REFERENCES

[1] Ofra Amir, Ece Kamar, Andrey Kolobov, and Barbara Grosz. 2016. Interactive Teaching Strategies for Agent Training. *IJCAI* (2016).

[2] Michael Borenstein, Larry V Hedges, Julian Higgins, and Hannah R Rothstein. 2009. Converting among effect sizes. *Introduction to meta-analysis* (2009), 45–49.

[3] Daniel J Brooks, Abraham Shultz, Munjal Desai, Philip Kovac, and Holly A Yanco. 2010. Towards State Summarization for Autonomous Robots.. In *AAAI Fall Symposium: Dialog with Robots*, Vol. 61. 62.

[4] Henian Chen, Patricia Cohen, and Sophie Chen. 2010. How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in StatisticsâĂŤSimulation and Computation®* 39, 4 (2010), 860–864.

[5] Sandra Devin and Rachid Alami. 2016. An implemented theory of mind to improve human-robot shared plans execution. In *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on*. IEEE, 319–326.

[6] Thomas Dodson, Nicholas Mattei, and Judy Goldsmith. 2011. A natural language argumentation interface for explanation generation in Markov decision processes. *Algorithmic Decision Theory* (2011), 42–55.

[7] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. (2017).

[8] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. 2008. Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces*. ACM, 227–236.

[9] Bradley Hayes and Julie A Shah. 2017. Improving Robot Controller Transparency Through Autonomous Policy Explanation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 303–312.

[10] S. Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 65-70 (1979), 1979.

[11] O Khan, Pascal Poupart, J Black, LE Sucar, EF Morales, and J Hoey. 2011. Automatically generated explanations for Markov decision processes. *Decision Theory Models for Applications in Artificial Intelligence: Concepts and Solutions* (2011), 144–163.

[12] Omar Zia Khan, Pascal Poupart, and James P Black. 2009. Minimal Sufficient Explanations for Factored Markov Decision Processes.. In *ICAPS*.

[13] Been Kim, Cynthia Rudin, and Julie A Shah. 2014. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in Neural Information Processing Systems*. 1952–1960.

[14] Meghann Lomas, Robert Chevalier, Ernest Vincent Cross II, Robert Christopher Garrett, John Hoare, and Michael Kopack. 2012. Explaining robot actions. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, 187–188.

[15] Karen L Myers. 2006. Metatheoretic Plan Summarization and Comparison.. In *ICAPS*. 182–192.

[16] Stefanos Nikolaidis and Julie Shah. 2013. Human-robot cross-training: computational formulation, modeling and evaluation of a human team training strategy. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 33–40.

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386* (2016).

[18] Philipp Rohlfshagen and Simon M Lucas. 2011. Ms pac-man versus ghost team cec 2011 competition. In *Evolutionary Computation (CEC), 2011 IEEE Congress on*. IEEE, 70–77.

[19] Bastian Seegebarth, Felix Müller, Bernd Schattenberg, and Susanne Biundo. 2012. Making hybrid plans more clear to human users-a formal approach for generating sound explanations. In *Twenty-Second International Conference on Automated Planning and Scheduling*.

[20] Juliet P. Shaffer. 1995. Multiple Hypothesis-Testing. *Annual Review of Psychology* 46 (1995), 561–584.

[21] Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, Press William, Saxenian AnnaLee, Shah Julie, Tambe Milind, and Teller Astro. 2016. Artificial intelligence and life in 2030. *One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel* (2016).

[22] Leila Takayama, Doug Dooley, and Wendy Ju. 2011. Expressing thought: improving robot readability with animation principles. In *Proceedings of the 6th international conference on Human-robot interaction*. ACM, 69–76.

[23] Lisa Torrey and Matthew Taylor. 2013. Teaching on a budget: Agents advising agents in reinforcement learning. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1053–1060.

[24] Alfredo Vellido, José David Martín-Guerrero, and Paulo JG Lisboa. 2012. Making machine learning models interpretable.. In *ESANN*, Vol. 12. 163–172.