ARTIFICIAL INTELLIGENCE

Reinforcement learning with artificial microswimmers

S. Muiños-Landin[1,2], A. Fischer[1], V. Holubec[3,4], F. Cichos[1]*

Artificial microswimmers that can replicate the complex behavior of active matter are often designed to mimic the self-propulsion of microscopic living organisms. However, compared with their living counterparts, artificial microswimmers have a limited ability to adapt to environmental signals or to retain a physical memory to yield optimized emergent behavior. Different from macroscopic living systems and robots, both microscopic living organisms and artificial microswimmers are subject to Brownian motion, which randomizes their position and propulsion direction. Here, we combine real-world artificial active particles with machine learning algorithms to explore their adaptive behavior in a noisy environment with reinforcement learning. We use a real-time control of self-thermophoretic active particles to demonstrate the solution of a simple standard navigation problem under the inevitable influence of Brownian motion at these length scales. We show that, with external control, collective learning is possible. Concerning the learning under noise, we find that noise decreases the learning speed, modifies the optimal behavior, and also increases the strength of the decisions made. As a consequence of time delay in the feedback loop controlling the particles, an optimum velocity, reminiscent of optimal run-and-tumble times of bacteria, is found for the system, which is conjectured to be a universal property of systems exhibiting delayed response in a noisy environment.

INTRODUCTION

Living organisms adapt their behavior according to their environment to achieve a particular goal. Information about the state of the environment is sensed, processed, and encoded in biochemical processes in the organism to provide appropriate actions or properties. These learning or adaptive processes occur within the lifetime of a generation, over multiple generations, or over evolutionarily relevant time scales. They lead to specific behaviors of individuals and collectives. Swarms of fish or flocks of birds have developed collective strategies adapted to the existence of predators (*1*), and collective hunting may represent a more efficient foraging tactic (*2*). Birds learn how to use convective air flows (*3*). Sperm have evolved complex swimming patterns to explore chemical gradients in chemotaxis (*4*), and bacteria express specific shapes to follow gravity (*5*).

Inspired by these optimization processes, learning strategies that reduce the complexity of the physical and chemical processes in living matter to a mathematical procedure have been developed (*6*). Many of these learning strategies have been implemented into robotic systems (*7*–*9*). One particular framework is reinforcement learning (RL), in which an agent gains experience by interacting with its environment (*10*). The value of this experience relates to rewards (or penalties) connected to the states that the agent can occupy. The learning process then maximizes the cumulative reward for a chain of actions to obtain the so-called policy. This policy advises the agent which action to take. Recent computational studies, for example, reveal that RL can provide optimal strategies for the navigation of active particles through flows (*11*–*13*), the swarming of robots (*14*–*16*), the soaring of birds (*3*), or the development of collective motion (*17*). The ability of how fish can harness the vortices in the flow field of others for energy-efficient swimming has been explored (*18*). Strategies of how to optimally steer active particles in a potential energy landscape (*19*) have been explored in simulations, and deep Q-learning approaches have been suggested to navigate colloidal robots in an unknown environment (*20*).

Artificial microswimmers are a class of active materials that integrate the fundamental functionality of persistent directed motion, common to their biological counterparts, into a user-designed microscopic object (*21*). Their motility has already revealed insights into a number of fundamental processes, including collective phenomena (*22*–*24*), and they are explored for drug delivery (*25*) and environmental purposes (*26*). However, the integration of energy supply, sensing, signal processing, memory, and propulsion into a micrometer-sized artificial swimmer remains a technological challenge (*27*). Hence, external control strategies have been applied to introduce sensing and signal processing, yet only schemes with rigid rules simulating specific behaviors have been developed (*28*–*31*). Combining elements of machine learning and real-world artificial microswimmers would considerably extend the current computational studies into real-world applications for the future development of smart artificial microswimmers (*32*).

Here, we incorporate algorithms of RL with external control strategies into the motion of artificial microswimmers in an aqueous solution. While the learning algorithm is running on a computer, we control a real agent acting in a real world subjected to thermal fluctuations, hydrodynamic and steric interactions, and many other influences. In this way, it is possible to include real-world objects in a simulation, which will help to close the so-called reality gap, i.e., the difference of pure in silico learning and real-world machine learning even at microscopic length scales (*27*). Our experimental investigation thus goes beyond previous purely computational studies (*3*, *11*–*13*, *20*). It allows us to observe the whole learning process optimizing parameters, which are not accessible in studies of biological species, to identify the most important ingredients of the real dynamics and to set up more realistic, but still simple, models based on this information. It also provides a glimpse of the challenges of RL for objects at those length scales for future developments.

[1]Molecular Nanophotonics Group, Peter Debye Institute for Soft Matter Physics, Universität Leipzig, 04103 Leipzig, Germany. [2]AIMEN Technology Centre, Smart Systems and Smart Manufacturing–Artificial Intelligence and Data Analytics Laboratory, Pl. Cataboi, 36418 Pontevedra, Spain. [3]Institute for Theoretical Physics, Universität Leipzig, 04103 Leipzig, Germany. [4]Department of Macromolecular Physics, Faculty of Mathematics and Physics, Charles University, 18000 Prague, Czech Republic.
*Corresponding author. Email: cichos@physik.uni-leipzig.de

RESULTS

Self-thermophoretic microswimmer

To couple machine learning with microswimmers, we used a light-controlled self-thermophoretic microswimmer with surface-attached gold nanoparticles (Fig. 1A and see the Supplementary Materials). For self-propulsion, the swimmer has to break the time symmetry of low Reynolds number hydrodynamics (*33*). This is achieved by an asymmetric illumination of the particle with laser light of 532-nm wavelength. It is absorbed by the gold nanoparticles and generates a temperature gradient along their surface, inducing thermo-osmotic surface flows and lastly resulting in a self-propulsion of the microswimmer suspended in water. The direction of propulsion is set by the vector pointing from the laser position to the center of the particle. The asymmetric illumination is maintained during the particle motion by following the swimmer's position in real time and steering the heating laser (see the Methods section below). As compared with other types of swimmers (*28*, *34*, *35*), this symmetric swimmer removes the time scale of rotational diffusion from the swimmer's motion and provides an enhanced steering accuracy (*36*, *37*) (see the Supplementary Materials).

Gridworld

To show RL with a real-world microscopic agent, we refer to the standard problem of RL, the gridworld. The gridworld problem allows us to have an experimental demonstration while being able to access the problem numerically. We coarse grain a sample region of 30 μm by 30 μm into a gridworld of 25 states (s, 5 × 5), each state having a dimension of 6 μm by 6 μm (Fig. 1B). One of the states is defined as the target state (goal), which the swimmer is learning to reach. The gridworld is surrounded by 24 boundary states according to Fig. 1B. The obtained real-time swimmer position is used to identify the state s in which the swimmer currently resides. To move between states, we define eight actions a. The actions are carried out by placing the heating laser at the corresponding position on the circumference of the particle (see Fig. 1C). A sequence of actions defines an episode in the gridworld, which ends when the swimmer either leaves the gridworld to a boundary state or reaches the target state. During an episode, rewards or penalties are given. Specifically, the microswimmer gets a reward once it reaches the target state and a penalty in other cases (see the Supplementary Materials for details on the reward definitions). The reward function R thus only depends on the state s, i.e., $R = R(s)$.

RL implementation

We have implemented the model-free Q-learning algorithm to find the optimal policy that solves the navigation problem (*38*). The gained experience of the agent is stored in the Q-matrix (*10*), which tracks the utilities of the different actions a in each state s. When the swimmer transitions between two states s and s' (see the Supplementary Materials for details on the choice of the next state), the Q-matrix is updated according to

$$Q_{t+\Delta t}(s,a) = Q_t(s,a) + \alpha\left[R(s') + \gamma \max_{a'} Q_t(s',a') - Q_t(s,a)\right] \quad (1)$$

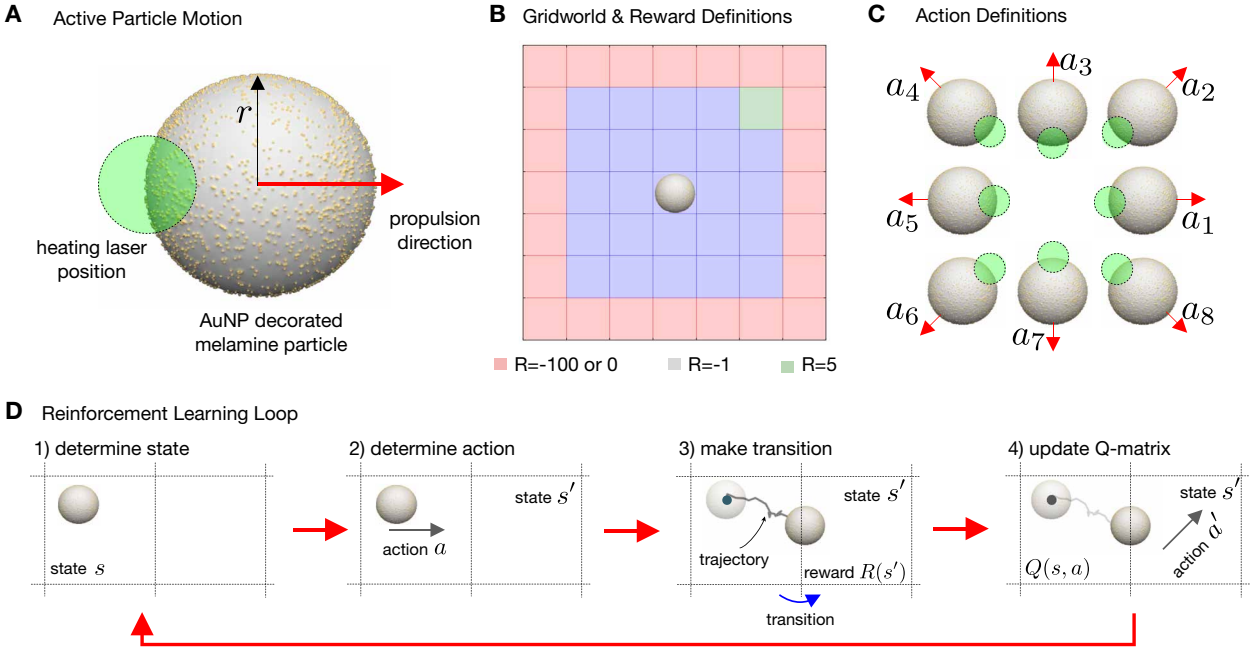

Fig. 1. Gold nanoparticle–decorated microswimmer, states, and actions. (**A**) Sketch of the self-thermophoretic symmetric microswimmer. The particles used have an average radius of r = 1.09 μm and were covered on 30% of their surface with gold nanoparticles of about 10 nm diameter. A heating laser illuminates the colloid asymmetrically (at a distance d from the center), and the swimmer acquires a well-defined thermophoretic velocity $\boldsymbol{v}$. (**B**) The gridworld contains 25 inner states (blue) with one goal at the top right corner (green). A set of 24 boundary states (red) is defined for the study of the noise influence. (**C**) In each of the states, we consider eight possible actions in which the particle is thermophoretically propelled along the indicated directions by positioning the laser focus accordingly. (**D**) The RL loop starts with measuring the position of the active particle and determining the state. For this state, a specific action is determined with the ϵ greedy procedure (see the Supplementary Materials for details). Afterward, a transition is made, the new state is determined, and a reward for the transition is given. On the basis of this reward, the Q-matrix is updated, and the procedure starts from step 1 until an episode ends by reaching the goal or exiting the gridworld to a boundary state.

taking into account the reward $R(s')$ of the next state, the utility of the next state $Q_t(s', a')$ after taking the best action a', and the current utility $Q_t(s, a)$. The influence of these values is controlled by two factors, the learning rate α and the discount factor γ. The learning rate defines the fraction at which new information is incorporated into the Q-matrix, and the discount factor determines the value of future events into the learning process. The reward function is the only feedback signal that the system receives to figure out what it should learn. The result of this RL procedure is the optimal policy function $\pi^*(s) \rightarrow a$, which represents the learned knowledge of the system, $\pi^*(s) = \mathrm{argmax}_a Q(s, a)$, $Q(s, a) = \lim_{t\to\infty} Q_t(s, a)$. Figure 1D highlights the experimental procedure of actuating the swimmer and updating the Q-matrix. As compared with computer models solving the gridworld with deterministic agents, there are four important differences to note. (i) The swimmer can occupy all positions within each state of 6 μm by 6 μm size. It can be arbitrarily close to the boundary. (ii) The swimmer moves in several steps through each state before making a transition. A swimmer velocity of $v = 3\ \mu\mathrm{m\ s}^{-1}$ leads to a displacement of about 6 μm within 2 s, corresponding to about 11 frames at an inverse frame rate $\Delta t_{\mathrm{exp}} = 180$ ms until a transition to the next state is made. (iii) The new state after a transition does not have to be the state that was targeted by the actions. The microswimmers are subject to Brownian motion with a measured diffusion coefficient of $D = 0.1\ \mu\mathrm{m}^2\ \mathrm{s}^{-1}$. The trajectory is therefore partially nondeterministic. With this respect, the system we consider captures a very important feature of active matter on small length scales that is inherent to all microscopic biological systems, where active processes have been optimized to yield robust functions in a noisy background. (iv) Due to a time delay in the feedback loop controlling the active particles, the action applied to the swimmer is not determined from its present position but from its position in the past, which is a common feature for all living and nonliving responsive systems.

Learning process

Figure 2 summarizes the learning process of our microswimmer for boundary states with $R = 0$ and a velocity of $v_{\parallel} = 3.0\ \mu\mathrm{m\ s}^{-1}$, $v_{\parallel} = \langle \Delta\mathbf{r} \cdot \mathbf{e}_{\parallel} \rangle / \Delta t_{\mathrm{exp}}$ where $\langle \Delta\mathbf{r} \cdot \mathbf{e}_{\parallel} \rangle$ is the mean projected displacement of the swimmer along the direction of the action $\mathbf{e}_{\parallel}$. Over the course of more than 5000 transitions (more than 400 episodes, about 7 hours

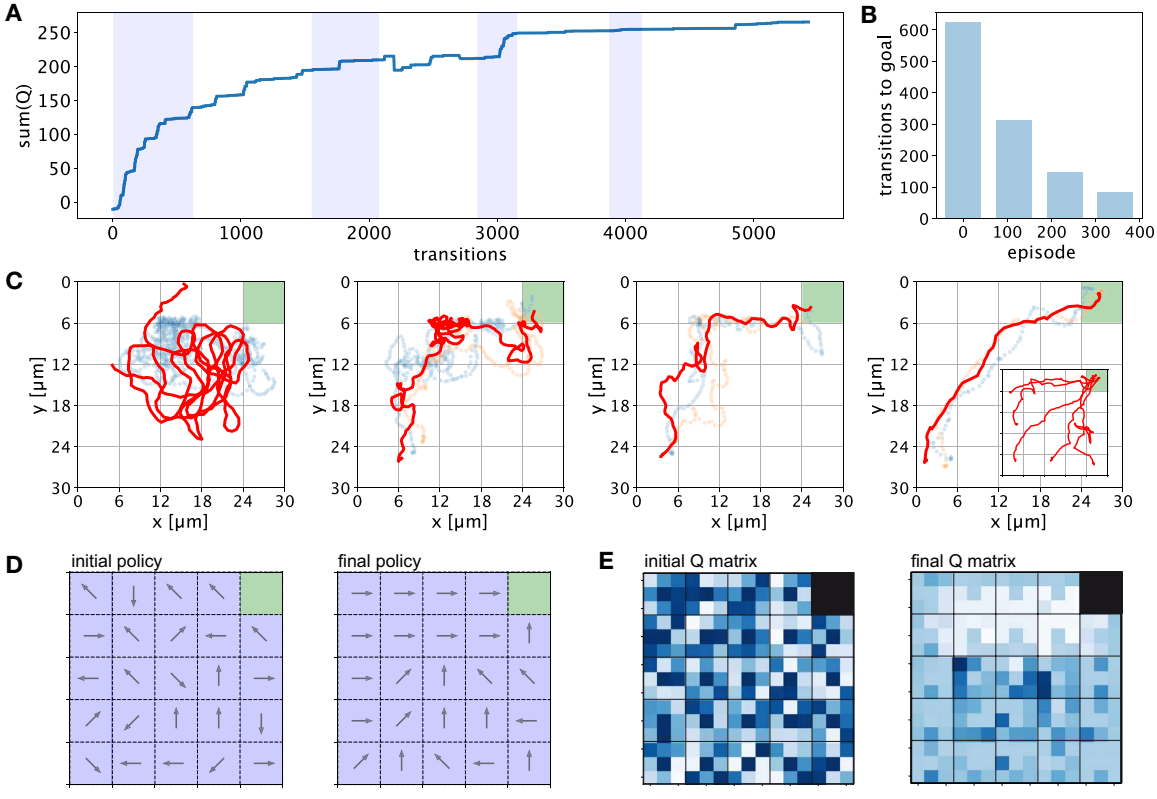

Fig. 2. Single microswimmer learning. (**A**) Learning progress for a single microswimmer in a gridworld at a velocity of $v_{\parallel} = 3.0\ \mu\mathrm{m\ s}^{-1}$. The progress is quantified by the sum of all Q-matrix elements at each transition of the learning process. The Q-matrix was initialized randomly. The shaded regions denote a set of 25 episodes in the learning process, where the starting point is randomly chosen. (**B**) Mean number of steps required to reach the target when starting at the lower left corner as the number of the learning episodes increases. (**C**) Different examples of the behavior of a single microswimmer at different stages of the learning process. The first example corresponds to a swimmer starting at the beginning of the learning process at an arbitrary position in the gridworld. The trajectory is characterized by a large number of loops. With an increasing number of learning episodes, the trajectories become more persistent in their motion toward the goal. This is also reflected by the decreasing average number of steps taken to reach the goal [see (B)]. The inset in the rightmost graph reveals trajectories from different starting positions. (**D**) Policies $\pi(s) = \mathrm{argmax}_a Q_t(s, a)$ defined by the Q-matrix before ($Q_t(s, a) = Q_0(s, a)$) and after ($Q_t(s, a) = Q(s, a)$) the convergence of the learning process. (**E**) Color representation of the initial and the final Q-matrix for the learning process. The small squares in each state represent the utility of the corresponding action (same order as in Fig. 1C) given by its Q-matrix entry, except for the central square. Darker colors show smaller utility, and brighter colors show a better utility of the corresponding action.

of experiment), the sum of all Q-matrix entries converges (Fig. 2A). During this time, the mean number of transitions to reach the goal state decreases from about 600 transitions to less than 100 transitions (Fig. 2B). Accordingly, the trajectories of the swimmer become more deterministic, and the swimmer reaches the goal state independent of the initial state (Fig. 2C and inset). As a result of the learning process, the initial random policy is changing into a policy driving the swimmer toward the goal state. In this respect, the final policy provides an effective drift field with an absorbing boundary at the goal state (Fig. 2D). During this process, which correlates the actions of neighboring cells, the average projected velocity $v_{\parallel}$ causing the drift toward the goal also increases. Although the obtained policy is reflecting the best actions only, the Q-matrix shown in Fig. 2E provides the cumulative information that the swimmer obtained on the environment. It delivers, for example, also information on how much better the best action in a state has been as compared with the other possible actions. The representation in Fig. 2E encodes the Q-matrix value in the brightness of eight squares at the boundary of each state (center square has no meaning). Brighter colors thereby denote larger Q-matrix value.

Because our gridworld is overlayed to the real-world sample, we may also define arbitrary obstacles by providing penalties in certain regions. Figure 3 (A and B) shows examples for trajectories and policies where the particles have been trained to reach a goal state close to a virtual obstacle. Similarly, real-world obstacles can be inserted into the sample to prevent the particle from accessing specific regions and thus realizing certain actions. More complex applications

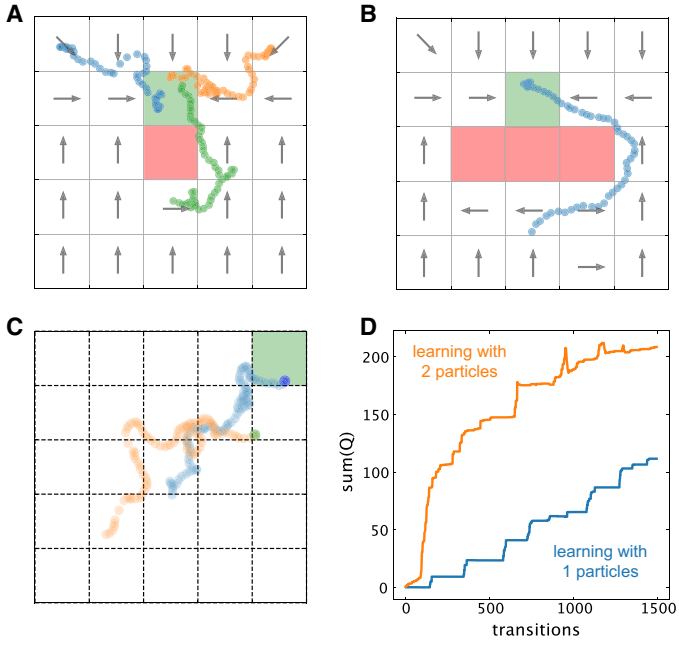

Fig. 3. Learning with obstacles and shared information. (**A**) Example trajectories for a learning process with a virtual obstacle (red square, $\boldsymbol{R} = -100$) next to the goal state ($\boldsymbol{R} = 5$) in the center of the gridworld. (**B**) Example trajectory for an active particle that has learned to reach a goal state ($\boldsymbol{R} = 5$) behind a large virtual obstacle (red rectangle, $\boldsymbol{R} = -100$). (**C**) Example trajectories for two particles sharing information during the learning process. The same rewards as in Fig. 2 have been used. (**D**) Sum of all Q-matrix elements at each transition comparing the learning speed with two particles sharing the information. In all the panels, the active particle speed during the learning process has been $\boldsymbol{v}_{\parallel} = 3.0\,\mu\mathrm{m\,s^{-1}}$.

can involve the emergence of collective behavior, where the motion of multiple agents is controlled simultaneously (*30*). Different levels of collective and cooperative learning may be addressed (*14*, *39*). A true collective learning is carried out when the swimmer is taking an action to maximize the reward of the collective, not only its individual one. Swimmers may also learn to act as a collective when positive rewards are given if an agent behaves like others in an ensemble (*17*). This mimics the process of developing swarming behavior implicated, for example, by the Vicsek model (*40*). Our control mechanism is capable of addressing multiple swimmers separately such that they may also cooperatively explore the environment. Instead of a true collective strategy, we are considering a low density of swimmers (number of swimmers ≪ number of states), which share the information gathered during the learning process by drawing their actions from and updating the same Q-matrix. The swimmers are exploring the same gridworld in different spatial regions, and thus, a speedup of the learning is expected. Figure 3C displays the trajectories of two particles sharing the same Q-matrix, which is updated in each learning step. As a result, the learning speed is enhanced (Fig. 3D). The proposed particle control therefore provides the possibility to explore a collective learning or the optimization of collective behavior and thus delivers an ideal model system with real physical interactions.

Influence of thermal fluctuations on the learning process

A notable difference between macroscopic agents, like robots, and microscopic active particles is the Brownian motion of microswimmers. There is an intrinsic positional noise present in the case of active particles, which is also of relevance for small living organisms like bacteria, cells, and all active processes on microscopic length scales. The advantage of the presented model system, however, is that the influence of the strength of the noise can be explored for the adaption process and the final behavior, whereas this is difficult to achieve in biological systems.

The importance of the noise in Brownian systems is commonly measured by the Peclet number, $\mathrm{Pe} = rv/2D$, comparing the product of particle radius r and the deterministic particle displacement $v\Delta t$ to the corresponding square displacements by Brownian motion $2D\Delta t$. To explore the influence of the noise strength, we change the speed of the active particle v, whereas the strength of the noise is given by the constant diffusion coefficient D. We further introduce a penalty in the boundary states $R = -100$ to modify the environment in a way that the influence of noise can introduce quantitative consequences for the transitions.

When varying the speed $v_{\parallel}$ between 2 and 5 $\mu\mathrm{m\,s^{-1}}$, we make four general observations. (i) Due to time delay in the feedback loop controlling the particles, the noise influence depends on the particle speed nonmonotonously (Fig. 4E and the Supplementary Materials). As a result, we find an optimal particle speed for which the noise is least important, as discussed in more detail in the following section. For the parameters used in the experiment, the optimal velocity is close to the maximum speed available. When increasing the speed in the limited interval of the experiment, the importance of the noise thus decreases. (ii) The Q-matrix converges considerably faster for higher particle speeds corresponding to a lower relative strength of the noise. This effect is intuitive because the stronger the noise, the lower the correlation between action and desired outcome. Figure 4A shows the convergence of the sum of the Q-matrix elements (summed over all entries for a given transition) for different

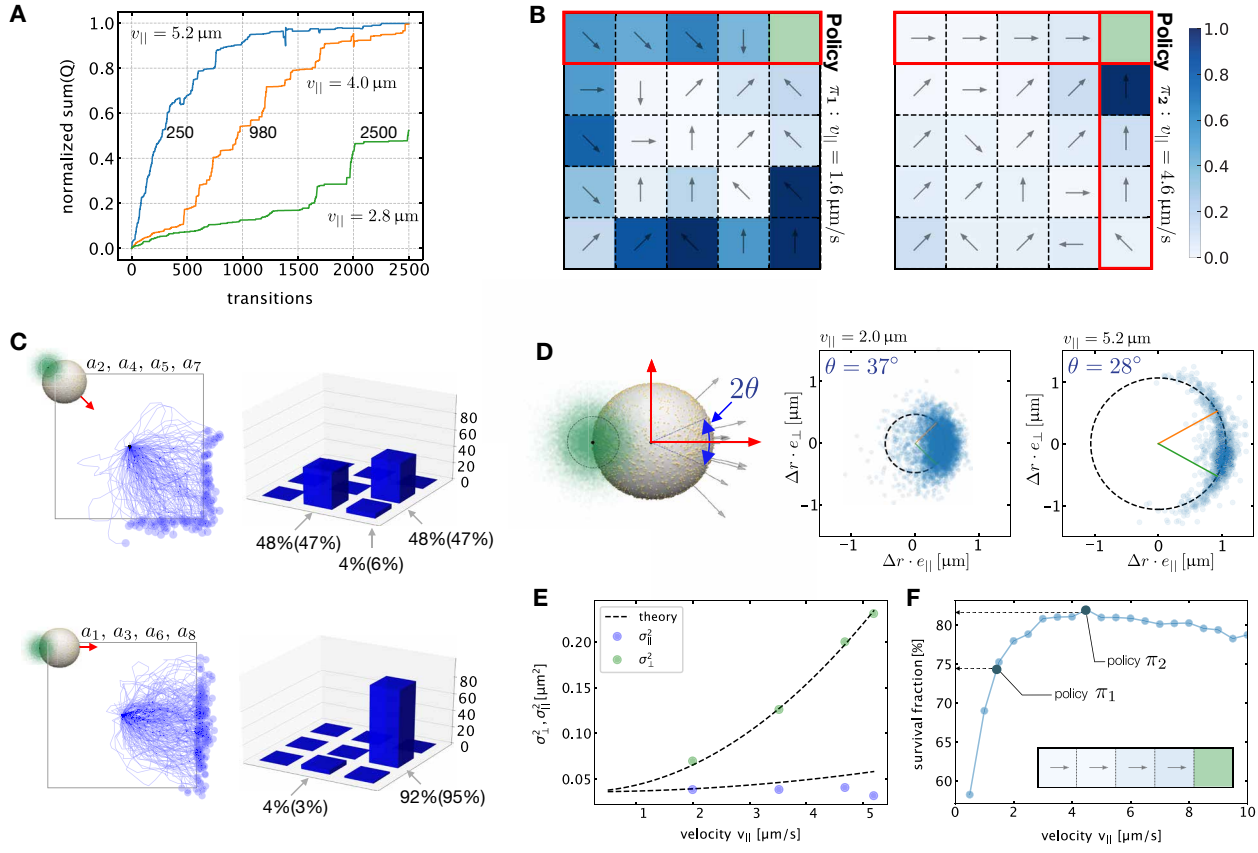

Fig. 4. Influence of Brownian motion on the learning process. (**A**) Sum of the Q-matrix elements as a function of the total number of transitions during the learning process. The different curves were obtained for learning with three different microswimmer speeds. (**B**) Policy obtained from learning processes at high noise (low velocity) ($\pi_1 : v_\parallel = 1.6\,\mu m\,s^{-1}$) and low noise (high velocity) ($\pi_2 : v_\parallel = 4.6\,\mu m\,s^{-1}$). The coloring of the states corresponds to the contrast between the value of the best action and the average of all other actions (Eq. 2). (**C**) Transition probabilities used in Bellman's Eq. 3 for diagonal and nondiagonal actions as determined from experiments with 500 trajectories for a velocity of 1.6 and 4.6 $\mu m\,s^{-1}$. The blue lines indicate example experimental trajectories, which yield equivalent results for actions a_2, a_4, a_5, a_7 (top) and a_1, a_3, a_6, a_8 (bottom). The blue dots mark the first point outside the grid cell. The histograms to the right show the percentage arriving in the corresponding neighboring states. The numbers below denote the percentages for the two velocities (value in parentheses for higher velocity). (**D**) Origin of directional uncertainty. The green dots indicate the possible laser position due to the Brownian motion of the particle within the delay time δt. The two graphs to the right display the experimental particle displacements of a single microswimmer within the delay time $\delta t = \Delta t_{exp} = 180$ ms, when starting at the origin for two different particle velocities. (**E**) Variances of the point clouds in (D) parallel and perpendicular to the intended direction of motion. The dashed lines correspond to the theoretical prediction according to Eq. 4 for the perpendicular motion ($\sigma_\perp^2$) and $\sigma_\parallel^2 = 2D\delta t + (\cosh(\sigma_\theta^2) - 1)\, v_\parallel^2\, \delta t^2$ for the tangential motion with $\sigma_\theta^2 \approx 0.23\ rad^2$, $D = 0.1\,\mu m^2\,s^{-1}$, and $t = \delta t = 180$ ms. (**F**) Survival fraction of particles moving in the upper states at the boundary toward the goal state in policy π_2 indicated in the inset. The survival has been determined from simulations for the same parameters as in (E).

microswimmer speeds ($v_\parallel = 2.8\,\mu m\,s^{-1}$, $v_\parallel = 4.0\,\mu m\,s^{-1}$, and $v_\parallel = 5.1\,\mu m\,s^{-1}$). Although the sum reaches 50% after 250 transitions for the highest velocity, this requires almost 10 times more transitions at about half the speed. (iii) The resulting optimal policy depends on the noise strength. In Fig. 4B, we show the policies obtained for two different velocities ($v_\parallel = 1.6\,\mu m\,s^{-1}$ and $v_\parallel = 4.6\,\mu m\,s^{-1}$). Differences in the two policies are, in particular, visible in the states close to the boundary. Most of the actions at the top and right edge of the low-velocity policy point inward, whereas actions parallel to the edge are preferred at the higher velocity (see highlighted regions in Fig. 4, B and C). (iv) The contrast between the best action and the average of the other actions, which we take as a measure of the decision strength, is enhanced upon increasing importance of the noise. This contrast for a given state s_k is measured by

$$\Delta G(s_k) = \frac{1}{\Psi}\left\{Q(s_k, a^b) - \langle Q(s_k, a_i)\rangle_i\right\} \qquad (2)$$

where a^b denotes the best action for the state and $\langle Q(s_k, a_i)\rangle_i = \sum_{i=1}^{8} Q(s_k, a_i)/8$. The result is normalized by a factor Ψ to make the largest contrast encoded in the color of the states in Fig. 4B equal to one.

DISCUSSION

Because the environment (gridworld with its rewards) stays constant for all learning processes at different velocities, all our above observations for varying particle speed are related to the importance of the noise strength. According to Bellman's equation (*10*)

$$Q(s,a) = \sum_{s'} P(s' \mid s,a)\,[R(s') + \gamma \max_{a'} Q(s',a')] \qquad (3)$$

the influence of the noise on the learning process is encoded in the transition probabilities $P(s' \mid s, a)$, i.e., the probabilities that an action a in the state s leads to a transition to the state s'. This equation

couples the element $Q(s, a)$ of the optimized Q-matrix, corresponding to a state s and action a, with the discounted elements $\pi^*(s') = \max_{a'} Q(s', a')$ of the optimal policy in the future states s' and the corresponding future rewards $R(s')$, weighted by transition probabilities $P(s' \mid a, s)$. Using this equation, one can obtain the Q-matrix and the optimal policy by a Q-matrix value iteration procedure if the transition probabilities are known. The transition probabilities thus contain the physics of the motion of the active particle, including the noise, and decide how different penalties or rewards of the neighboring states influence the value of Q.

We have measured the transition function for the two types of transitions (diagonal and nondiagonal) using 500 trajectories in a single grid cell. To obtain the transition function, we set the starting position of all the trajectories to the center of the grid cell, carried out the specific action, and determined the state in which the particle trajectory ended up. The results are shown in Fig. 4C with exemplary trajectories and a histogram to the right. The numbers below the histograms show the corresponding transition probabilities to the neighboring state in percent for a velocity of $v_\parallel = 1.6\ \mu m\ s^{-1}$ ($v_\parallel = 4.6\ \mu m\ s^{-1}$ for the values in parentheses). The two velocities show only weak changes in the transition probabilities for the nondiagonal actions, which appear to be responsible for the changes in the policies in Fig. 4B. Carrying out a Q-matrix value iteration confirms the changes in the policy in the marked regions for the measured transition probability range (see the Supplementary Materials).

The advantage of our experimental system is that we can explore the detailed physical behavior of each microswimmer in dedicated experiments. To this end, we find two distinct influences of the Brownian motion as the only noise source on the microswimmers' motion. Figure 4D shows the distribution of microswimmer displacement vectors within a time $\Delta t_{exp} = 180$ ms for two different velocities. Each displacement starts at the origin, and the point cloud reflects the corresponding end points of the displacement vectors. With increasing velocity, the particles increase their step length in the desired horizontal direction. The mean distance corresponds to the speed of the particle, and the end points are located close to a circle. At the same time, a directional uncertainty is observed where the angular variance σ_θ^2 is nearly constant for all speeds (see the Supplementary Materials for details). This directional noise is the result of a delayed action in the experiments (*30*, *41*), i.e., a time separation between sensing (imaging the position of the particle) and action on the particle position (placing the laser for propulsion). Both are separated by a delay time δt, which is the intrinsic delay of the feedback loop ($\delta t = \Delta t_{exp} = 180$ ms in our experiments). A delayed response is a very generic feature of all active responsive systems, including biological species. In the present case of a constant propulsion speed, it leads to an anisotropic noise. In the direction perpendicular to the intended action, the Brownian noise gets an additional component that is increasing nonlinearly with the particle speed, whereas the noise along the intended direction of motion is almost constant (Fig. 4E).

The increase in the variance perpendicular to the direction of motion can be analyzed with a simple model (see the Supplementary Materials for details), which yields

$$\sigma_\perp^2 = v_\parallel^2\, \delta t\, \sinh(\sigma_\theta^2)\, t + 2Dt \tag{4}$$

and corresponds well with experimental data (Fig. 4E) for $\sigma_\theta^2 \approx 0.23\ rad^2$ and fixed time $t = \delta t$. In particular, it captures the nonlinear increase of $\sigma_\perp^2$ with the particle speed v.

The increase has important consequences. When considering the motion in the top four states of policy π_2 (Fig. 4B), the particle would move horizontally toward the goal starting at an arbitrary position in the leftmost state. From all trajectories that started, only a fraction will arrive at the goal state before leaving these states through the upper, lower, or left boundaries of those four states. This survival fraction has been determined from simulations (also see the Supplementary Materials for an approximate theoretical description). Overall, a change between the two policies π_1 and π_2 is induced by an increase of the survival by less than 10% when going from $v_\parallel = 1.6\ \mu m\ s^{-1}$ to $v_\parallel = 4.6\ \mu m\ s^{-1}$. When further increasing the velocity, we find in simulations that an optimal velocity for maximum survival exists. This maximum corresponds to the minimum

$$v_{\parallel\,opt} = \sqrt{\frac{2D}{\sinh(\sigma_\theta^2)\,\delta t}} \tag{5}$$

in the variance (Eq. 4) for a fixed traveled distance $a = v_\parallel t$, which only depends on the diffusion coefficient D, the angular variance σ_θ^2, and the sensorial delay δt (see the Supplementary Materials for details). In the limit of instantaneous actions ($\delta t = 0$), an infinitely fast motion would yield the best results. Any nonzero delay will introduce a "speed limit" at which a maximum survival is ensured. We expect that the optimal policy for very high velocities should yield a similar policy as for low velocities. An experimental verification of this conjecture is currently out of reach, as Fig. 4F shows the results of the simulations.

The observed behavior of the survival probability, which exhibits a maximum for a certain particle velocity, implies that the probability to reach the target is maximal for the same optimal velocity. Moreover, because the underlying analysis is solely based on the competition of two noises omnipresent in (Brownian) active matter, namely the diffusion and the uncertainty in choosing the right direction, we conjecture that the observed type of behavior is universal. The precision of reaching the target (long time variance of the distance from the target) by the run-and-tumble motion of bacteria exhibits a minimum as a function of the run-and-tumble times (*42*, *43*) reminiscent of our results. These results also demonstrate that the combination of machine learning algorithms with real-world microscopic agents can help to uncover physical phenomena (such as time delay in the present work), which play important roles in the microscopic motion of biological species.

Concluding, we have demonstrated RL with a self-thermophoretic microswimmer carrying out actions in a real-world environment with its information processing and sensing capabilities externalized to a computer and a microscopy setup. Already with this hybrid solution, one obtains a model system, where strategies in a noisy environment with virtual obstacles or collective learning can be explored. Although our simple realization of a gridworld is based on a global position detection defining the state of the swimmer, future applications will consider local information, e.g., the response to a temporal sequence of local physical or chemical signals, to allow for navigation in unknown environments. As compared with a computer simulation, our system contains a nonideal control limited by the finite reaction time of the feedback loop, presence of liquid flows, imperfections of the swimmers or sample container, hydrodynamic interactions, or other uncontrolled parameters that naturally influence the learning process. In this way, it resembles a new form of computer simulation using real-world agents. An important advantage is that

the physics of the agent can be explored experimentally in detail to understand the learned strategies, and the real-world interactions in more complex environments can be used to adapt the microswimmer's behavior. In that sense, even the inverse problem of using the learned strategy to reveal the details of these uncontrolled influences may be addressed as a new form of environmental sensing. Similarly, the control of active particles by machine learning algorithms may be used in evolutionary robotics (*8*, *44*), where the interaction of multiple particles may be optimized to yield higher-order functional structures based on environmental interactions. Although the implementation of signaling and feedback by physical or chemical processes into a single artificial microswimmer is still a distant goal, the current hybrid solution opens a whole branch of new possibilities for understanding adaptive behavior of single microswimmers in noisy environments and the emergence of collective behavior of large ensembles of active systems.

MATERIALS AND METHODS

Materials

Samples consisted of commercially available gold nanoparticle–coated melamine resin particles of a diameter of 2.19 μm (microParticles GmbH, Berlin, Germany). The gold nanoparticles were covering about 30% of the surface area and were between 8 and 30 nm in diameter (see the Supplementary Materials for details.) Microscopy glass cover slides were dipped into a 5% Pluronic F127 solution, rinsed with deionized water, and dried with nitrogen. The Pluronic F127 coating prevented sticking of the particles to the glass cover slides. Two microliters of particle suspension was placed on the cover slides to spread about an area of 1 cm by 1 cm, forming a 3-μm-thin water film. The edges of the sample were sealed with silicone oil (polydimethylsiloxane) to prevent water evaporation.

Methods

Samples were investigated in a custom-built inverted dark-field microscopy setup based on an Olympus IX-71 microscopy stand. The sample was held by a Piezo stage (Physik Instrumente) that was mounted on a custom-built stepper stage for coarse control. The sample was illuminated by a halogen lamp (Olympus) using a dark-field oil-immersion condenser [Olympus, numerical aperture (NA), 1.2]. The scattered light was collected by an oil-immersion objective lens (Olympus, 100×, NA 1.35 to 0.6) with the NA set to 0.6 and captured with an Andor iXon emCCD camera. A $\lambda = 532$ nm laser was focused by the imaging objective into the sample plane to serve as a heating laser for the swimmers. Its position in the sample plane was steered by an acousto-optic deflector (AOD; AA Opto-Electronic) together with a 4-f system (two $f = 20$ cm lenses). The AOD was controlled by an ADwin realtime board (ADwin-Gold, Jäger Messtechnik) exchanging data with a custom LabVIEW program. A region of interest of 512 pixels by 512 pixels (30 μm by 30 μm) was used for the real-time imaging, analysis, and recording of the particles, with an exposure time of $\Delta t_{exp} = 180$ ms. The details of integrating the RL procedure are contained in the Supplementary Materials.

SUPPLEMENTARY MATERIALS

robotics.sciencemag.org/cgi/content/full/6/52/eabd9285/DC1
Fig. S1. Symmetric swimmer structure.
Fig. S2. Swimmer speed as a function of laser power.
Fig. S3. Directional noise as function of the swimming velocity measured in the experiment.
Fig. S4. Directional noise model.
Fig. S5. Results of the analytical model of the influence of the noise.
Fig. S6. Q-matrix value iteration result.
Movie S1. Single-swimmer free navigation toward a target during learning.
Movie S2. Single-swimmer free navigation toward a target after learning.
Movie S3. Navigation toward a target with virtual obstacles.
Movie S4. Multiple-swimmer free navigation toward a target.

REFERENCES AND NOTES

1. J. K. Parrish, W. M. Hamner, *Animal Groups in Three Dimensions* (Cambridge Univ. Press, 1997).
2. Y. Lin, N. Abaid, Collective behavior and predation success in a predator-prey model inspired by hunting bats. *Phys. Rev. E* **88**, 062724 (2013).
3. G. Reddy, A. Celani, T. J. Sejnowski, M. Vergassola, Learning to soar in turbulent environments. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E4877–E4884 (2016).
4. J. A. Kromer, S. Märcker, S. Lange, C. Baier, B. M. Friedrich, Decision making improves sperm chemotaxis in the presence of noise. *PLOS Comput. Biol.* **14**, e1006109 (2018).
5. B. ten Hagen, F. Kümmel, R. Wittkowski, D. Takagim, H. Löwen, C. Bechinger, Gravitaxis of asymmetric self-propelled colloidal particles. *Nat. Commun.* **5**, 4829 (2014).
6. L. P. Kaelbling, M. L. Littman, A. W. Moore, Reinforcement learning: A survey. *J. Artif. Intell. Res.* **4**, 237–285 (1996).
7. J. Kober, J. Peters, Reinforcement learning in robotics: A survey, in *Learning Motor Skills* (Springer Tracts in Advanced Robotics, 2014), vol. 97, pp. 9–67.
8. S. Doncieux, N. Bredeche, J. B. Mouret, A. E. Eiben, Evolutionary robotics: What, why, and where to. *Front. Robot. AI* **2**, 4 (2015).
9. M. Wiering, M. v. Otterlo, Reinforcement Learning, in *Adaptation, Learning, and Optimization* (Springer Berlin Heidelberg, 2012), vol. 12.
10. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, 1998).
11. S. Colabrese, K. Gustavsson, A. Celani, L. Biferale, Flow navigation by smart microswimmers via reinforcement learning. *Phys. Rev. Lett.* **118**, 158004 (2017).
12. K. Gustavsson, L. Biferale, A. Celani, S. Colabrese, Finding efficient swimming strategies in a three-dimensional chaotic flow by reinforcement learning. *Eur. Phys. J. E* **40**, 110 (2017).
13. J. K. Alageshan, A. K. Verma, J. Bec, R. Pandit, Machine learning strategies for path-planning microswimmers in turbulent flows. *Phys. Rev. E* **101**, 043110 (2020).
14. H. M. La, R. Lim, W. Sheng, Multirobot cooperative learning for predator avoidance. *IEEE Trans. Control Syst. Technol.* **23**, 52–63 (2015).
15. M. Birattari, A. Ligot, D. Bozhinoski, M. Brambilla, G. Francesca, L. Garattoni, D. Garzón Ramos, K. Hasselmann, M. Kegeleirs, J. Kuckling, F. Pagnozzi, A. Roli, M. Salman, T. Stützle, Automatic off-line design of robot swarms: A manifesto. *Front. Robot. AI* **16**, 59 (2019).
16. L. Pitonakova, R. Crowder, S. Bullock, Information flow principles for plasticity in foraging robot swarms. *Swarm Intell.* **10**, 33–63 (2016).
17. K. Ried, T. Müller, H. J. Briegel, Modelling collective motion based on the principle of agency: General framework and the case of marching locusts. *PLOS ONE* **14**, e0212044 (2019).
18. S. Verma, G. Novati, P. Koumoutsakos, Efficient collective swimming by harnessing vortices through deep reinforcement learning. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 5849–5854 (2018).
19. E. Schneider, H. Stark, Optimal steering of a smart active particle. *EPL* **127**, 64003 (2019).
20. Y. Yang, M. A. Bevan, B. Li, Efficient navigation of colloidal robots in an unknown environment via deep reinforcement learning. *Adv. Intell. Syst.* **2**, 1900106 (2020).
21. C. Bechinger, R. Di Leonardo, H. Löwen, V. Volpe, V. Volpe, Active particles in complex and crowded environments. *Rev. Mod. Phys.* **88**, 045006 (2016).
22. J. Palacci, S. Sacanna, A. P. Steinberg, D. J. Pine, P. M. Chaikin, Living crystals of light-activated colloidal surfers. *Science* **339**, 936–940 (2013).
23. I. Buttinoni, J. Bialké, F. Kümmel, H. Löwen, C. Bechinger, T. Speck, Dynamical clustering and phase separation in suspensions of self-propelled colloidal particles. *Phys. Rev. Lett.* **110**, 238301 (2013).
24. A. Aubret, M. Youssef, S. Sacanna, J. Palacci, Targeted assembly and synchronization of self-spinning microgears. *Nat. Phys.* **14**, 1114–1118 (2018).
25. Y. Wu, X. Lin, Z. Wu, H. Möhwald, Q. He, Self-propelled polymer multilayer Janus capsules for effective drug delivery and light-triggered release. *ACS Appl. Mater. Interfaces* **6**, 10476–10481 (2014).
26. M. Safdar, J. Simmchen, J. Jänis, Light-driven micro- and nanomotors for environmental remediation. *Environ. Sci. Nano* **4**, 1602–1616 (2017).
27. F. Cichos, K. Gustavsson, B. Mehlig, G. Volpe, Machine learning for active matter. *Nat. Mach. Intell.* **2**, 94–103 (2020).
28. B. Qian, D. Montiel, A. Bregulla, F. Cichos, H. Yang, Harnessing thermal fluctuations for purposeful activities: The manipulation of single micro-swimmers by adaptive photon nudging. *Chem. Sci.* **4**, 1420–1429 (2013).
29. A. P. Bregulla, H. Yang, F. Cichos, Stochastic localization of microswimmers by photon nudging. *ACS Nano* **8**, 6542–6550 (2014).

30. U. Khadka, V. Holubec, H. Yang, F. Cichos, Active particles bound by information flows. *Nat. Commun.* **9**, 3864 (2018).
31. F. A. Lavergne, H. Wendehenne, T. Bäuerle, C. Bechinger, Group formation and cohesion of active particles with visual perception-dependent motility. *Science* **364**, 70–74 (2019).
32. A. C. H. Tsang, E. Demir, Y. Ding, O. S. Pak, Roads to smart artificial microswimmers. *Adv. Intell. Syst.* **2**, 1900137 (2020).
33. E. M. Purcell, Life at low Reynolds number. *Am. J. Phys.* **45**, 3–11 (1977).
34. H.-R. Jiang, N. Yoshinaga, M. Sano, Active motion of a Janus particle by self-thermophoresis in a defocused laser beam. *Phys. Rev. Lett.* **105**, 268302 (2010).
35. I. Buttinoni, G. Volpe, F. Kümmel, G. Volpe, C. Bechinger, Active Brownian motion tunable by light. *J. Phys. Condens. Matter* **24**, 284129 (2012).
36. M. Selmke, U. Khadka, A. P. Bregulla, F. Cichos, H. Yang, Theory for controlling individual self-propelled micro-swimmers by photon nudging I: Directed transport. *Phys. Chem. Chem. Phys.* **20**, 10502–10520 (2018).
37. M. Selmke, U. Khadka, A. P. Bregulla, F. Cichos, H. Yang, Theory for controlling individual self-propelled micro-swimmers by photon nudging II: Confinement. *Phys. Chem. Chem. Phys.* **20**, 10521–10532 (2018).
38. J. C. H. Watkins, thesis, King's College, Cambridge (1989).
39. L. Busoniu, R. Babuška, B. De Schutter, Multi-agent reinforcement learning: A survey, in *Proceedings of the 9th International Conference on Control, Automation, Robotics and Vision (ICARCV 2006)* (Singapore, 2006), pp. 527–532.
40. T. Vicsek, A. Czirók, E. Behn-Jakob, I. Cohen, O. Shochet, Novel type of phase transition in a system of self-driven particles. *Phys. Rev. Lett.* **75**, 1226–1229 (1995).
41. D. Geiss, K. Kroy, V. Holubec, Brownian molecules formed by delayed harmonic interactions. *New J. Phys.* **21**, 093014 (2019).
42. P. Romanczuk, G. Salbreux, Optimal chemotaxis in intermittent migration of animal cells. *Phys. Rev. E* **91**, 042720 (2015).
43. A. Diz-Muñoz, P. Romanczuk, W. Yu, B. Bergert, K. Ivanovitch, G. Salbreux, C.-P. Heisenberg, E. K. Paluch, Steering cell migration by alternating blebs and actin-rich protrusions. *BMC Biol.* **14**, 74 (2016).
44. S. Jones, A. F. Winfield, S. Hauert, M. Studley, Onboard evolution of understandable swarm behaviors. *Adv. Intell. Syst.* **1**, 1900031 (2019).

Acknowledgments: Helpful discussion with P. Romanczuk is acknowledged in pointing out observations of directional noise for biological systems. Fruitful discussion and help with extrapolating the theory to the experiments by K. Ghazi-Zahedi are acknowledged. We thank A. Kramer for helping to revise the manuscript. **Funding:** The authors acknowledge financial support by the DFG Priority Program 1726 "Microswimmers" through project 237143019. F.C. is supported by the DFG grant 432421051. V.H. is supported by a Humboldt grant of the Alexander von Humboldt Foundation and by the Czech Science Foundation (project no. 20-02955J). **Author contributions:** F.C. conceived the research. S.M.-L. and F.C. designed the experiments. S.M.-L. implemented the system, and S.M.-L. and A.F. performed the experiments. S.M.-L., V.H., and F.C. analyzed and discussed the data. F.C., V.H., and S.M.-L. wrote the manuscript. **Competing interests:** The authors declare that they have no competing financial interests. **Data and materials availability:** All data needed to evaluate the conclusions are available in the paper or in the Supplementary Materials. Additional data and materials are available upon request.

Submitted 21 July 2020
Accepted 26 February 2021
Published 24 March 2021
10.1126/scirobotics.abd9285

Citation: S. Muiños-Landin, A. Fischer, V. Holubec, F. Cichos, Reinforcement learning with artificial microswimmers. *Sci Robot.* **6**, eabd9285 (2021).

Science Robotics

Reinforcement learning with artificial microswimmers

S. Muiños-Landin, A. Fischer, V. Holubec, and F. Cichos

Sci. Robot. **6** (52), eabd9285. DOI: 10.1126/scirobotics.abd9285

Science Robotics (ISSN 2470-9476) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Robotics* is a registered trademark of AAAS.