REFERENCES

[1] 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).

[2] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J. Russell, and Anca Dragan. 2017. Inverse reward design. In *Advances in Neural Information Processing Systems*.

[3] Subbarao Kambhampati. 2007. Model-lite planning for the web age masses: The challenges of planning with incomplete and evolving domain models. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*.

[4] Victoria Krakovna, Laurent Orseau, Miljan Martic, and Shane Legg. 2019. Penalizing side effects using stepwise relative reachability. In *IJCAI AI Safety Workshop*.

[5] Ramya Ramakrishnan, Ece Kamar, Debadeepta Dey, Julie Shah, and Eric Horvitz. 2018. Discovering Blind Spots in Reinforcement Learning. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*.

[6] Sandhya Saisubramanian and Shlomo Zilberstein. 2019. Adaptive outcome selection for planning with reduced models. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*.

[7] Richard S. Sutton. 1996. Generalization in reinforcement learning: Successful examples using sparse coarse coding. In *Advances in Neural Information Processing Systems*.

[8] Kyle Hollins Wray, Shlomo Zilberstein, and Abdel-Illah Mouaddib. 2015. Multi-objective MDPs with conditional lexicographic reward preferences. In *Proceedings of the 29th Conference on Artificial Intelligence*.

[9] Shun Zhang, Edmund H. Durfee, and Satinder P. Singh. 2018. Minimax-regret querying on side effects for safe optimality in factored Markov decision processes. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*.