scientific reports

Check for updates

OPEN

Learning to cooperate for low-Reynolds-number swimming: a model problem for gait coordination

Yangzhe Liu[1], Zonghao Zou[2], On Shun Pak[3]✉ & Alan C. H. Tsang[1]✉

Biological microswimmers can coordinate their motions to exploit their fluid environment—and each other—to achieve global advantages in their locomotory performance. These cooperative locomotion require delicate adjustments of both individual swimming gaits and spatial arrangements of the swimmers. Here we probe the emergence of such cooperative behaviors among artificial microswimmers endowed with artificial intelligence. We present the first use of a deep reinforcement learning approach to empower the cooperative locomotion of a pair of reconfigurable microswimmers. The AI-advised cooperative policy comprises two stages: an approach stage where the swimmers get in close proximity to fully exploit hydrodynamic interactions, followed a synchronization stage where the swimmers synchronize their locomotory gaits to maximize their overall net propulsion. The synchronized motions allow the swimmer pair to move together coherently with an enhanced locomotion performance unattainable by a single swimmer alone. Our work constitutes a first step toward uncovering intriguing cooperative behaviors of smart artificial microswimmers, demonstrating the vast potential of reinforcement learning towards intelligent autonomous manipulations of multiple microswimmers for their future biomedical and environmental applications.

In nature, animals like fish and bird use their fluid environment—and each other—to gain advantage for their locomotion[1,2], leading to fascinating pattern formation and adjustments of locomotory gaits observed in fish schooling and bird flocking. Such cooperative behaviors are also ubiquitous in the microscopic world, where swimming microorganisms exploit hydrodynamic interactions to enhance their locomotory performance[3,4]. Successful cooperative locomotion between microswimmers would require fine adjustments of not only their individual swimming gaits but also their spatial arrangements simultaneously. Biological microswimmers can evolve strategies to achieve such a complex coordination. For example, a pair of nearby sperm cells phase-lock and synchronize their flagellar beating patterns to swim cooperatively[5–8]. Yet, there are no evolved strategies available for cooperative locomotion of artificial microswimmers[9,10]. Moreover, strategies employed by biological microswimmers may not be directly applicable to artificial microswimmers, which have intrinsically different actuation mechanisms. Pioneering works have endowed artificial microswimmers with artificial intelligence (AI) to acquire effective locomotion strategies[11–17]. These advances prompt several general questions we set out to address here: When artificial microswimmers are equipped with adaptive decision making, what are the strategies for them to cooperate and achieve enhanced locomotion otherwise unattainable by isolated swimmers? Do these microswimmers adapt their strategy at different stages of cooperative swimming? How should the locomotory gaits of neighbouring swimmers be adjusted to exploit hydrodynamic interactions for maximizing their overall propulsion?

In this work, we present the first use of reinforcement learning (RL) to investigate cooperative locomotion of microswimmers. Recent studies have demonstrated the prowess of RL as a new approach to investigate locomotion problems in fluids. Different RL techniques have been utilized to empower simple reconfigurable microswimmers consisting of linked spheres to self-learn effective locomotory gaits based on interactions with the surrounding fluid[15,18–20]. Without any prior knowledge of locomotion at low Reynolds number (Re), these smart microswimmers are capable of evolving effective locomotory gaits to perform complex maneuvers such

[1]Department of Mechanical Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong, China. [2]Sibley School of Mechanical and Aerospace Engineering, Cornell University, Ithaca, NY 14850, USA. [3]Department of Mechanical Engineering, Santa Clara University, Santa Clara, CA 95053, USA. ✉email: opak@scu.edu; alancht@hku.hk

as targeted navigation and chemotactic responses[19,20]. Recent experimental studies have also begun to realize artificial microswimmers with control systems integrated with RL algorithms[16,21]. Other machine learning approaches have also been proposed to address locomotion problems in fish[22–27] as well as navigation problems of self-propelled particles[11,13,14,16,17]. There are initial efforts on extending machine learning approaches to cooperative locomotion problems at high Re such as fish locomotion[23–25]. Yet, the cooperative behavior of smart artificial microswimmers remains a largely unexplored area of research.

Here, as a first step, we consider a simple model problem of a pair of reconfigurable microswimmers consisting of three linked spheres aligned colinearly (Fig. 1a). The locomotory gait for a single three-sphere swimmer was first studied by Najafi and Golestanian, which generate net propulsion by modulating the relative distances between the spheres[28]. The three-sphere swimmer, together with pioneering work of Purcell's three-link swimmer[29], represent canonical examples on how to escape the constraints of the scallop theorem and generate self-propulsion at low Re. Recent works have demonstrated how RL enables the self-learning of effective locomotory gaits of a single three-sphere swimmer[15,19,20]. Here we employ a deep neural network with an Actor-Critic structure (Fig. 1b) to investigate how RL enables two three-sphere swimmers to coordinate their motions to enhance the overall locomotory performance. The swimmer pair will learn how to exploit hydrodynamic interactions by finely adjusting their individual locomotory gaits as well as modulating relative distances between each other. We show that the swimmers approach each other initially and both swimmers eventually swim in Najafi-Golestanian's strokes (N-G-strokes) with a constant phase mismatch between their gaits. The synchronized motions allow the swimmer pair to move together coherently with an enhanced locomotion performance unattainable by a single swimmer alone. This work constitutes a first step toward uncovering intriguing cooperative behaviors of smart artificial microswimmers.

Swimmer model and deep reinforcement learning framework

Model of a pair of reconfigurable microswimmers. We consider a pair of colinear, reconfigurable microswimmers comprising three spheres with radius R connected by extensible arms of negligible diameters (Fig. 1). The positions of the spheres are denoted by $\mathbf{r}_i$ ($i = 1$ to $i = 6$) and the lengths of the arms are denoted by L_i ($i = 1$ to $i = 4$). The swimmer transitions from one configuration to the other by extending or contracting one arm at a time (Fig. 1c), where we set the extended length and contracted length of the arms to be $10R$ and $6R$, respectively. A set of effective locomotory gaits for a single three-sphere swimmer was obtained by Najafi and Golestanian[28], which is featured by a periodic sequence of motions from configuration 1 to configuration 4 illustrated in Fig. 1c. Subsequent studies have used similar reconfigurable systems to generate net translation, rotation, and combined motion[30–37], including a recent application of deep RL to obtain effective locomotory gaits for complex maneuvers[20]. Instead of focusing on locomotion problems of a single microswimmer as in these previous studies, here we investigate effective strategies for cooperative locomotion of a pair of reconfigurable swimmers via RL approach.

Hydrodynamic interactions. The hydrodynamics of the reconfigurable microswimmers at low Re flow is governed by the Stokes equation ($\mu\nabla^2\mathbf{u} = \nabla p, \nabla\cdot\mathbf{u} = \mathbf{0}$). Here, p, μ and $\mathbf{u}$ denote the pressure, dynamic viscos-

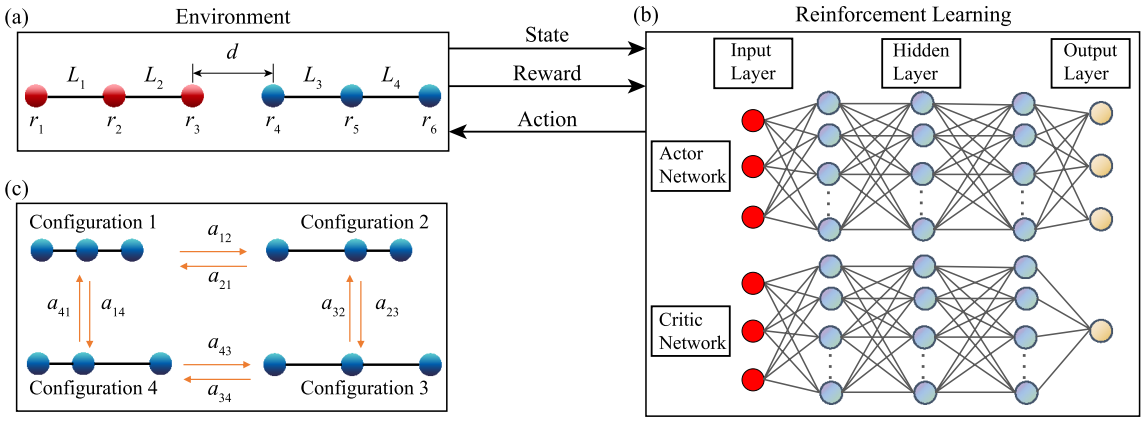

Figure 1. Schematics of a pair of three-sphere microswimmers with colinear arrangement and Actor-Critic neural network architecture. (**a**) Schematic of environment setup for reinforcement learning. Each swimmer consists of three rigid spheres with radius R and two extensible arms, and two identical swimmers are arranged colinearly. We distinguish the two swimmers by marking the spheres of the swimmer at the back as red and the spheres of the swimmer at the front as blue. The lengths of the extensible arms are denoted by L_i (i = 1,2,3,4) and the positions of the spheres' centers are denoted by $\mathbf{r}_i$ (i = 1,2,...,6). The closest distance between two swimmers is denoted as d, which is defined as the distance between $\mathbf{r}_3$ and $\mathbf{r}_4$. (**b**) The deep neural network has an Actor-Critic structure, in which the Actor-network memorizes and updates the learning policy, and the Critic-network estimates a value function to evaluate the performance of the policy. (**c**) Schematic showing the transition of the swimmer's configuration due to its actuation. The swimmer can either extend or contract one of its two links at a step and each swimmer has a total of 4 possible configurations.

ity, and velocity field, respectively. When the spheres are far apart from each other (i.e. in the limit of $R/L_i \ll 1$), the leading-order hydrodynamic interactions between the spheres in the fluids can be captured by the Oseen tensor[38–40]. The velocities of spheres $\mathbf{V}_i$ and the forces acting on each sphere $\mathbf{F}_i$ are related as

$$\mathbf{V}_i = \sum_{j=1}^{N} \mathbf{H}_{ij}\mathbf{F}_j, \tag{1}$$

where

$$\mathbf{H}_{ij} = \begin{cases} \mathbf{I}/6\pi\mu R & i = j \\ \left(1/8\pi\mu|\mathbf{r}_{ij}|\right)\left(\mathbf{I} + \mathbf{r}_{ij}\mathbf{r}_{ij}/|\mathbf{r}_{ij}|^2\right) & i \neq j \end{cases}. \tag{2}$$

Here $\mathbf{I}$ is an identity matrix, $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ is the vector from sphere i to sphere j. At low Re, each swimmer is subject to the force-free condition individually,

$$\sum_{i=1}^{3} \mathbf{F}_i = \mathbf{0}; \quad \sum_{i=4}^{6} \mathbf{F}_i = \mathbf{0}. \tag{3}$$

The colinear system considered here is also torque-free by axisymmetry. Equations (1)–(3) form a closed system of equations that describes the interaction dynamics of our model microswimmers. In this problem, we choose the radius of the sphere R as the characteristic length and use it to scale all lengths in the system. Hereafter we present results only with scaled lengths.

Reinforcement learning. We employ a deep neural newtork based on an Actor-Critic structure to examine the cooperative behavior of our model microswimmers (Fig. 1b)[41,42]. Both Actor and Critic networks consist of three hidden layers with sizes of 128, 128, and 64 respectively. We implement separated layers between Actor and Critic networks to avoid interference between the two networks. In the current problem, the configurations of the swimmers are discrete and the relative distance between the swimmers is continuous (Fig. 1a,c). This is in contrast with previous studies of locomotion of a single reconfigurable microswimmer that account for either fully discrete or continuous state and action spaces[15,19,20]. Here we extend the deep RL framework in previous studies to tackle state and action spaces with mixed discrete and continuous parameters. We implement a clipped version of the Proximal Policy Optimization (PPO) algorithm[41–43]. See Algorithm 1 in the "Methods" section for the pseudo-code of the PPO algorithm.

At a learning step n, the microswimmer pair (learning agent) with observed states $\mathbf{S}_n$ performs an action $\mathbf{a}_n$ to reach a new state $\mathbf{S}_{n+1}$ and obtains a corresponding reward $\mathbf{R}_n$ from the environment. In this work, we define our **State, Action, Reward** as follow:

1. **State $\mathbf{S}_n$**. The state consists of two parts: The first part corresponds to the discrete configurations of the swimmers' arm length. Each arm has two possible configurations, either being extended or contracted (Fig. 1c). For two swimmers with a total of 4 arms, the discrete configuration space has a size of $2^4 = 16$. The second part corresponds to the closest distance d between the two swimmers, which is defined as the distance between the closest spheres of the two swimmers (i.e., spheres 3 and 4), and $d = |\mathbf{r}_4 - \mathbf{r}_3|$ (Fig. 1c). The value of d changes continuously during the propulsion of the swimmers, which provides information for the swimmers to adjust their relative position.
2. **Action $\mathbf{a}_n$**. The swimmers can perform an action $\mathbf{a}_{ij}$ to transition from configuration i to configuration j (Fig. 1c). Here, we allow both swimmers to actuate at the same time. Namely, each swimmer can choose to actuate one of its arms or to not actuate at each step. We exclude the action where two swimmers both stop at a step. Thus, the action space is discrete with a size of $3^2 - 1 = 8$.
3. **Reward $\mathbf{R}_n$**. Our learning goal is to maximize the overall net displacement of the two swimmers. To this end, we define the reward r_n at each learning step as the sum of centroid displacement of the swimmer at the back ($D_B = \sum_{i=1}^{3} D_i/3$) and the centroid displacement of swimmer at the front ($D_F = \sum_{i=4}^{6} D_i/3$), where D_i denotes the displacement of the sphere $\mathbf{r}_i$. We note that the adjustment of d is a key component for effective cooperative locomotion, where the swimmers have to maximize the hydrodynamic interactions between each other and avoid collision at the same time. To avoid the collision of the swimmers and to maintain the validity of the Oseen tensor approximation, we introduce a lower bound d_{lower} for d between the swimmers. The training episode will terminate when $d \leq d_{lower}$. Similarly, we introduce an upper bound d_{upper} for d to avoid the swimmer getting too far away, which corresponds to ineffective cooperative locomotion. To ensure a full exploration of the relative distance between the swimmers, we introduce an additional soft bound d_{soft} that is slightly larger than d_{lower}. A soft penalization r_{soft} is applied when $d < d_{soft}$ before a sharp termination of learning episode and a hard penalization $r_{terminate}$ are applied at $d \leq d_{lower}$. This transition from soft penalization to hard penalization results in a more continuous reward over the learning process, which helps for searching the optimal relative position between the swimmers for effective cooperative locomotion. The same hard penalization $r_{terminate}$ is applied when $d \geq d_{upper}$. As a result, the reward R_n can be expressed as

$$\mathbf{R}_n = \begin{cases} D_F + D_B, & d_{upper} > d \geq d_{soft} \\ r_{soft}, & d_{soft} > d > d_{lower} \\ r_{terminate}, & d \leq d_{lower}, d \geq d_{upper} \end{cases} \tag{4}$$

Here, d_{upper}, d_{lower} and d_{soft} are set as 70, 5 and 6, respectively. d_{lower} and d_{soft} are selected such that the spheres are sufficiently far from each other and the Oseen tensor approximation remains valid (i.e., $d_{soft}, d_{lower} \ll 1$). The reward for the buffer region and the termination region are set as $r_{soft} = -0.3$ and $r_{terminate} = -1$. These penalties are set relative to $D_F + D_B$ which has a typical order of 10^{-1}.

We limit the length of each training episode to $16 \times 1024 = 16,384$ learning steps, which corresponds to 1024 times of the size of the discrete configuration spaces. This ensures a sufficient number of learning steps for the swimmers to explore the effects of hydrodynamic interactions at different configurations. A discount factor $0 \leq \gamma < 1$ was introduced to assign a weight to immediate reward over the future reward. We set $\gamma = 0.9997$ to ensure farsightedness of the agent. We randomize the initial configurations of the swimmers and the initial closest distance $d_{initial}$ between two microswimmers in each episode to ensure a full exploration of all possible state spaces over training episodes.

We collect all the training information and extract the policy obtained from the training process periodically with a frequency of 2×10^5 learning steps. The extracted policy is then evaluated in an isolated evaluation environment. Note that the evaluation environment is the same as the training environment. However, there is no additional training performed during the evaluation process. The policy obtained from the training process typically has a probability distribution of various possible actions at a given state of the agent. There are two ways to evaluate the training results, namely stochastic evaluation and deterministic evaluation (see Supplementary Materials for more details). A stochastic policy follows the probability distribution to select the action at each step, whereas a deterministic policy always follows the action with the highest probability. In order to avoid being trapped in undesirable solutions, here we evaluate the extracted policy in a stochastic manner. After sufficient training is performed (i.e., $N_t > 10^7$, where N_t is the total number of training step), we observe that continuous training may result in a drop in the performance of the resulting policy. Possible reasons for such a drop in performance are catastrophic forgetting or overfitting[44,45]. In order to select the best policy in the training process, we perform an early stop on training before the performance drops. Such an early stopping has been demonstrated as an efficient way to prevent aggravating policy performance through long-time training in other studies[21,45].

Result and discussion

Cooperative locomotion and gait coordination. We systematically investigate how deep RL achieves an effective strategy for cooperative locomotion. All hyperparameters corresponding to the deep RL algorithm are summarized in the "Methods" section. We train the agent with a control policy π_θ to maximize the total displacement of the swimmer pairs. We monitor the training process by considering the moving average of the episodic reward over the last 100 episodes with respect to the total training steps N_t (see Supplementary Fig. S1). We perform an early stop on the training process and select the best model at $N_t = 5.4 \times 10^6$ steps according to the evaluation result (see Supplementary Fig. S2).

We visualize the selected best policy of cooperative locomotion in an isolated evaluation environment (Fig. 2, Supplementary Movie S1). We place two microswimmers in their fully extended configurations with an initial $d = 20$. We note that the policy obtained by RL is insensitive to the choice of initial condition as we have trained the swimmers with different initial conditions during the training process. The policy of cooperative locomotion obtained by the RL comprises two distinct stages, which we refer to as the approach stage (yellow region in Fig. 2a) and the synchronization stage (pink region in Fig. 2b). The closest distance d first gradually decreases in the approach stage until it reaches a minimum value close to d_{soft}, after which d oscillates periodically in the synchronization stage (Fig. 2a). We note that d will still be adjusted slightly during the synchronization stage. After a prolonged simulation, d will eventually get almost equal to d_{soft} (see Supplementary Fig. S3). The transition from the approach stage to the synchronization stage is classified by a switch in the pattern of d and hence a switch in the locomotory gaits of the swimmers. We consider it to be a complete transition when a new pattern of d repeats for 5 gait cycles. The occasional disturbances in d in the two stages are effects due to stochastic evaluation. In the approach stage, the swimmer at the back has substantially larger net displacement (D_B, red line in yellow region of Fig. 2b) than the swimmer at the front (D_F, blue line in yellow region of Fig. 2b), therefore the swimmer at the back will "approach" the swimmer at the front. In the synchronization stage, the two swimmers have essentially the same net displacement, as indicated by the same slope and the same periodic pattern of D_B and D_F in the pink region of Fig. 2b.

Now we elaborate how the AI-advised policy achieves effective cooperative locomotion by analyzing the details of the locomotory gaits of the swimmers at the two stages (Fig. 2c,d). An effective cooperative locomotion is subject to two major challenges: first, the swimmers have to adopt a strategy to approach each other and swim together, while not getting too close to collide with each other; second, the swimmers have to finely coordinate their locomotory gaits to exploit hydrodynamic interactions to maximize their overall propulsion. Here we show that these two challenges are indeed tackled properly by the AI-advised locomotory gaits of the swimmers at the two stages correspondingly. During the approach stage (yellow region in Fig. 2a,b, Supplementary Movie S2), the swimmer at the back swims in N-G strokes (Fig. 2c, red spheres, and approaches the swimmer at the front, whereas the swimmer at the front only exhibits forward propulsion occasionally, as can be seen from the relatively flat slope of D_F in Fig. 2b. Most of the time the swimmer at the front does not perform any motion or performs reciprocal motions that lead to zero net self-propulsion at low Re (Fig. 2c, blue spheres)[29]. As a result, the swimmer at the front "waits" for the swimmer at the back to catch up. After the swimmers are in sufficiently close proximity (i.e., minimum $d \approx d_{lower}$), the swimmers start to synchronize their locomotory gaits. In the synchronization stage (Supplementary Movie S3), the swimmers propel forward with N-G strokes, where there is a constant phase difference in the N-G strokes adopted by the two swimmers. Namely, the swimmer at the front has a delay of 1 actuation step in the N-G strokes compared to the swimmer at the back (Fig. 2d). The synchronized swimmers maintain a fixed range of d and propel with the same overall displacement (pink region in Fig. 2a,b).

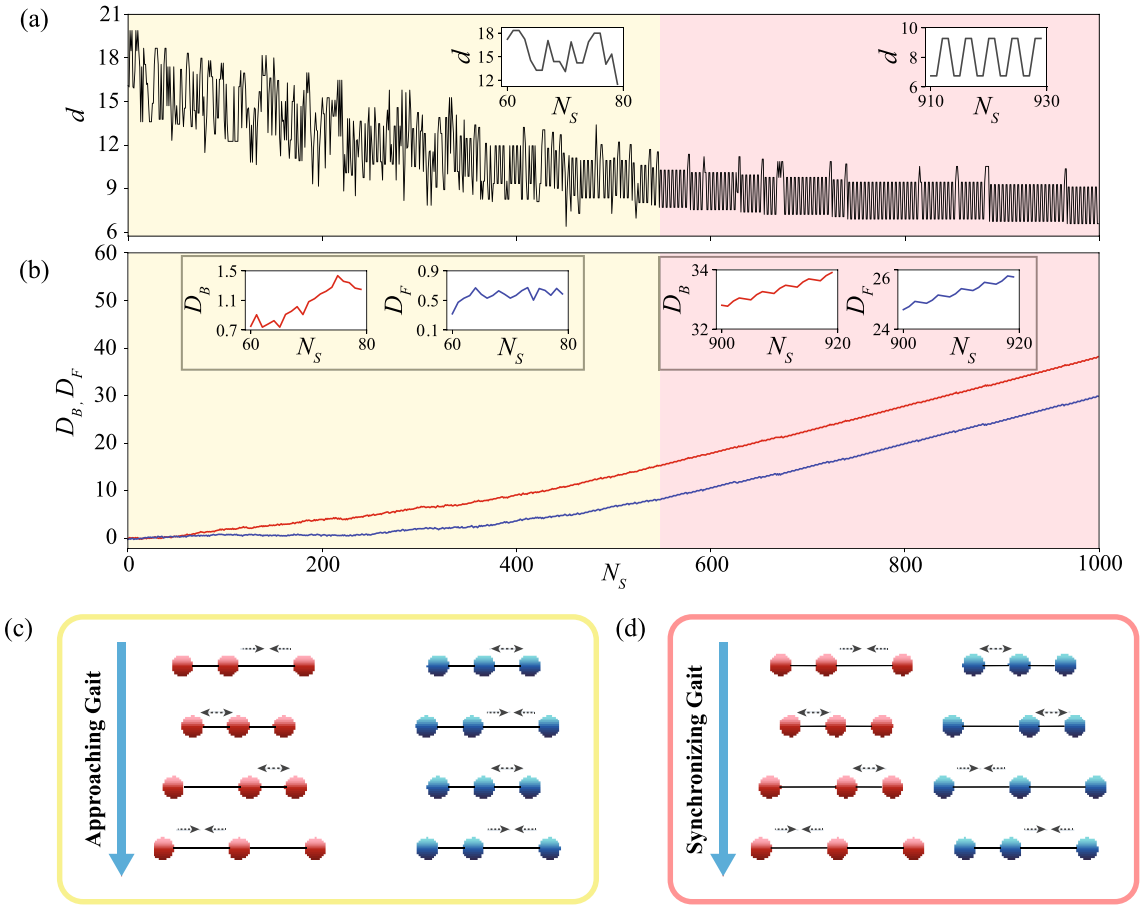

Figure 2. Visualization of AI-advised policy of cooperative locomotion in an isolated evaluation environment. (**a**) Change in d of the AI-advised policy with respect to the number of steps N_s. (**b**) Displacement of the swimmer at the front (D_F, blue line) and Displacement of the swimmer at the back (D_B, red line). In (**a**) and (**b**), the yellow and pink regions correspond to the approach stage and the synchronization stage, respectively. (**c**) Schematic of locomotory gaits for the approach stage, where the swimmer at the back follows N-G strokes and the swimmer at the front either generates zero net propulsion or propels with a small distance occasionally. (**d**) Schematics of locomotory gaits for the synchronization stage, where two swimmers swim cooperatively with N-G strokes at a constant phase mismatch. The swimmer at the front has a delay in 1 actuation step compared to the swimmer at the back.

We remark that in addition to the best model, the RL algorithm also acquires several suboptimal models throughout the training process, as can be seen from the local peaks in the average episodic rewards (Supplementary Fig. S1). These suboptimal models also exhibit locomotory gaits with approach and synchronization stages similar to the best model. However, these suboptimal models fails to adjust d properly in the approach stage and results in minimum d being larger than d_{soft} in the synchronization stage. Thus the suboptimal models fail to fully exploit hydrodynamic interactions between the swimmers for effective cooperative locomotion, leading to smaller episodic rewards. We note that the implementation of the soft bound d_{soft} in the reward in Eq. (4) is a key to obtain the best model with the minimum value of d being close to the lower bound d_{lower} in the synchronization stage. If the soft bound d_{soft} is not implemented, the policies obtained by RL will end up with a much larger d in the synchronization stage and the swimmers will fail to fully exploit the hydrodynamic advantage from their interactions.

Comparison between deterministic and stochastic policies. The AI-adivsed policy has successfully demonstrated how a swimmer pair achieves cooperative locomotion. Albeit being more robust in avoiding undesirable solutions, the policy achieved by stochastic evaluation has introduced random noises in the locomotory gaits. Here we investigate how such a stochasticity influences the overall locomotory performance by comparing the stochastic policy and the deterministic policy. In the deterministic policy, the swimmer at the front does not generate any net propulsion and waits for the swimmer at the back to get close in the approaching stage, instead of exhibiting random motions with a small overall displacement. After the swimmers get sufficiently close (i.e., $d \sim d_{soft}$), the swimmers enter the synchronization stage and propel with N-G strokes, where the strokes of the two swimmers have the same constant phase shift as the stochastic policy. We compare the average displacement of the two swimmers, i.e., $\langle D \rangle = (D_B + D_F)/2$ for the deterministic policy and the stochastic policy (Fig. 3, Supplementary Movie S4). The deterministic policy outperforms the stochastic policy at a short time scale ($N_s < 500$) as the swimmers have no random actions and enter the synchronization stage earlier. However, both policies perform approximately equal at a long time scale where they all enter the synchronization stage ($N_s > 500$). Nevertheless, the stochastic policy has a more stable performance during training, while the determinsitic policy can possibly be trapped in undesirable solutions with unexpectedly small net displacement during the training process.

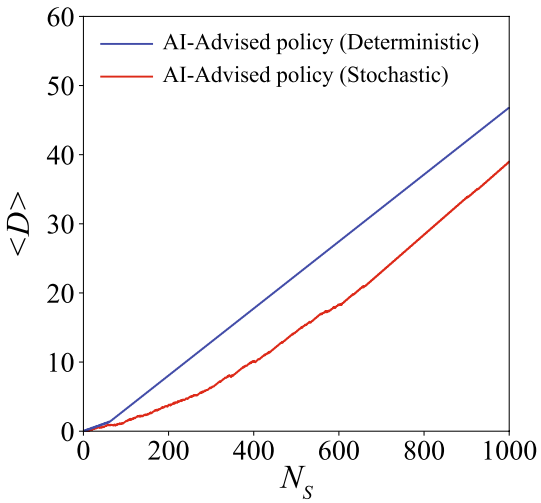

Figure 3. Comparison of average displacement $\langle D \rangle$ of the swimmer pair between the deterministic policy and the stochastic policy. The blue line and the red line denote the results for the deterministic policy and the stochastic policy, respectively.

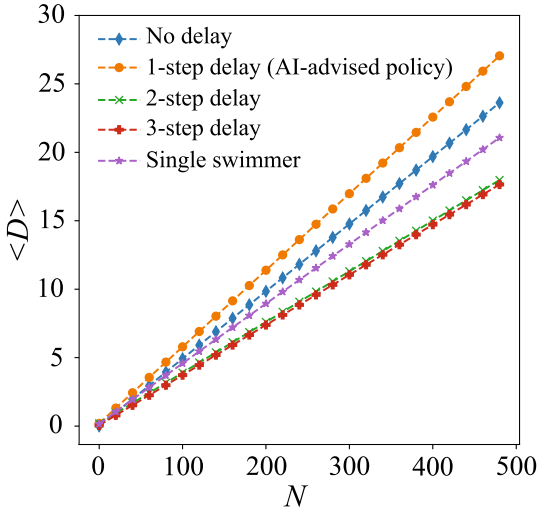

Figure 4. Comparison of average displacement $\langle D \rangle$ of the swimmer pair with prescribed N-G strokes and different phase mismatches. The colored lines denote the cases with different delays in N-G strokes of the swimmer at the front, including the cases of no delay (blue line), 1-step delay (orange line, AI-advised policy), 2-step delay (green), 3-step delay (red). Displacement of a single swimmer following N-G strokes (purple line) is added to benchmark the performance of cooperative locomotion of different prescribed locomotory gaits.

Comparison with other synchronized locomotory gaits. Our AI-advised policy suggests there exists a phase difference between the synchronized N-G strokes of the swimmer pair that maximizes the displacement. To further investigate how the phase difference influences the cooperative locomotion, we prescribe the locomotory gaits of the swimmer pair with N-G strokes with different mismatches in phase (i.e., different delays in actuation step for the swimmer at the front) and compare their locomotory performance. For a fair comparison, all prescribed gaits have the same minimum closest distance $d = 6$. We measure the locomotory performance of different prescribed gaits by the average displacements of the two swimmers, i.e., $\langle D \rangle$. The evolution of $\langle D \rangle$ for N-G strokes at different phase mismatches are displayed in Fig. 4 and Supplementary Movie S5. Our results demonstrate that the AI-advised locomotory gait (Fig. 4, orange dashed line) has indeed learnt the most effective phase mismatch of N-G strokes for cooperative locomotion among all mismatches considered. We also benchmark the locomotory performance of the prescribed locomotory gaits with a single swimmer with N-G strokes (Fig. 4, purple dashed line). We note that not all the synchronized N-G strokes outperforms the locomotory performance of a single swimmer. For the cases where the swimmer at the front has a delay of 2 or 3 steps in its N-G strokes (Fig. 4, green and red dashed lines), the synchronization of N-G strokes of the swimmer pair can be counterproductive, leading to a locomotory performance worse than that of a single swimmer alone. In contrast, the synchronized N-G strokes obtained by RL improves the propulsion speed by ~ 25% compared to the N-G strokes of a single swimmer. Here we demonstrate how RL successfully searches for effective locomotory gaits to achieve cooperative locomotion.

Conclusion

In this work, we present the first use of deep RL to empower a pair of microswimmers to cooperate for enhanced locomotion at low Re. The AI-advised policy of cooperative locomotion can be distinguished in two stages: an approach stage where the swimmers adjust their relative distance to get in close proximity for increasing their hydrodynamic interactions, followed by a gait coordination that optimizes the hydrodynamic interaction between the swimmers in a synchronization stage. The self-learning of this AI-advised policy involves the consideration of state and action spaces that include both discrete and continuous components. While fully continuous state and action spaces can be considered as in Refs.[15,18–20], this may significantly increase the number of learning steps to search for the optimal solution. We consider an axisymmetric, colinear configuration in this work as arguably the simplest model problem to explore the cooperative behavior of smart artificial microswimmers, while the deep RL framework here also applies to more general configurations. Subsequent works will build on the framework to investigate the cooperative behaviors of increased numbers of microswimmers with more complex spatial arrangements, probing the probable emergence of pattern formation[23,46] among smart artificial microswimmers, analogous to collective behaviors of fish, bird, and microorganisms observed in nature[47–49]. The consideration of non-colinear configurations may require more complex swimmer models that allow combined translational and rotational motion[20] to exhibit effective cooperative behaviors. Larger state and action spaces will have to be set up for the learning agent to incorporate these additional complex maneuvers. We envision that the RL approach here would be particularly relevant for coordinating a group of artificial microswimmers to perform collective microrobotic tasks that require fine adjustments of spatial locations within the group[50,51]. Here we reward the microswimmers for maximizing their overall net displacements; future works will explore the design of other reward functions to empower microswimmers to cooperate for different tasks such as chemotaxis[19,52]. Lastly, we also remark on the potential use of the centralized training and decentralized execution approach in multi-agent RL, which capitalizes on access to the full state and information during the training phase while addresses the challenge of individual agents not having access to the full state during the executive phase. Such an approach has been shown effective in different real-world applications[53,54] and may be considered as an alternative approach in future works.

To conclude, we have presented the first use of RL to achieve effective cooperative locomotion of artificial microswimmers endowed with AI. This proof-of-principle demonstration opens up new opportunities towards intelligent autonomous manipulation of multiple microrobots, laying the groundwork for their future biomedical and environmental applications[50,55,56].

Methods

PPO algorithm. We utilize a PPO clipped version[41], and calculate the advantage through Generalized Advantage Estimation[43]. The pseudo-code for the PPO clipped version is shown in Algorithm. 1. See Supplementary Materials for more details about the deep RL algorithm.

Algorithm 1 PPO (clipped version)

Input: Initial policy parameters θ_0, initial value function parameter ϕ_0

for $k = 1, 2, 3...K$ **do**

  Collect trajectories $\mathbf{D_k}$ by running policy $\pi_k = \pi(\theta_k)$

  Estimate advantage $\hat{A}_t^{\pi_k}$ using Generalized Advantage Estimation

  Compute clipped surrogate loss function with clipping parameter ϵ:

$$L\left(s,a,\theta_k,\theta\right)=\min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}A^{\pi_{\theta_k}}(s,a),\quad \text{clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)},1-\epsilon,1+\epsilon\right)A^{\pi_{\theta_k}}(s,a)\right)$$

  Update policy $\theta_{k+1}=\arg\max_\theta \underset{s,a\sim\pi_{\theta_k}}{\mathrm{E}}[L(s,a,\theta_k,\theta)]$ through Adam optimizer

  Update value function $\phi_{k+1}=\arg\min_\phi \underset{s,a\sim\pi_{\theta_k}}{\mathrm{E}}[(V_\phi(s_t)-\hat{R}_t)^2]$ through SGD

  $\theta_{old}\leftarrow\theta,\ \phi_{old}\leftarrow\phi$

end for

Hyperparameters. Here we elaborate on the hyperparameter settings we have used for training the cooperative locomotion of the microswimmer pair. The hyperparameters are tuned based on the training performance of different learning trails. It is conceivable that a better model can be learnt through systematic parameter tuning. We do not perform hyperparameter optimization due to the cost of computational expense. The hyperparameters are shown in Table 1.

Name of hyperparameter	Value	Description
Learning rate	0.0001	The learning rate used by Gradient descent
Neural network update frequency	16,384	Number of environment steps to run for each neural network update
Neural network architecture	Actor:128,128,64 Critic:128,128,64	Size of three hidden layers for the Actor-Critic network
Batch size	256	Number of training cases to be computed at each stochastic gradient descent (SGD)
Epoch	10	Number of time for the whole training cases computed at SGD
Discount factor	0.9997	Factor for computing discounted future reward
Clip range	0.2	Clipping parameter for PPO algorithm
Target KL divergence	N/A	Limit the difference between current policy and new policy

Table 1. Hyperparameters for deep reinforcement learning.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability

The codes that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 2 March 2023; Accepted: 31 May 2023
Published online: 09 June 2023

References

1. Weihs, D. Hydromechanics of fish schooling. *Nature* **241**, 290–291 (1973).
2. Sumpter, D. *Collective Animal Behavior* (Princeton University Press, 2010).
3. Lauga, E. & Powers, T. R. The hydrodynamics of swimming microorganisms. *Rep. Prog. Phys.* **72**, 096601 (2009).
4. Elgeti, J., Winkler, R. G. & Gompper, G. Physics of microswimmers-single particle motion and collective behavior: A review. *Rep. Prog. Phys.* **78**, 056601 (2015).

5. Yang, Y., Elgeti, J. & Gompper, G. Cooperation of sperm in two dimensions: Synchronization, attraction, and aggregation through hydrodynamic interactions. *Phys. Rev. E* **78**, 061903 (2008).
6. Woolley, D. M., Crockett, R. F., Groom, W. D. & Revell, S. G. A study of synchronisation between the flagella of bull spermatozoa, with related observations. *J. Exp. Biol.* **212**, 2215–2223 (2009).
7. Taylor, G. I. Analysis of the swimming of microscopic organisms. *Proc. R. Soc. Lond. Ser. A. Math. Phys. Sci.* **209**, 447–461 (1951).
8. Elfring, G. J. & Lauga, E. Hydrodynamic phase locking of swimming microorganisms. *Phys. Rev. Lett.* **103**, 088101. https://doi.org/10.1103/PhysRevLett.103.088101 (2009).
9. Martinez-Pedrero, F., Ortiz-Ambriz, A., Pagonabarraga, I. & Tierno, P. Colloidal microworms propelling via a cooperative hydrodynamic conveyor belt. *Phys. Rev. Lett.* **115**, 138301 (2015).
10. Martinez-Pedrero, F., Navarro-Argemí, E., Ortiz-Ambriz, A., Pagonabarraga, I. & Tierno, P. Emergent hydrodynamic bound states between magnetically powered micropropellers. *Sci. Adv.* **4**, eaap9379 (2018).
11. Colabrese, S., Gustavsson, K., Celani, A. & Biferale, L. Flow navigation by smart microswimmers via reinforcement learning. *Phys. Rev. Lett.* **118**, 158004 (2017).
12. Gustavsson, K., Biferale, L., Celani, A. & Colabrese, S. Finding efficient swimming strategies in a three-dimensional chaotic flow by reinforcement learning. *Eur. Phys. J. E* **40**, 1–6 (2017).
13. Schneider, E. & Stark, H. Optimal steering of a smart active particle. *Europhys. Lett.* **127**, 64003 (2019).
14. Yang, Y., Bevan, M. A. & Li, B. Micro/nano motor navigation and localization via deep reinforcement learning. *Adv. Theory Simul.* **3**, 2000034 (2020).
15. Tsang, A. C. H., Tong, P. W., Nallan, S. & Pak, O. S. Self-learning how to swim at low Reynolds number. *Phys. Rev. Fluids* **5**, 074101 (2020).
16. Muiños-Landin, S., Fischer, A., Holubec, V. & Cichos, F. Reinforcement learning with artificial microswimmers. *Sci. Robot.* **6**, eabd9285 (2021).
17. Gunnarson, P., Mandralis, I., Novati, G., Koumoutsakos, P. & Dabiri, J. O. Learning efficient navigation in vortical flow fields. *Nat. Commun.* **12**, 1–7 (2021).
18. Liu, Y., Zou, Z., Tsang, A. C. H., Pak, O. S. & Young, Y.-N. Mechanical rotation at low Reynolds number via reinforcement learning. *Phys. Fluids* **33**, 062007 (2021).
19. Hartl, B., Hübl, M., Kahl, G. & Zöttl, A. Microswimmers learning chemotaxis with genetic algorithms. *Proc. Natl. Acad. Sci.* **118**, e2019683118 (2021).
20. Zou, Z., Liu, Y., Young, Y.-N., Pak, O. S. & Tsang, A. C. H. Gait switching and targeted navigation of microswimmers via deep reinforcement learning. *Commun. Phys.* **5**, 1–9 (2022).
21. Behrens, M. R. & Ruder, W. C. Smart magnetic microrobots learn to swim with deep reinforcement learning. *Adv. Intell. Syst.* **4**, 2270049 (2022).
22. Gazzola, M., Hejazialhosseini, B. & Koumoutsakos, P. Reinforcement learning and wavelet adapted vortex methods for simulations of self-propelled swimmers. *SIAM J. Sci. Comput.* **36**, B622–B639 (2014).
23. Gazzola, M., Tchieu, A. A., Alexeev, D., de Brauer, A. & Koumoutsakos, P. Learning to school in the presence of hydrodynamic interactions. *J. Fluid Mech.* **789**, 726–749 (2016).
24. Novati, G. *et al.* Synchronisation through learning for two self-propelled swimmers. *Bioinspiration Biomimetics* **12**, 036001 (2017).
25. Verma, S., Novati, G. & Koumoutsakos, P. Efficient collective swimming by harnessing vortices through deep reinforcement learning. *Proc. Natl. Acad. Sci.* **115**, 5849–5854 (2018).
26. Jiao, Y. *et al.* Learning to swim in potential flow. *Phys. Rev. Fluids* **6**, 050505 (2021).
27. Yu, H. *et al.* Deep-reinforcement-learning-based self-organization of freely undulatory swimmers. *Phys. Rev. E* **105**, 045105 (2022).
28. Najafi, A. & Golestanian, R. Simple swimmer at low Reynolds number: Three linked spheres. *Phys. Rev. E* **69**, 062901 (2004).
29. Purcell, E. M. Life at low Reynolds number. *Am. J. Phys.* **45**, 3–11 (1977).
30. Dreyfus, R., Baudry, J. & Stone, H. A. Purcell's, "rotator": Mechanical rotation at low Reynolds number. *Eur. Phys. J. B-Condens. Matter Complex Syst.* **47**, 161–164 (2005).
31. Avron, J., Kenneth, O. & Oaknin, D. Pushmepullyou: An efficient micro-swimmer. *New J. Phys.* **7**, 234 (2005).
32. Golestanian, R. & Ajdari, A. Stochastic low Reynolds number swimmers. *J. Phys. Condens. Matter* **21**, 204104 (2009).
33. Alouges, F., DeSimone, A. & Lefebvre, A. Optimal strokes for low Reynolds number swimmers: An example. *J. Nonlinear Sci.* **18**, 277–302 (2008).
34. Nasouri, B., Vilfan, A. & Golestanian, R. Efficiency limits of the three-sphere swimmer. *Phys. Rev. Fluids* **4**, 073101 (2019).
35. Earl, D. J., Pooley, C., Ryder, J., Bredberg, I. & Yeomans, J. Modeling microscopic swimmers at low Reynolds number. *J. Chem. Phys.* **126**, 02B603 (2007).
36. Andrychowicz, M. *et al.* What matters in on-policy reinforcement learning? A large-scale empirical study. arXiv preprint arXiv:2006.05990 (2020).
37. Alouges, F., DeSimone, A., Heltai, L., Lefebvre-Lepot, A. & Merlet, B. Optimally swimming Stokesian robots. *Discrete Continuous Dyn. Syst. B* **18**, 1189–1215 (2013).
38. Happel, J. & Brenner, H. *Low Reynolds Number Hydrodynamics: With Special Applications to Particulate Media* Vol. 1 (Springer Science & Business Media, 2012).
39. Kim, S. & Karrila, S. J. *Microhydrodynamics: Principles and Selected Applications* (Courier Corporation, 2013).
40. Dhont, J. K. *An Introduction to Dynamics of Colloids* (Elsevier, 1996).
41. Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017).
42. Raffin, A. *et al.* Stable-baselines3: Reliable reinforcement learning implementations. *J. Mach. Learn. Res.* **22**, 1–8 (2021).
43. Schulman, J., Moritz, P., Levine, S., Jordan, M. & Abbeel, P. High-dimensional continuous control using generalized advantage estimation. arXiv preprint arXiv:1506.02438 (2015).
44. Kirkpatrick, J. *et al.* Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci.* **114**, 3521–3526 (2017).
45. Caruana, R., Lawrence, S. & Giles, C. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. *Adv. Neural Inf. Process. Syst.* **13**, 402–408 (2000).
46. Yigit, B., Alapan, Y. & Sitti, M. Programmable collective behavior in dynamically self-assembled mobile microrobotic swarms. *Adv. Sci.* **6**, 1801837 (2019).
47. Vicsek, T. & Zafeiris, A. Collective motion. *Phys. Rep.* **517**, 71–140 (2012).
48. Jeckel, H. *et al.* Learning the space-time phase diagram of bacterial swarm expansion. *Proc. Natl. Acad. Sci.* **116**, 1489–1494 (2019).
49. Cichos, F., Gustavsson, K., Mehlig, B. & Volpe, G. Machine learning for active matter. *Nat. Mach. Intell.* **2**, 94–103 (2020).
50. Tsang, A. C. H., Demir, E., Ding, Y. & Pak, O. S. Roads to smart artificial microswimmers. *Adv. Intell. Syst.* **2**, 1900137 (2020).
51. Gardi, G., Ceron, S., Wang, W., Petersen, K. & Sitti, M. Microrobot collectives with reconfigurable morphologies, behaviors, and functions. *Nat. Commun.* **13**, 1–14 (2022).
52. Pezzotta, A., Adorisio, M. & Celani, A. Chemotaxis emerges as the optimal solution to cooperative search games. *Phys. Rev. E* **98**, 042401. https://doi.org/10.1103/PhysRevE.98.042401 (2018).
53. Lowe, R. *et al.* Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31st International NeurIPS, NIPS'17*, 6382–6393 (Curran Associates Inc., 2017).

54. Foerster, J., Farquhar, G., Afouras, T., Nardelli, N. & Whiteson, S. Counterfactual multi-agent policy gradients. In *AAAI*, vol. 32, (2018). https://ojs.aaai.org/index.php/AAAI/article/view/11794.
55. Mushtaq, F. *et al.* Magnetically driven Bi_2O_3/BiOCl-based hybrid microrobots for photocatalytic water remediation. *J. Mater. Chem. A* **3**, 23670–23676 (2015).
56. Wang, X. *et al.* Mofbots: Metal-organic-framework-based biomedical microrobots. *Adv. Mater.* **31**, 1901592 (2019).

Acknowledgements
A.C.H.T. acknowledges funding support from the Croucher Foundation. O.S.P. acknowledges funding support by the National Science Foundation (Grant No. 1830958). The computations were performed using research computing facilities offered by Information Technology Services, The University of Hong Kong.

Author contributions
Y.L., Z.Z., O.S.P., and A.C.H.T performed research; Y.L., Z.Z., O.S.P., and A.C.H.T analyzed data; and Y.L., Z.Z., O.S.P., and A.C.H.T wrote the paper.

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/s41598-023-36305-y.

Correspondence and requests for materials should be addressed to O.S.P. or A.C.H.T.

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.