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*

Aquatic organisms can use hydrodynamic cues to navigate, find their preys and escape from predators. We consider a model of two competing microswimmers engaged in a pursue-evasion task while immersed in a low-Reynolds-number environment. The players have limited abilities: they can only sense hydrodynamic disturbances, which provide some cue about the opponent's position, and perform simple manoeuvres. The goal of the pursuer is to capture the evader in the shortest possible time. Conversely the evader aims at deferring capture as much as possible. We show that by means of Reinforcement Learning the players find efficient and physically explainable strategies which non-trivially exploit the hydrodynamic environment. This Letter offers a proof-of-concept for the use of Reinforcement Learning to discover prey-predator strategies in aquatic environments, with potential applications to underwater robotics.

Aquatic organisms can detect moving objects in their environment by sensing the induced hydrodynamic disturbances [1–3]. Such an ability is crucial in prey-predator interactions as well as for navigation, especially in murky water or in the dark as in the case of 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 inhabit a low-Reynolds-number environment and have antennae and setae to sense hydrodynamic signals produced by predators and preys [8, 9]. Bioinspired mechanosensors that can sense the hydrodynamic fields are used in underwater robots employed for search and recovery, surveillance and ship inspection [10, 11]. Thus, understanding how to exploit hydrodynamic cues is of interest both for mechanistic explanations of animal behavior and for underwater robotics.

Abstracting away from specific mechanisms developed by aquatic organisms or deployed for robots, the problem of pursue-evasion in microswimmers guided by hydrodynamic cues poses substantial difficulties that are rooted in the physics of the ambient medium. At low Reynolds numbers, flow disturbances are generally weak and characterized by symmetries [12] that create ambiguities about the location of the signal source especially when it is distant from the receiver [2, 3, 8]. Moreover, hydrodynamics has dynamical effects, since the disturbances generated by one microswimmer alter the motion of the other. Which pursuit-evasion strategies can be devised in such dynamic, partially observable environments? How do they compare with strategies based on visual cues? Can hydrodynamics be exploited and how?

In this Letter, we explore the use of Multi Agent Reinforcement Learning (MARL) [13] as a general model-free framework for discovering effective strategies for chasing and escaping at low Reynolds number. Reinforcement

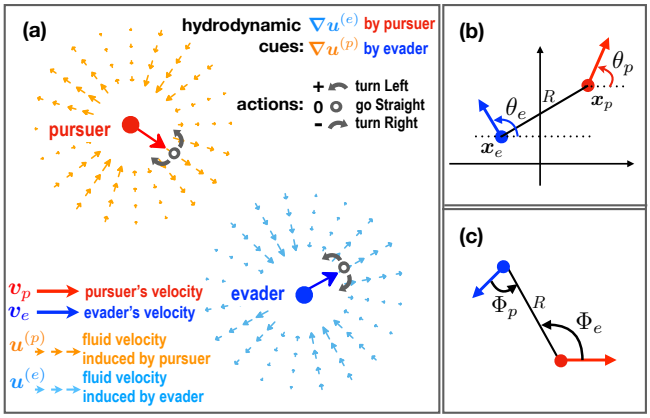

FIG. 1. Model illustration. (a) The pursuer (p, red)/ evader (e, blue) goal is to min/maximize the time their distance reaches the capture value R_c within a given time horizon. Agents move in the plane with speed $v_{p/e}$; every τ time-unit they choose to maintain or turn left/right their heading direction. By swimming an agent generates a velocity disturbance, $\boldsymbol{u}^{(p/e)}$, which drags the other and offers a cue to the other agent on the relative position and orientation via its gradients, $\boldsymbol{\nabla u}^{(p/e)}$. (b) Geometry of the problem in a fixed frame of reference with indicated the heading angles. (c) Bearing angle $\Phi_{e/p}$ corresponding to the angular position of an agent with respect to the heading direction of its opponent.

Learning (RL) approaches rely on trials and errors to improve the quality of the decisions made by an agent – here a microswimmer. The potential of RL for navigation in complex and dynamic fluid environments has been demonstrated in various tasks, both in silico [14–21] and experimentally [22, 23].

Here, inspired by classical pursue-evasion problems [24] and by recent applications of RL to hide-and-seek contests [25, 26], we frame the problem of prey-predator microswimmers in a game theoretic framework [27]. As-

suming limited manoeuvrability and partial information via hydrodynamic cues, the swimmers play the following zero-sum game (Fig. 1a): Agents start at distance R_0 with random heading directions; 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 upon capture (pursuer wins) or if its duration exceeds a given time T_{max} (evader wins). We train the agents in this adversarial setting via MARL and then try to decode the coevolving complex strategies discovered by them.

Modeling the agents. We model the agents as "pushers" producing a force dipole that moves with speed v_α with $\alpha = e, p$ (see Fig. 1a), which well approximates the far field of many microorganisms [28]. Besides self-propulsion each microswimmer (assumed spherical) is advected and reoriented by the velocity field generated by the other. For simplicity we assume that there is no external flow. At each τ time units, i.e at each decision time, agents can partly steer by imparting a torque. This results in an angular velocity Ω_α. Thus the position $\boldsymbol{x}_\alpha$ and heading orientation θ_α evolve according to

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

$$\dot{\theta}_\alpha = \Omega_\alpha + 1/2\, \omega^{(\beta)}\,, \tag{2}$$

with $\boldsymbol{n}_\alpha = (\cos\theta_\alpha, \sin\theta_\alpha)$ and where $\boldsymbol{u}^{(\beta)}$ and $\omega^{(\beta)}$ are the velocity and vorticity field at position $\boldsymbol{x}_\alpha$, generated by the opponent agent β in $\boldsymbol{x}_\beta$ with heading orientation θ_β, see Fig. 1b and Sec. I of [29] for details.

Modeling the hydrodynamic cues. We assume that each agent can sense the presence and the movement of the opponent only through the gradients of the velocity field it generates, similarly to what copepods do with sensory setae [8]. As detailed in Sec. I of [29] (see also Fig. 1b,c), such cues depend on agents' separation, $R = |\boldsymbol{x}_p - \boldsymbol{x}_e|$, relative heading, $\Theta_\beta = \theta_\beta - \theta_\alpha$, and the bearing angles, Φ_e and Φ_p, as they are called in the pursuit-evasion-games language [24]. The symmetries of the equations lead to ambiguities in the determination of the position of the source of the signal, akin to the 180^o ambiguity occurring in fish hearing [30]. Both the relative heading and bearing angle can indeed be transformed as $\Theta_\beta \to \Theta_\beta + \pi$ and $\Phi_\beta \to \Phi_\beta + \pi$ leaving the perceived gradients unchanged (see Eqs. (8-10) in [29]). Such ambiguities make it impossible to implement standard pursuit-evasion strategies such as the ones based on visual cues [24, 31]. Memory of past gradient detections and/or multipole effects at short distance can mitigate the ambiguities which, however, typically persist at larger distances [1, 2, 32].

Learning to pursue and evade through reinforcement. To properly set up a learning framework, we need to identify: a set of observations, o, that each agent can receive and use to infer the state of the opponent; the actions, a, through which it can implement its strategy; and the rewards, r, to evaluate the quality of 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 [33] through the velocity gradients. The actions are $a \in \mathcal{A} = \{0, +, -\}$ (Fig. 1a), and correspond to three angular velocities $\Omega_\alpha = 0, +\varpi_\alpha, -\varpi_\alpha$ that each agent can choose to control its heading direction. Once actions are taken, the agents evolve for a time τ with the dynamics (1-2) and a reward is issued. In this zero-sum game, the currency is given by the elapsing time: the pursuer/evader receives a reward $r = {}^{+1}/_{-1}$ at the end of each decision time. After each action, the agents update their policy by combining past and new information with the received 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 T exceeds T_{max}).

Reinforcement Learning algorithm. Among the many approaches to MARL we chose to adopt a Natural Actor-Critic architecture [34, 35] because of its theoretical guarantees and its connection with evolutionary game theory [27, 36] (see Sec. II of [29] for details). In this class of algorithms, locally optimal solutions are sought by means of stochastic gradient ascent in the space of the policies and natural gradients are used, i.e. covariant derivatives with respect to the metric defined by the Fisher information [37]. In real organisms environmental cues are processed by the nervous system which then encodes the policy, for example by means of dedicated neurons that control escape responses of fishes [38]. Such neural encoding can be emulated by artificial neural networks [39].

Here, in the interest of explainability, we opted for an explicit parameterization of the policy in terms of a few features of the observations. This is where the expert advice coming from physical intuition about the process at hand becomes relevant. Dropping the agent indexes for the sake of notation simplicity, we set $\pi(a|o) = \exp(\boldsymbol{\mathcal{F}}(o)\cdot\boldsymbol{\xi}_a)/\sum_{a'}\exp(\boldsymbol{\mathcal{F}}(o)\cdot\boldsymbol{\xi}_{a'})$, where $a, a' \in \mathcal{A}$, $\boldsymbol{\mathcal{F}}(o)$ are features that encode the observations o, and $\boldsymbol{\xi}_a$ the parameters to be learned. By combining the components of the velocity gradients we chose to extract the following observables, o (see Sec. IIA of [29]): the vorticity ω; a proxy for the distance from the other agent, $\hat{R} \propto 1/R^2$; and a linear combination of the 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 the angle γ. In order to partially encode for the heading direction, we include some short-term memory in the form of a combination of a few past observations (see Sec. IIA of [29]). We made some exploratory study with more features and we did not observe qualita-

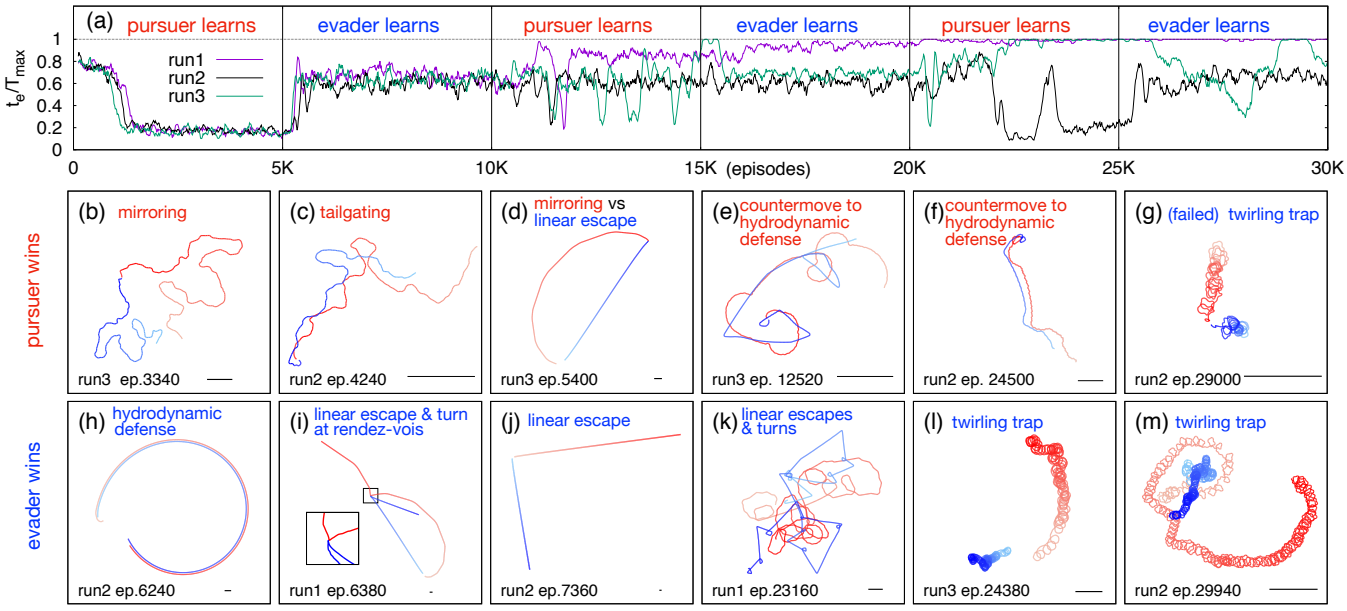

FIG. 2. History of first 6 training cycles and coevolving strategies. (a) Running average (over 100 episodes) of normalized episode duration T/T_{max} for 3 realizations of the learning process. (b-g) Winning pursuing strategies: (b) mirroring, (c) tailgating, (d) mirroring vs linear escape with a rendez-vous, (e,f) tailgating with different countermoves to hydrodynamic defense, (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 at distance, (l,m) twirling trap. Red/Blue denotes pursuer/evader trajectories, episode time runs from lighter to darker color; run and episode are labeled on each panel; the black segment on the bottom right displays the unit length.

tively different results from the minimal setting described above.

Training scheme. Agents start their training with a random policy, $\pi(a|o) = 1/|\mathcal{A}| = 1/3$ for all o. Learning is organized in phases where agents alternately improve their policies. At first, the pursuer learns while keeping the evader's policy frozen, then the evader learns against a pursuer policy held fixed to the one obtained at the end of the previous phase. Each phase consists of $M = 5 \cdot 10^3$ episodes. Episodes start with agents at a distance $R_0 = 1$ and random heading directions, and end either upon capture ($R \leq R_c = 0.05\ R_0$) or when time exceeds the threshold $T_{max} = 50\ T_0$, 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: they can make turns with the same curvature radius $v_p/\varpi_p = v_e/\varpi_e$. With this choice hydrodynamics disturbances dominate over swimming at distances $R \lesssim R_0$; the decision time is $\tau = 0.01 T_0$ for both agents. The qualitative results are quite robust as can be checked upon varying the parameters around these values. A detailed investigation of the dependency on different set of parameters will be reported elsewhere.

Results. Our results are summarized in Fig. 2: panel (a) shows the normalized game duration T/T_{max} during the first six learning phases for three independent learning experiments; panels (b-g) and (h-m) display some representative examples of pursuer and evader winning strategies, respectively. The first two cycles are quite reproducible with the pursuer discovering ways to rapidly catch its prey and the latter finding ways to counteract in its own learning cycle. 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 in which the evader eventually dominates the games (run1 in Fig. 2a). We hypothesize that such variability arises from a combination of insufficient tuning of hyperparameters [40] and/or subtle instabilities in the learning algorithm. Notwithstanding these limitations, many aspects of the learned strategies are robust and, to some extent, physically explainable as discussed in the following.

Pursuing strategies: mirroring and tailgating. In its first learning phase, the evader executes a random policy insensitive to any cue, while the predator learns to pursue its prey either 'mirroring' its actions (Fig. 2b) or 'tailgating' it (Fig. 2c). When the pursuer approaches the evader, a switch between the two strategies can also 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 en-

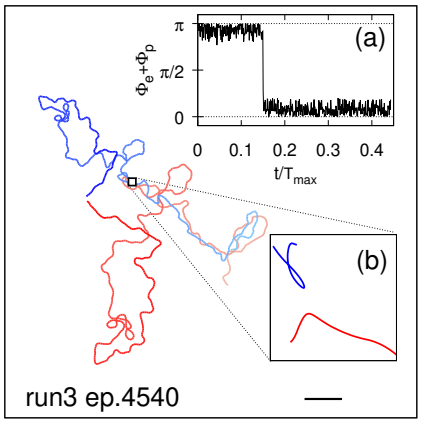

FIG. 3. Switching between tailgating to mirroring strategies. Inset (a) Sum of bearing angles $\Phi_e + \Phi_p$ vs normalized time, notice the switching from $\approx \pi$ (tailgating) to ≈ 0 (mirroring) at $t/T_{max} \approx 0.15$ corresponding to the close encounter and the evader turning shown in inset (b).

force over time specific relations between the bearing angles, namely $\Phi_e = -\Phi_p$ for mirroring and $\Phi_e = -\Phi_p + \pi$ for tailgating (see Fig. 3a). In Sec. III of [29], we show how such relations connect to the information gained by the pursuer from the gradients of the velocity field generated by the evader. The aforementioned 180^o ambiguity on the heading directions makes the pursuer unable to discern mirroring and tailgating just on the basis of instantaneous hydrodynamical cues: which strategy is chosen depends on the initial conditions, dynamical memory and hydrodynamic interactions. For example, in Fig. 3 we show a typical case when a close encounter (failed capture) triggered the pursuer to switch from a tailgating to a mirroring strategy allowing for a final successful capture.

Escaping strategies: hydrodynamic defense and linear flights. In its own training phase, the evader learns to contrast mirroring and tailgating. As for the latter, it finds a way to exploit hydrodynamics (Fig. 2h). 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 the hydrodynamic velocity essentially cancels the advantage of the latter (see Supplementary movie1 displaying pursuer's trajectory in the evader frame of reference). As a result, the pursuer is trapped at a fixed distance from its prey while following it. Another strategy adopted by the evader is a (almost) linear escape (Fig. 2i,j). However, as shown in Fig. 2d, this is not always successful as the pursuer can apply a mirroring strategy and intercept the evader to a *rendez-vous* point by performing a long smooth arch. Such arches correspond to adjusting the axis of mirroring in the course of time. However, either by making such *rendez-vous* point very far (Fig. 2j) or by exploiting hydrodynamics and turns upon close encounters (Fig. 2i), the evader can consistently make their evasion strategies quite efficient.

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

Understanding simple strategies. By using the equations of motion (1-2), upon neglecting hydrodynamical interactions, one can exactly derive the equations for the separation and bearing angle [31]. By imposing that the pursuer follows either mirroring or tailgating strategies, such equations read (see Sec. III of [29])

$$\dot{R} = -(v_p \pm v_e)\cos\Phi_e \tag{3}$$

$$\dot{\Phi}_e = \Omega_e - \frac{1}{R}(v_p \mp v_e)\sin\Phi_e\,, \tag{4}$$

where $\pm$ refers to mirroring or tailgating, respectively. From Eq. (3) we readily see that the tailgating strategy 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 (4) leads to a vanishing bearing angle $\Phi_h \to 0$ for small enough distances, which explains the origin of tailgating. Mirroring, instead, remains effective also for $v_p = v_e$ as it essentially maps the pursue into a first hitting problem of a one-dimensional random walker (when Ω_e is randomly chosen by the evader). Further tests with RL applied to the full unconstrained dynamics, including hydrodynamical interactions and equal velocities confirmed this scenario. We can then interpret mirroring as a random search with dimensionality reduction [41].

Conclusions. In this Letter, we have shown how microswimmers even if endowed with limited manoeuvring ability and poor positional information can discover complex strategies to pursue and evade from each other by exploiting the dynamics and signals provided by the hydrodynamic environment. It is interesting to compare the policies discovered by RL with the visual strategies found in pursuit-evasion games based on the knowledge of the line of sight [31]. The mirroring strategy that we observe has some similarities with "parallel navigation" where the line-of-sight direction is kept constant with respect to an inertial frame of reference. It has been conjectured that dragonflies follow this class of strategies for

predation purposes [42]. Instead, tailgating is similar to a "pure pursuit" where the heading is constantly directed toward the line of sight (zero bearing angle), as e.g. bats or some fishes appear to do [43, 44]. However, in tailgating the capture is prevalently realized from behind.

Our study is a first attempt at implementing a game-theoretic approach to interacting hydrodynamical agents. We see it as a preliminary step towards further research on the use of reinforcement learning algorithms with a twofold goal: rationalizing observed prey-predator interactions between aquatic organisms, and training underwater robots to accomplish complex tasks – e.g. artificial fishes imitating escape responses [45]. Here we did not discuss the effect of external flows and/or boundaries. Preliminary studies conducted in a circular arena show that, in spite of the confounding cues and more complex dynamics arising from the presence of the walls, the agents can nevertheless learn to exploit hydrodynamics in order to perform their pursue/evasion tasks. Results on this subject will be reported elsewhere. 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 [19].

F.B. acknowledges hospitality from ICTP. AC has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement N°956457. This work received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 882340)

* Corresponding author; massimo.cencini@cnr.it
† Corresponding author; celani@ictp.it

[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] 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).
[11] 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).
[12] J. Happel and H. Brenner, *Low Reynolds number hydrodynamics: with special applications to particulate media*, Vol. 1 (Springer Science & Business Media, 2012).
[13] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction* (MIT press, 2018).
[14] 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).
[15] 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).
[16] G. Reddy, A. Celani, T. J. Sejnowski, and M. Vergassola, "Learning to soar in turbulent environments," Proc. Nat. Acad. Sci. **113**, E4877 (2016).
[17] S. Colabrese, K. Gustavsson, A. Celani, and L. Biferale, "Flow navigation by smart microswimmers via reinforcement learning," Phys. Rev. Lett. **118**, 158004 (2017).
[18] 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).
[19] M. Mirzakhanloo, S. Esmaeilzadeh, and M.-R. Alam, "Active cloaking in stokes flows via reinforcement learning," J. Fluid Mech. **903** (2020).
[20] F. Cichos, K. Gustavsson, B. Mehlig, and G. Volpe, "Machine learning for active matter," Nature Mach. Intel. **2**, 94 (2020).
[21] J. Qiu, N. Mousavi, L. Zhao, and K. Gustavsson, "Active gyrotactic stability of microswimmers using hydromechanical signals," arXiv preprint arXiv:2105.12232 (2021).
[22] G. Reddy, J. Wong-Ng, A. Celani, T. J. Sejnowski, and M. Vergassola, "Glider soaring via reinforcement learning in the field," Nature **562**, 236–239 (2018).
[23] S. Muiños-Landin, A. Fischer, V. Holubec, and F. Cichos, "Reinforcement learning with artificial microswimmers," Sci. Robot. **6** (2021).
[24] P. J Nahin, *Chases and escapes: the mathematics of pursuit and evasion* (Princeton University Press, 2012).
[25] B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powell, B. McGrew, and I. Mordatch, "Emergent tool use from multi-agent autocurricula," in *International Conference on Learning Representations* (2019).
[26] B. Chen, S. Song, H. Lipson, and C. Vondrick, "Visual hide and seek," in *Artificial Life Conference Proceedings* (MIT Press, 2020) pp. 645–655.
[27] J. Hofbauer and K. Sigmund, *Evolutionary games*

and population dynamics (Cambridge University Press, 1998).

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

[29] See Supplemental Material [url] for a full description of the hydrodynamic fields generated by the microswimmers, for details on the implemented Reinforcement Learning algorithm, and for a derivation of Eqs. (3-4).

[30] 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).

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

[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 Publishers, 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] 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.

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

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

[43] 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).

[44] 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).

[45] 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).