WARWICK

From Minimal Cognition to Collective Dynamics

Sam Turley

Monday 18th May, 2026

WARWICK

Collective dynamics in animal flocking is generally considered to be a spectacle. Large-scale coordinated group behaviour arising from individuals is impressive.

Fig. 1: Flocking in birds

Objective of this work

WARWICK

Many models for collective dynamics require access to neighbours' positions or velocities, even if unrealistic. Here, we are looking to explore collective dynamics where...

- ... the decision process is fast
- ... the input data is only visual, with no neighbour identification

We use *Q-Learning*, a reinforcement learning algorithm. The goal is to learn a decision-making policy that evaluates the possible actions in the current state. As such, we need to define

- **Actions**. The things an agent can do, a_i^t
- **States**. What the agent currently observes, s_i^t
- **Rewards**. External feedback to promote or punish an action, r_i^t

Actions - a_i^t

WARWICK

We use N agents in infinite 2D space. Unit discs with an orientation and have some movement rules associated.

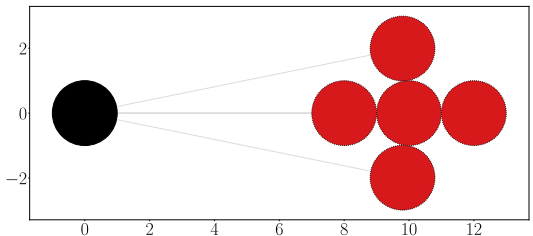


Fig. 2: The ways an agent is allowed to move, 1 time step

Parameters: $v_0, \Delta v$.

Using the Visual Input

Each agent can observe basic visual input:

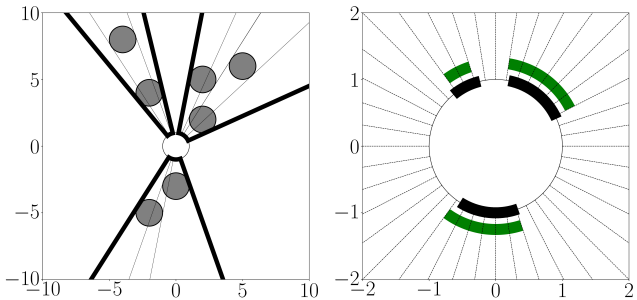


Fig. 3: The visual projection and the associated sensor activations (green)

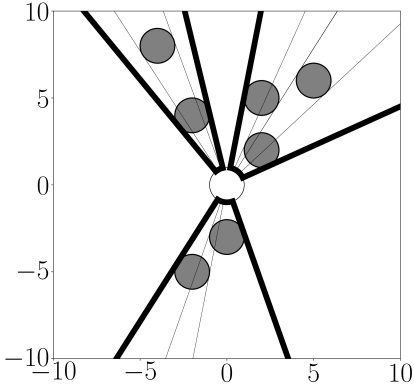
Parameters: n_s

Projection Representation

The projection function maps the true state of the system to the observed state by i

$$\phi_i^t(\mathbf{x}_{-i}) \mapsto \{0,1\}^{n_s}. \tag{1}$$

The problem is a Partially Observable Decision-Making Process.



We could just use

$$s_i^t = \phi_i^t \tag{2}$$

... but the decision becomes purely reactionary. One frame of visual information provides position. To infer orientation, we need to introduce memory:

$$s_i^t = (\phi_i^t, ..., \phi_i^{t-M+1}) \tag{3}$$

Introduction of memory helps to reduce partial observability - it makes the system "more" Markovian.

We define the opacity of a visual state as the proportion of active sensors

$$\Theta_i^t = \frac{n_i^t}{n_s}. \tag{4}$$

The reward scheme is given by

$$r^t = 1 - 4\left(\Theta_i^t - \frac{1}{2}\right)^2 \tag{5}$$

Rewards - r_i^t

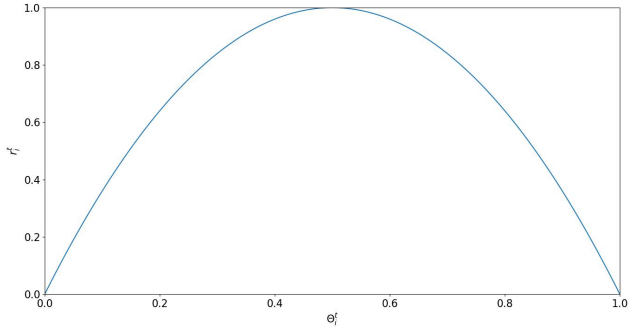


Fig. 5: Agents are rewarded for maintaining a marginally opaque visual state

Learning A Policy

WARWICK

Given the size of the state space, we use a Deep Q-Neural Network to estimate the Q-values. We use *Centralised Training* (think a hive mind), where all agents use the same QNN and their experiences are shared for training.

WARWICK

To ensure good diversity of training data, we initialise the agents in a box of size (L, L) where $L \sim U(N, 5N)$. Each agent has orientation $\theta_i = 3\Delta\theta \cdot \eta_i$, where $\eta_i \sim N(0, 1)$.

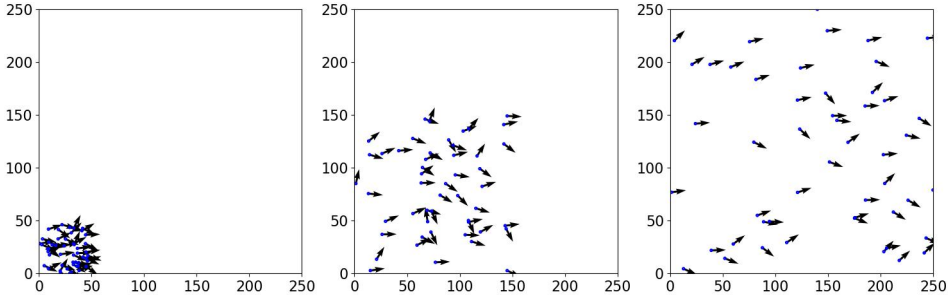


Fig. 6: A variety of different density initialisations

During Training Simulations

WARWICK

During training, one of the 5 movement actions is selected according to a soft-max (Boltzman) policy:

$$\mathbb{P}(a|s^t) \propto \exp\left(\frac{Q(s^t, a)}{T}\right) \tag{6}$$

Early simulations, $s \ll s_{\max}$, are ran "hot" with a large T and then cooled gradually.

$$T(s) = T_0 \exp\left(-k\frac{s}{s_{\max}}\right) \tag{7}$$

We also train a left-right symmetry-enforced network. During training, the Q-values are changed to policy is changed to

$$Q^*(s^t, a) = \frac{1}{2}\left(Q(s^t, a) + Q(s^t_m, a_m)\right). \tag{8}$$

Symmetric states then have a symmetric policy.

Training Parameters

WARWICK

After each simulation, we train the network on $(s^t, a^t, r^{t+1}, s^{t+1})$ and the mirror data point $(s_m^t, a_m^t, r^{t+1}, s_m^{t+1})$.

N	50	v_0	10
Δv	2	$\Delta\theta$	0.2
$t_{\max}$	250	n_s	40
k	10	Memory	2
$s_{\max}$	10000	T_0	10

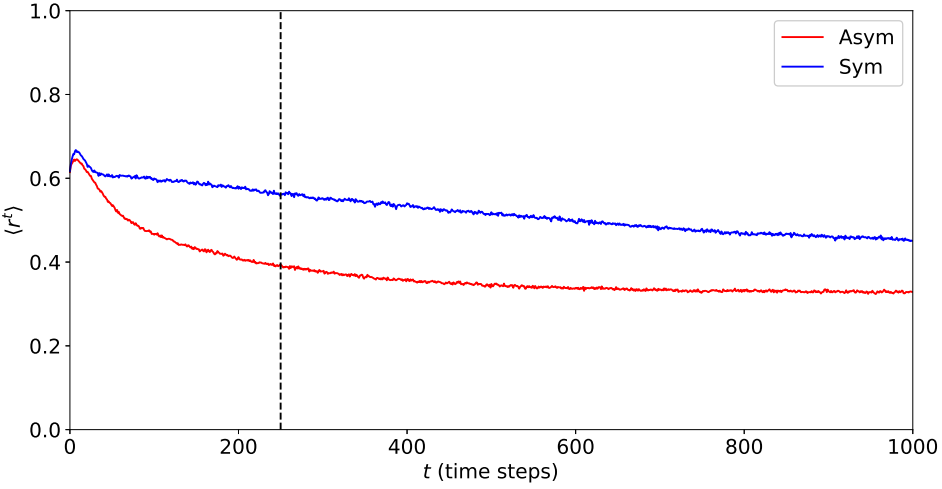
Table 1: Parameters for training

Results

WARWICK

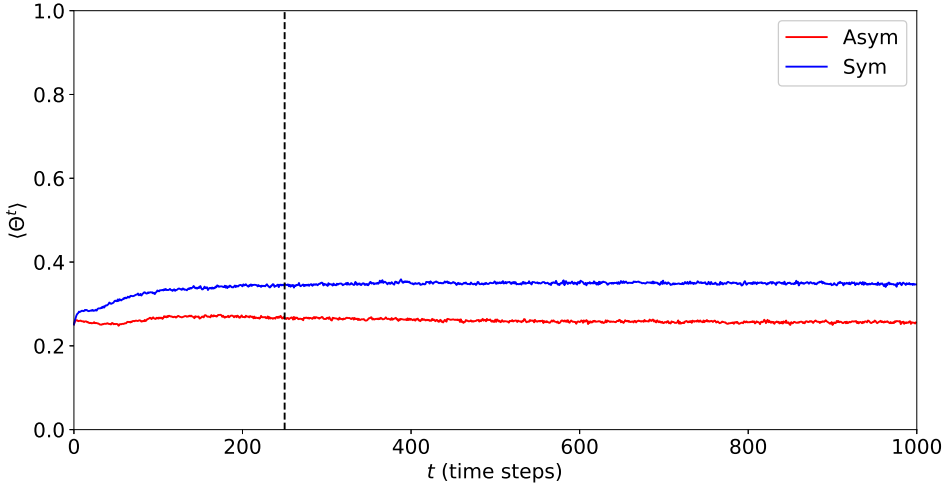
Videos

Results - Average Reward



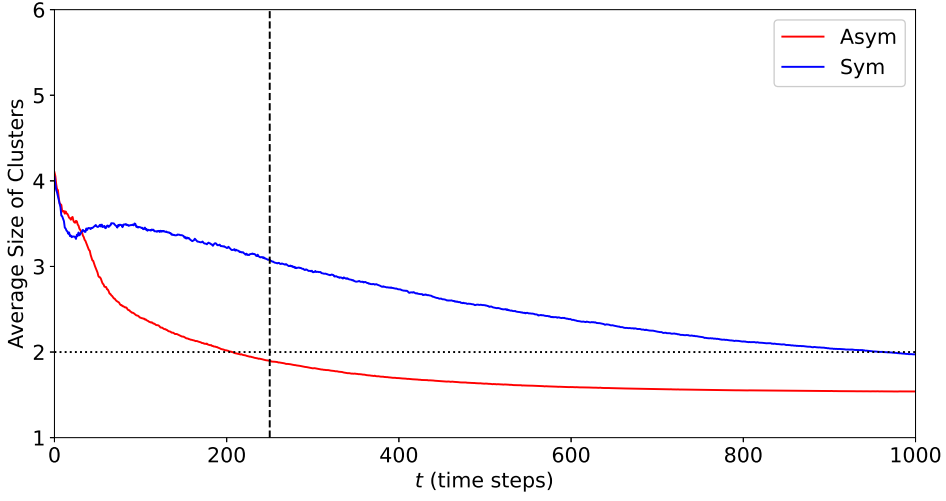
Results - Average Opacity

WARWICK

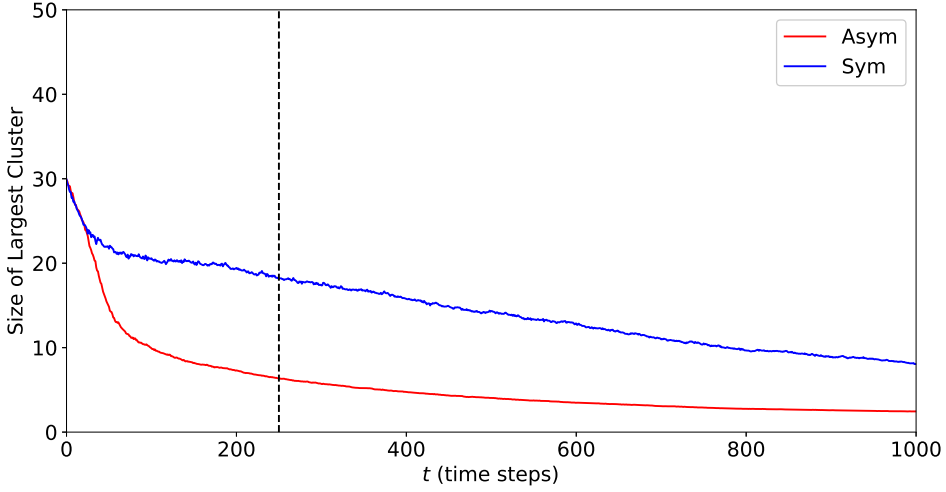


Results - Average Size of Clusters

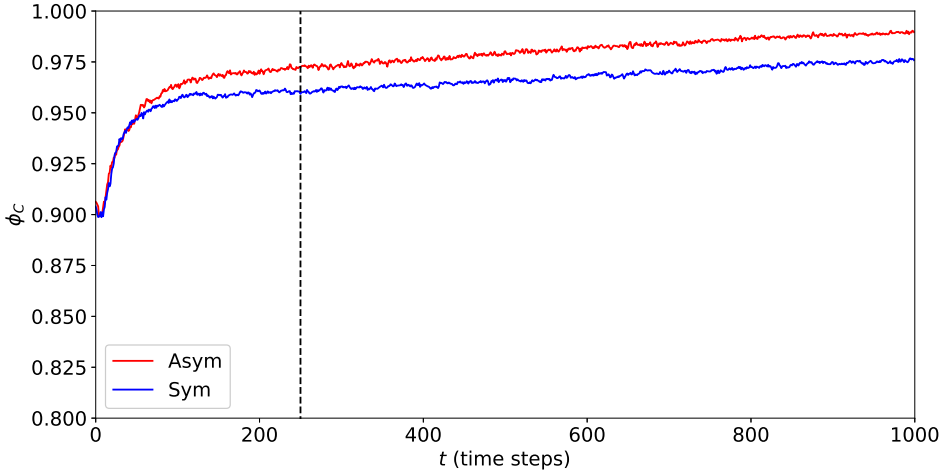
WARWICK



Results - Average Size of Largest Cluster



Results - Average Order of Largest Cluster



Observations

WARWICK

- We can generate *some* collective behaviour without neighbour identification
- The network has learnt what it was trained to do: consistently accumulate rewards (maintain marginal opacity) over time
- The network has found an "easy" solution; breaking up into small flocks means the visual states are more informative

Further Work

WARWICK

- **Utilitarian Training.** We are currently training networks where the reward for all agents at each time step is the average reward, $r = \langle r^t \rangle$. Are the flocks different if you care about everyones reward?
- **Reward Scheme**. We are looking at a more complicated reward scheme, based on entropy-maximising flocking. Can we generate cohesive, co-aligning flocks?
- **Memory**. We need to check whether the time step of memory is necessary. What policy emerges?
- **Decentralised Training.** Currently all the agents are identical. Does N unique QNNs have different dynamics as a flock?

Thank you, questions and sugggestions!