REFERENCES

[1] Dimitrios Antos and Avi Pfeffer. 2008. Identifying Reasoning Patterns in Games. (2008), 9–18.

[2] Ryan Carey, Eric Langlois, Tom Everitt, and Shane Legg. 2020. The Incentives that Shape Behaviour. *arXiv:2001.07118* (2020).

[3] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. 2009. *Introduction to Algorithms* (third ed.). The MIT Press, Cambridge, Massachusettss.

[4] Tom Everitt, Ryan Carey, Eric Langlois, Pedro Ortega, and Shane Legg. 2021. Agent Incentives: a Causal Perspective. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, (AAAI-21)*. Virtual. Forthcoming.

[5] Tom Everitt and Marcus Hutter. 2019. Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective. *arXiv:1908.04734* (2019).

[6] Tom Everitt, Ramana Kumar, Victoria Krakovna, and Shane Legg. 2019. Modeling AGI Safety Frameworks with Causal Influence Diagrams. *arXiv:1906.08663* (2019).

[7] Y. Gal and A. Pfeffer. 2008. Networks of Influence Diagrams: A Formalism for Representing Agents' Beliefs and Decision-Making Processes. *Journal of Artificial Intelligence Research* 33 (2008), 109–147.

[8] Koen Holtman. 2020. Towards AGI Agent Safety by Iteratively Improving the Utility Function. In *Artificial General Intelligence*. Springer International Publishing, 205–215.

[9] Ronald A. Howard and James E. Matheson. 2005. Influence Diagrams. *Decision Analysis* 2, 3 (2005), 127–143.

[10] Albert Xin Jiang, Kevin Leyton-Brown, and Avi Pfeffer. 2009. Temporal Action-Graph Games: A New Representation for Dynamic Games. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (Montreal, Quebec, Canada) *(UAI '09)*. AUAI Press, 268–276.

[11] Marek Mikolaj Kaminski. 2019. Generalized Backward Induction: Justification for a Folk Algorithm. *Games* 10, 3 (2019), 34.

[12] Daphne Koller and Brian Milch. 2001. Multi-agent Influence Diagrams for Representing and Solving Games. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2* (Seattle, WA, USA) *(IJCAI'01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1027–1034.

[13] Daphne Koller and Brian Milch. 2003. Multi-agent Influence Diagrams for Representing and Solving Games. *Games and Economic Behavior* 45, 1 (2003), 181–221.

[14] Harold William Kuhn and Albert William Tucker (Eds.). 1953. *Contributions to the Theory of Games (AM-28), Volume II.* Princeton University Press.

[15] Eric Langlois and Tom Everitt. 2021. How RL Agents Behave When Their Actions Are Modified. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, (AAAI-21)*. Virtual. Forthcoming.

[16] Michael Maschler, Eilon Solan, and Shmuel Zamir. 2009. *Game Theory*. Cambridge University Press.

[17] Richard D. McKelvey, Andrew M. McLennan, and Theodore L. Turocy. 2016. Gambit: Software Tools for Game Theory. http://www.gambit-project.org.

[18] Brian Milch and Daphne Koller. 2008. Ignorable Information in Multi-Agent Scenarios. MIT-CSAIL-TR-2008-029 (2008).

[19] J. F. Nash. 1950. Equilibrium Points in N-person Games. *Proceedings of the National Academy of Sciences* 36, 1 (1950), 48–49.

[20] Judea Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[21] Judea Pearl. 2009. *Causality*. Cambridge University Press, Cambridge, UK.

[22] Avi Pfeffer and Ya'akov Gal. 2007. On the Reasoning Patterns of Agents in Games. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 1* (Vancouver, British Columbia, Canada) *(AAAI'07)*. AAAI Press, 102–109.

[23] Michele Piccione and Ariel Rubinstein. 1997. The Absent-Minded Driver's Paradox: Synthesis and Responses. *Games and Economic Behavior* 20, 1 (1997), 121–130.

[24] Reinhard Selten. 1965. Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit: Teil i: Bestimmung des Dynamischen Preisgleichgewichts. *Journal of Institutional and Theoretical Economics* H. 2 (1965), 301–324.

[25] R. Selten. 1975. Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games. *International Journal of Game Theory* 4, 1 (1975), 25–55.

[26] Michael Spence. 1973. Job Market Signaling. Vol. 87. Oxford University Press (OUP), 355.

[27] Halbert White and Karim Chalak. 2009. Settable Systems: An Extension of Pearl's Causal Model with Optimization, Equilibrium, and Learning. *Journal of Machine Learning Research* 10 (2009), 1759—-1799.