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*. 6765–6774.

[3] Sarah Keren, Avigdor Gal, and Erez Karpas. 2014. Goal Recognition Design. In *Proceedings of the 24th International Conference on Automated Planning and Scheduling*.

[4] Jette Randløv. 2000. Shaping in Reinforcement Learning by Changing the Physics of the Problem. In *Proceedings of the 17th International Conference on Machine Learning*. 767–774.

[5] Sandhya Saisubramanian, Ece Kamar, and Shlomo Zilberstein. 2020. A Multi-Objective Approach to Mitigate Negative Side Effects. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*. 354–361.

[6] Sandhya Saisubramanian, Shlomo Zilberstein, and Ece Kamar. 2020. Avoiding Negative Side Effects due to Incomplete Knowledge of AI Systems. *CoRR* abs/2008.12146 (2020).

[7] 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*. 4867–4873.