Reasoning about Causality in Games

Lewis Hammond[1], James Fox[1], Tom Everitt[2], Ryan Carey[1], Alessandro Abate[1], Michael Wooldridge[1]
[1]University of Oxford, [2]DeepMind

Motivation

- We want to make AI systems safer, fairer, and better at cooperating (in multi-agent settings).
- Therefore, we want to predict the behaviour of agents as a result of their objectives and the environment.
- The causal structure of an agent's environment determines important aspects of an agent's behaviour.

Contributions

- We introduce (structural) causal games, generalising:
  - Causal Bayesian Networks and Structural Causal Models [4] to the game-theoretic domain.
  - Multi-agent influence diagrams [3] to the causal domain.
- We introduce mechanism variables to these models in order to represent strategic dependencies.
- We show how causal games can be used to answer various kinds of associative, interventional, and counterfactual queries.

Models

- A multi-agent influence diagram (MAID) *M = (G, θ)* specifies:
  - a graph *G = (N, V, E)* with players *N*, *vertices* $V = X \cup \{D^i\}_{i \in N} \cup \{U^i\}_{i \in N}$, and edges $E \subset V \times V$
  - *parameters* $\theta = \{\theta_v\}_{V \in X \cup U}$ *that define CPDs* $Pr(x,u : d) \coloneqq \Pi_{V \in X \cup U} Pr(v \mid pa_v; \theta_v)$ *for every non-decision variable.*
- *A (structural) causal game is a* MAID *M = (G, θ) such that for any (deterministic) parameterisation of the decision variable CPDs π, the induced model with distribution* $Pr^{\pi}(V)$ *is a CBN (SCM).*
- *Mechanised games explicitly represent the CPDs θ and the decision rules π.*

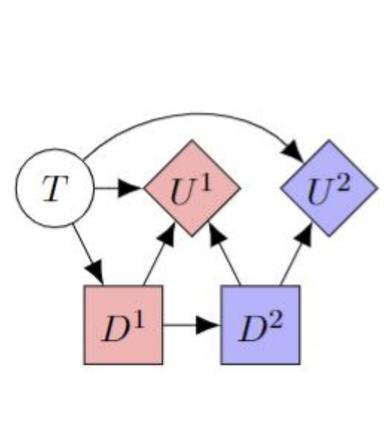
MAID Relevance Graph Subgames

Subgames and Equilibrium Refinements

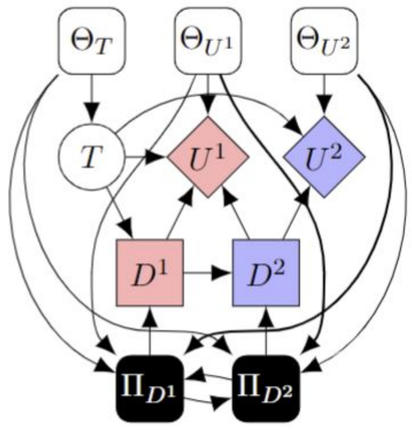
- A Nash equilibrium is a policy profile such that no agent has an incentive to unilaterally deviate.
- A subgame is a part of the full game that can be solved independently from the rest.
- A subgame perfect equilibrium is a nash equilibrium in every (feasible) subgame.
- Since more subgames can be identified in MAIDs than in extensive form games, subgame perfect equilibria can rule out more non-credible threats.

Causal Queries

- Unlike in standard causal models, queries in games:
  - Can be made with or without agents' awareness (characterised as pre- or post-policy queries in the mechanised game, respectively).
  - Are best conceptualised as first-order, where the policy profile π is a free variable, typically belonging to some set of rational outcomes, e.g., $\varphi(\pi) \equiv Pr^{\pi}(u^1_{d\_1})$ and $\max_{\pi \in NE(M)} \varphi(\pi) \geq p$



MAID



Mechanised MAID

Description of Examples

	Post-Policy	Pre-Policy
Associative	$\Pr^{\pi}(u^1 \mid d_1)$	$\Pr(u^2 \mid \bar{\pi}_{D^1})$
Interventional	$\Pr^{\pi}(u^1_{d_1})$	$\Pr(u^2_{\hat{\pi}_{D^1}})$
Counterfactual	$\Pr^{\pi}(u^1_{d_1} \mid \neg d_1)$	$\Pr(u^2_{\tilde{\pi}_{D^1}} \mid \tilde{\pi}_{D^1})$

Applications

- Formal definitions of important philosophical concepts such as agency, incentives, intention, blame, manipulation, signaling, social influence, harm, threats and offers, etc.
- Mechanism design and economic analysis.

References

1. A. P. Dawid, "Influence Diagrams for Causal Modelling and Inference," International Statistical Review (70:2), pp. 161–189. International Statistical Institute (ISI). 2002.
2. L. Hammond, J. Fox, T. Everitt, A. Abate, and M. Wooldridge, "Equilibrium Refinements for Multi-Agent Influence Diagrams: Theory and Practice," AAMAS-21, pp. 574–582. 2021.
3. D. Koller and B. Milch, "Multi-agent Influence Diagrams for Representing and Solving Games," Games and Economic Behavior (45:1), pp. 181–221. Elsevier. 2003.
4. J. Pearl, Causality. Cambridge University Press. 2009

Acknowledgements

Lewis Hammond was supported by an EPSRC Doctoral Training Partnership studentship (Reference: 2218880), James Fox was supported by the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems (Reference: EP/SO24050/1), and Michael Wooldridge was supported by a UKRI Turing AI World Leading Researcher Fellowship (Reference: EP/W002949/1)

Full paper available here

Contact:
james.fox@cs.ox.ac.uk

EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines & Systems