



OPEN

Collective foraging of active particles trained by reinforcement learning

Robert C. Löffler¹, Emanuele Panizon² & Clemens Bechinger^{1,3}✉

Collective self-organization of animal groups is a recurring phenomenon in nature which has attracted a lot of attention in natural and social sciences. To understand how collective motion can be achieved without the presence of an external control, social interactions have been considered which regulate the motion and orientation of neighbors relative to each other. Here, we want to understand the motivation and possible reasons behind the emergence of such interaction rules using an experimental model system of light-responsive active colloidal particles (APs). Via reinforcement learning (RL), the motion of particles is optimized regarding their foraging behavior in presence of randomly appearing food sources. Although RL maximizes the rewards of single APs, we observe the emergence of collective behaviors within the particle group. The advantage of such collective strategy in context of foraging is to compensate lack of local information which strongly increases the robustness of the resulting policy. Our results demonstrate that collective behavior may not only result on the optimization of behaviors on the group level but may also arise from maximizing the benefit of individuals. Apart from a better understanding of collective behaviors in natural systems, these results may also be useful in context of the design of autonomous robotic systems.

The self-organization of organisms into functional collective groups is one of the most remarkable examples of how dynamical spatio-temporal patterns can be achieved by only local interaction rules without external control. The abundance of such collective behaviors in many living systems such as birds¹, fish², insects^{3,4} and bacteria⁵ suggests system-independent overarching organization principles. It has been demonstrated that collective behaviors can be understood in terms of so-called social interaction rules which control local alignment with and attraction towards neighbouring peers^{6–8}. Even though such framework is able to reproduce flocking, milling and swarming behaviors, they do not provide an answer why individuals follow such rules. Opposed to a priori motional rules, collective behaviors can be also understood by asking for the specific goals individuals and how they are reached by specific motional behaviors. Examples for such goals are related to foraging⁹, heat preservation¹⁰ and anti-predation^{11–13}. Notably, even when such goals are defined only on the level of individuals, this may lead to advantages for the entire group¹³. The implementation of specific goals into a corresponding framework can be achieved by multi-agent reinforcement learning (MARL), where motional rules of individuals are varied according to a rewarding scheme to an achieve optimal behavior (policy) regarding a given goal^{14,15}. This approach has successfully used in computer simulations to investigate e.g. the efficiency of animal flocks^{16,17}, cooperative foraging strategies¹⁸ and predator avoidance^{19–21}. In addition to numerical simulations, recently MARL has been also experimentally applied to synthetic systems of active colloidal particles (APs) which mimic many aspects of living systems²².

Here, we present an experimental study where we investigate the optimal foraging strategy of 30 APs in presence of a randomly appearing food source. Opposed to previous studies where a priori motional rules have been applied to groups of APs, here we only define a specific task (foraging) which will be accomplished by the group within a MARL framework and without knowing the strategy beforehand. Another advantage of this approach is that system-specific interactions are automatically considered and only little knowledge regarding the details of interactions between APs is required. We demonstrate that the quest of APs to optimize their individual foraging strategy leads to a collective milling motion, similar as observed in living systems^{8,23}. Once particles have been trained towards their optimal strategy, a milling behavior is maintained even in absence of food. This suggests the robustness of an optimal strategy regarding variations in the environment.

Konstanzer Online-Publikations-System (KOPS)

URL: <http://nbn-resolving.de/urn:nbn:de:bsz:352-2-bqxlft012te0>

¹Fachbereich Physik, Universität Konstanz, 78464 Konstanz, Germany. ²The Abdus Salam International Centre for Theoretical Physics (ICTP), Strada Costiera 11, 34151 Trieste, Italy. ³Centre for the Advanced Study of Collective Behaviour, Universität Konstanz, 78464 Konstanz, Germany. ✉email: clemens.bechinger@uni-konstanz.de

Experimental system

In our experiments we have used light-responsive APs which are made from transparent silica particles (diameter $\sigma = 6.3 \mu\text{m}$) being coated on one side with a 80 nm light absorbing carbon cap. They are suspended in a water-lutidine mixture contained in a thin sample cell whose temperature is kept below the mixture's lower demixing point $T_c \approx 34 \text{ }^\circ\text{C}$. Due to gravity, the APs settle to the bottom plate of the sample cell where they perform a two-dimensional motion. When illuminated with a focused laser beam, the caps are selectively heated above T_c which results in a temperature gradient which leads to local demixing and eventually self-propulsion of APs²⁴. By scanning the laser beam across all particles and individually adjusting its intensity and position relative to the carbon cap, the magnitude and the direction of the propulsion velocity of APs can be controlled independently (Methods).

Because the positions and orientations of all particles are continuously determined via real-time tracking, the local configuration of all particles are permanently recorded. This information can be allocated to each AP to gain knowledge regarding its environment. To quantify the position and orientation of peers and the location of the food source, each AP virtually senses its environment by a vision cone which covers 180° (aligned with the AP orientation) which is divided in five equal sections (Fig. 1a). For each section, the particle determines the density and mean orientation of neighbors and the imaged fraction of the food source which has a diameter of $80 \mu\text{m}$ (Fig. 1a, orange region). To yield a realistic visual perception, we have chosen a metric perception model where objects in the environment contribute with their inverse distance. Such signal decay is motivated by its established role in the swarming of insects²⁵. To make this model even more realistic, we have considered visual obstruction effects due to the finite size of the APs (Fig. 1a, gray areas; Methods).

At each instance of time, every AP (agent) chooses one of three possible motional actions, depending on the instantaneous visual cues described above. These actions are: (i) move straight forward, (ii) turn left and (iii) turn right respectively. Even though such action space appears to be rather simple, it resembles the discrete motional behavior of several bacteria²⁶. During the turning motions, APs also exhibit a forward motion which results in a radius of curvature of about $10 \mu\text{m}$. The choice of a specific action is determined via an artificial neural network (ANN) which delivers the "policy", i.e. the probabilities for selecting one of the above actions (Fig. 1b). This ANN, called "actor", is optimized by the framework of clipped proximal policy optimization (PPO, Methods) to maximize the sum of future rewards of each agent, known as return, which is the primary optimization objective in reinforcement learning. In our study, the return $G = \sum_t \gamma^t R_{t+1}$ is discounted by $\gamma = 0.97$. To optimize the policy towards a given task (goal), one has to define the instantaneous reward. In our specific example of foraging the reward R_t is defined to be positive when the AP's center of mass is within the circular area of the food source (for a precise reward definition we refer to the Methods). To reproduce the consumption of food, the capacity of the food source decreases depending on the time and number of APs within this region. When the food source is exhausted, another food source appears at a new randomly chosen location within the experimental field of view (Methods).

Results

In general, each experiment starts with training sequences, where the ANN is initialized with random weights which leads to random actions of the APs. Over time, agents learn which actions are leading to an increased return depending on the observables. Such training is conducted until the weights within the ANN converge to an optimized policy. In the case of our experiments, such training takes about 60 h of measurement time. Figure 2a shows an example of the APs trajectories resulting from a learned policy when moving from a depleted food source on the left (dashed circle) to another one appearing on the right (solid circle). The brightness of

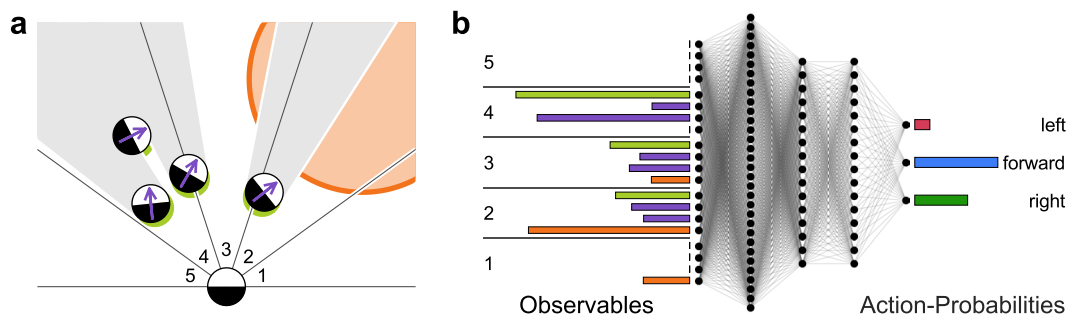


Figure 1. Multi-agent reinforcement learning. **(a)** APs retrieve information about their local environment by visual perception. A 180° vision cone is divided into five segments for each of which particles observe the density of neighboring particles (lime green), their mean orientation (purple) and the food source (orange). The observation strength is determined by the sum of imaged fractions of objects weighted with inverse distance (Methods). Neighboring particles also obstruct vision towards objects further away (gray areas). **(b)** The resulting set of 20 observables for neighbor presence, neighbor orientation (as two-element vector) and food source serves as input to the neural network (policy), which is modeled as a dense ANN with three hidden layers. The output of the policy represents the probability distribution for an appropriate action, being either left turn (red), forward motion (blue) or right turn (green). The figure was created using MatLab (version 2022B), <https://www.mathworks.com/products/matlab.html>.

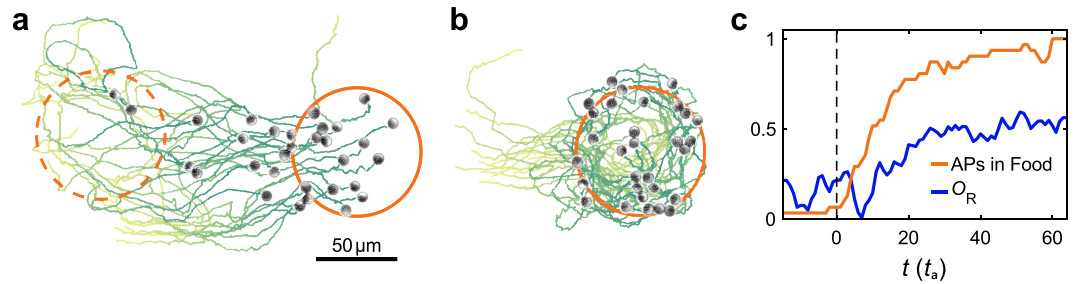


Figure 2. Collective behavior. (a, b) Experimental snapshots of trained APs, moving as a flock from one food source to the next and forming a milling state within the food source. Snapshots show microscope images of APs, annotated with trajectories during the last 40 actions (colored from bright to dark evolving in time) and previous (dashed orange) and current (solid orange) location of the food source (see also supplementary video S1). (c) Strength of milling characterized by order parameter O_R (blue line, see main text), which strongly increases with the fraction of APs located inside the food source (orange line) once the flock has arrived at the food source (dashed vertical line). The figure was created using MatLab (version 2022B), <https://www.mathworks.com/products/matlab.html>.

the trajectories (yellow green to dark green) indicates the evolution in time. As expected from the APs greedy strategy, they follow an almost direct path from the depleted to the new food source which results in high relative orientational alignment. Such behavior will be called a flocking state in the following. Once particles have arrived at the food source, their motion changes into a milling motion, i.e. rather circular trajectories within the food area (Fig. 2b, see also Suppl. Video S1). The group's milling motion can be quantified by a rotational order parameter

$$O_R = \frac{1}{N} \sum_i (\hat{r}_i \times \hat{u}_i) \cdot \hat{e}_z \quad (1)$$

where N is the number of particles, \hat{r}_i is a unit vector pointing from the group center to the i -th AP, \hat{u}_i is the unit vector denoting the APs orientation and \hat{e}_z is the unit vector perpendicular to the sample plane. Figure 2c shows the temporal evolution O_R together with the number of APs within the food source. In our experiments we observe that milling is maintained until the food becomes entirely depleted.

At first glance, the observation of a collective flocking and milling behavior is surprising because the reward (and the maximized return) is defined and optimized only on the level of single APs. To rationalize the flocking behavior, one has to consider that not all particles have a full, i.e. unobstructed view towards the food source due to particle sight-blocking by peers. To compensate for such lacking information, APs with zero (or limited) food perception can increase their chance of steering towards the food source by following their peers. This sounds a reasonable strategy because peers may have a better, i.e., more direct view towards the food. This idea is supported by Fig. 3a which shows the trajectories of APs moving towards a food source. The trajectories are labeled in red for those APs whose vision towards the food source is blocked by their peers. This applies to up

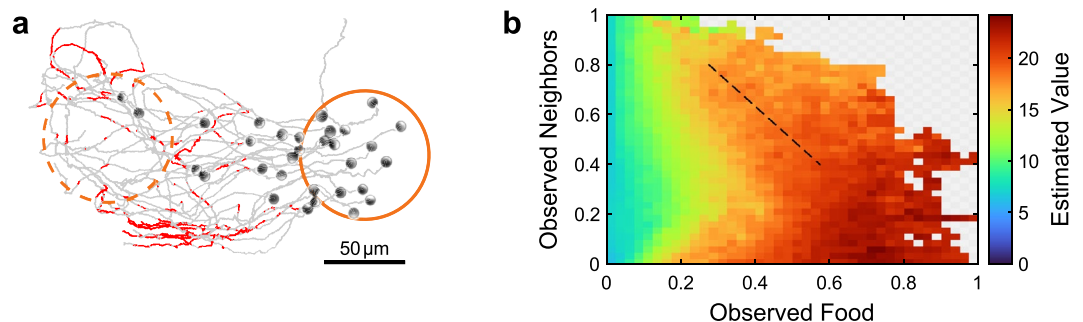


Figure 3. Vision obstruction and value function. (a) Experimentally measured trajectories (grey) of a group of trained APs which are moving towards a food source (orange circle). Red parts of the trajectories indicate instances where the vision of the corresponding AP towards the food source is blocked by peers. (b) Estimated return (value of state) in dependence of sum over neighbor density observables and sum of food observables, respectively. Although reward is only correlated to food, expected return is also correlated to neighboring particles (indicated by dashed line). All data is sampled from experimentally measured configurations, gray-checked background marks combinations of observables which did not occur in experiments. The figure was created using MatLab (version 2022B), <https://www.mathworks.com/products/matlab.html>.

to 10% of the APs in the tail of the group. As mentioned above, this lacking information is then compensated by following their peers which then leads to alignment, i.e. a flocking state.

The above arguments are quantitatively supported by the so-called value function (an auxiliary ANN which is part of the RL framework (Methods)) which provides an estimate for the return depending on a specific configuration. Figure 3b shows the measured estimated value obtained from trajectories as a function of the observables characterizing the amount of neighbors and food perceived by an individual. Obviously, this quantity is high for large perception of the food since under such conditions APs are getting closer to the rewarded food source. In addition, a proximity to peers also leads to an increase of the estimated value, even when the perception of the food source is low (see arrow). This must stem from instances where the APs are close to food but their vision is mostly blocked by other peers. APs then learn to align and follow peers—leading to flocking—driven by the realization the large perception of peers leads to higher gains.

After having explained why a flocking behavior is part of an ideal food-searching strategy, we now discuss the organization of the APs after they have reached the food source. Because the reward requires the permanently moving APs to be localized within the circular food source, this naturally leads to a milling motion of the group, in agreement with our experiments. Note, that a single sense of rotation is randomly selected during the training process and maintained afterwards.

The milling behavior is also reflected in the spatial distribution of actions performed by the particles. Figure 4 shows the measured actions of APs (within their frame of reference) for a counter-clockwise group rotations within the food source. In fact, the steering direction of the APs is slightly offset from the group center which is important to create a tangential motional component. Such behavior is in excellent agreement with previously observed milling based on social interaction rules^{27,28}. Interestingly, formation of milling motion is already initiated from the beginning when individuals enter the food source. This is shown in Fig. 4b which shows the spatially resolved actions of a single particle (i.e. without perception of peers) near the food source obtained from the optimized policy. Such conditions typically apply to the leading particles of the flock approach the food. As seen by the blue region (forward motion), the AP only moves straight towards the food source when located to the right hand side of the food source. Otherwise APs reorient accordingly, to enter the food always from the right. Thus behavior is ideal for the development of counter-clockwise milling.

As shown above, the optimal policy of the above discussed food-searching problem leads to a milling behavior within the food source. Notably, the identical policy leads to a milling behavior even in absence of food. Figure 5a shows the resulting trajectories of a group of APs which has been released from a hexagonal initial positional configuration with random orientation. Even though the food perception of all APs is set to zero, the particles show a very similar milling as reported in Fig. 2b (see also Suppl. Video S2). The corresponding rotational order parameter can reach values up to 0.6 which is similar to the value within a food source (Fig. 5b). In contrast to Fig. 2b, here the particle density is reduced near the center of rotation. This behavior can be rationalized by considering the above mentioned orientation alignment of APs with no view towards the food which is enforced by the policy. In absence of food, this results in a 'circular' flock where each particles follows its peers in front.

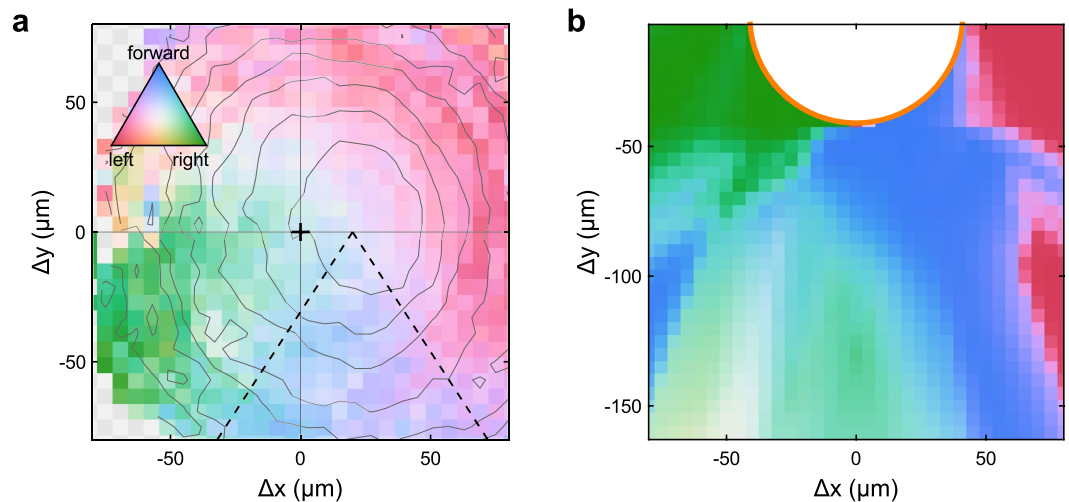


Figure 4. Policy map. Probabilities for an individual AP to choose one of the three actions (forward, left turn, right turn) in respect to its current configuration. **(a)** Action probabilities for a north-facing AP depending on its position relative to the center of mass of the particle group. Data is sampled from experimental trajectories, the gray contour lines indicate the number of occurrences of the respective configuration with most particles being found counter-clockwise aligned, i.e. to the right of the group center. The dashed lines serve as a guide to the eye. **(b)** Action probabilities for a single particle without perception of peers approaching a food source. The figure was created using MatLab (version 2022B), <https://www.mathworks.com/products/matlab.html>.

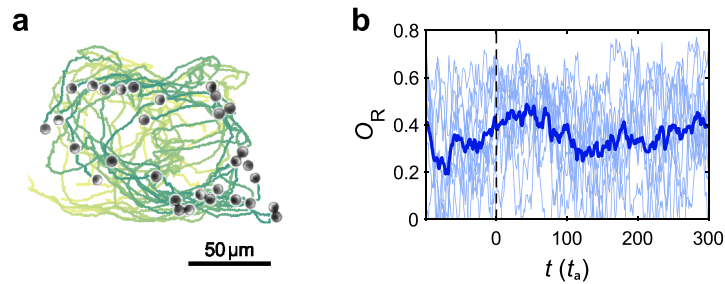


Figure 5. Behavior in the absence of food. (a) APs freely milling in the absence of food. The Snapshot shows a microscope image of APs, annotated with trajectories during the last 40 actions (colored from bright to dark evolving in time, see also Suppl. Video S2). (b) Orientational order parameter O_R for several time series (thin blue lines) and their average (blue line) for a scenario where particles mill inside a food source for $t < 0$, but also keep a stable milling formation after the food source is removed ($t > 0$).

Discussion

In our experiments we have investigated the ideal strategy of active particles to localize a randomly appearing finite food source using MARL. After training is complete, the optimized policy leads APs to move towards the food source in a aligned flock. Once the food source is reached, a milling motion develops within the circular area of the food. Such milling motion appears to be rather robust and also occurs in absence of the food source. Often collective states are described as a result of social interaction rules where particles aim to adjust their motion to that of peers. In our work, we demonstrate that similar states can also result from a mere selfish motivation which in our case is the individual (not that of the group) food uptake which is rewarded in our policy. The main reason why collective states may arise even under selfish conditions is, that individuals benefit from considering the motion of their peers. In particular when relevant information to reach a desired goal is missing (e.g. by visual obstruction of neighbors), this can be partially compensated by adjusting the behavior to that of peers which may have more (or different) information regarding their surrounding. The fact, that the above behavior is observed in an experimental system demonstrates its robustness regarding thermal noise, hydrodynamic and steric interactions but also unavoidable variations in the properties of the APs. Such deviations from an ideal and monodisperse behavior which is typically not included in numerical simulations is certainly of importance in living but also artificial robotic collectives comprised of hundreds or even thousands of group members. In our case, phoretic interactions of APs being in immediate contact is known to deteriorate their ability to rapidly change their steering direction. Although close particle distances have not been penalized within our reward definition, after the training has been completed, the optimized policy strongly avoided particle collisions which is important to enhance the ability of individuals to suddenly change their motion in response to a newly appearing food source. As a possible extension of our study, one could consider food sources which are distinguished by their physical properties. One possibility could be to rely on physical, impenetrable objects, such as additional disks, which then would be “virtually depleted” (in post-processing) as the particles enter in contact with them. This, however, would require major experimental modifications beyond the scope of the current work.

Methods

Active particles

Light activated active particles are fabricated from commercially obtained $6.2 \mu\text{m}$ silica spheres which are capped with a layer of 80 nm Carbon on one hemisphere. They are then suspended in a thin sample cell in a critical binary mixture of water and lutidine which is kept close to its lower demixing point at $34 \text{ }^\circ\text{C}$. Upon illumination of the particles, the capped hemisphere is heated above the critical point, leading to local demixing of the fluid and, thus, self-propulsion of the particle^{24,29}. To enable individual steering of particles, a feedback-loop is used: Images of the sample are taken at a rate of 5Hz; live image-analysis and particle tracking are performed to provide particle trajectories to the reinforcement-learning algorithm; finally an acousto-optic deflector, scanning the particles at a rate of 10MHz is used to illuminate individual particles with a slightly defocused 532 nm laser beam. The beam waist in the particle plain is about $4 \mu\text{m}$. In order to apply active steering to the particles, the laser spot is either offset to the capped side of the particles for a stabilized forward motion, or two laser spots per particle with different intensities to either side are used to generate a heat gradient within the carbon cap and therefore anisotropic demixing, resulting in an active torque^{27,30}. Note that due to the weak intensities and the defocused beam used in the experiment, no optical forces are applied to the particles. Throughout the measurements we keep the number of particles constant to $N = 30$. To ensure that no particles get lost and no new particles diffuse into the measurement area, effective boundary conditions are applied by means of the feedback loop: Particles entering the field of view get propelled back out, while particles reaching the boundary of the measurement area from inside get rotated and propelled back in, creating effective reflective boundary conditions³⁰. Note that these actions overrule the RL policy and consequently, the trajectories of particles reorienting due to the boundary condition are not included when training the policy. Before the start of each measurement, particles are propelled to homogeneously spaced starting positions at the center of the measurement area.

Reward and observables definition

The virtual food source used for the reward definition has a fixed diameter of 80 μm . The capacity is set to 1000 rewards, which equals to an average depletion time of about 15 min. Upon depletion, the food source is relocated to a new random location within the available experimental accessible space of about 300 μm by 400 μm , with a minimum distance to the boundary and at least 120 μm away from the last location. Particles at position \vec{r}_i are rewarded strictly if their center is within the food source,

$$R_{i,t} = \begin{cases} 1 & (|\vec{r}_i - \vec{r}_{\text{food}}| \leq 80 \mu\text{m}) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The observables $\vec{O}_{i,t}$ which represent the visual input of the agent i at time step t are calculated per section m of the vision cone in respect to particle orientation θ_i , as

$$\vec{O}_{i,t} = (\vec{f}_{i,t}, \vec{p}_{i,t}^{\text{d}}, \vec{p}_{i,t}^{\text{sin}}, \vec{p}_{i,t}^{\text{cos}}), \quad (3)$$

$$f_{i,t,m} = \min\left(\frac{\sigma_{\text{food}}}{|\vec{r}_i - \vec{r}_{\text{food}}|}, 1\right) g_{i,\text{food},t,m}, \quad (4)$$

$$p_{i,t,m}^{\text{d}} = \sum_j \frac{\sigma_{\text{part}}}{|\vec{r}_i - \vec{r}_j|} g_{i,j,t,m}, \quad (5)$$

$$p_{i,t,m}^{\text{sin}} = \sum_j \sin(\theta_i - \theta_j) \frac{\sigma_{\text{part}}}{|\vec{r}_i - \vec{r}_j|} g_{i,j,t,m}, \quad (6)$$

$$p_{i,t,m}^{\text{cos}} = \sum_j \cos(\theta_i - \theta_j) \frac{\sigma_{\text{part}}}{|\vec{r}_i - \vec{r}_j|} g_{i,j,t,m}, \quad (7)$$

where σ is the food/particle diameter and $g_{i,j,t,m}$ denotes the relative amount of particle j being visible within section m of particle i 's vision cone (see Fig. 1a). Note, that \vec{p}_i can not diverge, as particles cannot get closer than σ .

The multi agent reinforcement learning (MARL) algorithm

A natural theoretical framework to study how collective motion emerges from the solution of tasks is that of Multi Agent Reinforcement Learning. While a comprehensive overview of it is outside the scope of this work (see¹⁵ for a recent review), here we provide the basic information relevant for our study. In our setup agents perform actions in response to environmental cues, the observables $\vec{O}_{i,t}$ defined above. Actions are consequently drawn from the policies $a_{i,t} \sim \pi_i(a | \vec{O}_{i,t})$, which encode the behaviors of the agents. Here we employ the centralized-training, decentralized-execution paradigm¹⁴, where policies are shared, i.e. $\pi_i = \pi_j = \pi$. While the policy is shared, agents do not communicate and are fully independent in the choice of action, which depends only on the individual local environment (through the observables). The agent's actions therefore can be different at any time.

Through experience, the agents optimize their policies to a specific goal objective, defined through the reward functions. The past experience is stored in the "trajectories", which are the temporal series of observations, actions and rewards obtained by each agent during a whole episode. The process of optimization is done through stochastic policy-gradient algorithm: The agents' policies are learned using the Proximal Policy Optimization (PPO) actor-critic learning algorithm³¹, together with the generalized advantage estimation (GAE)³². The two artificial neural networks (ANNs) corresponding to the "actor" (i.e. the policy π) and "critic", share the same input and hidden layers, with the input having 20 nodes corresponding to the observable vector \vec{O}_i (see "Reward and observables definition"). The hidden layers are dense layers with 32, 16 and 16 nodes respectively and a ReLU non-linear activation function. The output layer of the critic is a scalar, while the actor ANN has a final layer of the size of the number of actions, with a softmax activation. Both ANNs are initialized with random weights. The parameter for the reward discount is set to $\gamma = 0.97$, and that for the estimator of the GAE as $\lambda = 0.97$. For a more detailed description of the code and algorithm used, see the Methods in¹³. While here we use an experimental realizations of the system to obtain trajectories of data, the algorithmic part is effectively equivalent.

The global state of the system at any given time is captured by the imaging and tracking algorithm in the feedback loop. The difference between states of consecutive time steps is given by the physical evolution of the experimental system. As the RL algorithm requires a significant difference between consecutive states in order to adjust the policy in a meaningful way, actions are drawn at a lower frequency than the steering feedback loop. Namely, new actions $a_{i,t}$ are assigned to all particles every 10 seconds based on their local observables $\vec{O}_i(s_t)$ and kept constant until the next update. This corresponds to a forward motion of approximately 1σ per action and a possible rotation of approximately 22 °C per action. If a particle reaches the boundary of the measurement area, it is considered lost to the RL framework and its trajectory is ended. After it is properly reoriented and swimming back into the measurement area, a new trajectory is started, such that the RL algorithm can not exploit the boundary condition for its strategy. Note that, while this situation is common in early training, due to the general attraction towards the food source, events of particles reaching the defined boundary area become less likely towards the end of policy optimization. Every 120 time steps (corresponding to 20 minutes of experimental time) all recorded trajectories are evaluated and the gathered experience is used to improve the policy.

It is important to note that the RL algorithm optimizes the policy in respect to a single particle to maximize reward gained over time. This particularly means that in some circumstances an action might be preferred which does not maximize the reward in the immediate next step, but only over the long term. In respect to the given task, it is not only important for a particle to stay inside the current food source, but also to be able to reach the next food location as fast as possible when the current one is depleted.

Data availability

The datasets generated and analysed during the current study are available from the corresponding author on reasonable request.

Received: 28 June 2023; Accepted: 5 October 2023

Published online: 10 October 2023

References

- Cavagna, A. *et al.* Scale-free correlations in starling flocks. *Proc. Natl. Acad. Sci.* **107**, 11865–11870. <https://doi.org/10.1073/pnas.1005766107> (2010).
- Parrish, J. K., Viscido, S. V. & Grünbaum, D. Self-organized fish schools: An examination of emergent properties. *Biol. Bull.* **202**, 296–305. <https://doi.org/10.2307/1543482> (2002).
- Buhl, J. *et al.* From disorder to order in marching locusts. *Science* **312**, 1402–1406. <https://doi.org/10.1126/science.1125142> (2006).
- Cavagna, A. *et al.* Dynamic scaling in natural swarms. *Nat. Phys.* **13**, 914–918. <https://doi.org/10.1038/nphys4153> (2017).
- Czirók, A., Ben-Jacob, E., Cohen, I. & Vicsek, T. Formation of complex bacterial colonies via self-generated vortices. *Phys. Rev. E* **54**, 1791–1801. <https://doi.org/10.1103/physreve.54.1791> (1996).
- Couzin, I. D., Krause, J., James, R., Ruxton, G. D. & Franks, N. R. Collective memory and spatial sorting in animal groups. *J. Theor. Biol.* **218**, 1–11. <https://doi.org/10.1006/jtbi.2002.3065> (2002).
- Sumpter, D. J. T., Buhl, J., Biro, D. & Couzin, I. Information transfer in moving animal groups. *Theory Biosci.* **127**, 177–186. <https://doi.org/10.1007/s12064-008-0040-1> (2008).
- Vicsek, T. & Zafeiris, A. Collective motion. *Phys. Rep.* **517**, 71–140. <https://doi.org/10.1016/j.physrep.2012.03.004> (2012).
- Detrain, C. & Deneubourg, J.-L. Collective decision-making and foraging patterns in ants and honeybees. *Adv. Insect Physiol.* **35**, 123–173. [https://doi.org/10.1016/S0065-2806\(08\)00002-7](https://doi.org/10.1016/S0065-2806(08)00002-7) (2008).
- Gilbert, C., Blanc, S., Le Maho, Y. & Ancel, A. Energy saving processes in huddling emperor penguins: From experiments to theory. *J. Exp. Biol.* **211**, 1–8. <https://doi.org/10.1242/jeb.005785> (2008).
- Krause, J. & Tegeeder, R. W. The mechanism of aggregation behaviour in fish shoals: Individuals minimize approach time to neighbours. *Anim. Behav.* **48**, 353–359. <https://doi.org/10.1006/anbe.1994.1248> (1994).
- King, A. J. *et al.* Selfish-herd behaviour of sheep under threat. *Curr. Biol.* **22**, R561–R562. <https://doi.org/10.1016/j.cub.2012.05.008> (2012).
- Monter, S., Heuthe, V.-L., Panizon, E. & Bechinger, C. Dynamics and risk sharing in groups of selfish individuals. *J. Theor. Biol.* **562**, 111433. <https://doi.org/10.1016/j.jtbi.2023.111433> (2023).
- Gupta, J. K., Egorov, M. & Kochenderfer, M. Cooperative multi-agent control using deep reinforcement learning. In *Autonomous Agents and Multiagent Systems* 66–83 (Springer International Publishing, Berlin, 2017). https://doi.org/10.1007/978-3-319-71682-4_5.
- Zhang, K., Yang, Z. & Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. In *Handbook of Reinforcement Learning and Control* 321–384 (Springer International Publishing, Berlin, 2021). https://doi.org/10.1007/978-3-030-60990-0_12.
- Verma, S., Novati, G. & Koumoutsakos, P. Efficient collective swimming by harnessing vortices through deep reinforcement learning. *Proc. Natl. Acad. Sci.* **115**, 5849–5854. <https://doi.org/10.1073/pnas.1800923115> (2018).
- Durve, M., Peruani, F. & Celani, A. Learning to flock through reinforcement. *Phys. Rev. E* **102**, 012601. <https://doi.org/10.1103/physreve.102.012601> (2020).
- López-Incera, A., Ried, K., Müller, T. & Briegel, H. J. Development of swarm behavior in artificial learning agents that adapt to different foraging environments. *PLoS ONE* **15**, e0243628. <https://doi.org/10.1371/journal.pone.0243628> (2020).
- Hahn, C., Phan, T., Gabor, T., Belzner, L. & Linnhoff-Popien, C. Emergent escape-based flocking behavior using multi-agent reinforcement learning. In *The 2019 Conference on Artificial Life* (MIT Press, 2019). https://doi.org/10.1162/isal_a_00226.
- Sunehag, P. *et al.* Reinforcement learning agents acquire flocking and symbiotic behaviour in simulated ecosystems. In *The 2019 Conference on Artificial Life* (MIT Press, 2019). https://doi.org/10.1162/isal_a_00148.
- Young, Z. & La, H. M. Consensus, cooperative learning, and flocking for multiagent predator avoidance. *Int. J. Adv. Robot. Syst.* **17**, 172988142096034. <https://doi.org/10.1177/1729881420960342> (2020).
- Muñoz-Landin, S., Fischer, A., Holubec, V. & Cichos, F. Reinforcement learning with artificial microswimmers. *Sci. Robot.* <https://doi.org/10.1126/scirobotics.abd9285> (2021).
- Delcourt, J., Bode, N. W. F. & Denoël, M. Collective vortex behaviors: Diversity, proximate, and ultimate causes of circular animal group movements. *Q. Rev. Biol.* **91**, 1–24. <https://doi.org/10.1086/685301> (2016).
- Gomez-Solano, J. R. *et al.* Tuning the motility and directionality of self-propelled colloids. *Sci. Rep.* <https://doi.org/10.1038/s41598-017-14126-0> (2017).
- Attanasi, A. *et al.* Collective behaviour without collective order in wild swarms of midges. *PLoS Comput. Biol.* **10**, e1003697. <https://doi.org/10.1371/journal.pcbi.1003697> (2014).
- Wadhwa, N. & Berg, H. C. Bacterial motility: Machinery and mechanisms. *Nat. Rev. Microbiol.* **20**, 161–173. <https://doi.org/10.1038/s41579-021-00626-4> (2021).
- Bäuerle, T., Löffler, R. C. & Bechinger, C. Formation of stable and responsive collective states in suspensions of active colloids. *Nat. Commun.* **11**, 2547. <https://doi.org/10.1038/s41467-020-16161-4> (2020).
- Löffler, R. C., Bäuerle, T., Kardar, M., Rohwer, C. M. & Bechinger, C. Behavior-dependent critical dynamics in collective states of active particles. *Europhys. Lett.* **134**, 64001. <https://doi.org/10.1209/0295-5075/ac0c68> (2021).
- Lozano, C., ten Hagen, B., Löwen, H. & Bechinger, C. Phototaxis of synthetic microswimmers in optical landscapes. *Nat. Commun.* <https://doi.org/10.1038/ncomms12828> (2016).
- Bäuerle, T., Fischer, A., Speck, T. & Bechinger, C. Self-organization of active particles by quorum sensing rules. *Nat. Commun.* <https://doi.org/10.1038/s41467-018-05675-7> (2018).
- Schulman, J., Wolski, E., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. *arXiv:1707.06347* (2017).
- Schulman, J., Moritz, P., Levine, S., Jordan, M. & Abbeel, P. High-dimensional continuous control using generalized advantage estimation. *arXiv:1506.02438* (2015).

Acknowledgements

We acknowledge fruitful discussions with Veit-Lorenz Heuthe. This work is financially supported by the DFG Centre of Excellence 2117 'Centre for the Advances Study of Collective Behaviour' (ID: 422037984) and the ERC Advanced Grant ASCIR (No. 693683).

Author contributions

C.B, R.C.L. and E.P. have conceived the experiments and implemented the computational framework, R.C.L. conducted the experiments and analyzed the results. All authors have contributed to the writing of the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-44268-3>.

Correspondence and requests for materials should be addressed to C.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023