Why Do Animals Swirl and How Do They Group?

Egor E. Nuzhin

Skolkovo Institute of Science and Technology

Maxim E. Panov

Skolkovo Institute of Science and Technology

Nikolai V. Brilliantov (✉ nb144@leicester.ac.uk)

University of Leicester

Research Article

Keywords: animals, biological function, swirling, reinforcement learning

Posted Date: April 13th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-401688/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Why do animals swirl and how do they group?

Egor E. Nuzhin¹, Maxim E. Panov¹ and Nikolai V. Brilliantov¹ ✉

¹*Skolkovo Institute of Science and Technology, 121205, Moscow, Russia.* ✉*e-mail: n.brilliantov@skoltech.ru and nb144@leicester.ac.uk*

We report a possible solution for the long-standing problem of the biological function of swirling motion, when a group of animals orbits a common center of the group. We exploit the hypothesis that learning processes in the nervous system of animals may be modelled by reinforcement learning (RL) and apply it to explain the phenomenon. In contrast to hardly justified models of physical interactions between animals, we propose a small set of rules to be learned by the agents, which results in swirling. The rules are extremely simple and thus applicable to animals with very limited level of information processing. We demonstrate that swirling may be understood in terms of the escort behavior, when an individual animal tries to reside within a certain distance from the swarm center. Moreover, we reveal the biological function of swirling motion: a trained for swirling swarm is by orders of magnitude more resistant to external perturbations, than an untrained one. Using our approach we analyze another class of a coordinated motion of animals – a group locomotion in viscous fluid. On a model example we demonstrate that RL provides an optimal disposition of coherently moving animals with a minimal dissipation of energy.

PACS numbers:

I. INTRODUCTION

Swirling is one of the most enigmatic phenomenon of the collective behavior of animals. The circular motion around a common center is observed in large groups of animals at different evolution stages ranging from insects and plant-animal worms to fish. The biological function of such motion is hardly understood and remains under debate till now [16, 18, 27, 30, 31, 38]. Moreover, even a realistic model of swirling is presently lacking. The existing approaches exploit artificial mechanical forces acting between animals. These interactions have a form of spring-like forces, gravitation-like forces, and forces of Morse potential [14, 15, 26, 29]. Certainly, such mechanical forces do not exist but serve to mimic an *intention* of moving animals to change their velocity. This intention is modeled in the form of the second Newton law, which describes the rate of change of animals' speed. The authors of Ref. [2] assumed that all particles move with a constant absolute velocity. The directions of the velocities vary in response to the torques originating due to interactions between the particles. Although the milling patterns arise in this model the functional form of such torques is rather artificial and may be hardly justified.

More realistic was the celebrated Vicsek model [11, 41]. Here an intention of a moving agent (animal) is explicitly formulated in the form of a *kinematic algorithm*. The algorithm dictates the change of the animals' velocity. It is assumed that all agents have a constant absolute velocity of varying direction. At each time step the direction is chosen as an average direction of motion of all neighbors, located within a certain distance from the agent. In spite of the simplicity, the model demonstrates a very rich behavior, predicting different modes of collective motion, pattern formation, like in systems of dissipative particles [8, 9] and even phase transitions [11, 12, 26, 34, 41].

Still Vicsek model is too simple to describe swirling motion, which neither arises spontaneously in this model, nor in the presence of a circularly moving leader; the reason for this is a fixed magnitude of the velocity [7]. Hence there is a challenge to propose an *intention*-based model with a simple kinematic algorithm resulting in swirling. Such an algorithm, based on the a priori knowledge about the agents, should be as realistic as possible. The aim of the present study is to propose a relevant intention-based model of collectively moving agents, which accounts for their perceptual and physical limitations. We expect that such a model will not only predict the emergence of the swirling motion, but also help to understand the biological function of this enigmatic phenomenon.

In a seminal paper [33], Reynolds mentioned that an adequate model should reflect a limited perception of animals performing coordinated motion (“fish in murky water”, “land animals that see only nearest herdmates”, etc. [33]). A problem of multi-agent connectedness under limited information access has been addressed in [22], where the existence of algorithms keeping the connectedness has been mathematically proven. This study demonstrates, however, that it is extremely difficult to formulate an explicit action algorithm for systems with a limited information access. Furthermore, the additional constraints, due to the physical limitations of the agents, make such a problem even harder and possibly unsolvable.

Nevertheless, nature finds the solution. Myriads of living beings – insects fish, etc. swirl. It seems extremely improbable that animals follow a sophisticated mathematical algorithm imprinted in their genes. More reasonable is to assume that animals learn to adopt their velocity – both the magnitude and the direction, by trial-and-error method, receiving a reward for a correct action and some form of punishment for an incorrect one. The reward and

punishment are regulated by internal chemical processes in animals' organisms. Hence, the most plausible assumption is that the response to an action (reward or punishment) and the action variability are the fundamental features of animals, imprinted in their genes. This is our main hypothesis, see also the discussion below.

Once we accept such a hypothesis we immediately recognize that the trial-and-error method, supplemented by a reward and penalty is in the heart of the reinforcement learning (RL) – one of the most powerful tools of machine learning (ML) techniques. This method has been successfully exploited for various transport problems of active particles [10, 32, 36]. The application of ML to communication problems, including animal communications has also demonstrated its efficiency [17, 19, 23, 37, 40, 42]. Moreover, the RL applied to biological systems gives a new insight into reward-based learning processes [28] and, as we show below, helps to reveal the biological function of swirling. In recent studies [20, 21], the RL has been exploited to solve the rendezvous and pursuit-evasion problem for communicating agents possessing only local (limited) information. The inverse RL, which allows to get an individual strategy of the agents yielding the observed collective behavior, was also proposed [21].

In the present article, we apply the RL approach to describe a collective motion in a swarm under constraints, which reflect a very limited perception and physical limitations of the agents, as dictated by their biological nature. Instead of applying the inverse RL, as suggested in [20, 21], we explicitly consider different individual action rules, that may stem from very limited abilities of the information processing of the animals. This reveals an interesting connection between the agents' strategy and informational completeness and/or their physical limitations.

As it follows from our analysis, the swirling motion may be understood as a specific form of an escort behavior, when an individual animal tries to remain within a certain distance from a swarm center. Therefore, to illustrate the basic ideas we start from the most simple escort problem, when only a few (in our case – a pair) of animals are involved. Such an escort behavior is observed mainly for mammals [6, 38] with a high information-processing abilities. For illustrative purposes, however, we consider the escort of animals with different levels of information processing.

The simplest one-to-one escort behavior may be formulated as follows: One animal – the "follower", tries to reside within a certain interval of distances from another, independently moving animal – the "leader". We demonstrate that depending on the degree of awareness and physical abilities, the follower may choose very different strategies. Some of these strategies look rather astonishing. Next we apply the same methodology for the collective motion, supplementing the rules of one-to-one escort motion by a few rules of collective motion in a large group of animals. We show that the swirling motion arises spontaneously and persists. Then we apply random external perturbation to a swarm and observe that the group of animals, "trained" to perform swirling, demonstrates much stronger resistance to the perturbations, than a non-trained group. This justifies our conjecture about the biological function of the swirling motion¹

To check whether the RL can find the best strategy, we investigate the optimal locomotion of a group of animals coherently moving in viscous fluid; they also experience physical interactions through the media. The optimal locomotion strategy may be found in this case from a straightforward solution of the coupled equations of motion for the agents. We demonstrate that the RL correctly reproduces all the results of the direct optimization.

II. RESULTS

A. One-to-one escort strategies

In the one-to-one escort a follower tries to reside within some range of distances from an independently moving leader. As we wish to model rather simple animals, we assume a very basic level of their perception. For instance, it is hard to believe that such simple beings can perceive an exact distance between themselves and other objects. At the same time, it is natural to assume the ability of a simpler perception – whether they reside within some distance range from one another. We call this interval of distances, which can be rather large, a "comfort zone", see Fig. 1a. The notion of the comfort zone will be later applied for swirling motion, see Fig. 1b and c.

Some additional information is needed to construct a trajectory of the follower (in what follows also the "agent"). We assume that the follower can perceive its own direction of motion and this of the leader and can also distinguish between approaching and moving off objects. Next, we consider three different scenarios: (A) The follower can perceive its absolute velocity, and there is no limitation for its acceleration; (B) the follower can perceive its absolute velocity and its acceleration is limited and (C) the follower does not perceive its absolute velocity but can perceive

¹ Note, that the dynamic programming (see e.g. [4]) may be an alternative way to find an algorithm yielding swirling motion under information-processing constraints. In application to living beings, however, the policy gradient RL approach seem to be more adequate, as it mimics natural learning processes.

FIG. 1: (a) Illustration of the escort problem – the leader moves independently on the outer circle, while the follower tries to reside within the comfort zone (shown shaded). (b) Illustration of the swirling motion in a swarm. It emerges spontaneously. The comfort zone w.r.t. the center of the swarm is shown (shaded). (c) Sketch of the swirling motion in a large colony of insects (ants); the individual comfort zone is zoomed. Similar patterns are observed in nature, see e.g. [1].

its acceleration, which is limited from above. The last scenario is the most realistic. Indeed, it is not easy to measure the velocity, however, even primitive beings can perceive the direction of motion and their acceleration, through muscle efforts or other biological mechanisms². The underlying physiology dictates the limits for the animals' acceleration.

Leaving the mathematical and algorithmic detail for the section Methods and Supplementary Material (SM), we discuss here the general ideas – how to implement the RL to the addressed problem. In the heart of the RL is the reward – the function of an agent action in a given system state [5, 28]. If the action brings the agent closer to the aim, the reward is positive, otherwise – negative. This is the same as the positive and negative stimuli in the nervous system of a living being. The efficiency of the action policy, is characterized by the average of the sum of all rewards (positive and negative) at all actions. The neural network is trained to choose the action policy that maximizes the reward and reaches in this way the desired goal. We wish to stress that the neural network and the according policy is associated with an individual agent. Moreover, we hypothesize that our RL-based model of the agent actions mimics the most prominent features of the real informational processes which determine the behavior of living beings. We investigate different trajectories of a leader – circle, ellipse, eight-curve, spiral, and triangle, which is not a smooth curve. For all these trajectories, the RL managed to train the network to develop successful strategies for the follower.

The results of the application of the RL to the escort problem are presented in Fig. 2. It is interesting to note that the optimal strategy drastically depends on the available information and physical limitations. Furthermore, the shape of the optimal trajectories of the follower is not necessarily smooth and sometimes looks very astonishing. For the case of “abundant” information, when the velocity of the follower is known, and there are no physical limitations (A-scenario), the follower reaches first the target distance to the leader and then applies the “frog strategy” of successive jumps: It waits until the leader moves far away and makes a long jump. In the new position, it waits again and then jumps, and so on, Fig. 2a. Noteworthy, the jerkily motion with non-smooth trajectories is not related to the discontinuity of the step-function, involved in the estimate of the reward. It persists for other smooth functions, see SM for more detail. When the limitation of the acceleration is imposed, but the agent still perceive its velocity (B-scenario) the follower changes the strategy and moves on smooth trajectories, Fig. 2b.

For the most realistic, C-scenario, when only acceleration is known (which is also limited), the follower starts to use the surprising wrapping trajectories, independently on the shape of the leader trajectory, Fig. 2c,d,e. In this way it is guaranteed that the follower always remains within a “comfort zone”, even if the information of the velocity and the distance to the leader is lacking. We have also considered an additional energy-saving condition for the C-scenario of the escort. The resulting optimized trajectory remains wrapping, although the wrapping circles become much smaller. The appearance of the curled trajectories for the most realistic C-scenario gives us a hint that these trajectories can transform into swirling for a multi-agents problem.

² Even plants “know” the direction of gravitation acceleration.

FIG. 2: (a) The typical trajectory of the follower after the training for A-scenario of escort – the follower can perceive its absolute velocity without a limitation for the acceleration. Note the “frog-strategy” of the follower (see the text). (b) The typical trajectory of the follower after the training for B-Scenario of escort – the follower can perceive its absolute velocity and its acceleration is limited. The follower moves on a relatively smooth trajectory. (c)-(e) The typical trajectory of the follower after the training for C-scenario of escort – the follower can perceive only its acceleration and the acceleration is limited. The trained follower moves on wrapping trajectories for all trajectories of the leader – circular (not shown), eight-curve (c), triangular (d) and elliptic (e).

B. The emergence of swirling

Now we address the collective motion (see Methods for mathematical and algorithmic details). As in Ref. [20] we assume that all animals (also “agents” here) are identical³ and follow the same action policy. We also assume that any information about an agent is not transmitted to any other agent. That is, n animals learn simultaneously and

³ That is, we do not have leader(s) and follower(s).

FIG. 3: (a) Swirling in a swarm of 100 agents. The large violet circle describes the target distance to the group center, white arrows indicate the speed direction, and the small shaded circles are the agents' comfort zones. The center of the swarm moves with a constant velocity. Inset: Time dependence of the average angular velocity Ω of the swarm (see Sec. III for the definition). Ω evolves from zero to non-zero steady-state value. (b) The dependence of the shift of a swarm center in the direction of the applied force as a function of the average external force. The results are shown for swarms trained for swirling and untrained swarms. The trained swarm (with swirling) can resist the external force up to 100 times more efficiently than the untrained one. The resistance of the trained swarm fades for very large external forces when they become comparable to the maximal force that an agent can exert. The base value marker indicates the reference value of the average force used in SM.

independently, receiving the reward individually. We believe that the application of the individual training describes more adequately the studied phenomenon and prevents a spurious information exchange between the agents.

To model the collective behavior, we supplement the individual perception rules, as for the escort problem by the collective rules. Namely the following information may be perceived and processed by an agent in a swarm: (i) whether the closest agent approaches or moves off; (ii) whether the closest agent is within a comfort zone; (iii) the direction of motion of the closest neighbor and of the self-motion; (iv) the acceleration of the agent, i.e., the exerted by the agent force; (v) whether the agent approaches or moves off from the center of a group; (vi) the direction of the entire group velocity; (vii) whether the agent resides within a target distance from the group center, which corresponds to the “comfort zone” with respect to (w.r.t.) the group center.

Note that while the rules (i)-(iv) describe the “one-to-one” escort motion, the rules (v)-(vii) describe the behavior in a group. This essentially corresponds to the escort of the group center by an individual animal. Indeed, the reward is given when an agent resides within the comfort zones, both w.r.t. to its nearest neighbor, as well as w.r.t. to the swarm center; otherwise, the agent is penalized. We wish to stress that animals possess a very fuzzy knowledge about the location of their neighbours and the swarm center – they perceive only whether they reside in the according comfort zone, which may be rather large.

Here we hypothesize that the rules (i) – (vii) have been imprinted in animals genes by evolution; they motivate all animals in a group to move in a way that fulfills the requested criteria. The agents choose an appropriate policy using (very limited) information at hand. We observe that starting with very different initial conditions, swirling motion of a swarm spontaneously emerges, see Fig. 3a. The swirling around a swarm center is commonly accompanied by a linear motion of the swarm as a whole. The swirling motion may be quantified by the average angular velocity of particles orbiting their common center of mass Ω (see the next Sec. III for the definition). As it may be seen from Fig. 3a, initially $\Omega = 0$. In the course of time the swarm self-synchronizes and a steady swirling with non-zero average angular velocity arises and persists.

C. The biological function of the swirling motion

Is it possible to understand the biological function of the swirling motion in swarms? We assume that the swirling helps to resist the external forces, which may jeopardize animals. For instance, it can be a wind that may blow an insects' swarm far from their inhabitation. To prove this conjecture, we conducted the following experiments. We incorporated an additional external force into the environment and measured the shift of the swarm center in the direction of the external force. During the training, we changed the direction of the external force with a uniform angular distribution. We also modulated the force strength with some periodicity, applying a stretched exponential distribution, widely used to describe natural phenomena, see e.g. [24, 35] and Sec. III for detail. We performed the averaging of the swarm shift (in the direction of the force) for different values of the average force. For each average

FIG. 4: Optimal configurations for the group of two, three, and four agents moving with the same velocity (see SM for the optimal configurations of larger swarms). The results are obtained by the direct minimization of the total dissipation power (see Methods). The average dissipated power per agent is indicated in units of the power dissipated by a solely moving agent. Note that optimal configuration for the triplet has two symmetric realizations. Also note that the larger the swarm, the smaller the specific dissipation.

force, we repeat our experiments ten times. We check that varying the parameters of the force distribution does not change the qualitative behavior of the system. Moreover, the behavior of the system persist when the size of the comfort zone or a number of agent in a swarm varies. More details are given in SM.

In Fig. 3b we plot the dependence of the shift of a swarm center on the average external force for the basic version of the stretched exponential distribution (see Sec. III); the results for the general case may be found in SM. We compare two groups of agents: The one has been trained to swirl, as described in the previous section, the other – untrained. As it may be seen from the figure, the group that performs swirling resists up to 100 times more efficiently than that without the swirling. The resistance fades for very strong external forces, comparable with the maximal force which may be exerted by an agent. Hence we come to an important, although surprising conclusion: The intention to move around a swarm center results in extremely high resistance to external perturbations. It seems astonishing that such a simple strategy helps living beings to cope with a hostile environment. In other words, we conclude that the enigmatic swirling motion is not at all a random arrangement or a behavioral error of a group of animals. In contrast, it plays a crucial role in their survival. Certainly, animals with a complex organization can learn more efficient strategies, but for very simple beasts, this strategy may be the most reliable to resist the environmental threats.

D. Optimal swarm locomotion

As we have illustrated above, the RL allows modeling complicated coordinated motion of swarms where the agents receive and process very restricted information about themselves and the system as a whole. Moreover, the RL allows to model systems without formulation of an explicit action strategy – the action policy is developed through the training with the use of an appropriately chosen reward function. Since the explicit action strategy was lacking for the case of swirling motion, it would be worth to check whether the RL provides a truly optimal strategy. Therefore we consider a problem where the optimal strategy is known beforehand and prove that the RL does find this strategy. Namely, we analyze an optimal locomotion of a group of animals coherently moving in viscous medium. The total energy dissipated by the moving group sensitively depends on the mutual disposition of the agents. The specific energy dissipation per an animal in a group may be significantly smaller than the dissipation of a solely moving agent. That's why the migrating birds form flocks and cyclists form a compact group in a cycle race – this helps to save a lot of energy. Here we demonstrate that the RL applied to the locomotion problem of a swarm, indeed, yields the optimal disposition of the swarm members.

Realistic modelling of a flock of birds is extremely difficult due to complicated form of birds and a complex flux structure in the air surrounding birds. Moreover, birds fly at high Reynolds numbers. Therefore, to illustrate the concept, we consider a simplified model, where a swarm is comprised of spherical agents moving with velocities corresponding to low Reynolds numbers. These constraints make the model tractable. Indeed, for low Reynolds numbers, there exists an analytical theory for the forces acting between spherical particles moving in viscous fluid (see Methods and SM). Using this theory the optimal configuration may be easily found for any number of agents, as illustrated in Fig. 4.

Next, we check whether the RL is able to find the optimal configurations for the same setup. Namely, we consider a group with a leading animal, which moves with a fixed velocity $\mathbf{v}_{lead}$ and is not affected by other agents of the group. Here we assume that more abundant information may be received and processed. Namely, we assume that the

FIG. 5: Optimal conformation for the group locomotion, obtained by RL for a system of two (a), three (b) and four agents (c). The red circle depicts the leader comfort zone, and circles of other color show the agent-followers comfort zones. The lines represent trajectories. Note that the configurations found by the RL practically coincides with those, obtained by a direct optimization, Fig. 4.

animals have a high perception level, that is, they can perceive: (i) relative positions of the agents and the leader; (ii) whether the nearest neighbor is within the comfort zone; (iii) self-velocity of an agent in the laboratory frame and the relative velocity with respect to other agents; (iv) the force exerted by an agent. The reward function tends to keep agent in their mutual comfort zones and minimize the total dissipation (see Methods and SM for detail).

Fig. 5 illustrates optimal configurations found by the RL, which practically coincides with those obtained by the direct optimization, Fig. 4. These results justify our a’p’riory trust that the RL manages to find the optimal solution, even if it is not known explicitly.

Conclusion and Outlook

We propose an explanation of the enigmatic phenomenon – swirling motion in large groups of animals at different evolution stage; possibly we reveal its biological function. Our approach is based on the hypothesis that the learning processes in nervous system of animals may be mimicked by the reinforcement learning (RL), with the reward for a beneficial action and punishment for a harmful one. We apply the RL to understand a collective motion of simple animals. They have very limited abilities to perceive and process information about their kinematic states. These limitations are associated with a rather basic level of their organization. We also consider physical limitations caused by the biological nature of the living beings. We formulate a small set of very simple rules which animals strive to follow to conform with the beneficial behavior. We hypothesize that such set of rules is imprinted in animals’ genes.

We assume, that among the main rules of the beneficial behavior in a swarm, is the rule to reside within a certain interval of distances (the comfort zone) from the center of the group and to reside within the comfort zone with respect to other neighbours. Such kind of the individual behavior corresponds to the escort behavior w.r.t. the group center. Therefore, we start with the analysis of the one-to-one escort problem, with a leader and a follower. The leader moves independently on various trajectories and the follower tends to reside within the respective comfort zone. We demonstrate that depending on the information-processing abilities of the follower and its physical limitations, very different escort strategies arise. These include such an amazing behavior as the leap-frog strategy and the strategy of wrapping trajectories.

We demonstrate that the escort strategy of individuals w.r.t the swarm center, results in a spontaneous swirling motion. We observe the emergence of swirling for all studied initial conditions and propose a criterion to quantify it. Interestingly, we reveal that the swirling motion leads to a dramatic increase of the resistance of a swarm to external perturbations. We believe that this proves the biological function of the swirling – animals swirl to protect themselves in a hostile environment.

We also consider the problem of optimized locomotion of a group of animals moving in viscous fluid. In this case the energy dissipation of a single animal drastically depends on the mutual disposition of the group members. That’s why the organization of the group is so important for coherently moving animals. Using a simplified model, we

demonstrate the ability of the RL to find the optimal configuration of a pack. We show that the agents disposition obtained by the RL practically coincide with these found by the direct minimization of the total dissipation, as it follows from the underlying physics. We believe that this result confirms that the RL is able to find the optimal solution, even for the case when such solutions may not be found by traditional mathematical methods.

In our study we use the assumption, that the exploited reward rules have been formed in the course of biological evolution – the process continuing on the evolution time-scale, comprising myriads of species generations. It is governed by random mutations of animals’ genes accompanied by natural selection [3]. This has a strong impact on the evolution of the rewards rules and drives them (by the natural selection) to the most optimal ones for a species survival. The result of the evolution is imprinted in genes of all animals of a species and persists as long as the species exists. In contrast, the training process, has a much shorter time-scale – just a time needed to train an animal. The result of the training is coded in neurons of a particular living being and disappears with its death. Naturally, it is desirable to have a complete model, which describes both, the formation of the reward rules during the biological evolution as well as a training of individual beings. This is however computationally very challenging task and we leave it for future studies.

III. METHODS

A. General principles of the RL

To model an agent (animal) information processing we firstly need to describe the surroundings. That is, we need to formulate the physical properties of the objects and the surrounding media. The respective model is called the “environment”. It describes the system dynamics using the predefined parameters as well as a distribution of these parameters, e.g., the distribution of the initial agents’ positions and velocities. Generally speaking, the environment contains all possible evolution scenarios of the system; its elaboration is in the heart of the approach. Secondly, we need to specify the level of the agents awareness of the environment. The available information about the surroundings refers to an agent state. It contains information about the medium and surrounding objects – their speed, location, etc. It reflects an agent’s understanding of itself as well. Thirdly, we need to define an action space – it determines the way, how the agents interact with the environment. The fourth component is called “policy”; it characterizes the strategy of the agents’ actions. It is quantified by a probability of a certain action at a given state. Finally, the fifth part refers to a reward. The reward is a mechanism to judge, how desirable is the present state as compared to another one. It also assesses a benefit of an action to reach a specific state right now, instead of doing this later.

Once all the above parts are specified, the most computationally intensive part – the learning of the optimal policy, may be performed. Such an approach is called “reinforcement learning” (RL). There are plenty of algorithms to obtain the optimal policy. Below we use the policy gradient algorithm [39]. To summarize, we use the following steps in our analysis:

1. Implement a dynamic simulator of an environment
2. Identify conditions of the environment available for an agent awareness
3. Propose a set of actions to interact with the environment and other agents
4. Limit possible strategies in a parametric class called a policy
5. Identify the motives/goals driving the agents in the form of rewards
6. Run learning to find out the optimal policy

In what follows we discuss the above steps in detail; at the end of the section we present mathematical formulation and algorithms.

B. Environment

We start with the random initialization and then continue with the dynamic simulations. Namely, we generate random initial positions and velocities of all agents. Also, an external force direction (wind direction) is specified as a random parameter. Random parameters as well as other specified variables remain constant during a specified time cycle of modeling, called an “epoch”. The force magnitude changes independently as a random process in time

according to specified probability distribution. In our study we exploit the stretched exponential distribution for the magnitude of the random force f , widely used to model natural phenomena [24, 35]:

$$P(f | f_0, \beta) = \frac{1}{f_0} \frac{\beta}{\Gamma(1/\beta)} \cdot e^{-(f/f_0)^\beta}.$$

Here $\Gamma(x)$ is the Gamma function, β is the stretching exponent and f_0 is the scale factor. We mainly use the simplest version with $\beta = 1$, which is also called exponential distribution. The parameter f_0 of the exponential distribution corresponds to the distribution of a mean value. Other distributions from this class with a wide range of β have been also tested. This does not change the qualitative behavior of the system, see SM for more detail. In our numerical experiments we model the evolution of the environment. Namely, we solve time-discretized equations, where the variables are calculated using their values on the previous time step and applying the action policy. Random factors that impact the process were also included. The simulations were repeated from one epoch to another, iteratively improving the action policy, according to the learning rules.

C. Policy

The action policy has been developed for various types of actions. All these require continuum random variables. Here we use normally distributed variables with a slight modification, that prevents an occurrence of very large quantities forbidden by the underlying physics. It is straightforward to derive the properties of such distributions (we call them “bounded normal distributions”); this is detailed in SM.

Denote $\pi_\theta(a_t | s_t)$ the policy with the parameters θ which is the probability density function of an action a_t for a given state s_t at time t . The probabilistic environment $p(s_{t+1} | s_t, a_t)$ describes the probability of the next state s_{t+1} at time $t + 1$ for a given state s_t and the applied action a_t at time t . The policy gradient algorithm [39] is based on a policy approximation by a class of parametric functions that encodes the probability distribution of actions in a given state. This class of functions iteratively approaches the optimal action policy with the use of Monte-Carlo gradient estimate which maximizes the expected reward. The reward $r_t = r_t(a_t, s_t)$ is a quantitative description of an advantage of a current state. It quantifies how desirable is the state s_t caused by the action a_t . The quality of the whole process, i.e., the quality of the action policy, can be assessed as an average sum of the successive (discounted) rewards:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots = \sum_{k=0}^T \gamma^k r_{t+k+1},$$

where $0 < \gamma < 1$ is the discount factor and T is the duration of the training epoch. The overall goal of an agent is to achieve the maximal cumulative reward. Note that the environment is random and the action policy is associated with the probability density function. Hence the maximization of a cumulative reward R_t corresponds to the maximization of the average expected reward, see SM for more detail. An analytical estimate of this quantity is usually not possible, while the numerical integration over an epoch is computationally demanding. Hence Monte-Carlo gradient estimate method is applied. Knowing the complete history of the state-action distribution, one can find an unbiased estimate of the policy gradient. Using the gradient descent, epoch after epoch, the optimal policy is eventually found. We refer to [39] and SM for the detailed description of this method.

D. Neural network architectures

The policy described above depends on two sets of parameters – the mean values of the original normal distribution and the according variances. Since the action space is two-dimensional (the motion occurs in 2D), we have two means and two variances. The number of observed states strongly depends on the problem, however the neural network architecture remains the same. For the action policy this comprises three fully connected layers and two ELU (Exponential Linear Unit) activation functions between them, see Fig. 6a. The exponential activation functions are used to avoid negative values of the variances. The architecture of the baseline function Fig. 6b has the similar structure.

FIG. 6: (a) The architecture of the neural network for the action policy. (b) The architecture of the neural network for the baseline network.

E. Algorithms of the Escort problem

First we consider the escort problem which consists of a single agent-follower moving in an inert two-dimensional space and a leader, moving on a predefined trajectory. The goal of the follower is to pursue the leader, more specifically – to reside within some range of distances from the leader.

Let the leader position be $\mathbf{x}_l$ and the follower – $\mathbf{x}_f$. Then the following information available for the follower reads:

1. Whether the follower resides within a range of distances from the leader:

$$S_d = H(e_t - |d_t - \|\mathbf{x}_l - \mathbf{x}_f\||),$$

where $H(x)$ is the Heaviside step function, d_t is the average distance from the leader and $e_t > 0$ specifies the range of acceptable distances.

2. Whether the follower approaches the leader:

$$S_a = H\left(-\frac{d}{dt}\|\mathbf{x}_l - \mathbf{x}_f\|\right).$$

The goal of the agent is to maximize the reward which is computed by the Algorithm 1. We encode the reward function in the way that the most desirable state is to reside within a range of target distances from the leader. However, if the agent is not close enough, the reward is given for the pursuit of the leader.

Algorithm 1 Reward – Escort problem

Require: Agent states:

1. Whether the follower resides within a range of target distances from the leader: S_d .
2. Whether the follower approaches the leader: S_a .

Ensure: Reward r .

1: $r = 1$ **if** $S_d = 1$ **else** (0 **if** $S_a = 1$ **else** -1)

The second important part of the policy refers to the actions which a follower can perform to maximize the reward. We define the leader velocity as $\mathbf{v}_l$ and the follower velocity as $\mathbf{v}_f$. We assume that the follower knows the direction of the leader velocity $\hat{\mathbf{v}}_l = \mathbf{v}_l / \|\mathbf{v}_l\|$ and the direction of its own velocity $\hat{\mathbf{v}}_f = \mathbf{v}_f / \|\mathbf{v}_f\|$. For the scenarios A and B detailed in the Section II the follower also knows the magnitude of $\mathbf{v}_f$; that is, the agent possesses a complete information about its velocity. Hence, the agent (follower) actions refer to the regulation of its own velocity. In real life, however, the velocity is limited, which implies the limitation of actions that needs to be reflected in the policy. For the scenario B there exist an additional limitation for an agent acceleration, which is also reflected in the policy, see SM.

Here we discuss in detail the most realistic escort scenario C, where the information about the velocity of an agent (follower) is not available, but the agent can control the force, that is, regulate its acceleration. Namely, the agent possesses the following information characterising the state of the environment: (i) the direction of the leader velocity $\hat{\mathbf{v}}_l = \mathbf{v}_l / \|\mathbf{v}_l\|$, (ii) the direction of the follower velocity $\hat{\mathbf{v}}_f = \mathbf{v}_f / \|\mathbf{v}_f\|$ and (iii) the follower force $\mathbf{F}_f$. The available information about the space location has been itemised above. The actions of an agent for the scenario C refer to the regulation of its force, instead of the velocity, as in the scenarios A and B. The force, and thus the acceleration is limited, as in the scenario B.

Noteworthy, without a knowledge of a follower/leader position and a follower/leader absolute velocity, it is extremely difficult (if possible) to obtain an explicit policy of the optimal pursuit. The reinforcement learning allows to construct an efficient stochastic policy for a very limited available information.

F. Algorithms of motion in a swarm and emergence of swirling

Similarly to the escort scenario each agent moving in a swarm exploits the available information of its state. We additionally assume that the agents strive to reside not far from the group center and supplement the agent state with extra parameters which specify the desired interval of distances from the group center located at $\mathbf{r}_{gc}$. That is, we use the concept of the target range of distances from the group center. Each agent strives to reside within the radius r_{cz} (i.e. in the comfort zone) around $\mathbf{r}_{gc}$. Thus, each i th agent in the swarm perceives the following data:

1. Whether the closest neighbour breaks into the comfort zone r_{cz} :

$$S_{cz}^i = H(r_{cz} - \|\mathbf{x}_i - \mathbf{x}_{c_i}\|), \quad c_i = \arg \min_{j \neq i} \|\mathbf{x}_i - \mathbf{x}_j\|,$$

where $\mathbf{x}_{c_i}$ is the coordinate of the nearest neighbour and $H(x)$ is the Heaviside step function.

2. Whether the agent approaches the closest agent:

$$S_{ca}^i = H\left(-\frac{d}{dt}\|\mathbf{x}_i - \mathbf{x}_{c_i}\|\right).$$

3. Whether the agent resides within a target distance r_{cz} to the group center $\mathbf{r}_{gc}$:

$$S_{gc}^i = H\left(r_{cz} - \|\mathbf{x}_i - \mathbf{r}_{gc}\|\right).$$

4. Whether the agent approaches the group center:

$$S_{ga}^i = H\left(-\frac{d}{dt}\|\mathbf{x}_i - \mathbf{r}_{gc}\|\right), \quad \mathbf{r}_{gc} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i,$$

where n is a number of the agents.

Based on these data we define the reward, see the Algorithm 2. It encodes our assumptions about an agent striving. The most desirable agent states are realized when the agent resides within a target distance to the group center. However, the agent is penalized if someone brakes into its comfort zone. Finally, if the agent is not within a target distance, we reward a pursuit of the group center.

Algorithm 2 Reward for the motion in a swarm

Require: The agent states:

1. Whether the closest agent breaks into the comfort zone: S_{cz}^i .
1. Whether the agent approaches to the group center: S_{ga}^i .
2. Whether the agent resides within a target distance from the group center: S_{gc}^i .

Ensure: Reward r^i .

- 1: $r_{gc}^i = 2$ if $S_{gc}^i = 1$ else (1 if $S_{ga}^i = 1$ else 0)
 - 2: $p_{cz}^i = 1$ if $S_{cz}^i = 1$ else 0
 - 3: $r^i = r_{gc}^i - p_{cz}^i$
-

Additionally, each agents knows the direction of the closest agent velocity $\hat{\mathbf{v}}_{c_i} = \mathbf{v}_{c_i}/\|\mathbf{v}_{c_i}\|$, the velocity direction of the agent itself $\hat{\mathbf{v}}_i = \mathbf{v}_i/\|\mathbf{v}_i\|$, the agent force (its action) $\mathbf{F}_i$ and the direction of the entire group velocity $\hat{\mathbf{v}}_{gc} = \mathbf{v}_{gc}/\|\mathbf{v}_{gc}\|$ for $\mathbf{v}_{gc} = \frac{1}{n} \sum_{i=1}^n \mathbf{v}_i$.

The emergent swirling motion is quantified by the average angular velocity of the swarm, which is defined as

$$\boldsymbol{\Omega} = \sum_{i=1}^n (\mathbf{v}_i - \mathbf{v}_{gc}) \times \frac{\mathbf{x}_i - \mathbf{r}_{gc}}{\|\mathbf{x}_i - \mathbf{r}_{gc}\|^2},$$

where the radius vector $\mathbf{r}_{gc}$ and velocity $\mathbf{v}_{gc}$ of the group center (i.e. of the center of mass) have been given above.

G. Physical background and algorithms for RL of efficient locomotion of a swarm

A system of spherical particles moving in viscous fluid at small Reynolds numbers may be described by the so-called Rodne-Prager theory [13], which goes beyond the classical Oseen theory [25]. This theory may be formulated in the form of the velocity of i -th agent, $\mathbf{v}_i$, resulting from the forces $\mathbf{F}_j^e$, $j = 1, 2, \dots, n$, applied to all n agents by the environment:

$$\mathbf{v}_i = - \sum_{j=1}^n \hat{\zeta}_{ij} \mathbf{F}_j^e \quad (1)$$

with the matrix of friction coefficients,

$$\hat{\zeta}_{ij} = \frac{1}{6\pi\eta a} \left[\frac{3a}{4x_{ij}} \left(\hat{I} + \mathbf{r}_{ij} \mathbf{r}_{ij}^T \right) + \frac{1}{2} \left(\frac{a}{x_{ij}} \right)^3 \left(\hat{I} - 3\mathbf{r}_{ij} \mathbf{r}_{ij}^T \right) \right]; \quad \hat{\zeta}_{ii} = \frac{1}{6\pi\eta a} \hat{I}, \quad (2)$$

where $\hat{I}$ is a unit matrix, a is the agent radius, η is the fluid viscosity and $\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ is the radius vector joining two agents. Finally, $\mathbf{r}_{ij} = \mathbf{x}_{ij}/\|\mathbf{x}_{ij}\|$ is the unit inter-agent vector.

We assume that the motion is always over-damped, so that the velocity of an agent immediately relaxes to a uniform, time-independent value, corresponding to the set of forces $\mathbf{F}_i^e$, as in Eqs. (1) and (2), see SM. Then the power dissipated by i -th agent may be written as $(\mathbf{v}_i - \mathbf{v}_{f,i}) \cdot \mathbf{F}_i^a$, where $\mathbf{F}_i^a$ is the force exerted by i -th agent on the medium and $\mathbf{v}_{f,i}$ is the fluid velocity at the location of i -th agent. By the third Newton's law the force exerted by a uniformly moving body on a medium equals to the force acting on the body back from the medium. Hence the total power dissipated by a swarm of n agents reads,

$$P = \sum_{i=1}^n P_i = \sum_{i=1}^n (\mathbf{v}_i - \mathbf{v}_{f,i}) \cdot \mathbf{F}_i^a = (6\pi\eta a)^{-1} \sum_{i=1}^n (\mathbf{F}_i^a)^2. \quad (3)$$

To obtain the last part of Eq. (3), we observe that the local fluid velocity follows from Eqs. (1) and (2): $\mathbf{v}_{f,i} = \mathbf{v}_i + \zeta_{ii} \mathbf{F}_i^e$, so that $(\mathbf{v}_i - \mathbf{v}_{f,i}) = -\mathbf{F}_i^e/(6\pi\eta a) = \mathbf{F}_i^a/(6\pi\eta a)$.

Consider a swarm where all animals move with the same absolute velocity in the same direction. Then, as it follows from Eqs. (1), (2) and (3), the total dissipation P will be determined by the set of vectors $\mathbf{x}_{ij}$, $i, j = 1, \dots, n$, that define the mutual disposition of the agents. To find the optimal swarm configuration, we need to minimize P . This is a well-posed optimization problem with a straightforward solution; the result is given in Fig. 4. Interestingly, the optimal configuration for a pack of three animals corresponds to a triangle with a base along the direction of motion, that is, two equivalent configurations are possible, see Fig. 4.

Our next goal is to check whether the RL is able to find the optimal configurations which are known. To solve this problem we consider a group of animals comprising of a leader moving in fixed direction with a constant velocity and several followers which aim to retain their comfort zone w.r.t other agents and the leader; simultaneously they strive to minimize the dissipation of energy. The agents (followers) are aware of the following information:

1. Whether the closest agent breaks into the comfort zone r_{cz} :

$$S_{cz}^i = H\left(r_{cz} - \|\mathbf{x}_i - \mathbf{x}_{c_i}\|\right), \quad c_i = \arg \min_{j \neq i} \|\mathbf{x}_i - \mathbf{x}_j\|,$$

where $\mathbf{x}_i$ is the i -th agent position and $H(x)$ is the Heaviside step function.

2. Whether the agent approaches the closest agent:

$$S_{ca}^i = H\left(-\frac{d}{dt}\|\mathbf{x}_i - \mathbf{x}_{c_i}\|\right).$$

3. Whether the agent resides within a target distance from the leader:

$$S_d^i = H\left(d_t - \|\mathbf{x}_i - \mathbf{x}_l\|\right),$$

where $\mathbf{x}_l$ is the leader position and d_t specifies the range of target distances to the leader.

We defined a reward, as explained in the Algorithm 3. The most desirable for the agent is to reside within the target distance to the leader. However, the agent is penalized if someone brakes into its comfort zone. If an agent is not within the target distance to the leader, we reward a pursuit of the leader. If an agent resides within the target distance to the leader, and there are no neighbours in the agent's comfort zone, it is penalized for the energy consumption. In our simulations the value of d_t was large enough to accommodate the whole swarm with an optimal configuration.

Algorithm 3 Reward: The locomotion efficiency increase

Require: The agent states:

1. Whether the closest agent brakes into the comfort zone: S_{cz}^i .
2. Whether the closest agent approaches to the leader: S_{ca}^l .
3. Whether the agent resides within a target distance to the leader: S_d^i .
4. Power dissipation of the i -th agent P_i .

Ensure: Reward r^i .

- 1: $r_l^i = 2 - \max(0, P_i)$ **if** $S_d^i = 1$ **else** (1 **if** $S_{ca}^l = 1$ **else** 0)
 - 2: $p_{cz}^i = 1$ **if** $S_{cz}^i = 1$ **else** 0
 - 3: $r^i = r_l^i - p_{cz}^i$
-

In this scenario, each agent is also aware about its velocity w.r.t. the “laboratory” system, that is, w.r.t. fluid at rest $\mathbf{v}_i$, the relative agents' velocities: $\mathbf{v}_{ij} = \mathbf{v}_j - \mathbf{v}_i$, the agent force (its action) $\mathbf{F}_i^a$ and the relative agents positions.

-
- [1] Amaze Lab Channel. Why army ants get trapped in 'death circles'. *Online video-sharing platform: YouTube*, 2019. URL: <https://www.youtube.com/watch?v=LEKwQxO4EZU>.
 - [2] L. Barberis and F. Peruani. Large-scale patterns in a minimal cognitive flocking model: incidental leaders, nematic patterns, and aggregates. *Physical review letters*, 117(24):248001, 2016.
 - [3] M. Begon, J. L. Harper, C. R. Townsend, et al. *Ecology. Individuals, populations and communities*. Blackwell scientific publications, 1986.
 - [4] D. P. Bertsekas, D. P. Bertsekas, D. P. Bertsekas, and D. P. Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena Scientific; 4th edition, Elsevier, 2017.
 - [5] M. Botvinick, S. Ritter, J. X. Wang, Z. Kurth-Nelson, C. Blundell, and D. Hassabis. Reinforcement learning, fast and slow. *Trends in cognitive sciences*, 23(5):408–422, 2019.
 - [6] M. D. Breed and J. Moore. *Animal behavior*. Academic Press, 2015.
 - [7] N. V. Brilliantov, H. Abutuqayqah, I. Y. Tyukin, and S. A. Matveev. Swirlonic state of active matter. *Scientific Reports*, 10:1, 2020.
 - [8] N. V. Brilliantov, A. Formella, and T. Pöschel. Increasing temperature of cooling granular gases. *Nature communications*, 9(1):797, 2018.
 - [9] N. V. Brilliantov and T. Pöschel. *Kinetic theory of granular gases*. Oxford University Press, 2004.
 - [10] S. Colabrese, K. Gustavsson, A. Celani, and L. Biferale. Flow navigation by smart microswimmers via reinforcement learning. *Physical review letters*, 118(15):158004, 2017.
 - [11] A. Czirók, H. E. Stanley, and T. Vicsek. Spontaneously ordered motion of self-propelled particles. *Journal of Physics A: Mathematical and General*, 30(5):1375, 1997.
 - [12] S. K. Das, S. A. Egorov, B. Trefz, P. Virnau, and K. Binder. Phase behavior of active swimmers in depletants: molecular dynamics and integral equation theory. *Physical Review Letters*, 112(19):198301, 2014.
 - [13] J. K. G. Dhont. *An introduction to dynamics of colloids*, volume 2. Elsevier, 1996.
 - [14] U. Erdmann, W. Ebeling, and V. S. Anishchenko. Excitation of rotational modes in two-dimensional systems of driven brownian particles. *Physical Review E*, 65(6):061106, 2002.

- [15] U. Erdmann, W. Ebeling, and A. S. Mikhailov. Noise-induced transition from translational to rotational motion of swarms. *Physical Review E*, 71(5):051904, 2005.
- [16] G. Flierl, D. Grünbaum, S. Levins, and D. Olson. From individuals to aggregations: the interplay between behavior and physics. *Journal of Theoretical biology*, 196(4):397–454, 1999.
- [17] J. Foerster, I. A. Assael, N. de Freitas, and S. Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2137–2145, 2016.
- [18] N. R. Franks, A. Worley, K. A. Grant, A. R. Gorman, V. Vizard, H. Plackett, C. Doran, M. L. Gamble, M. C. Stumpe, and A. B. Sendova-Franks. Social behaviour and collective motion in plant-animal worms. *Proceedings of the Royal Society B: Biological Sciences*, 283(1825):20152946, 2016.
- [19] J. K. Gupta, M. Egorov, and M. Kochenderfer. Cooperative multi-agent control using deep reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 66–83. Springer, 2017.
- [20] M. Hüttenrauch, S. Adrian, G. Neumann, et al. Deep reinforcement learning for swarm systems. *Journal of Machine Learning Research*, 20(54):1–31, 2019.
- [21] M. Hüttenrauch, A. Šošić, and G. Neumann. Local communication protocols for learning complex swarm behaviors with deep reinforcement learning. In *International Conference on Swarm Intelligence*, pages 71–83. Springer, 2018.
- [22] M. Ji and M. Egerstedt. Distributed coordination control of multiagent systems while preserving connectedness. *IEEE Transactions on Robotics*, 23(4):693–703, 2007.
- [23] M. Kabra, A. A. Robie, M. Rivera-Alba, S. Branson, and K. Branson. Jaaba: interactive machine learning for automatic annotation of animal behavior. *Nature methods*, 10(1):64, 2013.
- [24] J. Laherrere and D. Sornette. Stretched exponential distributions in nature and economy: fat tails with characteristic scales. *Eur. Phys. J. B*, 2:525, 1998.
- [25] L. D. Landau and E. M. Lifshitz. *Fluid Mechanics*, volume 2nd edition. Oxford University Press, Oxford, 1987.
- [26] H. Levine, W.-J. Rappel, and I. Cohen. Self-organization in systems of self-propelled particles. *Physical Review E*, 63(1):017101, 2000.
- [27] A.-C. Mailleux, J.-L. Deneubourg, and C. Detrain. How do ants assess food volume? *Animal behaviour*, 59(5):1061–1069, 2000.
- [28] E. O. Neftci and B. B. Averbeck. Reinforcement learning in artificial and biological systems. *Nature Machine Intelligence*, 1(3):133–143, 2019.
- [29] K. P. O’Keefe, H. Hong, and S. H. Strogatz. Oscillators that sync and swarm. *Nature communications*, 8(1):1–13, 2017.
- [30] J. K. Parrish and L. Edelstein-Keshet. Complexity, pattern, and evolutionary trade-offs in animal aggregation. *Science*, 284:99, 1999.
- [31] J. K. Parrish, S. V. Viscido, and D. Grünbaum. Self-organized fish schools: an examination of emergent properties. *The biological bulletin*, 202(3):296–305, 2002.
- [32] G. Reddy, J. Wong-Ng, A. Celani, T. J. Sejnowski, and M. Vergassola. Glider soaring via reinforcement learning in the field. *Nature*, 562(7726):236–239, 2018.
- [33] C. W. Reynolds. Flocks, herds and schools: A distributed behavioral model. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, pages 25–34, 1987.
- [34] M. Romensky, V. Lobaskin, and T. Ihle. Tricritical points in a vicsek model of self-propelled particles with bounded confidence. *Phys. Rev. E*, 90:063315, 2014.
- [35] D. F. Schmidt and E. Makalic. Universal models for the exponential distribution. *IEEE Transactions on Information Theory*, 55:3087, 2009.
- [36] E. Schneider and H. Stark. Optimal steering of a smart active particle. *EPL (Europhysics Letters)*, 127(6):64003, 2019.
- [37] S. Sukhbaatar, R. Fergus, et al. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems*, pages 2244–2252, 2016.
- [38] D. J. T. Sumpter. *Collective animal behavior*. Princeton University Press, 2010.
- [39] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.
- [40] J. J. Valletta, C. Torney, M. Kings, A. Thornton, and J. Madden. Applications of machine learning in animal behaviour studies. *Animal Behaviour*, 124:203–220, 2017.
- [41] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet. Novel type of phase transition in a system of self-driven particles. *Physical review letters*, 75(6):1226, 1995.
- [42] C. Zhang, C. Sun, L. Gao, N. Zheng, W. Chen, and X. Zheng. Bio-robots automatic navigation with graded electric reward stimulation based on reinforcement learning. In *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6901–6904. IEEE, 2013.

Figures

Figure 1

(a) Illustration of the escort problem { the leader moves independently on the outer circle, while the follower tries to reside within the comfort zone (shown shaded). (b) Illustration of the swirling motion in a swarm. It emerges spontaneously. The comfort zone w.r.t. the center of the swarm is shown (shaded). (c) Sketch of the swirling motion in a large colony of insects (ants); the individual comfort zone is zoomed. Similar patterns are observed in nature, see e.g. [1].

Figure 2

(a) The typical trajectory of the follower after the training for A-scenario of escort - the follower can perceive its absolute velocity without a limitation for the acceleration. Note the "frog-strategy" of the follower (see the text). (b) The typical trajectory of the follower after the training for B-Scenario of escort { the follower can perceive its absolute velocity and its acceleration is limited. The follower moves on a relatively smooth trajectory. (c)-(e) The typical trajectory of the follower after the training for C-scenario of escort - the follower can perceive only its acceleration and the acceleration is limited. The trained follower

moves on wrapping trajectories for all trajectories of the leader - circular (not shown), eight-curve (c), triangular (d) and elliptic (e).

Figure 3

(a) Swirling in a swarm of 100 agents. The large violet circle describes the target distance to the group center, white arrows indicate the speed direction, and the small shaded circles are the agents' comfort zones. The center of the swarm moves with a constant velocity. Inset: Time dependence of the average angular velocity of the swarm (see Sec. III for the definition). Ω evolves from zero to non-zero steady-state value. (b) The dependence of the shift of a swarm center in the direction of the applied force as a function of the average external force. The results are shown for swarms trained for swirling and untrained swarms. The trained swarm (with swirling) can resist the external force up to 100 times more efficiently than the untrained one. The resistance of the trained swarm fades for very large external forces when they become comparable to the maximal force that an agent can exert. The base value marker indicates the reference value of the average force used in SM.

Figure 4

Optimal configurations for the group of two, three, and four agents moving with the same velocity (see SM for the optimal configurations of larger swarms). The results are obtained by the direct minimization of the total dissipation power (see Methods). The average dissipated power per agent is indicated in units of the power dissipated by a solely moving agent. Note that optimal configuration for the triplet has two symmetric realizations. Also note that the larger the swarm, the smaller the specific dissipation.

Figure 5

Optimal conformation for the group locomotion, obtained by RL for a system of two (a), three (b) and four agents (c). The red circle depicts the leader comfort zone, and circles of other color show the agent-followers comfort zones. The lines represent trajectories. Note that the configurations found by the RL practically coincides with those, obtained by a direct optimization, Fig. 4.

Figure 6

(a) The architecture of the neural network for the action policy. (b) The architecture of the neural network for the baseline network.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplInfoWhySwirl060421.pdf](#)
- [TexfilesandfiguresEdited.zip](#)