information provided by standard evaluation tools.

While the results are limited to the specific robot-mediated collaborative activity introduced here; the conclusions drawn from them can be extended to other educational settings, specifically highlighting the need for similar baseline studies and multi-dimensional evaluation metrics when assessing the impact of various robot strategies. As a next step, we move on to exploring and modelling the behavioural patterns, collected in this user study, that could be indicative of higher understanding of the learning goal, and hence are indicative of an engagement which is beneficial for the learning process in an open-ended collaborative robot mediated educational setting. This will lead us to formally introduce the concept of *Productive Engagement*.

3 Productive Engagement

As stated in our introductory chapter, in educational HRI it is sometimes naively believed that a robot's behavior has a direct effect on the engagement of a user with the robot, and the task at hand. Increasing this engagement is then believed to lead to increased learning and productivity. State of the art studies usually investigate the relationship between the behavior of the robot and the engagement state of the user while assuming a linear relationship between user engagement and user 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. Besides being laborious, this introduces further subjectivity in the already subjective construct of engagement. Can we have data driven models for engagement detection where labels do not rely on human annotations? In this chapter, looking deeper at the behavioral patterns, learning outcomes and performance of the children involved in our user-study with *Ron*, we observe a hidden link between student's behavioral patterns and learning that we term as *Productive Engagement*. At this stage, we theorize that a robot incorporating this knowledge will be able to 1) distinguish teams based on engagement that is conducive of learning; and 2) adopt behaviors that might eventually lead the users to increased learning by means of being more productively engaged. This seminal link paves way for data-driven models of engagement in educational HRI.

This work corresponds to the following publications:

J. Nasir, B. Bruno, M. Chetouani, and P. Dillenbourg, "What if Social Robots Look for Productive Engagement?." in *Int Journal of Soc Robotics* (2021) (Nasir, Bruno, Chetouani, et al., 2021)

[Dataset] **J. Nasir**, U. Norman, B. Bruno, M. Chetouani, and P. Dillenbourg, "PE-HRI: a multimodal dataset for the study of productive engagement in a robot mediated collaborative educational setting." *Zenodo* (2020). (Nasir et al., 2020a)

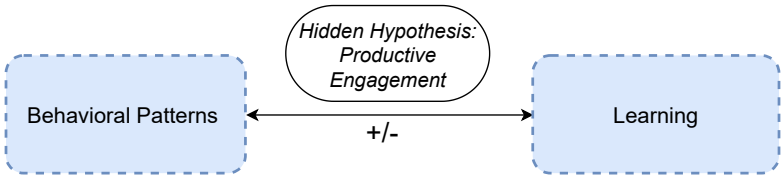


Figure 3.1: Overview - Productive Engagement

3.1 Introduction

As highlighted in challenge 1 in Chapter 1, the standard approaches in the literature look to *maximize* engagement. 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. We postulate that to maximize learning, engagement need not be *maximized*, rather *optimized*. This postulation draws inspiration from the idea of *Productive Failure* proposed by Kapur, 2008 where he theorizes that "Engaging students in solving complex, ill-structured problems without the provision of support structures can be a productive exercise in failure". More often that not, there are learners that consecutively fail in a constructivist design, apparently scoring low on engagement and yet ending up with high learning as demonstrated by our *Ron* study. Similarly, there are learners that seem to succeed in the activity but achieve lower learning. An example of this can be observed in Do-lenh, 2012 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 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 and induce them to reflect more. Going back to the idea of *engaging robot for education*, as pointed out by Belpaeme et al., 2018, designing one such robot is thus not an easy feat: indeed, even experienced human instructors struggle to always make the best choice for an intervention. We believe that the ability 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 key role in the effectiveness of interventions.

If optimal engagement exists, higher learning should be reflected by certain behavioral patterns of the users. These patterns can then be leveraged to inform the behavior of the robot. This chapter makes the following contributions:

- Validate the existence of "a *hidden hypothesis* that links multi-modal behaviors of the users to learning" that we term as **Productive Engagement** (See Figure 3.1).

- The existence of the hidden hypothesis paves way to machine-learning based engagement models for which the labels do not come from human annotators but instead can emerge from the data itself.

3.2 Background

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). While we did touch upon the literature from the HRI perspective in Chapter 1, we briefly look at it from the learning analytics perspective.

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 (Deci, 2017; Wolters et al., 1996), For example, in the work by Ramachandran, Huang, et al., 2019, 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 (Maslow, 1943). These 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 Dewan et al., 2019, include *affective*, *behavioral*, *cognitive*, *academic*, and *psychological*.

Various methods to *measure* engagement along these facets can then be found in the HRI and MLA literature. In Dewan et al., 2019, the authors 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 Dewan et al., 2019, we focus on the *manual* and *automatic* categories:

3.2.1 Manual

Two of the most popular manual methods found both in HRI and MLA engagement literature are: 1) *Self-Reporting*, where “the learners report their own levels of engagement, attention,

distraction, motivation, excitement, etc." (H. L. O'Brien & Toms, 2010; Whitehill et al., 2014); 2) *Observational Checklist*, where external observers complete questionnaires on learners engagement or annotate video or speech data (Kapoor & Picard, 2006; Parsons & LeahTaylor, 2011). While self-reporting is easy to administer and useful for "self-perception and other less observable engagement indicators" (Whitehill et al., 2014), it brings about the issue of validity that depends on several factors such as learners honesty, willingness, and self-perception accuracy, etc. (D'Mello et al., 2014). 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 pointed out in Whitehill et al., 2014, "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.

3.2.2 Automatic

Some of the most widely used methods in MLA and HRI for engagement modelling fall under this category. They can be further 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 disengagement, possibly alongside performance (in educational settings), via behavioral indicators like the frequency of doing a particular behavior, the time taken on a particular action or the confidence level associated with a submitted response, etc. (Alyuz et al., 2016; Castellano et al., 2012; Cocea & Weibelzahl, 2009; Maqsood et al., 2022). Various learning analytics and data mining approaches are used to perform *log-file analysis* in educational settings (R. Baker & Siemens, 2012) including prediction methods, structure discovery, relationship mining, etc. While interaction traces are relatively easy to log and, hence, result in considerable amount of data; they lack 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. (Anzalone et al., 2015; Benkaouar & Vaufreydaz, 2012; Castellano et al., 2009; Foster et al., 2017; Ishii & Nakano, 2010; Ishii et al., 2011; Kim et al., 2016; Papakostas et al., 2021; Salam, Celiktutan, et al., 2017; Sanghvi et al., 2011b). 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. (Chaouachi et al., 2010; Kulíc & Croft, 2007). While collecting video and audio data requires careful privacy considerations, the main advantage of relying on video and audio data only is that the setup can be made relatively unobtrusive. 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 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 (Gordon et al., 2016; Rich et al., 2010), 2) supervised classifiers where the labels come from human annotators (Castellano et al., 2009; Kim et al., 2016; Salam, Celiktutan, et al., 2017), and 3) deep-learning (Nezami et al., 2018) and deep reinforcement learning (Oggi et al., 2019; Rudovic et al., 2019) approaches. 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 methods of the first kind 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, they suffer from the problem of generalization and accuracy since they are modeled in a specific context and the labels are provided by multiple human annotators. We must also note that not many studies actually report the annotation protocol. Lastly, the latest deep learning approaches generally suffer from the lack of interpretability/explainability of results and require an abundance of data.

The brief 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 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 (Perugia et al., 2020). We propose an automatic machine learning method which relies on both log-files and video/audio data, analysed with clustering techniques. This method generates labels for engagement which can be utilized for training a supervised classifier.

Engagement research in HRI is usually studied as the standalone goal of an experiment and, to the best of our knowledge, no study tries to explicitly link it to learning. On the other hand, 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 (R. Baker & Siemens, 2012; R. S. Baker et al., 2008; Corbett & Anderson, 1995; Desmarais & Baker, 2012; Pardos & Heffernan, 2010) in addition to modelling meta-cognitive behaviors, affective states, engagement, and motivation (R. S. Baker et al., 2004; Beal et al., 2004; C. & H., 2009; Craig et al., 2007; Desmarais & Baker, 2012). The reported MLA literature supports our hypothesis that it is possible to "glimpse" learning and performance in the way learners engage with each other and the task at hand. This chapter 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.

3.3 Productive Engagement

In this section, we lay out our definitions for *Productive Engagement* (PE) in the light of the distinctions made in the literature when defining engagement, as highlighted in Chapter 1. Most of our definitions by necessity, due to the conceptions underlying PE where learning needs to be incorporated in the definition, differ from the already existing definitions. Our research is motivated by the following conceptions:

1. Maximizing engagement does not necessarily lead to increased learning outcomes, as first noted in section 3.1.
2. As discussed in section 3.2, evaluating engagement in light of domain specific measures like learning outcomes, that is a more objective construct, and relying upon multi-modal data, can be more effective in open-ended educational settings than using classifiers with labels from human observers.

We define Productive Engagement as the type 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 where the children might not be engaged in what they are supposed to be engaged in) 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 (Corrigan et al., 2013). *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 of a person with other entities (learners and robots).
2. **Task Engagement** that we define as the quality and quantity of interactions of a person with the task.

Furthermore, the choice to have two users in our setting as shown in Figure 2.1, introducing social engagement with a human, is to allow us to grasp all facets of engagement, since we do not know yet which ones better relate to learning. Particularly, 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 (Blaye, 1988; Dillenbourg et al., 1996), or conflict resolution (Glachan & Light, 1982; Schwarz et al., 2000). Since two humans in this setting introduce the concept of group engagement, first outlined in Chapter 1, in our human-human-robot setup, we adopt the definition by Oertel et al., 2011 that is a “group variable which is calculated as the average of the degree to which individual people in a group are engaged". Briefly, for the purpose of analyzing the hidden hypothesis highlighted in the introduction of this chapter, we want to consider multiple facets of engagement as well as have two human users in the setting to have richer interactions.

As seen in Figure 3.1, 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 distinction commonly adopted in HRI (*cognitive* and *affective*), as seen in the review by (Belpaeme et al., 2018), comes under the umbrella of both *task* and *social* engagement aspect of an interaction (elaborated more in our definitions below). To shed more light on the motivation to use this distinction, we include the outcomes (what the robot intervention targets and what the learning activity is designed for) classification from the aforementioned review by Belpaeme et al., 2018. They showed that in most of the studies carried out with robots in educational settings, the outcomes can be classified into *cognitive* and *affective* (Belpaeme et al., 2018).“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, a positive affective outcome does not imply a positive cognitive outcome or vice versa (Belpaeme et al., 2018; C.-M. Huang & Mutlu, 2014). Based on the definitions in the engagement literature (Chi & Wylie, 2014; Henrie et al., 2015; H. O'Brien et al., 2016; H. L. O'Brien & Toms, 2008; Whitehill et al., 2014) as well as our understanding that these distinctions fall under both social and task engagement aspect of an interaction, 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. This can be reflected in their in-task actions (task aspect) or conversations with their partner(s) (social aspect).
2. **Affective engagement** encompasses feelings, enjoyment, attitude, the mood of the learners, etc. This can be considered w.r.t how they feel towards the task itself (task aspect) or how their partner(s) makes them feel (social aspect).

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 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 and/or social space) we are using it to measure.

3.4 Research Questions

We consider our definition of Productive Engagement described in the previous section as a **hidden hypothesis** that “links multi-modal behaviors of the users to learning and performance". Briefly, the analysis in this chapter investigates the following research questions:

- **RQ1:** Given the behavioral patterns, whether cognitive or affective, social- or task-related, can we reveal a quantitative relationship that links them to learning or in-task performance? i.e., do people that differ in their behavior also differ in their learning or in-task 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 contributions of this chapter, 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 (Sharma & Giannakos, 2020)) and learning. In this chapter, we postulate that it is time to reunite the two sides of the equation: robot behavior with user engagement with 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 automatically infer the user's *Productive Engagement* (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.

3.5 Generating an Open-Source Dataset: PE-HRI

From the Ron study described in the previous chapter, where we recorded video, audio and log data of the children interacting with each other and the robot in the context of the *JUSThink* activity, we generated an open-source multi-modal data set called PE-HRI (Nasir et al., 2020a). To ensure that data used for this analysis 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 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 chapter.

[1] This definition is inspired by Dillenbourg's way of defining collaborative learning in (Dillenbourg et al., 1996)

3.6 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 Section 3.6.1. Following that, in Section 3.6.2, we first discuss the engagement-related features extracted from video, audio and log data (see Table 3.1), then investigate the existence of a link between behavior and learning and performance, 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 (Pedregosa et al., 2011).

3.6.1 Backward Analysis

We make use of the following learning outcomes and performance metric (first outlined in Chapter 2, whose definitions are here reiterated for an easier read:

- **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 last 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 grasps the amount of knowledge acquired together by the team members during the activity. This is related to the notion of *shared understanding*.

We calculate these measures for each team, normalize them to have unit variance, and then

perform a K-means clustering on the metrics which yields the results shown in Figure 3.2. The $k = 4$ 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 Figure 3.3. 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* and *Productive Success* (Kapur, 2008, 2016)[2]), 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 Figure 3.3).
- **Productive Failure**, i.e. teams that did not perform well in the task but ended up learning; hence, with higher last errors and higher learning gains (BA cluster = PF in orange in Figure 3.3).
- **non-Productive Failure**, i.e. teams that neither performed well in the task nor ended up learning; hence, with higher last errors and lower learning gains (BA cluster = non-PS in green in Figure 3.3).
- **Productive Success**, i.e. teams that performed well in the task and also ended up learning; hence, with lower last errors and higher learning gains (BA cluster = PS in red in Figure 3.3).

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, i.e., do backward and forward analyses match?

[2] Our groups can be considered to have experienced, at some level, the various phenomenon, outlined in (Kapur, 2016) even though we did not design for it.

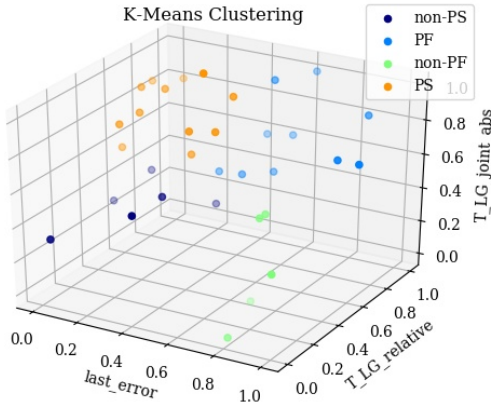


Figure 3.2: Clustering of teams in the PE-HRI dataset based on their learning and performance.

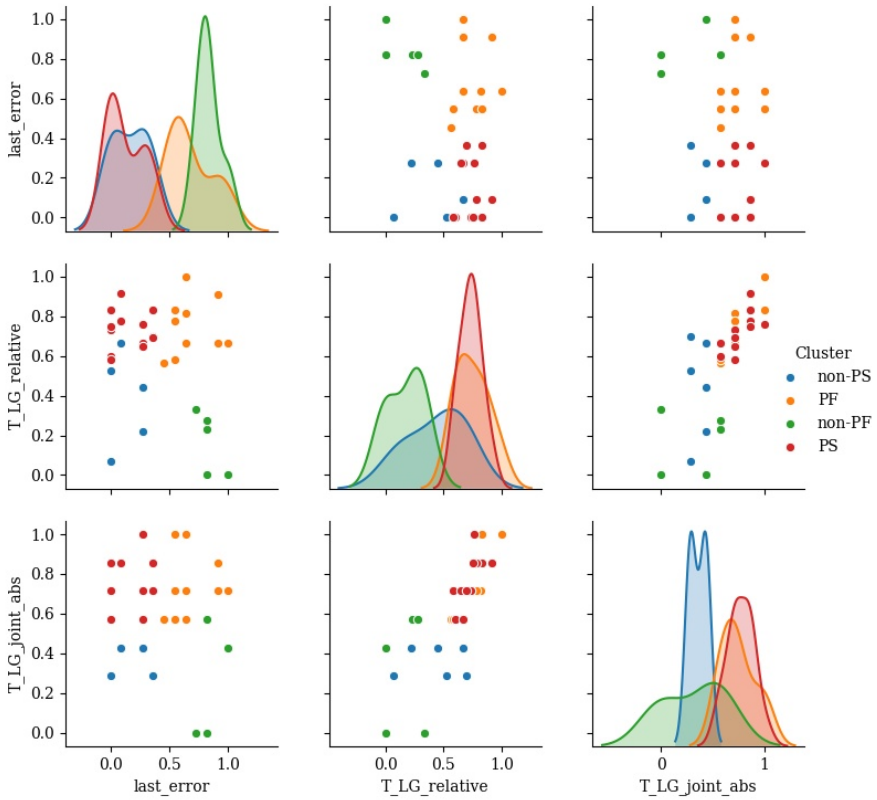


Figure 3.3: Pair plots of the clusters obtained through the backward approach. According to their relative placement w.r.t. learning and performance (and in line with terms and concepts used in Education), we can label the clusters as: *non-Productive Success* (non-PS), *Productive Failure* (PF), *non-Productive Failure* (non-PF) and *Productive Success* (PS).

Table 3.1: 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/Affective
Gaze at Robot	The average percentage of time a team is looking at the robot	Social/Cognitive/Affective
Gaze (Other)	The average percentage of time a team is looking in the direction opposite to the robot	Social/Cognitive/Affective
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

3.6.2 Forward Analysis

Joint analysis of video, audio and log features

As explained in Section 3.2, in this work we focus on video, audio and log features since some of the most commonly used features for engagement detection, such as speech, affective states, and gaze come from such data. Table 3.1 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 section 3.3. As a pre-processing 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 learning task duration only.

We briefly elaborate on how the behaviors are operationalised. We extract log behaviors from the recorded rosbags while the behaviors related to gaze and affective states are computed through the open-source library OpenFace (Baltrušaitis et al., 2016) that returns facial actions units (AUs) and gaze angles. In Baltrušaitis et al., 2016, the authors validate their tool both in terms of AU recognition and eye gaze estimation among other features. Facial Action Coding System (FACS), first presented by Ekman and Friesen, 1978, is considered a major step in the research on facial expressions and is also considered to be the most widely used method for analysing facial expressions (Cohn, 2006). Facial Action Coding System made it possible to map facial muscle movements, indicated by the AUs, to a corresponding displayed facial expression. A detailed table on each AU, its description, the facial muscle it corresponds to, and an example can be found at the IMotions blog[3]. The process of detecting AUs from human faces is now automated by tools such as OpenFace. Certain combinations of these AUs can then be used to infer an emotional state (Baltrusaitis et al., 2011; Benitez-Quiroz et al., 2016; El Kaliouby & Robinson, 2004). We make use of the two dimensions of emotional states, *valence* and *arousal*, often used in emotion research. Valence refers to the pleasantness and unpleasantness of an emotional stimulus (Kauschke et al., 2019). Further each emotional state is also linked to physiological arousal, such as anger and happiness being linked to increased autonomic response while sadness and boredom, are linked to decreased autonomic response (Herman et al., 2018). To generate quantitative values for positive and negative valence, we build on AUs that correspond to positive and negative emotions, respectively, based on the findings from IMotions. Authors in Benitez-Quiroz et al., 2016 also conclude on similar findings. The AUs that we employ for positive and negative valence as well as their description and the emotional states they correspond to are shown in Table 3.2. After smoothening the data for each AU by employing exponential moving average, we take an average of the AUs to return the valence values. To calculate arousal, we use the average of all of the AUs listed in Table 3.2. For the smile extraction based on AUs, we base it on the findings from a smile authenticity study conducted by Korb et al., 2014. OpenFace not only returns the presence of an AU but also its intensity on a 5 point scale. Lastly, the gaze angles generated by OpenFace can be used to determine the eye gaze direction in radians in world coordinates. This means that the given x and y gaze angles in radians are relative to the position of the camera. In our case, a camera

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

Table 3.2: Actions units employed for the calculation of positive and negative valence

Constructs	Action Units (AUs)	Corresponding Description	Corresponding Emotional States
Positive Valence	1, 2, 5, 6, 12, 26	Inner Brow Raiser, Outer Brow Raiser, Upper Lid Raiser, Cheek Raiser, Lip Corner Puller, Jaw Drop	happiness, amusement, surprised
Negative Valence	1, 2, 4, 5, 7, 15, 20, 23, 26	Inner Brow Raiser, Outer Brow Raiser, Brow Lowerer, Cheek Raiser, Lid Tightener, Lip Corner Depressor, Lip Stretcher, Lip Tightener, Jaw Drop	sad, angry, fear

is placed straight in front of a student, i.e., in total two cameras were used, one for each team member. Using these gaze angles, it can be approximated if a person is looking straight ahead, left or right as described in the wiki of OpenFace[4].

For voice activity detection (VAD), that classifies if a piece of audio is voiced or unvoiced, we made use of the python wrapper for the open-source Google WebRTC VAD. WebRTC is a project that provides real-time communication capabilities for many different applications. This project is actively maintained by the Google WebRTC team[5] and due to it being open-source as well as reportedly one of the best and well maintained, there are several wrappers for it now, including for Python and Matlab. With the classification of voiced versus unvoiced frames for each student's audio channel, we can thus generate all the audio features listed in Table 3.1.

Assessing Forward Clusters

To cluster teams based on their behavior pattern, as captured by the 28 features listed in Table 3.1, 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). We use the first three principal components identified by the PCA that account for 50% of the variance within the features dataset. Please note, our criterion for selecting the number of PCs is based on the variance in the dataset explained by each individual PC (see Figure 3.4), visualization possibility, and the motivation to reduce the size of the feature set given a relatively smaller sample size compared to the number of features. 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 Figure 3.5, 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 Figure 3.6. 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 performance, they significantly differ in terms

[4]https://github.com/TadasBaltrusaitis/OpenFace/wiki/Output-Format
[5]https://webrtc.org/

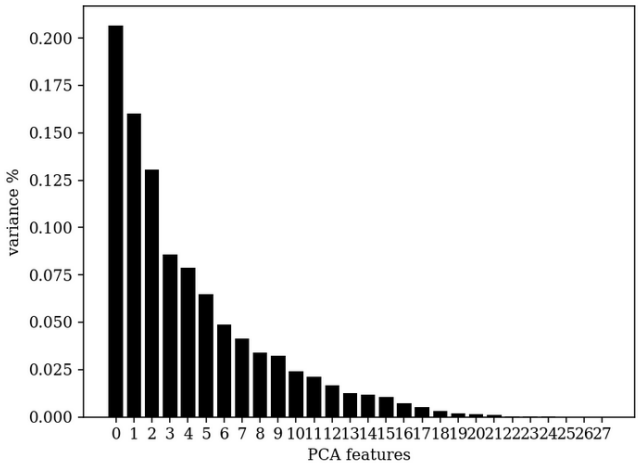


Figure 3.4: Percentage of variance explained by each individual PC.

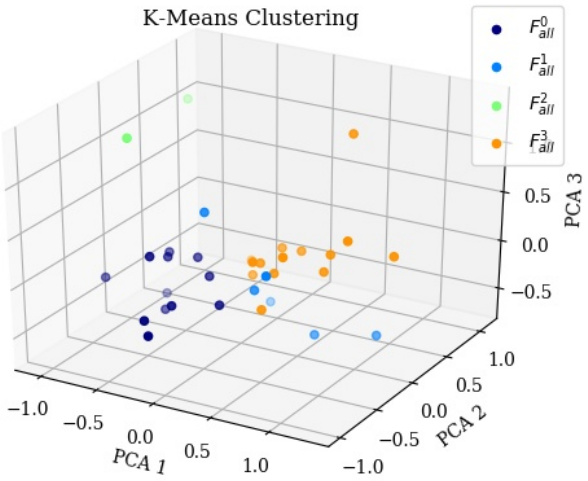


Figure 3.5: Clustering of teams based on their behavioural pattern (extracted from video, audio and log features).

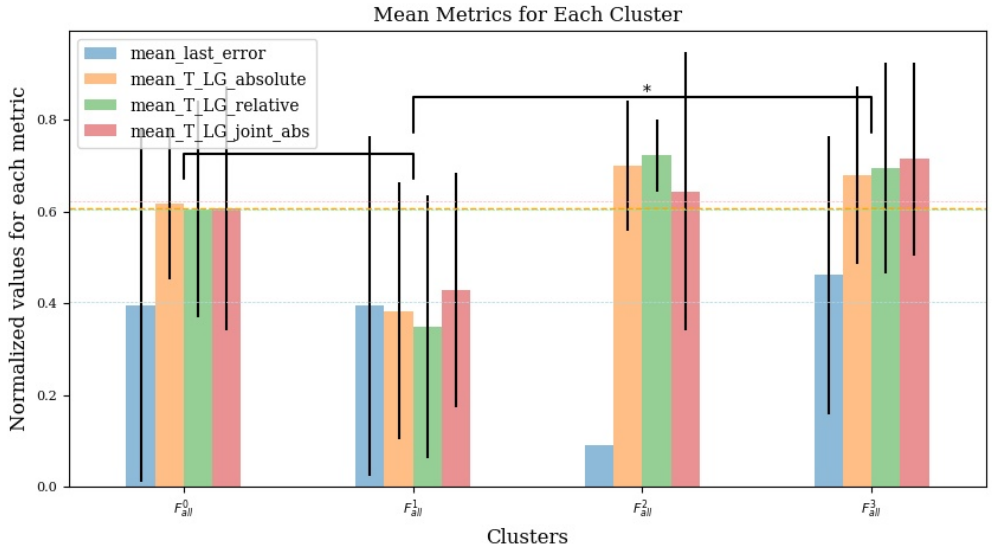


Figure 3.6: Learning outcomes and performance metric (averaged within cluster) for the clusters computed with the forward approach. Stars denote statistically significant differences ($p < 0.05$) which exist for the pair (F^1_{all}, F^3_{all}). For the pair (F^0_{all}, F^1_{all}), the differences are only marginally significant. Dashed horizontal lines indicate the metrics' global averages.

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 lower learning). To validate these differences statistically, we perform a Kruskal-Wallis (KW) test on these metrics between each pair of clusters. In addition to the learning outcomes first defined in Section 3.6.1, 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 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, 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 from the teams ending up with lower learning gains. In other words, this suggests that participants' behavior is indicative of the teams' learning. This, in turn, supports our hypotheses of the existence of a link between engagement and learning (RQ1) and its representability with features that do not require human annotation (RQ2).

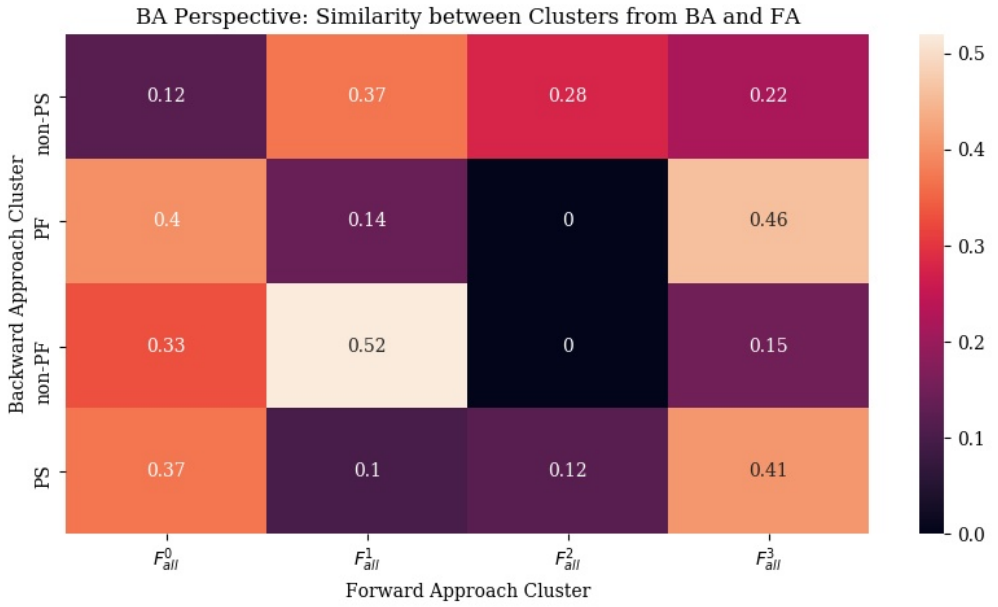


Figure 3.7: Similarity matrix between the clusters computed on the learning outcomes and performance metric (backward analysis - rows) and those computed on the engagement features listed in Table 3.1 (forward analysis - columns).

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_B^F for each backward cluster B with each forward cluster F as:

$$S_B^F = \frac{common\ teams\ in\ both\ clusters}{total\ teams\ in\ both\ clusters} \tag{3.1}$$

which generates the *Similarity Matrix* shown in Figure 3.7. It must be noted here that in Figure 3.7, the naming order of the clusters on the axes is arbitrary, 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 Figures 3.3 and 3.6, along with Figure 3.7.

- 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 Figure 3.3) fall in the forward cluster F^1_{all} ($S^1_{non-PS} = 0.37$, $S^1_{non-PF} = 0.52$), which in fact is the one with lowest average learning gain values (see Figure 3.6 and Figure 3.7).

- Similarly, the majority of the teams belonging to the high-learning clusters (i.e., cluster PF - *Productive Failure* and cluster PS - *Productive Success* in Figure 3.3) 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 Figure 3.6 and Figure 3.7).

This analysis shows that there are similarities in the composition of clusters generated by evaluating the teams' learning and performance and those generated by considering their behavior, as 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 distinguished from those of teams that do not end up learning. In accordance with the definition put forth in Section 3.3, 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-Productively Engaged*). 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.

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 3.1, mostly cognitive, while the other two data sources provide mostly social features (both cognitive and affective). Hence, an additional motivation for the analysis is to check whether features of one type contribute more than the other to explaining the results seen in Section 3.6.2.

Performing PCA and K-means clustering on the log features (first section of Table 3.1), returns 3 clusters along 2 significant PCs (accounting for 55% of the variance within the features dataset) as shown in Figure 3.8. The similarity matrix given in Figure 3.10 between the backward (on learning outcomes and performance metric) and forward (on log-based 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 Figure 3.3) fall more in the forward cluster F^1_{logs} ($S^1_{non-PS} = 0.44$, $S^1_{non-PF} = 0.59$).

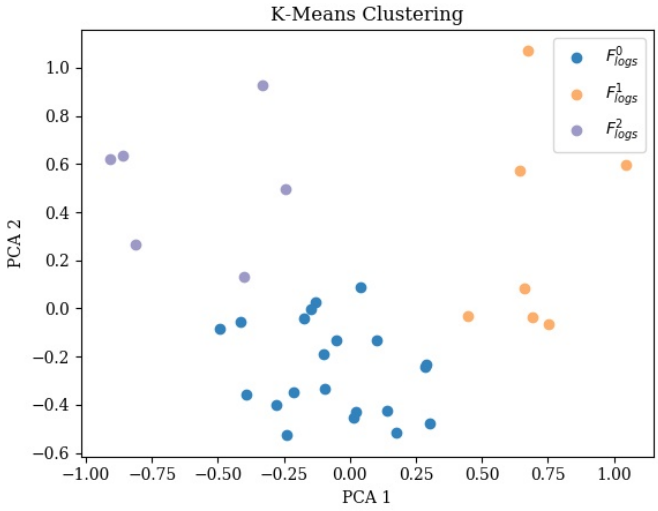


Figure 3.8: Clustering of teams based on their behavioural pattern (extracted from log features only).

- The high-learning backward clusters (i.e., cluster PF - *Productive Failure* and cluster PS - *Productive Success* in Figure 3.3) 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 Figure 3.9 and Figure 3.10).
- However, a Kruskal-Wallis test run pairwise for the forward clusters over the learning outcomes shown in Figure 3.9 reports no statistically significant difference, with only near-significant results 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 Figure 3.11, Figure 3.12, and Figure 3.13) with 3 clusters along 3 significant PCs, we see the same conclusion as previously seen.

- The low-learning backward clusters (i.e., cluster non-PS - *non-Productive Success* and cluster non-PF - *non-Productive Failure* in Figure 3.3) fall more in the forward cluster $F^2_{v\_a}$ ($S^2_{non-PS} = 0.54$, $S^2_{non-PF} = 0.44$) 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 Figure 3.3) 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 Figure 3.12 and Figure 3.13) that have higher learning gain values.
- However, a Kruskal-Wallis test run pairwise for the forward clusters over the learning outcomes shown in Figure 3.12 reports no statistically significant difference.

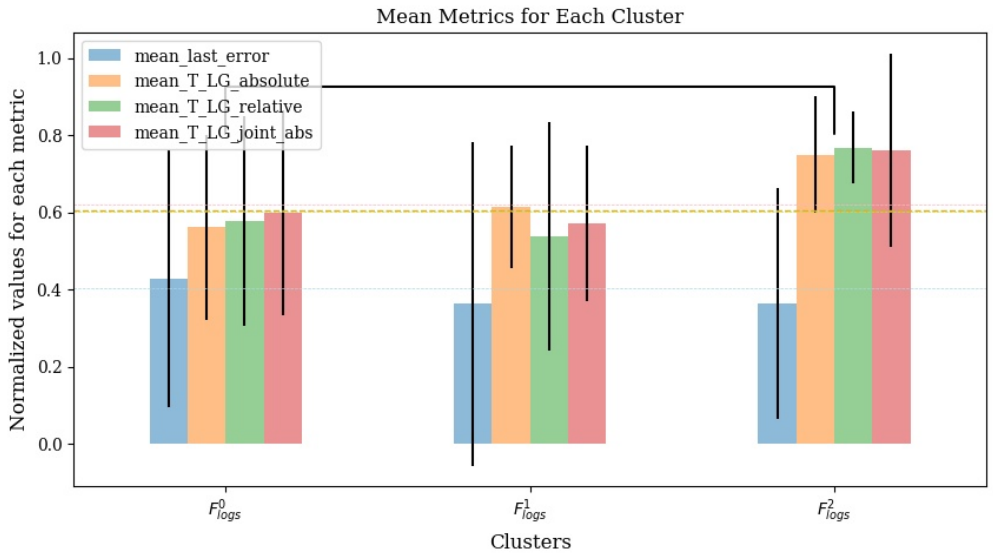


Figure 3.9: Learning outcomes and performance metric (averaged within cluster) for the clusters computed with the forward approach using log features only. Dashed horizontal lines indicate the metrics' global averages. No statistically significant difference between clusters is found.

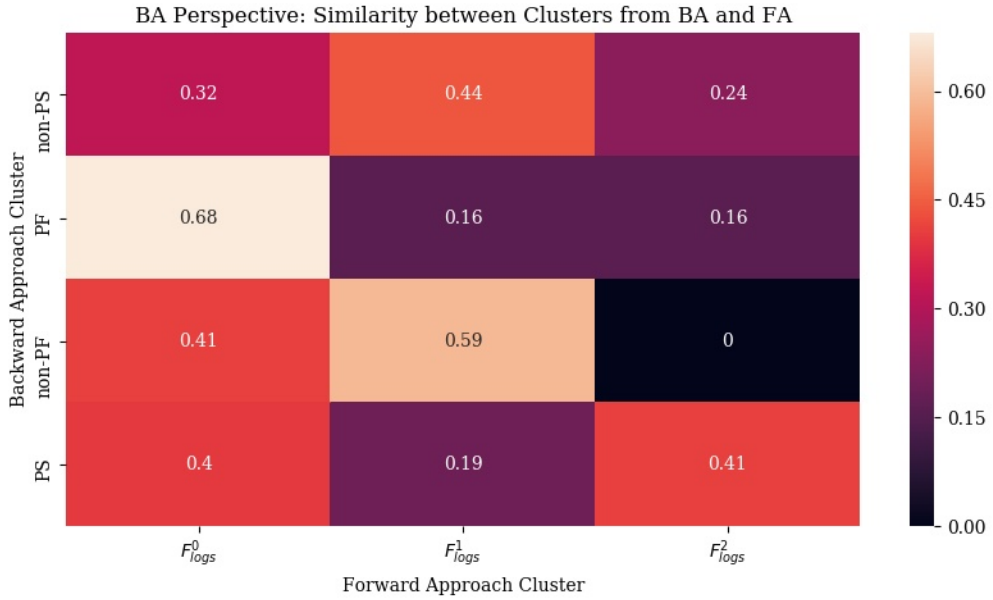


Figure 3.10: Similarity Matrix between the clusters computed on the learning outcomes and performance metric (backward analysis - rows) and those computed on the log features listed in the top section of Table 3.1 (forward analysis - columns).

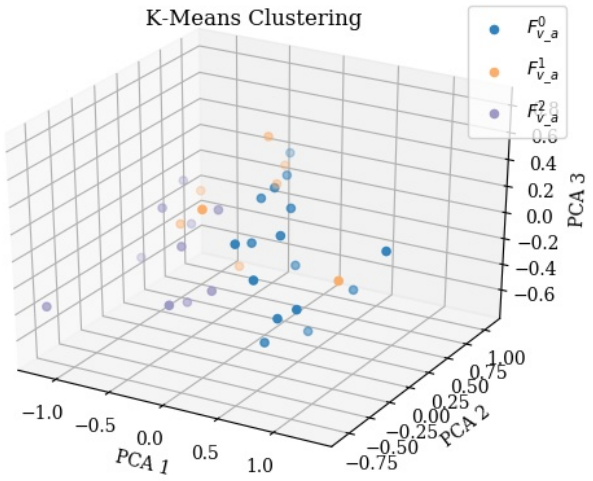


Figure 3.11: Clustering of teams based on their behavioural pattern (extracted from video and audio features only).

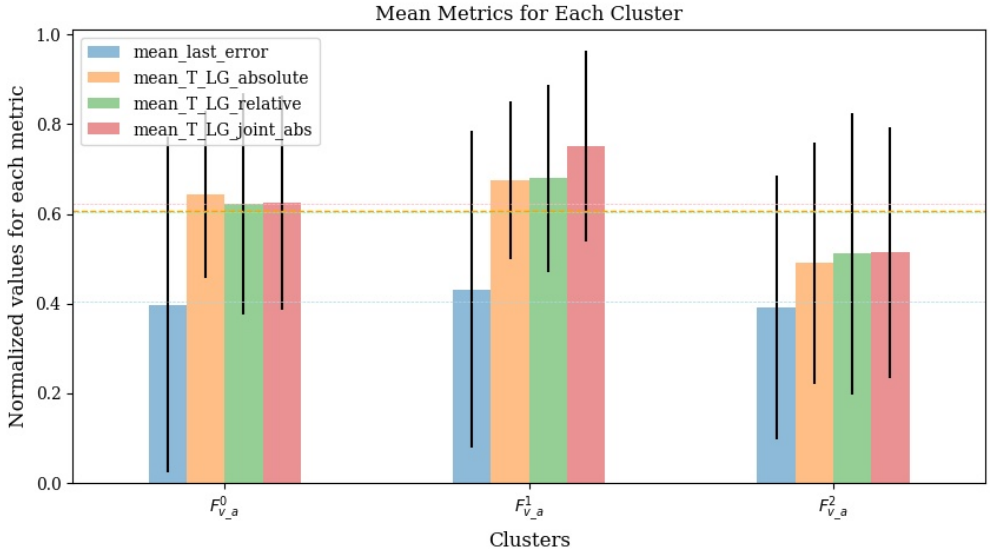


Figure 3.12: Learning outcomes and performance metric (averaged within cluster) for the clusters computed with the forward approach using video and audio features only. Dashed horizontal lines indicate the metrics' global averages. No statistically significant difference between clusters is found.

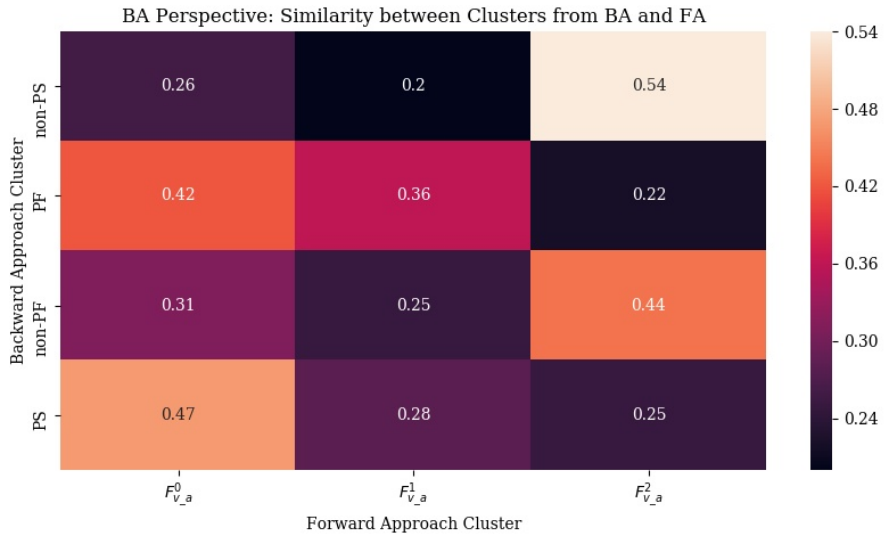


Figure 3.13: Similarity Matrix between the clusters computed on the learning outcomes and performance metric (backward analysis - rows) and those computed on the video and audio features listed in the middle and bottom sections of Table 3.1 (forward analysis - columns).

The results of the type-specific analyses suggest that (1) the results obtained in the global analysis of Section 3.6.2 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.

3.7 Conclusion

As outlined in Section 3.3 and more generally in the motivations for this thesis, 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 (1) the possibility to compute an approximation of user engagement which is devoid of human intervention, to allow for its automatic extraction (RQ2); (2) the preservation of the link between the operationalization of engagement obtained in step 1 and user learning (RQ1). The results we have obtained, reported in Section 3.6, support both hypotheses. The analysis in this chapter explores the link between engagement and learning and, proposes the concept of *Productive Engagement*, its validation in an HRI data set (made publicly available), and considerations on its consequences.

Firstly, we conclude that there are behavioral features, pertaining to task or/and social engagement, that correlate with 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 Section 3.3, 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 Figure 3.3) in the JUSThink activity may or may not 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 Figure 3.3). 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 the 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 comes from the data itself. Specifically, the link between the behavioral patterns and the learning outcomes and performance metric, in the form of statistically significant differences found with KW and the similarity matrix shown in section 3.6.2, 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 characterizing high-learners than low-learners (see results in Section 3.6.2 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.

Additionally, while in-task performance is usually a biasing factor for humans when annotating a subjective construct like engagement in collaborative learning 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.

To determine *what behavior to induce in the user* while designing for effective robot interventions, the immediate next logical step for this research is the characterization of the forward clusters obtained in Section 3.6.2 in terms of the contributions of the single features, and emerging differences between high- and low-learners. This takes us to our following two chapters where we delve deeper into understanding how the *productively engaged* and *non-productively engaged* groups look like in terms of the behaviors they exhibit. 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* that can be practically put in use by a more sophisticated robot.

4 Identifying multi-modal behavioral profiles of collaborative learning in constructivist activities

Now that we have surfaced multiple behavioral profiles depicting productive or unproductive engagement, in this chapter, we formalize our technique that is a combined multi-modal learning analytics and interaction analysis method. Concretely, this technique uses video, audio and log data to *identify multi-modal collaborative learning behavioral profiles* of the 32 dyads that we have from our user study with *Ron*. These profiles, that we name *Expressive Explorers*, *Calm Tinkerers*, and *Silent Wanderers*, confirm previous findings that in a collaborative setting, the amount of speech interaction and the overlap of speech between a pair of learners are highly discriminating behaviors between learning and non-learning pairs. In other words, overlapping speech while turn-taking can indicate engagement that is conducive to learning. However, additionally considering learner affect and actions during the task helps us identify that there exist multiple behavioural profiles exhibited even among those who learn. Specifically, we discover that those who learn vary in their behaviors along the two dimensions of problem solving strategy (actions) and emotional expressivity (affect), suggesting that there is a relation between problem solving strategy and emotional behaviour; one strategy leads to more frustration compared to another. These findings have implications for the design of real-time learning interventions that support productive collaborative learning in open-ended tasks.

This work corresponds to the following publications:

J. Nasir, B. Bruno, and P. Dillenbourg, "Is There 'ONE way' of Learning? A Data-driven Approach." In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion)*. Association for Computing Machinery, New York, NY, USA, 388–391 (Nasir, Bruno, & Dillenbourg, 2020).

J. Nasir, A. Kothiyal, B. Bruno, and P. Dillenbourg, "Many Are The Ways to Learn: Identifying multi-modal behavioral profiles of collaborative learning in constructivist activities" in *International Journal of Computer-Supported Collaborative Learning (IJCSCL)*, 2021 (Nasir,

Kothiyal, et al., 2021).

[Dataset] Norman, U., Dinkar, T., **Nasir, J.**, Bruno, B., Clavel, C., and Dillenbourg, P. (2021). JUSThink Dialogue and Actions Corpus (v1.0.0). Zenodo.

4.1 Introduction

For effective collaborative learning to occur in open-ended learning environments, learners need to share and regulate their own and each others cognition, metacognition, affect and motivations (Järvelä et al., 2020). This learning process is complex and its success has been evaluated based on indicators of discourse, gestures, gaze, cognition and social skills (Spikol et al., 2017; Stahl et al., 2013). As first highlighted in Chapter 3, recent research has suggested that multi-modal data, i.e. integrating multiple of the behavioral indicators listed above, provides an opportunity to more comprehensively characterize learning in open-ended learning environments such as those involving engineering design (Blikstein & Worsley, 2016; Spikol et al., 2017). In this thesis, we consider a behavior as *an action or expression (verbal or facial) of the learner while interacting with the learning environment or their team member*. Further, we refer to multi-modality as *the application and interplay of multiple semiotic modes in order to help understand a specific process, in this case, learning.* In our previous chapter (more specifically in (Nasir, Bruno, Chetouani, et al., 2021)), we found that in an open-ended collaborative learning activity, multi-modal behaviors better distinguish those who learn from those who do not as compared to when only a single modality was used. Further, we argue that it is not straightforward to classify a certain behavior as absolutely good or bad for learning. For example, D'Mello and Graesser, 2012 propose a model to explain the dynamics of affective states that emerge during deep learning. Based on their studies, they suggest that frustration regulation in learners is important as it is considered a negative state (D'Mello & Graesser, 2012; Hone, 2006; Klein et al., 2002). On the other hand, R. S. Baker et al., 2010 suggest that remediation of boredom is more important than frustration. This is also supported by the work of Mentis et al., 2007 who proposes that frustration only needs to be remediated when it occurs due to events that are not under the control of the user, for example, a system bug. Further, the literature on learning by failure suggests that "productive confusion" is conducive to learning as it enables learners to become aware of knowledge gaps and identify deep features (Lodge et al., 2018; Loibl et al., 2017).

The findings above together suggest that there is an interplay between behaviours in their effect on collaborative learning; specifically, the role of a behaviour, such as an affective state, in collaborative learning depends on the context and the accompanying behaviours. This points to the need to examine multiple behaviours together to build a more robust understanding of learning, instead of relying on a single behavior. This is especially important when we have to intervene and scaffold learners appropriately during an activity. This motivates us to explore the use of multi-modal behavioral data to build comprehensive learning vs non-learning profiles in an open-ended collaborative learning setting. In this chapter, we present

an approach for identifying the *collection* of behaviors associated with learning. Specifically, we consider the corpus of multi-modal behavioral data collected during *JUSThink* that follows the *problem-based learning* paradigm (Barron et al., 1998). Our goal in this chapter is to explore the role of multi-modality and identify specifically the multi-modal behaviours which characterize learning and non-learning. We argue that a collection of multi-modal behaviours may offer a richer characterization of collaborative learning in an open-ended activity, so that we may then use these learning profiles to build real-time robot interventions which can scaffold learners.

We investigate the following research question:

> **RQ:** What do learners' visible behaviours reveal about learning in a collaborative open-ended learning activity?

4.2 Related Work

Research on problem-based learning suggests that learners collaboratively working on authentic, open-ended problems is effective for conceptual understanding (Barron et al., 1998; Kirschner et al., 2011). Furthermore, impasses have been shown to play an important role in learning (Kapur, 2008; Schwartz & Bransford, 1998; Schwartz & Martin, 2004; VanLehn et al., 2003); for instance, during coached problem solving, more often than not learning happens when learners reach an impasse (VanLehn et al., 2003). Similarly, when learners solve authentic, open-ended problems collaboratively they often fail, but this failure is productive for learning and leads to deep conceptual understanding and improved transfer (Kapur, 2008; Schwartz & Martin, 2004). Therefore in this work, we broadly adopt the impasse-driven theories of learning such as productive failure which suggest that

- Performance in problem solving is not necessarily an indicator of learning (Loibl & Rummel, 2014).
- Learning is driven by the mechanisms of becoming aware of ones' knowledge gaps, followed by recognition of deep knowledge structures that is engendered in moments of failure during problem solving (Lodge et al., 2018; Loibl et al., 2017).
- Learning while working on activities collaboratively and encountering failures, requires learners to sustain and regulate their own and the teams' cognition, meta-cognition, emotions and behaviours towards completing the task and learning through impasses (Järvelä et al., 2020).

The theory above highlights the need to identify the multiple constructs that are together responsible for the success of collaborative impasse-driven learning. The effectiveness of collaborative learning depends on many factors such as team members speech, their actions

within the learning environment and their eye gaze (Spikol et al., 2017; Stahl et al., 2013). Further, in impasse-driven learning paradigms, as learners work on complex problems, there is a "zone of optimal confusion" (Lodge et al., 2018) where learners become aware of their knowledge gaps and subsequently recognize the deep features of the underlying concept (Loibl et al., 2017). In this zone, confusion can be productive. However if learners' confusion persists, it can become unproductive and lead to frustration and then disengagement (D'Mello & Graesser, 2012). Thus the regulation of emotions becomes crucial in impasse-driven learning situations to ensure that learners do not transcend into disengagement. Putting these factors together we argue that learning while collaborating in a technology-based open-ended activity depends on sharing and regulating learners *speech*, *actions*, *gaze* and *emotions*. Based on this theoretical framing, we choose to focus on these four indicators and their interplay to characterize collaborative learning. Below we elaborate on the literature related to the effect of each of these indicators on collaborative learning and then argue why it is necessary to integrate these indicators to build comprehensive profiles.

4.2.1 Indicators of collaborative learning

While collaboration can make learning more effective, especially in open-ended learning activities, several researchers stress that this depends on the quality of the interaction. Dillenbourg et al., 2009 emphasize that in collaborative settings, particular forms of interactions among people, such as productive verbal elaborations, are expected to occur, which could trigger learning mechanisms, but there is no guarantee that the expected interactions will actually occur. Other work (Barron, 2003; Lou et al., 2001; Meier et al., 2007) similarly suggests that the conditions under which collaborative learning is effective are diverse and complex. Hence, researchers have attempted to understand the collaborative learning mechanisms using various indicators of collaboration such as learners' speech (for instance, (Weinberger & Fischer, 2006)), eye gaze (Jermann & Nüssli, 2012), physiological measures (Schneider et al., 2020) and actions (Popov et al., 2017), and identified conditions for productive collaborative learning. Below we describe some of these indicators and their relationship to productive collaborative learning.

Speech: Speech plays a very important role in collaborative learning as it is primarily through dialogue that learners build a joint understanding of the shared problem space and engage in knowledge construction (Barron, 2003; Roschelle & Teasley, 1995; Teasley, 1997). Within learner dialogue (speech or chats), it has been found that the quantity (eg, number and length of utterances, and talk time) and heterogenity and transactivity of verbal participation (eg, turn taking and building on each others reasoning), along with features of speech such as voice inflection, are indicative of good collaboration (Martinez et al., 2011; Reilly & Schneider, 2019; Viswanathan & VanLehn, 2017; Weinberger & Fischer, 2006). Pauses are also considered an essential part of speech and dialogue as sometimes one pauses to breathe, to plan, or to check whether someone else wants to speak (Fors, 2015; Maroni et al., 2008). Research has shown that shorter pauses (200 - 500 ms), relative to longer pauses (>1000 ms), tend to

be linked with positive perception of speech, the ease of understanding speech as well as memorisation (Fors, 2015). All of these aspects help with better communication, that is related to better collaboration and learning.

Eye gaze: Eye gaze has been used, often along with dialogue, to evaluate collaborative learning (Jermann et al., 2011; Schneider & Pea, 2013; Sharma et al., 2021). Research has shown that measures of joint visual attention, such as cross-recurrence (Jermann et al., 2011; Jermann & Nüssli, 2012; Schneider et al., 2016) and gaze similarity (Sharma et al., 2015; Sharma et al., 2021) are related to increased collaboration quality and learning outcomes. On the other hand, a measure of gaze dispersion is found to be related to misunderstandings (Cherubini et al., 2008) and unbalanced gaze participation is negatively correlated with learning outcomes (Schneider et al., 2018). Similarly, sharing gaze among collaborators is related to improved collaboration (improved transactivity in learner dialogue) and learning gains (Schneider & Pea, 2013, 2015).

Actions: Interaction logs within technology-enhanced learning environments are used to examine the state of learners performance and learning in both individual and collaborative conditions. In collaborative learning, learners clickstream or touch traces are used, often along with their dialogue, to identify productive actions and patterns (Evans et al., 2016; Martinez-Maldonado et al., 2013a; Popov et al., 2017; Rodríguez & Boyer, 2015; Viswanathan & VanLehn, 2017). Research has shown that analytics of task-specific actions when learners collaborate in complex problem-solving environments can be used to distinguish high and low performers in collaborative learning (Emara et al., 2018; Kapur, 2011; Perera et al., 2008). For instance, while collaborating around an interactive tabletop, while the number or symmetry (participation of each member of a team) of actions and speech was not found to relate to collaboration quality, certain sequences of actions and speech were found to be indicative of quality of collaboration (Martinez-Maldonado et al., 2013a). Specifically, low collaborating groups were found to act in parallel, without discussing, while high collaborating groups were found to work together on task-related objects while discussing. Other work has found that the combination of touches to unrelated objects on the screen and multiple users interacting with the screen at the same time can predict collaboration quality (Evans et al., 2016). However, in a chat-based collaborative learning environment, researchers (Popov et al., 2017) found that neither alignment of learner actions (synchrony) nor learners building on each others' reasoning (transactivity) was related to performance on the task, but other factors such as group dynamics and prior knowledge played a more critical role. Thus the role of symmetry, synchrony and transactivity in actions during collaborative learning appears to depend on the context.

Affect: Affect play an important role in learning and so investigating the role of affect or emotions during collaborative learning is an important area of research within collaborative learning (Järvelä & Hadwin, 2013). Arousal and valence, which indicate affect (Russell, 2003), can be inferred from video data and used to evaluate collaborative learning (Dindar et al., 2020; Hayashi, 2019). For instance, Dindar et al., 2020 attempted to characterize collabora-

tion quality by identifying leaders and followers in a collaborative task using the degree of emotional mimicry. Hayashi, 2019 identified that the process of developing mutual understanding during a collaborative task is correlated with negative emotions. Additionally, the relationship between physiological synchrony and collaboration quality has been explored (Malmberg, Haataja, et al., 2019; Pijeira-díaz et al., 2019; Schneider et al., 2020) and initial results suggest that physiological synchrony can be an indicator for collaboration quality. For instance, Schneider et al., 2020 used electrodermal data and identified a metric related to the number of cycles between low and high synchronization to be significantly correlated with collaboration quality and learning outcomes. Together this research suggests that the role of affective and physiological indicators on collaborative learning is still unclear and mediated by other factors. Hence it is important to look at these indicators along with other indicators such as speech and actions, while evaluating collaborative learning.

4.2.2 Building multi-modal models of collaborative learning

The literature above shows that several indicators impact collaborative learning, sometimes in contradictory ways. For instance, while some research suggests that transactivity in actions is not related to good collaboration (Popov et al., 2017), other research shows that transactivity in dialogue is indeed related to collaborative learning outcomes (Schneider & Pea, 2015). Another example is that while Popov et al., 2017 suggest that synchrony in actions is not related to good collaboration, other research suggests that synchrony in gaze is indicative of high quality of collaboration (Schneider & Pea, 2013). These complicated findings suggest that the effectiveness of collaborative learning in open-ended activities depends on multiple interconnected indicators. Recent research therefore investigates collaborative learning by combining multiple indicators obtained through multi-modal data sources in order to develop a richer and more comprehensive understanding of the learning mechanisms. Empirical results suggest that combining multiple sources of data can provide better predictions of collaborative learning outcomes than any single modality of data alone (Emerson et al., 2020a; Giannakos et al., 2019; K. Huang et al., 2019; Liu et al., 2018; Malmberg, Järvelä, Holappa, et al., 2019; Olsen et al., 2020a; Spikol et al., 2018; Vrzakova et al., 2020; Worsley & Blikstein, 2018). Vrzakova et al., 2020, for instance, examined collaborative problem solving among triads and explored combinations of speech, actions and body posture patterns which correlate with task performance. They found that certain multi-modal patterns are better than unimodal patterns for predicting performance. Olsen et al., 2020a investigated collaborative learning outcomes in an intelligent tutoring system and found that combining modalities such as dual gaze, tutor log, audio and dialog provides more accurate prediction of learning gains than models using a single modality.

While multi-modal learning analytics explores different combinations of data streams along with various machine learning methods, what is not yet clear is how these combinations of indicators characterize collaborative learning. In order to develop a richer understanding of the collaborative learning processes, it is necessary to develop multi-modal learning profiles

of groups of learners collaborating. K. Huang et al., 2019 did this by combining eye gaze, physiological sensor and motion sensing data, and identified three multi-modal states and the transitions between them, that are significantly correlated with task performance and learning gains. *In this part of the thesis, we add to this line of research by proposing an approach to build multi-modal collaborative learning profiles of dyads as they work on an open-ended task around interactive tabletops with a robot mediator.*

4.3 Methods

4.3.1 Dataset and preprocessing

We use our PE-HRI multi-modal dataset, first presented in Chapter 3; hence we have 28 multi-modal behaviors extracted from log, video and audio data, alongside performance metrics and various learning gains of the 32 teams. Although, the features themselves have already been described in the previous chapter, we describe them here again in Table 4.1 as in this work, we group them under certain constructs found to be linked to learning.

As discussed in the previous section, several learner measures can be used as indicators of collaborative learning. These can be divided into 1) behaviors, and 2) constructs. As described earlier, we consider a behavior as *an action or expression (verbal or facial) of the learner while interacting with the learning environment or their team member*. We extract them from the log, audio, and video data streams of each participating dyad. These behaviors are representative of constructs that are *non-observable but have been linked to the process of learning*, such as *attention*, *exploration*, *reflection*, *frustration*, *confusion*, *excitement*, *synchrony* or *turn-taking* (Cherubini et al., 2008; Dindar et al., 2020; Hayashi, 2019; Martinez et al., 2011; Nasir et al., 2019; Sharma et al., 2015; Sharma et al., 2020; Weinberger & Fischer, 2006).

To begin with, the popular Russel's Core Affect Framework (Russell, 2003) states that an affect has a valence as well as an arousal component. Based on this widely adopted framework, negative valence and moderate to high levels of arousal are often linked with confusion and frustration, respectively, whereas positive valence and high arousal are indicative of excitement (R. S. Baker et al., 2010; Sharma et al., 2020). Inspired by this we consider four features (*Positive Valence*, *Negative Valence*, *Difference in Valence* and *Arousal*) related to the emotional state of the team. The feature *Difference in Valence* is of interest as it immediately highlights that a team with a higher value has a positive emotional state. Please note that in this work, we do not distinguish between conceptual confusion and frustration as it is not straightforward to separate these accurately on the basis of the values of valence and arousal alone. For these reasons, we use the terms interchangeably when discussing our findings.

Similarly, gaze patterns have often been analyzed to gauge the attention of learners in collaborative settings (Schneider et al., 2016; Sharma et al., 2021). Therefore, here we extract the attention of the team to various parts of the screen, their partner, and the robot. Furthermore, in collaborative settings, speech measures have been widely used to measure the dynamics

of the collaboration between the team members (Bassiou et al., 2016; Martinez et al., 2011; Viswanathan & VanLehn, 2017). We make use of several of these speech measures such as *Speech Activity*, *Short Pauses*, *Long Pauses*, *Speech Overlap* and *Overlap_to_Speech_Ratio* to capture talk time, and heterogeneity of verbal participation. The length of the *Short Pauses* and *Long Pauses* is based on findings from Campione and Véronis, 2002 that, when analysing pauses in various languages, found that pauses seem to support a categorization into brief (< 200 ms), medium (200-1000 ms), and long (>1000 ms) pauses. This is also echoed by the work of Heldner and Edlund, 2010.

When it comes to interaction with a learning activity, log data such as frequency of actions has been used as an approximation for various constructs such as attention, engagement, interest, exploration, etc (Martinez-Maldonado et al., 2013a; Popov et al., 2017; Viswanathan & VanLehn, 2017). With our activity, we make use of frequency of actions such as additions, deletions, redundant edges on the map (*T_add*, *T_remove*, *T_ratio_add_del*, *Redundant_exist*). Furthermore, we are interested in actions or patterns that can indicate reflection. Consulting previously explored solutions is an indicator of reflection (Veenman, 2013). In addition to that, certain action patterns can also be indicative of reflecting on self or partner's actions. Hence, we also consider such behaviors of looking at past solutions (*T_hist*), and correcting one's own or partner's actions on the go (*T1_T1_add*, *T1_T2_add*, *T1_T1_delete*, and *T1_T2_delete*) as indicators of reflection. Please note that we use *T_help* as an indicator of *usability confusion*, i.e., confusion with regards to the user interface and not as an indicator of conceptual confusion that has been discussed previously.

In addition to these behaviors, we make use of one performance metric *last_error* and several types of learning gains that have already been defined in the thesis. For convenience of the reader, we reiterate here. The performance metric gives the error of the last submitted solution where error can be defined as the difference between the cost of a submitted solution and the cost of an optimal/correct solution (optimal cost), normalised by the optimal cost (for an optimal solution, error will then be 0). Further, the three types of learning gains *absolute*, *relative*, and *joint absolute* learning gains that, respectively, measure how much the participant learned of all the knowledge available, how much the participant learned of the knowledge he/she did not possess before the activity, and the amount of knowledge acquired together by the team members during the activity. The team level values for the first two learning gains are calculated by taking the average of the individual learner values. It is important to mention that we distinguish between performance and learning such that performance measures the success/failure in the task itself via *last_error* whereas learning (*absolute*, *relative*, and *joint absolute*) measures the amount of knowledge gained during the interaction via a pre- and a post-test. Both the tests are composed of 10 multiple-choice questions assessing various concepts for the minimum spanning tree problem. The three types of learning gains are plotted versus the last error for all 32 teams in Figure 4.1.

Table 4.1: Multi-modal features that represent behaviors and constructs

Construct	Marker	Behavior
		Log Features
Exploration	T_add	The number of times a team added an edge on the map
Exploration	T_ratio_add_rem	The ratio of addition of edges over deletion of edges by a team
Exploration	T_action	The total number of actions taken by a team (add, delete, submit, presses on the screen)
Exploration	Redundant_exist	The number of times they had redundant edges in their map
Reflection (Metacognition)	T_remove	The number of times a team removed an edge from the map
Reflection (Metacognition)	T_hist	The number of times a team opened the sub-window with history of their previous solutions
Reflection (Metacognition)	T1_T1_add	The number of times a team, either member, followed the pattern consecutively: I delete, I add back
Reflection (Metacognition)	T1_T1_delete	The number of times a team, either member, followed the pattern consecutively: I add, I then delete
Reflection (Metacognition)	T1_T2_add	The number of times a team, either member, followed the pattern consecutively: I delete, You add back
Reflection (Metacognition)	T1_T2_delete	The number of times a team, either member, followed the pattern consecutively: I add, You then delete
Usability Confusion	T_help	The number of times a team opened the instructions manual
		Video Features: Affective states and Gaze
Emotional State	Positive Valence	The average value of positive valence for the team
Emotional State	Negative Valence	The average value of negative valence for the team
Emotional State	Difference in Valence	The difference of the average value of positive and negative valence for the team
Emotional State	Arousal	The average value of arousal for the team
Emotional State	Smile	The average percentage of time of a team smiling
Attention	Gaze at Partner	The average percentage of time a team has a team member looking at their partner
Attention	Gaze at Robot	The average percentage of time a team is looking at the robot
Attention	Gaze (Other)	The average percentage of time a team is looking in the direction opposite to the robot
Attention	Gaze at Screen_Left	The average percentage of time a team is looking at the left side of the screen
Attention	Gaze at Screen_Right	The average percentage of time a team is looking at the right side of the screen
Attention	Gaze Ratio of Screen_Right and Screen_Left	The ratio of looking at the right side of the screen over the left side
		Audio Features: Speech
Communication	Speech Activity	The average percentage of time a team is speaking over the entire duration of the task
Communication	Silence	The average percentage of time a team is silent over the entire duration of the task
Communication	Short Pauses	The average percentage of time a team pauses briefly (0.15 sec) over their speech activity
Communication	Long Pauses	The average percentage of time a team makes long pauses (1.5 sec) over their speech activity
Communication	Speech Overlap	The average percentage of time the speech of the team members overlaps over the entire duration of the task
Communication	Overlap to Speech Ratio	The ratio of the speech overlap over the speech activity

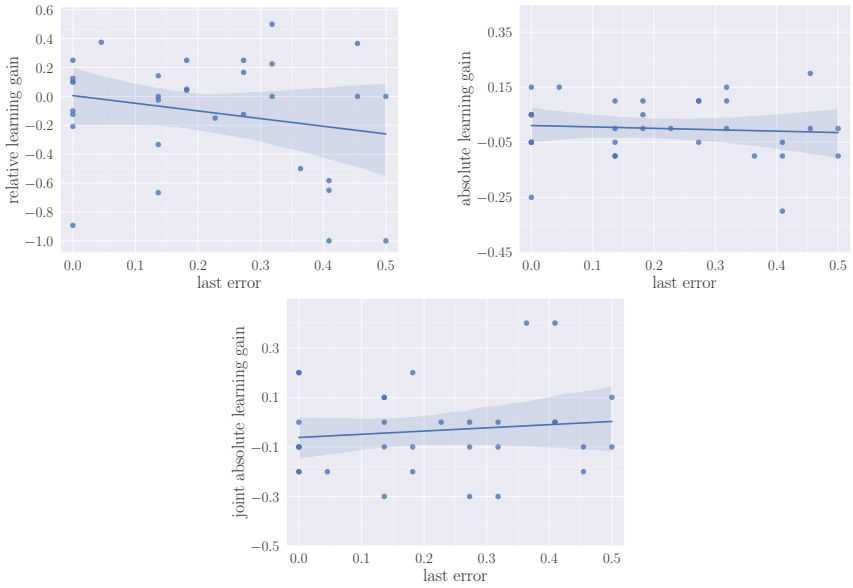


Figure 4.1: Learning gains vs performance. All values here are non-normalized.

This dataset, with multi-modal behaviors as well as performance metric and learning gains, has been publicly made available (Nasir, Norman, Bruno, Chetouani, et al., 2021). It must be noted that, in our dataset, all behaviors are treated as averages and frequencies over the entire duration of the task and that it is not temporal data. The average value for the team for various behaviors is calculated by taking an average of the individual behaviors by each team member. Furthermore, for all the behaviors, data has been normalized across the teams with each behavior having values between 0 and 1. This would mean that a value of 0 would be the lowest value of a behavior across all teams. Similarly, a value of 1 would be the highest value of a behavior across all teams. With respect to our work in Chapter 3 (Nasir, Bruno, Chetouani, et al., 2021), for the analysis in this chapter, we made slight changes to two of the behaviors *Short Pauses* and *Long Pauses* in the original dataset (Nasir et al., 2020a). Originally, the two pause behaviors were not normalized with respect to the teams speech activity over the interaction. The change is motivated by the belief that normalizing the pause time gives a more accurate measure.

4.3.2 Analysis Approach

The goal of this work is to build and understand comprehensive multi-modal profiles of dyads who learn and those who don't as they work on JUSThink. To this end, we developed an analysis approach consisting of two parts: a quantitative approach and a qualitative approach. The quantitative approach is a learning analytics technique that we extend, and formalize here, from our forward and backward approach first presented in the previous chapter (Nasir, Bruno, Chetouani, et al., 2021). It helps identify groups of learners who have learning gains and those who don't. Using this approach we are able to build the multi-modal behavioral

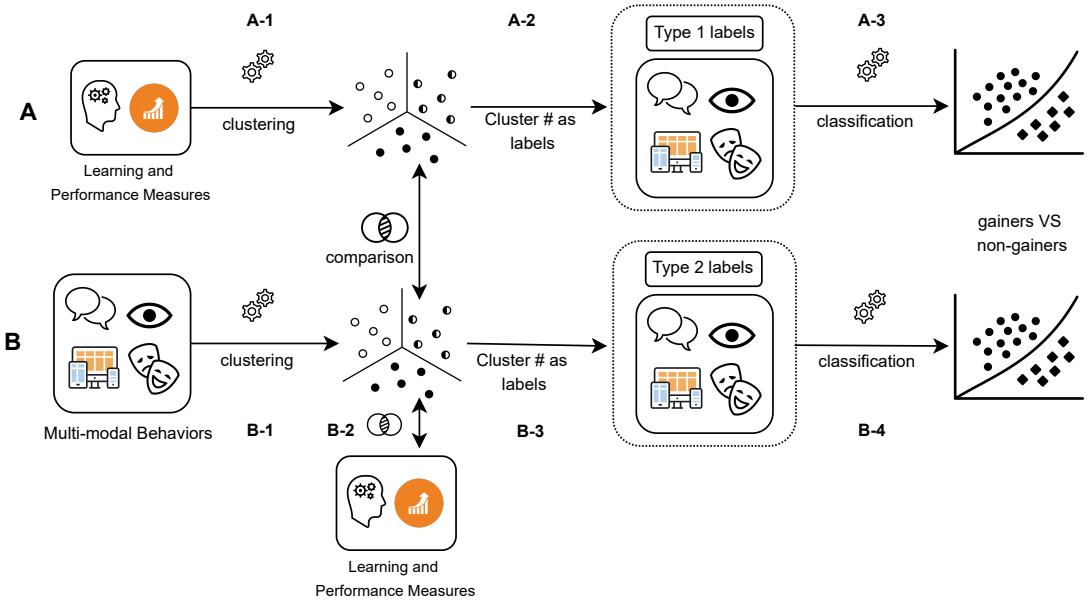


Figure 4.2: Overview of our technique in Nasir, Bruno, and Dillenbourg, 2020.

profiles for each group of learners. The goal of the qualitative approach is to allow us to better interpret the multi-modal profiles and understand the learning mechanisms at play within each group of learners previously identified. We do this by interaction analysis of cases wherein we study the multi-modal behaviours from dyads within specific episodes of activity in each group of learners. We then unpack the likely multi-modal learning mechanisms at play. The choice of episodes will be based upon the findings of the quantitative approach; specifically we will focus on episodes where certain behaviours of interest identified in the quantitative approach are highlighted. Further details are in 4.3.2.

Multi-modal Learning Analytics

The technique is visually presented in Figure 4.2. It consists of two approaches: 1) *approach A*, which can be considered as a backward approach as it connects the learning outcomes back to the behaviors observed during the learning process and 2) *approach B*, that can be considered as a forward approach as it helps to move from multi-modal behaviors to learning outcomes. For the remainder of our thesis, we will generally refer to them as *approach A* and *approach B*. This technique adopts a data-driven approach to identify labels linking behavioural profiles and learning. It must be noted that in our work, we use the term *gainers* to refer to learners who end up having learning gains while the term *non-gainers* refers to learners that do not have positive learning gains.

Approach A:

This approach starts with clustering on the learning as well as performance metrics of the teams as shown in step A-1. We then use these cluster labels as the ground truth for a classifier trained on multi-modal behaviors of the learners as shown in A-2 and A-3. This approach has been applied within learning analytics to identify the behavioural profiles of gainers vs non-gainers or high vs low performers (Kinnebrew et al., 2013; Worsley & Blikstein, 2011). In our case, the clustering reveals four clusters (more details on the four clusters in appendix A). Then, as a step towards building profiles, we perform a Kruskal-Wallis analysis on each pair of clusters to identify the significantly discriminating behaviors between each pair. However, we observe no significantly discriminating behavior between each pair. Such analysis on the four clusters from *approach A* can raise a misunderstanding that all learners, irrespective of learning or performance, exhibit similar multi-modal behaviors. It must be noted however, that this approach assumes *by design* that each of the learning and performance profiles (given by one cluster) is associated with a unique set of behaviors. However, what if teams with similar learning and performance actually exhibit two or more different sets of behaviors? This is the motivation for adopting *Approach B*, which represents a perspective shift to take such a possibility into account.

Approach B:

As depicted in step B-1 (Figure 4.2), this approach begins with clustering the teams based on their multi-modal behaviors in order to identify the different behavioural profiles existing within the data. We then compare these behavioral clusters in terms of the learning gains and performance metric of the teams (B-2) in order to identify differences between the behavioural profiles in terms of their learning and performance. This is followed by comparing the clusters obtained in both approaches with respect to the teams they consist of. If, as we hypothesized, there are indeed multiple sets of behaviors associated with learning then 1) we should observe significant differences among some of the approach B clusters in regards to their learning gains (requirement 1) as well as 2) there should to be a one-to-many or many-to-many comprehensible mapping between the clusters from both approaches (requirement 2). This second requirement would mean that approach B provides us with distinct variants of behavioral profiles for the same learning profiles. To reiterate, while requirement 1 highlights that indeed gainers and non-gainers have different behaviors, requirement 2 is necessary to validate the existence of multiple behavioral profiles for the same type of learning profile, which is the motivation behind this approach. If the two requirements are met, the cluster labels from approach B can be employed as ground truth for a classifier as shown in steps B-3 and B-4. Thus this approach allows us to unearth the differences that exist within both learning and behaviour data, and align them to create multiple learning profiles.

The classification results are reported in the appendix A. As we obtained excellent classification results from *approach B*, in our work, we focus in-depth on building behavioral profiles from the clusters resulting from this approach. As shown in Table 4.2, *Approach B* gives 3 behavioral clusters with the first two exhibiting high learning and the third lower learning; hence, with

respect to learning, the groups can be named as *type 1 gainers*, *type 2 gainers*, and *non-gainers*, respectively. Note that the performance (*Last_Error*) in the task is very similar for each group. As done with the *Approach A* clusters, we now proceed to compare the resulting clusters in terms of their multi-modal behaviors by first performing a variance analysis on the three clusters obtained in *approach B* and then by performing a Kruskal-Wallis analysis on each pair of clusters to identify the significantly discriminating behaviors between them. Indeed, we observe several discriminating behaviors between each pair, and so with respect to these behaviours that will be seen in more detail in the upcoming sections, we name the groups of *type 1 gainers*, *type 2 gainers*, and *non-gainers* as *Expressive Explorers*, *Calm Tinkerers*, and *Silent Wanderers*, respectively. From this point on, we will use the two types of names interchangeably. Based on these quantitative findings, we then qualitatively analyse each group of learners as described in the following subsection.

Interaction Analysis of Multi-modal Cases

In order to better interpret the multi-modal behavioural profiles identified above and elaborate the likely learning mechanisms occurring in each group of learners, we qualitatively analyse a learning episode from each group. To do the analysis, we select episodes when a "behaviour of interest" is high. The exact behaviour of interest will depend on the results of the quantitative analysis which we described in Section 4.3.2, but the rationale is to unpack a behaviour which discriminates students who learn from those who don't, and whose effect on learning is not straightforwardly understood. We analyse three episodes, one each from one randomly selected dyad belonging to a different group of learners. As our quantitative results aggregate behaviours over the entire activity, these cases are meant to be illustrative of the likely underlying learning mechanisms during certain episodes when a behaviour of interest is high and so we choose a random dyad from each group in order not to bias this illustration.

We begin by extracting the dialogue of the learners during this episode. The full transcripts can be found in the publicly available JUSThink dialogue and actions corpus by Norman, Dinkar, Nasir, et al., 2021. This corpus relies on manual transcription, due to the poor performance of state-of-the-art automatic speech recognition systems on this dataset which consists of children's speech with music playing in the background. A graduate student completed two passes on each transcript, which were then checked by another native English speaking graduate student with experience in transcription/annotation tasks. We augment the dialogue transcript with average values of other behaviours during this episode to build a multi-modal transcript. We then interleave the dialogue, action and affective states to unpack how learning is happening within each episode. We perform interaction analysis of each episode with the analytic focus of turn-taking. The goal is to understand how turn-taking leads to learning during the episode, specifically the relationship between the content of the speech, the actions, the affect of the learners and their learning outcomes. Thus, with both the quantitative and qualitative methods aforementioned, we make an attempt to answer the question, what do learners' visible behaviours reveal about how learning happens in a collaborative constructivist

Table 4.2: The three clusters in approach B with mean values for learning gains (LG) as well as the last error. The significantly different learning gains are represented in bold.

Cluster name	Last_Error	Absolute_LG	Relative_LG	Joint_Absolute_LG	N
		Approach B			
Expressive Explorers	0.461	0.678*	0.693*	0.714*	14
Calm Tinkerers	0.393	0.616	0.604	0.607	12
Silent Wanderers	0.393	0.383*	0.348*	0.428*	6

learning activity.

4.4 Results

4.4.1 Pairwise Significantly Distinct Behaviors

From Figure 4.3, we observe that the behaviours with the highest variance among the three clusters come from all three modalities pertaining to log, speech and affective features. Overall, it is clear that the manner in which each of the group interacted with the task (*T1_T2_add*, *T_remove*, *T_action*) is unique. Speech behavior (*speech_overlap*, *speech_activity*, *silence*, *overlap_to_speech_ratio*), on the other hand, is similar in the two gainer groups but very different from the *Silent Wanderers*. An interesting observation with regards to the affective features (*negative_valence*, *arousal*) is that the *Silent Wanderers* exhibit very similar arousal and negative valence behaviors to *Expressive Explorers* and both these groups differ from the *Calm Tinkerers* in this behavior. We elaborate on the differences between each pair of groups below. Note that in the upcoming figures, for the ease of comprehension, each modality is represented by a unique pattern and each behavior within a modality by several shades of the same color.

Expressive Explorers and Silent Wanderers

Figure 4.4 shows the features that are significantly distinct between *Expressive Explorers* and *Silent Wanderers*. The corresponding p-values are listed in Table 4.3. Concerning log features, we observe that *Expressive Explorers*, relative to *Silent Wanderers*, do significantly fewer actions of the sort where one team member deletes an edge and the other adds it back (*T1_T2_add*). At the same time, they look at their previous solutions (*T_hist*) significantly more than *Silent Wanderers*. This suggests that *Expressive Explorers* perform more global reflection, i.e., reflection on their previously constructed solutions, while *Silent Wanderers* do more local reflection, i.e., reflection on their most recent actions.

Apart from this difference, the two groups are also significantly different in their speech behavior. *Expressive Explorers* not only speak more between themselves (*Speech Activity*), but

Table 4.3: p-values for the Kruskal-Wallis analysis on each pair with significance level of 0.05

Markers	Expressive Explorers and Silent Wanderers	Calm Tinkerers and Silent Wanderers	Expressive Explorers and Calm Tinkerers
Log Features			
T_add	0.14	0.85	0.03*
T_remove	0.13	0.42	0.00*
T_ratio_add_rem	0.05*	0.16	0.00*
T_action	0.07	0.37	0.00*
T_hist	0.04*	0.60	0.01*
T_help	0.45	0.14	0.46
T1_T1_remove	0.50	0.03*	0.00*
T1_T1_add	0.92	0.73	0.76
T1_T2_remove	0.80	0.39	0.04*
T1_T2_add	0.01*	0.19	0.00*
Redundant_exist	0.07	0.00*	0.83
Video Features: Affective states and Gaze			
Positive Valence	0.74	0.07	0.00*
Negative Valence	0.80	0.00*	0.00*
Difference in Valence	0.62	0.22	0.16
Arousal	0.93	0.00*	0.00*
Smile	0.93	0.05*	0.01*
Gaze at Partner	0.28	0.45	0.12
Gaze at Screen_Left	0.11	0.01*	0.23
Gaze at Screen_Right	0.02*	0.22	0.53
Gaze Ratio of Screen_Right and Screen_Left	0.28	0.45	0.83
Gaze at Robot	0.50	0.22	0.16
Gaze (Other)	0.45	0.16	0.04*
Audio Features: Speech			
Speech Activity	0.00*	0.00*	0.71
Silence	0.00*	0.00*	0.71
Short Pauses	0.04*	0.16	0.23
Long Pauses	0.01*	0.01*	0.60
Speech Overlap	0.00*	0.00*	0.68
Overlap to Speech Ratio	0.00*	0.00*	1.00

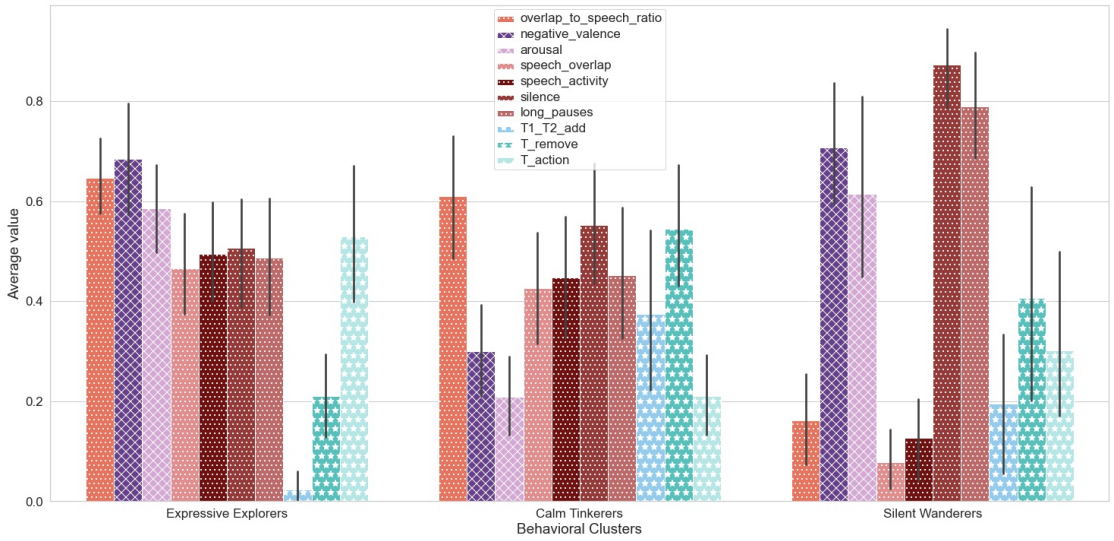


Figure 4.3: Features with highest variance between all three behavioral clusters. For the ease of comprehension, each modality is represented by a unique pattern and each behavior within a modality by several shades of the same color.

also have lower number of short and long pauses (*Short Pauses, Long Pauses*) when they speak and a higher degree of overlap (*Speech Overlap, Overlap_to_Speech_Ratio*) when interacting. Finally the two groups show no significant difference in their affective features, as seen by the fact that both *Expressive Explorers* and *Silent Wanderers* displayed very similar valence and arousal behaviors (which is high *arousal* and high *negative valence*).

Calm Tinkerers and Silent Wanderers

Looking at the significantly distinct behaviors between *Calm Tinkerers* and *Silent Wanderers* (see Figure 4.5 and Table 4.3 for the p-values of the KW tests), we observe that the differences lie in the way in which they interact with the task itself, their speech behavior and also their affective features. Relative to *Silent Wanderers*, the *Calm Tinkerers* do more of local reflective actions, where a team member adds an edge and then removes it right after (*T1_T1_rem*). Moreover, while *Calm Tinkerers* carefully minimize the number of redundant edges (i.e., two alternative paths connecting location A with location B) present at any time on their map in the task, *Silent Wanderers* allow for such redundancies to be present on the map significantly more.

In terms of their speech behavior, *Calm Tinkerers* have higher speech activity (*Speech Activity*), lower number of long pauses (*Long Pauses*) and higher speech overlap (*Speech Overlap, Overlap_to_Speech_Ratio*) than *Silent Wanderers* (who are non-gainers in terms of learning). It is important to remark that the same difference was observed between *Expressive Explorers* and *Silent Wanderers*, thus suggesting that speech behaviours can allow for distinguishing gainers from non gainers. Lastly, this group of gainers displays significantly lower negative

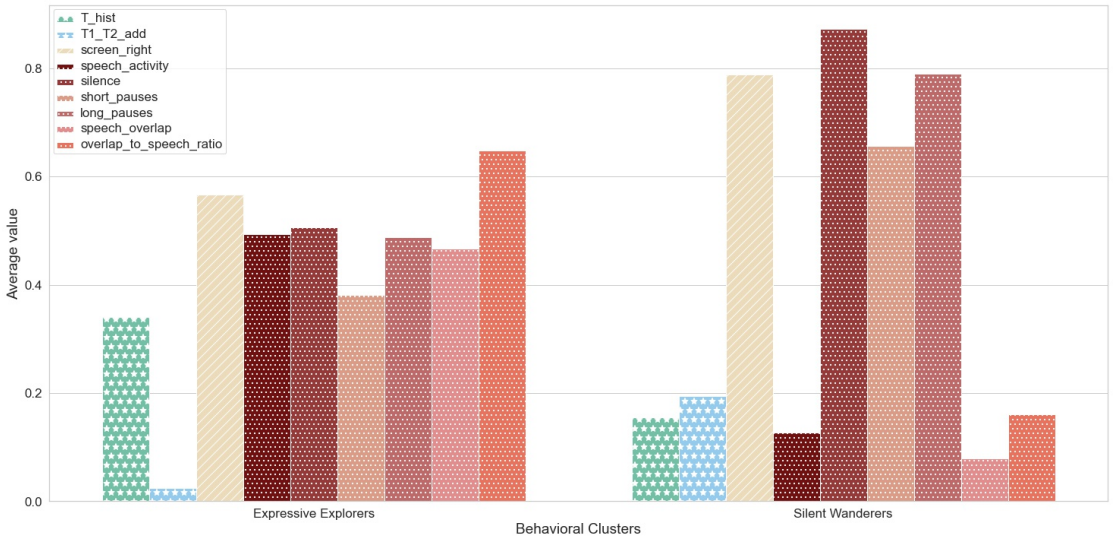


Figure 4.4: Significantly distinctive features between the *Expressive Explorers* and the *Silent Wanderers*.

valence and arousal (*negative_valence*, *arousal*) compared to the *Silent Wanderers*, indicating that *Calm Tinkerers* are relatively calmer.

Expressive Explorers and Calm Tinkerers

Lastly, we compare the significantly distinct behaviors between two types of gainers (see Figure 4.6 and Table 4.3 for the p-values of the KW tests). We observe that the two groups of gainers significantly differ in most of their log behaviors. If we look closely at these behaviors, we observe that *Expressive Explorers* do more actions (*T_action*) in general, specifically doing more edge additions (*T_add*) and, consequently, displaying a higher ratio of adding to deleting edges (*T_ratio_add_rem*). Furthermore, they open their history significantly more times (*T_hist*). *Calm Tinkerers*, on the other hand, have more deletion actions (*T_remove*) and a higher number of addition-deletion action patterns of the type *T1_T1_rem*, *T1_T2_rem* and *T1_T2_add*. These findings suggest that *Expressive Explorers* enact a global exploratory approach characterized by global reflection on previous solutions while *Calm Tinkerers* exhibit a local exploratory approach where they carry out in-the-moment reflection and correct their own and their partners' actions on the go, which can be described as local reflection. For example, *Expressive Explorers* successively add edges on the map and then look at the cost effectiveness of their constructed map by comparing it with their past solutions, while *Calm Tinkerers* show a pattern of adding an edge and then deleting it right after or vice versa which may be triggered due to reflection. A specific example will follow in the case studies discussed in section 4.4.2.

Moreover, *Expressive Explorers* have higher average values of *valence* and *arousal* compared to the *Calm Tinkerers*, suggesting that they were more expressive in their interactions. These

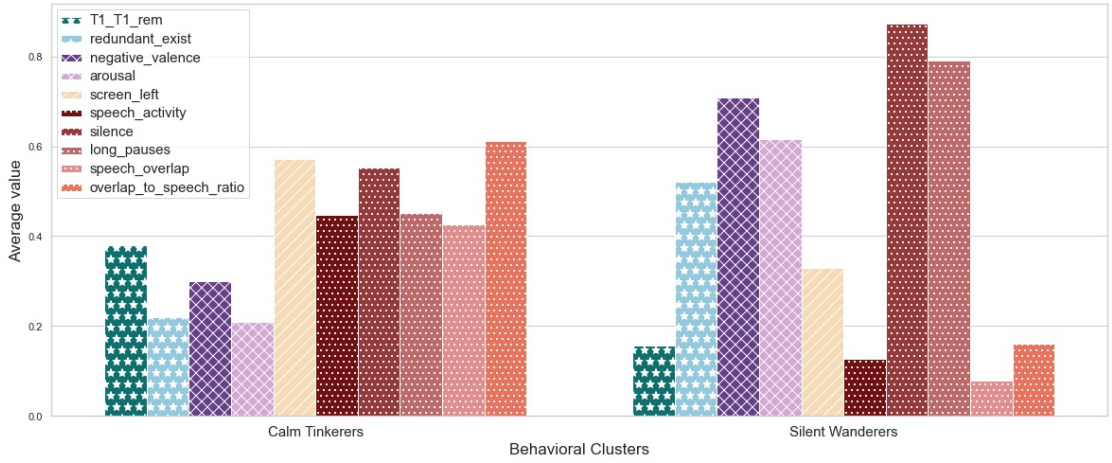


Figure 4.5: Significantly distinctive features between the *Calm Tinkerers* and the *Silent Wanderers*.

results show that gainers can exhibit a frustrated profile or a calm one. Lastly, notice how none of the speech behaviors is significantly different between the two types of gainers, once again pointing to the fact that gainers, irrespective of their other behaviors, all had a similar speech behavior quantitatively.

4.4.2 Interaction Analysis of Multi-Modal Cases

As shown above, *speech overlap* is a behaviour which distinguishes *Silent Wanderers* (who do not overlap with one another as much) from both types of gainers (who overlap significantly more). Specifically, our results suggest that a high amount of overlapping speech can be more productive for learning relative to when there is less speech overlap. It has been reported in literature (Bassiou et al., 2016) that speech overlap is one of the speech features that distinguishes the quality of collaboration. However, literature also suggests that the frequency of overlaps is negatively correlated with collaboration in children (Kim et al., 2015). Given these contradictory findings on the role of overlapping speech in collaborative learning, we consider "speech overlap" as a behaviour of interest for qualitative analysis. We seek to understand the nature of overlapping speech and turn-taking during the task. Specifically, for one randomly selected team from each group of learners, we pick a chunk of dialogue of a few seconds, that corresponds to the first time a team reaches the highest level of speech overlap to speech ratio (*overlap_to_speech_ratio*) consecutively for the whole duration of the chunk. We report below the dialogues taking place between the team members, along with the averages of their actions and affect in this duration. The blue and red colored rectangles, in the upcoming figures highlighting dialogue, indicate the duration in which learner A and B are speaking, respectively; hence, highlighting speech overlap when the rectangles overlap. The start and the end time for the dialogues (in seconds) are also indicated in the figures. Right next to the dialogues, in these figures, we also report other behaviors for each chunk. Our temporal data

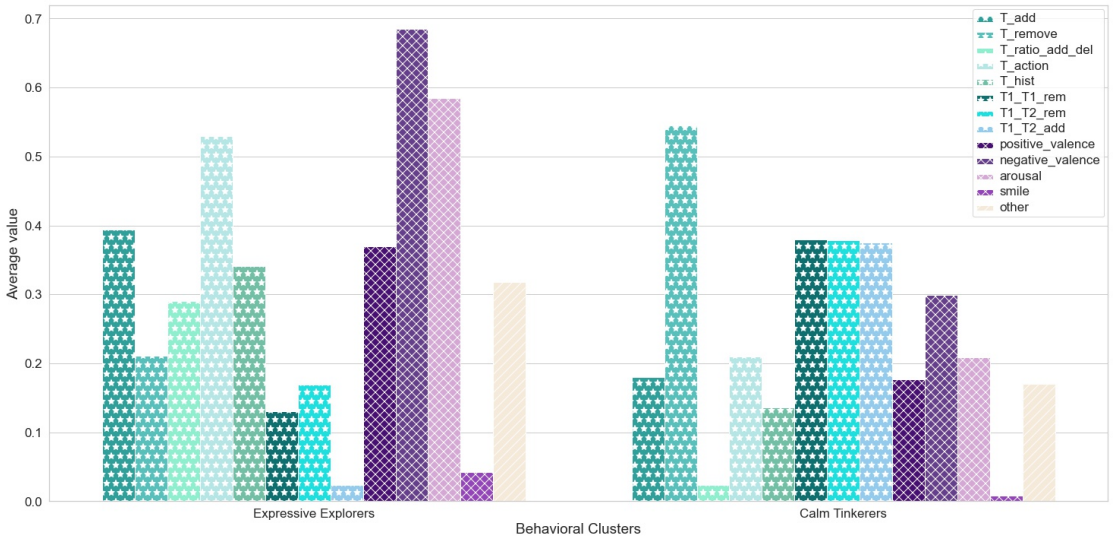


Figure 4.6: Significantly distinctive features between the two type of gainers.

for this qualitative analysis is organized in 10 second windows. We use the values in these windows to report both the average of these behaviors over the *entire interaction* and *within the chosen chunks* (that range from 30-60 seconds), one for each team in the case studies. Note that we do not include gaze behaviors as gaze was not found to be a significant behavior in our quantitative analysis.

Episode from Expressive Explorers

This dialogue excerpt, shown in Figure 4.8, occurs right after the team submitted a solution and were informed by the robot that it is not the optimal solution yet. Hence, what the participants see on their screens at the time when this dialogue starts is an empty map, as shown in Figure 4.7, i.e. a map that has been cleared after submitting a solution. The team can now start building a new solution on this empty map.

We observe that both team members interject each other. However, the content of the dialogue builds on their partner's conceptual ideas, which is indicative of the emergence of novel solution ideas. The high speech overlap is thus not caused by a lack of collaboration but a high degree of understanding between the team members, owing to which they are "completing each other's sentences". In addition we observe that the average values of arousal and negative valence during this exchange are lower (0.22 and 0.20 respectively) than the average values (0.34 and 0.28) of this team over the entire task, suggesting a shift towards low arousal states such as "neutral", "boredom" or even "sadness" right after hearing feedback on their solution. This is interesting because *Expressive Explorers* exhibit a higher level of frustration overall.

Now looking at the log actions of this team during this chunk with respect to the whole task, we observe that the team employs a more global exploration strategy with an increase in both

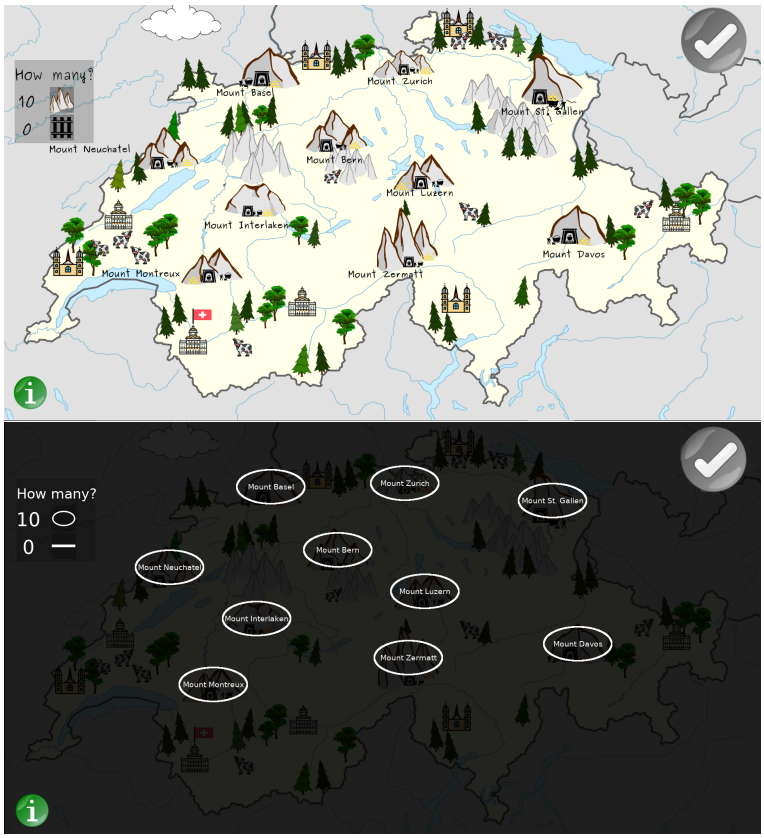
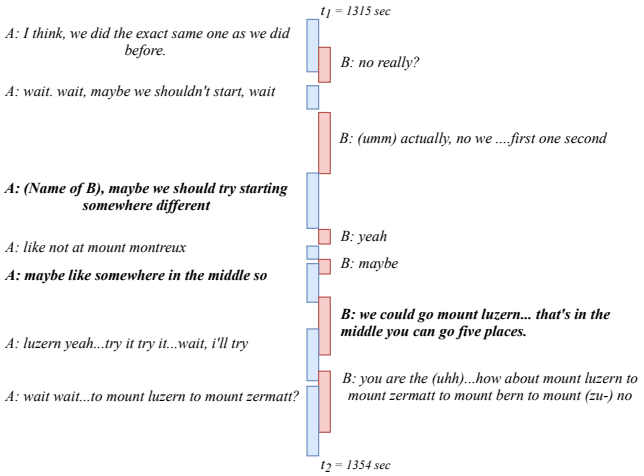


Figure 4.7: The two views of the JUSThink game, namely *figurative* and *abstract*, as shown on the screens of the participants when they are empty.

addition actions and reflection in terms of looking more at their history. As seen in the bold section of the dialogue, the team reassesses the foundations on their approach and revises it. Further, looking at the ratio between additions and deletions actions in this chunk versus over the whole task, we note that the team is only doing additions. This maybe because in this time the team is starting from an empty map and building a new solution. Connecting these observations back to the overall solution strategies of these types of gainers, this episode provides deeper multi-modal insights for how these type of gainers learn through a more global exploratory approach and reflection on their overall solution strategy.

Episode from Calm Tinkerers

This snippet of dialogue, shown in Figure 4.10, from a random team of *Calm Tinkerers* occurs approximately one minute after they submitted their first solution and were told it is not the optimal solution yet. The two views on their respective screens, at the time this dialogue starts, are as shown in Figure 4.9.



	Average over entire interaction	Average over the chunk
Arousal	0.34	0.22
Negative Valence	0.28	0.20
Addition	0.22	0.4
Deletions	0.13	0
History	0.09	0.12
Local Reflective actions	0	0

Figure 4.8: The dialogue for an *Expressive Explorers* team where the blue and red rectangles indicate the duration in which learner A and B are speaking, respectively. Speech overlap is indicated by the overlapping rectangles. Other relevant log and affective features are also shown in a parallel table.

In this excerpt, the team members are attempting to optimize the solution by adding a particular edge ("Bern to Interlaken") to the solution. Firstly, when both team members agree upon the overall strategy, they both speak over each other to complete the steps to be taken towards the solution. Secondly, when there is disagreement about the next action, there is a high overlap of speech; however the dialogue leads to an agreement on the action to be taken. Thus the high degree of overlap seems to be related to these cycles of proposal-negation-agreement, which could be one mechanism by which the locally reflective problem solving strategy is manifested in this group of learners. Indeed, as the dialogue shows, the team members immediately reflect and correct each others actions. This is a sign of negotiation that is inherent in a collaborative problem solving session and that leads to mutual understanding of the solution space.

Zooming into the teams' affective state during this exchange, we find that the average arousal and negative valence in this chunk was 0.38 and 0.30 respectively which is higher than the team's average arousal and negative valence (0.32 and 0.23 respectively) over the entirety of the interaction. This indicates that during this period of high speech overlap, the team was in a higher state of arousal, which could possibly indicate a state of disequilibrium as suggested by previous research (D'Mello & Graesser, 2012; Lodge et al., 2018). Recall that *Calm Tinkerers* overall exhibit lesser frustration than the other two types of learners.

In terms of actions, we see that in this chunk the teams' ratio of deletions to additions is higher than over their entire interaction. This could be because by this time the team had already added several edges towards a potential solution and were deleting edges through the negotiation and optimization process seen in the dialogue above. Further we see that none of the other actions signifying reflection, such as looking at their history or deleting

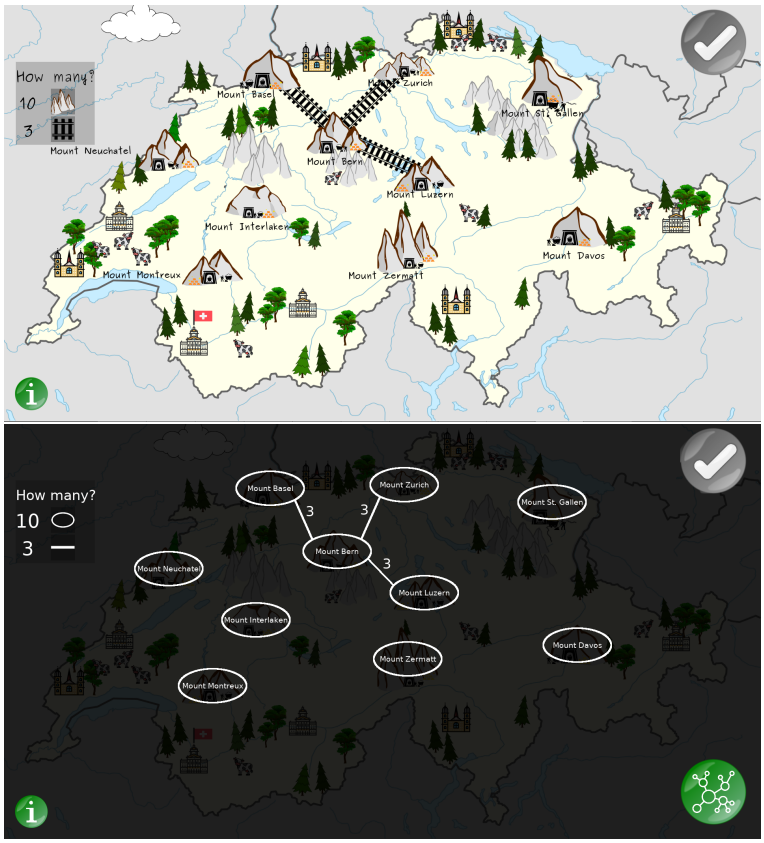


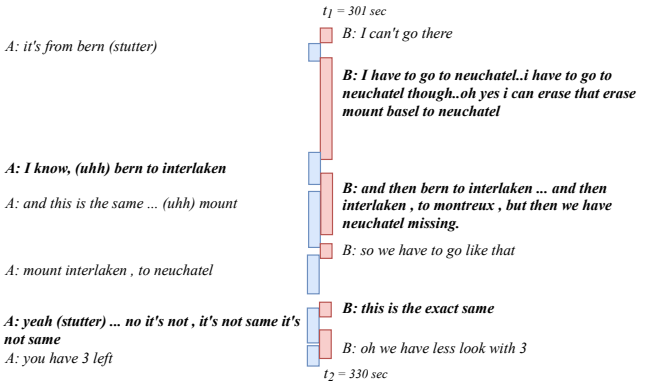
Figure 4.9: The two views of the JUSThink game, namely *figurative* and *abstract*, as shown on the screens of the participants from a team belonging to the group of *Calm Tinkerers*.

their own or their partners edges is seen here. Connecting back to *Calm Tinkerers* overall solution strategy, we see that this chunk demonstrates how these teams learn through a local exploration strategy of additions and deletions, rather than reflecting on overall strategy.

Episode from Silent Wanderers

This dialogue, shown in Figure 4.11, takes place within a team of *Silent Wanderers* right after they submitted a solution and were told by the robot that it is not optimal. Hence, when the dialogue starts, the screens for this team also shows empty maps, as in Figure 4.7. The team can now start building a new solution.

In this dialogue we observe that the team first agrees on the goal to achieve (a solution cost of 24). However the initial idea put forth by a team member (B) is not taken up by A, which leads to a cycle of proposal-negation-agreement. The negotiation between the team takes longer compared to *Calm Tinkerers* but they eventually come to an agreement about the first action to take while building a new solution (connecting “Basel to Zurich”). During this negotiation the team members speak over each other, as also seen with the *Calm Tinkerers* and it indicates



	Average over entire interaction	Average over the chunk
Arousal	0.32	0.38
Negative Valence	0.23	0.30
Addition	0.31	0.25
Deletions	0.12	0.12
History	0.10	0
Local Reflective actions	0.75	0

Figure 4.10: The dialogue for a *Calm Tinkerers* team.

	Average over entire interaction	Average over the chunk
Arousal	0.51	0.50
Negative Valence	0.37	0.40
Addition	0.24	0.2
Deletions	0.12	0
History	0.07	0.07
Local Reflective actions	0	0

Figure 4.11: The dialogue for a non-gainer team of *Silent Wanderers*.

constructive collaboration because this non-gainer team also reaches an agreement on a path forward to the solution. Overall, however, as seen from our quantitative analysis, the duration of such speech overlap is significantly lesser in the *Silent Wanderers*.

It is interesting to note that the arousal of this non-gainer team sees an increase followed by a dip (ranging from 0.57 to 0.45) during this exchange, compared to the teams' average arousal of 0.51. The dip in arousal during that occurs right after getting the feedback on their solution (which is very far from the optimal) suggests a tendency towards low arousal emotions such as “neutral", “sadness" or “boredom". On average, however, this teams arousal and negative valence over this chunk (0.5 and 0.4 respectively) is similar to their arousal and negative valence over the entire task (0.51 and 0.37 respectively). We recall that non-gainer teams on average exhibit higher frustration than the gainer teams.

In terms of actions, we observe that in this chunk where they begin from an empty map the team performs only additions and no deletions as they try to negotiate and build a better solution. Further their reflection actions (looking at history and deleting their own or their partners actions) are similar to their reflection actions across the entire task. Recall that non-gainer teams on average do fewer reflective actions of any type.

To summarize, while the non-gainer team, similar to the two gainer teams, also exhibits constructive communication during an episode of high speech overlap, they do not demonstrate any change in their reflective actions during this chunk right after a “failure” or any change in their affective states. Further, they have significantly lesser duration of such speech overlap over the entire task duration compared to both types of gainers. This, along with the fact that they have fewer reflective actions overall, could be a reason for their learning process not being as effective as the gainers.

4.5 Discussion

The goal of this paper is to build a multi-modal understanding of learning vs non-learning as it happens in a collaborative open-ended activity. Our combined multi-modal learning analytics and interaction analysis methodology enabled us to identify two multi-modal profiles of learners who have learning gains and one multi-modal profile of learners who do not have learning gains. Now that we have quantitatively compared the profiles pair-wise in section 4.4.1 and qualitatively compared three teams, one from each profile, separately in section 4.4.2, in this section, we begin by discussing each of the three profiles of learners with respect to each modality. Next, we discuss how multi-modality furthers our understanding of collaborative learning and how the outcomes contribute to designing effective interventions in similar computer-supported collaborative learning (CSCL) settings.

4.5.1 Speech Behaviors

In terms of speech behaviours, both types of gainers exhibit a very similar behavior quantitatively, that is significantly different from the one displayed by the *Silent Wanderers*. The same is true for other speech behaviors including speech overlap between team members, the overlap to speech activity ratio, and short and long pauses over the entire speech activity. Overall we find that there is a lot more verbal interaction within the teams that end up with higher learning gains, as observed in previous research on collaborative learning (Bassiou et al., 2016; Praharaj et al., 2021; Weinberger & Fischer, 2006). This is not surprising because the nature of the collaborative activity requires the learners to communicate, share information and build a common ground to construct a solution (Barron, 2003; Roschelle & Teasley, 1995). As we highlighted in the episodes of high speech overlap dialogue, we observe two mechanisms of verbal interaction that support collaborative learning. In one case, the dyad demonstrates a high degree of transactivity, which is known to be good for learning (Teasley, 1997). This is seen by completion of each others' sentences and the speech overlap is a way to align on their plan for solution building. In the other two cases, we observe proposal-negation-agreement cycles (Barron, 2003; Roschelle, 1992) in the team members' dialogue during these periods of high speech overlap, indicating that the process of proposal discussion and uptake was happening, which is also indicative of good collaboration (Barron, 2003). Hence, contrary to the literature that suggests that the frequency of overlaps is negatively correlated with collaboration in children Kim et al., 2015, speech overlap in children seems to be an indicator of the negotiation that is inherent in the collaborative learning process as also found by (Bassiou et al., 2016; Praharaj et al., 2021). The difference in the learning of the *Silent Wanderers* could be because of fewer such productive collaborative episodes within this group. Lastly, both types of gainers show significantly lesser percentage of long pauses in their speech relative to *Silent Wanderers* which again, as suggested by previous research (Fors, 2015), tends to be indicative of better communication which is essential for good collaboration.

4.5.2 Log Actions

In terms of actions, it is clear that the two types of gainers do not exhibit the same exploratory approach with *Expressive Explorers* showcasing a more global exploratory approach of building a solution, testing and reflecting on their previous solutions before building a new one, while *Calm Tinkerers* displaying a more local exploratory approach of adding edges, reflecting on and possibly deleting an edge in-the-moment, as they build the solution. On the other hand, *Silent Wanderers* seem to not be adhering strictly to either of the two strategies and rather are displaying a mix of both. However, as we observe, both approaches incorporate some form of reflection, that is generally lesser in the non-learning group both in terms of reflective-in-the moment and reflection-on-prior actions. This may be why there are more redundancies present on the map for them at a given point in time. Hence, in terms of interaction with the task, it is the act of regulating their solution building approach through reflection that is differentiating the gainers from the *Silent Wanderers*. This is not surprising since reflection

has been found to play a pivotal role in learning from problem based learning environments (Barron et al., 1998; Do-lenh, 2012; Etkina et al., 2010; Hmelo-Silver, 2004).

As suggested by research, regulating ones' own and a partners' cognition, metacognition, behaviours and emotions is important for productive collaborative learning (Järvelä et al., 2016). Our findings related to speech and actions together suggest that the gainers regulated their learning by verbally interacting with each other and reflecting on their solution approach, thus obtaining learning gains. The *Silent Wanderers*, on the other hand, had less verbal interaction and reflection, which could be the reason for not having learning gains.

4.5.3 Affective Behaviors

When it comes to affective behaviors, we observe that *Expressive Explorers* exhibit high arousal and negative valence (possible confusion/frustration) similar to the non-learning group *Silent Wanderers* and significantly different from the second group of gainers *Calm Tinkerers*. This suggests that confusion/frustration itself may not be the reason for not learning and that it is rather the set of other behaviors, which accompany this frustration, that define if a team would end up learning or not in an open-ended collaborative activity. This outcome is contrary to the more popular belief that views frustration as something to alleviate (D'Mello & Graesser, 2012; Hone, 2006; Klein et al., 2002) but rather is in line with the work of R. S. Baker et al., 2010 and Mentis et al., 2007 that have suggested that in some cases, frustration may not need remediation. However, an important question that arises here is whether the *Expressive Explorers* end up learning *despite* frustration or *because* of it. The answer to this question is out of the scope of this thesis; however, it can be an interesting question to explore for the community.

Further, as highlighted by the interaction analysis, both types of gainers show a change in their average emotional states right after submitting a sub-optimal solution, together with a phase of high speech overlap. The team of *Expressive Explorers* show a dip in their emotional state while the team of *Calm Tinkerers* show an increase in their emotional state. The latter case can be explained by the model proposed by D'Mello and Graesser, 2012 for the dynamics of affective states during complex learning, where the authors suggest that learners states oscillate between a state of equilibrium (flow) and disequilibrium (confusion) when an impasse is detected. In the episode we analysed, as the *Calm Tinkerers* discover that their solution is incorrect, this can lead to confusion (higher emotional states). The case of the *Expressive Explorers* team is interesting because it is not directly explained by the model of D'Mello and Graesser, 2012. However, it must also be noted that this team in general showed higher frustration during the activity and thus this can be considered *their* state of equilibrium. Hence, on receiving feedback about the sub-optimality of their solution, they switched to a lower emotional state, which for them is a state of disequilibrium. In the case of both types of gainers however, we see an attempt to regulate the state of disequilibrium via effortful reasoning and problem solving (Järvelä et al., 2016). This leads to an increase in

verbal interaction with interjections while discussing revised problem solving strategies. It is interesting that the *Silent Wanderers* team showed no change in their affective state in the episode of high speech occurring after submitting a sub-optimal solution. It is worth exploring further what this lack of change in affective state at a moment of impasse means for learning.

4.5.4 Gaze Behaviors

When it comes to gaze patterns, we did not observe any significant differences between the two gainer groups, suggesting that they have a very similar behavior when paying attention to the screen as well as when looking at their partner or the robot. Moreover, when comparing the two types of gainers with the *Silent Wanderers*, the only significant difference observed was with respect to looking more on the right (where the previous solutions can be displayed upon clicking on a button) or the left side of the screen, while there are no differences among the gaze patterns when looking towards their partner, the robot or the opposite side of the robot. This suggests that, for the gaze behaviours we considered, a "productive" gaze pattern does not emerge from the data.

4.5.5 Tying it All Together: How the Different Modalities Interplay?

Going back to our research question on multi-modal behavioural profiles of learning in a collaborative constructivist activity, we have identified two types of gainer profiles based on our pair-wise analysis in section 4.4. The first gainer profile, *Expressive Explorers* consists of *effective communication* as seen by their high amount of verbal interaction between the team members, periods of high overlap in speech of the team members, fewer longer pauses in the speech; a *global exploratory approach* consisting of adding a lot more edges while solving the task followed by *reflection* by opening their past solutions; and exhibiting a state of *frustration* seen by high arousal and negative valence. The second gainer profile, *Calm Tinkerers*, similar to the first one, is characterized by *effective communication*. However, differently from the first one, it consists of a *local exploration approach* in which team members remove a lot more edges while constructing a solution; *local reflection* or reflection-in-the moment, represented by a higher number of sequence actions such as a team member adding or removing their own or their partners' recently added edge; and a *relatively calm emotional state* characterized by low arousal and negative valence. Finally, the non-gainer profile, i.e., that of *Silent Wanderers*, is characterized by *poorer communication* meaning significantly less verbal interaction and less speech overlap, and more long pauses compared to the two types of gainer profiles. In addition, similar to *Expressive Explorers*, *Silent Wanderers* exhibit *frustration*; however, compared to both the gainer profiles, they *reflect* less both on prior solutions (open their history less) and recent actions (have less sequence actions such as a team member adding or removing their own or their partner's actions). This third profile lends further support for the need of regulation of learners' problem solving strategies and frustration via reflection and verbal communication in order for effective collaborative learning to happen (Järvelä et al., 2016).

The fact that only two out of three identified multi-modal behavioural profiles learned, is in line with literature which suggests that while collaboration can scaffold learning, it is contingent upon the quality of the interactions (Dillenbourg et al., 2009), and diverse and complex conditions (Lou et al., 2001; Meier et al., 2007). Furthermore we found, similar to literature, that while impasses and failures can offer the conditions for learning to happen, whether it actually does happen depends on learners' cognitive (Barron, 2003; Lodge et al., 2018; Loibl et al., 2017), social (Weinberger & Fischer, 2006) and emotional behaviours (D'Mello & Graesser, 2012) as a response to the moment of encountering an impasse. Our work identifies two possible collections of actions, speech and affective behaviours under which effective collaborative impasse-driven learning can occur and one collection of behaviours under which it does not. Thus through this work, we provide a more holistic assessment of the behaviours underlying collaborative impasse-driven learning that can contribute to refining the theories of both collaborative learning and impasse-driven learning as we elaborate below.

Our findings confirm some of the findings in the CSCL literature in the context of an open ended collaborative activity: (1) *verbal interaction, not just in terms of amount of speech but also overlap of speech between team members, in a constructivist collaborative activity emerges to be a discriminatory factor between gainers and non-gainers* but (2) *it is not always one single behavior that discriminates gainers from non-gainers; rather it is a set of behaviors* which may not always be obvious when observed by experts such as a teacher or observer in such exploratory collaborative activities. Furthermore, it must also be noted that half of the gainers in *Expressive Explorers* and *Calm Tinkerers* groups actually *fail* at the task, and the same ratio holds in the *Silent Wanderers* group, suggesting once more that (3) *performance in the task, which often influences human experts in their evaluation of a learner's progress, is not always a reliable predictor of learning.*

What is relatively less clear from literature is when high and low reflection or emotions are productive for learning. Our work makes a step in that direction, as the aggregate multi-modal behavioural profiles of learners highlight that certain kinds of reflection (reflection-in-the-moment) is accompanied by calmer emotions, while other kinds of reflection (reflection-on-action) is accompanied by more expressive emotions. That is, in our work, we discover that there exists a *relationship* between two of the modalities, i.e., *problem-solving strategy* and *emotional expressivity*, that can discriminate multiple ways of achieving the learning goal. The fact that the strategies differ among the two types of gainers is not a surprise as problem-solving strategies have been studied in CSCL literature; however, *the fact that the arousal and valence are interplaying with the different types of strategies is a novel contribution of this work.* More specifically, we observe the interplay in the diagonal shown in Figure 4.12, that suggests that *expressivity of emotions could be related to the problem-solving strategy.* A certain strategy leads to more episodes of frustration than the other, and examining multiple modalities simultaneously allows us to unearth this relationship. It also raises an interesting question as to why there are no gainer teams in the cross diagonal. Is this where the non-gainers lie? While the *Silent Wanderers* do exhibit a higher emotional expressivity, they do not strictly adhere to either of these two problem-solving strategies as they exhibit lower level

		Emotions	
		Expressive	**Calm**
Strategies	**Explorers**	X	
	Tinkerers		X

Figure 4.12: The interplay between the problem-solving strategies and the emotional expressivity for the gainer teams.

of both local and global reflection. Hence, they too do not lie in this cross diagonal. Then the question to consider is whether the cross-diagonal would have learning or non-learning profiles.

Hence, the insights from our current results can inform CSCL designers regarding *what* interplay between problem solving strategies and emotional expressivity may be more conducive to learning in such a CSCL setup in *addition* to the more obvious behavior of speech activity. This can help in making a more informed design of a robot or an autonomous agent for adaptive interventions which can first use simple speech activity measures to identify non-gainers. Other speech measures such as semantics of speech, that might be more descriptive, need manual work by humans that cannot always be done in real time. Hence, the easier automatic assessment in real-time with speech activity measures makes them a great choice for guiding effective interventions by intelligent systems. Once an 'unproductive state' is identified via speech, the agent/robot can use information provided by the other modalities to try and scaffold the learners towards either of the gainer profiles. For example, if a team is following a more tinkering problem-solving strategy and they continuously start displaying higher levels of frustration on average, there may be a need to remediate this frustration, as it could push them to a non-gainer profile. Conversely, frustration displayed by a team displaying a more global exploratory problem strategy may not need remediation.

We must point out that our analysis in this chapter is limited in certain aspects. The data driven clusters are imbalanced meaning that with our pipeline, the non-learning cluster that emerges has lesser number of teams. This may be one of the reasons why the gainer profiles are clearer compared to the *Silent Wanderers* profile. Secondly, while current learning profiles tie back to literature both in terms of behaviors and constructs as we see above, the limitation lies in the fact that the profiles are only based on a *snapshot* of learning *at the end of the process*. Ultimately, these behaviours are not constant across the activity and learning is inherently characterized by episodes of both reflection-on-action and reflection-in-action (Lavoué et al., 2015) and both positive and negative emotions (T. Sinha, 2021). To better understand the *evolution* of these behaviors and constructs, i.e. to elaborate the *process of learning* and to further build theories of impasse-driven collaborative learning, in our next two chapters, we

aim to investigate temporal data from the same *Ron* study to develop temporal understanding of the learning process. If by analysing deeper at temporal level using the *needed* modalities for the goal at hand, we obtain similar findings, this could further strengthen the intervention framework. The eventual goal is then to incorporate these insights for real-time intervention in constructivist collaborative activities.

5 Temporal Pathways to Learning

How Learning Emerges in an Open-ended Collaborative Activity

Now that we have established 3 profiles of collaborative learning in the last chapter, when looking at the aggregate team level data; in this chapter, we advance our understanding of the *process* of learning by focusing on temporal data. We report on the findings of employing a multi-modal Hidden Markov Model (HMM) to investigate the temporal learning processes of the gainers and non-gainers. Considering log data, speech behavior, affective states and gaze patterns, we find that all learners start from a similar state of non-productivity, but once out of it they are unlikely to fall back into that state, especially in the case of the learners that have learning gains. Unlike what we concluded in the last chapter, gainer groups actually shift between both the problem solving strategies, each characterized by both exploratory and reflective actions, as well as demonstrate speech and gaze patterns associated with these strategies, that differ from those who don't have learning gains. Further, gainers also differ between themselves in the manner in which they employ the problem solving strategies over the interaction, as well as in the manner they express negative emotions while exhibiting a particular strategy. These outcomes contribute to understanding the multiple pathways of learning in an open-ended collaborative learning environment, and provide actionable insights for designing effective interventions.

This work corresponds to the following publications:

J. Nasir, M. Abderrahim, A. Kothiyal, and P. Dillenbourg, "Temporal Pathways to Learning: How Learning Emerges in an Open-ended Collaborative Activity." in *Computers Education: Artificial Intelligence*, 2022 (Nasir, Abderrahim, et al., 2022).

[Dataset] **Jauwairia Nasir**, Barbara Bruno, Pierre Dillenbourg. (2021). PE-HRI-temporal: A Multimodal Temporal Dataset in a robot mediated Collaborative Educational Setting. Zenodo. https://doi.org/10.5281/zenodo.5576058. (Nasir, Bruno, & Dillenbourg, 2021a).

5.1 Introduction

Learning does not occur in a single moment, but is rather a dynamic *process* that evolves over time (Kapur, 2011; Reimann, 2009). This process, especially in open-ended learning environments such as inquiry-based learning and problem-based learning environments, is non-linear (Brooks & Brooks, 1993; Chow et al., 2015; Schulte, 1996). Researchers have proposed that learning contexts are in fact complex systems where elements at different levels, such as cognitive, intrapersonal and interpersonal, interact and this results in the emergence of learning (Jacobson et al., 2016). Therefore, understanding the conditions for emergence of learning in this complex system is important, as this will help identify those moments when an intervention could potentially be effective to improve learning. Within computer-supported collaborative learning (CSCL) research, there is now an emphasis to focus on how the CSCL *process* unfolds (Lämsä et al., 2021).

While pre and post-tests help ascertain how much knowledge a learner has gained, they do not help understand *how* this knowledge was *gained* in a particular context, i.e., the temporal and multi-modal aspects of the learning process. These aspects of the learning process have been previously studied using methods such as microgenetic analysis (Siegler & Crowley, 1991), interaction analysis (B. Jordan et al., 1995) and interactional ethnography (Castanheira et al., 2000) of learner discourse and actions, which track students conceptual development across an individual or collaborative learning activity. However these qualitative methods can be time intensive. With technology-based learning contexts and multisensory data becoming increasingly widespread, researchers are making use of multiple sources of behavioral data such as interaction logs, audio, video, eye gaze and physiological data, along with machine learning methods, to understand the process of learning as a function of time (Engelmann & Bannert, 2021; Olsen et al., 2020b). For example, in Lämsä et al., 2020, the authors make use of log data and lag sequential analysis to highlight the potential of temporal analysis to identify differences in the inquiry-based learning processes of scaffolded and non-scaffolded groups. Specifically, they discover three temporally distinct inquiry-based learning transition patterns among the three experimental groups that indicate different ways of using the scaffolds that could explain their learning. Further, in Csanadi et al., 2018, the authors show that their proposed methodology accounting for temporality, provides more insights than the traditional code-and-count strategies to characterize the socio-cognitive activities of learning in CSCL environments. Specifically, they found that 'evaluating evidence' was a core epistemic practice for dyads but not for individuals, suggesting that students collaborating argued in a more evidence-focused manner compared to individuals.

To reiterate, in this chapter, our goal is to develop a temporal and multi-modal model of the learning process in our open-ended collaborative activity. Towards this goal, *we propose a Hidden Markov Model (HMM) based temporal analysis of multi-modal behavioral data to identify the differences and similarities between the learning processes of those who learn and those who do not.* Our choice of using HMMs is motivated by the fact that HMMs allow us to model learning as a latent process based on our observations of student interaction with the

learning activity.

In the upcoming section, we will review literature regarding *temporal and multi-modal* analysis methods for learning. Then in Section 5.3, we elaborate on the participants, the activity and the dataset used in this work, the experimental setup, as well as the adopted analysis methodology. This is followed by results, discussion, and conclusion in Section 5.4 and 5.5, respectively.

5.2 Literature Review

When embedded in a learning activity, intelligent agents must intervene at the right moment and in the right manner to enhance the learners' learning gains. To do so, the system must have an ongoing comprehensive and deep understanding of the learners and learning situation. Temporal analysis of learners data, either performance or behaviors, can provide such an understanding.

5.2.1 Performance Based Systems

Knowledge Tracing

In Knowledge Tracing (KT) systems, temporal learner understanding is developed by estimating the learner's knowledge from their performance on past problems (Corbett & Anderson, 1994; Desmarais & Baker, 2012). **Bayesian Knowledge Tracing (BKT)** determines if and when the learning of a skill occurs during problem-solving steps (Desmarais & Baker, 2012). It assumes a two-state learning model where each skill is either in the learned or unlearned state. Assuming that each step of each problem calls for a single skill, the student can either succeed or fail the step, and the tutor updates its estimate of the learners knowledge on the skill accordingly (Corbett & Anderson, 1994; Desmarais & Baker, 2012). BKT has been applied both in the form of a *Hidden Markov Model* as well as in the form of a *Knowledge Tracing algorithm* (van de Sande, 2013). While these approaches have been applied successfully to model student knowledge in well-structured problem-solving, they fail at more complex open-ended learning activities (Wang et al., 2021). Hence, to increase the representational power and better model complex problem structures, Käser et al., 2017 suggest a **Dynamic Bayesian Network (DBN)** model that incorporates skill topologies. In this, different skills of a learning domain are considered within a single model capturing the dependencies between them. Incorporating skill hierarchies yields a significant improvement in predicting students' knowledge during complex problem solving, more accurately compared to the traditional KT models.

Further, **Deep Knowledge Tracing (DKT)** (Piech et al., 2015), an application of recurrent neural networks, has been shown to be able to learn the latent structure in skill concepts without the need for explicit human coding of domain knowledge. For this reason, it demonstrates a drastic

improvement on the well-known BKT models over several data sets. Nonetheless, similar to BKT, the DBN model as well as DKT assume that each problem-solving step or action maps to an underlying skill that could be either learned or unlearned, which is not necessarily the case in open-ended learning environments. Moreover, these approaches assume that an incorrect answer implies not learning or "slipping". However, it has been found that learners' actions that may seem to suggest failure vis-à-vis conventional standards of efficiency, accuracy, and performance quality may still lead to learning gains (Kapur & Kinzer, 2009). Thus, indicators other than performance should be considered to model the learning process in open-ended learning activities. In Ramachandran, Huang, et al., 2019, the authors suggest a link between motivation, actions, and the learning outcomes that underlies the learning process. They propose creating more effective tutoring interactions by finding observable behaviors that correspond to motivational factors and employing a robot to respond to these behaviors. In Nasir, Bruno, Chetouani, et al., 2021, the authors found that teams achieving higher learning gains in a robot-mediated human-human collaborative learning activity, may not necessarily perform well in the task. However, their speech, actions and emotions are distinctive as compared to the teams with lower learning gains. Thus, behavioral analysis could allow for better discrimination between high and low learners which will be the focus of our next sub-section.

5.2.2 Behavior Based Systems

Qualitative Methods

When analyzing the learning process using learners' behaviors, both qualitative and quantitative approaches have been employed. Qualitative methods have been used to analyze, mainly, learners' gestures and speech to see how their learning is evolving. For instance, M. E. Jordan and McDaniel Jr, 2014 employ discourse analysis to describe the issues about which learners experienced uncertainty as they pursue collaborative learning projects that include a cognitive feeling of uncertainty. They identified how language was used in these particular social contexts to create and reflect meaning and structure. In Voutsina et al., 2019, authors used microgenetic task analysis to analyze the change in children's verbal reports when their overall solving approach appears to remain stable during a mathematical problem-solving task. They found that in fact the phases of stability are underlain by dynamic changes in the way the same strategy is communicated and conceptualized.

Although qualitative methods make it possible to contextualize and interpret the data based on human perception and analysis of the learning scenario, they sometimes overlook hidden factors that human observation cannot capture. Additionally, these methods are time and effort intensive, and as a result, do not scale up efficiently. With the development of sensors that capture data that is not perceivable by humans and the advancement in machine learning analysis techniques, there has been an increase in the deployment of quantitative approaches. Desmarais and Baker, 2012 argue that as more and more learner data becomes available and

methods for exploiting that data improve, there is potential for constant improvement of learner models. In this regard, researchers have attempted to gain an understanding of the learning process by considering multiple modalities and machine learning (ML) techniques as discussed below.

Quantitative Methods

Perera et al., 2009 apply **sequential pattern mining (SPM)** on learners' log actions in a collaborative learning environment to extract sequences of frequent events. This analysis revealed interesting patterns, such as the presence of frequent task-focused communication, characterizing the teams ending up with positive and negative outcomes. Successful groups exhibit patterns suggestive of members giving frequent updates to the group while working on a task; such patterns are not present in the weaker groups. Kinnebrew et al., 2014 used SPM algorithms along with an hierarchical clustering algorithm to study the temporal evolution of the sequential patterns throughout the intervention, and compare the similarities and differences of their use between experimental groups interacting with distinct versions of the software. The mined patterns allow for identifying and interpreting students' cognitive skills and learning behaviors. Besides, comparing these mined patterns with performance and context information, and tracking their temporal evolution better characterizes these behaviors as effective versus ineffective learning strategies. For instance, the importance of solution evaluation behaviors in complex learning tasks, is identified as one of the effective learning strategies.

Process Mining (PM) has also been applied to behavioral data to examine the learning process. This technique was adopted to discover the underlying problem solving or learning process model from the learning activity interaction sequence. Paans et al., 2019 employs a fuzzy miner algorithm, on sequences of encoded verbal utterances within dyads in a collaborative learning activity and find that repeated occurrences of social challenges during collaboration harm the learning outcomes. Here social challenges are defined as the failure to get along, a lack of joint attention, being highly critical, and so on. In fact, pairs, who repeatedly have disagreements, are more easily distracted, more easily go off-topic, have trouble getting back on topic again, and thus, are at risk for lower assignment quality.

Further, research suggests including more than one modality in the analysis because incorporating **multi-modal techniques** would allow researchers to examine unscripted, constructionist, complex tasks in more holistic ways (Blikstein, 2013). Emerson et al., 2020b investigate this by analyzing log actions, facial expression of emotions, and eye gaze both separately and combined, and find that models utilizing multi-modal data either perform equally well or outperform models utilizing unimodal data to predict learners' posttest performance and interest in a game-based learning environment. Olsen et al., 2020a further incorporate data temporality by using a **Long Short-Term Memory (LSTM)** model on log, gaze, audio, and dialog temporal data to predict teams' performance in a collaborative learning activity. The results indicate that combining various data streams from different time scales may be more

beneficial than unimodal data. They also highlight the value of accounting for temporal aspects of the learning process as the temporal analysis of the gaze and audio measures provided accurate prediction of the normalized learning gain, while the averages and counts based analysis on the same features provided no information. Further, Giannakos et al., 2019 highlight how fused multi-modal data, consisting of eye tracking, EEG, video, and wrist band data in addition to click stream data, can considerably reduce the prediction error for learning performance as compared to when only click streams are used in the design of learning technology. Lastly, in Yang et al., 2021, the authors have modelled the joint visual attention and with that the cognitive engagement of dyads using eye gazes and eye blinks data, and suggest that this multi-modal temporal approach gives more and accurate insights into the collaborative problem solving engagement.

Another ML technique that has been used to temporally model the learning process with multi-modal data is the **Hidden Markov Model (HMM)**. In Sharma and Giannakos, 2020, the authors use a combination of HMMs and the Viterbi algorithm to predict learners' effortful behaviors throughout the learning activity. They consider the effort categories as the hidden states and multi-modal data-driven clusters as the observations. Results show that the suggested method outperforms the contemporary classification algorithms in classifying learners' behavioral patterns as effortful or effortless. Furthermore, this methodology highlights the exact moments when feedback is needed during the learning activity.

Literature suggests several data-driven multi-modal ML approaches that could be used to analyze temporal data. Choosing a particular approach depends on the assumptions made about the measured data and the learning process underlying it, the nature of the data, the volume of available data, the purpose of the analysis, and the interpretability of the obtained models. The purpose of our analysis is to build a multi-modal temporal model of the underlying process of learning as it happens in an open-ended collaborative learning activity. Sequence mining, sequential pattern analysis, and stochastic methods such as lag-sequential analysis, for instance, do not include the assumption of a latent learning process governing the sequence of observations (Bannert et al., 2014). Thus, we do not consider such methods for our temporal multi-modal behavioral data analysis. Process mining, on the other hand, does account for latent processes; however, it is usually used to identify, confirm, or extend process models on sequential event data, which are sequences of discrete data, and thus, are different in nature from the data we investigate, which includes multivariate continuous features. Then, Recurrent Neural Networks (RNN), particularly LSTMs, have been broadly employed to analyze temporal multi-modal behavioral data while complying with the assumption of a hidden process controlling the sequence of observations. Although promising (Spikol et al., 2018), these neural networks lack the interpretability for multi-variable data regarding variable importance and variable-wise temporal importance due to their opaque hidden states (Guo et al., 2019). HMMs however offer more interpretability as the hidden states are well defined by their transition probabilities and emissions distributions. Therefore, they allow for a better understanding of the latent learning process during the learning activity.

Therefore, in our work, we adopt the approach of building a Hidden Markov Model of the learning process, trained on learners' multi-modal behavioral data. Our goal is to examine how these behaviors evolve throughout the activity and lead to learning gains during an open-ended collaborative learning activity. Broadly, our research question in this chapter is, "*How do the learning behaviors of different types of learners evolve across an open-ended collaborative learning activity*?"

5.3 Methods

5.3.1 Dataset

We make use of our open-source temporal dataset *PE-HRI-Temporal* (Nasir, Bruno, & Dillenbourg, 2021b) generated from the data collected in the *Ron* study that has been elaborated previously in Chapters 3 and 4. In this data set, for each team, the interaction of around 20-25 minutes is organized in windows of 10 seconds; hence, we have a total of 5048 windows of 10 seconds each. We report team level log actions, speech behavior, affective states, and gaze patterns for each window. More specifically, within each window, 26 features are reported in two formats; hence, giving a total of 52 values. We make use of the *non-incremental* format of the 26 features which means we look at the value of a feature in that particular time window without carrying any information from previous time windows. For more details, please see Nasir, Bruno, and Dillenbourg, 2021b. The 26 features are listed in Tables. 5.1, 5.2, and 5.3. The rationale for using these features to analyse learning are explained in our previous chapters (specifically referring to the chapters 3 and 4).

In addition to the features in the aforementioned tables, each window also includes a *normalized_time* feature which refers to the time when this window occurred with respect to the total duration of the task for a particular team. The dataset also consists of team level learning and performance metrics, where performance is measured based on the cost of a current solution relative to the optimal solution, while learning gains (absolute, relative or joint-absolute) are calculated by looking at the difference between the students scores on their post-tests and pre-tests. More detailed definitions are provided at Nasir, Bruno, and Dillenbourg, 2021b. Please note again that this dataset provides data for 34 teams, but for our current analysis we make use of data from 32 teams giving us 4676 windows. We removed two teams that were outliers in terms of their behaviors (based on data driven behavior profiles that were generated in the last chapter). Lastly, considering learning analytics and/or educational human-robot interaction studies with a robot, similar or even lower sample sizes are the norm (Belpaeme et al., 2018; Gordon et al., 2016; Ramachandran, Huang, et al., 2019), as is the case with the type of analysis that we do in this work (for example, see Sharma and Giannakos, 2020.

Table 5.1: Log features from our PE-HRI-Temporal dataset

Log Features	
Feature Name	**Description**
T_add	The number of times a team added an edge on the map in that window
T_remove	The number of times a team removed an edge from the map in that window
T_ratio_add_rem	The ratio of addition of edges over deletion of edges by a team in that window
T_action	The total number of actions taken by a team (add, delete, submit, presses on the screen) in that window
Redundant_exist	The number of times the team had redundant edges in their map in that window
T_hist	The number of times a team opened the sub-window with history of their previous solutions in that window
T1_T1_add	The number of times either of the two members in the team followed the pattern consecutively: I delete an edge, I add it back in that window
T1_T1_rem	The number of times either of the two members in the team followed the pattern consecutively: I add an edge, I then delete it in that window
T1_T2_add	The number of times the members of the team followed the pattern consecutively: I delete an edge, you add it back in that window
T1_T2_rem	The number of times the members of the team followed the pattern consecutively: I add an edge, you then delete it in that window
T_help	The number of times a team opened the instructions manual in that window

Table 5.2: Video based features from our PE-HRI-Temporal dataset

Video Features: Affective states and Gaze	
Feature Name	**Description**
Positive_Valence	The average value of positive valence for the team in that window
Negative_Valence	The average value of negative valence for the team in that window
Difference_in_Valence	The difference of the average value of positive and negative valence for the team in that window
Arousal	The average value of arousal for the team in that window
Gaze_at_Partner	The average of the the two team member's gaze when looking at their partner in that window where each individual member's gaze is calculated as a percentage of time in that window.
Gaze_at_Robot	The average of the the two team member's gaze when looking at the robot in that window where each individual member's gaze is calculated as a percentage of time in that window.
Gaze_other	The average of the the two team member's gaze when looking in the direction opposite to the robot in that window where each individual member's gaze is calculated as a percentage of time in that window.
Gaze_at_Screen_Left	The average of the the two team member's gaze when looking at the left side of the screen in that window where each individual member's gaze is calculated as a percentage of time in that window.
Gaze_at_Screen_Right	The average of the the two team member's gaze when looking at the right side of the screen in that window where each individual member's gaze is calculated as a percentage of time in that window.
Gaze Ratio of Screen_Right and Screen_Left	The average ratio of a team member looking at the right side of the screen over the left side in that window

Table 5.3: Audio based features from our PE-HRI-Temporal dataset

Audio Features: Speech	
Feature Name	**Description**
Speech_Activity	The average of the two team member's speech activity in that window where each individual member's speech activity is calculated as a percentage of time that they are speaking in that window.
Silence	The average of the two team member's silence in that window where each individual member's silence is calculated as a percentage of time in that window.
Short_Pauses	The average of the two team member's short pauses over their speech activity in that window. Each individual member's short pause refers to a brief pause of 0.15 seconds and is calculated as a percentage of time in that window.
Long_Pauses	The average of the two team members long pauses over their speech activity in that window. Each individual member's long pause refers to a pause of 1.5 seconds and is calculated as a percentage of time in that window.
Speech_Overlap	The average percentage of time the speech of the team members overlaps in that window.
Overlap_to_Speech_Ratio	The ratio of the speech overlap over the speech activity of the team in that window.

5.3.2 Analysis Methodology

In our previous chapter, we generated behavioral profiles based on the same features described above in section 5.3.1, but aggregated across the entire activity. We found differences in the behaviors between those who learn, i.e., *gainers* and those who do not end up learning, i.e., *non-gainers*. Further, we also observed behavioral differences in the two types of gainers (Chapter 4). We saw that while *speech behavior* was a discriminatory factor between gainers and non-gainers, it was actually the interplay between problem solving strategies and emotional expressivity that distinguished the different ways in which gainers learned. Based on that, we identified the two types of gainers as *Expressive Explorers* and *Calm Tinkerers*, and the non-gainers as *Silent Wanderers*. In this Chapter, we retain the same terminology. While the aforementioned behavioral profiles highlight the aggregate differences between all types of learners, in order to identify the differences between the *learning process* of those who learn and those who do not, we employ HMMs to generate multi-modal *temporal* behavioral profiles for each type of learners. This enables us to understand how the multi-modal behaviors of each type of learners evolve throughout the interaction.

An HMM is a doubly stochastic model with an underlying stochastic process that is not observable, but can only be observed through another set of stochastic processes that produce the sequence of observed symbols. It is specified by a set of N states, an initial probability distribution, a transition probability matrix, and a sequence of emission probabilities. Additionally, HMMs require three assumptions: firstly, that the next state is dependent only on the current state, secondly, that the state transition probabilities are independent of the time of transition and finally, that the current observations are statistically independent of the previous outputs. In our case, our data is grouped into independent 10 second windows, with each window containing behaviors occurring in those 10 seconds alone, and thus assumption 3 holds. Further, each hidden state of the HMM manifests a set of significantly different behaviors by which the

state is characterized; this set of behaviors together signify a particular *approach to learning*. Hence, the next state or the approach to learning taken next by a pair of learners depends only on the current state (assumption 1) and the probability of transitioning to a different approach to learning is independent of when in the activity it occurs (assumption 2). Thus all the assumptions required to do an HMM analysis are valid for our data and learning context; hence, allowing us to proceed with HMM modeling. Our analysis consists of four main steps:

Step1: Preprocessing

As our features come from different kinds of behavioral modalities, they are on different scales. So we begin by applying a min-max scaler to normalize our data.

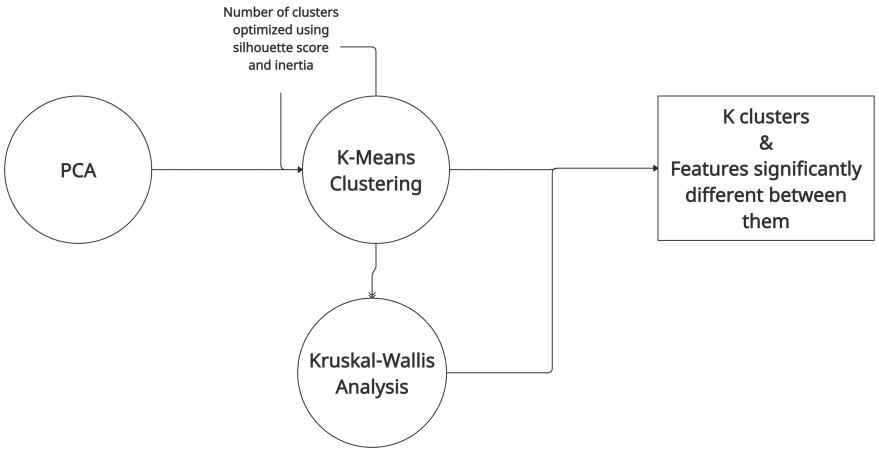


Figure 5.1: Behaviors Clustering step

Step2: Behaviors Clustering

In order to have a starting point for the number of states of the HMM, we perform a clustering of the temporal behavioral features to identify significantly different behavioral clusters. We then assume that these clusters are emitted by distinct hidden states, and so the number of states is the same as the number of behavioral clusters. For clustering, a Principal Component Analysis (PCA) is conducted to compute the principal components, the first components are kept based on the elbow method on the proportion of variance explained. The Principal Components are then clustered using the K-Means algorithm. The number of clusters is optimized based on the elbow method on inertia and the silhouette score. In order to confirm that the obtained clusters are actually different in terms of multi-modal behaviors, we perform a Kruskal-Wallis test on the clusters' behavioral features. This test further serves as a means to identify behaviors that significantly distinguish a cluster from the other. This step is summarized in Figure 5.1.

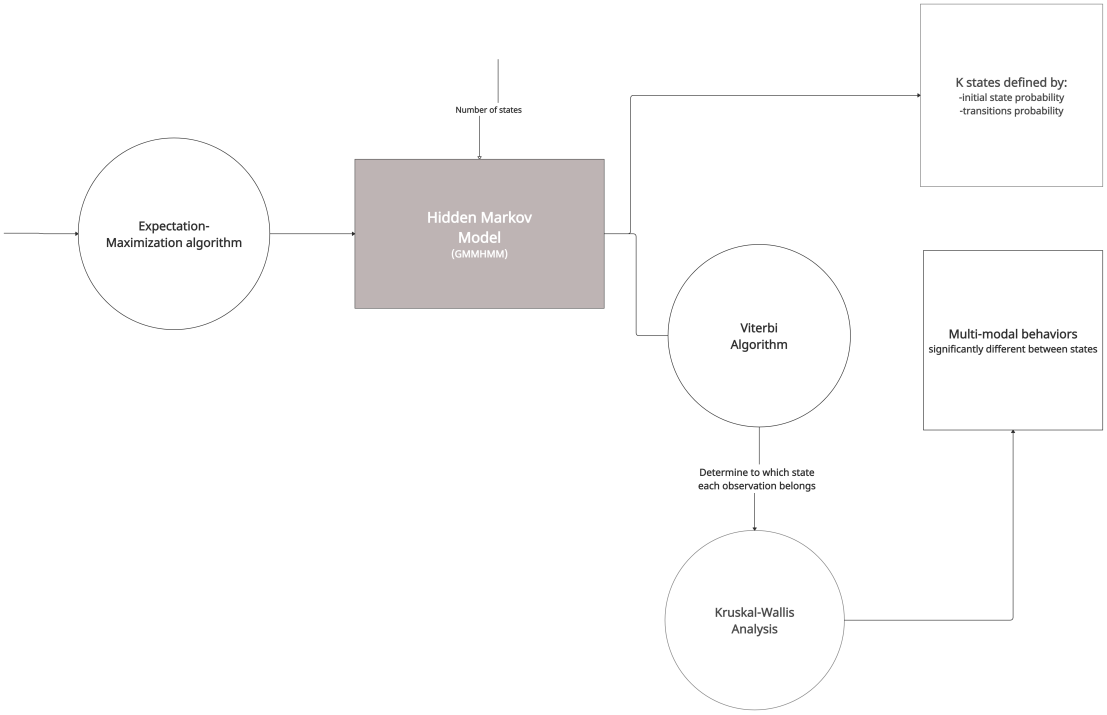


Figure 5.2: The HMM step

Step3: the HMM

Since our temporal behavioral features are multivariate and most of them have continuous values, our emission probability distribution should be continuous multivariate. Thus, for this step, we use the GMMHMM model provided by the hmmlearn library[1], as it accounts for the aforementioned condition by representing the emission distribution as a mixture of multiple Gaussian densities.

We set the number of hidden states to the number of clusters found in the previous step. The HMM is then trained using the Expectation-Maximization algorithm on the set of the teams' sequences. Each sequence consists of all the observations of a team sorted in increasing order of time, where an observation consists of the normalized multi-modal behavioral features and time at a given time window. We then apply the Viterbi algorithm on these sequences to recognize at which hidden state each observation is emitted. As a result, for each hidden state, we can construct the set of observations emitted by that state. Finally, we perform a Kruskal-Wallis test on each feature between each pair of these sets with the significance threshold set to 0.01. For each of the significantly different features between a pair of sets, we further compare the mean values across the sets and label the mean value of each set with one of the labels {Highest, High, Medium, Low, Lowest} based on a generated score in the following manner:

[1]hmmlearn is a set of algorithms for unsupervised learning and inference of Hidden Markov Models, https://hmmlearn.readthedocs.io/

For a significantly different feature x, we first define:

$min(x)$ = minimum of mean values of x across all sets

$max(x)$ = maximum of mean values of x across all sets

Then, for a set i, we generate a score for the feature x as:

$$score(x,i) = \frac{(\text{mean of } x \text{ in } i - min(x))}{(max(x) - min(x))}$$

Lastly, the feature x in i is labeled with:

- 'Highest', if $score(x,i) = 1$.
- 'High', if $2/3 \leq score(x,i) < 1$.
- 'Medium', if $1/3 \leq score(x,i) < 2/3$.
- 'Low', if $0 < score(x,i) < 1/3$.
- 'Lowest', if $score(x,i) = 0$.

The significantly different features and their labels for a set i represent the manifestation of the hidden state corresponding to the set i and we subsequently use these labeled features to represent the state. This enables us to interpret the progression of the hidden learning states in terms of the values of the significantly differing observed behaviors. Figure 5.2 outlines the processes employed to train and interpret the model.

In conclusion, in this step, the HMM is trained in order to learn the hidden states that emit the observed multi-modal behavioral features, and the significantly different features that characterize each state are identified. Interpreting these results allows for building the learning profiles that dyads go through during the activity. Furthermore, the model allows for learning the initial probability distribution as well as the probabilities to transition from one state to the other, which allows for building the temporal profile.

This entire pipeline, as summarized in Figure 5.3, is adopted to identify the temporal profiles for each type of learners separately.

5.4 Results

This section presents the results of the analysis methodology applied to the temporal multi-modal datasets of the *Expressive Explorers*, the *Calm Tinkerers*, and the *Silent Wanderers*. The clustering analysis, as discussed in the previous section, applied for the *Expressive Explorers*,

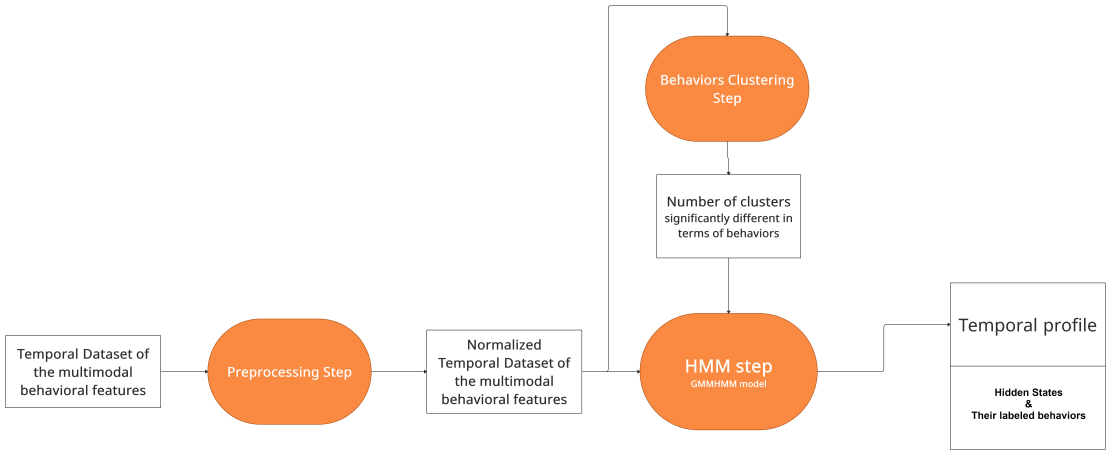


Figure 5.3: The Analysis Methodology

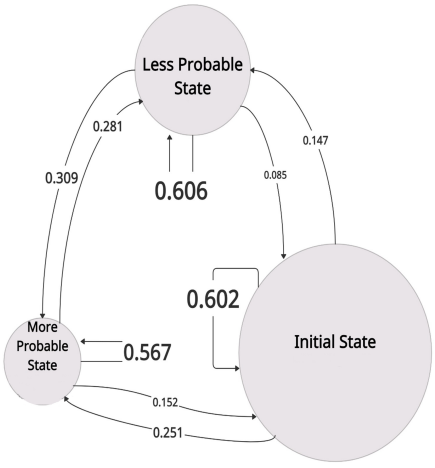
the *Calm Tinkerers*, and the *Silent Wanderers* suggests the following number of components [PCs = 4, PCs = 4, PCs = 5 respectively] and the following number of clusters [K=2, K=3, K=3 respectively], based on the elbow method on inertia and the silhouette scores. These are considered as a starting point for the number of hidden states, and we further train Hidden Markov models with K+1 states to identify whether other non trivial states exist or not, that eventually suggests that we have three hidden states for each of these groups. Hence, we define the following naming convention for the hidden states in each of the groups' models:

- ***InitialState***: the state with the highest initial probability.
- ***MoreProbableState***: the state with the highest transition probability from the initial state.
- ***LessProbableState***: the state with the lowest transition probability from the initial state.

We further define the following conventions for the state diagrams:

- The size of a state in the state diagrams is representative of its initial state probability. That is, the bigger the circle representing the state, the bigger its initial probability is.
- The size of the font of the transition probabilities in the state diagrams is illustrative of its magnitude. Explicitly, higher transition probabilities have bigger font sizes.

For each of the three groups, their HMM model, trained on sequences of observations of the respective group and the number of states set to three, is represented by the state diagrams in Figure 5.4, 5.5, and 5.6, respectively. For all groups, the probabilities suggest that once in *InitialState*, staying in that state has the highest probability compared to other possible transitions. However, once out of this state, going back to the *InitialState* from the *LessProbableState* and *MoreProbableState* generally has lower transition probabilities. The probabilities



Featuraes	States		
	Initial State	More Probable State	Less Probable State
Log Features:			
T_add	Lowest	Highest	Low
T_remove	Lowest	Low	Highest
T1_ratio_add_rem	Lowest	Highest	Low
T_action	Lowest	Highest	Low
Redundant_exist	Low	Highest	Lowest
T_hist	Highest	High	Lowest
T1_T1_rem	Low	Lowest	Highest
T1_T2_rem	Lowest	Low	Highest
T_help	Highest	Lowest	Low
Video Features:			
Negative_Valence	Lowest	Medium	Highest
Arousal	Lowest	Medium	Highest
Gaze_at_Partner	Low	Lowest	Highest
Gaze_at_Robot	High	Lowest	Highest
Gaze_at_Screen_Left	Highest	High	Lowest
Gaze_at_Screen_Right	Highest	High	Lowest
Audio Features:			
Speech_Activity	Lowest	High	Highest
Silence	Highest	Low	Lowest
Short_Pauses	Highest	High	Lowest
Long_Pauses	Low	Highest	Lowest
Speech_Overlap	Lowest	High	Highest
Overlap_to_Speech_Ratio	Lowest	Medium	Highest
Normalized_Time	Lowest	High	Highest

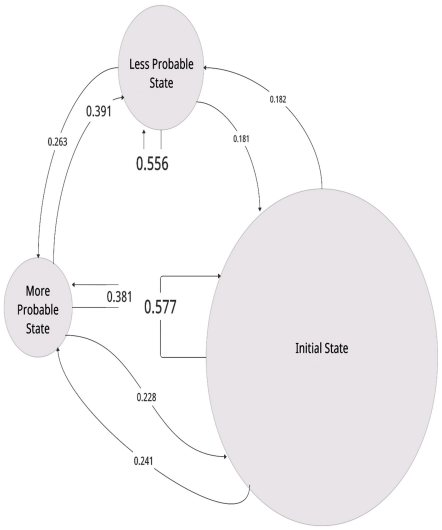
Figure 5.4: HMM State diagram for the Expressive Explorers

are especially low in the case of *Expressive Explorers* from both of the other states, and for both *Calm Tinkerers* and *Silent Wanderers* from the *LessProbableState*. On the other hand, the *Silent Wanderers* can still transition from *MoreProbableState* to *InitialState* with a non-trivial probability of *0.305* which is higher than the probability of going to *LessProbableState* from *MoreProbableState*. Similarly, the *Calm Tinkerers* also have a relatively higher transition probability to go back to the *InitialState* from their *MoreProbableState*; however, they still have a higher probability to transition to their *LessProbableState* from this state. Furthermore, the findings from the Kruskal-Wallis analysis comparing the values of the multi-modal behavioral features between each pair of states, for each group of learners, is shown in tables next to the respective HMM models. The tables include the features which represent the manifestation of the hidden states. Note that the features that do not differ significantly between the states are not shown in these tables. This does not mean the absence of that feature in a state, rather that the feature does not differ significantly between states, i.e., the value of that feature does not oscillate between states significantly. We discuss further on these results in the upcoming section.

5.5 Discussion

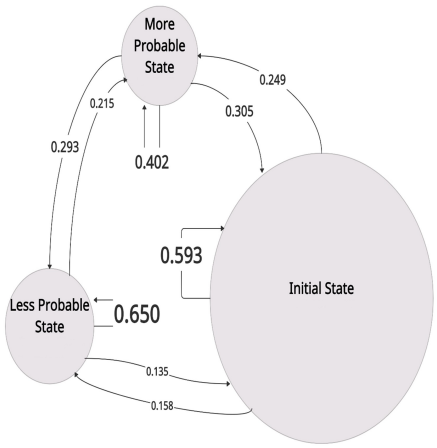
5.5.1 Temporal Multi-modal behavioral Profiles

In this section, we describe the higher level understanding that the temporal analysis, based on the HMMs identified in the previous section, provides us of how the multi-modal behaviors of each group of learners evolve during the collaborative learning activity and what this says



Featuraes	States		
	Initial State	More Probable State	Less Probable State
Log Features:			
T_add	Lowest	Highest	Low
T_remove	High	Lowest	Highest
T_ratio_add_rem	Lowest	Highest	Low
T_action	Lowest	Highest	Low
Redundant_exist	Lowest	Highest	Low
T1_T1_rem	Low	Lowest	Highest
T1_T2_rem	Low	Lowest	Highest
T_help	High	Lowest	Highest
Video Features:			
Positive_Valence	Lowest	Highest	Medium
Negative_Valence	Lowest	Highest	High
Difference_in_Valence	Lowest	Highest	Low
Arousal	Lowest	Highest	High
Gaze_at_Partner	Highest	Lowest	Medium
Gaze_at_Robot	Medium	Lowest	Highest
Gaze_at_Screen_Left	Lowest	High	Highest
Gaze_at_Screen_Right	High	Highest	Lowest
Audio Features:			
Speech_Activity	Lowest	High	Highest
Silence	Highest	Low	Lowest
Short_Pauses	Highest	High	Lowest
Long_Pauses	Highest	Medium	Lowest
Speech_Overlap	Lowest	High	Highest
Overlap_to_Speech_Ratio	Lowest	High	Highest
Normalized_Time	Lowest	Highest	High

Figure 5.5: HMM State diagram for the Calm Tinkerers



Featuraes	States		
	Initial State	More Probable State	Less Probable State
Log Features:			
T_add	Low	Highest	Lowest
T_remove	High	Lowest	Highest
T_ratio_add_rem	Low	Highest	Lowest
T_action	Low	Highest	Lowest
Redundant_exist	Medium	Highest	Lowest
T_help	Highest	Low	Lowest
Video Features:			
Positive_Valence	Lowest	Highest	High
Difference_in_Valence	Low	Lowest	Highest
Arousal	Lowest	Highest	Medium
Gaze_at_Partner	Lowest	Highest	Medium
Gaze_at_Screen_Left	Highest	Medium	Lowest
Gaze_at_Screen_Right	Lowest	Highest	High
Audio Features:			
Speech_Activity	Lowest	Medium	Highest
Silence	Highest	Medium	Lowest
Speech_Overlap	Lowest	Low	Highest
Overlap_to_Speech_Ratio	Lowest	Low	Highest
Normalized_Time	Lowest	Medium	Highest

Figure 5.6: HMM State diagram for The Silent Wanderers

about their learning process. Based on the findings in Section 5.4, we observe two kinds of problem solving (PS) strategies namely:

- Global PS Strategy: This strategy includes global level exploration and/or reflection characterized by addition actions and looking at past solutions (history).
- Local PS Strategy: This strategy includes local level exploration and/or reflection characterized by deletion actions and addition followed by deletion actions or vice versa.

Previously, in the results section, we name our states on the basis of initial probability (*Initial-State*) or transition probabilities from the initial state (*LessProbableState*, *MoreProbableState*). In this section, we try to understand the nature of the states and consequently, we name them based on their:

1. Productivity
2. Problem solving strategy

With respect to 1, in the previous chapter (specifically, Chapter 4), we found that the quantity and quality of speech was able to discriminate between productive and non-productive teams in terms of learning. Additionally, we found that when the behaviors were averaged across the entire interaction for each team, there were two problem solving strategies (Global PS Strategy and Local PS Strategy) that emerged and overall, one group of gainers displayed only one strategy, while the other group of gainers displayed the other. However, the temporal profiles of each group of learners help elaborate these findings further.

Please note that in the upcoming figures of the profiles, the strength of the transition probabilities is represented by the strength of the arrows and the unproductive, semi-productive and productive states and transitions are represented by different colors as described in the legend of the figures.

Expressive Explorers

The temporal profile for *Expressive Explorers* is shown in Figure 5.7 from which we see that these learners start, with the highest probability, at a state characterized by more technical help-seeking, fewer actions with the learning activity, and high silence. For these reasons, it appears to be a state of non-productivity. As opposed to the averages and frequency analysis in Chapter 4, which suggests that *Expressive Explorers* learned by following a more global problem solving strategy, this temporal analysis indicates that once they go out of the non-productive state, they employ both of the problem solving strategies: in the more probable state they follow a global problem solving strategy of adding edges and looking more at their previous solutions, and in the less probable state they follow a local problem solving strategy consisting

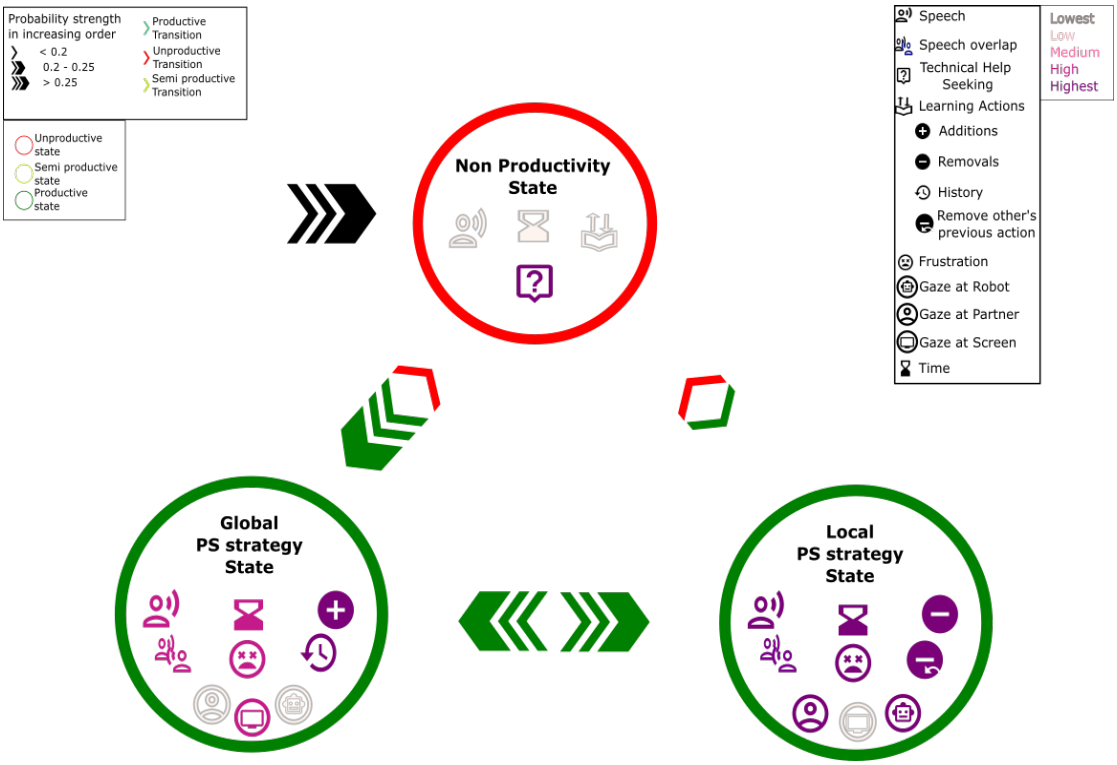


Figure 5.7: Temporal profile for Expressive Explorers

of more removals in general, and removing each other's last added edges in particular. What is interesting is that the latter state is more likely to occur at later times in the activity than the global problem solving state, suggesting that these students begin with a more global problem solving approach and move on to a more local strategy of making quick changes. This transition is also characterized by increasing negative emotions, such as frustration, that is perhaps brought on by the awareness of reaching the end of the activity and the allotted time. In the states of non-productivity (while trying to understand the activity) and global problem solving (while adding edges), the learners gaze at the screen is high, while in the state of local problem solving while removing edges, and in particular each others' edges, the learners gaze at their partners is highest. However, both of the problem solving states are characterized by high speech and speech overlap which signifies good collaboration (Viswanathan & Vanlehn, 2018). Once *Expressive Explorers* reach a productive state, it is highly unlikely to get back to the non-productive one.

Calm Tinkerers

Calm Tinkerers as shown in Figure 5.8 start, with the highest probability, at a state characterized by high technical help-seeking, fewer actions, and high silence. Due to these behaviors, it seems to be a state of non-productivity. Similar to *Expressive Explorers*, the temporal analysis done in this chapter gives a richer insight into these learners behaviors. Contrary to the

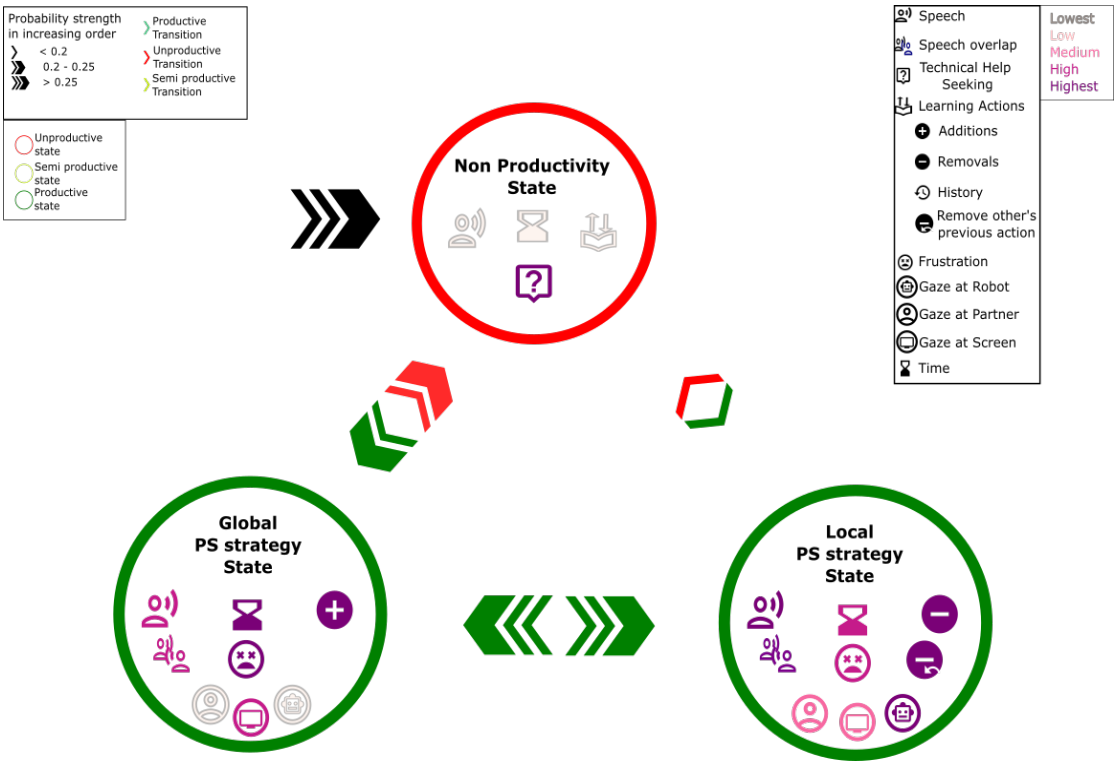


Figure 5.8: Temporal profile for Calm Tinkerers

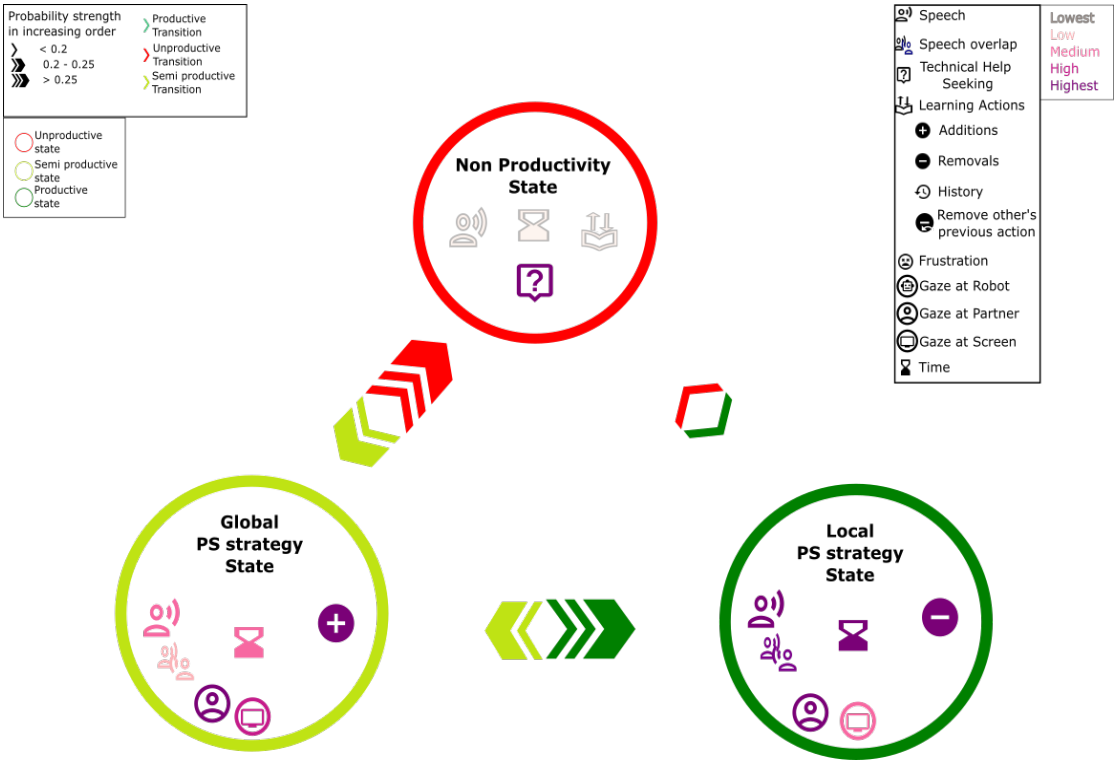


Figure 5.9: Temporal profile for Silent Wanderers

Table 5.4: Interplay between stages of problem solving strategies and behaviors of speech, gaze, and affect

When employing a global problem solving strategy			
Behavior	Expressive Explorers	Calm Tinkerers	Silent Wanderers
Speech	High	High	Medium
Gaze towards partner and/or robot	Lowest	Lowest	Highest
Gaze towards the screen	High	High	High
Affect	Medium Negative	Highest both	Highest Positive
When employing a local problem solving strategy			
Behavior	Expressive Explorers	Calm Tinkerers	Silent Wanderers
Speech	Highest	Highest	Highest
Gaze towards partner and/or robot	Highest	High	Medium
Gaze towards the screen	Lowest	Medium	Medium
Affect	Highest Negative	Medium both	High Positive

aggregate analysis which suggested that these learners adopt a local problem solving strategy, this analysis suggests that these type of gainers too go through two states of productivity: a less probable state of local problem solving and a more probable state of global problem solving. In the state of local problem solving, *Calm Tinkerers* do most removal actions, particularly removing each other's last added edges, show lesser negative emotions, and their speech is at its highest. In the state of global problem solving, these learners do more addition actions, are more frustrated and their speech decreases but is still relatively high. Contrary to *Expressive Explorers*, we find that in *Calm Tinkerers* the state of local problem solving is more likely to occur earlier in the activity than the state of global problem solving, suggesting that these learners begin with a local problem solving approach. However, similar to the *Expressive Explorers*, these learners change in problem solving strategies is also accompanied with an increase in negative emotions.

In the state of non-productivity while trying to understand the activity, the *Calm Tinkerers* gaze at their partner as well as the right side of the screen is high. In the state of global problem solving, while adding edges, the learners gaze on both sides of the screen is high. In the state of local problem solving, while removing edges, including each others' edges, the learners gaze at the robot and the left side of the screen is highest. We must note that the only difference between the left and the right sides of the screen is that if a previous solution is opened, it is displayed on the right side; whereas, the information on the total number of nodes and the number of edges currently present on the map is on the left side. Similar to *Expressive Explorers* both productive states are characterized by high speech signifying good collaboration is both states (Viswanathan & Vanlehn, 2018). Further, similar to *Expressive Explorers*, the speech in the local PS state is highest and this is likely because this state involves the highest removal of each others' edges which requires discussion and agreement among both partners, thus increasing the speech activity. Lastly, different from *Expressive Explorers*, these learners still have a medium probability to fall back to the unproductive state from the state of global problem solving strategy.

Silent Wanderers

Similar to the two gainer groups, the *Silent Wanderers* start with the highest probability at a non-productive state characterized by more technical help-seeking, high silence, and low actions with the learning activity. They go through a more probable state, occurring in the middle of the activity (suggested by medium normalized time), where they adopt a global problem solving strategy in which their speech increases and they do more addition actions. However there is no change in their reflective actions in this state, either in terms of looking at their previous solutions or removing their own or their partners added edges. Even from this state of productivity, they can still fall back to the state of non-productivity with a high transition probability. In the less probable state, which is more likely to occur towards the end of the activity and is characterized by a more local problem solving strategy, non-gainers do more removals and few additions. We may infer that this is a more reflective phase although their reflection, unlike the gainers, does not include a significant increase in the use of the solution history or each other's last actions. However, this state is characterized by their highest speech.

In terms of gaze, in the non-productive state while trying to understand the activity these learners gaze at the left side of the screen is highest and this could be because the information on the number of nodes and number of edges currently present on the map is located on the left. In the more probable state of doing additions their gaze at their partner and the right side of the screen is highest, where the history is also located and it could be that learners were accessing their past solutions. Finally, in the less probable state of removing edges their gaze at the right side of the screen is high, which could again indicate learners accessing their history. Interestingly, we find no difference in the learners frustration between the three states, indicating that their negative emotions were relatively stable regardless of whatever they were doing in the activity. Thus our analysis reveals that non-gainers go through a "slower" learning pathway characterized by an intermediate semi-productive state where actions on the activity and speech increases, but reflection is generally unchanged. While they do reach a productive state of reflective problem solving and higher amount of discourse, it is reached late in the activity. However, this suggests that given time even the non-gainers could achieve higher learning gains since once they reach this productive state, similar to gainers, the probability of going back to the non-productive states is low. We hypothesize that the lack of reflection in the intermediate state could be the reason why non-gainers do not have higher learning gains as it is known that reflection plays a crucial role in learning from problem solving (Do-lenh, 2012; Hmelo-Silver, 2004).

Together our findings suggest that not only are there multiple behavioral profiles of learning (Chapter 4), there are multiple behavioral pathways for learning, and learners who have learning gains do not adopt a single problem solving strategy, global or local, but indeed a combination of both. Further, they modify strategies based on the status of the problem solving and feedback obtained from the environment. Our findings also suggest an interplay between PS strategies and other behaviors which we explore in-depth in the next section.

5.5.2 Interplay between PS Strategies and other behaviors

Now that the temporal learning profiles have been explained for each group, we would like to focus on how speech, affect and gaze evolve for each of these groups and interplay with the global vs the local problem solving strategies i.e., while performing addition actions predominantly or when removal actions are more frequent, respectively. This *interplay* between the *problem solving strategies* and behaviors of *speech*, *gaze*, and *affect* is shown in Table. 5.4, which has been synthesized based on our results described in section 5.5.1. We note that this table does not include those behaviors that stayed consistent for a certain group of learners between the two strategies. For example, for *Silent Wanderers*, the fact that we do not see negative affect in the table indicates that there were not any significant oscillations for their negative valence between the two strategies, i.e., their negative emotions were more consistent irrespective of which problem strategy they used.

When doing global problem solving consisting predominantly of additions, the two gainer groups *Expressive Explorers* and *Calm Tinkerers* have high speech, while *Silent Wanderers* speak relatively less. In this phase, the two gainer groups gaze at their screen is high, while the gaze towards their partner or the robot is lowest. On the other hand, for the non-gainer group *Silent Wanderers*, while the gaze towards the screen is high, their gaze towards their partner is highest in this phase. Lastly, in terms of affect, *Expressive Explorers* express medium level of negative emotions, *Calm Tinkerers* display both highest levels of positive as well as negative emotions in this phase, while the non-gainer group *Silent Wanderers* are associated with their highest levels of positive emotions in this phase.

Next, we observe that when using local problem solving strategy, i.e., more removals, an action indicative of reflection, each group's speech activity is at their highest. In terms of gaze behavior, the two gainer groups *Expressive Explorers* and *Calm Tinkerers* gaze at their partners as well as the robot is high in this phase, while *Silent Wanderers* gaze towards their partner is lesser. Furthermore, *Expressive Explorers* gaze towards the screen is the lowest in this phase, while the other two groups gaze at the screen is medium. Lastly, *Expressive Explorers* show most negative emotions during this strategy, *Calm Tinkerers* are associated with medium emotions, while *Silent Wanderers* lean towards high positive emotions while removing.

It is interesting to note that irrespective of the phase of problem solving, both gainer groups maintain a high level of verbal interaction as opposed to the non-gainer group *Silent Wanderers* who speak less during global problem solving and speak the most while in the local problem solving phase. This suggests that verbal interactions are important to be maintained during both the global and local problem solving phases, i.e. both when making additions, as well as when doing removals. The need for communication itself is not surprising as the collaborative problem solving task requires learners to share information for building a common ground and improving their understanding to construct a solution, monitor and reflect on the solution (Barron, 2003; Chang et al., 2017; Hausmann et al., 2004; Roschelle & Teasley, 1995). Our analysis reiterates the need for communication throughout collaborative problem

solving, regardless of the PS strategy being applied. Nevertheless some phases may demand a higher level of interaction between partners. For instance, literature suggests an increase in interaction between participants during phases of socially shared regulation of learning which involves reflection, monitoring the solution (Isohätälä et al., 2017; Rogat & Linnenbrink-Garcia, 2011; S. Sinha et al., 2015). We also find similar behaviors in that we see an increase in speech activity of all learners in their most reflective phase of problem solving, which in our case is the local problem solving that involves continuously evaluating whether an added edge satisfies the requirement of minimising cost and removing it if not. This requires partners to share the information on their respective screens and discuss it with respect to the overall solution, thus leading to increase in speech.

In terms of affect, all groups oscillate between different affective states and/or different *levels of affect. Expressive Explorers* oscillate between medium and very high negative valence levels during global and local phases respectively, i.e., *showing a higher frustration during the local strategy*. On the other hand, the second type of gainers, *Calm Tinkerers* oscillate between higher to medium level of arousal, with a mix of both positive and negative valence, when moving respectively between global and local problem solving, i.e, *displaying higher levels of both excitement and frustration during the global strategy*. Lastly, for *Silent Wanderers*, the oscillation is more in terms of arousal, that shifts between their relative levels of highest to high positive valence between global and local problem solving, respectively, i.e, *being more excited during global problem solving*. The changing dynamics of affective states over the entire problem solving is supported by the work of D'Mello and Graesser, 2012; however, what is interesting is that *both gainer groups experience negative emotions during both global and local problem solving phases*. A meta-analysis of discrete affective states during learning with technology indicates that negative states such as anger, contempt, sadness, anxiety, fear, etc. are relatively infrequently experienced when students engage with technology-enhanced learning contexts (D'Mello, 2013). However, these learning contexts are guided discovery learning contexts that usually employ success-driven scaffolding to nudge the learners towards the correct solution. T. Sinha, 2021, in a recent work suggested that in a problem-solving followed by instruction (PS-I) context, where the problem-solving phase is "naturally designed to be ill- structured and afford the generation of multiple suboptimal solutions (Kapur & Bielaczyc, 2012)", some levels of negative emotions can in fact be beneficial as they can "keeps one alerted of challenges requiring more focused attention, and assists in comprehending conflicting information (Ivtzan et al., 2015; Kashdan & Biswas-Diener, 2014)". Since our open-ended activity is also designed as a PS-I activity, the surfacing of absolute medium levels of negative emotions among gainers (the mean values can be seen in the Tables in the appendix B; note that the labels highest, high, medium, low, lowest are relative within a group) can be considered as supporting what was reported in T. Sinha, 2021. In this work, we additionally point out *when* negative emotions increase during problem-solving, relative to other phases, for the different types of gainers.

Another point of interest is that while the interplay between problem solving strategy and affect was highlighted in our previous Chapter 4, this work highlights that a particular affect

is not strictly associated with a *type* of problem solving strategy but it also depends on the *phase* of the activity. A particular problem solving strategy applied at the later stages of the activity can lead to more negative emotions than would be otherwise observed. In D'Mello and Graesser, 2012, the authors highlight that moving from a state of equilibrium or flow to a state of disequilibrium results in negative emotions such as confusion and frustration. Our findings of gainers emotions also suggests a similar behavior; for instance when *Expressive Explorers* change strategies from a global to a local one, it is accompanied by an increase in negative emotions and when *Calm Tinkerers* shift from a local to a global strategy they show an increase in negative emotions. This *change* in negative emotions in not very prominent among *Silent Wanderers* which could be because they did not pay as much attention to the task at hand or notice the gaps in their prior knowledge and the need for reflection (T. Sinha, 2021).

Oscillation of gaze between the partner and the screen, and the robot and the screen, is particularly interesting as we observe that for both gainer groups, they look the least at their partner or at the robot when employing the global PS strategy but highest during the local PS strategy. On the contrary, the non-gainer group looks more to their partner and the robot when exhibiting global PS strategy compared to the local PS strategy. Literature suggests that gaze is a means of action monitoring, predicting intention, action co-ordination and planning in order to establish a common ground that can lead to better collaboration (C.-m. Huang et al., 2015; Sebanz et al., 2006). Together our findings and literature suggest that in an environment that has both social (a partner) and task elements (screens), looking at your partner during the local PS strategy, which involves mostly removing what the team has already built and requires agreeing on which edges to remove, can support joint action. Since in this work we do not distinguish between moments when both partners are looking at each other and when one partner is looking at the other (both are considered when computing the feature "gaze at partner"), eye gaze could either be a way to confirm agreement on a bilaterally decided course of action or a way to negotiate to reach a consensus when a unilateral decision was taken. On the other hand, during the global PS strategy which involves series of additions, it is more productive to look at the screen rather than at the partner as the plan is already agreed on (global reflection/planning).

5.5.3 Connections to Computer-supported Collaborative Learning Literature

Within CSCL literature the temporal analysis of computer-supported collaborative learning (Lämsä et al., 2021) has predominantly focussed on the content of learners verbal communication/interaction/discussion and how it evolves during the learning activity, with the non-verbal activities such as actions within the technology-based learning environment, serving to complement the analysis of verbal communication. In our work, we employ multimodal features to understand how pairs of students learn by working on an open-ended scripted collaborative problem-solving activity. For this, we consider the pair as a single unit and examine how their collective behaviors (speech activity, problem-solving actions, eye

gaze and affect) change across the activity as they learn by problem-solving. Our analysis does not include any measure of the quality of the verbal discussion, but studies the temporal evolution of this units' learning behaviors using only fully quantitative data and methods. Similar methods have been used in (Martinez-Maldonado et al., 2013b) where the authors were able to distinguish between high and low collaborating groups based on their action and speech sequences and our work adds to this literature by additionally considering affect and eye gaze, and modeling the temporal learning process of different types of learners.

Further, using the quality of speech, with and without problem-solving actions, has allowed researchers to understand how learners temporally regulate their open-ended problem-solving (Chang et al., 2017; Emara et al., 2021; Kapur, 2011; Malmberg et al., 2015; Sobocinski et al., 2017) in face-to-face collaborative conditions. For instance, researchers identified that increased socially shared regulation across time corresponded with increased use of more systematic action sequences (Emara et al., 2021) and higher performance (Malmberg et al., 2015). Similarly, Sobocinski et al., 2017 found that in low challenge sessions, learners transitioned between the forethought and performance phases of self-regulated learning only once, while in high challenge sessions they transitioned between forethought and performance phases more frequently. Chang et al., 2017 identified that successful groups discourse transitioned more frequently from monitoring to formulating and exploring, along with doing exploratory actions, as opposed to less successful groups whose discourse suggested a more trial-and-error strategy. While we did not explicitly identify socially shared regulation, our findings did agree with the above findings in that increased speech activity was overall associated with increased reflective problem-solving actions, both global and local. In addition, our work offers a complementary view of how collaborative open-ended problem-solving proceeds, in terms of problem-solving strategies (local vs global) rather than problem-solving phases (exploring, formulating, planning and monitoring). The global problem solving strategy can be considered as one in which planning, exploring, formulating and monitoring happens on the scale of the entire problem. The local problem solving strategy is one in which the planning, exploring, formulating and monitoring happens on the scale of the next step towards the solution. Our work thus adds to CSCL literature by suggesting that learners seamlessly intertwine these two strategies in their productive collaborative problem-solving, and that neither is at the outset "*better*" than the other.

5.5.4 Implications for Design of Adaptive Learning Interventions

In this subsection, we highlight some implications of the findings discussed above for the design of adaptive learning interventions, both at a broader level for the CSCL community, and at the specific level of the intervention in our setting. To summarize our observations from the temporal profiles, we find that:

1. All learner groups have the highest probability to start with and stay in a state of non-productivity. However, once out of it, all learners have the lowest probability to return to

this state.

2. The non-gainers transition between states of non-productivity and productivity in a smoother manner with an intermediate semi-productive state in terms of time. In contrast, gainers' transitions are sharper, in that they transition from the non-productive state to one of two productive states.

3. *Expressive Explorers* and *Calm Tinkerers* do not exclusively adopt a global or a local PS approach respectively throughout the activity, as suggested by the aggregate behavioral profiles in Chapter 4. This analysis reveals that both these gainer types switch between the two approaches throughout the interaction. One key difference is the stages of the interaction in which the two groups employ the strategies, with the *Expressive Explorers* adopting the global strategy earlier and then the local strategy, while the *Calm Tinkerers* adopting the reverse approach.

4. Further, for the two gainer groups, each of the two problem solving strategies is associated with speech, gaze and affect in a unique way, that is in some ways comparable (speech and gaze) and in other ways opposing (affect). Diving deeper, the relationship of affect with a particular problem solving strategy does not seem to be as straightforward as suggested by aggregate behavioral analysis in Chapter 4. Both types of gainers seem to have increased emotional behavior relative to themselves towards the later part of the interaction irrespective of which problem solving strategy they are using.

Following up from the above observations, (1) suggests that adaptive interventions should start early in the interaction, irrespective of the group. For example, all groups speak the least in the non-productive state and have yet not established either of the problem solving strategies. An effective intervention could then be to try to induce communication between the dyad earlier in the interaction, that eventually could help with mitigating confusion, building a common ground, resolving conflict and pushing the team towards a more reflective set of behaviors, i.e., to follow either a global or a local problem solving strategy.

Further, going back more often (i.e., with a higher probability) into a non-productive state of low speech (as *Silent Wanderers* as well as *Calm Tinkerers* did) might suggest that the students have not yet established a shared understanding of the problem. Without an appropriate intervention, the relevant team may take longer to have productive interactions or transition to a productive state. Such an unstable behavior of moving back and forth between the non-productive and productive states need to be mitigated by an intervention targeted at inducing behaviors that would increase the chances of building a shared understanding. Further, observation of *Silent Wanderers* suggests that it is the lack of reflective actions such as looking back at their previous solutions and observing their own or their partners action, that might be the cause of a delayed shared understanding of the problem. Hence, such actions can be additionally suggested by an intelligent agent if the team is observed to be going back often to a state of lower speech that suggests being in a non-productive state.

Lastly, as highlighted by (3) and (4), identification of a team as following a local or global PS strategy at the early stages of the interaction should be taken with caution. Instead continuous identification of the teams current PS strategy is necessary as the teams shift between multiple PS strategies and each problem solving strategy elicits different speech, gaze and affective behavior in learners. Therefore, it is important to inform the mechanism behind interventions of this sophisticated interplay and suggest interventions accordingly. For example, *Expressive Explorers* increase in their intensity of negative emotions as they move from global to local PS strategy and vice versa for the *Calm Tinkerers*; however, when looking at the time axis, in both cases this increase is towards the later phase of the interaction. Hence, the adaptive intervention system does not always need to mitigate frustration, especially towards the end of the interaction as this level of frustration may be conducive to more productive behaviors. This can be an interesting avenue for further investigation by the community. As another example, both gainer groups looking more at the partner when moving from global to local PS strategy seems to suggest better collaboration quality; therefore, the adaptive intervention system can try to induce relevant gaze behaviors when the associated PS strategy is detected among learners potentially by sharing gaze among the peers as has been shown to be effective (Schneider et al., 2018).

Concluding on our discussion, in this chapter we contribute by applying an HMM based methodology to model and understand the *collaborative learning process* of gainer and non-gainer teams. However, there are some limitations with the current analysis some of which have also been highlighted in previous chapters. Firstly, in order to generalize the outcomes and inferences to collaborative settings in open-ended environments, there is a need of carrying out even more extensive studies, i.e., with more teams. Then, the current data is skewed when it comes to non-gainer teams, that is we have lesser non-gainer teams in our data than gainer teams and that can add to making our results less straightforward to generalize. Lastly, since the study is done at international schools in Switzerland, the students are from a selective pool coming from a certain economic and social background; hence, this requires us to be careful about the group we generalize it to.

Now that we have explored and investigated in depth the collaborative behavioral profiles, both at an aggregate level as well as at the temporal level, extracted through our forward and backward technique proposed in the *Productive Engagement* framework; in the upcoming chapters, we will focus on how we can utilize this information to design our envisioned adaptive goal-centric robot for real-time interventions.

6 A Speech-based Productive Engagement Metric for Real-time Human-Robot Interaction in Collaborative Educational Contexts

"Simplicity is a great virtue but it requires hard work to achieve it and education to appreciate it. And to make matters worse: complexity sells better."

— Edsger Wybe Dijkstra

Our motivation to move forward is as follows: to construct a simplistic measure that is both *sufficient* to gauge the state of the learners as well as *efficient* in real time. In this Chapter, we propose and validate a metric for the real-time analysis of the behaviour of learners, allowing to assess whether they are engaged in meaningful learning behaviours. Specifically, building on our proposed concept of Productive Engagement, that inherently links learning with engagement, we hereby propose methods to quantify and compute it reliably in real-time. The training and testing of the methods is done using the open access PE-HRI-temporal dataset introduced in the previous Chapter.

This work corresponds to the following publications:

J. Nasir, B. Bruno, M. Chetouani, and P. Dillenbourg, "A Speech-based Productive Engagement Metric for Real-time Human-Robot Interaction in Collaborative Educational Contexts." under revision in *IEEE Transactions on Affective Computing* (Nasir, Bruno, Chetouani, et al., 2022).

6.1 Introduction

The learning assessment, i.e., the process by which available information is used to inform the behavior and/or interventions of the robot, is currently approached from multiple angles. According to a survey by Belpaeme et al., 2018, the type of information typically used by the robot for the assessment is *cognitive* or *affective* or both. In some cases, with the aforementioned

type of information in addition to other behavioral data, the robot assesses constructs such as attention, motivation, etc. that, by the HRI and learning analytics literature, are broadly classified under the umbrella of *engagement* or interchangeably referred to as engagement. These constructs are then considered to be representative of learning.

Among the former group, i.e. *cognitive*, in Leyzberg et al., 2014, the robot uses the *in-task performance* of students solving grid-based logic puzzles to build personalization models allowing it to provide the most relevant lesson corresponding to their weakest skill. In a comparison against a non-personalizing approach, the authors saw a one-sigma improvement in the post-tests between the experimental and control group. Similarly, in Ramachandran et al., 2017, a robot that looks at the students' performance in mathematical concepts to provide personalized, non-task related, breaks for cognitive rest, reveals benefits both in terms of efficiency and accuracy in completing the problems. In-task performance is also used in Ramachandran, Sebo, et al., 2019, where the robot maintains a belief about students' mastery of the mathematical problems at hand as well as their engagement in the task using an Assistive Tutor POMDP, which drives its decisions concerning whether and how to provide help. Higher learning gains were found for students who interacted with such adaptive robot, compared to those who interacted with a robot adopting a fixed help action strategy. While these references agree on and highlight the benefits arising from taking *in-task performance* into account in the robot's real-time intervention strategy, this metric might not always be available. Indeed, *in-task performance* as a learning assessment measure assumes that most (if not all) actions taken by the learner during the learning activity can unequivocally be labelled as either right or wrong, an assumption which does not necessarily hold true for constructive and/or exploratory problem solving activities.

Further, in Leite et al., 2014, a social robot, in a chess playing long-term scenario with primary school students, adapts its actions (expressive behaviors and speech utterances) according to the affective state of the child. It does so based on both visual cues in addition to a cognitive source of information that is the state of the game. The robot is evaluated on children's perception of social presence, social support and engagement. Then in the context of second language learning in Gordon et al., 2016, a robot that personalizes its motivational strategies based on its affective reinforcement learning algorithm that takes the child's valence and engagement as input signals, was found to be increasing the valence of the students significantly more than those who interacted with a non-personalizing robot. Another long-term study (Park et al., 2019) in story telling context found that an affective policy trained using reinforcement learning approach successfully personalized to each child and led to a boost in their learning outcomes and engagement. Szafir and Mutlu, 2012 show that a robot that evaluates user attention, in a memory task, using EEG and tries to regain diminishing attention levels improves student recall abilities by 43% over the baseline. Similarly, Bourguet et al., 2020 proposed an affect (using facial images) and behavior (using pose estimation) recognition system to capture the dynamics of a classroom setting whereby the state of each student in the classroom is used by a robot teacher to decide an appropriate action to choose. They found that while such a robot leads to no significant improvement over the baseline conditions when

it comes to understanding of the lectures speech content, it does lead to an increase in, what they define as, engagement. Then, in Ramachandran, Huang, et al., 2019, the authors show how motivation in learning, evaluated through a self-reported questionnaire, is linked with observable sub-optimal help seeking behaviors in a mathematical problem and later demonstrate how a tutoring robot uses this information to improve behavior and learning outcomes. A lot more supervised engagement models, annotated by labels provided by human observers, can be found in non-educational and/or non-HRI settings (Atamna & Clavel, 2020; Foster et al., 2017; Ishii & Nakano, 2010; Ishii et al., 2011; Kim et al., 2016; Salam, Celiktutan, et al., 2017) where the idea is to build them for the robot to automatically assess user engagement in real-time.

As the afore-discussed literature suggests, measures of engagement could be the way to obtain a real-time assessment of a learner's status that is, at least to some extent, independent from the learning activity at hand. However, as was pointed out at the start of this thesis, state-of-the-art models of engagement in learning scenarios often look to increase engagement *with* the agent as a way to mediate the learning outcomes.

In this Chapter, we investigate method(s) for computing *Productive Engagement* reliably and online, allowing a robot to gauge in real-time the level of *engagement that is conducive to learning* and adapt its behaviour accordingly. Please note that real-time in this work means the inputs for audio and video are received every second, for log data every time an action is performed, while the output, information that a robot can use, is calculated every ten seconds.

Concretely, the contribution of this Chapter is twofold: (1) we investigate how real-time metrics allowing a robot to assess whether learners are engaged in meaningful learning behaviours can be constructed, using the *Productive Engagement* framework as a reference; (2) we implement several data-driven methodologies for the computation of such *Productive Engagement metrics*, measuring and discussing their performance on our publicly available dataset. Thus, this makes the Chapter predominantly a methodology one.

6.2 Revisiting Productive Engagement

Here, we would like to remind the reader about the gaps and challenges that motivated the concept of *Productive Engagement* (PE) and then how did we tackle some of those challenges and what is left to be tackled; thus, forming the focus of this Chapter.

1. While it is often assumed that engagement and learning, as previously discussed, have a linear relationship, this is not proven.
2. Many automatic models of engagement rely on human annotators and often suffer from low inter- and intra-rater agreement because of the subjective nature of this construct.
3. Open-ended learning environments typically envision failure as a means towards learning and do not allow for the definition of straightforward in-task performance metrics.

4. Learning happens in real-time and cannot wait just because the sensors or the robot need more time. As a consequence, reliable and fast real-time assessment of the students' learning is crucial.

In our previous Chapters we define engagement *with* the *learning process*, i.e. *Productive Engagement*, as the engagement that should be maximized in order to maximize learning (Nasir, Bruno, Chetouani, et al., 2021). Endowing a robot with the ability to track such an engagement measure would also inform it about when *not* to intervene and let the children be, i.e., ensuring that all of its interventions are effective.

With respect to the different approaches for engagement definition outlined in Chapter 1 section 1.1, since learning is a process (Reimann, 2009), we consider *productive engagement* as a process and consisting of both a social and a task element (Nasir, Bruno, Chetouani, et al., 2021). More specifically, we define the social element of *productive engagement* as *the evolution of the quality and quantity of the verbal and non-verbal social interactions with other entities (learners and robots)* whereas its task element is defined as *the evolution of the quality and quantity of the interaction with the task*.

Concretely, building on our multi-modal dataset (Nasir, Norman, Bruno, Chetouani, et al., 2021), in Chapter 3, we validated, the existence of a hidden link between students' behaviours and their learning; hence, moving away from performance driven metrics catering for (3). Simply put, students closer to understanding indeed behaved differently from those who did not end up learning. More specifically, targeting challenge (1) highlighted above, in Chapter 4, we identify that out of the three distinct sets of behaviours displayed by the students during the activity, two sets are linked with higher learning, i.e., are displayed by those students who are productively engaged with the activity (denoted as *Type 1 gainers* and *Type 2 gainers*), while the latter set is displayed by the non-productively engaged teams, denoted as the *non-gainers*. High speech activity, high speech overlap and fewer longer pauses are the behavioral characteristics that distinguish both types of gainers from the non-gainers. Conversely, the difference between the two types of gainers seem to relate to the interplay between their problem solving strategies and emotional states, with one type of gainers exhibiting more global patterns of exploration-exploitation and expressing more frustration than the other. Based on these characteristics, we termed the two types of gainers and non-gainers as *Expressive Explorers*, *Calm Tinkerers*, and *Silent Wanderers*, respectively.

Going back to the the analysis method of automatic assessment of engagement described in section 3.2.2, we use data-driven methods to surface the labels automatically which can either be used to build models like in (2) or directly be used as an assessment metric depending on the constraints of the setup; hence, reducing to a metric that is generated based on certain cues found conducing to the engagement we require the robot to seek. Keeping the aforementioned methods of automatic assessment of engagement as well as our definition for productive engagement, *quality of interactions* would then refer to *those behaviors that discriminate learning* whereas *quantity of interactions* would refer to *where on the continuum do those*

behaviors lie, i.e., should that behavior be more or less.

This is where we start off by breaking down our research question, in the next section, to navigate better through the problem at hand that is reliable real-time assessment of productive engagement catering for (4).

6.3 Problem Statement

As defined above, *Productive Engagement* can be considered as *behavioral patterns conducive to learning*. In the following, we will focus on each of the terms, also referred to as factors of our problem statement, in this definition individually to clarify the boundaries of the problem at hand, i.e.,:

- ways of representing **learning**
- ways of mapping **behavioral patterns** to learning

6.3.1 Treatment of Learning

Students' learning during the activity can be represented in two different ways:

- as an ***undivided process with a single label***. Intuitively, this approach assumes that learning occurs in unforeseeable modes throughout the activity and thus no discretization is meaningful. Systems for the analysis of learning should thus consider the whole interaction as a single data point. This also implies that such systems would then have one output label for the whole interaction.
- as a ***collection of individually labelled time-stamped windows***. The intuition behind this approach is that learning "builds up over time" during the activity, and thus discretization is meaningful. Under this light, each time-stamped window on its own can provide meaningful information about the amount of progress in learning made within that time. Systems for the analysis of learning thus view a single interaction as composed of multiple, sequential data points. This also implies that each data point would then have an individual output label.

In the latter case, there are multiple ways to consider the link between one window and the next, making it more or less strong. *Non-incremental* approaches envision the data corresponding to a window to only refer to what happened *within* that window, with the timestamp being the only feature retaining information about the ordering of windows. *Incremental* approaches, conversely, envision the data corresponding to a window to refer to everything that happened *until* that window, with all features thus retaining information about the ordering of windows.

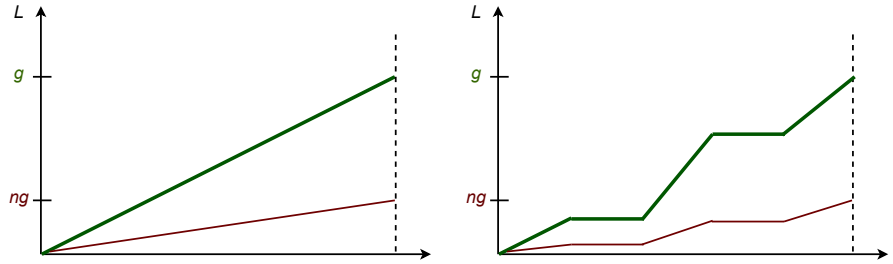


Figure 6.1: Representation of evolution of learning: in the left hand side figures, it is assumed that learning evolves linearly while in the right hand figures, the assumption is that learning evolves non-linearly where t, L, g, and ng represent time, learning, gainers and non-gainers. The thicker/green lines correspond to the gainers.

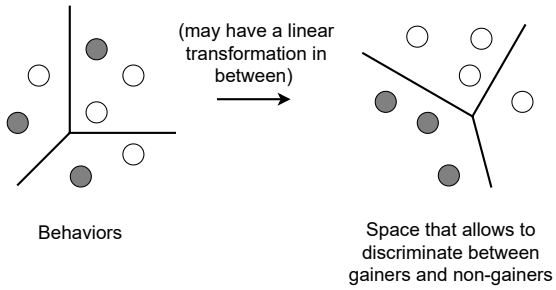


Figure 6.2: Representation of a direct mapping between behavioral patterns and learning.

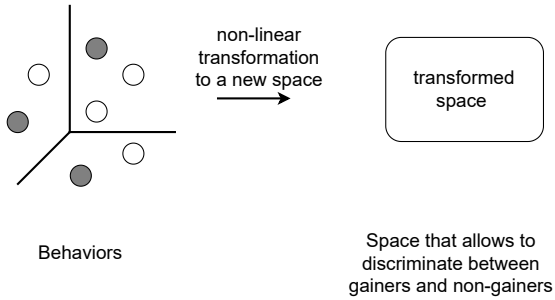


Figure 6.3: Representation of an indirect mapping between behavioral patterns and learning.

While the view of learning as a continuous, indivisible process is closer to the perspective of learning science (Kapur, 2011; Reimann, 2009), interventions need to occur *before* the end of the activity to be meaningful. Hence, the second approach is closer to the perspective of practitioners and designers of systems, such as social robots, aiming to have a direct impact on the learning process while it occurs. The use of incremental approaches might be an effective compromise between the two perspectives.

When we talk about the ways in which behavioral patterns can be conducive to learning, we can assume learning to evolve either ***linearly*** or ***non-linearly***, as represented in Figure 6.1.

To elaborate, linear modelling would entail that we assume the learning process to be linear; and hence, use the information available *at the end of the process*, i.e., if the learners ended up learning or not, to generate a label for the full sequence of the team (when looked as a whole) as well as for each window in that sequence. With such labels, classifiers can tell if a team would end up as gainers or not. However, even those who are predicted to be gainers may oscillate between productive and unproductive phases, i.e., here the assumption is that learning evolves non-linearly as represented in the right-hand side graphs in Figure 6.1; hence, extracting information at the end of the process is not enough. For this, we require a continuous quantity for each team that is affected by the information available in the current point in time.

For the former case, we make use of our previous results in Chapter 4 where the label generating approach returns one label for each team. The label indicates if the team will end up being productively engaged or not. On the other hand, to generate a continuous quantity, we utilize those features that we found to be discriminatory between gainers and non-gainers (see Chapter 4 for more details) to generate a value $\in [0,1]$ that tries to quantify productive engagement. This will be elaborated in the upcoming sections.

6.3.2 Treatment of Behavioral Patterns

As shown in Figures 6.2 and 6.3, there are also two ways to envision the mapping between behavioural patterns and learning:

- the ***direct mapping*** approach, as shown in Figure 6.2, envisions differences in behaviors to directly relate to differences in learning. Under this assumption, the analysis of differences in learning can be done in the space of the behavioural data itself via clustering methods such as K-nearest neighbours (KNNs) or in a transformed space but such that the transformation is linear, for example, with dimensionality reduction methods such as PCA followed by clustering methods.

- the ***indirect mapping*** approach, as shown in Figure 6.3, envisions differences in behaviors to indirectly relate to differences in learning, thus necessarily requiring a transformation from the space of the behavioural patterns to the space of learning where we

Table 6.1: Characterization of Productive Engagement

Characterization of PE	**Description**	**Outcome**
PE Labels	By generating one label for each sequence or all windows in that sequence as explained in Chapter 4	A team is predicted to *end up* as a productively or non-productively engaged team.
PE Score	By generating a different score for each window in a sequence	A team is assessed/predicted to be productively or non-productively engaged at *this moment in time*.

Table 6.2: Factors of our problem statement and their associated Characteristics

Factors	**Values they can assume**	**Assumption(s)**	**Methods**	**How is PE characterized?**
Mapping between behaviors and Learning	Direct	Raw input data itself can surface meaningful representations	Clustering with or without dimensionality reduction techniques like KNN, PCA, respectively	Using either direct representations of behavioral clusters or PE Labels
	Indirect	Mapping to a new space can introduce generalizability	Classification techniques like SVM, RF, LSTM	
Learning	Undivided process with a single label	Learning occurs in unpredictable times and modes	Systems that analyze sequential data (classifiers)	PE Labels
	Collection of individually labelled time-stamped windows	Each smaller window holds information about the process of learning	Systems that analyze discrete data (classifiers, regressors, dynamic assessment, clustering)	PE labels/PE Score
Evolution of Learning	Linear	Learning process is linear	Classifiers	PE Labels
	Non-linear	Learning process evolves non-linearly	Regressors, dynamic assessment	PE Score

have little to no idea about how the transformation looks like and the transformation is non-linear. Classification methods such as Support Vector Machine (SVM), Random Forests (RF) and Long Short-Term Memory (LSTM) Neural Networks inherently perform such transformation.

Tying everything altogether in this section, we first recap the ways of characterizing *Productive Engagement* (PE) in Table 6.1. Then, all the factors in the problem statement, values that they can take, the assumptions those values carry, methods to formalize the values, and how PE is being characterized for each of them is summarized in Table 6.2. Briefly, the afore-discussed section provides the motivation for selecting the various methods for our analysis.

Table 6.3: Log features from our PE-HRI-Temporal dataset

Feature Name	Description
	Log Features
T_add/(_inc)	The number of times a team added an edge on the map in that window/(until that window)
T_remove/(_inc)	The number of times a team removed an edge from the map in that window/(until that window)
T_ratio_add_rem/(_inc)	The ratio of addition of edges over deletion of edges by a team in that window/(until that window)
T_action/(_inc)	The total number of actions taken by a team (add, delete, submit, presses on the screen) in that window/(until that window)
Redundant_exist/(_inc)	The number of times the team had redundant edges in their map in that window/(until that window)
T_hist/(_inc)	The number of times a team opened the sub-window with history of their previous solutions in that window/(until that window)
T1_T1_add/(_inc)	The number of times either of the two members in the team followed the pattern consecutively: I delete an edge, I add it back in that window/(until that window)
T1_T1_rem/(_inc)	The number of times either of the two members in the team followed the pattern consecutively: I add an edge, I then delete it in that window/(until that window)
T1_T2_add/(_inc)	The number of times the members of the team followed the pattern consecutively: I delete an edge, you add it back in that window/(until that window)
T1_T2_rem/(_inc)	The number of times the members of the team followed the pattern consecutively: I add an edge, you then delete it in that window/(until that window)
T_help/(_inc)	The number of times a team opened the instructions manual in that window/(until that window)

Table 6.4: Video features from our PE-HRI-Temporal dataset

Feature Name	**Description**
	Video Features: Affective states and Gaze
Positive_Valence/(_inc)	The average value of positive valence for the team in that window/(until that window)
Negative_Valence/(_inc)	The average value of negative valence for the team in that window/(until that window)
Difference_in_Valence/(_inc)	The difference of the average value of positive and negative valence for the team in that window/(until that window)
Arousal/(_inc)	The average value of arousal for the team in that window/(until that window)
Gaze_at_Partner/(_inc)	The average of the the two team member's gaze when looking at their partner in that window/(until that window). Each individual member's gaze is calculated as a percentage of time in that window/(until that window).
Gaze_at_Robot/(_inc)	The average of the the two team member's gaze when looking at the robot in that window/(until that window). Each individual member's gaze is calculated as a percentage of time in that window/(until that window).
Gaze_other/(_inc)	The average of the the two team member's gaze when looking in the direction opposite to the robot in that window/(until that window). Each individual member's gaze is calculated as a percentage of time in that window/(until that window).
Gaze_at_Screen_Left/(_inc)	The average of the the two team member's gaze when looking at the left side of the screen in that window/(until that window). Each individual member's gaze is calculated as a percentage of time in that window/(until that window).
Gaze_at_Screen_Right/(_inc)	The average of the the two team member's gaze when looking at the right side of the screen in that window/(until that window). Each individual member's gaze is calculated as a percentage of time in that window/(until that window).

Table 6.5: Audio features from our PE-HRI-Temporal dataset

Feature Name	**Description**
	Audio Features: Speech
Speech_Activity/(_inc)	The average of the two team member's speech activity in that window/(until that window). Each individual member's speech activity is calculated as a percentage of time that they are speaking in that window/(until that window).
Silence/(_inc)	The average of the two team member's silence in that window/(until that window). Each individual member's silence is calculated as a percentage of time in that window/(until that window).
Short_Pauses/(_inc)	The average of the two team member's short pauses over their speech activity in that window/(until that window). Each individual member's short pause refers to a brief pause of 0.15 seconds and is calculated as a percentage of time in that window/(until that window).
Long_Pauses/(_inc)	The average of the two team members long pauses over their speech activity in that window/(until that window). Each individual member's long pause refers to a pause of 1.5 seconds and is calculated as a percentage of time in that window/(until that window).
Speech_Overlap/(_inc)	The average percentage of time the speech of the team members overlaps in that window/(until that window).
Overlap_to_Speech_Ratio/(_inc)	The ratio of the speech overlap over the speech activity of the team in that window/(until that window).

6.4 Methods

6.4.1 Dataset

For this paper, we rely on our multi-modal temporal dataset (Nasir, Bruno, & Dillenbourg, 2021b) that is described in Chapter 5. Linking with Section 6.3.1, the features in our dataset are represented in two ways:

- **non-incremental**: A non-incremental type would mean the value of a feature *in* that particular time window while
- **incremental**: an incremental type would mean the value of a feature until that particular time window. The incremental type is indicated by an "_inc" at the end of the feature name.

Without going into the details of the dataset already provided in the last Chapter, here, we only define the features both in their incremental and non-incremental version in Tables 6.3, 6.4, 6.5 as that was not done previously.

6.4.2 Analysis

Clustering

With the clustering approach we treat learning as a *collection of time-stamped windows* using non-incremental features and assume a *direct mapping* between behavioural data and learning. The reasoning behind not treating learning as an *undivided process* (sequence as a data point) here is because with 26 features per team when the total data points are only 32 would lead to a very complex search space for which the data would be insufficient. We follow the following steps for clustering analysis on the data where each window is considered as a data point:

- We start off with Principal Component Analysis (PCA) on the normalized features that returns 4 principle components (PCs) which account for approximately 75 percent variance within features.
- Then, on the 4 PCs, we apply k-mean clustering where the number of clusters k is given by the inertia score. Based on this menthod, k=3 is chosen.
- The 3 clusters are shown in Figure 6.7, as a 3D figure with the first 3 PC's as the 3 axis, where we perform statistical tests to identify the significantly distinct behaviors between the clusters.

Classification

Depending on the classification model employed, classification can be considered both as *direct mapping* or *indirect mapping* as first described in Section 6.3. Furthermore, for classification, we can treat learning either as a *continuous process* or as *sequence of time-stamped windows*. As for the evolution of learning, we start off with assuming that it evolves *linearly*; hence, one label is returned for each team indicating if the team is a gainer or a productively engaged team, i.e., those who end up learning, or a non-gainer team (non-productively engaged) that is those who do not end up learning (Chapter 4). We use the label both for the full sequence and for each window in that sequence.

- Regardless whether the entire sequence or a window is one data point, we feed them as inputs to several classifiers.
- For each classifier, we tune the parameters by a grid search. With the best parameters returned, we perform a k-fold cross validation (k= 5 folds) as well as generate the accuracy and F1-score for the test set.
- In the latter case when each window is a data point, we further differentiate while dividing the training and testing data: windows from all team are *randomly mixed* between training and testing data or windows from some teams are not at all present in the training data (*leave some out* strategy).
- Motivated by Section 6.3, we then test with commonly employed classifiers such as State Vector Machines (SVM), Random Forests (RF), K-nearest Neighbors (KNNs), as well as LSTMs. When using sequences as data points, we utilize the library tslearn (Tavenard et al., 2020) for SVM and KNN and the library keras (Chollet et al., 2015) for LSTMs. In the other case, we use classifiers from the sklearn library (Pedregosa et al., 2011). We tested with various forms of LSTMs such as a vanilla LSTM, stacked LSTM, bi-directional LSTM, and multi-step LSTM. For conciseness reasons, we report the results for the best LSTM model only.

Dynamic Assessment

Going back to our framing of the problem, when learning is assumed to evolve *non-linearly*, we highlighted the need of having a more dynamic assessment and for that, we proposed a productive engagement score. More precisely, we generate a linear combination of those features that are discriminatory between non-gainers and both types of gainers to give us a productive engagement score *PE_Score*. The *PE_Score* in itself can either 1) serve as the continuous quantity a regressor predicts or better even 2) it can stand on its own because to drive an intervention in an educational scenario, dynamic assessment by itself can be enough; we do not need the *prediction* necessarily. For generating this Productive Engagement (PE) score:

- We start off with identifying the features that significantly distinguish both types of gainers from non-gainers. Tying it to our definition of PE in Section 6.2, this links with the *quality of interaction*.
- We then generate an equation which is a linear combination of the identified features (S = Speech, SO = Overlap_to_Speech_Ratio, LP = Long_Pauses). Interestingly, note that while our analysis is multi-modal, the features that end up being significantly discriminating between both gainer types and non-gainers are speech based. Further, the equation is scaled based on variance analysis done for the two groups: gainer teams and non-gainer teams. This variance analysis gives us the contribution of each feature to the variance in the data.

The PE score is then given as follows:

$$PE\_Score = S * \frac{\alpha(SO) + \beta(1 - LP)}{\alpha + \beta} \tag{6.1}$$

where $\beta = \alpha/2$ as LP contributes half as much as SO to the variance in the data. Furthermore, the signs with SO and LP in this equation are based on the fact that gainer teams are linked to higher amount of SO and lower amount of LP and vice versa. This scaling as well as assigning signs links back to the *quantity of interaction*. As mentioned previously, the *PE_Score* can take a value $\in [0, 1]$. We generate this score for each time window within a sequence; hence, giving a sequence of output values for each team. Figures 6.4, 6.5, and 6.6, in which we analyse our equation, further illustrate how our proposed equation for the *PE_Score* behaves as a function of its 3 contributing factors, respectively. As observed in Figure 6.4, *PE_Score* is most sensitive to the amount of speech activity which is a desirable and an expected behavior since the amount of speech is the basis of all the other features too. This is illustrated by the lowest values of the *PE_Score* when the controlling variable is at 0.5 (in the middle sub-figure of Figure 6.4) as we traverse through the space with the values of LP and SO ranging from 0 to 1. Furthermore, Figures 6.5, and 6.6 highlight that the score is positively affected by the increasing amount of overlaps in the speech of the two team members and negatively affected by the long pauses exhibited in the speech, respectively. Note that the effect of long pauses is lesser than the effect overlaps in the speech have on the *PE_Score*, as desired. Specifically, this can be observed clearly when SO is 0 in Figure 6.5 and LP is 1 in Figure 6.6.

Furthermore, we perform three tests to validate if the score can be considered a legitimate form of evaluating the *productively engaged* state of the teams. For the first test, we do a Kruskal Wallis analysis between the averages of the *PE_Scores* as well as all the points in a *PE_Score* sequence for all the gainer teams verses the non-gainer teams. Secondly, we generate a Dynamic Time Warping (DTW) distance matrix where the DTW distance is calculated between the *PE_Score* sequence of each team with every other team. This DTW distance matrix is then given as an input to hierarchical agglomerative clustering to see if the gainer and non-gainer

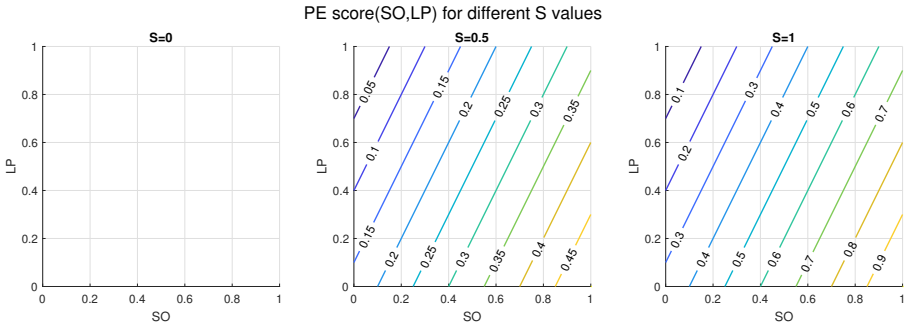


Figure 6.4: Behavior of the proposed *PE_Score* when keeping speech level at 0, 0.5 and 1, respectively

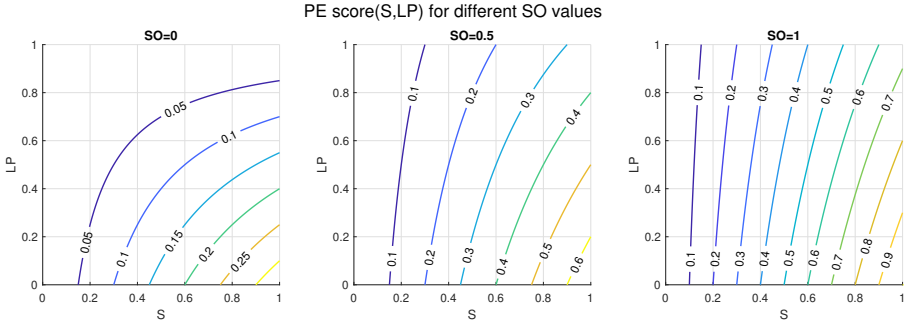


Figure 6.5: Behavior of the proposed *PE_Score* when keeping the Overlap_to_Speech_Ratio level at 0, 0.5 and 1, respectively

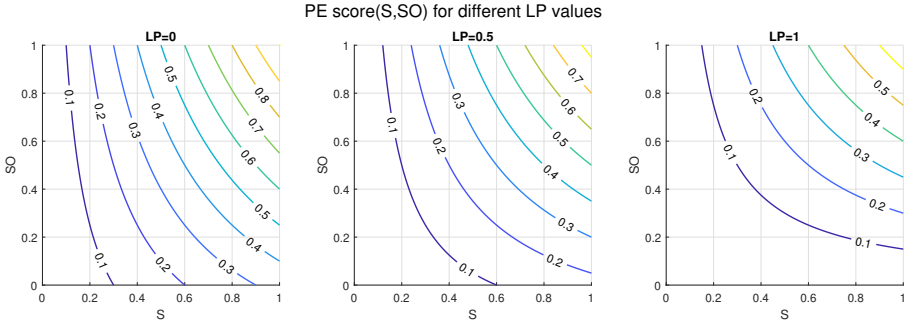


Figure 6.6: Behavior of the proposed *PE_Score* when keeping the Long_Pauses level at 0, 0.5 and 1, respectively

teams are clustered separately indicating that the *PE_Score* sequences of gainer teams are different than those of the non-gainer teams. As a third test, we do a Kruskal Wallis test between the DTW distances of every gainer team 1) with every non-gainer team and 2) with every gainer team, as well as between the DTW distances of every non-gainer team 1) with every gainer team and 2) with every non-gainer team.

Lastly, notice that while the PE labels incorporate information from all modalities (audio, video and logs), the PE score is reduced to one modality (speech), as it aims to incorporate only *necessary* and *sufficient* information to discriminate the desired engagement making it computationally lighter in real-time.

6.5 Results

6.5.1 Clustering

The clustering provided 3 clusters as shown in Figure 6.7. Upon inspection, utilizing Kruskal-Wallis test to identify the significantly different behaviors in each pair of clusters, we find that cluster C_2 is significantly different from the other two in terms of the amount of speech (Speech_Activity) and speech overlap (Speech_Overlap), with the mean values for both features being significantly lower in this cluster than in the other two C_0 and C_1.

- For C_2: Speech_Activity = 0.302, Speech_Overlap = 0.132
- For C_0: Speech_Activity = 0.583, Speech_Overlap = 0.443
- For C_1: Speech_Activity = 0.479, Speech_Overlap = 0.325

C_2 is also associated with lower average normalized time (For C_2: 0.357) relative to the other two clusters (For C_0: 0.577 and for C_1: 0.523) indicating that most of the data points in this cluster occur at an earlier point in time in the interaction. Lastly, most data points in this cluster are contributed by the non-gainer teams. Please note that the values for all features are normalized and hence lie within the range [0,1] as well as the p-values for all the tests are less than 0.01.

6.5.2 Classification Models

Sequence as a data point

In this case, we have $n = 32$. The accuracy and F1-scores for both the validation and test sets can be seen in Table 6.6. We observe that SVM and KNN perform similarly on the test set, with an accuracy of 0.75 and an F1-score of 0.69. While the results appear to be good, upon closer inspection it becomes evident that the two gainer classes (*Expressive Explorers* and *Calm Tinkerers*, which constitute about 80 percent of the total teams) are identified correctly, while

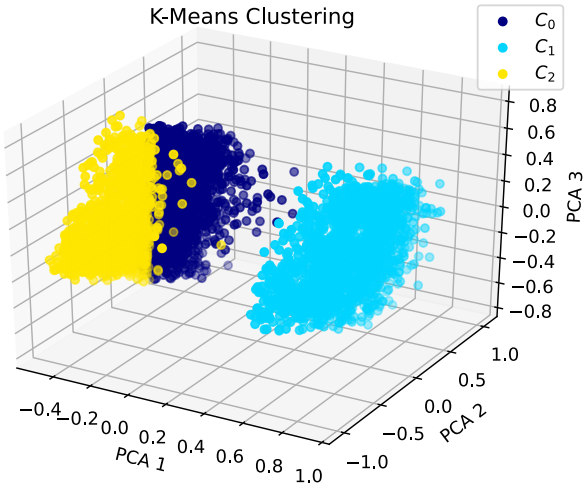


Figure 6.7: k-means clustering on the principle components generated from the behavioural windows.

Table 6.6: Classification Results I for when we consider each sequence as a data point. Please note that there is one multi-variate sequence per team

n=32 sequences				
Classifier	k-fold cross-validation		test-set	
	Accuracy	F1-score	Accuracy	F1-score
SVM	0.75	0.73	0.75	0.69
KNN	0.83	0.80	0.75	0.69
Multi-step LSTM	0.54	0.36	0.57	0.51

the non-gainer class, *Silent Wanderers*, is almost always misclassified. It is also interesting to notice that, contrary to expectation, the results with LSTM variants on sequence data is never above 0.60 both in terms of accuracy as well as F1-score. This may be due to our dataset being too small for a complex model such as LSTMs.

Each time window in a sequence as a data point

When treating each window from the 32 sequences as a data point, we end up with $n = 4676$. Table 6.7 and Table 6.8 show results for both the validation and test sets when using *non-incremental* and *incremental* type features, respectively.

As shown in Table 6.7, the classification results with non-incremental features seem to be marginally better when the data is *randomly mixed* compared to the *leave some out* situation, albeit being generally poor. This suggests that when the classifier has seen some windows from a team, it is able to recognize and thus better classify new incoming windows from the

Table 6.7: Classification Results II for when we consider each window (non-incremental) in a sequence as a data point

n=4676 windows				
Classifier	k-fold cross-validation		test-set	
	Accuracy	F1-score	Accuracy	F1-score
Randomly mixed				
SVM	0.60	0.60	0.59	0.59
KNN	0.66	0.64	0.65	0.65
RF	0.60	0.58	0.61	0.61
Leave some out				
SVM	0.59	0.59	0.60	0.60
KNN	0.67	0.66	0.59	0.59
RF	0.62	0.61	0.55	0.55

Table 6.8: Classification Results III for when we consider each window (incremental) in a sequence as a data point

n=4676 windows				
Classifier	k-fold cross-validation		test-set	
	Accuracy	F1-score	Accuracy	F1-score
Randomly mixed				
SVM	0.98	0.98	0.98	0.98
KNN	1	1	1	1
RF	0.99	0.99	1	1
Leave some out				
SVM	0.98	0.98	0.74	0.75
KNN	1	1	0.78	0.78
RF	0.99	1	0.60	0.59

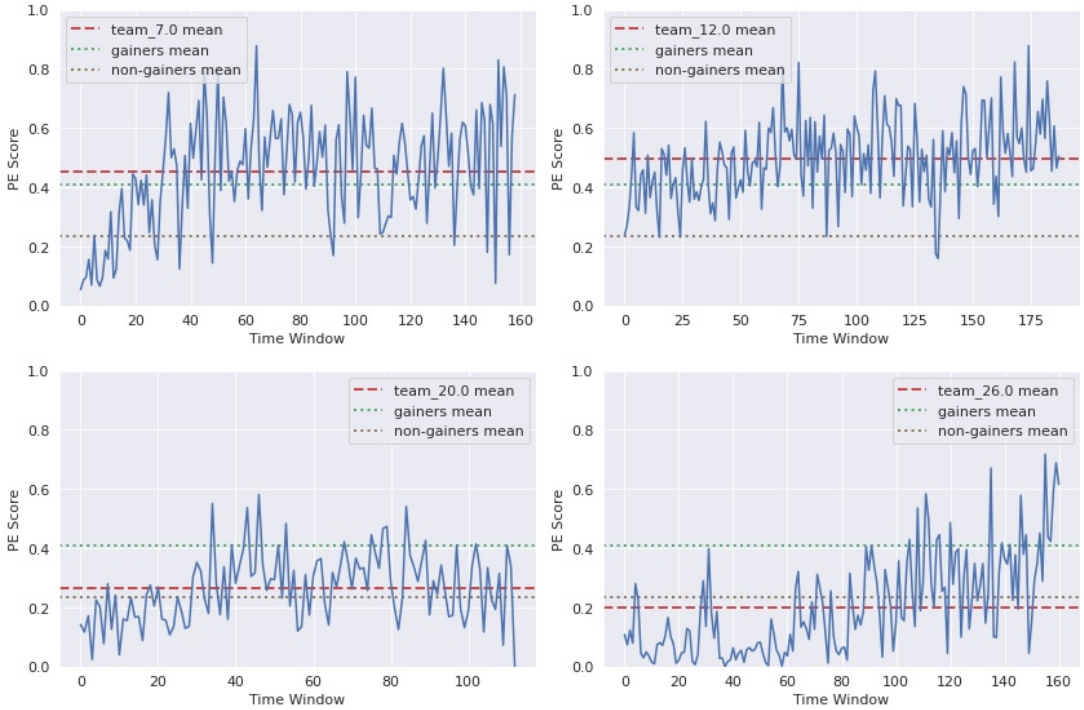


Figure 6.8: Raw PE scores for two random gainer (top) and non-gainer teams (bottom)

same team. With incremental features, the results improve drastically (see Table 6.8) for the *randomly mixed* case. Upon inspection, it is realized that this is due to the *incremental* nature of the features. One can attribute this to *the closeness between values in the windows as the features are incremental* as well as the fact that the training set has windows from all teams. However, in the *leave some out* case, this property cannot be taken advantage of as the teams in the test set do not appear in the training set nor in the validation test. In this case, the performance is still higher than when using non-incremental data.

6.5.3 PE score

In Figure 6.8, we show the *PE_Score* of two random teams from the gainers and the non-gainers group. In general going over all 32 teams, we see a positive slope indicating that the *PE_Score* increases as the interaction proceeds. This is suggestive of more speech overlap among the team members as well as lesser number of long pauses, as the interaction unfolds. Another observation is that generally the slope is more steep for the non-gainer teams. This suggests that with time, the non-gainer teams exhibited a more contrasting behavior with regards to their speech behavior compared to how they were at the beginning of the interaction.

In addition to these qualitative observations, as mentioned in Section 6.4.2, we perform 3 validation tests of the *PE_Score*.

Table 6.9: Validation test 1: Kruskal Wallis tests for the averages (test 1a) as well as for all the points (test 1b) in PE score sequences of the gainers (G) and non-gainers (NG)

test 1a			
Group	mean	standard deviation	n
G	0.40	0.09	26
NG	0.23	0.03	6
p-value= 0.000353			
test 1b			
Group	mean	standard deviation	n
G	0.40	0.18	3741
NG	0.23	0.15	935
p-value= $4.823950e^{-140}$			

Validation test 1

For the first test, we do a Kruskal Wallis analysis between the averages of the *PE_Scores* for the gainer (G) and non-gainer (NG) teams (G vs NG) as well as all the points in a *PE_Score* sequence for the gainer teams versus the non-gainer teams (G vs NG), denoted as test 1a and test 1b, respectively, in Table. 6.9. In both cases, we get statistically significant results.

Validation test 2

As our second validation test, we perform a hierarchical agglomerative clustering technique on a 32 x 32 DTW distance matrix where the DTW distance is calculated between the *PE_Score* sequence of each team with every other team. As shown in Figure 6.9, we observe one of the team being clustered separately and all others being clustered into two clusters when the threshold, minimum distance required to be a separate cluster, is set between the values of 6 and 8 on the y-axis. On closer inspection, we observe that all the teams from the non-gainer group are present in the same cluster (smaller cluster on the left). One of the gainer teams is clustered together with the non-gainer teams, which indicates the need to be careful of outliers.

Validation test 3

Lastly, we again employ Kruskal Wallis, with test 3a being done between the DTW distances of every gainer team 1) with ever non-gainer team (NG) and 2) with every gainer team (G) while test 3b is done between the DTW distances of every non-gainer team 1) with every gainer team (G) and 2) with every non-gainer team (NG). The results are shown in Table. 6.10, and similar to test 1, we find statistically significant results in both.

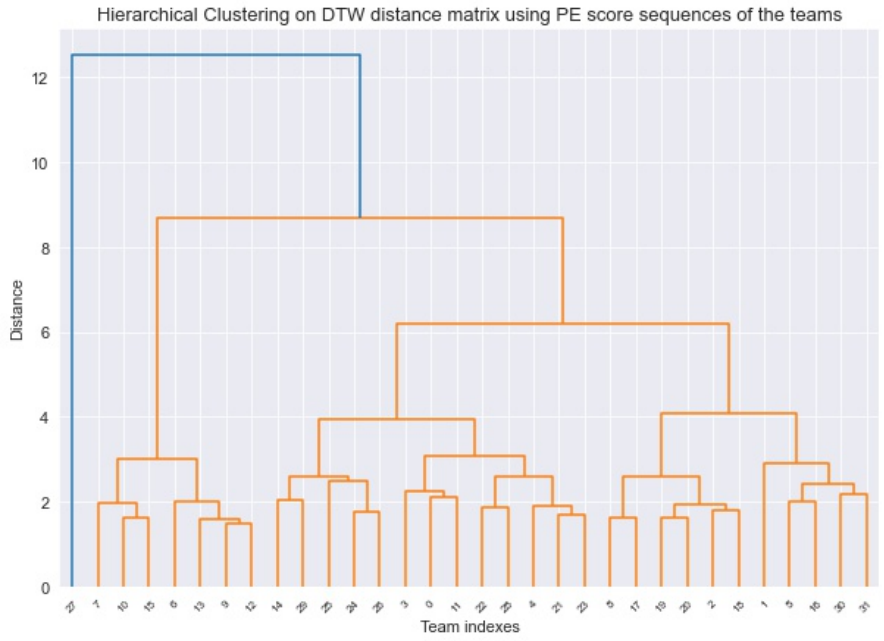


Figure 6.9: Two clusters returned by agglomerative clustering where the numbers on the x-axis represent the team index. The smaller cluster on the left consists all of the non-gainer teams

Table 6.10: Validation test 3: Kruskal Wallis tests between the DTW distances of gainers (G) with the two groups (test 3a) as well as the non-gainers (NG) with the two groups (test 3b)

test 3a			
Group	mean	standard deviation	n
NG	1.51	0.46	676
G	1.69	0.54	156
p-value= 0.000134			
test 3b			
Group	mean	standard deviation	n
G	1.69	0.54	156
NG	1.01	0.47	36
p-value= $5.897200e^{-15}$			

Based on the results of the 3 validation tests, we can conclude that 1) the *PE_Scores* of the gainer and non-gainer teams are significantly different, 2) clustering based on the DTW distance between the *PE_Score* sequences of the teams divides non-gainers and gainers in two separate groups indicating that indeed the way the *PE_Score* of the gainers evolves is different from that of non-gainers, and 3) the DTW distance between each gainer team with a non-gainer team *PE_Score* is significantly different than the DTW distance between each gainer team with another gainer team.

6.6 Discussion

The clustering analysis provides us meaningful representations of behaviors and how these representations link with learning. However, such *direct mapping* still comes with limitations when aiming to assess the *productively engaged* state of the learners, such as the need to set an upper bound on the number of clusters. The insights provided by these methods regarding the learning process, however, can still be leveraged when building an intervention strategy.

Classification methods, falling under *direct mapping* as well as *indirect mapping*, can be used in real-time to assess learners. Classifiers that take a full sequence as an input, by their nature, may not be very effective when trying to predict early on in the interaction if the team would end up as a *productively engaged* team or not, since the model is trained on long sequences encapsulating the entire interaction. However, this limitation can be overcome by treating learning as a *sequence of time-stamped windows* as shown with both the incremental and non-incremental features when assuming that learning evolves linearly. Indeed, such methods can go as far as predicting where the team will end up (gainer or non-gainer) given their current behavior. When assuming that learning evolves non-linearly, we can use a regressor instead in the same way to predict a continuous quantity, for example, a *Productive Engagement* score. This can enable to better predict the *current* state of the learners; however, to drive the intervention, we do not necessarily need this prediction, but in fact just the dynamic assessment itself,i.e., generating the score directly, can be sufficient. This is to say that we get the same outcomes but without having to put up with the computational cost of regression in real-time as that cost would still be more compared to a linear equation that generates the score directly. Hence, we did not train regressors. The validation analyses conducted on the *PE_Score* allow us to conclude that it can serve as an efficient, fast and lightweight way of tracking the teams' "Productive Engagement" state, that can serve as the first indicator of when and whether an intervention is needed as it has the ability to discriminate between moments that might be productively engaging versus those that are not.

6.7 Conclusion

In this work, we first investigate how real-time metrics allowing a robot to assess whether learners are engaged in meaningful learning behaviors can be constructed, using the *Pro-*

ductive Engagement framework as a reference. Secondly, we implement several data-driven methodologies for the computation of such *Productive Engagement* metrics, among which is the *Productive Engagement* score. These methods are then evaluated on our publicly available dataset.

Our key findings and take-away conclusions can be summarised as follows. Firstly, the quantity and quality of speech is *sufficient* for assessing *Productive Engagement*, implying that we could use simpler uni-modal features instead of multi-modal features with great computational benefits for real-time systems. Secondly, teams change over the course of their interaction, i.e., they alternate between moments of high productive engagement with moments of low (or lower) productive engagement. A system that is to provide effective interventions is more concerned with *when* the team is *not* productively engaged. As highlighted in our results section, such dynamics are surfaced in the PE score, unlike methods relying on the PE labels where the prediction is about what the team may *end* up being instead of what they are *in* the moment. At the same time, the PE labels can indicate which of the three identified groups a team may belong to, while the PE score does not provide that information. Lastly, following the previous point and considering the requirements of an intervention system, one can build systems that only assesses the PE score, or only predicts the PE label, or does both. This choice would then affect the computational cost of the system accordingly.

We emphasize here that our focus when developing the *Productive Engagement* score is not to give a *complete* representation of what learning looks like in our multi-modal open-ended collaborative activity context, but rather provide with what is *necessary* or *sufficient* to allow the robot to guide its actions, i.e., *when* the robot should intervene. With this information at hand, the next question is then *what* suggestions should the robot provide to the learners when it identifies their PE score to be low. Eventually, the idea is for a social robot to use this information with an action selection algorithm and then test the effectiveness of its interventions in a user study, that we present in the next chapter.

7 Designing and Evaluating Autonomous Social Robots using the Productive Engagement Framework

In this chapter, we focus on the design of the *Hermione* and *Harry* robots that would incorporate and leverage different levels of information acquired via the *Productive Engagement* framework. With the design of the robots, we try to answer our research question 4 (see Chapter 1) that is *how can a robot make use of the representations of Productive Engagement learned so far to induce the relevant behaviors in the learner?* We also present an HRI user study, that evaluates the two robots, conducted across multiple schools in Switzerland with 136 students aged 9-14 years old. More precisely, in this Chapter, we will make use of the outcomes of Chapters 4, and 6: behavioral profiles, and the *Productive Engagement* score.

7.1 Theoretical Description of the Robots

In Chapter 1, the three robot versions were introduced (Figure 7.1 shown here again) while then in Chapter 2, we went into the details of the role and capabilities of the *Ron* version. To reiterate briefly, *Ron* helps to automate the entire interaction, guides the learners between the various phases of the activity, and provides basic motivational feedback as well as the scores of submitted solutions, and it is least aware of its surroundings. *Harry* has all the capabilities that *Ron* has and additionally it has an idea of *what* behaviors could be useful for learning in the context of this activity. Hence, it suggests randomly one among those behaviors at fixed times. *Hermione* too has all the capabilities of *Ron* and additionally it not only has the knowledge of *what* behaviors could be useful for learning, like *Harry*, but it also has an idea of *when* to suggest a particular behavior and *why* to suggest that specific behavior.

It must be noted that even with these capabilities, both of the two new robot versions are unaware of the solution to the learning task at hand. Therefore, one cannot call them as *informed peers* as typically done in HRI. If we were to have two different axis, as shown in Figure 7.2, one on domain knowledge and the other on knowledge on student behaviors that

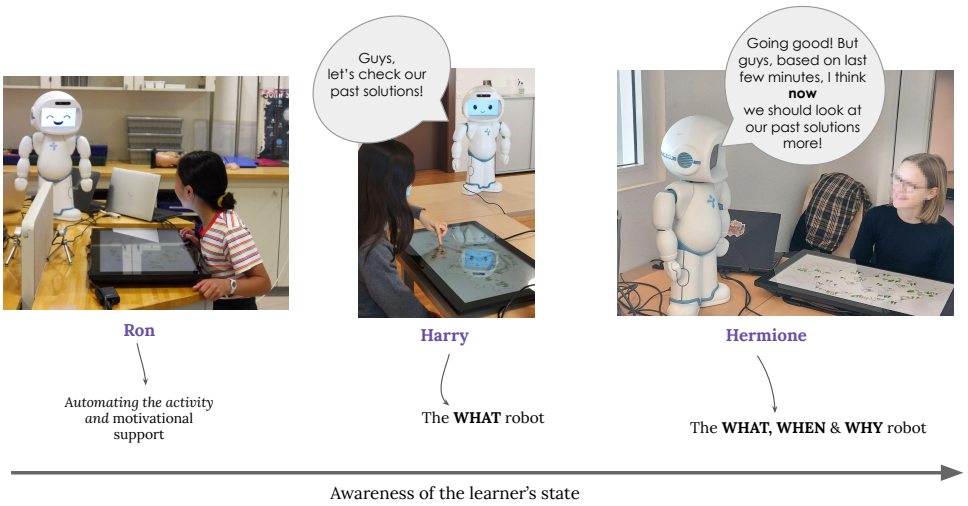


Figure 7.1: Theoretical description of the three robot versions

are conducive to learning, generally the most common robots in HRI happen to fall on the left side of this 2d space, or on the vertical axis, such as *a novice robot*, *an informed peer*, *a tutor*, etc. In our case, we envision robots leaning further on the behavioral knowledge axis, i.e., having an understanding of student behaviors that are conducive to learning; thus classifying the robots as *skilled ignorant peers*. This means they have the needed skills, i.e., an understanding of what behaviors could help us to do better in such a computational thinking and collaborative task, but are as aloof and novice as the students to the solution of the underlying learning problem at hand. In that regard, we show where the three of the robots designed in this thesis are placed on the horizontal axis. We must note that social robots that perceive and try to influence the affective states of the children to provide social support based on the assumption that being in a certain affective state will help improve learning, for example (Gordon et al., 2016; Leite et al., 2014), would also fall on this horizontal axis and may or may not have a y-component depending on the domain knowledge they possess.

The difference between the two robots *Harry* and *Hermione* is mainly motivated by the argument that the *timing* of the intervention is just as important as the *content* of the intervention. While the content of the intervention is shaped by the robot having an understanding of what learner behaviors might help in better understanding the learning concepts, the timing of providing that suggestion is shaped by the robot's ability to access the absence of the desired behaviors at the right time and fast (as highlighted in our challenge number four introduced in Chapter 1) and then respond accordingly. This is represented by the *when* aspect of *Hermione*. In order to experimentally evaluate this aspect, *Harry* then serves as a hard baseline for *Hermione* as it only focuses on the *content*.

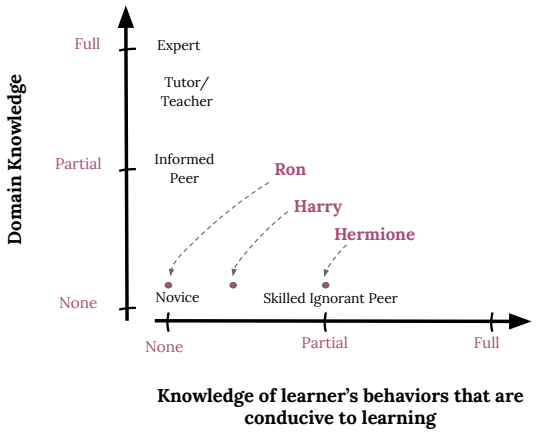


Figure 7.2: The placement of our robots *Ron*, *Harry* and *Hermione* on the space of domain knowledge and behavior knowledge

7.2 Designing Harry and Hermione

In this section, we first describe smaller components of the full architecture that enables the intended interaction for both *Harry* and *Hermione* and then towards the end we bring it altogether. Briefly, in section 7.2.1, we describe the pool of robot behaviors that will be used by both *Harry* and *Hermione* followed by the generation of the *PE score* in section 7.2.2 as well as comparison of learner profiles in real-time in section 7.2.3 where the outcomes of both of these methods will be employed by *Hermione*. Next, we outline the robot control architecture for both of the robots in section 7.2.4 that also includes the action selection techniques for each of the robots. Please note that in this Chapter, by training data we mean the data generated in the *Ron* study (see Chapter 2 and the associated datasets (Nasir, Bruno, & Dillenbourg, 2021a; Nasir et al., 2020a)) which we have used for validating the concept of *Productive Engagement*, for generating behavioral profiles, as well as for designing and generating the PE score in the previous chapters.

7.2.1 Designing Pool of Robot Behaviors

Each robot behavior, that serves as an intervention/suggestion, is comprised of verbal and non-verbal components where the non-verbal component consist of gestures and facial expressions (some of them are shown in Figure 7.3). The content of these interventions is designed in a way as to induce those behaviors in the team that have been found to be conducive to learning in this activity (see Chapter 4 for the details of the behaviors). In this regard, the interventions

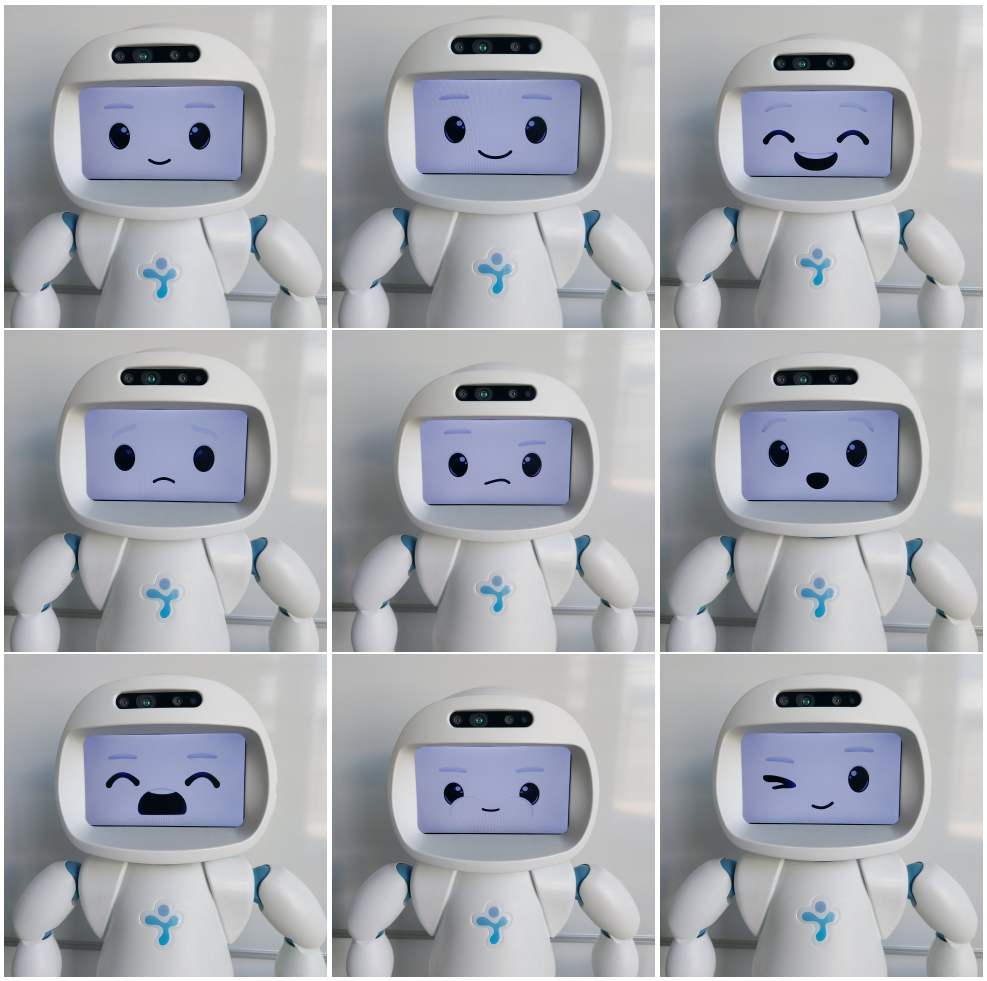


Figure 7.3: Facial expressions of QTrobot in horizontal order from top left corner: neutral, smiling, happy, sad, confused, surprised, bored/yawning, puffing cheeks/being cute, and winking.

can be categorized into the following three types where we list some examples for each type in Table 7.1:

1. **Exploration Inducing:** these suggestions by the robot are explicitly designed to induce the behavior of *Edge Addition* in the learners.
2. **Reflection Inducing:** these suggestions are explicitly designed to induce the behaviors of *Edge Deletion*, *History*, *A_A_add*, *A_A_delete*, *A_B_add*, and *A_B_delete* that we believe are accompanied with some form of reflection (see Chapter 4).
3. **Communication Inducing:** these suggestions are explicitly designed to induce *Speech Activity* or more generally communication among the team.

It must be noted that any of the three types of the suggestions given by the robots can implicitly induce other behaviors in the learners too as all the behaviors are intertwined. For example, *Exploration Inducing* or *Reflection Inducing* suggestions can indirectly induce an increase in communication between the team members that could lead to an increase in the behaviors that these intervention types are explicitly designed for or vice versa. Similarly, *Communication Inducing* suggestions can indirectly induce certain exploration or reflection related actions in the learners as they start communicating more about the next steps or their internal understanding of what may need to be removed or added. While the three types above define the content of the robot's interventions, their style is always supportive, i.e. the robot always conveys the suggestion using positive and supportive language. We have designed 8 interventions to induce communication related learner behaviors. For all other learner behaviors (eg. *Edge Addition* or *Edge Deletion*), we have designed 4 interventions. More number of interventions for the communication related learner behaviors is because it is the intervention type that might get triggered more than the others so we aim to avoid repetitions (details in upcoming section 7.2.4). This gives us a total of 36 interventions with 8, 4, and 24 interventions in the type *Communication Inducing*, *Exploration Inducing*, and *Reflection Inducing*, respectively.

Additionally, we also design a pool of idle behaviors that are non-verbal robot actions, consisting of gestures and facial expressions. These behaviors are triggered every few seconds to give the feeling of a lively robot as well as to provide a more natural feel to the interaction. These behaviors are only executed when no other task of a higher priority is being executed (more on this later). A few examples of such behaviors include: 1) the robot looking side to side to the two team members, 2) the robot scratching its head, 3) the robot looking confused, 4) the robot folding arms behind its back as if observing the situation, etc.

7.2.2 Generation of the PE Score in Real-Time

For the generation of the *PE score* in real-time, we employ the pipeline shown in Figure 7.4. Within a team, each learner's speech is fed through the laptops microphone to a Voice Activity

Table 7.1: Examples of Robot Interventions

Type	Robot's Speech	Facial Expression	Gesture
Exploration Inducing	Guys, we may not be exploring all the rail-tracks. Why don't we connect more gold mines to see how much they cost?	Puffing its cheeks	Putting both arms ahead to gesture while moving head side to side to look at both learners
Exploration Inducing	Are there some tracks we haven't explored yet? If yes, why don't we explore other tracks too?	Smile	Moving head side to side to convey looking at both learners
Reflection Inducing	Guys! I have this idea. Why don't we remove the rail-tracks we do not need? What do you think?	Puffing its cheeks	Moving head side to side while swiping its arm from left to right
Reflection Inducing	Oh hey, may be we have already explored some of these rail-tracks. Should we check our history? I think we did not look at it much in the last few minutes	Smile	Moving head side to side while pointing at the front
Reflection Inducing	Guys, I am a bit lost. I would like you to tell me why is it that you removed the last rail-track?	Confused	Moving head side to side while putting its arms at the back on the hips
Communication Inducing	So Alice, why dont you tell us about what you think we need to do, and then Bob, you tell us what you think.	Neutral expression	Moving head side to side while swiping the right arm
Communication Inducing	So my friends, based on the last few minutes, I feel like we are not communicating much with each other, and that may be important for us to solve this problem.	Confused	Moving head side to side while shifting the left arm in a natural movement

Detector (VAD) for which we used the open-source python wrapper for Google WebRTC VAD[1]. The VAD returns a vector for each team member that consists of voiced and unvoiced frames. These vectors are then used by a feature extraction module to generate all the relevant features such as *Speech Activity*, *Speech Overlap*, and *Long Pauses* (see Chapter 3 for details on VAD, and our proposed feature extraction). Following that, the features are normalized with respect to the training data which is the corpus of data collected in the *Ron* study. Finally, with these normalized features, the productive engagement score is calculated as described by the equation 6.1 in Chapter 6 section 6.4.2.

As will be seen in section 7.2.4, the *PE score* of a team is compared against a threshold which decides if an intervention is needed or not. This threshold is generated as:

$$\tau_{PE} = \frac{a+b}{2} \tag{7.1}$$

where a = average *PE score* of the gainers, teams with higher learning gains, and b is the average *PE score* of the non-gainers, teams with lower learning gains, from the training data. The value for τ_{PE} is then set to 0.32 according to equation 7.1.

[1] https://webrtc.org/

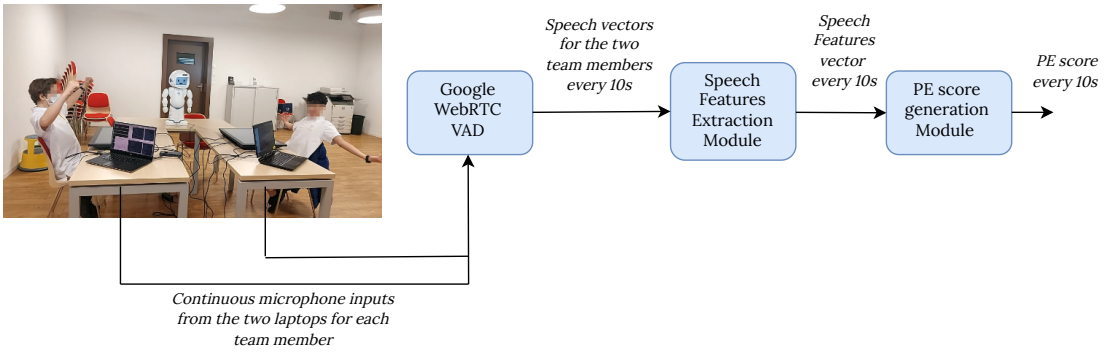


Figure 7.4: Pipeline for the generation of the PE score

We must note that in this pipeline, by real-time, we mean every 10 seconds. Hence, we have a *PE score* for the team every 10 seconds. Before using this pipeline in real-time, we verified that the different way we process audio offline and online does not yield a difference in the output. We do this by comparing the outcome generated by the *PE score* pipeline online with direct sound input through a microphone versus when the same speech content is stored as a .wav file and then fed as an input for post-hoc analysis. This latter method is exactly what was employed in all our post-hoc analysis in the previous chapters. For validation, we used a 30 minute long audio session giving us 180 windows, each of ten seconds, where for each of the windows we have a *PE Score* value generated online as well as offline. Between the two *PE score* vectors, we observe a mean difference (offline vector subtracted from online) of -0.162 and a standard deviation of 0.243.

7.2.3 Profile Comparison in Real-Time

The profile comparison pipeline can be seen in Figure 7.5. Every time an action takes place on a participants application, the application shares that with the log features extraction module that generates all the relevant features such as *Edge Addition*, *Edge Deletion*, *History*, *A_A_add*, *A_A_delete*, *A_B_add*, *A_B_delete* (see Chapter 3). These features are then fed to a profile generation and comparison module that buffers the incoming features. Then every 5 minutes, it averages all the features until that point in time and normalizes them with respect to the training data. After that, the module does an euclidean distance based comparison between the normalized log features vector and the reference log features vectors of the profiles of the two gainer types *Expressive Explorers* (EE), and *Calm Tinkerers* (CT) generated from our training data. This allows to classify the team's current problem solving strategy into one of the strategies identified previously in the thesis, i.e., global exploratory approach, or local exploratory approach.

To elaborate, at time t, the euclidean distance d_t^g of the current feature vector cv_t of a team is calculated with the two reference profiles p_t^g where $g \in [EE, CT]$. Based on the lower distance returned for $g \in [EE, CT]$, the current feature vector cv_t is classified as that gainer profile cp_t

if and only if the distance d_t^g is lower than a threshold τ_t^g. This means not only do we check for which gainer profile is the current feature vector *closest* to but if it is also *close enough* to be classified as that gainer profile. Precisely at each $t \in [10, 15, 20, 25]$ minutes:

$$cp_t = \text{argmin}\,[d_t^g(cv_t, p_t^g)] \iff \min d_t^g < \tau_t^g \tag{7.2}$$

Once a profile is chosen, the profile comparison module returns an ordered list of features, from the one in which the incoming vector is farthest to the reference, to the one where it is closest (more on this in the upcoming section 7.2.4). In the case the distance d_t^g is not *close enough*, the cv_t is not classified as belonging to any gainer type profile.

While the previous paragraphs elaborate on what happens in the pipeline for profile comparison in run-time, the aforementioned thresholds τ_t^g are generated a priori using the training data. For each gainer type $g \in [EE, CT]$:

$$\tau_t^g = \frac{d_t^{intra} + d_t^{inter}}{2} \tag{7.3}$$

where d_t^{intra} is the average intra group (teams within the gainer group) euclidean distance with the average profile vector v_t for the type of gainers at time t and d_t^{inter} is the average inter group (teams within the other gainer group) euclidean distance with the average profile vector v_t for the type of gainers at time t. All the thresholds for τ_t^g are listed in Table 7.2 under the column 'Original Value' where as the column 'After Validation' will be explained later in section 7.4.4. The idea to generate an average profile vector at ever 5 minutes is to incorporate temporal changes. In the training data collected in the *Ron* study, we noticed that for each gainer type, the profile generated every 5 minutes after time t = 10 minutes is consistent with the average profile of the gainer type over the entire activity. The consistency is in terms of the differences between the two types of profiles (EE, CT), i.e., the differences among the two gainer types remain consistent at every additional 5 minute mark. However, the values of each individual feature within a gainer type profile still oscillate. In order to consider that, we choose to check a profile every 5 minutes. For example, the feature of opening up history (*T_hist*) always has a higher value in EE profiles compared to CT profiles at every 5 minute mark. However, within the EE profiles at the different time marks, the value for the feature changes.

Now that we are aware of how a profile is chosen at run-time as well as the thresholds, we conclude this section with an example. Let's suppose at time t = 15 minutes, the distances d_{15}^{EE} and d_{15}^{CT} of cv_{15} are 0.57 and 0.83, respectively. The distance 0.57 of cv_{15} with $g = EE$ is lower and it is also lower than the τ_{15}^{EE} that is 0.613. Hence, at 15 minutes into the game

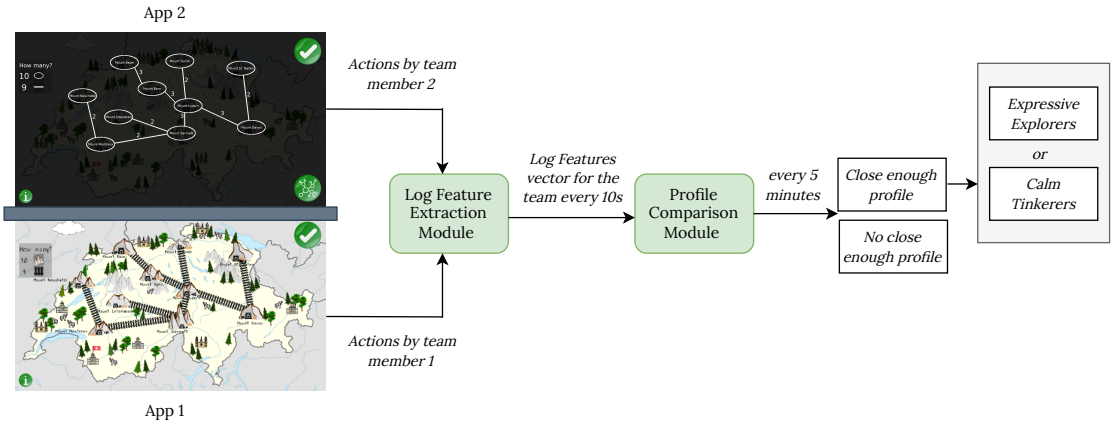


Figure 7.5: Pipeline for the generations of profiles

Table 7.2: Threshold values for *Hermione* where *EE* and *CT* stand for *Expressive Explorers* and *Calm Tinkerers* respectively and the numbers represent the minutes into the game.

Threshold	Original Value	After Validation
τ_{10}^{EE}	0.532	0.638
τ_{15}^{EE}	0.613	0.735
τ_{20}^{EE}	0.703	0.843
τ_{25}^{EE}	0.638	0.765
τ_{10}^{CT}	0.818	0.981
τ_{15}^{CT}	0.759	0.910
τ_{20}^{CT}	0.771	0.925
τ_{25}^{CT}	0.797	0.956

interaction, the team should be classified as exhibiting the global exploratory strategy based on their current actions in the game.

7.2.4 Robot Control Architecture

For both *Harry* and *Hermione*, their behavior is controlled via two modules: 1) a *basic* module, and 2) a *control* module. Each robot module is responsible for specific tasks for the robot where every task is *blocking* as well as has a *priority*. By blocking we mean that once the task has started to execute, it would go on to completion without getting interrupted regardless of any other task being triggered. In regards to priority, a task with a higher priority is selected if two tasks are triggered at the same time. Except the idle behaviors of the robot that have a lower priority, all the other tasks have the same priority. This means that if two tasks of the same priority are triggered at the same time, they will be executed one after the other. A summary on the division of the tasks by the two modules is shown in Table 7.3.

The *basic* module is responsible for automating the entire activity and for handling fixed

events occurring during the game play while the *control* module controls the selection and execution of all interventions by the robot, *Harry* or *Hermione*, as well as the idle behaviors during the game play. Automating the entire activity means, based on the low level activity events such as what is pressed, submitted, etc., the robot takes the team through the different stages of the activity pipeline (explained in Chapter 2 and Table 2.1) as well as gives supportive comments at various points. During the game play, every time a solution is submitted by a team, the *basic* module generates the score and then the robot iterates the score as if it is reading the score from a screen. In addition to this, the *basic* module helps the robot to generate reminders for the team on the possibility of submitting multiple solutions as well as on the remaining time (the game play is for limited time, i.e., 30 minutes). The same module is also responsible for pausing the game whenever an intervention is triggered by the *control* module. As for the *control* module, all the interventions as well as idle behaviors are selected and executed through it. The communication between the two modules allows for handling all the tasks in a smooth manner while taking into consideration the blocking nature as well as the priority of the tasks. Also notice that if the *control* module is removed, the robot is essentially reduced to *Ron*.

These two modules interact with the environment as well as with each other, as shown in the Figure 7.6, through the Robot Operating System (ROS). Precisely, for both *Harry* and *Hermione*, the *basic* module receives information from the two apps as well as from the *control* module. For executing the tasks assigned to the *basic* module when a task is triggered, it sends the chosen robot task to the built-in service controllers of the robot. Additionally, it also sends a message to the *control* module to let it know that the robot resources are busy and thus pausing any other commands the *control* module might want to trigger.

While the *basic* module for the two robots has the exact same functionality, the *control* modules differ. For the *control* module for *Harry*, it implements the algorithm described below in section 7.2.4 and algorithm 1 for selecting an intervention. In the case of *Hermione*, the *control module* receives information from the *PE score* generation module as well as the profile comparison module. With this information, it is able to identify the course of action for *Hermione* according to the algorithm described later in section 7.2.4. In the case of both of the robots, the *control* module also generates idle behaviors, communicates its assigned task to the built-in service controllers of the robot and also notifies the *basic* module about the resources being busy.

After an intervention is executed by a robot, *Harry* or *Hermione*, a pop-up is displayed on the screens of the two students prompting them to individually select either *"I found the robot's suggestion useful"* or *"I found the robot's suggestion not useful"*. A *suggestion_usefulness* score su_i for intervention i is generated which can be 0, 0.5, or 1 depending whether both of the team members found the suggestion not useful, one of them found it useful, or both of them found it useful, respectively. The robot then responds correspondingly to their feedback such as *"Good to know we all agree on the suggestion"* or *"Oh, so you guys do not agree with my suggestion"*. The following metrics, detailed out later in this section, are evaluated and

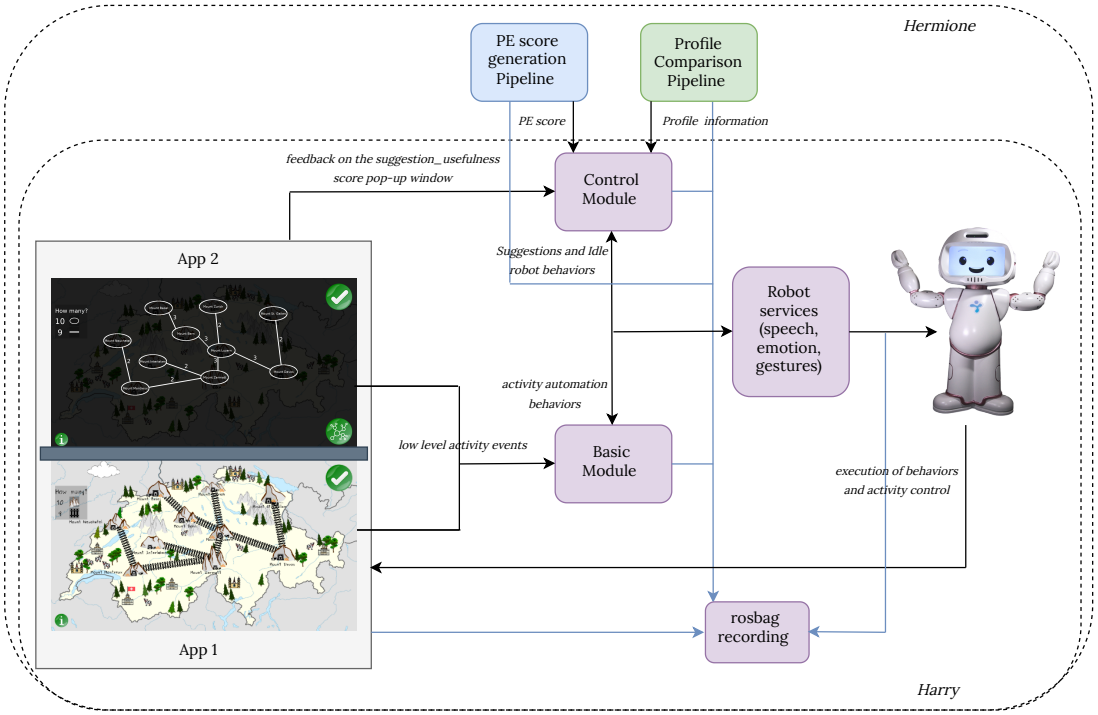


Figure 7.6: Robot Control Architecture for both *Harry* and *Hermione*

stored after each intervention i: the gain in *PE score* PE_i^{gain}, the associated weight w_i, and the *suggestion_usefulness* score su_i. These metrics are then used in the post-hoc evaluation of the two robots.

With the general architecture for both of the robots laid out, we now dive deeper into the robot specific action selection techniques.

Action Selection Technique for Harry

The action selection technique for *Harry*, described in Algorithm 1, is simple: the counter c for the intervention is set to $rand(0,2)$, i.e., every 0 to 2 minutes, an intervention is randomly

Table 7.3: Robot tasks for both *Harry* and *Hermione*

Basic Module			**Control Module**		
Task	**Priority**	**Activity Stage**	**Task**	**Priority**	**Activity Stage**
Automates the entire activity	1	All	Selects and executes Interventions	1	Game Play
Generates score feedback	1	Game Play	Selects and executes Idle behaviors	2	Game Play
Generates reminders	1	Game Play	Manages feedback on suggestions usefulness	1	Game Play
Pauses the game when an intervention is taking place	1	Game Play			

selected from the pool of behaviors and is executed. After an intervention is executed, *Harry* will always wait for a minimum of 2 minutes before the counter is reset to a new value $c = rand(0,2)$. In principle, this means the interventions will happen every 2 to 4 minutes except the first one which can happen within 2 minutes after the start of the game session. The choice of a gap of a minimum of 2 minutes between two interventions is arbitrary and based on the idea of letting a reasonable time pass in between interventions. This time allows to gauge the effectiveness of an intervention keeping in mind the granularity of all the behaviors under question (learner's speech behaviors are measured every 10 seconds; however their actions on the game are sparser than speech behaviors).

Lastly, we must note that while *Harry* does not make use of the two pipelines (*PE score* generation module and the profile comparison module) in its action selection technique, we do still run these pipelines for post-hoc analysis.

Algorithm 1 Action Selection Technique for Harry

1: a = *Exploration inducing* interventions
2: b = *Reflection inducing* interventions
3: c = *Communication inducing* interventions
4: **Every** rand(0,2) minutes:
5: Pick an intervention i by rand(a, b, c)
6: *Harry* executes i
7: Wait for 2 minutes
8: Calculate and store w_i, PE_i^{gain}, su_i

Action Selection Technique for Hermione

The action selection technique for *Hermione* is described in the algorithm 2 whereas a simplified visualization of the technique is shown in Figure 7.7. The decision process for *Hermione* is relatively more complex as it takes a layered approached. At the first layer, in order to decide whether the robot should intervene or not, the exponentially weighted moving average (EWMA) of the team's *Productive Engagement* score is calculated. The EWMA of the *PE score* of the team is calculated with a sliding window of 2 minutes to keep coherent with our choice of a minimum gap of 2 minutes between interventions. This average is then compared to the τ_{PE} and if it is above the threshold, the robot does not intervene in order to not provide unnecessary distractions to the learning process. In the case the *PE score* is lower than the threshold, and depending on the phase of the activity (less than 10 minutes since the start of the game play), or if the game play is at a later stage but the incoming profile cp_t of the team is not close enough to any reference profile, *Hermione* randomly picks one of the *communication inducing* behavior. On the other hand, after 10 minutes into the interaction, when the team is matched to either *Expressive Explorers* or *Calm Tinkerers* (distance returned is less than the τ_t^g), *Hermione* chooses an *exploration inducing* or a *reflection inducing* behavior. This behavior is based on the weakest log action based feature (with highest distance as explained in section 7.2.3) of the matched profile. Initially, all the interventions start with a default

Algorithm 2 Action Selection Technique for Hermione

a = *Exploration inducing* interventions
b = *Reflection inducing* interventions
c = *Communication inducing* interventions
$\forall i,\ w_i = 0,\ S_i = 0$
if *PE Score* $\geq \tau_{PE}$ **then**
    Do nothing
else if *PE Score* $< \tau_{PE}$ **then**
    if $t \leq 10$ minutes **or** $cp_t \neq$ any $g \in [EE, CT]$ **then**
        if $\forall i \in c,\ S_i = 1$ **then**
            Sort i based on w_i in descending order
            Set $S_i = 0$ for $\forall i \in c$
        else if $\forall i \in c,\ S_i \neq 1$ **then**
            Pass
        end if
        Pick the first intervention i of type c such that $S_i = 0$
        Hermione executes i
        Wait for 2 minutes
        Update w_i, PE_i^{gain}, su_i
        Set $S_i = 1$
    else if $t > 10$ minutes **and** $cp_t =$ any $g \in [EE, CT]$ **then**
        if $\forall i \in a$ **or** $b,\ S_i = 1$ **then**
            Sort $i \in a$ or b based on w_i in descending order
            Set $S_i = 0$ for $\forall i \in a$ or b
        else if $\forall i \in a$ **and** $b,\ S_i \neq 1$ **then**
            Pass
        end if
        Identify the weakest log action based feature of the matched profile
        Pick the first corresponding intervention i of type a or b such that $S_i = 0$
        Hermione executes i
        Wait for 2 minutes
        Update w_i, PE_i^{gain}, su_i
        Set $S_i = 1$
    end if
end if

weight of 0. To avoid repeatability, every time an intervention of a particular type is used, it is not used again (a flag S_i is raised to 1) until the robot has made a full pass over the set of that intervention type (exploration, reflection or communication). Once all interventions of that type have been used once by the robot (which is quite rare in a 30 minute activity), all the flags are reset to 0. After that, the interventions of one particular type are chosen based on the decreasing order of the weight associated with the intervention, i.e., the intervention with the highest w_i is chosen. The weight w_i of an intervention when it is used for the ith time for $i \in [1,2,3,..,n]$ is updated as follows:

$$w_i = w_{i-1} + PE_i^{gain} \tag{7.4}$$

where

$$PE_i^{gain} = PE_i^{after} - PE_i^{before} \tag{7.5}$$

where w_0 is the default weight of 0, PE_i^{after} and PE_i^{before} are the values of the *PE score* calculated as an exponentially weighted moving average in the 2 minutes window after and before an intervention, respectively. The choice of an exponentially weighted moving average in this technique instead of a simple average is to give more weight to the recent quantity and quality of the communication between the team members.

7.2.5 Validation of robot behaviors with Harry in a small online study

Before the final *Harry* and *Hermione* study, we did a small online study with *Harry* to mainly get feedback on the content of our pool of robot suggestions to refine them. The aforementioned pool of robot suggestions in section 7.2.1 is a result of the feedback received in this online study.

Due to the COVID-19 pandemic and with the schools not allowing researchers on campus, we needed to adapt the original setup of *JUSThink* for online experiments. The online setup is shown in Figure 7.8. In the online setup, *Harry* and the experimenter are situated in the lab with two laptops and two RGB cameras. The student team as well as a teacher (in the background) are present together in a classroom at the school with two laptops, one for each child. They sit in a way that does not allow them to see each other's screens. Each child is connected to one of the laptops in the lab on Zoom and can see and hear *Harry* on the video and audio stream, respectively. On each of the two laptops in the lab, there is also an instance of the *JUSThink* application running that can be controlled by the student via the remote control setting provided by zoom. This setup allowed for the schools to participate in

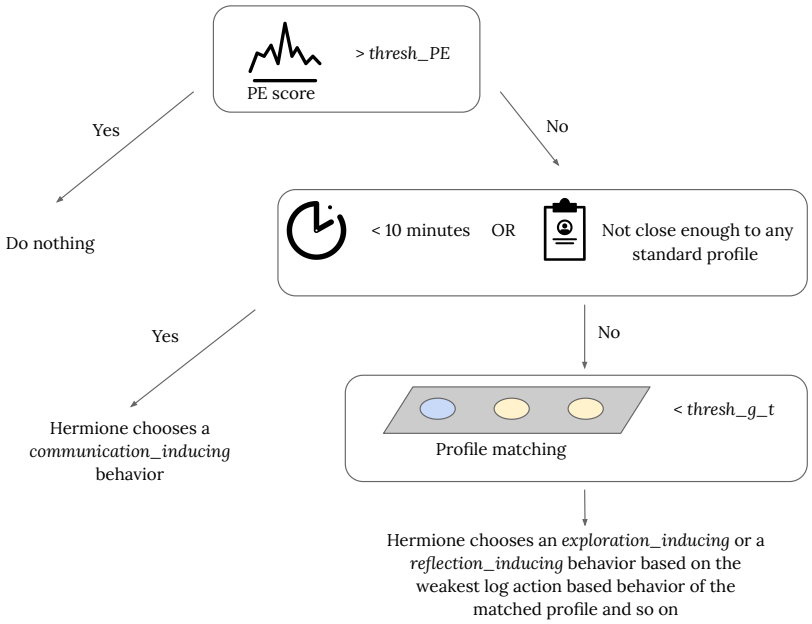


Figure 7.7: A simplified visualization of the action selection technique for *Hermione*

the online study without installing anything at their end except for the video conferencing software Zoom.

22 participants, (gender: 12 males and 8 females with a median age of 10.5), interacted with the activity from two Swiss schools. All the students had prior experience with STEM activities including robotics.

In addition to all the evaluation metrics listed in Chapter 2, we conducted open-ended semi-structured interviews with each team which typically lasted between 5-10 minutes depending on how long their answers were. We addressed questions on similar topics as those in the quantitative questionnaire (see Chapter 2), e.g., their thoughts on the activity, the interaction with the robot, the suggestions by the robot, their trust in the robot's suggestions and in general. More precisely, the questions are listed below:

- What did you think of the activity/robot?
- What did you think about the robot's suggestions?
- Which suggestions helped you to think harder or carefully?
- Did you trust the robot's suggestions?
- Do you think you can trust the robot generally?

The motivation behind doing an open-ended questionnaire was to have unconstrained user responses that would help understand better how the robot interventions are perceived. This

Figure 7.8: Setup adapted for an online study with *Harry*

perception would eventually facilitate improving the robot interventions as well as the overall activity. Here we highlight some of the changes taken into account when refining the pool of robot behaviors based on the most recurrent feedback in the interviews as well as based on our analytical observations. Please note that for the questions on trust, albeit receiving interesting answers, we will not discuss the outcomes in this thesis for the sake of keeping things to the point.

1. The number of interventions within each type of interventions were increased as well as worded in very different ways to avoid the robot sounding repetitive. This was based on answers by teams such as *"Early on suggestions were good later on it repeated"* and *"Sometimes they were repetitive that would distract me"*.

2. Originally, within each type of interventions, we also had suggestions that were meant to *reduce* a relevant behavior, i.e., the robot saying things like "I think we should not add any more rail-tracks for a few minutes, and think about what we already added." Through analytical observations, as well as feedback like *"Could have been better if the robot gave better suggestions .. about what to do"*, we realized that it is easier for people to *do* rather than *not do*. If the robot tells the students what not to do in order to improve, they have multiple possibilities of what to do instead and thus might feel confused about which action to pick. Hence, we removed those interventions that were suggesting to the students to reduce a certain behavior and rather we only kept the ones suggesting to the students to increase a certain behavior.

3. We increased the interrogative style of the suggestions where the robot asks the team members to explain their previous action or what they are about to do. This was based on how well the students responded to such suggestions in terms of verbalizing their reasoning as well as pointing to such interrogative suggestions when asked which

suggestions helped them think harder or carefully. This question was asked to the students as a follow up question to their answers such as *"Suggestions pushed us to think harder and made us to think more carefully", "Robots suggestions helped us to think harder, probably even overthink at times"*.

4. We incorporated the notion of time within some of the suggestions such as "Based on the last few minutes,...". This was to highlight that the current suggestion by the robot is based on just a slice of time and not the entire past duration. An example feedback that inspired this is *"(The robot) was kind of like a mother or a father, giving helpful suggestions at times, sometimes just reminding what we are already doing or have done"*

7.3 Hypotheses

Moving on to our final user study with *Harry* and *Hermione*, we make the following hypotheses:

- H1: (a) *Hermione* will lead to higher learning gains as well as (b) a higher number of teams achieving a higher learning gain as compared to *Harry*.
- H2: Teams that interact with *Hermione* will display higher *Productive Engagement* scores compared to the teams that interact with *Harry*.
- H3: *Hermione* will be rated higher on competence as compared to *Harry*.
- H4: Robot interventions will have an effect on the *PE score* in both the robots.
- H5: Robot interventions will have the desired effect on learner behaviors, i.e., we will observe a visible increase in the desired behaviors.

7.4 User Study

While generally the collaborative learning activity is the same as first outlined in Chapter 2, the version of the activity used in *Harry* and *Hermione* study will be referred to as *JUSThink-Pro*. This is due to the addition of real-time assessment modules based on our *Productive Engagement* framework, replacement of *Ron* with either *Harry* or *Hermione*, and a few refinements such as addition of a pop-up box for feedback on the perception of the usefulness of each intervention.

7.4.1 Participants

As mentioned before, the COVID-19 pandemic made it increasingly difficult to conduct HRI user studies especially with children. With an intensive effort over multiple months to reach out to international schools, especially targeting boarding schools, we were successful in

Figure 7.9: Children interacting with *JUSThink-Pro* at the six schools that participated in the *Harry* and *Hermione* study

taking our setup to six schools across Switzerland for the *Harry* and *Hermione* study[2]. Figure 7.9 shows students from each school interacting with *JUSThink-Pro* with a zoomed in view of the setup in Figure 7.10. More specifically, in this two-month long user study, 136 students (74 male, 62 female) with the age range 9-14 years (median age: 12 years old) interacted with our *JUSThink-Pro* activity over 70 hours. This gave us a total of 68 teams out of which the first 19 teams were used for validation (more on this below in section 7.4.4) leaving us with 54 teams for the experimental evaluation. For technical reasons such as data completeness, 2 more teams were discarded so we have a total of 52 teams with 26 teams in each condition. We must mention that as part of our agreement with some of the participating schools, after the study, we got back to the schools that were interested with a personalized feedback report on how their students performed in the study with the different robots (see appendix C for an anonymized example of such a report).

7.4.2 Real-time setup

Just like in *JUSThink* setup, here too, each participant interacts with an instance of the *JUSThink-Pro* participant application that is written in Python and uses pyglet as the windowing and multimedia library. Hence, a separate instance of the application is run for each participant in a team. The robot behaviour applications are also developed in Python and govern what the robot does and when as explained in section 7.2.4. All the real-time assessment applications for both log and speech features are also developed in Python. Lastly, all the applications communicate via the Robot Operating System (ROS). For the sound input, we make use of the built-in microphones of the laptops that are just placed next to the participants as shown in Figure 7.10.

[2]Ethical approval for this study was obtained from EPFL Human Research Ethics Committee (051-2019/05.09.2019)

Figure 7.10: A zoomed in view of the *JUSThink-Pro* setup at one of the schools

7.4.3 Evaluation Metrics

For evaluation purposes, in addition to the usual learning gains and in-task performance such as joint absolute learning gain *T_LG_joint_abs*, and last error as well as the robot perception questionnaire (see Chapters 2 and 3 for more details); we also made use of *PE score*, and *suggestion_usefulness* score. In the robot perception questionnaire, for this study, we added a few more questions to our questionnaire on robot competence that fall under the group *Robot (Godspeed-like)* and the category *Robot Behavior* based on Table 2.2 in Chapter 2:

- I think the robot was giving us the right suggestions.
- I think the robot gave us suggestions at the right time.

The first question is specifically targeting the *WHAT* aspect (what is it that the robot suggested us) while the second question targets the timing of the suggestion, i.e., the *WHEN* aspect. Please note that the choice to focus on joint absolute learning gain *T_LG_joint_abs* is because it captures the shared understanding between the team that, as established previously in the thesis, is relevant for collaborative learning.

7.4.4 Validation of the thresholds

We utilized the first 19 teams out of the 68 teams for validating our thresholds (τ_{PE}, τ_t^g). Please note that 5 of these teams in which *Harry* was used were also kept in the experimental data set since the the action selection technique for *Harry* does not utilize these thresholds; hence, the

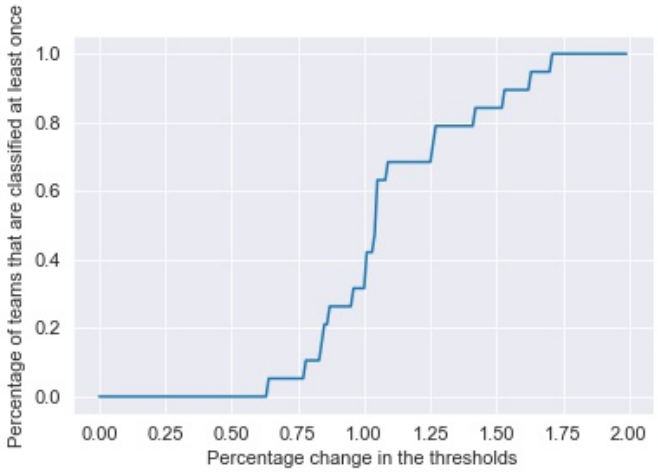


Figure 7.11: Validation of the thresholds for profile classification

validation of these thresholds only matters for *Hermione*. Naturally, the remaining 14 out of 19 teams that interacted with *Hermione* were discarded from the experimental data. For τ_{PE}, we wanted to make sure that the values in the validation data span between the entire range of 0 and 1 (see Chapter 6)) as it did in the training data. For the various τ_t^g, we were interested to observe the number of times an incoming profile was detected to be close enough. This was to ensure that the system was not too strict and never classifying an incoming profile as either *Expressive Explorers* or *Calm Tinkerers*. The Figure 7.11 shows that for the original thresholds (represented by the value of 1 on the x-axis), around 30-40% teams were classified at least once as either of the two gainer profiles. As the thresholds are increased or decreased by a percentage on the x-axis, the percentage of teams that are classified as belonging to a gainer profile at least once are depicted on the y-axis. We see that the original thresholds seem to be almost at the vertical asymptote of the curve. At the end of this validation, we chose to increase the thresholds of the type τ_t^g by 20% to allow for at least 50% of the teams being classified as either of the two gainer type at least once during the course of interaction. The Table 7.2 in section 7.2.3 lists down the thresholds after this increase in the 'After Validation' column.

7.5 Results

First of all in order to validate the relationship between the *Productive Engagement* score and the learning gain *T_LG_joint_abs*, which serves as a basis of our design and reasoning behind the *Productive Engagement* framework, we perform a linear regression analysis. We do so by using ordinary least squares (OLS) methods with the statsmodels library (Seabold & Perktold, 2010). The results are shown in Figure 7.12 between the two variables in both conditions with the *PE score* as the independent variable and the learning gain as the dependent variable. In case of *Harry*, it seems that *PE score* significantly predicts the learning gain (β: 0.39, p-value:

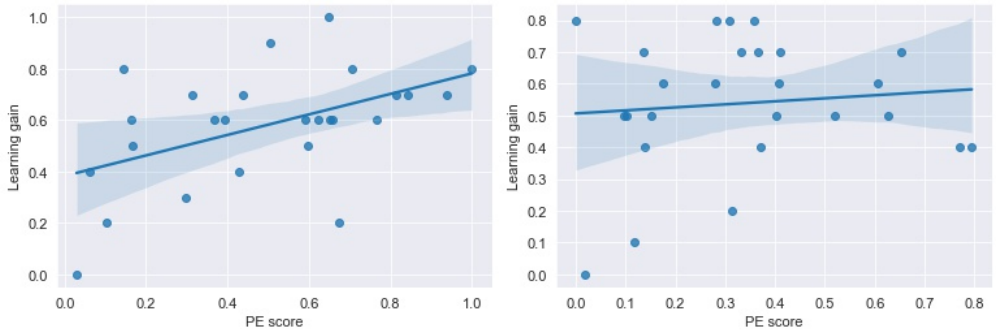


Figure 7.12: Linear regression between the the PE scores and the learning gains of the teams that interacted with *Harry* (on the left) and *Hermione* (on the right). For the former, the *PE score* significantly predicts the learning gain with a β of 0.39 and a p-value of 0.01 while for the latter, we do not find a significant result.

0.01) with the fitted regression model as 0.38 + (0.39**PE score*); however, it is not such in the case of *Hermione* (β: 0.09, p-value: 0.625). This means that for *Harry*, an increase of one in the *PE score* is associated with an increase of 0.39 in the learning gain.

7.5.1 Comparing Harry and Hermione on the Evaluation Metrics

Next, to evaluate our hypotheses H1-H3, we start off by doing a Kruskal Wallis test between the two conditions for the aforementioned evaluation metrics. As shown in Figure 7.13, there is no difference in terms of the joint absolute learning gain, that signifies shared understanding, between the two conditions. This means that the teams achieve similar learning gains in both conditions. Hence, H1(a) is not supported. Furthermore, contrary to our expectations, we observe that the *Productive Engagement* score is significantly higher (p-value: 0.03, H: 4.66) for the teams that interacted with *Harry* than those who interacted with *Hermione*. Thus hypothesis H2 is rejected. On the other hand, the teams that interacted with *Hermione* did rate the robot higher both on the suggestions being right as well as being at the right time, albeit non-significantly; however, the rating for the *suggestion_usefulness* score is higher for *Hermione* with marginal significance (p-value: 0.06, H: 3.52) than for *Harry*. Hence, hypothesis H3 is partially supported.

High and Low Learning Groups Between conditions

In order to dive a bit deeper to understand and explain the afore-discussed outcomes and hypotheses, we are interested to unveil where does the difference arise between the two robots in terms of the *Productive Engagement* score as well as the *suggestion_usefulness* score. Our first intuition is to observe whether the differences come from the differences in the learning gains within each condition. We split the two conditions into high and low learning gain groups. For this, we use a mean split by calculating the average *T_LG_joint_abs* of the entire

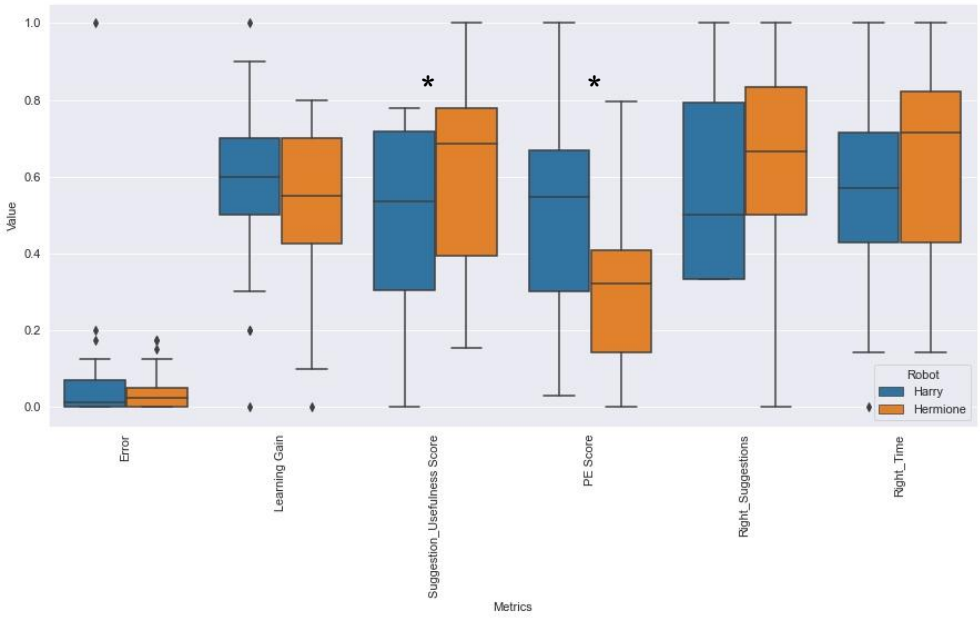


Figure 7.13: Comparison of Harry and Hermione in terms of our evaluation metrics where the asterisk on the graph represents a significant difference on the Kruskal Wallis test. There is a significant difference between the two robots in terms of the *PE score* (p-value: 0.03) and the *suggestion_usefulness* score (p-value: 0.06).

data set, which is 0.559 (normalized between 0 and 1), and then we use that to split the teams in each condition into high and low learning groups. To validate this mean split, we observe via Kruskal Wallis tests, that indeed the learning gains of the low learning groups in *Harry* as well as *Hermione* conditions differ significantly from the high learning groups in the respective conditions (For *Harry*, p-value: $4.77e^{-05}$, H: 16.53; for *Hermione*, p-value: $1.28e^{-05}$, H: 19.03). Interestingly, we note that in the group that interacted with *Harry*, 18 out of 26 teams ended up with higher learning gains while in the group that interacted with *Hermione*, 13 out of 26 teams ended up with higher learning gains. Thus, H1(b) is rejected.

Now that the split based on the learning gain has been validated, we compare the low and high learning groups between the two conditions on the same metrics as in section 7.5.1 using Kruskal-Wallis tests. As shown in Figures 7.14 and 7.15, between the two conditions, there is no difference in terms of any metric between the low learning groups, i.e., low learning groups behave similarly irrespective of the robot they interact with. However, as suspected, the difference in the *Productive Engagement* score as well as the *suggestion_usefulness* score is indeed as a result of the high learning groups, i.e., the high learning group in the *Harry* condition displays a significantly higher PE score (p-value: 0.004, H: 8.07) while rate the robot significantly lower on the usefulness of the suggestions (p-value: 0.05, H: 3.81) as compared to the high learning group in the *Hermione* condition (see Figure 7.15).

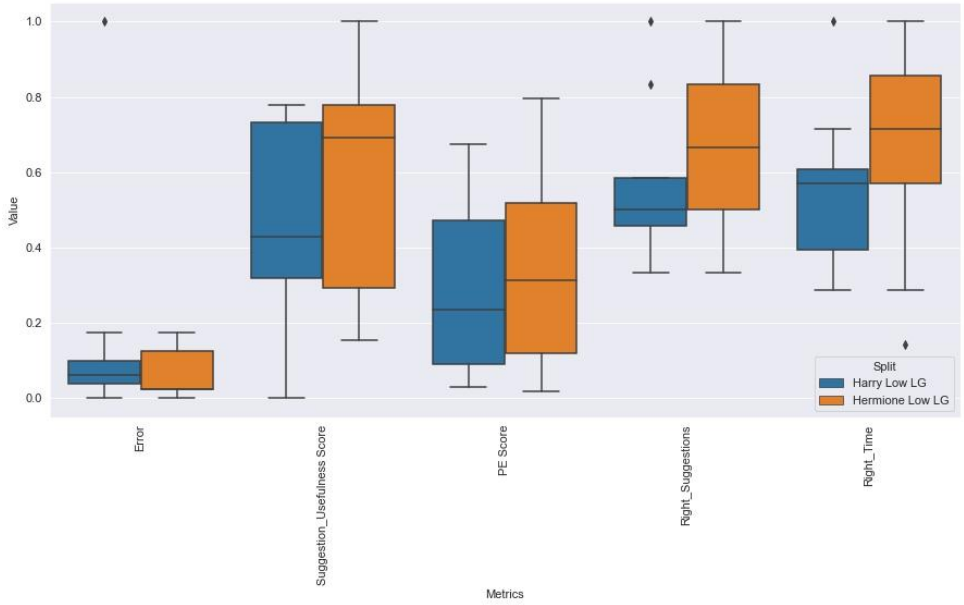


Figure 7.14: Comparison between the low learning teams in the two conditions. None of the metrics differ with statistical significance.

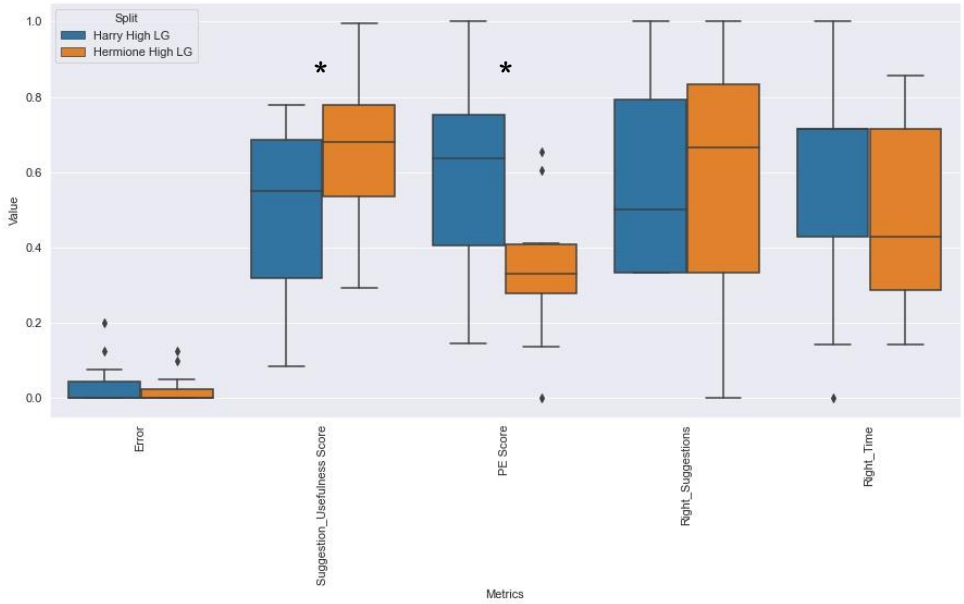


Figure 7.15: Comparison between the high learning teams in the two conditions where the asterisk on the graph represents a significant difference on the Kruskal Wallis test in terms of the *PE score* (p-value: 0.004) and the *suggestion_usefulness* score (p-value: 0.05).

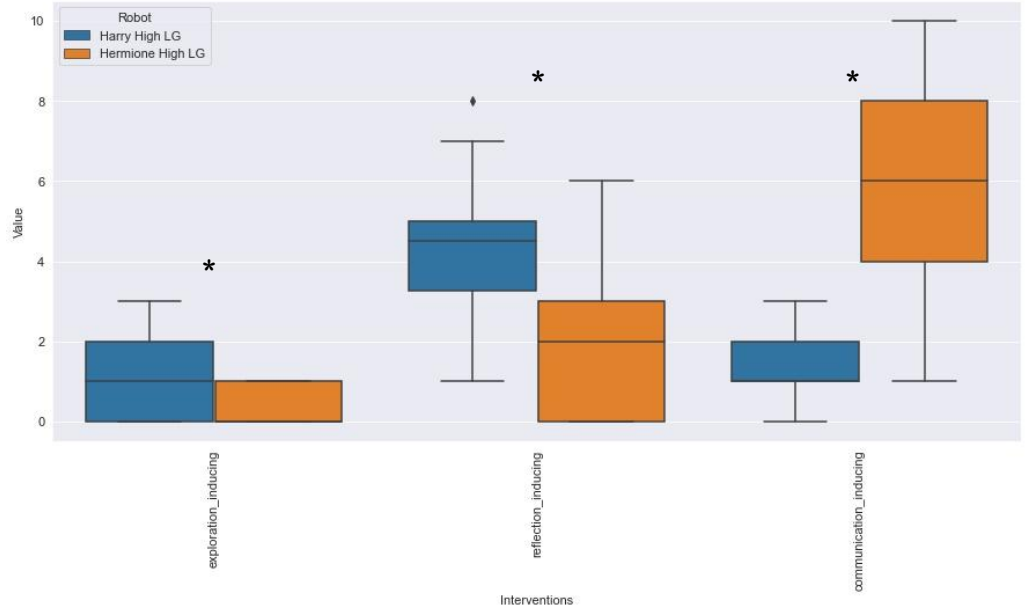


Figure 7.16: Comparison of the intervention types received by the high learning teams in the two conditions where *Harry* received significantly more *exploration inducing* (p-value: 0.03) and *reflection inducing* (p-value: 0.0019) interventions while *Hermione* received significantly more *communication inducing* (p-value: $1.64e^{-05}$) interventions

7.5.2 Comparing Harry and Hermione on Robot Interventions

In order to answer the hypothesis H4, we are first interested to identify what kind of robot interventions were received by the students in each condition. More specifically, we observe the two high learning groups who interacted with *Harry* and *Hermione* as that is where we see the differences surfacing from in terms of *PE score* and *suggestion_usefulness* score. We inspect this again with Kruskal Wallis test, where a box plot is shown in Figure 7.16. Indeed, the teams interacting with *Harry* received significantly more *exploration inducing* interventions (p-value: 0.03, H: 4.32) as well as *reflection inducing* interventions (p-value: 0.0019, H: 9.60) while the teams that interacted with *Hermione* received significantly more *communication inducing* interventions (p-value: $1.64e^{-05}$, H: 18.56). These tests highlight the differences in terms of the interventions the teams in high learning groups of the two conditions received. However, the relationship between these type of interventions and the PE score cannot yet be established as a general rule. Hence, at this stage, we cannot claim anything definitive regarding whether it is indeed the differences in these types of interventions that leads to the differences in the PE score for the two robot conditions.

To unearth this, we perform linear regression analysis, between the type of interventions as the independent variable and the *PE score* as the dependent variable for all teams in each condition. Referring to the Figure 7.17 and the Table 7.4, we observe that for *Harry* with all the teams, none of the intervention types (*exploration inducing*, *reflection inducing*,

Table 7.4: Regression tests for both *Harry* and *Hermione*

Independent variable	dependent variable	coefficient	p-value	intercept	p-value
Regression tests for Harry					
Exploration inducing	PE score	-0.42	0.317	0.56	0.00*
Reflection inducing	PE score	0.36	0.37	0.25	0.35
Communication inducing	PE score	0.03	0.93	0.49	0.00*
Regression tests for Hermione					
Exploration inducing	PE score	-1.48	0.14	0.37	0.00*
Reflection inducing	PE score	-0.46	0.02*	0.41	0.00*
Communication inducing	PE score	0.42	0.02*	-0.007	0.96

communication inducing,) are statistically significant predictors of the *PE score*. On the other hand, for *Hermione*, when looking at Figure 7.18 and the Table 7.4, we observe that both the intervention types of *reflection inducing* and *communication inducing* are statistically significant predictors of the *PE score*; however, they effect the *PE score* in opposite ways. The increase of one in *communication inducing* intervention type seems to be associated with an average increase of 0.42 in the *PE score* while the increase of one in the *reflection inducing* intervention type seems to be associated with an average decrease of 0.46 in the *PE score*. This is quite interesting as, in accordance to our design principle behind the action selection technique of *Hermione*, the interventions are designed with the goal of increasing the *PE score* and an intervention is triggered when the *PE score* is detected to be lower than the τ_{PE}. We see that happening in the case of *communication inducing* interventions, which by design specifically target effective communication between the team members. However, we also observe that the increase in the *reflection inducing* interventions is negatively affecting the PE score of the team members. With these outcomes for *Harry* and *Hermione*, H4 is partially supported as the interventions do have an effect on the *PE score* in the case of *Hermione*.

For H5, we evaluate the effectiveness of the interventions, i.e., if the relevant learner behavior increases after the intervention is suggested in the following two minutes compared to the two minutes that preceded the intervention. If that is the case, the intervention is considered effective. In this way, for the three types of interventions, we calculate the percentage of interventions that were effective as shown in Figure 7.19. For both robots, while the *communication inducing* and the *exploration inducing* interventions are effective in a medium range (30-60 %); however, very few (6-10 %) of the *reflection inducing* interventions seem to have been effective. Hence, H5 is only partially supported.

7.6 Discussion

To interpret our results, we go back to the analogy of the cosco ladder we presented in Chapter 3 that linked robot behaviors to user engagement to user learning via the concept of *Productive Engagement*. For the second part of the ladder that goes from engagement to learning, while we indeed observe a linear relationship between the *PE score* and the learning gain when

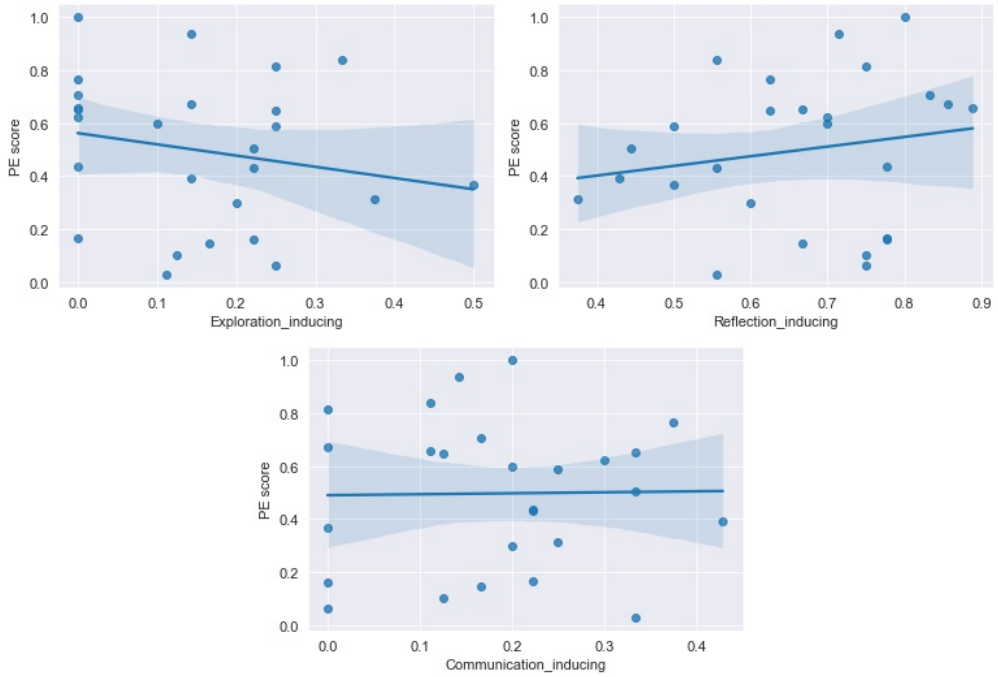


Figure 7.17: Linear regression between the three intervention types and the *PE score* for the teams that interacted with *Harry*. None of the intervention types are statistically significant predictors of the *PE score*

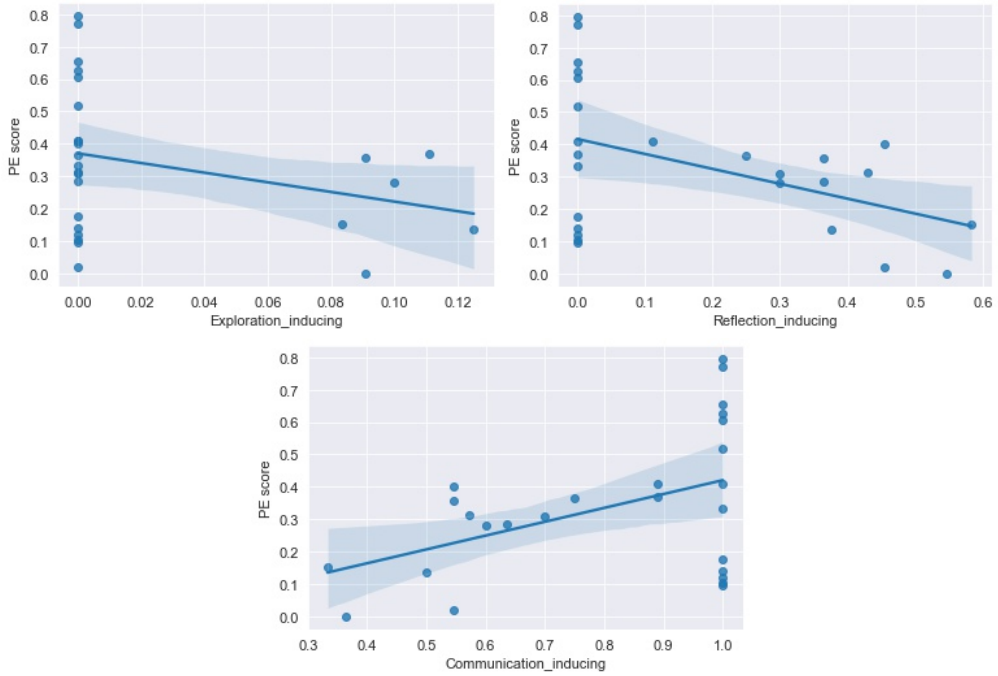


Figure 7.18: Linear regression between the three intervention types and the PE score for the teams that interacted with *Hermione. communication inducing* and are statistically significant predictors of the *PE score* with p-values of 0.02 and 0.02, respectively.

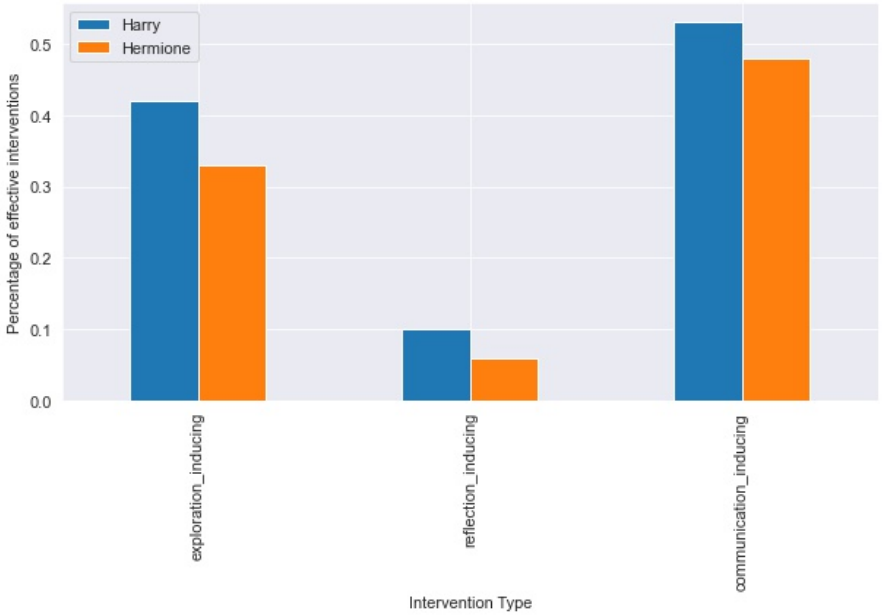


Figure 7.19: Percentage of effective interventions for the two robots: 42% and 33% of the intervention type *Exploration inducing*, 6% and 10% of the intervention type *Reflection inducing*, and 53% and 48% of the intervention type *Communication inducing* are effective for *Harry* and *Hermione*, respectively.

teams interacted with *Harry*, we do not see such a relationship when teams interact with *Hermione*. We make the following two hypotheses that may provide possible explanations for this behavior:

1. *PE score* needs to be above a certain level to have that linear relationship with the learning gain. Note that in the condition with *Hermione*, the *PE score* is not only significantly lower than *Harry*, but generally lies in a lower range with a mean value of 0.33 which is just around the same value as the τ_{PE}, i.e., 0.32 (see Table 7.2 in section 7.2.3).

2. There is a need to refine our definition of the *Productive Engagement* score that currently may not be capturing the complete dynamics of the relationship with learning. The new refined definition may then not have a linear relationship.

We pose both of the aforementioned hypotheses as open questions for the community as well as our own future work. Further, for the first part of the ladder that connects robot behaviors to user engagement, interestingly, we find that in the case of *Harry* that has a very simplistic model for making suggestions, none of the types of intervention significantly predict the *PE score*. On the other hand with *Hermione*, indeed the *reflection inducing* as well as *communication inducing* intervention types significantly predict the *PE score*; however, in opposite ways. The *communication inducing* interventions have a positive correlation with the *PE score* indicating that as these interventions increase, the mean of the dependent

variable, i.e., *PE score*, also increases. This directly validates: (a) the *communication inducing* interventions that are explicitly designed to induce communication as well as (b) our design choice for *Hermione* providing this intervention type at the initial phase of the game as well as when no profile is close enough, i.e., when the robot is unsure of the problem solving strategy the learners are exhibiting. Conversely, the *reflection inducing* interventions have a negative correlation meaning that as these interventions increase, the mean of the *PE score* decreases. While reflection related behaviors (*Edge Deletion*, *History*, *A_A_add*, *A_A_delete*, *A_B_add*, and *A_B_delete*) that this intervention type is supposed to induce in the learners have been established to be effective for learning (see Chapter 4), it is possible that: (a) either the suggestions are not effective in inducing those behaviors, or (b) the timing is off. For (a), indeed as seen in Table 7.19, we found that only 6% of *reflection inducing* interventions were effective. In both cases (a) and (b), this could lead to a negative correlation of *reflection inducing* interventions with the *PE score*.

Tying our main findings altogether, both robots induce similar learning outcomes (H1a) and similar level of effective interventions (H5). *Harry* generates higher *Productive Engagement* in the students (H2) while *Hermione's* interventions are perceived more useful (H3) and have an effect on the *Productive Engagement* (H4). To reiterate, for *Harry*, a robot that leverages much less information than *Hermione*, there exists a relationship between the *PE Score* and the learning gain as well as more teams interacting with *Harry* end up with higher learning gains (H1b). However, the *PE Score* does not seem to be explained by the interventions of the robot (H4) as well as the robot's suggestions are perceived less useful by the learners (H3). On the other hand, for *Hermione*, there exists a relationship between some of the robot's interventions and the *PE Score* (H4) as well as the suggestions are perceived more useful by the learners (H3). However, there is a lack of a relationship between the *PE Score* and the learning gain. These results raise some interesting reflections and hypotheses for us regarding the 1) relationship between the *Productive Engagement* score currently defined and learning, 2) design and effectiveness of the robot interventions, and 3) action selection techniques in particular for *Hermione*. For (1), we have laid down some hypotheses earlier in the section. For (2) and (3), we put forth the following comments:

1. A more *conscious* action selection strategy, i.e., that of *Hermione*, indeed significantly influences the variable of interest *Productive Engagement* showing the potential of such robots over a hard baseline.

2. In order to truly establish relationships between intervention types and *Productive Engagement*, there is a need to refine the design of interventions in a way that the interventions are effective to some extent in surfacing learner behaviors they are explicitly designed for. Currently, we did not see that for the *reflection inducing* intervention type.

3. Based on the findings and specifically what we found in terms of interventions and their relation or lack of relation with the *PE score*, we hypothesize that timing could be extremely crucial to define this relationship. Specifically, in *Harry*, the interventions

effects seem to even out while for *Hermione*, the interventions that had a good timing or were more effective accumulated their positive effect while the ones that were badly timed accumulated a negative effect on the *PE score*.

With this chapter, we close the loop by evaluating the effect of robots, endowed with various levels of understanding of *Productive Engagement*, on the learning of the students in our collaborative activity. In the upcoming chapters, we synthesize the main findings and contributions of this thesis, as well as the ongoing and future work.

8 Broadening The Horizon

At the end of the previous chapter, we hinted at some intended future work in regards to extending our understanding of *Productive Engagement*, and its possible refinements. In this chapter, we briefly discuss some expansions related to this thesis. More specifically we broaden our understanding on the two branches: 1) *when* the robot should intervene or 2) *how* the robot should intervene. In total, there are three ideas (listed below) covered in this chapter where the first two relate with point (1) and the last with point (2):

- An alternate design for the robot *Hermione*,
- Personalization models for productive engagement, and
- Incorporating personality in an educational robot.

8.1 An Alternate Design for the robot Hermione

In Chapter 6, with the *PE score*, we showed one possible way to construct a metric for the assessment of the *Productive Engagement*, i.e., a way to link learner's behaviors with learning. Then, in Chapter 4, we identified aggregate learner profiles. Together, the score and the profiles were used to design an action selection strategy for *Hermione* in Chapter 7. We observed that the *PE score* based system worked in some ways while still surfacing some shortcomings or needed refinements. In our ongoing effort, we are working towards building an alternate system with a focus on a new action selection strategy for the robot *Hermione* (the *what* and *when* robot). This alternate system makes use of a *non-Productive Engagement score*, thereafter referred to as *nPE score*, that is constructed using an HMM based methodology. This methodology is inspired by the outcomes of Chapter 5 in this thesis where HMM based temporal profiles were built for the three group of learners. To differentiate the *Hermione* robot previously discussed and the robot making use of this new system, we will refer to the robot here as *Snape*. Theoretically, while the intentions of the robot *Snape* are good just like *Hermione*, its actions are triggered by the presence of unproductive events.

This system differs in several ways from the already designed and evaluated strategy for *Hermione*:

1. In the PE based system, the speech generates the *PE score and then* the log features generate the profiles that are then used to intervene accordingly. In the alternate system, the speech *and* log features *together* generate a hidden state, and looking at a sequence of hidden states, the system intervenes accordingly. In short, we employ an HMM based technique that generates a current hidden state, given the team's behaviors.
2. Instead of looking at the *PE score* as a deciding factor for intervening, the system relies on an *nPE score*, proposed in this ongoing work. We consider the *nPE score* as another metric for characterizing *Productive Engagement* in addition to the *PE labels* and the *PE score*. Based on a particular sequence of unproductive states, a *non-Productive Engagement* score is incremented and eventually leads to an intervention (more details later). Therefore, the actions of a robot incorporating an *nPE score* based system are triggered by the presence of unproductive events while the actions of a robot incorporating a *PE score* based system are triggered by the absence of productive events.

These differences in how the system utilizes speech and log information form the motivation behind the new strategy. In particular, we are interested to investigate:

- How does a robot incorporating this alternate system, alongside the *when* aspect like *Hermione*, affect the children's learning?
- And how effective are its interventions?
- How similar and dissimilar are the two systems (*PE* and *nPE* based) in terms of inducing interventions? , i.e., how do the two systems *validate* and *complement* each other?

8.1.1 Construction of the nPE score

The *nPE score* is aimed at looking for patterns, a sequence of states, that are indicative of possible unproductiveness. To begin, we use our publicly available dataset PE-HRI-Temporal (Nasir, Bruno, & Dillenbourg, 2021a) as the training data just like for *Hermione*. Here, without going into details, we will briefly highlight the analysis methodology that leads to the construction of the nPE score. We briefly walk through the steps shown in a flow diagram in Figure 8.1:

- **steps 1 and 2**: First of all, we train an HMM using speech and log features to obtain the set of observations emitted at each state. Similar to our results in Chapter 5, 3 hidden states exist that could be interpreted as an unproductive state, a global problem solving strategy state, a local problem solving strategy state.

- **steps 3, 4, and 5**: Upon inspecting patterns of state transitions of length 6 (pertaining to one minute) using sequence mining and frequency analysis, we identify some patterns that are significantly more present among the non-gainers (p-values < 0.01) than in the gainer teams while some patterns only exist for the non-gainers. We notice that most of such patterns of state transitions that were more prevalent or only existing for non-gainers include the unproductive state (low communication, low exploration, low reflection). Another interesting observation we make is in terms of when such patterns appear. We observe that at the very beginning of the interaction, both gainers and non-gainers have very similar distribution of these patterns which then changes drastically for the gainers particularly after 3 minutes into the interaction. However, the non-gainers still display such patterns. Please note that the aforementioned choice of one minute is arbitrary and based on having an appropriate time period that neither corresponds to a very short period nor a very long phase of interaction.
- **steps 6**: As a further step, we study the effect of such patterns on the learning gains. We do a regression analysis, using Ordinary Least Square (OLS) method, with the total number of occurrences of all significant patterns as the predictor and each of the learning gains as the outcome. We observe that the total number of occurrences explains around 20% of the variance in each learning gain and has a negative, statistically significant coefficient (with all p-values < 0.01). We then perform regression for each individual pattern and observe similar results for 8 particular patterns in terms of explaining the variance and having negative significant coefficient. We term these 8 patterns as the particularly significant patterns.

Based on this analysis, we define an *nPE score* that is an alternate metric to characterize being *productively engaged*. In short, the score starts at zero and it updates as follows: 1) For each occurrence of a particularly significant pattern (the 8 patterns), the score is incremented by the sum of the significant coefficients of its OLS models, 2) For each occurrence of other significant patterns, the score is incremented by the sum of the coefficients of the global OLS model. Since the coefficients are negative as the correlation between these patterns and learning gain is negative, the score is thus negative. This means the smaller the score gets, the less *productively engaged* the team is gauged to be.

The robot *Snape* aims at *minimizing* the *absolute* value of the *nPE score* in real-time, i.e., will try to bring the *nPE score* towards zero. For that, the robot intervenes when the score of the team drops below a certain threshold τ_{nPE}. This is a dynamic threshold generated as the weighted average of the average *nPE score* of the gainer teams at each time window and that of the non-gainer teams from the training data.

8.1.2 nPE based real-time Control Architecture for Snape

The real-time control architecture for *Snape*, based on ROS, consists of three main phases that briefly are:

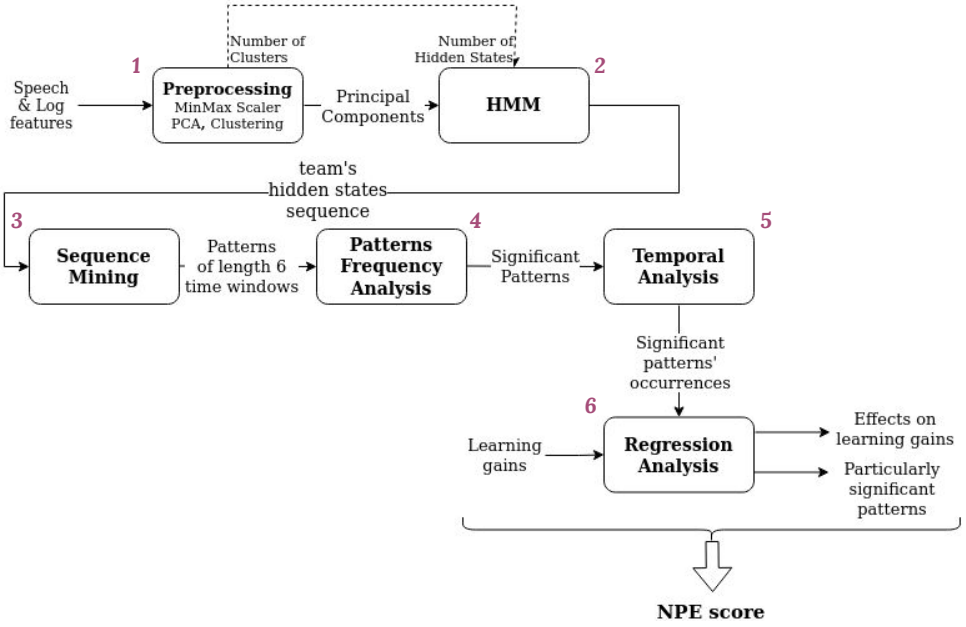


Figure 8.1: The analysis methodology that led to the construction of the nPE score

1. **A preprocessing phase**: Every 10 seconds, with the log and speech features as inputs, the current hidden state is generated for the team.

2. **An *nPE* update phase**: The current hidden state is buffered until 60 seconds giving us a pattern of 6 hidden states. The pattern is compared against the significant patterns and if it is one of the significant patterns, the *nPE score* is updated as explained previously. Following that, if the updated score is below τ_{nPE}, an intervention is triggered.

3. **An intervention phase**: The pool of interventions consists of suggestions intended to increase learner behaviors found to be lacking in the unproductive state. These include the behaviors of addition of edges, looking at their past solutions, and communicating with each other. We employed an exploration-exploitation policy where an intervention is chosen based on the outcome of a Bernoulli trial with a probability of exploration. When the trial is successful, the intervention is chosen randomly among all possible interventions; however, when the trial is a failure, the intervention with the highest weight is chosen. The weight for an intervention is updated based on the effect it has on the slope of the *nPE score* in the two minutes before and after the intervention. After the execution of an intervention by *Snape*, the score is reset to 0.

8.1.3 Pilot Study

We conducted a pilot study at an international school in Switzerland with 22 children aged 9 to 12 years, divided into teams of 2 (shown in Figure 8.2). One team is omitted from the analysis due to technical problems during the experiment. The resulting dataset consists of 10 teams,

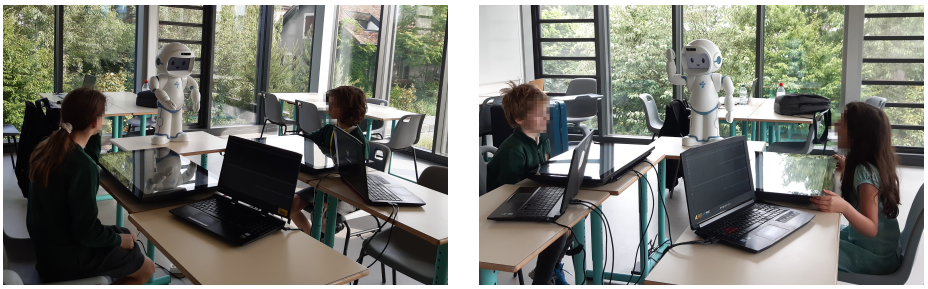


Figure 8.2: Children interacting with *JUSThink-Pro* along with the *Snape* robot at a Swiss school

totaling 20 children (7 females: M=10.14, SD=0.69; 13 males: M=10.53, SD=0.77).

8.1.4 Preliminary Results

Currently, we are in the process of analyzing the data. Some of our preliminary observations are:

1. Indeed, as seen when designing the *nPE* based system, the number of occurrences of the significant patterns is negatively correlated with the relative and joint learning gains with marginal significance (p-values = 0.08 and 0.07, respectively). This provides validation for the *nPE score*.

2. The effectiveness of interventions in terms of inducing the desired behavior is 40%, 70%, and 57% for the type *exploration inducing*, *reflection inducing*, and *communication inducing* interventions, respectively. Thus all of the interventions at least had an effect more than 40% of the time. Unlike with *Hermione*, the pool of *reflection inducing* interventions only include *T_hist* inducing suggestions for *Snape* as the unproductive state in this data driven methodology dictated that.

3. One very interesting observation is that all interventions together increase speech in 54% of the cases, speech overlap in 50% of the cases, and reduce the long pauses in 34% of the cases. These behaviors are exactly what defined the *PE score* in Chapter 6 where both speech and speech overlap had a positive effect on the *PE score* and the long pauses had a negative effect on the *PE score*.

4. Another observation of note is related to a comparison between teams that interacted with *Ron* in the *Ron* study (baseline study for the thesis) and the teams that interacted with *Snape* in this study. For the teams from the *Ron* and *Snape* study matched based on propensity score formed on their pre-test scores: the median of the post-test scores and all the learning gains is higher for the teams from the *Snape* study. For post-test scores, the difference is marginally significant at p-value = 0.09 with a Kruskal Wallis test. However, none of the differences for learning gains are significant.

Apart from an ongoing in-depth analysis, for the future we envision a larger user study possibly with a comparison between *Hermione* and *Snape*.

This work is a collaborative effort with Mortadha Abderrahim who the author of this thesis supervised both in their bachelor's semester project as well as in their role at the lab as a Research Fellow. This work is under preparation for a publication.

8.2 Personalization models for Productive Engagement

The line of research this thesis focused on was more on equipping robots with the ability to infer whether the learners are engaged in the learning activity at hand. Another popular line of research in educational Human Robot Interaction (HRI) and Intelligent Tutoring Systems (ITS) is to use robots in order to personalize learning strategies to needs of an individual in order to cater for the learning goal as not everyone learns in the same way (Cordova & Lepper, 1996; Leyzberg et al., 2014; Ramachandran et al., 2017). We believe engagement modelling goes hand in hand with personalization as *the better the robot is aware about the students individual characteristics, the better it can detect the engagement state of the learners, which itself can be manifested in several ways.* Similarly, *the better this engagement state is inferred, the better personalized learning interactions the robot can offer.*

In this collaborative work, we propose and compare personalized models for *Productive Engagement* recognition. Briefly, we use the aggregated profiles and the *PE score*, from the thesis, within an AutoML deep learning framework to personalize *Productive Engagement* models.

We investigate two approaches for this purpose: (1) Single-task Deep Neural Architecture Search (ST-NAS), and (2) Multitask NAS (MT-NAS). In the former approach, personalized models for each learner profile are learned from multimodal features and compared to non-personalized models. In the latter approach, we investigate whether jointly classifying the learners' profiles with the engagement score through multi-task learning would serve as an implicit personalization of the productive engagement. Moreover, we compare the predictive power of the two types of features in our dataset: incremental and non-incremental features.

8.2.1 Methodology

Generally, the methodology consists of three steps starting from learner's profiling, followed by feature extraction, and eventually then the use of an efficient neural architecture system (ENAS) as shown in the Figure 8.3 (taken from (Vikashini et al., 2022)).

The learner's profiling comes from the profiles we have generated in our work in this thesis and the extracted features are those log, speech, facial expressions, and gaze features that are available in the PE-HRI-Temporal dataset (Nasir, Bruno, & Dillenbourg, 2021a), also generated in this thesis. For the last step, as mentioned at the begining of this section, we either have

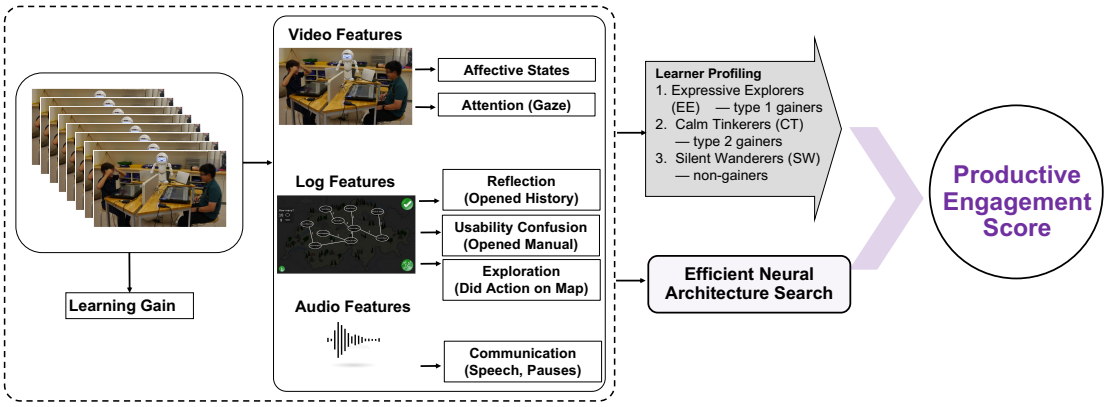


Figure 8.3: The general architecture for the personalization models for Productive Engagement starting with learner's profile, feature extraction, and then efficient neural architecture system

profile level personalization (ST-NAS) or *multi-task learning personalization* (MT-NAS).

In *profile level personalization*, for a sub-set of the original dataset corresponding to each learner profile, an adaptive neural architecture is automatically designed and trained using an efficient neural architecture system (ENAS). In this work, we make use of the Auto-Keras library (Jin et al., 2019).

In *multi-task learning personalization*, a multitask ENAS learns an adaptive model to predict both the engagement score as well as the profiles of the learners. This means it implicitly uses the information of the profiles while learning to predict the *PE score* due to the property of weight sharing.

8.2.2 Initial Results

Briefly, our experimental results show that:

1. Personalized models improve the recognition performance with respect to non-personalized models when training models for the gainer vs. non-gainer groups,
2. Multitask NAS (implicit personalization) also outperforms non-personalized models
3. The speech modality has high contribution towards prediction
4. Non-incremental features outperform the incremental ones overall

For future work, the idea is to explore the direction of comparing with other personalized strategies, such as those based on SVMs like Selective Transfer Machines.

This work is a collaborative effort with Hanan Salam, Vetha Vikashini, and Oya Celiktutan with whom the following paper is accepted at ICMI, 2022:

H. Salam, V. Vikashini, J. Nasir, B. Bruno, and O. Celiktutan."Personalized Productive Engagement Recognition in Robot-Mediated Collaborative Learning". accepted in the 24th ACM *International Conference on Multimodal Interaction (ICMI)*, 2022 (Vikashini et al., 2022).

8.3 Incorporating Personality in an Educational Robot

Our personality, at some level, has the power to affect how people perceive us. This perception can then directly change people's level of attention, engagement and trust in what we have to say and how people react to each other in social settings (Driskell et al., 2006; Peeters et al., 2006). This becomes especially critical in positions of responsibility, such as human/robot teachers/tutors, where their personality may translate into their pedagogical strategy and hence, influence the learning process for a child.

On this note, previous research has suggested that indeed the personality of a robot can influence the quality of human-robot interaction (Kiderle et al., 2021; Robert et al., 2020). While the most commonly employed robot personality is considered to be an extroverted personality (Robert et al., 2020; Speranza et al., 2020; Staffa et al., 2021), in the context of educational HRI, we believe that some other examples of a strong robot personality, inspired by psychology and learning theories, can include: an *adversarial robot* that induces conflict among the team members as a way to raise the cognitive load of the students or a *Socratic robot* that asks questions for the same purpose of increasing the cognitive load of the students or a *Supportive robot* with excessive positive reinforcements to motivate the students towards the learning process. Currently, most robots in educational HRI seem to embody the *Supportive* personality.

In our ongoing work, we are interested to endow robots with *adversarial* and *socratic* personalities in the context of *JUSThink* as the two personalities represent two different approaches in learning. The *adversarial* robot disagrees with the students and has a clear idea of what the students should do instead. Thus its suggestions challenge the opinion of the students, so that they can think back on their solution and choose to keep it, or to move towards the other opinion. The *adversarial* personality in our work is inspired by cognitive psychology, and in particular, in the Cognitive Load Theory (Sweller, 2011). This term refers to the amount of working memory resources used when learning. We can differentiate three types of cognitive load.

- *The Intrinsic cognitive load*: It is induced by the difficulty of the task, and therefore, it cannot be influenced by the teaching methods.
- *The Extraneous cognitive load*: It is precisely generated by the teaching method itself, so it varies depending on how the information is presented to the learner.
- *The Germane cognitive load*: It is the result of the effort that is put into creating a mental schema, when students are aware of what they just learned, and are able to link it with

	Supportive	Socratic	Adversarial
Opinionated	Yes	No	Yes
Positive/Negative	Positive	Positive	Negative
Agree	Yes	Yes	No
Disagree	No	No	Yes
Refer to desired situation	Yes	No	Yes
Refer to the current situation	Yes	Yes (but in the form of a question - no opinions of its own)	Yes

Figure 8.4: Comparison matrix between the 3 personalities' suggestions

THE BIG FIVE	Socratic	Adversarial
Openness to experience/Curious	Yes	No
Conscientiousness	Both	Both
Extroversion	Yes	Yes
Agreeableness	Yes	No
Neuroticism (Not really important here)	-	-

Figure 8.5: Description of the personalities in terms of the OCEAN traits

the information of their long time memory.

The fact that the adversarial suggestions are very clear and induce a critical reflection of the students about their own choices may decrease the extraneous cognitive load while increasing the germane cognitive load. It has been shown that raising the germane load also improves learning. Another way of increasing this cognitive load is by using the *socratic* questioning (Carey & Mullan, 2004), named after Socrates. This is an educational method that focuses on discovering answers by asking questions to students, and our *socratic* robot is directly inspired by this method. In an educational setting, it would entail questioning the students, so that they become aware of their non-understanding of a problem, and to be able to start finding solutions to the problem at hand.

So our broader research question in this direction is: *How do the two personalities differ in terms of their effect on the learning gain of the students, in-task performance of the student, the student's perception of the robot's suggestions and other characteristics as well as the students trust in the robot and engagement in the activity?*

Adversarial robot's suggestions	Socratic robot's suggestions
"Hey, Alice and Bob, we really need to talk more to each other to understand the game."	"So, Alice and Bob, how could we see if there are better costs somewhere?"
"Hey Alice and Bob, if we don't want to repeat the same mistakes, we should look at the past solutions."	"Hey Alice and Bob, I am trying to remember the costs of the rail-tracks we added in the past. Do you have an idea?"
"Come on guys, I think we had enough reflection for this one. Let's move-on now!"	"Hey friends! I am wondering how fast we are, to go from one move to the other."
"You know, the only way to find better costs is to connect more gold mines."	"Just a minute, Alice, and bob, what is our goal when we look at our past solutions? Have we reached this goal yet?"
"Now guys, I think it is time to delete some rail-tracks!"	"Hey guys, do you feel like we are taking too much time to make decisions?"

Figure 8.6: Examples of the suggestions for each robot personalities

8.3.1 Methodology

For designing the two personalities, we make use of the two matrices as shown in Figures 8.4 and 8.5. Without going into the details, the first matrix is inspired directly by the brief literature we discussed above on the very different pedagogical ways in which a robot can declare a problem, and orient the participants in the right direction to correct the situation. The second matrix is adapted from the Big Five personality traits (OCEAN traits) which is widely used to describe human personality traits. Briefly, this means that the designed *adversarial* robot is straightforward, strongly opinionated, extroverted, and expresses its opinion in the form of a disagreement. On the other hand, the *socratic* robot is cooperative, not opinionated, extroverted, vague as it does not explicitly advice the participants on what to do but rather asks questions, and expresses itself in a positive/agreeable manner.

The two robots are incorporated in the *JUSThink-Pro* platform where they have the same action selection technique as that for *Harry*. Some examples of the suggestions given by the two robots are shown in Table 8.6.

Validation Study

Before conducting a preliminary user study, we conducted an independent online validation study. This was to ensure that people can clearly differentiate the two robots, i.e., making the upcoming comparative user study meaningful. For the validation study, we created an online form where the two robots, namely *Kauri* and *Zuri*, corresponding to the *socratic* and *adversarial* personalities, were presented and described. Then in the context of the *JUSThink-Pro* setup which was also explained in the form, 10 short videos of the robots, 5 each, with the

Figure 8.8: Adults interacting with *JUSThink-Pro* at EPFL

robot giving a suggestion, were shown to the participants in a random order. For each video, the participants had to select which robot they thought they saw in the video: *Kauri* or *Zuri*.

As a result of publishing the study online for relevant communities, we received 57 responses with participants ages ranging between 15 to 55 years old. The percentage of correct answers in the form varied between 54.4 % and 86 % depending on the suggestions presented. Thus, for all the suggestions most of the participants recognized the personality well. A confusion matrix is also shown in Figure 8.7 that highlights an interesting observation: the *adversarial* robot *Zuri* is guessed correctly much more than the *socratic* robot *Kauri*. This may suggest that designing a *socratic* personality may be relatively less straightforward.

		Predicted	
		Kauri	Zuri
Actual	Kauri	196 (68 %)	94 (32 %)
	Zuri	53 (18 %)	237 (82 %)

Figure 8.7: Confusion matrix for the recognition of Kauri and Zuri (N = 290)

8.3.2 User study with Adults

Due to the limitations caused by the pandemic and the time constraints of the student project within which this work was conducted, we targeted students from EPFL computer science department. We had a total of 40 participants, i.e. 20 teams. Of these, 45% participants were women and the overall average age was 23.3 years. Due to data incompleteness, we had to discard 4 teams leaving us with 16 teams. The study was setup in an open space in the Rolex Learning Center on EPFL campus as shown in Figure 8.8.

8.3.3 Results

Briefly, when performing a statistical analysis using Kruskal Wallis to compare differences between the two robot groups, we did not find any statistically significant differences (see

Figure 8.9). It is however interesting to notice that some p-values are quite low especially for the parameters concerning the perception of the participants in terms of *distraction* and *social trust*. Contrary to our expectations, *socratic robot* was perceived to be more distracting while on the other hand the *adversarial robot* was perceived to be more trustworthy in terms of social trust.

Learning Gain	T_LG_ absolute: **0.15**	T_LG_ relative: **0.18**	T_LG_ joint_abs: **0.22**		
Perception of the robot	perception_ comp: **0.79**	perception_ like: **0.83**	competence_ trust: **0.17**	attention_to_ robot : **0.27**	
Interventions usefulness	intervention_ indiv1: **0.11**	intervention_ indiv2: **0.59**	intervention_ union: **0.29**	intervention_ intersection: **0.71**	
Performance	range_error: **0.63**	avg_error: **0.53**	std_error: **0.60**	nb_actions: **0.46**	duration : **0.75**
Participants behavior	engagement: **0.15**	think_harder: **0.29**	distraction: **0.10**	social_trust: **0.06**	play_again: **0.17**

Figure 8.9: Matrix showing p-values obtained with Kruskal Wallis test for each parameter

Wrapping up, as opposed to our hypothesis, there were no significant differences between the two robot conditions. It is therefore possible that the two personalities do not influence the metrics under question differently. However, there are a few factors that should be considered: 1) The study was suppose to be with children aged between 9 to 14 years old as the *JUSThink-Pro* platform is designed with that age range in mind, 2) The number of participants is limited for a stronger statistical analysis. As a future work, we plan to conduct a study with a larger group of younger participants which could allow for a better and a more intended comparative analysis.

This work, which is currently unpublished, is done with William Ouensanga who was supervised by the author of this thesis for their bachelor's semester project.

9 Synthesis

9.1 Overview

This thesis investigated the relationship between learning and engagement, in robot-mediated learning activities. More precisely, we (i) delved deep in learning analytics, machine learning and statistical methods, to identify and quantify the relationship between learning and engagement, which we termed as *Productive Engagement* (PE), (ii) designed and developed a framework for the robot's real-time autonomous monitoring of *Productive Engagement*, selection and execution of appropriate interventions, (iii) validated the effects of such a framework in multiple user studies for which we (iv) designed and developed a robot-mediated, open-ended collaborative learning activity aiming to help children hone their computational thinking skills.

To this end, in **Chapter 1**, we began with understanding the challenges in the current state of the art in terms of modelling learners' engagement and its manipulation by an autonomous robot/agent. Among them, we focused on four challenges in this thesis, listing them here again for convenience:

- **C1**: While it is often assumed that engagement and learning have a linear relationship, this is not proven
- **C2**: Many automatic models of engagement rely on human annotators and often suffer from low inter- and intra-rater agreement because of the subjective nature of this construct
- **C3**: Open-ended learning environments typically envision failure as a means towards learning and do not allow for using the straightforward in-task performance metrics as measures for learning, at least, not in a linear or monotonic way.
- **C4**: Learning happens in real-time and cannot wait just because the sensors or the robot need more time. As a consequence, reliable and fast real-time assessment of the students' learning is crucial

With these challenges in mind, we formulated four broad research questions that paved the path of this work:

1. **Research Question 1**: Given the learners behavioral patterns, can we reveal a quantitative relationship that links them to learning?
2. **Research Question 2**: Which human behaviors are predictive of learning and how?
3. **Research Question 3**: Can we build representations of engagement using the behaviors identified in RQ2 that can then be used for its detection in real-time?
4. **Research Question 4**: How can a robot make use of these representations to induce the relevant behaviors, found as a result of RQ2, in the learners?

In the remainder of this section, we briefly revisit the research work in each of the previous chapters and highlight how they contribute to each of the aforementioned challenges and research questions. In order to develop, design and evaluate our *Productive Engagement* framework, we started off with building a rich robot mediated collaborative learning activity, *JUSThink*, in **Chapter 2**. The choice to have two users in our setting, introducing social engagement with a human, was because we wanted to grasp all facets of engagement, since we did not know a priori which ones would better relate to learning. The activity itself was designed with the goal to hone the computational thinking skills as well as collaborative skills of the learners. For the former, the activity presents a Minimum Spanning Tree (MST) problem implicitly as an optimization problem where the learners need to connect gold mines with rail-tracks by spending as little money as possible in a fictional scenario situated in Switzerland. For the latter, the activity incorporates a *collaborative script* in the design of the activity that enforces collaboration through the choice of *partial information* and *role switching* among other features. The collaborative script was motivated by our own previous iteration of a collaborative design activity *Cellulo City* and a user study with it that highlighted the implications the *design* of an activity has on enforcing collaboration: one cannot merely put two students together and expect them to collaborate.

Within the *JUSThink* platform, we then introduced the robot *Ron* who was intended to only intervene to automate the activity and to give some basic motivational support, without causing unnecessary distractions. We took our system to two schools in Switzerland where 98 students interacted with the system. With this *Ron* study, we showed that in-task performance and learning are not correlated, and despite *Ron's* rudimentary behaviour, participants perceived it as highly competent, intelligent, friendly, likeable, not distracting, and reported not feeling a need for more feedback from the robot. Linking to **C3**, the lack of correlation between learning and performance metrics highlighted the importance of moving away from robot interventions that affect (and refer to) only superficial measures of students' learning, e.g. in-task performance. Instead, it emphasized on focusing on learner's behavioural patterns that could more solidly indicate whether participants would end up learning or not.

The *JUSThink* platform then served as the basis for all our future studies as well as the multi-modal data collected in the *Ron* study served as the basis for several kinds of analyses. The collected data was also made available openly and publicly in the form of two datasets.

Specifically, with regards to **C1** and **RQ1**, we moved on to formally define and validate the concept of *Productive Engagement* in **Chapter 3**. For the purpose of defining it, we drew inspiration from engagement literature from both perspectives of HRI and Multi-modal Learning Analytics (MLA). From the multi-modal data collected in the *Ron* study, we extracted several behavioral features (log, speech, facial and gaze) using open-source tools like OpenFace and Google WebRTC VAD. With these behavioral features, we validated the proposed concept via a *forward-backward clustering technique*, that was discussed in the same chapter. This technique allowed to observe whether the teams cluster similarly in terms of their learning and performance and in terms of their behaviors. In terms of learning and performance, we saw four types of clusters emerge: *Productive Success*, *Productive Failure*, *non-Productive Success*, *non-Productive Failure*. With regards to behaviors, we observed three clusters. Upon performing a similarity analysis between the two cluster types in terms of the teams they shared, we showed that it is possible to compute an approximation of user engagement in a data driven manner and that the operationalization of engagement obtained preserves the link with user learning, i.e., *Productive Engagement* was validated.

For a robot to make effective interventions, i.e., to know what behaviors should be encouraged in the learners and when, we needed to dive deeper into what multi-modal behavioral sets were found to be associated with higher learning in Chapter 3. Hence, in **Chapter 4**, we extended our *forward-backward clustering technique* to *forward-backward clustering and classification technique* to identify multi-modal behavioral profiles of collaborative learning in constructivist activities. This work that consisted of the aforementioned quantitative technique as well as interaction analysis, a qualitative technique, specifically targeted **RQ2** and **C2**. The quantitative approach allowed: 1) to build the multi-modal behavioral profiles for each group of learners, 2) to take into account that teams with similar learning and performance might actually exhibit two or more different sets of behaviors, and 3) the use of cluster labels as ground truth for classifiers, thus, allowing for methods that are devoid of human intervention. On the other hand, the qualitative approach allowed us to better interpret the multi-modal profiles and understand the learning mechanisms at play within each group of learners.

Our classification results showed that the use of multi-modal behavioural labels, i.e., PE labels, proven solidly linked with learning, seem to allow for a better discrimination between high and low gainers than the direct use of learning labels. With our methodology, we found two types of gainer profiles and one type of non-gainer profile that we termed as *Expressive Explorers*, *Calm Tinkerers*, and *Silent Wanderers*, respectively. While the two gainers types differed from the non-gainers in terms of their speech behaviors, the two types of gainers differed from each other based on their problem solving strategies (local vs global) as well as their emotional expressivity. Among other results, one of the most important findings of this work was that verbal interaction, not just in terms of amount of speech but also overlap of speech between

team members, in a constructivist collaborative activity emerges to be a discriminatory factor between gainers and non-gainers. Furthermore, we discovered that there exists a *relationship* between the type of *problem-solving strategy* and *emotional expressivity*, that can discriminate multiple ways of achieving the learning goal.

As this analysis led us to gather some insights for **RQ2** while tackling **C2**, we realized that the temporal aspect of the learning process cannot be ignored. Therefore, we advanced our understanding of the *process* of learning in our particular context by focusing on temporal data from the *Ron* study in **Chapter 5**. We employed a multi-modal Hidden Markov Model (HMM) based methodology, with 4676 datapoints, to investigate the temporal learning processes of the gainers and non-gainers profiles identified earlier. The temporal analysis allowed for additional insights w.r.t. the previous findings: 1) that the gainer groups actually shifted back and forth between the two problem solving strategies, each characterized by both exploratory and reflective actions, 2) a particular affect is not strictly associated with a *type* of problem solving strategy but it also depends on the *phase* of the activity, i.e., irrespective of the strategy, the negative emotions increase towards the later stages of the activity.

Our work also offered a complementary view of how collaborative open-ended problem-solving proceeds, in terms of problem-solving strategies (local vs global) rather than problem-solving phases (exploring, formulating, planning and monitoring). The global problem solving strategy can be considered as one in which planning, exploring, formulating and monitoring happens on the scale of the entire problem. The local problem solving strategy is one in which the planning, exploring, formulating and monitoring happens on the scale of the next step towards the solution. Our work thus adds to CSCL literature by suggesting that learners seamlessly intertwine these two strategies in their productive collaborative problem-solving, and that neither is at the outset "*better*" than the other.

While Chapters 3-5 focused on defining, validating and identifying *Productive Engagement* in a collaborative robot-mediated context, in order for a robot to assess *Productive Engagement* in real-time, it was necessary to build representations of such engagement of the learners with least computational resources, for example, sensors or modalities. Hence, our focus shifted to **RQ3**, **C4** and again **C2**. In **Chapter 6**, we first investigated how real-time metrics allowing a robot to assess whether learners are engaged in meaningful learning behaviors can be constructed, using the *Productive Engagement* framework as a reference. Then, based on that, we implemented several data-driven methodologies for the computation of such *Productive Engagement* metrics, among which was the *Productive Engagement* score. Upon validation of this score, defined as a weighted linear combination of previously established discriminatory speech behaviors, we found that this score can indeed serve as an efficient, fast and lightweight way of tracking the teams' "Productive Engagement" state. In particular it can do so by being the first indicator of when and whether an intervention is needed as it has the ability to discriminate between moments that might be productively engaging versus those that are not.

In our view, the most important outcome from the construction of PE in terms of the PE score was that the quantity and quality of speech was found to be *sufficient* for assessing *Productive Engagement*, implying that we could use simpler uni-modal features instead of multi-modal features with great computational benefits for real-time systems.

Finally, we designed the fully autonomous robots *Harry* and *Hermione* utilizing varying degrees of insights gathered through our proposed framework in **Chapter 7**; thus considering **RQ4**. The design of *Hermione* was motivated by the idea that the timing of an intervention by the robot is equally crucial as knowing what kind of suggestions to make to the learners. *Harry* on the other hand only focused on the content of the suggestions while the timing was randomly decided; hence, it served then as a hard baseline to evaluate *Hermione*. Unlike most robots in educational HRI settings whose role is defined based on having varying degrees of domain knowledge, both of our robots were manipulated on a proposed 2d space, specifically on the axis of knowledge about learner's behaviors that are conducive to learning as shown in Figure 7.2.

Then with the two robots, we conducted an extensive user study with 136 students from 6 international schools in Switzerland. What we found was that both the robots led to similar learning gains. Furthermore, *Harry*, the robot that randomly made suggestions only focusing on the *content*, might randomly generate a ratio of the three types of interventions that, albeit having no significant relation individually with the *PE Score* and being perceived less useful, as a whole led to higher *Productive Engagement*. On the other hand *Hermione*, the robot with a more carefully designed action strategy that specifically and consciously focuses on the *timing* as well as *content*, has a significant relationship between the intervention types and *Productive Engagement* and is perceived more useful; thus, validating its action selection technique. However, the level of *Productive Engagement* induced may just not be enough for higher learning gains.

In short, for the core of this thesis we conducted 4 iterative HRI studies globally involving ~300 students excluding the pre-experiments, 9 international schools in Switzerland, and ~130 hours of data collection.

9.2 Contributions

Now that we have gone through an overview of the thesis in the previous section, here we explicitly briefly highlight some of our contributions. For details, each of the contributions has been discussed thoroughly in the light of the related literature in the respective discussion sections in the chapter where it is first highlighted.

This thesis lies at the intersection of two communities, namely the community of Human-Robot Interaction (HRI), and the community of Learning Analytics (LA), particularly to multi-modal and collaborative learning analytics communities and by extension to the community of Computer-supported collaborative learning (CSCL). Some aspects of our work provide insights

for all communities while other aspects are more relevant for one of them. Additionally, this thesis tackles both the aspects of *perception* and *behavior design* for autonomous educational robots.

Previous work on educational HRI and LA rely on the hypothesis 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 LA literature has investigated the relation between learners behaviors (indicative of constructs like engagement, effortful behavior, etc.) and learning. This thesis contributes to both communities by taking a step in the direction of reuniting the two sides of the equation: robot behavior to user engagement to user learning via proposing the concept of *Productive Engagement* that emerged by investigating such domains in parallel. We believe this conceptualization could be of interest to both communities towards understanding engagement in various learning scenarios as well as building better skilled autonomous agents to help induce that engagement in learners.

We contribute to the LA literature by proposing a *forward and backward clustering and classification technique* to build multi-modal collaborative learning profiles of dyads as they work on an open-ended task around interactive tabletops with a robot mediator. The proposed technique is envisioned to be applicable in other educational contexts as well as with individual or multiple participants.

Through this technique, we identified three learner profiles that gave insights for both the CSCL and HRI designers regarding how the various multi-modal behaviors of learners differ among those who learn and those who don't. In particular, we showed the most significant discriminator to be the overlapping and interjecting speech between learners. Then, we showed that those who learn exhibited two particular kinds of problem solving strategies, both of which consisted of an exploratory and a reflective element. We also identified *which* interplay between problem solving strategies and emotional expressivity may be more conducive to learning in such a CSCL setup in *addition* to the more obvious behavior of speech activity.

In this direction, we also contributed by generating temporal profiles, employing a proposed HMM-based technique, to further show how the temporal processes of the three groups of learners evolve. This provided insights that explained further the previous outcomes as well as gave additional outcomes that contribute to understanding the multiple *pathways* of learning in an open-ended CSCL and robot-mediated environment, and provide actionable insights for designing effective interventions. We showed that learners alternate between both problem solving strategies over the course of their interaction instead of sticking to one strategy; however, non-gainers lack reflective behaviors in both the problem solving strategies. Additionally, affect is not only influenced by the learner's problem solving strategy but also the phase of the activity.

Another contribution of the *forward and backward clustering and classification technique* is that it gives a possibility to surface data driven labels for engagement; thus, circumventing

the tedious and often subjective process of human annotations. Hence, it contributes to both the fields of HRI and LA in the direction of building data driven machine learning models that can make it easier to be more objective in decision making. In the same direction, the construction process of the simplistic *PE Score* also contributes in the same way. While on the one hand, it can be used as an assessment metric directly, on the other hand, it can also be used as data driven labels for building more sophisticated regression models.

Then, with the design of our three robots *Ron*, *Harry*, *Hermione*, we provided a complementary perspective to the two communities (HRI and LA) regarding the role of educational social robots. In this regard, this thesis demonstrated a proof of concept by designing a fully autonomous *skilled ignorant peer* in the form of *Hermione*, i.e., a robot that is ignorant about domain knowledge but is aware of: 1) the skills needed to navigate successfully through the exploratory collaborative learning space as well as 2) the state of the learner. A study conducted with such a robot aware of *what* to suggest and *when* to suggest versus a baseline robot demonstrated the potential of a more intentional robot action selection strategy in terms of manipulating the variable of interest, i.e., *Productive Engagement* in this case.

Furthermore, we contribute by the design from scratch of our experimental platform *JUSThink*, incorporating a collaboration script, for enhancing the computational thinking as well as collaboration skills of learners. Additionally, it serves as a platform for the design and development of autonomous educational robots. While this thesis utilizes the platform to build robots with a particular skill set, it can be used to develop and evaluate other robot skills as well. For instance, my colleague Utku Norman, who was involved in the design of this original platform, has used derivations of the same platform for developing and testing *mutual modelling* skills of a robot (Norman et al., 2022; Norman, Dinkar, Bruno, et al., 2021). With the same platform, we are also currently working in the direction of designing robot personalities and investigating their pedagogical effects (see 8).

Moreover, as established by the underlying motivation of this thesis, *perceiving* certain subjective constructs such as what being *engaged in the learning process looks like* is not straight forward even for domain experts and is often prone to subjectivity. This is where data driven machine learning methods can provide more objective representations, the understanding of which can then be used by domain experts to select appropriate robot interventions to drive the robot policy. Most of the times, the domain expert is responsible both for *perceiving* the environment as well as *intervening* in the learning scenario as one would envision a future robot to be, for example, as in the end to end Participatory Design (PD) methodology in HRI (Winkle et al., 2021). Our iterative methodology demonstrates how the use of machine learning can be used to complement the skills of domain experts so that the automation part (transparent AI) takes care of the *perception* and the human expert takes care of the *intervention* side. Ironically the same human characteristics that can make them subjective on the perception side would allow them to be more empathic and well-rounded on the intervention side.

Lastly, the thesis contributed to open science practices by publishing the data utilized for the

bulk of our thesis in the form of two open-source datasets (Nasir, Bruno, & Dillenbourg, 2021b; Nasir et al., 2020a). These datasets can be relevant, but are not limited in use, for researchers from all the aforementioned communities that are looking to explore/validate theoretical models of engagement in learning scenarios. There are already two ongoing collaborations that stemmed out from these datasets with Hanan Salam at the Smart Lab at New York University Abu Dhabi together with Oya Celiktutan at the Center for Robotics Research at King's College London, and the other collaboration with Justine Cassell at Articulabo at INRIA, Paris, France.

9.3 Take-aways

With the contributions of this thesis in mind, our achievements as well as failures, we list down some of the take-aways for the intended communities:

- When building autonomous robots for educational purposes, there is a need for more systematic data driven investigation to:
  - understand the relationship between a learner's engagement and learning. Possibly, this can start with data collection in the intended context with a baseline robot that carries as little assumptions in its behaviors as possible.
  - move away from human annotated labels in order to create less subjective and less biased models of automatic assessment.
- In exploratory/constructivist learning scenarios, the assessment of learner's states such as engagement, motivation, effort, etc. should not be limited to relatively superficial measures of in-task performance.
- The possibility of multiple sets of very varied learner behaviors associated with the same learning profile (high or low learning) should be considered when designing pedagogical interventions.
- Seemingly negative behaviours, such as interjecting speech and negative emotions, should not be dismissed as having a negative influence on the learning process without experimental proof.
- While multi-modal behaviors can provide a better understanding of the underlying learning mechanisms, not all modalities useful for understanding learning might be necessary for assessing learning in real-time. In short, trade-offs should be considered between accuracy and fast/lighter systems.
- An educational social robot can help advance learning with little to no domain knowledge by focusing on the behavioral skills needed by a learner to advance in the learning activity.
- For an educational social robot, it is as important to know when it should *not* intervene as to know when it needs to intervene.

- When designing action selection strategies for autonomous robots, intentional, well-informed and conscious choices driven by data could allow for a more transparent evaluation and interpretation of the robot's effectiveness.
- When assessing the effectiveness of robots, perception questionnaires should be complemented by data driven metrics that could objectively highlight the effect the robot's interventions had on the learner's run-time behavior.
- When conducting user studies with schools, getting back to the interested schools with a personalized feedback report on how their students performed in the study (see appendix C for an anonymized example of such a report in the context of our *Harry* and *Hermione* study) is, in our experience, an effective practice to increase the transparency in educational HRI that also enhances mutual trust between researchers and the non-roboticist stakeholders such as teachers, school directors, etc.

9.4 Limitations

This thesis in only a first step in the direction of defining, conceptualizing and constructing the concept of *Productive Engagement*. This is both a strength of the thesis and also its limitation since by no means are the current outcomes complete. As seen by our *Harry* and *Hermione* study, we already can see some directions (see end of Chapter 7) which we may need to further explore in order to refine the definition of *Productive Engagement*. More iterative studies in other learning contexts with varied student demographics will help in this refinement process. Additionally, our qualitative informal interviews at the end of that study (not included in Chapter 7) revealed that when most of the students thought the robot was useful, the reason why they thought so was *'because it makes us think harder and gives us reminders'*. This user feedback highlights that the students do not attribute *reflection* explicitly to a limited set of robot interventions but broadly to any intervention. More exploration is required in this direction. Then, when most of the students thought the robot was not useful, the reason they stated was *'because it told us what we were already doing..that took time from us'*. This points to the need of the robot knowing *when not* to intervene even though we did incorporate this in our current strategy; however, it needs more careful consideration. Furthermore, when checking the effectiveness of the interventions, we only measured the learner behavior the intervention was explicitly designed for. It is possible that the interventions may have desired or undesired indirect effects on other learners behaviors too. This also needs to be investigated. In this regard, we also invite the interested community to join in investigating and extending the understanding and characterization of *Productive Engagement* with the eventual goal of making the concept more robust and concrete.

Although it is an achievement in itself to do user studies with multiple schools in a time of pandemic, some limitations need to be highlighted. To not to be allowed in-person studies for more than an year caused a longer than planned **time difference between the user studies**. Further, we must note that since the studies are done at international schools in Switzerland,

the students are from a **very specific pool coming from a certain economic and social background**; hence, this requires us to be careful about the group we generalize our results on learning profiles and on the interaction with the robot in this thesis. Beside the disruption that the pandemic brought, we got the chance to move our setup online for one of our studies, which albeit being challenging, seemingly a limitation, and an effort not foreseen in the plan for the thesis; it ended up giving us a very unique experience. We conducted an online HRI study with a physical robot in a creative zoom setup. This opened the doors for imagining newer and more portable ways of conducting user studies in a post-COVID world as well as with target populations that would otherwise not be an option.

Furthermore, a point that we realized early on in our Chapter 3 was that the data driven clusters were imbalanced meaning that with our pipeline, the non-learning cluster that emerged had lesser number of teams. **Data imbalance** is inherent in the real world and will inevitably lead to skewed distributions; however machine learning algorithms are known to work best with balanced data. Taking this into account, in order to deal with the imbalanced data, we employed a variety of machine learning algorithms including decision trees that frequently perform well with imbalanced data.

Another important aspect of developing educational robots is to test their effectiveness on students learning in the long term. This is currently not a very common practice in educational HRI and similarly is one of the limitations of this thesis too. Ideally, we would have liked to have **longitudinal studies** where we would have: 1) had participants interact with our *JUSThink-Pro* setup in at least two sessions separated by a week and, 2) tested their learning, *Productive Engagement*, robot perception periodically over this time period and additionally a few days after the second session to measure their retention of knowledge. However, this was not possible due to the large amount of resources required in terms of time and effort from the schools as well as the teachers reluctance for their students to miss multiple classes. These concerns were already raised when arranging single session studies used in this thesis.

Furthermore, our ***Productive Engagement*** **framework was only tested in the context of the *JUSThink* activity**. The design methodology to reach an autonomous robot equipped with the concept of *Productive Engagement* required us to go through a rigorous and iterative process over the course of this thesis to build and evaluate the autonomous robots in the same context. However, we don't know how *Productive Engagement* and the computation we put forth for it in this thesis would generalize to other tasks and learning activities. For instance, our task relies on a shared visual workspace which has an influence on both the problem solving strategies and the interactions. Other tasks might not have the same characteristics. Therefore, in the future, we would like this framework to be adopted and evaluated in other learning activities as well as other learning contexts.

A Appendix A

For clustering, we utilize k-means in both approaches where the value of *k* is selected based on entropy analysis. Approach A gives four clusters corresponding to high/low combinations of learning gains and performance metric named accordingly as *Productive Success* (high learning and performance), *Productive Failure* (high learning but low performance), *non-Productive Success* (high performance but low learning), *non-Productive Failure* (low learning and performance) abbreviated as PS, PF, non-PS, non-PF, respectively. On the other hand, approach B gives 3 behavioral clusters with the first two exhibiting high learning and the third lower learning; hence, named as *type 1 gainers*, *type 2 gainers*, and *non-gainers*, respectively.

When comparing the three behavioral clusters from *approach B*, cluster 1 and 2 both have learning gains that are significantly higher than the learning gains exhibited by the third behavioral cluster, while the average performance of all 3 behavioral clusters is very similar. When comparing the similarity between the forward and the backward clusters in terms of the teams they consist of (Figure A.1), we observe that the first two behavioral clusters have more than 70% teams from both the *Productive Failure* and *Productive Success* groups, while the third behavioral cluster mostly has teams from the *non-Productive Failure* and *non-Productive Success* groups. Concretely, this implies that learners who end up with a learning gain regardless of their performance in the task exhibit two kinds of behaviors. With the two requirements mentioned in section 4.3.2 being met, we can proceed with the labels surfaced from the two approaches to be used by classifiers trained on multi-modal behaviors. We made use of two commonly used classifiers, SVM and Random Forests with our dataset (Nasir, Norman, Bruno, Chetouani, et al., 2021). Please notice that the classifiers were trained and tested on this newer dataset version with the two slightly modified features, hence providing slightly different results from those reported in our paper Nasir, Bruno, and Dillenbourg, 2020. As can be seen in Table A.1, we achieve much higher accuracy and recall on the validation and test set with labels from *approach B*; thus, lending further support to our argument that this approach is better than *approach A* in identifying the behavioral profiles of gainers and *Silent Wanderers*.

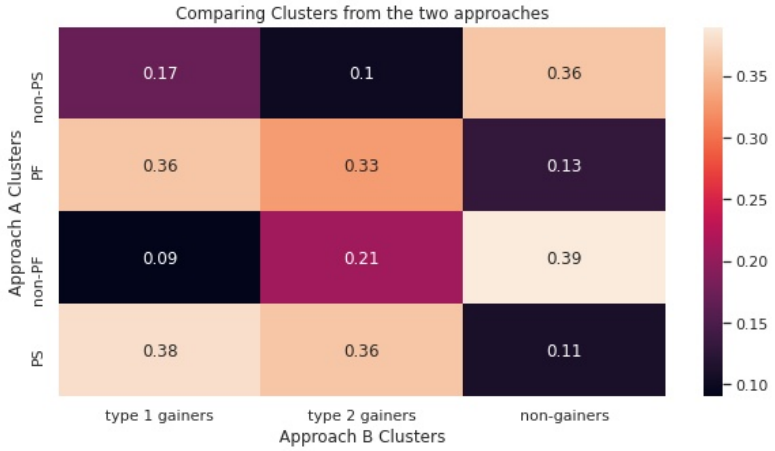


Figure A.1: Comparison between the clusters of the two approaches in terms of the teams they consist of.

Table A.1: Classification Results

Classifier	k-fold cross-validation		test-set	
	Accuracy	F1-score	Accuracy	F1-score
Approach A				
SVM	0.28	0.23	0.44	0.34
RF	0.36	0.26	0.33	0.39
Approach B				
SVM	0.80	0.76	0.88	0.89
RF	0.72	0.68	0.88	0.82

B Appendix B

Table B.1: Features' Mean values in each of the Expressive Explorers' states

Feature	**InitialState**	**MoreProbableState**	**LessProbableState**
T_add	1.102384×10^{-9}	3.573773×10^{-1}	4.0838×10^{-2}
T_ratio_add_rem	4.183066×10^{-9}	9.999993×10^{-1}	1.8960×10^{-2}
T_action	3.870495×10^{-2}	9.699460×10^{-2}	4.4758×10^{-2}
normalized_time	2.029434×10^{-1}	5.387081×10^{-1}	6.26040×10^{-1}
Speech_Overlap	2.723118×10^{-1}	4.758125×10^{-1}	5.67795×10^{-1}
Overlap_to_Speech_Ratio	5.461976×10^{-1}	6.965724×10^{-1}	8.04408×10^{-1}
Speech_Activity	4.099696×10^{-1}	5.986744×10^{-1}	6.65616×10^{-1}
Silence	6.541648×10^{-1}	4.866740×10^{-1}	4.10024×10^{-1}
T_remove	9.325293×10^{-10}	3.182062×10^{-7}	1.26406×10^{-1}
Gaze_at_Robot	4.343753×10^{-2}	9.518764×10^{-3}	4.5563×10^{-2}
redundant_exist	3.763263×10^{-3}	6.830674×10^{-3}	2.477×10^{-3}
T1_T1_rem	1.027991×10^{-17}	3.735072×10^{-20}	1.16683×10^{-1}
Gaze_at_Partner	7.156486×10^{-2}	6.737566×10^{-2}	1.17361×10^{-1}
T_help	7.807358×10^{-2}	6.557370×10^{-3}	1.4712×10^{-2}
T1_T2_rem	1.478677×10^{-15}	4.773094×10^{-7}	4.3755×10^{-2}
T_hist	5.047627×10^{-3}	5.044131×10^{-3}	1.290×10^{-3}
Gaze_at_Screen_Right	5.915012×10^{-1}	5.912295×10^{-1}	5.85986×10^{-1}
Gaze_at_Screen_Left	3.447677×10^{-1}	3.441521×10^{-1}	3.06201×10^{-1}
Long_Pauses	4.414569×10^{-3}	1.723356×10^{-2}	2.917×10^{-3}
Arousal	2.705875×10^{-1}	3.101827×10^{-1}	3.75027×10^{-1}
Short_Pauses	1.685203×10^{-1}	1.542912×10^{-1}	1.16228×10^{-1}
Negative_Valence	2.056995×10^{-1}	2.568086×10^{-1}	3.08619×10^{-1}
Positive_Valence	3.375673×10^{-1}	3.469566×10^{-1}	4.12408×10^{-1}
Gaze_Other	8.812433×10^{-2}	5.841153×10^{-2}	6.2550×10^{-2}
T1_T2_add	0.000000	0.000000	0.000000
Difference_in_Valence	5.507043×10^{-1}	5.013340×10^{-1}	5.13887×10^{-1}
T1_T1_add	0.000000	0.000000	0.000000

Table B.2: Features' Mean values in each of the Calm Tinkerers' states

Feature	InitialState	MoreProbableState	LessProbableState
T_ratio_add_rem	2.673037×10^{-10}	1.000000	4.172444×10^{-3}
T_add	6.682594×10^{-10}	3.166667×10^{-1}	1.043111×10^{-2}
Speech_Overlap	3.282486×10^{-1}	5.413534×10^{-1}	5.921923×10^{-1}
Speech_Activity	4.855440×10^{-1}	6.827465×10^{-1}	7.096177×10^{-1}
Silence	5.495654×10^{-1}	3.864979×10^{-1}	3.520730×10^{-1}
T_action	1.260936×10^{-2}	5.866667×10^{-2}	2.662700×10^{-2}
Overlap_to_Speech_Ratio	6.216050×10^{-1}	7.520540×10^{-1}	7.951161×10^{-1}
normalized_time	3.205108×10^{-1}	5.447131×10^{-1}	5.415407×10^{-1}
T_remove	1.266331×10^{-2}	3.999698×10^{-15}	1.701398×10^{-2}
T1_T1_rem	1.168974×10^{-12}	8.479291×10^{-18}	6.258666×10^{-2}
T1_T2_rem	1.347584×10^{-7}	2.299210×10^{-21}	2.086218×10^{-2}
redundant_exist	1.747415×10^{-3}	1.458333×10^{-2}	5.336480×10^{-3}
Positive_Valence	3.665011×10^{-1}	4.497916×10^{-1}	4.121070×10^{-1}
Arousal	3.269324×10^{-1}	3.862492×10^{-1}	3.702702×10^{-1}
Gaze_at_Robot	1.162740×10^{-2}	5.383023×10^{-3}	1.723160×10^{-2}
Negative_Valence	2.549870×10^{-1}	2.911126×10^{-1}	2.905847×10^{-1}
T_help	1.132622×10^{-2}	7.855360×10^{-22}	1.394327×10^{-2}
Gaze_at_Screen_Right	5.173883×10^{-1}	5.228333×10^{-1}	4.819030×10^{-1}
Short_Pauses	6.129101×10^{-2}	6.038230×10^{-2}	5.266563×10^{-2}
Difference_in_Valence	5.511903×10^{-1}	6.026717×10^{-1}	5.586536×10^{-1}
Gaze_at_Partner	1.790690×10^{-1}	1.355978×10^{-1}	1.555657×10^{-1}
Gaze_at_Screen_Left	4.294050×10^{-1}	4.452297×10^{-1}	4.510707×10^{-1}
Long_Pauses	1.474565×10^{-2}	9.933266×10^{-3}	1.916058×10^{-3}
T1_T2_add	3.027555×10^{-32}	1.666667×10^{-2}	9.423054×10^{-19}
Gaze_Other	5.388054×10^{-2}	7.411003×10^{-2}	6.656883×10^{-2}
T_hist	9.891350×10^{-3}	8.333333×10^{-3}	2.176802×10^{-2}
T1_T1_add	0.000000	0.000000	0.000000

Table B.3: Features' Mean values in each of the Silent Wanderers' states

Feature	**InitialState**	**MoreProbableState**	**LessProbableState**
T_ratio_add_rem	1.312014×10^{-2}	9.999978×10^{-1}	6.729559×10^{-3}
T_add	3.644306×10^{-2}	3.281263×10^{-1}	1.682390×10^{-2}
Speech_Overlap	6.135501×10^{-2}	1.679682×10^{-1}	4.213296×10^{-1}
Speech_Activity	2.082582×10^{-1}	3.460734×10^{-1}	5.755465×10^{-1}
Overlap_to_Speech_Ratio	1.891342×10^{-1}	3.170220×10^{-1}	6.084412×10^{-1}
Silence	7.586753×10^{-1}	6.372338×10^{-1}	4.604999×10^{-1}
T_action	5.638728×10^{-2}	1.191416×10^{-1}	5.272196×10^{-2}
normalized_time	3.157809×10^{-1}	4.738481×10^{-1}	7.326249×10^{-1}
T_remove	1.092886×10^{-1}	1.552110×10^{-6}	1.347397×10^{-1}
redundant_exist	3.997135×10^{-2}	5.468935×10^{-2}	2.257087×10^{-2}
Gaze_at_Screen_Right	5.556238×10^{-1}	6.204418×10^{-1}	6.182313×10^{-1}
Gaze_at_Screen_Left	2.746859×10^{-1}	2.511985×10^{-1}	2.227564×10^{-1}
Positive_Valence	2.501105×10^{-1}	2.826254×10^{-1}	2.784481×10^{-1}
T_help	3.497000×10^{-2}	6.249981×10^{-3}	1.707482×10^{-10}
Gaze_at_Partner	1.141224×10^{-1}	1.443724×10^{-1}	1.290232×10^{-1}
Difference_in_Valence	3.721043×10^{-1}	3.680229×10^{-1}	3.831570×10^{-1}
Arousal	2.464960×10^{-1}	3.058044×10^{-1}	2.857545×10^{-1}
T1_T1_rem	2.914167×10^{-2}	1.625856×10^{-13}	1.646490×10^{-12}
T1_T2_rem	6.827181×10^{-5}	4.079701×10^{-11}	3.360833×10^{-2}
Gaze_Other	4.831607×10^{-2}	1.046721×10^{-1}	5.910927×10^{-2}
Gaze_at_Robot	8.571055×10^{-2}	4.328244×10^{-2}	4.276117×10^{-2}
T1_T1_add	0.000000	0.000000	0.000000
Negative_Valence	2.408355×10^{-1}	3.125240×10^{-1}	2.792128×10^{-1}
Short_Pauses	2.230099×10^{-1}	1.495351×10^{-1}	1.287077×10^{-1}
T1_T2_add	5.158890×10^{-15}	6.249980×10^{-2}	6.094868×10^{-10}
Long_Pauses	2.606827×10^{-2}	6.442913×10^{-3}	9.601131×10^{-3}
T_hist	7.199166×10^{-3}	2.083376×10^{-2}	3.378669×10^{-2}

Table B.4: p-values from Kruskal-Wallis test on the Expressive Explorers' states

Feature	LessProbableState-MoreProbableState	LessProbableState-InitialState	MoreProbableState-InitialState	LessProbableState-MoreProbableState-InitialState
T_add	$6.459889 \times 10^{-272}$	7.786821×10^{-7}	$7.321670 \times 10^{-241}$	0.000000
T_ratio_add_rem	0.000000	8.011061×10^{-7}	$6.043280 \times 10^{-301}$	0.000000
T_action	3.229241×10^{-62}	1.116775×10^{-12}	$1.255941 \times 10^{-128}$	$2.724522 \times 10^{-134}$
normalized_time	6.573843×10^{-13}	$1.588430 \times 10^{-118}$	2.121637×10^{-76}	$2.268858 \times 10^{-127}$
Speech_Overlap	1.823268×10^{-20}	4.726497×10^{-96}	4.636181×10^{-33}	1.145578×10^{-91}
Overlap_to_Speech_Ratio	2.359913×10^{-16}	4.714581×10^{-82}	2.061077×10^{-29}	8.026889×10^{-78}
Speech_Activity	1.878127×10^{-18}	1.021272×10^{-76}	9.454754×10^{-27}	1.719850×10^{-74}
Silence	4.052784×10^{-11}	1.423945×10^{-69}	4.624967×10^{-33}	8.671185×10^{-69}
T_remove	1.923088×10^{-30}	8.584127×10^{-11}	3.403098×10^{-8}	1.086990×10^{-34}
Gaze_at_Robot	2.408710×10^{-20}	2.522376×10^{-1}	1.119077×10^{-13}	4.107882×10^{-21}
redundant_exist	6.323133×10^{-13}	5.619846×10^{-1}	5.410339×10^{-12}	5.495916×10^{-18}
T1_T1_rem	1.071271×10^{-9}	1.175112×10^{-6}	NaN	7.707447×10^{-14}
Gaze_at_Partner	8.078957×10^{-11}	4.279389×10^{-8}	8.546495×10^{-1}	1.426036×10^{-11}
T_help	5.077449×10^{-2}	1.210030×10^{-5}	5.427370×10^{-10}	1.167470×10^{-10}
T1_T2_rem	3.871818×10^{-7}	6.822129×10^{-4}	7.513610×10^{-2}	2.112531×10^{-8}
T_hist	9.259736×10^{-6}	2.101519×10^{-6}	3.011715×10^{-1}	1.159495×10^{-7}
Gaze_at_Screen_Right	3.147718×10^{-7}	6.138099×10^{-2}	1.809002×10^{-3}	8.358584×10^{-7}
Gaze_at_Screen_Left	5.380571×10^{-6}	3.960269×10^{-4}	6.152767×10^{-1}	8.665531×10^{-6}
Long_Pauses	3.314372×10^{-4}	1.312236×10^{-3}	9.461985×10^{-1}	4.948046×10^{-4}
Arousal	4.132118×10^{-3}	4.371238×10^{-4}	3.136678×10^{-1}	7.750580×10^{-4}
Short_Pauses	1.013922×10^{-2}	2.814769×10^{-4}	1.729218×10^{-1}	8.445037×10^{-4}
Negative_Valence	3.444524×10^{-3}	5.710819×10^{-3}	8.727249×10^{-1}	3.942160×10^{-3}
Positive_Valence	1.202900×10^{-1}	8.595192×10^{-3}	1.684401×10^{-1}	2.711117×10^{-2}
Gaze_Other	6.718909×10^{-1}	6.410131×10^{-2}	1.878263×10^{-2}	5.268550×10^{-2}
T1_T2_add	1.782952×10^{-1}	NaN	2.607401×10^{-1}	2.148112×10^{-1}
Difference_in_Valence	1.769718×10^{-1}	7.301056×10^{-1}	3.636468×10^{-1}	3.731152×10^{-1}
T1_T1_add	7.361626×10^{-1}	2.372005×10^{-1}	1.682336×10^{-1}	4.040213×10^{-1}

Table B.5: p-values from Kruskal-Wallis test on the Calm Tinkerers' states

Feature	InitialState-MoreProbableState	InitialState-LessProbableState	MoreProbableState-LessProbableState	InitialState-MoreProbableState-LessProbableState
T_ratio_add_rem	$1.046633 \times 10^{-214}$	3.203160×10^{-15}	$1.448010 \times 10^{-209}$	$2.208200 \times 10^{-300}$
T_add	$8.825740 \times 10^{-194}$	2.671724×10^{-15}	$1.084962 \times 10^{-136}$	$2.779006 \times 10^{-241}$
Speech_Overlap	2.486453×10^{-26}	$1.140660 \times 10^{-105}$	4.815798×10^{-22}	$3.079549 \times 10^{-103}$
Speech_Activity	2.699760×10^{-26}	4.539962×10^{-96}	4.068440×10^{-19}	1.789244×10^{-94}
Silence	6.369222×10^{-31}	1.859685×10^{-91}	9.868460×10^{-16}	7.739133×10^{-92}
T_action	2.530858×10^{-96}	9.718553×10^{-15}	3.560970×10^{-28}	2.236893×10^{-84}
Overlap_to_speech_ratio	1.266769×10^{-19}	1.470592×10^{-84}	9.053385×10^{-20}	1.396038×10^{-82}
Normalized_time	9.962378×10^{-23}	1.277583×10^{-57}	2.040765×10^{-10}	3.391638×10^{-59}
T_remove	4.195097×10^{-14}	3.443724×10^{-19}	4.953212×10^{-44}	3.040368×10^{-51}
T1_T1_rem	3.566311×10^{-1}	6.479934×10^{-18}	4.254322×10^{-16}	2.402922×10^{-30}
T1_T2_rem	1.921375×10^{-1}	1.517800×10^{-9}	2.073554×10^{-9}	6.086552×10^{-16}
Redundant_exist	7.350425×10^{-13}	2.140998×10^{-2}	2.082055×10^{-7}	1.988810×10^{-13}
Positive_Valence	1.140000×10^{-4}	8.498930×10^{-11}	2.452327×10^{-2}	5.318031×10^{-10}
Arousal	5.355177×10^{-2}	1.451686×10^{-9}	1.509734×10^{-4}	5.760692×10^{-9}
Gaze_at_robot	9.302748×10^{-8}	4.740781×10^{-3}	4.274607×10^{-3}	5.047403×10^{-7}
Negative_Valence	7.412760×10^{-1}	3.148931×10^{-6}	3.576009×10^{-5}	1.455941×10^{-6}
T_help	6.274225×10^{-5}	2.649523×10^{-4}	3.834952×10^{-1}	5.631461×10^{-6}
Gaze_at_screen_right	3.407637×10^{-4}	9.740129×10^{-2}	3.593597×10^{-6}	5.796425×10^{-6}
Short_pauses	4.721115×10^{-3}	9.175213×10^{-7}	9.540181×10^{-2}	6.114326×10^{-6}
Difference_in_Valence	4.211367×10^{-6}	6.285657×10^{-3}	5.459578×10^{-2}	2.763204×10^{-5}
Gaze_at_partner	1.282025×10^{-1}	6.511073×10^{-5}	2.696452×10^{-2}	3.242851×10^{-4}
Gaze_at_screen_left	1.130881×10^{-3}	4.568822×10^{-1}	1.584547×10^{-2}	4.372545×10^{-3}
Long_pauses	6.461623×10^{-1}	1.588477×10^{-2}	3.468741×10^{-3}	8.647671×10^{-3}
T1_T2_add	4.872548×10^{-3}	2.015286×10^{-2}	5.295830×10^{-1}	2.237594×10^{-2}
Gaze_other	8.545218×10^{-1}	8.659117×10^{-2}	1.429300×10^{-1}	1.664223×10^{-1}
T_hist	7.147372×10^{-1}	3.024626×10^{-1}	1.804699×10^{-1}	3.499533×10^{-1}
T1_T1_add	8.703039×10^{-1}	5.027353×10^{-1}	4.142353×10^{-1}	7.077864×10^{-1}

Table B.6: p-values from Kruskal-Wallis test on the Silent Wanderers' states

Feature	**InitialState-LessProbableState**	**InitialState-MoreProbableState**	**LessProbableState-MoreProbableState**	**InitialState-LessProbableState-MoreProbableState**
T_ratio_add_rem	4.589301×10^{-1}	$5.220515 \times 10^{-122}$	$6.663973 \times 10^{-130}$	$5.535321 \times 10^{-183}$
T_add	4.577846×10^{-1}	$6.010859 \times 10^{-100}$	$2.610590 \times 10^{-114}$	$2.388349 \times 10^{-160}$
Speech_overlap	5.892624×10^{-97}	7.100573×10^{-18}	9.712381×10^{-31}	2.760344×10^{-98}
Speech_Activity	8.477783×10^{-90}	2.493802×10^{-20}	7.562644×10^{-23}	1.735427×10^{-89}
Overlap_to_speech_ratio	7.595498×10^{-72}	3.505213×10^{-11}	6.190938×10^{-30}	1.645284×10^{-75}
Silence	3.187753×10^{-70}	1.142510×10^{-13}	2.080269×10^{-19}	4.649497×10^{-69}
T_action	1.056353×10^{-1}	3.188625×10^{-32}	1.079551×10^{-44}	3.200090×10^{-49}
Normalized_time	1.518080×10^{-46}	3.510905×10^{-9}	5.506350×10^{-15}	2.159810×10^{-46}
T_remove	6.971572×10^{-1}	1.116435×10^{-11}	6.037973×10^{-11}	1.127712×10^{-10}
Redundant_exist	3.512804×10^{-1}	1.635185×10^{-6}	3.081867×10^{-9}	1.046462×10^{-9}
Gaze_at_screen_right	4.238118×10^{-1}	2.854616×10^{-7}	1.300446×10^{-5}	4.386002×10^{-7}
Gaze_at_screen_left	1.959077×10^{-4}	2.097955×10^{-6}	1.816661×10^{-1}	3.475313×10^{-6}
Positive_Valence	4.477664×10^{-5}	8.188481×10^{-1}	3.719569×10^{-5}	8.771086×10^{-6}
T_help	2.145970×10^{-4}	4.619297×10^{-3}	5.533612×10^{-1}	1.305032×10^{-4}
Gaze_at_partner	6.836966×10^{-5}	3.239169×10^{-1}	8.317920×10^{-3}	2.405678×10^{-4}
Difference_in_Valence	5.798869×10^{-4}	7.765849×10^{-1}	4.280245×10^{-3}	9.162789×10^{-4}
Arousal	3.133786×10^{-2}	2.050905×10^{-1}	4.771730×10^{-4}	2.053134×10^{-3}
T1_T1_rem	2.168625×10^{-1}	4.151916×10^{-3}	3.627620×10^{-2}	1.568344×10^{-2}
T1_T2_rem	6.483012×10^{-1}	1.048949×10^{-2}	4.416490×10^{-3}	2.137952×10^{-2}
Gaze_other	1.069325×10^{-2}	4.460858×10^{-2}	8.001324×10^{-1}	2.297604×10^{-2}
Gaze_at_robot	1.107108×10^{-1}	8.746928×10^{-3}	2.324421×10^{-1}	2.894125×10^{-2}
T1_T1_add	NaN	1.155100×10^{-1}	9.639491×10^{-2}	7.286920×10^{-2}
Negative_Valence	9.084048×10^{-1}	6.888763×10^{-2}	3.284085×10^{-2}	8.004833×10^{-2}
Short_pauses	3.175136×10^{-1}	9.025873×10^{-2}	2.226728×10^{-1}	1.799852×10^{-1}
T1_T2_add	5.001169×10^{-1}	2.773726×10^{-1}	8.430488×10^{-2}	1.889137×10^{-1}
Long_pauses	6.990590×10^{-1}	1.992827×10^{-1}	2.787486×10^{-1}	3.964593×10^{-1}
T_hist	3.746288×10^{-1}	9.370430×10^{-1}	3.645077×10^{-1}	5.606479×10^{-1}

C Appendix C

School of Computer and Communication Sciences
CHILI - Computer-Human Interaction in Learning and Instruction

Jauwairia Nasir
EPFL
Lausanne
Switzerland

Lausanne, May 18th, 2022

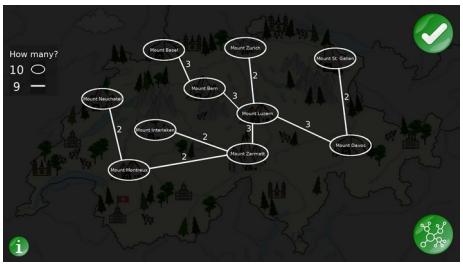
EPFL JUSThink Study 1 at School - [redacted] Report

Dear responsible for the school [redacted]

With this report, it is our pleasure to share with you a few details about the study that took place in March 2022 at [redacted] where students in pairs interacted with our social humanoid robot named QTrobot.

The study was part of the JUSThink project[1], developed in CHILI lab at EPFL. It involves a collaborative problem solving activity for school children, mediated by the robot (see details [1]). The aim is to improve their computational thinking skills, by exercising abstract reasoning on graphs as well as collaboration skills. The activity also serves as a platform to build intelligent autonomous social robots that can promote childrens learning by assisting teachers through complementary activities.

In this version of the activity called JUSThink Pro the students interacted in pairs with each other as well as with a robot for about an hour. The learning concept underlying the robot mediated collaborative activity is Minimum Spanning Trees that the students try to learn through an optimization problem. They were shown a fictional map of Switzerland with gold mines and they were asked to connect all the gold mines with rail-tracks by spending as little money as possible. Each student alternates between two different views (as shown below) with two different functionalities - the partial information as well as different roles enforce collaboration by design. The robot provides suggestions to the students on what kind of behaviors they can adopt that may help them to learn better. The students have a choice to follow the suggestion or not. Before and after the game play session, which is of 30 minutes, the students individually answer 10 questions on the learning concept (pre and post test). At the very end, they also answer a questionnaire on their self and robot perception. During the game play, they can submit a solution as many times as they want.

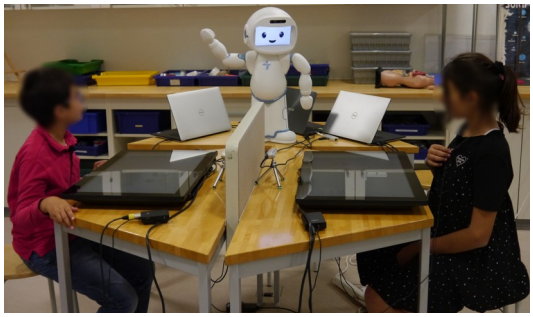


[1] https://www.epfl.ch/labs/chili/index-html/research/animatas/justhink/

School of Computer and Communication Sciences

Jauwairia Nasir
EPFL IC IINFCOM CHILI
RLC D1 740
Station 20
CH–1015 Lausanne

+41 77 997 89 58
jauwairia.nasir@epfl.ch
http://people.epfl.ch/jauwairia.nasir



More precisely, in this study with 6 schools and around 140 students, we evaluated the effectiveness of two robot variants, **Harry** and **Hermione**, that try to infer, through students' speech behaviors as well as their actions on the activity, if the team is *productively engaged* in the learning process or not and then suggests behaviors accordingly. The robots were designed building on our research of the last three years on the concept of ***Productive Engagement*** (see details in [2,3]). *Productive Engagement* (PE) is referred to as the engagement that is conducive to learning where we validated the existence of PE through a data driven machine learning pipeline. We observed that indeed some behavioral profiles are linked to higher learning gains versus others using previously collected data from several schools in Switzerland [2,3]. Specifically, we found that the quantity and quality of speech activity between a pair is most discriminatory in separating those who learn from those who do not end up learning. Further, among those who learn, they display two types of problem solving strategies linked with different emotional profiles. The two robots, Harry and Hermione, acting as better-skilled peers, use different levels of information from the outcomes of our Productive Engagement framework to suggest behaviors to the students to help them learn better based on the speech behavior of the team as well as their actions on the activity. Both robots use different techniques for selecting a suggestion where out of the two, Hermione is developed with more sophisticated decision making abilities. We divide the students randomly among the two conditions (either they interact with Harry or with Hermione). We report on the following metrics:

- *Error_in game* refers to how far is the cost of the last solution found by the students with respect to the optimal cost. A zero error indicates they successfully found the solution.
- *LG_relative* and *LG_joint* refers to the learning gain of the team based on their pre-test and post test scores. *LG_relative* grasps how much the participant learned of the knowledge that they did not possess before the activity (we calculate for each student and then take the average). *LG_joint* grasps the amount of knowledge acquired together by the team members during the activity.
- *Reflection* refers to how much the team looked at their past solutions.
- *Task and Social Engagement, Self Competence, Tensed* refers to how the team perceived their engagement with the task, their partner, their competence in the activity, and their own stress levels, respectively.
- *Robot Likeability, Competence Trust* refers to how the team rated the robot in terms of its likeability, and its competence, respectively.
- *Suggestions Usefulness* refers to how the team rated the usefulness of a suggestion, after every suggestion, during the game.

School of Computer and Communication Sciences

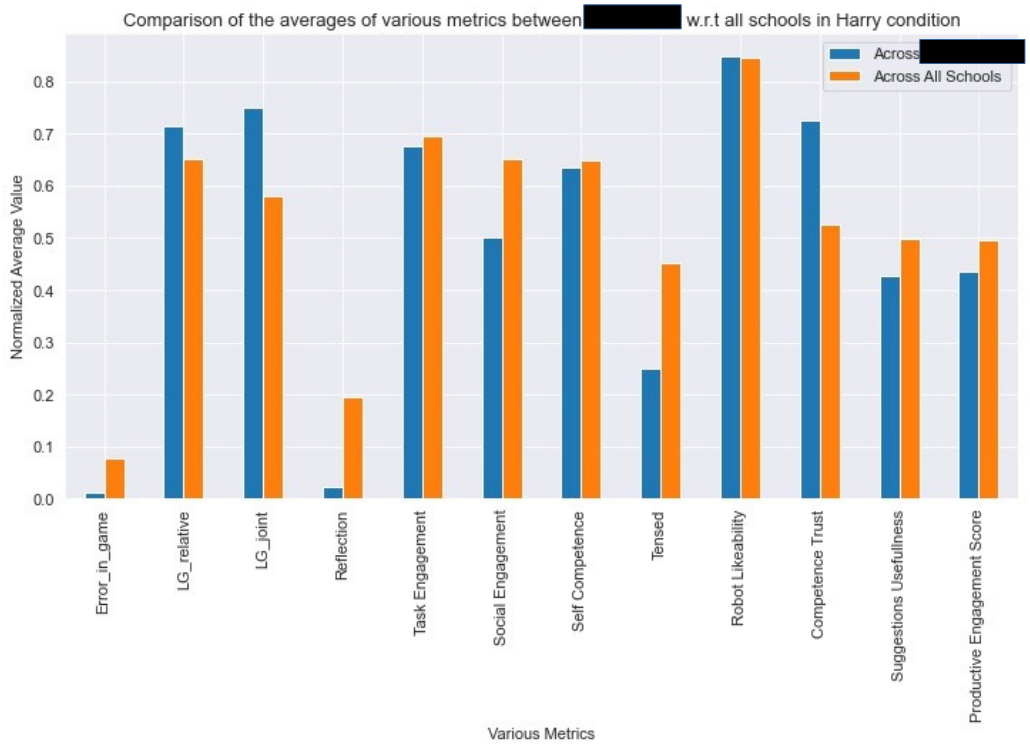
Jauwairia Nasir
EPFL IC IINFCOM CHILI
RLC D1 740
Station 20
CH–1015 Lausanne

+41 77 997 89 58
jauwairia.nasir@epfl.ch
http://people.epfl.ch/jauwairia.nasir

- *Productive Engagement Score* refers to the average students' productive engagement. Please note, the higher the score, our system considers them to be more productively engaged.

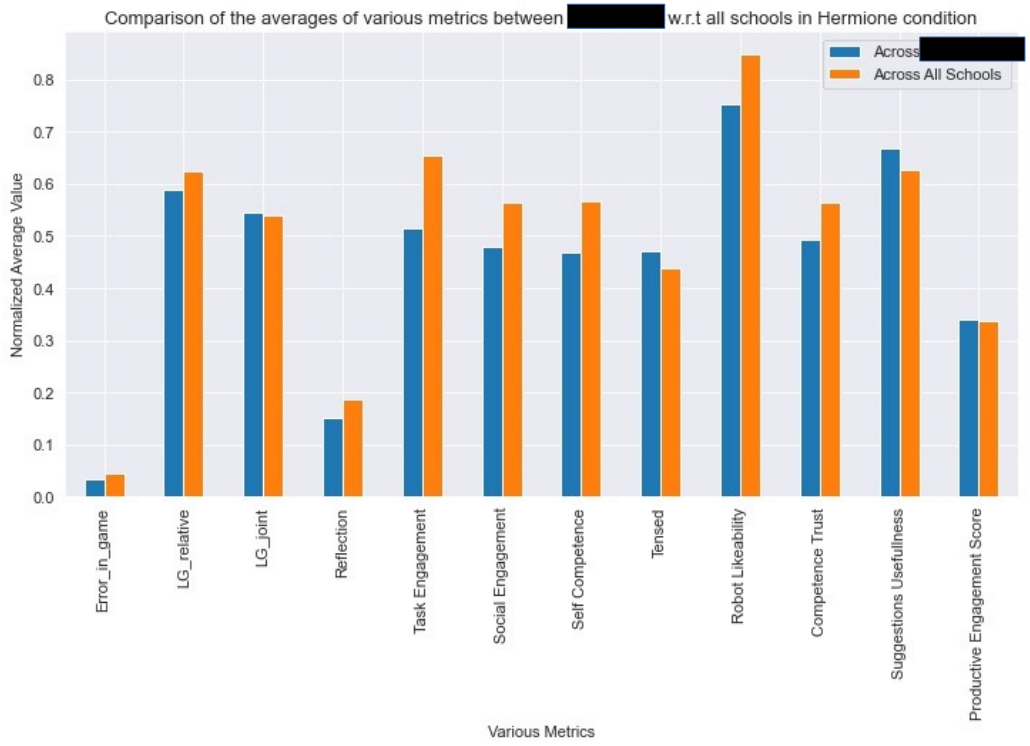
Personalized Feedback for [redacted]

The school performed as follows on all the above listed metrics with respect to all the students across all the schools. We show the results separately for the two groups, i.e., the group with students that interacted with Harry and the group with students that interacted with Hermione.



School of Computer and Communication Sciences
Jauwairia Nasir
EPFL IC IINFCOM CHILI
RLC D1 740
Station 20
CH–1015 Lausanne
+41 77 997 89 58
jauwairia.nasir@epfl.ch
http://people.epfl.ch/jauwairia.nasir

School of Computer and Communication Sciences
CHILI - Computer-Human Interaction in Learning and Instruction



1. The students at [redacted] especially with the strategy adopted by Harry, seem to have higher learning gains than the rest of the schools. **This shows the potentially positive effect of exploratory activities, incorporated with a carefully designed robot, on the learning gain of the students** [redacted]
2. For both cases, we observe that the social engagement displayed by the teams is lower than the average of other schools. **This motivates the need of collaborative activities that could enforce and improve social engagement.**
3. While with Harry, the students acheived very low errors in the game, experienced less stress as well as rated the robot high on competence compared to other schools; with Hermione, they achieved higher levels of productive engagement score as well as found the suggestions more useful. **This shows the importance of having portfolio of learning activities with personalized feedback strategies that target different aspects of the interaction.**

The analysis of the team's interaction with the robot is currently ongoing. Please contact Jauwairia Nasir (jauwairia.nasir@epfl.ch) if you wish to be informed of the results of the analysis or if you have any question. We would be happy to share scientific publications on this work with the schools as soon as they are published.

We immensely thank you for participating in our research.

Sincerely,

Jauwairia Nasir, on behalf of the EPFL research team

References

[1] Nasir*, J., Norman*, U., Bruno, B., and Dillenbourg, P. (2020); When Positive Perception of the Robot Has No Effect on

School of Computer and Communication Sciences

Jauwairia Nasir
EPFL IC IINFCOM CHILI
RLC D1 740
Station 20
CH–1015 Lausanne

+41 77 997 89 58
jauwairia.nasir@epfl.ch
http://people.epfl.ch/jauwairia.nasir

Learning. IEEE RO-MAN 2020 - 28th IEEE International Symposium on Robot and Human Interactive Communication, 2020.

[2] Nasir, J., Bruno, B., Chetouani, M., and Dillenbourg, P. What if Social Robots Look for Productive Engagement?. International Journal of Social Robotics, 2021.

[3] Nasir, J., Kothiyal, A., Bruno, B., and Dillenbourg, P. Many are the ways to learn: Identifying multi-modal behavioral profiles of collaborative learning in constructivist activities. Int. Journal of Computer-Supported Collaborative Learning (IJCSCL), 2021.

School of Computer and Communication Sciences

Jauwairia Nasir
EPFL IC IINFCOM CHILI
RLC D1 740
Station 20
CH–1015 Lausanne

+41 77 997 89 58
jauwairia.nasir@epfl.ch
http://people.epfl.ch/jauwairia.nasir

Bibliography

Akkila, A. N., Almasri, A., Ahmed, A., Masri, N., Sultan, Y. A., Mahmoud, A. Y., Zaqout, I., & Abu-naser, S. S. (2019). Survey of Intelligent Tutoring Systems Up To the End of 2017. *International Journal of Academic Information Systems Research*, *3*(3), 71–81. www.ijeais.org/ijaisr

Alves-Oliveira, P., Sequeira, P., Melo, F. S., Castellano, G., & Paiva, A. (2019). Empathic robot for group learning: a field study. *ACM THRI*, *8*(1). https://doi.org/10.1145/3300188

Alyuz, N., Okur, E., Oktay, E., Genc, U., Aslan, S., Mete, S. E., Stanhill, D., Arnrich, B., & Esme, A. A. (2016). Towards an emotional engagement model: Can affective states of a learner be automatically detected in a 1:1 learning scenario? *CEUR Workshop Proceedings*, *1618*(1), 1–7.

Anzalone, S. M., Boucenna, S., Ivaldi, S., & Chetouani, M. (2015). Evaluating the Engagement with Social Robots. *International Journal of Social Robotics*, *7*(4), 465–478. https://doi.org/10.1007/s12369-015-0298-7

Atamna, A., & Clavel, C. (2020). Hri-rnn: a user-robot dynamics-oriented rnn for engagement decrease detection, 4198–4202. https://doi.org/10.21437/Interspeech.2020-1261

Ausubel, D. P. (1960). The use of advance organizers in the learning and retention of meaningful verbal material. *Journal of Educational Psychology*, *51*(5), 267–272. https://doi.org/10.1037/h0046669

Baker, R., & Siemens, G. (2012). Educational Data Mining and Learning Analytics. In R. K. Sawyer (Ed.), *Chls* (pp. 253–272). Cambridge University Press. https://doi.org/10.1017/CBO9781139519526.016

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom, 383–390. https://doi.org/10.1145/985692.985741

Baker, R. S., Corbett, A. T., & 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*, 406–415. https://doi.org/10.1007/978-3-540-69132-7-44

Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments.

International Journal of Human Computer Studies, *68*(4), 223–241. https://doi.org/10.1016/j.ijhcs.2009.12.003

Baltrusaitis, T., McDuff, D., Banda, N., Mahmoud, M., Kaliouby, R., Robinson, P., & Picard, R. (2011). Real-Time Inference of Mental States from Facial Expressions and Upper Body Gestures, 909–914. https://doi.org/10.1109/FG.2011.5771372

Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit, 1–10. https://doi.org/10.1109/WACV.2016.7477553

Bannert, M., Reimann, P., & Sonnenberg, C. (2014). Process mining techniques for analysing patterns and strategies in students' self-regulated learning. *Metacognition and Learning*, *9*(2), 161–185. https://doi.org/10.1007/s11409-013-9107-6

Barron, B. (2003). When smart groups fail. *The journal of the learning sciences*, *12*(3), 307–359.

Barron, B., Schwartz, D., Vye, N., Moore, A., Petrosino, A., Zech, L., & Bransford, J. (1998). Doing with understanding: lessons from research on problem-and project-based learning. *Journal of the learning sciences*, *7*(3-4), 271–311.

Bartneck, C., Croft, E., & Kulic, D. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *IJSR*, *1*(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Bassiou, N., Tsiartas, A., Smith, J., Bratt, H., Richey, C., Shriberg, E., D'Angelo, C., & Alozie, N. (2016). Privacy-preserving speech analytics for automatic assessment of student collaboration. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *08-12-September-2016*, 888–892. https://doi.org/10.21437/Interspeech.2016-1569

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*, *12*(5). https://doi.org/10.1371/journal.pone.0178126

Beal, C. R., Qu, L., & Lee, H. (2004). Basics of Feedback Control - Elements of Feedback control | Instrumentation and Control Engineering, 151–156.

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: a review. *Science Robotics*, *3*(21), eaat5954.

Benitez-Quiroz, C. F., Srinivasan, R., & Martinez, A. M. (2016). Emotionet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5562–5570. https://doi.org/10.1109/CVPR.2016.600

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 at. (May 2014), 45–52.

Blaye, A. (1988). *Confrontation socio-cognitive et résolution de problèmes* (Doctoral dissertation). Centre de Recherche en Psychologie Cognitive, Université de Provence. 13261 Aix-en-Provence, France.

Blikstein, P. (2013). Multimodal learning analytics. https://doi.org/10.1145/2460296.2460316

Blikstein, P., & Worsley, M. (2016). Multimodal Learning Analytics and Education Data Mining: using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, *3*(2), 220–238. https://doi.org/10.18608/jla.2016.32.11

Bourguet, M.-L., Jin, Y., Shi, Y., Chen, Y., Rincon-Ardila, L., & Venture, G. (2020). Social robots that can sense and improve student engagement. *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 127–134. https://doi.org/10.1109/TALE48869.2020.9368438

Brooks, J. M., & Brooks, M. (1993). In search of understanding: the case for constructivist classrooms.

Brown, L. V., Kerwin, R., & Howard, A. M. (2013). Applying behavioral strategies for student engagement using a robotic educational agent. *Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013*, 4360–4365. https://doi.org/10.1109/SMC.2013.744

C., C., & H., M. (2009). Empirically Building and Evaluating a Probabilistic Model of User Affect. *User Modeling and User-Adapted Interaction*, *19*, 267–303.

Campione, E., & Véronis, J. (2002). A large-scale multilingual study of pause duration. *Speech Prosody 2002. Proceedings of the1st International Conference on Speech Prosody*, 199–202. http://www.isca-speech.org/archive/sp2002/sp02_199.html

Carey, T. A., & Mullan, R. J. (2004). What is socratic questioning? *Psychotherapy: theory, research, practice, training*, *41*(3), 217.

Castanheira, M. L., Crawford, T., Dixon, C. N., & Green, J. L. (2000). Interactional Ethnography: An Approach to Studying the Social Construction of Literate Practices. *Linguistics and Education*, *11*(4), 353–400. https://doi.org/10.1016/S0898-5898(00)00032-2

Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., & McOwan, P. W. (2012). Detecting engagement in hri: An exploration of social and task-based context. *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*, 421–428. https://doi.org/10.1109/SocialCom-PASSAT.2012.51

Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., & Mcowan, P. W. (2014). Context-Sensitive Affect Recognition for a Robotic Game Companion. *ACM Transactions on Interactive Intelligent Systems*, *4*(2), 1–25. https://doi.org/10.1145/2622615

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, 119–126. https://doi.org/10.1145/1647314.1647336

Chalmers, C. (2018). Robotics and computational thinking in primary school. *IJCCI*, *17*, 93–100.

Chandra, S., Alves-Oliveira, P., Lemaignan, S., Sequeira, P., Paiva, A., & Dillenbourg, P. (2015). Can a child feel responsible for another in the presence of a robot in a collaborative learning activity? *RO-MAN 2015*, 167–172. https://doi.org/10.1109/ROMAN.2015.7333678

Chang, C.-j., Chang, M.-h., Chiu, B.-c., Liu, C.-c., Chao, P.-y., Lai, C.-h., Wu, S.-w., Chang, C.-k., & Chen, W. (2017). An analysis of student collaborative problem solving activities

mediated by collaborative simulations. *Computers & Education*, *114*(300), 222–235. https://doi.org/10.1016/j.compedu.2017.07.008

Chaouachi, M., Chalfoun, P., Jraidi, I., & Frasson, C. (2010). Affect and Mental Engagement: Towards Adaptability for Intelligent Systems. *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, (Flairs), 355–360.

Chase, C. C., Chin, D. B., Oppezzo, M. A., & Schwartz, D. L. (2009). Teachable agents and the protege effect: increasing the effort towards learning. *J Sci Educ Tech*, *18*(4), 334–352.

Cherubini, M., Nüssli, M.-A., & Dillenbourg, P. (2008). Deixis and gaze in collaborative work at a distance (over a shared map) a computational model to detect misunderstandings. *Proceedings of the 2008 symposium on Eye tracking research & applications*, 173–180.

Chi, M. T., & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist*, *49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Chollet, F. et al. (2015). *Keras*. https://github.com/fchollet/keras

Chow, J. Y., Davids, K., Button, C., & Renshaw, I. (2015). *Nonlinear pedagogy in skill acquisition: an introduction*. Routledge.

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

Cohn, J. (2006). Foundations of Human Computing: Facial Expression and Emotion. *ICMI'06: 8th International Conference on Multimodal Interfaces, Conference Proceeding*, 233–238. https://doi.org/10.1007/978-3-540-72348-6_1

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of student knowledge. http://act-r.psy.cmu.edu/papers/893/CorbettAnderson1995.pdf

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278. https://doi.org/10.1007/BF01099821

Cordova, D., & Lepper, M. (1996). Intrinsic motivation and the process of learning: beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, *88*, 715–730. https://doi.org/10.1037/0022-0663.88.4.715

Corrigan, L. J., 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.

Craig, S. D., Witherspoon, A., D'Mello, S. K., Graesser, A., & McDaniel, B. (2007). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*, *18*(1-2), 45–80. https://doi.org/10.1007/s11257-007-9037-6

Crescenzi-Lanna, L. (2020). Multimodal learning analytics research with young children: a systematic review. *British Journal of Educational Technology*, *51*. https://doi.org/10.1111/bjet.12959

Csanadi, A., Eagan, B. R., Kollar, I., Shaffer, D. W., & Fischer, F. (2018). When coding-and-counting is not enough: using epistemic network analysis (ena) to analyze verbal data

in cscl research. *International Journal of Computer-Supported Collaborative Learning*, *13*, 419–438.

Deci, E. (2017). Intrinsic motivation and self-determination. https://doi.org/10.1016/B978-0-12-809324-5.05613-3

Desmarais, M. C., & Baker, R. S. J. d. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, *22*(1), 9–38. https://doi.org/10.1007/s11257-011-9106-8

Desmarais, M. C., & Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, *22*(1-2), 9–38. https://doi.org/10.1007/s11257-011-9106-8

Dewan, M. A. A., Murshed, M., & Lin, F. (2019). Engagement detection in online learning: a review. *Smart Learning Environments*, *6*(1), 1–20. https://doi.org/10.1186/s40561-018-0080-z

Dillenbourg, P. (1999). What do you mean by collaborative learning? In P. Dillenbourg (Ed.), *Collaborative-learning: cognitive and computational approaches* (pp. 1–19).

Dillenbourg, P., Baker, M., Blaye, A., & O'Malley, C. (1996). The evolution of research on collaborative learning. In H. Spada & P. Reimann (Eds.), *Learning in humans and machines: towards an interdisciplinary learning science* (pp. 189–211). Oxford, Elsevier.

Dillenbourg, P., Järvelä, S., & Fischer, F. (2009). The Evolution of Research on Computer-Supported Collaborative Learning:From Design to Orchestration. *Technology-enhanced learning* (pp. 3–19). https://doi.org/10.1007/978-1-4020-9827-7

Dindar, M., Jarvela, S., Ahola, S., Huang, X., & Zhao, G. (2020). Leaders and followers identified by emotional mimicry during collaborative learning: a facial expression recognition study on emotional valence. *IEEE Transactions on Affective Computing*.

D'Mello, S. (2013). A Selective Meta-Analysis on the Relative Incidence of Discrete Affective States During Learning With Technology. *Journal of Educational Psychology*, *105*, 1082. https://doi.org/10.1037/a0032674

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001

D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, *29*, 153–170. https://doi.org/10.1016/j.learninstruc.2012.05.003

Do-lenh, S. (2012). Supporting Reflection and Classroom Orchestration with Tangible Tabletops. *5313*, 241. https://doi.org/10.5075/epfl-thesis-5313

Driskell, J., Goodwin, G., Salas, E., & O'Shea, P. (2006). What makes a good team player? personality and team effectiveness. *Group Dynamics: Theory, Research, and Practice*, *10*, 249–271. https://doi.org/10.1037/1089-2699.10.4.249

Ekman, P., & Friesen, W. (1978). Facial action coding system: manual.

El Kaliouby, R., & Robinson, P. (2004). Real-time inference of complex mental states from facial expressions and head gestures. *2004 Conference on Computer Vision and Pattern Recognition Workshop*, 154–154. https://doi.org/10.1109/CVPR.2004.427

Elgarf, M., Calvo-Barajas, N., Alves-Oliveira, P., Perugia, G., Castellano, G., Peters, C., & Paiva, A. (2022). "and then what happens?": promoting children's verbal creativity using a robot. *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, 71–79.

Emara, M., Hutchins, N., Grover, S., Snyder, C., & Biswas, G. (2021). Examining Student Regulation of Collaborative, Computational, Problem-Solving Processes in Open-Ended Learning Environments. *Journal of Learning Analytics*, *8*(1), 49–74. https://doi.org/10.18608/jla.2021.7230

Emara, M., Rajendran, R., Biswas, G., Okasha, M., & Elbanna, A. A. (2018). Do students' learning behaviors differ when they collaborate in open-ended learning environments? *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–19.

Emerson, A., Cloude, E. B., Azevedo, R., & Lester, J. (2020a). Multimodal learning analytics for game-based learning. *British Journal of Educational Technology*, *51*(5), 1505–1526. https://doi.org/10.1111/bjet.12992

Emerson, A., Cloude, E. B., Azevedo, R., & Lester, J. (2020b). Multimodal learning analytics for game-based learning. *British journal of educational technology.*, *51*(5).

Engelmann, K., & Bannert, M. (2021). Analyzing temporal data for understanding the learning process induced by metacognitive prompts. *72*, 101205. https://doi.org/10.1016/j.learninstruc.2019.05.002

Etkina, E., Karelina, A., Ruibal-Villasenor, M., Rosengrant, D., Jordan, R., & Hmelo-Silver, C. E. (2010). Design and reflection help students develop scientific abilities: Learning in introductory physics laboratories. *Journal of the Learning Sciences*, *19*(1), 54–98. https://doi.org/10.1080/10508400903452876

Evans, A. C., Wobbrock, J. O., & Davis, K. (2016). Modeling collaboration patterns on an interactive tabletop in a classroom setting. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, *27*, 860–871. https://doi.org/10.1145/2818048.2819972

Fors, K. L. (2015). *Production and perception of pauses in speech* (Doctoral dissertation). University of Gothenburg. https://gupea.ub.gu.se/bitstream/2077/39346/1/gupea%5C_2077%5C_39346%5C_1.pdf

Foster, M. E., Gaschler, A., & Giuliani, M. (2017). Automatically Classifying User Engagement for Dynamic Multi-party Human–Robot Interaction. *International Journal of Social Robotics*, *9*(5), 659–674. https://doi.org/10.1007/s12369-017-0414-y

Fry, P. S. (1976). Success, failure, and self-assessment ratings. *Journal of Consulting and Clinical Psychology*, *44*(3), 413–419.

Gatica-Perez, D., McCowan, L., Zhang, D., & Bengio, S. (2005). Detecting group interest-level in meetings. *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, *1*, I–489.

Giannakos, M. N., Sharma, K., Pappas, I. O., Kostakos, V., & Velloso, E. (2019). Multimodal data as a means to understand the learning experience. *International Journal of Information Management*, *48*(March), 108–119. https://doi.org/10.1016/j.ijinfomgt.2019.02.003

Glachan, M., & Light, P. (1982). Peer interaction and learning: can two wrongs make a right. *Social cognition: studies of the development of understanding* (pp. 238–262). Harvester Press.

Gordon, G., Spaulding, S., Westlund, J. K., Lee, J. J., Plummer, L., Martinez, M., Das, M., & Breazeal, C. (2016). Affective personalization of a social robot tutor for children's second language skills. *Proceedings of the 30th Conference on Artificial Intelligence (AAAI 2016)*, (2011), 3951–3957.

Guo, T., Lin, T., & Antulov-Fantulin, N. (2019). Exploring interpretable lstm neural networks over multi-variable data.

Hamamsy, L. E., Johal, W., Asselborn, T., Nasir, J., & Dillenbourg, P. (2019). Learning by collaborative teaching: an engaging multi-party CoWriter activity. *RO-MAN 2019*, 1–8. https://doi.org/10.1109/RO-MAN46459.2019.8956358

Hausmann, R. G., Chi, M. T., & Roy, M. (2004). Proceedings of the Annual Meeting of the Cognitive Science mechanisms. *Proceedings of the Annual Meeting of the Cognitive Science Society, 26*, 547–552.

Hayashi, Y. (2019). Detecting collaborative learning through emotions: an investigation using facial expression recognition. *International conference on intelligent tutoring systems*, 89–98.

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*(4), 555–568. https://doi.org/10.1016/j.wocn.2010.08.002

Henrie, C. R., Halverson, L. R., & Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: A review. *Computers and Education, 90*, 36–53. https://doi.org/10.1016/j.compedu.2015.09.005

Herman, A. M., Critchley, H. D., & Duka, T. (2018). The role of emotions and physiological arousal in modulating impulsive behaviour. *Biological Psychology*, *133*, 30–43. https://doi.org/https://doi.org/10.1016/j.biopsycho.2018.01.014

Hmelo-Silver, C. E. (2004). Problem-based learning: what and how do students learn? *Educational psychology review*, *16*(3), 235–266.

Hone, K. (2006). Empathic agents to reduce user frustration: the effects of varying agent characteristics. *Interacting with Computers*, *18*(2), 227–245. https://doi.org/10.1016/j.intcom.2005.05.003

Huang, C.-m., Andrist, S., Sauppé, A., & Mutlu, B. (2015). Using gaze patterns to predict task intent in collaboration. *6*(July), 1–12. https://doi.org/10.3389/fpsyg.2015.01049

Huang, C.-M., & Mutlu, B. (2014). Learning-based modeling of multimodal behaviors for humanlike robots, 57–64. https://doi.org/10.1145/2559636.2559668

Huang, K., Bryant, T., & Schneider, B. (2019). Identifying collaborative learning states using unsupervised machine learning on eye-tracking, physiological and motion sensor data. *EDM 2019 - Proceedings of the 12th International Conference on Educational Data Mining*, (Edm), 318–323.

Ioannou, A., & Makridou, E. (2018). Exploring the potentials of educational robotics in the development of computational thinking: a summary of current research and practical

proposal for future work. *Education and Information Technologies*, *23*(6), 2531–2544. https://doi.org/10.1007/s10639-018-9729-z

Ishii, R., & Nakano, Y. I. (2010). An empirical study of eye-gaze behaviors. *Proceedings of the 2010 workshop on Eye gaze in intelligent human machine interaction - EGIHMI '10*, (April 2016), 33–40. https://doi.org/10.1145/2002333.2002339

Ishii, R., Shinohara, Y., Nakano, I., & Nishida, T. (2011). Combining Multiple Types of Eye-gaze Information to Predict User ' s Conversational Engagement. *Human Factors*.

Isohätälä, J., Järvenoja, H., & Järvelä, S. (2017). Socially shared regulation of learning and participation in social interaction in collaborative learning. *International Journal of Educational Research*, *81*, 11–24. https://doi.org/10.1016/j.ijer.2016.10.006

Ivtzan, I., Lomas, T., Wong, P., & Niemiec, R. (2015). Second wave positive psychology: Embracing the dark side of life.

Jacobson, M. J., Kapur, M., & Reimann, P. (2016). Conceptualizing Debates in Learning and Educational Research: Toward a Complex Systems Conceptual Framework of Learning. *Educational Psychologist*, *51*(2), 210–218. https://doi.org/10.1080/00461520.2016.1166963

Järvelä, S., Gašević, D., Seppänen, T., Pechenizkiy, M., & Kirschner, P. A. (2020). Bridging learning sciences, machine learning and affective computing for understanding cognition and affect in collaborative learning. *British Journal of Educational Technology*, *51*(6), 2391–2406.

Järvelä, S., & Hadwin, A. F. (2013). New frontiers: regulating learning in cscl. *Educational psychologist*, *48*(1), 25–39.

Järvelä, S., Kirschner, P. A., Hadwin, A., Järvenoja, H., Malmberg, J., Miller, M., & Laru, J. (2016). Socially shared regulation of learning in CSCL : understanding and prompting individual- and group-level shared regulatory activities. *International Journal of Computer-Supported Collaborative Learning*, *11*, 263–280. https://doi.org/10.1007/s11412-016-9238-2

Jermann, P., Mullins, D., Nüssli, M.-A., & Dillenbourg, P. (2011). Collaborative gaze footprints: correlates of interaction quality. *Connecting Computer-Supported Collaborative Learning to Policy and Practice: CSCL2011 Conference Proceedings.*, (CONF), 184–191.

Jermann, P., & Nüssli, M.-A. (2012). Effects of sharing text selections on gaze cross-recurrence and interaction quality in a pair programming task. *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 1125–1134.

Jin, H., Song, Q., & Hu, X. (2019). Auto-keras: an efficient neural architecture search system. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1946–1956.

Johal, W. (2020). Research Trends in Social Robots for Learning. *Current Robotics Reports*, *1*, 1–9. https://doi.org/10.1007/s43154-020-00008-3

Jordan, B., Henderson, A., Jordan, B., & Henderson, A. (1995). Interaction Analysis : Foundations and Practice. *The Journal of the Learning Sciences*, *4*(1), 39–103.

Jordan, M. E., & McDaniel Jr, R. R. (2014). Managing uncertainty during collaborative problem solving in elementary school teams: the role of peer influence in robotics engineering

activity. *Journal of the Learning Sciences*, *23*(4), 490–536. https://doi.org/10.1080/10508406.2014.896254

Kapoor, A., & Picard, R. W. (2006). Multimodal affect recognition in learning environments, 677. https://doi.org/10.1145/1101149.1101300

Kapur, M. (2008). Productive Failure. *Cognition and Instruction*, *26*(3), 379–424. https://doi.org/10.1080/07370000802212669

Kapur, M. (2011). Temporality matters: advancing a method for analyzing problem-solving processes in a computer-supported collaborative environment. *International Journal of Computer-Supported Collaborative Learning*, *6*(1), 39–56.

Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist*, *51*, 1–11. https://doi.org/10.1080/00461520.2016.1155457

Kapur, M., & Bielaczyc, K. (2012). Designing for Productive Failure. *Journal of the Learning Sciences*, *21*(1), 45–83. https://doi.org/10.1080/10508406.2011.591717

Kapur, M., & Kinzer, C. K. (2009). Productive failure in CSCL groups. *International Journal of Computer-Supported Collaborative Learning*, *4*(1), 21–46. https://doi.org/10.1007/s11412-008-9059-z

Käser, T., Klingler, S., Schwing, A. G., & Gross, M. (2017). Dynamic bayesian networks for student modeling. *IEEE Transactions on Learning Technologies*, *10*(4), 450–462. https://doi.org/10.1109/TLT.2017.2689017

Kashdan, T., & Biswas-Diener, R. (2014). *The upside of your dark side: why being your whole self–not just your "good" self–drives success and fulfillment*. Penguin Publishing Group. https://books.google.ch/books?id=C5QxAwAAQBAJ

Kauschke, C., Bahn, D., Vesker, M., & Schwarzer, G. (2019). The Role of Emotional Valence for the Processing of Facial and Verbal Stimuli—Positivity or Negativity Bias? *Frontiers in Psychology*, *10*, 1654. https://doi.org/10.3389/fpsyg.2019.01654

Kiderle, T., Ritschel, H., Janowski, K., Mertes, S., Lingenfelser, F., & Andre, E. (2021). Socially-aware personality adaptation. *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 1–8. https://doi.org/10.1109/ACIIW52867.2021.9666197

Kim, J., Co, H., Truong, K., Evers, V., & Truong, K. P. (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

Kim, J., Truong, K. P., Charisi, V., Zaga, C., Lohse, M., Heylen, D., & Evers, V. (2015). Vocal turn-taking patterns in groups of children performing collaborative tasks : an exploratory study. *INTERSPEECH 2015*, 1645–1649.

Kinnebrew, J. S., Loretz, K. M., & Biswas, G. (2013). A contextualized, differential sequence mining method to derive students' learning behavior patterns. *Journal of Educational Data Mining*, *5*(1), 190–219.

Kinnebrew, J. S., Segedy, J. R., & Biswas, G. (2014). Analyzing the temporal evolution of students' behaviors in open-ended learning environments. *Metacognition and Learning*, *9*(2), 187–215. https://doi.org/10.1007/s11409-014-9112-4

Kirschner, F., Paas, F., & Kirschner, P. A. (2011). Task complexity as a driver for collaborative learning efficiency: The collective working-memory effect. *Applied Cognitive Psychology*, *25*(4), 615–624. https://doi.org/10.1002/acp.1730

Klein, J., Moon, Y., & Picard, R. (2002). This computer responds to user frustration: theory, design, and results. *Interacting with Computers*, *14*(2), 119–140. https://doi.org/10.1016/S0953-5438(01)00053-4

Kollar, I., Fischer, F., & Hesse, F. (2006). Collaboration scripts – a conceptual analysis. *Educational Psychology Review*, *18*. https://doi.org/10.1007/s10648-006-9007-2

Korb, S., With, S., Niedenthal, P., Kaiser Wehrle, S., & Grandjean, D. M. (2014). The perception and mimicry of facial movements predict judgments of smile authenticity [ID: unige:84135]. *PLOS ONE*, *9*(6), e99194.

Krishna, S., & Pelachaud, C. (2022). Impact of error-making peer agent behaviours in a multi-agent shared learning interaction for self-regulated learning, 337–344. https://doi.org/10.5220/0010881400003116

Kulíc, D., & Croft, E. (2007). Affective state estimation for human-robot interaction. *IEEE Transactions on Robotics*, *23*(5), 991–1000. https://doi.org/10.1109/TRO.2007.904899

Lämsä, J., Hämäläinen, R., Koskinen, P., Viiri, J., & Lampi, E. (2021). What do we do when we analyse the temporal aspects of computer-supported collaborative learning? A systematic literature review. *Educational Research Review*, *33*, 100387. https://doi.org/https://doi.org/10.1016/j.edurev.2021.100387

Lämsä, J., Hämäläinen, R., Koskinen, P., Viiri, J., & Mannonen, J. (2020). The potential of temporal analysis: combining log data and lag sequential analysis to investigate temporal differences between scaffolded and non-scaffolded group inquiry-based learning processes. *Computers & Education*, *143*, 103674. https://doi.org/https://doi.org/10.1016/j.compedu.2019.103674

Lavoué, É., Molinari, G., Prié, Y., & Khezami, S. (2015). Reflection-in-action markers for reflection-on-action in computer-supported collaborative learning settings. *Computers & Education*, *88*, 129–142.

Leite, I., Castellano, G., Pereira, A., Martinho, C., & Paiva, A. (2014). Empathic robots for long-term interaction. *International Journal of Social Robotics*, *6*(3), 329–341.

Leyzberg, D., Spaulding, S., & Scassellati, B. (2014). Personalizing robot tutors to individuals' learning differences. *ACM/IEEE International Conference on Human-Robot Interaction*, 423–430. https://doi.org/10.1145/2559636.2559671

Liu, R., Stamper, J. C., & Davenport, J. (2018). A Novel Method for the In-Depth Multimodal Analysis of Student Learning Trajectories in Intelligent Tutoring Systems. *Journal of Learning Analytics*, *5*(1), 41–54. https://doi.org/10.18608/jla.2018.51.4

Lodge, J. M., Kennedy, G., Lockyer, L., Arguel, A., & Pachman, M. (2018). Understanding Difficulties and Resulting Confusion in Learning: An Integrative Review. *Frontiers in Education*, *3*(June), 1–10. https://doi.org/10.3389/feduc.2018.00049

Loibl, K., Roll, I., & Rummel, N. (2017). Towards a Theory of When and How Problem Solving Followed by Instruction Supports Learning. *Educational Psychology Review*, *29*(4), 693–715. https://doi.org/10.1007/s10648-016-9379-x

Loibl, K., & Rummel, N. (2014). The impact of guidance during problem-solving prior to instruction on students' inventions and learning outcomes. *Instructional Science*, *42*(3), 305–326.

Lou, Y., Abrami, P. C., & d'Apollonia, S. (2001). Small group and individual learning with technology: a meta-analysis. *Review of educational research*, *71*(3), 449–521.

Malmberg, J., Haataja, E., Seppänen, T., & Järvelä, S. (2019). Are we together or not? the temporal interplay of monitoring, physiological arousal and physiological synchrony during a collaborative exam. *International Journal of Computer-Supported Collaborative Learning*, *14*(4), 467–490.

Malmberg, J., Järvelä, S., Holappa, J., Haataja, E., Huang, X., & Siipo, A. (2019). Going beyond what is visible: What multichannel data can reveal about interaction in the context of collaborative learning? *Computers in Human Behavior*, *96*(May 2018), 235–245. https://doi.org/10.1016/j.chb.2018.06.030

Malmberg, J., Järvelä, S., Järvenoja, H., & Panadero, E. (2015). Promoting socially shared regulation of learning in CSCL: Progress of socially shared regulation among high- and low-performing groups. *Computers in Human Behavior*, *52*, 562–572. https://doi.org/10.1016/j.chb.2015.03.082

Maqsood, R., Ceravolo, P., Romero, C., & Ventura, S. (2022). Modeling and predicting students' engagement behaviors using mixture markov models. *Knowledge and Information Systems*, *64*. https://doi.org/10.1007/s10115-022-01674-9

Maroni, B., Gnisci, A., & Pontecorvo, C. (2008). Turn-taking in classroom interactions: Overlapping, interruptions and pauses in primary school. *European Journal of Psychology of Education*, *23*(1), 59–76. https://doi.org/10.1007/BF03173140

Martinez, R., Wallace, J. R., Kay, J., & Yacef, K. (2011). Modelling and identifying collaborative situations in a collocated multi-display groupware setting. *International conference on artificial intelligence in education*, 196–204.

Martinez-Maldonado, R., Dimitriadis, Y., Martinez-Monés, A., Kay, J., & Yacef, K. (2013a). Capturing and analyzing verbal and physical collaborative learning interactions at an enriched interactive tabletop. *International Journal of Computer-Supported Collaborative Learning*, *8*(4), 455–485.

Martinez-Maldonado, R., Dimitriadis, Y., Martinez-Monés, A., Kay, J., & Yacef, K. (2013b). Capturing and analyzing verbal and physical collaborative learning interactions at an enriched interactive tabletop. *International Journal of Computer-Supported Collaborative Learning*, *8*(4), 455–485. https://doi.org/10.1007/s11412-013-9184-1

Maslow, A. (1943). A THEORY OF HUMAN MOTIVATION. (13), 370–396.

Meier, A., Spada, H., & Rummel, N. (2007). A rating scheme for assessing the quality of computer-supported collaboration processes. *International Journal of Computer-Supported Collaborative Learning*, *2*(1), 63–86.

Menon, D., Bp, S., Romero, M., & Viéville, T. (2019). Going beyond digital literacy to develop computational thinking in K-12 education. In L. Daniela (Ed.), *Smart Pedagogy of Digital Learning*. Taylor&Francis (Routledge).

Mentis, H. M. et al. (2007). Memory of frustrating experiences. *Information and Emotion: The Emergent Affective Paradigm in Information Behavior Research and Theory, Eds. Diane Nahl and Dania Bilal*, 197–210.

Nasir, J., Abderrahim, M., Kothiyal, A., & Dillenbourg, P. (2022). Temporal pathways to learning: how learning emerges in an open-ended collaborative activity. *Computers &; Education: Artificial Intelligence*. http://infoscience.epfl.ch/record/296043

Nasir, J., Bruno, B., Chetouani, M., & Dillenbourg, P. (2021). What if social robots look for productive engagement? *International Journal of Social Robotics*. https://doi.org/https://doi.org/10.1007/s12369-021-00766-w

Nasir, J., Bruno, B., Chetouani, M., & Dillenbourg, P. (2022). A Speech-based Productive Engagement Metric for Real-time Human-Robot Interaction in Collaborative Educational Contexts. *IEEE Transactions on Affective Computing*. http://infoscience.epfl.ch/record/294035

Nasir, J., Bruno, B., & Dillenbourg, P. (2020). Is there 'one way' of learning? a data-driven approach. *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 388–391. https://doi.org/10.1145/3395035.3425200

Nasir, J., Bruno, B., & Dillenbourg, P. (2021a). *PE-HRI-temporal: A Multimodal Temporal Dataset in a robot mediated Collaborative Educational Setting*. Zenodo. https://doi.org/10.5281/zenodo.5576058

Nasir, J., Bruno, B., & Dillenbourg, P. (2021b). *PE-HRI-temporal: A Multimodal Temporal Dataset in a robot mediated Collaborative Educational Setting*. Zenodo. https://doi.org/10.5281/zenodo.5576058

Nasir, J., Kothiyal, A., Bruno, B., & Dillenbourg, P. (2021). Many are the ways to learn identifying multi-modal behavioral profiles of collaborative learning in constructivist activities. *International Journal of Computer-Supported Collaborative Learning*. https://doi.org/https://doi.org/10.1007/s11412-021-09358-2

Nasir, J., Norman, U., Bruno, B., Chetouani, M., & Dillenbourg, P. (2021). *PE-HRI: A Multimodal Dataset for the study of Productive Engagement in a robot mediated Collaborative Educational Setting*. Zenodo. https://doi.org/10.5281/zenodo.4633092

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*. Zenodo. https://doi.org/10.5281/zenodo.4288833

Nasir, J., Norman, U., Bruno, B., & Dillenbourg, P. (2020a). When positive perception of the robot has no effect on learning. *29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*.

Nasir, J., Norman, U., Bruno, B., & Dillenbourg, P. (2020b). You tell, I do, and we swap until we connect all the gold mines! *ERCIM News*, *2020*(120).

Nasir, J., Norman, U., Johal, W., Olsen, J., Shahmoradi, S., & Dillenbourg, P. (2019). Robot analytics: what do human-robot interaction traces tell us about learning? *28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*.

Nezami, O. M., Hamey, L., Richards, D., & Dras, M. (2018). Engagement Recognition using Deep Learning and Facial Expression. *2013*.

Norman, U., Chin, A., Bruno, B., & Dillenbourg, P. (2022). Efficacy of a 'misconceiving' robot to improve computational thinking in a collaborative problem solving activity: a pilot study. *2022 31st IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, 8. http://infoscience.epfl.ch/record/294825

Norman, U., Dinkar, T., Bruno, B., & Clavel, C. (2021). Studying alignment in a collaborative learning activity via automatic methods: the link between what we say and do. https://doi.org/10.48550/ARXIV.2104.04429

Norman, U., Dinkar, T., Nasir, J., Bruno, B., Clavel, C., & Dillenbourg, P. (2021). *Justhink dialogue and actions corpus* (Version v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.4627104

Nwana, H. S. (1990). Intelligent tutoring systems: an overview. *Artificial Intelligence Review*, *4*(4), 251–277. https://doi.org/10.1007/BF00168958

O'Brien, H., Freund, L., & Kopak, R. (2016). Reading Environments. *In Proceedings of the 2016 ACM on conference on human information interaction and retrieval*, 71–80. https://doi.org/10.1145/2854946.2854973

O'Brien, H. L., & Toms, E. (2008). What is user engagement? a conceptual framework for defining user engagement with technology. *JASIST*, *59*, 938–955.

O'Brien, H. L., & Toms, E. (2010). The development and evaluation of a survey to measure user engagement. *JASIST*, *61*, 50–69.

Oertel, C., Castellano, G., Chetouani, M., Nasir, J., Obaid, M., Pelachaud, C., & Peters, C. (2020). Engagement in human-agent interaction : an overview. *Frontiers in Robotics and AI*, *7*, Article 92. https://doi.org/10.3389/frobt.2020.00092

Oertel, C., Scherer, S., & Campbell, N. (2011). On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation. *Twelfth Annual Conference of the International Speech Communication Association*.

Oggi, O. (, Rudovic,) Park, H. W., Busche, J., Schuller, B., Breazeal, C., & Picard, R. W. (2019). Personalized Estimation of Engagement from Videos Using Active Learning with Deep Reinforcement Learning.

Olsen, J. K., Sharma, K., Rummel, N., & Aleven, V. (2020a). Temporal analysis of multimodal data to predict collaborative learning outcomes. *British Journal of Educational Technology*, *51*(5), 1527–1547. https://doi.org/https://doi.org/10.1111/bjet.12982

Olsen, J. K., Sharma, K., Rummel, N., & Aleven, V. (2020b). Temporal analysis of multimodal data to predict collaborative learning outcomes. *British Journal of Educational Technology*, *51*(5), 1527–1547. https://doi.org/https://doi.org/10.1111/bjet.12982

Olusegun, S. (2015). Constructivism Learning Theory: A Paradigm for Teaching and Learning. *IOSR Journal of Research & Method in Education Ver. I*, *5*(6), 2320–7388. https://doi.org/10.9790/7388-05616670

Paans, C., Onan, E., Molenaar, I., Verhoeven, L., & Segers, E. (2019). How social challenges affect children's regulation and assignment quality in hypermedia: a process mining study. *Metacognition and Learning*, *14*(2), 189–213. https://doi.org/10.1007/s11409-019-09204-9

Papakostas, G., Sidiropoulos, G., Lytridis, C., Bazinas, C., Kaburlasos, V., Kourampa, E., Karageorgiou, E., Kechayas, P., & Papadopoulou, M. (2021). Estimating children engagement interacting with robots in special education using machine learning. *Mathematical Problems in Engineering*, *2021*, 1–10. https://doi.org/10.1155/2021/9955212

Pardos, Z. A., & Heffernan, N. T. (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*, 255–266. https://doi.org/10.1007/978-3-642-13470-8_24

Park, H. W., Grover, I., Spaulding, S., Gomez, L., & Breazeal, C. (2019). A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education. *AAAI*.

Parsons, J., & LeahTaylor. (2011). *Student Engagement: What do we know and what should we do?* (Tech. rep.). University of Alberta.

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. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peeters, M., Tuijl, H., Rutte, C., & Reymen, I. (2006). Personality and team performance: a meta-analysis. *European Journal of Personality*, *20*, 377–396. https://doi.org/10.1002/per.588

Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaïane, O. R. (2008). Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(6), 759–772.

Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaïane, O. R. (2009). Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(6), 759–772. https://doi.org/10.1109/TKDE.2008.138

Perugia, G., Boladeras, M., Català, A., Barakova, E., & Rauterberg, M. (2020). Engage-dem: a model of engagement of people with dementia. *IEEE Transactions on Affective Computing*, *PP*, 1–1. https://doi.org/10.1109/TAFFC.2020.2980275

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, 505–513.

Pijeira-díaz, H. J., Drachsler, H., Järvelä, S., & Kirschner, P. A. (2019). Sympathetic arousal commonalities and arousal contagion during collaborative learning : How attuned are triad members ? *Computers in Human Behavior*, *92*(May 2018), 188–197. https://doi.org/10.1016/j.chb.2018.11.008

Poggi, I. (2007). *Mind, hands, face and body: a goal and belief view of multimodal communication* [OCLC: ocn143609341]. Weidler.

Popov, V., van Leeuwen, A., & Buis, S. (2017). Are you with me or not? temporal synchronicity and transactivity during cscl. *Journal of Computer Assisted Learning*, *33*(5), 424–442.

Praharaj, S., Scheffel, M., Drachsler, H., & Specht, M. (2021). Literature Review on Co-Located Collaboration Modeling Using Multimodal Learning Analytics — Can We Go the Whole Nine Yards ? *IEEE Transactions on Learning Technologies*, *PP*(10), 1. https://doi.org/10.1109/TLT.2021.3097766

Ramachandran, A., Huang, C.-M., & Scassellati, B. (2017). Give me a break! personalized timing strategies to promote learning in robot-child tutoring. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 146–155. https://doi.org/10.1145/2909824.3020209

Ramachandran, A., Huang, C.-M., & Scassellati, B. (2019). Toward Effective Robot–Child Tutoring: Internal Motivation, Behavioral Intervention and Learning Outcomes. *ACM Transactions on Interactive Intelligent Systems*, *9*(1), 1–23. https://doi.org/10.1145/3213768

Ramachandran, A., Sebo, S. S., & Scassellati, B. (2019). Personalized Robot Tutoring using the Assistive Tutor POMDP (AT-POMDP). *Proceedings of The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, 1–8.

Reilly, J. M., & Schneider, B. (2019). Predicting the quality of collaborative problem solving through linguistic analysis of discourse. *EDM 2019 - Proceedings of the 12th International Conference on Educational Data Mining*, (Edm), 149–157.

Reimann, P. (2009). Time is precious: variable- and event-centred approaches to process analysis in cscl research. *I. J. Computer-Supported Collaborative Learning*, *4*, 239–257. https://doi.org/10.1007/s11412-009-9070-z

Rich, C., Ponsler, B., Holroyd, A., & Sidner, C. L. (2010). Recognizing engagement in human-robot interaction. *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 375–382. https://doi.org/10.1109/HRI.2010.5453163

Robert, L., Alahmad, R., Esterwood, C., Kim, S., You, S., & Zhang, Q. (2020). A review of personality in human–robot interactions. *Available at SSRN 3528496*.

Rodríguez, F. J., & Boyer, K. E. (2015). Discovering Individual and Collaborative Problem-Solving Modes with Hidden Markov Models. *Artificial Intelligence in Education: Proceedings of the World Conference on AI in Education 2015*, 408–418. https://doi.org/10.1007/978-3-319-19773-9

Rogat, T. K., & Linnenbrink-Garcia, L. (2011). Socially shared regulation in collaborative groups: an analysis of the interplay between quality of social regulation and group processes. *Cognition and Instruction*, *29*(4), 375–415.

Roschelle, J. (1992). Learning by Collaborating: Convergent Conceptual Change. *Journal of the Learning Sciences*, *2*(3), 235–276. https://doi.org/10.1207/s15327809jls0203-1

Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. *Computer supported collaborative learning*, 69–97.

Rossi, A., Raiano, M., & Rossi, S. (2021). Affective, cognitive and behavioural engagement detection for human-robot interaction in a bartending scenario. *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, 208–213. https://doi.org/10.1109/RO-MAN50785.2021.9515435

Rudovic, O., Zhang, M., Schuller, B., & Picard, R. (2019). Multi-modal active learning from human data: a deep reinforcement learning approach. *2019 International Conference on Multimodal Interaction*, 6–15.

Russell, J. (2003). Core Affect and the Psychological Construction of Emotion. *Psychological review*, *110*, 145–172. https://doi.org/10.1037//0033-295X.110.1.145

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, *55*(1), 68–78.

Sabourin, J., McQuiggan, S., & De Waal, A. (2016). SAS tools for educational data mining. *CEUR Workshop Proceedings*, *1633*(1), 85–106. https://doi.org/10.3102/1076998616666808

Salam, H., & Chetouani, M. (2015a). A multi-level context-based modeling of engagement in human-robot interaction. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, *03*, 1–6. https://doi.org/10.1109/FG.2015.7284845

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.

Salam, H., Çeliktutan, 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. https://doi.org/10.1109/ACCESS.2016.2614525

Salam, H., & Chetouani, M. (2015b). Engagement detection based on mutli-party cues for human robot interaction. *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, 341–347. https://doi.org/10.1109/ACII.2015.7344593

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., & Paiva, A. (2011a). Automatic analysis of affective postures and body motion to detect engagement with a game companion. *Proceedings of the 6th International Conference on Human-Robot Interaction*, 305–312. https://doi.org/10.1145/1957656.1957781

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., & Paiva, A. (2011b). Automatic analysis of affective postures and body motion to detect engagement with a game companion. *Proceedings of the 6th international conference on Human-robot interaction - HRI '11*, 305. https://doi.org/10.1145/1957656.1957781

Schneider, B., Dich, Y., & Radu, I. (2020). Unpacking the relationship between existing and new measures of physiological synchrony and collaborative learning: a mixed methods study. *International Journal of Computer-Supported Collaborative Learning*, *15*(1), 89–113.

Schneider, B., & Pea, R. (2013). Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *International Journal of Computer-supported collaborative learning*, *8*(4), 375–397.

Schneider, B., & Pea, R. (2015). Does seeing one another's gaze affect group dialogue? a computational approach. *Journal of Learning Analytics*, *2*(2), 107–133.

Schneider, B., Sharma, K., Cuendet, S., Zufferey, G., Dillenbourg, P., & Pea, R. (2018). Leveraging mobile eye-trackers to capture joint visual attention in co-located collaborative

learning groups. *International Journal of Computer-Supported Collaborative Learning*, *13*(3), 241–261.

Schneider, B., Sharma, K., Cuendet, S., Zufferey, G., Dillenbourg, P., & Pea, R. (2016). Using mobile eye-trackers to unpack the perceptual benefits of a tangible user interface for collaborative learning. *ACM Trans. Comput.-Hum. Interact.*, *23*(6). https://doi.org/10.1145/3012009

Schulte, P. L. (1996). A definition of constructivism. *Science Scope*, *20*(3), 25–27.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and instruction*, *16*(4), 475–5223.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: the hidden efficiency of encouraging original student production in statistics instruction. *Cognition and instruction*, *22*(2), 129–184.

Schwarz, B. B., Neuman, Y., & Biezuner, S. (2000). Two wrongs may make a right ... if they argue together! *Cognition and Instruction*, *18*(4), 461–494. https://doi.org/10.1207/S1532690XCI1804_2

Seabold, S., & Perktold, J. (2010). Statsmodels: econometric and statistical modeling with python. *9th Python in Science Conference*.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in cognitive sciences*, *10*(2), 70–76.

Sharma, K., Caballero, D., Verma, H., Jermann, P., & Dillenbourg, P. (2015). Looking at versus looking through: a dual eye-tracking study in mooc context. *Proceedings of 11th International Conference of Computer Supported Collaborative Learning*, *1*(CONF), 260–267.

Sharma, K., & Giannakos, M. (2020). Multimodal data capabilities for learning: What can multimodal data tell us about learning? *British Journal of Educational Technology*, *51*(5), 1450–1484. https://doi.org/https://doi.org/10.1111/bjet.12993

Sharma, K., Olsen, J. K., Aleven, V., & Rummel, N. (2021). Measuring causality between collaborative and individual gaze metrics for collaborative problem-solving with intelligent tutoring systems. *Journal of Computer Assisted Learning*, *37*(1), 51–68.

Sharma, K., Papamitsiou, Z., Olsen, J. K., & Giannakos, M. (2020). Predicting learners' effortful behaviour in adaptive assessment using multimodal data. *ACM International Conference Proceeding Series*, 480–489. https://doi.org/10.1145/3375462.3375498

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., & Rich, C. (2005). Explorations in engagement for humans and robots. *Artificial Intelligence*, *166*(1-2), 140–164. https://doi.org/10.1016/j.artint.2005.03.005

Siegler, R. S., & Crowley, K. (1991). The Microgenetic Method : A Direct Means for Studying Cognitive Development. *American Psychologist*, *46*(June), 606–620. https://doi.org/10.1037/0003-066X.46.6.606

Siemens, G., & Baker, R. S. J. d. (2012). Learning analytics and educational data mining, 252. https://doi.org/10.1145/2330601.2330661

Sinha, S., Rogat, T. K., Adams-Wiggins, K. R., & Hmelo-Silver, C. E. (2015). Collaborative group engagement in a computer-supported inquiry learning environment. *International Journal of Computer-Supported Collaborative Learning, 10*(3), 273–307.

Sinha, T. (2021). Enriching problem-solving followed by instruction with explanatory accounts of emotions. *Journal of the Learning Sciences*, 1–48.

Sinha, T., & Kapur, M. (2021). *When Problem Solving Followed by Instruction Works: Evidence for Productive Failure* (Vol. 20). https://doi.org/10.3102/00346543211019105

Sobocinski, M., Malmberg, J., & Järvelä, S. (2017). Exploring temporal sequences of regulatory phases and associated interactions in low- and high-challenge collaborative learning sessions. *Metacognition and Learning, 12*(2), 275–294. https://doi.org/10.1007/s11409-016-9167-5

Speranza, S., Recchiuto, C. T., Bruno, B., & Sgorbissa, A. (2020). A model for the representation of the extraversion-introversion personality traits in the communication style of a social robot. *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 75–81. https://doi.org/10.1109/RO-MAN47096.2020.9223537

Spikol, D., Ruffaldi, E., & Cukurova, M. (2017). Using multimodal learning analytics to identify aspects of collaboration in project-based learning. *Computer-Supported Collaborative Learning Conference, CSCL, 1*(June), 263–270. https://doi.org/10.22318/cscl2017.37

Spikol, D., Ruffaldi, E., Dabisias, G., & Cukurova, M. (2018). Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. *Journal of Computer Assisted Learning, 34*(4), 366–377.

Staffa, M., Rossi, A., Bucci, B., Russo, D., & Rossi, S. (2021). Shall I Be Like You? Investigating Robot's Personalities and Occupational Roles for Personalised HRI. In H. Li, S. S. Ge, Y. Wu, A. Wykowska, H. He, X. Liu, D. Li, & J. Perez-Osorio (Eds.), *Social robotics* (pp. 718–728). Springer International Publishing.

Stahl, G., Law, N., & Hesse, F. (2013). Reigniting CSCL flash themes. *International Journal of Computer-Supported Collaborative Learning, 8*(4), 369–374. https://doi.org/10.1007/s11412-013-9185-0

Stower, R., & Kappas, A. (2021). Cozmonaots: designing an autonomous learning task with social and educational robots. *Interaction Design and Children*, 542–546. https://doi.org/10.1145/3459990.3465210

Sweller, J. (2011). Chapter two - cognitive load theory. Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Szafir, D., & Mutlu, B. (2012). Pay Attention! Designing Adaptive Agents that Monitor and Improve User Engagement. *Conference on Human Factors in Computing Systems (CHI)*. https://doi.org/10.1145/2207676.2207679

Tatarian, K., Wallkötter, S., Buyukgoz, S., Stower, R., & Chetouani, M. (2020). Mobiaxis: an embodied learning task for teaching multiplication with a social robot. *ArXiv*. https://doi.org/10.48550/arXiv.2004.07806

Tavenard, R., Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., Payne, M., Yurchak, R., Rußwurm, M., Kolar, K., & Woods, E. (2020). Tslearn, a machine learning toolkit for

time series data. *Journal of Machine Learning Research*, *21*(118), 1–6. http://jmlr.org/papers/v21/20-091.html

Teasley, S. D. (1997). Talking about reasoning: how important is the peer in peer collaboration? *Discourse, tools and reasoning* (pp. 361–384). Springer.

Tulli, S., Couto, M., Vasco, M., Yadollahi, E., Melo, F., & Paiva, A. (2020). Explainable Agency by Revealing Suboptimality in Child-Robot Learning Scenarios. In A. R. Wagner, D. Feil-Seifer, K. S. Haring, S. Rossi, T. Williams, H. He, & S. Sam Ge (Eds.), *Social robotics* (pp. 23–35). Springer International Publishing.

van de Sande, B. (2013). Properties of the bayesian knowledge tracing model. *EDM 2013*.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction*, *21*(3), 209–249.

Veenman, M. V. J. (2013). Assessing metacognitive skills in computerized learning environments. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 157–168). Springer New York. https://doi.org/10.1007/978-1-4419-5546-3-11

Vikashini, V., Salam, H., Nasir, J., Bruno, B., & Celiktutan, O. (2022). Personalized productive engagement recognition in robot-mediated collaborative learning. *24th ACM International Conference on Multimodal Interaction*. http://infoscience.epfl.ch/record/296044

Viswanathan, S. A., & VanLehn, K. (2017). Using the tablet gestures and speech of pairs of students to classify their collaboration. *IEEE Transactions on Learning Technologies*, *11*(2), 230–242.

Viswanathan, S. A., & Vanlehn, K. (2018). Using the Tablet Gestures and Speech of Pairs of Students to Classify Their Collaboration. *IEEE Transactions on Learning Technologies*, *11*(2), 230–242. https://doi.org/10.1109/TLT.2017.2704099

Vogel, F., Wecker, C., Kollar, I., & Fischer, F. (2017). Socio-cognitive scaffolding with computer-supported collaboration scripts: a meta-analysis. *Educational Psychology Review*, *29*. https://doi.org/10.1007/s10648-016-9361-7

Voutsina, C., George, L., & Jones, K. (2019). Microgenetic analysis of young children's shifts of attention in arithmetic tasks: underlying dynamics of change in phases of seemingly stable task performance. *Educational Studies in Mathematics*, *102*(1), 47–74. https://doi.org/10.1007/s10649-019-09883-w

Vrzakova, H., Amon, M. J., Stewart, A., Duran, N. D., & D'Mello, S. K. (2020). Focused or stuck together: Multimodal patterns reveal triads' performance in collaborative problem solving. *ACM International Conference Proceeding Series: Learning Analytics and Knowledge, 2020*, 295–304. https://doi.org/10.1145/3375462.3375467

Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., & André, E. (2013). The social signal interpretation (ssi) framework: multimodal signal processing and recognition in real-time. *Proceedings of the 21st ACM International Conference on Multimedia*, 831–834. https://doi.org/10.1145/2502081.2502223

Wang, C., Sahebi, S., Zhao, S., Brusilovsky, P., & Moraes, L. O. (2021). Knowledge tracing for complex problem solving: granular rank-based tensor factorization. *Proceedings of the 29th acm conference on user modeling, adaptation and personalization* (pp. 179–188). Association for Computing Machinery. https://doi.org/10.1145/3450613.3456831

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & education*, *46*(1), 71–95.

Whitehill, J., Serpell, Z., Lin, Y. C., Foster, A., & Movellan, J. R. (2014). The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*, *5*(1), 86–98. https://doi.org/10.1109/TAFFC.2014.2316163

Winkle, K., Senft, E., & Lemaignan, S. (2021). LEADOR: A Method for End-To-End Participatory Design of Autonomous Social Robots. *Frontiers in Robotics and AI*, *8*. https://doi.org/10.3389/frobt.2021.704119

Wolters, C. A., Yu, S. L., & Pintrich, P. R. (1996). The relation between goal orientation and students' motivational beliefs and self-regulated learning. *Learning and Individual Differences*, *8*(3), 211–238. https://doi.org/10.1016/S1041-6080(96)90015-1

Worsley, M., & Blikstein, P. (2011). What's an expert? using learning analytics to identify emergent markers of expertise through automated speech, sentiment and sketch analysis. *EDM*, 235–240.

Worsley, M., & Blikstein, P. (2018). A Multimodal Analysis of Making. *International Journal of Artificial Intelligence in Education*, *28*, 385–419.

Yadollahi, E., Johal, W., Paiva, A., & Dillenbourg, P. (2018). When deictic gestures in a robot can harm child-robot collaboration. *IDC '18*, 195–206. https://doi.org/10.1145/3202185.3202743

Yang, C. W., Cukurova, M., & Porayska-Pomsta, K. (2021). Dyadic joint visual attention interaction in face-to-face collaborative problem-solving at K-12 Maths Education: A Multimodal Approach. *CEUR Workshop Proceedings*, *2902*.

Jauwairia Nasir

Professional Goals

I aspire to create a social difference with my research on leveraging AI and Human-Robot Interaction in domains such as Education and Healthcare. I have 7+ years of research experience.

Email:
jauwairianasir@gmail.com

Languages Spoken:
English, Urdu, Punjabi, Korean (B2), French (A1/A2)

Nationality:
USA/Pakistani

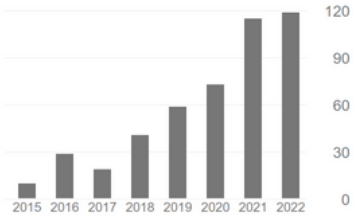
Technical Skills:
Python, ROS, Pandas, Jupyter notebook, Latex, Git, sklearn, keras

Transversal Skills:
Leadership, Public Speaking, Determination, Adaptability, Collaboration, Communication.

Jauwairia Nasir

Cited by VIEW ALL

	All	Since 2017
Citations	487	427
h-index	8	8
i10-index	8	7



Education

École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
PhD in Robotics Control and Intelligent Systems | EU Horizon Marie Curie ITN Fellow at ANIMATAS | September 2018 - October 2022
- Supervisor: Prof. Pierre Dillenbourg, co-supervisor: Dr. Barbara Bruno
- Led or co-led user studies with ~360 participants

Sorbonne University, Paris, France
Visiting Researcher at ISIR | October - December, 2019
- Worked with: Prof. Mohamed Chetouani

Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea
Masters in Electrical Engineering and Computer Science | September 2014 - August 2016
- Supervisor: Prof. Jong-Hwan Kim

NUST School of Electrical Engineering and Computer Sciences (NUST-SEECS), Islamabad, Pakistan
BSc. in Electrical Engineering | October 2008 - August 2012
- Final Year project supervisor: Prof. Yasar Ayaz

Professional Experience

Consultant
Four | April 2017 - November 2017
A startup (fourgroup.co), at the intersection of AI and healthcare, that focuses on developing sustainable healthcare systems. I assumed several roles including technical and design feedback, qualitative interviews with the targeted medical audience, representing the company at events, meetings with potential future employees.

Research Scholar
KAIST | September 2016 - February 2017
Worked on cognitive architectures for service robots for task planning, and motion planning for autonomous agents

Publications

Journal Publications

- J. Nasir, B. Bruno, M. Chetouani, and P. Dillenbourg. *A Speech-based Productive Engagement Metric for Real-time Human-Robot Interaction in Collaborative Educational Contexts*. under revision in IEEE Transactions on Affective Computing.
- J. Nasir, M. Abderrahim, A. Kothiyal, and P. Dillenbourg. *Temporal Pathways to Learning: How Learning Emerges in an Open-ended Collaborative Activity*, in Computers & Education: Artificial Intelligence, 2022.
- J. Nasir, A. Kothiyal, B. Bruno, and P. Dillenbourg, *Many Are The Ways to Learn: identifying multi-modal behavioral profiles of collaborative learning in constructivist activities.* in International Journal of Computer-Supported Collaborative Learning (IJCSCL) 2021.
- J. Nasir, B. Bruno, M. Chetouani, and P. Dillenbourg, *What if Social Robots Look for Productive Engagement?.* in Int Journal of Social Robotics (2021).
- C. Oertel, G. Castellano, M. Chetouani, J. Nasir, M. Obaid, C. Pelachaud, and C. E. Peters, *Engagement in Human-Agent Interaction: An Overview*, in Frontiers in Robotics and AI (2020).
- P. Dillenbourg, K. Kim, J. Nasir, T. Yeo, J. Olsen, *Applying IDC theory to education in the Alps region: a response to Chan et al.'s contribution.* Research and Practice in Technology Enhanced Learning, 2019.
- J. Nasir, Yong-Ho. ; Kim D.-H.; Kim Jong-Hwan, *User Preference-based Dual-Memory Neural Model with Memory Consolidation Approach*, in IEEE Transactions on Neural Networks and Learning Systems (2017).
- J. Nasir,. ; Kim D.-H.; Kim Jong-Hwan, *ART neural network-based integration of episodic memory and semantic memory for task planning for robots*. Autonomous Robots, 2019.
- J. Nasir, ; Islam, F.; Malik, U. ; Ayaz, Y. ; Hasan, O. , *RRT*-Smart: Rapid Convergence Implementation of RRT**, in Int. J. Adv. Robot. Syst, Jun. 2013.
- J. Nasir,. ; Islam, F.; Ayaz, Y. , *Adaptive rapidly-exploring-random-tree-star (RRT*)-smart: algorithm characteristics and behavior analysis in complex environments.* Asia-Pacific Journal of Information Technology and Multimedia, 2013.

Conference Publications

- J. Nasir, P. Oppliger, B. Bruno, and P. Dillenbourg, *Questioning Wizard of Oz: Effects of Revealing the Wizard behind the Robot.* in 31st IEEE International Conference on Robot & Human Interactive Communication (RO-MAN), 2022.
- H. Salam, V. Vikashini, J. Nasir, O. Celiktutan, and B. Bruno. *Personalized Productive Engagement Recognition in Robot-Mediated Collaborative Learning*. in the 24th ACM International Conference on Multimodal Interaction (ICMI), 2022
- J. Nasir*, U. Norman*, B. Bruno, and P. Dillenbourg, *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), Naples, Italy, 2020.
- J. Nasir, B. Bruno, and P. Dillenbourg. *Is There `ONE way' of Learning? A data-driven approach*. Companion publication of the 2020 International Conference on Multimodal Interaction (ICMI) 2020.
- J. Nasir, U. Norman, B. Bruno, and P. Dillenbourg,*You tell, I do, and we swap until we connect all the gold mines!.* Educational Technology, 2020.
- J. Nasir*, U. Norman*, W. Johal, J. K. Olsen, S. Shahmoradi and P. Dillenbourg, *Robot Analytics: What Do Human-Robot Interaction Traces Tell Us About Learning?,* 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), New Delhi, India, 2019, pp. 1-7
- L. Hamamsy, W. Johal, T. Asselborn, J. Nasir, P. Dillenbourg. *Learning by collaborative teaching: an engaging multi-party cowriter activity.* 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), New Delhi, India, 2019,
- S. Shahmoradi, J. Olsen, S. Haklev, W. Johal, U. Norman, J. Nasir, P. Dillenbourg. *Orchestration of robotic activities in classrooms: challenges and opportunities*. European Conference on Technology Enhanced Learning, 2019.
- Islam, F.; Nasir, J. ; Malik, U. ; Ayaz, Y. ; Hasan, O. , *RRT*-Smart: Rapid Convergence Implementation of RRT* Towards Optimal Solution*", in proceedings of International Conference of Mechatronics and Automation, China 2012.

*equal contribution
Note: First author is the main contributor for the research

Open Source Datasets

- Nasir, J., Norman, U., Bruno, B., Chetouani, M., & Dillenbourg, P. (2021). PE-HRI: A Multimodal Dataset for the study of Productive Engagement in a robot mediated Collaborative Educational Setting [Data set]. Zenodo.
- Nasir, J., Bruno, B., & Dillenbourg, P. (2021). PE-HRI-temporal: A Multimodal Temporal Dataset in a robot mediated Collaborative Educational Setting [Data set]. Zenodo.
- Norman, U., Dinkar, T., Nasir, J., Bruno, B., Clavel, C.,, & Dillenbourg, P.. (2021). JUSThink Dialogue and Actions Corpus (v1.0.0) [Data set]. Zenodo.

Teaching Experience

Guest Lecturer
Machine Learning for Behavioral Data | EPFL | April 2022

Teaching Assistant
École Polytechnique Fédérale de Lausanne (EPFL) | 2019 - 2022

- Introduction to Visual Computing (Spring 2019, Spring 2020)
- Robotics Practicals (Spring 2021, Spring 2022)

Advising

Supervised Bachelors/Masters Semester Projects and Research Fellows
École Polytechnique Fédérale de Lausanne (EPFL) | 2019 - 2022

- Mortadha Abderrahim, Research Scholar for 1 year
- Laura Mathex on 'Does a Conversational Robot's Gaze Behavior Affect the Robots Perception and How Users Distance Themselves from the Robot Psychologically?'
- Anne Donnet on 'In a competitive setting, is an explicit robot better than an implicit robot?'
- Mortadha Abderrahim on 'Trends in multi-modal behavioral state transitions for learners in a robot mediated human-human collaborative activity'
- EL Mekki Malek on 'Sequence mining to analyse learner's action patterns in a robot mediated human-human collaborative activity'
- Leandro Graziano on 'Developing a Natural Interaction Robot Behavior Module for QTrobot'
- Haoyu Sheng on 'To speak or not to speak when doing task actions collaboratively – does it matter?'
- Pierre Oppliger on 'Master of Puppets - Social Robotics version'
- William Ouensanga on 'Robots with Personality!'

Honors and Awards

- Nominated and Finalist for '***Hidden Figures Award 2020***' by TechFace, Switzerland
- Selected for EU Horizon ***Marie Curie ITN Fellowship*** (2018-2022)
- Won ***Korean Government Scholarship*** (KGSP) 2013-2016
- Won Erasmus Mundus Scholarship
- Won the Pakistan National Engineering Robotic Competition 2011.
- Nominated for ***Rectors Gold Medal Award*** for Best Final Year Project.
- Received three NUST SEECS Student Appreciation Awards in 2012.
- Finished ***Second*** in my Batch of around 200 Electronics students
- Received GPA Based Scholarship in all Semesters of Bachelors

Invited Talks

In Pursuit of Goal-centric Social Robots in Educational HRI
AFAR Lab, University of Cambrdige | March 2022

What if Social Robots Look for Productive Engagement?
Talking Robotics | June 2021

In Pursuit of Goal-centric Social Robots in Educational HRI
NLP Zurich | June 2021

Interview Contribution to a white paper on Advancing Women in AI
ImpactIA and Women in Digital Switzerland 2021

Can AI Help in Understanding Cognitive and Motor Skills of Learners?
WeTechTogether Conference 2020

Leveraging AI to Understand Cognitive and Motor Skills of Learners
Global WAI Summit 2020

Research Collaborations

Mohamed Chetouani | Sorbonne University, October 2019 - ongoing
Justine Cassell | CMU/INRIA, March 2022 - July 2022
Hanan Salam | NYU Abu Dhabi, Nov 2021 - May 2022
Oya Celiktutan | King's College London, Nov 2021 - May 2022
Catharine Oertal | TU Delft, 2019 - Aug 2020
Ginevra Castellano | Uppsala University, Feb 2020 - Aug 2020
Christopher Peters | KTH Royal Institute of Technology, Feb 2020 - Aug 2020
Catherine Pelachaud | Sorbonne University, Feb 2020 - Aug 2020
Mohammad Obaid | Chalmers University of Technology, Feb 2020 - Aug 2020

Academic Service

- PC member for *ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2023*, Sweden.
- Co-organized a global Syposium in Schwarzsee Switzerland on '*Human-Machine Interaction: Perception, Social Learning, Personalised Adaptation in Educational Settings*', in an hybrid format gathering ~80 participants in total from 12-14 October 2021.
- Co-organized winter school on '*Developing Technologies in Educational Settings*' at EPFL, 5-8 February, 2019.
- Reviewer for several conferences and journals such as HRI 2019, 2020, 2021, ROMAN 2022, IEEE RA-L, ICRA 2021, IEEE Transactions on Neural Network and Learning Systems 2017, International Journal of Child-Computer Interaction 2020, 2021

Leadership & Equal Opportunities Work

Education & Research Ambassador Switzerland
Women in AI | May 2019 - Present
It is a community-driven global initiative where we organize events & articles for promoting unbiased AI that benefits global society.

- Lead the organization of WAICamp 2020: *A Day of Exchange and Discovery for Girls* (cancelled due to covid)
- Co-organized the flagship event WAITalk on *The Dark side of AI* 2020 targeting AI professionals,
- Co-organized a workshop/hackathon on *AI for Equality* at WeTechTogether Conference, 2021
- Participated in writing an article on *AI landscape in Switzerland*
- Interview Contribution to a white paper on *Advancing Women in AI* by ImpactIA and Women in Digital Switzerland
- Invited for several talks

Co-founder and President (Alumni)
International Student and Scholar Academic Council (ISSAC) | KAIST, Republic of Korea
A platform for connecting, networking and generating ideas for the international students at KAIST with its alumni, academia and industry.

- Organized several events including *industry trips to Hyundai Motor Company*
- Workshops on Python, Java and web development
- Invited as founders to represent International students in Korea at the *International Student Symposium in Turkey*, 2016

Personal Roles and Interests

- *Happiness Manager at CHILI Lab, EPFL*, 2019-2022
- Travel photography and blogging
- Nature
- Painting and other artwork