PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Research

Cite this article: Kaminka GA. 2025 Swarms can be rational. *Phil. Trans. R. Soc. A* **383**: 20240136.
https://doi.org/10.1098/rsta.2024.0136

Received: 15 July 2024
Accepted: 25 November 2024

One contribution of 12 to a theme issue 'The road forward with swarm systems'.

Subject Areas:
artificial intelligence, robotics

Keywords:
swarms, rationality, game theory, multi-robot systems, distributed intelligence, collective intelligence

Author for correspondence:
Gal A. Kaminka
e-mail: galk@cs.biu.ac.il

THE ROYAL SOCIETY PUBLISHING

Swarms can be rational

Gal A. Kaminka

Department of Computer Science & Gonda Brain Science Center & BINA Nano-Technology Center Bar Ilan University, Bar Ilan University, Israel

GAK, 0000-0001-8341-322X

The emergence of collective order in swarms from local, myopic interactions of their individual members is of interest to biology, sociology, psychology, computer science, robotics, physics and economics. *Cooperative swarms*, whose members unknowingly work towards a common goal, are particularly perplexing: members sometimes take individual actions that maximize collective utility, at the expense of their own. This seems to contradict expectations of individual rationality. Moreover, members choose these actions without knowing their effect on the collective utility. I examine this puzzle through game theory, machine learning and robots. I show that in some settings, the *collective utility* can be transformed into *individual rewards* that can be measured locally: when interacting, members individually choose actions that receive a reward based on how quickly the interaction was resolved, how much individual work time is gained and the approximate effect on others. This internally measurable reward is individually and independently maximized by learning. This results in a equilibrium, where the learned response of each individual maximizes both its individual reward and the collective utility, i.e. both the swarm and the individuals are rational.

This article is part of the theme issue 'The road forward with swarm systems'.

1. The perplexing nature of cooperative swarms

Individual members of collectives must interact to achieve their goals, yet their *interaction capacity* is bounded. Individual interaction bounds arise from logical and physical limits on perception range, latency

and bandwidth, from analogous limits on the reach and impact of actions and from computational limits on processing incoming and outgoing signals.

By nature, individual interaction bounds do not change with the size of the collective, or its inherent requirements for interaction complexity. When a collective is small, or its inherent interaction requirements are sufficiently low, individuals have no problem interacting with others as needed. However, as collectives grow in size or interactions grow in complexity, individuals are prohibited, by their bounds, from keeping up. Discord and disorder ensue. These may be alleviated when individuals interact explicitly to *organize* the collective, forming *hierarchies* (management, command) and *divisions* (ministries, departments). Otherwise, however, collectives may become *swarms*.

For the purposes of this paper, I define *swarms* intuitively, as collectives of agents whose individual interaction capabilities permit them to interact with a few others (local interactions), but not with all (global interactions). They are flat organizations (no hierarchy), whose members are essentially anonymous and replaceable. Swarms permeate our natural, societal and artificial surroundings. They are found in the collective motion of animals [1–3], human pedestrians [4–6] and traffic [7,8] and multi-robot applications [9–17]. They inspire novel methods for optimization [18,19], safe collision-avoidance [20–24] and future medicinal molecular robots [25–29].

Swarms are investigated not (only) for their ubiquity but for their puzzling nature: they are fantastic exemplars of an ordered system-wide phenomenon arising out of synergistic local interactions between components. Swarm members cannot possibly know the collective state of the swarm, nor can they take actions that directly impact the collective state. Yet time and again, despite the limitations of their constituent members, we see examples of swarms achieving and maintaining an ordered (coordinated) collective state.

Cooperative swarms, whose behaviour is understood in terms of collective objectives, are particularly perplexing. Given a measure of collective utility, one could describe a cooperative swarm by its *collective rationality*, the pursuit of the swarm's members, in aggregate, to maximize the expected collective utility. This follows the definition of the *principle of rationality* [30–33] at the swarm level [34]. For example, the order of a swarm's collective motion can serve as a measure of collective utility, as could the total amount of food gathered by a foraging swarm, or the rate at which the food is gathered, or the number of items found as swarm members carry out a collective search. Cooperative swarms maximize these measures.

However, introducing rationality as a methodological lens raises significant difficulties when it comes to individuals. While the collective objectives of the swarm may be understood and formalized, their decomposition to individual, self-interested, rational decision making is not at all clear. Agents in cooperative swarms may sometimes need to take actions at their own resource expense, to benefit the swarm. This seems to contradict expectations of *individual rationality*, whereby agents maximize their individually perceived reward. Moreover, even assuming that agents simply adopt the swarm's collective utility as their own (thus they benefit with others from their individual expenses), swarm agents cannot possibly perceive (measure) the collective utility, as they are limited to local perception. They cannot perceive the collective effects of their actions, and cannot verify that their actions improve the collective utility.

For instance, suppose the collective utility is measured by the total number of food items brought to the nest by foragers. An individual agent can (at best) measure this only when it is at the nest. As it forages, it cannot use the number of food items as a guide to its decision making. Moreover, suppose an agent leaving the nest to forage is about to collide with an incoming forager. Just as it is likely better to let people off the elevator before attempting to move in, the incoming forager may need to back off; but this depends on whether it is holding a food item, how close it is to depositing it and what others (occluded or far) are doing. In swarms, the individuals do not know this information.

Given this difficulty in accounting for individual rationality in cooperative swarms, *Game Theory*, one of the fundamental principled tools for studying multi-agent interactions, is often

ignored in cooperative swarm research (see [35,36] for exceptions). Instead, existing approaches rely on task-specific black-box procedures that mechanistically proscribe individual behaviour that increases collective rewards [16], without regarding individual costs and gains. Examples include procedures for collective motion [37–42], area coverage [43–46] and foraging [47–50]. These bypass the question of individual self-interest, and are understood and studied in terms of the collective phenomenon [38,51–58]. They do not lend themselves to analysis from a game-theoretic point of view.

In this article, I present *rational swarms*, a game-theoretic model of *cooperative robot swarms*, as infinite-horizon fully cooperative Markov games (also known as Markov team games), with significant restrictions on the knowledge of the agents. Agents in such games are awarded the collective utility resulting from the joint actions, and thus have an individual incentive to maximize the collective utility. Given their limitations, swarm agents can only receive local rewards, partial proxies of the collective utility. As a result, when they seek to maximize their own local rewards, they can cause the collective utility to decrease, a phenomenon generally referred to as *the price of anarchy*, made famous in the game of Prisoner's Dilemma.

Focusing on robot swarms, I begin by showing how under modest assumptions, collective utility can be approximated by *aggregating the individual work times of the robots in the swarm*: the times in which they are engaged in their individual tasks, rather than the overhead of coordination. As embodied agents (animals, robots) can be assumed to measure time, this allows swarm members to all use a common, always-accessible, measure of utility. It only requires them to differentiate time and resources spent on their task, from those spent on coordination.

Naively, if each agent maximizes its operational time, and minimizes interference due to miscoordination (e.g. collisions), then the aggregated times increase, and so does the abstract swarm utility [51,56]. However, as the agents know nothing about the effects of their actions on others, attempts to increase their own working time may actually hurt others' efforts.

To address this, I next show that it is possible to reduce collective rewards to individual rewards that are *aligned*, by approximating the individual *difference rewards* [36,59,60]. These turn the Markov game into a potential game [61], where maximizing the individual rewards maximizes the collective rewards [35,62].

Using the rational swarms difference reward with distributed multi-agent reinforcement learning [63–67], the robots learn to coordinate and resolve collisions efficiently, maximizing the swarm's collective utility. The technique has been demonstrated in extensive experiments, in various tasks and settings (figure 1), over the last 15 years.

This article presents a synthesis of the rational swarms model. It begins by showing how the cooperative swarm can be represented as Markov team games, and how the inherent locality of perception of swarm agents can prevent them from individually acting rationally (§2). Then, the rational swarms model is introduced in §3. The §4 discusses lessons learned and insights gained from 15 years of research and development of the model. Finally, §5 discusses the way forward for rational swarms research, and concludes.

2. Cooperative swarms as Markov team games

Preliminaries. We consider swarms composed of set N of embodied agents, that perceive and act locally, i.e. within some bounded range of their position. The swarm is active for a duration T, which is unknown to us. During this time, each agent $i \in N$ takes actions a_t^i (action a^i taken at time $t < T$), drawn from a set of possible actions A^i by the *strategy* π^i, which determines the action taken by agent i at any given time. Given the local perception of the agents, each individual agent $i \in N$ knows only the action it has taken a_t^i; the effects of the action are localized, affecting only $n \ll |N|$ agents (their *social neighbourhood*).

We follow up on previous work [68,69] in distinguishing actions taken by each individual carrying out its task independently of others, and actions taken by agents to coordinate with each other. The need for coordination in swarms arises in two distinct situations. First,

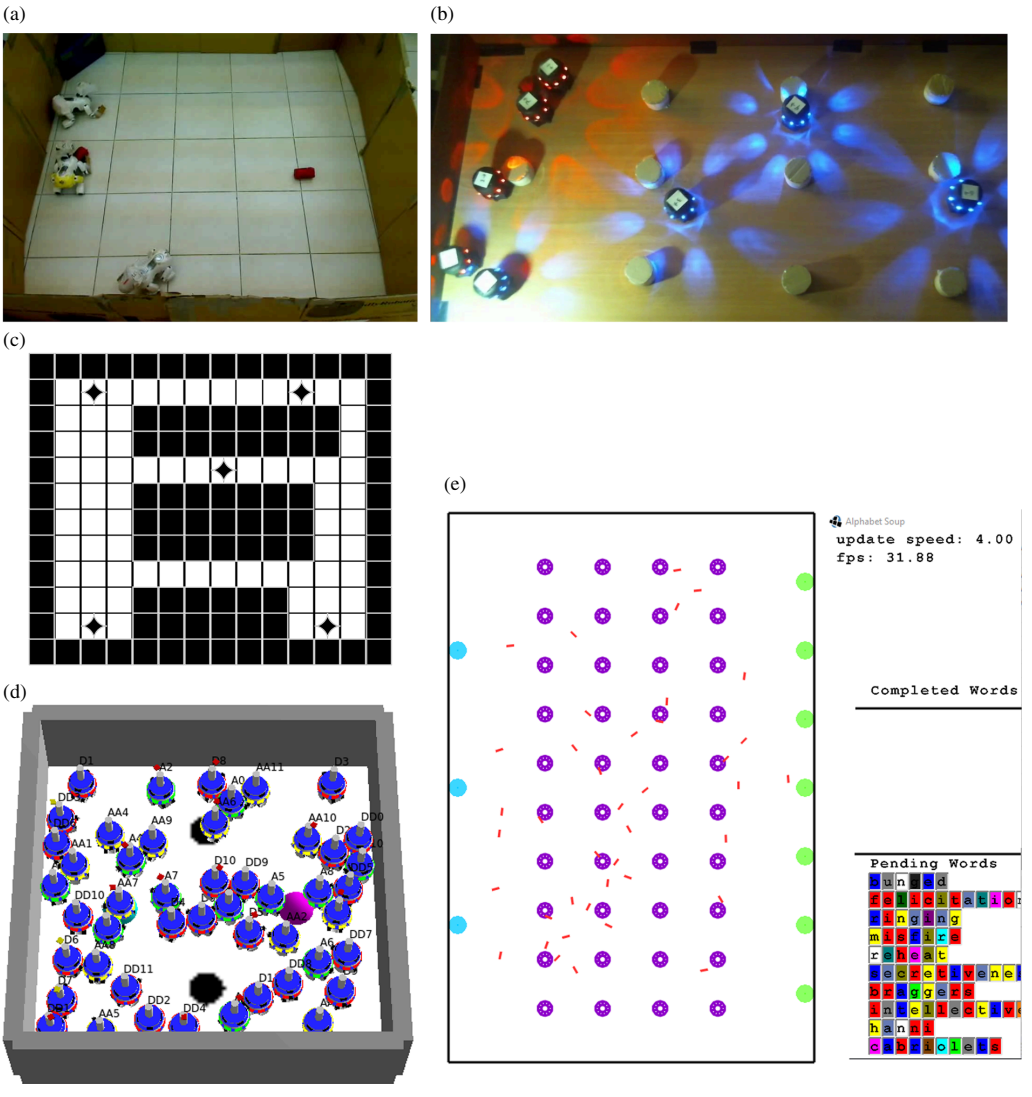

Figure 1. Agents and robots in environments used in experiments with the *rational swarms* model described herein; the agents used reinforcement learning. (a, b) Real robots; (c) agents moving on a discrete grid; (d, e) simulated robots moving in continuous spaces. In all, agents sensed and acted locally, with no communications with their neighbours. Learning was only applied in selecting actions for collision handling. (a) AIBO robots collecting cans [68]. (b) Krembot robots searching for objects [69]. (c) Foraging on a grid [70]. (d) ARGoS [71]-simulated Krembot robot swarms in a search task [72,73]. (e) Alphabet Soup [74] simulation of robots in a warehouse [69].

agents may inadvertently interfere with each other, i.e. they are *in conflict* [23,51,52,56,68]. This happens for example when their motion trajectories intersect and they (are about to) collide. A second form of coordination may be needed materially for the task, when robots cannot perform a component of the task independently of each other, i.e. more than a single robot is needed to carry out an atomic component of a task (e.g. lifting a long table from both ends requires two robots).

We focus on swarm tasks where coordination is used to resolve conflicts. In these, the individual agent repeatedly switches between two abstract modes of operation, back and forth: an individual *task*-execution mode (called *Program* mode) where each individual agent is independently carrying out its task within the swarm, and an *Avoidance* mode, where it takes coordination actions to resolve conflicts with other agents, e.g. avoiding collisions. The

total operating duration of a swarm member is segmented into *stages*, by conflicts requiring coordination to resolve. Each stage starts with a conflict. Then, coordination actions are taken in avoidance mode to resolve the conflict. This allows the agent to go back to work in program mode, until a new conflict occurs (see figure 2 for illustration).

Multi-robot foraging [49,50,56,75] is a good example. Here, the robots leave the nest in search for food items, which they individually bring back to the nest. Their individual program mode would consist of actions that do not require coordination with others: searching for items, picking them up, finding the way home, dropping collected items. In the course of these activities, they inevitably come near each other (e.g. when some agents are heading with items into the nest area, while others are heading away). Impending collisions trigger a conflict state, which switches the robots to coordination mode, where they take actions to resolve or prevent the collision. Once done, they switch back to program mode.

For now, let us assume that all agents are involved in every collision, i.e. the swarm works from one stage to the next. While all swarm members enter a collision together, they do not necessarily resolve it at the same time: different joint actions may result in some agents leaving the avoidance mode sooner than others. For example, this can occur when three agents collide, and two remain entangled for a while longer, while the third turns away immediately (see figure 2, in the second stage: Robot 1 resolved the conflict earlier than Robots 2 and 3).

We distinguish and focus on cooperative swarms, where a *collective objective* is defined, that agents collectively try to achieve. For example, in collective foraging, the goal of the swarm is to maximize the total number of items collected from the work area. In common collective-motion models, the goal is to establish a common movement heading for the agents. In collective search, the goal is to maximize the number of items discovered. Agents in cooperative swarms individually select actions that maximize achievement of the swarm goals.

This view of the robot's timeline allows us to position our work with respect to others. Many methods focus on improving the efficiency of the program-mode actions (see e.g. [47,76,77] in foraging). Others focus on improving the entire stage (avoidance and program), by restricting the behaviour of the robot during both program and avoidance, such that collisions and conflicts are minimized [51], e.g. by pre-allocating robots to different areas [78] or tasks [79]. We go beyond the single-agent stage, by examining the stages of all agents, collectively.

Team (Fully Cooperative) Markov Games. Each stage (collision resolution and subsequent task activity; figure 2, top) can be viewed as a *normal-form team game*, where the agents all receive an identical payoff for their joint actions. During the avoidance-mode interval of the conflict, agents (*players*) individually select conflict-resolution actions, synthesizing joint actions. In team games, *all agents receive this identical payoff for their joint actions*. Different joint actions can yield different joint rewards, but whatever the joint reward, every agent individually receives it.

The sequence of stages—*subgames*—forms a view of the swarm's agents as engaged in a *stochastic team game* (also called a Markov team game) [80–83]. A Markov team game is defined by a tuple $G := \langle N, S, A, D, R \rangle$, where N is the set of agents, S is the set of states (each representing a subgame type, associated with a conflict that could begin a stage), A the set of potential conflict-resolution *joint* actions, D a state-transition probability function and R a function determining the swarm reward from taking an action at a state S. The components of G are explained in detail below. The swarm seeks to maximize the accumulated reward R over many stages. As their number is unknown, this is formally an *infinite-horizon* game [81,82].

States S. Conflicts are events that interrupt an individual agent's program mode, and require coordination to resolve. S is the set of recognizable types of conflicts. Collisions are a good example: we can define all collisions to be the same (i.e. a single type, as done in [68,69]), or we can distinguish different types of collisions, especially when the distinction matters to the actions taken. For example, a collision front-to-front may require a different response to a collision front-to-rear, and so the relative position of agents may distinguish conflict states

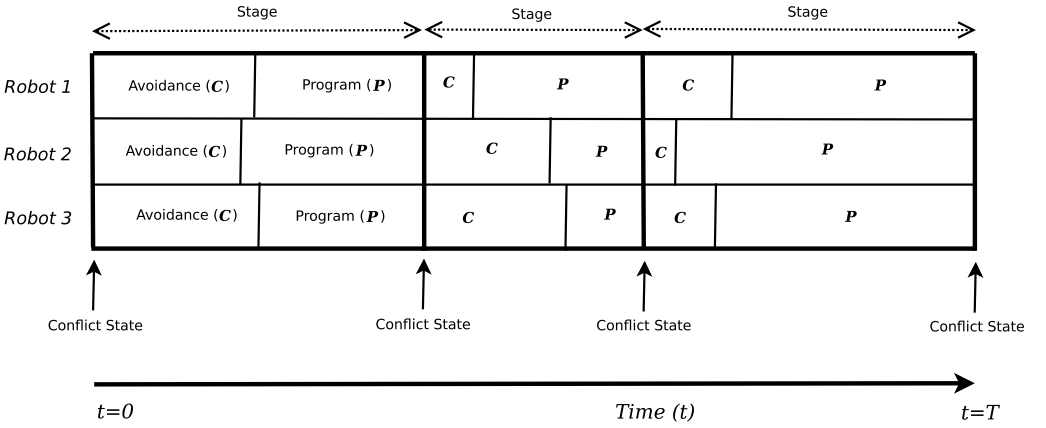

Figure 2. A visual illustration of a swarm member's activity timeline. Duration of *Avoidance* and *Program* modes are implicitly shown by the length of the respective boxes along the horizontal axis. Each stage begins when a conflict state is detected, sending all agents to switch to avoidance mode. Switching back to program mode is done when the conflict is deemed resolved, by each individual separately. Conflict states are assumed to be recognized by all agents.

[72]. Similarly, the position of agents in the work area [70], or the local density [84], can also distinguish states.

The joint actions set A**.** The set of all possible joint actions is defined by the sets of potential individual actions $A = A^1 \times A^2 \times \cdots \times A^{|N|}$. Each agent i selects its own action $a^i \in A^i$. Then, a specific joint action $a \in A$ is a tuple $(a^1, a^2, \ldots, a^{|N|})$.

State-Transition Probability Function D**.** A joint action $a_t \in A$ can be applied in any state $s \in S$. It transitions the swarm from the state s to the next state $s' \in S$. This transition corresponds to a single stage (figure 2): the agents begin in a conflict of type s, individually apply π^i to select avoidance action $a^i \in A^i$. The individual actions of all agents compose a joint action a, which resolves the conflict and allows each agent to continue in program mode (drawing actions from $\pi^i_{\mathcal{P}}$). The next state s' is determined stochastically, by the transition probability function $D : S \times A \times S \to [0,1]$, measuring the probability of transitioning from state s to state s' after applying the joint action a.

The Collective Reward R**.** In the common definition of stochastic games, there are $|N|$**individual** reward functions $\{R^i\}, \forall i \in N$, where function R^i returns the player's individual payoff, having been in state s, where the joint action a was taken, and reaching state s', determined by the state-transition probability D.

The definition we use here, of team games, differs from the common definition. Instead, we use a single collective reward, $R : S \times A \times S \to \mathbb{R}$, describing *the collective payoff* from performing the joint action a at state s (resolving the conflict), and reaching a new state s'. All agents share the payoff $R(s, a, s')$, as this is a fully cooperative (team) game. Note that R is *stationary*; it does not vary with time, but only with state and action.

Maximizing the Swarm Collective Reward Accumulated Over Time. Given a conflict state s, every agent $i \in N$ selects its conflict resolution action a^i using its strategy π^i, i.e. $a^i := \pi^i(s)$. The individual actions of all agents compose the joint action a that is then applied to the conflict state s to resolve it. This allows the agents to go back to their independent individual programs, until the next conflict state occurs (determined stochastically by D), and a new stage begins. The *joint strategy* of the swarm is defined by the agents' $\pi := (\pi^1, \ldots, \pi^{|N|})$.

We use s_t to denote the state of conflict that had begun at time t. At that time, a joint action a_t is taken, whose durative effects last until a new conflict arises at time $s_{t'}$. The associated stage is denoted $(s_t, \pi(s_t), s_{t'}) = (s_t, a_t, s_{t'})$. The state $s_{t'}$ is a stochastic result of applying the action a_t; the probability of reaching $s_{t'}$ is given by $D(s_t, \pi(s_t), s_{t'})$, and the reward for this stage is given by $R(s_t, \pi(s_t), s_{t'})$.

A *possible play* $\eta(\pi)$ is a sequence $\langle (s_{t_0}, a_{t_0}), (s_{t_1}, a_{t_1}), \ldots, (s_{t_k}, a_{t_k}), \ldots \rangle$, generated by repeatedly applying a strategy π to states, beginning with state s^0 (i.e. $t_0 = 0$), such that for any $k \in \mathbb{N}$, t_k is a time-stamp, and $D(s_{t_{k-1}}, a_{t_{k-1}}, s_{t_k}) > 0$. The accumulating reward over K stages of a *specific* play η_0 is $\mathcal{R}_K(\eta_0) := \sum_{k=1}^{K} R(s_{t_{k-1}}, a_{t_{k-1}}, s_{t_k})$, where t_k are specified by η_0.

As the new state s_{t_k} resulting from applying $a_{t_{k-1}}$ is determined stochastically by D, a better characterization of the rewards of π is by the *expected* collective reward over K stages, i.e. all of the possible plays π may induce, with K stages:

$$\mathcal{R}_K(\pi) := \mathbb{E}[\mathcal{R}_K(\eta(\pi))] \tag{2.1}$$

$$= \sum_{k=1}^{K} \sum_{s_{t_k} \in S} [R(s_{t_{k-1}}, \pi(s_{t_{k-1}}), s_{t_k}) \cdot D(s_{t_{k-1}}, \pi(s_{t_{k-1}}), s_{t_k})]\,. \tag{2.2}$$

As a notational shortcut, we used $R_k(\pi)$ for $R(s_{t_{k-1}}, a_{t_{k-1}}, s_{t_k})$, and analogously, $D_k(\pi)$ for $D(s_{t_{k-1}}, \pi(s_{t_{k-1}}), s_{t_k})$. Therefore,

$$= \sum_{k=1}^{K} \sum_{s_{t_k} \in S} [R_k(\pi) \cdot D_k(\pi)]\,. \tag{2.3}$$

We seek a strategy π^* that maximizes the expected collective reward generated, in comparison with alternative strategies. However, as K tends towards infinity, as long as R is positive, $\mathcal{R}_K(\pi)$ will grow unbounded for *any* strategy π. In other words, when K tends to infinity, strategies might not be differentiated by their collective rewards.

We therefore use a standard alternative objective, where the swarm seeks to maximize the *average collective rewards* (limit of means) of $\mathcal{R}_K$:

$$\mathcal{R}(\pi) := \lim_{K \to \infty} \frac{1}{K} \mathcal{R}_K(\pi) = \lim_{K \to \infty} \frac{1}{K} \sum_{k=1}^{K} \sum_{s_{t_k} \in S} [R_k(\pi) \cdot D_k(\pi)]\,. \tag{2.4}$$

A strategy π^* is then one that maximizes $\mathcal{R}(\pi)$:

$$\pi^* := \arg\max_{\pi} \mathcal{R}(\pi)\,.$$

If all the agents know how $\mathcal{R}$ changes as a result of their individual actions (i.e. they all receive $\mathcal{R}$), it is straightforward to show (i) that their individual self-interested selections will lead to a Nash equilibrium (i.e. the strategy guarantees stability), and (ii) that the Nash equilibrium also maximizes $\mathcal{R}$ (i.e. it guarantees collective reward optimality). Allowing the agents to be rational and self-interested in maximizing their rewards will necessarily maximize the collective rewards, as they are one and the same.

The rewards of the collective are said to be *aligned* with those of the individual [36,60]. In this case, by utilizing reinforcement learning, the agents may evaluate different strategies π by the collective rewards $\mathcal{R}(\pi)$, until converging to the optimal strategy π^*.

Locality of Perception Leads to Loss of Collective Information. Unfortunately, the condition that all agents know $\mathcal{R}$ does not hold in swarms: the agents cannot measure the collective reward $\mathcal{R}$, as they can only perceive in their immediate local environment. Having no ability to measure $\mathcal{R}$, they naturally also have no way to learn or predict how any joint action, affects it. Individually, they may follow their own self-interested, rational decision-making preferring actions that appear best *from their perspective*, but we lose all guarantees that this will maximize the collective reward.

It may therefore appear that framing cooperative swarms as fully cooperative games yields little insight as to how individual agents should select their actions to maximize the collective reward. As I show next, this is not the case.

3. Rational swarms: overcoming locality of perception

The key difficulty with the view of cooperative swarms as Markov team games lies in the fact that swarm agents cannot be assumed to measure the collective reward $\mathcal{R}$, as they can only perceive in their immediate local environment. Thus the first step we take (§3(a)) is to transform the arbitrary collective reward measure into a proxy measure that is individually accessible to all agents. In §3(b) we then show how agents can—in principle—maximize their individually accessible reward, such that it maximizes the collective reward $\mathcal{R}$.

(a) Local measurement of swarm utility measures

The reward of a cooperative swarm depends on its task, and so is measured in terms that, in general, agents cannot measure directly. While we may model the swarm as seeking to maximize the *average* swarm utility over an infinite horizon (equation 2.4), no agent can realistically measure it. A foraging agent away from the nest cannot keep track of the number of items collected, nor can they assess whether resolving a collision using a specific action is better (improves the $\mathcal{R}$) than another. We therefore seek to transform the utility measure into a form that is accessible to every agent, regardless of where it is.

We focus on *time* as a proxy to the utility gained by swarm members. In particular, we consider the relation between time spent by the swarm members on the task, and the utility resulting from it; we expect maximization of the former to be equivalent to maximization of the latter. As the measurement of time can be carried out individually, without even knowing what the task is, the swarm agents can focus on increasing the time spent on the task, rather than maximizing some abstract notion of utility (this is not trivial to do, see next section).

We remind the reader that we view the swarm's operation as a series of stages, wherein each agent has two modes of operation: an *avoidance mode*, triggered by collisions or other interference requiring coordination (for which they choose a joint action), and a resulting *program mode* where they carry out their individual tasks, undisturbed.

Here, we examine the agents' individual intervals of avoidance and program execution, whose duration depends on the individual action taken by each agent (figure 2). The duration of the avoidance interval for agent $i \in N$ is denoted $\mathcal{C}^i_k(\pi^i) = \mathcal{C}^i(s_{t_{k-1}}, \pi^i(s_{t_{k-1}}), s_{t_k})$, and the duration of the subsequent program interval is denoted $\mathcal{P}^i_k(\pi^i) = \mathcal{P}^i(s_{t_{k-1}}, \pi^i(s_{t_{k-1}}), s_{t_k})$. Their sum is the total duration of the stage $\left[\mathcal{C}^i_k(\pi^i) + \mathcal{P}^i_k(\pi^i)\right] = t_k - t_{k-1}$. While $\mathcal{C}^i$ and $\mathcal{P}^i$ generally vary between agents, their total is common to all agents (as we assume all agents are involved in every conflict). For any stage k, the following holds: $\forall i, j \in N\left[\mathcal{C}^i_k(\pi^i) + \mathcal{P}^i_k(\pi^i)\right] = \left[\mathcal{C}^j_k(\pi^j) + \mathcal{P}^j_k(\pi^j)\right]$.

It is reasonable to assume that the collective reward grows as agents spend more time in *program mode*, and less in *avoidance* mode, i.e. that it monotonically grows with the collective time spent in program mode. For simplicity, we assume proportionality; there exist constants $\beta^i, \alpha^i \in \mathbb{R}$ where $\beta^i > 0$, $\alpha^i \geq 0$, such that for any stage $(s_{t_{k-1}}, \pi(s_{t_{k-1}}), s_{t_k})$,

$$R_k(\pi) = \sum_{i \in N}\left[\beta^i \mathcal{P}^i_k(\pi^i) - \alpha^i \mathcal{C}^i_k(\pi^i)\right]. \tag{3.1}$$

From this we derive the expected time-based reward $\Gamma_K(\pi)$ of a policy π over K stages from $\mathcal{R}_K(\pi)$:

$$\mathcal{R}_K(\pi) = \sum_{k=1}^{K} \sum_{s_{t_k} \in S} [R_k(\pi) \cdot D_k(\pi)] \quad \text{from equation (2.3)} \tag{3.2}$$

$$= \sum_{k=1}^{K} \sum_{s_{t_k} \in S} \sum_{i \in N} \left[\left[\beta^i \mathcal{P}^i_k(\pi^i) - \alpha^i \mathcal{C}^i_k(\pi^i)\right] \cdot D_k(\pi)\right] \quad \text{from equation (3.1)} \tag{3.3}$$

and since D is defined for the joint action $\pi(s_{t_{k-1}})$, it is the same for all agents $i \in N$,

$$= \sum_{k=1}^{K} \sum_{s_{t_k} \in S} D_k(\pi) \sum_{i \in N} \left[\beta^i \mathcal{P}^i_k(\pi^i) - \alpha^i \mathcal{C}^i_k(\pi^i)\right]. \tag{3.4}$$

$$=: \Gamma_K(\pi). \tag{3.5}$$

It is tempting to then redefine $\mathcal{R}(\pi)$ as the limit of $\frac{1}{K}\Gamma_K(\pi)$, as $K \to \infty$. However, this is incorrect, as it averages over the number of conflicts, giving equal weight to short and long intervals, regardless of their accumulating contributions to the reward. Instead, we should use the average over the accumulating time spent by the swarm agents working through the K stages. The accumulated time T is $t_k - t_o$ (absolute time of the current conflict, minus the absolute time of the initial conflict that marks the beginning of the swarm task). Simply $T = t_k$, if we measure the beginning of the swarm task as $t_0 = 0$. However, the conflict state s_{t_k} and its time t_k are determined stochastically by D. Therefore, the total duration T varies, depending on π and D.

$$\Gamma(\pi) := \lim_{K \to \infty} \sum_{k=1}^{K} \sum_{s_{t_k} \in S} D_k(\pi) \sum_{i \in N} \frac{\beta^i \mathcal{P}^i_k(\pi^i) - \alpha^i \mathcal{C}^i_k(\pi^i)}{t_k - t_0}. \quad \text{Plugging in equation (3.5)} \tag{3.6}$$

$$= \lim_{K \to \infty} \sum_{k=1}^{K} \sum_{s_{t_k} \in S} \frac{D_k(\pi)}{T} \sum_{i \in N} \left[\beta^i \mathcal{P}^i_k(\pi^i) - \alpha^i \mathcal{C}^i_k(\pi^i)\right]. \quad T = t_k - t_0 \text{ is same for any } i, j \in N \tag{3.7}$$

We now have a measure of the collective reward, which is accessible in principle to any agent $i \in N$, since it relies solely on *time*, which any of the agents can measure. Noting that the term $\lim_{K \to \infty} \sum_{k=1}^{K} \sum_{s_{t_k} \in S} \frac{D(s_{t_{k-1}}, \pi(s_{t_{k-1}}), s_{t_k})}{T}$ is common to all agents, we denote by $\Gamma^i_k(\pi)$ the contribution of agent $i \in N$ to $\Gamma(\pi)$ in stage $k-1$:

$$\Gamma^i_k(\pi^i) := \left[\beta^i \mathcal{P}^i_k(\pi^i) - \alpha^i \mathcal{C}^i_k(\pi^i)\right], \tag{3.8}$$

i.e. $\Gamma(\pi) = \lim_{K \to \infty} \sum_{k=1}^{K} \sum_{s_{t_k} \in S} \frac{D_k(\pi)}{T} \sum_{i \in N} \Gamma^i_k(\pi)$.

$\Gamma^i_k(\pi)$ (equation 3.8) can be independently computed by agents, and is therefore a step towards compensating for missing information: with every collision, every agent, independently of others, can record the time it spent resolving the previous conflict $\mathcal{C}_k$ and the time spent working in program mode since the conflict was resolved $\mathcal{P}^i_k$. The terms α^i, β^i are also knowable to the agent i. In addition, it also knows T, which is the total time the swarm has been operating, and is therefore the same for all agents.

However, some information remains beyond the immediate perception of the agent, preventing it from computing the collective reward (equation 3.7) from the individual $\Gamma^i_k(\pi)$ (equation 3.8). In particular, the local perception of each swarm agent i prevents it from knowing $\mathcal{P}^j, \mathcal{C}^j$ for any agent $j \neq i \in N$. It cannot perceive N, and so does not know how many other agents there are. As a result, agents seem to have little choice but to attempt to rationally maximize their own perceived individual reward $\Gamma_{\pi|i}$, based on their own local knowledge, substituting π^i for the strategy π, their own locally perceived state s^i for s, etc.

As the reward to one agent may be at the expense of another, each agent's individual reward is no longer *aligned* with the actual collective reward: maximization of one does not lead to maximizing the other. To intuitively see why this happens, imagine some agents are attempting to leave the nest after dropping collected items off, while others are attempting to enter the nest. Those attempting to enter should ideally back off, allowing those inside the nest to go out. However, backing off adds to the duration of the avoidance mode, and reduces the duration of the program mode. Thus those agents are motivated to push forward. This hinders the swarm from collecting items. The next section addresses this challenge.

(b) Computing a swarm-aligned reward

We can align the individual rewards with those of the collective by finding a *utility potential function* [61]. A utility potential function assigns a scalar (*potential*) to the joint actions of agents,

such that any change to the local reward of player $i \in N$, stemming from its unilateral preference of an individual action a^i over a different individual action a^i, is reflected by a matching change to the potential: if an agent seeks greater individual reward, it will necessarily cause an increase in the potential. Games that possess such a function are called *potential games* [61], and have the property that rational strategic individual preferences for greater local rewards necessarily result in a Nash equilibrium that also maximizes the rewards of others.

Such a potential function was proposed by Wolpert & Tumer [36]. Initially called *Wonderful Life Utility*, it was later renamed and extended as *difference rewards* [59,60], and shown to be related to the economic concept of *marginal contribution* [62,85].

The difference reward $\Delta(i)$ of agent $i \in N$ captures the contribution of agent i to the collective reward of the swarm. It is defined as $\Delta^i(\pi) := \delta^{+i}(\pi) - \delta^{-i}(\pi)$, where $\delta^{+i}(\pi)$ denotes the collective reward achieved in the presence (involvement) of the agent, and $\delta^{-i}(\pi)$ denotes the hypothetical counterfactual reward achieved in the agent's absence. We discuss this in detail below, and point the reader elsewhere for analytical discussion of the difference reward as a potential [35,62].

(i) The collective reward when agent i is present, $\delta_k^{+i}(\pi)$.

When agent $i \in N$ is involved (and utilizing strategy π), it is straightforward to take $\delta_k^{+i}(\pi)$ to be the collective time-based reward $\Gamma_k(\pi)$ at any given stage $k-1 = (s_{t_{k-1}}, \pi(s_{t_{k-1}}), s_{t_k})$. We rewrite it using $\Gamma^i(\pi)$ to emphasize the role of agent i:

$$\delta_k^{+i}(\pi) := \Gamma_k(\pi) = \sum_{j \in N} \Gamma_k^j(\pi) \tag{3.9}$$

$$= \Gamma_k^i(\pi) + \sum_{j \in N \setminus \{i\}} \Gamma_k^j(\pi), \tag{3.10}$$

where $N \setminus \{i\}$ is the set of agents N, without the agent i. Substituting by the respective definitions, this yields:

$$= \left[\beta^i \mathcal{P}_k^i(\pi^i) - \alpha^i \mathcal{C}_k^i(\pi^i) + \sum_{j \in N \setminus \{i\}} \left[\beta^j \mathcal{P}_k^j(\pi^j) - \alpha^j \mathcal{C}_k^j(\pi^j)\right]\right]. \tag{3.11}$$

(ii) The collective reward when agent i is not present, $\delta^{-i}(\pi)$.

$\delta^{-i}(\pi)$ is a counterfactual: it asks what the collective reward would have been, had the agent *not* participated or contributed. Computing the counterfactual takes into account detailed information about the task at hand [60,86], and sometimes can be be computed directly [60]. However, the locality of perception, and thus the limited information available to each agent, makes analytical computation of the counterfactual difficulty. Our one aid in this discussion is that all possibilities are reflective, rather than predictive: any assessment of $\delta_k^{-i}(\pi)$ is made at time t_k, looking back at the duration from the conflict state at time t_{k-1}. Thus, the action taken and the resulting duration of the stage τ_k are known at the point of assessment.

We may naively believe $\delta^{-i}(\pi)$ is equal to $\sum_{j \in N \setminus \{i\}} \Gamma^j(\pi) + 0$, where the 0 component marks the lack of contribution, of any kind, by agent i (i.e. it is the result of some null action, and $\Gamma^i(\pi) = 0$). In that case,

$$\Delta_k^i(\pi) = \underbrace{\Gamma_k^i(\pi) + \sum_{j \in N \setminus \{i\}} \Gamma_k^j(\pi)}_{\delta_k^{+i}(\pi),\ \textit{equation}\ 3.10} - \underbrace{\sum_{j \in N \setminus \{i\}} \Gamma_k^j(\pi) - 0}_{\delta_k^{-i}(\pi)} = \Gamma_k^i(\pi).$$

However, this is incorrect. The agents we discuss are *embodied*, necessarily having physical mass and body geometry, and existing over time. When they interact with others, they necessarily affect them. An agent that is not affecting others in a collision is one that simply goes through them, undisturbed and undisturbing. Thus while it is possible for an agent to have

zero individual contribution over the duration of a stage ($\Gamma^i(\pi) = \beta^i \mathcal{P}^i_k - \alpha^i \mathcal{C}^i_k = 0$), its physical embodied existence affects others during the interval. It is perceived by others, and it can block or facilitate their movement. Furthermore, even if we consider the agent's hypothetical physical removal from the system altogether, its *absence* would still affect others.

Lacking an analytical model of interactions that is parameterized by information about each individual agent (e.g. its position and velocity, its communicative acts when interacting, etc.), we are left with essentially philosophical approaches for building such a model. We distinguish two components: (i) a model, denoted $\Gamma^{-j}_k(\pi)$, of how other agents $j \neq i$ are affected when agent i is hypothetically absent from the conflict; and (ii) a model, denoted $\Gamma^{-i}_k(\pi)$, for how agent i is affected in this case.

How others are hypothetically affected. We argued above that in principle, it is unreasonable to expect all other $N-1$ agents participating in a collision to be generally unaffected by the absence of agent i. Instead, I consider the direction of the effect.

In principle, the others can do worse without i: their respective avoidance periods grow longer, and necessarily their productivity decreases, i.e. $\Gamma^{-j}_k(\pi) < \Gamma^j_k(\pi)$. In the extreme case, they hypothetically remain in avoidance mode for the duration of the stage, and so,

$$\Gamma^{-j}_k(\pi) := -\alpha^j \tau_k = -\alpha^j \left[\mathcal{P}^j_k(\pi^j) + \mathcal{C}^j_k(\pi^j) \right]. \tag{3.12}$$

Adapting this interpretation requires a narcissist view of agent i's importance to the interaction: without its involvement as a benefactor, others are unable to proceed.

The opposite view is that without agent i, the conflict did not occur (i.e. agent i is a disruptor, the only cause for a conflict). In that case, the other agents would have spent the duration of their now-hypothetical avoidance interval working their own task (program mode), and the avoidance mode duration for all agents $j \in N \setminus \{i\}$ would have been as rewarding as the program mode. Instead of reducing the reward by $-\alpha^j \mathcal{C}^j_k$, we *increase* the reward by $\beta^j \mathcal{C}^j_k$, as the entire duration of stage k is considered to have been spent by any agent j carrying out its program:

$$\Gamma^{-j}_k(\pi) := \beta^j \tau_k = \beta^j \left[\mathcal{P}^j_k(\pi^j) + \mathcal{C}^j_k(\pi^j) \right]. \tag{3.13}$$

The two models mark extreme points that bound the true values from above and below. If agent i was responsible for the conflict, its removal from the system increases everyone's productivity up to $\beta^j \left[\mathcal{P}^j_k(\pi^j) + \mathcal{C}^j_k(\pi^j) \right]$ (equation 3.13). If instead agent i was the benefactor, whose presence allowed the conflict to be resolved, then everyone's productivity would be reduced to $-\alpha^j \left[P^j_k(\pi^j) + C^j_k(\pi^j) \right]$ (equation 3.12) in its absence.

How agent i is hypothetically affected. Agent i knows its own actual Γ^i_k. There are several ways of approaching the question of how this value changes when i is absent from the system. Clearly, one possibility is indeed to assume that it in its absence, its own counterfactual contribution is 0 (as discussed above):

$$\Gamma^{-i}_k := 0 \leq \Gamma^i_k. \tag{3.14}$$

I see two other interpretations, that bound the unknown real value from above and below. One possibility is that the agent i had spent the entire stage in avoidance mode, i.e. it is absent from the stage in the sense that it is not producing anything:

$$\Gamma^{-i}_k := -\alpha^i \tau_k = -\alpha^i \left[\mathcal{P}^i_k(\pi^i) + \mathcal{C}^i_k(\pi^i) \right]. \tag{3.15}$$

The other interpretation is that in its absence from the conflict, it never went through the avoidance interval, and so,

$$\Gamma^{-i}_k := \beta^i \tau_k = \beta^i \left[\mathcal{P}^i_k(\pi^i) + \mathcal{C}^i_k(\pi^i) \right]. \tag{3.16}$$

Putting $\Gamma^{-j}_k(\pi)$ and Γ^{-i}_k together. We can now define the counterfactual using the above. While multiple possibilities exist, most recent work [72,73] empirically demonstrated that it is useful

to take the so-called *benefactor* view, whereby without agent i, the conflict did not occur (or at least, the avoidance duration decreased [69,87]) (equation 3.13), while the agent's hypothetical contribution in its absence is 0 (equation 3.14):

$$\delta_k^{-i}(\pi) := \Gamma_k^{-i}(\pi) + \sum_{j \in N \setminus \{i\}} \Gamma_k^{-j}(\pi) \tag{3.17}$$

$$= \underbrace{0}_{\text{equation (3.14)}} + \sum_{j \in N \setminus \{i\}} \underbrace{\left[\beta^j \mathcal{P}_k^j(\pi^j) + \beta^j \mathcal{C}_k^j(\pi^j)\right]}_{\text{equation (3.13)}}. \tag{3.18}$$

(iii) Putting it all together: an aligned individual reward, $\Delta_k^i(\pi)$

$\Delta_k^i(\pi)$ can now be defined as follows:

$$\Delta_k^i(\pi) := \delta_k^{+i}(\pi) - \delta_k^{-i}(\pi) \tag{3.19}$$

$$= \overbrace{\beta^i \mathcal{P}_k^i(\pi^i) - \alpha^i \mathcal{C}_k^i(\pi^i) + \sum_{j \in N \setminus \{i\}} \left[\beta^j \mathcal{P}_k^j(\pi^j) - \alpha^j \mathcal{C}_k^j(\pi^j)\right]}^{\delta_k^{+i}(\pi), \text{ equation 3.10}} - \overbrace{\sum_{j \in N \setminus \{i\}} [\beta^j \mathcal{P}_k^j(\pi^j) + \beta^j \mathcal{C}_k^j(\pi^j)]}^{\delta_k^{-i}(\pi), \text{ equation 3.18}} \tag{3.20}$$

$$= \beta^i \mathcal{P}_k^i(\pi^i) - \alpha^i \mathcal{C}_k^i(\pi^i) + \sum_{j \in N \setminus \{i\}} \left[\beta^j \mathcal{P}_k^j(\pi^j) - \alpha^j \mathcal{C}_k^j(\pi^j) - \beta^j \mathcal{P}_k^j(\pi^j) - \beta^j \mathcal{C}_k^j(\pi^j)\right] \tag{3.21}$$

$$= \beta^i \mathcal{P}_k^i(\pi^i) - \alpha^i \mathcal{C}_k^i(\pi^i) - \sum_{j \in N \setminus \{i\}} (\alpha^j + \beta^j) \mathcal{C}_k^j(\pi^j). \tag{3.22}$$

Note that when the agent i is alone (no neighbours, i.e. $N \setminus \{i\} = \emptyset$), $\Delta_k^i(\pi)$ is simply its own time-based reward $\Gamma_k^i(\pi) = \beta^i \mathcal{P}_k^i(\pi^i) - \alpha^i \mathcal{C}_k^i(\pi^i)$. The aligned difference reward elegantly simplifies to the natural individual reward in this case. There is no need for the agent to switch between reward functions depending on its social settings.

(c) Determining π^* by reinforcement learning

Through a completely distributed learning process, every agent $i \in N$ uses its own computed Δ_k^i to learn its own optimal strategy π^i, maximizing

$$\lim_{K \to \infty} \sum_{k=1}^{K} \sum_{s_{t_k} \in S} \frac{D_k(\pi)}{T} \Delta_k^i(\pi^i),$$

such that the composed *joint* policy is optimal π^*. While equation (3.22) removes some of the knowledge requirements from the individual agent (e.g. it does not need to know $\mathcal{P}^j$ and T), it still leaves some unknowns ($N, \alpha^j, \beta^j, \mathcal{P}^j, \mathcal{C}^j$). These will be approximated below. The transition probability function D is not known to the agent, and will be addressed through the learning process.

First, assuming agents are homogeneous, we may set $\alpha^i = \alpha^j, \beta^i = \beta^j$, and rewrite equation (3.22):

$$\Delta_k^i(\pi^i) \approx \beta^i \mathcal{P}_k^i(\pi^i) - \alpha^i \mathcal{C}_k^i(\pi^i) - \sum_{j \in N \setminus \{i\}} \underbrace{(\alpha^i + \beta^i)}_{\text{replacing } \alpha^j, \beta^j} \mathcal{C}_k^j(\pi^j). \tag{3.23}$$

Next, we can estimate the unknown duration $\mathcal{C}^j$ by the mean duration of $\mathcal{C}^i$ in previous stages, i.e. $\overline{\mathcal{C}_k^i}(\pi^i) := \frac{1}{K} \sum_{k=1}^{K} c_k^i(\pi^i)$. Both last assumptions follow a common approach in multi-agent reinforcement learning [88], where the learners assume others are similar. Using this last estimate, we can continue rewriting,

$$\approx \beta^i \mathcal{P}_k^i(\pi^i) - \alpha^i \mathcal{C}_k^i(\pi^i) - \sum_{j \in N \setminus \{i\}} (\alpha^i + \beta^i) \overline{\mathcal{C}_k^i}(\pi^i) \quad (\mathcal{C}_k^j(\pi^j) \approx \overline{\mathcal{C}_k^i}(\pi^i)) \tag{3.24}$$

and finally, as $N \setminus \{i\}$ is unknown, we use the locally perceived social neighbourhood, of size $n \ll |N|$ (n here includes the agent i) as a sufficient estimator [72,73,84]. The intuition for this is that in collisions, agents that are not sensed are not part of the collision. They are therefore not affected by it, nor do they affect it, and so their $\mathcal{C}$ duration is 0, having no impact on $\Delta_k^i(\pi)$ (their $\mathcal{P}$ duration is not needed for the computation in any case). We therefore finally rewrite

$$\approx \quad \beta^i \mathcal{P}_k^i(\pi^i) - \alpha^i \mathcal{C}_k^i(\pi^i) - (n-1)(\alpha^i + \beta^i)\overline{\mathcal{C}_k^i}(\pi^i) \quad \text{(local neighbours)}. \tag{3.25}$$

This last step touches on a key assumption in the model, that conflicts involve all agents in N. This assumption is generally violated in swarms, and most certainly when we examine collisions as the source of conflicts: there is no reason, and no possibility, that all robots collide together. The intuition provided above for the use of only the n robots locally perceivable by the robot argued on practical grounds that the $\mathcal{C}, \mathcal{P}$ can be ignored or nullified. However, this is a practical approximation. Theoretically, the assumption still needs to be addressed.

From a mathematical point of view, Douchan [84] shows that it is enough to consider agents that are not involved in a conflict to have selected a special individual action a_P, which has a $\mathcal{C}$ duration of 0, and therefore is maximally productive (has a program interval of length τ_k). However, this mathematical equivalence is not amenable to local approximation. In future work, the rational swarms model should be developed to address *sparse interactions*, by appropriate shaping of the reward [89] or using specialized algorithms [90].

Necessarily, each agent uses its own *estimated* $\Delta_k^i(\pi^i)$ (equation 3.25), using no information about others. Agents are therefore *independent learners* in multi-agent reinforcement learning. Despite the well-known inherent challenges of using reinforcement learning with independent learners [63–66,82,91–94], many algorithms have been developed to allow cooperative learning in multi-agent settings [67,81,92,95,96]. Surprisingly, the approach presented here has been repeatedly demonstrated to reach stable and highly optimized collective swarm rewards in many experimental settings, using the most basic of algorithms (see discussion below). Note that as the rational swarm model is stated—naturally—as an average (undiscounted) rewards optimization criterion, rather than discounted rewards, some algorithms may be better suited than others [97–99].

4. Lessons from experiments with artificial rational swarms

Over the last 15 years, the rational swarms approach was investigated in different experimental settings, involving embodied agents and robots (figure 1). The discussion above presents an up-to-date perspective, which allows us to view earlier work as special cases and approximations. In this section, I highlight insights from experiments carried out as part of the research. The next section (§5) will discuss future work, informed by this perspective.

Working in continuous spaces, with no communications between the agents, the rational swarms approach used with reinforcement learning (§3) has been consistently demonstrated to achieve superior swarm collective rewards when compared with manually tuned collision-avoidance methods (including methods allowing stochasticity in algorithm selection) [68–70,72,73,87,100,101]. A bird's eye view of the experiments reveals the importance of both components of the model: the use of time as a measure of utility (§3(a)), as well as the alignment of individual and collective rewards (§3(b)).

Earlier formulations and experiments [68,70] did not account for alignment of collective and individual rewards, instead nullifying the impact on others. They focused on minimizing the impact of the avoidance interval ($\mathcal{C}$), ignoring the program interval and its gains for the most part. Later experiments used various alignment approximations, and produced superior results [69,87], though still focusing on the avoidance interval. These also demonstrated that

the aligned time-based rewards worked on-par, or even better, than aligned rewards using the utility measure used in the task (e.g. number of items picked). We believe that this is due to the time-based difference rewards offering a more sensitive, finer-resolution measure; similar observations were also made by others [60,86].

More recently [72,73] the rational swarms model was clarified and simplified, as a result of mathematical derivation of the time-based collective reward from the perspective of the swarm. The model presented in §3 is a generalization and synthesis of these latter investigations. It builds on the empirical evidence to prefer the so-called *benefactor* view, whereby without agent i, the conflict did not occur (or at least, the avoidance duration decreased [69,87]), and also the use of equation (3.14) ($\Gamma_k^{-i} = 0$) for the agent's own counterfactual contribution. However, the other models are useful as analytical bounds [87].

The success of the rational swarms model in *continuous* settings is particularly noteworthy. Most of the work on Markov games and much of the literature on game-theory in general is carried out in discrete settings and discrete time (differential games being the notable exception [102,103]). It is therefore not straightforward to be able to apply a model based on Markov games, with its assumptions of discrete actions and states, to continuous settings.

I believe the observed success in continuous action spaces owes much to the use of *collision-avoidance algorithms* as *actions* used by the individual (i.e. composing the set of actions A^i). Each algorithm works as a *macro-action* or *option* [104]: once selected, it takes over control of the agent, generating motion actions at a fine resolution to resolve the collision. Deciding on a macro-action is carried out at a higher level of abstraction than the fine resolution of continuous motion. Robotics literature reports on many such algorithms [20,21,23,24,105–107]. In fact, this research direction began by observing that no one algorithm could be shown to be superior [23,52]. It seemed reasonable to let the swarm learn *when* and *which* algorithm to use.

More specifically, as agents learn independently of each other, each agent i learns its own strategy π^i that determines which algorithm to apply, and when. The strategy may differ from that of other agents. A repeating observation is that the swarm *rarely* converged to a homogeneous choice of collision-avoidance algorithms. Rather, in the great majority of experiments, the post-learning swarm was composed of groups, each distinguished by the fact that its member agents had learned to deal with collisions using a specific macro-action, different from those of other groups. In other words, the swarm has learned to diversify, in terms of collision-handling responses. Repeatedly, it had become *behaviourally heterogeneous* [53,108], even if the robots are physically homogeneous (contrast with investigations on heterogeneous robots [109–111]). Kaminka & Douchan [87] present detailed results analysing the behavioural heterogeneity of the swarm, albeit using a dated variant of the rational swarms model.

The use of a limited set of macro-actions also greatly simplifies the requirements from the learning mechanism used. Although there has been much progress in utilizing reinforcement learning in continuous spaces, and especially in robotics [112–116], the model results in simple formulations of the learning settings: so-called *stateless* settings (equivalent to classic multiarm bandits), where every conflict state is considered to be identical to any other (i.e. $|S| = 1$, in which case the Markov game model above is reduced to a *repeating games* model [68,69]); or settings including only a handful of states, distinguished by the local density of the social neighbourhood of the agents [84], the position of the agents [70], or the side of the collision [72]. In these simple settings, even the simplest classic algorithms for reinforcement learning work well in practice: *UCB1* [99] for multiarm bandits, and *Q-Learning* [117,118] for multi-state settings.

Some of the lessons from these 15 years of investigations are negative in nature, and point the way towards needed improvements in the theory and practice of rational swarms. Most importantly, while the use of the rewards is robust (i.e. the agents learn), it is necessary to carefully check that the proportionality assumptions underlying equation (3.1) (relating the rewards to the time spent in program and avoidance modes) are maintained. We had, on occasion, discovered that robots learned to move about doing essentially nothing when the

underlying program was ineffective (i.e. β^i was close to 0). A related challenge for future research is to allow associating actions $\pi^i(s)$ with their time constants α or β. This, for instance, would allow us to differentiate collision-avoidance algorithms that take the same amount of time (i.e. their resulting $\mathcal{C}^i(\pi^i)$) is the same), but they have other costs differentiating them (e.g. energy consumption).

Two important challenges are raised in using the rational swarms model with real robots. First, robots may face difficulties distinguishing collisions with walls, from collisions with other robots. Such a distinction is critical: collisions with robots signify conflict, while collisions with walls are handled as part of the robot program or avoidance modes, without changing the conflict state. Second, attempting to introduce states into the learning process (e.g. distinguishing actions taken in different spatial arrangements of neighbours around the robot) results in a combinatorial explosion in the number of states, a problem that is exacerbated when we consider continuous spaces. The use of state-decomposition or neural network methods may be useful.

5. The way forward: natural and artificial rational swarms

To the best of my knowledge, the rational swarms model offers the first game-theoretic view of cooperative robot swarms, bridging a perplexing gap between the rationality of the swarm as a collective, and the individual rationality of its members. Continued development of the model is informed through empirical work. The model is currently being investigated as a basis for analysis and development of simulated and real robot swarms carrying out various cooperative tasks. It has already been extended and demonstrated in swarm-competitive foraging [73], where two or more cooperative swarms compete with one another with respect to the number of items collected.

New tasks point the way towards needed generalization and sophistication, that promise to broaden the applicable scope of the approach, and deepen its impact. For example, *collective transport* is a foraging variant, where some items require more than one robot to carry: some items require multiple robots to carry. This simple variation changes the underlying assumption of the model: here, robots coordinate not only to avoid interfering with each other, but also to carry out their task. In terms of the rational swarms approach, this requires extending the approach to address coordination during program mode. Such extensions will also offer an opportunity to interact with orthogonal investigations by others, for improving individual foraging: better search patterns, localization capabilities, homing, etc. [47,49,119].

A primary open question, raised empirically to be addressed theoretically, deals with bounds on errors due to estimations used in the calculation of the difference reward, in particular of the counterfactual collective reward without the participation of the agent. A few steps towards such bounds are discussed in §3(b)-ii: upper and lower bounds for counterfactual values are presented, and may be used in future theoretical analysis.

However, other components in the rational swarms model are also approximated. For instance, the specific agent's mean experience (mean duration of avoidance) is used as an estimate for the contributions of others. This follows in the general spirit of assumptions, that the agents are homogeneous, anonymous and replaceable. However, it has no formal basis, and an empirical investigation of this estimator versus others (e.g. min or max) in [84,87] proved largely inconclusive. These issues remain open for future investigations.

Perhaps the most urgent open question is that of the convergence of the aggregated individual difference reward (equation 3.22), or its approximation in equation (3.25), to the time-based collective reward (equation 3.7). To my best knowledge, no theoretical proof exists that the difference reward, when used in a distributed fashion, convergence to the maximum-payoff social–welfare equilibrium in which the collective reward is maximized. Yet empirically, it seems to do just that across a wide variety of environments. The challenge is exacerbated because in independent learners settings (the case here), multi-agent distributed learning is

hampered by the non-stationary nature of the rewards, and the potential existence of multiple equilibria [66,81,92,96].

More generally, despite the promise of the rational swarm model in terms of providing an individually rational account of cooperative swarms, I note that its current state offers no predictions at a fine resolution (e.g. the time that it would take for the swarm to reach maximum mean collective reward), nor tools for analysis (e.g. would the swarm end up being heterogeneous in its individual strategies, and to what degree). These are questions that are open for many models of swarms, and restrict both our understanding of natural swarms and our use of synthetic swarms in important application areas. A detailed predictive model of foraging is presented by Lerman & Galstyan [56], demonstrating that such predictions are possible.

The use of robots as exploratory tools for investigating swarms is a direct connection to applications of swarm robotics. It is also an approach to investigating swarms synthetically. Robots and animals are both embodied, and share design constraints: energy use, geometrical and kinematic constraints restricting motions, noise in sensing and actuation, computational processing limitations and more. Understanding of animal swarms can and does inform our understanding of synthetic swarms. However, the reverse can also be true [120–122].

I wish to highlight two example swarm research areas, one of natural swarms, and one of synthetic swarms, to illustrate both the promise of the rational swarms model and the questions it leaves open:

1. *Are Swarm Animals Individually Rational?* The nymbot–locust hybrid swarm [123,124] mixes locust nymphs and robots in laboratory settings. The investigation seeks to answer fundamental questions about individual locust decision making, by using robots to conduct controlled experiments; controlling robot swarm motions, we measure the animal responses. It also attempts to construct algorithmic models of natural behaviour (see [4,6,122], for like-minded modelling attempts).

In principle, the rational swarms model is applicable here (to guide robot motions). However, a key open question touches on its suitability for *modelling the animals*, not *driving the robots*. Is there a detailed account of individual animal decision making, that is both individually rational, as well as collectively optimal? In other words, can we demonstrate that the locust (or other animals) perceive their neighbours through the transformation imposed by a difference reward? Do they use internal time measurements as part of their reward?

2. *Molecular Medicinal Robot Swarms.* There are many investigations of nanometer-scale molecular devices, some as simple as particles whose size and shape yield medically useful results, some as complex three-dimensional structures with local actuation [125]. These so-called *nanobots* offer an opportunity for clinical targeting of specific organs or biosites, which are not typical of more familiar types of medicine. While most investigations focus on the affinity between the device and its target location, there is growing evidence that by combining different nano-devices, i.e. creating a *heterogeneous drug swarm*, better results can be achieved [25–29,126].

These advanced therapies necessarily require consideration of the interactions of nano-devices within the body, by direct chemical reactions [26–28] or through synergistic interactions in the bio-chemical environment [25,29,126]. The extreme limitations of nano-devices inherently mean that they are inherently and myopically 'selfish', following chemical gradients and reactions, with no capacity for prediction or foresight. As medical applications require the devices to serve a collective medical goal, a method is needed to align the greedy, self-interested behaviour of the nano-devices (viewed here as nanobots) with the goals of the swarm. I envision a biochemical version of the rational swarm model—if developed for biochemical use—whereby it is used to plan both the construction and reactions of nanobots such that the affinity of different particles with respect to each other or target areas is guaranteed to achieve a clinical collective result. In other words, the nanobots would be designed such that they follow a collectively aligned gradient.

6. Conclusions

The research presented in this paper, and the future directions described above, should be understood in the context of a call to arms, for multidisciplinary research into *rational* swarms, to fill our world in its natural as well as in its urban, technological aspects. Natural and synthetic swarms are similarly constrained and can share analysis and modelling approaches, as have been demonstrated. The focus on rationality of swarms, from a multidisciplinary perspective, can generate significant impact on all disciplines involved in swarm research.

Data accessibility. This article has no additional data.

Declaration of AI use. I have not used AI-assisted technologies in creating this article.

Authors' contributions. G.A.K.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, writing—original draft, writing—review and editing.

Conflict of interest declaration. I declare I have no competing interests.

Funding. Partial support was provided by Israel Science Foundation grants #2306/18 and #1373/24

Acknowledgements. I thank colleagues and students who have, over the years, shared my passion for understanding the nature of swarms: Noa Agmon, Asaf Shiloni, Dan Erusalimchik, Sarit Kraus, Ido Bachelet, Inbal Wiesel-Kapah, Chana Weitman, Yinon Douchan, Amir Ayali, Luca Giuggioli, Eden R. Hartman, Karen Katz, Lee-or Alon, Erel Shtossel, Peleg Shefi and Aaron Reuven. I thank Michael Wooldridge and the editors for the opportunity to present this work. As always, I am indebted to K. Ushi.

References

1. Ariel G, Ayali A. 2015 Locust Collective Motion and Its Modeling. *PLoS Comput. Biol.* **11**, e1004522. (doi:10.1371/journal.pcbi.1004522)
2. Lopez U, Gautrais J, Couzin ID, Theraulaz G. 2012 From behavioural analyses to models of collective motion in fish schools. *Interface Focus* **2**, 693–707. (doi:10.1098/rsfs.2012.0033)
3. Aplin LM, Farine DR, Mann RP, Sheldon BC. 2014 Individual-level personality influences social foraging and collective behaviour in wild birds. *Proc. R. Soc. B* **281**, 20141016. (doi:10.1098/rspb.2014.1016)
4. Fridman N, Kaminka GA. 2010 Modeling pedestrian crowd behavior based on a cognitive model of social comparison theory. *Comput. Math. Organ. Theory* **16**, 348–372. (doi:10.1007/s10588-010-9082-2)
5. Warren WH. 2018 Collective Motion in Human Crowds. *Curr. Dir. Psychol. Sci.* **27**, 232–240. (doi:10.1177/0963721417746743)
6. Kaminka GA, Fridman N. 2018 Simulating Urban Pedestrian Crowds of Different Cultures. *ACM Trans. Intell. Syst. Technol.* **9**, 1–27. (doi:10.1145/3102302)
7. Gipps PG. 1981 A behavioural *car*-following model for computer simulation. *Transp. Res. Part B* **15**, 105–111. (doi:10.1016/0191-2615(81)90037-0)
8. Oyler DW, Yildiz Y, Girard AR, Li NI, Kolmanovsky IV. 2016 A game theoretical model of traffic with multiple interacting drivers for use in autonomous vehicle development. In *2016 American Control Conference (ACC)*, pp. 1705–1710.
9. Şahin E. 2005 Swarm Robotics: From Sources of Inspiration to Domains of Application. In Swarm robotics (eds E Şahin, WM Spears), pp. 10–20. Berlin, Heidelberg: Springer Berlin Heidelberg.
10. Wurman P, D'Andrea R, Mountz M. 2008 Coordinating Hundreds of Cooperative, Autonomous Vehicles in Warehouses. *AI Mag* **29**, 9–20. (doi:10.1609/aimag.v29i1.2082)
11. Correll N, Martinoli A. 2009 Towards Multi-Robot Inspection of Industrial Machinery: From Distributed Coverage Algorithms to Experiments with Miniature Robotic Swarms. *IEEE Robot. Autom. Mag.* **16**, 103–112.
12. Brambilla M, Ferrante E, Birattari M, Dorigo M. 2013 Swarm robotics: a review from the swarm engineering perspective. *Swarm Intell.* **7**, 1–41. (doi:10.1007/s11721-012-0075-2)

13. Song Z, Vaughan RT. 2013 Sustainable robot foraging: adaptive fine-grained multi-robot task allocation for maximum sustainable yield of biological resources. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE.
14. Bayındır L. 2016 A review of swarm robotics tasks. *Neurocomputing* **172**, 292–321. (doi:10.1016/j.neucom.2015.05.116)
15. Slowik A, Kwasnicka H. 2018 Nature Inspired Methods and Their Industry Applications—Swarm Intelligence Algorithms. *IEEE Trans. Ind. Informatics* **14**, 2017. (doi:10.1109/tii.2017.2786782)
16. Hamann H. 2018 *Swarm robotics: a formal approach*. Cham, Switzerland: Springer.
17. Dorigo M, Theraulaz G, Trianni V. 2021 Swarm Robotics: Past, Present, and Future [Point of View]. *Proc. IEEE* **109**, 1152–1165. (doi:10.1109/jproc.2021.3072740)
18. Dorigo M, Maniezzo V, Colorni A. 1996 Ant system: optimization by a colony of cooperating agents. *IEEE Trans. Syst. Man Cybern. Part B* **26**, 29–41. (doi:10.1109/3477.484436)
19. Bonabeau E, Dorigo M, Theraulaz G. 2000 Inspiration for optimization from social insect behaviour. *Nature* **406**, 39–42. (doi:10.1038/35017500)
20. Vaughan R, Støy K, Sukhatme G, Matarić M. 2000 Go ahead, make my day: robot conflict resolution by aggressive competition. In *Proceedings of the 6th int. conf. on the Simulation of Adaptive Behavior*, Paris, France. (doi:10.7551/mitpress/3120.003.0052)
21. Zuluaga M, T.Vaughan R. 2005 Reducing spatial interference in robot teams by local-investment aggression. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Edmonton, Alta., Canada, pp. 2798–2805. (doi:10.1109/IROS.2005.1545099)
22. Godoy J, Karamouzas I, J.Guy S, Gini M. 2015 Adaptive learning for multi-agent navigation. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*, pp. 1577–1585.
23. Rosenfeld A, Kaminka GA, Kraus S, Shehory O. 2008 A study of mechanisms for improving robotic group performance. *Artif. Intell.* **172**, 633–655. (doi:10.1016/j.artint.2007.09.008)
24. Balch T, Arkin RC. 1998 Behavior-based formation control for multirobot teams. *IEEE Trans. Robot. Autom.* **14**, 926–939. (doi:10.1109/70.736776)
25. von Maltzahn G *et al.* 2011 Nanoparticles that communicate in vivo to amplify tumour targeting. *Nat. Mater.* **10**, 545–552. (doi:10.1038/nmat3049)
26. Wiesel-Kapah I, A.Kaminka G, Hachmon G, Agmon N, Bachelet I. 2016 Rule-based programming of molecular robot swarms for biomedical applications. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 3505–3512.
27. Kaminka GA, Spokoini-Stern R, Amir Y, Agmon N, Bachelet I. 2017 Molecular Robots Obeying Asimov's Three Laws of Robotics. *Artif. Life* **23**, 343–350. (doi:10.1162/artl_a_00235)
28. Llopis-Lorente A, Díez P, Sánchez A, Marcos MD, Sancenón F, Martínez-Ruiz P, Villalonga R, Martínez-Máñez R. 2017 Interactive models of communication at the nanoscale using nanoparticles that talk to one another. *Nat. Commun.* **8**, 15511. (doi:10.1038/ncomms15511)
29. Wu J *et al.* 2020 An immune cocktail therapy to realize multiple boosting of the *cancer*-immunity cycle by combination of drug/gene delivery nanoparticles. *Sci. Adv.* **6**, c7828. (doi:10.1126/sciadv.abc7828)
30. von Neumann J, Morgenstern O. 1944 *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.
31. Popper KR. 1994 *The myth of the framework: defence of science and rationality*. Routledge Press, UK.
32. Newell A. 1982 The knowledge level. *Artif. Intell.* **18**, 87–127. (doi:10.1016/0004-3702(82)90012-1)
33. Newell A. 1990 Unified Theories of Cognition. Cambridge, Massachusetts: Harvard University Press.
34. Feinerman O, Korman A. 2017 Individual versus collective cognition in social insects. *J. Exp. Biol.* (eds JD Levine, DJC Kronauer, MH Dickinson), **220**, 73–82. (doi:10.1242/jeb.143891)

35. Marden JR, Shamma JS. 2017 Game-Theoretic Learning in Distributed Control. In *Handbook of dynamic game theory* (eds T Basar, G Zaccour). Cham: Springer International Publishing. (doi:10.1007/978-3-319-27335-8_9-1)
36. Wolpert DH, Tumer K. 2002 Collective Intelligence, *Data* Routing and Braess' Paradox. *J. Artif. Intell. Res.* **16**, 359–387. (doi:10.1613/jair.995)
37. Aoki I. 1982 A simulation study on the schooling mechanism in fish. *Nippon SUISAN GAKKAISHI* **48**, 1081–1088. (doi:10.2331/suisan.48.1081)
38. Reynolds CW. 1987 Flocks, herds and schools: A distributed behavioral model. *ACM SIGGRAPH Comput. Graph.* **21**, 25–34. (doi:10.1145/37402.37406)
39. Mataric MJ. 1994 Interaction and Intelligent Behavior. PhD thesis, Massachusetts Institute of Technology.
40. Balch T, Hybinette M. 2000 Social potentials for scalable multirobot formations. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA-00)*.
41. Turgut AE, Çelikkanat H, Gökçe F, Şahin E. 2008 Self-organized flocking in mobile robot swarms. *Swarm Intell.* **2**, 97–120. (doi:10.1007/s11721-008-0016-2)
42. Moshtagh N, Michael N, Jadbabaie A, Daniilidis K. 2009 Vision-Based, Distributed Control Laws for Motion Coordination of Nonholonomic Robots. *IEEE Trans. Robot.* **25**, 851–860. (doi:10.1109/tro.2009.2022439)
43. Batalin MA, Sukhatme GS. 2002 Spreading out: A local approach to multi-robot coverage. In *International symposium on distributed autonomous robotic systems (dars)*, pp. 373–382. Tokyo, Japan: Springer.
44. Shakhatreh H, Khreishah A, Chakareski J, B.Salameh H, Khalil I. 2016 On the continuous coverage problem for a swarm of uavs. In *IEEE 37th Sarnoff Symposium*, pp. 130–135. IEEE. (doi:10.1109/SARNOF.2016.7846742)
45. Rothermich JA, Ecemiş Mİ, Gaudiano P. 2004 Distributed localization and mapping with a robotic swarm. In *International Workshop on Swarm Robotics*, pp. 58–69. Springer.
46. Hsiang TR, Arkin EM, Bender MA, Fekete SP, Mitchell JSB. 2004 Algorithms for rapidly dispersing robot swarms in unknown environments. In *Springer tracts in advanced robotics algorithmic foundations of robotics v*, pp. 77–93. Berlin, Germany: Springer. (doi:10.1007/978-3-540-45058-0_6)
47. Shell DA, Mataric MJ. 2006 On foraging strategies for large-scale multi-robot systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2717–2723. IEEE. (doi:10.1109/IROS.2006.281996)
48. Dirafzoon A, Lobaton E. 2013 Topological mapping of unknown environments using an unlocalized robotic swarm. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5545–5551. IEEE.
49. Winfield AF. 2009 Foraging Robots. In Encyclopedia of complexity and systems science (ed. RA Meyers), pp. 3682–3700. New York, NY: Springer New York.
50. Talamali MS, Bose T, Haire M, Xu X, Marshall JAR, Reina A. 2020 Sophisticated collective foraging with minimalist agents: a swarm robotics test. *Swarm Intell.* **14**, 25–56. (doi:10.1007/s11721-019-00176-9)
51. Goldberg D, J.Mataric M. 1997 Interference as a tool for designing and evaluating multi-robot controllers. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*, pp. 637–642. Providence, RI: AAAI Press.
52. Rybski P, A.Larson M, Gini M. 1998 Performance evaluation of multiple robots in a search and retrieval task. In *Proceedings of the Workshop on Artificial Intelligence and Manufacturing*, Albuquerque, NM, pp. 153–160.
53. Balch T. 1999 The impact of diversity on performance in multi-robot foraging. In *Proceedings of the third annual conference on Autonomous Agents*, pp. 92–99.
54. Czirók A, Barabási AL, Vicsek T. 1999 Collective Motion of Self-Propelled Particles: Kinetic Phase Transition in One Dimension. *Phys. Rev. Lett.* **82**, 209–212. (doi:10.1103/physrevlett.82.209)
55. Vicsek T, Zafeiris A. 2012 Collective Motion. *Phys. Rep* **517**, 71–140. (doi:10.1016/B978-1-55860-307-3.50049-6)
56. Lerman K, Galstyan A. Mathematical model of foraging in a group of robots: Effect of interference. *Auton. Robot* **13**, 127–141. (doi:10.1109/tsmcc)

57. Calovi DS, Lopez U, Ngo S, Sire C, Chaté H, Theraulaz G. 2014 Swarming, schooling, milling: phase diagram of a *data*-driven fish school model. *New J. Phys.* **16**, 015026. (doi:10.1088/1367-2630/16/1/015026)
58. Gazi V, Passino KM. 2004 Stability Analysis of Social Foraging Swarms. *IEEE Trans. Syst. Man Cybern. Part B* **34**, 539–557.
59. Tumer K, Agogino AK, Wolpert DH. 2002 Learning sequences of actions in collectives of autonomous agents. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1*, pp. 378–385. New York, NY, USA: ACM.
60. Devlin S, L.Yliniemi D, Tumer K. 2014 Potential-based difference rewards for multiagent reinforcement learning. In *Proceedings of the Thirteenth International Joint Conference on Autonomous Agents and Multiagent Systems*, Paris, France.
61. Monderer D, Shapley LS. 1996 Potential Games. *Games Econ. Behav.* **14**, 124–143. (doi:10.1006/game.1996.0044)
62. Marden JR, Wierman A. 2013 Distributed Welfare Games. *Oper. Res.* **61**, 155–168. (doi:10.1287/opre.1120.1137)
63. Tan M. 1993 Multi-agent reinforcement learning: independent vs cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning*, pp. 330–337. (doi:10.1016/B978-1-55860-307-3.50049-6)
64. Busoniu L, Babuska R, De Schutter B. 2008 A Comprehensive Survey of Multiagent Reinforcement Learning. *IEEE Trans. Syst. Man Cybern. Part C* **38**, 156–172. (doi:10.1109/tsmcc.2007.913919)
65. Nowé A, Vrancx P, De Hauwere YM. 2012 Game Theory and Multi-Agent Reinforcement Learning. In Reinforcement learning, pp. 441–470. Springer. (doi:10.1007/978-3-642-27645-3_14)
66. Hernandez-Leal P, Kaisers M, Baarslag T, de Cote EM. 2019 A Survey of Learning in Multiagent Environments: Dealing with Non-Stationarity. *arXiv* doi:10.48550/arxiv.1707.09183.
67. Fatima S, Jennings NR, Wooldridge M. 2024 Learning to Resolve Social Dilemmas: A Survey. *J. Artif. Intell. Res.* **79**, 895–969. (doi:10.1613/jair.1.15167)
68. Kaminka GA, Erusalimchik D, Kraus S. 2010 Adaptive multi-robot coordination: a game-theoretic perspective. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA-10)*, Anchorage, AK. (doi:10.1109/ROBOT.2010.5509316)
69. Douchan Y, Wolf R, Kaminka GA. 2019 Swarms can be rational. In *Proceedings of the International Joint Conference on Autonomous Agents and Multi-Agent Systems*.
70. Erusalimchik D. 2009 Reinforcement Learning of Multi-Robot Coordination Based on Resource Spending Velocity. Master's thesis, Bar Ilan University.
71. Pinciroli C *et al.* 2012 ARGoS: a modular, parallel, multi-engine simulator for multi-robot systems. *Swarm Intell.* **6**, 271–295. (doi:10.1007/s11721-012-0072-5)
72. Hartman ER. 2022 Swarming Bandits: A Rational and Practical Model of Swarm Robotic Tasks. Master's thesis, Bar Ilan University.
73. Katz K. 2023 Competitive Multi-Swarm Systems. Master's thesis, Bar Ilan University.
74. Hazard CJ, R.Wurman P. 2006 Alphabet soup: a testbed for studying resource allocation in multi-vehicle systems. In *Proceedings of the 2006 AAAI Workshop on Auction Mechanisms for Robot Coordination*, pp. 23–30.
75. Zedadra O, Jouandeau N, Seridi H, Fortino G. 2017 Multi-Agent Foraging: state-of-the-art and research challenges. *Complex Adapt. Syst. Model.* **5**. (doi:10.1186/s40294-016-0041-8)
76. Hoff NR, Sagoff A, Wood RJ, Nagpal R. 2010 Two foraging algorithms for robot swarms using only local communication. In *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Tianjin, China, pp. 123–130. IEEE. (doi:10.1109/ROBIO.2010.5723314)
77. Pitonakova L, Crowder R, Bullock S. 2014 Understanding the role of recruitment in collective robot foraging (eds H Lipson, H Sayama, J Rieffel, S Risi, R Doursat). In *Artificial Life XIV: Proceedings of the Fourteenth International Conference on the Synthesis and Simulation of Living Systems*, pp. 264–271. Massachusetts Institute of Technology (MIT) Press.
78. Schneider-Fontan M, Matarić M. 1998 Territorial Multi-Robot Task Division. *IEEE Trans. Robot. Autom* **14**, 815–822. (doi:10.1109/70.720357)

79. Sung C, Ayanian N, Rus D. 2013 Improving the performance of multi-robot systems by task switching. In *2013 IEEE International Conference on Robotics and Automation*, pp. 2999–3006. IEEE.
80. Shapley LS. 1953 Stochastic Games. *Proc. Natl. Acad. Sci.* **39**, 1095–1100. (doi:10.1073/pnas.39.10.1095)
81. Matignon L, Laurent GJ, Le Fort-Piat N. 2012 Independent reinforcement learners in cooperative Markov games: a survey regarding coordination problems. *Knowl. Eng. Rev.* **27**, 1–31. (doi:10.1017/s0269888912000057)
82. Claus C, Boutilier C. 1998 The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, pp. 746–752. USA: American Association for Artificial Intelligence.
83. Wang X, Sandholm T. 2002 Reinforcement learning to play an optimal nash equilibrium in team markov games. In Advances in neural information processing systems (eds S Becker, S Thrun, K Obermayer), vol. **15**. Cambridge, MA: MIT Press.
84. Douchan Y. 2018 Reinforcement Learning in Multi-Robot Swarms. Master's thesis, Tel Aviv University.
85. Marden JR, Wierman A. 2008 Distributed welfare games with applications to sensor coverage. In *47th IEEE Conference on Decision and Control*, pp. 1708–1713. Cancun, Mexico: IEEE.
86. Tumer K, Welch Z, Agogino A. 2008 Aligning social welfare and agent preferences to alleviate traffic congestion. In *Proceedings of the Seventh International Joint Conference on Autonomous Agents and Multiagent Systems*, Estoril, Portugal.
87. Kaminka GA, Douchan Y. 2025 Heterogeneous Foraging Swarms Can Be Better. *Front. Robot. AI* (doi:10.3389/frobt.2024.1426282)
88. Raileanu R, Denton E, Szlam A, Fergus R. 2018 Modeling others using oneself in multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pp. 4257–4266.
89. de Hauwere YM, Devlin S, Kudenko D, Nowé A. 2016 Context-sensitive reward shaping for sparse interaction multi-agent systems. *Knowl. Eng. Rev.* **31**, 59–76. (doi:10.1017/s0269888915000193)
90. Hauwere YM. 2011 Sparse Interactions in Multi-Agent Reinforcement Learning. PhD thesis, [Brussels]: Vrije Universiteit Brussel.
91. Littman ML. 2001 Value-function reinforcement learning in Markov games. *Cogn. Syst. Res.* **2**, 55–66. (doi:10.1016/s1389-0417(01)00015-8)
92. Kapetanakis S, Kudenko D. 2002 Reinforcement Learning of Coordination in Cooperative Multi-Agent Systems. In *AAAI/IAAI*, pp. 326–331.
93. Zhang C, Lesser V. 2013 Coordinating Multi-Agent Reinforcement Learning with Limited Communication. In *Proceedings of the International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pp. 1101–8IFAAMAS. http://mas.cs.umass.edu/paper/520.
94. Zhang K, Yang Z, Başar T. 2021 Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. In Handbook of reinforcement learning and control, pp. 321–384. (doi:10.1007/978-3-030-60990-0_12)
95. Lauer M, Riedmiller M. 2000 An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *Proceedings of the Seventeenth International Conference on Machine Learning*. Citeseer.
96. Hao J, Leung H. 2013 The Dynamics of Reinforcement Social Learning in Cooperative Multiagent Systems. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 184–190.
97. Schwartz A. 1993 A Reinforcement Learning Method for Maximizing Undiscounted Rewards. In Proceedings of the Tenth International Conference on Machine Learning (ICML), University of Massachusetts, Amherst, pp. 298–305. Elsevier. (doi:10.1016/b978-1-55860-307-3.50045-9)
98. Mahadevan S. 1996 Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Mach. Learn.* **22**, 159–195. (doi:10.1007/bf00114727)

99. Auer P, Cesa-Bianchi N, Fischer P. 2002 Finite-Time Analysis of the Multiarmed Bandit Problem. *Mach. Learn.* **47**, 235–256. (doi:10.1023/A:1013689704352)
100. Erusalimchik D, Kaminka GA. 2008 Towards Adaptive Multi-Robot Coordination Based on Resource Expenditure Velocity. In *Proceedings of the Tenth Conference on Intelligent Autonomous Systems (IAS-10)*. IOS Press.
101. Douchan Y, Kaminka GA. 2016 The Effectiveness Index Intrinsic Reward for Coordinating Service Robots (eds S Berman, M Gauci, E Frazzoli, A Kolling, R Gross, A Martinoli, F Matsuno). In *13th International Symposium on Distributed Autonomous Robotic Systems (DARS-2016)*. Springer.
102. Yıldız A. 2016 Foraging Motion of Swarms as Nash Equilibria of Differential Games. PhD thesis, [Turkey]: Bilkent University.
103. Friedman A. 1972 Stochastic Differential Games. *J. Differ. Equ.* **11**, 79–108. (doi:10.1016/0022-0396(72)90082-4)
104. Sutton RS, Precup D, Singh S. 1999 Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artif. Intell.* **112**, 181–211. (doi:10.1016/s0004-3702(99)00052-1)
105. van den Berg J, Guy SJ, Lin M, Manocha D. 2011 Reciprocal n-Body Collision Avoidance. In Robotics Research, pp. 3–19. Berlin, Heidelberg: Springer. (doi:10.1007/978-3-642-19457-3_1)
106. Bouraine S, Fraichard T, Azouaoui O, Salhi H. 2014 Passively safe partial motion planning for mobile robots with limited field-of-views in unknown dynamic environments. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, Hong Kong, China. (doi:10.1109/ICRA.2014.6907375)
107. Fox D, Burgard W, Thrun S. 1997 The dynamic window approach to collision avoidance. *IEEE Robot. Autom. Mag.* **4**, 23–33. (doi:10.1109/100.580977)
108. Kengyel D, Hamann H, Zahadat P, Radspieler G, Wotawa F, Schmickl T. 2015 Potential of heterogeneity in collective behaviors: a case study on heterogeneous swarms. In *International Conference on Principles and Practice of Multi-Agent Systems*, pp. 201–217. Cham: Springer.
109. Dorigo M *et al.* 2012 Swarmanoid: A Novel Concept for the Study of Heterogeneous Robotic Swarms. *IEEE Robot. Autom. Mag.* **20**, 60–71. (doi:10.1109/mra.2013.2252996)
110. Prorok A, A.Hsieh M, Kumar V. 2016 Formalizing the impact of diversity on performance in a heterogeneous swarm of robots. In *Proceedings of IEEE International Conference on Robotics and Automation*, Stockholm, Sweden, pp. 5364–5371. IEEE. (doi:10.1109/ICRA.2016.7487748)
111. Harada K, Corradi P, Popesku S, Liedk J. 2010 Reconfigurable Heterogeneous Mechanical Modules (Ch 2.1). In *Symbiotic multi-robot organisms: reliability, adaptability, evolution*. Berlin: Springer-Verlag. (doi:10.1007/978-3-642-11692-6_3)
112. Matarić MJ. 1997 Reinforcement Learning in the Multi-Robot Domain. *Auton. Robot.* **4**, 73–83. (doi:10.1023/A:1008819414322)
113. Yang E, Gu D. 2004 *Multiagent Reinforcement Learning for Multi-Robot Systems: A Survey. Technical Report CSM-404*. University of Essex Department of Computer Science.
114. Riedmiller M, Gabel T, Hafner R, Lange S. 2009 Reinforcement learning for robot soccer. *Auton. Robot.* **27**, 55–73. (doi:10.1007/s10514-009-9120-4)
115. Kober J, Bagnell JA, Peters J. 2013 Reinforcement Learning in Robotics: A survey. *Int. J. Robot. Res* 1238–1274. (doi:10.1177/0278364913495721)
116. La HM, Lim R, Sheng W. 2015 Multirobot Cooperative Learning for Predator Avoidance. *IEEE Trans. Control Syst. Technol.* **23**, 52–63. (doi:10.1109/tcst.2014.2312392)
117. Watkins CJCH. 1989 Learning from Delayed Rewards. PhD thesis, [Cambridge United Kingdom]: King's College.
118. Sutton RS, Barto AG. 2018 *Reinforcement learning: an introduction*. MIT press.
119. Alers S, Tuyls K, Ranjbar-Sahraei B, Claes D, Weiss G. 2014 Insect-inspired robot coordination: foraging and coverage. In *Artificial Life 14: Proceedings of the Fourteenth International Conference on the Synthesis and Simulation of Living Systems*, pp. 761–768. (doi:10.1162/978-0-262-32621-6-ch123)

120. Balch T, Dellaert F, Feldman A, Guillory A, Isbell CL, Khan Z, Pratt SC, Stein AN, Wilde H. 2006 How Multirobot Systems Research will Accelerate our Understanding of Social Animal Behavior. *Proc. IEEE* **94**, 1445–1463. (doi:10.1109/jproc.2006.876969)
121. Horsevad N, Kwa HL, Bouffanais R. 2022 Beyond Bio-Inspired Robotics: How Multi-Robot Systems Can Support Research on Collective Animal Behavior. *Front. Robot. AI* **9**. (doi:10.3389/frobt.2022.865414)
122. Goldshtein A *et al.* 2020 Reinforcement Learning Enables Resource Partitioning in Foraging Bats. *Curr. Biol.* **30**, 4096–4102.(doi:10.1016/j.cub.2020.07.079)
123. Ayali A, Kaminka GA. 2023 The *hybrid* bio-robotic swarm as a powerful tool for collective motion research: a perspective. *Front. Neurorobotics* **17**, 1215085. (doi:10.3389/fnbot.2023.1215085)
124. Shefi P, Ayali A, A.Kaminka G. 2024 Pausing Makes Perfect: Intermittent Pauses for Resilient Swarming. In *International Symposium on Distributed Autonomous Robotic Systems (DARS)*. Springer.
125. Nummelin S, Shen B, Piskunen P, Liu Q, Kostiainen MA, Linko V. 2020 Robotic DNA Nanostructures. *ACS Synth. Biol.* **9**, 1923–1940. (doi:10.1021/acssynbio.0c00235)
126. Alon L, Weitman H, Shleyfman A, Kaminka GA. 2024 Planning to be healthy: towards personalized medication planning. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*. (doi:10.3233/FAIA240996)