Reinforcement learning for pursuit and evasion of microswimmers at low Reynolds number[*]

Francesco Borra,[1, †] Luca Biferale,[2] Massimo Cencini,[3, ‡] and Antonio Celani[4, §]

[1] *Dipartimento di Fisica, Università "Sapienza" Piazzale A. Moro 5, I-00185 Rome, Italy*

[2] *Department of Physics and INFN, University of Rome Tor Vergata, 00133 Rome, Italy*

[3] *Istituto dei Sistemi Complessi, CNR, 00185 Rome, Italy and INFN "Tor Vergata"*

[4] *Quantitative Life Sciences, The Abdus Salam International*
Centre for Theoretical Physics - ICTP, Trieste, 34151, Italy

We consider a model of two competing microswimming agents engaged in a pursue-evasion task within a low-Reynolds-number environment. Agents can only perform simple maneuvers and sense hydrodynamic disturbances, which provide ambiguous (partial) information about the opponent's position and motion. We frame the problem as a zero-sum game: The pursuer has to capture the evader in the shortest time, while the evader aims at deferring capture as long as possible. We show that the agents, trained via adversarial reinforcement learning, are able to overcome partial observability by discovering increasingly complex sequences of moves and countermoves that outperform known heuristic strategies and exploit the hydrodynamic environment.

[*] Version accepted for publication (postprint) on Phys. Rev. Fluids 7, 023103 (2022) – Published 23 February 2022 DOI:10.1103/PhysRevFluids.7.023103

[†] Present address: Laboratory of Physics of the Ecole Normale Supérieure, CNRS UMR8023, 24 rue Lhomond, 75005 Paris, France

[‡] Corresponding author; massimo.cencini@cnr.it

[§] Corresponding author; celani@ictp.it

I. INTRODUCTION

Aquatic organisms can detect moving objects by sensing the induced hydrodynamic disturbances [1–3]. Such an ability is crucial in prey-predator interactions and for navigation, especially in murky or dark waters, as for the blind Mexican cavefish [4]. Fishes have developed the lateral line, a mechanosensory system very sensitive to water motions and pressure gradients [5–7]. Planktonic microorganisms, inhabiting a low-Reynolds-number environment, have antennae and setae to sense hydrodynamic signals produced by predators and preys [8, 9].

Abstracting away from specific mechanisms developed by aquatic organisms, the problem of pursue-evasion in microswimmers guided by hydrodynamic cues poses substantial difficulties rooted in the physics of the ambient medium. At low Reynolds numbers, flow disturbances are generally weak and rich of symmetries [10] leading to ambiguities about the signal source location especially if distant from the receiver [2, 3, 8]. Moreover, hydrodynamics has dynamical effects, as the disturbances generated by one microswimmer alter the other motion. Consequently, an agent's strategy inevitably affects the opponent dynamics and strategies. It is thus crucial to understand how agents' strategies co-evolve by competing against one another [11], which necessitates going beyond just escaping from a prescribed pursuit strategy or pursuing a nonresponsive moving target [12, 13]. Which pursuit-evasion strategies can be devised in such dynamic, partially observable environments? How do they coevolve while competing? Can hydrodynamics be exploited and how? How do they compare with strategies based on visual cues?

Here, we formulate the problem of prey-predator microswimmers in a game-theoretic framework [14], a natural setting to model the emergence of adversarial strategies [11]. As we are interested in the learning and evolution of strategies and not in fine tuning on specific details of the two microswimmers, we choose a simplified hydrodynamics. Inspired by recent applications of multi agent reinforcement Learning (MARL) [15] to hide-and-seek contests [16, 17], we explore its use as a general model-free framework for discovering effective chase-and-escape strategies at low Reynolds number. Reinforcement learning (RL) approaches rely on trial and error to improve the quality of the decisions made by an agent – here a microswimmer – and has been already applied in numerical and experimental study of navigation in complex fluid environments [18–27]. We show that RL is able to discover complex strategies, evolving during the different phases of the adversarial learning, and thus depending on the combined training history. The discovered strategies efficiently overcome the limitations imposed by the partial observability. In particular, pursuer strategies are shown to outperform a heuristic baseline policy. Moreover, we show that the main strategies discovered by

RL are explainable and for some of them we provide an analytical description, which allows us to rationalize how the pursuer overcomes the difficulties due to partial information.

The material is organized as follows. In Sec. II we present the model. In Sec. III we discuss the basic ideas of reinforcement learning applied to our model and some detail on the implementation. In Sec. IV we present the results, while Sec. V is devoted to discussions and conclusions. Some more technical material is presented in the Appendices and details on the numerical implementation of the reinforcement learning algorithm are discussed in supplementary material [28].

II. MODEL

A. Game theoretic formulation

The basic settings of the game-theoretic formulation of the problem are shown in Figs. 1(a) and 1(b). Agents have a limited maneuverability and partial information on the opponent via hydrodynamic cues, which we choose to be the gradients of the velocity field [Fig. 1(a)]. The two swimming agents play the following zero-sum game [Fig. 1(b)]: They start at distance R_0 with random heading directions. At each decision time τ each agent senses the hydrodynamic field and chooses an action (steer left/right or go straight). The pursuer (p) aims at reaching the capture distance R_c from the evader (e) in the shortest possible time, while the latter has to keep the pursuer at bay (at distance $R > R_c$). The game terminates either on capture (pursuer wins) or if its duration exceeds a given time T_{max} (evader wins). While playing many games the agents are trained via via reinforcement learning (see Sec. III)

B. Modeling the agents

For simplicity, we model the agents as “pusher” discoids in an idealized two-dimensional environment disregarding any effect due to walls or confinements and in the absence of external flows. By swimming, they generate a velocity field modeled as a force dipole moving with speed v_α with $\alpha = e, p$ [Fig. 1(a)]. The force-dipole approximates well the far field of many microorganisms [29]. Near-field corrections, depending on details, are not implemented as we are not interested in tuning the model to a specific swimming mechanism, though they can matter in close encounters [30]. Besides self-propulsion each microswimmer is advected and reoriented by the flow generated by the other. Every τ time units, i.e., at each decision time, agents can steer by imparting a torque, resulting in an angular velocity Ω_α. Thus the position $\boldsymbol{x}_\alpha$ and heading direction $\boldsymbol{n}_\alpha = (\cos\theta_\alpha, \sin\theta_\alpha)$

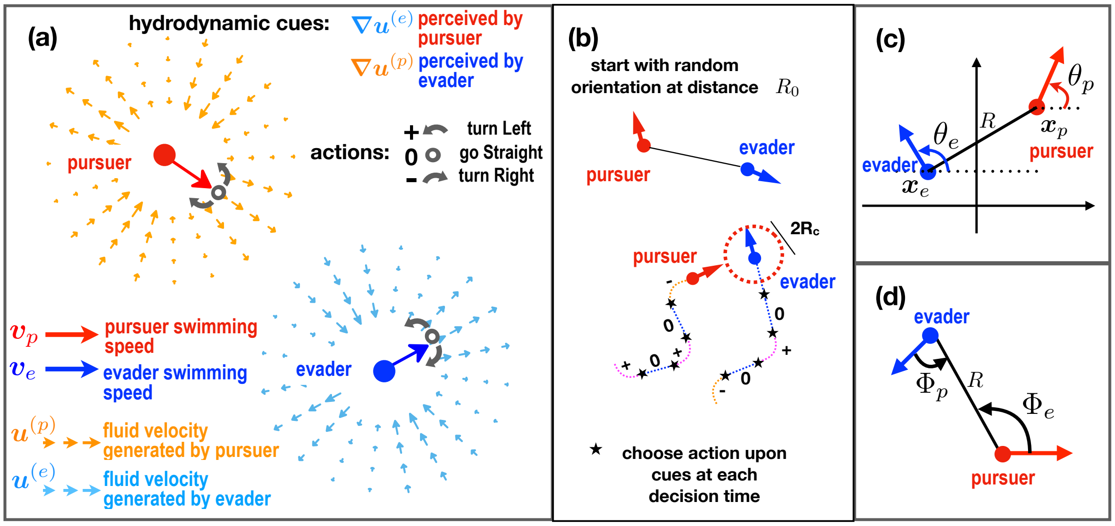
(a)
hydrodynamic cues:
$\nabla u^{(e)}$ perceived by pursuer
$\nabla u^{(p)}$ perceived by evader
pursuer
actions:
+ turn Left
0 go Straight
- turn Right
v_p pursuer swimming speed
v_e evader swimming speed
evader
$u^{(p)}$ fluid velocity generated by pursuer
$u^{(e)}$ fluid velocity generated by evader
(b)
start with random orientation at distance R_0
pursuer
evader
2Rc
pursuer
evader
0
+
-
★ choose action upon cues at each decision time
(c)
θ_p
R
x_p
pursuer
θ_e
evader
x_e
(d)
evader
Φ_p
R
Φ_e
pursuer

FIG. 1. (Color online) Model illustration. (a) Basic elements: The pursuer (p, red)/evader (e, blue) swims with speed $v_{p/e}$ generating a velocity field $\boldsymbol{u}^{(p/e)}$, which drags the other agent and offers a cue to the other agent on the relative position and orientation via its gradients, $\boldsymbol{\nabla u}^{(p/e)}$. Agents only have a limited control on their heading directions – the actions. (b) Sketch of a game: The game starts with the agents at distance R_0 and the pursuer/evader goal is to min/maximize the time their distance reaches the capture value R_c within a given time horizon. Agents move in the plane, every τ time-unit they choose to maintain or turn left/right their heading direction on the basis of the cues they receive. (c) Geometry of the problem in a fixed frame of reference with indicated the heading angles. (d) Bearing angle $\Phi_{e/p}$ corresponding to the angular position of an agent with respect to the heading direction of its opponent.

evolve as

$$\dot{\boldsymbol{x}}_\alpha = v_\alpha \boldsymbol{n}_\alpha + \boldsymbol{u}^{(\beta)} \tag{1}$$

$$\dot{\theta}_\alpha = \Omega_\alpha + \omega^{(\beta)}/2\,, \tag{2}$$

where $\boldsymbol{u}^{(\beta)}(\boldsymbol{x})$ and $\omega^{(\beta)}(\boldsymbol{x})(= \boldsymbol{\nabla}\times\boldsymbol{u}^{(\beta)}(\boldsymbol{x}))$ are the velocity and vorticity field at position at position $\boldsymbol{x}_\alpha$, generated by the opponent agent β in $\boldsymbol{x}_\beta$ with heading orientation θ_β, see Fig. 1c.

As detailed in Appendix A, we can write $\boldsymbol{u}^{(\beta)}(\boldsymbol{x}) = (\partial_y, -\partial_x)\Psi(\boldsymbol{x}-\boldsymbol{x}_\beta;\theta_\beta)$, with the stream function $\Psi = {}^{D_\beta}\!/_2 \sin(2\phi - 2\theta_\beta)$, where $\boldsymbol{x}-\boldsymbol{x}_\beta = |\boldsymbol{x}-\boldsymbol{x}_\beta|(\cos\phi,\sin\phi)$ and D_β denotes the dipole intensity of agent β. We consider $D_\beta > 0$, i.e., pusher like microswimmers [29]. The velocity in Eq. (1) is then obtained by deriving the stream function in $\boldsymbol{x} = \boldsymbol{x}_\alpha$, corresponding to $\phi = \phi_\alpha$ [see Fig. 5(a) for the notation on the angles with respect to a fixed frame of reference]. While the vorticity in Eq. (2) is $\omega^{(\beta)} = 2D_\beta/R^2 \sin(2\phi_\alpha - 2\theta_\beta) = 2D_\beta/R^2 \sin(2\phi_\beta - 2\theta_\beta)$, with $R = |\boldsymbol{x}_\alpha - \boldsymbol{x}_\beta|$ and the second equality stemming from $\phi_\beta = \phi_\alpha + \pi$ [Fig. 5(a)].

C. Modeling the hydrodynamic cues

As already discussed, we assume an agent can only sense the gradients of the velocity field generated by its opponent, similarly to what copepods do with sensory setae [8]. Since agents have no notion of an external frame of reference we assume that they perceive the gradients in their own frame of reference, i.e., projected along the angents' swimming direction. In this frame of reference the three independent components (vorticity and longitudinal and shear strain) of the velocity gradients read

$$\omega^{(\beta)} = \partial_x u_y^{(\beta)} - \partial_y u_x^{(\beta)} = \frac{2D_\beta}{R^2}\sin(2\Phi_\beta - 2\Theta_\beta) \tag{3}$$

$$\mathcal{L}^{(\beta)} = \partial_x u_x^{(\beta)} = -\partial_y u_y^{(\beta)} = -\frac{D_\beta}{R^2}\cos(4\Phi_\beta - 2\Theta_\beta) \tag{4}$$

$$\mathcal{S}^{(\beta)} = \frac{1}{2}(\partial_x u_y^{(\beta)} + \partial_y u_x^{(\beta)}) = -\frac{D_\beta}{R^2}\sin(4\Phi_\beta - 2\Theta_\beta)\,. \tag{5}$$

They depend on agents' distance (R), relative heading $\Theta_\beta = \theta_\beta - \theta_\alpha$ [see Fig. 5(b)] and angular position of β with respect the heading direction of α, $\Phi_\beta = \phi_\beta - \theta_\alpha$, i.e., the bearing angle [Fig. 1(d)], as called in the pursuit-evasion-games language [13].

In general, gradients are symmetric with respect to parity, i.e., to the combined transformation $\Theta_\beta \to \Theta_\beta + \pi$ and $\Phi_\beta \to \Phi_\beta + \pi$. The force dipole case is even more degenerate as, owing to the fore-aft symmetry, either of the two transformations leaves the gradients unchanged, due to the nematic nature of dipoles. Such symmetries result in ambiguities in the identification of the position and orientation of the opponent, akin to the 180^o ambiguity in fish hearing [31]. Memory of past detections and/or multiple hydrodynamical cues can, in principle, mitigate such ambiguities which, however, typically persist at large distances [1, 2, 32]. In spite of its simplicity the model is thus rich enough to represent the typical observability limitation inherent to organisms that can only perceive the gradients of the velocity field.

III. LEARNING TO PURSUE AND EVADE THROUGH REINFORCEMENT

To set up a learning framework, we need to identify: a set of observations, o, that an agent perceives and uses to infer the opponent's state; the actions, a, through which it can implement its strategy; and the rewards, r, to evaluate its actions. The learning task here is to find an optimal reactive policy, $\pi^*(a|o)$, that associates actions to observations in order to maximize the expected cumulative rewards. In our setting, the environmental state (relative position and heading) is only partially observable through the velocity gradients[33]. The actions $a \in \mathcal{A} = \{0, +, -\}$ [Fig. 1(a)] correspond to the three angular velocities $\Omega_\alpha = 0, +\varpi_\alpha, -\varpi_\alpha$ agent α can choose to control its orientation. Once actions are taken, the agents evolve for a time τ with the dynamics (1) and (2) and a reward is issued. In this zero-sum game, the currency is the elapsing time: The pursuer/evader receives a reward $r = \mp 1$ at the end of each decision time. After each action, the agents update their policy by combining past and new information with the issued reward. In the new state, gradients are sensed again, new actions are taken and rewards received; the cycle repeats itself until the terminal state is achieved, with either the pursuer (if $R \leq R_c$) or the evader winning (if the game duration exceeds T_{max}). The total return accumulated by the pursuer/evader in an episode is therefore $\mp T$ where T is the duration of the episode itself.

A. Reinforcement Learning algorithm

Among the many approaches to MARL we adopt a natural actor-critic architecture (see [34, 35] and supplementary material [28] for details) because of its theoretical guarantees and connection with evolutionary game theory [14, 36]. In this class of algorithms, locally optimal solutions are sought by means of stochastic gradient ascent in policy space. Natural gradients are used, by virtue of their covariance with respect to the metric defined by the Fisher information [37]. Real organisms process the environmental cues with their nervous system that encodes the policy, e.g., in fishes dedicated neurons control escape responses [38]. Such neural encoding can be emulated by artificial neural networks [39]. Here, in the interest of explainability, we opted for an explicit parametrization of the policy in terms of few selected features of the observations. Dropping the agent indices for simplicity, we set

$$\pi(a|o) = \frac{\exp(\boldsymbol{\mathcal{F}}(o) \cdot \boldsymbol{\xi}_a)}{\sum_{a'} \exp(\boldsymbol{\mathcal{F}}(o) \cdot \boldsymbol{\xi}_{a'})} ,$$

where $a, a' \in \mathcal{A}$, $\boldsymbol{\mathcal{F}}(o)$ are features that encode the observations o, and $\boldsymbol{\xi}_a$ the learning parameters.

By combining the velocity-gradient components, we chose to extract the following observables (o) (see Appendix B): the vorticity ω; a proxy for the agents' distance, $\hat{R} \propto 1/R^2$; and a linear combination of heading and bearing angle $\gamma = 4\Phi - 2\Theta$. As features, $\boldsymbol{\mathcal{F}}(o)$, we used the raw observables ω and $\hat{R}$, and the first and second harmonics of angle γ. To encode for the heading direction, we include some short-term memory by combining a few past observations. As discussed in Appendix B, exploratory studies with more features did not give qualitatively different results from the minimal setting described above. Moreover, eliminating memory yields the same strategies, which indicates that a more sophisticated exploitation of memory is needed.

B. Training scheme

To better interpret the evolution of strategies and counter-strategies, we organized learning in phases (each made of $M = 5 \times 10^3$ episodes) where agents alternately improve their policies. Assuming no prior knowledge, agents start their training with a random policy, $\pi(a|o) = 1/|\mathcal{A}| = 1/3$ for all o. At first, the pursuer learns with the evader's policy frozen, and then the evader learns against the pursuer policy from the previous phase, and so on. Episodes start with agents at a distance $R_0 = 1$ and random heading directions, and end either on capture ($R \leq R_c = 0.05R_0$) or when time exceeds the cap $T_{max} = 50T_0$, where $T_0 = R_0/v_e$ is estimated in terms of the evader speed and initial distance. We fixed the evader speed at $v_e = 0.1$ and angular velocity $\varpi_e = 3$. For the pursuer, we chose $(v_p, \varpi_p) = (0.15, 4.5)$ which gives a slight speed advantage maintaining the same steering ability (same curvature radius $v_p/\varpi_p = v_e/\varpi_e$). The intensity of the force dipole is taken to be equal for both agents $D_p = D_e = 0.03$. With this choice hydrodynamic velocity dominates over swimming at distances $R \lesssim R_0$. The decision time is $\tau = 0.01T_0$ for both agents.

IV. RESULTS

Our main results are summarized in Fig. 2: Figure 2(a) shows the running average of normalized game duration T/T_{max} [40] in the first six learning phases for three independent learning experiments; Figs. 2(b)-2(g) and Figs. 2(h)-2(m) display some representative examples of pursuer and evader winning strategies, respectively. Cycles 1 and 2 are quite reproducible: The pursuer discovers ways to rapidly catch the evader which, in turn, finds ways to counteract. Conversely, cycles 3-6 are characterized by a higher variability: Agents seem to acquire and lose good policies also within their own learning turn, and we see cases (run1 in Fig. 2a) in which the evader eventually dominates the game. We hypothesize that such variability arises from a combination of insufficient hyperparameters tuning [41] and/or subtle instabilities in the learning algorithm. Notwithstanding these limitations, many aspects of the learned strategies are reproducible and, to some extent, physically explainable as discussed below. The effect of a variation of the parameters and the addition of rotational noise is discussed in Sec. II of supplementary material [28].

A. Pursuit strategies: Mirroring and tailgating

In its first learning phase, the evader executes a random cue-insensitive policy, while the predator learns to pursue its prey either "mirroring" its actions [Fig. 2(b)] or "tailgating" it [Fig. 2(c)]. When

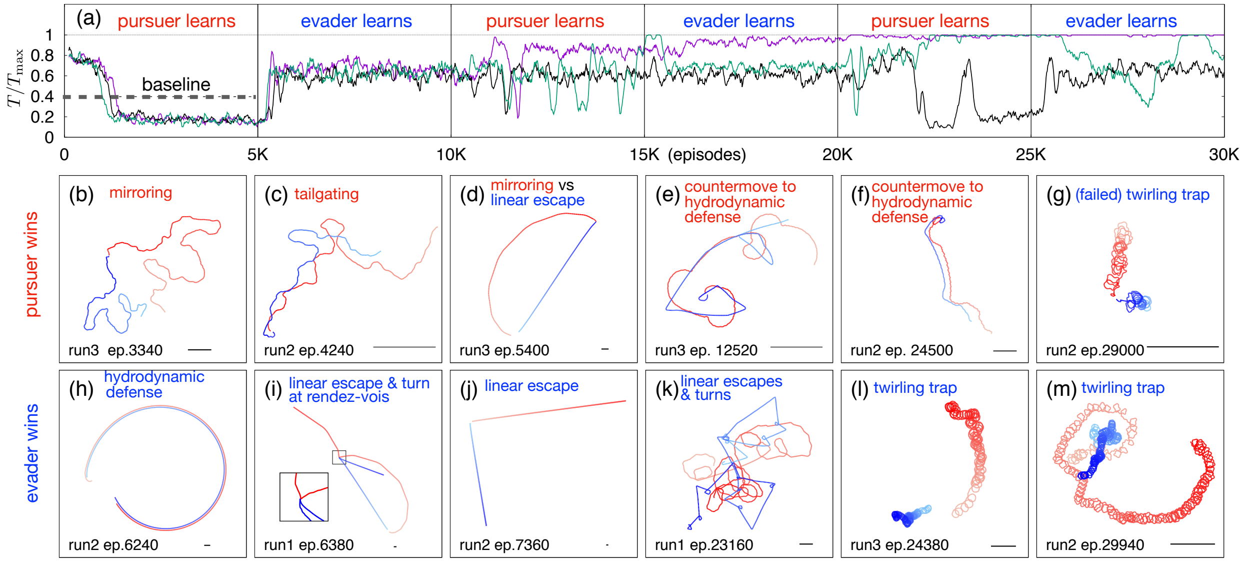

FIG. 2. (Color online) Co-evolving strategies in the first six training cycles. (a) Running average (over 100 episodes) of normalized episode duration T/T_{max} for three realizations of learning: run1, run2 and 3 (purple, green and black curves). The dashed horizontal gray line represents a heuristic baseline value as discussed in Sec. IV A 1. [(b)-(g)] Winning pursuit strategies: (b) mirroring, (c) tailgating, (d) mirroring vs linear escape with a rendez-vous, [(e) and (f)] tailgating with countermoves to hydrodynamic defense, and (g) failing twirling on mirroring. [(h)-(m)] Winning evasion strategies: (h) hydrodynamic defense, (i) linear escape with turn and hydrodynamic collision at rendez-vous, (j) linear escape against mirroring, (k) linear escapes and turns inducing pursuer switches between mirroring and tailgating at distance, and [(l) and (m)] twirling trap. Red/blue denotes pursuer/evader trajectories, time runs from lighter to darker color (apparent close encounters actually take place at different times); run/episode labeled on each panel; the bottom-right bar displays the unit length.

the pursuer approaches the evader, a switch between the two strategies can sometimes be observed presumably due to hydrodynamical effects overcoming self-swimming at these distances combined to evader turning [Fig. 3]. Close inspection reveals that the pursuer orchestrates its actions in such a way to enforce over time specific relations (linked to the hydrodynamical cues as discussed in Appendix C) between the bearing angles, namely $\Phi_e = -\Phi_p$ for mirroring and $\Phi_e = -\Phi_p + \pi$ for tailgating [Fig. 3(a)]. Due to the aforementioned 180^o ambiguities, the pursuer cannot discern mirroring and tailgating just on the basis of instantaneous hydrodynamical cues: The strategy chosen depends on initial conditions and hydrodynamic interactions [as, e.g., in Fig. 3(b)]. The two strategies emerge from the same policy in response to partial observability and can be analytically described, as detailed in Appendix C and briefly summarized in the following. On neglecting hydrodynamical interactions in Eqs. (1) and (2), we can derive the equations for the separation

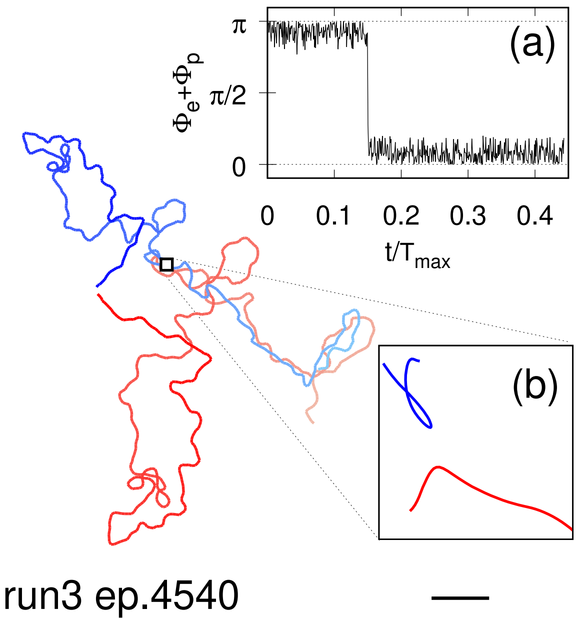

FIG. 3. (Color online) Switching between tailgating to mirroring. Inset (a): Sum of bearing angles $\Phi_e + \Phi_p$ vs normalized time. Proximity and evader turning [inset (b)] triggers the switch $\Phi_e + \Phi_p \approx \pi \to 0$ (tailgating → mirroring) at $t/T_{max} \approx 0.15$.

and bearing angle [42]. By imposing that the pursuer follows either mirroring or tailgating, such equations read

$$\dot{R} = -(v_p \pm v_e) \cos \Phi_e \tag{6}$$

$$\dot{\Phi}_e = \Omega_e - R^{-1} \, (v_p \mp v_e) \sin \Phi_e \, , \tag{7}$$

with ± for mirroring/tailgating. Equation (6) shows that tailgating is doomed to fail when $v_p = v_e$ as $\dot{R} = 0$, while for $v_p > v_e$ it becomes an efficient strategy as the dynamics (7) leads to $\Phi_e \to 0$ for small enough distances, and (6) implies $\dot{R} < 0$. Mirroring remains effective also for $v_p = v_e$ (and Ω_e random) as it essentially maps the pursue into a first hitting problem for a random search with dimensionality reduction [43]. Tests with RL and the full dynamics [Eqs. (1) and (2)] for $v_p = v_e$ confirmed the scenario.

1. *Comparison with a heuristic strategy based on visual cues adapted to partial information*

It is interesting to compare the pursuit policies discovered by RL against well-established visual pursuit strategies based on the knowledge of the line of sight with the target [13, 42]. Mirroring bears some similarities with *parallel navigation*, where the line-of-sight direction is kept constant with respect to an inertial frame of reference, a strategy that appears to be applied by dragonflies

[44]. Tailgating resembles *pure pursuit*, where heading is constantly directed toward the line of sight (zero bearing angle), as bats or some fishes appear to do [45, 46]. Such strategies cannot be directly implemented here because of the 180^o ambiguities inherent to perceiving only the gradients. However, we can introduce a heuristic strategy in the form of a randomized pure pursuit: The pursuer heads either toward the evader ($\Phi_e = 0$) or to its "image" ($\Phi_e = \pi$) with equal probability with some persistency in time. As a limiting case, it could randomly choose its target once for all at the beginning, in which case it is bound to fail half of the times so that $\langle T \rangle / T_{max} > 1/2$; however, the pursuer may instead randomly choose either targets every N_p decision times (we call N_p persistency). By scanning $\langle T \rangle / T_{max}$ as a function of N_p, at $N_p \approx 400$ we numerically found the minimum $\langle T \rangle / T_{max} \approx 0.4$ [see Fig. 4 and dashed line in Fig. 2(a)] which is slightly more than twice the value obtained with the mirroring-tailgating strategy. The policy discovered by RL clearly outperforms the randomized pure pursuit offering a more efficient way to overcome the ambiguities due to the partial information provided by the hydrodynamic cues.

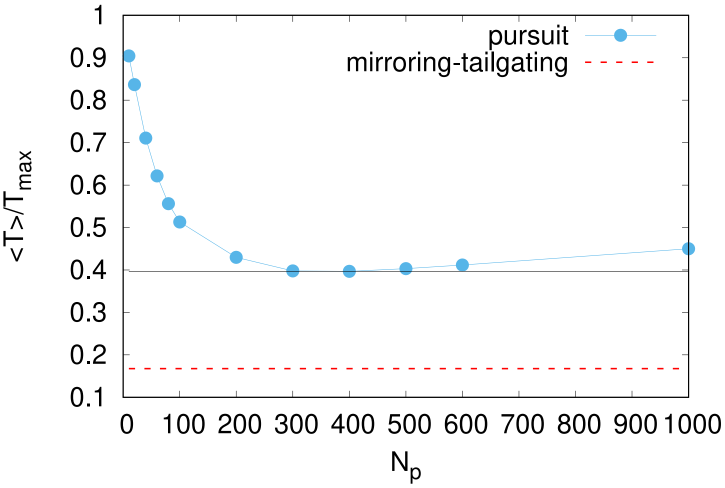

FIG. 4. (Color online) $\langle T \rangle / T_{max}$ as a function of the persistency N_p (circles) of the randomized pure pursuit strategy described in text. The black line shows the normalized episode duration for the optimal N_p, while the red dashed line shows the average value obtained with the mirroring-tailgating strategy obtained from Fig. 2(a).

B. Evader strategies: Hydrodynamic defense and linear flights

In its training phase, the evader learns to contrast mirroring and tailgating. As for the latter, it finds a way to exploit hydrodynamics [Fig. 2(h)]. In many episodes of this kind, the pursuer approaches its opponent from behind with small bearing angle (tailgating). The evader reacts by placing itself in a position relative to its predator such that its backward push cancels the speed advantage of the pursuer and keeps it at bay at a fixed distance (supplementary movie1 displays the pursuer's trajectory in the frame of reference of the evader). In principle, near-field corrections to the force dipole could modify the hydrodynamic defense, and it would be interesting to explore this aspect when focusing on specific microswimmers. Another strategy adopted by the evader takes the form of an almost linear escape trajectory [Figs. 2(i) and 2(j)]. As shown in Fig. 2(d), this is not always successful as, via mirroring, the pursuer can intercept the evader to a *rendezvous* point by performing a long smooth arc. Such arcs correspond to adjusting the axis of mirroring in the course of time. However, either by making such *rendezvous* point very far [Fig. 2(j)] or by exploiting hydrodynamics and turns on close encounters [Fig. 2(i)], the evader can consistently make its evasion strategies quite efficient.

C. Refining strategies

As training proceeds both agents learn more complex strategies in response to the ones described above. We now briefly discuss some examples that stand out because of their repeated occurrence and explainability. Interestingly, the pursuer discovers different ways to contrast the hydrodynamic defense of its opponent [Figs. 2(e) and (f), see also supplementary movie2]. Remarkably, the evader learns to devise diverse winning manouvers as in Fig. 2(k), which consist in linear escapes and turnings which make the predator switching from mirroring to tailgating before capture (see supplementary movie3). The evader also discovers that twirling can trap the pursuer [Fig. 2(l) and 2(m)] in a looping motion induced by its own mirroring-tailgating strategy. Trapping is not always successful though [Fig. 2(g)]. With small variations, the basic strategic patterns discussed above are found also with different parameter choices and will be reported elsewhere.

V. CONCLUSIONS

In this study, we have shown how microswimmers can discover complex strategies to pursue and evade from each other, even if endowed with limited maneuvering ability and inherently equivocal information about their relative position and orientation. Our study presents a novel game-theoretic approach to pursuit and evasion in an aquatic microenvironment. We expect it to spur further research on the use of reinforcement learning algorithms to rationalize observed prey-predator interactions in more general contexts [11]. Owing to the simplicity of our model we have been able to analytically describe some of the strategies discovered by RL and show why they are effective in overcoming partial observability: For instance, mirroring and tailgating allow to reduce the dimensionality of the search by mapping the search into a first hitting problem. In this respect it would be interesting to study a three dimensional version of the problem to understand which dimensionality reduction could emerge in that case and if it can be still reduced to a one-dimensional hitting time problem.

The present model can be easily generalized to ellipsoidal swimmers, by adding to Eq. (2) rotation by the strain-rate tensor. It can also be extended to "pullers" as well as to other specific microswimmers by including the appropriate near-field hydrodynamics. Indeed while, e.g., mirroring and tailgating are expected to maintain their efficiency in the far field, the pursuer policy may need some refinement in the near-field in order to account for more complex hydrodynamic interactions and near-field corrections would also modify the response of the evader (e.g., the kind of hydrodynamic defense it can develop). With suitable modifications of the hydrodynamics, the approach that we developed here can be used to train underwater robots which can sense the hydrodynamic fields with bioinspired mechanosensors[47, 48] and thus accomplish complex tasks – for instance, artificial fishes that imitate escape responses [49]. Here we did not discuss the effect of external flows and boundaries. Preliminary results in a circular arena confirm that the agents can learn to exploit hydrodynamics to perform their pursue/evasion tasks in spite of the confounding cues and complex dynamics arising from the presence of the walls.

Exciting and formidable challenges still lie ahead, and among them stands out the emergence of collective pursue strategies like wolf-packing, and collective escape responses such as hydrodynamic cloaking [23].

ACKNOWLEDGMENTS

We thanks Simone Pigolotti for useful comments on the manuscript, and Xu Zhuoqun for a very careful reading of our manuscript. F.B. acknowledges hospitality from ICTP. AC has received funding from the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie Grant No. N956457. This work received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant No. 882340).

Appendix A: Force-dipole hydrodynamic fields

As described in Sec. II B, we model the agents as two swimming discoids which generate a force dipole, in the sequel we detail the hydrodynamic fields, which enter the dynamics of the agents [see Eqs. (1) and (2)], generated by a force dipole in two dimensions.

We start considering a Stokeslet, i.e., the fundamental solution of the Stokes equation for a point force, $\boldsymbol{F} = F\boldsymbol{n}$, which, for the sake of simplicity, we locate in the origin, and thus solving the equation

$$\nu \Delta \boldsymbol{u} - \boldsymbol{\nabla} p = \boldsymbol{F}\delta(\boldsymbol{x})\,, \tag{A1}$$

where $\boldsymbol{u}$ and p are the velocity and pressure field, and ν the fluid viscosity. The fundamental solution to Eq. (A1) in two dimensions is

$$u_i(\boldsymbol{x}) = G_{ij}(\boldsymbol{x})F_j\,, \tag{A2}$$

where G is the Green function:

$$G_{ij}(\boldsymbol{x}) = \frac{1}{4\pi\nu}\left[-\delta_{ij}\ln\left(\frac{|\boldsymbol{x}|}{L}\right) + \frac{x_i x_j}{|\boldsymbol{x}|^2}\right] \tag{A3}$$

with L being an arbitrary length. The pressure field takes the form $p(\boldsymbol{x}) = \boldsymbol{F}\cdot\boldsymbol{x}/(4\pi|\boldsymbol{x}|^3) + p_\infty$, with p_∞ a constant.

Considering two point forces $\boldsymbol{F}^{\pm} = \pm F\boldsymbol{n}$ located in $\boldsymbol{x}^{\pm} = \pm\epsilon\boldsymbol{n}$, with $\epsilon \ll 1$ and using Eq. (A2), we can express the velocity field generated by this couple as

$$\boldsymbol{u}(\boldsymbol{x}) = G_{ij}(\boldsymbol{x}-\boldsymbol{x}^{+})F_j^{+} + G_{ij}(\boldsymbol{x}-\boldsymbol{x}^{+})F_j^{-} \simeq -2Fn_k\partial_k G_{ij}(\boldsymbol{x})n_j \tag{A4}$$

where $F_i^{+} = -F_i^{-} = Fn_i$ and we retained only the first order, to obtain an expression which well approximates the velocity field for large distances $|\boldsymbol{x}| \gg \epsilon$. Working out the algebra yields:

$$\boldsymbol{u}(\boldsymbol{x}) = \frac{D}{|\boldsymbol{x}|}\left[2\left(\frac{\boldsymbol{n}\cdot\boldsymbol{x}}{|\boldsymbol{x}|}\right)^2 - 1\right]\frac{\boldsymbol{x}}{|\boldsymbol{x}|} = \frac{D}{|\boldsymbol{x}|}\cos(2\phi-2\theta)(\cos\phi,\sin\phi) \tag{A5}$$

where $D = F\epsilon/(2\pi\nu)$ measures the dipole intensity ($D > 0$ corresponding to pushers and $D < 0$ to pullers [29]) and, in the second equality, $\boldsymbol{x} = |\boldsymbol{x}|(\cos\phi,\sin\phi)$ and $\boldsymbol{n} = \boldsymbol{n}(\theta) = (\cos\theta,\sin\theta)$. Notice that the velocity (A5) can equivalently be derived as $\boldsymbol{u} = (\partial_y\Psi, -\partial_x\Psi)$ where Ψ is the stream function which can be written as

$$\Psi(\boldsymbol{x}) = \frac{D}{2}\sin(2\phi-2\theta)\,. \tag{A6}$$

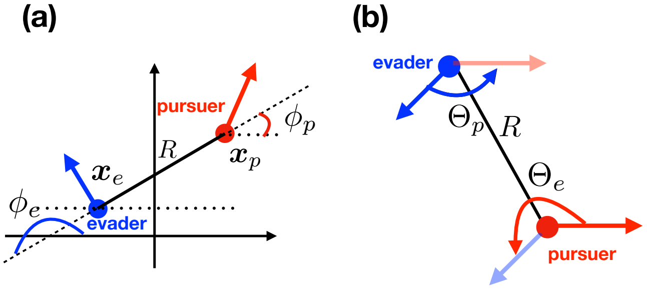

FIG. 5. (Color online) Description of relevant angles entering the agent dynamics. (a) In a fixed frame of reference it is shown the angular position of the pursuer ϕ_p and of the evader ϕ_e, heading angles are shown in Fig. 1(b). (b) We show the relative heading angles $\Theta_e = \theta_e - \theta_p$ and $\Theta_p = \theta_p - \theta_e$, needed together with the bearing angles [Fig. 1(d)] to express the velocity gradients in the pursuer and evader frame of reference, respectively.

The velocity field due to agent β and advecting agent α [see Eq. (1] is simply obtained from Eq. (A5) substituting $\phi = \phi_\alpha$ and $\theta = \theta_\beta$, where θ_β are the heading directions shown in Fig. 1(c), and ϕ_α the angular position with respect to a fixed frame of reference shown in Fig. 5(a). The vorticity field due to agent β and rotating the heading orientation θ_α of agent α instead can be easily derived to be:

$$\omega^{(\beta)} = \frac{2D_\beta}{R^2}\sin(2\phi_\alpha - 2\theta_\beta) = \frac{2D_\beta}{R^2}\sin(2\phi_\beta - 2\theta_\beta) \tag{A7}$$

with $R = |\boldsymbol{x}_\alpha - \boldsymbol{x}_\beta|$ and the second equality stemming from $\phi_\beta = \phi_\alpha + \pi$ [Fig. 5(a)].

Appendix B: Choice of observables and features

Equations (3)-(5) express the gradients in the frame of reference the observing agent. The three independent components of the gradients $\omega, \mathcal{L}$ and $\mathcal{S}$ can be mapped one to one onto the space of the following quantities $o = \{\omega, \hat{R}, \gamma\}$, which we assume are observables. Here, ω is the vorticity itself, which combines information about the agent distance and about $\sin(2\Phi - 2\Theta)$. The other two quantities can easily obtained combining the expression of the longitudinal and shear strain, as $\hat{R} = (\mathcal{L}^2 + \mathcal{S}^2)^{1/2} \propto 1/R^2$ and $\gamma = \arctan2(\mathcal{S}/\mathcal{L})$. The observations o offer partial information about the other agent due to the 180^o ambiguities in Θ and Φ discussed in Sec. II C. The policy is parameterized by the features $\mathcal{F}(o)$, i.e., functions of the observables. Notice that even if o was precisely identifying the reciprocal position of the agents, in order to have access to the full space of possible policies, the features should be chosen as a complete functional basis of the observation, which is not practicable if not using deep reinforcement learning techniques, an option that we did not adopt to have a better understanding of the discovered policies. So we encode the observations o by using a set of only N_F features and, consequently, the agents must decide their actions in condition of *partial observability* [15, 33]).

We tested different choices of the features and the results presented in Fig. 2 correspond to the choice of the $N_F = 13$ features summarized in Table I. While the first six features are clearly related to the information that can be extracted from gradients, the features $i = 7, 12$ are introduced to provide the agents with some memory, which may mitigate some of the aforementioned ambiguities. Finally, the 13^{th} feature is unrelated to the gradients and it is chosen to allow the agents to adopt strategies independent of the percepts.

$\mathcal{F}_1(o_t) = \hat{R}(t)$
$\mathcal{F}_2(o_t) = \sin(\gamma(t))$
$\mathcal{F}_3(o_t) = \cos(\gamma(t))$
$\mathcal{F}_4(o_t) = \sin(2\,\gamma(t))$
$\mathcal{F}_5(o_t) = \cos(2\,\gamma(t))$
$\mathcal{F}_6(o_t) = \omega(t)$
$\mathcal{F}_i(o_t) = (1-\mu)\ \mathcal{F}_i(t-\tau) + \mu\ \mathcal{F}_{i-6}(t-\tau) \quad \text{for} \quad i = 7, 12$
$\mathcal{F}_{13}(o_t) = 1$

TABLE I. Implemented Features. Note that features $7 - 12$ provide some memory of the values of the previous features, in the implementation we chose $\mu = 0.3$ to retain some memory about the last $2 - 4$ past observations approximately).

We remark that the mirroring and tailgating strategies discussed in Sec. IV A can be obtained also removing memory (i.e., features from 7 to 12) while they seem to crucially depend on features 4 and 5 which, as explained before, are derived from the strain components Eqs. (4) and (5). These two features are clearly related to mirroring and tailgating strategies (see also Appendix C). Indeed, by removing them, such basic strategies are lost. In order to assess the robustness of our results, we tested the algorithm with different choices of features. Specifically, we tried using powers (both positive and negative) of $\hat{R}$, higher harmonics of γ and various products of the percepts. For instance, we tried to add the following features $\cos(3\gamma)$, $\sin(4\gamma)$, $\cos(3\gamma)$, $\cos(4\gamma)$, $\omega\cos(\gamma)/\hat{R}$, $\omega\sin(\gamma)/\hat{R}$, $\omega\sin(2\gamma)/\hat{R}$, $\omega\cos(2\gamma)/\hat{R}$, $\hat{R}\cos(\gamma)$, $\hat{R}\sin(\gamma)$, $\hat{R}\sin(2\gamma)$, $\hat{R}\cos(2\gamma)$, $\hat{R}\sin(3\gamma)$, $\hat{R}\cos(3\gamma)$, $|\omega|$ and $\omega/\hat{R}$. In a separate batch of tests, we tried to use $\hat{R}^2$ and $\log\hat{R}$. While it cannot be excluded that we did miss a specific combination of features or that we did not run enough tests, the strategies emerging from these additional trials were qualitative equivalent to those presented in Fig. 2.

Appendix C: Analytical description of mirroring and tailgating strategies

In this Appendix we discuss the mirroring [Fig. 2(b)] and tailgating ([Fig. 2(c)] strategies and derive Eqs. (6) and (7).

First we recall the definitions [see also Figs. 1(c) and 1(d) and Figs. 5(a) and 5(b)] of the relative heading angle $\Theta_e = \theta_e - \theta_p$, bearing angle from the point of view of the pursuer, $\Phi_e = \phi_e - \theta_p$, and of the evader, $\Phi_p = \phi_p - \theta_e$; moreover, we recall that $\phi_e = \phi_p + \pi$ [see Fig. 5(a)]. As discussed in Sec. IV A in these strategies the pursuer chooses its actions in such a way to approximatively keep the following relations between the two bearing angles:

$$\begin{cases} \Phi_p = -\Phi_e & \text{Mirroring} \\ \Phi_p = -\Phi_e + \pi & \text{Tailgating}\,. \end{cases} \tag{C1}$$

As discussed in Appendix B, one of the key observables available to the pursuer is the angle $2\Phi_e - \Theta_e$, with simple algebra one can recognize that

$$2\Phi_e - \Theta_e = \Phi_p + \Phi_e - \pi\,, \tag{C2}$$

so that we can re-express (C1) as

$$2\Phi_e - \Theta_e = \Gamma_\pm \mod 2\pi \tag{C3}$$

or equivalently, in the laboratory frame of reference

$$\theta_p + \theta_e = 2\phi_e - \Gamma_\pm\,, \tag{C4}$$

with $\Gamma_+ = \pi$ for mirroring and $\Gamma_- = 0$ for tailgating respectively. Note that $(\theta_p + \theta_e)/2$ identifies (for $v_p = v_e$ exactly, and otherwise approximatively) the axis of symmetry with respect to which the pursuer trajectory mirrors the one of the evader.

Notice that, at least in the absence of memory, $\Gamma_\pm$ cannot be discriminated from observing gradients alone due to the fore-aft symmetry of the swimming dipole. Therefore, depending on the initial condition the agent will pick one of two strategies. Unless the pursuer can resolve the aforementioned ambiguity, it must learn both strategies or neither. In the actual hydrodynamic simulations we have added memory effects (see features $\mathcal{F}_i$ for $i = 7, 12$ in Table I) to possibly allow the agents to break the fore-aft symmetry in observations. Though tests in the absence of memory suggest that the way memory was implemented is likely not sufficient to eliminate such ambiguities. Anyway, we will ignore possible memory effects in the following.

In order to explore the basic features of such strategies we will neglect also the hydrodynamic effects (i.e., we will not consider the effects on the pursuer due to the velocity field induced by the evader) meaning that we approximate Eqs. (1) and (2) as

$$\dot{\boldsymbol{x}}_\alpha = v_\alpha \boldsymbol{n}(\theta_\alpha) \tag{C5}$$

$$\dot{\theta}_\alpha = \Omega_\alpha \,. \tag{C6}$$

Moreover, while for the evading agent we assume the dynamics as given by Eqs. (C5) and (C6) with Ω_e chosen random among the values $0, \pm\varpi_e$, as in Ref. [42] for the pursuer we enforce the constraint (C4) exactly. It is worth underlying that the above kinematic equations remain valid also for Ω_e non random. Then we can derive – in polar coordinates $R = |\boldsymbol{x}_e - \boldsymbol{x}_p|$ and Φ_e – the motion of the pursuer in the pursuer frame of reference.

Computing $\dot{R}$: In order to compute $\dot{R}$, we have to project both velocities on the direction (ϕ_e) connecting the two agents. We can use Eqs. (C3)-(C1) (i.e., we can either project velocities onto the pursuer to evader direction of motion or compute $\dot{R}$ with the chain rule). This procedure leads to $\dot{R} = v_e \cos(\theta_e - \phi_e) - v_p \cos(\theta_p - \phi_e) = v_e \cos(\Phi_e - \Gamma_\pm) - v_p \cos(\Phi_e)$, and thus to $\dot{R} = -(v_p \pm v_e)\cos(\Phi_e)$.

Computing $\dot{\Phi}_e$: Let us now focus on $\dot{\Phi}_e$. By using Eq. (C4) and that $\dot{\theta}_e = \Omega_e$ (being Ω_e the angular velocity selected by the pursuer), we can deduce that $\dot{\theta}_p = 2\dot{\phi}_e - \Omega_e$ which implies that

$$\dot{\Phi}_e = -\dot{\phi}_e + \Omega_e \tag{C7}$$

On the other hand, a direct computation yields

$$\dot{\phi}_e = \frac{1}{R}[v_e \sin(\theta_e - \phi_e) - v_p \sin(\theta_p - \phi_p)] \,. \tag{C8}$$

Now using Eqs. (C7) and (C8) and noticing that $\theta_e - \phi_e = \phi_e - \theta_p - \Gamma_\pm = \Phi_e - \Gamma_\pm$ from (C4), we can deduce that

$$\dot{\Phi}_e = \Omega_e - \frac{1}{R}[v_e \sin(\Phi_e - \Gamma_\pm) + v_p \sin(\Phi_e)] = \Omega_e - \frac{1}{R}(v_p \mp v_e)\sin(\Phi_e) \,. \tag{C9}$$

We can then summarize the previous results in the set of equations

$$\begin{cases} \dot{R} & = -(v_p \pm v_e)\cos(\Phi_e) \\ \dot{\Phi}_e & = \Omega_e - \dfrac{1}{R}(v_p \mp v_e)\sin(\Phi_e) \,, \end{cases} \tag{C10}$$

where we recall the upper sign choice applies to mirroring and the lower choice to tailgating. Notice that Eq. (C10) essentially coincides with Eq. (17) of Ref. [42] but is specialized to the mirroring/tailgating constraint (C4) on the angles.

[1] M. S. Triantafyllou, G. D. Weymouth, and J. Miao, "Biomimetic survival hydrodynamics and flow sensing," Ann. Rev. Fluid Mech. **48**, 1 (2016).

[2] D. Takagi and D. K. Hartline, "Directional hydrodynamic sensing by free-swimming organisms," Bull. Math. Biol. **80**, 215 (2018).

[3] L. J. Tuttle, H. E. Robinson, D. Takagi, J R. Strickler, P. H. Lenz, and D. K. Hartline, "Going with the flow: hydrodynamic cues trigger directed escapes from a stalking predator," J. Royal Soc. Interface **16**, 20180776 (2019).

[4] E. Lloyd, C. Olive, B. A. Stahl, J. B. Jaggard, P. Amaral, E. R. Duboué, and A. C. Keene, "Evolutionary shift towards lateral line dependent prey capture behavior in the blind mexican cavefish," Develop. Biol. **441**, 328 (2018).

[5] J. C. Montgomery, C. F. Baker, and A. G. Carton, "The lateral line can mediate rheotaxis in fish," Nature **389**, 960 (1997).

[6] H. Bleckmann and R. Zelick, "Lateral line system of fish," Integr. Zool. **4**, 13 (2009).

[7] M. J. Kanter and S. Coombs, "Rheotaxis and prey detection in uniform currents by lake michigan mottled sculpin (cottus bairdi)," J. Experm. Biol. **206**, 59 (2003).

[8] T. Kiørboe and A. W. Visser, "Predator and prey perception in copepods due to hydromechanical signals," Mar. Ecol. Progr. Ser. **179**, 81 (1999).

[9] M. Doall, J. Strickler, D. Fields, and J. Yen, "Mapping the free-swimming attack volume of a planktonic copepod, euchaeta rimana," Mar. Biol. **140**, 871 (2002).

[10] J. Happel and H. Brenner, *Low Reynolds Number Hydrodynamics: With Special Applications to Particulate Media*, Vol. 1 (Springer Science & Business Media, New York, 2012).

[11] A. M. Hein, D. L. Altshuler, D. E. Cade, J. C. Liao, B. T. Martin, and G. K. Taylor, "An algorithmic approach to natural behavior," Curr. Biol. **30**, R663 (2020).

[12] P. Domenici, J. M. Blagburn, and J. P. Bacon, "Animal escapology i: Theoretical issues and emerging trends in escape trajectories," J. Exp. Biol. **214**, 2463 (2011).

[13] P. J Nahin, *Chases and Escapes: The Mathematics of Pursuit and Evasion* (Princeton University Press, Princeton, NJ, 2012).

[14] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics* (Cambridge University Press, Cambridge, UK, 1998).

[15] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction* (MIT Press, Cambridge, MA, 2018).

[16] B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powell, B. McGrew, and I. Mordatch, "Emergent tool use from multi-agent autocurricula," in *Proceeding of the International Conference on Learning*

Representations (2019).

[17] B. Chen, S. Song, H. Lipson, and C. Vondrick, “Visual hide and seek,” in *Artificial Life Conference Proceedings* (MIT Press, Cambridge, MA, 2020) pp. 645–655.

[18] L. Biferale, F. Bonaccorso, M. Buzzicotti, P. Clark Di Leoni, and K. Gustavsson, “Zermelo’s problem: Optimal point-to-point navigation in 2d turbulent flows using reinforcement learning,” Chaos **29**, 103138 (2019).

[19] J. K. Alageshan, A. K. Verma, J. Bec, and R. Pandit, “Machine learning strategies for path-planning microswimmers in turbulent flows,” Physical Review E **101**, 043110 (2020).

[20] G. Reddy, A. Celani, T. J. Sejnowski, and M. Vergassola, “Learning to soar in turbulent environments,” Proc. Nat. Acad. Sci. **113**, E4877 (2016).

[21] S. Colabrese, K. Gustavsson, A. Celani, and L. Biferale, “Flow navigation by smart microswimmers via reinforcement learning,” Phys. Rev. Lett. **118**, 158004 (2017).

[22] S. Verma, G. Novati, and P. Koumoutsakos, “Efficient collective swimming by harnessing vortices through deep reinforcement learning,” Proc. Nat. Acad. Sci. **115**, 5849–5854 (2018).

[23] M. Mirzakhanloo, S. Esmaeilzadeh, and M.-R. Alam, “Active cloaking in stokes flows via reinforcement learning,” J. Fluid Mech. **903** (2020).

[24] F. Cichos, K. Gustavsson, B. Mehlig, and G. Volpe, “Machine learning for active matter,” Nature Mach. Intel. **2**, 94 (2020).

[25] J. Qiu, N. Mousavi, L. Zhao, and K. Gustavsson, “Active gyrotactic stability of microswimmers using hydromechanical signals,” Phys. Rev. Fluids **7**, 014311 (2021).

[26] G. Reddy, J. Wong-Ng, A. Celani, T. J. Sejnowski, and M. Vergassola, “Glider soaring via reinforcement learning in the field,” Nature (Lond.) **562**, 236–239 (2018).

[27] S. Muiños-Landin, A. Fischer, V. Holubec, and F. Cichos, “Reinforcement learning with artificial microswimmers,” Sci. Robot. **6** (2021).

[28] See supplementary material [url] for supplementary figures, details on the implemented Reinforcement Learning algorithm including a pseudo-code, and for the captions of supplementary movies.

[29] E. Lauga and T. R. Powers, “The hydrodynamics of swimming microorganisms,” Rep. Progr. Phys. **72**, 096601 (2009).

[30] K. Ishimoto, E. A. Gaffney, and B. J. Walker, “Regularized representation of bacterial hydrodynamics,” Phys. Rev. Fluids **5**, 093101 (2020).

[31] R. J. Wubbels and N. A. M. Schellart, “Neuronal encoding of sound direction in the auditory midbrain of the rainbow trout,” J. Neurophysiol. **77**, 3060 (1997).

[32] A. B. Sichert, R. Bamler, and J. L. van Hemmen, “Hydrodynamic object recognition: When multipoles count,” Phys. Rev. Lett. **102**, 058104 (2009).

[33] T. Jaakkola, S. P. Singh, and M. I. Jordan, “Reinforcement learning algorithm for partially observable markov decision problems,” in *Advances in Neural Information Processing Systems*, Vol. 8, edited by D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (Morgan Kaufmann, San Francisco, CA, 1995) p.

345.

[34] S. Bhatnagar, R. S. Sutton, M. Ghavamzadeh, and M. Lee, "Natural actor–critic algorithms," Automatica **45**, 2471–2482 (2009).

[35] I. Grondman, L. Busoniu, G. A. D. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," IEEE Trans. Syst. Man. Cybern. Part C **42**, 1291–1307 (2012).

[36] D. Hennes, D. Morrill, S. Omidshafiei, R. Munos, J. Perolat, M. Lanctot, A. Gruslys, J.-B. Lespiau, P. Parmas, E. Duenez-Guzman, and K. Tuyls, "Neural replicator dynamics," arXiv:1906.00190 [cs.LG] (2019).

[37] S.-I. Amari, "Natural gradient works efficiently in learning," Neural Comput. **10**, 251–276 (1998).

[38] R. C. Eaton, R. K. K. Lee, and M. B. Foreman, "The Mauthner cell and other identified neurons of the brainstem escape network of fish," Progr. Neurobiol. **63**, 467–485 (2001).

[39] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," IEEE Signal Proces. Mag. **34**, 26–38 (2017).

[40] Note that each point represents the average over the previous and following 50 episodes, so it is not immediate to recognize those episodes in which the evader wins, i.e., in which $T/T_{max} = 1$.

[41] We use a fixed learning rate instead of an adaptive one, and possibly, due to the need to explore, a larger number of episodes per turn would be necessary.

[42] F. Belkhouche, B. Belkhouche, and P. Rastgoufard, "Parallel navigation for reaching a moving goal by a mobile robot," Robotica **25**, 63–74 (2007).

[43] G Adam and M Delbrück, "Reduction of dimensionality in biological diffusion processes," Structural chemistry and molecular biology **198**, 198–215 (1968).

[44] R. M. Olberg, A. H. Worthington, and K. R. Venator, "Prey pursuit and interception in dragonflies," J. Compar. Physiol. A **186**, 155 (2000).

[45] C. Chiu, P. V. Reddy, W. Xian, P. S. Krishnaprasad, and C. F. Moss, "Effects of competitive prey capture on flight behavior and sonar beam pattern in paired big brown bats, eptesicus fuscus," J. Exper. Biol. **213**, 3348 (2010).

[46] B. S. Lanchester and R. F. Mark, "Pursuit and prediction in the tracking of moving food by a teleost fish (acanthaluteres spilomelanurus)," J. Exper. Biol. **63**, 627 (1975).

[47] A. G. P. Kottapalli, M. Asadnia, J. Miao, and M. Triantafyllou, "Soft polymer membrane micro-sensor arrays inspired by the mechanosensory lateral line on the blind cavefish," J. Intell. Mat. Syst. Struct. **26**, 38 (2015).

[48] B. A. Free, J. Lee, and D. A. Paley, "Bioinspired pursuit with a swimming robot using feedback control of an internal rotor," Bioinsp. Biomim. **15**, 035005 (2020).

[49] A. D. Marchese, C. D. Onal, and D. Rus, "Autonomous soft robotic fish capable of escape maneuvers using fluidic elastomer actuators," Soft Robot. **1**, 75 (2014).

Supplementary material: Reinforcement learning for pursuit and evasion of microswimmers at low Reynolds number

Francesco Borra,[1] Luca Biferale,[2] Massimo Cencini,[3] and Antonio Celani[4]

[1] *Dipartimento di Fisica, Università "Sapienza" Piazzale A. Moro 5, I-00185 Rome, Italy*
[2] *Department of Physics and INFN, University of Rome Tor Vergata,*
Via della Ricerca Scientifica 1, 00133, Rome, Italy
[3] *Istituto dei Sistemi Complessi, CNR, via dei Taurini 19, 00185 Rome, Italy and INFN "Tor Vergata"*
[4] *Quantitative Life Sciences, The Abdus Salam International Centre for Theoretical Physics - ICTP, Trieste, 34151, Italy*

OUTLINE OF SUPPLEMENTARY INFORMATION

In Sec. I we detail the implementation of the Reinforcement Learning algorithm (including the pseudocode). In Sec. II we briefly comment on different choices of the parameters and on the addition of rotational noise, showing that these variations leave qualitatively unchanged the results presented in main text. Finally, Sec. III contains the captions of the supplementary movies.

I. DETAILS ON THE IMPLEMENTATION OF REINFORCEMENT LEARNING

As explained in main text, the two agents play a zero-sum game with opposite goals: the pursuer needs to find (i.e. to be at distance $\leq R_c$ from) the evader in the shortest possible time, while the latter needs to remain at distance $> R_c$ from the opponent as long as possible. In order to accomplish their goals they can only act on their angular velocity which can take three discrete values: these constitute the set $\mathcal{A}$ of *actions* a each agent can take (here and in the following we drop any reference to the specific agent unless necessary). At each time t (multiple of the decision time τ) the agent has access to some (partial) observation o_t of the *state* of the *environment* via the perceived gradients of the velocity field which are summarized in the features $\boldsymbol{\mathcal{F}}(o_t)$ (observables and features are discussed in Appendix B of main text). Such features are used to parameterize the *policy* $\pi(a|o)$ of each agent that is the probability to take action $a \in \mathcal{A}$ given the observation o, parameterized as a softmax function:

$$\pi(a|o) = \frac{\exp\left(\sum_{i=1}^{N_F} \mathcal{F}_i(o)\xi_{ia}\right)}{\sum_{a'\in\mathcal{A}} \exp\left(\sum_{j=1}^{N_F} \mathcal{F}_j(o)\xi_{ja'}\right)}\,. \tag{1}$$

In the above expression $\boldsymbol{\xi} = \{\xi_{ia}\}_{i=1,N_F\ a\in\mathcal{A}}$ are the parameters defining the policy that the agent needs to (learn) optimize to achieve its goal and N_F the number of features used. Once an action is taken the agents receive a reward. To perform such optimization we will use a *reinforcement learning (RL) scheme* to update the parameters $\boldsymbol{\xi}$, which is specified below. For a detailed introduction to RL we refer the reader to the book [6].

A. Reinforcement Learning Algorithm

We have now to specify the reinforcement-learning update scheme for the policy $\pi(a|o)$ of the learning agent within each episode. We have used a *natural actor-critic* (NAC) scheme [2] (see also Ref. [3] for a general review on actor-critic algorithms). This choice motivated by the fact that our problem setting basically defines a zero-sum game and in Ref. [4] it has been demonstrated that,in stateless zero-sum games, NAC algorithms are able to reproduce the replicator dynamics and thus to find evolutionary stable solutions to the game.

In a nutshell the basic idea of Actor-Critic algorithms is to learn simultaneously both the policy – *actor* step –and the observation value function $V(o)$ – *critic* step –, the latter corresponds to the expected future reward given a certain observation. Figure S 1 schematically describes the working principle, which can be summarized as follows: The agent gets an observation o_t of the environmental state, picks an action a_t according to the current policy $\pi(a_t|o_t)$, gets a reward r_t (for the reward scheme see main text) and observes the environment again $o_{t+\tau}$ then it

***critic* step:** updates with a bootstrap procedure the approximation of the observation value function that, similarly to π, is parameterized, using the features of table.I in Appendix B of main text, as

$$V(o) = \sum_{i=1}^{N_F} \mathcal{F}_i(o)\kappa_i\,, \tag{2}$$

where $\boldsymbol{\kappa} = \{\kappa_i\}_{i=1,N_F}$ are the parameters to be optimized. Such optimization is realized using *temporal difference* [6], whose basic idea is to produce a new approximation of V evaluating the difference between the actual reward r_t with the expected value according to the approximation of V based on the previous values of the parameters $\boldsymbol{\kappa}$.

***actor* step:** updates the policy parameter with a gradient ascent algorithm (policy gradient theorem in which $V(o)$ is used as a baseline [6]). To improve convergence, we use gradients covariant with respect to the metrics on the parameters defined by the Fisher information G – *natural gradients* [1, 2]. Such natural gradients are approximated by an auxiliary stochastic process as described in Ref. [2].

B. Pseudocode

We denote with η_A, η_C the learning rates of the actor and critic, respectively. While η_G is the learning rates of the estimator, g_{ia} (depending on feature i and action a), of the natural gradients. The following pseudocode describes the whole training and policy updating scheme structure. A sample code of the algorithm can be made available upon reasonable request. Policy and value-function update closely follows the natural actor-critic algorithm 3 of [2], adapted to our partial observability and adversarial setting (see e.g. [5]). Upper indices p/e refer to pursuer and evader respectively.

```
Parameter initialization
    $\boldsymbol{\xi}^{p/e} \leftarrow \mathbf{0}$;    $\boldsymbol{\kappa}^{p/e} \leftarrow \mathbf{0}$;    $\boldsymbol{g}^{p/e} \leftarrow \mathbf{0}$
    $\eta_C^p = \eta_C^e \leftarrow \eta_C$;    $\eta_G^p = \eta_G^e \leftarrow \eta_G$
Loop on learning cycles: $(c = 1, \ldots)$
    IF $c$ is ODD    $\eta_A^p \leftarrow \eta_A$    $\eta_A^e \leftarrow 0$    i.e. pursuer learns
    ELSE             $\eta_A^p \leftarrow 0$    $\eta_A^e \leftarrow \eta_A$    i.e. evader learns
        Loop on episodes: $(e = 1, M)$
            Initialize agents' positions and swimming orientations
            observations $o_0^{p/e}$
            Loop on time $t$: $(\{t \leq T_{max}$ OR $t|R(t) \leq R_c\})$
                pick action $a_t^{p/e} \sim \pi^{p/e}(a_t^{p/e}|o_t^{p/e})$
                advance dynamics using Eqs. (1)-(2) of main text from $t$ to $t+\tau$
                observations $o_{t+\tau}^{p/e}$
                Learning updates based on $o_t^{p/e}, o_{t+\tau}^{p/e}, a_t^{p/e}$:
                Compute temporal difference:
                    $\delta^{p/e} = \begin{cases} -V^{p/e}(o_t^{p/e}) & \text{if the state is terminal} \\ r_t^{p/e} + V^{p/e}(o_t^{p/e}) - V^{p/e}(o_{t+\tau}^{p/e}) & \text{otherwise.} \end{cases}$
                update parameters:
                    $\kappa_i^{p/e} \leftarrow \kappa_i^{p/e} + \eta_C^{p/e} \delta^{p/e} \nabla_{\kappa_i} V^{p/e}$
                    $g_{ib}^{p/e} \leftarrow g_{ib}^{p/e} + \eta_G^{p/e} \left[ \delta^{p/e} - \sum_{jc} \nabla_{\alpha_{jc}^{p/e}} \ln \pi^{p/e}(a_t^{p/e}|o_t^{p/e}) g_{jc}^{p/e} \right] \nabla_{\alpha_{ib}^{p/e}} \ln \pi^{p/e}(a_t^{p/e}|o_t^{p/e})$
                    $\xi_{ib}^{p/e} \leftarrow \xi_{ib}^{p/e} + \eta_A^{p/e} g_{ib}^{p/e} - (\xi_{ib}^{p/e}/\xi_0)^3$    $(\star)$
                $t \leftarrow t + \tau$
```

In equation $(\star)$ we have added a cubic regularization term to constrain the dynamics of policy parameter since they would otherwise diverge in norm, preventing efficient adversarial updating. The rationale for our choice is that the cubic term does not interfere with the parameter update for $\|\boldsymbol{\xi}\| \ll \xi_0$ but becomes relevant for $\|\boldsymbol{\xi}\| \gg \xi_0 = 20$. The choice of a soft hypercube rather than hard clipping can partially accommodate different intrinsic scales of the features and was empirically more performing. Note that this regularization breaks the covariant structure but was still performing well.

In table I we summarize the parameters used in the definition of episodes and turns, and the learning rates, which as shown in Ref. [2] should be chosen such that $\eta_A \ll \eta_G \ll \eta_C$ to ensure convergence.

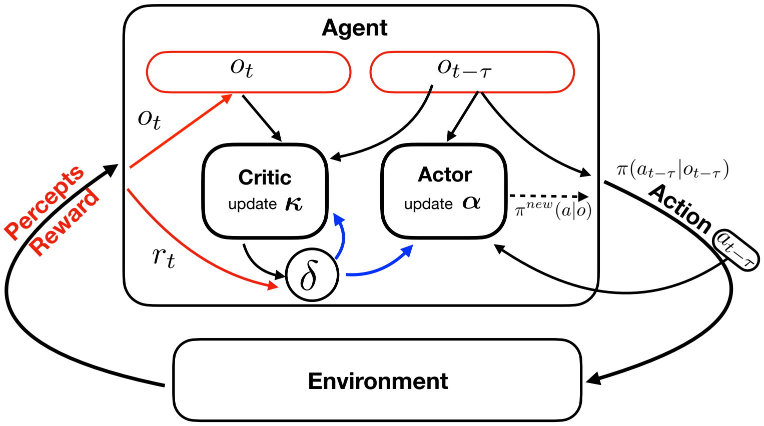

Fig-SI. 1. Pictorial representation of the actor-critic algorithm. The agent interacts with the environment acting on it and modifies its own behavior by looking at the consequences of its actions. Let us start at time $t-\tau$: the agent observes $o_{t-\tau}$ and picks its action $a_{t-\tau}$ according to its current policy $\pi(a_{t-\tau}|o_{t-\tau})$. After decision time τ, the environment responds with a reward $\mathcal{R}_t$ and yields a new observable signal o_t. The agent then uses o_t, $o_{t-\tau}$ and $\mathcal{R}_t$ to compares the actual outcome of its action with previous expectations (critic) and computes the mismatch δ (time difference, the circle in the sketch). This mismatch is used to update both expectations about future rewards V and its policy π. With the new π, the agent chooses the next action given o_t. Color coding: red links show available information, blue links the parameter updating paths.

The dynamics (1)-(2) of main text is integrated with a 4^{th} order Runge-Kutta scheme with time-step $dt = 0.02$ time unit, while the decision time is $\tau = 0.1$.

$T_{max} = 500$	$M = 5000$	$\eta_A = 10^{-5}$	$\eta_G = 10^{-4}$	$\eta_C = 10^{-3}$	$\xi_0 = 20$

TABLE I. Parameters used in the reinforcement learning.

II. TEST WITH DIFFERENT PARAMETERS AND ADDING ROTATIONAL NOISE

In the main text we studied one choice of the parameters (agents' speeds and angular velocities) moreover we did not include any stochastic effect which can affect the dynamics of the microswimmers.

We did studied several variations of the parameters around the values presented in the main text. We always obtained qualitatively similar results: the main pursuit and evasion strategies discussed in Fig.2 of main text were found also in other cases, with some variability in the order in which they are learned which is not surprising given that the learning is intrinsically stochastic. In this section, we show the results for a pursuer faster by a factor 2 than in main text and with larger curvature radius, i.e. $(v_p, \varpi_p) = (0.3, 3)$ and reconsider the example shown in Fig. 2 of main text by adding also stochastic effects in the orientation dynamics. In all cases we keep fixed the force dipole strength to the main text value, i.e. $D_p = D_e = 0.03$.

Figure-SI 2 shows the equivalent of Fig. 2 of main text for the case $(v_p, \varpi_p) = (0.3, 3)$. Also in this case we observe some variability in the history of the rewards and some instabilities in the learning cycles after the second (see Fig-SI. 2a), which again we can interpret in problems related to exploration and/or in the stability of the algorithm as discussed in main text. Panels (b) and (c) display the characteristic mirroring and tailgating strategies learnerd by the pursuer in the first cycle. Differently from the case presented in main text, in its first learning cycle the evader first learn the very effective twirling strategy (panels (h) and (i) show two instances). In the its second turn the pursuer changes its strategy and sometimes is able to win (e.g. panel (d)), its new strategy reveals that in the policy of the evader is also coded other behaviors than twirling such as linear escapes (see panel (j)). For run 1 and 3 anyway, a part from a few intervals for run 3, the evader's policy seems to be very strong or, equivalently, the pursuer has some difficulty in finding an effective counter-strategy. In run2, however, in the second learning cycle of the evader, maybe due to the change of policy of the pursuer, some instability leads to more variegate behaviors and we find capture arches (e) in which the pursuer wins and many episodes in which the evader use the hydrodynamic defense as in (k). These strategies and linear escapes remain for the rest of the cycles, see panels (f,g) and (l,m). Notice that in the same episode the can be combined, e.g. panel (g) shows an instance in which the pursuer can win with an arch

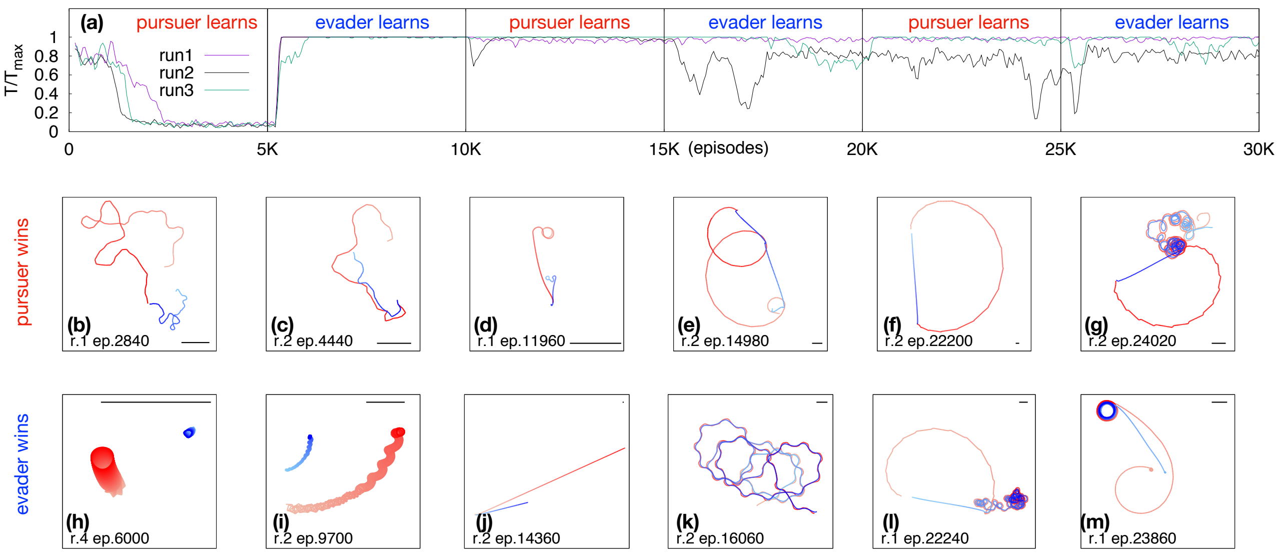

Fig-SI. 2. Co-evolution of pursuit-evasion strategies with $(v_e, \varpi_e) = (0.1, 3)$ and $(v_p, \varpi_p) = (0.3, 3)$, the other parameters as in Fig. 2 of main text. (a) The history of the reward, namely the running average (over 100 episodes) of T/T_{max} for three different realizations as labeled. (b-g) Samples of winning pursuit strategies: (a) mirroring, (b) tailgating, (c) capture upon escape from twirling, (e,f) capture arch similar to panel (d) of Fig.2 main text, (g) successful capture arch after escape from hydrodynamic defense. (h-m) Samples of winning escape strategies: (g,h) two forms of twirling, (j) linear escape, (k) hydrodynamic defense, (l) failed capture arch followed by hydrodynamic defense, (k) similar to (l) but with some variation on the path. Colors coding of the trajectories as in Fig. 2 of main text. Note that the unit length bar is on the bottom/top right for the winning pursuer/evader panels.

after defeating the hydrodynamic defense of the evader, while panels (l) and (m) show two successful hydrodynamic defenses following an attempt of capture of the pursuer.

We also considered a slightly modified dynamics to account for possible rotational noise on the angular dynamics of the microswimmers, i.e. we modified Eq. (2) of main text in

$$\dot{\theta}_\alpha = \Omega_\alpha + \frac{1}{2}\omega^{(\beta)} + \sqrt{2D_r}\eta\,,$$

where η is a zero mean, delta-correlated in time Gaussian noise and D_r the rotational diffusion constant. As shown in Fig-SI. 3, obtained with the same parameters of Fig. 1 of main text and $D = 0.025$ we qualitatively found the same results. In particular, notice that the evolution of the reward (Fig-SI. 3a) closely follows the average one of the deterministic case in the first two learning cycles. In particular, panels (b,c) display the typical features of mirroring (b) and tailgating (c), and in panel (d) we show an instance of switching between the two. In the second cycle, where the evader learns, we can recognize the features of the hydrodynamic defense (g) which, due to the rotational noise, however is more curved and irregular with respect to the deterministic case. Panel (h) shows a successful attempt of linear escape which, however, due to the noise is not perfectly linear. Indeed in most of the cases the evader trajectory is irregular and curved so that mirroring is more effecting than in the deterministic case: we can indeed see a failed escape in (e) and a successful one in (i) which is basically indistinguishable from a failed mirroring. Later on (j) the prey learns to use twirling which sometimes fail (f). Possibly due to similar instabilities to those observed in the deterministic case, however, the evader is unable to fix twirling as a winning strategy and in the other cycles we observe a phenomenology similar to that of the second cycle, of these cycles we only show panel (k) which displays an interesting instance in which the pursuer starts with mirroring and the prey is able to make it switch to tailgating which can be contrasted with a hydrodynamic defense.

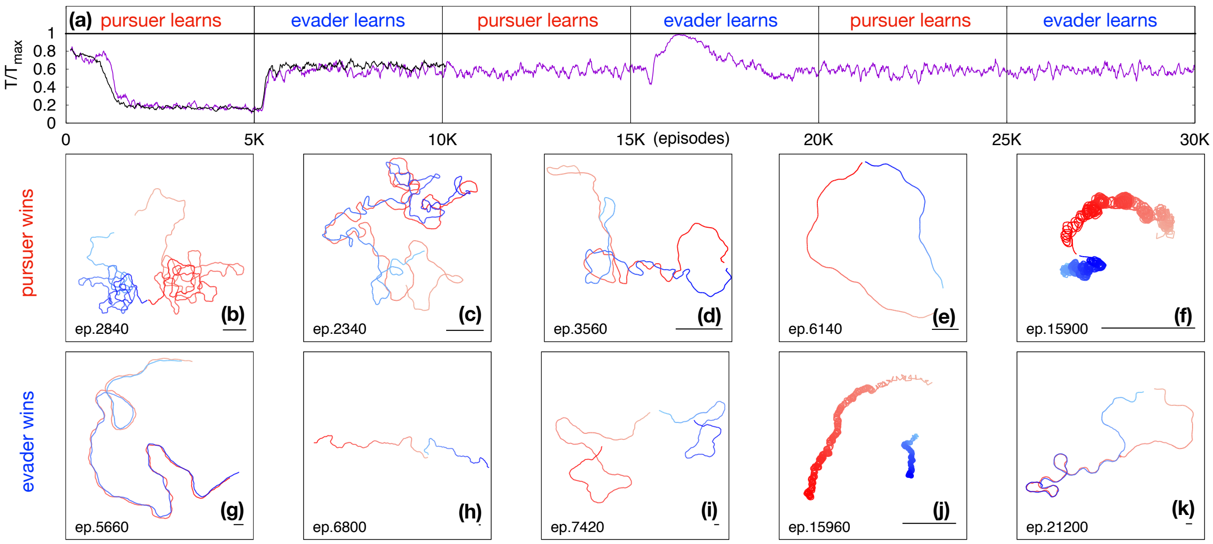

Fig-SI. 3. Co-evolution of pursuit-evasion strategies in the case in which rotational noise with $D_r = 0.025$ is added (values around this one gives similar results within the variability of the learning process which is intrinsically stochastic). (a) The history of the reward, namely the running average (over 100 episodes) of T/T_{max}. The purple curve shows one instance of the learning for the case with rotational noise while the black one shows the average, for the first two cycles, of the learning curve of the deterministic case (i.e. the average of run 1, 2 and 3 in Fig.2 of main text. (b-f) Samples of winning pursuit strategies: (a) mirroring, (b) tailgating, (c) switching between the latter two, (e) similar to panel (d) of Fig.2 main text, (f) failed twirling defense. (g-k) Samples of winning escape strategies: (g) hydrodynamic defense, (h) and (i) attempts of linear escapes which due to the noise are almost indistinguishable with mirroring, (j) twirling, (k) successful hydrodynamic defense after a capture attempt with mirroring. Colors coding of the trajectories as in Fig. 2 of main text.

III. CAPTIONS OF SUPPLEMENTARY MOVIES

movie1.gif: Time evolution of pursuer trajectories in run2 episode 6240 (shown in Fig.2h of main text) as seen in the frame of reference of the evader, i.e. with the evader in the origin (the blue circles of size R_c shows the evader position) and oriented with its heading direction along the x-axis. The arrow shows the pursuer relative heading orientation with respect to the evader. Notice that soon the pursuer trajectory becomes basically stationary in this frame of reference, meaning that its speed relative to the prey is basically vanishing.

movie2.gif: Similar to movie1.gif but referring to run3 episode 12520 (Fig.2e of main text). Notice that at some instants the trajectory tends to be trapped but suddenly the pursuer finds a way to escape the hydrodynamic trap.

movie3.gif: (Top) Evolution of the trajectory of pursuer (red) and evader (blue) in run1 episode 23160 (Fig.2k of main text). (Bottom) evolution of the angle $\Phi_p + \Phi_e$. Notice the approximate switches between mirroring ($\Phi_p + \Phi_e \approx 0$) and tailgating ($\Phi_p + \Phi_e \approx 0$) triggered by the evader turns.

[1] S.-I. Amari. Natural gradient works efficiently in learning. *Neural Comput.*, 10(2):251–276, 1998.

[2] S. Bhatnagar, R. S. Sutton, M. Ghavamzadeh, and M. Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

[3] I. Grondman, L. Busoniu, G. A. D. Lopes, and R. Babuska. A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE Trans. Syst. Man. Cybern. Part C*, 42(6):1291–1307, 2012.

[4] D. Hennes, D. Morrill, S. Omidshafiei, R. Munos, J. Perolat, M. Lanctot, A. Gruslys, J.-B. Lespiau, P. Parmas, E. Duenez-Guzman, and K. Tuyls. Neural replicator dynamics. *arXiv:1906.00190 [cs.LG]*, 2019.

[5] T. Jaakkola, S. P. Singh, and M. I. Jordan. Reinforcement learning algorithm for partially observable markov decision problems. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, page 345. Morgan Kaufmann Publishers, 1995.

[6] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.