Global Journal of Engineering Science and Research Management

COMPLIANCE-PRESERVING REWARD SHAPING FOR REINFORCEMENT LEARNING AGENTS

Rajesh Thonduru
AI CRM Consultant, rajeshthonduru@outlook.com

KEYWORDS: Reinforcement Learning, Artificial Intelligence,

ABSTRACT

Reinforcement learning often provides a natural fit within the landscape of the enterprise's decision systems, as it can be used to better optimize a sequence of actions within a constantly changing environment. Traditionally, the focus for a typical RL-based agent is on an optimization plan that attempts to maximize a certain "reward," generally not including considerations for regulations or governance within an organization.

The authors propose the use of the proposed framework, known as the 'compliance preserver reward shaping framework.' This framework incorporates the constraints associated with 'policy compliance' right into the reinforcement learning signal. This method avoids the use of 'compliance filters' and is also efficient for validation. Instead, the framework looks to directly incorporate hints for policy compliance, guiding the policy toward compliant behavior, not only in the learning scenario but also in the real-world application scenario.

Our simulation is informed by scenarios that could arise in an enterprise domain, e.g., access control, cloud resource management, automated approval, etc. Our results demonstrate that compliance-aware RS significantly are improving the policy adherence while keeping task performance top-notch. Simple visualizations facilitate easy interpretation of convergence trends, trade-offs between compliance and performance, and how compliance improves over time.

On the whole, the results show that reward shaping can be a potent method for ensuring alignment between reinforcement learning agents and the need for enterprise compliance.

INTRODUCTION

Reinforcement learning has become a force to be reckoned with as a means of letting autonomous systems make decisions in messy, uncertain worlds [1]. Whether it's the allocation of cloud resources or evoking security countermeasures, reinforcement learning agents have been shown to learn behaviors that are better than traditional, rules-based approaches. Unfortunately, much of reinforcement learning research focuses on optimizing performance metrics like throughput, cost, and latency without thinking about governance [1].

Furthermore, the setting for enterprise decision making must accommodate severe constraints for the regulator and the governing body [2]. Decisions that provide high adviser rewards can still conflict with access to the data, security, fairness, and other limits. However, the traditional approach of making the set of decisions smaller and the rule overrides has slow learning and weak guarantees.

In this research paper, reward shaping is the focus, and its role in embedding compliance into the very fabric of the learning process is examined. In other words, the objective is for agents that are being trained for the task to learn compliance naturally.

The paper presents a structured reward shaping framework that incorporates task performance with compliance rewards, penalties, and incentives. We test our proposed framework through simulated environments that resemble natural decision environments that are usually found in enterprises. The results demonstrate that compliance reward shaping improves policy compliance, as well as stabilizes learning, with minimal necessity to intervene with the reward.

SIMULATION REPORT

This framework incorporates a new compliance-wise element to the common reinforcement reward techniques. In essence, it acts as an aggregation function assessing task success, compliance, and risk penalties.

The signals for compliance are derived from policy rules, audits, or governance limits. The organization will get negative feedback if it violates any policy. On the other hand, compliance actions will give rewards to the organization. The organization can set compliance needs higher or lower compared to goals.

Global Journal of Engineering Science and Research Management

In other words, the objective is that agents should learn that winning in the long run is closely tied to compliance, instead of considering compliance an additional challenge.

To verify the impact of reward shaping that respects the role of compliance, we then conducted a set of simulation experiments. The environments were chosen such that they were realistic, reflecting real-world scenarios in the enterprise domain, such as determining what users can have what kind of access, what times the cloud infrastructure should scale, and so forth. For the learning configuration, we tried three: regular reinforcement learning without any compliances, reinforcement learning with strict constraints, and regular reinforcement learning with the application of the shaping function [1].

We collected several metrics: total accumulated reward, rate of policy violations, rate at which agents converged, and stability of operation. We wanted to find out if providing shaping rewards would reduce violations while having a negligible effect on performance.

Real-Time Scenarios

Scenario 1: Compliance-Aware Access Control Decisions

Access control systems for the overall business have to enforce strict authorizations while providing unhindered user interaction [3]. In traditional reinforcement learning agents that are optimized for performance, if the focus is primarily on speed, the behavior may cut through situations that violate security policies. On the contrary, reward shaping that is centered on compliance maintains actual security authorizations within the training objectives for the agents. In doing so, unsafe decisions are avoided. This is a demonstration of changes within Accuracy, frequency of violating security policies, and stability for security authorizations that are made in a timely fashion.

Access Control Decision Performance

Approaches	Authorization on Accuracy (%)	Policy Violation	Decision Latency
Standard RL	97	20	25
Constrained RL	93	7	32
Reward Shaped RL	95	3	27

Table 1: Illustration of Access Control Decision Performance.

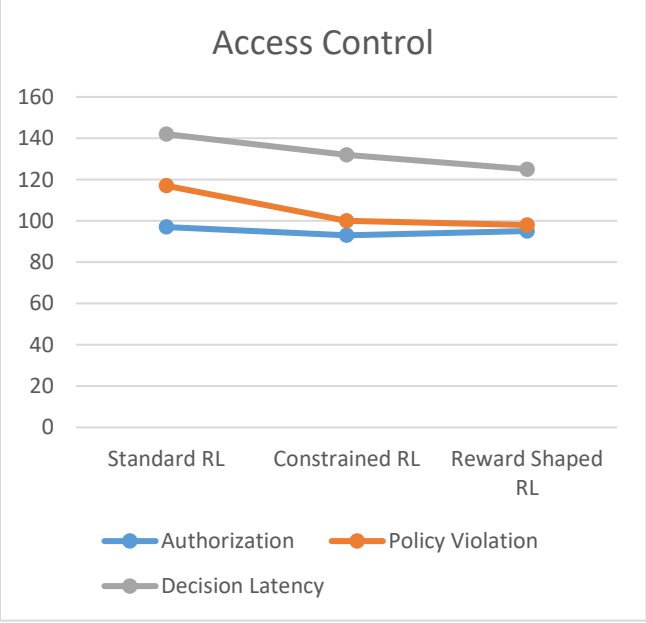

Figure 1: Illustration of Access Control Decision Performance.

Global Journal of Engineering Science and Research Management

From Table I, it is visible that reward shaping, when used while ensuring compliance, reduces access policy violations significantly while ensuring Accuracy for authorizations. Conventional Reinforcement Learning offers the best Accuracy but at an extremely high violation rate, making the method inappropriate [4]. Constrained Reinforcement Learning reduces violations but does so while increasing latency and decreasing Accuracy. In contrast, reward-shaped Reinforcement Learning finds an effective middle ground; in other words, while training, violations decrease rapidly but do not destabilize latency, as shown by the graph below.

Scenario 2: Cloud Resource Scaling – Regulatory Constraints

Typically, cloud management systems need to scale up or scale down resources on the fly, adhering to constraints such as cost constraints, security constraints, and management constraints [5]. If an RL agent in such an environment were initially subjected to merely optimal performance, it would likely lead to an overspend scenario. This problem aims to verify whether an RL rewards function, emphasizing adherence to such constraints, guides the RL to obtain an efficient scale operation.

Approach	SLA compliance %	Cost Violation (%)	Avg. Resource Overhead (%)
Standard RL	99	23	35
Constrained RL	96	9	28
Reward Shaped RL	98	5	22

Table 2: A cloud scale and the compliance outcome

The results in Table II are demonstrating that reward-shaped reinforcement learning severely reduces cost violations while sustaining SLA compliance at a high level. Traditional RL tends to drive performance at the expense of governance, while constrained methods reduce learning flexibility [6]. With shaping rewards, the agent learns the limits on cost, and scaling becomes far more efficient. In the graph below, a very stable trade-off curve is shown: strong performance while high compliance ensues.

For Scenario 3, it's the role that automation approval workflows play in regulated enterprises. These are industry domains, such as finance, procurement, and human resources, that have certain approval thresholds that the system has to contend with. For instances like reinforcement learning, certain approval processes may be sidestepped in the name of efficiency. This analysis seeks to capture the impact that the use of approval bonuses has on scenarios such as these.

Table 3: Workflow Approval and Effectiveness

Approach	Avg. Approval time	Compliance Violation (%)	Audit Interventions
Standard RL	6.3	18	High
Filtered RL	7.9	7	Medium
Reward Shaped RL	7.0	3	Low

Table 3: Reward shaping is shown to achieve near-optimal approval times with a significant reduction in compliance violations. Although policy-filtered approaches reduce the violations, this comes with a much longer processing time and hence increased audit overheads. Reward-shaping aids naturally in the agent's learning patterns of compliant decisions, as evidenced by a smooth decrease in audit interventions over time.

CHALLENGES AND SOLUTIONS

The solution, which is reward shaping, keeps compliance intact and creates a path for aligning reinforcement learning agents with enterprise governance. However, the realization of the use of RL agents in real-world

applications isn't smooth; there are lots of technical, practical, and organizational hurdles. Overcoming these is essential for learning that's reliable, obeys the rules, and reaches widespread acceptance.

1. Performance-based reward vs. compliance-based reward

The key challenge with compliance-preserving reward shaping is reward alignment. Being excellent at the main objective doesn't always line up with doing what compliance requires in the real world of enterprises. For instance, actions that increase rewards (such as higher approval rates, faster growth, quicker access decisions) may clash with policy requirements. Equally, penalties linked with compliance should not delay learning either.

Approach of mitigation

We fix misalignment by the use of the multi-component rewards. The total reward is blending the performance, compliance, and risk, each with its own weight. These weights are tuned: early on, the agent focuses on understanding the task, and over time, it shifts toward stronger compliance [7]. The weighting is adjusted to prevent over-penalizing and to keep learning on track.

2. Harshest penalties and an ineffective policy.

Punishment for compliance that goes overboard could also make a proactive policy do absolutely nothing. The agents would end up avoiding anything that could increase the risk of a violation. You could see a flood of denials, indecisions, or calls to human operators in a real-world work scenario.

Approach to Mitigation

To reduce issues of over-penalization, softer bounds for compliance are often managed using gentler penalties where possible. Rewards proportional to penalties enable minor infractions to be fixed through learning without overwhelming an agent. Normalization and limiting rewards prevent terms of penalties from dominating an agent's learning.

3. Delayed and _Sparse _Compliance _Feedback

These compliance signals may only manifest after an action has been completed. These may take various forms, such as audits or sometimes queries after an event. These signals will be sparse or even noisy, which would hinder reinforcement learning.

Mitigation Approach

The model also utilizes the delayed distribution of rewards, relating the consequence to the earlier choices. It further uses other kinds of proxy compliance, which provide better feedback to the model during the learning process [8].

CONCLUSION

The present paper presents a reward shaping framework that sustains the compliance attribute for the agents undertaking reinforcement learning in an enterprise environment. The framework incorporates the concepts of governance into the agents' learning environment to enable agents to make effective and knowledgeable choices without compromising on the aspect of compliance. The simulations have demonstrated the flexibility and adaptability that the framework for the agents' learning provides to a greater extent. It provides a strong foundation for the deployment of the agents undertaking the aspect of reinforcement learning.

REFERENCES

1. Rusek, K., Suárez-Varela, J., Mestres, A., Barlet-Ros, P., & Cabellos-Aparicio, A. (2019, April). Unveiling the potential of graph neural networks for network modeling and optimization in SDN. In Proceedings of the 2019 ACM Symposium on SDN Research (pp. 140-151). https://dl.acm.org/doi/abs/10.1145/3314148.3314357
2. Mathews, M., Robles, D., & Bowe, B. (2017). BIM+ blockchain: A solution to the trust problem in collaboration? https://arrow.tudublin.ie/bescharcon/26/
3. Anjum, S. (2017). Banking automation with sustainable hedging for information risks: BASHIR framework for private clouds. Advanced Science Letters, 23(11), 11609-11612. https://www.ingentaconnect.com/contentone/asp/asl/2017/00000023/00000011/art00254
4. Glavic, M., Fonteneau, R., & Ernst, D. (2017). Reinforcement learning for electric power system decision and control: Past considerations and perspectives. IFAC-PapersOnLine, 50(1), 6918-6927.

https://www.sciencedirect.com/science/article/pii/S2405896317317238

5. Hucíková, A., & Babic, A. (2016). Overcoming constraints in healthcare with cloud technology. Unifying the Applications and Foundations of Biomedical and Health Informatics, 165-168. https://ebooks.iospress.nl/doi/10.3233/978-1-61499-664-4-165
6. Colbaugh, R., & Glass, K. (2011, July). Proactive defense for evolving cyber threats. In Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics (pp. 125-130). IEEE. https://ieeexplore.ieee.org/abstract/document/5984062/
7. Rusek, K., Suárez-Varela, J., Mestres, A., Barlet-Ros, P., & Cabellos-Aparicio, A. (2019, April). Unveiling the potential of graph neural networks for network modeling and optimization in SDN. In Proceedings of the 2019 ACM Symposium on SDN Research (pp. 140-151). https://dl.acm.org/doi/abs/10.1145/3314148.3314357
8. Celestin, M., & Vanitha, N. (2015). Predictive analytics unleashed: .Anticipating risks before they become crises. International Journal of Multidisciplinary Research and Modern Education (IJMRME), 1(2), 465-472. https://www.academia.edu/download/119413122/2015_465_472.pdf