Agent-Specific Effects: A Causal Effect Propagation Analysis in Multi-Agent MDPs

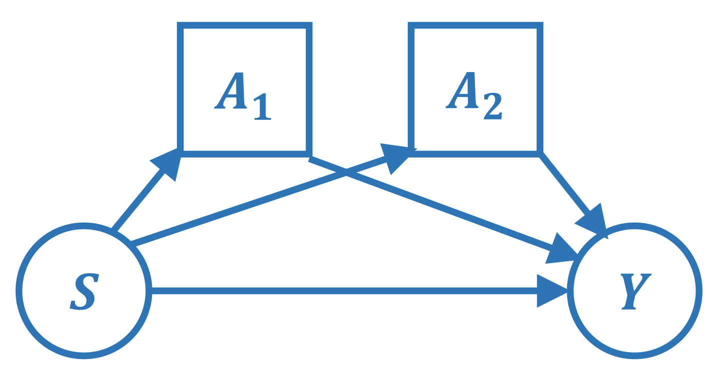
MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS
WARWICK THE UNIVERSITY OF WARWICK

Stelios Triantafyllou strianta@mpi-sws.org
Aleksa Sukovic asukovic@mpi-sws.org
Debmalya Mandal Debmalya.Mandal@warwick.ac.uk
Goran Radanovic gradanovic@mpi-sws.org

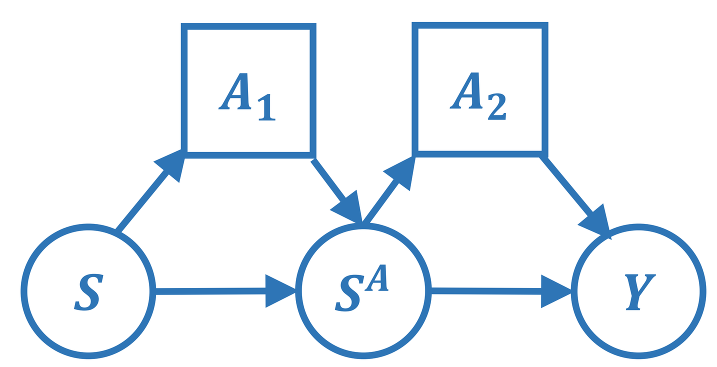
ICML International Conference On Machine Learning

Research Question

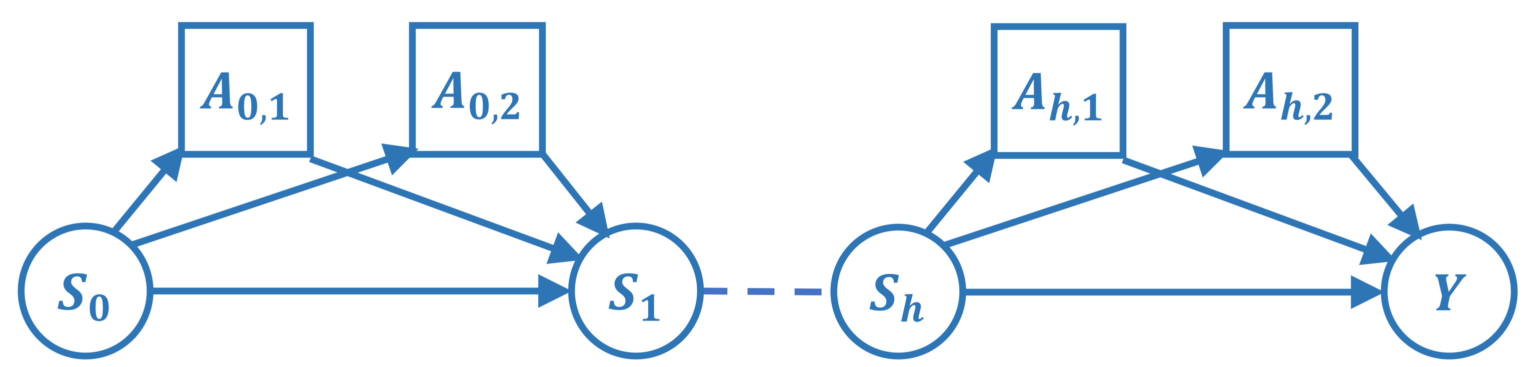
How to measure the effect of an action that propagates through a set of agents in multi-agent sequential decision making?



Multi-Agent One-Step



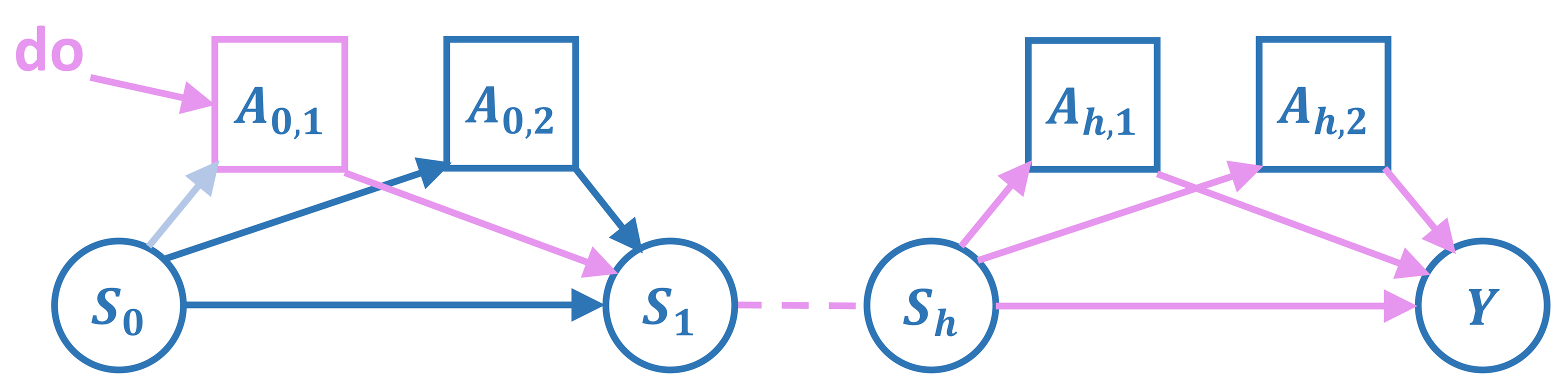
Turn-Based One-Step



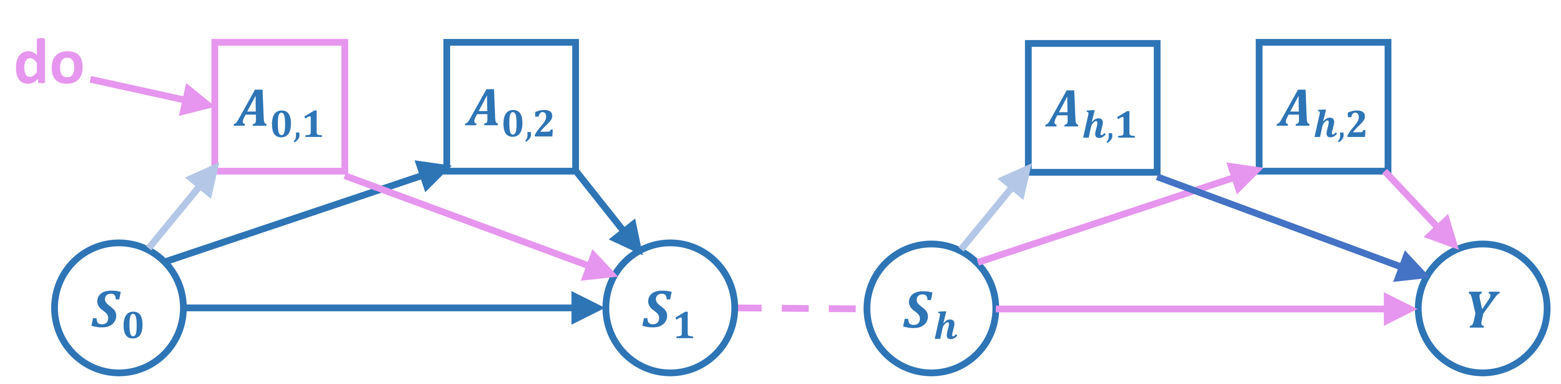
Multi-Agent Sequential

Setup & Prior Work

Framework: Multi-Agent Markov Decision Processes (MMDPs) & Structural Causal Models (SCMs) with **categorical** observed variables.



Total Counterfactual Effect (TCFE)



Path-Specific Effect (PSE) [1]

Problem: The PSE approach in this setting can lead to counter-intuitive results. For example, in scenarios where the actions of Agent 2 do not affect the environment state, PSE might still have a positive evaluation. PSE also fails to capture higher-order dependencies between agents' actions.

References

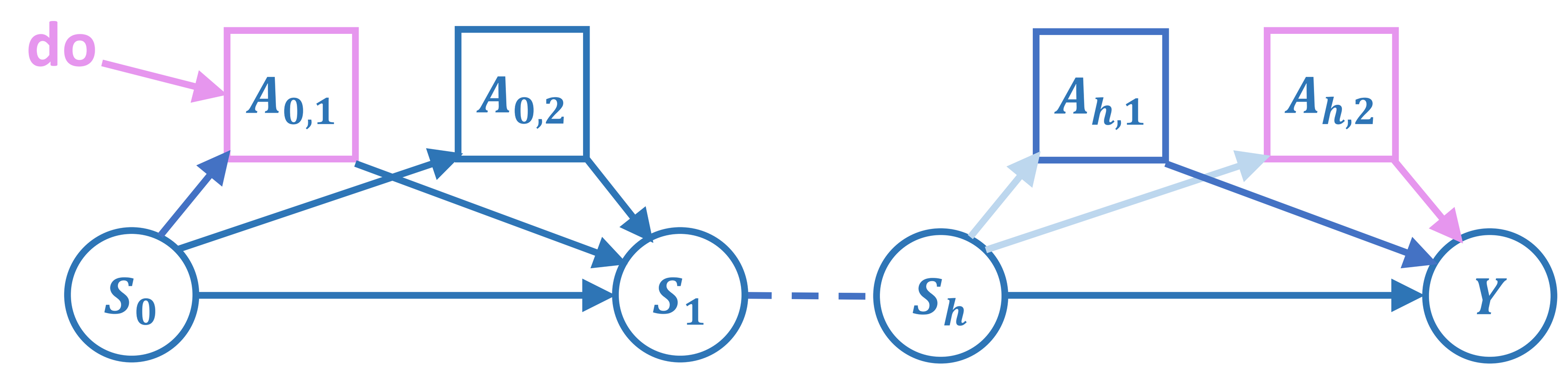
[1] Pearl J., 2001. Direct and indirect effects. UAI.

[2] Lu, C., Huang, B., Wang, K., Hernandez-Lobato, J. M., Zhang, K., & Scholkopf, B., 2020. Sample-efficient reinforcement learning via counterfactual-based data augmentation. NeurIPS Workshop

[3] Pearl, J., 2009. Causality. Cambridge University Press.

This Work

Main Idea: *Effect agents* should behave as if other agents' actions are **not fixed**, but rather responsive to the considered intervention.



Agent-Specific Effect (ASE)

Distinctions to PSE:

- The actions of the **effect agents** are fixed to the values that they would **naturally** take under the intervention.
- The effect is measured w.r.t. the **factual/reference value** of $A_{0,1}$.

Remark: ASE **cannot** be expressed by PSE.

Identifiability Results (Informal)

Problem: ASE is in general **non-identifiable** without further assumptions.

*Noise Monotonicity: Given an SCM M with causal graph G, we say that V^i is noise-monotonic in M w.r.t. to a **total ordering** $\leq_i$ on $dom\{V^i\}$, if for any pa^i and u_1^i, u_2^i s.t. $u_1^i < u_2^i$ it holds that $f^i(pa^i, u_1^i) \leq_i f^i(pa^i, u_2^i)$.*

Theorem: Every MMDP can be **represented** by an SCM whose observed variables V^i satisfy noise-monotonicity w.r.t. to some total ordering $\leq_i$.

Theorem*: ASE and its **counterfactual counterpart cf-ASE** are **identifiable** under the **assumptions** of *exogeneity* and noise monotonicity.

*[2] shows a similar result but assuming **strong** noise monotonicity.

Other Results

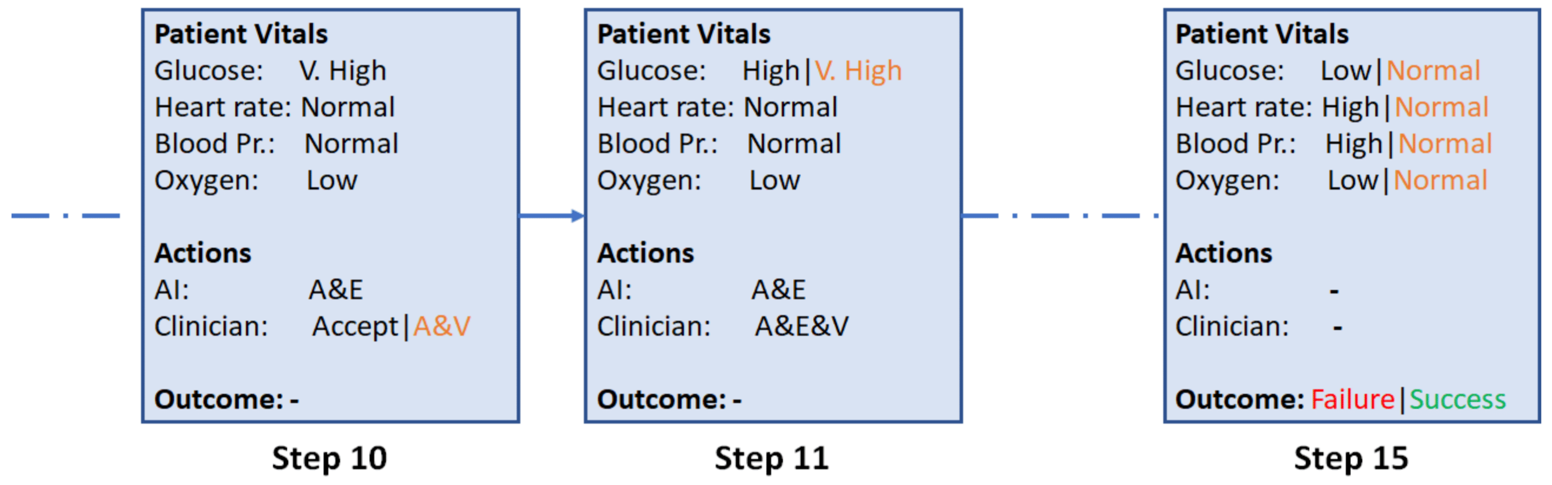
Algorithm: ASE is measured following the standard *abduction-action-prediction* methodology for counterfactual inference [3]. Our algorithm makes use of observational data to output an **unbiased estimator** of ASE, given that noise monotonicity holds.

Connections to PSE: We introduce the **fixed path-specific effects** (FPSE), a causal notion that generalizes PSE by reasoning across 3 (instead of 2) alternative worlds. **Importantly**, FPSE can be used to express ASE.

Experiments

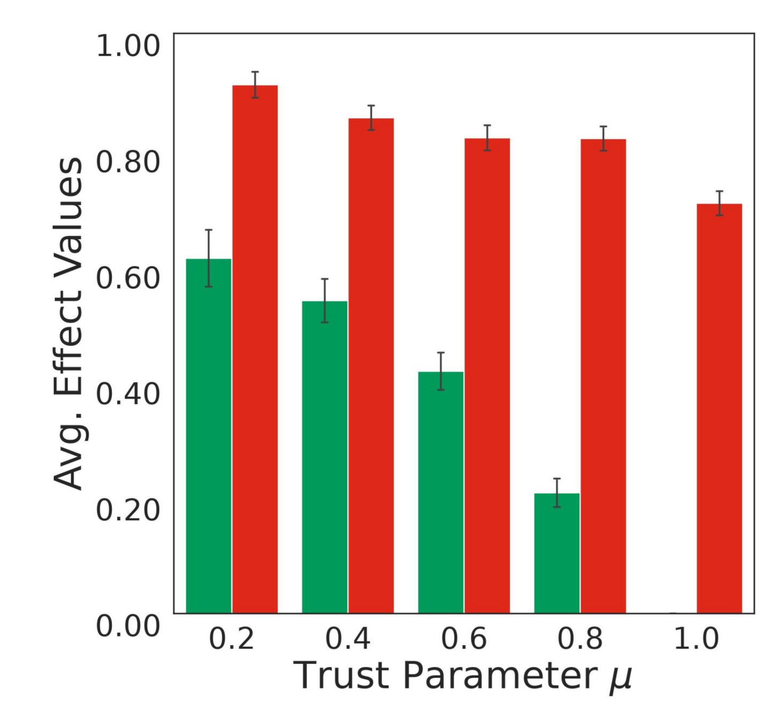
Environments: Two-agent Sepsis management simulator **(poster)** with *AI* (actor) and *clinician* (supervisor), and a Graph environment.

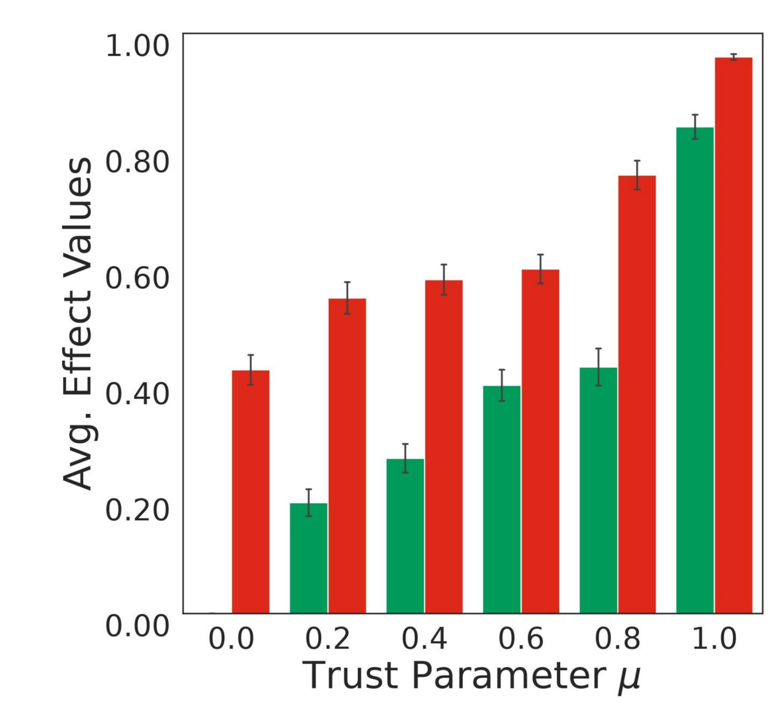
Evaluation Criteria: Practicality **(poster)** and robustness to uncertainty.



Example Scenario. We estimate that if the clinician had overridden the AI's action at time-step 10 with A&V, the treatment would have been successful with an **82% likelihood**, i.e., $TCFE = 0.82$. The **AI-specific effect** in this scenario, as measured by ASE, is equal to $\mathbf{0.23}$.

Trust Parameter μ: Models the clinician's level of trust in the AI's actions. Greater values of μ correspond to higher levels of trust and lower probabilities of action override from the clinician.





Average Effects. For various values of μ, we estimate the effects that propagate through the clinician **(left)** and AI **(right)**, as measured by **PSE** and **ASE**, respectively. **Results** indicate that:

- The effect of an agent's action on the patient outcome can be frequently attributed to the behavior of the other agent.
- **ASE** aligns better with standard intuition compared to **PSE**.

Future Work

A *causal explanation formula*, tailored to MMDPs, that **decomposes** the **TCFE** of an agent's action by attributing to each **agent** and **state** variable a score reflecting their contributions to the effect, **utilizing ASE**.