Modeling of human group coordination

Hannes Hornischer,[1,2] Paul J. Pritz,[3] Johannes Pritz,[4] Marco G. Mazza,[2,5,*] and Margarete Boos[6,†]

[1]*Institute of Systems Sciences, Innovation, and Sustainability Research, University of Graz, Graz, Austria*
[2]*Max Planck Institute for Dynamics and Self-Organization (MPIDS), Am Faßberg 17, 37077 Göttingen, Germany*
[3]*Department of Computing, Imperial College London, London, United Kingdom*
[4]*Courant Research Centre Evolution of Social Behaviour, University of Göttingen, Göttingen, Germany*
[5]*Interdisciplinary Centre for Mathematical Modelling and Department of Mathematical Sciences, Loughborough University, Loughborough, Leicestershire LE11 3TU, United Kingdom*
[6]*Faculty for Biology and Psychology, Department of Social and Communication Psychology, University of Göttingen, Göttingen, Germany*

(Received 22 September 2021; accepted 23 March 2022; published 12 April 2022)

We study the coordination in a group of humans by means of experiments and simulations. Experiments with human participants were implemented in a multiclient game setting, where players move on a virtual hexagonal lattice, can observe their and other players' positions on a screen, and receive a payoff for reaching specific goals on the playing field. Flocking behavior was incentivized by larger payoffs if multiple players reached the same goal field. We choose two complementary simulation methods to explain the experimental data: a minimal cognitive force approach, based on the maximization of future movement options in the agents' local environment, and multiagent reinforcement learning (RL), which learns behavioral policies to maximize reward based on past observations. Comparison between experimental and computer simulation data suggests that group coordination in humans can be achieved through nonspecific, information-based strategies. We also find that although the RL approach can capture some key aspects of the experimental results, it achieves lower performance compared to both the cognitive force simulation and the experiment, and matches the observed human behavior less closely.

DOI: 10.1103/PhysRevResearch.4.023037

I. INTRODUCTION

Human group coordination [1–6] is a ubiquitous but challenging problem, as it spans multiple scales, from few individuals to large-scale organizations such as political elections and international treaties [1,7,8]. Group coordination necessarily strikes a balance between individual strategies and social flocking behavior. Different modeling approaches have been developed to describe emergent patterns on a group level, based on the interaction of individual-level strategies [9–15]. "High-level models" [16] directly encode empirically observed features of individual or collective behavior. Another class of models attempts to explain emergent patterns in group behavior by making assumptions on the individuals' underlying mental processes and encoding these in a group model [17–19]. In contrast to these, we present a functional approach, where we aim to reproduce observed human behavior, by applying simple formalized decision making rules. We address the question of whether high-level models, based on empirical data or psychological assumptions, can be replaced by "low-level" [16,20], functional models of group coordination with potentially explanatory power for the collective pattern.

It is commonly assumed that an agent's decision process involves two components: an estimation phase, and a decision rule [4]. In this work, we evaluate two paradigms for the decision process, i.e., (1) choosing the action that maximizes future options [21–24] and (2) choosing the action that leads to the highest expected reward learned from past experience [25]. To examine these two simple decision making rules, we employ the following methods: (i) computer simulations of agents subject to an interaction called "cognitive force," which is based on the maximization of future options, and (ii) multiagent reinforcement learning (RL), which learns reward-maximizing policies from past observations collected from the environment. As the basis for our evaluation, we use experimental data, which was collected using a multiplayer game with human participants. The goal of this work is to compare and evaluate two complementary methods to explain the experimental data. Their complementarity consists in how estimation and decision rules are implemented. In the cognitive force model, decision making is based on the maximization of future movement options in the agents' local environment, and evaluation of *future* options; in the RL model, *past* observations are used to create behavioral policies in order to maximize reward.

On the one hand, the cognitive force approach using maximum entropy arguments has already proven fruitful in modeling distributions of organisms [26], identifying models for flocks of birds [27], or phase transitions in pedestrian

[*] Corresponding authors: m.g.mazza@lboro.ac.uk
[†] mboos@gwdg.de

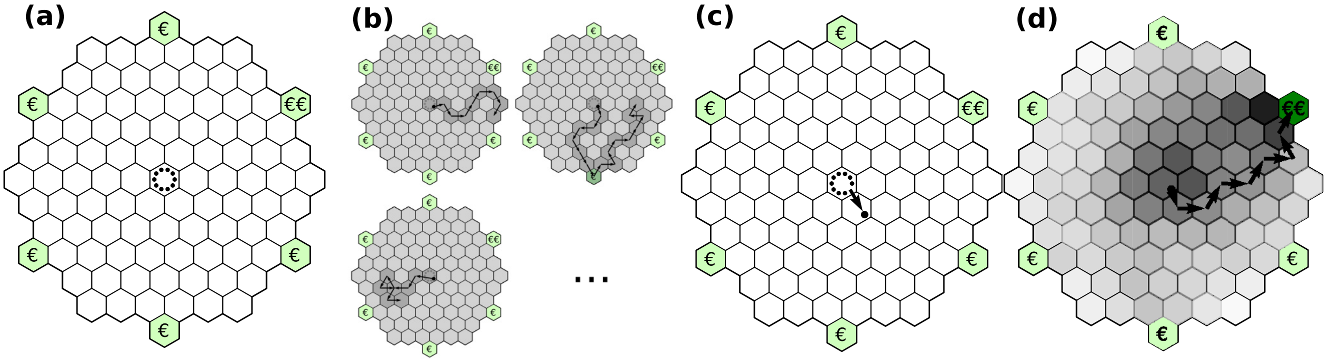

FIG. 1. (a) Schematic illustration of the systems' setup. Each human player or simulated agent is represented by a slightly larger black dot in the experiment or simulation, respectively (not visible in these schematics). The six green fields with € symbols represent the reward fields, while the one with two € symbols represents a reward field where biased players/agents receive relatively higher payoff. (b) Cognitive force model. Three exemplary hypothetical random walks of a simulated agent located at the center field. Visited fields are marked in dark gray. As shown in the top-right figure, when an agent reaches a reward field for the first time, it receives additional moves. Hence, the agent continues its hypothetical walk. The three walks, toward the right (r) shown in the top-left panel, down-right (d-r) shown in the top-right panel, and left (l) shown in the bottom-left panel, are associated with a number of visited fields of $A_{\mathrm{r}} = 9$, $A_{\mathrm{d-r}} = 17$, and $A_{\mathrm{l}} = 8$, respectively. Accordingly, the agent would choose to move to the adjacent field down-right, as illustrated in (c). (c) Based on its hypothetical random walks an agent determines which adjacent field to move toward. (d) The relative frequency of finding an agent on a certain hexagon throughout a simulation, where the darker the gray scale, the higher the frequency. One can observe how agents tend to move to the higher-reward field in the top right, marked with €€. One exemplary trajectory of an agent is indicated via arrows. The frequency distribution is averaged over 100 independent simulation runs.

dynamics [11]. In this paper we present a novel application of this approach, which is grounded in physics [20,23] and cognitive and artificial intelligence sciences [28–30], by using it to model human group behavior.

Multiagent reinforcement learning, on the other hand, has been successfully applied to solve group optimization tasks such as swarm navigation, communication-based tasks, and energy grid management [31–33]. However, to the best of our knowledge, it has not yet been used to model observed human behavior.

This article is organized as follows. In Sec. II we describe the experimental setup and the information given to human participants. In Sec. III both modeling approaches used here are described: the cognitive force model and the multiagent RL approach. In Sec. IV we present our results and critically compare the two modeling methods. Finally, in Sec. V we draw our conclusions.

II. EXPERIMENTS

We build both simulation approaches in a setup conceptually identical to the experiment described below. In the following, we refer to human participants in the experiment as "players" and to simulated entities as "agents." To define what information the players and simulated agents have access to, and what external influences have an impact on the players'/agents' behavior, we construct a well-controlled environment. The experiments and related computer simulations presented here are conducted using the HoneyComb paradigm [34].

Human participants in groups of ten move in a virtual environment, i.e., a multiclient game setup through which they only have information about the in-game behavior of other participants, preventing real-life visual or auditory interactions. The virtual playground of the game consists of hexagonal fields as schematically illustrated in Fig. 1(a). Each participant is represented by an avatar, and has solely insight into the current position and movement of their co-participants. From their current positions, each participant can move their avatar to one of the adjacent hexagons. At the beginning of the game, all players' avatars are positioned in the center of the honeycomb. Each player has a fixed number $M_{\mathrm{exp}} = 15$ of available moves to play a game. Within this setup, we implemented incentives at the individual level via six reward fields, granting payoffs (marked in green, with € symbols). The players were given the following information at the start of the game:

(1) A majority of eight players was informed of six equal lower-reward goal fields.

(2) A minority of two randomly selected players was informed that one specific goal field held a higher reward [marked with €€ in Fig. 1(a)], while the other goal fields held a lower reward. The higher-reward field is randomly positioned, but the same for both biased and unbiased players.

(3) Players were not informed about whether they were in the unbiased or in the biased group or even that there was a difference in information among the players.

(4) Players were informed that payoff for arriving at a reward field was multiplied by the number of co-players at the same reward field by the end of the game.

By limiting the information players receive in this way, we induce goal-directed behavior, as well as group cohesion. However, we do not motivate any specific behavioral strategies, but only install a reward structure. While players strive to maximize their individual reward, this depends on the other players' actions, rendering coordination advantageous. In this paper, we use experimental data collected by Boos *et al.* [35], who conducted 40 iterations of this experiment, each of which with a distinct group of 10 participants, i.e., a total of 400 participants.

III. MODELING

We now describe the two modeling approaches and compare their results with the experiments. In the cognitive force approach, the agents' motion obeys the following basic algorithm: (1) compute number of options; (2) make one move; (3) repeat until the number of moves runs out or a goal field is reached. Each agent has a fixed number $M_{\text{sim}} = 25$ of moves to complete a game. The number of moves available in the experiment and the cognitive force simulation differs; please refer to Sec. III in the Supplemental Material [46] for more details. In order to make a move, the number of options is calculated according to the steps described in the following. Given the current position $\boldsymbol{x}^j_{t,k}$ of agent $j = 1, \ldots, 10$, there are six adjacent fields $\boldsymbol{x}^j_{t,1}, \ldots, \boldsymbol{x}^j_{t,6}$, at time t with $t = 1, \ldots, 25$ (except at the boundary of the playing field). The agent will move to the field $\boldsymbol{x}^j_{t,\tilde{k}}$ associated with the largest number of options. Options associated with each adjacent field are evaluated by simulating the hypothetical trajectories emanating from each adjacent field $\boldsymbol{x}^j_{t,1}, \ldots, \boldsymbol{x}^j_{t,6}$. This is illustrated in Figs. 1(b)–1(d). Each hypothetical trajectory is a random walk $\mathcal{T} = (\boldsymbol{x}^j_a, \boldsymbol{x}^j_b, \ldots)$ of finite length M_{hyp} with equal probabilities to move to one of six neighboring fields at each step. M_{rem}, the remaining number of moves an agent has, with $1 \leqslant M_{\text{rem}} \leqslant M_{\text{sim}}$, starts at $M_{\text{rem}} = M_{\text{sim}}$ at the beginning of the game and decreases to $M_{\text{rem}} = 1$ just before the end of the game.

In order to encourage agents to move to a reward field, we introduce the payoff M_{rew} that is obtained upon reaching a reward field and which is calculated so as to mirror the reward structure in the experiment with human players, and described as follows. When n agents reach the same reward field during a random walk, they obtain a payoff $M_{\text{rew}} = n\mathcal{P}$, where $\mathcal{P}$ is a default payoff. A biased minority of two agents is furnished with double the amount of additional moves when reaching the higher-reward field (€€) during hypothetical trajectories; this multiplication is also applied when other agents are present on that higher-reward field. We consider additional moves a meaningful currency for payoff when reaching any of the six reward fields; i.e., we increase M_{rew} if an agent visits a reward field during a simulated random walk. This payoff allows the agents to extend their hypothetical trajectory and thus to visit more fields. During a hypothetical trajectory agents receive a payoff only once, i.e., upon reaching a reward field for the second time (either the same or another); no additional payoff in moves is received.

Then, starting from the current position $\boldsymbol{x}^j_t$, each neighboring field $\boldsymbol{x}^j_{t,1}, \ldots, \boldsymbol{x}^j_{t,6}$ is the beginning of $N_{\text{walk}} = 5000$ random walks of finite length M_{hyp}. The number of fields visited during each random walk equals the number of options associated with the field $\boldsymbol{x}^j_{t,k}$, $k = 1, \ldots, 6$, and trajectory $\mathcal{T}$, denoted as $N_f(\boldsymbol{x}^j_{t,k}, \mathcal{T})$. Every visited field is counted only once, i.e., $N_f \leqslant M_{\text{hyp}}$.

In its motion, agent j will choose the neighboring field $\boldsymbol{x}^j_{t,\tilde{k}}$ associated with the single hypothetical trajectory with the highest $N_f(\boldsymbol{x}^j_{t,k}, \mathcal{T})$. When agents reach a reward field during actual movement, analogously to the experiment, they will remain at the reward field until the end of the game. The calculation of $N_f(\boldsymbol{x}^j_{t,k}, \mathcal{T})$ is repeated for each subsequent step on the playing field, and each agent independently of each other. During the simulation of these trajectories for any single agent, the other agents remain stationary. This means that rewards for visiting a reward field are only multiplied by the number of agents already on that field.

We show three exemplary, hypothetical trajectories for an agent located on the central hexagon in Fig. 1(b). The fields visited by those hypothetical trajectories are marked in dark gray. Based on the three hypothetical trajectories shown in Fig. 1(b) an agent would decide to move down-right, because the trajectory starting from the down-right field explored the largest number of fields, as illustrated in Fig. 1(c). During the hypothetical trajectory in Fig. 1(b), top-right panel, an agent reaches a € reward field.

The simulation explained above mirrors the experimental setup by incentivizing goal-directed behavior and group cohesion. Goal-directed behavior is motivated through the additional moves received from visiting a reward field, while group cohesion is promoted by multiplying payoffs by the number of agents on a visited reward field.

IV. RESULTS

In our first simulation approach, namely the cognitive force simulation, we find that the agents exhibit a strong tendency to move toward the higher-reward field, Fig. 1(d). The agents' relative occupation frequency is shown in Fig. 1(d) with the gray shading. This result suggests that a biased minority has a significant impact on the majority, i.e., the unbiased agents, and can "lead" them to the higher-reward field [Fig. 2(a)]. This is in agreement with previous work [3,9,35,36]. To further compare the simulation and human experiment we quantify the tendency of players/agents to coordinate as a group by measuring the distribution of the size of the largest group of players/agents that reached any reward field by the end of a game; see Fig. 2(b). In the experiments, the relative frequency is highest for a group of $n = 8$ players, followed by groups of $n = 9$ and all 10 players on the same field. The corresponding cognitive force model results also exhibit large peaks for $n \geqslant 8$, and we find the maximum in relative frequency for $n = 10$, closely matching the experiment's results. Figure 2(c) captures the general pattern of choices in the game. It shows the relative frequency of finding n_{u} unbiased players/agents on the higher-reward field. The distribution exhibits peaks for $n_{\text{u}} = 0$ and $n_{\text{u}} = 8$, corresponding to no unbiased agents and all unbiased agents reaching the higher-reward field, respectively. The cognitive force model agents reproduce the features of the experimental data.

Assuming there is no biased minority in the game, the probability to find an agent at a certain reward field by the end of a game is $p = \frac{1}{6}$, since all reward fields are equivalent. However, the biased minority is able to influence the unbiased majority such that the relative frequency of agents at the higher-reward field by the end of a game is significantly higher than the lower-reward field, i.e., $f_{€€,\text{cogF}} \approx 0.39$ for the

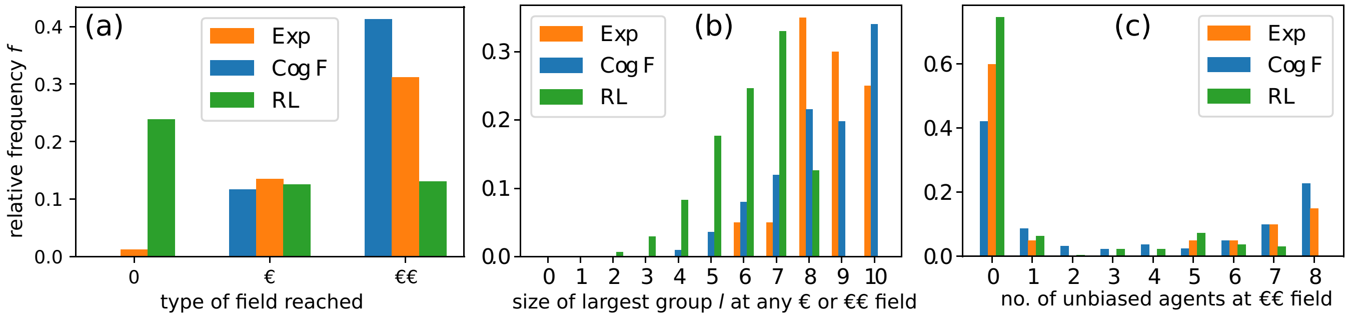

FIG. 2. (a) The relative frequency of players/agents reaching no reward field (0), a lower-reward field [€], or a higher-reward field [€€] by the end of a game. The relative frequency of reaching the higher-reward field is significantly larger than the relative frequency of reaching any single lower-reward field. The relative frequency of players/agents on a lower-reward field at the end of a game is averaged over all of the five lower-reward fields; that is, data are normalized such that $f_0 + 5f_{€} + f_{€€} = 1$. (b) The relative frequency of the largest group or the largest number of players/agents reaching the same (higher- or lower-) reward field. For both experiment and simulation the tendency of coordinating as a group in ultimately reaching the same reward field emerges, while such a tendency is only partially observable for the RL agents. (c) Relative frequency of finding n_{u} number of unbiased players/agents on the higher-reward field from experimental (orange), simulation (blue), as well as RL (green) results. The highest relative frequency is to find no unbiased agent on the higher-reward field. Second-highest relative frequency is for $n_{\mathrm{u}} = 8$, indicating that all unbiased agents moved as a group to the higher-reward field.

cognitive force simulation. This closely replicates the results obtained from the experiment with human players, where this frequency was $f_{€€,\mathrm{exp}} \approx 0.32$. We also observe a similarity between the experiment and the simulation in that the biased minority either leads most (in about 15% of all games with human players compared to around 25% in the simulation) or no unbiased players (60% in the games with humans vs 40% in the simulation) toward the higher-reward field [Fig. 2(c)]. We may call this an all-or-nothing collective response [37].

This means that the simulated agents, although they lack higher cognitive abilities, can reproduce the same shift of probabilities as observed in the human experiment.

As a second approach, alternative to the simulations based on option-maximizing agents, we implement a multiagent RL algorithm [38] to learn a behavioral policy for each agent that can be directly applied in our system. We express our simulation setup using the framework of Markov games, an extension of the Markov decision process [39]. We define the state set S, where the agent-specific state is given by the positions of all agents, as well as the positions of the goal fields and the remaining moves. The biased agents, additionally, observe the location of the higher-reward fields as part of their state. The action set for each of the N agents, i.e., $A_1, A_2, \ldots, A_N$, consists of the moves to the adjacent hexagons and a void action that allows agents to remain on their current field. The state transition function $T : S \times A_1 \times \ldots \times A_N \mapsto S$ defines the transition to next states given the current state and all agents' actions. In our simulations, this is deterministic and simply updates all agents' states using the moves selected in the last time step, which are executed concurrently. Lastly, the reward function $R_i : S \times A_1 \times \ldots \times A_N \mapsto \mathbb{R}$ defines the payoff received by each agent, given a state and all agents' actions. The reward function reflects the setup used for the experiments with human players; i.e., agents receive zero reward unless they reach a payoff field, in which case the reward is computed according to the rules outlined above. Our goal is then to learn a policy $\pi_{\theta_i} : S \mapsto P(A_i)$ for each agent i, parametrized by θ_i, that maximizes its respective reward obtained in a game. Note that the dependence of the state transition and the reward function on the actions chosen by the other players introduces the problem of nonstationarity to our simulation, rendering single-agent RL algorithms invalid [40]. Consequently, we employ a state-of-the-art multiagent RL algorithm proposed by Iqbal and Sha [38]. They propose an actor-critic algorithm, where a centralized critic learns state-action value functions for each agent based on all agents' states and actions. At the same time, each agent has its own policy (actor) that only requires its own state as an input. Thus, the agents do not rely on explicit sharing of information, i.e., of chosen actions, at test time, while reducing the nonstationarity problem in the training process. The components of the RL algorithm, i.e., (1) the centralized critic and (2) the agents' individual policies, are learned as follows. We denote agent i's state, action, and reward as s_i, a_i, and r_i, respectively, agent i's next state as s'_i and next action as a'_i, the parameters of the state-action value function as ψ, and the parameters of agent i's policy as θ_i. Both the critic and the policies are parametrized by neural networks, the details of which can be found in Sec. I in the Supplemental Material [46]. To increase training stability, we also use a target critic and a set of target policies, which are effectively delayed versions of the critic and policies; i.e., their parameters are updated more slowly. The parameters of these target networks are denoted by $\overline{\psi}$ and $\overline{\theta_i}$, respectively. The critic as well as the policies are learned using data sampled from a replay buffer D; i.e., we use data not only from the current game but from a set of games played by the agents. A centralized critic, providing an estimate of the state-action value function for all agents, is then learned by minimizing the loss $\mathcal{L}_Q$,

$$\mathcal{L}_Q(\psi) = \sum_{i=1}^{N} \mathbb{E}_{(s,a,r,s')\sim D}\big\{\big[Q_i^{\psi}(s, a) - y_i\big]^2\big\}, \qquad (1)$$

$$y_i = r_i + \gamma \mathbb{E}_{a'\sim\pi_{\overline{\theta}}(s')}\big\{Q_i^{\overline{\psi}}(s', a') - \alpha \ln[\pi_{\overline{\theta_i}}(a'_i|s'_i)]\big\}, \qquad (2)$$

where Q_i^{ψ} is the current state-action value function, $\ln_{10}[\pi_{\overline{\theta_i}}(a'_i|s'_i)]$ is an entropy term used as a regularization, α

is a temperature parameter setting the balance between maximizing entropy and rewards, and $\mathbb{E}_{(s,a,r,s')\sim D}$ is the expectation over tuples sampled from the replay buffer [38]. The target y_i measures agent i's immediate payoff r_i obtained from taking action a given state s as well as the expected payoff thereafter, captured by the state-action value function of the next state s' and next action a', i.e., $Q_i^{\overline{\psi}}(s', a')$. For learning the state-action value function, actions and states from all agents are used. The individual agents learn their respective policy by ascending the gradient of their performance function $J(\pi_\theta)$,

$$\nabla_{\theta_i} J(\pi_\theta) = \mathbb{E}_{s\sim D, a\sim\pi}(\nabla_{\theta_i} \ln[\pi_{\theta_i}(a_i|s_i)] \\ \{-\alpha \ln[\pi_{\theta_i}(a_i|s_i)] + Q_i^{\psi}(s, a) - b(s, a_{\setminus i})\}), \quad (3)$$

where $\pi_{\theta_i}(a_i|s_i)$ is agent i's policy, $\mathbb{E}_{s\sim D,a\sim\pi}$ is the expectation over states sampled from the replay buffer and actions sampled from the agent's policy given these states, and $b(s, a_{\setminus i})$ is a baseline calculated as $\mathbb{E}_{a_i\sim\pi_i(o_i)}[Q_i^{\psi}(s, (a_i, a_{\setminus i}))]$. This baseline helps determine the value of taking a particular action by comparing the state-action value function with the value of taking the "average" action with all other agents fixed [38]. We refer the reader to Iqbal and Sha [38] for more details on this algorithm.

In our RL-based simulations, we train the agents for up to 2×10^6 episodes and then use the set of models that performed best across these to evaluate their behavior in our system. One episode consists of one complete game, where all agents start in the center of the honeycomb and the game is finished once all agents have arrived at a goal field or reached the maximum number of moves [41].

While one would expect the agents to converge to optimal behavior, i.e., jointly arriving on the higher-reward field in each game, we in fact do not observe such results in our RL-based simulations. The agents apparently fail to discover a set of cooperative policies following the biased agents that would lead to the reward-maximizing outcome. Instead we find that smaller groups of up to 8 agents tend to jointly arrive on lower-reward fields, with a small fraction failing to arrive at any reward field or arriving at a higher-reward field (see Fig. 2). We consequently conjecture that the RL agents do not learn to follow the biased agents to higher-reward fields, as observed in both the human experiments and the cognitive force simulation, and instead converge to locally optimal policies, leading them to a lower reward field. These results can be partially explained with the rules of the game underlying both simulations, making the application of RL difficult. The simulation contains 10 agents and the rewards are sparse; i.e., the agents only receive a payoff at the end of each game, and the reward structure creates a high degree of interdependency between agents as their individual rewards are significantly impacted by their ability to coordinate. In addition, our setting is a mixed competitive-cooperative environment, since two biased agents may compete for the followership, while the optimal outcome is a cooperative one; i.e., all agents move to the same higher-reward field. Iqbal and Sha [38] focus on cooperative problems, which could be a factor explaining the difference in observed performance. Our RL simulations also exhibited high brittleness, i.e., failed to converge to a stable optimum and instead alternated between phases of moderate and poor performance over the course of the training process. For an example of this, please see Fig. S1 in the Supplemental Material [46]. All of these aspects make it particularly difficult to learn optimal policies via RL and might thus explain our results. It should be noted that we were able to reproduce the basic results reported by Iqbal and Sha [38], which confirms both the applicability of their algorithm to certain problems, as well as the correctness of our implementation. The remaining steps we took to ensure correctness of our implementation can be found in the Supplemental Material [46].

The ability to reach consensus has a cost. In our setting, this cost is quantified by the number of moves employed to coordinate successfully. To separate the impact of coordination from leader-follower dynamics, we analyzed the experiment, cognitive force model simulations, and RL-based simulations in the absence of biased players and agents, respectively. Figure 3 shows the dependence of the mean distance d of players and agents from the reward field (which is eventually reached) in each game on the normalized group-cumulative number of moves μ of all players/agents; μ is the sum of all the moves used by all unbiased players/agents in a given game, and is normalized by $M_{\text{exp}}N$ or $M_{\text{sim}}N$, for experiments and simulations, respectively, where N is the number of players/agents. For the human participants, the mean distance d decreases linearly until $\mu \approx 0.6$, after which it slowly converges to $d = 0$. This shows that in the experiment players tend to move directly to a reward field. In the cognitive force model simulations, the mean distance d initially decreases slowly, and then for $\mu > 0.3$, d decreases more steeply when agents move toward the reward field. In the cognitive force model simulations, agents first move outward from the starting point and later move toward a reward field. Contrary to this, RL agents' distance decreases almost linearly over the entire number of moves available to them, suggesting that their movement is less directed than that of human players and that they take longer to reach a reward field.

Our model explicitly avoids *ad hoc* assumptions about the dynamical behavior; thus we do not expect a perfect agreement of the dynamical features of the movement behavior. There is an incentive for group cohesion on the reward field manifesting itself in the payoff scheme but no incentive to flock at every step.

V. CONCLUSIONS

We show how a general behavioral principle—maximization of options—leads to the emergence of coordination in groups with a leader-follower dynamic. Namely, letting computer-simulated agents maximize their number of possible options produces collective patterns which match the results of the experiment studied here and of equivalent experiments on coordination and leadership in human groups rather well [35]. Specifically, we observe that a biased minority is able to significantly influence a majority and also confirm common patterns such as first-mover effects in the simulation data [37] (see Secs. IV and V of the Supplemental Material [46]).

We conclude that high-level models and tailored proximate principles for the reproduction of collective patterns can be

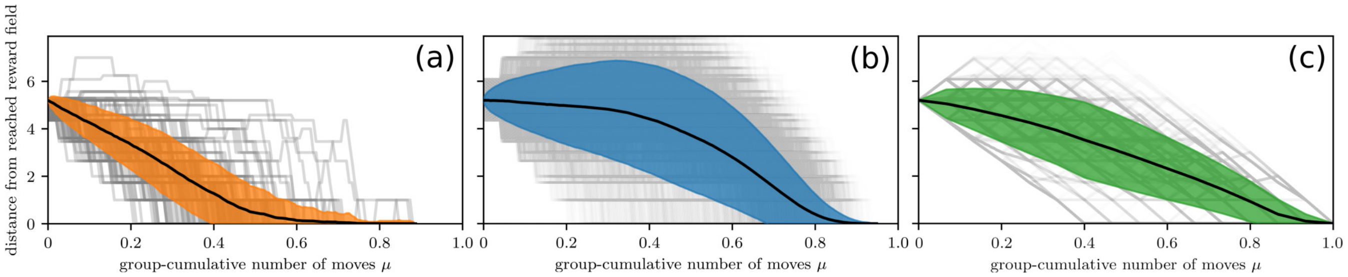

FIG. 3. Dependence of the distance from a reward field on the normalized group-cumulative number of moves of all players/agents for (a) unbiased players in the experiment, (b) unbiased cognitive-force simulated agents, and (c) unbiased RL agents. Movement of unbiased players/agents is shown only for games in which 7 or more players/agents reached the same reward. Players/agents which did not reach any reward field are not shown. The percentage of games where fewer than 7 players/agents reached the same reward field in experiments, cognitive-force simulation, and RL is 10%, 10%, and 22%, respectively. Individual players'/agents' trajectories are indicated in gray. The black line represents the mean distances from the reward field. The shaded areas indicate the standard deviation from the mean. The group-cumulative number of moves is normalized by the maximum number of possible moves in a game $\mu_{\max} = M_{\exp}N$ or $\mu_{\max} = M_{\text{sim}}N$, respectively.

replaced by a general ("low-level") principle which might serve as an explanation of emergent collective patterns. In the experiment, we found that the success of the biased minority players in leading the unbiased majority to their € € reward field was achieved by a specific locomotive pattern, i.e., moving their avatars analogously into the same direction, and making use of a first-mover effect [35,37,42,43].

In the cognitive force simulations, we were able to produce results similar to those in the experiment by applying the basic principle of maximization of options to the environmental setup. This combination of a mechanistic and a functional analysis [7] can contribute to a reality-based understanding of social behavior.

Interestingly, the RL-based simulations, despite following a similarly simple decision-making rule, did not yield results that match the experimental data as closely. Multiagent reinforcement learning is still an area of active research that suffers from instability in the learning algorithms. Our results do therefore not invalidate the underlying decision-making rule as a model for human behavior in its entirety.

In summary, we apply the general principle of option maximization [20,28,29,44,45] to explain human group coordination and successfully validate it by comparison to data obtained from human experiments. Due to the generality of the principle, it constitutes a promising approach for explaining a wide range of group behavioral phenomena.

ACKNOWLEDGMENTS

MGM is grateful to the Max Planck Society for funding. We thank Stephan Herminghaus for useful discussions.

[1] L. Conradt and C. List, Group decisions in humans and animals: A survey, Philos. Trans. R. Soc. B **364**, 719 (2009).

[2] F. A. Heller, *Decision Making and Leadership* (Cambridge University Press, Cambridge, UK, 1992).

[3] J. R. G. Dyer, A. Johansson, D. Helbing, I. D. Couzin, and J. Krause, Leadership, consensus decision making and collective behaviour in humans, Philos. Trans. R. Soc. B **364**, 781 (2009).

[4] R. P. Mann, Collective decision making by rational individuals, Proc. Natl. Acad. Sci. USA **115**, 10387 (2018).

[5] R. P. Mann, Collective decision-making by rational agents with differing preferences, Proc. Natl. Acad. Sci. USA **117**, 10388 (2020).

[6] J. Freeman, J. A. Baggio, and T. R. Coyle, Social and general intelligence improves collective action in a common pool resource system, Proc. Natl. Acad. Sci. USA **117**, 7712 (2020).

[7] H. A. Hofmann, A. K. Beery, D. T. Blumstein, I. D. Couzin, R. L. Earley, L. D. Hayes, P. L. Hurd, E. A. Lacey, S. M. Phelps, N. G. Solomon *et al.*, An evolutionary framework for studying mechanisms of social behavior, Trends Ecol. Evol. **29**, 581 (2014).

[8] A. Tavoni, A. Dannenberg, G. Kallis, and A. Löschel, Inequality, communication, and the avoidance of disastrous climate change in a public goods game, Proc. Natl. Acad. Sci. USA **108**, 11825 (2011).

[9] I. D. Couzin, J. Krause, N. R. Franks, and S. A. Levin, Effective leadership and decision-making in animal groups on the move, Nature (London) **433**, 513 (2005).

[10] H. Shirado and N. A. Christakis, Locally noisy autonomous agents improve global human coordination in network experiments, Nature (London) **545**, 370 (2017).

[11] D. Helbing and P. Molnar, Social force model for pedestrian dynamics, Phys. Rev. E **51**, 4282 (1995).

[12] A. Johansson, D. Helbing, and P. K. Shukla, Specification of the social force pedestrian model by evolutionary adjustment to video tracking data, Adv. Complex Syst. **10**, 271 (2007).

[13] M. Moussaïd, D. Helbing, S. Garnier, A. Johansson, M. Combe, and G. Theraulaz, Experimental study of the behavioural mechanisms underlying self-organization in human crowds, Proc. R. Soc. B **276**, 2755 (2009).

[14] M. Moussaïd, D. Helbing, and G. Theraulaz, How simple rules determine pedestrian behavior and crowd disasters, Proc. Natl. Acad. Sci. USA **108**, 6884 (2011).

[15] U. Lopez, J. Gautrais, I. D. Couzin, and G. Theraulaz, From behavioural analyses to models of collective motion in fish schools, Interface Focus **2**, 693 (2012).

[16] H. J. Charlesworth and M. S. Turner, Intrinsically motivated collective motion, Proc. Natl. Acad. Sci. USA **116**, 15362 (2019).

[17] A. Pérez-Escudero and G. G. de Polavieja, Collective animal behavior from Bayesian estimation and probability matching, PLOS Comput. Biol. **7**, 1 (2011).

[18] S. Arganda, A. Pérez-Escudero, and G. G. de Polavieja, A common rule for decision making in animal collectives across species, Proc. Natl. Acad. Sci. USA **109**, 20508 (2012).

[19] A. Pérez-Escudero and G. G. de Polavieja, Adversity magnifies the importance of social information in decision-making, J. R. Soc. Interface **14**, 20170748 (2017).

[20] H. Hornischer, S. Herminghaus, and M. G. Mazza, Structural transition in the collective behavior of cognitive agents, Sci. Rep. **9**, 12477 (2019).

[21] O. Morgenstern and J. Von Neumann, *Theory of Games and Economic Behavior* (Princeton University Press, Princeton, New Jersey, USA, 1953).

[22] K. Friston, The free-energy principle: A unified brain theory? Nat. Rev. Neurosci. **11**, 127 (2010).

[23] A. D. Wissner-Gross and C. E. Freer, Causal Entropic Forces, Phys. Rev. Lett. **110**, 168702 (2013).

[24] R. P. Mann and R. Garnett, The entropic basis of collective behaviour, J. R. Soc. Interface **12**, 20150037 (2015).

[25] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, Massachusetts, USA, 2018).

[26] J. Harte, T. Zillio, E. Conlisk, and A. B. Smith, Maximum entropy and the state-variable approach to macroecology, Ecology **89**, 2700 (2008).

[27] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, Statistical mechanics for natural flocks of birds, Proc. Natl. Acad. Sci. USA **109**, 4786 (2012).

[28] H. Von Foerster, *Observing Systems* (Intersystems Publications, Salinas, California, USA, 1984).

[29] C. Salge, C. Glackin, and D. Polani, Empowerment: An introduction, in *Guided Self-Organization: Inception* (Springer, Berlin, Heidelberg, 2014), pp. 67–114.

[30] T. Parr, L. Da Costa, and K. Friston, Markov blankets, information geometry and stochastic thermodynamics, Philos. Trans. R. Soc. A **378**, 20190159 (2020).

[31] A. Prasad and I. Dusparic, Multi-agent deep reinforcement learning for zero energy communities, in *2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)* (IEEE, Bucharest, Romania, 2019), pp. 1–5.

[32] X. Fang, J. Wang, G. Song, Y. Han, Q. Zhao, and Z. Cao, Multi-agent reinforcement learning approach for residential microgrid energy scheduling, Energies **13**, 123 (2020).

[33] M. Roesch, C. Linder, R. Zimmermann, A. Rudolf, A. Hohmann, and G. Reinhart, Smart grid for industry using multi-agent reinforcement learning, Appl. Sci. **10**, 6900 (2020).

[34] M. Boos, J. Pritz, and M. Belz, The honeycomb paradigm for research on collective human behavior, J. Vis. Exp. **143**, e58719 (2019).

[35] M. Boos, J. Pritz, S. Lange, and M. Belz, Leadership in moving human groups, PLoS Comput. Biol. **10**, e1003541 (2014).

[36] J. R. G. Dyer, C. C. Ioannou, L. J. Morrell, D. P. Croft, I. D. Couzin, D. A. Waters, and J. Krause, Consensus decision making in human crowds, Anim. Behav. **75**, 461 (2008).

[37] O. Petit, J. Gautrais, J.-B. Leca, G. Theraulaz, and J.-L. Deneubourg, Collective decision-making in white-faced capuchin monkeys, Proc. R. Soc. B **276**, 3495 (2009).

[38] S. Iqbal and F. Sha, Actor-attention-critic for multi-agent reinforcement learning, in *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 97, edited by K. Chaudhuri and R. Salakhutdinov (PMLR, Long Beach, California, 2019), pp. 2961–2970.

[39] M. L. Littman, Markov games as a framework for multi-agent reinforcement learning, in *Proceedings of the Eleventh International Conference on International Conference on Machine Learning*, ICML'94 (Morgan Kaufmann Publishers Inc., San Francisco, 1994), p. 157–163.

[40] K. Zhang, Z. Yang, and T. Başar, Multi-agent reinforcement learning: A selective overview of theories and algorithms, in *Handbook of Reinforcement Learning and Control*, edited by K. G. Vamvoudakis, Y. Wan, F. L. Lewis, and D. Cansever (Springer International Publishing, Cham, 2021), pp. 321–384.

[41] The complete set of hyperparameters used in our experiments can be found in the Supplemental Material [46], and the code of our environment can be found at https://github.com/pauljpritz/HoneyComb-Python-Environment.

[42] C. Sueur and O. Petit, Signals use by leaders in Macaca tonkeana and Macaca mulatta: Group-mate recruitment and behaviour monitoring, Anim. Cognition **13**, 239 (2010).

[43] O. Petit and R. Bon, Decision-making processes: The case of collective movements, Behav. Proc. **84**, 635 (2010).

[44] S. H. Cerezo and G. D. Ballester, Fractal AI: A fragile theory of intelligence, arXiv:1803.05049.

[45] H. Hornischer, J. C. Varughese, R. Thenius, F. Wotawa, M. Füllsack, and T. Schmickl, Cimax: Collective information maximization in robotic swarms using local communication, Adapt. Behav. **29**, 297 (2021).

[46] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevResearch.4.023037 for more information on the parameters used in the experiments and both numerical approaches. We also assess how the cognitive force model results depend on the choice of parameters.