Generalization in Reinforcement Learning from Logical Specifications

Vignesh Subramanian[1]

[1]School of Computer Science, Georgia Institute of Technology, Atlanta, Georgia, USA

Abstract

Reinforcement learning policies often fail to generalize beyond the specific tasks on which they are trained, and this limitation becomes especially severe in long-horizon settings where success depends on satisfying structured constraints over time. In these settings, task objectives are naturally compositional, reward signals are sparse and delayed, and existing approaches often lack both a clear target notion of generalization and a principled way to evaluate whether the underlying generalization pattern has been learned. This dissertation studies how logical structure can be used to address these challenges.

The central research goal is to develop a framework for *reliable long-horizon generalization* in reinforcement learning from temporal-logic specifications. The key idea is to use logical specifications not only to describe complex tasks, but also to define structured families of related tasks in which generalization can be formalized, learned, and eventually verified. In this setting, tasks evolve according to a fixed update rule, and the objective is to train on a small subset of task instances and produce policies that succeed zero-shot on unseen ones. This perspective opens up several linked research directions such as learning compact policy-evolution rules that capture transfer across task families, scaling such learning to long-horizon problems through decomposition and stable supervision, and developing certificate-based tools that move evaluation beyond empirical rollout success toward more structured notions of correctness.

In our current progress, we have formalized inductive generalization over specification-defined task families, developed a scalable decoupled approach for learning policy-evolution templates, and introduced certificate-based methods for evaluating and diagnosing generalization. Taken together, these components motivate a broader thesis program aimed at integrating specification-guided task structure, compositional planning, and certificate-guided reasoning into a unified approach for scalable and reliable reinforcement learning generalization.

Keywords

reinforcement learning, generalization, long-horizon tasks, temporal-logic specifications, inductive task families, zero-shot transfer, policy evolution templates, certificates

1. Introduction and Problem Statement

Reinforcement learning has achieved strong performance on individual control tasks, but policies learned in this way often fail to transfer even under small changes in the task [1, 2, 3]. This weakness becomes more severe in long-horizon settings, where success depends on satisfying structured objectives over many stages of execution [4, 5]. In such problems, reward signals are often sparse and delayed, and learning can become highly brittle when the agent must coordinate safety, sequencing, and branching requirements over time. As a result, both *generalization* and *long-horizon reasoning* remain central obstacles to deploying reinforcement learning in structured domains [3, 6].

A further difficulty is that generalization in reinforcement learning is often studied through loosely related benchmark variations, such as changed goals, shifted initial states, modified layouts, or altered dynamics [1, 2, 7]. While these settings are all useful, they do not by themselves provide a precise target notion of what kind of transfer should be learned. This dissertation focuses on a more structured setting in which tasks are not arbitrary variations, but members of a family connected by a systematic rule. The goal is to identify forms of task variation for which generalization is both meaningful and learnable.

Doctoral Consortium of the 23rd International Conference on Principles of Knowledge Representation and Reasoning (KR 2026 DC), July 20-23, 2026, Lisbon, Portugal

vignesh@gatech.edu (V. Subramanian)

https://vigneshs10.github.io/personal-website/ (V. Subramanian)

0009-0004-1088-6981 (V. Subramanian)

Our framework uses temporal-logic specifications to represent long-horizon tasks [4, 5, 8]. Logical specifications are useful in this setting because they expose the compositional structure of a task, such as reachability, safety, sequencing, and disjunction, in a way that is difficult to express cleanly through rewards alone. They also make it possible to define families of related tasks that share the same high-level structure while differing in the instantiation of goals, predicates, or initial conditions. This leads to a concrete notion of *inductive task families*, where tasks are indexed and evolve according to a fixed update rule [9].

The central problem studied is to learn from a small subset of tasks in a structured family and produce policies that succeed on the remaining unseen tasks without additional interaction. This is a zero-shot generalization problem over structured task families. The key hypothesis is that when tasks evolve systematically, their policies should also evolve systematically [9]. Rather than learning each task from scratch, we can instead attempt to learn a compact *policy-evolution rule* that captures how the policy changes as the task index changes. This perspective gives rise to two main research questions:

- **RQ1.** How can structured generalization be formalized and learned in a way that scales to long-horizon tasks?
- **RQ2.** How can we evaluate whether a learned policy family has truly generalized beyond the observed training instances?

Our recent work takes initial steps toward RQ1 by formalizing inductive generalization from specifications and by developing a scalable decoupled framework for learning policy-evolution templates [9]. It also takes initial steps toward RQ2 by developing certificate-based methods for evaluating and diagnosing generalization behavior beyond standard rollout success rates. These directions support the thesis that logical structure can be used not only to specify long-horizon reinforcement learning tasks, but also to define, learn, and eventually verify meaningful forms of generalization across task families. The remainder of this paper outlines this research direction, summarizes progress to date, and describes planned work on scalable long-horizon planning and certificate-guided generalization.

2. Background and Setting

We study reinforcement learning modeled as Markov decision processes (MDPs) [10]. An MDP is given by (S, A, P, η), where S is the state space, A is the action space, P denotes the transition dynamics, and η is the initial-state distribution. The agent interacts with the environment through trajectories and aims to learn a policy that maximizes task success. In the settings considered here, the dynamics are not known in advance, so policies must be learned from sampled interaction.

The tasks of interest are long-horizon and structurally rich. Instead of describing them only through scalar reward functions, we represent them using compositional logical specifications [4, 11, 12, 8]. More precisely, throughout this proposal I use the phrase "temporal-logic specification" in a broad sense to refer to task specifications with temporal structure. The concrete specification language used in the current work is a SPECTRL-style compositional language, rather than standard LTL or LTLf syntax. This language contains operators for reachability, safety, sequential composition, and disjunction, and is well suited for expressing long-horizon reinforcement learning tasks such as reach-avoid, sequencing, and branching objectives [4, 5].

This distinction is important. Standard LTL and LTLf provide temporal operators such as $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{U}$, while the SPECTRL-style syntax used here includes program-like constructs such as achieve, ensuring, and sequential composition. Thus, the formulas in this proposal should not be read as literal LTL formulas. Rather, they are compositional task specifications inspired by temporal logic and dynamic-logic-style task composition. At a high level, the framework only requires a specification template whose predicates or regions can be re-grounded across task indices; in principle, similar ideas could be instantiated with other finite-trace specification formalisms. In the present work, however, the examples and experiments use the SPECTRL-style language.

A reach-avoid task requires the agent to reach a goal set while staying within a safe region [5, 12]. Although the full specifications considered in our experiments are often more complex, they can be decomposed into simpler reach-avoid components [5, 9]. This decomposition is important because it exposes the structure of the task and provides a basic unit on which generalization can be studied more systematically.

Using this specification-based view, we consider families of related tasks that share the same high-level logical form but differ in their low-level grounding. For example, a family may keep the same sequencing structure while shifting the goal regions, the initial conditions, or the relevant predicates with the task index. This structured variation is the setting in which we define and study inductive generalization in the next section.

Relation to logic-guided reinforcement learning. This dissertation is closely related to the growing literature on logic-guided reinforcement learning. Reward machines expose the internal structure of non-Markovian reward functions using automaton-like representations, enabling reward shaping, task decomposition, and more sample-efficient learning [13]. Related work on LTL and formal-language-based reward specification studies how temporal and regular-language specifications can be translated into reward machines or other structured reward representations for reinforcement learning [14]. Temporal-logic reward approaches also use logical specifications, such as truncated linear temporal logic, together with quantitative semantics to define reward signals for complex robotic tasks [15]. In a complementary direction, shielding methods use formal verification to restrict unsafe actions during exploration or deployment, thereby enforcing safety constraints while the reinforcement learning algorithm optimizes its performance objective [16]. My work shares the broader goal of using logical structure to make reinforcement learning more reliable. However, the primary focus is different: rather than using logic only to specify, shape, or shield rewards for a single task, I use logical specifications to define indexed families of related tasks and study zero-shot generalization across the family. In this setting, the specification is not only a task description, but also the structure that induces the generalization problem.

3. Research Vision: Inductive Generalization as a Target Notion

Generalization in reinforcement learning can refer to many kinds of transfer, including new goals, new initial conditions, new obstacles, or new dynamics [3, 1, 2]. Our research focuses on a more structured setting in which tasks are not arbitrary variants, but members of a family connected by a fixed update rule. This makes the target notion of generalization more precise and also gives structure that learning algorithms can exploit. Formally, we study an indexed family of tasks

$$R = \{R_i\}_{i=0}^{L},$$

where each task R_i consists of an MDP M_i together with a SPECTRL-style compositional task specification ϕ_i. The family is *inductive* if all task instances share the same high-level specification structure and differ only in the instantiation of predicates or the initial state distribution [9]. Equivalently, there exists a fixed specification template $\phi(\cdot)$ and a sequence of predicate instantiations $\{b_i\}_{i=0}^{L}$ such that

$$\phi_i = \phi(b_i),$$

where the grounding evolves with the task index according to a structured update rule. This captures settings where the logical form of the task remains the same, while the low-level task parameters shift predictably across indices.

As one concrete example, consider an indexed family of reach-avoid tasks in which the agent starts from an index-dependent initial region $Init_k$, must reach an index-dependent goal region $goal_k$, and may first pass through one of two intermediate regions g_1 or g_2 while always avoiding obstacles. The corresponding specification can be written as

$$\varphi_k \triangleq \Big((\text{achieve reach}(g_1)) \text{ or } (\text{achieve reach}(g_2))\Big); \text{achieve reach}(\text{goal}_k) \text{ ensuring } S.$$

where S denotes the safe region. Here, the compositional structure is fixed, while $Init_k$ and $goal_k$ evolve with k. Given such a family, the inductive generalization problem is the following. Let

$$Train \subseteq \{0, \ldots, L\}, \qquad Unseen = \{0, \ldots, L\} \setminus Train,$$

with $0 \in Train$ and typically $L \in Train$. The goal is to train only on the tasks indexed by $Train$ and produce policies for indices in $Unseen$ without any further environment interaction. In other words, the objective is zero-shot transfer across a structured indexed family [9].

The key hypothesis is that if tasks evolve systematically, then their policies may also evolve systematically [9]. Let π_i denote the policy for task R_i. Prior work models this through a higher-order policy-evolution rule

$$\pi_{i+1} = \Omega(\pi_i),$$

where $\Omega : \Pi \to \Pi$ maps the policy of one task instance to the policy of the next. Since learning a completely general higher-order map is difficult, Ω is approximated by an m-degree polynomial template

$$\Omega(\pi_i) = \kappa_m \odot [\pi_i]^m + \cdots + \kappa_1 \odot [\pi_i] + \kappa_0,$$

where $\kappa_0, \ldots, \kappa_m$ are template coefficients, $\odot$ denotes elementwise multiplication, and $[\pi_i]^r$ denotes elementwise powers of the policy parameter vector [9].

This reduces inductive generalization to learning the template coefficients. More precisely, given an inductive family of reach-avoid tasks

$$R = \{R_i\}_{i=0}^{L},$$

a training index set

$$Train = \{i_1 < \cdots < i_n\},$$

and a base policy π_0 for the base task R_0, the central learning problem is

$$\kappa_0^*, \ldots, \kappa_m^* \in \arg\max_{\kappa_0, \ldots, \kappa_m} \sum_{i \in Train} \Pr[\pi_i \models R_i],$$

where each π_i is obtained by unrolling the κ-template from π_0 [9]. This formulation turns generalization into a concrete prediction problem over task families, rather than an informal hope that function approximation will transfer.

This view is attractive for two reasons. First, it gives a precise target notion of zero-shot generalization across structured task families. Second, it aligns naturally with temporal-logic specifications, since the same logical structure that defines a family also constrains how policies should evolve. Our earlier work introduced this formulation and showed that it can support zero-shot transfer on long-horizon tasks derived from logical specifications [9]. At the same time, it exposed an important bottleneck: learning the policy-evolution template becomes increasingly unstable as the number of training indices grows. This motivates the scalable learning framework described in the next section.

4. Progress to Date I: Scalable Learning of Policy-Evolution Templates

The formulation in the previous section reduces inductive generalization to learning the template coefficients $\kappa_0, \ldots, \kappa_m$. Our recent work takes an initial step toward making this learning problem scalable on long-horizon tasks. The main difficulty is that, in the original GenRL framework, learning the template is tightly coupled with reinforcement learning itself. Starting from a base policy π_0, the current template is unrolled to generate policies for the training indices, these generated policies are executed on their tasks, and the template coefficients are updated using reward feedback aggregated across all training tasks [9]. As the number of training indices grows, this loop becomes unstable because the reward signals become noisy and conflicting, while the generated policies are still weak and changing.

This creates two main problems. First, the quality of the template update depends on the quality of the policies currently produced by the template. If those policies are poor early in training, the resulting reward feedback is also poor, which leads to weak updates to κ. Second, the learning procedure is closely tied to the RL backbone used inside the loop. In practice, prior work relies on ARS-style optimization [17, 9], and when this backbone struggles on harder long-horizon tasks, template learning also degrades. This limits how well the method scales as the training set grows.

To address this, we developed a decoupled framework that separates per-index policy learning from template learning. Instead of learning κ directly from multi-task reward aggregation, the method first learns a strong policy for each training index and then uses those policies to train the shared template. This changes the problem from unstable reward-based learning of the template into a more stable supervised learning problem.

In the first stage, for each training index $i \in Train$, the method learns a strong task-specific teacher policy $\hat{\pi}_i$ for task R_i. To make these teacher policies easier to fit with one shared template, neighboring teachers are encouraged to stay close to one another in parameter space. After training these teachers, the method builds a dataset of candidate states for each index. These states are collected from several sources, including successful teacher rollouts, replay or exploration states, and reset or randomized initial states. This gives broader coverage than using only successful rollouts.

However, not all of these candidate states are reliable for supervision. Some lie in regions where the teacher may not provide stable or useful labels. To address this, the method uses a confidence model that scores whether a state is suitable for training. Only high-confidence states are kept, and those states are labeled with the teacher's action. This produces a filtered dataset D_i^e for each training index.

In the second stage, the template coefficients are learned by fitting the template-generated policies directly to these teacher-labeled datasets. In other words, the learned template is trained so that the policy it produces at each training index matches the action choices of the corresponding teacher. This is much simpler than the original coupled RL loop, since the template is no longer trained through noisy aggregated rewards, but through stable supervision from strong per-index experts.

This decoupling changes the role of reinforcement learning in the overall framework. RL is used only to learn strong local experts, while the cross-index policy-evolution rule is learned separately through supervised fitting. Empirically, this leads to much better training scalability and stronger zero-shot generalization across several long-horizon benchmarks. It also makes the framework less brittle with respect to the choice of RL backbone, since stronger single-task learners can be used to train the teachers without changing the template-learning objective.

5. Progress to Date II: Evaluating Generalization with Certificates

While the previous section focuses on how to learn policies that generalize across structured task families, an equally important question is how to *evaluate* whether such generalization is actually correct. In most reinforcement learning work, generalization is assessed through rollout-based success rates on held-out tasks [1, 2, 3]. Although this is useful, it gives only a coarse empirical picture. A policy may succeed on some unseen tasks and fail on others, but rollout success alone does not explain *why* generalization holds, where it breaks, or whether there is a more principled way to reason about progress across a family of tasks.

Our recent work takes initial steps toward this problem through *certificates of correct generalization*. The goal is to learn a function that captures structured progress across demonstrations from multiple task instances and can be used to evaluate whether the observed behavior is consistent with correct generalization. Intuitively, such a certificate should be positive on safe states, should decrease as the agent makes progress within a task, should relate the completion of one task to the start of the next task in the family, and should become negative at unsafe states. This gives a notion of progress that is shared across task indices, rather than being tied to a single rollout outcome. Formally, for a family of tasks and a set of demonstration trajectories, the certificate is a function

$$\mathcal{C} : S \times \mathbb{N}_0 \to \mathbb{R},$$

where $\mathcal{C}(s, i)$ measures progress at state s for task index i. The desired properties are that $\mathcal{C}$ remains non-negative on safe states visited along demonstrations, decreases along successful trajectories within each task, decreases across consecutive task instances, and is negative on unsafe states. At a high level, these conditions turn the certificate into a shared progress measure over the indexed family.

This provides a more structured way to evaluate generalization than simply measuring average success over unseen tasks. If a learned policy family generalizes correctly, then one expects its trajectories to exhibit a consistent notion of progress across task indices. Certificates also provide diagnostic information when generalization fails. Violations of the desired monotonicity or safety conditions can reveal whether failure comes from poor within-task progress, poor transfer across indices, or unsafe behavior. In this sense, the certificate acts not only as an evaluation tool, but also as a lens for understanding the geometry of generalization.

The current results in this direction focus on using certificates to assess and compare generalization behavior. A longer-term goal is to move beyond post hoc evaluation and use certificates inside the learning loop itself. This could enable a learner–verifier style framework in which certificate violations produce counterexamples or refinement signals for the policy-evolution template. Such a framework would make generalization not only more measurable, but eventually more reliable.

6. Proposed Dissertation Plan

The goal of the remaining dissertation is to build a broader framework for reliable long-horizon generalization in reinforcement learning. So far, the work has focused on defining inductive generalization, improving the scalability of learning policy-evolution rules, and developing certificate-based tools for evaluation. The next steps extend these ideas into a more complete learning and reasoning framework.

Long-horizon planning over generalizable subtasks. Many generalizable reinforcement learning methods work best on short-horizon tasks [18, 19, 20], while real tasks are often much longer horizon and contain multiple stages or branching choices. We plan to develop a long-horizon planner that breaks a complex task into smaller subtasks and applies a generalizable reinforcement learning method to each subproblem. The main question is whether these locally generalizable policies can be composed into a globally successful long-horizon solution.

Richer policy-evolution models. The current framework uses a polynomial template, which is a useful starting point but may be too limited for more complex task families [9]. In particular, branching behavior, mode switches, and sharper nonlinear changes across indices may require more expressive update rules. We plan to study richer evolution mechanisms that can better capture these structured changes while preserving zero-shot transfer across the task family.

Certificate-guided refinement. The current certificate framework helps assess and diagnose whether a policy family exhibits consistent progress and safety across task indices. The next step is to incorporate these ideas into a learner–verifier loop, where certificate violations guide refinement of the learned policy-evolution rule. This could make the overall framework more reliable by combining empirical learning with structured feedback about where generalization fails.

Together, these directions aim to move from isolated examples of zero-shot transfer toward a more general framework for long-horizon reinforcement learning generalization from specifications.

7. Conclusion

This doctoral work studies how to make reinforcement learning generalize more reliably in long-horizon settings. The central idea is to use temporal-logic specifications not only to describe complex tasks, but also to define structured families of related tasks in which generalization can be studied in a precise

way [4, 9]. This leads to inductive generalization as a target notion, where the goal is to train on a small subset of task indices and produce policies that succeed zero-shot on unseen indices [9].

The work completed so far takes initial steps toward this goal along two directions. First, it develops scalable methods for learning policy-evolution rules across structured task families, including a decoupled framework that improves stability by separating per-task policy learning from shared template learning. Second, it develops certificate-based methods for evaluating and diagnosing generalization beyond aggregate rollout success. Together, these results suggest that logical structure can play a useful role in both learning and evaluating generalization.

The remaining dissertation work will build on these foundations by studying long-horizon planning over generalizable subtasks, richer policy-evolution mechanisms, and tighter integration of certificates into learning and refinement. More broadly, the goal is to move toward a framework in which reinforcement learning policies can generalize across structured task families in a way that is scalable, interpretable, and eventually more reliable.

Acknowledgements

I am grateful to my advisor, Prof. Suguman Bansal at Georgia Institute of Technology, for her guidance and support during my Ph.D. Her mentorship has been central to shaping my research direction and the ideas presented in this doctoral consortium paper. I also sincerely thank Prof. Subhajit Roy at IIT Kanpur and Prof. Djordje Zikelic at Singapore Management University for their valuable discussions, feedback, and collaboration on the research projects described here. Their insights have helped refine these works and have influenced the directions outlined in this paper.

Generative AI Statement

Generative AI tools were used only for language-level assistance during the preparation of this manuscript, such as grammar correction, spelling correction, and minor phrasing improvements. They were not used to generate the research ideas, technical content, results, analysis, or conclusions of the paper. All scientific claims, formulations, and final manuscript content were reviewed and verified by the author.

References

[1] K. Cobbe, O. Klimov, C. Hesse, T. Kim, J. Schulman, Quantifying generalization in reinforcement learning, in: International conference on machine learning, PMLR, 2019, pp. 1282–1289.

[2] K. Cobbe, C. Hesse, J. Hilton, J. Schulman, Leveraging procedural generation to benchmark reinforcement learning, in: International conference on machine learning, PMLR, 2020, pp. 2048–2056.

[3] R. Kirk, A. Zhang, E. Grefenstette, T. Rocktäschel, A survey of zero-shot generalisation in deep reinforcement learning, Journal of Artificial Intelligence Research 76 (2023) 201–264.

[4] K. Jothimurugan, R. Alur, O. Bastani, A composable specification language for reinforcement learning tasks, Advances in Neural Information Processing Systems 32 (2019).

[5] K. Jothimurugan, S. Bansal, O. Bastani, R. Alur, Compositional reinforcement learning from logical specifications, Advances in Neural Information Processing Systems 34 (2021) 10026–10039.

[6] T. Hospedales, A. Antoniou, P. Micaelli, A. Storkey, Meta-learning in neural networks: A survey, IEEE transactions on pattern analysis and machine intelligence 44 (2021) 5149–5169.

[7] C. Zhang, O. Vinyals, R. Munos, S. Bengio, A study on overfitting in deep reinforcement learning, arXiv preprint arXiv:1804.06893 (2018).

[8] R. Alur, S. Bansal, O. Bastani, K. Jothimurugan, A framework for transforming specifications in reinforcement learning, in: Principles of Systems Design: Essays Dedicated to Thomas A. Henzinger on the Occasion of His 60th Birthday, Springer, 2022, pp. 604–624.

[9] V. Subramanian, R. Kushwah, S. Roy, S. Bansal, Inductive generalization in reinforcement learning from specifications, in: International Symposium on Automated Technology for Verification and Analysis, Springer, 2025, pp. 277–298.
[10] R. S. Sutton, A. G. Barto, et al., Reinforcement learning: An introduction, volume 1, MIT press Cambridge, 1998.
[11] G. De Giacomo, L. Iocchi, M. Favorito, F. Patrizi, Foundations for restraining bolts: Reinforcement learning with LTLf/LDLf restraining specifications, in: Proceedings of the International Conference on Automated Planning and Scheduling, volume 29, 2019, pp. 128–136.
[12] M. Hasanbeig, A. Abate, D. Kroening, Logically-constrained reinforcement learning, arXiv preprint arXiv:1801.08099 (2018).
[13] R. T. Icarte, T. Q. Klassen, R. Valenzano, S. A. McIlraith, Reward machines: Exploiting reward function structure in reinforcement learning, Journal of Artificial Intelligence Research 73 (2022) 173–208.
[14] A. Camacho, R. T. Icarte, T. Q. Klassen, R. A. Valenzano, S. A. McIlraith, Ltl and beyond: Formal languages for reward function specification in reinforcement learning., in: IJCAI, volume 19, 2019, pp. 6065–6073.
[15] X. Li, C.-I. Vasile, C. Belta, Reinforcement learning with temporal logic rewards, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2017, pp. 3834–3839.
[16] B. Könighofer, R. Bloem, S. Junges, N. Jansen, A. Serban, Safe reinforcement learning using probabilistic shields, in: International Conference on Concurrency Theory: 31st CONCUR, 2020.
[17] H. Mania, A. Guy, B. Recht, Simple random search of static linear policies is competitive for reinforcement learning, Advances in neural information processing systems 31 (2018).
[18] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: International conference on machine learning, PMLR, 2017, pp. 1126–1135.
[19] L. Zintgraf, K. Shiarlis, M. Igl, S. Schulze, Y. Gal, K. Hofmann, S. Whiteson, Varibad: A very good method for bayes-adaptive deep rl via meta-learning, in: International Conference on Learning Representations, 2019.
[20] J. Oh, S. Singh, H. Lee, P. Kohli, Zero-shot task generalization with multi-task deep reinforcement learning, in: International Conference on Machine Learning, PMLR, 2017, pp. 2661–2670.