dr hab. inż. Krzysztof Dembczyński
Instytut Informatyki (Institute of Computing Science)
Politechnika Poznańska (Poznań University of Technology)
ul. Piotrowo 2, 60-965 Poznań
tel: (+48) 61 665 2936
kdembczynski@cs.put.poznan.pl

January 31, 2026

Review of the Doctoral Dissertation

Title: Subgoal Search and Efficient Exploration in Reinforcement Learning
Author: Michał Zawalski

1 A short summary of the thesis

The dissertation concerns the problem of Reinforcement Learning (RL), aiming to address challenges such as low sample efficiency, limited generalization, and poor interpretability. The author introduces and discusses the integration of different structured reasoning and planning mechanisms into RL frameworks, drawing inspiration from human cognitive capabilities. The dissertation was submitted as a collection of five scientific articles preceded by an extended summary of contributions. Both the summary and articles are written in English. The list below contains bibliographical information concerning the articles:

- [P1] Zawalski, M., Osiński, B., Michalewski, H., and Miłoś, P. (2022). Off-policy correction for multi-agent reinforcement learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1774–1776
- [P2] Czechowski, K., Odrzygóźdź, T., Zbysiński, M., Zawalski, M., Olejnik, K., Wu, Y., Kuciński, Ł., and Miłoś, P. (2021). Subgoal search for complex reasoning tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 624–638
- [P3] Zawalski, M., Tyrolski, M., Czechowski, K., Odrzygóźdź, T., Stachura, D., Piękos, P., Wu, Y., Kuciński, Ł., and Miłoś, P. (2023). Fast and precise: Adjusting planning horizon with adaptive subgoal search. In *The Eleventh International Conference on Learning Representations*
- [P4] Zawalski, M., Chen, W., Pertsch, K., Mees, O., Finn, C., and Levine, S. (2025a). Robotic control via embodied chain-of-thought reasoning. In Agrawal, P., Kroemer, O., and Burgard, W., editors, *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 3157–3181. PMLR
- [P5] Zawalski, M., Góral, G., Tyrolski, M., Wiśnios, E., Budrowski, F., Cygan, M., Łukasz Kuciński, and Miłoś, P. (2025b). What matters in hierarchical search for combinatorial reasoning problems? *Submitted to Transactions on Machine Learning Research*

The central thesis of the dissertation is that RL systems can efficiently recognize abstract problem structures, and conduct interpretable and structured reasoning. The introduced methods are original, solving different challenging tasks such as multi-agent RL, robotic control, combinatorial puzzles like Rubik's Cube and Sokoban, or inequality theorem proving. The extended summary discusses two contributions of the author. The first one involves hierarchical search algorithms that operate on high-level concepts to solve complex reasoning tasks by decomposing them into manageable subgoals. The second contribution is robotic control based on Embodied Chain-of-Thought (ECoT), which enables policies to reason explicitly through tasks using multi-step textual reasoning grounded in multimodal perception. Interestingly, the summary does not mention any results on multi-agent RL published in [P1]. Below, I shortly summarize all five articles being the core of the thesis.

Article [P1] introduces MA-Trace, a new on-policy actor-critic algorithm that utilizes an importance sampling mechanism, derived from V-Trace,[1] to provide off-policy corrections for training data. As a theoretical contribution, the authors prove a fixed-point theorem ensuring convergence. Notably, the authors identify a specific technical gap in the original V-Trace convergence proof and correct it within their own result.

The second article [P2] introduces the Subgoal Search method, a hierarchical search algorithm for solving combinatorial tasks by using a high-level policy to generate subgoals at a fixed distance k. [P3] extends this approach to use an adaptive horizon instead of the constant k-step to dynamically adjust to problem complexity. The work on hierarchical search algorithms is further continued in [P5], which investigates their key characteristics to determine under which conditions these methods perform well.

Article [P4] addresses the robotic control problem by introducing ECoT. Unlike standard models that map images directly to actions, ECoT requires the model to generate an explicit natural language reasoning trace, based on visual and physical cues, before predicting an action. This reasoning process grounds the high-level task in the physical environment, ultimately outputting precise low-level control actions.

2 General remarks and comments

The dissertation addresses highly relevant and contemporary issues in the field of Artificial Intelligence. The author demonstrates a deep understanding of the limitations of modern RL, such as the requirement for millions of interactions and brittleness of pattern-matching policies. The thesis introduces several key algorithmic contributions that significantly advance the state-of-the-art:

- The extension of V-Trace to multi-agent settings, which is a significant contribution for scaling on-policy RL algorithms in distributed environments, ensuring that off-policy data from multiple actors can be integrated stably.
- The novel hierarchical algorithms for subgoal search that can adaptively decide when to search for subgoals, allowing the agent to allocate computational budget where the transition uncertainty is highest. The experimental results show very promising performance compared to competitive baselines.
- The ECoT-based robotic control that combines an explicit natural language reasoning with precise low-level control actions of a robot in a physical environment. This last contribution is particularly innovative, as it provides a mechanism for human operators to understand and even correct a robot's reasoning through natural language. This advances the goal of building "Explainable AI" in physical embodiments.

The contribution of the thesis is primarily conceptual and empirical. While efforts were made in [P1] and [P5] to support algorithmic solutions with theoretical justifications, these results are surprisingly not emphasized in the extended summary of author's contributions.

The publications included in the thesis are of high quality, appearing in prestigious venues such as AAMAS, NeurIPS, ICLR, and CoRL. Particularly noteworthy is the ICLR 2023 publication, which was recognized as a **Notable Top 5%** paper. The author's contributions across these papers are substantial, ranging from 15% to 60%. Unfortunately, the status of article [P5] is unclear. It seems it has not yet been published. This fact should be properly emphasized in the submitted thesis, not only by marking it as *Submitted to: Transactions on Machine Learning Research* in the list of publications. I need to state that it is somewhat uncommon to include an unpublished paper in a collection-based thesis.

[1] Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., Legg, S., and Kavukcuoglu, K. (2018). Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. *Proceedings of the 35th International Conference on Machine Learning*, 80:1407–1416

Another weakness of the submitted thesis is its extended summary which lacks discussion on [P1] and ignores the obtained theoretical results. Furthermore, the description of the Subgoal Search is on a very high-level making it very hard to understand. The attached diagram in Figure 3.1 and the pseudocode in Algorithm 1 do not help much. Only by reading the original papers, the concept behind this contribution becomes fully clear. I would expect the extended summary to contain more details (but still presented in a concise way) on training data preparation and learning procedures. Basically, the subgoal generation seems to be reduction to a supervised learning on expert data consisting of ordered state pairs (s_i, s_j), $i < j$, with the objective to accurately predict action a taken in state s_i that led to future state s_j. Details of this kind should be included in the summary.

Personally, I prefer dissertations that are in the form of a cohesive book that synthesizes the achievements of a doctoral student. If a student decides to submit a collection of articles preceded by a extended summary, this summary should provide a self-contained picture, appropriately combining the attached articles into a single and consistent story. In this particular case, the summary could try to place a common umbrella over the diverse achievements of the author. Unfortunately, this is only partly done in the introduction and conclusion chapters. I am missing a separate section that discusses a general AI framework consisting of planning, human interaction, and the autonomous execution of produced plans. With such a general framework, the particular contributions of the author could be presented as specific instances of it. Otherwise, the result is only a collection of loosely connected articles without achieving the social goal of PhD dissertations, which is disseminating and promoting knowledge.

3 Specific comments and questions

The list below contains specific comments and questions, some of which are directly connected to my general remarks presented in the previous section. It also includes a few minor issues.

- What is the current status of article [P5]? Is it still in the review process for Transactions of Machine Learning Research?
- Why are the contributions of [P1] regarding multi-agent reinforcement learning completely ignored in the summary?
- Similarly, why theoretical results from [P1] and [P5] are not emphasized and discussed in the summary?
- If the author were to create a common procedural umbrella over the different achievements collected in this thesis, how could it be presented in the form of high-level pseudocode?
- To what extent does the requirement for generating multi-step textual reasoning chains increase inference latency compared to traditional image-to-action models, and how can this overhead be mitigated for high-frequency control tasks beyond the approaches discussed in [P4]?
- Several future directions have been mentioned in the concluding section of the summary. One of them is applying Subgoal Search to robotics. What would be the connection between such an approach and the one based on ECoT introduced in [P4]?
- How do the results achieved in this work correspond to current advances in "System 2 AI" designed for deliberate and analytical reasoning?
- It appears that references, Collaboration et al., 2024 and Embodiment Collaboration et al., 2024, are duplicates. Furthermore, citations of the from "Open X-Embodiment Collaboration et al. [...]" should be preferred as they are more informative.
- While general discussions on search algorithms can be referenced using citation to the seminal book of Russel and Norvig, references to specific algorithms should point to the original papers, for example, the A* algorithm should be cited as Hart et al. (1968).[2]

It would be great if the above questions (besides the last two) could be answered during the public defense.

[2] Hart, P. E., Nilsson, N. J., and Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107

4 Final Conclusion

The presented dissertation provides impressive results across a wide spectrum of problems related to structured reasoning in Reinforcement Learning. The author has demonstrated exceptional research intuition and extensive knowledge across diverse areas such as machine learning, artificial intelligence, and robotics. In my opinion, the thesis meets all the requirements for a doctoral degree in computer science. I recommend that mgr Michał Zawalski be admitted to the further stages of the doctoral proceedings.

While I am highly impressed by the obtained scientific results, I must also note my disappointment regarding the final form of the submitted thesis and the structural issues mentioned in the previous sections. To properly balance these factors, I have decided not to directly recommend the thesis for distinction at this time. However, I remain open to supporting such a motion should the other reviewers reach that conclusion based on their evaluations.

dr hab. inż. Krzysztof Dembczyński