Identifying Secondary School Students' Misconceptions about Machine Learning: An Interview Study

Erik Marx*
Center for Scalable Data Analytics
and Artificial Intelligence (ScaDS.AI)
Dresden/Leipzig, Germany
erik.marx@tu-dresden.de

Clemens Witt
University of Technology Dresden
Dresden, Germany
clemens.witt@tu-dresden.de

Thiemo Leonhardt
RWTH Aachen University
Aachen, Germany
leonhardt@cs.rwth-aachen.de

ABSTRACT

Since students are familiar with machine learning (ML)-based applications in their everyday lives, they already construct mental models of how these systems work. This can result in misconceptions that influence the learning of correct ML concepts. Therefore, this study investigates the misconceptions students hold about the functionality of ML-based applications. To this end, we conducted semi-structured interviews with five students, focusing on their understanding of facial recognition and ChatGPT. The interviews were analyzed using an inductively developed code system and qualitative content analysis. This process identified six key misconceptions held by students: "Programmed Behavior," "Exactness," "Data Storage," "Continuous Learning," "User-trained Model," and "Autonomous Data Acquisition". These misconceptions include the notion that AI learns continuously during application or that training data is saved and reused later. This paper presents the identified misconceptions and discusses their implication for the design and evaluation of effective learning activities in the context of ML.

CCS CONCEPTS

• **Social and professional topics → K-12 education**; **Computer science education**; **Computational thinking**; • **Computing methodologies → Machine learning**; **Artificial intelligence**.

KEYWORDS

students conceptions, mental models, machine learning, artificial intelligence, interview study, qualitativ research

ACM Reference Format:
Erik Marx, Clemens Witt, and Thiemo Leonhardt. 2024. Identifying Secondary School Students' Misconceptions about Machine Learning: An Interview Study. In *Proceedings of The 19th WiPSCE Conference on Primary and Secondary Computing Education (WiPSCE '24).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

A considerable number of publications have appeared on the subject of ML education in which the desired learning objectives, competencies and corresponding curricula are presented [14, 15, 19, 26, 34]. However, divergent opinions exist regarding the extent to which ML concepts are understandable for students. Sanusi et al. posit that AI is challenging to convey due to the abstract nature of many concepts [26]. Conversely, various studies indicate that children are capable of learning basic ML concepts [8].

One aspect that affects the learning of new concepts is the existence of mental models [21, 29]. It can be expected that students already possess mental models of ML when they come into contact with the subject at school. Students construct mental models based on everyday experiences [31, 37]. They frequently encounter ML applications, such as facial recognition on smartphones, recommendation systems on platforms like TikTok, or even when completing assignments with ChatGPT. Additionally, analogies to familiar concepts are used when constructing mental models [5, 6], so general conceptions of AI need to be considered. Finally, these mental models are related to existing computer science knowledge. For instance, students resort to classical computational thinking (CT) concepts when unable to explain the functionality of ML applications [8]. However, the domain of ML is distinct, representing a paradigm shift that deviates from established concepts of classical computational thinking [33, 37], potentially hindering understanding due to existing knowledge [20]. For example, according to Tedre et al., ML problems entail different notions of correctness, as programs are no longer evaluated based on the correctness of a pre-designed algorithm but rather on their level of confidence determined and assessed by developers using appropriate metrics. Additionally, ML problems differ in employed problem-solving and debugging strategies as well as in their transparency [33].

Ultimately, the question arises as to what pre-instructional mental models do students bring into the classroom, particularly those that can influence the learning of ML. In a comprehensive scoping review, we identified that considerable research gaps still exist in this area. Previous studies have solely taken a superficial look at the subject area of AI. Additionally, investigations into the mental models students hold regarding specific applications and the misconceptions that may arise remain scarce. Furthermore, we found that a traditional method of conception research, interviews, has been underutilized [16]. In this study, we aim to address these existing research gaps guided by the following research question:

> *What misconceptions arise from students' mental models on machine learning?*

[*] Also with University of Technology Dresden.

WiPSCE '24, September 16 – September 18, 2024, Munich, GER
© 2024 Association for Computing Machinery.
ACM ISBN XXX-X-XXXX-XXXX-X/XX/XX…$XX.XX
https://doi.org/XXXXXXX.XXXXXXX

In the following, we will first present related work and clarify the theoretical foundations by presenting a theoretical overview over mental model theory and the technological foundations. Next, we will outline the methodology of our interview study and the data analysis procedure. Subsequently, we will present the misconceptions identified regarding the research question and discuss their implications for designing learning activities.

2 RELATED WORK

In our scoping review we found that the majority of mental model research in the field of ML is focused on general conceptions of AI, yet resulting misconceptions and their influence on the learning process remain underexplored [16]. As scoping reviews provide only an overview of research approaches, we will summarize the relevant results of related work below.

The applications students associate with AI or ML include recommender systems [1, 3], knowledge-based systems [12], cookies/web browsers [12], computers [3, 27, 36, 37], voice assistants [12, 17, 31, 36], and smartphones [12, 26]. In addition to everyday applications, numerous studies have also identified entertainment media, films, and science fiction as influential factors [1, 8, 17, 30]. It is generally associated with AI that: AI is programmable or programmed [3, 25, 36], requires internet access or a cloud [12, 17, 27, 37] and is capable of learning [17, 20, 36]. More specific results can be found on the learning process. For example, learning is understood as repeated attempts to improve the system or increase the range of features [3, 20]. There are also references to the idea that data is stored during learning [12] or that human behavior is learned [3], but these are research outlines rather than fully-fledged studies. Finally, there is also the opposite idea that AI systems cannot learn at all, but their behavior is hard-coded [11, 12, 17, 18].

In addition to the assumption of learning human behavior, there are other findings pertaining the anthropomorphization of AI. For example, it is assumed that AI functions like a human brain or acts like a human [1, 17, 18, 30]. The existence of consciousness, emotions and general characteristics of a strong AI have also been assessed [11, 17]. Results on conceptions of the data used by ML systems also exist, albeit sporadically. For example, Kim et al. identified the notions that AI can use any form of data and only needs a large amount of data [11]. Similarly, Sanusi et al. observed the notion that more data increases the accuracy of the system [26]. Finally, it is important to consider the findings on controllability and explainability. For instance, there are ideas that AI is (completely) controllable or (completely) uncontrollable [1, 18, 20, 30]. Additionally, AI is sometimes perceived as impartial, fair, or objective [11, 14].

At this point, we focus on the studies which were explicitly concerned with identifying misconceptions in AI or ML (also called naive conceptions by Kim et al. [11]). Mertala & Fagerlund conducted a qualitative online questionnaire with 195 Finnish 5th and 6th graders to identify their misconceptions about AI. The questionnaire included five open-ended questions for the children to answer. The authors then identified three categories of misconceptions. The first category included misconceptions in which the term AI was not understood as a technical term and therefore reflected a complete lack of knowledge about the concept. The second and most prevalent category of misconceptions pertained to anthropomorphic AI, wherein AI was ascribed human capabilities or behaviors. The third misconception was that AI possessed knowledge or intelligence that had been pre-installed by developers [17]. An interesting counterpoint is made by Mühling & Große-Bölting, who found no profound misconceptions in their study, only anthropomorphizing statements made by students, which they explicitly did not classify as misconceptions. We take up the discussion of anthropomorphizing language in section 5.

Schaper et al. identified three key misconceptions in their analysis of drawings and accompanying explanations by 11 Danish eighth graders about the use of technology and ML in the future [27]. First, some students assumed that every robot uses ML. Second, some students assumed that devices connected to a computer use the Internet and therefore also use ML. Third, some students assumed that AI works without human input.

Kim et al. qualitatively analyzed learning artifacts and videos from 14 6th-8th graders who participated in a summer camp on AI, and presented the progression of their beliefs. They identified five themes of naive conceptions, such as that AI can use any data, or that AI is the same as automation and robots [11].

As can be observed, the studies to date do not delve deeply into the specifics of AI or ML. Consequently, technological foundations or a conceptual model are not employed to assess the presence of misconceptions, presumably because the ideas identified are typically so superficial that no uniform model can be applied. From this vantage point, it is only possible to identify ideas that are characterized by a strong lack of knowledge or which cannot be classified as generally incorrect. The survey methods used also left little room for understanding students' specific thought processes in detail. As a result, only limited hypotheses can be derived from these studies' findings regarding the impact of these misconceptions on the learning of ML concepts. Therefore, the objective of our study is to identify specific misconceptions that relate more directly to individual steps of the ML workflow.

3 THEORETICAL BACKGROUND

The field of student conceptions research is rich in both theory and terminology. A foundational theory in this field is that of mental models, which has significantly shaped research on student conceptions [21]. Mental models are based on the constructivist premise that knowledge cannot simply be transmitted in a passive manner, but must be independently "constructed" by learners, integrating it into their existing cognitive framework [28, 32]. Mental models are cognitive representations of situations or domains that aid in reasoning, learning, inference, or prediction [5, 6]. They are used to transfer phenomena from the external world into a mental representation. Mental models retain the structure and dynamics of the phenomenon (spatial, temporal, causal) and are available for mental operations (sometimes referred to as "simulation"). Since they have the same structure as the actual phenomenon, conclusions about the phenomenon can be drawn by manipulating the model [6, 21]. One challenge in investigating mental models is the relationship between language and cognition, which is intricate and multifaceted. Language affects cognitive processes and vice versa. The precise manner in which language interacts with cognition remains largely unknown [23]. Nevertheless, interpretive analysis of interviews is a a well-established tool in mental model research [5].

The exact properties and functionalities of mental models depend on the theoretical viewpoint being considered. The field of mental models is characterized by two somewhat divergent understandings of mental models. Greca & Moreira refer to these as the "theoretical approach" and the "instructional approach" [6]. The objective of the *theoretical approach* is to present a unified theory capable of explaining various cognitive phenomena, such as reading and language comprehension [6, 21]. According to this perspective, mental models are constructs formed temporarily at the moment of their utilization to accomplish tasks in the present moment [5, 10, 21]. The *instructional approach* describes the knowledge that individuals develop about physical or technological systems, or in other words knowledge-rich domains, without the goal of proposing a universally applicable theory [5, 6]. Thus, mental models represent subjective functional models for specific domains [21]. In practice, however, there is less distinction made between them [21], and many theories regarding student conceptions and conceptual change derived from these incorporate characteristics of both approaches. Therefore, at this point, we do not intend to further differentiate but rather consider the characteristics arising from both approaches.

Primarily, mental models are subjective, as already inferred from the constructivist theory [6]. In addition, Norman introduces several important properties of mental models. Mental models are incomplete and lack clearly defined boundaries. Consequently, they do not fully represent a problem domain, and similar systems are often confused with each other [22]. This is consistent with the observation that the less similar two systems are at a surface level, the poorer the transfer between mental models, even if the systems are isomorphic [28]. Additionally, mental models are also unscientific, meaning they reflect "superstitious" beliefs or conceptions that may not necessarily be factually correct [22]. Therefore, they can yield both incorrect and correct results [5, 6]. Furthermore, according to the theoretical approach, mental models are abstract, dynamic, adaptive, and continuously adjusted [6, 21, 24]. Finally, attention should be drawn to the "parallelism" of mental models. Individuals may possess multiple mental models pertaining to a given topic as mental models lack clearly defined boundaries. Furthermore, these models may conflict with one another, particularly for complex problems (like ML) [6, 28, 29].

As demonstrated , mental models are intricate and often challenging to clearly characterize. However, given the relevance of both the theoretical and instructional approach for didactic research, there are various theories in conceptual change research that integrate the two [21]. Given these numerous interconnections, it is not surprising that terms from conceptual change research like "conception" are often used synonymously with mental models and in many cases can be regarded as such [4, 21, 25]. Some authors posit that the two concepts can be considered distinct levels of abstraction [20]. However, Franco et al. argue that there is no universally applicable relationship between the two terms, but that they are highly domain-specific [4]. In the following, we define the term "mental model" as the respondents' internal cognitive framework. In relation to this, under the term "conception" we summarize the observable results of the use of mental models (e.g. in the form of students' responses). Finally, we specify the term "misconception", which is also referred to as "alternative" or "naive conception" by some authors. While Bewersdorf et al. refer to misconceptions as "flawed mental models" [1], we wish to provide a more concrete definition with the use of a conceptual model. A conceptual model is a precise and complete representation that is coherent with scientifically accepted knowledge [6]. This can range from a simple analogy to a complex explanation [28]. Conceptual models are mostly used by educators to promote the construction of favorable mental models [6, 22]. We consider statements to be a misconception when they are not compatible with our chosen conceptual model of the ML workflow.

3.1 The ML-Workflow as a Conceptual Model

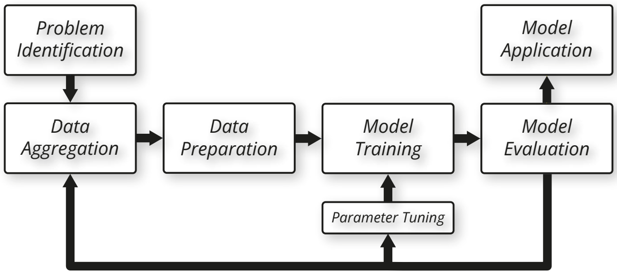

Figure 1: The ML-Workflow

Despite the wide variety of technical approaches employed in the field of ML, a number of overarching commonalities in the ML process can be identified. The term *ML-Workflow* represents an abstraction of the problem-solving steps in the field of ML (figure 1). Zimmerman presents the prototypical ML-Workflow in seven steps [39]. The initial step of problem identification precedes the actual problem-solving process. It involves evaluating the suitability of ML methods as a potential solution for the given problem. If ML methods prove to be an adequate approach, the subsequent step would be the aggregation of data by the developers. Once a sufficient quantity and quality of data has been acquired, the data is checked for inconsistencies, redundancies, and errors, and is accordingly cleaned for training (step data preparation). Prior to the initiation of the training process, it is essential to select the model to be trained, the training algorithm and the training parameters. While training a statistical model is fitted to the data. Upon completion of the training, the performance of the model is evaluated using previously unseen data. In the event that the performance of a trained model does not meet the specified requirements, it is necessary to adjust the selection of data or the learning parameters, often in repeated iterations. The model is only released for use in the intended application scenarios once the adaptations have led to desired model behavior. No further modifications are made to the model during its application. Any subsequent alterations are only implemented once the model is retrained and the ML workflow is initiated anew.

The didactic potential of the ML workflow has been widely discussed in several publications in the field of AI education. For example, Tedre et al. emphasize its importance in illustrating the differences between problem-solving processes in ML compared to classical computer science problems [33]. In their competency framework, Long & Magerko identify the comprehension of the steps of the ML-Workflow and the associated practices and challenges as a central competency for learners engaging with AI [15]. Furthermore, Michaeli et al. delineate a number of competences within the domain of ML that can be attributed to the ML-Workflow [19]. Similarly, Touretzky et al. provide guidelines for the third Big Idea in the field of AI – Learning. Here, the authors highlight "essential insights" of the ML-Workflow as specific competency development goals [34].

3.2 Technological Background

The objective of *facial recognition systems* is to identify, authenticate or classify individuals. To achieve this, the individual must first be recognized in a picture or video frame. Secondly, in order to normalize a face when it deviates from the focal point, the algorithm must be trained to identify general facial features and then center the face. In a third step key features are extracted from the image by a convolutional neural network (CNN). In the fourth step, a person is authenticated by comparing the extracted features with the ones saved during setup, typically using a Euclidean distance matrix in a one-to-one comparison [2]. In conclusion, it is essential to emphasize that when facial recognition is set up by the user, no machine learning (ML) model is trained; rather, only pre-trained models are utilised to extract a feature vector, which is later used to authenticate the individual in question.

Another area in which ML is employed is the use of *large language models* (LLM). These models utilize a specialized type of neural network known as a transformer specifically developed for processing sequences, such as text or code. Mechanisms, such as attention, are employed by the model to enable it to recognize and focus on important elements within an input text. This technique is particularly useful for understanding context and retaining information over longer sections of a sequence [35]. The model is trained in two phases. In the initial phase, training is conducted through self-supervised learning on a large body of text data. During this stage, the statistical relationships inherent in language, such as spelling and grammar, are trained. In the subsequent phase, the model is trained for a specific task using transfer learning and a smaller data set selected for this purpose [9]. In the case of ChatGPT, for example, this could be the conversational style of a chatbot. In both steps, model parameters, such as weights, are modified during training, resulting in a change in the model's structure. Conversely, this is not the case when the user interacts with the model, as using the model does not alter its underlying structure.

4 METHODOLOGY

A semi-structured interview format was chosen to explore students' thoughts on ML, which is a classic instrument for investigating mental models that has been used sparingly to date in the area of ML, although it allows for in-depth exploration of students' thought processes [5, 16]. Due to the vast and largely abstract nature of the field of ML, we decided to explore the topic through ML-based technologies, because everyday experiences were identified as one influencing factor. Furthermore, Jones et al. have demonstrated that respondents' answers are more nuanced and detailed when they have direct interaction with the phenomenon under investigation [10]. The selected applications were *facial recognition in smartphones* and *ChatGPT*. Both applications were therefore presented to the participants, who interacted with them with the goal of stimulating reflection on their functionality. The specific choice of these two applications warrants further explanation. The advantage of these applications is that the students are already familiar with them. In addition, their functionality is straightforward to understand and demonstrate, in contrast to other applications such as recommendation systems. Finally, as described earlier, conceptions of AI represent another potential influencing factor in mental model construction. We, therefore, used these two applications to moderate the influence of AI conceptions by selecting one application that is strongly associated with AI (ChatGPT) and one in which the use of ML is more concealed (Facial Recognition) [26].

4.1 Participants

In March 2023, we interviewed five German students who were recruited from an internship at the faculty or through an extracurricular program, indicating their inherent interest in the subject. Despite this shared interest in computer science, the participants exhibited varying levels of prior knowledge in ML. One student had previously attended a ML workshop, while two others had observed such a workshop during their internship. The two remaining students lacked formal education in AI or ML. Their knowledge, if any, likely stemmed from informal engagement with these topics during their free time. These diverse backgrounds provided valuable insights into how newly acquired concepts can influence explanations. In total, the study involved one female (age 14) and four male participants (ages 10, 15, 15 and 17). The interviews ranged in duration from 44 to 69 minutes.

4.2 Interview Structuring

A structured interview guide, developed beforehand, served as a framework for the interview. The interview guide was composed of questions and potential follow-up inquiries designed to elicit in-depth explanations from the participants[1]. When formulating questions for the interview guide, emphasis was placed on developing "generative questions," open-ended prompts that encouraged extended explanations and in-depth reflections [21]. Whenever feasible, the teach-back technique was employed, in which the interviewer reformulates the respondent's response to facilitate reflection and correction [16]. The ML-Workflow also served as a guiding framework for clarifying students responses, particularly regarding the distinction between the learning and application phases, as well as the evaluation process.

The themes discussed in the interview were derived from the presumed influencing factors that shape mental models of ML. These factors included traditional CT concepts, such as correctness, debugging, and problem-solving as described by Tedre et al. [33]. Additionally, questions were tailored to align with the common associations that students hold regarding AI. These inquiries delved into the students' understanding of programming requirements

[1] The interview guide as well as the untranslated original transcripts can be found at https://osf.io/uf3an/

Table 1: Codebook for Dimension "General Perspectives on ML"

Code	Definition	Example
Blackbox	Certain parts of a question cannot be answered due to a lack of knowledge.	*I don't know exactly, it's in my laptop at home. So ... I don't know exactly*
Anthropomorphic Thinking		
User-like Acting	Inner workings are explained in terms of human behavior.	*Well, let's assume that it has now hypothetically found 30 tabs where you have something with "Mail" and "Apply for student internship" ... then it extracts the information from them or simply drags it over ... in here [points to chat GPT tab] and then perhaps writes it up nicely.*
Human-like Cognition	Inner workings are explained in terms of human cognition.	*It's like: "no, the answer wasn't quite right", or "yes", then I [the AI] think about it again. So I think about: "What could I say next time?" or something like that. Or: "How do I do it differently?", "How can I do it better?"*
Algorithmic Thinking	Problem-solving strategies are described algorithmically.	*Maybe you could imagine it a bit like Scratch with this if-then command. So "if this and this", meaning if all the features match, then "access allowed to the mobile phone" or "unlock".*
Data-based Thinking		
Pattern-based Processing	Information derived from data is used for problem solving.	*The smartphone scans - or rather, the camera scans the face for various features that have been saved for the image - and this allows you to unlock the phone.*
Raw Data Processing	Raw data is used to solve a problem.	*The image is already saved and that's essentially the only thing it needs to access when it compares the two images to unlock [the smartphone].*

[3, 11, 17, 25, 36], the necessity of internet connectivity [17, 27, 37], the concept of consciousness in AI systems [11, 17, 25], and the data handling practices of ML applications [11, 20, 25, 26].

The first application discussed was facial recognition. Initially, the technology was demonstrated to the students, followed by a walk through of the setup. Subsequently, the students were engaged in a structured interview to explore their understanding of the application's functioning. Accordingly, to minimize the influence of any preconceived notions or biases associated with AI when facial recognition would maybe not be recognized as such, students were not informed that the interview would focus on AI. As an example of an open question, students were asked to consider whether the facial recognition system could accurately identify the interviewer without glasses or recognize their twin.

After the students had thoroughly explained their understanding of facial recognition, the topic of AI was openly introduced, and they were briefly questioned about their general perception of AI. Subsequently, they were asked to reflect on whether, based on their own understanding, they would classify facial recognition as an AI application and to draw connections between their notions of AI and the specific characteristics of facial recognition technology. Finally, the students were engaged in a structured discussion about the functioning of ChatGPT. As with facial recognition, an example was presented at the outset, showcasing ChatGPT's ability to generate an email for applying to a student internship. Afterwards, the same questions were asked but adapted to ChatGPT with help of the interview guide.

4.3 Analysis and Code System

The recorded interviews were digitally transcribed and subsequently analyzed using the method of qualitative content analysis [13], aided by the software MAXQDA24. The systematic description of the verbalized concepts of the Participants is best achieved by qualitative content analysis, as it offers a reproducible and systematic approach [13]. In preparation for the analysis, we followed Kuckartz's suggestion and used main categories derived from the interview guide in the first phase of our analysis, aiming to inductively refine these categories [13]. Examples of these categories included themes such as "Anthropomorphism" and "Programming". Subsequently, initial coding of the transcribed interviews was conducted individually by two researchers. To enhance the quality of the coding, discrepancies between the coders were meticulously addressed through a second round of consensus coding [13]. Throughout this iterative process, the category system underwent continuous refinement, with main categories further subdivided and refined, and new categories being inductively incorporated. During the coding stage of the third interview, there was increased agreement among coders and minimal requirement for adjustments to the coding system. Consequently, the last two interviews were initially processed by a single coder. In the subsequent consensus coding step, both researchers also read the other interview and focused on codings that warranted discussion, based on the experience from the first three interviews. The category system developed delineates two key dimensions for analyzing students' conceptions on ML: *General Perspectives on ML* (table 1) and *System Perspective* (table 2).

General perspectives on ML. This group comprises three categories, each addressing different dimensions on the perception of ML systems. *Anthropomorphic Thinking* groups notions that the behavior of ML systems resembles human actions *(Human-like Acting)* or arises from processing mechanisms similar to human cognition *(Human-like Cognition). Algorithmic Thinking* refers to considerations of ML systems akin to traditional computer systems, whose problem-solving strategies can be algorithmically described. The codes categorized under *Data-based Thinking* elucidate the foundational role of data in the development of ML systems. They comprise notions that the functionality of these systems hinges on either the direct processing of unaltered data *(Raw Data Processing)* or the extraction of patterns and information from provided data *(Pattern-based Processing).*

System perspective. The codes in this group encompass central technical and structural characteristics of ML systems. The

Table 2: Codebook for Dimension "System Perspective"

Code	Definition	Example
Adaption/Learning		
Immutability	A software or explicit parts of a software do not change.	*No, [the instruction keywords] don't have to be learned [...] that's provided, that information.*
Development Stage	A software learns during its development.	*Actually, [self-driving cars] want to learn on their own. At first in the learning phase, so before they are used for actual traffic.*
Application Stage	A software learns while in use.	*Well, in a sense, the smartphone learns to know the owner, so to speak ... if you save your face there ... and then it knows a bit what they look like. I would say it learns the entire time.*
Data Aggregation		
User-provided	The user provides data to enable the functionality of a software.	*The data that was saved right at the beginning during the setup of the app.*
Developer-provided	The developer provides data to develop a software.	*In this case, Open AI provided [the data], meaning the developer of the AI.*
Autonomous Acquisition	A software purposefully collects new data to solve a problem.	*You can write questions or requests as a message and it [...] searches for information from the whole Internet, so to speak, and summarizes it for the respective topic that you want.*
Programmability		
Non-programmed	Parts of a software are explicitly not programmed.	*So the questions are not programmed by the programmer, I think. And the answers are definitely [not] either.*
Hard-coded	Specific features of a software are programmed entirely by developers.	*I believe that it was in fact programmed by the developers themselves so that it always knows that it is itself an AI.*
Foundational Code	Specific features of a software are programmed by developers and extended or enhanced by the software itself.	*It simply develops from the foundational code into a large intelligence that can then provide information well, safely and quickly.*
Architecture		
Data Storage	Data (images, texts ...) is saved and may be reused later.	*Well, I would think that the cell phone would then simply compare the two pictures - that is, what is currently in front of the camera with the picture already taken [during setup] - and then see if they match.*
Model	Processing rules are saved or changed.	*And with chat GPT, it's mainly characters, the letters and these rules that it memorizes.*
Data characteristics		
Variance of Data	Data exists in various forms (e.g. different nose shapes, eye colors, ...)	*Faces change a bit over the years and so that you don't have to reconfigure it every three months, there's a tolerance.*
Quality of Data	Data exists in varying quality (e.g. blurred images)	*So the tolerance arises [...] from flawed images [...] and from imperfect scanning of the photos.*
Amount of Data	Data exists in various quantities	*But also with the memory part, there is an extremely large amount of data stored there. There are millions of data, billions, quadrillions, in other words gigantic amounts of data.*
Forms of Data	Data exists in various forms (e.g. video, image, text)	*But with the right face recognition, it has a 3D model and the dots. With [this application here] it depends on the pixels.*
Confidence		
Exactness	The solution to a problem can be determined exactly.	*Well, because if [...] there's a little shadow from above or a light from the side, then it says: "No, there's light here, it's not 100%, so it's not you."*
Uncertainty	The result of software is subject to a degree of uncertainty.	*So the AI is supposed to function like a human and humans are not always [predictable]. So that means the AI will also have this randomness at some point. Especially one of this size.*

category *Adaptation/Learning* groups statements regarding the timing of potential adaptation processes *(Development Stage vs. Application Stage)* or contrary statements on *Immutability*. The codes within the category *Data Aggregation* relate to statements about the data sources used by ML systems to solve problems. They cover the concepts that ML systems receive necessary data either from users *(User-provided)*, developers *(Developer-provided)*, or through autonomous acquisition from various sources *(Autonomous Acquisition)*. The category *Programmability* examines perceptions regarding the extent to which the functionalities of ML systems can be implemented through methods of traditional programming. The boundaries of this notion encompass views that ML systems are either fully programmed like regular software *(Hard-coded)* or that certain aspects of them are not programmable at all *(Non-programmed)*. Within this spectrum of perceptions, learners may believe that developers provide a foundational codebase that the system autonomously adjusts and extends *(Foundational Code)*. *Architecture* directs attention to the structure of ML systems: *Data Storage* refers to the assumption that ML systems store data such as texts or images and retrieve them as required later on. Conversely, *Model* emphasizes the idea that ML systems store processing rules derived from data and may adjust them later if needed. As described in Section 2, the functionality and applicability of ML systems are decisively influenced by the characteristics of the data used in their creation: They can vary in terms of quantity *(Amount of Data)* and exist in different formats *(Forms of Data)*, particularly differing in their quality *(Quality of Data)* and the manifestation of specific features *(Variance of Data)*. The category *Confidence* comprises notions regarding the reliability of results from ML systems. Since their decisions are based on statistical assumptions, they are inherently associated with a degree of uncertainty *(Uncertainty)*. A prevalent, contrasting notion is the assumption that ML systems are capable of precisely determining solutions to considered problems *(Exactness)*.

5 RESULTS

In terms of participants' familiarity with the selected applications, it was evident that all interviewees were acquainted with them. Additionally, ChatGPT was unanimously recognized as AI by all participants. All in all we assigned 638 codes across all interviews. During the consensus coding and the development of the code system, six misconceptions were identified by the two coders. Subsequently, both researchers independently reviewed and analyzed the corresponding codes. The resulting interpretations were then compiled again in a joint discussion. The six identified misconceptions are presented below, along with exemplary statements from the participants.

Programmed Behavior. Students mistakenly assumed that all behaviors or features are directly programmed by developers, overlooking the nature of ML, where decision rules emerge through model training, not explicit programming. This misunderstanding particularly affects the training step of the ML workflow, as well as aspects of data collection and data cleaning, since these may not be considered at all. Additionally, it impairs the step of problem identification, as understanding which problems can be directly implemented is crucial for deciding on the benefits or downsides of using ML.

This misconception is evident in statements suggesting that developers program the rules for facial recognition or the grammar in ChatGPT. It is reflected in codes from the *Programmability* group and the *Algorithmic Thinking* code. For example, Student A states, *"Grammar and spelling are fixed rules that have been determined by algorithms. Therefore, a [dictionary] can also be specified directly."* For facial recognition, Student C describes, *"There is a lot of programming involved, and we have to program it. For example, these buttons and how [the application] should respond to the face."* In its strongest form, this misconception may include the belief that the program cannot learn at all, but that its entire behavior is programmed, as Student C further elaborates: *"You program it like that. [...] That's why the app cannot learn."*

A novel finding, which we term "foundational code", reveals that students integrate classical programming with concepts of intelligence and learnability. They view some parts of a program as fixed, while others as developed by the AI. To illustrate, student A states that *"the grammar is an algorithm. It has nothing to do with intelligence. It's just the way it is, it's done that way. The content is determined by the intelligence."* An alternative interpretation is that AI requires programmed guidelines under which it operates. When asked how ChatGPT knows how to structure an email, Student A responded that the program needs a template with rules, because when the AI *"puts things together"* itself, it is prone to errors. He further explains that the rules for spelling and grammar must therefore be programmed. However, not all students share this view. For instance, Student E explicitly states, *"If a new language is added [...] then it must first learn how this language works."*

Exactness. According to this misconception, the outcomes of ML systems can be determined with absolute precision, which is contrary to the nature of ML. This contradicts the training step in the ML workflow, where a statistical model is fitted to the data. These models make predictions and decisions based on probabilities. Statements reflecting this misconception are found in the *Exactness* code. For instance, in face recognition, Student E states, *"No, there's light here, it's not 100%, so it's not you."* For ChatGPT, Student C says, *"I think it has different options [of answers] and chooses one of them. [...] I think it takes the very best one."* The notion varies: some think a single, certain solution exists, while others recognize multiple potential solutions but believe one can be identified as optimal. This view persists even when acknowledging uncertainty, as illustrated when Person A notes, *"So when I put my face into the app, it compares it with the pictures I took at the beginning. And then it works in 99 percent of cases. So it only doesn't work if there were impurities in the pictures I put in or if the camera is blocked."* This implies the notion that the software can compare two faces with exact precision and that only external factors, such as user error, cause inaccuracies. The examples also show that notions of exactness often coincide with ideas about programming. This is also reflected in the response of Student B who states, *"Maybe you can imagine it a bit like with Scratch with this if-then command. So 'if this and that,' if all the features match, then 'access allowed to the phone' or 'unlock'."* Lastly, it should be noted that students are aware that ML are subject to uncertainty. However, when considering statements in the *Uncertainty* codes, this is not always attributed to the inherent architecture of the ML-models but rather to the complexity of ML problems, which precludes finding an optimal solution. Student A explains, *"[ChatGPT] has to somehow figure out how to structure it. What is the content? It has to generate all that anew. And it's unlikely that exactly the same thing that it has already generated will be generated again."* Notions of exactness are almost exclusively found in face recognition. ChatGPT is more often associated with uncertainty or randomness by the respondents. Student D summarizes this by saying, *"AI is supposed to function like a human, and humans are also random. So that means the AI will also have this randomness at some point."*

Data Storage. In this misunderstanding, students erroneously assume that raw data utilized in learning is permanently stored and later reused. For example, they believe that facial recognition software stores images from the setup phase to be used for comparison during the unlocking process. In reality, during the *Training* step of the ML workflow, a model is generated from the training data, which approximates these data. The training data are relevant only before the training step; afterwards, the generated model is used (steps *Evaluation* and *Application*).

This misconception is reflected in the codes *Data Storage* and *Raw Data Processing*. Students believe that data is saved in various forms and later reused. For instance, Student C states, *"And it simply stores your face on the phone, so it has it in the memory card."* Student B, referring to ChatGPT, responds, *"Then it would look in the saved things that it learned and stored itself and then search further on* *the internet."* The type of data presumed to be stored varies. Thus, original data is assumed to be saved, as Student C replies when asked what ChatGPT stores: *"All the questions that were asked [and the answers]."* Additionally, information derived from the data is believed to be stored such as face features or keywords. Student A says, *"[Images of the face] are stored, and some features of the face are scanned and stored and remembered."* Lastly, the storage of rules is also mentioned (Student D: *"It develops rules from what it writes and from the people who respond. [...] It also stores that."*). The storage of derived data and rules cannot be directly classified as a misconception as this is essentially what occurs during the training step, even if the "storage" in the form of, for example, weights in neural networks is no longer directly understandable to developers. However, the storage of raw data for the ML training process does not occur. The use of raw data is highlighted in the *Raw Data Processing* code. For example, Student B says, *"The phone simply compares the two pictures – that is, the one currently in front of the front camera – with the already taken picture from the recognition and then checks if they match."* Regarding image generation, he further states the application selects various individual images and skillfully combines them. The autonomous search for data is further explored in the misconception *Autonomous Data Acquisition*.

Continuous Learning. This misconception involves the belief that an application continuously learns and improves during use, as succinctly expressed by Student D: *"I would say it learns all the time."* This view conflicts with the ML workflow, which clearly distinguishes between training and application phases. During the training phase, ML models adjust structures and weights through an iterative process. After training, the model is deployed for use but does not adapt or "learn" further on its own. While data collected during use can improve the model, this requires a separate training phase. As Heuer et al. argue, although ML systems can learn during use through a process known as "online learning," such instances are rare and typically not relevant to the applications or algorithms students encounter daily or in educational settings [7]. Misconceptions about continuous improvement during the application phase are common, particularly when paired with the code *Autonomous Acquisition*. For example, Student B stated, *"And then at some point it will have understood, perhaps, how it works with forming sentences, because it had these examples and models on the Internet of how to do the whole thing when it was looking for information."*

Many respondents believe that ChatGPT continuously learns, but their understanding of how this learning occurs varies. Student B believes learning happens through constant observation of the environment. He states, *"Every time it unlocks, it memorizes the face a bit better, finds new features to recognize, and compares it faster and better, thus learning and improving its performance."* He associates this with the application's ability to reflect. Similarly, Student E explains that an AI *"can think independently and ask questions, thus learning on its own."* The autonomy of the application is frequently emphasized. Student D asserts, *"So it has to learn independently."* Student D attended one of the ML workshops, where the distinction between the training and application phases was specifically addressed. However, this did not dispel his misconception; instead, he integrated it into his existing mental model. In the interview, he says, *"The AI is actually constantly in the learning phase,"* indicating the persistence of the misconception. Finally, Student A describes learning as a result of action and reaction: *"Continuous development based on decisions, action, and reaction."* He adds that through his interactions, he has contributed to ChatGPT's learning process. Learning through user interaction will be examined in more detail in the next section.

User-trained Model. With this misconception it is assumed that ML models are generated using the user's data during application, similar to the continuous learning misconception. Both stem from a misunderstanding of the ML workflow, where models are actually trained in a dedicated training phase by developers, not during user interaction. This misconception is reflected in the codes related to the application phase, often combined with the code *User-provided (Data)*.

For the two applications, the misconception manifests differently. For ChatGPT, the continuous learning misconception persists, with user interaction seen as an additional learning opportunity. Student A expresses this idea, saying *"I have actually contributed a tiny bit to the learning. I tried to recreate a relatively old text-based game somehow. [...] And then I said no, that sounds wrong, I actually start there. It then saw, 'Oh yes, that's how it starts'."* Student E highlights another form of user feedback: *"Here [points to the generated response] there are buttons Thumbs up, Thumbs down. [...] You can indicate whether it is correct or incorrect, or add notes."* The first example refers to ChatGPT's ability to adapt to conversational contexts, while the second points to the built-in feedback feature. In both cases, structural learning of the application, as assumed by the students, does not actually occur.

The second variant of this misconception is that the application learns directly from the data provided by the user, with the user training the model themselves instead of the developers. This is particularly evident in facial recognition. Student B explains: *"Well, in a sense, the phone kind of learns to recognize the owner when you store your face and it sees what you look like."* Student E describes this process more concretely: *"Yes, you take these pictures. So, recordings of the face [holds his hand in front as if holding a phone and moves his head in all directions to 'scan' his face] [...] Now you are training the AI. This is how the AI gets it."*

Autonomous Data Acquisition. This misconception suggests that the application is capable of autonomously searching and selecting data based on its own metrics and using it for learning. However, according to the ML workflow, data is gathered in the development phase. In reality, it is the responsibility of developers to collect, clean, and prepare data for training ML models. Even in applications not covered by this study, such as reinforcement learning (RL), where an agent might interact autonomously with its environment, developers still determine which data is collected and how it is processed.

This misconception is not limited to the concept of learning. For instance, Student C asserts, *"I believe that when we enter the question, it just asks Google and takes the text from there and simply writes it."* This assertion implies that the participant is aware of the significance of data, yet sees the application's role as merely finding data. Building on this, Student B describes the possibility that the application uses internet searches to verify its own statements: *"ChatGPT could, for example, check the internet before it writes something to see if it is okay."* He later adds the capability of learning, stating, *"And then it will have understood at some point, maybe, how it works* *with forming sentences, because it had these templates and models on the internet, how to do it all, when it was searching for information."* This illustrates how closely this misconception is linked to the idea of continuous learning.

This misconception emerged primarily with ChatGPT and once when Student A described Google Lens functionality and its dependence on internet searches. Notably, participants did not consider the possibility of facial recognition autonomously capturing images. This highlights the connection between this misconception and a system's internet access and underscores the distinct perceptions between the two applications. Some respondents believe ChatGPT uses internet data not only to complete tasks but also for independent learning and self-reflection. Student B says, *"It could be possible in the near future, if it finds reports on the internet about why something is the way it is and then adapts and learns and asks itself this question."* Additionally, the misconception of data storage appears here, such as in Student C's statement that ChatGPT searches its internal memory, where all questions and the generated answers are stored.

6 DISCUSSION

When analysing for misconceptions anthropomorphic statements were of particular interest. Mertala et al. classify these as misconceptions [17]. Mühling & Große-Bölting argue that the use of anthropomorphic language does not necessarily indicate a lack of understanding, but may instead reflect a lack of technical vocabulary [20]. This is also supported by the fact that AI experts also use anthropomorphic language [25]. In our coding system, we distinguish between *User-like Acting* and *Human-Like Cognition* within the "Anthropomorphic Thinking" category. While the former is clearly linked to misconceptions such as *Autonomous Data Acquisition*, the latter's classification is not as straightforward. Consider Student B, who described the program as actively improving itself with questions like, "What did I just do?", "How can I do this better?", and "How can I apply this somewhere else to make the other thing better?". The first two thoughts are compatible with the training step in the ML-Workflow. However, whether the last statement is correct depends on the context and the type of application. In our view, relying on human cognition as a means of communication does not directly express a misconception. However, anthropomorphic language may indicate a misconception when attempting to explain the functioning of a system in terms of human behavior.

Our findings regarding misconceptions in ML education have significant implications for the design of learning activities and curricula. Currently, various approaches to teaching ML have been proposed with the central question being whether and how to teach learners about the inner workings of ML models and algorithms [38]. Our findings contribute to a more nuanced discussion on this topic. To illustrate this, we will examine the popular unplugged activity "Hexapawn", used by Mühling & Große-Bölting [20] in their study.

Mühling & Große-Bölting argue "it may not be necessary to 'open the black box' in a K12 setting in order to counter hard-to-erase misconceptions of 'real intelligence' happening in machines", as their study showed a positive shift in students' understanding of "learning" in ML systems through the activity *Hexapawn.* In this activity, students engage in a simple 3×3 chess game against a set of matchboxes. After each game, the matchboxes are adjusted to improve their performance. This interaction risks reinforcing misconceptions rather than dispelling them. The matchboxes' improvement after each game can reinforce the idea that the AI continuously learns (*Continuous Learning*). Additionally, the system learns through user interaction, potentially reinforcing the *User-trained Model* misconception. Therefore, to effectively use this game in a learning setting, the system's training phase must be clearly distinguished from the application phase. Learners must understand that the unplugged activity simulates the training phase. This is particularly crucial in RL, where many learning activities explain the RL concepts through interaction with the learner. Another hurdle in RL is, that the agent collects data independently, potentially reinforcing the *Autonomous Data Acquisition* misconception. It is important to emphasize that even in RL, developers define data selection and processing. Accordingly, students should be able to define and modify data aggregation parameters to counter this misconception, as is also proposed by Touretzky et al. [34]. Finally, we want to emphasize that this activity is by no means unsuitable for conveying ML concepts. It is also perfectly adequate to address the probably most critical but also easiest to change misconception, *Programmed Behavior*, as it largely stems from a lack of understanding of the process itself. It is, therefore, essential to emphasize that an understanding of potential misconceptions is vital in order to utilize specific teaching concepts effectively. Nevertheless, we support the view that ML black boxes need to be opened, as our analysis shows that even after a seemingly successful intervention, significant misconceptions might still be present. In summary, the results presented here contain a series of new insights. Through the use of semi-structured interviews, we were able to identify and characterize six misconceptions in detail. This supports the development and evaluation of appropriate learning interventions.

7 LIMITATIONS AND OUTLOOK

The interview study is subject to the typical limitations of qualitative research. The results presented here cannot be generalized and primarily serve to develop theories and hypotheses. In addition, the sample size of five people is relatively small and heterogeneous. Although we have only presented conceptions that were evident across several respondents, it is unclear how these vary across the full range of possible prior knowledge and age and how representative they are. A further limitation of this study is that the interview guide was not pilot tested prior to the main study. Consequently, the interview procedure could not be adapted before the main study, for example by changing the order of questions.

The results of this interview study can be explored in more depth, for example by investigating the extent to which mental models are dependent on the respective application. Moreover, concepts which are valid for the entire field of ML can be derived from the presented misconceptions and serve as structuring aids for the construction of appropriate mental models. Finally, the results presented here can also be verified quantitatively by means of a questionnaire in which the misconceptions serve as distractors and with which learning activities can be evaluated.

REFERENCES

[1] Arne Bewersdorff, Xiaoming Zhai, Jessica Roberts, and Claudia Nerdel. 2023. Myths, Mis- and Preconceptions of Artificial Intelligence: A Review of the Literature. *Computers and Education: Artificial Intelligence* 4 (Jan. 2023), 100143. https://doi.org/10.1016/j.caeai.2023.100143

[2] D. Ciresan, U. Meier, and J. Schmidhuber. 2012. Multi-Column Deep Neural Networks for Image Classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Providence, RI, 3642–3649. https://doi.org/10.1109/CVPR.2012.6248110

[3] Ignacio Evangelista, Germán Blesio, and Emanuel Benatti. 2018. Why Are We Not Teaching Machine Learning at High School? A Proposal. In *2018 World Engineering Education Forum - Global Engineering Deans Council (WEEF-GEDC)*, Institute of Electrical and Electronics Engineers (IEEE) (Ed.). Institute of Electrical and Electronics Engineers (IEEE), Albuquerque, New Mexico, USA, 1–6. https://doi.org/10.1109/WEEF-GEDC.2018.8629750

[4] Creso Franco, Henrique Lins de Barros, Dominique Colinvaux, Sonia Krapas, Glória Queiroz, and Fátima Alves. 1999. From Scientists' and Inventors' Minds to Some Scientific and Technological Products: Relationships between Theories, Models, Mental Models and Conceptions. *International Journal of Science Education* 21, 3 (March 1999), 277–291. https://doi.org/10.1080/095006999290705

[5] D. Gentner. 2001. Mental Models, Psychology Of. In *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier, Amsterdam, 9683–9687. https://doi.org/10.1016/B0-08-043076-7/01487-X

[6] Ileana Maria Greca and Marco Antonio Moreira. 2000. Mental Models, Conceptual Models, and Modelling. *International Journal of Science Education* 22, 1 (Jan. 2000), 1–11. https://doi.org/10.1080/095006900289976

[7] Hendrik Heuer, Juliane Jarke, and Andreas Breiter. 2021. Machine Learning in Tutorials – Universal Applicability, Underinformed Application, and Other Misconceptions. *Big Data & Society* 8, 1 (Jan. 2021), 1–13. https://doi.org/10.1177/20539517211017593

[8] Tom Hitron, Yoav Orlev, Iddo Wald, Ariel Shamir, Hadas Erel, and Oren Zuckerman. 2019. Can Children Understand Machine Learning Concepts? The Effect of Uncovering Black Boxes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–11.

[9] Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 328–339. https://doi.org/10.18653/v1/P18-1031

[10] Natalie A. Jones, Helen Ross, Timothy Lynam, and Pascal Perez. 2014. Eliciting Mental Models: A Comparison of Interview Procedures in the Context of Natural Resource Management. *Ecology and Society* 19, 1 (2014), 7 pages. jstor:26269480

[11] Keunjae Kim, Kyungbin Kwon, Anne Ottenbreit-Leftwich, Haesol Bae, and Krista Glazewski. 2023. Exploring Middle School Students' Common Naive Conceptions of Artificial Intelligence Concepts, and the Evolution of These Ideas. *Education and Information Technologies* 28, 8 (Jan. 2023), 9827–9854. https://doi.org/10.1007/s10639-023-11600-3

[12] Moritz Kreinsen and Sandra Schulz. 2021. Students' Conceptions of Artificial Intelligence. In *The 16th Workshop in Primary and Secondary Computing Education*. ACM, Virtual Event Germany, 1–2. https://doi.org/10.1145/3481312.3481328

[13] Udo Kuckartz and Stefan Rädiker. 2023. *Qualitative Content Analysis: Methods, Practice and Software* (second ed.). SAGE Publications, Thousand Oaks.

[14] Irene Lee, Safinah Ali, Helen Zhang, Daniella DiPaola, and Cynthia Breazeal. 2021. Developing Middle School Students' AI Literacy. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (SIGCSE '21)*. Association for Computing Machinery, New York, NY, USA, 191–197. https://doi.org/10.1145/3408877.3432513

[15] Duri Long and Brian Magerko. 2020. What Is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–16. https://doi.org/10.1145/3313831.3376727

[16] Erik Marx, Thiemo Leonhardt, and Nadine Bergner. 2023. Secondary School Students' Mental Models and Attitudes Regarding Artificial Intelligence - A Scoping Review. *Computers and Education: Artificial Intelligence* 5 (2023), 100169. https://doi.org/10.1016/j.caeai.2023.100169

[17] Pekka Mertala and Janne Fagerlund. 2024. Finnish 5th and 6th Graders' Misconceptions about Artificial Intelligence. *International Journal of Child-Computer Interaction* 39 (March 2024), 100630. https://doi.org/10.1016/j.ijcci.2023.100630

[18] Pekka Mertala, Janne Fagerlund, and Oscar Calderon. 2022. Finnish 5th and 6th Grade Students' Pre-Instructional Conceptions of Artificial Intelligence (AI) and Their Implications for AI Literacy Education. *Computers and Education: Artificial Intelligence* 3 (Jan. 2022), 100095. https://doi.org/10.1016/j.caeai.2022.100095

[19] Tilman Michaeli, Ralf Romeike, and Stefan Seegerer. 2023. What Students Can Learn About Artificial Intelligence – Recommendations for K-12 Computing Education. In *Towards a Collaborative Society Through Creative Learning*, Therese Keane, Cathy Lewin, Torsten Brinda, and Rosa Bottino (Eds.), Vol. 685. Springer Nature Switzerland, Cham, 196–208. https://doi.org/10.1007/978-3-031-43393-1_19

[20] Andreas Mühling and Gregor Große-Bölting. 2023. Novices' Conceptions of Machine Learning. *Computers and Education: Artificial Intelligence* 4, 100142 (2023), 1–11. https://doi.org/10.1016/j.caeai.2023.100142

[21] Sandra Nitz and Sabine Fechner. 2018. Mentale Modelle. In *Theorien in der naturwissenschaftsdidaktischen Forschung*, Dirk Krüger, Ilka Parchmann, and Horst Schecker (Eds.). Springer, Berlin, Heidelberg, 69–86. https://doi.org/10.1007/978-3-662-56320-5_5

[22] Donald A. Norman. 1983. Some Observations on Mental Models. In *Mental Models*, Dedre Gentner and Albert L. Stevens (Eds.). Psychology Press, New York, 15–22. https://doi.org/10.4324/9781315802725-5

[23] Leonid Perlovsky. 2011. Language and Cognition Interaction Neural Mechanisms. *Computational Intelligence and Neuroscience* 2011 (2011), 1–13. https://doi.org/10.1155/2011/454587

[24] David N Rapp. 2005. Mental Models: Theoretical Issues for Visualizations in Science Education. In *Visualization in Science Education*, John K. Gilbert (Ed.). Springer Netherlands, Dordrecht, 43–60. https://doi.org/10.1007/1-4020-3613-2_4

[25] Michael T. Rücker and Niels Pinkwart. 2016. Review and Discussion of Children's Conceptions of Computers. *Journal of Science Education and Technology* 25, 2 (April 2016), 274–283. https://doi.org/10.1007/s10956-015-9592-2

[26] Ismaila Temitayo Sanusi, Solomon Sunday Oyelere, Henriikka Vartiainen, Jarkko Suhonen, and Markku Tukiainen. 2023. Developing Middle School Students' Understanding of Machine Learning in an African School. *Computers and Education: Artificial Intelligence* 5 (Jan. 2023), 100155. https://doi.org/10.1016/j.caeai.2023.100155

[27] Marie-Monique Schaper, Mariana Aki Tamashiro, Rachel Charlotte Smith, Maarten Van Mechelen, and Ole Sejer Iversen. 2023. Five Design Recommendations for Teaching Teenagers' about Artificial Intelligence and Machine Learning. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. ACM, Chicago IL USA, 298–309. https://doi.org/10.1145/3585088.3589366

[28] Juha Sorva. 2013. Notional Machines and Introductory Programming Education. *ACM Transactions on Computing Education* 13, 2 (June 2013), 1–31. https://doi.org/10.1145/2483710.2483713

[29] Nancy Staggers and A. F. Norcio. 1993. Mental Models: Concepts for Human-Computer Interaction Research. *International Journal of Man-Machine Studies* 38, 4 (April 1993), 587–605. https://doi.org/10.1006/imms.1993.1028

[30] Elisabeth Sulmont, Elizabeth Patitsas, and Jeremy R. Cooperstock. 2019. Can You Teach Me To Machine Learn?. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. ACM, Minneapolis MN USA, 948–954. https://doi.org/10.1145/3287324.3287392

[31] Jessica Szczuka, Clara Strathmann, Natalia Szymczyk, Lina Mavrina, and Nicole Krämer. 2022. How Do Children Acquire Knowledge about Voice Assistants? A Longitudinal Field Study on Children's Knowledge about How Voice Assistants Store and Process Data. *International Journal of Child-Computer Interaction* 33 (Jan. 2022), 100460. https://doi.org/10.1016/j.ijcci.2022.100460

[32] Keith S. Taber. 2017. The Nature of Student Conceptions in Science. In *Science Education*, Keith S. Taber and Ben Akpan (Eds.). SensePublishers, Rotterdam, 119–131. https://doi.org/10.1007/978-94-6300-749-8_9

[33] Matti Tedre, Peter Denning, and Tapani Toivonen. 2021. CT 2.0. In *21st Koli Calling International Conference on Computing Education Research*. ACM, Joensuu Finland, 1–8. https://doi.org/10.1145/3488042.3488053

[34] David Touretzky, Christina Gardner-McCune, and Deborah Seehorn. 2022. Machine Learning and the Five Big Ideas in AI. *International Journal of Artificial Intelligence in Education* 33, 2 (Oct. 2022), 1–34. https://doi.org/10.1007/s40593-022-00314-1

[35] Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2023. *Natural Language Processing with Transformers: Building Language Applications with Hugging Face* (revised edition ed.). O'Reilly Media, Sebastopol, CA.

[36] Jessica Vandenberg and Bradford Mott. 2023. "AI Teaches Itself": Exploring Young Learners' Perspectives on Artificial Intelligence for Instrument Development. In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*. ACM, Turku Finland, 485–490. https://doi.org/10.1145/3587102.3588778

[37] Henriikka Vartiainen, Tapani Toivonen, Ilkka Jormanainen, Juho Kahila, Matti Tedre, and Teemu Valtonen. 2021. Machine Learning for Middle Schoolers: Learning through Data-Driven Design. *International Journal of Child-Computer Interaction* 29 (Sept. 2021), 100281. https://doi.org/10.1016/j.ijcci.2021.100281

[38] Jane Waite, Ethel Tshukudu, Veronica Cucuiat, Robert Whyte, and Sue Sentance. 2023. Towards a Framework for Learning Content Analysis in K-12 AI/ML Education. In *2023 IEEE Frontiers in Education Conference (FIE)*. IEEE, College Station, TX, USA, 1–5. https://doi.org/10.1109/FIE58773.2023.10343368

[39] Michelle Zimmerman. 2018. *Teaching AI: Exploring New Frontiers for Learning*. Internation Society for Technology in Education, Portland, Oregon.