REFERENCES

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. (2016). arXiv:1606.06565

[2] Lucian Busoniu, Robert Babuska, and Bart De Schutter. 2008. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38, 2 (2008), 156–172.

[3] Bart Bussmann, Jacqueline Heinerman, and Joel Lehman. 2019. Towards Empathic Deep Q-Learning. In *Proceedings of the Workshop on Artificial Intelligence Safety 2019 co-located with the 28th International Joint Conference on Artificial Intelligence, AISafety@IJCAI (CEUR Workshop Proceedings, Vol. 2419)*. CEUR-WS.org, Aachen. http://ceur-ws.org/Vol-2419/paper_19.pdf

[4] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. Cooperative AI: machines must learn to find common ground. *Nature* 593 (2021), 33–36. https://doi.org/10.1038/d41586-021-01170-0

[5] Yuqing Du, Stas Tiomkin, Emre Kiciman, Daniel Polani, Pieter Abbeel, and Anca Dragan. 2020. AvE: Assistance via Empowerment. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc.

[6] Richard G. Freedman, Steven J. Levine, Brian C. Williams, and Shlomo Zilberstein. 2020. Helpfulness as a Key Metric of Human-Robot Collaboration. (2020). arXiv:2010.04914

[7] León Illanes, Xi Yan, Rodrigo Toro Icarte, and Sheila A. McIlraith. 2020. Symbolic Plans as High-Level Instructions for Reinforcement Learning. In *Proceedings of the Thirtieth International Conference on Automated Planning and Scheduling*. AAAI Press, 540–550.

[8] Toryn Q. Klassen and Sheila A. McIlraith. 2021. Planning to Avoid Side Effects (Preliminary Report). In *IJCAI Workshop on Robust and Reliable Autonomy in the Wild (R2AW)*. http://rbr.cs.umass.edu/r2aw/papers/R2AW_paper_15.pdf

[9] Toryn Q. Klassen, Sheila A. McIlraith, Christian Muise, and Jarvis Xu. 2022. Planning to Avoid Side Effects. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)*. To appear.

[10] Victoria Krakovna, Laurent Orseau, Miljan Martic, and Shane Legg. 2019. Penalizing Side Effects using Stepwise Relative Reachability. In *Proceedings of the Workshop on Artificial Intelligence Safety 2019 co-located with the 28th International Joint Conference on Artificial Intelligence, AISafety@IJCAI 2019 (CEUR Workshop Proceedings, Vol. 2419)*. CEUR-WS.org, Aachen. http://ceur-ws.org/Vol-2419/paper_1.pdf

[11] Victoria Krakovna, Laurent Orseau, Richard Ngo, Miljan Martic, and Shane Legg. 2020. Avoiding Side Effects By Considering Future Tasks. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc.

[12] Jim Lebans. 2020. The threat from AI is not that it will revolt, it's that it'll do exactly as it's told. CBC Radio. URL https://www.cbc.ca/radio/quirks/apr-25-deepwater-horizon-10-years-later-covid-19-and-understanding-immunity-and-more-1.5541299/the-threat-from-ai-is-not-that-it-will-revolt-it-s-that-it-ll-do-exactly-as-it-s-told-1.5541304.

[13] Stuart Russell. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Publishing Group, New York.

[14] Sandhya Saisubramanian, Ece Kamar, and Shlomo Zilberstein. 2020. A Multi-Objective Approach to Mitigate Negative Side Effects. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. 354–361. https://doi.org/10.24963/ijcai.2020/50

[15] Amartya Sen. 1974. Rawls Versus Bentham: An Axiomatic Examination of the Pure Distribution Problem. *Theory and Decision* 4, 3-4 (1974), 301–309. https://doi.org/10.1007/BF00136651

[16] Maayan Shvo. 2019. Towards Empathetic Planning and Plan Recognition. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 525–526.

[17] Maayan Shvo, Toryn Q. Klassen, and Sheila A. McIlraith. 2020. Towards the Role of Theory of Mind in Explanation. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: Second International Workshop, EXTRAAMAS 2020*. Springer-Verlag, Berlin, Heidelberg, 75–93. https://doi.org/10.1007/978-3-030-51924-7_5

[18] Maayan Shvo and Sheila A McIlraith. 2019. Towards empathetic planning. (2019). arXiv:1906.06436

[19] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (second ed.). MIT Press, Cambridge, MA. http://incompleteideas.net/book/the-book.html

[20] Richard S. Sutton, Doina Precup, and Satinder P. Singh. 1999. Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning. *Artificial Intelligence* 112, 1-2 (1999), 181–211. https://doi.org/10.1016/S0004-3702(99)00052-1

[21] Alex Turner. 2019. Reframing Impact. Blog post, https://www.lesswrong.com/s/7CdoznhJaLEKHwvJW.

[22] Alex Turner, Neale Ratzlaff, and Prasad Tadepalli. 2020. Avoiding Side Effects in Complex Environments. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc.

[23] Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. 2020. Conservative Agency via Attainable Utility Preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, United States, 385–391. https://doi.org/10.1145/3375627.3375851

[24] Carroll Wainwright and Peter Eckersley. 2020. SafeLife 1.0: Exploring Side Effects in Complex Environments. In *Proceedings of the Workshop on Artificial Intelligence Safety (SafeAI 2020) co-located with 34th AAAI Conference on Artificial Intelligence (AAAI 2020)*. 117–127. http://ceur-ws.org/Vol-2560/paper46.pdf

[25] Christopher J. C. H. Watkins and Peter Dayan. 1992. Q-Learning. *Machine Learning* 8 (1992), 279–292. https://doi.org/10.1007/BF00992698

[26] Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. 2021. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. (2021). arXiv:1911.10635

[27] 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 Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*. 4867–4873. https://doi.org/10.24963/ijcai.2018/676