Explaining Decisions of Agents in Mixed-Motive Games

Maayan Orner[1], **Oleg Maksimov**[1], **Akiva Kleinerman**[1], **Charles Ortiz**[2], **Sarit Kraus**[1]

[1]Department of Computer Science, Bar-Ilan University, Israel
[2]SRI International, USA
maayanorner@gmail.com

Abstract

In recent years, agents have become capable of communicating seamlessly via natural language and navigating in environments that involve cooperation and competition, a fact that can introduce social dilemmas. Due to the interleaving of cooperation and competition, understanding agents' decision-making in such environments is challenging, and humans can benefit from obtaining explanations. However, such environments and scenarios have rarely been explored in the context of explainable AI. While some explanation methods for cooperative environments can be applied in mixed-motive setups, they do not address inter-agent competition, cheap-talk, or implicit communication by actions. In this work, we design explanation methods to address these issues. Then, we proceed to establish generality and demonstrate the applicability of the methods to three games with vastly different properties. Lastly, we demonstrate the effectiveness and usefulness of the methods for humans in two mixed-motive games. The first is a challenging 7-player game called no-press Diplomacy. The second is a 3-player game inspired by the prisoner's dilemma, featuring communication in natural language.

Extended version — https://arxiv.org/abs/2407.15255

1 Introduction

Many important real-world scenarios resemble mixed-motive games. In these settings, agents' interests can be partly aligned or opposed, resulting in varying motivations to cooperate and compete. In recent years, the study of automated agents that act in mixed-motive games has gained renewed attention within the community, using a 7-player game called Diplomacy as a research testbed (Paquette et al. 2019; Peskov and Cheng 2020; Anthony et al. 2020; FAIR et al. 2022; Wongkamjan et al. 2024).

Meanwhile, the explainability of AI systems has been extensively studied (Simonyan, Vedaldi, and Zisserman 2013; Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017; Schleibaum et al. 2024). Moreover, the focus on explainability has also expanded to various types of multi-agent systems (Ciatto et al. 2019; Kraus et al. 2020; Boggess, Kraus, and Feng 2022; Qing et al. 2022; Guo et al. 2023). However, the unique aspects of explaining agents' decisions in mixed-motive games have not been adequately addressed.

In mixed-motive games and analogous real-world situations (Gnyawali and Park 2011), although agents may possess conflicting interests (Gallo Jr and McClintock 1965), there is a competitive advantage for some cooperation (Kraus, Ephrati, and Lehmann 1994). Furthermore, in mixed-motive games involving more than two agents, the decision of whom to cooperate with or compete against is crucial. In these scenarios, agents' tendencies, the alignment of their interests, the actions they take, and the way they communicate shape cooperation patterns, which in turn affect the payoffs they receive (Parkhe 1993; Sally 1995).

To tackle these unique challenges, we define a framework of *three conceptual levels* that explanations should address:

1. **Strategic**: The utility the agent obtains depends on its decisions, but also the decisions of other agents.
2. **Situational**: The state of the environment and the policies of the participating agents can motivate cooperation or competition among them.
3. **Diplomatic:** Agents' actions convey information to other agents, which can influence future outcomes.

Based on that, our (main) research question is as follows:

> ***Main Research Question:*** Can we design explanation methods that address the strategic, situational, and diplomatic levels in mixed-motive games?

The main contribution of this work is the development of methods for explaining agents' decisions in mixed-motive games, addressing the above research question.

We apply the methods in three environments with vastly different properties (see section 3 and appendix). To further demonstrate the applicability of the methods, agents in each environment employ a different type of policy: neural network in no-press Diplomacy, black-box (GPT-4) in Communicate Out of Prison (COP), and heuristic policy in Risk.

Furthermore, we show the usefulness of the methods for humans through two user studies: the first in no-press Diplomacy and the second in COP, a 3-player game inspired by the prisoner's dilemma, featuring communication in natural language.

To the best of our knowledge, this is the first work to propose explanation methods specifically designed for mixed-motive games.

2 Related Work

During the last few years, researchers have proposed a new research area called Explainable Decisions in Multi-Agent Environments (Kraus et al. 2020). Various aspects of the topic have begun to be explored (Boggess, Kraus, and Feng 2022; Qing et al. 2022; Guo et al. 2023; Zahedi, Sengupta, and Kambhampati 2023). However, most of that work has focused only on explaining agent interactions in cooperative settings (Nizri, Azaria, and Hazon 2022; Heuillet, Couthouis, and Díaz-Rodríguez 2022; Yang et al. 2022; Mahjoub et al. 2023; Angelotti and Díaz-Rodríguez 2023).

Other studies have included experimentation in mixed-motive games, but these efforts remain partial. For example, one study applied existing explanation methods (Khlifi et al. 2023) in a mixed-motive game that focuses on coordination (Level-Based Foraging), while the other applied novel methods designed for cooperative setups to the same game (Boggess, Kraus, and Feng 2023). A different study, which also focuses on coordination, does not evaluate the explanations relative to humans (Milani et al. 2022). None of these studies aims to address the unique challenges we discussed in section 1.

Although solution concepts from cooperative game theory, such as Shapley values (Heuillet, Couthouis, and Díaz-Rodríguez 2022) or Myerson values (Angelotti and Díaz-Rodríguez 2023) can provide a framework to examine the contribution of each agent to a cooperating team, they are insufficient for games with mixed motives. In such games, payoffs are typically not collective. Moreover, agents are not necessarily part of a team; they cooperate when beneficial but may become adversaries when motivated to compete.

Explanations of feature attribution (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017) often contain many irrelevant details while lacking relevant information. For example, in Diplomacy, such methods can be utilized to estimate the contribution of each unit movement (see section 3.1) to the utility value of a strategy (see section 4.1). Based on preliminary experimentation, such explanations are difficult to interpret. More importantly, they do not address questions related to agents' interactions. Therefore, new explanation methods are needed.

We drew a considerable amount of inspiration and designed the estimation algorithm for Strategy-based Utility Explanations (section 4.2, figure 1 upper heatmap) based on the Simulation Action Value Estimation (SAVE) algorithm from (Kramár et al. 2022). Shared Interests Correlation Analysis (section 4.4, figure 1 lower heatmap), although developed independently, is similar to a technique utilized by (Zhang et al. 2021) for a different task. Presenting the probable actions other agents might take (section 4.3, figure 1 arrows) is a special case of example-based explanations (Cai, Jongejan, and Holbrook 2019).

Large language model (LLM) agents are capable of acting in mixed-motive games (FAIR et al. 2022). To further examine the diplomatic level, we apply our explanation methods in setups with communication; it requires the development of LLM agents' game simulations, a topic that has received considerable attention (Yan et al. 2023; Akata et al. 2023; Xu et al. 2023; Mukobi et al. 2023; Gemp et al. 2024).

3 Environments

To examine different explanation methods, we explored alternative game settings. The first was no-press Diplomacy, a mixed-motive game with a large action space that introduces *situations* that encourage cooperation or competition between agents. In addition to tactical and *strategic* considerations, the game requires *diplomatic* understanding, making it suitable for studying our research question.

Since research has shown that cheap-talk can alter cooperation rates in games similar to the prisoner's dilemma (Sally 1995), we designed the Communicate Out of Prison (COP) game. COP is strategically simple to minimize confounding factors but features communication in unconstrained natural language. The LLM agents we developed for COP can adopt different personality types. This allows us to examine the explanations in settings where heterogeneous agent policies lead to different cooperation patterns. COP is defined as a 3-player game, to introduce rich social dynamics.

To broaden the scope of our evaluation, we also applied the methods in a simplified version of Risk (see appendix).

3.1 No-Press Diplomacy

Diplomacy is a simultaneous game in which each player controls one of the seven great powers of Europe in the years leading to World War 1. The goal of the game is to control at least 18 out of the 36 strategic locations on the map ("supply centers"). An action in Diplomacy is composed of multiple unit sub-actions (e.g., Naples fleet to Ionian Sea, Rome army to Apulia,...), yielding a large combinatorial action space. The game is estimated to have 10^{21} - 10^{64} joint actions per turn and a game tree size that can be infinitely large (median size $\approx 10^{896.8}$) (Anthony et al. 2020).

There are two notably popular versions of Diplomacy, one permits explicit communication (full-press) (Kraus and Lehmann 1995), while the other relies on implicit communication through in-game actions (no-press).

For our experiments, we use the game environment from (Paquette et al. 2019), along with a neural policy network and value function from (Anthony et al. 2020).

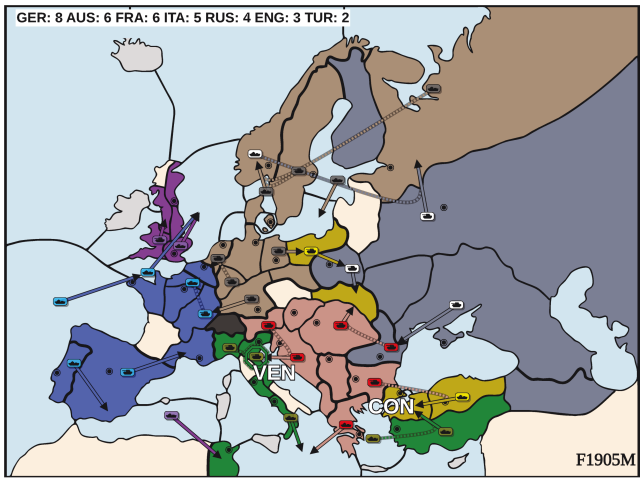
3.2 Communicate Out of Prison Game

In this game, which draws inspiration from the prisoner's dilemma, three agents (denoted as $\{a, b, c\}$) attempt to avoid punishment for a robbery.

The game starts with a communication stage, in which agents exchange private messages sequentially. After the communication stage is over, every agent announces whether each of the other agents is innocent or guilty; all agents announce simultaneously. For example, agent a can announce $(b = guilty, c = innocent)$. The payoffs, which are determined by the announcements, were designed to motivate both cooperation and competition.

For our experiments, we defined three types (private information) and prompted the LLM agents to play accordingly:

- **con-artist:** cruel, manipulative, and deceitful.
- **"simple-person":** nice, trusting, honest, and hates lies.
- **politician:** a political genius, selfish but rather honest, prefers "simple and nice" agents, dislikes manipulators.



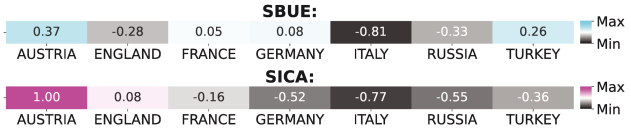


Figure 1: Explanation for Austria's strategy a^i, where it assists Turkey in preventing Italy from taking over Constantinople while attacking Venice. *SICA* detects animosity with Italy; *SBUE* explains that a^i implicitly communicates hostility to Italy and friendliness to Turkey. Austria's arrows visualize a^i; arrows of others present their *probable actions*.

4 Explanation Methods

In this section, we present the explanation methods. In the appendix, we provide additional pseudo-code and a method to find counterfactual actions in Diplomacy. Assuming we play as agent i, and would like to get an explanation for action a^i in state s, we propose three different explanation methods: **(1)** A utility-based method that presents the expected utility associated with a^i both for agent i and the other agents in the environment (section 4.2). **(2)** A probable actions-based method that specifies the most probable actions other agents can take when agent i performs a^i (section 4.3). **(3)** An estimation of the relationships between the agents in state s (section 4.4). A visual example of all three types of explanations in Diplomacy is provided in figure 1.

4.1 Preliminary Definitions

Environment and agents: An environment of p agents incorporates an action space A, state space S, transition function $T : S \times A^p \to S$, and reward function $R : S \times S \to \mathbb{R}^p$. Considering the current state as s_t (the subscript is occasionally omitted), we denote $R(s_{t-1}, s_t)$ as R_t, and the corresponding reward for agent i as R^i_t.

The set P consists of agents that act according to policies $(\pi_1, ..., \pi_p)$. A policy $\pi_i : S \times A \to [0, 1]$ is a joint probability mass function. For each agent, we define a value function $V_i : S \to \mathbb{R}$, which estimates the expected return of the agent given π; i.e. $V_i(s_t) \approx \mathbb{E}[R^i_{t+1} + ... + \gamma^{n-1} R^i_n | s_t]$, where $\gamma \in [0, 1]$ is a discount factor. For convenience, we define $V : S \to \mathbb{R}^p$ as a function that applies $(V_1, ..., V_p)$, and $\pi : S \times A^p \to [0, 1]^p$ as a function that returns the probability of agents' actions according to $(\pi_1, .., \pi_p)$. We denote $a^i \sim \pi_i$ as drawing an action from the distribution π_i given state s (s is implied), and $a \sim \pi$ as a vectorized version of it. We use $-i$ to denote all agents that are not i.

Utility of an outcome: Starting from state s_t, the utility vector of a specific outcome is defined as $\gamma V(s_{t+1}) + R_{t+1}$.

Expected utility of an action: The expected utility vector of action $a^i \in A$ is a vector of size p, in which the element with index j corresponds to the expected utility of action a^i for agent j. Formally, it is defined as $\mathbb{E}_{a^{-i} \sim \pi_{-i}}[\gamma V(T(s, (a^i, a^{-i}))) + R_{t+1}]$, where π_{-i} are the policies of agents $-i$ or the policies we assume they employ.

Algorithm 1: Simulate

```
1:  procedure SIMULATE(s ∈ S, k ∈ ℕ, d ∈ ℕ, C)
2:      // C is a set of tuples (a^i_e, d_e) where a^i_e ∈ A and d_e ∈ ℕ
3:      X ← k · d × p matrix; s_0 ← s
4:      for j ← 0 to k − 1 do // K simulations
5:          s ← s_0 // restore initial state
6:          r_cumulative ←< 0, ..., 0 > // vector of size p
7:          for t ← 0 to d − 1 do // to depth d
8:              a ← a^1, .., a^p ∼ π // draw actions for all agents
9:              for (a^i_e, d_e) ∈ C do
10:                 if t = d_e then a^i ← a^i_e // replace with a^i_e.
11:             s' ← T(s, a)
12:             r_cumulative ← r_cumulative + γ^t R(s, s')
13:             utility ← γ^{t+1} V(s') + r_cumulative
14:             X[j · d + t, :] ← utility; s ← s'
15:     return X
```

4.2 Strategy-Based Utility Explanations (SBUE)

Motivation and description: When engaging in mixed-motive games with more than two agents, it is advantageous to consider not only the action's benefit to the agent itself (*strategic*) but also its influence on other agents observing this action (*diplomatic*). This is crucial due to the role of implicit communication in these games. For example, a "friendly" action can communicate willingness for future cooperation, whereas a "hostile" action usually does not.

SBUE addresses both the strategic and diplomatic levels by explaining how an action influences the game outcomes for all agents, *presenting the expected utility value for each agent* conditioned on that action.

Explanation estimation: To explain action a^i_e (e denotes explained) given state s, the following steps are performed:

1. Simulate the next turn from s for k times, where agent i performs action a^i_e, and all other agents follow their respective policies.
2. Estimate the utility values of each outcome using the value functions and rewards (algorithm 1 line 13).
3. Estimate the expected utilities of a^i_e by computing the mean utility of each agent (column); return a vector of size p.

Steps 1 and 2 are equivalent to using algorithm 1:

$$Simulate(s, k, d = 1, C = \{(a^i_e, 0)\})$$

In cases where the value function is difficult to interpret, we estimate μ^i and σ^i for all $i \in P$ via unconstrained simulation (algorithm 1 with $C = \emptyset$), and perform Z-score standardization to each column before step 3.

4.3 Probable Actions-Based Explanations

Motivation and description: In any environment with more than two agents, understanding the policies of the other agents can be useful. Therefore, we present the most probable actions of agents $-i$, assuming agent i selects action a_e^i. This explanation was (primarily) designed to address the *strategic* level. It can be viewed as an example-based explanation, as it presents an example of a probable outcome assuming the agent plays a_e^i.

Explanation estimation: As in SBUE, we run k simulations from state s_t, where agent i performs action a_e^i and all other agents follow their respective policies. Then, we extract the most commonly used action of each agent accordingly. This explanation can be extended to multi-turn trajectories by greedily repeating the process where all agents, including i, follow their respective policies. In our Diplomacy user study, we present one turn since understanding longer trajectories is cognitively challenging (see figure 1, arrows).

Limitations for LLM agents: Finding the most commonly used actions of LLM agents is challenging because of the complexity of natural language. As a workaround, we decode each action (by decoding tokens) using temperature $\tau = 0$. While this solution is consistent with our greedy approach, it can lead to incorrect explanations (see section 5.3).

4.4 Shared Interests Correlation Analysis (SICA)

Motivation and description: In mixed-motive setups, a central question is whether the state of the environment and agents' policies can facilitate effective cooperation. To explain the cooperation tendencies and alignment of interests among pairs of agents in a given state, we introduce Shared Interests Correlation Analysis (SICA).

The SICA value of agents i, j is the Pearson correlation coefficient of their obtained utilities. When causal relationships can be assumed (i.e., the actions of i can influence the payoff of j), the SICA value can be interpreted as a measure of the "friendliness" or cooperativeness of i and j.

SICA is typically accompanied by an action-based explanation. This pairing reveals the impacts of actions on friends and enemies, presenting a fuller diplomatic picture. For example, it highlights how hostile actions toward friends differ from those toward enemies, addressing the *situational* and *diplomatic* levels in our framework.

Explanation estimation: To estimate SICA, we perform k unconstrained simulations to depth d according to algorithm 1, where all agents act according to their respective policies, i.e., $a \sim \pi$. This step is equivalent to calling:

$$X \leftarrow Simulate(s, k, d, C = \emptyset)$$

The resulting dataset X, considering the case of depth-1 ($d = 1$) without loss of generality, is a $k \times p$ matrix, in which element $X_{x,y}$ is the utility agent y obtains in simulation x (algorithm 1 line 13). Lastly, we compute the sample Pearson correlation coefficient for each pair of columns (i.e., agents), which results in a $p \times p$ correlation matrix.

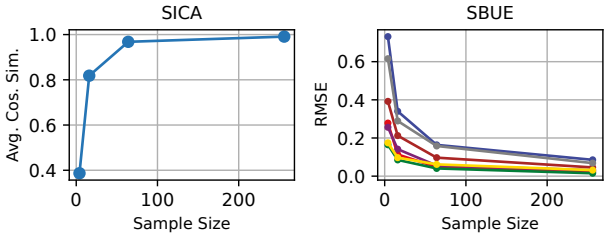


Figure 2: Convergence of SICA and SBUE in Diplomacy. On the right side, each line corresponds to a different agent.

Interpretation of SICA: The correlation coefficients are determined by the environment and policies of the agents, which we refer to as *conditions*. A *strong positive correlation coefficient* between agents i and j indicates that the conditions lead to mutual benefit and cooperation. A *strong negative correlation coefficient* suggests that the conditions lead to conflicting interests, resulting in competition or antagonism. A *weak correlation coefficient* (near zero) may result from conditions that lead to inconsistent impacts of one agent's actions on another, reflecting interactions that range from indifference to complex, nonlinear patterns.

Interdependence theory: The definition of SICA is analogous to the concept of covariation of interests (Van Lange and Balliet 2015), drawing a connection to interdependence theory (Kelley et al. 1959; Kelley 1978).

5 Modules Evaluation

This section presents the results of our evaluation experiments in Diplomacy and Risk. The experiments conducted in the COP game are summarized here and described in detail in the appendix.

5.1 Convergence of Estimation of SICA and SBUE in Diplomacy

Since the run-time of both methods increases linearly with the number of samples and sampling is expensive, we examine the error-runtime trade-off for the methods using several sample sizes. For both methods, we perform 50 independent repeated estimations for each sample size k.

For SBUE, we estimate once with $k = 2,500$ and define it as ground truth. Then, we repeatedly estimate for each k. The RMSE (root mean square error) for each agent's value is computed independently relative to its corresponding ground truth. Small RMSE values indicate that k is sufficient to provide reliable estimates. For SICA, for each k, we repeatedly estimate and compute the average cosine similarity among all pairs of (flattened) correlation matrices. When the average cosine similarity is high, it implies convergence.

In Diplomacy, although the action space is large, we observe that SICA and SBUE require reasonable sample sizes in the game states we examined, as shown in figure 2 (plot of typical results in a middle-game state). However, the convergence rates are highly dependent on the specific setup.

5.2 Evaluation of SICA in Diplomacy and Risk

We hypothesize that *SICA's estimation of the relationships between agents is well-aligned with humans' opinions*, as long as the agents play similarly to humans.

Setup: To test our hypothesis, we conducted a human-based experiment, in which we asked human players to annotate the most friendly and hostile agents in multiple board states, assuming they play one of the roles (e.g., Austria).

Specifically, we randomly generated 30 Diplomacy game states (ensuring representation from different game stages) and asked human players to annotate the top two hostile and friendly agents for each board. If an annotator was unable to decide the identity of the second most friendly or hostile agents, we allowed the determination to be left unfilled. For Risk (4-player version), we generated 12 board states, and a human selected the top enemy and top friend for each board.

We used two different annotators for diplomacy: one was considered a strong player, and the other was an intermediate player (introducing a variation of skill level). For Risk, which is not our main focus, the dataset was annotated by the authors.

Metric and evaluation: To evaluate the alignment of the annotations with SICA, we used MAP@K (Manning 2009). We ranked the other agents by the SICA value they shared with the agent, high to low, and reversed the ranking to rank agents by hostility. Then, we evaluated SICA using the annotated datasets. This methodology was built upon the assumption that agents play similarly to humans, and a violation of it is likely to result in a decrease in the MAP@K values.

Inter annotator agreement (IAA): To compare SICA to the agreement between annotators, we computed lower and upper bounds for MAP@K for the annotations of A compared to the partial ranking of annotator B (called IAA-rank in table 1). Note, $MAP@K(A, B_{rank}) \neq MAP@K(B, A_{rank})$, but we do not expect symmetry in this setup.

Results: The results (see table 1) suggest that SICA is well-aligned with human intuition and performs better than random rankings in both Diplomacy and Risk. In Diplomacy, SICA outperforms a heuristic-based ranking, which sorts agents by the number of centers they own (a proxy for strength), by a considerable margin. This is achieved without relying on non-generalizable properties of the environment. To clarify, the results for the random rankings of friends (0.58) and enemies (0.55) differ because the expert, as allowed, selected different numbers of friends and enemies.

The range of IAA-rank was higher than SICA in 2 out of 4 cases, lower in one case, and overlapped in another. In other words, in this setup, the degree of alignment between SICA and humans is comparable to the degree of agreement among humans themselves.

Category	Rand.	Cen.	SICA	IAA-rank	N
(D, e) E	0.55	0.63	**0.74**	0.86 ± 0.01	30
(D, e) F	0.58	0.60	**0.66**	0.725 ± 0.015	30
(D, p) E	0.56	0.74	**0.81**	0.82 ± 0.01	30
(D, p) F	0.56	0.58	**0.78**	0.72 ± 0.02	30
(R) E	0.61	-	**0.81**	-	12
(R) F	0.61	-	**0.72**	-	12

Table 1: MAP@K score, Diplomacy (D, e) - expert annotation, (D,p) - non-expert annotation, Risk (R); E - enemies, F - friends; Rand. - Random, Cen. - Centers.

5.3 LLM Evaluation Experiments (COP)

Types consistency: First, to validate that LLM agents mimic the types we defined consistently, we asked two humans to detect the type of each agent in anonymized games (based on communication only; payoffs were hidden). We observed that humans can distinguish between the types (accuracy: annotator 1 - 86.66%; annotator 2 - 93.33%).

Preliminary experiments: We conducted experiments to examine the methods in a setup with LLM agents.

(1) We defined the agents' personalities to encourage cooperation between the politician and the "simple-person" against the con-artist (see section 3.2). Therefore, we hypothesized that SICA would explain it.

First, we generated the explanation in a standard setup with a politician, a "simple-person", and a con-artist. The explanation was aligned with our hypothesis. To compare that explanation with another explanation and to ensure that SICA would reflect changes in the agents' dynamics, we repeated the process in a different setup involving two politicians and one "simple-person". In this case, as we expected, the explanation showed balanced cooperation and competition patterns among all the agents.

(2) To examine SBUE, we manually curated two messages from the politician to the "simple-person": a friendly message (m_f) and a hostile message (m_h). The SBUE explanations indicated that m_f was more beneficial to both agents, while m_h provided an advantage to the con-artist, which was consistent with our hypothesis.

(3) For the probable actions-based explanations, we examined how the temperature parameter affected the game outcomes. We found that using a temperature $\tau = 0$, which corresponds to greedy decoding (our approach), sometimes led to outcomes that were not probable when using $\tau = 0.7$. This highlights a significant limitation: the explanations produced using our (simplistic) approach can be misleading.

COP large language model study: Lastly, we studied how SICA and SBUE influence LLM agents' decisions. This study mirrors (to a large extent) the study described in section 6.2, but here the "participants" are LLM agents instead of humans. Our findings suggest that agents follow the explanations in most cases. We demonstrated that the explanations can help mitigate a well-known bias in LLMs, where decision-making is affected by the order of presented options (Wang et al. 2023). This highlights the potential of the explanations to enhance LLM agents' decision-making in the setups we study.

6 User Studies

We conducted two complementary studies with humans in two different environments. Our *Diplomacy user study* in-

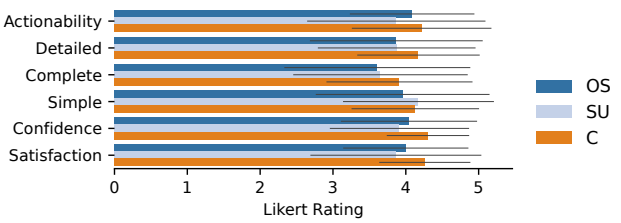


Figure 3: Mean and SD of participant ratings of explanations. Participants almost consistently prefer explanation *C*.

cludes a comparison of multiple explanation types and a general examination of the effectiveness of each explanation type with respect to our research question.

The *COP user study* was designed based on the results of the Diplomacy study. Here, we further examine the effectiveness of SICA and SBUE in a setup with explicit communication and heterogeneous policies — focusing on specific aspects of the situational and diplomatic levels.

Questionnaire design: In both experiments, users answered questionnaires where each response was on a Likert scale (1-5). They were asked the same set of questions before and after being presented with an explanation, and we examined how the explanation modified their answers. We tested our hypotheses using the Wilcoxon signed-rank test with a α value defined as 0.05. In the Diplomacy user study, we asked additional questions after presenting each explanation to assess how it was perceived (see section 6.1).

6.1 Diplomacy User Study

We recruited 26 subjects (24 males, 1 female, 1 other), with the requirement that players were familiar with the rules of Diplomacy. Participants who provided inconsistent answers to our consistency check questions or failed a basic knowledge exam were discarded automatically (3 out of 26).

We presented users with 3 types of explanations:

- *OS* (others' strategies): presents the probable actions of other agents (see section 4.3; figure 1 - arrows).
- *SU* (shared interests and utilities): a combination of SICA and SBUE (sections 4.4, 4.2; figure 1 - heatmaps).
- *C* (combined): combines explanation types *OS* and *SU*.

Experiment setup: The experiment covers three different Diplomacy boards (states); in each of them, the user is presented with a strategy (unit movements) with higher expected utility together with a counterfactual. Each board is paired with a different explanation type, where each type is presented only once. We followed a round-robin scheme to determine *(state, explanation)* allocations and temporal explanation type order, to preserve balance and prevent biases.

Hypotheses: We make two hypotheses that cover the conceptual levels we defined:

- H1: The explanation method improves users' performance at the strategic level.
- H2: The explanation method enhances users' understanding of the situational and diplomatic levels.

H	Exp.	Lower	Higher	p-val.	N-subj.
H1 (+)	OS	0.17	0.57	0.005	23
	SU	0.09	**0.65**	0.001	23
	C	0.00	**0.65**	0.001	23
H1 (-)	OS	0.52	0.13	0.012	23
	SU	0.43	0.00	0.004	23
	C	**0.57**	0.13	0.007	23
H2 (+)	OS	0.05	0.23	0.102*	22
	SU	0.05	0.36	0.021	22
	C	0.05	**0.41**	0.012	22

Table 2: Diplomacy - the proportion of instances where users raised/lowered their answers after seeing the explanation. In H(+) we hypothesized an increase and in H(-) a decrease.

To test H1: Users were asked to assess their willingness to choose each strategy before and after seeing the explanation.

To test H2: Users were asked to rank their understanding of the effects of their preferred strategy on "friends and enemies" before and after seeing the explanation. This requires understanding which agents are considered either friends or enemies (*situational*), and how the strategy influences agents within these identified groups (*diplomatic*).

Results: As seen in table 2, the effect of the explanations is statistically significant in all cases, except for H2 explanation *OS*. For the *strategic* level, H1($\pm$), the ratio of cases where users follow the explanations and raise (lower) their willingness to select the superior (inferior) strategy is at least as high as 43% for all types (*SU*, *OS*, and *C*).

Notably, the proportion of subjects who rated their understanding of agents' alignment of interests, and how a strategy influences other agents (H2), is nearly twice as high when SICA and SBUE are provided with or without example-based explanations, in comparison to example-based explanations alone (*C*: 41%, *SU*: 36% vs *OS*: 23%).

In all cases, combined explanations *C* are at least as effective as explanations *SU* or *OS* individually, as shown by the bolded values in the table.

Comparing explanations: In addition to the main part of the experiment, we defined a set of favorable properties based on previous work (Hoffman et al. 2018; Boggess, Kraus, and Feng 2022, 2023). For each explanation, we asked participants to rate these properties on a Likert scale.

Given the sample size, the statistical power was insufficient to detect significant differences per property. However, aligned with our intuition, *explanation C (combined) is rated higher in every category except for simplicity* (6 out of 7, see figure 3). That result further demonstrates the complementary nature of the methods and is consistent with the main results of the experiment (table 2).

6.2 Communicate Out of Prison User Study

We recruited 38 subjects (31 males, 7 females; all are graduate-level computer science students) who did not participate in the Diplomacy user study. Subjects who stated they did not understand the explanations were discarded automatically (5 out of 38).

The focus of this study is to examine whether SICA and SBUE[1] are useful for humans in environments where agents communicate via natural language. To evaluate this, we selected Communicate Out of Prison (COP) as a testbed. Recall that in COP, three agents communicate with each other privately, trying to stay out of prison by convincing others to side with them.

Experiment setup: We presented users with one to three dialog states in a randomized order. The number of states varied because users were allowed to finish the experiment early.

To generate the states, we simulated the game until it included some chat history of the playing agent with both other agents, to enable users to form opinions about the other agents' personalities and tendencies. The agents played according to the personality types we defined in section 3.2.

In each state, we presented users with an action (message to one of the other agents) with a higher expected utility and a counterfactual. For example, a possible action is *"A to B: Listen here, mate. C's been tryin' to convince me it's you ... But I ain't buyin' it. You've been straight with me, ... We stick together ... Let's point at C."*

Hypotheses: We make three different hypotheses, one is related to the effectiveness of the explanation methods at the *strategic* level, and the others refer to aspects of the *diplomatic* and *situational* levels:

- H3: The explanation method improves users' performance at the strategic level in mixed-motive games with explicit communication via natural language.
- H4: The explanation method improves the understanding of inter-agent relationships (with respect to the chat history and the policies of the agents) in mixed-motive games with explicit communication via natural language.
- H5: The explanation method improves the understanding of the diplomatic influence of explicit communication actions in mixed-motive games.

To test H3: Users were asked which of two messages should be sent, before and after observing the explanation.

To test H4: Users were asked which agent out of the other two was more friendly and trustworthy, before and after being presented with the explanation.

To test H5: Users were asked to rate their understanding of the influence of the messages on the other agents, before and after being presented with the explanation.

Results: As seen in table 3, the explanations modified the opinions of humans significantly, both about the question of which message should be sent (73%), as well as their understanding of inter-agent relationships (59%). The explanations also subjectively improved users' understanding of the influence of each message on the other agents (30%).

H	Lower	Higher	p-val.	N-subj.
H3 (+)	0.05	0.73	3×10^{-9}	33
H4 (+)	0.09	0.59	5×10^{-7}	33
H5 (+)	0.076	0.30	0.001	33

Table 3: COP - the proportion of instances where users raised/lowered their answers after seeing the explanation. H(+) indicates we hypothesized the answer would be raised.

7 Discussion, Limitations, and Future Work

Discussion: In this work, we presented methods to explain the decisions of agents that act in mixed-motive environments, focusing on games with more than two agents. First, we briefly discussed some of the challenges of mixed-motive games and described what explanation methods should address, using a three-level framework. Based on that, we designed explanation methods, including a method to explain relationships between agents in mixed-motive games. We applied these methods in three environments, one of which involved LLM agents with distinct personality types.

We conducted two user studies to evaluate the usefulness of the methods for humans and found the explanations helpful. We observed that SICA and SBUE are effective for the tasks they were designed for but likely enhanced when combined with example-based explanations. We conclude that the proposed methods are complementary to some extent.

Limitations: The main limitation of this work has to do with the fact that the subject is understudied, making a comparison with appropriate well-established baselines difficult. Since our methods are simple to apply, we believe they can be adopted as baselines in mixed-motive games.

In both user studies, participants were sampled from groups of potential users, resulting in gender imbalances that reflect the groups' demographics. For example, 85 out of 87 participants at the 2014 World DipCon were males (Hill 2014). Similarly, the COP study mirrors the male-to-female ratio in many subfields of computer science (Yamamoto and Frachtenberg 2022).

Future work: The proposed solutions are designed to be simple and general, inviting further extensions and improvements. For example, in SICA, any association function can be substituted for instead of Pearson's r.

Future work can also examine our (or novel) explanation methods for agents that act in mixed-motive games involving natural language and strategic elements, such as full-press Diplomacy. Additionally, developing LLM agents that act in simulated environments is an expanding research area (see related work), and investigating the explanation methods in complex LLM-based environments can be valuable.

Acknowledgments

This research has been partially supported by the Israel Ministry of Innovation, Science & Technology grant 1001818511.

References

Akata, E.; Schulz, L.; Coda-Forno, J.; Oh, S. J.; Bethge, M.; and Schulz, E. 2023. Playing repeated games with Large Language Models. *arXiv preprint arXiv:2305.16867*.

[1]The limitation we discuss in section 4.3, and the results of our Diplomacy user study motivated us to exclude the probable actions-based explanations from this study.

Angelotti, G.; and Díaz-Rodríguez, N. 2023. Towards a more efficient computation of individual attribute and policy contribution for post-hoc explanation of cooperative multi-agent systems using Myerson values. *Knowledge-Based Systems*, 260: 110189.

Anthony, T.; Eccles, T.; Tacchetti, A.; Kramár, J.; Gemp, I.; Hudson, T.; Porcel, N.; Lanctot, M.; Pérolat, J.; Everett, R.; et al. 2020. Learning to play no-press Diplomacy with best response policy iteration. *Advances in Neural Information Processing Systems*, 33: 17987–18003.

Boggess, K.; Kraus, S.; and Feng, L. 2022. Toward Policy Explanations for Multi-Agent Reinforcement Learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Boggess, K.; Kraus, S.; and Feng, L. 2023. Explainable Multi-Agent Reinforcement Learning for Temporal Queries. *arXiv preprint arXiv:2305.10378*.

Cai, C. J.; Jongejan, J.; and Holbrook, J. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*, 258–262.

Ciatto, G.; Calegari, R.; Omicini, A.; et al. 2019. Towards XMAS: explainability through multi-agent systems. In *CEUR WORKSHOP PROCEEDINGS*, volume 2502, 40–53. Sun SITE Central Europe, RWTH Aachen University.

FAIR; Bakhtin, A.; Brown, N.; Dinan, E.; Farina, G.; Flaherty, C.; Fried, D.; Goff, A.; Gray, J.; Hu, H.; Jacob, A. P.; Komeili, M.; Konath, K.; Kwon, M.; Lerer, A.; Lewis, M.; Miller, A. H.; Mitts, S.; Renduchintala, A.; Roller, S.; Rowe, D.; Shi, W.; Spisak, J.; Wei, A.; Wu, D. J.; Zhang, H.; and Zijlstra, M. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378: 1067 – 1074. Meta Fundamental AI Research Diplomacy Team.

Gallo Jr, P. S.; and McClintock, C. G. 1965. Cooperative and competitive behavior in mixed-motive games. *Journal of Conflict Resolution*, 9(1): 68–78.

Gemp, I.; Bachrach, Y.; Lanctot, M.; Patel, R.; Dasagi, V.; Marris, L.; Piliouras, G.; and Tuyls, K. 2024. States as strings as strategies: Steering language models with game-theoretic solvers. *arXiv preprint arXiv:2402.01704*.

Gnyawali, D. R.; and Park, B.-J. R. 2011. Co-opetition between giants: Collaboration with competitors for technological innovation. *Research Policy*, 40(5): 650–663.

Guo, Y.; Campbell, J.; Stepputtis, S.; Li, R.; Hughes, D.; Fang, F.; and Sycara, K. 2023. Explainable action advising for multi-agent reinforcement learning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 5515–5521. IEEE.

Heuillet, A.; Couthouis, F.; and Díaz-Rodríguez, N. 2022. Collective explainable AI: Explaining cooperative strategies and agent contribution in multiagent reinforcement learning with shapley values. *IEEE Computational Intelligence Magazine*, 17(1): 59–71.

Hill, D. 2014. The board game of the alpha nerds.

Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Kelley, H. H. 1978. Interpersonal relations: A theory of interdependence. *(No Title)*.

Kelley, H. H.; et al. 1959. *The social psychology of groups*. Transaction Publishers.

Khlifi, W.; Singh, S.; Mahjoub, O.; de Kock, R.; Vall, A.; Gorsane, R.; and Pretorius, A. 2023. On Diagnostics for Understanding Agent Training Behaviour in Cooperative MARL. *arXiv preprint arXiv:2312.08468*.

Kramár, J.; Eccles, T.; Gemp, I.; Tacchetti, A.; McKee, K. R.; Malinowski, M.; Graepel, T.; and Bachrach, Y. 2022. Negotiation and honesty in artificial intelligence methods for the board game of Diplomacy. *Nature Communications*, 13(1): 7214.

Kraus, S.; Azaria, A.; Fiosina, J.; Greve, M.; Hazon, N.; Kolbe, L.; Lembcke, T.-B.; Muller, J. P.; Schleibaum, S.; and Vollrath, M. 2020. AI for explaining decisions in multi-agent environments. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 13534–13538.

Kraus, S.; Ephrati, E.; and Lehmann, D. 1994. Negotiation in a non-cooperative environment. *Journal of Experimental & Theoretical Artificial Intelligence*, 3(4): 255–281.

Kraus, S.; and Lehmann, D. 1995. Designing and building a negotiating automated agent. *Computational Intelligence*, 11(1): 132–171.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Mahjoub, O.; de Kock, R.; Singh, S.; Khlifi, W.; Vall, A.; Tessera, K.-a.; and Pretorius, A. 2023. Efficiently Quantifying Individual Agent Importance in Cooperative MARL. *arXiv preprint arXiv:2312.08466*.

Manning, C. D. 2009. *An introduction to information retrieval*. Cambridge university press.

Milani, S.; Zhang, Z.; Topin, N.; Shi, Z. R.; Kamhoua, C.; Papalexakis, E. E.; and Fang, F. 2022. MAVIPER: Learning Decision Tree Policies for Interpretable Multi-agent Reinforcement Learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 251–266. Springer.

Mukobi, G.; Erlebach, H.; Lauffer, N.; Hammond, L.; Chan, A.; and Clifton, J. 2023. Welfare diplomacy: Benchmarking language model cooperation. *arXiv preprint arXiv:2310.08901*.

Nizri, M.; Azaria, A.; and Hazon, N. 2022. Explaining Fair Allocations and Recommendations.

Paquette, P.; Lu, Y.; Bocco, S. S.; Smith, M.; O-G, S.; Kummerfeld, J. K.; Pineau, J.; Singh, S.; and Courville, A. C. 2019. No-press Diplomacy: Modeling multi-agent gameplay. *Advances in Neural Information Processing Systems*, 32.

Parkhe, A. 1993. Strategic alliance structuring: A game theoretic and transaction cost examination of interfirm cooperation. *Academy of management journal*, 36(4): 794–829.

Peskov, D.; and Cheng, B. 2020. It takes two to lie: One to lie, and one to listen. In *Proceedings of ACL*.

Qing, Y.; Liu, S.; Song, J.; Wang, H.; and Song, M. 2022. A survey on explainable reinforcement learning: Concepts, algorithms, challenges. *arXiv preprint arXiv:2211.06665*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Sally, D. 1995. Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and society*, 7(1): 58–92.

Schleibaum, S.; Feng, L.; Kraus, S.; and Müller, J. P. 2024. ADESSE: Advice Explanations in Complex Repeated Decision-Making Environments. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, 7904–7912. ijcai.org.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *CoRR*, abs/1312.6034.

Van Lange, P. A.; and Balliet, D. 2015. Interdependence theory.

Wang, P.; Li, L.; Chen, L.; Zhu, D.; Lin, B.; Cao, Y.; Liu, Q.; Liu, T.; and Sui, Z. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Wongkamjan, W.; Gu, F.; Wang, Y.; Hermjakob, U.; May, J.; Stewart, B. M.; Kummerfeld, J. K.; Peskoff, D.; and Boyd-Graber, J. L. 2024. More Victories, Less Cooperation: Assessing Cicero's Diplomacy Play. *arXiv preprint arXiv:2406.04643*.

Xu, Y.; Wang, S.; Li, P.; Luo, F.; Wang, X.; Liu, W.; and Liu, Y. 2023. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.

Yamamoto, J.; and Frachtenberg, E. 2022. Gender differences in collaboration patterns in computer science. *Publications*, 10(1): 10.

Yan, M.; Li, R.; Zhang, H.; Wang, H.; Yang, Z.; and Yan, J. 2023. LARP: Language-Agent Role Play for Open-World Games. *arXiv preprint arXiv:2312.17653*.

Yang, M.; Moon, J.; Yang, S.; Oh, H.; Lee, S.; Kim, Y.; and Jeong, J. 2022. Design and implementation of an explainable bidirectional lstm model based on transition system approach for cooperative ai-workers. *Applied Sciences*, 12(13): 6390.

Zahedi, Z.; Sengupta, S.; and Kambhampati, S. 2023. 'Why didn't you allocate this task to them?'Negotiation-Aware Explicable Task Allocation and Contrastive Explanation Generation. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 2292–2294.

Zhang, Y.; Yang, Q.; An, D.; and Zhang, C. 2021. Coordination between individual agents in multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 11387–11394.