Discovering Agents (Abstract Reprint)

Zachary Kenton[1], Ramana Kumar[1], Sebastian Farquhar[1], Jonathan Richens[1], Matt MacDermott[2], Tom Everitt[1]

[1]DeepMind, United Kingdom of Great Britain and Northern Ireland
[2]Imperial College London, United Kingdom of Great Britain and Northern Ireland

Abstract Reprint. This is an abstract reprint of a journal article by Kenton, Kumar, Farquhar, Richens, MacDermott, and Everitt (2023).

Abstract

Causal models of agents have been used to analyse the safety aspects of machine learning systems. But identifying agents is non-trivial – often the causal model is just assumed by the modeller without much justification – and modelling failures can lead to mistakes in the safety analysis. This paper proposes the first formal causal definition of agents – roughly that agents are systems that would adapt their policy if their actions influenced the world in a different way. From this we derive the first causal discovery algorithm for discovering the presence of agents from empirical data, given a set of variables and under certain assumptions. We also provide algorithms for translating between causal models and game-theoretic influence diagrams. We demonstrate our approach by resolving some previous confusions caused by incorrect causal modelling of agents.

References

Kenton, Z.; Kumar, R.; Farquhar, S.; Richens, J.; MacDermott, M.; and Everitt, T. 2023. Discovering agents. *Artificial Intelligence*, 322: 103963.