Learning to Explore when Mistakes are Not Allowed

Extended Abstract

Charly Pecqueux-Guézénec
Sorbonne Université, CNRS, ISIR
F-75005 Paris, France
pecqueuxguezenec@isir.upmc.fr

Stéphane Doncieux
Sorbonne Université, CNRS, ISIR
F-75005 Paris, France
doncieux@isir.upmc.fr

Nicolas Perrin-Gilbert
Sorbonne Université, CNRS, ISIR
F-75005 Paris, France
perrin@isir.upmc.fr

ABSTRACT

Goal-Conditioned Reinforcement Learning (GCRL) enables learning unified controllers but its trial-and-error process poses risks in real-world applications. We propose a method that allows agents to explore while avoiding harmful mistakes. Since environment dynamics are often uniform in space, a policy trained for safety without exploration purposes can still be exploited globally. Our approach has two phases: first, pretraining a safety policy using safe reinforcement learning and distributional techniques; second, ensuring safe exploration by selecting the action to perform on the environment either from the safety policy or from the learning goal-conditioned (GC) policy, depending on current state. In simulated environments, we show that it covers most of the goal-space while minimizing mistakes during exploration, unlike traditional GCRL. We also perform an ablation study and failure analysis, providing insights for future research.

KEYWORDS

Safe Exploration ; Safe Reinforcement Learning ; Goal-Conditioned Reinforcement Learning

ACM Reference Format:
Charly Pecqueux-Guézénec, Stéphane Doncieux, and Nicolas Perrin-Gilbert. 2025. Learning to Explore when Mistakes are Not Allowed: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Goal-conditioned reinforcement learning (GCRL) enables a single controller to handle diverse tasks and adapt [3–6, 16, 17, 19], but often relies heavily on primitives for exploration and safety. Safe exploration methods fall into two categories [14]: auxiliary reward-based [8, 12, 15] and human knowledge-based approaches [9, 11, 18, 21], the latter relying on strong assumptions like emergency stops [22] or linearized models [7]. We propose a safe exploration framework[1] at the intersection, inspired by viability theory [23], requiring fewer assumptions and focusing on empirically showing that an agent can explore safely without mistakes, at the cost of theoretical guarantees. The agent alternates between a safety policy that maintains viability and a goal-conditioned policy that explores and reaches new goals, with a risk measure guiding the switches (Figure 1). Our contributions are threefold. (1) a distributional safe RL framework to pretrain a safety policy, (2) an action selection mechanism ensuring safe exploration, (3) analyzing key components of the method and failure modes to orient future research.

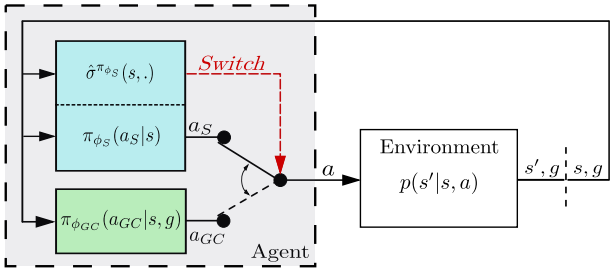


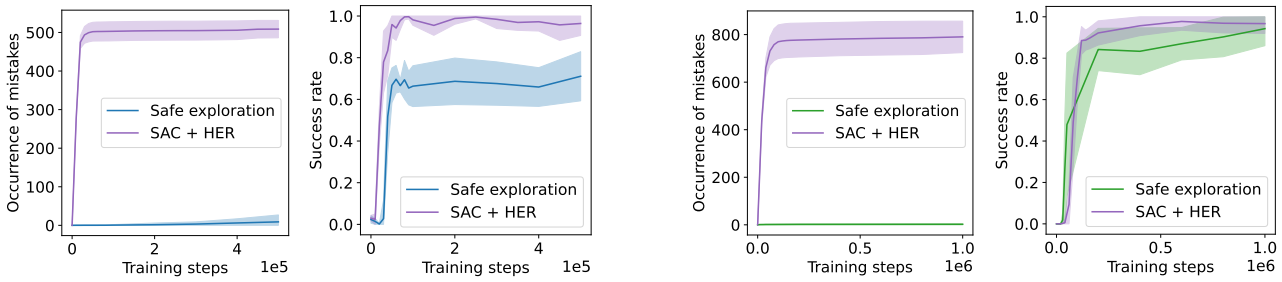
Figure 1: Action selection mechanism to guarantee safe exploration. $\pi_{\phi_S}(.|s)$ is the parameterized safety policy, $\pi_{\phi_{GC}}(.|s,g)$ is the goal-conditioned policy, and $\sigma^{\pi_{\phi_S}}$ the function estimating the level of confidence in the safety policy's ability to avoid potential future errors. If it is too low, the safe action a_S is executed to keep the system safe. Otherwise, a_{GC} is executed, allowing the agent to explore.

2 METHOD

The safe exploration problem can be seen as the combination of a CMDP $(\mathcal{S}, \mathcal{A}, p, p_0, r_S, h)$ [1], solved by the safety policy, and a multi-goal MDP $(\mathcal{S} \times \mathcal{G}, \mathcal{A}, p, p_0, p_{\mathcal{G}}, r_{\mathcal{G}})$ [19], solved by the GC policy. The safety reward r_S is 1 within a set N_0 around an equilibrium point and 0 elsewhere, enabling the safety policy's critics to estimate the steps needed to reach this set from any given state. We use thresholds on this estimate to guide action selection, as safety decreases with step count. Due to neural networks continuity, terminal states may appear safe if near a safe state. To address this, we assume the agent receives a cost at each step, defined by a continuous constraint function h that verifies $h(s) > 0$ for terminal states [24]. Safety pretraining is inspired by TQC [13] and incorporate a regularization term based on reachability like RCRL [24]. Using quantiles allows to prevent overestimation and to compute a risk measure (Figure 1). During the safe exploration phase, the safety policy and critics are fixed, and a randomly initialized GC policy is trained. The replay buffer starts empty and is filled during safe exploration with transitions from both policies. SAC-N [2], a variant of SAC [10], is used to train the GC policy while preventing overestimation bias in the critic. To decide when to switch, we use the mean of the worst 10% of safety critic quantiles.

[1] See the full version of this paper for more details: https://arxiv.org/abs/2502.13801

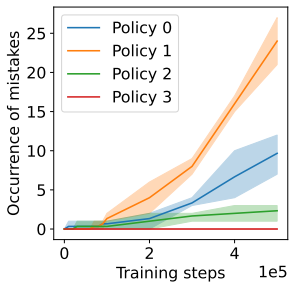
Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).



(a) CartPoleGC (b) SkydioX2GC

Figure 2: Comparison between our safe exploration method and the baseline (SAC+HER) in terms of safety during exploration and coverage on CartPoleGC (a) and SkydioX2GC (b) environments. To avoid cherry picking, we used multiple seeds both for pretraining and safe exploration: 4 of pretraining times 3 of safe exploration for CartPoleGC, 3 times 3 for SkydioX2GC.

Figure 3: Occurrence of mistakes obtained during safe exploration for different pre-trained policies with the CartPoleGC environment.



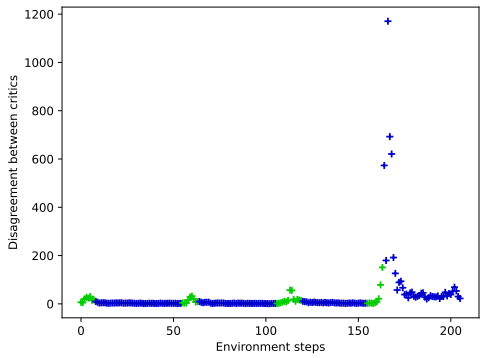


Figure 4: Critic disagreement in L_1 norm between the quantile critics along a failed episode of the CartPoleGC environment. Dots are green when the GC policy is active and blue when the safety policy is active.

3 EXPERIMENTS

We conducted experiments on a custom goal-based version of the gymnasium CartPole environment [20] with continuous actions that we call *CartPoleGC*, and a goal environment based on the Skydio X2 drone from the Mujoco menagerie [25] that we call *SkydioX2GC*. We performed an ablation study to validate our design choices and compared our approach with a classical GC method combining SAC and HER, with 80% of relabelling and *future* strategy [3]. As our general objective is zero error during exploration, we also analyze the causes of the few mistakes we obtained during safe exploration using our method. In terms of coverage, the baseline obtains a better success rate, around 98% on average against 70% for CartPoleGC (Figure 2a) while on SkydioX2GC safe exploration offers almost the same performance but with higher variance (Figure 2b). However, our approach considerably reduces the number of mistakes during exploration in comparison to the baseline which does not take safety into account (Figure 2). On CartPoleGC we obtained at most 27 mistakes, and 7 mistakes at most on SkydioX2GC, but for some runs, we obtained 0 mistakes for the whole training. These differences, within the same environment, stem from the strong reliance of safe exploration on the pretrained safety policy, and consequently, on the quality of the pretraining (Figure 3). This dependence on the safety policy is further confirmed, and even more clearly demonstrated, in Figure 4 that shows disagreement between quantile critics along a failed episode of CartPoleGC. Disagreement between critics is large before the mistake occurs, indicating insufficient training of the safety policy on those states and actions. This suggests two research directions: incorporating disagreement into the switching mechanism and fine-tuning the safety policy during safe exploration.

4 CONCLUSION

Our experiments show that our safe exploration framework effectively trains a goal-conditioned policy while preventing most of the mistakes. Failures mainly stem from insufficient safety policy pretraining, which could be improved by leveraging critic disagreement or fine-tuning during exploration. Beyond trusting the safety policy, gaining a deeper understanding of its limitations will be key to refining our approach and guiding future improvements in safe exploration.

ACKNOWLEDGMENTS

This work has received funding from the European Commission's Horizon Europe Framework Program under grant agreement No 101070381 (PILLAR-Robots project).

REFERENCES

[1] Eitan Altman. 1995. *Constrained Markov Decision Processes.* Number RR-2574. 13–16 pages. https://inria.hal.science/inria-00074109

[2] Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. 2021. Uncertainty-Based Offline Reinforcement Learning with Diversified Q-Ensemble. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 7436–7447. https://proceedings.neurips.cc/paper_files/paper/2021/file/3d3d286a8d153a4a58156d0e02d8570c-Paper.pdf

[3] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. 2017. Hindsight Experience Replay. https://proceedings.neurips.cc/paper_files/paper/2017/file/453fadbd8a1a3af50a9df4df899537b5-Paper.pdf

[4] Elliot Chane-Sane, Cordelia Schmid, and Ivan Laptev. 2021. Goal-conditioned reinforcement learning with imagined subgoals. In *International conference on machine learning*. PMLR, 1430–1440.

[5] Jongwook Choi, Archit Sharma, Honglak Lee, Sergey Levine, and Shixiang Shane Gu. 2021. Variational Empowerment as Representation Learning for Goal-Conditioned Reinforcement Learning. , 1953–1963 pages. https://proceedings.mlr.press/v139/choi21b.html

[6] Cédric Colas, Tristan Karch, Olivier Sigaud, and Pierre-Yves Oudeyer. 2022. Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: a short survey. *Journal of Artificial Intelligence Research* 74 (2022), 1159–1199.

[7] Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. 2018. Safe Exploration in Continuous Action Spaces. arXiv:1801.08757 [cs.AI]

[8] Mehdi Fatemi, Shikhar Sharma, Harm Van Seijen, and Samira Ebrahimi Kahou. 2019. Dead-ends and Secure Exploration in Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 1873–1881. https://proceedings.mlr.press/v97/fatemi19a.html

[9] J. Garcia and F. Fernandez. 2012. Safe Exploration of State and Action Spaces in Reinforcement Learning. *Journal of Artificial Intelligence Research* 45 (Dec. 2012), 515–564. https://doi.org/10.1613/jair.3761

[10] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. 2019. Soft Actor-Critic Algorithms and Applications. arXiv:1812.05905 [cs.LG]

[11] Nathan Hunt, Nathan Fulton, Sara Magliacane, Nghia Hoang, Subhro Das, and Armando Solar-Lezama. 2020. Verifiably Safe Exploration for End-to-End Reinforcement Learning. arXiv:2007.01223 [cs.AI] https://arxiv.org/abs/2007.01223

[12] Thommen George Karimpanal, Santu Rana, Sunil Gupta, Truyen Tran, and Svetha Venkatesh. 2020. Learning transferable domain priors for safe exploration in reinforcement learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–10.

[13] Arsenii Kuznetsov, Pavel Shvechikov, Alexander Grishin, and Dmitry Vetrov. 2020. Controlling Overestimation Bias with Truncated Mixture of Continuous Distributional Quantile Critics. , 5556–5566 pages. https://proceedings.mlr.press/v119/kuznetsov20a.html

[14] Pawel Ladosz, Lilian Weng, Minwoo Kim, and Hyondong Oh. 2022. Exploration in deep reinforcement learning: A survey. *Information Fusion* 85 (Sept. 2022), 1–22. https://doi.org/10.1016/j.inffus.2022.03.003

[15] Zachary C. Lipton, Kamyar Azizzadenesheli, Abhishek Kumar, Lihong Li, Jianfeng Gao, and Li Deng. 2018. Combating Reinforcement Learning's Sisyphean Curse with Intrinsic Fear. arXiv:1611.01211 [cs.LG] https://arxiv.org/abs/1611.01211

[16] Soroush Nasiriany, Vitchyr H. Pong, Steven Lin, and Sergey Levine. 2019. Planning with Goal-Conditioned Policies. arXiv:1911.08453 [cs.LG] https://arxiv.org/abs/1911.08453

[17] Vitchyr H. Pong, Murtaza Dalal, Steven Lin, Ashvin Nair, Shikhar Bahl, and Sergey Levine. 2020. Skew-Fit: State-Covering Self-Supervised Reinforcement Learning. arXiv:1903.03698 [cs.LG] https://arxiv.org/abs/1903.03698

[18] William Saunders, Girish Sastry, Andreas Stuhlmueller, and Owain Evans. 2017. Trial without Error: Towards Safe Reinforcement Learning via Human Intervention. arXiv:1707.05173 [cs.AI] https://arxiv.org/abs/1707.05173

[19] Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. 2015. Universal Value Function Approximators. In *Proceedings of the 32nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 37)*, Francis Bach and David Blei (Eds.). PMLR, Lille, France, 1312–1320. https://proceedings.mlr.press/v37/schaul15.html

[20] Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. 2024. Gymnasium: A Standard Interface for Reinforcement Learning Environments.

[21] Matteo Turchetta, Felix Berkenkamp, and Andreas Krause. 2016. Safe Exploration in Finite Markov Decision Processes with Gaussian Processes. arXiv:1606.04753 [cs.LG] https://arxiv.org/abs/1606.04753

[22] Akifumi Wachi, Wataru Hashimoto, Xun Shen, and Kazumune Hashimoto. 2024. Safe exploration in reinforcement learning: A generalized formulation and algorithms. *Advances in Neural Information Processing Systems* 36 (2024).

[23] Pierre-Brice Wieber. 2008. Viability and Predictive Control for Safe Locomotion. In *IROS 2008-IEEE-RSJ International Conference on Intelligent Robots & Systems*. IEEE, 1103–1108.

[24] Dongjie Yu, Haitong Ma, Shengbo Li, and Jianyu Chen. 2022. Reachability Constrained Reinforcement Learning. , 25636–25655 pages. https://proceedings.mlr.press/v162/yu22d.html

[25] Kevin Zakka, Yuval Tassa, and MuJoCo Menagerie Contributors. 2022. MuJoCo Menagerie: A collection of high-quality simulation models for MuJoCo. http://github.com/google-deepmind/mujoco_menagerie