Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

What you need to know about a learning robot: Identifying the enabling architecture of complex systems

Helen Beierling, Phillip Richter, Mara Brandt, Lutz Terfloth, Carsten Schulte, Heiko Wersing, Anna-Lisa Vollmer

2024

Preprint:

Copyright Elsevier 2024. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/

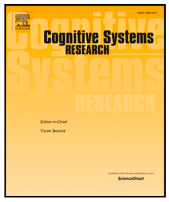
Cognitive Systems Research 88 (2024) 101286



Contents lists available at ScienceDirect

Cognitive Systems Research

journal homepage: www.elsevier.com/locate/cogsys



What you need to know about a learning robot: Identifying the enabling architecture of complex systems

Helen Beierling [a,*], Phillip Richter [a], Mara Brandt [a], Lutz Terfloth [b], Carsten Schulte [b], Heiko Wersing [c], Anna-Lisa Vollmer [a]

[a] Bielefeld University, Medical School EWL, Interactive Robotics, Universitätsstraße 25, 33615 Bielefeld, Germany
[b] Paderborn University, Faculty of Computer Science, Electrical Engineering and Mathematics, Didactics of Informatics, Warburger Straße 100, 33098 Paderborn, Germany
[c] Honda Research Institute Europe GmbH, Carl-Legien-Str. 30, 63073 Offenbach am Main, Germany

ARTICLE INFO

Action editor: Mog Stapleton

Keywords:
Robotics
HRI
Explainability
Didactics
Didactic reconstruction

ABSTRACT

Nowadays we deal with robots and AI more and more in our everyday life. However, their behavior is not always apparent to most lay users, especially in error situations. This can lead to misconceptions about the behavior of the technologies being used. This in turn can lead to misuse and rejection by users. Explanation, for example through transparency, can address these misconceptions. However, explaining the entire software or hardware would be confusing and overwhelming for users. Therefore, this paper focuses on the 'enabling' architecture. It describes those aspects of a robotic system that may need to be explained to enable someone to use the technology effectively. Furthermore, this paper deals with the 'explanandum', i.e. the corresponding misunderstandings or missing concepts of the enabling architecture that need to be clarified. Thus, we have developed and are presenting an approach to determine the 'enabling' architecture and the resulting 'explanandum' of complex technologies.

1. Introduction

In our everyday life we come more and more into contact with robots. Examples include PARO Robots, Inc. (2014) in the care sector, Intuitive Surgical, Inc (2023) in the operating room, temi (temi USA inc., 2023) in hospitals, Pepper (Aldebaran, 2022) in retail and iRobot vacuum cleaners (iRobot Corporation, 2022) in our own homes. These are just a few examples of robots that are used in environments where the tasks to be performed can hardly be learned in advance, as they depend heavily on the user's preferences.

As robots increasingly take over tasks in our close environment and assist us with everyday tasks, it is essential that they be able to adapt to our personal preferences. This requires users to be able to effectively train robots during human–robot interaction. A common method for this is interactive robot learning, which incorporates user feedback into the robot's learning process, either by directly using human feedback as a reward or through demonstrations (Akalin & Loutfi, 2021; Argall, Chernova, Veloso, & Browning, 2009; Thomaz & Breazeal, 2008; Thomaz et al., 2006).

However, such methods can suffer if users have misconceptions in their mental model of the robot's inner workings (Hindemith, Vollmer, Göpfert, Wiebel-Herboth, & Wrede, 2020). For example, if the robot has human features such as eyes and is taught by demonstration when the robot does not know what the user wants or expects, the user tends to do additional movements in order for the robot to follow them better (Vollmer et al., 2014). These movements are then potentially learned as part of the target movement, which slows down the learning process. Another example of these misconceptions in learning based on direct reward is the misuse of reward for motivation, which outweighs the rated behaviors (Thomaz & Breazeal, 2008). This leads to misunderstandings and ineffective learning processes. One approach to resolving these misunderstandings is to explain the underlying architecture of the robot (Thomaz & Breazeal, 2008). However, the internal structures of the robot are often complex, and users are not willing to invest much time in training (Hindemith et al., 2020). Furthermore, explainability approaches would profit from a reduced search space of relevant aspects for explanations. Therefore, a structured procedure is needed to identify those parts of the architecture that, when understood, enable the user to train the robot. We call these parts the 'enabling architecture' of a system. On this basis, the contribution of this paper focuses on three key aspects:

* Corresponding author.
E-mail address: helen.beierling@uni-bielefeld.de (H. Beierling).

https://doi.org/10.1016/j.cogsys.2024.101286
Received 30 July 2023; Received in revised form 12 July 2024; Accepted 16 September 2024
Available online 28 September 2024
1389-0417/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

- 1. Interdisciplinary integration of didactics: We have incorporated didactic methods into our approach and used the methodology of didactic reconstruction to identify key aspects of the enabling architecture.
- 2. Transferring didactic methods to HRI: We transfer didactic methods to the field of HRI to explore an approach for identifying the enabling architecture. If XAI has knowledge of the enabling architecture, it is able to provide explanations that enable the explainer to train the system.
- 3. Implementation of a use case as a basis for future research: Our third contribution is the realization of a use case which serves as a blueprint for future, more detailed research. This use case study demonstrates our approaches in an exemplary practical implementation in an HRI teaching scenario, and provides a practical application in the field of XAI.

It should be noted, however, that we do not provide ready-made explanations but rather an approach to systematically find the important areas for explanations.

Our paper begins with a review of the relevant literature on explainability in human–robot interaction and the development of mental models. We then discuss the concept and application of didactic reconstruction, distinguishing between classical reconstruction, which includes the factual situation, reduction, the lay user's mental model, and its synthesis. We further explore the adaptation of this methodology to the field of Explainable Artificial Intelligence (XAI), detailing the process from factual situation and reduction to synthesis, aligned with the learner's perspective. A case study is presented to illustrate these concepts in practice, focusing on the factual situation, the user's mental model, and the subsequent synthesis. Finally, this paper concludes with a discussion of the findings, implications for future research, and potential directions for further exploration in this area.

2. Related work

This paper explores misconceptions in human mental models within the context of human–robot interaction (HRI) and their subsequent effects on explainable artificial intelligence (XAI). Human mental models are cognitive representations constructed from individual experiences and environmental interactions, shaping one's perception of themselves and their surroundings (Gentner & Stevens, 2014). In robotics, these mental constructs extend to technical systems, particularly robots. Factors such as a robot's social cues, human-like movements, and anthropomorphism significantly influence human mental models of robots. Media portrayals further contribute to misunderstandings about a robot's capabilities (Bruckenberger et al., 2013; Kiesler, 2005).

In the domain of HRI, the forming of mental models has significant implications. Research shows that people develop distinct mental models for different types of robots (Lee, Lau, Kiesler, & Chiu, 2005), which subsequently influence their behavior towards these robots (Mumm & Mutlu, 2011). The increasingly human-like design of robots often leads to assumptions about similar ability and learning methods, resulting in unrealistic expectations of their functioning and behavior. This discrepancy between expectation and reality can result in frustration and poor interactions in HRI (Berzuk & Young, 2023; Malle, Fischer, Young, Moon, & Collins, 2020).

The investigation of human mental models of robots is a complex task that has been investigated in various studies (Kiesler & Goetz, 2002; Lee et al., 2005; Powers & Kiesler, 2006). Traditional methods such as direct querying (Brooks, 2019) pose risks of intrusion and inaccuracies due to lack of expertise. However, recent research suggests that low-level actions by humans towards robots can provide insights into these mental models (Brooks & Szafir, 2019; Pérez-D'Arpino & Shah, 2015).

The concept of interactive task learning (ITL) in HRI is a further key aspect. ITL refers to an agent learning from a human trough immediate feedback while performing a task (Kirk & Laird, 2014; Ramaraj, Ortiz, & Mohan, 2021). There are various approaches within ITL, including teaching by demonstration (Argall et al., 2009), learning by advice (Kuhlmann, Stone, Mooney, & Shavlik, 2004), and learning by reward-based feedback (Knox & Stone, 2009; Thomaz et al., 2006). However, for some interactive learning algorithms, research has shown how humans adapt their teaching behavior as they develop a mental model of a robot's capabilities, limitations, and how it learns (Hindemith et al., 2020; Thomaz & Breazeal, 2008; Vollmer et al., 2009).

To improve the human mental mode, we aim to identify architectures that enable lay users to use the system correctly. Therefor, we introduce a novel approach in which didactic reconstruction is used to help identify the enabling architecture that differs from existing methodologies. In general, the concept of didactic reconstruction is unexplored in the context of XAI. Previous research has effectively utilized didactic reduction in the setting of metaphorical narratives to simplify AI models for domain experts (Jentner et al., 2018). Their work, which aims to create understandable and confidence-building narratives, uses didactic reduction as a tool to abstract complex AI models into more accessible forms. Our research, however, diverges by targeting lay users and applying didactic reconstruction instead of reduction. This shift in both target audience and methodological approach characterizes our work as a unique and necessary extension in the field of XAI, with the goal of bridging the understanding gap for non-experts.

3. Didactic reconstruction

The field of Didactics provides structured methods to simplify complex topics, structure them and present the result to lay users of the topic.

3.1. Classical reconstruction

The didactic reconstruction was designed by Kattmann et al. (1997) in 1997 and has been developed further since then. It addresses the problem of conveying complex information with the lay users' knowledge in mind. Instead of merely reducing complex information for the lay user, it starts with their prior knowledge and builds upon it. This is important, since the human mental model is also formed on a base domain of prior knowledge (Staggers & Norcio, 1993). The didactic reconstruction considers three aspects: The factual situation, the pupils' mental models and the form of presentation used to convey the information. The factual situation refers to the information to be taught, which will be presented along with the description of the pupils' mental model in the following sections. However, the form of presentation that best conveys the information and fits the target audience is significantly influenced by the explanandum, determined by the synthesis the factual situation and the pupils' mental models. It allows for extending the pupils' conceptions and adapt and classify them within the factual situation. Thus, the factual situation and the pupils' mental model must be identified and synthesized first. Hence, we focus on three aspects to derive the relevant concepts for an explanation: the factual situation, the pupils' mental models. The form of presentation is future work for this paper. In our approach, synthesis is explicitly added, which in classical reconstruction is implicitly part of the presentation form, as the content presented there also needs to be synthesized.

3.1.1. Factual situation and reduction

First, the approach takes an overview of the factual situation of the selected information to teach (Duit, Gropengiesser, Kattmann, Komorek, & Parchmann, 2012). This is accomplished through a qualitative content analysis of state-of-the-art resources, such as textbooks or academic papers. However, the challenge with this literature is its dense presentation of information, making it not directly usable for pupils.

Thus, it is not directly usable for pupils. Furthermore, this approach is also not valuable when only a single source exists for the topic, such as the development of a specific algorithm or method. However, another part of the overview of the factual situation is to access experts' knowledge on the topic. Therefore, the analysis of the experts' ideas is generally helpful, but especially in a situation where the textual information is limited.

Expert. An expert in this context is someone who provides knowledge. This can be an expert with procedural or conceptual knowledge. An expert also does not necessarily need to have a deep knowledge about the topic. For example, a university student can be an expert for first-graders. However, in comparison to a professor, the student would assume the role of the pupil.

Reduction. After identifying the factual situation, it is necessary to reduce this information. The reduction also eases the synthesis of the factual overview of the expert's and the pupils' mental models since the overlap of aspects, and especially the missing concepts, can be systematically evaluated. The reduction process is based on didactic reduction, which is an older approach compared to didactic reconstruction and focuses solely on reducing information. It provides more refined reduction techniques than the didactic reconstruction.

The didactic reduction was developed from 1958 to 1991 through the collaborative efforts of various authors who complimented and continued each other's work. Central figures in this development include Dietrich Hering and Gustav Grüner. For this paper, we used the more recent work by Lehner (2020) as the basis for our reduction approach. Didactic reduction occurs when extensive and complex issues are processed to make them clear and comprehensible for the learners, as described on p.9 of Lehner (2020). In contrast to the didactic reconstruction, didactic reduction is based on the factual situation and reduces it without considering the pupils from the beginning. Regarding the application of didactic reduction, there are different types and focuses. Firstly, there is a distinction between horizontal and vertical reduction. Horizontal reduction aims to reduce the form of presentation. This can be done by visualizations, for example. Vertical reductions intend to reduce the content that is presented. The explanation scope of the factual situation might not very large, but its complexity is. Therefore, we aim to perform a vertical reduction of the explanation. Another distinction is made in the didactic reduction between quantitative and qualitative reduction. Quantitative reduction reduces the amount of information. Qualitative reduction reduces the complexity of the content. Thus, we focus on vertical qualitative reduction. However, most of the suggested reduction techniques for vertical qualitative reduction are based on everyday examples of the phenomenon to be explained. We also focus on systematic methods that are as minimally biased and are reproducible, since the approach should also apply to other architectures. One option that fulfills these criteria is the systematic creation of a concept map. This approach is suggested by Lehner (2020) on p.155 as a method for didactic reduction, and the didactic reconstruction also supports it. Although we focus on conceptual knowledge in this paper, it is also possible to represent procedural knowledge with concept maps. They offer a good approach to depict knowledge in general as for example explored by Sims-Knight et al. (2004), Passmore (1999), and Kinchin, Möllits, and Reiska (2019).

3.1.2. Pupil

A pupil is someone who seeks knowledge, whether practical or theoretical. In the context of didactic reconstruction, a pupil is typically understood in a classical school or university setting. However, the presented approach does not restrict the domain of pupils. Therefore, anyone with missing knowledge or existing misunderstandings qualifies as a pupil.

Pupils' mental models. Besides the factual situation, the prior experience and knowledge of the target audience needs to be accessed. This can be done by interviewing the pupils regarding their knowledge of the topic as described by Niebert and Gropengiesser (2013). From the interviews, the pupils' mental models can be assessed. Johnson-Laird (1983) defines mental models as cognitive frameworks that guide how we perceive, understand, and interact with the world. These models simplify complex information and help us make decisions and predict outcomes based on our experience and knowledge. When evaluating mental models, it is crucial to examine them in three phases: before, during, and after interaction with the subject to be explained. This is especially interesting since during the interaction, the initial mental model changes continuously, as discussed by Thomaz and Breazeal (2008). Afterward, the obtained model needs to be analyzed, structured and generalized. The same reduction technique can be applied here as for the experts. This simplifies the synthesis of the two models, as they will be in the same format.

3.2. Synthesis

Given the assessed factual situation and the pupils' mental models, they both need to be synthesized into a teaching concept. This concept should consider the gaps in knowledge that were revealed by the analysis of the pupils' mental models. Those gaps should then be filled by the results of the analysis of the factual situation.

3.3. Transfer of reconstruction to XAI

Another contribution of this paper is the transfer of the presented didactic reconstruction to the field of XAI and explainable robots. We will structure the presentation of this transfer the same way as in the initial presentation of the didactic reconstruction. First, we will first discuss how the factual situation can be assessed in a XAI context. Then, we will present the mental models of pupils in this context. Finally, we will describe how synthesis can be achieved in an XAI setting.

In didactic literature, the term 'pupil' is commonly used. We aim to transfer these concepts into the field of HRI. Therefore, what is traditionally referred to as a 'pupil' in didactics, we will now call a 'lay user' in the context of HRI.

3.3.1. Factual situation and reduction

In the case of XAI, the explanation is focused on one algorithm that is to be explained. The available textbooks and scientific papers might be limited for some algorithms. Furthermore, the corresponding papers can be very complex and assume a significant amount of prior knowledge. This is particularly true for novel AI approaches, which are developed rapidly and in diverse ways, making textbook publication difficult. Therefore, we suggest using the method of expert interviews, which is an alternative for the assessment of the factual situation as presented in Section 3.1.1. For this, we propose expert interviews that query explanations of the approach addressed to lay users. Those explanations are already partially reduced by the explainer and should focus on the essential parts from the expert's perspective.

Expert. In the context of this paper, experts in the field of XAI include those who have contributed to the paper, textbook, related work, or who have practical experience with the approach. However, the approach can also be applied when interviewing experts outside academia, for example with practical knowledge only. Using the interviews as a base is also beneficial when using an expert with practical knowledge, since there also might be a lack of textbooks and papers in their field of expertise.

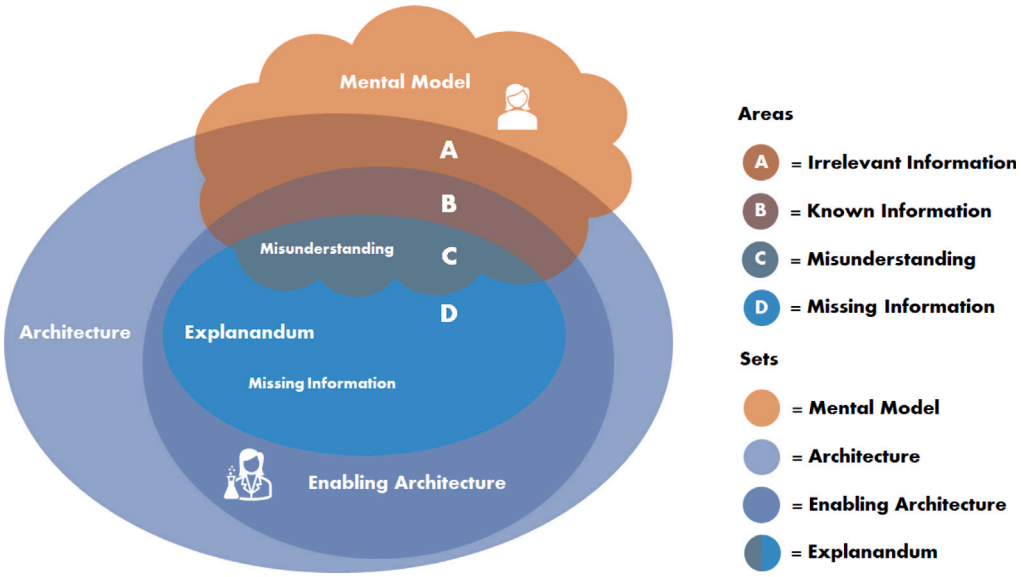


Fig. 1. A representation of areas of information regarding the architecture, the lay users and experts.

Reduction. One method of simplification suggested by didactic reduction is the creation of concept maps, as described in Section 3.1.1. It falls under the visual methods of reduction. This method is predestined for the domain of XAI since it can be done in a systematic and computable way. Since the concept maps are based on expert interviews, it is also essential to reduce them. The literature for example by Lehner (2020), Mietzel (2007), and Bromme and Rambow (1998) suggest, that the raw expert explanation are too complex for lay users, even though they have lay users as target audience. For this, we defied an analysis to systematically reduce and prepare the experts interviews. This method ensures that the creation of the concept map, based on systematic analysis, is as reproducible as possible. Thus, the reduction transferred to XAI is divided into four steps. First, identify experts, second, interview experts regarding the topic, third, prepare and reduce the interview, and fourth, create the concept maps. In the following, we will describe the four steps of the analysis:

1. **Preprocessing** The experts' explanations needs to be preprocessed by removing all function words.
2. **Quantitative Analysis** The emerging concepts and interactions must be analyzed quantitatively by determining their frequency. This included the number of times each expert mentions a concept. It also needs to be noted how frequently all experts mentioned the concepts and interactions. Both frequency are required, since the first might suggest that one concept was used very often, even though this might have been only one expert that used the concept in the explanation very frequent. Thus, it is necessary to also consider, the frequency over all experts.
3. **Merge Syntactical** After the quantitative analysis, concepts and interactions that are identical in meaning need to be merged. This step involves a syntactical merge, uniting interactions and concepts in different grammatical forms, such as plural and singular, as well as different grammatical cases.
4. **Merge Semantical** The final step of the analysis is a semantic merge, which involves merging interactions and concepts that are similar in meaning. This includes linguistic synonyms, where the more abstract version is retained. Additionally, synonyms that arise from context should be merged, keeping the more accurate term according to the literature. In both cases, the frequency of both terms is summed up, and only the retained term is considered for the concept map.

We have summarized the steps in Fig. 2.

With the extracted concepts and interactions, the concept maps can be created. The nodes of the concept map represent the reduced concepts and interactions. Thus, the arrangement and the edges between the concepts and interactions need to be specified.

For our approach, we based the edges and relations on the concepts of designing object-oriented programming classes. This is based on the idea that object-oriented programming already has a structured form for creating map-based methods for software design, for example unified modeling language (UML) by the Object Management Group, Inc. (2023). Which is used to model even abstract concepts and domains for systematic development of software. For the edges, we defined the relations: "has", "does", "gets", and "produces". This is based on the idea that a class in object-oriented programming has attributes, methods that get an input and produce an output. "has" represents the attributes the class would have. For example, an *apple* class would have a *size* attribute. An apple has a size. The word "does" relates a class concept to a method that class has. An example here would be a *dog* class could have method *run*. "gets" relates the input of a class method to the method it would be inputted to. "produces" on the other hand relates a class method to the output it would produce. Considering the run method from the dog example, it could get a time and destination and then produce the speed the dog is running. Furthermore, we set a dashed line to indicate that a concept or interaction is a hypernym. This relates to the idea of superclass and subclasses in the realm of object-oriented programming. For example, fruit is a hypernym for apple. Thus, an apple is a fruit and could as an object-oriented class inherit the attributes and methods from a fruit. This edge indication allows for a better overview of how deep of a knowledge of some topics the expert assume. These would be the additional attributes and methods an apple has compared to a fruit.

With the given relations, the arrangement of the concepts is still open. The first criteria would be that classes, methods and attribute that belong together should be clustered together. This improves clarity and highlights central topics. Also, attributes should grouped per class for the same reasons as for the classes. Method interactions should be aligned and placed close to their inputs and outputs, ideally from input to the left and output to the right, or alternatively, from top to bottom. The gathering of the expert interviews, analysis, and the described relations allow the transfer of producing the concept maps for the didactic reconstruction in the field of XAI systematically.

This concludes the analysis of the expert interviews. With this reduction, we can proceed with the didactic reconstruction. The next step is to transfer the idea to identify the lay user's mental model.

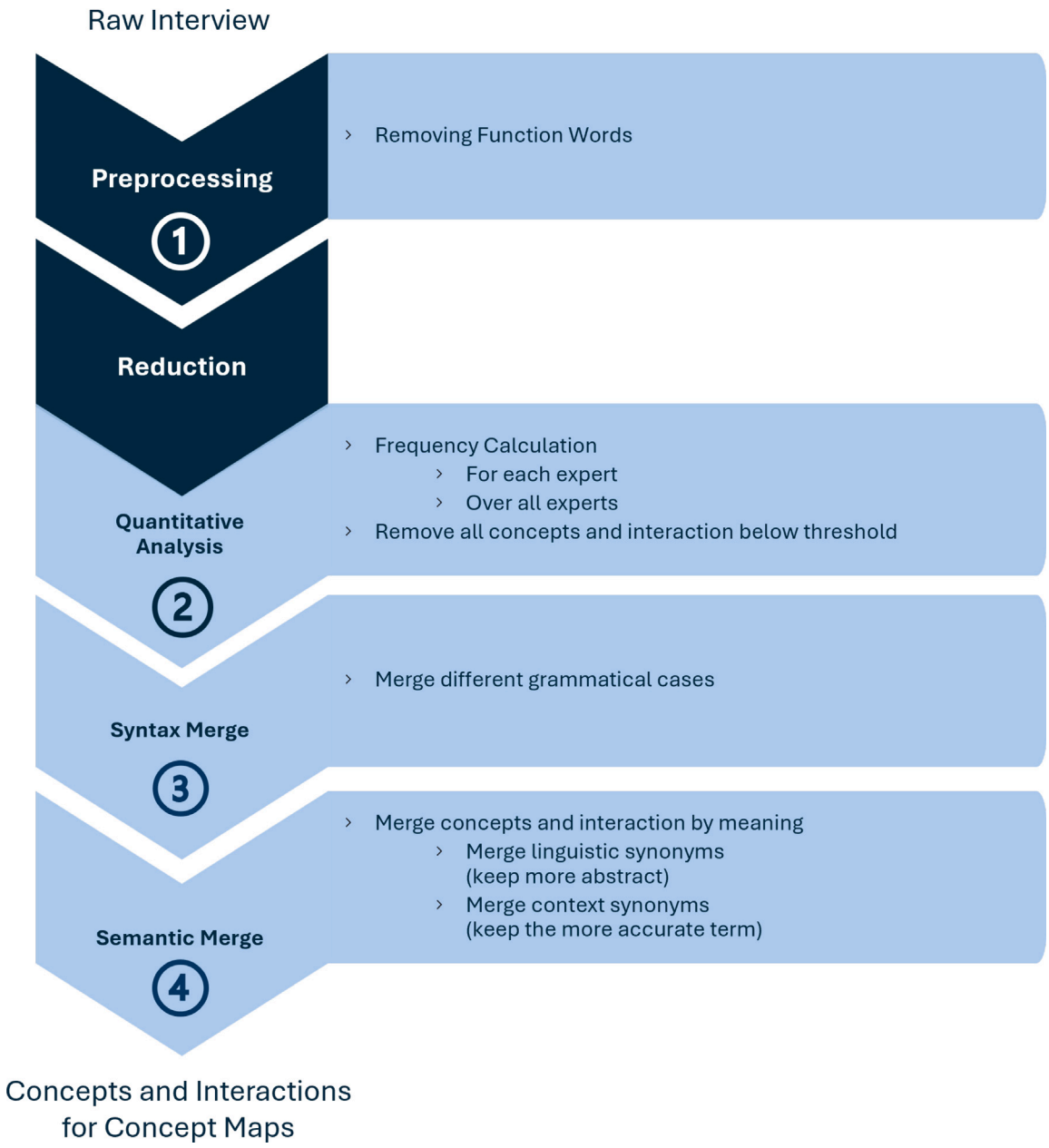


Fig. 2. A depiction of the analysis steps to process the raw expert to extract the concepts and interactions for the concept maps.

3.3.2. Lay user

In the context of XAI a lay user is a user or developer of the AI System. However, the lay user can seek theoretical as well as practical knowledge about the AI. For example, either to use the AI as an end-user or as a developer of an AI.

Reduction. The definition of a mental model, as understood by the didactic reconstruction, aligns more closely with the concept of prior knowledge, as described under the definition of mental models in Section 2. Therefore, we aim to systematically bringing the mental model, influenced by the users' prior knowledge, into a form that can be synthesized with the factual situation extracted from the expert explanations. To achieve this, we captured the users' explanations for their interactions with the AI in order to reduce them, similar to the approach taken with the expert explanations. In the following, we will describe the steps involved in collecting the users' explanations.

As shown by Norman (2014) the execution of a task is impacted by the prior knowledge. However, the mental model based on the context of Section 2 changes over the course of the interaction, as discussed by Thomaz and Breazeal (2008). Therefore, the lay user's mental model should not only be assessed once before the session or only after but rather in the three phases before, during, and after the interaction with the AI. For the assessment of the before the interaction, a questionnaire can be used, as suggested by the didactic reconstruction. Then the users interact with the AI. Afterward, a video recall of that interaction afterward can be performed to access the changes during the interaction. This approach offers several advantages. During the recall, the lay users can be probed when they directly or indirectly mention an expert concept extracted as described in Section 3.1.1. This probing allows clarifying that the lay users refers to the same concepts as the experts, besides naming it the same way. Furthermore, it eases the synthesis of expert concepts and lay user's mental models. Another advantage of the recall technique is that it allows the lay users to focus more on the subject of explanation since they are less involved after the interaction. On the one hand, the mental load of the lay user is lower, as they do not have to act and describe their actions. On the other hand, lay users

not only describe what they do, but can also describe the causality of their actions or even comment on the underlying system. This reduced focus of the users on the causality, their beliefs of systems and their own actions during the video recall also eases the synthesis because by the probing and the high level thoughts, their mentioned concepts are closer to the experts statements. After the action and interview, the lay users should again be quarried to access their mental model of a learning AI after the interaction.

In this way, the mental model is assessed before, during (via recall), and after the interaction. This method results in three explanations per lay user: Before, during, and after the interaction However, this leads to a rich lay user explanations, which also need to be reduced. This reduction of the explanation can then be conducted analogously to the reduction of expert explanations as described in Section 3.3.1. Afterward, the reduced expert architecture and lay user's mental models can be synthesized, which is the last step in the didactic reconstruction.

3.3.3. Synthesis

The approach results in two concept maps that each represent an average of the concepts and interactions of experts as well as lay users. These do not represent the individual mental models and might be more complex than the actual model. They also do not represent the concepts and interactions that are mentally present at all time of the interaction. However, they allow an overview of the possible gaps of knowledge that an individual might have and as described in Section 3.2 the gaps between the lay user's mental model and factual situation must be identified for effective explanations. To achieve this, the factual situation and the lay user's mental models need to be matched to find these gaps. The matching process requires as much structure as possible to be reproducible and applied to other contexts. With the two sets of reduced explanations in the same concept map format, the synthesis can be done.

Categories. Initially, we have examined diverse types of information that occur. That are given by the corresponding architecture, the reduced explanations of the lay users and experts, the subject of explanation, the mental model of the lay users as well as the enabling architecture. To illustrate this, we have depicted it in a Venn-diagram (Fig. 1), which we will explain in more detail below.

On the one hand, there is all the information about the AI system. This is the architecture which is shown in light blue. Part of the architecture is the enabling architecture, shown in dark blue. This includes all the information a lay user needs to use or develop the system effectively. The part we focus on is the explanandum highlighted in turquoise. This represents the part of the enabling architecture that still keeps the lay user from using or develop the system efficiently and needs to be explained. Since, the experts are able to use or develop the system effectively, the enabling architecture is part of their knowledge. Thus, the enabling architecture is included in the reduced expert explanations. Adjacent to this is the lay user's overall mental model, which is shown in orange.

Area A The mental model contains the irrelevant information that is not needed in the context of the interaction with the AI, marked in dark orange.

Area B Furthermore, there exists information within the lay user's mental model that is already accurately memorized with regard to the enabling architecture. This is the already known information located in area B and illustrated in blue–orange.

Area C The explanandum intersects the lay user's mental model, which are concepts and interactions that the lay user already knows, but with the wrong assumption or in the wrong context. However, these are not irrelevant to the efficient usage of the architecture. These are misunderstandings and represent area C marked in blue–orange. Misunderstandings within the relevant architecture are part of the explanandum, as they need to be clarified to allow efficient usage of the system.

Area D Lastly, there is area D representing the concepts and interactions the user is missing from the enabling architecture. These are part of the explanandum as well. The users need this information to effectively interact with the system.

Thus, the enabling architecture is part of the overall architecture and contains the information that a lay user might need to use the system effectively. Part of the enabling architecture is the explanandum, which is the information the lay user actually missing. It excludes the parts of the enabling architecture that are already known. The already-known information is part of the lay user's mental model and the enabling architecture. The explanandum itself is divided into misunderstandings and missing information. The misunderstandings are part of the enabling architecture and the mental model of the lay user since they are information about the enabling architecture the lay user knows partially. The missing information is not part of the mental model of the lay user but part of the enabling architecture. The information that is not part of the enabling architecture, but part of the architecture, is considered irrelevant for the effective usage of the system.

Categorization. The concepts and interactions of both experts and lay users within their concept maps can be categorized by assigning them to the elaborated classes A to D. This classification, however, can only be done by considering both concept maps, as the experts and lay users together define the categories A to D. With the classification, one can see the missing information in the classified classes within the experts concept map. This, however, is only the set of possible topics to explain and would represent the average need for explanation, which does not mean that every lay user needs all the missing information or the same information to use the system effectively. This also holds for the misunderstandings in the lay user's concept map. Both the knowledge gap and misunderstandings serve as the foundation for developing explanations for lay users in the XAI setting, mirroring their role in didactics for teaching concepts. In the following, we will describe the matching process for the synthesis of the two concept maps.

Initially, it is necessary to identify all concepts and interactions within area B, which represent the already-known information. This simplifies the categorization of the remaining concepts and interactions into their respective areas. The identification entails determining all concepts and interactions that are present in both concept maps, sharing the same context and interpretation. The concepts and interactions that were classified by this are part of area B of the already-known information c.f. Fig. 1.

The next step is to identify missing information and irrelevant information, both of which are exclusive to one type of concept map.

The irrelevant information is only part of the lay user's concept maps. This is due to the fact that the lay user's concept maps were created based on the interaction with the AI. Thus, the base explanations are with the interaction in mind. However, with the assumption that explanations of the underlying architecture improves the effectivity of interaction with an AI as for example shown by Hindemith et al. (2020) the target of the explanation is thus the underlying architecture and therefore interaction specific are irrelevant information for explanations. This does not mean that the concepts and interactions cannot be used in an explanation as examples for underlying concepts and integrations, which are supported by the didactic reconstruction. In these cases however those are not the target of the explanation but rather a tool for the explanation. Therefore, all concepts and interactions that are part of the lay user's concept map but not of the experts are considered irrelevant since they do not referrer to the underlying architecture. These irrelevant concepts and interactions of the lay users are then categorized in group A c.f. Fig. 1.

On the other hand, there are concepts and interactions that are only in the expert's concept map. Those are considered missing information since they are potential topics of explanations. Those concepts and interactions are then categorized as missing information and therefore in group D c.f. Fig. 1.

Following that, one can proceed to identify the concepts and interactions situated within area C, which represent misunderstandings. These entail concepts and interactions discussed in incorrect contexts or based on false assumptions.

As for the concepts and interactions of the area B the misunderstanding are identified by concepts and interactions existing in both maps. However, these are concepts and interactions that exist in both concept maps but do not share the same context and interpretation. The misunderstandings are then marked in the concept map of the lay users. As the didactic reconstructions suggests, those need to clarified and are thus ground of possible explanations. This does not mean that experts might not also have misconceptions. Those concepts and interactions are the ones that are left besides the matching and the missing ones in each diagram.

By this, all concepts and interactions of both concept maps can be categorized. This classification can now serve as a basis for explaining the AI system accordingly. The known information can be considered given. The misunderstandings need to be clarified, and the missing information needs to be supplemented. Those explanations could then be the base of a dynamic XAI System. This method does not cover the sequence and structuring of the explanation. However, it does provide a clear and structured approach to define and trim down the pool of possible topics. This concludes the transfer of didactic reconstruction to XAI, providing an approach to identify the enabling architecture. This architecture includes what needs to be explained for a user to use the AI system more effectively.

Overall, the approach has three structured phases that can be applied in other domains within XAI as well as other fields. First, gather explanations from experts and users. Second, reduce both according to the approach described in Section 3.3.1 and connecting them to acquire the concept maps. Third, synthesize them according to Section 3.3.3. In the following section, we applied the approach to an actual AI system for a feasibility check.

4. Case study

In this chapter we will go through the three phases of the approach using an example architecture. We consider the example scenario of a human teaching a robot arm. The robot arm is controlled by the PI^{BB} algorithm (Stulp & Sigaud, 2012). The learning will be examined more precisely in a mini-golf task setup. We will present two studies in this context where we collect explanations from both experts and users, reduce these explanations, and then synthesize the resulting concept maps.

4.1. Factual situation and reduction

First, we queried architecture experts to get our example architecture that forms the concept map's base, as described in Section 3.3.1.

4.1.1. Method

The experts. As an expert, we define the scientists that worked on the initial paper of PI^{BB} or were referenced as related work in those papers. We contacted over 60 experts via mail and got 19 responses. Of those 19 experts (all male), only 9 provided evaluable and complete feedback. The other experts responded to the request but did not access the questionnaire, did not complete it, or their responses were not linguistically coherent or relevant to the questions asked. All nine experts had recently worked with the algorithms. On average, they worked with the algorithm a year ago with a standard deviation of 0.93. Furthermore, they worked with the architecture for a long period of their careers (M=575 years, SD=570 years). They modestly self-reported the complexity as medium (M=3, SD=1) on a five point Likert scale.

Questionnaire. Since most experts are active in industry or research, they are busy, making appointment scheduling difficult. Additionally, the experts are located worldwide, which adds the difficulty of different time zones. We thus chose a questionnaire over direct interviews. This also allowed the experts to fill out the questionnaire in multiple sessions. As we have also observed, response rates from experts tend to be low, making it all the more important to keep the threshold for participation as low as possible.

The questionnaire was divided into two parts. In the first part of the questionnaire, the experts stated their state of knowledge with our example architecture. This included the amount of experience they had with the algorithm and when they last worked with it. In the questionnaire's second section, we asked for an explanation of our example algorithm. We queried explanations targeted to lay users. Besides the context, we stated that the scope of explanations could be individually defined. This means the experts could give bullet points or full-text answers, and had the choice of including the own figures. This specific questionnaire consists of general questions but with a focus on PI^{BB}, for example, "Imagine explaining PI^{BB} to a lay user". While the questionnaire could be adapted for other algorithms, a similar questionnaire could also be used to generate expert explanations.

4.1.2. Reduction

However, the raw explanations themselves do not yet serve as the enabling architecture, due to their varying scope and the complexity level, which is too high for lay users. Regarding the didactic reconstruction, the expert statements form the basis for the factual situation to determine the architecture of our example algorithm, as described in Section 3.3.1. Accordingly, the explanations need to be reduced as described in Section 3.3.1. The results of our study and the reduction is the concept map depicted in (Fig. 3) based on the explanation of PI^{BB}.

The results suggest that the experts focus on two main concepts: algorithms in interactive robotics contexts in general, and learning, specifically PI^{BB}. This is evident in the resulting division of the diagram into two halves: the left side focusing on algorithms and the right side on learning. The algorithms section includes properties that are most characteristic, such as "Input", "Parameters", and "Output".

The reason these subject areas were highlighted is that experts differentiated the PI^{BB} algorithm from other algorithms. This is exemplified in explanations such as

> "*PI^{BB} is a learning algorithm that* ⋯" [T40] or "*PI^{BB} is an optimization algorithm that* ⋯" [T36].

Thus, an understanding of algorithms is presupposed to comprehend the workings of PI^{BB}. Moreover, the application of PI^{BB} was discussed to highlight its advantages, which explains the focus on interactive robotics. This can be seen in explanations like:

> "⋯ *to achieve a certain goal/movement with a robot*" [T44] or "*The robot performs a task several times.*" [T52]

Regarding the learning algorithm, general terms such as "Exploration" and "Optimization" were mentioned to distinguish PI^{BB}, which also accounts for the description of specific PI^{BB} properties like "Weight" and "Mean". Examples of this include

> "*reward-weighted optimization approach* ⋯" [36] or "⋯ *the mean is computed from the parameters*" [T58].

4.2. Lay user's mental model

As described in Section 3.3.2 we conducted another study to identify the lay users' mental models of our learning algorithm PI^{BB}.

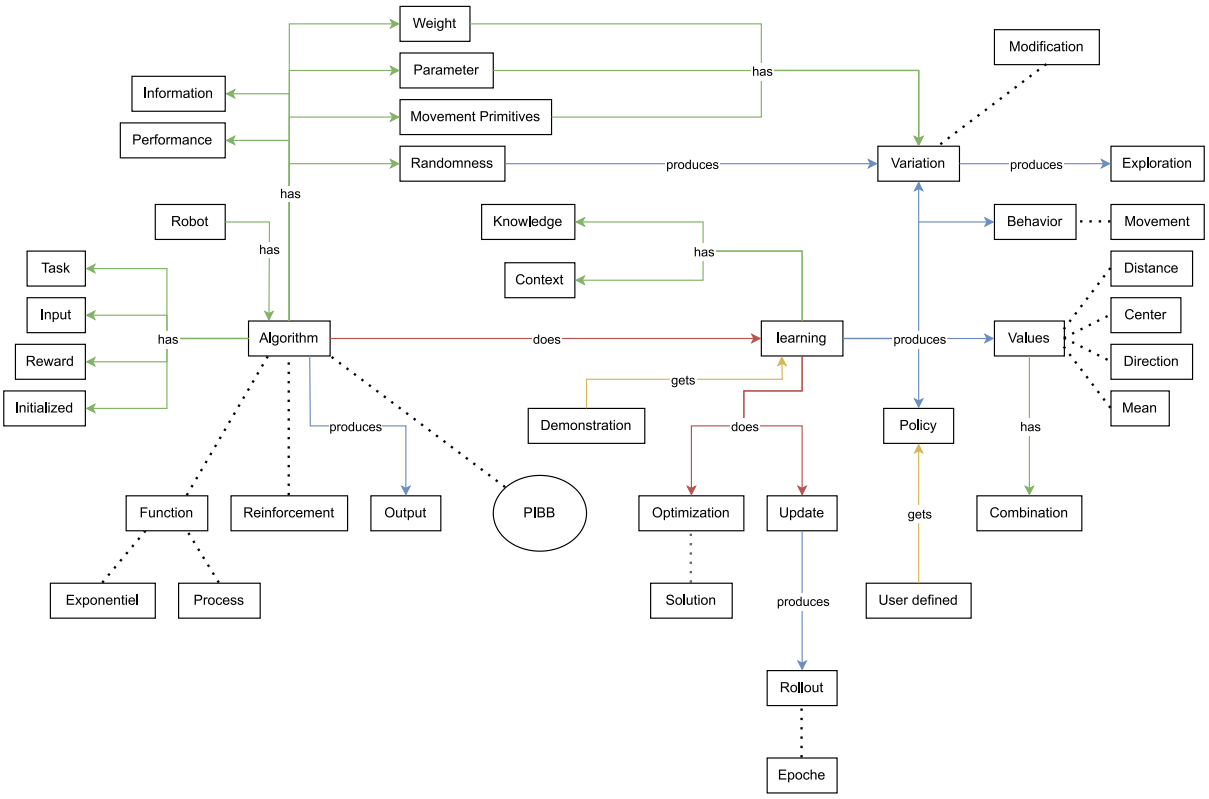


Fig. 3. The concept map of PI^{BB}, based on the analysis and reduction of the expert explanation. Solid arrows ('has', 'gets', 'does', 'produces') represent direct relationships. Dotted lines link conceptually analogous terms.

4.2.1. Method

Lay users. Overall, in our study, we had 10 participants who were acquired through our working group's network. The study had slightly more male participants than female (m=7, f=3). In accordance with the definition of a lay user in XAI (Section 3.3.2), we define a lay user as an individual with limited background knowledge in robotics and machine learning in the context of our use case. We queried the participants about their background in robotics. Half of the participants assessed their knowledge in robotics as very low, while the majority of the remaining participants rated their understanding as medium on a 5 point likert scale ($M = 2, SD = 0,94$). The knowledge was never graded higher than medium on a five-point scale from very low to very high. Thus, the participants can be considered lay users according to our definition. We also asked about their educational level, which varied widely. It ranged from high school diploma to university degree ($N = 5$). Half of the participants possessed university degrees in fields other than robotics or computer science, for example, engineering or economics.

Setup and robot. The modalities were implemented on a Kinova Gen2/Jaco (Jaco2) robotic arm (Kinova inc., 2023) using ROS Noetic (Open Robotics, 2020). To represent movements, we used probabilistic movement primitives (ProMP) more specifically used the implementation by Fabisch and Klemm (2020), Paraschos, Daniel, Peters, and Neumann (2013). ProMP depict movements based on parameterized distributions. This makes them adaptable as they cover entire distributions rather than singular trajectories. Moreover, they accommodate changes in trajectories by changes in parameters. For learning purposes, We implemented a version of the PI^{BB} black-box algorithm (Stulp & Sigaud, 2012). The PI^{BB} algorithm learns the correlation between ProMP weights and provided preference feedback, thereby refining a distribution of ProMP weights utilized to generate new, improved movements. Regarding the covariance decay factor (= 0.973) and with the eliteness factor $h = 10$. The preference reward was realized by the function

$$Preference \begin{cases} (100, 100) & \texttt{Both preferred} \\ (100, -100) & \texttt{First action preferred} \\ (-100, 100) & \texttt{Second action preferred} \\ (-100, -100) & \texttt{None preferred} \end{cases}$$



Fig. 4. Our example graphical user interface.

In our study, we also developed a graphical user interface (GUI) to showcase a preference feedback modality, depicted in Fig. 4. The title says "*Rating*", and both hearts are titled "*Movement 1*" and "*Movement 2*". The central button is a submit button for the rating and is labeled "*Submit*".

Task. For our study, we selected a miniature golf scenario featuring a track without obstacles (Fig. 5) as the task. We chose this task as it is supposed to be straightforward enough for users to assess without being overly abstract, ensuring it does not detract from the learning experience. Additionally, its engaging nature should help maintain users' interest throughout the process. This setup was designed to focus on direct user input for robot training, avoiding the complexities

that visual learning might introduce. While there are classic tasks for human–robot interaction such as pick-and-place, these do not focus on the movement itself but rather on the objective of the movement. For tasks that focus on movement, playful activities like table tennis are often used (Ding et al., 2022; Peters, Kober, Mülling, Krämer, & Neumann, 2013). We have adapted these tasks to fit the hardware used, considering factors such as movement speed, space requirements, etc.

Procedure. First, we informed the participants that the study was concerned with robot training. We also stated that we wanted to evaluate how users with little knowledge about the system and algorithm use them. Participants were informed about the payment of 10€ for the study. We communicated to the users that their data would be collected and processed, and all subjects gave their informed consent for inclusion before participating in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the local Ethics Committee. Subsequently, we surveyed the participants through a questionnaire about their background in robotics and their educational background.

Then we showed them our example setup as shown in Fig. 5 and explained to them that this is the robot they will be working with. In this context, we explained the tasks to the participants. We made it clear that participants need to hit in a single stroke and that there are 20 rounds; if they succeed earlier, they should continue to hit the ball. We explained that the training is based on their ratings. After that, we presented our example graphical user interface and explained how they could use it to grade the robot c.f. Fig. 4. We stated that the robot will show them two movements, and they could rate them by liking a movement. The interface also allows the user to rate both movements as good or both as bad. They were also made aware that although a camera would record the interaction for our observation, it would not be utilized by the robot during its training phase. Afterward, we performed a semi-structured video recall with the participants immediately after the interaction with the system as described in Section 3.3.2. We guided them to a separate table to show them their interaction with the robot. They were shown all loops of learning, where one loop consists of the movement that was presented and the reward the participants gave for that movement. Those loops were shown one after another, and the users stated their thoughts regarding the loop. The participants were informed that the interviews would be recorded. We instructed them that they were supposed to explain to us why they were acting as they did in the video. Also, we informed them that we would not judge their reasoning and that it is not essential that their statements are correct, and that it is most important that they answer the questions as honestly as possible. When they mention a concept stated by the experts or a related concept, we probed them further to understand what they meant. Care was taken to ensure that the participants were not given a predefined answer. Finally, we thanked the participants for their participation and sent them off.

We analyzed the results by using the same analysis as for the expert explanation. The concepts and interactions were counted and reduced as described in Section 3.1.1.

4.2.2. Reduction

The analysis resulted in the concept map depicted in Fig. 6

The results indicate that the participants' explanations were clearly focused on the task and its execution. For example, the reason that

> *"… Distanz vom Rand des Schlägers zum Ball"*[translated: "Distance from the edge of the bat to the ball"] [T1] or *"… der Schläger sozusagen einen Teil vom Ball berührt hat."*[translated:"the club has touched a part of the ball, so to speak."] [T2]

When participants referred to the algorithm, they often focused on the observable behavior of the robot, like variations or randomness.

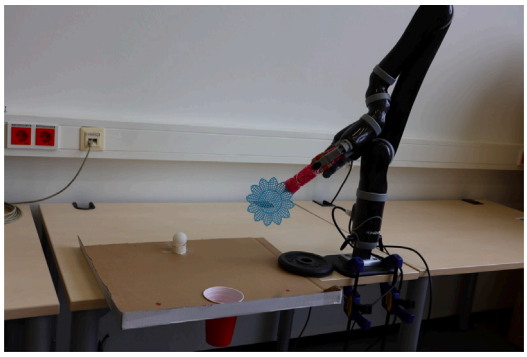


Fig. 5. Our example setup. The Kinova Jaco 2 with an obstacle-free self-made golf course. The ball is a light ping-pong ball.

> *"… dass dann einfach was Neues versucht …"*[translated: "that then simply try something new"] [T8] or *"… das war jetzt einfach vielleicht Glück …"* [translated:"maybe that was just luck"] [T2]

The last major part of the explanations often relied on common knowledge about machine learning. For example,

> *"… dass er sich das merkt … "*[translated: "that he (the robot) remembers that"] [T1] or *"… wahrscheinlich ein neuronales Netz oder so etwas …"* [translated:"probably a neural network or something"] [T5]

With these results, we could synthesize the expert's concepts and the user's mental models in a structured way.

4.3. Synthesis

As described in Section 3.3.3, we have considered different types of knowledge that are given by the corresponding architecture and the lay users. We categorized the concept maps as described in Section 3.3.3. First, all concepts and interactions that appear in both concept maps depicted in Figs. 3 and 6 were marked in the same context and interpretation in orange as Area B in Fig. 1. Afterward, the concepts covered in the wrong context or incorrect assumptions were classified as misunderstandings. Lastly, the missing and irrelevant information were categorized. The missing information was marked in turquoise as Area D in Fig. 1. Irrelevant information was highlighted in orange, completing the classification of all states. The synthesis of results from the two concept maps is visualized through the color-coding in Figs. 7 and 8, showcasing the distinctions and overlaps between expert and lay user perspectives.

5. Discussion

The discussion is structured along the areas of knowledge presented in Fig. 1.

We will first consider the area of irrelevant information (Area A). The algorithm can be applied to different tasks without changes to its functioning. Thus, the experts explained the algorithm independent of the task, and the enabling architecture is task-independent. However, the algorithm is utilized in the context of a task. Therefore, the lay users are introduced to the architecture in a task context. As a result, lay users tend to refer a lot to the actual task, which is irrelevant to their understanding of the underlying algorithm. This task-focused perspective is evident in our case study, where the lay users focus more on the physical environment in the form of 'cup', 'ball' or 'goal' or individual characteristics of the robot arm stroke, such as the 'direction', 'speed' or 'height'.

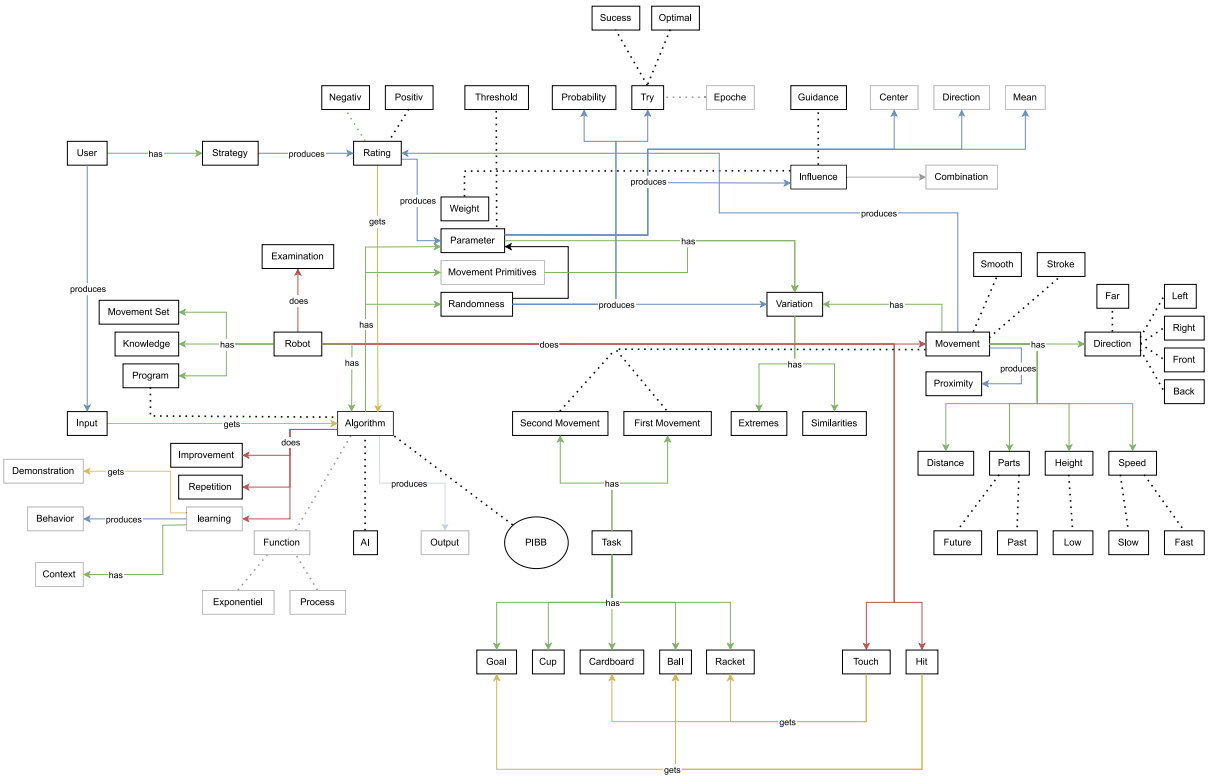


Fig. 6. The concept map of PI^{BB} based on the analysis of the lay user explanation. Solid arrows ('has', 'gets', 'does', 'produces') represent direct relationships. Dotted lines link conceptually analogous terms.

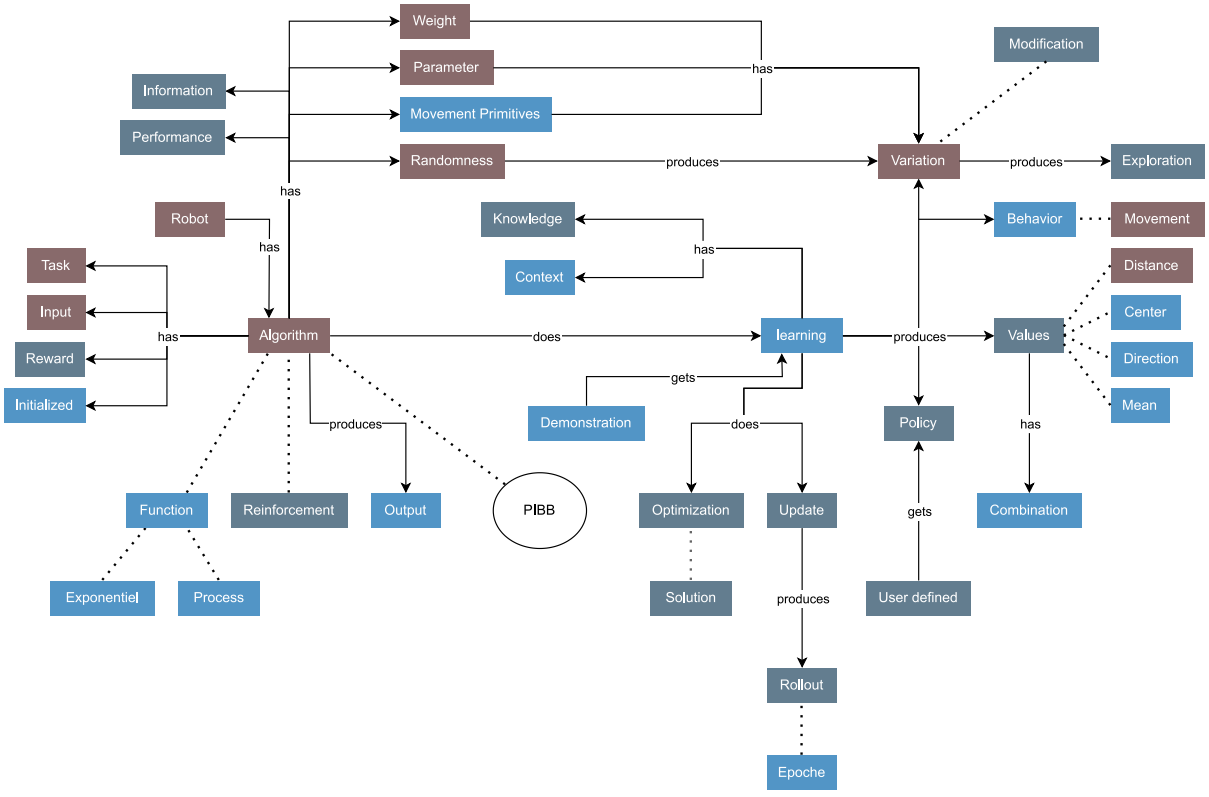


Fig. 7. The classified version of the concept map of the expert's concept map. Solid arrows ('has', 'gets', 'does', 'produces') represent direct relationships. Dotted lines link conceptually analogous terms. Colors of concepts in the diagram signify: light blue for missing, dark blue for misunderstood and wine red for known information.

Next we will consider the area B of known information, including aspects mentioned by the lay users but also stated in the expert study. For instance, the lay users have described that the robot has an algorithm with some kind of randomness and thus provides movement variation.

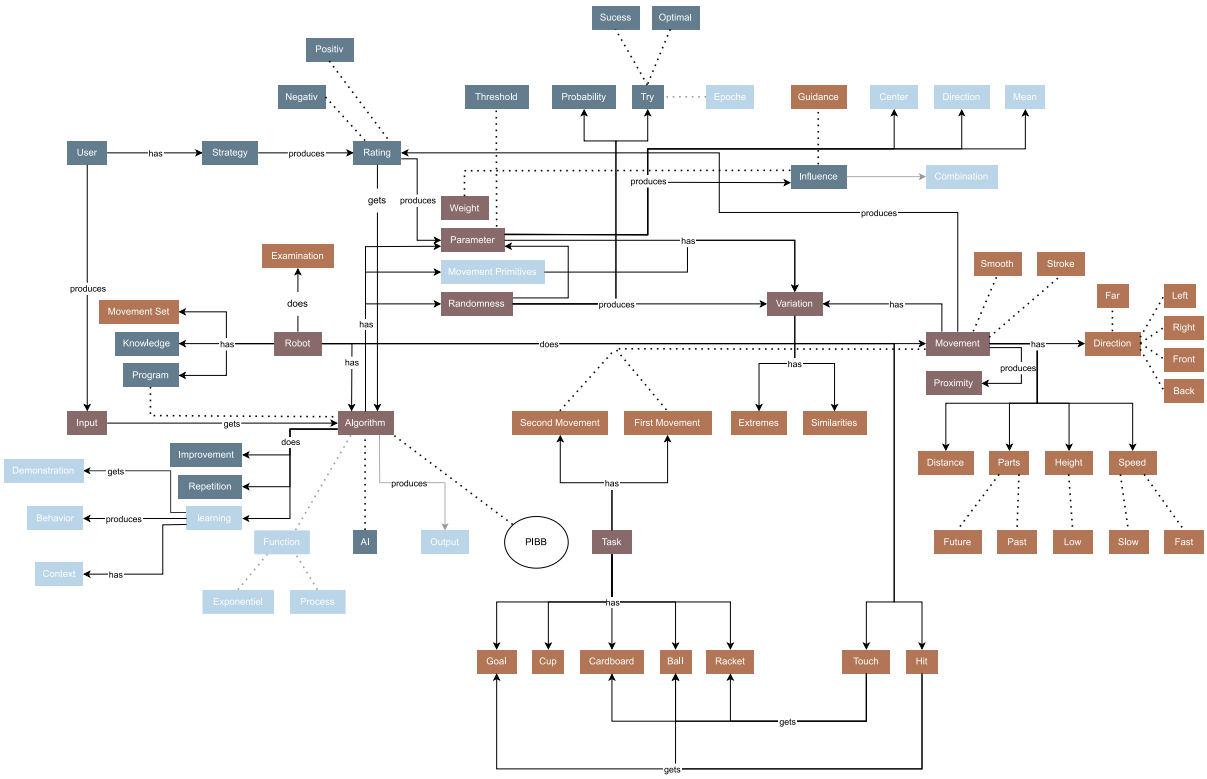


Fig. 8. The classified version of the concept map of the lay user's mental model. Solid arrows ('has', 'gets', 'does', 'produces') represent direct relationships. Dotted lines link conceptually analogous terms. Colors of concepts in the diagram signify: light blue for missing, dark blue for misunderstood, wine red for known, and orange for irrelevant information.

In area C, concerning misunderstandings, there were instances among lay users where concepts were either partially understood or incorrectly placed in the wrong context. Misinformation seems to arise when lay users are aware of the existence of a concept or interaction but lack understanding regarding its processing, utility, functionality, or context. In contrast to the known information (Area B), those are concepts and interactions that are not derivable by the perceived situation or are less commonly known and thus not part of the prior knowledge of the lay users. For example, the existence of the rating was apparent since it was part of the task and user interface. However, its processing and utility as a reward in a reinforcement learning context were not derivable by observation and not part of the lay users' prior knowledge. Nevertheless, the user would probably benefit from a deeper understanding of the reward concept, since it would allow for more precise manipulation of the learning process.

In the realm of missing information (Area D), it became apparent that lay users predominantly lacked understanding of the robot's operating principles. Certain concepts were completely absent from user discussions, suggesting that users might either be unaware of these concepts within this context or oblivious to their existence altogether. This lack of association or recognition could root from the concepts being less commonly known, hence not being a part of the users' pre-existing knowledge base. An illustrative example of such a concept, likely unknown to the users, is 'Movement Primitive'. This indicates a gap in conceptual understanding that needs to be addressed to enhance comprehension and engagement with the technology.

Overall, it is evident that lay users primarily focused on physically observable aspects, such as the movement of the robot arm, rather than delving into the underlying mechanisms that control such movements. For these users, the algorithm functioned similarly to a black box; they observed inputs and outputs, as well as components like weights and randomness, but seldom engaged in reasoning about the interactions and causal relationships between these elements. This observation highlights a gap in understanding the complex interplay that governs the behavior of robotic systems, emphasizing the need for more intuitive explanations of these underlying structures.

Concluding this classification could serve as a foundation for the subsequent explanation of the learning system. The known information can be treated as established, while the misunderstandings will need clarification, and the missing information should be supplemented. These explanations could then be applied in contexts like explainable artificial intelligence in robotics, enhancing transparency and understanding of the underlying mechanisms driving robotic behaviors.

Throughout this paper, we have transferred the principle of didactic reconstruction to the area of XAI. This allows for the identification of the set of information that potentially should be explained. The specific subset needed and how it should be explained differs individually. Finally, we applied this methodology to an example architecture to demonstrate that this approach can systematically lead to a basis of concepts and interaction that might need explanations.

5.1. Future work

The questions surrounding the influence of modalities on users' mental models, and consequently on the enabling architecture, remain a significant area of inquiry. In our study, we deliberately chose to exclude modalities, except preference, to focus our analysis. However, it is plausible to suggest that different modalities could substantially alter both the mental model of users and the structure of the enabling architecture. For instance, scalar evaluations demand from users the creation of a global scalar, applicable across all interactions, as opposed to preferences that are only relevant within a singular interaction context.

Another critical question that our research has yet to address fully is the determination of the appropriate level of abstraction, which we anticipate tackling in the synthesis step of our approach. The chosen level of abstraction will dictate which concepts and interactions from

the reduced architecture should be adapted to meet the user's needs. Here, abstraction refers to simplifying the complexity of expert concepts to make them comprehensible to lay users. Preliminary insights into structuring the level of abstraction can be gleaned from analyzing different knowledge areas, using the already-known information as a foundation for further explanations. Misunderstandings – where users are aware of certain concepts or interactions but lack detailed knowledge – might indicate the minimum level of abstraction necessary for user comprehension, while missing information could represent concepts below their current level of abstraction.

Furthermore, the form of the presentation itself is an open question. This also includes how the explanations are structured. In our approach, the information is not prioritized or graded. This could be explored further in the future. Nonetheless, we are actively exploring the structure of explanations within our example architecture, aiming to enhance clarity and accessibility for users. This endeavor not only seeks to bridge the gap between expert and lay understanding, but also to refine the methods through which complex architectures are communicated and understood. Lastly, the automation of our method is an important future task. Our systematic approach could, for example, enable the use of knowledge graph generations or large language models in order to reduce the factual situation.

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 – 438445824 and SAIL, SAIL is funded by the Ministry of Culture and Science of the State of North Rhine-Westphalia under the grant no NW21-059A.

CRediT authorship contribution statement

Helen Beierling: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Phillip Richter:** Writing – review & editing, Writing – original draft, Validation, Investigation, Data curation. **Mara Brandt:** Conceptualization. **Lutz Terfloth:** Conceptualization. **Carsten Schulte:** Conceptualization. **Heiko Wersing:** Supervision. **Anna-Lisa Vollmer:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

Akalin, N., & Loutfi, A. (2021). Reinforcement learning approaches in social robotics. *Sensors*, *21*(4), 1292, Publisher: MDPI.

Aldebaran (2022). Pepper. URL: https://www.aldebaran.com/en/pepper.

Argall, B. D., Chernova, S., Veloso, M., & Browning, B. (2009). A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, *57*(5), 469–483.

Berzuk, J. M., & Young, J. E. (2023). Clarifying social robot expectation discrepancy: Developing a framework for understanding how users form expectations of social robots. In *Companion of the 2023 ACM/IEEE international conference on human-robot interaction* (pp. 231–233).

Bromme, R., & Rambow, R. (1998). *Die Verständigung zwischen Experten und Laien: Das Beispiel Architektur*. Springer.

Brooks, C., & Szafir, D. (2019). Building second-order mental models for human-robot interaction. arXiv preprint arXiv:1909.06508.

Bruckenberger, U., Weiss, A., Mirnig, N., Strasser, E., Stadler, S., & Tscheligi, M. (2013). The good, the bad, the weird: Audience evaluation of a "real" robot in relation to science fiction and mass media. In *Social robotics: 5th international conference, ICSR 2013, Bristol, UK, October 27–29, 2013, proceedings 5* (pp. 301–310). Springer.

Ding, T., Graesser, L., Abeyruwan, S., D'Ambrosio, D. B., Shankar, A., Sermanet, P., et al. (2022). Learning high speed precision table tennis on a physical robot. In *2022 IEEE/RSJ international conference on intelligent robots and systems* (pp. 10780–10787). IEEE.

Duit, R., Gropengiesser, H., Kattmann, U., Komorek, M., & Parchmann, I. (2012). The model of educational reconstruction–A framework for improving teaching and learning science. In *Science education research and practice in Europe* (pp. 13–37). Brill.

Fabisch, A., & Klemm, J. (2020). Movement primitives. URL: https://github.com/dfki-ric/movement_primitives, https://github.com/dfki-ric/movement_primitives.

Gentner, D., & Stevens, A. L. (2014). *Mental models*. Psychology Press.

Hindemith, L., Vollmer, A.-L., Göpfert, J. P., Wiebel-Herboth, C. B., & Wrede, B. (2020). Why robots should be technical: Correcting mental models through technical architecture concepts. arXiv preprint arXiv:2011.02731.

Intuitive Surgical, Inc (2023). Intuitive. URL: https://www.intuitive.com/en-us.

iRobot Corporation (2022). Irobot. URL: https://www.irobot.com/.

Jentner, W., Sevastjanova, R., Stoffel, F., Keim, D. A., Bernard, J., & El-Assady, M. (2018). Minions, sheep, and fruits: metaphorical narratives to explain artificial intelligence and build trust. In *Workshop on visualization for AI explainability at IEEE*.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*, (no. 6), Harvard University Press.

Kattmann, U., Duit, R., Gropengießer, H., Komorek, M., et al. (1997). Das modell der didaktischen rekonstruktion. *Zeitschrift für Didaktik der Naturwissenschaften*, *3*(3), 3–18.

Kiesler, S. (2005). Fostering common ground in human-robot interaction. In *ROMAN 2005. IEEE international workshop on robot and human interactive communication, 2005* (pp. 729–734). IEEE.

Kiesler, S., & Goetz, J. (2002). Mental models of robotic assistants. In *CHI'02 extended abstracts on human factors in computing systems* (pp. 576–577).

Kinchin, I. M., Möllits, A., & Reiska, P. (2019). Uncovering types of knowledge in concept maps. *Education Sciences*, *9*(2), 131.

Kinova inc. (2023). Kinova assistiv. URL: https://assistive.kinovarobotics.com/product/jaco-robotic-arm.

Kirk, J. R., & Laird, J. E. (2014). Interactive task learning for simple games. *Advances in Cognitive Systems*, *3*(13–30), 5.

Knox, W. B., & Stone, P. (2009). Interactively shaping agents via human reinforcement: The TAMER framework. In *Proceedings of the fifth international conference on knowledge capture* (pp. 9–16).

Kuhlmann, G., Stone, P., Mooney, R., & Shavlik, J. (2004). Guiding a reinforcement learner with natural language advice: Initial results in RoboCup soccer. In *The AAAI-2004 workshop on supervisory control of learning and adaptive systems* (pp. 30–35).

Lee, S.-l., Lau, I. Y.-m., Kiesler, S., & Chiu, C.-Y. (2005). Human mental models of humanoid robots. In *Proceedings of the 2005 IEEE international conference on robotics and automation* (pp. 2767–2772). IEEE.

Lehner, M. (2020). *Didaktische reduktion*. utb GmbH.

Malle, B. F., Fischer, K., Young, J., Moon, A., & Collins, E. (2020). Trust and the discrepancy between expectations and actual capabilities. *Human-Robot Interaction: Control, Analysis, and Design*, 1–23.

Mietzel, G. (2007). *Pädagogische Psychologie des Lernens und Lehrens. 8., überarb. und erw*, Aufl. Göttingen: Hogrefe.

Mumm, J., & Mutlu, B. (2011). Human-robot proxemics: physical and psychological distancing in human-robot interaction. In *Proceedings of the 6th international conference on human-robot interaction* (pp. 331–338).

Niebert, K., & Gropengiesser, H. (2013). The model of educational reconstruction: A framework for the design of theory-based content specific interventions. The example of climate change. *Educational Design Research–Part B: Illustrative Cases*, 511–531.

Norman, D. A. (2014). Some observations on mental models. In *Mental models* (pp. 15–22). Psychology Press.

Object Management Group, Inc. (2023). Kinova assistiv. URL: https://www.uml.org/.

Open Robotics (2020). ROS noetic ninjemys. URL: http://wiki.ros.org/noetic.

Paraschos, A., Daniel, C., Peters, J. R., & Neumann, G. (2013). Probabilistic movement primitives. *Advances in Neural Information Processing Systems*, *26*.

PARO Robots, Inc. (2014). PARO therapeutic robot. URL: http://www.parorobots.com/.

Passmore, G. J. (1999). *Concept maps and processes of comprehension, explicating cognition and metacognition, structural knowledge and procedural knowledge* (Ph.D. thesis).

Pérez-D'Arpino, C., & Shah, J. A. (2015). Fast target prediction of human reaching motion for cooperative human-robot manipulation tasks using time series classification. In *2015 IEEE international conference on robotics and automation* (pp. 6175–6182). IEEE.

Peters, J., Kober, J., Mülling, K., Krämer, O., & Neumann, G. (2013). Towards robot skill learning: From simple skills to table tennis. In *Machine learning and knowledge discovery in databases: European conference, ECML PKDD 2013, Prague, Czech Republic, September 23–27, 2013, proceedings, part III 13* (pp. 627–631). Springer.

Powers, A., & Kiesler, S. (2006). The advisor robot: tracing people's mental model from a robot's physical attributes. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on human-robot interaction* (pp. 218–225).

Ramaraj, P., Ortiz, C. L., & Mohan, S. (2021). Unpacking human teachers' intentions for natural interactive task learning. In *2021 30th IEEE international conference on robot & human interactive communication* (pp. 1173–1180). IEEE.

Sims-Knight, J. E., Upchurch, R. L., Pendergrass, N., Meressi, T., Fortier, P., Tchimev, P., et al. (2004). Using concept maps to assess design process knowledge. In *34th annual frontiers in education, 2004*. IEEE, F1G–6.

Staggers, N., & Norcio, A. F. (1993). Mental models: concepts for human-computer interaction research. *International Journal of Man-Machine Studies*, *38*(4), 587–605.

Stulp, F., & Sigaud, O. (2012). Policy improvement methods: Between black-box optimization and episodic reinforcement learning.

temi USA inc. (2023). Temi. URL: https://www.robotemi.com/.

Thomaz, A. L., & Breazeal, C. (2008). Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, *172*(6–7), 716–737.

Thomaz, A. L., Breazeal, C., et al. (2006). Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In *Aaai*: *vol. 6*, (pp. 1000–1005).

Vollmer, A.-L., Lohan, K. S., Fischer, K., Nagai, Y., Pitsch, K., Fritsch, J., et al. (2009). People modify their tutoring behavior in robot-directed interaction for action learning. In *2009 IEEE 8th international conference on development and learning* (pp. 1–6). http://dx.doi.org/10.1109/DEVLRN.2009.5175516.

Vollmer, A.-L., Mühlig, M., Steil, J. J., Pitsch, K., Fritsch, J., Rohlfing, K. J., et al. (2014). Robots show us how to teach them: Feedback from robots shapes tutoring behavior during action learning. *PLoS One*, *9*(3), Article e91349.