# Counterfactual Effect Decomposition in Multi-Agent Sequential Decision Making

**Stelios Triantafyllou**[1] **Aleksa Sukovic**[1, 2] **Yasaman Zolfimoselo**[1] **Goran Radanovic**[1]

[1]Max Planck Institute for Software Systems
[2]Saarland University

## Abstract

We address the challenge of explaining counterfactual outcomes in multi-agent Markov decision processes. In particular, we aim to explain the total counterfactual effect of an agent's action to some realized outcome through its influence on the environment dynamics and the agents' behavior. To achieve this, we introduce a novel *causal explanation formula* that decomposes the counterfactual effect of an agent's action by attributing to each agent and state variable a score reflecting its respective contribution to the effect.

## 1 INTRODUCTION

Applying counterfactual reasoning to retrospectively analyze the impact of different actions in decision making scenarios is pivotal for accountability. One popular such measure is the notion of *total counterfactual effects*, which quantifies the extent to which an alternative action would have affected the outcome of a realized scenario.

In multi-agent sequential decision making, an agent's action typically affects the outcome indirectly. To illustrate this, consider the problem of AI-assisted decision making in healthcare Lynn [2019], where a clinician and their AI assistant treat a patient over a period of time. Fig. 1a depicts a specific example, where the treatment fails. We estimate that if the clinician had not followed the AI's recommendation at step 10 and administered vasopressors (V) instead of mechanical ventilation (E), the treatment would have been successful with an $82\%$ likelihood. This counterfactual effect, however, propagates through all subsequent actions of the clinician and the AI, as well as the subsequent changes in the patient's state. This makes the interpretability of the total counterfactual effect more nuanced, as the change from action to outcome can be transmitted by multiple distinct causal mechanisms. Hence, in this work we ask:

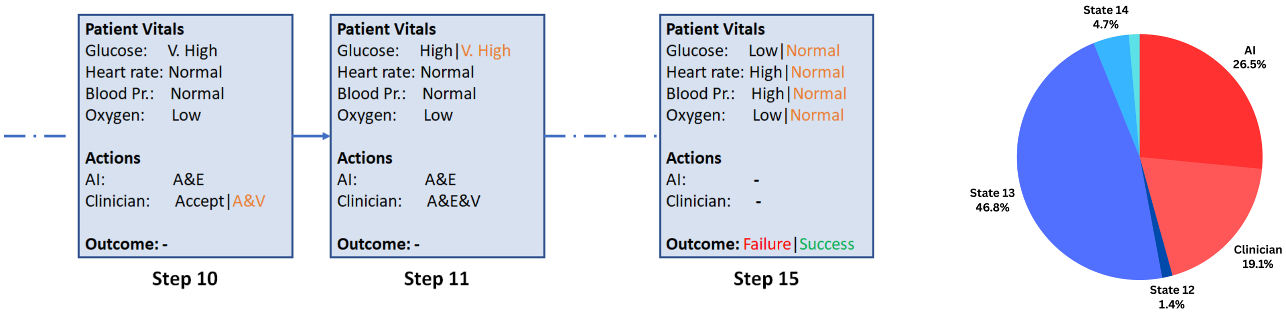
*How to explain the total counterfactual effect of an action in multi-agent sequential decision making?*

Much prior work in causality has focused on decomposing causal effects Pearl [2001], Zhang and Bareinboim [2018a,b] under the rubric of *mediation analysis* Imai et al. [2010, 2011], VanderWeele [2016], which aims to understand how effects propagate through causal paths. However, such an approach would not yield interpretability in multi-agent sequential decision making. There can be exponentially many paths connecting an action to the outcome, and not all of them have a clear operational meaning to help explain the effect intuitively. We instead posit that it is more natural to interpret the effect of an action in terms of its influence on the agents' behavior and the environment dynamics. In the previous example, the total counterfactual effect of the considered action can be decomposed as shown in Fig. 1b. This approach explains the effect by attributing a score to each doctor (clinician and AI) and patient state, reflecting their respective contributions to the overall effect.

**Contribution.** Focusing on Multi-Agent Markov Decision Processes (MMDPs) Boutilier [1996] and Structural Causal Models (SCMs) Pearl [2009], we provide a systematic approach to attributing the total counterfactual effect of an agent's action based on a novel bi-level decomposition.

## 2 LEVEL 1

We utilize the MMDP-SCM framework Triantafyllou et al. [2024] to express MMDPs as SCMs. Given an MMDP-SCM $M$ with $n$ agents, a trajectory $\tau$ of $M$ and a response variable $Y$, we denote as $\text{TCFE}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M$ the total counterfactual effect of action $a_{i,t}$ on $Y$, relative to the factual action $\tau(A_{i,t})$. Here, subscripts $i$ and $t$ denote the corresponding agent and time-step of an action, respectively. The first step of our approach entails a *causal explanation formula* that establishes a relationship between TCFE and the counterfactual effects of $a_{i,t}$ on $Y$ that propagate only through the agents and state transitions of $M$, respectively.



(a) Factual and Counterfactual Trajectory

(b) Decomposition

Figure 1: 1a depicts (part of) a simulated scenario from the two-agent *Sepsis* environment proposed by Triantafyllou et al. [2024], where the patient's treatment fails. In the same figure, we have also included the values from a sampled counterfactual scenario (different values are shown in orange), where the clinician's action is fixed to override the AI at step 10. Hence, the patient receives treatment A&V instead of A&E. 1b shows the results of our decomposition approach for this scenario.

We formulate these effects based on two causal quantities inspired from prior work. Formal definitions and the proof of Theorem 2.1 can be found in the Supplementary Material.

**Agent-specific effects Triantafyllou et al. [2024].** The **N**-specific effect $\mathrm{ASE}^{\mathbf{N}}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M$ quantifies the counterfactual effect of $a_{i,t}$ on $Y$ that propagates only through a subset of agents $\mathbf{N}$. ASE measures the counterfactual value of $Y$ in $\tau$, had all subsequent actions of agents in $N$ been fixed to the values that they would naturally take under $a_{i,t}$, and all other actions were fixed to their factual values.

**State-specific effects.** $\mathrm{SSE}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M$ measures the counterfactual effect of $a_{i,t}$ on $Y$ in a modified model, where all subsequent agents' actions are fixed to their factual values, i.e., their actions in $\tau$. As such, SSE quantifies the effect that propagates only through the changes in the subsequent state variables and can be seen as a special case of the standard *path-specific effects* Avin et al. [2005].

**Theorem 2.1.** $\mathrm{TCFE}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M$ *is equal to* $\mathrm{ASE}^{\{1,...,n\}}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M - \mathrm{SSE}_{\tau(A_{i,t}),a_{i,t}}(Y|\tau)_M$.

In words, Theorem 2.1 states that the total counterfactual effect associated with the transition from the factual action $\tau(A_{i,t})$ to the counterfactual action $a_{i,t}$ is equal to the *total* agent-specific effect of the same transition *minus* the state-specific effect of the reverse transition.

## 3 LEVEL 2: AGENTS

To further decompose the total agent-specific effect (t-ASE), we propose an axiomatic framework based on agent-specific effects for attributing the total effect to individual agents. The set of axioms includes *efficiency*, which requires that the agents' contributions sum to t-ASE (for the complete list see the Supplementary Material). Our attribution method, ASE-SV, operationalizes Shapley value Shapley [1953] with agent-specific effects, and uniquely satisfies the set of proposed axioms. More specifically, ASE-SV assigns to each agent $j \in \{1, ..., n\}$ a contribution score for t-ASE equal to

$$\sum_{S\subseteq\{1,...,n\}\setminus\{j\}} w_S \cdot \Big[\mathrm{ASE}^{S\cup\{j\}}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M - \mathrm{ASE}^{S}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M\Big],$$

where coefficients $w_S$ are set to $w_S = \frac{|S|!(n-|S|-1)!}{n!}$.

## 4 LEVEL 2: STATES

To further decompose the state-specific effect of the reverse transition (r-SSE), we utilize the notion of *intrinsic causal contributions* (ICC) Janzing et al. [2024], which enables us to quantify the informativeness of the individual state variables regarding the counterfactual outcomes related to the computation of r-SSE. Our proposed method attributes r-SSE to the subsequent state variables, **without** modifying the causal mechanisms of the underlying environment. Moreover, our method is *efficient* under a relatively mild assumption: at least one state variable has non-zero ICC (i.e., is informative about the counterfactual outcomes).

## 5 CONCLUSION

We introduce a causal explanation formula tailored to MMDPs, which decomposes the total counterfactual effect of an agent's action by attributing it to the agents and dynamics of the environment. In the extended version of this article, we experimentally validate the interpretability of our approach using two multi-agent environments with heterogeneous agents: a grid-world environment, where two actors trained with RL are instructed by an LLM planner to complete a sequence of tasks, and a two-agent sepsis management simulator, depicted in Fig. 1.

## Acknowledgements

This research was, in part, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 467367360.

## References

Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. In *International Joint Conference on Artificial Intelligence*, pages 357–363, 2005.

Craig Boutilier. Planning, learning and coordination in multiagent decision processes. In *Conference on Theoretical Aspects of Rationality and Knowledge*, pages 195–210, 1996.

Juan Correa, Sanghack Lee, and Elias Bareinboim. Nested counterfactual identification from arbitrary surrogate experiments. *Advances in Neural Information Processing Systems*, 34:6856–6867, 2021.

Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71, 2010.

Kosuke Imai, Luke Keele, Dustin Tingley, and Teppei Yamamoto. Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105 (4):765–789, 2011.

Dominik Janzing, Patrick Blöbaum, Lenon Minorics, Philipp Faller, and Atalanti Mastakouri. Quantifying intrinsic causal contributions via structure preserving interventions. In *International Conference on Artificial Intelligence and Statistics (to appear)*, 2024.

Lawrence A. Lynn. Artificial intelligence systems for complex decision-making in acute care medicine: A review. *Patient safety in Surgery*, 2019.

Judea Pearl. Direct and indirect effects. In *Conference on Uncertainty and Artificial Intelligence*, pages 411–420, 2001.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Lloyd S Shapley. A value for n-person games. *Annals of Mathematics Studies*, (28):307–317, 1953.

Stelios Triantafyllou, Aleksa Sukovic, Debmalya Mandal, and Goran Radanovic. Agent-specific effects: A causal effect propagation analysis in multi-agent MDPs. In *International Conference on Machine Learning (to appear)*, 2024.

Tyler J VanderWeele. Explanation in causal inference: developments in mediation and interaction. *International journal of epidemiology*, 45(6):1904–1908, 2016.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *AAAI Conference on Artificial Intelligence*, 2018a.

Junzhe Zhang and Elias Bareinboim. Non-parametric path analysis in structural causal models. In *Conference on Uncertainty in Artificial Intelligence*, 2018b.

# Supplementary Material

**Stelios Triantafyllou**[1] **Aleksa Sukovic**[1, 2] **Yasaman Zolfimoselo**[1] **Goran Radanovic**[1]

[1]Max Planck Institute for Software Systems
[2]Saarland University

## 6 FORMAL DEFINITIONS

In this section we provide the formal definitions of *total counterfactual effects*, *state-specific effects* and *agent-specific effects*, using standard terminology and notation from the SCM framework Pearl [2009]. Similar to Correa et al. [2021], when random variables have subscripts we will use square brackets to denote *do* interventions.

**Definition 6.1** (TCFE). Given an MMDP-SCM $M$ and a trajectory $\tau$ of $M$, the *total counterfactual effect* of intervention $do(A_{i,t} := a_{i,t})$ on $Y \in \mathbf{V}$, relative to reference $\tau(A_{i,t})$, is defined as

$$\begin{aligned}\mathrm{TCFE}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M &= \mathbb{E}[Y_{a_{i,t}}|\tau]_M - \mathbb{E}[Y_{\tau(A_{i,t})}|\tau]_M \\ &= \mathbb{E}[Y_{a_{i,t}}|\tau]_M - \tau(Y).\end{aligned}$$

**Definition 6.2** (SSE). Given an MMDP-SCM $M$ and a trajectory $\tau$ of $M$, the *state-specific effect* of intervention $do(A_{i,t} := a_{i,t})$ on $Y \in \mathbf{V}$, relative to reference $\tau(A_{i,t})$, is defined as

$$\begin{aligned}\mathrm{SSE}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M &= \mathbb{E}[Y_{a_{i,t}}|\tau; M]_{M^{do(I)}} - \mathbb{E}[Y_{\tau(A_{i,t})}|\tau]_M \\ &= \mathbb{E}[Y_{a_{i,t}}|\tau; M]_{M^{do(I)}} - \tau(Y),\end{aligned}$$

where $I = \{A_{i',t'} := A_{i',t'[\tau(A_{i',t'})]}\}_{i' \in \{1,\ldots,n\},t'>t}$.

Furthermore, the state-specific effect associated with the reverse transition can be defined as follows

$$\begin{aligned}\mathrm{SSE}_{\tau(A_{i,t}),a_{i,t}}(Y|\tau)_M &= \mathbb{E}[Y_{\tau(A_{i,t})}|\tau; M]_{M^{do(I)}} - \mathbb{E}[Y_{a_{i,t}}|\tau]_M \\ &= \mathbb{E}[Y|\tau; M]_{M^{do(I)}} - \mathbb{E}[Y_{a_{i,t}}|\tau]_M,\end{aligned} \tag{1}$$

where $I = \{A_{i',t'} := A_{i',t'[a_{i,t}]}\}_{i' \in \{1,\ldots,n\},t'>t}$.

**Definition 6.3** (ASE). Given an MMDP-SCM $M$, a non-empty subset of agents $\mathbf{N}$ in $M$ and a trajectory $\tau$ of $M$, the $\mathbf{N}$-*specific effect* of intervention $do(A_{i,t} := a_{i,t})$ on $Y \in \mathbf{V}$, relative to reference $\tau(A_{i,t})$, is defined as

$$\begin{aligned}\mathrm{ASE}^{\mathbf{N}}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M &= \mathbb{E}[Y|\tau; M]_{M^{do(I)}} - \mathbb{E}[Y_{\tau(A_{i,t})}|\tau]_M \\ &= \mathbb{E}[Y|\tau; M]_{M^{do(I)}} - \tau(Y),\end{aligned}$$

where $I = \{A_{i',t'} := \tau(A_{i',t'})\}_{i' \notin \mathbf{N},t'>t} \cup \{A_{i',t'} := A_{i',t'[a_{i,t}]}\}_{i' \in \mathbf{N},t'>t}\}$.

## 7 AXIOMS FROM SECTION 3

In this section, we formally state the axioms included in the framework described in Section 3 for attributing the total agent-specific effect to the individual agents.

**Efficiency:** *The total sum of agents' contribution scores is equal to the total agent-specific effect*. Formally,

$$\sum_{j \in \{1,\dots,n\}} \phi_j = \mathrm{ASE}^{\{1,\dots,n\}}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M.$$

**Invariance:** *Agents who do not marginally contribute to the total agent-specific effect are assigned a zero contribution score.* Formally, if for every $S \subseteq \{1, ..., n\} \backslash \{j\}$

$$\mathrm{ASE}^{S \cup \{j\}}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M - \mathrm{ASE}^{S}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M = 0,$$

then $\phi_j = 0$.

**Symmetry:** *Agents who contribute equally to the total agent-specific effect are assigned the same contribution score.* Formally, if for every $S \subseteq \{1, ..., n\} \backslash \{j, k\}$

$$\mathrm{ASE}^{S \cup \{j\}}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M - \mathrm{ASE}^{S}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M = \mathrm{ASE}^{S \cup \{k\}}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M - \mathrm{ASE}^{S}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M,$$

then $\phi_j = \phi_k$.

**Contribution monotonicity:** *The contribution score assigned to an agent depends only on its marginal contributions to the total agent-specific effect and monotonically so.* Formally, let $M_1$ and $M_2$ be two MMDP-SCMs with $n$ agents, if for every $S \subseteq \{1, ..., n\} \backslash \{j\}$

$$\mathrm{ASE}^{S \cup \{j\}}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_{M_1} - \mathrm{ASE}^{S}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_{M_1} \geq \mathrm{ASE}^{S \cup \{j\}}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_{M_2} - \mathrm{ASE}^{S}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_{M_2},$$

then $\phi_j^{M_1} \geq \phi_j^{M_2}$.

## 8 PROOF OF THEOREM 2.1

In this section, we restate and prove Theorem 2.1.

*The total counterfactual effect, total agent-specific effect and state-specific effect obey the following relationship*

$$\mathrm{TCFE}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M = \mathrm{ASE}^{\{1,\dots,n\}}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M - \mathrm{SSE}_{\tau(A_{i,t}),a_{i,t}}(Y|\tau)_M. \tag{2}$$

*Proof.* Eq. (2) follows directly from Definition 6.1, Definition 6.3 and Eq. (1):

$$\begin{aligned}
\mathrm{TCFE}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M &= \mathbb{E}[Y_{a_{i,t}}|\tau]_M - \tau(Y) \\
&= \mathbb{E}[Y_{a_{i,t}}|\tau]_M - \tau(Y) + \mathbb{E}[Y|\tau; M]_{M^{do(I)}} - \mathbb{E}[Y|\tau; M]_{M^{do(I)}} \\
&= \mathrm{ASE}^{\{1,\dots,n\}}_{a_{i,t},\tau(A_{i,t})}(Y|\tau)_M - \mathrm{SSE}_{\tau(A_{i,t}),a_{i,t}}(Y|\tau)_M,
\end{aligned}$$

where $I = \{A_{i',t'} := A_{i',t'[a_{i,t}]}\}_{i' \in \{1,\dots,n\}, t' > t}$.

□