Distributed on-line reinforcement learning in a swarm of sterically interacting robots

Matan Yah Ben Zion, Nicolas Bredeche, and Olivier Dauchot

1 Gulliver UMR CNRS 7083, ESPCI Paris, PSL Research University, Paris 75005, France
2 Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, ISIR, F-75005 Paris, France
3 School of Physics, and the Center for Physics and Chemistry of Living Systems, Tel Aviv University, Tel Aviv 6997801, Israel

While naturally occurring swarms thrive when crowded, physical interactions in robotic swarms are either avoided or carefully controlled, thus limiting their operational density. Designing behavioral strategies under such circumstances remains a challenge, even though it may offer an opportunity for exploring morpho-functional self-organized behaviors. In this paper, we explicitly consider dense swarms of robots where physical interactions are inevitable. We demonstrate experimentally that an a priori minor difference in the mechanical design of the robots leads to important differences in their dynamical behaviors when they evolve in crowded environments. We design Morphobots, which are Kilobots augmented with a 3D-printed exoskeleton. The exoskeleton not only significantly improves the motility and stability of the Kilobots, it also allows to encode physically two contrasting dynamical behaviors in response to an external force or a collision. This difference translates into distinct performances during self-organized aggregation when addressing a phototactic task. Having characterized the dynamical mechanism at the root of these differences, we implement a decentralized on-line evolutionary reinforcement learning algorithm in a swarm of Morphobots. We demonstrate the learning efficiency and show that the learning reduces the dependency on the morphology. We present a kinetic model that links the reward function to an effective phototactic policy. Our results are of relevance for the deployment of robust swarms of robots in a real environment, where robots are deemed to collide, and to be exposed to external forces.

I. INTRODUCTION

Swarm robotics involves a population of robots with limited computation and communication capabilities, that self-organize to achieve complex tasks at the level of the collective [1, 2]. It is looked upon as a promising venue for many applications [3-5]. Previous work demonstrated potential applications to collective construction [6], coordinated motion for flying UAVs [7-10], patrol in open water [11] or underwater exploration [12, 13]. These applications are characterized by the deployment of swarms where robots do not physically come into contact with one another [14]. There are however some situations where robots need to form dense assemblies or cohesive groups such as self-assembling structure and morphogenesis [15-17] or coordination as a multicellular robotic body [18, 19]. Formulation of robust algorithms for swarm robotics in crowded environments are therefore needed.

One can take inspiration from nature, which is rich with examples of dense swarms that rely on morphological interaction to thrive, accomplishing many tasks that could also be relevant in swarm robotics. Larvae self-organize into a living fountain to quickly consume a localized food source [20], ants form bridges [21], self-assemble into rafts [22] and carry heavy cargo [23]. At a smaller scale, bacteria colonies display internal turbulent flow [24] thereby mixing in nutrients [25, 26]. And at a larger scale, self-organized collective motion is observed in many animals, from fish schools to bird flocks and herds of mammals [27], as well as with pedestrians where crowd dynamics can emerge from body collisions [28]. Interestingly, self organization in dense assemblies is not a prerogative of living beings. It has been shown that collective motion can emerge from independent mindless robots physically interacting with one another [29, 32], considering a stripped-down flavour of swarm robotics as a particular kind of active matter [33, 34], with no explicit decision computed at the level of the robots.

In this paper, we bridge the gap between swarm robotics and active matter, by considering a swarm of robots that uses physical interactions in order to achieve a collective task. In order to leverage physical interactions, we use evolutionary reinforcement learning for swarm robotics [35, 36]. As noted by others [37], learning in swarm robotics suffers from the famous reality gap inherent to using simulation, which is generally used to accelerate learning. This is even more true in our case given that a simulation that is physically accurate is out of reach without an astronomical amount of experimental, numerical and theoretical work. To exploit relevant physical interactions, we propose to distribute learning over the robot swarm, which implies individual robots to continuously innovate and diffuse efficient strategies to their current neighbours, mimicking a form of social learning at the level of the swarm [38, 39].

For the purpose of concreteness, we consider an environmental-mediated aggregation task where robots self-organize to position themselves close to one another within an illuminated area, the rest of the environment being dark [1]. The expected behavior is phototaxis which, similarly to galvanotaxis and chemotaxis, is a common behavior observed in the animal kingdoms [40, 41]. As noted by several authors [1, 2, 42, 43], stimulus-induced locomotory movements (or taxis) constitute an
important building block in swarm robotics, and several algorithmic implementations have been proposed for phototaxis using ambient light sensors [13, 14, 17] and phototaxis using microphone and speakers [48, 49]. Phototaxis has been addressed by implementing behaviors observed in nature [50–52], virtual physics law [53, 54] or by using evolutionary reinforcement learning [55–57]. While they vary in implementation details, all works to date implement strategies mediated by the nearby population, which requires robots to feature sensors capable of detecting other robots, for instance to measure the density of robots nearby.

In our setup, the light source has no specific orientation and does not generate any light gradient. The robots are simply equipped with a light intensity detector. When the lit area is so large that the robots can enter it with a very weak probability of encountering each other, the robots can safely be considered physically independent. Provided that they can translate the measure of the light intensity into the logical information — to be or not to be in the lit area (which requires the sole knowledge of a suited light intensity threshold) — the problem reduces to a series of independent exploration problems. The threshold itself can either be set in an ad-hoc way, or be the result of a learning process.

Here, we are interested in the case where the size of the lit area is too small to host the whole swarm. As a result, the robots must pack together to achieve the task. In such a context, it is clear that the steric interactions between the robots, the fact that robots cannot overlap, are likely to play a significant role. Typically, one expects the locomotion of the robots and the communication amongst them to be affected by the local high density of robots in a non trivial way. In the case where the robots policy results from a reinforcement learning algorithm distributed across the swarm, the learning process itself is prone to be impacted. Dealing with this complexity is a major challenge for the future of swarm robotics.

We investigate the above issues experimentally in a swarm of \(N = 64\) robots performing the phototactic task described above. To do so, we mount Kilobots, a low cost centimetric robots specifically designed for swarm robotics experiments [15], inside 3D printed exoskeletons, to form what we call a Morphobot. The exoskeletons explicit the steric and mechanical interactions between robots upon collision and increase the mobility of the Kilobots. As a first step, we provide evidences of the role of the steric interactions in the realization of the phototactic task and demonstrate that, in a setting without learning, a small morphological change of the exoskeleton can significantly alter the result of the task. A physical modeling of the robot dynamics allows us to capture the underlying mechanism responsible for this effect of the morphology. As a second step, we deploy a decentralized reinforcement learning algorithm on the swarm. As a proof of concept, the learning task is simplified to that of the light detection, while the dynamical behaviors are preset. We experimentally demonstrate the learning efficiency and show that the learning reduces the dependency on the morphology. Interestingly, our results point to the fact that the best learning is not necessarily obtained with the morphological and dynamical settings performing the best in the absence of learning.

II. THE MORPHOBOTS

The Morphobots are Kilobots [15] augmented with 3D printed exoskeletons. The Kilobot is a 3.4 cm tall, 3.3 cm diameter units standing on 3 rigid legs made of thin metallic rods. The Kilobots are equipped with two vibrators. When either is activated, the Kilobot turns at about 45 deg per second. When both are activated, the Kilobot moves forward. The power source is a rechargeable lithium-ion battery that can power the Kilobot for a few hours. The individual Kilobot is equipped with an infrared transmitter and receiver so that they can communicate with each other. The transmitter of a robot sends light toward the surface which reflects up to the receiver of another nearby robot. Obtaining straight motion over long distances is difficult and requires lengthy calibration of the vibrators. The calibration itself drifts in time. Also the Kilobot is relatively slow.

To both improve the robot’s nominal speed, and eliminate the need for individual calibration, we augmented the kilobots with 3D printed exoskeletons. We made two tripod based designs (a pair of flexible legs and an opposing stiff leg), with both designs having a circular frame (diameter \(d = 4.8\)cm). The two designs differ in the fore-versus-aft positioning of their flexible legs. (See Fig. 1 A,B,C and Methods). The natural vibration of the semi-flexible legs is tuned near resonance with the vibration motors, in order to maximize the coupling with their drive. Using anisotropic flexible leg design (1mm thick, 6 mm wide and 20 mm long, see Supplementary Information), vibrations are coupled predominately to the forward stick-slip motion of the robots. As we shall see below, both designs differ in the way they orient their motion in response to an external force and we respectively call them fronters and aligners.

By tracking the motion of individual Morphobots, we find that aligners and fronters show near identical motility. Motility measurement were carried out by tracking the two dimensional coordinates of individual robots as a function of time \((\vec{r}(t) = (x(t), y(t)))\) [58] when running in the experimental arena. Only a low concentration of robots was used (less than 1% filling fraction) excluding the effect of collisions, and robots were tracked only when away from the walls of the arena. The mean square displacement, \(\text{MSD} = \langle \Delta r^2 (\Delta t) \rangle_N\), of an ensemble of noisy walkers, presents two dynamical regimes [59]. On short time intervals \(\Delta t\), the displacement of the robots are aligned with the direction of their individual orientation and the amplitude of the displacements is proportional to \(\Delta t\). The motion is said
to be ballistic: \( \langle \Delta r^2(\Delta t) \rangle_N = v_0^2 \Delta t^2 \), where \( v_0 \) is the instantaneous speed. On large time intervals, the noise has reoriented the robots several times and the robots explore the space randomly. The motion is said to be diffusive and \( \langle \Delta r^2(\Delta t) \rangle_N = 2D_0 \Delta t \), where \( D_0 \) is called the diffusion constant. The two regimes are separated by a crossover, which takes place on a time scale \( \tau \), to which corresponds a persistence length \( \langle \Delta r^2(\tau) \rangle_N^{1/2} \). Fig. 1D displays the mean square displacement measured for a set of Morphobots equipped with the two types of exoskeleton, together with their ensemble average. One sees that the difference in motility between the two designs is no larger than the robot-to-robot variation within each design. The observed ballistic nature of the mean square displacement \( \langle \Delta r^2(\Delta t) \rangle_N \sim v_0^2 \Delta t^2 \) indicates that the persistence length of the motion is larger than the size of the circular arena (diameter \( D = 150 \) cm). Fitting the ensemble mean square displacement in the ballistic regime, we find that the mean nominal speeds are \( 5.2 \pm 0.9 \) cm/s and \( 4.8 \pm 0.1 \) cm/s for the aligners and fronters respectively.

Altogether, the anisotropic leg design allows us to use higher motor drive while keeping robots trajectories straight, alleviating the need for individual robot calibration, and resulting in a nominal speed of \( v_0 \approx 5 \) cm/s, a ten-fold increase, relative to the bare Kilobot design. A similar feat was achieved in bristle bots, where rows of soft legs were used [61] and in the commercially available HEXBUG [61, 63]. Using the differential drive, robots can pivot at 45 deg per second. Programming alternated straight and pivoting motion of random duration and random direction, we can set the Morphobots into a run and tumble motion [64], where, on average, the robots run straight during a run time \( \tau_{\text{run}} = 2 \) s, with a velocity to be specified, and tumble during a tumble duration \( \tau_{\text{tumble}} = 4 \) s. The total duration of a run-and-tumble sequence sets \( \tau_{\text{total}} = \tau_{\text{run}} + \tau_{\text{tumble}} \).

III. COLLECTIVE PHOTOTAXIS IN THE PRESENCE OF STERIC INTERACTIONS

To study the collective ability of a swarm (\( N = 64 \) bots) of either design to execute a phototactic task, we devised a light patterned circular arena (diameter \( D = 150 \) cm), with an un-oriented light spot (diameter \( b = 36 \) cm) free of local gradients (see Fig. 2A). The absence of local gradients exclude the traditional steepest decent search algorithm [17, 61], as the light field is essentially flat. The lit region covers about \( \sigma = 6\% \) of the arena and can contain a maximum of 50 Morphobots, a little bit less than the whole swarm.

As a first step we consider ad hoc calibrated swarms, without learning, where the morphbots policy are defined as follows. Each Morphobot continuously monitors local light intensity: when in the dark (below a predefined intensity threshold, \( P_{th} \)) the robot performs run-and-tumble with a running velocity \( v_0 \). When entering a bright region (light intensity above \( P_{th} \)), the robot stops, \( V_1 = 0 \). At early times, when there are only a few robots in the lit region, the two designs show similar performance, that can be quantitatively captured as a diffusion limited process (see Fig. 2F), with the kinetic reaction constant found using Smoluchowski first
FIG. 2. The collective ability of a robotic swarm to phototaxis is constrained by steric interactions. A a swarm of phototactic robots is placed in an arena patterned with an un-oriented light field. As there are no local gradients to guide a robot to the light source, robots perform a simple run-and-tumble in the dark, and stand in the light. B When a fronter arrives to the light spot, it collides with robots already taxied at the perimeter, but progressively pushes its way in. Repeating this process results in a the complete coverage of the light spot. (C) When an aligner arrives to the light spot D it slides along the wall of robots and leaves. This leads to only partial filling of the light spot (E), leaving an empty void in the center. F The resulting fraction of swarm in the light as a function of time shows that at early times (10mins), both design show equal performance (consistent with a diffusion limited reaction), at later times, fronters out performs aligners. Scale bars 20 cm.

The reason is that the collective success of each swarm depends on the morphological interactions of a robot impinging from the dark region on the first layer of robots stopped at the perimeter of the light region (Fig. 2B,D). From this point on-wards, the phototaxis rate slows down and deviates from the simple diffusion limited reaction, as crowding of taxiing robots at the perimeter of the light spot physically prevent further robots from entering the lit region (Fig. 2C,E). This steric effect is quantitatively more pronounced for the aligner design, as can be seen for the overall performance given by the fraction of the swarm that has entered the lit region on long times (Fig. 2F).

IV. FRONTERS VS. ALIGNERS RESPONSE TO AN EXTERNAL FORCE

We measure the response of the two Morphobots designs to an external body force by placing the robots on a plane tilted by a small angle, $\varphi$. The two designs have a striking different response: aligners go down hill, while fronters climb against gravity (Fig. 3A,B). The effect is consistent, regardless of the initial orientation of the robot being along the positive y direction (down-hill to its right), or negative y direction (down-hill to its left), excluding the effect of an internal left/right bias. We systematically increase the slope (see Fig. 3C), and find that with increasing $\varphi$, the trajectories turn increasingly curved, until the robot orients down-hill (aligner) or up-hill (fronter).

To quantify the different responses of the robots to an external body force, we use a model introduced in the context of active matter to describe self-aligning particles [62]. A Morphobot has two translational $\vec{r} = (x, y)$ and one orientational $\hat{n} = (\cos(\theta), \sin(\theta))$ degree of freedom. In the over damped limit, the deterministic dynamics of the velocity $\vec{v}$ and orientation $\hat{n}$ of a Morphobot with a nominal speed $v_0$ subjected to a body force $\vec{f}$ (see Fig. 3D) obey the following equations:

$$\frac{d\vec{r}}{dt} = \vec{v} = v_0 \hat{n} + \mu \vec{f}$$  
$$\frac{d\hat{n}}{dt} = \kappa (\hat{n} \times \vec{v}) \times \hat{n},$$

where $\mu$ is the particle mobility and $\kappa$ is the re-orientation strength. The first equation simply states that the velocity is the additive combination of the propulsion velocity $v_0 \hat{n}$ and the speed induced by the external force, $\mu \vec{f}$. The second equation contains the key ingredient of the model. It describes the reorientation of the Morphobots along
the direction of its motion. $\kappa$ controls the strength of this reorientation process. It is signed and has units of curvature (1/length). When $\kappa$ is positive, respectively negative, the Morphobots align towards, respectively opposite to, the direction of the external force. For a constant force in the $x$ direction $\vec{f} = f\hat{x}$, with $f = mg\sin\phi$, where $m$ is the mass of the Morphobot and $g$ is the acceleration of gravity, the dynamical evolution of the orientation, $\theta$, relative to the $x$ axis obeys a closed form equation:

$$\frac{d\theta}{dt} = -\kappa \mu f \sin\theta.$$  

Eq. 3 is the equation for a simple pendulum, which can be solved directly to give $\tan\frac{\theta}{2} = e^{-\kappa\mu ft}$ (for initial condition $\theta (t = 0) = 90^\circ$). In the limit where $v_0 \ll \mu f$, the orientation can be expressed as a function of the arclength, $\theta (s)$, and is found to be independent of the external force and mobility (see Supplementary Information),

$$\theta (s) = 2\tan^{-1} (e^{-\kappa s}).$$  

To find the alignment parameter, $\kappa$, we tracked the orientation of individual robots as they proceed across the inclined plane (see Fig. 3E and Methods). In the absence of external force (zero slope, $\phi = 0^\circ$), both fronters and aligners move on roughly straight lines. For steeper inclines ($\phi = 6^\circ$ and $\phi = 6^\circ$), both designs show similar rate for the convergence of the orientation. Fitting the evolution of the orientation to Eq. 3 we extract the alignment parameter, $\kappa$. The magnitude of $\kappa$ for both designs is similar, but with opposite signs: $\kappa_{\text{aligner}} \approx -\kappa_{\text{fronter}} \approx 0.3/d$. At a slope near zero, response is dominated by the robots internal bias (left/right curving, and internal noise), but averages to zero. Scale bar 20 cm.

The measured response of the orientation to an external force allows us to understand the different phototactic ability of the two designs: morphological interaction with a wall (or with a stationary barrier of robots) can be seen as a series of collisions, at each of which the robot experiences an external force (see Fig. 4A). With each collision, a Morphobot slightly re-orient: aligners turn parallel with the applied force (and away from the wall) while fronters turn anti parallel with the applied force (and towards the wall). This is confirmed by the direct observation of Morphobots sent to move towards a wall of stationary robots (see Fig. 4B). On the average, both design spend similar time ($\sim 5\tau$) near the wall, but aligners perform a path along the wall twice longer than that of the fronters (see Fig. 4C and Supplementary Information).
A normal force (red). Once aligned along the wall, there is no external force. When a Morphobot collides with a wall, each impingement can be seen as a momentary normal force. A. Aligners turn to align their orientation (black arrow) with the normal force (red). Once aligned along the wall, there is no further rotation as there are no more collisions. B. When a fronter collides with a wall, its orientation (black) turns anti-parallel with the normal force (red), and the robot continues to push into the wall. When programmed to run-and-tumble, aligners and fronters interact differently with a wall of stationary robots where aligners (A) cover much larger distance along the wall relative to fronters (B). C. Multiple collision experiments show that aligners and fronters spend similar time near a wall (3τ) yet aligners move twice as fast.

V. PHOTOTAXIS USING DISTRIBUTED ON-LINE REINFORCEMENT LEARNING

So far we considered an ad hoc calibrated swarm, where the light intensity threshold for phototaxis, $P_{th}$, was pre-programmed. However, the threshold value is not accessible beforehand when deploying a swarm in a real environment [12, 42, 43]. As several authors before [55, 56, 66], we propose to resort to machine learning to learn an efficient phototactic behavior using an Artificial Neural Network as controller. Learning is implemented as an evolutionary algorithm, which incrementally refines behavioral policies. Such algorithm requires a large number of trials and errors. Those are usually conducted by a central computer, and is made possible by the use of robotic simulations that both speed up the process and provide an easy access to data. Several works have shown that aggregation strategies evolved in simulation can then be transferred to real robots, with adjustments required to cope for the discrepancy between simulated and real robots [57, 67].

As compared to such classical evolutionary learning schemes, we face a critical problem as physical interactions between robots are here extremely challenging to simulate. The mathematical model we proposed earlier meets its limits where many-to-many simultaneous collisions occur, which makes it unlikely to resort to robotic simulations. This is a well-known problem in robotics that as been coined under the term of reality gap [68], and has also been recently exposed in swarm robotics [57]. In our case, this entails that learning should be conducted directly in the real world. Moreover, we consider deployment in an unknown environment, where a central coordinator for learning is missing. This departs from a fundamental hypothesis in evolutionary robotics that learning is orchestrated by a single central computer, who accesses performance assessment and individual policies, even though execution is decentralized over the swarm [39, 70, 71].

In order to provide a solution that can be deployed in the real world without a global supervisor (human or computer), we resort to a distributed on-line evolutionary reinforcement learning scheme, where each robot implement part of the learning algorithm [38, 39]. Each individual robot runs two processes in parallel (see Figure 5). First, a robot’s sense-act cycle corresponds to decision making, computing motor outputs from sensory inputs by using a control function (here, a Perceptron) with parameters $w^i$. Second, a robot’s learning cycle manages communication with nearby robots that are within reach of the communication apparatus (here, an infrared communication device with a range of approx. 7 cm around the robot, see Fig. 5C, and Supplementary Information), as well as message management. Each robot also performs a self-assessment of its performance with respect to the task at hand, here the amount of light sensed through its light sensor in a bounded time window:

$$\rho^i = \frac{1}{M} \sum_{k=0}^{M-1} P^i_k,$$  \hspace{1cm} (5)

where $P^i_0$, $P^i_{M-1}$ are the past $M = 250$ instantaneous light intensities measured by robot $i$ at a light sampling rate of 2Hz (see Methods). Performance assessment $\rho^i$ and control parameters $w^i$ are exchanged among nearby robots, and each robot updates its own control function parameters with that of its neighbours if they fare a better performance. In practise, for any two robots $i$ and $j$, robot $i$ (resp. robot $j$) transfers all its control parameters $w^i$ (resp. $w^j$) and performance assessment $\rho^i$ (resp. $\rho^j$) to the other robot. Values of all control parameter $w^i$ for robot $i$ are overwriten only if $\rho^i < \rho^j$, and symmetrically for robot $j$ (see Supplementary Information). As both robots communicate with the other, it means that whenever $\rho^i \neq \rho^j$, communication between two robots will result in one robot completely copying the strategy of the other by taking its control parameter values. As noted by other authors [71], this learning process is similar to social learning in nature, where successful survival and mating strategies compete with one another to invade the population, hopping from one individual to another [74, 75].

We implement the decentralized learning algorithm on a swarm of Morphobots and evaluate its performance in several conditions. There are two main differences between the experiments, depending on the morphology used, and the implementation of the dynamical behaviors. In terms of morphology, we compare performances obtained with a swarm of fronters and with a swarm of
FIG. 5. A distributed on-line reinforcement learning algorithm. A The embedded algorithm for updating the phototactic policy of each robot, is composed of two main moduli: a sense act cycle and a learning cycle. B The sense act cycle is encoded into a Perceptron: external stimuli are weighted and summed, evaluated through a sigmoid function, the output of which is binarized to select an action. The weights, \( w \), define the policy \( \pi \) of the robot. C The learning cycle is based on the conditional diffusion of the policies across the population, biased according to a reward function \( \rho \) evaluated by each robot individually: when robot 2 receives a message containing the policy control parameters \( \{w_1\} \) and the reward \( \rho_1 \) of robot 1, robot 2 will inherit the new policy, if it is associated with a higher reward \( \rho_1 > \rho_2 \). D-G A robot (black curve) already in the light source (C) departs E and broadcasts its reward and policy (F) to a robot in the dark (red curve), which inherits the new policy (G). Note that the broadcasting robot may have a wrong policy, effectively spreading “fake news”, with neither robot knowing what is the best policy for phototaxis.

FIG. 6. A robotic swarm initialized with a random set of policies can collectively learn a successful phototactic strategy using a decentralized reinforcement learning algorithm. When robots are programmed to completely stop when the photoperceptron is triggered (\( V_1/V_0 = 0 \), dashed curves) learning is slower than when robots are moving slightly when the photoperceptron is triggered (\( V_1/V_0 = 1/15 \), solid curves).

aligners. In terms of dynamical behaviors, robots choose between a "fast" and "slow" running velocity for their run and tumble dynamics. The fast running speed is set, as before, to \( V_0 \approx 5 \text{ cm/s} \). The slow running velocity is either set to \( V_1 = 0 \) or \( V_1 = V_0/15 \) (the value of which is discussed below). The other parameters of the run and tumble motion remain unchanged.

The control function is a Perceptron with two inputs, \( x_0 = 1 \), the bias and \( x_1 = P \), the light intensity, and one output \( y = \Theta(h(w_0 + w_1 P)) \) the binary choice among the two dynamical behaviors (fast or slow run and tumble), where \( \Theta \) and \( h \) respectively denote the Heaviside and the sigmoid functions. Due to the Morphobot’s very limited computation and communication capabilities, inherited from the Kilobots, decision making is performed by a Perceptron using only integer values, both for input and output values and weights. In particular, weights are discretized and take a limited number of values (2 bits per weight). Weights are randomly initialize on each robot before the experiment starts, implying that the swarm starts with robots that display different behavioral strategies. Moreover, robots are initially randomly located in the arena.

We report here results obtained with a swarm of 64 Morphobots. Each experiment runs for one hour (300\( \tau \)) and the proportion of swarm in the light is monitored and traced. Figure 6 show the learning curves of a typical run for each of the four setups evaluated. In all four setups, the number of robots in the light starts from a similar value, which corresponds to a uniform distribution of robots in the whole arena, and quickly climbs up as learning successfully leads to aggregation in the light. We observe a stark difference between using \( V_1/V_0 = 0 \) and \( V_1/V_0 = 1/15 \), with the latter notably improving phototactic performance. Interestingly, the difference in performance related to the two different morphologies, which we reported in the absence of learning, (Fig 2F) appears to be washed away when learning is involved.
VI. DECENTRALIZED LEARNING AS A CONDITIONAL DIFFUSION

Setting $V_1$ to zero quenches the learning kinetics. There is a rather intuitive explanation as robots which have learned the correct threshold cannot propagate their policy as they remain standing in the light. Also, robots with a too low threshold will remain in the dark, not moving, simply because $V_1 = 0$. While setting $V_1 > 0$ could be enough to avoid this pitfall, the balance between $V_1$ and $V_0$ can be optimally set for learning, by describing learning as a diffusion process.

We first quantify the slow and fast run and tumble dynamics. Fig. 7 displays the mean square displacements, $\langle \Delta r^2 (\Delta t) \rangle$ plotted as a function of $\tau = \Delta t/\tau_{\text{total}}$, measured for the two dynamical regimes. In both cases we observe the expected crossover between the ballistic $\langle \Delta r^2 (\Delta t) \rangle \sim \Delta t^2$ and the diffusive $\langle \Delta r^2 (\Delta t) \rangle \sim \Delta t$ regime for timescales respectively smaller and larger than $\tau = 1$, corresponding to $\Delta t = \tau_{\text{run}}$. The long time mean square displacement in both cases is diffusive ($\propto t^1$), with largely different diffusion constants, $D_0 \approx 5.5cm^2/s$ and $D_t \approx 0.06cm^2/s$. The inset clearly illustrates the two different dynamical behaviors. In the case of the slow running speed, the robots have hardly explored space in 20 minutes, while they cover distances comparable to the arena size in the case of the fast running speed.

We further proceed by monitoring the performance of individual robots, preset with different policies. The light threshold is set by manually adjusting the weights $\{w^0_i, w^1_i\}$, for the robot’s sense-act neural network, which defines the robot’s policy. This allows to set the shape of the sigmoid function relating the light intensity to the output of the perceptron before binarizing it. Fig. 7B displays the three preset sigmoid functions together with $\Phi = (T_{\text{in light}}) / T$ the fraction of time spent in the light by an individual robot preset accordingly. The fraction of time spent in the light for a robot with a wrong threshold, $\Phi_\text{w}$, are roughly the same regardless of the threshold being too high ($P^\text{high} = 1.5klux$) or too low ($P^\text{low} = 0.3klux$): $\Phi_\text{w} \approx \Phi^\text{high} \approx \Phi^\text{low} \approx 0.06$. In both cases, the robots’ dynamics are equally agnostic to the light and their fraction of time spent in the light is roughly the area fraction covered by the light $\Phi_\text{w} \approx \sigma$. Robots with a correctly calibrated threshold, $P^\text{th} \approx P^\text{light}$, show better than random performance, spending a greater fraction of the time in the light, $\Phi_C \approx 0.35$. We note that the reward function (Eq. 5) can be mapped to a temporal average over the past memory kernel of duration $T_M = M/\nu \approx 120$ s, where $\nu = 2Hz$ is the robots’ light sampling rate. In the case where $T_M$ is large enough, the reward is $r^\text{s} \approx \Phi$.

We next turn to analyze the collective performance of the preset swarm by monitoring the time evolution of $F(t) = \langle N_{\text{in light}} \rangle / N$, the fraction of the swarm that is in the light. Figure 7C displays the swarm fraction in the light as a function of time for a phototactic swarm with individuals’ policies preset to high/low/correct/or random thresholds. It shows that also collectively, both high and low are equally agnostic to the light spot. It also shows that the randomly calibrated swarm is agnostic too. More importantly we note that the measured fraction of the swarm, $F$, converges towards a value consistent with the individual robots’ fraction of time spent in the light $\Phi$, indicating a good mixing property of the dynamics. Finally, the significance of the velocity ratio $V_1/V_0$, on the collective performance of a preset swarm is obtained by conducting a series of phototactic experiments. Starting with 32 or 64 robots, we preset all robots with a correct policy, and vary the speed ratio between 0 and 1 (see Fig. 7D). In order to exclude the steric effects, and focus on the effect of the velocity ratio, the lit region is here set to allow the whole swarm inside. At steady state, when $V_1/V_0 = 0$ nearly all the swarm enters the light ($F_{\text{ss}} \approx 1$). Conversely, when the speed ratio approaches unity $V_1/V_0 = 1$, the steady state percentage of the swarm that reaches the light is roughly equal to the light area fraction $F_{\text{ss}} \approx \sigma$ meaning that the swarm is agnostic to the light. Between speed ratios of 0 and 1, $F_{\text{ss}}$ displays a monotonic decrease. This dependence of $F_{\text{ss}}$ on $V_1/V_0$ is captured from a flux balance argument. At the steady state, the entrant flux in the light region $J_{\text{in}} \approx c_0 V_0$ equals the exiting flux $J_{\text{out}} \approx c_1 V_1$. Using the definitions $c_0 = 9 N_{0,1} / N$, with $N_{0,1}$ the number of robots respectively outside and inside the light, one finds:

$$F_{\text{ss}} = \frac{1}{c_1} (\frac{1}{2} - 1) + 1,$$

which links individual behavior (ratio of the speeds in and out of the light) and performance (swarm in light). The curve in Fig. 7D shows that $F(V_1/V_0)$ is a tight upper bound of the values taken by the swarm in the steady state.

We are now in position to discuss how the choice of $V_1/V_0$ results from a delicate balance between two constraints, the need for (i) a good performance as quantified by a large fraction of the swarm in the lit region in the steady state and (ii) a fast diffusion of the correct policies, so that learning can take place during the life cycle of the robots, here the batteries life. It is worth mentioning that solving for the exact temporal evolution of the fraction of robots inside the lit region, taking into account the steric interaction, and the correlations it induces in the probability density of presence of the robots is an important challenge, that lies beyond the scope of the present work. We shall thus discuss how the velocity ratio $V_1/V_0$ impacts the overall performance of our swarm, following a simplified description.

On one hand, from the point of view of the asymptotic performance, and following the data obtained for the preset swarm in the absence of steric interactions (Fig. 7D), it is legitimate to consider that the best choice consist in setting $V_1/V_0 = 0$. On the other hand, on top of the intrinsic dynamics observed on Fig. 7C, when only ad-hoc robots are present, the pace at which the asymptotic performance is reached is controlled by the
learning rate, defined as the rate at which the fraction of robots with the correct policy increases. We have noticed that robots with too low or too high threshold values are equally wrong. We therefore limit the analysis to the situation where there are only two populations of robots, the robots with a wrong policy, \((W)\), and those with a correct one, \((C)\). In the simplest setting, the concentration of robots with correct policy then obeys the following kinetic equation:

\[
\frac{dc_C}{dt} = k (f_{CW} - f_{WC}) c_C c_W,
\]

where \(c_C, c_W = c_0 - c_C\) are respectively the concentration of robots with correct and wrong policy; \(c_0\) is the overall robot concentration in the arena, \(k\) is the encounter rate constant, and \(f_{CW}\), respectively \(f_{WC}\), are the transfer probability of \(W\), respectively \(C\), becoming \(C\), respectively \(W\), upon encounter. Optimizing the learning rate consist in ensuring \(\frac{dc_C}{dt} > 0\) and maximizing it. The transfer probabilities only depend on the difference in rewards \(f_{CW} = f(\Delta \rho_{CW})\), with \(\Delta \rho_{CW} = \rho_C - \rho_W\). The rewards are stochastic variables and their values do not always reflect the true quality of their associated policy. There should thus be the largest possible separation between the average reward of robots with correct policy and that of robots with wrong policy. By identifying the
reward \( r \) with the fraction of time spent by a robot in the light \( \Phi \), and using the empirically observed ergodic link relating \( \Phi \) and \( \mathcal{F} \), we obtain that the mean reward for the correct policy \( \rho_C \) decreases from 1 when \( V_1/V_0 = 0 \) towards \( \sigma \) when \( V_1/V_0 = 1 \). From this dependence, one conclude that maximizing \( f_{CW} - f_{WC} \), hence the learning rate, also leads to the choice of \( V_1/V_0 = 0 \).

However the learning rate also depends on the kinetic rate constant, \( k \propto D_{wc} = D_c + D_w \), where \( D_c, D_w \) are the diffusion constants of the robots with the respectively correct and wrong policy. The robots obey meta-behaviors with two distinct dynamics: running at velocity \( V_0 \), respectively \( V_1 \), with the effective diffusion constants \( D_{0,1} = \frac{V_{0,1}^2}{2} \), as given by the long time diffusion of a run and tumble process, where \( \tau \), the persistent time, takes the same value in both dynamics. A wrong robot always moves at the same speed, and, assuming low and high thresholds are equally probable, the mean diffusion of a wrong robot is \( D_w = \frac{1}{2} (D_0 + D_1) \). A correct robots switch dynamical behavior, according to its location in or outside the light region. Accordingly, we estimate its effective diffusion constant as the weighted average \( D_c = (1 - \mathcal{F}_{ss}) D_0 + \mathcal{F}_{ss} D_1 \), where \( \mathcal{F}_{ss} \) depends on \( V_1/V_0 \) as given by equation (6). Regrouping the above results, we find \( k \propto D_{wc} \propto D_0 K(V_1/V_0; \sigma) \), where the function \( K \), is a 3rd degree rational function (see Sec. IX.F.2 in Supplementary Information), is plotted on Fig. 8. The best learning rate is obtained for \( V_1/V_0 = 1 \) and, conversely, the best asymptotic performance \( \mathcal{F}_{ss} \rightarrow 1 \), achieved when \( V_1/V_0 = 0 \), results in the slowest learning rate. The selection of \( V_1/V_0 \) must therefore obey a compromise. From the shape of \( K \), one sees that the fastest growth of the learning rate takes place for small values of \( V_1/V_0 \), meaning that the learning swarm will arrive at its steady state phototaxis with only a small penalty to the performance, while for larger \( V_1/V_0 \), there is a much slower improvement to the learning rate. The selection of \( V_1/V_0 = 1/15 \) results from this delicate balance where \( V_1/V_0 \approx 0.06 \).

While this particular value is specific to phototaxis, the general conclusion is that robots should always physically diffuse (i.e. they should never stop) so that efficient strategies can spread throughout the swarm. Of course, as shown previously, there is a cost to distributed online learning as the optimal behavioral strategy should both address the user-defined task and ensure its diffusion throughout the swarm, which can hinder a sub-optimal task performance when compared to a pre-programmed swarm where behavioral strategy diffusion is not an issue.

VII. CONCLUSION

In this paper, we considered dense swarms of robots where collisions between robots are bound to occur on a frequent basis. We show that the robots’ physical morphology plays an important role in terms of collective behaviors, which is extremely difficult to model as small changes in the morphology can lead to very different outcomes in terms of dynamical behaviors even when one-to-one collisions are considered. Applying tools from statistical physics of active matter to swarm robotics, we show that collective self-aggregation can be attained with limited use of software-controlled behaviors by mainly capitalizing on the morphological properties of robots, demonstrating morphological computation at work within the swarm.

We also showed that distributed on-line evolutionary reinforcement learning can be implemented in a dense robot swarm, to learn how to exploit or mitigate physical contingencies after deployment in the open. We show that this kind of artificial social learning, when robots locally exchange information about their behavioral strategies, can be modelled as a conditional diffusion process. In particular, we reveal that there exists an optimal diffusion flow that balances between exploration (diffusing winning strategies) and exploitation (maximising efficiency), whose relevance goes beyond the self-aggregation with phototaxis task considered here.

Future physical models of a learning swarm should go beyond the mean-field approximation used here, and account for the spatial density fluctuations (as is done for describing the motility induced phase separation observed in active fluids [64]). From that point of view, our work shows that building upon the recent theoretical progress made in describing population of active particles, one can envision a rapid development of a statistical mechanic description of smart active matter, for future swarm engineering.

From a robot learning viewpoint, we can also extend social learning in swarm robotics by going beyond innovation and diffusion of behavioral strategies. A natural extension can be motivated by observing the cumulative cultural evolution at work in the natural world [75], which is built on the reformulation and combination of previously learned behaviors in order to continuously discover and retain new behaviors of growing complexity.
Finally, we can consider distributed on-line evolutionary reinforcement learning on the most common setting for swarm robotics, where robots can communicate in their vicinity but do not physically interact with one another. An extreme case is that of sparse robot swarms, where individual robots or groups may be temporarily disconnected from the whole [76]. While dense and sparse robot swarms may be considered as very different configurations, it is likely that a robot swarm in the open may have to switch from one configuration to another depending on the current context and objectives, therefore requiring versatile social learning capabilities to make the best of each situation.
ing, swirling and stasis in sequestered bristle-bots. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **469**, 20120637 (2013). URL: https://royalsocietypublishing.org/doi/10.1098/rspa.2012.0637 1302.5952.

[61] Debais, A. *et al.*. Boundaries Control Collective Dynamics of Inertial Self-Propelled Robots. *Physical Review Letters* **120**, 188002 (2018). URL: https://link.aps.org/doi/10.1103/PhysRevLett.120.188002

[62] Dauchot, O. & Démery, V. Dynamics of a Self-Propelled Particle in a Harmonic Trap. *Physical Review Letters* **122**, 1–5 (2019). 1810.13303.

[63] Boudet, J. F. *et al.*. From collections of independent, mindless robots to flexible, mobile, and directional superstructures. *Science Robotics* **6**, eabd0272 (2021).

[64] Tailleur, J. & Cates, M. E. Statistical Mechanics of Interacting Run-and-Tumble Bacteria. *Physical Review Letters* **100**, 218103 (2008). URL: https://link.aps.org/doi/10.1103/PhysRevLett.100.218103 abs/0803.1069.

[65] Smoluchowski, M. V. Drei Vortrage uber Diffusion, Brownsche Bewegung und Koagulation von Koloidteilchen. *Physik Zeitscher* **17**, 557–585 (1916).

[66] Dorigo, M. *et al.*. Evolving self-organizing behaviors for a swarm-bot. *Autonomous Robots* **17**, 223–245 (2004).

[67] Gauci, M., Chen, J., Li, W., Dodd, T. J. & Groß, R. Self-organized aggregation without computation. *The International Journal of Robotics Research* **33**, 1145–1161 (2014).

[68] Jakobi, N., Husbands, P. & Harvey, I. Noise and the Reality Gap: The Use of Simulation in Evolutionary Robotics. *Lecture Notes in Computer Science* **929**, 704–720 (1995). URL: citeseer.ist.psu.edu/jakobi95noise.html

[69] Nolfi, S. & Floreano, D. Evolutionary Robotics: the Biology, Intelligence, and Technology (MIT Press, Cambridge MA, USA, 2000).

[70] Floreano, D. & Mattiussi, C. Bio-inspired artificial intelligence: theories, methods, and technologies (MIT press, 2008).

[71] Heinerman, J., Drupsteen, D. & Eiben, A. E. Threefold Adaptivity in Groups of Robots: The Effect of Social Learning. In Silva, S. (ed.) *Proceedings of the 17th annual conference on Genetic and evolutionary computation*, GECCO ’15, 177–183 (ACM, 2015).

[72] Heinerman, J., Rango, M. & Eiben, A. E. Evolution, individual learning, and social learning in a swarm of real robots. In *Proceedings - 2015 IEEE Symposium Series on Computational Intelligence, SSCI 2015*, 1055–1062 (IEEE, 2016).

[73] Fontbonne, N., Dauchot, O. & Bredeche, N. Distributed on-line learning in swarm robotics with limited communication bandwidth. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, 1–8 (2020).

[74] Heyes, C. M. Social learning in animals: categories and mechanisms. *Biological Reviews* **69**, 207–231 (1994).

[75] Whiten, A. The burgeoning reach of animal culture. *Science* **372** (2021).

[76] Tarapore, D., Groß, R. & Zauner, K.-P. Sparse robot swarms: Moving swarms to real-world applications. *Frontiers in Robotics and AI* **7**, 83 (2020). URL: https://www.frontiersin.org/article/10.3389/frobt.2020.00083

[77] Edelstein, A. D. *et al.* Advanced methods of microscope control using μManager software. *Journal of Biological Methods* **1**, e10 (2014). URL: https://jbmmethods.org/jbm/article/view/109

[78] Rueden, C. T. *et al.* ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics* **18**, 529 (2017).

[79] Howse, J. R. *et al.* Self-Motile Colloidal Particles: From Directed Propulsion to Random Walk. *Physical Review Letters* **99**, 048102 (2007). URL: https://link.aps.org/doi/10.1103/PhysRevLett.99.048102 0706.4406.

**Acknowledgements**

We thank Y. Lahini and N. Oppenheimer. This work was supported by the MSR project funded by the Agence Nationale pour la Recherche under Grant No ANR-18-CE33-0006.

**VIII. METHODS**

**A. Exoskeleton design and manufacturing**

1. **Exoskeleton Manufacturing**

Exoskeletons were 3D printed using either one of the following 3D printers and materials to give similar results: *startAsys Objet350 Connex3* printer, using vero-Clear (modulus of elasticity $E = 1−2\, \text{GPa}$, density $\rho = 1.2\, \text{g/cm}^3$), and *3D systems projet 2500plus* printer using *VisiJet M2R-CL (MJP)* (density $\rho = 1.14\, \text{g/cm}^3$, Elastic modulus $E \approx 1\, \text{GPa}$), showing similar performance.

**B. Data acquisition and experimental setups**

1. **Phototaxis experiments**

Phototaxis experiments were performed on a 5 mm milky plexiglass sheet placed on the floor. A circular Lit region 6% of the arena was produced using a projector (EPSON EB-1795F), at RGB = (10,255,255), at approximately the spectral sensitivity peak (570nm) of the kilobot’s light sensor (TEPT5700, Vishay Semiconductors). To reduce interference of ambient light with the light sensor, the arena was lit using four desk lamps directed at the surrounding walls (for a homogeneous diffusive illumination) and covered with a red cellophone (where the robot’s light sensor’s sensitivity is at a minimum). A red cellophone was also covering the camera used for acquisition (PIXELINK.PL-D734MU), to reduce the saturation from the lit region, and imaging the robots. Image acquisition was carried out using *μManager* [77], at 2 frames per second.
2. Inclined plane experiments

Coupling between external force and robot’s orientation was measured by letting robots run on a plain made of 5 mm thick plexiglass placed at an incline, and images using SonyAlphaS7 camera, with video rate image acquisition at 30 frames per second.

3. Data analysis

Raw images were preprocess using ImageJ [78], or ffmpeg, followed by particle locating using trackpy package [58], or custom code, and trajectories were linked using trackpy [58].

IX. SUPPLEMENTARY INFORMATION

A. Exoskeleton design

Exoskeletons were design as a round chassis (diameter \( d = 4.8 \) cm) with an off-center cavity where the kilobot is pressed in. The exoskeleton stands on three, \( L = 2 \) cm tall legs — one leg is stiff and round, and two opposing legs that are flexible and with a rectangular cross-section (thickness \( T = 1 \) mm, and width \( W = 0.6 \) cm, see Fig. 1 in main text). The flexible legs were designed to have their first natural vibration near resonance with the vibration motors frequency (Pololu 10X2.0 mm vibration motors, \( f_{\text{motor}} \approx 250 \) Hz). This was done by estimating the legs to be elastic beams which are described dynamically using the Euler–Bernoulli beam theory for an homogeneous thin beam

\[
EI \frac{\partial^4 h}{\partial x^4} = -\mu \frac{\partial^2 h}{\partial t^2},
\]

where \( h \) is the deflection, \( E \) is the elastic modulus, \( \mu \) is the linear mass density (given by \( \mu = TW\rho \)), and \( I \) is the beam moment of inertia (given by \( I = \frac{WXT^3}{12} \)). In Fourier space, Eq. 8 becomes

\[
\omega = \sqrt{\frac{EI}{\mu} k^2},
\]

where \( \omega \) is the angular frequency, and \( k \) the wave number. The wavelength of the first harmonic \( \lambda_1 \) of a beam of length \( L \) fixed at one end is \( \lambda_1 = 4L \), along with Eq. 9 allows us to find the natural first harmonic of the leg is to leading order

\[
f_1 = \frac{\pi}{16\sqrt{6}} \sqrt{\frac{E T}{\rho L^2}}.
\]

Note the Eq. 10 is independent of the width. Given the materials and geometry used (see previous section), \( f_1 \approx 250 \) Hz \( \approx f_{\text{motor}} \).

B. Individual robot dynamics

1. Speed of an individual persistent particle

The effective diffusion constant, \( D_{\text{eff}} \), of a run-and-tumble particle, that covers a distance \( l_p \), during the time interval \( \tau_{\text{total}} \) is \( D = \frac{1}{4} \frac{l_p^2}{\tau} \) [79]. In the run and tumble mode, each Morphobot is programmed to run for \( \tau_{\text{run}} = 2 \) s at its nominal speed of \( v = 5 \) cm/s, then perform a tumble where it pivots on the average for \( \tau_{\text{pivot}} = 4 \) s. The step size is thus \( l_p = v \cdot \tau_{\text{run}} \approx 10 \) cm, and the total duration is \( \tau_{\text{tot}} = 6 \) s. The resulting effective diffusion constant is \( D_{\text{eff}} \approx \frac{1}{4} (10 \text{ cm})^2 / 6 \text{ s} = 4 \text{ cm}^2 / \text{s} \).

2. Particle orientation as a function of trajectory length under constant force

The dynamics of the orientation, \( \theta \), relative to the positive \( \hat{x} \) direction, for an active particle in the over-damped limit at the absence of noise is found to be that of a simple pendulum (Eq. 3 in main text):

\[
\dot{\theta} = -\kappa \beta \mu \sin \theta.
\]

It is helpful to formulate Eq. 11 as a function of \( s \), the trajectory’s traveled. In 2D, \( s \) is defined locally by the Euclidean distance:

\[
ds^2 = dx^2 + dy^2
\]
Dividing Eq. 12 by $dt^2$ reminds us that
\[
\left( \frac{ds}{dt} \right)^2 = \left( \frac{dx}{dt} \right)^2 + \left( \frac{dy}{dt} \right)^2 = v^2 \tag{13}\]

We can find $v^2$ by multiplying Eq. 1 by itself:
\[
v^2 = v_0^2 + (\beta \mu)^2 + v_0 \beta \mu \cos \theta. \tag{14}\]

Using the chain rule, $\frac{d\theta}{dt} = \frac{d\theta}{ds} \frac{ds}{dt}$, and plugging into Eqs. 13 and 14, we get an ODE describing $s$, the arclength:
\[
\frac{d\theta}{ds} = \frac{\kappa \beta \cos \theta}{\sqrt{v_0^2 + \beta^2 + v_0 \beta \cos \theta}}. \tag{15}\]

Equation 15 can be solved exactly, however it is more instructive to examine the limit of $\frac{\partial s}{\partial \mu} \ll 1$ which then simplifies to
\[
\frac{d\theta}{ds} = \kappa \cos \theta. \tag{16}\]

For $\theta(s = 0) = \frac{\pi}{2}$, Eq. 16 gives Eq. 4 in the main text:
\[
\theta(s) = 2 \arctan \left( e^{-\alpha s} \right). \tag{17}\]

C. Phototaxis without learning

1. Photoperceptron

Each robot has a photoperceptron encoded to measure environmental cues, take a specific action given that the perceptron fired or not. The minimal architecture for the photoperceptron implemented has a neural network of a single neuron (perceptron), with two inputs, bias $b = x_0$, and the locally measure light power $p = x_1$, and their associated weights $w_0$ and $w_1$, which together effectively define the threshold for firing. The perceptron output is, $p$, a logistic function of the form
\[
p = \frac{1}{1 + e^{-\mu}}. \tag{18}\]

where $h = \sum_{i=0}^{1} w_i x_i$

2. Sense-act cycle

The sense act-cycle uses the robot’s local light intensity measurement, where the robot chooses between two meta-behaviors: run and tumble or walk and tumble, given the photoperceptron is below threshold $P < 0.5$, or above threshold $P \geq 0.5$, respectively. In both cases the motion is a sequence of a moving phase (run/walk) and tumble phase. In the tumble phase, the robots randomly chooses a direction (clockwise or counterclockwise) and a random duration with a mean of 6 seconds. Randomizing the duration compensates for the robots’ internal dynamical variability, and environmental variability. For example, robots with stronger/weaker motors will take less/more time to tumble and complete the same reorientation. Similarly, a robot near a wall (or another robot) or in solitude. When in run mode, the robots sets both motor values to high, approximately moving at a straight line (persistence length much larger than the robot’s size), at its nominal speed $V_0 = v_0$. When walking, the robot moves for a duration $\tau_{\text{walk}}$, and then stands for a duration $\tau_{\text{stand}}$. The ratio between the two controls the mean speed of the robot when walking $V_1 = \frac{\tau_{\text{walk}}}{\tau_{\text{walk}} + \tau_{\text{stand}}} V_0$. When $\tau_{\text{walk}} = 0$ the robot stands when photoperceptron fires.

3. Phototaxis rate for low concentration non interacting robots estimated using first arrival time

The rate of robot arrival to the light spot at early times (where the lit region is mostly empty), can be found from the reaction rate constant, $k$, using Schmolokowski first order rate constant $k$, $\exp(-kt)$, of a Brownian particles with diffusion constant $D$, arriving at a target of size $b$ in an arena of diameter $d$ is given by
\[
k = \frac{2 \pi D}{\log \frac{d}{b}}. \tag{19}\]

The rate at which robot are subtracted from the dark region $\frac{dN_{\text{bulk}}}{dt}$, is given from a first order reaction rate
\[
\frac{dN_{\text{bulk}}}{dt} = -k N_{\text{bulk}}. \tag{20}\]

Which solves readily to $N_{\text{bulk}} = A \exp(-kt)$. Given an initial number of robots, $N_0$, the number of robots in the light, $N$, is given from the robot conservation $N_0 = N + N_{\text{bulk}}$,
\[
N(t) = N_0 (1 - e^{-kt}). \tag{21}\]

which is plotted in Fig. 2.

D. Steady state experiments with large lit region

To test how the speed ratio $V_1/V_0$ effects the number of bots in the light, phototaxis experiments were done by increasing the size of the lit region to 25% of the arena. Experiments were performed with either 64 bots or 32 bots (see inset in Fig. 7D). The speed ratio was set by controlling the $\tau_{\text{stand}}$ as described in Section IX C 2 changing the speed from standing in light $V_1/V_0 = 0$ to effectively being agnostic to the light $V_1/V_0 = 1$. The relative number of robots in the light was tracked over
the course of up to 2 hours, and the mean and standard deviation of the last 10% of the experiment duration were used to evaluate the steady state fraction of the swarm in the light, $F_{ss}$.

E. Phototaxis with learning

1. Performance of the learning swarm

We finally implement the decentralized learning algorithm on a real a randomly initialized swarm of $N = 64$ Morphobots and evaluate its performance in several conditions. Figure 4 provides the learning curves of two morphologically different swarms, composed respectively of aligners and fronters, under two dynamical conditions $V_1/V_0 = 0$ and $V_1/V_0 = 1/15$, and compare them with the performance of the same swarms under the same dynamical conditions when they are preprogrammed with the ad-hoc policies. First, in the case of the ad-hoc policies (continuous curves) we confirm our initial observations, namely that fronters perform better than the aligners. Second, a central observation is that the swarms do learn: the decentralized reinforcement learning algorithm operates in real swarm of physical robots despite their heterogeneities and all the imperfections inherent to a real experiment. We can also make more quantitative statements. As anticipated, and rationalized above, when $V_1/V_0 = 0$ (Fig. 4-A, the learning swarms have a hard time reaching the same performance as the swarms with ad-hoc policies. Setting the exploration of robots to zero $V_1 = 0$, quenches the learning kinetics ($k → 0$). Robots with correct threshold can not propagate their policy as they stand in the light; robots with low threshold remain standing in the dark. By contrast, when $V_1/V_0 = 1/15$ (Fig. 4-B, the learning swarms perform as well as the ad hoc calibrated swarm. Also, we note that the steady state difference between the swarm composed of aligners and that composed of fronters is washed away, suggesting that the aligners take an advantage during the learning process. We conjecture that the larger mobility of aligners along the boundary wall composed of robots, which rapidly surrounds the lit region (Fig. 4), allows for a faster diffusion of their correct policy.

F. Distributed online learning algorithm for phototaxis

Each robot in the population runs Algorithm 1. This Algorithm builds from the HIT algorithm originally presented in [73], with two important modifications. Firstly, the algorithm is designed with phototaxis as an objective (the $\rho$ objective function on line 9 measures the amount of light measured in the last $T$ time steps). Secondly, transfer of all control parameters is performed, which implies that whenever a robot receives a parameter set which is deemed to be performing better than its own, this parameter set is fully copied and overwrites existing parameter values. Mutation is not used in this implementation, as the number of robots used (64) and the low dimensionality of the search space ($\mathbb{R}^2$) allows for a sufficient amount of behavioral diversity in the initial population of robots (see [73] for a comprehensive study of transfer and mutation rate and operators).

It should also be noted that performance assessment ($\rho^i$) is not reset to zero after a parameter update, which implies that performance assessment of a newly updated policy is under-estimated for $T$ time steps. While this may temporarily lead to theoretically best-performing control parameters being overwritten before $T$ steps elapsed, the elitist selection scheme ensures that the original best-performing individual is never lost.

Finally, note that in the algorithm the observable $o$ contains information about the current light intensity $r$. They are separated for semantic reasons as $o$ is used as an input value for the policy $\pi$, while $r$ is used to update the performance self-assessment $\rho^i$.

Algorithm 1: The phototactic-HIT algorithm

Data:

- $i$: the unique identifier of the current robot,
- $T$: evaluation time,
- $\pi$: Policy function,
- $w^i$: Random uniform initialisation of policy parameters
- $R[T]$: Empty reward buffer of size $T$,
- $r$: Current light intensity,
- $\rho^i$: performance self-assessment during $T$,
- $a$: Null action vector,
- $o$: Null observation vector

1. begin
2. $t = 0$
3. while True do
4.     $o, r = \text{sense}()$
5.     $R[t \mod T] = r$
6.     $a = \pi(o|w^i)$
7.     $\text{act}(a)$
8.     if $t > T$ then
9.         $\rho^i = \sum_{k=0}^{T-1} R[k]$
10.        $\text{broadcast}(w^i, \rho^i)$
11.       if new message then
12.           $w^j, \rho^j = \text{decode message}$
13.           if $\rho^j > \rho^i$ then
14.               $w^i = w^j$
15.       end
16. end
17. $t = t + 1$
18. end
19. end
1. Translating an embedded reward function into a positive average decentralized learning

The goal is to link policy and environment to find a route for a successful decentralized learning scheme on the average. Learning kinetics:

The time evolution of the probability $p_C$, respectively $p_W$, of finding a robot $i$ with a correct, respectively wrong, policy is described by the following master equation:

$$
\frac{d}{dt} p_i = \omega_{C,W}^i p_i - \omega_{W,C}^i p_i, \tag{22}
$$

where $p_i = 1 - p_i^0$, and $\omega_{C,W}^i$, respectively $\omega_{W,C}^i$, are the transition rates at which the robot $i$ with a wrong policy turns into a correct policy and conversely. We assume the swarm is homogeneous and identify the probability of a given policy for a given robot with the concentrations and the transition elements read $\omega_{C,W} \approx k_{WC} C$, and symmetrically $\omega_{W,C} \approx k_{CW} W C$, where $k$ is the encounter rate, and $f_{CW}$, respectively $f_{WC}$, is the transfer probability of $W$, respectively $C$, becoming $C$, respectively $W$, upon encounter. As a result,

$$
\frac{d}{dt} c_C = k (f_{CW} - f_{WC}) c_C (c_0 - c_C), \tag{23}
$$

where $c_0$ is the overall concentration of robots in the arena.

To achieve mean collective, the factors of the product in Eq. 23 have to be non-negative, which amounts to

$$
f_{WC} > f_{CW}
$$

As the learning is decentralized, $f$ is an embedded function, and can depend only on information locally accessible to the robot. This includes the robot’s policy (here, the weights of its neural network, $\{w_i\}$), its measurement history, and the message received. Here we exclude a direct functional dependence of $f$ upon the policy as it will artificially bias the learning to a given trait (say turn left) instead of a desired fitness (say measure light). Therefor $f$ will depend on some function of the history of its own sensor input and the history of an encountered robot. In our minimal system the measurement history contains the $M$ most recent light intensity measurements, $I_{i \in 1..M}$ (stored as a rolling buffer, i.e FIFO). Since the robots have a finite bandwidth, in order to allow asynchronous performance, where communication does not slow down dynamics, message should be kept succinct. This means robots should only exchange a reduced version of their memory that can be contained within a message. This will be a function the memory, and will be called the reward function, $\rho = \rho (\{I_i\})$. In our implementation scheme, robots continuously broadcast a message with their reward and policy; a robot receiving a message compares the Broadcasted reward, $\rho_B$, to its self-reward, $\rho_S$, and if $\rho_B > \rho_S$ the policy will be deterministically adopted. The question then becomes what functional form does the local $\rho$ take, to satisfy global positive learning on the average: $\langle \rho_C \rangle > \langle \rho_W \rangle$.

Empirically we found that an ad-hoc calibrated swarm at steady state, individual robots with a correct policy spend more time in the light $\phi_C > \phi_W$, where the mean time in the light of an individual robot is measured by $\phi = \frac{1}{T} \int_0^T dt I(t)$. Identifying the reward function as the mean time in the light, $\rho \approx \phi$, allows us to bias the swarm towards learning the correct policy. We do this by computing the reward as the average of the light intensity buffer, $\rho = \frac{1}{M} \sum_{i=1}^M I_i$.

Recalling that the transfer probability function depends only on the received and embedded reward, the condition for learning simply becomes that on the average, correct robots have a greater reward:

$$
\rho_C > \rho_W
$$

We found that $\phi_C = \frac{1}{\tau_b (\frac{1}{10} - 1)}$ and $\phi_W = \sigma$, and so upon simplification, the condition for positive learning simply becomes:

$$
\frac{V_1}{V_0} < 1
$$

Note that this does not depend on the area fraction of the light spot, $\sigma$, nor the absolute magnitude of the velocities (only their ratio).

2. Learning rate constant

Knowing the mutual diffusion constant, $D_{WC}$, of Correct and Wrong particles, the learning rate constant $k$ in Eq. 23 is given by [65]:

$$
k = \frac{2 \pi D_{WC}}{\log \frac{1}{\tau}}.
$$

The mutual diffusion is given by the sum of the diffusion constants $D_{WC} = D_W + D_C$. The robots have two metabehaviors, moving fast and slow with the respective effective diffusion constants, $D_0$, $D_1$. The effective diffusion constant is given from the mean speed, $V_{i}$ of each behavior $D_i = \frac{V_i^2}{\tau}$, where $\tau$ is the persistent time (and equal across behaviors). A wrong robot always moves at the same speed, and assuming low and high thresholds.
are equally probable, the mean diffusion of a wrong robot is:

\[ D_W = \frac{1}{2} \left( D_0 + D_1 \right) = \frac{1}{2} D_0 \left( 1 + x^2 \right), \]

where we define \( x \equiv \frac{V_1}{V_0} \) as the relative slow down of the slower metabehavior. At steady state, the effective diffusion constant of the correct robots has to weigh the different time a robot stays at each behavior:

\[ D_C = (1 - \mathcal{F}_{ss}) D_0 + \mathcal{F}_{ss} D_1, \]

where \( \mathcal{F}_{ss} \) is given in Eq. [6]. Defining \( A \equiv \frac{1}{\sigma} - 1 \), the learning rate constant takes the following form:

\[ k = \frac{2\pi}{\log^2 B} \frac{A x^3 + 3 x^2 + 3 A x + 1}{2 A x + 2} \] (24)