

References

- [1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané, ‘Concrete problems in ai safety’, *arXiv preprint arXiv:1606.06565*, (2016).
- [2] Daniel Bethell, Simos Gerasimou, and Radu Calinescu, ‘Robust uncertainty quantification using conformalised monte carlo prediction’, in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 20939–20948, (2024).
- [3] Homanga Bharadhwaj, Aviral Kumar, Nicholas Rhinehart, Sergey Levine, Florian Shkurti, and Animesh Garg, ‘Conservative safety critics for exploration’, in *International Conference on Learning Representations*, (2020).
- [4] Vivek S Borkar, ‘An actor-critic algorithm for constrained markov decision processes’, *Systems & control letters*, **54**(3), 207–213, (2005).
- [5] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra, ‘Continuous control with deep reinforcement learning’, *arXiv preprint arXiv:1509.02971*, (2015).
- [6] Shahaf S Shperberg, Bo Liu, and Peter Stone, ‘Learning a shield from catastrophic action effects: Never repeat the same mistake’, *arXiv preprint arXiv:2202.09516*, (2022).
- [7] Chen Tessler, Daniel J Mankowitz, and Shie Mannor, ‘Reward constrained policy optimization’, in *International Conference on Learning Representations*, (2018).
- [8] Masaki Waga, Ezequiel Castellano, Sasinee Pruekprasert, Stefan Klikovits, Toru Takisaka, and Ichiro Hasuo, ‘Dynamic shielding for reinforcement learning in black-box environments’, in *International Symposium on Automated Technology for Verification and Analysis*, pp. 25–41. Springer, (2022).