ACKNOWLEDGMENTS

We wish to acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada CIFAR AI Chairs Program, and Microsoft Research. The first author also received funding from Open Philanthropy. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute for Artificial Intelligence (https://vectorinstitute.ai/partners).

REFERENCES

[1] Parand Alizadeh Alamdari, Toryn Q. Klassen, Rodrigo Toro Icarte, and Sheila A. McIlraith. 2022. Be Considerate: Avoiding Negative Side Effects in Reinforcement Learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022)*. 18–26.

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565* (2016). https://doi.org/10.48550/arXiv.1606.06565

[3] Andrea Baisero and Christopher Amato. 2022. Unbiased Asymmetric Reinforcement Learning under Partial Observability. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022)*. 44–52.

[4] Nick Bostrom. 2011. Information Hazards: A Typology of Potential Harms from Knowledge. *Review of Contemporary Philosophy* 10 (2011), 44–79.

[5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712* (2023). https://doi.org/10.48550/arXiv.2303.12712

[6] 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. http://ceur-ws.org/Vol-2419/paper_19.pdf

[7] Charles Evans and Atoosa Kasirzadeh. 2021. User Tampering in Reinforcement Learning Recommender Systems. In *4th FAccTRec Workshop on Responsible Recommendation*. https://arxiv.org/abs/2109.04083

[8] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. 1995. *Reasoning About Knowledge*. MIT Press. https://doi.org/10.7551/mitpress/5803.001.0001

[9] Joseph Y. Halpern. 2005. *Reasoning about Uncertainty*. MIT Press.

[10] Dan Hendrycks and Mantas Mazeika. 2022. X-Risk Analysis for AI Research. *arXiv preprint arXiv:2206.05862* (2022). https://doi.org/10.48550/arXiv.2206.05862

[11] Toryn Q. Klassen, Sheila A. McIlraith, Christian Muise, and Jarvis Xu. 2022. Planning to Avoid Side Effects. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI 2022)*. 9830–9839. https://doi.org/10.1609/aaai.v36i9.21219

[12] Michal Kosinski. 2023. Theory of Mind May Have Spontaneously Emerged in Large Language Models. *arXiv preprint arXiv:2302.02083* (2023). https://doi.org/10.48550/arXiv.2302.02083

[13] 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. http://ceur-ws.org/Vol-2419/paper_1.pdf

[14] 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 33 (NeurIPS 2020)*. https://papers.nips.cc/paper/2020/file/dc1913d422398c25c5f0b81cab94cc87-Paper.pdf

[15] Robert C. Moore. 1980. *Reasoning about Knowledge and Action*. Technical Note 191. SRI International. https://apps.dtic.mil/sti/citations/ADA126244

[16] Andrew Y. Ng and Stuart Russell. 2000. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*. Morgan Kaufmann, 663–670.

[17] Alexander Peysakhovich and Adam Lerer. 2018. Prosocial Learning Agents Solve Generalized Stag Hunts Better than Selfish Ones. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2018)*. 2043–2044.

[18] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1, 4 (1978), 515–526. https://doi.org/10.1017/S0140525X00076512

[19] Duncan Pritchard, John Turri, and J. Adam Carter. 2022. The Value of Knowledge. In *The Stanford Encyclopedia of Philosophy* (Fall 2022 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2022/entries/knowledge-value/

[20] Sandhya Saisubramanian, Ece Kamar, and Shlomo Zilberstein. 2022. Avoiding Negative Side Effects of Autonomous Systems in the Open World. *Journal of Artificial Intelligence Research* 74 (2022), 143–177. https://doi.org/10.1613/jair.1.13581

[21] Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 3762–3780. https://aclanthology.org/2022.emnlp-main.248

[22] Manisha Senadeera, Thommen George Karimpanal, Sunil Gupta, and Santu Rana. 2022. Sympathy-based Reinforcement Learning Agents. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022)*. 1164–1172.

[23] Robert Stalnaker. 1991. The Problem of Logical Omniscience, I. *Synthese* 89, 3 (1991), 425–440. https://doi.org/10.1007/BF00413506

[24] 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)*. 385–391. https://doi.org/10.1145/3375627.3375851

[25] Tomer Ullman. 2023. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. *arXiv preprint arXiv:2302.08399* (2023). https://doi.org/10.48550/arXiv.2302.08399

[26] Peter Vamplew, Cameron Foale, Richard Dazeley, and Adam Bignold. 2021. Potential-based multiobjective reinforcement learning approaches to low-impact agents for AI safety. *Engineering Applications of Artificial Intelligence* 100 (2021), 104186. https://doi.org/10.1016/j.engappai.2021.104186

[27] Wiebe van der Hoek and Michael J. Wooldridge. 2002. Tractable Multiagent Planning for Epistemic Goals. In *The First International Joint Conference on Autonomous Agents & Multiagent Systems, AAMAS 2002*. 1167–1174. https://doi.org/10.1145/545056.545095

[28] Audrey Wang, Rohan Chitnis, Michelle Li, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. 2020. A Unifying Framework for Social Motivation in Human-Robot Interaction. In *The AAAI 2020 Workshop on Plan, Activity, and Intent Recognition (PAIR 2020)*.

[29] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William S. Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 214–229. https://doi.org/10.1145/3531146.3533088

[30] Chengwei Zhang, Xiaohong Li, Jianye Hao, Siqi Chen, Karl Tuyls, Wanli Xue, and Zhiyong Feng. 2019. SA-IGA: a multiagent reinforcement learning method towards socially optimal outcomes. *Autonomous Agents and Multi-Agent Systems* 33, 4 (2019), 403–429. https://doi.org/10.1007/s10458-019-09411-3

[31] 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