# Counterfactual Effect Decomposition in Multi-Agent Sequential Decision Making

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

**Stelios Triantafyllou** strianta@mpi-sws.org

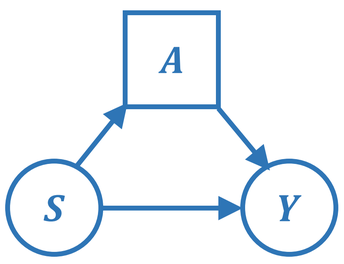
**Aleksa Sukovic** asukovic@mpi-sws.org

**Yasaman Zolfimoselo** yasamanzolfi4@gmail.com

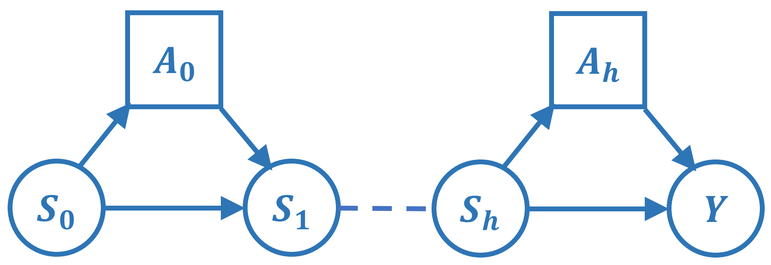
**Goran Radanovic** gradanovic@mpi-sws.org

## Research Question

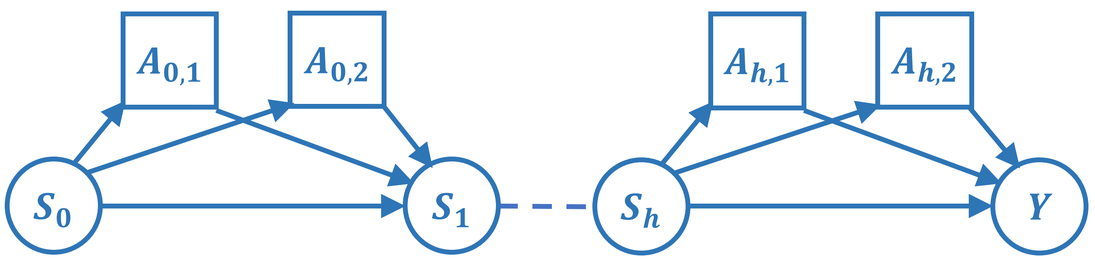
***How to explain the total counterfactual effect of an action in multi-agent sequential decision making?***




Single-Agent




Single-Agent **Sequential**




**Multi-Agent Sequential**

## Prior Work

**Mediation Analysis** aims to understand how causal effects propagate through different **paths** in the causal graph. Much prior work [1] focuses on **decomposing** causal effects under this rubric.

**Problem:** In the **multi-agent sequential** decision making setting, the causal graph can contain exponentially many paths connecting an action to the outcome. Furthermore, not all of these paths have a clear operational meaning to help explain the effect intuitively.

## This Work

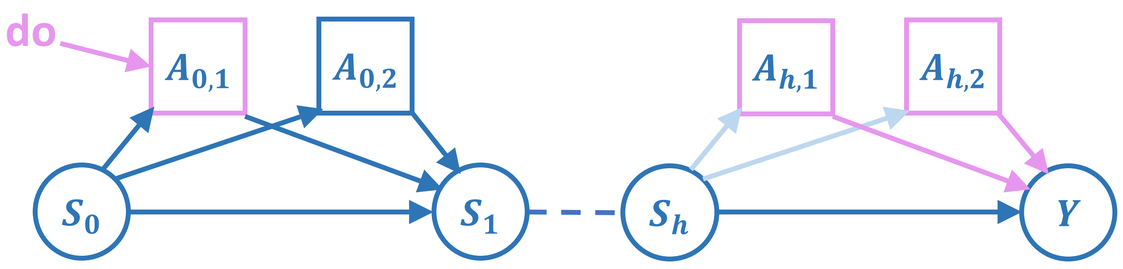
**Main Idea:** It is more natural to interpret the effect of an action in terms of its influence on the **agents' behavior** and the **environment dynamics**.

**Framework:** Multi-Agent Markov Decision Processes (MMDPs) and SCMs.

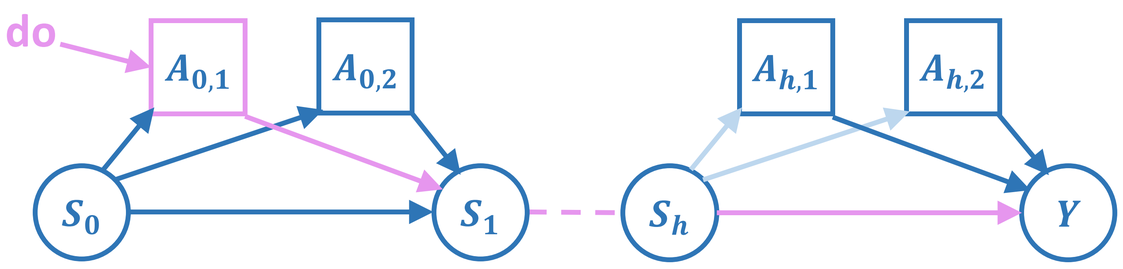
**Bilevel Decomposition Approach:** Attribute to each **agent** and subsequent **state variable** a score reflecting its respective contribution to the TCFE.

### (Level 1) Causal Explanation Formula

**Theorem:** TCFE is equal to the t-ASE *minus* the SSE of the **reverse** transition.

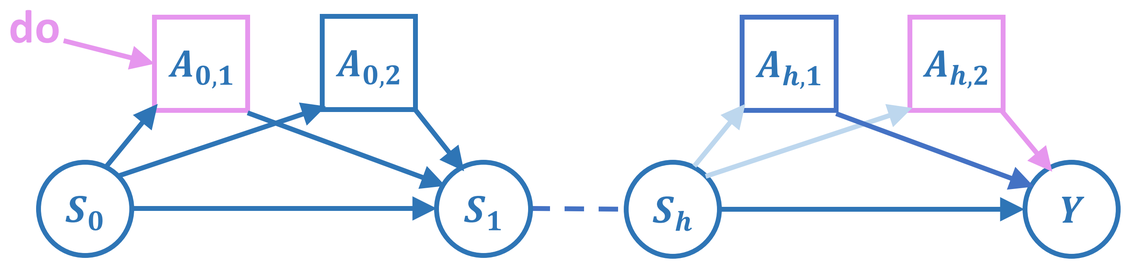



**Total Agent-Specific Effect (t-ASE)**




**State-Specific Effect (SSE)**

### (Level 2a) Decomposing the t-ASE




**Agent-Specific Effect (ASE) [2]**

**ASE-SV:** Uses **Shapley value** to attribute t-ASE to the **agents** based on ASE.

**Theorem:** ASE-SV is a **unique** attribution method for t-ASE that satisfies *efficiency, invariance, symmetry* and *contribution monotonicity.*

### (Level 2b) Decomposing the (reverse) SSE

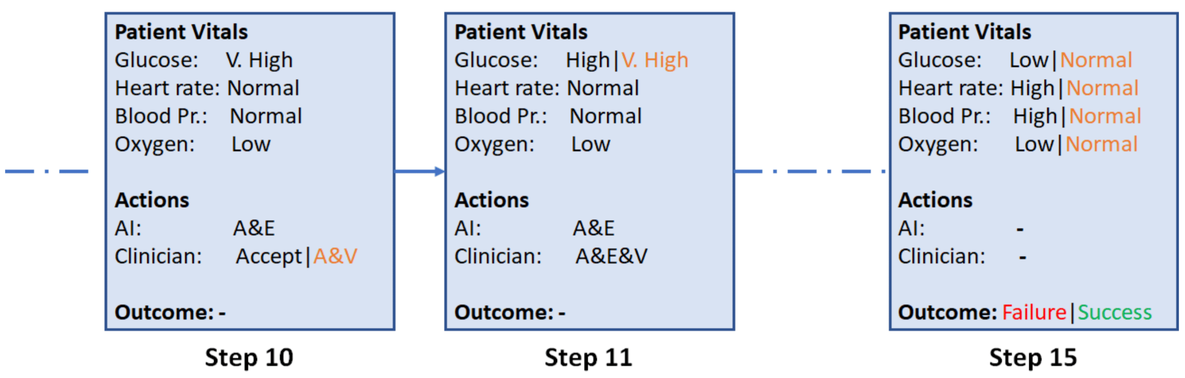
*Intrinsic Causal Contribution (ICC) [3]: The ICC of an observed variable $X$ to a target variable Y measures the reduction of uncertainty, here **variance**, in $Y$ when conditioning on the noise variable $U^X$.*

**r-SSE-ICC:** Attributes r-SSE to the **state variables** based on their marginal ICC to the counterfactual outcomes related to the computation of the effect.

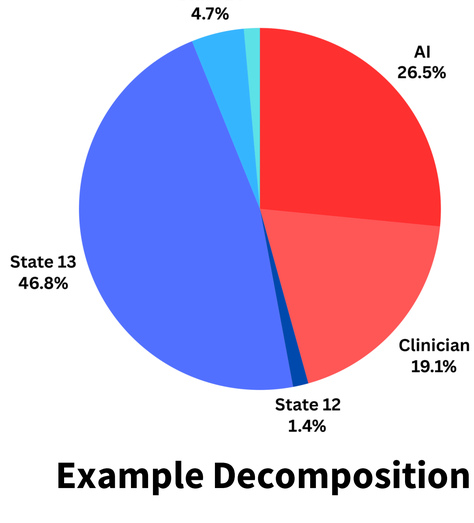
**Properties:** r-SSE-ICC is *efficient* and **does not require modifying the causal mechanisms of the underlying environment**.

## Experiments

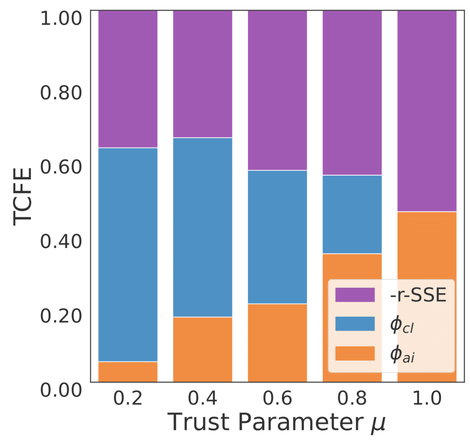
**Environments:** Two-agent ***Sepsis management*** simulator and a Gridworld environment with LLM-assisted RL agents.



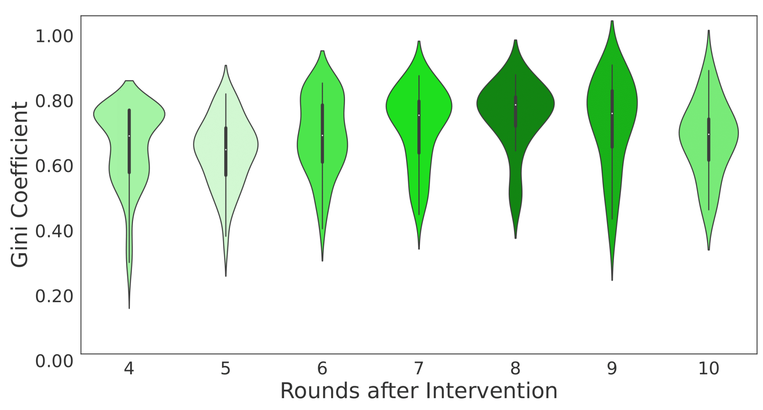

**Example Scenario**. We estimate that if the clinician had not followed the AI's recommendation at time-step 10, the treatment would have been successful with an **82% likelihood**, i.e., $TCFE = 0.82$.




**Example Decomposition**




**Avg. % Decomposition**




**Gini coefficient distribution over r-SSE-ICC scores**

**References**

[1] Zhang, J. & Bareinboim, E., 2018. Non-parametric path analysis in structural causal models. UAI.

[2] Triantafyllou, S., Sukovic, A., Mandal D. & Radanovic G., 2024. Agent-Specific Effects. ICML.

[3] Janzing, D., Blöbaum, P., Mastakouri, A. A., Faller, P. M., Minorics, L., & Budhathoki, K. 2024. Quantifying intrinsic causal contributions via structure preserving interventions. AISTATS.