Investigating the emergence of complex behaviours in an agent-based model using reinforcement learning.

Sedar Olmez, Daniel Birks and Alison Heppenstall

UNIVERSITY OF LEEDS

The Alan Turing Institute

Research question

Can **Machine Learning** techniques be adopted in **agent-based models** to aide in the development of better **human decision-making**?

Reinforcement learning

- A branch of Machine Learning where a computer program (Agent) is trained over time to make decisions in an environment that increases its reward function.

Agent-based modelling

- A computation model used to simulate the actions and interactions of autonomous agents with the ability to trace each individual agent's behavior.

RATIONALE Few approaches exist that can **quantify** and formalise key individual **behaviours.**

RATIONALE Increased interest in simulating intelligent adaptive behaviours.

PROBLEM Pre-existing decision-making frameworks are **not reflective of real-world scenarios** and fail to capture how behaviours can develop in response to **stimuli.**

METHOD A suitable method Reinforcement Learning (RL).

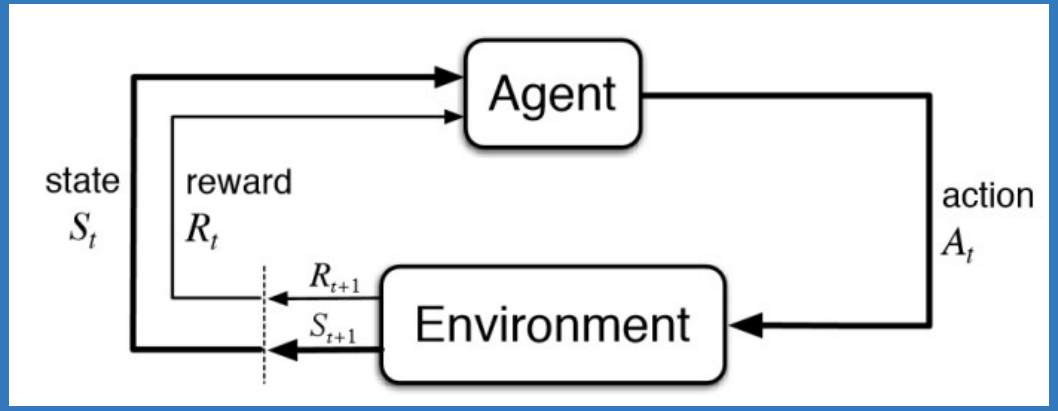
QUESTION How can RL be used to simulate emergent agent behaviours using a simple predator-prey like model.

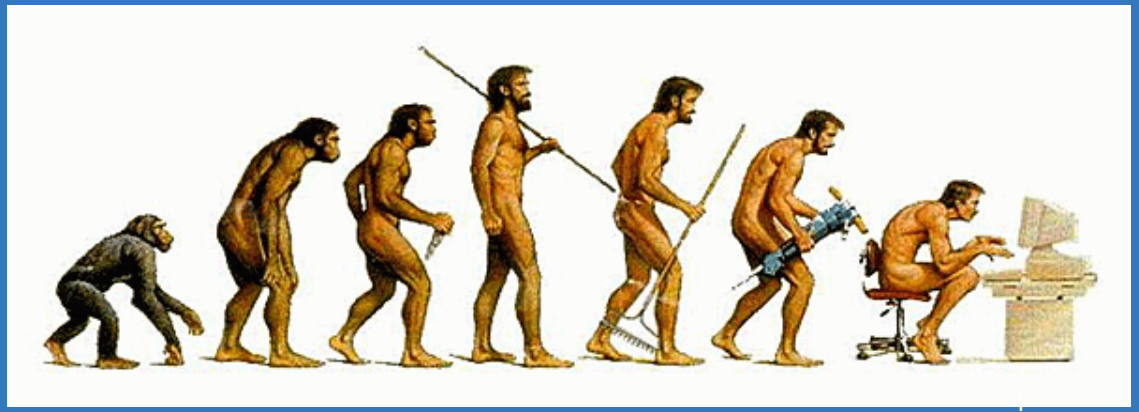
RESULT We found that agents exhibit properties of real-world **intelligent adaptive behaviours**.

Reinforcement learning (decision making framework)

"An agent (autonomous entity) performing actions in some environment and receives a reward; the learning challenge is to find a course of actions that maximise the reward received."

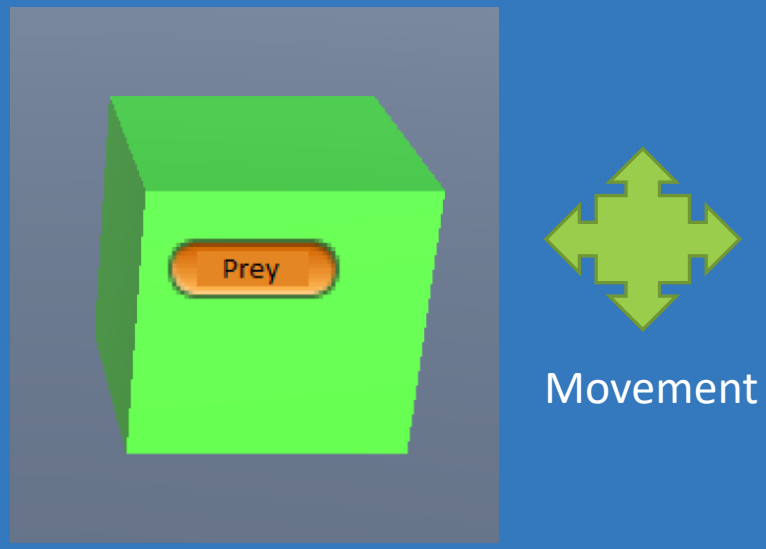
Wooldridge, 2020.





TIME

Prey (agent)



Movement

Variable	Description
Movement_speed	The speed of movement
Rigidbody	The 3D agent object has a rigid body property
Cube	The 3D agent's shape property
Box_Collider	A component used to trigger an event when it comes into contact with another object
Prey.cs	The C# script component allows the agent to perform actions in the environment
Camera	A camera component traces the movement of the agent in first person
Velocity	The velocity of the agent
Turn_speed	The speed of which the agent turns
Move_speed	The speed of which the agent moves on the X, Z axis
Normal_material	Normal material (agent is neither rewarded or penalised if it interacts with these)
Good_material	Good material (agent is rewarded if it interacts with these)
Bad_material	Bad material (agent is penalised if it interacts with these)
Use_Vector_Obs	If checked, the agent will send information to the neural network during training
Ray_Perception_Sensor	A sensor which identifies objects within a given perimeter

Table 7: Prey agent's parameters.

Goal:

- Collect as many positive points as possible.
- Avoid contact with predator.
- Avoid contact with negative points

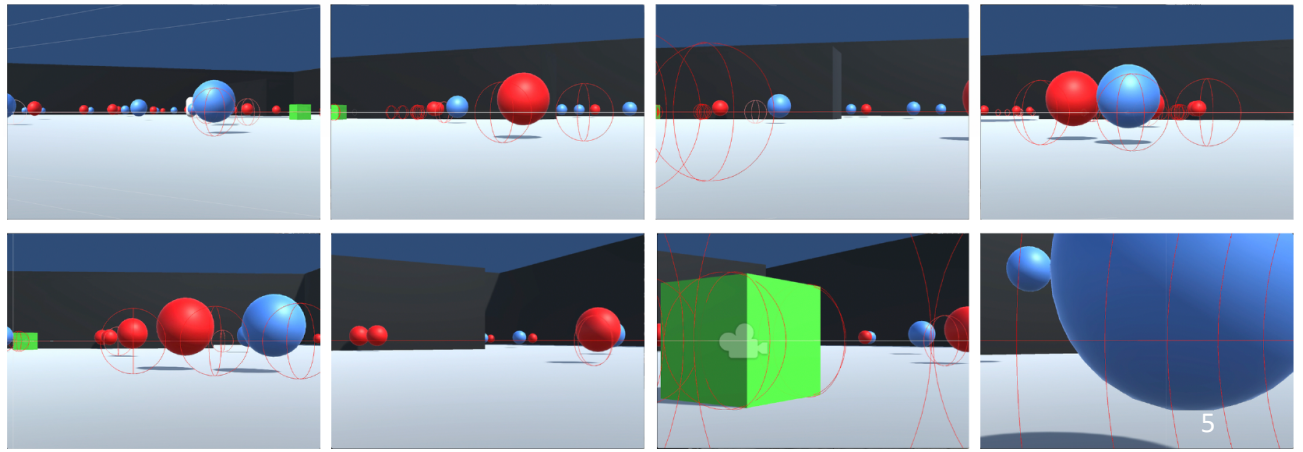


Figure 14: First person view from Prey agent's perspective.

Predator (agent)

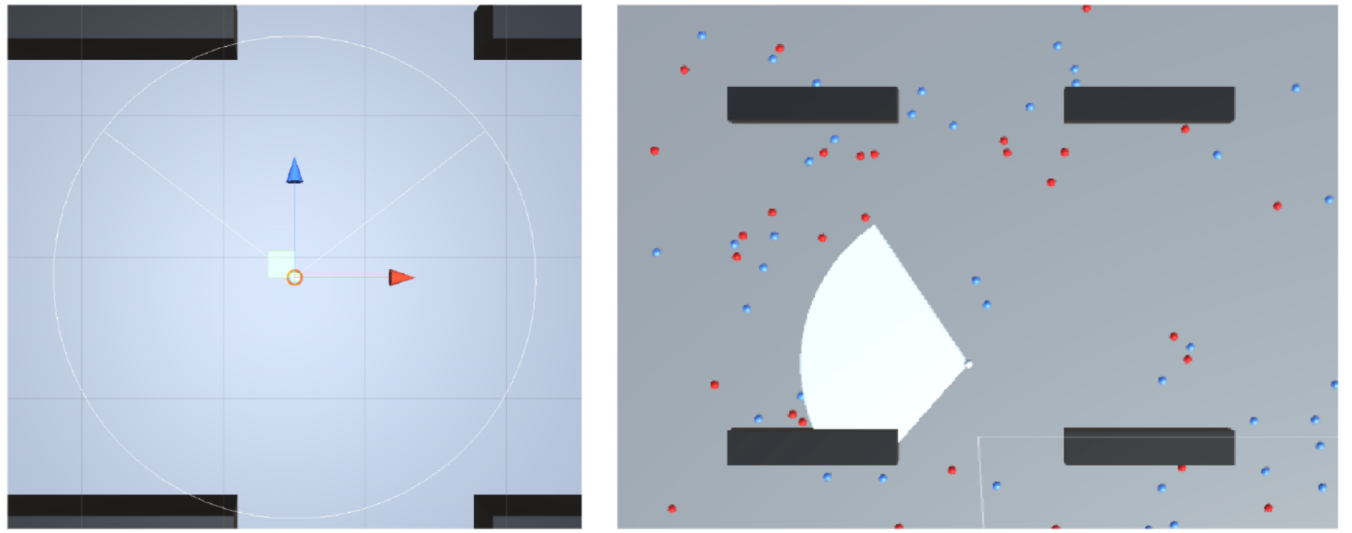
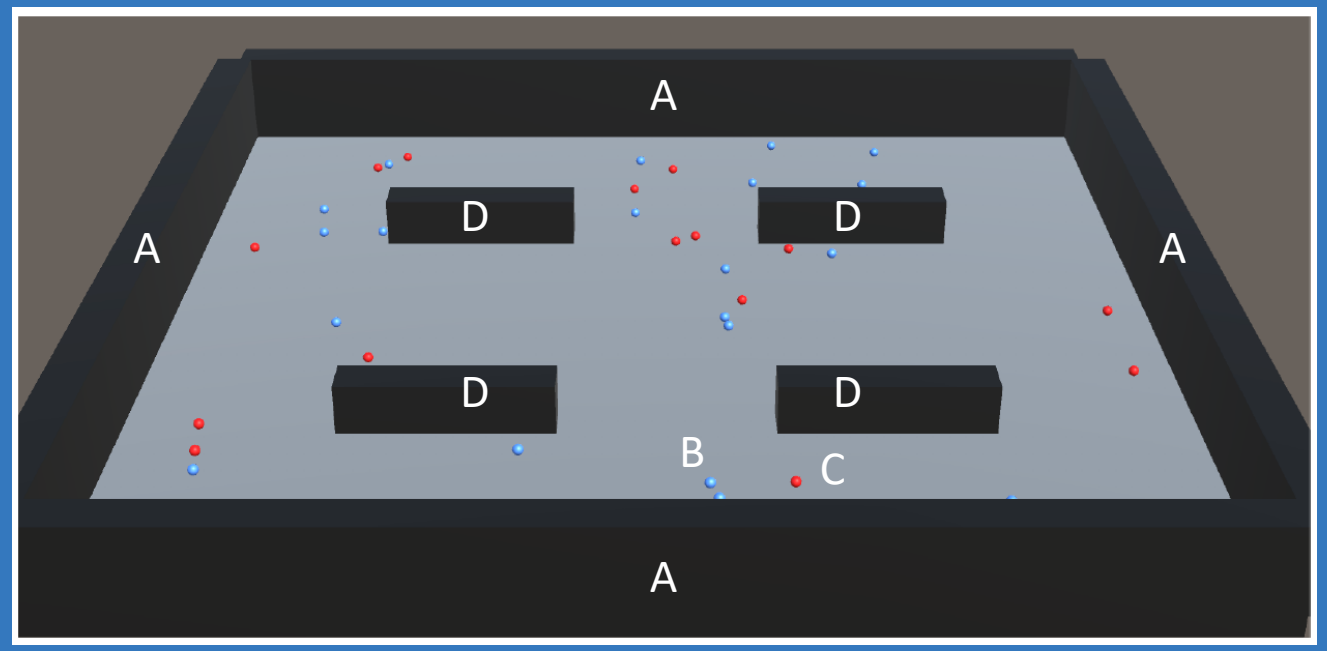


Figure 3: The area in which the predator can see (A). The model environment and the vision cone of the predator looking for prey agents (B).

Goal: Capture as many prey agents as possible

Environment configuration



(A) The environment contains several **barriers** to prevent the agents from falling off the surface plane.

The **blue spheres** (B) are positive points, while the **red spheres** (C) are negative points. These are randomly generated and re-spawn (reappear) when they are consumed.

(D) The environment also contains **four barriers** within the agent interaction space, these were added to test if prey agents will evolve to utilize these barriers as a means of preventing the predator from catching them.

Hyper-parameter tuning (pre-training)

- To successfully train a RL algorithm, hyper-parameter tuning must take place to ensure the performance of learning processes and quality of generated motions (Kim & Lee 2019; Juliani et al. 2018).
- According to Poole & Mackworth (2010), an RL model is performing well if the **cumulative reward is increasing** during training.

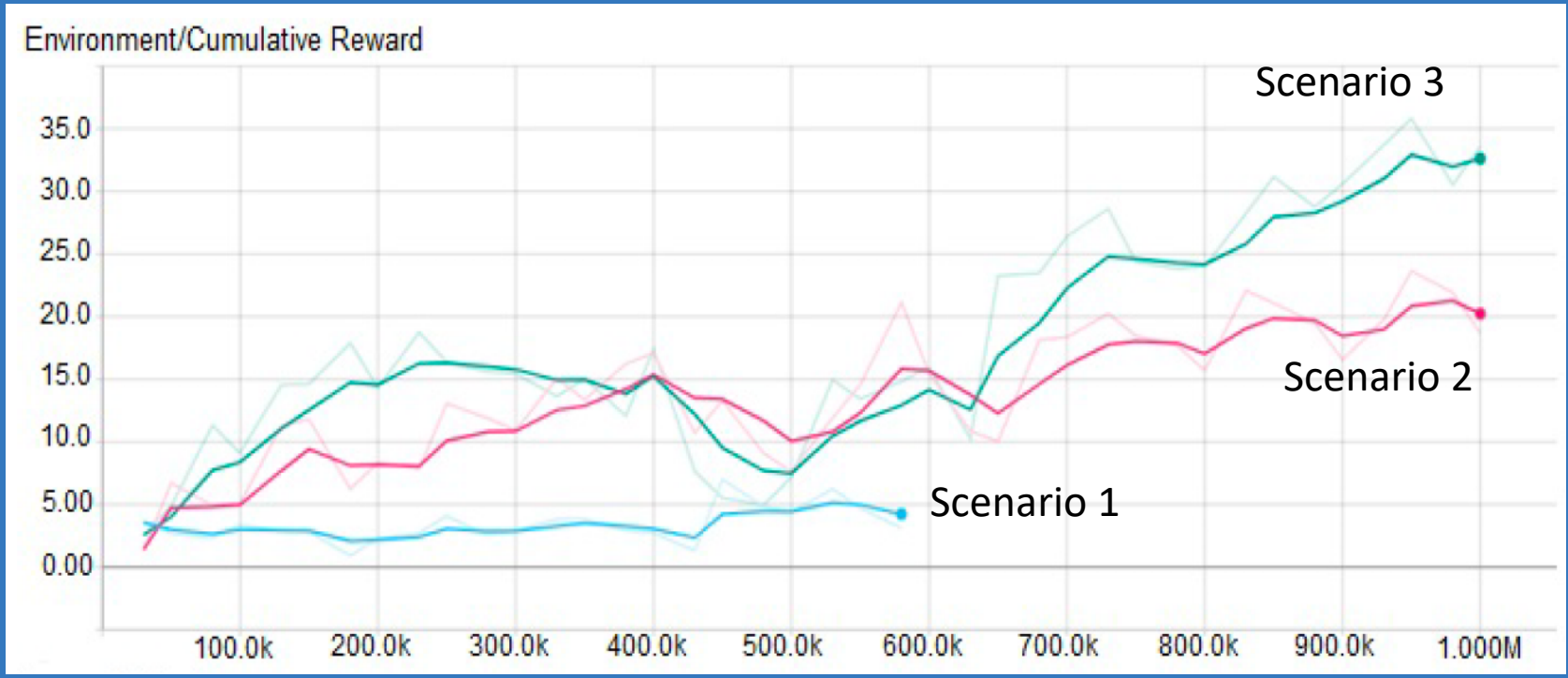
Parameter	Scenario 1 (w/predator)	Scenario 2 (w/predator)	Scenario 3 (wo/predator)
Trainer	PPO	PPO	PPO
Batch_size	1024	1024	1024
Beta	1.0e-2	1.0e-2	1.0e-2
Buffer_size	10240	10240	10240
Epsilon	0.2	0.2	0.2
Hidden_units	128	128	128
Lambd	0.95	0.95	0.95
Learning_rate	3.0e-4	3.0e-4	3.0e-4
Learning_rate_schedule	Linear	Linear	Linear
Max_steps	580000	1.0e6	1.0e6
Memory_size	256	256	256
Normalize	false	false	false
Num_epoch	3	3	3
Num_layers	2	2	2
Time_horizon	64	64	64
Sequence_length	64	64	64
Summary_freq	10000	10000	10000
Use_recurrent	false	false	false
Vis_encode_type	simple	simple	simple
Reward_signals	extrinsic: strength: 1.0, gamma: 0.99	extrinsic: strength: 1.0, gamma: 0.99	extrinsic: strength: 1.0, gamma: 0.99

Table 1: PPO hyper-parameters for all three scenarios of training.

Cumulative reward (training)

The training process of 1 million iterations where the **predator is not present** in the environment, led to a **higher cumulative reward**.

As expected, it seems like prey agents perform better when the predator is out of the picture.



The mean cumulative episode reward over all training scenarios, should increase during a successful training session.
y-axis: average reward, x-axis: number of epochs (training steps).

Experiments (post-training)

- Prior to examining the behaviours learnt by the prey agents operating under PPO, several **experiments** are developed to explore the behavioural capabilities of RL.
- Analysis of the model is conducted through a series of experiments that are analysed in two distinct ways. These include:
  - the **length of time** agents train for (specified in time steps) and,
  - the **stimuli presented** to prey agents during the training phase (in this case the presence or absence of the predator agent) and, in turn, the impact of this stimulus on their subsequent behaviour in the testing phase.
- A **task efficiency** measure was developed to compare the outputs of each experiment model condition. The formula contains: the mean of the total positive points collected, let us call this **PosTotal**, mean of the total negative points, known as **NegTotal**, finally the mean of the total number of times prey are caught **CaughtTotal.**

$$(PosTotal \times 1) + (NegTotal \times -0.2) + (CaughtTotal \times -1)$$

QUANTITATIVE ANALYSIS OF EXPERIMENT RESULTS

Collectively the results of both experiments indicate that within the simulation:

- Agents that **train for longer** are more effective in devising **goal-oriented strategies** (experiment one model condition two, experiment two all conditions).
- Agents that weigh the risks between **multiple penalties** (negative points and being caught by the predator) perform **sub-optimally** in achieving a goal (experiment one, both model conditions) compared to agents that focus solely on a single reward and penalty (experiment two, model condition one).

Experiment 1: exploring the impact of **training length** on task efficiency

Model Condition	1	2
Training Cycles	580k	1 mil
Predator in Training	Present	Present
Predator in Testing	Present	Present
Positive Points	326.88 (34.231)	779.4 (57.434)
Negative Points	320.44 (35.595)	706.88 (53.059)
Caught by Predator	55.48 (15.931)	72.7 (11.784)
Task Efficiency	207.312 (33.835)	565.324 (53.164)

Table 2: Summary of the mean and (std) for each variable including task efficiency measure over all experiment one model conditions.

Experiment 2: exploring the **impact of stimuli** on task efficiency

Model Condition	1	2	3
Training Cycles	1 mil	1 mil	1 mil
Predator in Training	Not Present	Present	Not Present
Predator in Testing	Not Present	Present	Present
Positive Points	1455.18 (111.235)	779.4 (57.434)	1476.62 (122.026)
Negative Points	491.24 (42.519)	706.88 (53.059)	504.98 (44.173)
Caught by Predator	0	72.7 (11.784)	102.08 (14.475)
Task Efficiency	1356.93 (108.803)	565.324 (53.164)	1273.544 (124.072)

Table 4: Summary of the mean and (std) for each variable including task efficiency measure over all experiment two model conditions.

QUALITATIVE ANALYSIS OF BEHAVIOURS

- No approach for analysing agent behaviours existed at the time of writing the paper; therefore, a systematic approach was devised to capture, analyse and interpret the behaviours observed:
- 1. Recording the experiment.
- 2. Play back the experiment recording.
- 3. Watch the movement of the predator agent, to see if it interacts with a prey.
- 4. If the predator interacts with the prey, the video is paused and played frame-by-frame.
- 5. The behaviour of the prey is captured frame-by-frame during the encounter until it has moved away and continues collecting points (known as the idle behaviour).
- 6. The behaviours observed are noted and compared with the quantitative analysis from the Experiments sub-sections to try interpret how the prey agent has behaved.

Observed behaviours

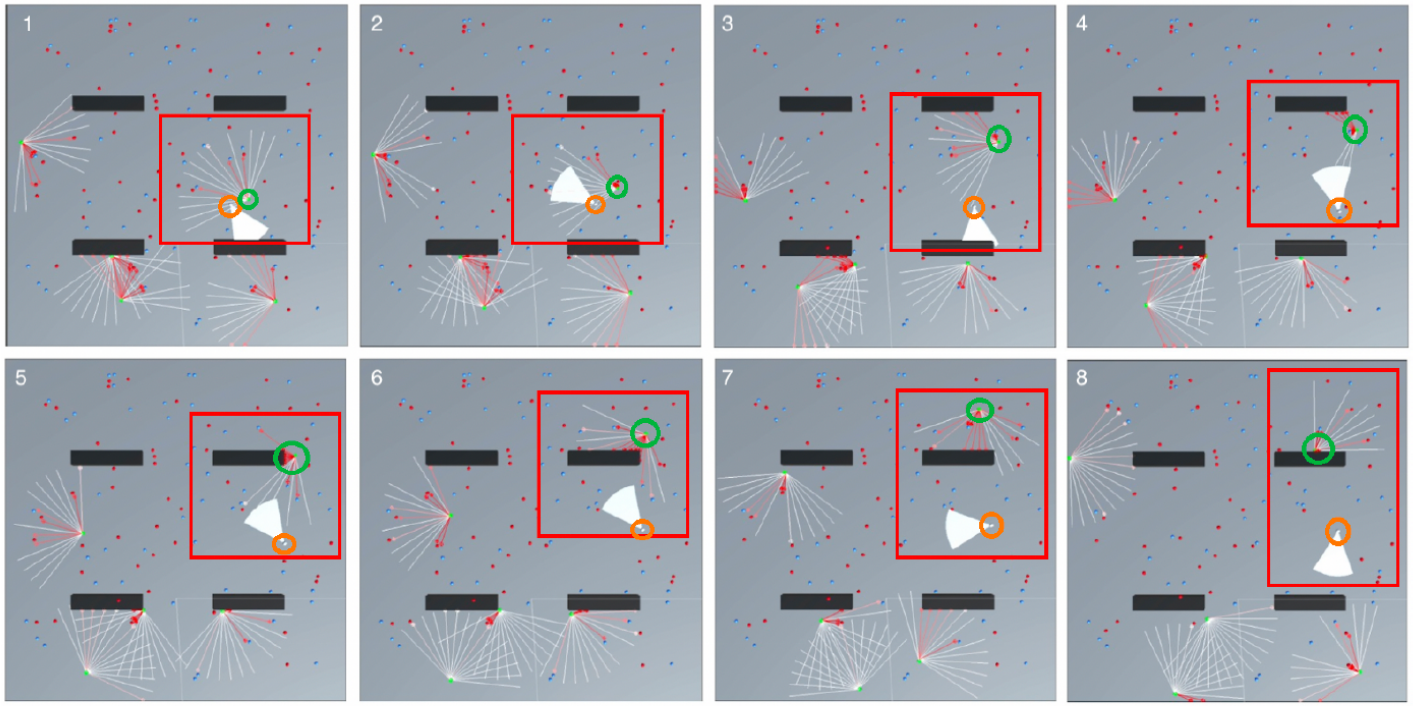
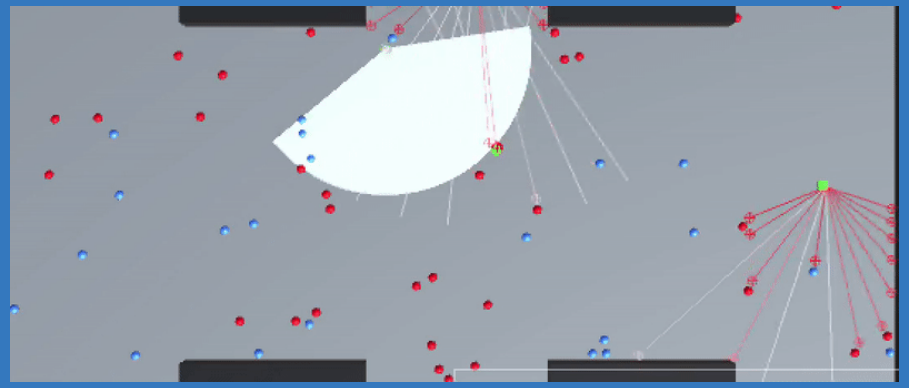


Figure 8: Experiment one, model condition one; a prey agent looking for a wall to hide behind to avoid the predator



Prey agents learn to cooperate in avoiding the Predator

Findings

- The RL technique supports agents in the **behavioural evolution** of intelligent decision making that resemble those observed in the **real-world**.
- Agents that devise policies influenced by **multiple penalties** weigh the impact of these penalties and try to **minimise** the most **impactful** compared to **less impactful**.
- The research highlights the ability for agents to operate under conditions in which they were **never trained to encounter** and continue to perform **relatively well**.
- These attributes that agents develop can be observed in real-world situations, i.e., when people weigh the risks of being captured by the police before attempting a crime, or predatory animals weighing the risks of the prey escaping before deciding to pursue or ignore.

Future research

- Explore the relationships between **people** and **enforced rules** to prevent a contagious virus from moving through the population.
- This research is currently being applied to quantify **driver behaviour** in a **real-world environment** to assess the impact **speed limits** and **traffic rules** have on a **person's decisions** while driving.
- RL can provide researchers with the ability to test various **behaviourally influenced hypotheses** while **minimising the costs** of conducting the research i.e., **collecting data, applying for ethical approval** etc.

Vehicle agent driving around the built environment of San Francisco.