FINAL REPORT

1. General Information

DFG reference number: EP143/4-1, WE1468/10-1

Project number: 433323019

Project title: Learning Conversational Action Repair for Intelligent Robots (LeCAREbot)

Name(s) of the applicant(s): Dr. Manfred Eppe, Prof. Stefan Wermter

Official address(es):

- Dr. Manfred Eppe, Data Science Foundations, Hamburg University of Technology (TUHH), Blohmstraße 15, 20179 Hamburg, Germany
- Prof. Stefan Wermter, Knowledge Technology, Universität Hamburg, Vogt-Kölln Straße 30, 22527 Hamburg, Germany

Name(s) of the co-applicants: N/A

Name(s) of the cooperation partners: Prof. Jerry Feldman, Berkeley, USA.

Reporting period (entire funding period): 1.1.2020 - 31.12.2024

2. Summary

English Summary:

What are the principal mechanisms required to capture the robustness and interactivity of human communication, given the situational, noisy and often ambiguous nature of natural language? And how, and to what extent, can we integrate these mechanisms within an embodied functional model that is computationally and empirically verifiable? We addressed these research questions by investigating the linguistic phenomenon of conversational repair (CR) -- a method to edit and re-interpret previously uttered sentences that were not correctly understood by the hearer.

Previous computational models for human-robot dialog consider non-understandings, but they do not consider misunderstandings. Misunderstandings are common in natural language communication: they can result from inconsistent world models, erroneous perceptions, or ambiguous instructions. Addressing misunderstandings is important because they can cause a robot to execute unintended potentially irreversible and destructive actions. For example, given the instruction “bring me the bottle of water”, a robotic listener's vision system might confuse the water with an accidentally nearby bottle of cleaning detergent. In this case, the operator should

be able to utter an interrupting repair command such as ``No, erm... stop! No, not the detergent! I mean the water, to your right!"

We refer to such commands as conversational action repair (CAR) commands. Previous dialog models for human-robot interaction did not support such commands.

Our first step to address CAR was to develop a goal-conditioned reinforcement learning approach based on hindsight learning. This improved the grounding capabilities for instruction-following (Röder et al., 2022). Our surprising main result was that our new self-speech feedback method can catalyze the learning process (cf. Fig. 1). Our second step was to extend the self-speech-based instruction-following by action repair commands (Röder & Eppe, 2022), and we found that self-speech also improves the learning process in this case.

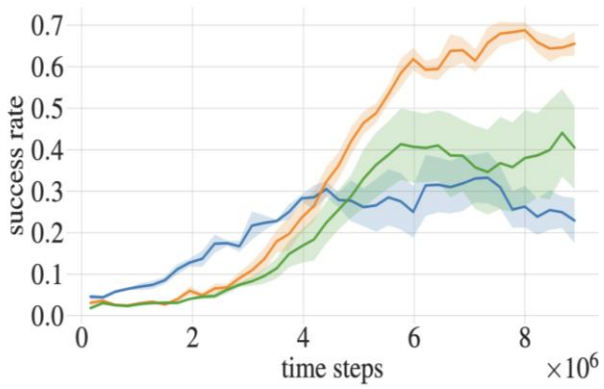

Fig 1: <u>English:</u> Results from our work (Röder et al., 2022) show that our innovative self-speech approach (orange) is superior to the expert feedback (green) and the baseline (blue). <u>Deutsch:</u> Ergebnisse aus unserer Arbeit (Röder et al., 2022) zeigen, dass unser innovativer Selbstgesprächsansatz (orange) dem Expertenfeedback (grün) und der Basislinie (blau) überlegen ist.

In addition to these results, we improved the Neuro-Inspired COLlaborator (NICOL), an adult-sized semi-humanoid based on our established NICO robot (Kerzel et al., 2023). We integrated our new ELMiRA (Embodying Language Models in Robot Action) architecture, merging speech, vision-language, and object detection with robot-specific spatial and motion models (Gäde, Özdemir, et al., 2024). This integration enables human-robot interaction and object manipulation tasks. To enhance sim-to-real transfer and imitation learning, we developed neural architectures using image-to-image transfer and differentiable forward kinematics (Gäde, Habekost, et al., 2024; Habekost et al., 2024; Spisak et al., 2024a).

<u>Deutsche Zusammenfassung:</u>

Was sind die wichtigsten Mechanismen, um die Robustheit und Interaktivität der menschlichen Kommunikation zu erfassen, angesichts der situativen, verrauschten und oft mehrdeutigen Natur der natürlichen Sprache? Und wie können wir diese Mechanismen in ein funktionales Modell integrieren, das rechnerisch und empirisch überprüfbar ist? Wir untersuchten diese Fragen, indem wir das sprachliche Phänomen der konversationalen Reparatur (KR)

betrachteten – eine Methode, um zuvor geäußerte Sätze zu bearbeiten und neu zu interpretieren, die nicht richtig verstanden wurden.

Frühere computationale Modelle für Mensch-Roboter-Dialoge berücksichtigen Nicht-Verständnisse aber keine Missverständnisse. Diese treten in natürlicher Sprache aber häufig auf. Sie können aus inkonsistenten Weltmodellen, fehlerhaften Wahrnehmungen oder mehrdeutigen Anweisungen resultieren. Die Behandlung von Missverständnissen ist wichtig, da sie zu unbeabsichtigten, potenziell irreversiblen Handlungen eines Roboters führen können. Zum Beispiel könnte das visuelle System eines Roboters bei der Anweisung „bring mir die Wasserflasche“ das Wasser mit einer Flasche Reinigungsmittel verwechseln. In diesem Fall sollte der Operator in der Lage sein, einen Reparaturbefehl wie „Nein, halt! Nicht das Reinigungsmittel, ich meine das Wasser, rechts von dir!“ zu äußern.

Wir bezeichnen solche Befehle als konversationale Aktionsreparatur (KAR)-Befehle. Vorherige Dialogmodelle für die Mensch-Roboter-Interaktion unterstützten solche Befehle nicht.

Unser erster Schritt zur Behandlung von KAR war die Entwicklung eines zielgerichteten Verstärkungslernansatzes basierend auf Hindsight Learning. Dies verbesserte die Fähigkeiten zur Anweisungsbefolgung erheblich (Röder et al., 2022). Unser überraschendes Ergebnis war, dass unser neues Selbstgesprächs-Feedback-Verfahren den Lernprozess katalysieren kann (vgl. Abb. 1). Unser zweiter Schritt war es, das Anweisungsbefolgen durch Aktionsreparaturbefehle zu erweitern (Röder & Eppe, 2022), und wir stellten fest, dass Selbstgespräche auch hier den Lernprozess verbessern.

Zusätzlich zu diesen Ergebnissen haben wir den Neuro-Inspired COLlaborator (NICOL), einen semi-humanoiden Roboter basierend auf unserem etablierten NICO-Roboter (Kerzel et al., 2023), weiter verbessert. Wir haben unsere modulare ELMiRA (Embodying Language Models in Robot Action) Architektur integriert, die Sprache, Vision-Language und Objekterkennung mit roboter-spezifischen räumlichen und Bewegungsmodellen kombiniert (Gäde, Özdemir, et al., 2024). Diese Integration ermöglicht Mensch-Roboter-Interaktionen und Objektmanipulationsaufgaben. Um den Sim-to-Real-Transfer und das Nachahmungslernen zu verbessern, haben wir neuronale Architekturen entwickelt, die Bild-zu-Bild-Transfer und differenzierbare Vorwärtskinematik verwenden (Gäde, Habekost, et al., 2024; Habekost et al., 2024; Spisak et al., 2024a).

3. Progress Report

3.1. Background and objectives

Both non-understandings and misunderstandings are very common in human-to-human communication. Previously existing robotic dialog systems focused on non-understanding, but there was a lack of systems that investigated misunderstandings and the repair of

misunderstandings. The main objective of this work was to fill this gap and to investigate conversational repair of misunderstandings for robotic agents. This central objective has been successfully achieved, as we show in our main publications (Kerzel et al., 2023; Röder et al., 2022; Röder & Eppe, 2022).

In the proposal, we formulated two overarching research questions, Q1 and Q2, that we found to be most critical to achieving our main objective.

Q1: How can we realize scalable representations that integrate the physical world with the dialog state?

Q2: How can we realize a data-efficient learning method for a hybrid neuro-symbolic semantic parser that is robust enough to cope with noisy and ungrammatical spoken language?

3.2. Project-specific results and findings

In the following, we describe how we addressed research questions Q1and Q2 in the corresponding work packages WP1 And WP2. In addition to these results, we have generated additional unexpected and valuable results and findings that we will also discuss in this subsection.

Q1/WP1: Scalable representations that integrate the physical world with the dialog state

To address this first work package, we extended goal-conditioned reinforcement learning towards temporally extended goals that reflect the dialog state, by integrating lingual goal commands into a verbal/nonverbal interaction between an agent and the environment/teacher. This resembles an instruction-following setting (Hermann et al., 2017; Luketina et al., 2019) but for our proposed environment LANRO (Röder et al., 2022) involves a complex control setting of a robot with raw language commands that need to be parsed by the agent end-to-end, to ground the language based on sparse rewards. In comparison to former grid world settings (Chevalier-Boisvert et al., 2019) and environments without realistic physics (Hermann et al., 2017), our self-speech goal annotation can translate complex and extended goal-driven behaviors into lingual goals, autonomously.

For our reinforcement learning approach, we used goal-conditioned learning in hindsight. Herein, the goals were given as natural language embeddings. They were combined with the raw world state data by concatenating both representations in a neural deep reinforcement learning network. The particular method is described in (Röder et al., 2022; Röder & Eppe, 2022), and, in more detail, in the PhD thesis by Frank Röder (Röder, 2025) which has been submitted in February 2025.

Q2/WP2: Language-based neural semantic parser.

To address this second work package, we used a self-attention-based neural network (Transformer) (Vaswani et al., 2017) and introduced a new self-speech method that provides the system with implicit semantic grounding. We show that the self-attention-based neural network layer (Bahdanau et al., 2015) effectively parses general instructions, particularly CAR commands. These layers build word-to-word representations that help resolve misunderstandings. Our results show that the success signal of our RL approach is fully sufficient for semantic parsing, and no additional parser is needed. This was possible by coupling our method with language generation for self-speech, as described in the following.

Additional results and findings

1. Self-speech for semantic language grounding improves learning to follow and repair instructions. The main purpose of our language generation approach is not verbal communication with the instructor, but communication with the agent itself. We show that this enables the robot to ground language in action, which we consider our most important unexpected and surprising result. More precisely, we convert a trajectory of sensorimotor states into an abstract representation to generate the corresponding language command that matches the agent's behavior, as depicted in Figure 2 below.

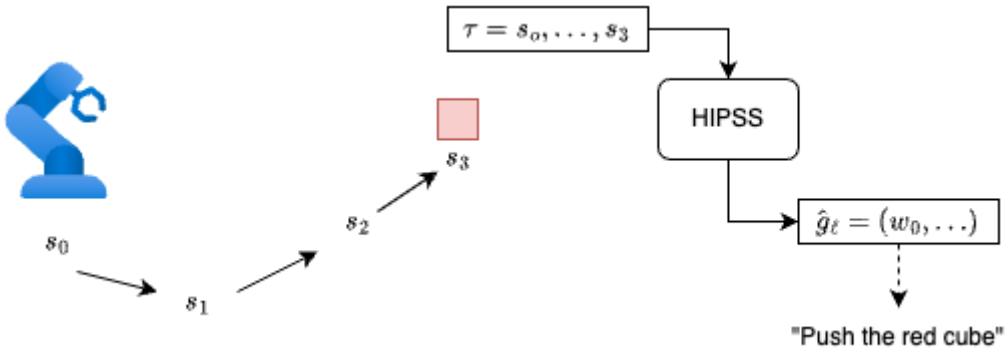

Fig. 2: The agent generates a trajectory of four states that our model HIPSS gets as input to generate an appropriate hindsight language goal, e.g, "Push the red cube".

Our proposed method HIPSS (Hindsight Instruction Prediction from State Sequences) not only is an approach to improve the sample efficiency of language-grounding in sparse reward tasks in RL but also improves the learning significantly by exploiting the compositional structure of language to artificially generate more experiences for rare situations the agent encountered in the environment. E.g., by knowing how an interaction with a "red cube" and a "green cylinder" looks like, the module can systematically generalize to generate out-of-

distribution samples of instructions containing “red cylinders” and “green cubes” without encountering them during the training phase.

Our method further underpins the idea of self-narration in child language development, where it improves language learning and grounding significantly. We assume that the systematic generalization of our egocentric speech approach is key to resembling the “vocabulary spurt” observed in early language learning of children (Plunkett et al., 1992).

2. Results in robotics

Our main research objective is intended to improve the interaction between humans and robots. Therefore, we require a robotic experimentation platform that we have significantly improved during this project.

Specifically, we have significantly improved our, adult-sized semi-humanoid robot, the Neuro Inspired COLlaborator (NICOL) (Kerzel et al., 2023). We developed our neural IK solver CycleIK to achieve competitive grasping accuracy at a very low runtime (Habekost et al., 2024). We also addressed sim-to-real transfer (Gäde et al., 2022; Gäde, Habekost, et al., 2024) and imitation learning (Spisak et al., 2024a, 2024b). Finally, we have developed ELMiRA, a dialog system based on large language models and visual language models (Allgeuer et al., 2024; Gäde, Özdemir, et al., 2024). ELMiRA combines pretrained ASR, TTS and object detection with robot specific visuospatial and motion models governed by a central dialogue manager. In general, ELMiRA can interact with any arbitrary object, provided the underlying action model can manipulate it. At the time of finishing the LeCAREbot project, the published architecture implemented rudimentary push, touch and point actions, but it provides an extensible modular framework to integrate our other findings and enable more complex interaction. Future work involves the integration of ELMiRA with HIPSS, our self-speech-based approach for language grounding described before.

3. The role of language grounding in intelligent problem-solving

Addressing misunderstandings and conversational action repair is very important for human-robot interaction, but it is a rather applied and practical research problem. In addition to the practical perspective on this problem, we expanded our view and asked, in how far language and language grounding plays a role in the general problem-solving abilities of intelligent agents. Our review and perspective published in Nature Machine Intelligence (Eppe et al., 2022) shows that compositional abstraction of sensorimotor data into abstract mental concepts is a fundamental mechanism required for transfer learning and one-shot problem-solving in biological and artificial agents. Language mirrors the compositional abstraction abilities of humans, by grounding sensorimotor experiences in an abstract compositional linguistic representation. Hence, our study underpins the importance of language, even if it is not used for communication.

3.3. Deviations from the original concept

We submitted our proposal in May 2019 and started the project in January 2020. Since then, there were many external factors, including the Corona pandemic and the extremely fast pace of new technologies around large language models and reinforcement learning that led to additional synergies and changes in our project design.

In WP1 of our proposal, we planned to use Embodied Construction Grammar (ECG) (Feldman, 2010) (Feldman et al., 2009) as a parsing tool to generate abstract symbolic representations to address the first research question of our proposal. This was meant a preparatory step for WP2 and the second research question, where we planned to substitute the ECG-related tools with a neural learning-based approach.

We planned to implement the ECG-related parts of WP1 in collaboration with our partner, Prof. Feldman from UC Berkeley. However, due to the Corona pandemic, we could not conduct our exchange visit as planned, which led to a delay in the work packages related to integrating ECG. Therefore, we directly moved on to WP2 and used attention-based Transformer architectures (Vaswani et al., 2017), without the intermediate ECG-based approach. As presented in our publications (Röder et al., 2022), it turned out that this method is very successful to address our main objective of enabling action repair, and that ECG is not necessary to achieve our main objective. Therefore, we continued with our focus on this approach. This decision was supported by the ongoing difficulties to conduct research exchange travels, and also by excessive work overhead that would have been problematic with using ECG because of the outdated software tools from our partners in Berkeley. Furthermore, LeCAREBot's association to the Active Self priority programme by the DFG offered very synergistic effects with our successful but unexpected focus on self-speech as a method to improve action repair dialogs, which is independent of ECG (see also Sec. 3.5).

3.4. Description of research data, methods and software, and how they are used to enhance the research quality

In LeCAREbot we developed and improved two software tools for our experiments: Our virtual robotics lab Scilab-RL and the physical NICO-L robot.

1. Scilab-RL: Our virtual robotics lab is a modular platform for integrating different robotic simulators and reinforcement learning algorithms, that are connected to a rich suite of data visualization and analysis tools (Benad et al., 2025). This includes reproducible experiments, statistical methods to measure performance, and benchmarking with other established approaches. Scilab-RL is intended to give researchers and students new to the field a quick start in developing new reinforcement learning methods. In LeCAREbot, we primarily added functionalities for language processing and rewards

conditioned on natural language goals. The project website and the code for Scilab-RL is freely available on github (https://scilab-rl.github.io/Scilab-RL/).

2. NICO-L: The Neuro Inspired COLlaborator is an adult-sized semi-humanoid robot mounted above a 100 x 200 cm² table environment, designed for research requiring both social interaction and physical collaboration abilities (Kerzel et al., 2023). Based on the Robot Operating System (ROS) middleware, the NICOL API provides a Python-based framework to control the hardware and access sensor data. It integrates the MoveIt planning framework as well as the Gazebo and CoppeliaSim simulation environments using a URDF model for compatibility with various other simulators and environments. More information on NICOL can be found on our website (https://www.inf.uni-hamburg.de/en/inst/ab/wtm/research/neurobotics/nicol.html). The software for the NICOL dialog manager is freely available on github (https://github.com/pallgeuer/chatty_robots), and the software for the ELMiRA architecture is available on our website (https://knowledgetechnologyuhh.github.io/ELMiRA/).

3.5. Scientific events and science communication

The project has been accepted as an associated project of the Active Self Priority programme, funded by the DFG. The Active Self is an interdisciplinary priority programme with researchers from various fields including Psychology, Biology, Mathematics, Cognitive Science, Mathematics, Computer Science, and Robotics. We used the infrastructure of the Active Self community for scientific exchange. This happened primarily during the bi-annual meetings of the programme, where all project members participated. In addition, we regularly participated in the programme's summer school to further foster the exchange and education of the PhD students.

3.6. Bibliography

(This part of the bibliography only contains external references. For literature authored by the project members, please consider Section 4 of this report.)

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *International Conference on Learning Representations (ICLR)*. http://arxiv.org/abs/1409.0473

Chevalier-Boisvert, M., Bahdanau, D., Lahlou, S., Willems, L., Saharia, C., Nguyen, T. H., & Bengio, Y. (2019). BabyAI: A Platform to Study the Sample Efficiency of Grounded Language Learning | OpenReview. *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=rJeXCo0cYX

Feldman, J. (2010). Embodied language, best-fit analysis, and formal compositionality. *Physics of Life Reviews*, *7*(4), 385–410. https://doi.org/10.1016/j.plrev.2010.06.006

Feldman, J., Bryant, J. E., & Dodge, E. (2009). Embodied Construction Grammar. In *The Oxford Handbook of Computational Linguistics* (pp. 38–111). Oxford University Press.

Gallagher, S. (2000). Philosophical Conceptions of the Self: Implications for Cognitive Science. *Trends in Cognitive Sciences*, *4*, 14–21.

Hermann, K. M., Hill, F., Green, S., Wang, F., Faulkner, R., Soyer, H., Szepesvari, D., Czarnecki, W. M., Jaderberg, M., Teplyashin, D., Wainwright, M., Apps, C., Hassabis, D., & Blunsom, P. (2017). *Grounded Language Learning in a Simulated 3D World*. https://arxiv.org/abs/1706.06551v2

Lúčny, A., Malinovská, K., & Farkaš, I. (2023). Robot at the Mirror: Learning to Imitate via Associating Self-supervised Models. *International Conference on Artificial Neural Networks (ICANN)*, *14254 LNCS*, 471–482. https://doi.org/10.1007/978-3-031-44207-0_39

Luketina, J., Nardelli, N., Farquhar, G., Foerster, J., Andreas, J., Grefenstette, E., Whiteson, S., & Rocktäschel, T. (2019). A Survey of Reinforcement Learning Informed by Natural Language. *International Joint Conferences on Artificial Intelligence (IJCAI)*, 6309–6317. https://doi.org/10.24963/ijcai.2019/880

Plunkett, K., Sinha, C., Møller, M. F., & Strandsby, O. (1992). Symbol Grounding or the Emergence of Symbols? Vocabulary Growth in Children and a Connecfionist Net. *Connection Science*, *4*(3–4), 293–312. https://doi.org/10.1080/09540099208946620

Surtees, A., Apperly, I., & Samson, D. (2013). Similarities and differences in visual and spatial perspective-taking processes. *Cognition*, *129*(2), 426–438. https://doi.org/10.1016/J.COGNITION.2013.06.008

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Conference on Neural Information Processing Systems (NIPS)*, 5998–6008. https://arxiv.org/pdf/1706.03762.pdf

4. Published Project Results

4.1. Publications with scientific quality assurance

(Open access publications are marked with blue font.)

Allgeuer, P., Ali, H., & Wermter, S. (2024). When Robots Get Chatty: Grounding Multimodal Human-Robot Conversation and Collaboration. *International Conference on Artificial Neural Networks (ICANN)*, 306–321. https://doi.org/10.1007/978-3-031-72341-4_21

Benad, J., Röder, F., & Eppe, M. (2025). Scilab-RL : A software framework for efficient reinforcement learning and cognitive modeling research. *SoftwareX*. https://doi.org/10.1016/j.softx.2025.102064

Eppe, M., Gumbsch, C., Kerzel, M., Nguyen, P. D. H., Butz, M. V, & Wermter, S. (2022). Intelligent problem-solving as integrated hierarchical reinforcement learning. *Nature Machine Intelligence*, *4*(1). https://doi.org/10.1038/s42256-021-00433-9

Gäde, C., Habekost, J.-G., & Wermter, S. (2024). Domain Adaption as Auxiliary Task for Sim-to-Real Transfer in Vision-based Neuro-Robotic Control. *International Joint Conference on Neural Networks (IJCNN)*. https://doi.org/10.1109/IJCNN60899.2024.10650722

Gäde, C., Kerzel, M., Strahl, E., & Wermter, S. (2022). Sim-to-Real Neural Learning with Domain Randomisation for Humanoid Robot Grasping. *International Conference on*

Artificial Neural Networks (ICANN), *13529 LNCS*, 342–354. https://doi.org/10.1007/978-3-031-15919-0_29

Gäde, C., Özdemir, O., Weber, C., & Wermter, S. (2024). Embodying Language Models in Robot Action. *European Symposium on Artificial Neural Networks*, 625–630. https://www.esann.org/sites/default/files/proceedings/2024/ES2024-143.pdf

Habekost, J.-G., Gaede, C., & Allgeuer Philipp and Wermter, S. (2024). Inverse Kinematics for Neuro-Robotic Grasping with Humanoid Embodied Agents. *International Conference on Intelligent Robots and Systems (IROS)*, 7315–7322. https://doi.org/10.1109/IROS58592.2024.10802010

Kerzel, M., Allgeuer, P., Strahl, E., Frick, N., Habekost, J.-G., Eppe, M., & Wermter, S. (2023). NICOL: A Neuro-Inspired Collaborative Semi-Humanoid Robot That bridges Social Interaction and Reliable Manipulation. *IEEE Access*, *11*, 123531–123542. https://doi.org/10.1109/ACCESS.2023.3329370

Röder, F., & Eppe, M. (2022). Language-Conditioned Reinforcement Learning to Solve Misunderstandings with Action Corrections. *Second Workshop on Language and Reinforcement Learning @ NeurIPS*. https://openreview.net/forum?id=lWd0qiv9E-

Röder, F., Eppe, M., & Wermter, S. (2022). Grounding Hindsight Instructions in Multi-Goal Reinforcement Learning for Robotics. *International Conference on Development and Learning (ICDL)*. https://doi.org/10.48550/arxiv.2204.04308

Spisak, J., Kerzel, M., & Wermter, S. (2024a). Diffusing in Someone Else's Shoes: Robotic Perspective Taking with Diffusion. *International Conference on Humanoid Robots (Humanoids)*, 141–148. https://doi.org/10.1109/Humanoids58906.2024.10769830

Spisak, J., Kerzel, M., & Wermter, S. (2024b, June). Robotic Imitation of Human Actions. *IEEE International Conference on Development and Learning (ICDL)*. https://doi.org/10.1109/ICDL61372.2024.10644215

4.2. Other publications and published results

Frank Röder. (2025). *Language Grounding in Deep Reinforcement Learning for Dynamic Goal-Oriented Robotics* [Ph.D. thesis (in review)]. Hamburg University of Technology.

4.3 Patents (applied for and granted)

The project had a focus on foundational science, and applications were secondary. Therefore, we did not apply for any patents.