Fostering Collective Action in Complex Societies using Community-Based Agents

Jonathan Skaggs[*] , **Michael Richards**[*] , **Melissa Morris** ,
Michael A. Goodrich and **Jacob W. Crandall**

Computer Science Department, Brigham Young University, Provo, UT, USA
{jbskaggs12, michael.richards256, mel4college}@gmail.com, {mike, crandall}@cs.byu.edu

Abstract

As AI integrates into human societies, its ability to engage in collective action is increasingly important. Human social systems have large and flexible strategy spaces, conflicting interests, power asymmetry, and interdependence among members, which together make it challenging for agents to learn collective action. In this paper, we explore the ability of community-based agents to learn collective action within a novel model of complex social systems. We first present this social model, called the Junior High Game (JHG). The JHG embodies key elements of human social systems that require players to act collectively. We then describe an agent, called CAB, which is based on community detection and formation algorithms. Via simulations and user studies, we evaluate the ability of CAB agents to interact in JHG societies consisting of humans and AI agents. These evaluations both identify requirements for successful collective behaviors in the JHG and identify important unsolved problems for developing AI agents capable of collective action in complex social systems.

1 Introduction

As AI integrates into human society, its ability to engage in collective action (i.e., the ability to coordinate and work toward a common goal) with humans and other AI is important. Despite its importance, developing algorithms that produce collective action is challenging. State-of-the-art AI algorithms often fail to learn to coordinate behavior, even in well-defined teaming scenarios (e.g., [Fosong *et al.*, 2023]). As such, a systematic study of collective action in human societies is needed to better understand how to create AI that engages in effective collective action.

Achieving collective action in human societies is challenging for several reasons. First, individual members of human societies do not always share the same goals and preferences. Despite these differences, individuals must find ways to work together to be successful. Second, human societies have an intricate inter-agent network structure. This structure includes power dynamics [Freeman, 1977; Bonacich and Lloyd, 2001; Katz, 1953; Salancik, 1978; Barabási and Albert, 1999] and laws of connectivity, including structural balance [Heider, 1946; Granovetter, 1973], assortativity [Newman, 2002; McPherson *et al.*, 2001], and cascading effects [Centola, 2018]. This network structure creates feedback effects as individuals reason about how they can work together [Easley and Kleinberg, 2010]. Finally, human societies are constantly evolving and tend not to perpetually converge to a stable state [Acemoglu and Robinson, 2013], meaning that AI agents must continually adapt to other members of society. Together, these complexities make it challenging for AI agents to learn effective collective action.

The complexities of human societies create non-stationary environments with vast state and action spaces. As a result, AI agents must often act in scenarios they have never seen before. To deal with these uncertainties, we consider using principles from the network science literature to guide agent behavior. In particular, we study how AI agents can develop mechanisms for collective action by forming dyads [Newman *et al.*, 2002], triads [Holland and Leinhardt, 1971; Watts and Strogatz, 1998], and groups using community detection and formation algorithms [Newman, 2010; Blondel *et al.*, 2008; Newman, 2006; Brandes *et al.*, 2007].

This paper makes two contributions. First, we present the Junior High Game (JHG) to model challenges of collective action in complex societies. Second, we present and evaluate a community-based agent, called CAB (*C*ommunity-*A*ware *B*ehavior), which uses community detection and formation algorithms to guide its behavior. Results provide insights into creating agents that foster effective collective action.

2 A Model of Social Systems

To effectively study collective action, test-beds that adequately model relevant attributes of human societies are needed. In this section, we first consider properties a test-bed should have for studying collective action and consider existing test-beds with respect to these properties. Second, we describe a new test-bed designed to satisfy these properties, and provide an example to illustrate the test bed's dynamics.

2.1 Test-bed Requirements for Collective Action

Test-beds for studying collective action should abstract attributes of human societies. Toward this end, Table 1 pro-

[*] Equal contribution

Table 1: Desirable properties of a test-bed for studying collective actions in complex social systems.

Property	**Description and details of the property**
Large & flexible strategy space	The test-bed's state space should be large enough that agents often encounter states they have never experienced before. Additionally, the test-bed's action space should provide agents the ability to interact (both positively and negatively) with other individual agents and subgroups.
Conflicting interests	Incentives are general sum, such that the players do not share all of the same goals and preferences, but their goals and preferences are not necessarily fully competitive either.
High interdependence	Agents are reliant on each other, such that an agent must coordinate its behavior with some (but not necessarily all) agents in society to achieve its goals. Agents are interdependent in both their need to help each other positively (e.g., trade) and to eliminate common threats (e.g., war).
Power asymmetry	Players have asymmetric abilities to impact the society. These abilities arise, at least in part, from the prior actions taken by individuals in the society, and thus can fluctuate over time.
Non-convergence	The game's incentives and dynamics encourage the structure of society to continually evolve. In general, the game's state does not perpetually converge to an equilibrium, and no single strategy dominates other strategies independent of other agents' behaviors.
Scalability	The test-bed scales to societies of tens to hundreds of individuals.

poses a set of properties that exist in typical human societies and that make collective action interesting and challenging. Test-beds used to evaluate multi-agent decision-making do not typically simultaneously satisfy all of these properties.

A *large state space* in *non-convergent* and *scalable* environments helps produce scenarios that encourage general-purpose reasoning, rather than brute-force approaches that allow agents to memorize solutions to specific scenarios. Furthermore, a *flexible* strategy space gives agents sufficient freedom to interact, which provides greater richness to the society. *Power asymmetry* arising from game play provides further incentives, dynamics, and feedback effects that can greatly alter how agents collaborate with each other. Finally, *conflicting interests* and *interdependence* together make it challenging, but necessary, for agents to cooperate and coordinate with each other. The additional stipulation of having multiple forms of interdependence makes the problem of collective action both richer and more general.

A variety of test-beds have been useful to developing AI, though they do not adequately model the properties listed in Table 1. For example, Chess [Liaqat *et al.*, 2020; Campbell *et al.*, 2002; Silver *et al.*, 2018] and Poker [Brown and Sandholm, 2019; Bowling *et al.*, 2017; Rubin and Watson, 2011] provide interesting challenges, but do not model conflicting interests and interdependence. Team games, such as Soccer (e.g., [Fosong *et al.*, 2023]), Overcooked [Bishop *et al.*, 2020; Baek *et al.*, 2022; Rosero *et al.*, 2021], and Hanabi [Walton-Rivers *et al.*, 2019; Bard *et al.*, 2020] provide interesting coordination and challenging problems, but do not model conflicting interests within non-convergent and scalable societies. Additionally, trading-agent competitions (e.g., [Wellman *et al.*, 2003]) model complex multi-agent systems, but typically do not require an agent to coordinate with other agents within an intricate society (i.e., high interdependence), nor do they model power asymmetry. Colored Trails [De Jong *et al.*, 2011] also offers a compelling trading scenario, but does not require multiple forms of interdependence, nor does it appear to be easily scaled to large societies.

Social dilemmas have been used to test the ability of AI agents to coordinate and cooperate within societies that model conflicting interests. These dilemmas include two-player repeated games (e.g., Stag Hunt [Skyrms, 2003], Prisoner's Dilemmas [Axelrod, 1984], and the Coin Game [Lerer and Peysakhovich, 2017]), which model fascinating dynamics related to collective action, but do not directly encode many-player (non-convergent) societies with large and flexible strategy spaces. Public-goods games (e.g., [Fehr and Gächter, 2002]) place agents within potentially large societies, but do not have action spaces that allow agents to interact with other individuals – players only interact with the group at large. More sophisticated versions of these games, such as prisoner's dilemmas on networks [Shi *et al.*, 2022; Biely *et al.*, 2007], give players greater flexibility to interact with subsets of agents, but still do not give players the ability to interact positively and negatively with other individual agents, nor do they seem to adequately model power asymmetry, non-convergence, and multiple forms of interdependence.

Other models potentially consider broader property sets. For example, agent-based models sometimes consider highly interconnected societies (e.g., [McCoy *et al.*, 2013]), though it is unclear how they would be used to study agent strategies for fostering collective action. Alternatively, Diplomacy [Kraus and Lehmann, 1988; Paquette *et al.*, 2019; De Jonge *et al.*, 2019; Dafoe *et al.*, 2021; (FAIR) *et al.*, 2022] offers a potential avenue to study collective action, though it is a zero-sum game. Welfare Diplomacy, made public after this work was initiated, alters Diplomacy to be a general-sum game [Mukobi *et al.*, 2023]. In addition to the Junior High Game (JHG) presented in the next section, this variation potentially meets many of the properties listed in Table 1.

2.2 The Junior High Game

The JHG is played by a set of players I, $|I| \geq 2$. Over a sequence of rounds, each player $j \in I$ seeks to become *popular*. Initially, each player is assigned a popularity. This popularity changes over time based on the interactions among the players, which are abstractly represented by token exchanges. In every round, each player j allocates N tokens, each of which j can keep, give to another player, or use to attack (i.e., take from) another player. Keeping tokens positively impacts j's subsequent popularity, while tokens j gives to player i positively impact i's subsequent popularity but do nothing for j. Finally, when j attacks i, it negatively impacts i's subsequent popularity while positively impacting j's. The amount that token transactions impact popularity is dependent on the popularity of the player allocating the tokens.

Formally, let $x_{i,j}(\tau)$ denote the tokens that player j allocates to player i in round τ, such that $\sum_{i \in I} |x_{i,j}| = 1$. That is, $|x_{i,j}(\tau)|$ is the proportion of its tokens that j uses to interact with i in round τ. $x_{i,j}(\tau)$ can be positive or negative. Let $x^+_{i,j}(\tau) = \max(0, x_{i,j}(\tau))$ and $x^-_{i,j}(\tau) = -\min(0, x_{i,j}(\tau))$ be the proportion of their tokens that player j gives to and takes from player i, respectively. Note that only one of $x^+_{i,j}(\tau)$ and $x^-_{i,j}(\tau)$ can be non-zero since a player cannot both help and attack the same player in a round.

Let $\mathcal{P}_i(\tau)$ denote the popularity of player i at the beginning of round $\tau > 0$. Initially, each player is assigned a popularity $\mathcal{P}_i^{\text{init}}$. Popularity then changes over time based on token allocations made by players in the game as describe by Eqs. 1-3:

$$\mathcal{P}_i(\tau) = \max\left(0, (1-\alpha)^{(\tau-1)}\mathcal{P}_i^{\text{init}} + \sum_{j \in I} \mathcal{I}_{i,j}(\tau-1)\right) \quad (1)$$

$$\mathcal{I}_{i,j}(\tau) = \begin{cases} 0, & \text{if } \tau = 0 \\ \alpha\mathcal{V}_{i,j}(\tau) + (1-\alpha)\mathcal{I}_{i,j}(\tau-1), & \text{else} \end{cases} \quad (2)$$

$$\mathcal{V}_{i,j}(\tau) = \begin{cases} \mathcal{P}_i(\tau)[c^{\text{keep}} x^+_{i,i}(\tau) + \sum_{k \in I} c_k^{\text{take}}(\tau) x^-_{k,i}(\tau)], & \text{if } i{=}j \\ \mathcal{P}_j(\tau)[c^{\text{give}} x^+_{i,j}(\tau) - c_i^{\text{take}}(\tau) x^-_{i,j}(\tau)], & \text{else} \end{cases} \quad (3)$$

$\mathcal{V}_{i,j}(\tau)$ (Eq. 3) defines how token allocations made by j in round τ impact i's subsequent popularity. The top case computes the benefit that i gets from keeping tokens and attacking others. The bottom case computes the benefit (or loss) that i receives from tokens $j \neq i$ gives to (or takes from) them.

In both cases of Eq. (3), the impact of token allocations is weighted by two elements. First, token allocations are weighted by the popularity of the allocator. Token allocations made by more popular players have greater influence than those made by less popular players. Second, token allocations are also impacted by three augmenters. Augmenters model the ability for resource transfer to have positive-sum impact on the receiver (or the taker), while keeping resources for one's self may have less of an impact. Three scalar values, c^{keep}, c^{give}, and c^{take}, define these augmenters. Typically, $c^{\text{keep}} < c^{\text{give}} < c^{\text{take}}$ so that taking has the highest direct impact on popularity, followed by giving and then keeping. Furthermore, $c^{\text{give}} > 1$ to model the positive impact of *trade*.

In addition to positively impacting a player's own popularity, keeping tokens has the additional benefit of shielding a player against attacks. In Eq. 3, the impact of taking tokens (both as the attacker and the defender) is multiplied by the coefficient $c_k^{\text{take}}(\tau)$, which defines how much attacks on player k impact popularity. This coefficient is computed as:

$$c_k^{\text{take}}(\tau) = c^{\text{take}} \max\left(0, 1 - \frac{x^+_{k,k}(\tau)\mathcal{P}_k(\tau)}{\sum_{j \in I} x^-_{k,j}(\tau)\mathcal{P}_j(\tau)}\right) \quad (4)$$

If $\nexists j : x^-_{k,j}(\tau) \neq 0$, then $c_k^{\text{take}}(\tau) = 0$. In words, Eq. (4) states that the amount of defense that player k gains by keeping tokens depends on the ratio of the strength of the defense of k (i.e., $x^+_{k,k}(\tau)\mathcal{P}_k(\tau)$) to the strength of the total attack on k (i.e. $x^-_{k,j}(\tau)\mathcal{P}_j(\tau)$). If the strength of the attack is less than the strength of k's defense, then player k does not receive any damage from attacks in the round (and likewise, attackers gain no value from their attacks). However, if the sum of attacks exceeds this value, it will result in k losing some of their popularity (and other players receiving it).

After computing the impact of token allocations in round τ, the overall influence of player j on i's popularity through round τ is computed (by Eq. 2) as a convex combination (weighted by the popularity-update rate $\alpha \in [0, 1]$) of $\mathcal{V}_{i,j}(\tau)$ and influence arising from token allocations in past rounds. Finally, Eq. (1) defines $\mathcal{P}_i(\tau)$, which is a function of player i's initial popularity and the influences, $\mathcal{I}_{i,j}(\tau-1)$, from each player up to that point in the game.

Parameters and conditions can be varied to model many different scenarios (SM-1.3). Throughout this paper, we set $\alpha = 0.2$, $c^{\text{keep}} = 0.95$, $c^{\text{give}} = 1.3$, $c^{\text{take}} = 1.6$. We allow all players to observe (for all i, j, and τ) $\mathcal{I}_{i,j}(\tau)$ and $\mathcal{P}_i(\tau)$. However, player i can only view tokens they allocate or receive (i.e., $\forall j \in |I|\ x_{i,j}(t)$ and $x_{j,i}(t)$). We also assume that all players have $N = 2|I|$ tokens to allocate in each round. Additional details about the JHG are provided in SM-1.

The JHG models each of the properties listed in Table 1. Token allocations (which allow players to interact both positively and negatively with other players), influence, and popularity model a large and flexible strategy space. The goal to maximize one's own popularity gives conflicting interests. Furthermore, the weighting of tokens by one's popularity creates power asymmetry that in turn produces scenarios that typically do not perpetually converge within the time-scales we consider in this paper. The game can be easily scaled to any number of players, though in this paper we consider societies that have on the order of ten players. Finally, as illustrated in the next subsection, the game also encodes multiple forms of interdependence.

2.3 An Illustrative Example

To facilitate understanding of the JHG, we present an example game played by seven experienced human players (Figure 1). As depicted by the network corresponding to Round 2 in Figure 1, the players began the game by (primarily) giving tokens, seemingly at random. By Round 9, the players had segmented into two disconnected groups, wherein group members built up each other through trade (a form of interdependence). While all players had risen in popularity through collective action within their group, members of the larger group (p2, p3, p4, and p6) had higher popularity (and power) than those in the smaller group (p0, p1, and p5).

In Round 13, p1 attempted to weaken the larger group by attacking p6. This resulted in a momentary increase in the popularity of p1. However, in the next round, p6's friends (p2 and p3) joined with p6 in retaliating against p1. We contrast this with the smaller group, where p5 began attacking the other group, but both p1 and p0 did not. A war ensued over the next several rounds, which resulted in the larger group rendering the smaller group powerless by Round 17. This mitigation of an external threat through collective action illustrates a second form of interdependence.

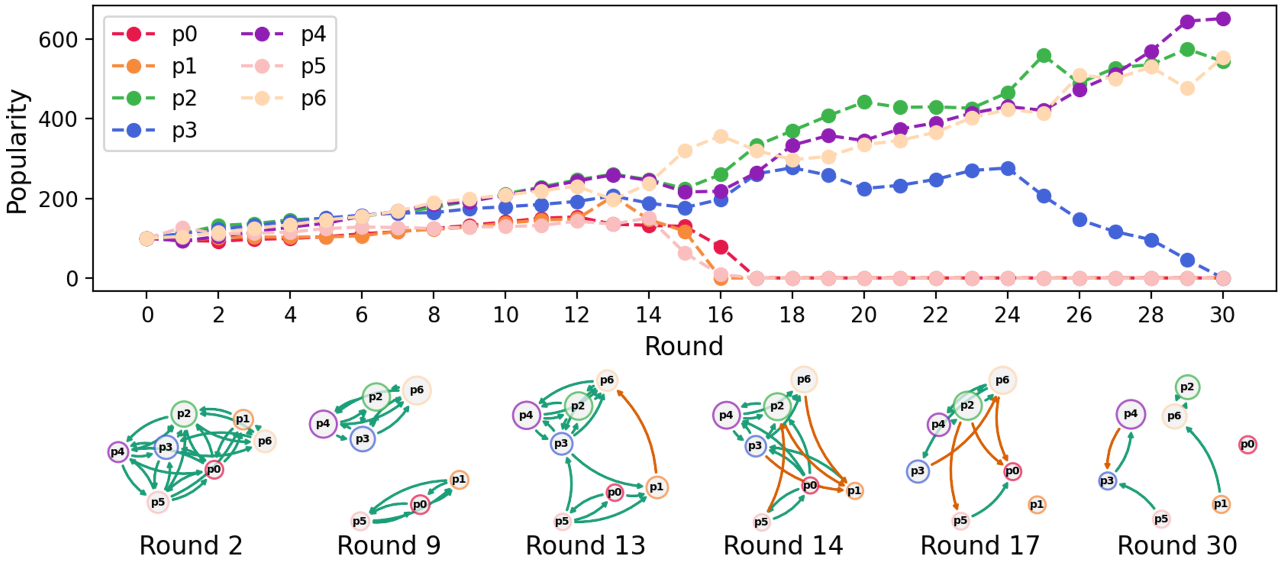

Figure 1: An example of a JHG game played among seven humans. (Above) Player popularities ($\mathcal{P}_i(\tau)$) in each round. (Below) Visualizations of the game network in selected rounds. Uniquely colored nodes, representing players, are positioned by influence, such that players i and j tend to be in close proximity when $\mathcal{I}_{i,j}(\tau)$ and $\mathcal{I}_{j,i}(\tau)$ are high. Larger nodes represent higher relative popularity. Directed edges denote token allocations in the round (i.e., $x_{i,j}(\tau)$), which are green if positive and orange if negative.

The altered social structure next led to a fracturing of the larger group. p3 dissented from the group by keeping most of their tokens in the rounds following the war, while p2, p4, and p6 continued to trade. Eventually, p2 (and later p4) began attacking p3, resulting in p3 eventually losing popularity.

While this particular game resulted in the formation of two groups, one of which destroyed the other, this need not have happened. In fact, all players could have achieved high popularity had they all chosen to act collectively in giving equally to each other throughout the entire game. In general, we have found, via playing and observing many JHG games, that each game is distinct from the others.

Despite variations between games, this example illustrates common challenges for how AI agents must act collectively to be successful in the JHG. First, like human players, AI agents must effectively form and join groups they can trust and which support them. Second, AI agents must determine when and how to work with other members of their group to stave off attacks from others. Finally, AI agents must be able to adapt their behavior to ever shifting power dynamics and community structures as groups form, change, and dissolve.

3 An Algorithm for Playing the JHG

In this section, we overview CAB (*C*ommunity-*A*ware *B*ehavior), a parameterized algorithm defining the behavior of an agent in the JHG. Details and code are given in SM-2.

3.1 Algorithm Description

A CAB agent, summarized in Algorithm 1, takes as input a parameter set Θ_i and observations of game play $\mathcal{G}_i(\tau)$, which contains, for all $t \in [0, \tau)$, the influence matrix ($\mathcal{I}(t)$), popularity functions $\mathcal{P}(t)$, and token allocations to and from the agent (i.e., $\forall j \in |I|\ x_{i,j}(t)$ and $x_{j,i}(t)$). It begins each round by identifying, based on prior token allocations, communities to which the players in the society belong. Based on these communities and its own preferences, a CAB agent then determines the community it would like to belong to. It then allocates its tokens to both form the desired community and to make its community successful. Θ_i determines how player i selects its desired community and allocates its tokens. These parameters can be varied to produce a wide range of behaviors. We overview each step of the algorithm in turn.

Algorithm 1 CAB token allocations in round τ for player i.

```
procedure ALLOCATETOKENS($\mathcal{G}_i(\tau)$, $\Theta_i$)
    $\mathbf{c}(\tau) \leftarrow$ detectCommunities($\mathcal{I}(\tau), \Theta_i$)
    $C_i^*(\tau) \leftarrow$ getDesiredCommunity($\mathbf{c}(\tau), \mathcal{G}_i(\tau), \Theta_i$)
    return computeAllocations($C_i^*(\tau), \mathbf{c}(\tau), \mathcal{G}_i(\tau), \Theta_i$)
end procedure
```

Step 1: Detect Communities

CAB first detects the communities based on the influence matrix $\mathcal{I}(\tau)$ (Eqs. 2-3), which is derived from all previous transactions up to round τ. CAB agents detect communities using one pass (including both phase 1 and 2) of the Louvain Method [Blondel *et al.*, 2008], which partitions agents into communities by greedily evaluating how changes in community structure increase modularity.

The Louvain Method requires a non-directed, non-negative graph. However, the influence matrix $\mathcal{I}(\tau)$ violates both of these assumptions. Thus, we compute modularity as a weighted sum of the modularity computed using the positive (i.e., $\mathcal{I}^+_{k,j}(\tau) = \max(0, \mathcal{I}_{k,j}(\tau))$) and negative (i.e., $\mathcal{I}^-_{k,j}(\tau) = |\min(0, \mathcal{I}_{k,j}(\tau))|$) influence matrices, respectively. The weighting is determined by Θ_i.

Step 2: Determine Desired Community

The community that player i is assigned in $\mathbf{c}(\tau)$ may not be their desired community. Thus, player i considers a subset of possible alterations to this community. These include communities that form from i moving to a different community, adding or subtracting single members from i's current community, or by combining i's current community with another community. Let $C_i^*(\tau) \subseteq I$ denote the set of agents that player i would like to be in its community.

To determine $C_i^*(\tau)$, agent i scores each possible set C based on five attributes: (1) *Modularity*, determined based on a community allocation in which all members $k \in C_i^*(t)$ are allocated to that community and other players are assigned to communities based on the Louvain Method; (2) *Target Group Strength*, which measures the collective strength of C relative to the group strength i desires (specified by Θ_i); (3) *Prominence* of agent i in C (i favors groups where it has high relative popularity); (4) *Familiarity*, the degree to which agents in C give to i; and (5) *Prosocial Behavior*, how much agents in C are reciprocating with each other. Each potential community is scored between 0 and 1 in each category. A weighted (by values in Θ_i) sum is then computed – $C_i^*(\tau)$ is the community with the highest weighted sum.

Step 3: Compute Token Allocations

Once CAB (player i) has determined $C_i^*(\tau)$, it then determines how to use its tokens to form and protect it. Player i does this in three steps. First, it determines how many tokens it should keep to protect itself from attack. Second, it considers attacking other players in three different forms: (1) attacks to retaliate against players that have taken popularity from i; (2) attacks on individuals that have attacked i's friends; and (3) unprovoked attacks on other players, designed to strengthen i's own target community $C_i^*(\tau)$ and weaken other communities. Fourteen different parameters from the set Θ are used to define these attacks and to ultimately pick which attack to carry out (if any). Third, i uses its remaining tokens to give to members of its community. The number of tokens it gives to each member of the desired community is based on (a) how much group members have reciprocated in the past and (b) its parameters Θ_i. Finally, player i keeps tokens if it cannot find anyone to give them to.

3.2 Learning Parameter Values

In parameterizing CAB agents using Θ_i, we considered that more popular (i.e., more powerful) players may encode different strategies than less popular players. Thus, we defined a CAB agent using three different sets of parameters. CAB uses one set of parameters to define its behavior when it is *poor* (i.e., $\mathcal{P}_i(\tau) < 0.75\bar{\mathcal{P}}(\tau)$, where $\bar{\mathcal{P}}(\tau)$ is the mean popularity in the society at time τ), another when it is *rich* (i.e., $\mathcal{P}_i(\tau) > 1.25\bar{\mathcal{P}}(\tau)$), and a third set when it is *middle* class.

The parameter set Θ_i, then, consists of 90 different parameters (30 for each popularity class as defined in Table 1 of the SM). This parameter set modulates the behavior of each aspect of a CAB agent. Different parameter values can cause CAB to be aggressive, collaborative, isolated, etc. As such, the success of a CAB agent depends on how these parameters are set. In this paper, we consider learning an effective parameter set using a combinatorial optimization algorithm. Given its simplicity and effectiveness, we use evolutionary simulations to learn effective parameter sets.

Our evolutionary simulations were conducted using a standard genetic algorithm. Initially, a random population of CAB agents (defined by their parameter values) are created. These CAB agents then interact in JHG games. After 200 games a new pool of CAB agents is created based on fitness (based on absolute and relative popularity standing), mutation, and crossover. The parameter settings of the top-performing agents from the 200th generation then define the parameterization of CAB agents used in our studies.

4 Associating with Human Players

Ultimately, AI agents should foster effective collective action in human societies. Thus, we begin by analyzing how CAB agents act collectively when playing the JHG with experienced human players. In particular, we consider two aspects of collective action: how well CAB agents (1) form and join groups that support them (group formation) and (2) work with members of their group to stave off threats (threat mitigation).

4.1 Experiment Design

We conducted an experiment in which experienced human players and CAB agents interacted in the JHG via the online platform found at `juniorhighgame.com`. Games were played under three conditions: (1) *Majority Human* (2 CAB agents and 6 humans); (2) *Even* (4 CAB agents and 4 humans); and *Majority Bot* (6 CAB agents and 2 humans).

Twenty-four people, participating in groups of eight, volunteered for the study. Each participant played three games (lasting 21-25 rounds), one in each condition, such that six games were played in each condition. To mitigate possible learning effects, conditions were counter-balanced across sessions. Participants were informed that approximately half of their competitors they encountered would be bots, but not who the bots were. As incentive, humans received monetary compensation proportional to their ending popularity.

To evaluate collective action with respect to group formation and threat mitigation, we analyze the performance and behavior of both individuals and society as a whole. As a measure of overall performance, we compare the popularity of human and CAB players across these games. We also consider the overall social welfare (as measured by average popularity) and wealth distribution (as measured by the Gini index [Gini, 1921] computed on popularity values) of societies in each condition. To measure behavior of societies and individuals, we use two classes of metrics. First, we consider the distribution of the type of token allocations used by the players. Second, we assess society behaviors by analyzing their mixing patterns at varying granularity. We analyze the formation of dyads (via *reciprocity*), triads (via the *clustering coefficient*), and groups (via *Louvain modularity*).

4.2 Results

Results of these experiments are summarized in Figure 2 and Table 2. As shown in Figures 2a-c, CAB agents' popularity levels were nearly on par with those of humans. Overall, humans had higher ending popularities than CAB players, but

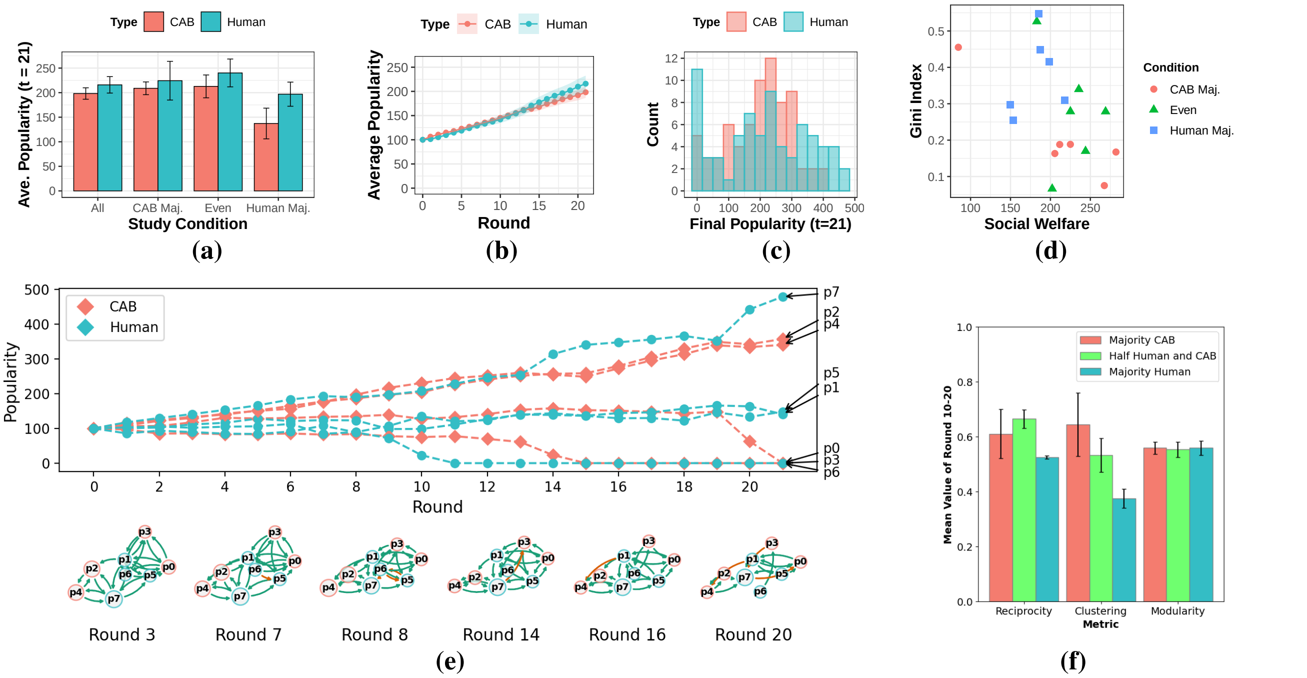

Figure 2: Results from the human-bot study. (a) The mean ending popularity of humans and CAB in each conditions. (b) The mean popularity of humans and CAB over time across all conditions. (c) A histogram of popularities of humans and CAB agents across conditions. (d) Social welfare (average popularity) vs. Gini Index in each game of the human-bot study. (e) Depicts one of 18 games played by humans and bots in the JHG. The four bots are depicted in orange and humans in blue. See Figure 1 for additional information about how to interpret the figure. (f) Average Mixing Patterns based on the last 10 rounds in each condition. Error bars and ribbons show the standard error of the mean.

Table 2: % token allocations by transaction type across study conditions. See Section 3.2 for classifications (poor, middle, and rich).

Player	% Keep	% Take	% Give
CAB (all)	37.07	0.08	62.85
CAB (poor)	63.07	0.43	36.49
CAB (middle)	31.16	0.01	68.83
CAB (rich)	33.53	0.00	66.47
Human (all)	15.82	8.38	75.80
Human (poor)	23.58	12.23	64.20
Human (middle)	13.82	7.15	79.03
Human (rich)	12.08	6.90	81.02

the difference is both relatively small and not statistically significant ($F(1, 140) = 2.293; p = 0.132$). This suggests that CAB agents were somewhat effective, though CAB agents had lower average popularity in Human-Majority games.

While the mean popularity of CAB players was nearly as high as human players across all conditions, the distribution of popularities between these two groups appear to be distinct (Figure 2c). Human players tended to have extreme (high or low) popularity, while CAB agents more often had moderate popularity levels. This suggests that strategies employed by human and CAB players were, on average, distinct, a trend that is verified in Table 2. Interestingly, CAB players rarely attacked other players, but frequently kept substantial amounts of tokens, especially when they were poor. On the other hand, human players frequently attacked others. They kept fewer tokens overall, though the same trend of keeping more tokens when poor is manifest. Since attacking other players both hurts overall social welfare and increases wealth disparity, it is not surprising that societies with more CAB agents tended to have better societal outcomes (Figure 2d, where lower Gini index and higher social welfare is better.).

These results paint a picture of the ability of CAB agents to form groups and mitigate threats. Greater understanding is gained by considering the game depicted in Figure 2e, which was played by four humans and four CAB agents. We highlight two aspects of this game. First, we consider the attack by p6 (human) on p5 (human) in Round 7. Whereas both p5 and p7 (a human friend of p5) retaliate against p6 in subsequent rounds, p0 (a bot who was also a friend to p5) did not. This suggests a failure by CAB agents to work with those in their group to mitigate outside threats. Second, we consider the relationship among p7 (human), p2 (bot), and p4 (bot). These players formed a strong relationship by Round 3, a relationship that largely persisted throughout the game. However, p7 occasionally attacked other players, which resulted in p7 becoming strong. While p2 and p4 reduced support to p7 after such attacks, they did not join in with the attacks, nor did they consider the potential threat that p7 was becoming to them.

This example highlights two attributes of CAB agents. First, as indicated by mixing patterns (Figure 2f), CAB agents were effective in forming groups in which they provided positive support to each other (through trade). This group formation, done in a way that humans could not identify them (SM-5.2), can largely explain the relatively high performance of CAB agents. However, CAB agents were less adept at threat mitigation, which we highlight in the next section.

Table 3: Mean popularity after 30 rounds in a society of eight CABs and two CATs. Numbers in parenthesis give the standard error.

Parameters	Trained w/ CATs?	CAB Popularity	CAT Popularity
Evolved (from random)	No	15.7 (±4.2)	146.6 (±5.0)
Evolved (from random)	Yes	71.5 (±1.1)	65.3 (±3.4)
Handcoded	N/A	262.4 (±11.0)	13.2 (±6.6)
Evolved (from handcoded)	No	9.4 (±0.4)	144.9 (±2.8)
Evolved (from handcoded)	Yes	19.6 (±4.2)	125.1 (±7.3)

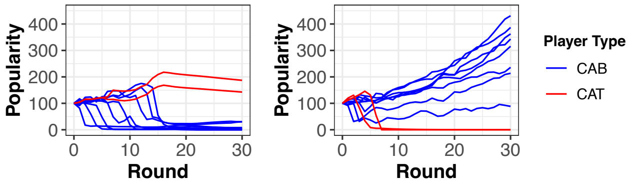

Figure 3: Popularities over time in prototypical games. (Left) Failure to mitigate CATs with evolved parameters. (Right) Successful mitigation of CATs with handcoded parameters.

5 Mitigating Adversarial Coalitions

To further explore how well CAB agents work together to mitigate threats, we consider societies with adversarial agents called CATs (*C*onspiring *A*utonomous *T*hieves).

5.1 Experiment Design

CATs work together to pillage non-CATs players. In the first round, a CAT signals their presence to other CATs by keeping all their tokens. Subsequent rounds consist of identifying and evaluating the community of CATs, then selecting the weakest non-CAT agent as a target. A CAT calculates the tokens to take from their target such that the total amount taken will reduce the non-CAT's popularity to zero. The CAT then keeps any unused tokens. If no targets are available, the CAT keeps all its tokens. Algorithmic details are provided in SM-4.

Given that two CATs are sufficient to overpower a group of eight other players that fail to work together, we evaluate the ability of eight CAB agents, parameterized in various ways, to overcome two CATs Specifically, we vary the CAB agents' training process with respect to initial parameters (random vs. handcoded) and whether the training scenarios include CATs.

5.2 Results

We first consider CAB agents evolved from randomly generated parameters sets and without training with CATs, which was how the CAB agents used in the study described in the previous section were trained. As shown by the first row of Table 3, the CATs are easily able to overpower these CAB agents. Popularity dynamics for a representative game in this condition are shown in Figure 3(Left). Clearly, these CAB agents fail to collectively mitigate this threat.

On the other hand, when these CAB agents are trained with CATs, they learn to exclusively keep (second row of Table 3), thus individually protecting themselves from attack. While this behavior is effective for avoiding attacks (producing higher average popularity than the prior group of agents), the resulting solution is inefficient.

The failure of these CAB agents to work together when facing CATs is due to the parameters learned by the optimization process rather than the ability of CAB agents to encode successful collective action. To demonstrate this, we considered handcoded CAB agents, which have parameter values set so that they retaliate against players that attack their friend (details provided in SM-4). When many members of the society have this behavior, they can disempower CATs and achieve high popularity (third row of Table 3). Figure 3(Right) shows a prototypical encounter between these CABs and CATs. While several of the CAB agents have their popularity substantially reduced initially, they all survive the coordinated attack and then effectively trade with each other thereafter in order to increase in popularity.

Given the existence of parameter settings that allow CAB agents to mitigate the threat posed by CATs, it is interesting to consider why the evolutionary algorithm does not learn them. We explore this with CAB agents evolved from the handcoded parameter setting. Results, shown in rows four and five of the table, indicate that the training process causes the agents to lose the ability to overcome the CATs. The optimization processes, which tune parameters based on the fitness of individual CAB agents, push the agents away from attacking others. This suggests that the reason the CAB agents are susceptible to exploitation is due to evolutionary instability in the parameter space, which is attracted to local minima in which they fail to mitigate threats. This highlights interesting future work that addresses how agents can learn to both form helpful groups and act together to mitigate threats.

6 Conclusion

In this paper, we explored the capabilities of community-based agents to learn collective action. We described the properties of a test-bed needed to study these capabilities, including a large and flexible state space, conflicting interests, interdependence, power asymmetry, non-convergence, and scalability. We then proposed a novel test-bed, the JHG, to model these properties. By simulating these properties, the JHG provides a test-bed for studying the ability of AI agents to foster collective action in complex societies.

To begin studying the capabilities of AI agents, we proposed the CAB algorithm, and evaluated its behavior in the JHG through simulations and user studies. The results indicate that CAB agents effectively form relationships and communities with humans and with other CAB agents. However, they sometimes fail to learn collective actions to defend their group against adversarial coalitions. These results give insights for future efforts to develop AI agents that foster collective action in complex social systems.

Supplementary Material (SM)

Supporting documentation, results, and code are available at: https://github.com/jakecrandall/IJCAI2024_SM.git

Acknowledgments

This work has been supported in part by the Office of Naval Research (N00014-18-1-2503).

References

[Acemoglu and Robinson, 2013] Daron Acemoglu and James A. Robinson. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. Profile Books, 2013.

[Axelrod, 1984] Robert Axelrod. *The Evolution of Cooperation*. Basic Books, New York, 1984.

[Baek *et al.*, 2022] In-Chang Baek, Tae-Gwan Ha, Tae-Hwa Park, and Kyung-Joong Kim. Toward cooperative level generation in multiplayer games: A user study in overcooked! In *2022 IEEE Conference on Games (CoG)*, pages 276–283. IEEE, 2022.

[Barabási and Albert, 1999] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[Bard *et al.*, 2020] Nolan Bard, Jakob N. Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H. Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. The Hanabi challenge: A new frontier for AI research. *Artificial Intelligence*, 280:103216, 2020.

[Biely *et al.*, 2007] Christoly Biely, Klaus Dragosits, and Stefan Thurner. The prisoner's dilemma on co-evolving networks under perfect rationality. *Physica D: Nonlinear Phenomena*, 228(1):40–48, apr 2007.

[Bishop *et al.*, 2020] Justin Bishop, Jaylen Burgess, Cooper Ramos, Jade B Driggs, Tom Williams, Chad C Tossell, Elizabeth Phillips, Tyler H Shaw, and Ewart J de Visser. Chaopt: a testbed for evaluating human-autonomy team collaboration using the video game overcooked! 2. In *2020 Systems and Information Engineering Design Symposium (SIEDS)*, pages 1–6. IEEE, 2020.

[Blondel *et al.*, 2008] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienn Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10, 2008.

[Bonacich and Lloyd, 2001] Phillip Bonacich and Paulette Lloyd. Eigenvector-like measures of centrality for asymmetric relations. *Social networks*, 23(3):191–201, 2001.

[Bowling *et al.*, 2017] Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. Heads-up limit hold'em poker is solved. *Communications of the ACM*, 60(11):81–88, 2017.

[Brandes *et al.*, 2007] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Gorke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. On modularity clustering. *IEEE transactions on knowledge and data engineering*, 20(2):172–188, 2007.

[Brown and Sandholm, 2019] Noam Brown and Tuomas Sandholm. Superhuman AI for multiplayer poker. *Science*, 365(6456):885–890, 2019.

[Campbell *et al.*, 2002] Murray Campbell, A. Joseph Hoane Jr, and Feng-hsiung Hsu. Deep blue. *Artificial intelligence*, 134(1-2):57–83, 2002.

[Centola, 2018] Damon Centola. *How behavior spreads: The science of complex contagions*, volume 3. Princeton University Press Princeton, NJ, 2018.

[Dafoe *et al.*, 2021] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. Cooperative AI: machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021.

[De Jong *et al.*, 2011] Steven De Jong, Daniel Hennes, Karl Tuyls, and Ya'akov Gal. Metastrategies in the colored trails game. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 551–558. Citeseer, 2011.

[De Jonge *et al.*, 2019] Dave De Jonge, Tim Baarslag, Reyhan Aydoğan, Catholijn Jonker, Katsuhide Fujita, and Takayuki Ito. The challenge of negotiation in the game of diplomacy. In *Agreement Technologies: 6th International Conference, AT 2018, Bergen, Norway, December 6-7, 2018, Revised Selected Papers 6*, pages 100–114. Springer, 2019.

[Easley and Kleinberg, 2010] David Easley and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010.

[(FAIR) *et al.*, 2022] Meta Fundamental AI Research Diplomacy Team (FAIR), Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

[Fehr and Gächter, 2002] Ernst Fehr and Simon Gächter. Altruistic punishment in humans. *Nature*, 415(6868):137–140, 2002.

[Fosong *et al.*, 2023] Eliott Fosong, Arrasy Rahman, Ignacio Carlucho, and Stefano V. Albrecht. Learning complex teamwork tasks using a sub-task curriculum. *ArXiv preprint arXiv:2302.04944*, 2023.

[Freeman, 1977] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[Gini, 1921] Corrado Gini. Measurement of inequality of incomes. *The economic journal*, 31(121):124–125, 1921.

[Granovetter, 1973] Mark S. Granovetter. The strength of weak ties. *American journal of sociology*, 78(6):1360–1380, 1973.

[Heider, 1946] Fritz Heider. Attitudes and cognitive organization. *The Journal of psychology*, 21(1):107–112, 1946.

[Holland and Leinhardt, 1971] Paul W. Holland and Samuel Leinhardt. Transitivity in structural models of small groups. *Comparative group studies*, 2(2):107–124, 1971.

[Katz, 1953] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[Kraus and Lehmann, 1988] Sarit Kraus and Daniel Lehmann. Diplomat, an agent in a multi agent environment: An overview. In *IEEE International Performance Computing and Communications Conference*, pages 434–435. IEEE Computer Society, 1988.

[Lerer and Peysakhovich, 2017] Adam Lerer and Alexander Peysakhovich. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *ArXiv preprint arXiv:1707.01068*, 2017.

[Liaqat *et al.*, 2020] Aisha Liaqat, Muddassar Azam Sindhu, and Ghazanfar Farooq Siddiqui. Metamorphic testing of an artificially intelligent chess game. *IEEE Access*, 8:174179–174190, 2020.

[McCoy *et al.*, 2013] Josh M. McCoy, Mike Treanor, Ben Samuel, Aaron A. Reed, Michael Mateas, and Noah Wardrip-Fruin. Prom week: Designing past the game/story dilemma. In *Proceedings of the 8th International Conference on the Foundations of Digital Games*, 2013.

[McPherson *et al.*, 2001] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[Mukobi *et al.*, 2023] Gabriel Mukobi, Hannah Erlebach, Niklas Lauffer, Lewis Hammond, Alan Chan, and Jesse Clifton. Welfare diplomacy: Benchmarking language model cooperation. *arXiv preprint arXiv:2310.08901*, 2023.

[Newman *et al.*, 2002] Mark Newman, Stephanie Forrest, and Justin Balthrop. Email networks and the spread of computer viruses. *Physical Review E*, 66(3):035101, 2002.

[Newman, 2002] Mark Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[Newman, 2006] Mark Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[Newman, 2010] Mark Newman. Networks: An introduction, 2010.

[Paquette *et al.*, 2019] Philip Paquette, Yuchen Lu, Seton Steven Bocco, Max Smith, Satya O-G, Jonathan K Kummerfeld, Joelle Pineau, Satinder Singh, and Aaron C Courville. No-press diplomacy: Modeling multi-agent gameplay. *Advances in Neural Information Processing Systems*, 32, 2019.

[Rosero *et al.*, 2021] Andres Rosero, Faustina Dinh, Ewart J de Visser, Tyler Shaw, and Elizabeth Phillips. Two many cooks: Understanding dynamic human-agent team communication and perception using overcooked 2. *arXiv preprint arXiv:2110.03071*, 2021.

[Rubin and Watson, 2011] Jonathan Rubin and Ian Watson. Computer poker: A review. *Artificial intelligence*, 175(5-6):958–987, 2011.

[Salancik, 1978] Gerald R Salancik. *The external control of organizations: A resource dependence perspective*. New York: Harper & Row, 1978.

[Shi *et al.*, 2022] Zhenyu Shi, Wei Wei, Matjaž Perc, Baifeng Li, and Zhiming Zheng. Coupling group selection and network reciprocity in social dilemmas through multilayer networks. *Applied Mathematics and Computation*, 418:126835, 2022.

[Silver *et al.*, 2018] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018.

[Skyrms, 2003] Brian Skyrms. *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press, 2003.

[Walton-Rivers *et al.*, 2019] Joseph Walton-Rivers, Piers R. Williams, and Richard Bartle. The 2018 Hanabi competition. In *2019 IEEE Conference on Games (CoG)*, pages 1–8. IEEE, 2019.

[Watts and Strogatz, 1998] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684):440–442, 1998.

[Wellman *et al.*, 2003] Michael P. Wellman, Shih-Fen Cheng, Daniel M. Reeves, and Kevin M. Lochner. Trading agents competing: Performance, progress, and market effectiveness. *IEEE Intelligent Systems*, 18(6):48–53, 2003.