ASP-Driven Emergency Planning for Norm Violations in RL (Extended Abstract)

Sebastian Adam[1][0009−0000−9905−0128] and Thomas Eiter[1][0000−0001−6003−6345]

Technische Universität Wien, Vienna, Austria
{sebastian.adam,thomas.eiter}@tuwien.ac.at

Abstract. Reinforcement learning (RL) is a widely used approach for training an agent to maximize rewards in a given environment. Action policies learned with this technique see a broad range of applications in practical areas like games, healthcare, robotics, or autonomous driving. However, enforcing ethical behavior or norms based on deontic constraints that the agent should adhere to during policy execution remains a complex challenge. Especially constraints that emerge after the training can necessitate to redo policy learning, which can be costly and, more critically, time consuming. To mitigate this problem, we present a framework for policy fixing in case of a norm violation, which allows the agent to stay operational. Based on answer set programming (ASP), emergency plans are generated that exclude or minimize cost of norm violations by future actions in a horizon of interest. By combining and developing optimization techniques, efficient policy fixing under real-time constraints can be achieved.

Motivation Enforcing ethical behavior or norms based on deontic constraints that a trained agent should adhere to in operation is a complex challenge that has garnered significant attention in research related to RL [8, 7], deontic logic [6, 3] and planning [4, 1]. Especially norms and constraints that emerge only after the training phase are difficult to factor into the agent's behavior.

To address this challenge, we introduce the notion of a policy fix, which consists in an alteration of the RL policy such that subsequent actions do not violate deontic constraints. We further introduce a framework in which policy fixing is expressed as a planning problem in ASP, which allows for computing optimal policy fixes using ASP solvers. Using a modular blueprint of an ASP program, the framework is adapted to deterministic and nondeterministic domains.

Policy Fixing & Framework Informally the goal of a policy fix is to alter a policy π, such that following the new policy π', subsequent actions of the agent minimize further norm violations, while accounting for the action preferences of π, balancing norm violations and policy adherence.

Suppose that violation of a norm Φ in a state s at any time t' has a real-valued penalty $p(\Phi, s, t') \geq 0$, and that $\pi' \preceq_Q \pi$ is the preorder of policies according to a learned Q-table, i.e., for each state s, $\pi'(s)$ is ranked better or equal than $\pi(s)$.

(a) Traditional RL framework (b) ASP enhanced framework

Fig. 1: Decision-making framework

Definition 1. *Given a transition system T, a maintenance norm Φ, an agent policy π, and a time point t, a* policy fix *for π at t is a policy π' such that*

$$\pi' = \text{argmin}_{\preceq_Q} \max_{\sigma \in Tr(\pi'',t)} \sum_{t'>t} p(\Phi, s_{t'}, t'), \tag{1}$$

where $Tr(\pi'', t)$ denotes the set of all trajectories σ whose suffix from t, i.e., $s_t, a_{t+1}, s_{t+1}, \ldots$, complies with π''.

Ideally, we select a *strict policy fix*, where the worst case total norm violations is zero. However, instead of giving priority to norm obedience, we may change equation (1) into joint optimization of reward and norms, s.t. norm violations are viewed as a reward discount. This *utility-based policy fix* can be preferred in situations where zero worst case norm violations cannot be guaranteed, or the benefits of following the policy outweigh potential norm violations.

Given that solving (1) amounts to a special conformat planning problem, which is already Σ_2^P-complete when restricted to plans of polynomial length in conventional settings (cf. [2]), we resort to approximate policy fixes of bounded length. Compared to traditional RL (cf. Figure 1a) the proposed framework (Figure 1b) computes such a *k-policy fix* using ASP after every agent action, helping the agent to choose a norm compliant path.

Evaluation & Conclusion Among others, we consider the popular video game *Pacman*, where we apply the same norms as in [6] and [7]. Using a utility-based policy fix, we were able to improve wins of Pacman, while keeping the same high norm compliance as achieved in the comparable approach described in [6].

Beyond grid games, we show the application of the framework for the caching problem of routers in a Content Centric Network (CCN). We use the sota CCN simulator *ndnSIM 2.9* [5] to simulate a small network, in which we define additional norms based on package age. In our testing, the framework drastically reduced the norm violations while achieving a similar cache performance (measured by cache hits) compared to an unaltered LRU policy.

We have thus shown the potential of increasing the norm compliance using the proposed framework in multiple deterministic and nondeterministic domains. Further performance improvements, potentially using different approaches for dealing with coNP problems, are planned for future work.

Acknowledgments. This project is supported by Vienna Science and Technology Fund (WWTF) project ICT22-023.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aminof, B., Giacomo, G.D., Murano, A., Rubin, S.: Planning under LTL Environment Specifications. Proceedings of the International Conference on Automated Planning and Scheduling **29**, 31–39 (2019). https://doi.org/10.1609/icaps.v29i1.3457
2. Baral, C., Kreinovich, V., Trejo, R.: Computational complexity of planning and approximate planning in the presence of incompleteness. Artificial Intelligence **122**(1-2), 241–267 (Sep 2000). https://doi.org/10.1016/S0004-3702(00)00043-6
3. Giordano, L., Martelli, A., Dupré, D.T.: Temporal deontic action logic for the verification of compliance to norms in ASP. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law. pp. 53–62. ACM, Rome Italy (Jun 2013). https://doi.org/10.1145/2514601.2514608
4. Kasenberg, D., Scheutz, M.: Norm Conflict Resolution in Stochastic Domains (Nov 2017)
5. Mastorakis, S., Afanasyev, A., Zhang, L.: On the Evolution of ndnSIM: An Open-Source Simulator for NDN Experimentation. ACM SIGCOMM Computer Communication Review **47**(3), 19–33 (Sep 2017). https://doi.org/10.1145/3138808.3138812
6. Neufeld, E.A., Bartocci, E., Ciabattoni, A., Governatori, G.: Enforcing ethical goals over reinforcement-learning policies. Ethics and Information Technology **24**(4), 43 (Dec 2022). https://doi.org/10.1007/s10676-022-09665-8
7. Noothigattu, R., Bouneffouf, D., Mattei, N., Chandra, R., Madan, P., Varshney, K., Campbell, M., Singh, M., Rossi, F.: Interpretable Multi-Objective Reinforcement Learning through Policy Orchestration (Sep 2018)
8. Thananjeyan, B., Balakrishna, A., Nair, S., Luo, M., Srinivasan, K., Hwang, M., Gonzalez, J.E., Ibarz, J., Finn, C., Goldberg, K.: Recovery RL: Safe Reinforcement Learning With Learned Recovery Zones. IEEE Robotics and Automation Letters **6**(3), 4915–4922 (Jul 2021). https://doi.org/10.1109/LRA.2021.3070252