Unplugged Activities on Machine Learning and Their Evaluation Through Mental States Attribution

Matteo Baldoni[1], Cristina Baroglio[1], Monica Bucciarelli[2,3], Sara Capecchi[1, 4], Leonardo Castellani[1, 2], Elena Gandolfi[2, 5], Francesco Ianì[2, 3], Elisa Marengo[1], Roberto Micalizio[1]

[1]Università di Torino, Dipartimento di Informatica
[2]Università di Torino, Dipartimento di Psicologia
[3]Center for Logic, Language and Cognition
[4]Laboratorio Informatica e Scuola CINI
[5]Universitas Mercatorum, Dipartimento di Scienze Umane e Sociali
matteo.baldoni@unito.it, cristina.baroglio@unito.it, monica.bucciarelli@unito.it, sara.capecchi@unito.it, l.castellani@unito.it, elena.gandolfi@unimercatorum.it, francesco.iani@unito.it, elisa.marengo@unito.it, roberto.micalizio@unito.it

Abstract

Theory of mind refers to the attribution of mental states that humans ascribe to other humans or objects (such as computer-based systems). Recently, the attribution of mental states has been investigated toward Artificial Intelligence (AI) as a basic manner to capture people's engagement toward it, and people's perception about AI social skills and AI capabilities. In line with this idea, mental state attribution can be used as an indirect measure of students' understanding of AI functioning, and in particular of the kind of interactions students may have with AI systems. Too often is the case of people using generative AI systems in ways that exceed their actual ways of functioning. In our study, children of age in the range 9-12 were involved in one-shot unplugged activities concerning data and models. The unplugged activities were not aimed at teaching the theory of Machine Learning, but rather they were designed so as to provide awareness on some basic mechanisms and help developing a correct use of tools that are becoming more and more present in everyday life. This paper introduces the activities and reports the results that were achieved.

1 Introduction

In recent years, Artificial Intelligence (AI) and Machine Learning (ML) have become ubiquitous in everyday life. The adoption and the exposure is so wide and fast evolving that it is difficult to predict the level of understanding that people have of AI systems, as well as the awareness about the mechanisms, opportunities, and threats of ML, especially among children (UNICEF 2019). It is, thus, useful to develop popular-scientific activities, that allow people to acquire a correct perception of how ML tools work without explaining ML theory, which many persons, because of age or of background, would not have the means to understand. The reason is that these same persons are exposed to generative AI and other systems and, unless endowed of the right conceptual tools, they would be exposed to risks (e.g., thinking the system is an infallible oracle, using the system as a psychologist or as an advisor, developing feelings for the system). For a wide diffusion, the activities should be simple, fast, easily deployable in various contexts (e.g., exhibitions, public engagement events, schools, fairs), and the focus should be on evaluating the (change of) perception people have towards ML systems rather than on the competence to use ML tools, which is not a target.

In this work, we describe a set of unplugged activities, that we developed in order to provide awareness on some basic mechanisms of ML (related to classification tasks), and help developing a correct use of tools that are becoming more and more present in everyday life. We also report results on their evaluation, that was carried out by verifying the mental states that children, of age in the range 9-12, ascribe to ML systems before and after the activities. This evaluation builds upon the literature concerning the *theory of mind*, that we briefly introduce hereafter, and in particular on works like (Frischknecht 2021), which explain that anthropomorphizing and ascribing agency to non-human entities is a common strategy that humans use to make sense of the environment, and like (Shank and DeSanti 2018) that observes an increase of mental states attribution when people have more information of how the used algorithm works, interpreting it as a perception of the system of being capable of planning and having intentions.

The term *theory of mind* refers to the ability of a person to infer other persons' mental states and emotions (Brüne and Brüne-Cohrs 2006). The attribution of mental states that one makes toward another person is at the basis of the interaction the person establishes, or is willing to establish. Therefore, it can be used to predict one's behaviors and social interaction attitudes. Similar interest to human-human mental state attribution has been directed toward human-computer mental state attribution, to predict and explain the interactions people establish with computer-based systems, and to potentially understand the level of confidence, trust, and such like people place in them (e.g., (McEneaney 2009)).

More recently, scholars started investigating people's mental state attribution to AI systems, as a basic manner to capture people perception of AI social skills, AI capa-

bilities, and people's engagement in using these technologies (Frischknecht 2021). Their findings are in line with other studies in the literature (Shank and DeSanti 2018; Thellman, de Graaf, and Ziemke 2022; de Graaf and Malle 2019; Epley, Waytz, and Cacioppo 2007; Levin et al. 2013; Marchesi et al. 2019; Thellman 2021), showing that the attribution of more mental states to an artificial device (e.g., a robot) correlates to a deeper understanding of how it functions, which, in turns, allows people to predict and explain the behavior of the device. Frischknecht (2021) investigates the relation between autonomy and mental states attribution: the more autonomous a system is, the more *agency* (i.e., the capability to act, make decisions, and achieve goals) is ascribed to it by people. This is well-known also in AI, and in particular in the Multiagent Systems literature (Wooldridge and Jennings 1995, Section 2.1).

In our work we measure children's mental state attribution to an AI-based software before and after an educational intervention. The measurement exploits the AMS scale (Manzi et al. 2020). The educational intervention, of the duration of about one hour, consists of a set of activities, and targets kids of age 9-12. The activities do not require the use of a computer or other electronic device (they are *unplugged)*, and they are structured as small tasks that children are asked to solve and whose solutions provide the starting points for explanations and discussions. A detailed description of the intervention is included in this paper. The expectation is that after having more information on how ML functions, children ascribe more agency to the AI and less of experience (i.e., capabilities to feel). This result would be along the line of (Shank and DeSanti 2018; Thellman, de Graaf, and Ziemke 2022) which, however, did not involve an educational intervention.

In the rest of the paper, we first present the design of the experiment and the involved participants. Then, we describe the activities we administered in roughly one hour of intervention. Finally, we comment the collected results and relate the work with the literature.

2 Method

For our study, we measured mental states attribution by way of a subset of the standard AMS scale (Manzi et al. 2020) in a pre- and post-test approach so to assess the effects of our intervention. In this section, we describe in detail our procedure, the involved participants, and the educational material we designed.

Procedure and Measures

The team of researchers, composed of computer scientists and psychologists, instructed a moderator to lead the educational intervention that develops as follows. As a first step, children are shown a video, extracted from the *AI for Ocean* activity (code.org - AI for oceans), offered by the code.org platform (code.org). The video shows an AI-empowered robot that successfully cleans the ocean of garbage (see supplementary material). The AI is able to distinguish without making any mistake between various types of fish, which are left swimming, and garbage items, which are, instead, thrown away. After the video, the moderator proposes the children to teach a new robot to do the same.

As a second step, children are asked to fill a scale, consisting of a subset of questions from the AMS scale (Manzi et al. 2020). We opted for this questionnaire since it is well known in the literature (Thellman, de Graaf, and Ziemke 2022) and children fill it quickly, which avoids them getting bored and thus distracted. At the top of the sheet of the AMS questionnaire, we reported the picture of the robot children have just seen in the video. The form contains 26 questions. For example, one question is "Do you think that the AI can understand?". For each question, the children can choose among the possible answers "a lot" (2 points), "a little" (1 point) or "not at all" (0 points). The total score is the sum of all responses (ranging from zero to 52). The lower the result, the fewer mental states are attributed to the AI system.

The moderator, then, introduces the unplugged activities by explaining to the children that, before teaching the ocean robot, the children themselves need to be instructed on how a robot learns. The moderator presents one task at a time, giving children enough time to solve it in autonomy. Then, the moderator collects some answers from the children, and comments on them, explaining why they were correct or not. This interaction gives us the possibility to explain the concepts, but also to grasp what are the aspects that children have more difficulties to understand, and also whether the tasks we are proposing them are well formulated and not too easy or too difficult for children to solve.

Once all tasks are completed, the moderator asks children to fill the AMS questionnaire again. Since the AMS was also administered at the very beginning, we are able to capture whether our intervention changes the children's perspectives on ML and AI.

The last step consists in asking the children to cooperate for training the robot they have seen at work in the initial video. The AI for Ocean exercise from code.org is then performed collectively, allowing children to practice what they have learned. Children discuss possible justifications for those cases in which the robot classifies wrongly fishes or garbage, based on the examples the AI was trained on (e.g., a fish was misclassified because of its color, similar to many garbage items saw in the training phase, or a piece of garbage is recognized as a fish because of its shape similar to many fishes). Children also decide when to stop showing the robot examples of fish and of garbage.

Participants

We experimented our set-up at two public events: a book fair, and a scientific fair for schools, both held in Italy. The intervention took about one hour. The groups of children (all volunteers) were heterogeneous in gender, provenance, and age, ranging from 9 to 12. The heterogeneity of the groups makes the analysis independent from children background and contexts.

Overall, we collected more than 120 AMS forms, but only 104 were completely filled both at the pre- and at the post-test. So, only these 104 forms were used in the analysis. In fact, due to the extra-curricular nature of the events, children were free to fill their questionnaires just partially.

Before administering the activities to children, these were approved by our University ethical committee. Moreover, children could at any time reject the consent to process the results of their questionnaires.

Material

When designing the educational intervention, we had to face some challenges. We wanted the intervention to be of short duration, not to tire children too much and thus loose their attention. At the same time, however, we wanted to cover a number of basic concepts that are recurrent in ML problems, so that the tasks can be sufficiently general. To keep children engaged, we opted for an approach that required children to be active. Therefore, first we propose children a task to solve and then we provide explanations starting from children answers. Given the format that we chose, i.e., asking children to solve tasks before giving explanations, we also had to carefully select the proper abstraction level so that children can perform the tasks even without any previous knowledge about ML without getting frustrated.

Concerning the topics addressed in the unplugged tasks, we focused on *classification*, which is one of the major ML areas, and shares issues concerning the construction of datasets also with other ML areas, like regression and unsupervised learning, but is more intuitive to children, since classification is an activity every person performs, usually inadvertently, plenty of times every day.

To increase the level of engagement, we devised unplugged playful activities based on images, in most cases made up with fictional characters. This allows us to simplify the images so as to include only the elements that are needed for the explanation that follows the task. Note that the use of images is motivated by our goal of devising simple classification tasks, where features are immediately recognized by children, and have a finite, discrete set of possible values. It is not our goal, here, to cope with image processing ML algorithms (e.g. CNNs). Rather, we aim at a more abstract and simpler language that children can naturally understand.

In the next section we describe the tasks in detail.

3 The Educational Intervention

Concerning *classification*, we focused on *datasets* and *models*, and cover issues that are typical of ML tasks. Specifically, we built upon the principle that ML can only happen when a set of examples (the *dataset*) is available, but it is necessary to avoid *bias* by selecting representative examples of the domain under consideration. Moreover, the learning activity generates a *model* that can be used to make predictions, but models may also be imprecise and give unexpected results (the *accuracy* problem). Finally, machine learning often relies on estimating the distance between two or more examples; such a distance depends on the example features, but not all the features may have the same importance (the problem of *dimensionality*).

As said, the tasks are *unplugged*, so they do not require any specific device. Moreover, they do not require any previous knowledge about ML and AI, not even about coding or programming. The choice of not relying on software or hardware tools is supported by the fact that, as noted in some recent works (Baldoni et al. 2024a, 2022; Tedre et al. 2021), the acquisition of competencies on how to use a tool does not necessarily correspond to the acquisition of awareness and understanding of how its underlying complex mechanisms work. We opted to focus on the methodologies underlying ML in an abstract, general way. Dealing with unplugged activities also simplifies the way the intervention can be administered and is less distracting for children.

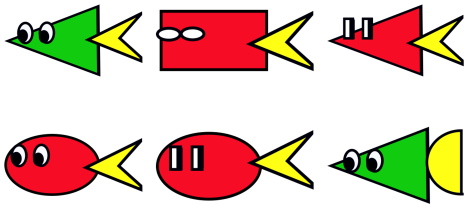

Figure 1: Excerpt of the fictional characters used in the classification tasks.

Since the target of our study are children whose age ranges from 9 to 12 years, we had to find a way to make these concepts both accessible (in terms of abstraction level and language) and engaging. As a consequence, we decided to work with datasets of images only, and to abstract from any mathematical aspect, with features and classes intuitively captured by the representations in the pictures. For practical reasons, the number of instances that we propose for each learning task is very small. Our datasets are not meant to be actually used for training, but are an exemplification of possible scenarios.

The Activities

We have defined ten tasks that children are expected to complete in one hour. The ten tasks are divided into four categories, namely: *i) Identifying Regularities*, *ii) Applying Models*, *iii) Classification with Real Images*, and *iv) Dataset Construction*. We now introduce each category.

Identifying Regularities. This category contains three tasks. Each task provides the children with images that are explicitly labeled as instances of some classes, but no explicit model is given (only the instances of the class). The first task focuses on one target class only: children are asked to label four new instances as belonging or not belonging to the class at issue. The second task involves two classes. This time, children are asked to decide to which of the two classes each of the four new items belongs. Also the third task concerns two classes, but this time children are asked whether each of the four new items belongs to one of the two target classes or to none.

The images that we used for this task category depict fictional characters inspired by (code.org - Computational Thinking Lesson) and (code.org - AI for oceans) (although there they are used in a completely different way). Figure 1 shows an example of items provided for the second task, where we use fictional fishes. Relying on fictional characters allows us simplifying the representation, reducing the num-

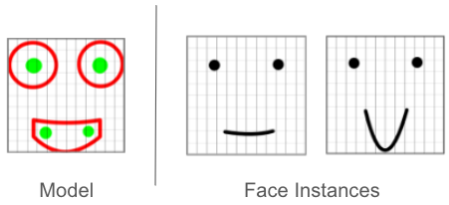

Figure 2: Example from the "Applying Models" set of tasks tackling model accuracy.

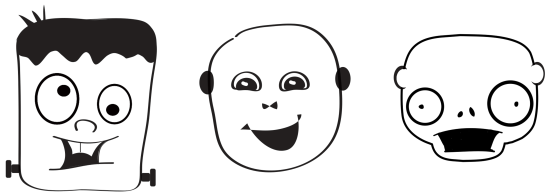

Figure 3: Fictional characters models.

ber of descriptive features. Fishes are characterized just by their *i)* body shape, *ii)* body color, *iii)* eyes shape, and *iv)* tail shape. To correctly solve the exercise, children must understand that not all these features contribute to identifying the class. In particular, only the combination of body color and eyes shape is relevant to correctly classify the fish. This task is, thus, related to the dimensionality problem, and forces children to identify the relevant (combinations of) features and to reason on them.

Applying Models. In this second group of tasks, the models identifying different classes are given explicitly. For instance, Figure 2, taken from (Bebras), on the left-hand side represents a model that has to be used by children to classify smiles. It shows areas for the mouth and for the eyes. When the model is overlaid with an instance of a face (as those represented in the same figure on the right-hand side), the face is classified as smiling if the mouth falls completely within the red area and the origin of the mouth touches the green spots. Similarly, the eyes must fall within the red areas and touch the green spots.

Children are then asked to classify a set of face instances by applying the model. With this exercise, children face situations where the predictions made by the model can provide very unexpected answers, and disappoint expectations. In fact, some face instances are clearly smiling for humans, but are not classified as such by the model, and vice versa (e.g., the right-hand side face in Figure 2 is not a smiling face for the model). This task raises the problem of model *accuracy*: in general, automatically built models will not be perfect, and they might fail the classification of some items.

A second task provides the models for three families of fictional characters, inspired by (code.org - Computational Thinking Lesson) and reported in Figure 3. The challenge, here, is to realize that all the features are weighted the same, and thus a new individual belongs to the family with which it has more features in common. This is unnatural because of the human tendency to weight features differently. For instance, human tend to weight the shape of the head as more important than the shape of the ears, which leads children to assign a new individual to the family with the same head shape, if not observing carefully models and individuals. This is another example of dimensionality problem.

Classification with Real Images. Similarly to the *Identifying Regularities* category, children are given a set of example images – of birds in this case. Then, they are asked to determine whether, on the basis of such examples, a machine would be able to recognize birds never seen before.

The challenge of this group of tasks is to understand that the answers should be given not by applying the human understanding of birds, but rather by applying a notion of similarity, by comparing each image to be classified with those in the dataset.

Of course, the fictional characters used in the *Identifying Regularities* category made the task easier, since the features to be considered were clearly identified. With real images, instead, the identification of the relevant high-level features and the similarity between two images is more challenging and, indeed, it was left to the children. Note also that all the images that we asked to classify contained birds (to the human eye), but some of the pictures captured just details (e.g., the eyes). So, the children must be capable of understand that a machine can only synthesize very partial models of birds, and in some cases these models may not be able to recognize birds.

Dataset Construction. The last two tasks concern the construction of a representative dataset. Also for these tasks, we used real images. The aim is to convey the problem that if the dataset from which a model is learned contains biases, then the predictions made by the model may be wrong, or not as expected.

For instance, we provide children with six images representing either a single dog, details of a dog (e.g., part of the face or the tail), groups of dogs with humans, or a single dog but in unusual positions (such as laying or jumping). Children were asked to choose the three images that they believed to be the best to help recognize dogs. The intuition is that even if a dog is depicted in a not so common position (laying or jumping), it is still a dog and should be included in a dataset. On the contrary, if an image represents only a detail (e.g., the tail), or the image is mixed, containing groups of dogs, or dogs being caressed by people, a machine may find it more difficult to come up with a good model.

4 Analysis and Discussion of the Results

As mentioned, we performed the AMS before and after educational intervention, collecting 120 forms, out of which 104 were completely filled both at the pre- and post-tests.

The AMS questions can be grouped into five categories (Manzi et al. 2020): *Epistemic*, *Emotive*, *Intentions and Desires*, *Perceptive*. To explore whether our intervention had an impact on attribution of mental states by category, we

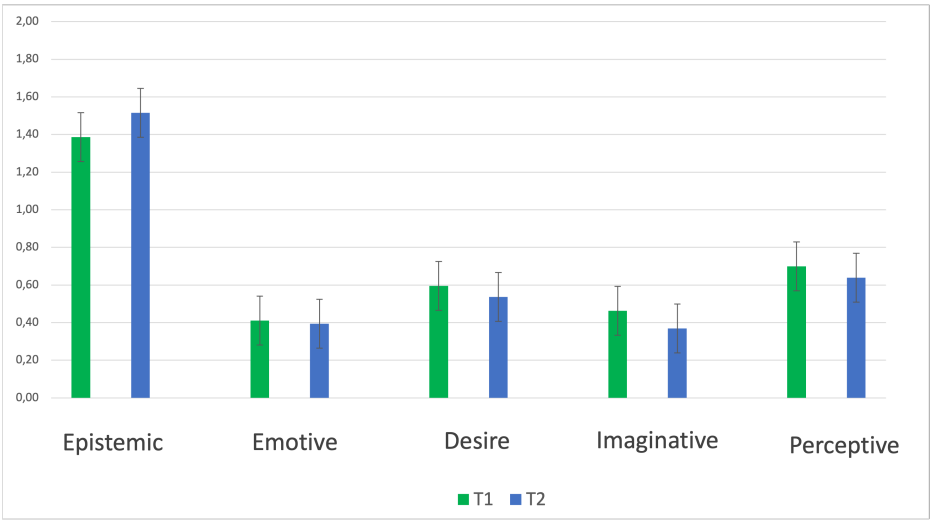

Figure 4: AMS mean values for the five categories at pre-test (T1) and post-test (T2)

performed a mixed repeated measure ANOVA 5x2 (5 categories × 2 administrations) on the AMS scale. It showed a main effect of the category [F(4,103) = 200.66, $p < .001$, $\eta_p^2 = .661$] but no main effect of administration [F(1,103) = 0.43, $p < .516$, $\eta_p^2 = .004$]. These results show that the scores attributed by children to the categories were statistically different. In particular, all category scores significantly differ from each other ($p < .010$ for all pair comparisons) except for those given to Emotive and Imaginative categories which did not differ ($p = .752$). Interestingly, the score attributed to the Epistemic category were significantly higher than the scores given for all the other categories [$p < .001$] (higher scores corresponds to the attribution of more mental states). There is also a tendency for the Imaginative category to decrease, though not in a statistically significant way. The results also show a significant effect of administration by category interaction [F(4,103) = 3.98, $p = .004$, $\eta_p^2 = .037$], indicating that from pre-test to post-test assessment the attribution to specific categories significantly changed. To explore this interaction a series of paired comparisons within each category between pre-test and post-test were performed. The results show that the Epistemic category scores were the only scores that significantly increased from pre- to post-test assessment [t (107) = 2.98, p = .004, Cohen's d = 0.286]. For the other category scores no statistically significant change was found [Emotive: t (107) = -0.03, p = .738; Intentions and Desires: t (107) = -1.17, p = .247; Imaginative: t(105) = -1.080, p = .074; Perceptive: t (103) = -1.39, p = .169].

Figure 4 shows at a glance the variations of the mean scores for the five AMS categories, at pre- and post-test. The picture shows clearly that the Epistemic category mean scores significantly increased in the post-test, whereas all the other means scores decrease, though these latter results did not reach statistical significance. Specifically, after our intervention, children attribute to an AI system more capabilities about learning, taking decisions, memorizing, and understanding (Manzi et al. 2020). We can thus observe that children ascribe to a machine the ability to learn and accomplish a task. On the contrary, as the category mean scores show, children attribute less Emotive and Imaginative traits, such as being happy or sad, feeling cold, or being afraid. Although only the increase on the Epistemic category is statistically significant, the overall trend of these results suggests that children correctly perceive the boundaries of a machine even though it exhibits an intelligent behavior. In fact, in carrying out the activities, children are asked to put in practice what they have learned; they have to "think like" a machine. After our intervention, children attribute to the machines more agency abilities. This may be justified by the fact that they experienced how to train an AI system to make a simple decision. This result goes in accordance with the literature (Frischknecht 2021; Thellman, de Graaf, and Ziemke 2022), where a better understanding of the functioning and a better prediction of a robot behaviors correlates with more mental states attribution. On the other hand, all the other category mean scores decrease. Although this decrease is not statistically significant, this sounds correct for the Emotive, Intentions and Desires, Imaginative, and Perceptive categories, which can be seen as the AMS experience capabilities (children do not humanize machines too much). Indeed, it must be noted that none of these aspects were stimulated in the intervention. For instance, the conductor did not use physical robots (but just a sprite character), and hence children were not exposed to the issue of how robot perceive the world.

We also investigated any gender difference on the attribution of mental states to the ocean sprite robot. Out of the 104 children participating in the trial, only 65 of them provided the gender information. Therefore, we could not perform an in depth analysis on the entire sample. The results we present are thus to be interpreted as a preliminary study to guide further investigations. We performed a mixed repeated measures ANOVA 5x2x2 (5 categories x gender x 2 administrations) on the AMS scale. The results show a

main effect of category [F(4,63) = 93.17, p = .001, η_p^2 = .597], indicating a statistically significant difference among all category scores ($p < .05$ for all pair comparisons) except for Emotive and Imaginative categories which did not differ one another [$p = .848$]. With this small number of participants, no administration by category interaction was found [F(4,63) = 0.88, p = .479, η_p^2 = .014]. Interestingly, a main effect of gender [F(1,63) = 7.80, p= .007, η_p^2 = .110] and a significant gender by category interaction [F(3,63) = 2.77, p = .028, η_p^2= .042] were found. These results show that boys and girls attributed different scores to the various categories, independent of pre- to post-test assessment. In particular, boys attributed higher scores than girls to Epistemic [t(63) = 3.17, p=.002, Cohen's d= 0.775] and Perceptive [t(63) = 4.24, p=.001, Cohen's d= 1.121] categories. No significant difference between boys and girls was found in Emotive [t(63) = 0.22, p = .828], Intentions and Desires [t(63) = 1.90, p = .062], and Imaginative [t(63) = 0.88, p= .382] categories. We can consider that the intervention that we performed did not have an effect in reducing the gender differences, as no significant administration by gender and by category interaction was found. We will investigate in future works whether this difference correlates to the results from the literature (see (Kuhn and Holling 2009)) reporting that boys show better attitudes in science, reasoning and abstract knowledge (Kuhn and Holling 2009) and that boys are in general more engaged in using AI tools (Denner et al. 2005; Teague 2002).

Our analysis also considers qualitative dimensions, and in particular the children's feedback during the training of the ocean sprite robot. During this last phase, the conductor let children decide on whether to continue with the training of the robot (by giving additional examples), or to stop the training and start the testing phase (thus making it classify new samples). In addition, the conductor asked children to explain the robot mistakes (misclassifying fish and garbage). Interestingly, it emerged that children were able to justify the mistakes. For instance, a particular fish had similar color or shape to many garbage samples seen in the learning phase, and so was wrongly considered garbage itself. In providing these justifications, children adequately used the same terminology adopted in the educational intervention (for instance, we used the term *typical model* to identify an image representing the model of a family of individuals. Children were correctly using the term.). These answers suggest that children had internalized the notions explained, such as those of model and training set, and its connected problem of bias. Of course, we take these reactions just as positive feedback about the educational power of the activities, but cannot give them statistical value.

5 Limitations

A first limitation that we acknowledge to this study is that it has been performed in Italy only. As mentioned in the literature, the attribution of mental states to AI depends on several factors, such as the context and the background of the participants. Therefore, currently we are not able to say how generalizable the results are to other countries.

Concerning context and background, we did not considered them as variables in this study and mitigated their effects by considering a high number of participants. However, it would be interesting, in future works, to consider them explicitly and investigate their effects on agency and experience mental states attribution.

Due to time and children's attention limits, we had to select a small number of topics to present in our activities. So we focused on classification, use of models, and dataset construction since these are foundational topics of ML. More advanced topics, such as unsupervised learning, can be introduced in the future.

Finally, we decided to adopt the AMS scale to assess the effects of our activities for two main reasons. First, children can fill it very easily without getting tired, and hence invalidating the test. Second, the AMS questionnaire is standardized and validated in the literature. However, by using the AMS, a direct comparison with other approaches may become more difficult since it does not categorize items into *agency* and *experience*, as other approaches in the literature. In fact, the AMS provides a more fine-grained categorization that we mapped as follows: the Emotive, Imaginative, Perceptive sections of the AMS correspond to the experience category; while the Epistemic section corresponds to agency. The Intentions and Desire section of the AMS, instead, does not fit well neither with agency nor with experience. Questions falling in this section, in fact, do not clearly refer to *acting and planning* capabilities (agency), nor to *perceiving and feeling* (experience).

6 Related Works

Over the last few years, research on how to introduce children and young adults to AI, and ML, has drawn the interest of many researches, resulting in so many proposals that it is not possible to exhaustively cover all of them.

One line of research focuses on the integration of AI concepts into K-12 curricula (e.g., (Zimmerman 2018)), addressing aspects such as which AI concept should be addressed and at which age. Among these, the AI4K12 Initiative (AI4K12.org), with its "five big ideas" of AI, proposes a reference framework listing the AI concepts that every K12 student should know (Touretzky et al. 2019; Touretzky, Gardner-McCune, and Seehorn 2023). The framework points out learning objectives in relation to the children's age, and introduces general guidelines for AI teaching paths. In (Sanusi et al. 2023), a systematic review of ML teaching and ML learning in K-12 is presented. The work highlights three key findings about effective ML teaching. These are: *i)* creating more ML activities for kindergarten through middle school and education in an informal context; *ii)* incorporating ML ideas into subject domains other than computer science to promote the integration of ML in schools (thinking strategies training can potentially be integrated in any school subject); and *iii)* developing assessments for ML that can be relevant across grade levels to compare students' ML understanding in different learning environments (achievable by not requiring prior knowledge).

Other works focus on tools, such as (code.org - AI for oceans; Google - Teachable Machines; ML for Kids), to

demonstrate the functioning of ML algorithms by allowing children to make their own experience in training a model. In fact, such tools, by way of an appealing graphical interface, help children to define a set of classes to create a training set of labeled instances, and to test their set-up. However, they hardly provide any insight into the processes that are involved, and how the hidden mechanisms actually work. As explained in (Gresse von Wangenheim et al. 2021), that reviews 16 interactive ML tools for K-12, the validity of the content of some of these tools is yet to be proved, but in our view, the greatest general pitfall of this kind of approaches is their lack of generalization and explanation (Tedre et al. 2021): showing that using a ML algorithm in a certain way is effective on some domain does not explain why it is effective. Thus, children have no clues on how to approach a different application domain, or a different ML algorithm: will the same approach still be effective?

Other approaches propose *unplugged* AI or Machine Learning activities. For instance, (Lindner, Seegerer, and Romeike 2019) presents two activities on ML to make secondary schools student compare ML and traditional approaches to AI; (Lehner and Landman 2024) approach includes an unplugged activity on decision trees; (Ossovski and Brinkmeier 2019) propose an activity on linear classification. All the above target secondary school students. The idea of using unplugged activities for teaching Computational Thinking and AI is not new, and often unplugged activities complement hands-on activities with educational robots and tools (see e.g., (Bell, Rosamond, and Casey 2012; Alamer et al. 2015; Fowler 2017; Lindner, Seegerer, and Romeike 2019)). These works are closer to our approach, since our tasks are conceived as unplugged activities. However, the activities that we propose can be used in two ways: *i)* As an educational tool inserted in a proper narrative framework, as discussed in this paper, where we have shown the effectiveness of this approach by measuring mental states attribution in children; *ii)* As a tool for evaluating children performances in solving ML tasks. In this case, the activities can be used to assess the effectiveness of other tools or educational paths by measuring children performances in solving the tasks (not addressed in the paper).

As mentioned in the introduction, the idea of leveraging mental state attribution to grasp people perception about a system is not new. Thellman, de Graaf, and Ziemke (2022) present a literature review on the attribution of mental states to robots. Wischnewski (2025), instead, focused on non-embodied autonomous systems. In this area, Shank and DeSanti (2018) investigate the attribution of mental states and the attribution of morality to AI systems by people. Frischknecht (2021) investigate the correlation of mental state attribution with the level of autonomy of a system. None of these approaches, however, investigate how and whether an educational intervention is able to modify the mental states attribution in children.

7 Conclusion

We presented the results of a study where we investigate how mental state attribution to an AI changes in children after the administration of an educational intervention consisting of a set of unplugged, machine learning tasks that are suitable for children from 9 through 12 years old. The tasks concern classification, including aspects such as dataset, models, and biases in datasets. For measuring the attribution of mental states by children, we leverage the AMS scale (Manzi et al. 2020), a well known questionnaire which is fast to be filled by children and that can be replicated in our studies.

Results show that in just one-hour of activity, children changed their mental state attribution, attributing less mental states for the categories *Emotive*, *Intentions and Desires*, *Imaginative*, and *Perception*, that is they better recognize ML tools as machines, and avoid humanizing them. For the AMS epistemic category, that includes capabilities such as *learn*, *think*, *remember*, *decide*, and *understand*, instead, the mental state attribution tend to increase. This can be interpreted as a better understanding of what AI can do and how it works, in line with (Frischknecht 2021; Thellman, de Graaf, and Ziemke 2022). Indeed, the proposed activities focus on aspects such as how a machine can "learn" whether a new instance is part of a class or "decide" the class of a new instance. Thus, we associate an improved understanding of ML functioning with an increase of epistemic mental states attribution.

As a next step along this line of research, we want to investigate whether different types of educational intervention (e.g., by using educational robots, or by stimulating basic children's reasoning abilities) have an effect on children mental state attribution (Baldoni et al. 2024b) and how these compare with the intervention described in this paper. Specifically, the literature shows that there is a tight connection between thinking strategies (such as slow vs fast thinking) and ML understanding. Therefore, we want to explore whether by training children slow thinking, their mental states attribution and their understanding of ML improves. The advantage would be that, instead of teaching one specific concept one can train children to develop the right mind attitude and intuition to interpret the functioning of AI-based systems in general.

Additionally, we want to investigate the possible correlation between the perception each child has of his/her capabilities compared to their performance in solving ML tasks. In particular, we want to use the educational activities as a testing tool and see if there exists a relationship between gender, self-confidence, performance in ML tasks, and mental state attribution. Understanding these aspects could be useful to delineate a learning path tailored to the needs of each child (e.g., to be used as a complement to approaches such as (Li et al. 2023)). The tasks could be used to understand which abilities are weak, in order to reinforce them, personalizing the training. We also plan to apply the proposal to adults, instead of children. In fact, the digital divide between those who master AI tools and those who do not also affects adults, especially in their working life. Thus, taking inspiration from the activities we have developed, we intend to investigate the teaching of ML to adults.

The tasks and material we have produced will be available for anyone interested in replicating the activities. To this aim, we are composing a teaching manual for instructors.

Acknowledgments

This work was supported by the project *AI-LEAP: LEArning Personalization with AI and of AI* (D13C23001280007) financed by Fondazione Compagnia di San Paolo and Cassa Depositi e Prestiti.

References

Alamer, R. A.; Al-Doweesh, W. A.; Al-Khalifa, H. S.; and Al-Razgan, M. S. 2015. Programming Unplugged: Bridging CS Unplugged Activities Gap for Learning Key Programming Concepts. In *2015 Fifth International Conference on e-Learning (ECONF)*.

Baldoni, M.; Baroglio, C.; Bucciarelli, M.; Capecchi, S.; Gandolfi, E.; Gena, C.; Ianì, F.; Marengo, E.; Micalizio, R.; Rapp, A.; and Nabil Ras, I. 2024a. Does Any AI-Based Activity Contribute to Develop AI Conception? A Case Study with Italian Fifth and Sixth Grade Classes. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI*. AAAI Press.

Baldoni, M.; Baroglio, C.; Bucciarelli, M.; Gandolfi, E.; Ianì, F.; Marengo, E.; and Nabil Ras, I. 2022. Empowering AI Competences in Children: The First Turning Point. In *Methodologies and Intelligent Systems for Technology Enhanced Learning, Workshops, 12th International Conference, MIS4TEL*, LNNS 538. Springer.

Baldoni, M.; Baroglio, C.; Bucciarelli, M.; Micalizio, R.; Gandolfi, E.; Iani, F.; Marengo, E.; and Capecchi, S. 2024b. Thinking Strategies Training to Support the Development of Machine Learning Understanding. A study targeting fifth-grade children. In *Proceedings of the 2024 9th International Conference on Information and Education Innovations*, ICIEI '24, 85–92. New York, NY, USA: Association for Computing Machinery. ISBN 9798400716409.

Bebras. 2017. Give me a Smile. Bebras dell'Informatica. Laboratorio di Didattica e Divulgazione dell'Informatica. Università degli Studi di Milano, Dipartimento di Informatica. https://bebras.it/platform/html/player_teacher.html?class=screenshot&code=2017_DE-02_GiveMeASmile.

Bell, T.; Rosamond, F.; and Casey, N. 2012. *Computer Science Unplugged and Related Projects in Math and Computer Science Popularization*, 398–456. Springer Berlin Heidelberg.

Brüne, M.; and Brüne-Cohrs, U. 2006. Theory of Mind—Evolution, Ontogeny, Brain Mechanisms and Psychopathology. *Neuroscience & Biobehavioral Reviews*, 30(4): 437–455.

code.org. 2020. Code.org. (Last access August 2025).

code.org - AI for oceans. 2020. Code.org – AI and Machine Learning. AI for Oceans. https://studio.code.org/s/oceans (Last access August 2025).

code.org - Computational Thinking Lesson. 2020. Code.org – Computational Thinking Lesson. https://studio.code.org/unplugged/unplug2.pdf (Last access August 2025).

de Graaf, M. M.; and Malle, B. F. 2019. People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 239–248.

Denner, J.; Coyle, K.; Robin, L.; and Banspach, S. 2005. Integrating Service Learning into a Curriculum to Reduce Health Risks at Alternative High Schools. *Journal of School Health*, 75(5): 151–156.

Epley, N.; Waytz, A.; and Cacioppo, J. T. 2007. On Seeing Human: a Three-Factor Theory of Anthropomorphism. *Psychological review*, 114(4): 864–886.

Fowler, A. 2017. Engaging Young Learners in Making Games: an Exploratory Study. In *International Conference on the Foundations of Digital Games*, FDG '17. Association for Computing Machinery. ISBN 9781450353199.

Frischknecht, R. 2021. A Social Cognition Perspective on Autonomous Technology. *Computers in Human Behavior*, 122: 106815.

Google - Teachable Machines. 2020. Teachable Machine – Train a Computer to Recognize Your Own Images, Sounds, & Poses. https://teachablemachine.withgoogle.com./ (Last access August 2025).

Gresse von Wangenheim, C.; Hauck, J. C. R.; Pacheco, F. S.; and Bertonceli Bueno, M. F. 2021. Visual Tools for Teaching Machine Learning in K-12: A Ten-Year Systematic Mapping. *Educ. Inf. Technol.*, 26(5).

Kuhn, J.-T.; and Holling, H. 2009. Exploring the Nature of Divergent Thinking: A Multilevel Analysis. *Thinking Skills and Creativity*, 4(2): 116–123.

Lehner, L.; and Landman, M. 2024. Unplugged Decision Tree Learning–A Learning Activity for Machine Learning Education in K-12. In *International Conference on Creative Mathematical Sciences Communication*. Springer.

Levin, D. T.; Killingsworth, S. S.; Saylor, M. M.; Gordon, S. M.; and Kawamura, K. 2013. Tests of Concepts About Different Kinds of Minds: Predictions About the Behavior of Computers, Robots, and People. *Human–Computer Interaction*, 28(2).

Li, T.; Wang, X.; Zhang, S.; Yang, F.; and Lu, W. 2023. A Personalized Learning Path Recommendation Method for Learning Objects with Diverse Coverage Levels. In *Artificial Intelligence in Education*. Cham: Springer Nature Switzerland.

Lindner, A.; Seegerer, S.; and Romeike, R. 2019. Unplugged Activities in the Context of AI. In *International Conference on Informatics in Schools: Situation, Evolution, and Perspectives (ISSEP)*. Springer.

Manzi, F.; Peretti, G.; Di Dio, C.; Cangelosi, A.; Itakura, S.; Kanda, T.; Hiroshi, I.; Massaro, D.; and Marchetti, A. 2020. A Robot Is Not Worth Another: Exploring Children's Mental State Attribution to Different Humanoid Robots. *Frontiers in Psychology*, 11: 1–12.

Marchesi, S.; Ghiglino, D.; Ciardo, F.; Perez-Osorio, J.; Baykara, E.; and Wykowska, A. 2019. Do We Adopt the Intentional Stance Toward Humanoid Robots? *Frontiers in Psychology*, 10.

McEneaney, J. E. 2009. Agency Attribution in Human-Computer Interaction. In Harris, D., ed., *Engineering Psychology and Cognitive Ergonomics*, 81–90. Berlin, Heidelberg: Springer Berlin Heidelberg.

ML for Kids. 2023. Machine Learning for Kids. https://machinelearningforkids.co.uk/ (Last access August 2025).

Ossovski, E.; and Brinkmeier, M. 2019. Machine Learning Unplugged - Development and Evaluation of a Workshop About Machine Learning. In *Informatics in Schools. New Ideas in School Informatics*. Springer International Publishing.

Sanusi, I. T.; Oyelere, S. S.; Vartiainen, H.; Suhonen, J.; and Tukiainen, M. 2023. A Systematic Review of Teaching and Learning Machine Learning in K-12 Education. *Educ. Inf. Technol.*, 28(5).

Shank, D. B.; and DeSanti, A. 2018. Attributions of Morality and Mind to Artificial Intelligence After Real-World Moral Violations. *Computers in Human Behavior*, 86: 401–411.

Teague, J. 2002. Women in Computing: What Brings Them To It, What Keeps Them In It? *SIGCSE Bull.*, 34(2): 147–158.

Tedre, M.; Toivonen, T.; Kahila, J.; Vartiainen, H.; Valtonen, T.; Jormanainen, I.; and Pears, A. 2021. Teaching Machine Learning in K-12 Classroom: Pedagogical and Technological Trajectories for Artificial Intelligence Education. *IEEE access*, 9.

Thellman, S. 2021. *Social Robots as Intentional Agents*. Ph.D. thesis, PhD dissertation, Linköping University Electronic Press.

Thellman, S.; de Graaf, M.; and Ziemke, T. 2022. Mental State Attribution to Robots: A Systematic Review of Conceptions, Methods, and Findings. *J. Hum.-Robot Interact.*, 11(4).

Touretzky, D.; Gardner-McCune, C.; Martin, F.; and Seehorn, D. 2019. Envisioning AI for K-12: What should every child know about AI? *33rd AAAI Conference on Artificial Intelligence, AAAI 2019 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*.

Touretzky, D.; Gardner-McCune, C.; and Seehorn, D. 2023. Machine Learning and the Five Big Ideas in AI. *International journal of artificial intelligence in education*.

UNICEF. 2019. Workshop Report: AI and Child Rights Policy.

Wischnewski, M. 2025. Attributing Mental States to Non-Embodied Autonomous Systems: A Systematic Review. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713958.

Wooldridge, M.; and Jennings, N. 1995. Intelligent Agents: Theory and Practice. *Knowledge Engineering Review*, 10(2): 115–152.

Zimmerman, M. 2018. *Teaching AI: Exploring new frontiers for learning*. International society for technology in education.