Safe But Not Sorry: Reducing Over-Conservatism in Safety Critics via Uncertainty-Aware Modulation

Extended Abstract

Daniel Bethell
University of York
York, UK
daniel.bethell@york.ac.uk

Simos Gerasimou
Cyprus University of Technology, Cyprus
University of York, York, UK
simos.gerasimou@cut.ac.cy

Radu Calinescu
University of York
York, UK
radu.calinescu@york.ac.uk

Calum Imrie
University of York
York, UK
calum.imrie@york.ac.uk

ABSTRACT

Ensuring safe exploration in reinforcement learning is essential for real-world deployment. Existing methods, however, often trade safety for performance by producing overly conservative policies or diffuse cost estimates that weaken policy gradients. We propose the Uncertain Safety Critic (USC), which modulates conservatism using critic uncertainty and refines under-covered regions, reducing safety violations by ≈ 40% while maintaining competitive or higher rewards and cutting cost-gradient error by ≈ 83%.

KEYWORDS

Safe Reinforcement Learning; Safety Critics; Safety-Critical Systems

ACM Reference Format:
Daniel Bethell, Simos Gerasimou, Radu Calinescu, and Calum Imrie. 2026. Safe But Not Sorry: Reducing Over-Conservatism in Safety Critics via Uncertainty-Aware Modulation: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/ERDQ6501

1 INTRODUCTION

Reinforcement learning (RL) learns sequential decision-making through exploration [22], and has achieved strong results in domains such as games [4] and robotics [10, 17]. However, in safety-critical settings (e.g., healthcare and embodied robotics), exploration can lead an agent into hazardous states where unsafe actions have unacceptable physical or ethical consequences [6]. Safe RL, therefore, must enforce constraints without eliminating the exploratory behaviour needed to discover high-performing policies [2], and must do so not only during training but also under deployment, where decisions directly affect stakeholders [5].

Prior approaches are commonly grouped into external-knowledge and cost-based methods. External-knowledge approaches constrain behaviour using formal specifications [9], expert demonstrations [13], or interactive oversight [14, 19], but often depend on substantial domain expertise or supervision [3]. Cost-based methods instead formulate safety as a constrained optimisation problem in CMDPs [1], typically via Lagrangian or primal–dual updates [15, 18, 21, 24]. In practice, these methods rely on safety critics to estimate expected cumulative cost; however, critic learning is fragile under sparse or noisy cost signals [20]. Conservative safety critics reduce underestimation by biasing costs upward [7], but this often inflates or flattens the cost landscape, weakening or saturating the gradients passed to the policy and stalling improvement, especially under strong dual penalties [8].

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/ERDQ6501

2 UNCERTAIN SAFETY CRITIC APPROACH

We study continuous-control CMDPs and train an actor–critic agent with reward critic Q_R and safety critic Q_C under a standard Lagrangian constraint. USC modifies the safety critic update to avoid the over-conservatism of conservative safety critics by concentrating conservatism where the critic is uncertain and risk is high. Concretely, for each replay sample, we compute an influence-based epistemic uncertainty score via Gauss–Newton influence [12]. Let θ_C denote the (frozen) safety critic parameters after the latest update, and let δI be a damped identity term. The uncertainty scalar is:

$$u(s,a) = \operatorname{diag}\Big(\big[\nabla_{\theta_C} Q_C(s,a;\theta_C^*)\big] \times \Big(\sum_{i=1}^{B} \nabla_{\theta_C} Q_C(s_i,a_i;\theta_C^*) \nabla_{\theta_C} Q_C(s_i,a_i;\theta_C^*)^\top + \delta I\Big)^{-1} \times \big[\nabla_{\theta_C} Q_C(s,a;\theta_C^*)\big]^\top\Big) \tag{1}$$

Intuitively, larger $u(s,a)$ indicates that small perturbations to the sample would induce a larger change in the critic parameters, i.e., higher epistemic uncertainty. We convert this into an uncertainty-adjusted conservatism weight:

$$\tilde{u}(s_t,a_t) = \log(1+u(s_t,a_t)) \times (1+\mathbf{1}\{Q_C(s_t,a_t) > \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|} Q_C(s_i,a_i)\}) \tag{2}$$

The logarithmic transformation of u ensures that influence values are stabilised, preventing extreme magnitudes from dominating

Table 1: Mean episodic reward and cost for comparative methods across evaluated benchmarks. Light blue and light green highlighted cells indicate the lowest cost and highest reward for each task, respectively.

	Method									
Environment	DDPG		Safety Critic		Conservative Safety Critic		Uncertain Safety Critic (NR)		Uncertain Safety Critic	
Safety Gymnasium	Reward ↑	Cost ↓	Reward ↑	Cost ↓	Reward ↑	Cost ↓	Reward ↑	Cost ↓	Reward ↑	Cost ↓
CarGoal1 ($\chi = 1.0$)	**6.57 ± 0.45**	1.05 ± 0.39	6.60 ± 0.45	1.02 ± 0.37	6.56 ± 0.40	1.03 ± 0.39	6.56 ± 0.39	1.06 ± 0.41	6.55 ± 0.39	**0.96 ± 0.33**
CarGoal2 ($\chi = 5.0$)	8.74 ± 0.72	5.36 ± 1.47	7.54 ± 1.38	5.65 ± 2.00	8.65 ± 0.75	5.28 ± 1.44	8.84 ± 0.64	5.44 ± 1.43	**8.84 ± 0.62**	**5.05 ± 1.27**
CarButton1 ($\chi = 5.0$)	6.20 ± 0.74	9.50 ± 3.01	5.12 ± 1.89	5.96 ± 2.68	5.48 ± 1.83	**5.57 ± 2.63**	6.31 ± 0.55	6.37 ± 2.05	**6.38 ± 0.62**	6.01 ± 2.13
CarButton2 ($\chi = 5.0$)	7.51 ± 0.67	10.30 ± 3.32	2.48 ± 3.32	**4.04 ± 4.46**	6.45 ± 2.44	6.49 ± 3.19	7.49 ± 0.79	8.02 ± 2.52	**7.69 ± 0.62**	6.32 ± 2.23
Gymnasium Robotics	Reward ↑	Cost ↓	Reward ↑	Cost ↓	Reward ↑	Cost ↓	Reward ↑	Cost ↓	Reward ↑	Cost ↓
FetchReach ($\chi = 5.0$)	−4.36 ± 16.04	8.19 ± 7.70	1.98 ± 4.65	**3.18 ± 1.05**	−1.53 ± 6.51	3.91 ± 3.07	0.55 ± 4.00	3.34 ± 1.36	**3.92 ± 2.51**	3.20 ± 1.11
Mujoco	Reward ↑	Cost ↓	Reward ↑	Cost ↓	Reward ↑	Cost ↓	Reward ↑	Cost ↓	Reward ↑	Cost ↓
HalfCheetah ($\chi = 0.1$)	5.71 ± 2.48	21.82 ± 23.20	3.19 ± 3.43	0.32 ± 0.63	4.13 ± 1.29	2.71 ± 5.70	**6.64 ± 2.26**	0.62 ± 0.51	3.76 ± 2.70	**0.20 ± 0.40**

the loss, while the indicator term accentuates penalties for transitions whose predicted cost exceeds the batch mean. Accordingly, conservatism is selectively concentrated on regions that are both uncertain and costly, rather than being spread uniformly across the state–action space. The safety critic is then updated by minimising

$$\mathcal{L}_C = \mathbb{E}_{(s_t,a_t,r_t,s_{t+1},c_t)\sim\mathcal{B}}\left[\frac{1}{2}\Big(Q_C(s_t,a_t) - \big(c_t + \gamma Q_C(s_{t+1},\pi(s_{t+1}))\big)\Big)^2\right.$$
$$\left. + \frac{1}{2}\,\tilde{u}(s_t,a_t)\ \times \log \sum_{a'\sim\mathrm{Unif}(\mathcal{A})} \exp\left(\frac{Q_C(s_t,a') - \mu}{\sigma+\epsilon}\right) - \frac{Q_C(s_t,a_t)-\mu}{\sigma+\epsilon}\right] \tag{3}$$

where μ and σ denote the batch mean and standard deviation of the safety critic outputs, and ϵ is a small constant for numerical stability.

Finally, USC includes an uncertainty refinement step: after each critic update, we rank replay samples by the Gauss–Newton influence score in Equation 1 and select the top-n most uncertain state–action pairs. For each selected pair, we form a synthetic cost target by interpolating from its nearest confidently predicted neighbours in the joint (s, a) space, then update Q_C toward these targets using a trust-region-style regulariser that limits deviation from the previous critic in high-uncertainty regions. This reduces epistemic uncertainty and sharpens cost estimates in sparsely covered areas without destabilising training.

3 RESULTS AND DISCUSSION

We evaluate USC[1] on standard continuous-control safe RL benchmarks drawn from Safety Gymnasium [11], Gymnasium Robotics [16], and MuJoCo [23], covering both navigation-style hazard avoidance and high-dimensional robot control. The suite spans varying constraint densities and dynamics, including CarGoal1/2 and CarButton1/2 (goal reaching and button pressing under increasing numbers of hazards), FetchReach (7-DoF manipulation under safety constraints), and HalfCheetah (locomotion with a safe-velocity constraint), enabling a consistent comparison against prior safety-critic baselines used in similar settings [7].

Table 1 reports the average episodic reward and cost across all benchmarks. Overall, USC achieves a more favourable reward–cost trade-off than the comparative methods. The base agent (DDPG) attains high rewards but consistently violates safety constraints, while standard safety critics reduce violations at the cost of weaker rewards. Conservative safety critics further lower costs but often over-penalise actions, leading to degraded performance.

Table 2: Predictive cost map errors in CarGoal2 compared to ground truth, showing gradient alignment, safe–unsafe contrast, and hazard boundary sharpness.

	Method		
Metric	Safety Critic	Conservative Safety Critic	Uncertain Safety Critic
Gradient MSE ↓	0.58 ± 0.29	1.95 ± 1.37	**0.10 ± 0.06**
Contrast Error ↓	0.09 ± 0.09	**0.04 ± 0.03**	0.08 ± 0.05
Entropy Error ↓	2.50 ± 0.67	3.02 ± 0.27	**2.16 ± 0.26**

In contrast, USC maintains strong reward while reducing safety violations across tasks. For example, in CarGoal2, USC lowers the average episodic cost from 5.65 ± 2.00 with the standard safety critic to 5.05 ± 1.27, while also achieving a higher reward (8.84 ± 0.62). Similarly, in CarButton1, USC attains the highest reward among all methods while reducing cost by nearly 40% relative to the unconstrained DDPG baseline. In more complex environments such as FetchReach, USC achieves the best reward while maintaining competitive costs, whereas conservative critics suffer from reduced performance.

We also evaluate an a USC variant, USC (NR), which removes the uncertainty refinement step. While USC (NR) attains similar rewards, it produces higher and less stable costs, indicating that uncertainty refinement is necessary to stabilise safety in poorly explored regions. Overall, the results show that USC mitigates the over-conservatism of prior safety critics while preserving informative gradients that support effective reward–safety trade-offs.

Table 2 corroborates these findings. USC yields substantially better gradient alignment with the true hazard boundaries (Gradient MSE 0.10 ± 0.06 vs. SC 0.58 ± 0.29 and CSC 1.95 ± 1.37), while maintaining strong safe–unsafe separation (Contrast Error 0.08 ± 0.05) without CSC's uniform cost inflation. It also produces the sharpest, least diffuse hazard maps (Entropy Error 2.16 ± 0.26), indicating more informative cost landscapes for policy optimisation.

Our evaluation demonstrates that USC consistently reduces safety violations while maintaining or exceeding state-of-the-art performance, particularly in larger and more complex environments. Future work will investigate extending USC to support partially observable and adversarial environments.

[1]Code available at: https://github.com/team-daniel/USC

REFERENCES

[1] Eitan Altman. 1998. Constrained Markov decision processes with total cost criteria: Lagrangian approach and dual linear program. *Mathematical methods of operations research* 48, 3 (1998), 387–417.

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[3] Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. 2018. Verifiable reinforcement learning via policy extraction. *Advances in neural information processing systems* 31 (2018).

[4] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. 2019. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680* (2019).

[5] Daniel Bethell, Simos Gerasimou, Radu Calinescu, and Calum Imrie. 2025. Learning to Navigate Under Imperfect Perception: Conformalised Segmentation for Safe Reinforcement Learning. *arXiv preprint arXiv:2510.18485* (2025).

[6] Daniel Bethell, Simos Gerasimou, Radu Calinescu, and Calum Imrie. 2025. Safe reinforcement learning in black-box environments via adaptive shielding. *28th European Conference on Artificial Intelligence* 413 (2025), 2450 – 2457. https://doi.org/10.3233/FAIA251092

[7] Homanga Bharadhwaj, Aviral Kumar, Nicholas Rhinehart, Sergey Levine, Florian Shkurti, and Animesh Garg. 2021. Conservative Safety Critics for Exploration. (2021).

[8] Agustin Castellano, Hancheng Min, Juan Andrés Bazerque, and Enrique Mallada. 2023. Learning safety critics via a non-contractive binary Bellman operator. In *2023 57th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 814–821.

[9] Dennis Gross, Nils Jansen, Sebastian Junges, and Guillermo A Pérez. 2022. COOL-MC: a comprehensive tool for reinforcement learning and model checking. In *International Symposium on Dependable Software Engineering: Theories, Tools, and Applications*. Springer, 41–49.

[10] Nicolas Heess, Dhruva Tb, Srinivasan Sriram, Jay Lemmon, Josh Merel, Greg Wayne, Yuval Tassa, Tom Erez, Ziyu Wang, SM Eslami, et al. 2017. Emergence of locomotion behaviours in rich environments. *arXiv preprint arXiv:1707.02286* (2017).

[11] Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. 2023. Safety gymnasium: A unified safe reinforcement learning benchmark. *Advances in Neural Information Processing Systems* 36 (2023), 18964–18993.

[12] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*. PMLR, 1885–1894.

[13] Jinning Li, Xinyi Liu, Banghua Zhu, Jiantao Jiao, Masayoshi Tomizuka, Chen Tang, and Wei Zhan. 2024. Guided online distillation: Promoting safe reinforcement learning by offline demonstration. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 7447–7454.

[14] Haritz Odriozola-Olalde, Maider Zamalloa, and Nestor Arana-Arexolaleiba. 2023. Shielded reinforcement learning: A review of reactive methods for safe learning. In *2023 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 1–8.

[15] Santiago Paternain, Miguel Calvo-Fullana, Luiz FO Chamon, and Alejandro Ribeiro. 2022. Safe policies for reinforcement learning via primal-dual methods. *IEEE Trans. Automat. Control* 68, 3 (2022), 1321–1336.

[16] Matthias Plappert, Marcin Andrychowicz, Alex Ray, Bob McGrew, Bowen Baker, Glenn Powell, Jonas Schneider, Josh Tobin, Maciek Chociej, Peter Welinder, Vikash Kumar, and Wojciech Zaremba. 2018. Multi-Goal Reinforcement Learning: Challenging Robotics Environments and Request for Research. arXiv:arXiv:1802.09464

[17] Nikita Rudin, David Hoeller, Philipp Reist, and Marco Hutter. 2022. Learning to Walk in Minutes Using Massively Parallel Deep Reinforcement Learning. In *Proceedings of the 5th Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 164)*, Aleksandra Faust, David Hsu, and Gerhard Neumann (Eds.). PMLR, 91–100. https://proceedings.mlr.press/v164/rudin22a.html

[18] Harsh Satija, Philip Amortila, and Joelle Pineau. 2020. Constrained markov decision processes via backward value functions. In *International Conference on Machine Learning*. PMLR, 8502–8511.

[19] William Saunders, Girish Sastry, Andreas Stuhlmueller, and Owain Evans. 2017. Trial without error: Towards safe reinforcement learning via human intervention. *arXiv preprint arXiv:1707.05173* (2017).

[20] Krishnan Srinivasan, Benjamin Eysenbach, Sehoon Ha, Jie Tan, and Chelsea Finn. 2020. Learning to be safe: Deep rl with a safety critic. *arXiv preprint arXiv:2010.14603* (2020).

[21] Adam Stooke, Joshua Achiam, and Pieter Abbeel. 2020. Responsive safety in reinforcement learning by pid lagrangian methods. In *International Conference on Machine Learning*. PMLR, 9133–9143.

[22] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[23] 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. *arXiv preprint arXiv:2407.17032* (2024).

[24] Ming Yu, Zhuoran Yang, Mladen Kolar, and Zhaoran Wang. 2019. Convergent policy optimization for safe reinforcement learning. *Advances in Neural Information Processing Systems* 32 (2019).

ACKNOWLEDGMENTS

This research has received funding from the Doctoral Centre for Safe, Ethical and Secure Computing (SEtS) at the University of York, UK, the European Union's Horizon projects GuardAI and AI4Work (grant agreements No 101168067 and 101135990, respectively), and the UK Advanced Research and Invention Agency's Safeguarded AI project ULTIMATE. This work was also supported by the Centre for Assuring Autonomy, University of York, UK. The authors are grateful to Charmaine Barker for her help and insight on this paper.