International Journal of Mechanical, Industrial and Aerospace Sciences
ISSN: 2517-9950
Vol:4, No:8, 2010

Biologically Inspired Controller for the Autonomous Navigation of a Mobile Robot in an Evasion Task

Dejanira Araiza-Illan and Tony J. Dodd
Department of Automatic Control and Systems Engineering
University of Sheffield, S1 3JD, UK
{d.araiza,t.j.dood}@sheffield.ac.uk

***Abstract*—A novel biologically inspired controller for the autonomous navigation of a mobile robot in an evasion task is proposed. The controller takes advantage of the environment by calculating a measure of danger and subsequently choosing the parameters of a reinforcement learning based decision process. Two different reinforcement learning algorithms were used: Q-learning and Sarsa (λ). Simulations show that selecting dynamic parameters reduce the time while executing the decision making process, so the robot can obtain a policy to succeed in an escaping task in a realistic time.**

***Index Terms*—Autonomous navigation, mobile robots, reinforcement learning.**

I. Introduction

The use of mobile robots to substitute human beings in high risk tasks has been multiplied due to the increase of flexibility, reliability, robustness, speed, accuracy, and many other advantages that advances in technology have brought. In the last ten years, many researchers have focused their efforts on the analysis of multi-robots with autonomous behaviours. They combine the advantages of multi-agent systems and tools provided by artificial intelligence methods and control analysis [2].

In a dynamic challenging environment, finding hostile conditions (e.g. collapsing structures, opponents) may become a possibility. Therefore, the necessity of designing and implementing robust systems that respond intelligently to these threats and can replace human beings becomes a priority task. A realistic and available alternative is the use of mobile robots. However, this implies several challenges not only in terms of the design and assembly of the hardware, but even more in the control structure.

The complexity of the analysis of a high risk and dynamic challenging environment which includes threats, in order to build realistic and fully useful control systems for autonomous mobile robots, is the main motivation behind this paper.

Animals are often under risky circumstances that can be interpreted as a high risk environment when predators are trying to feed. Many studies have been conducted to identify the behaviours of both prey and predators, including formations (flocking, schooling) [9], irregular changes in movement (protean behaviour) [7], or risk assessment factors [12]. Nevertheless, the possibilities of using this biological knowledge to improve control strategies in mobile robotics result in a broad and interesting area that is explored in the work presented.

The study of pursuit and evasion in general is justified from different points of view. Pursuit and evasion games are common in real life and scientifically interesting to evaluate evolution of behaviour, as well as robust and adaptive behaviours due to their presence in dynamic, stochastic and continuous environments. The broad range of obtained behaviours implies that some areas of knowledge such as behavioural biology, neuroethology and game theory have been focusing their efforts in finding the mechanics and models behind them. Also the applications of pursuit and evasion behaviours have been extended to robotics, video games, virtual environments and multi-agent domains, to mention some examples [10], [16].

In biology, escaping behaviours from various animal species have been deeply reviewed, leading to the identification of three main prey responses: freezing, fleeing and fighting back [4]. Although a random combination of freezing and fleeing is also present in individuals, groups also observe randomness in movement and roles. This kind of erratic behaviour has been named *protean* [7].

In robotics and multi-agent systems, pursuit and evasion behaviours have been targeted from different points of view. The simplest approaches include evasion of static and dynamic obstacles with different resources in terms of acquisition, analysis and processing information, also sometimes trying to find the shortest path with a specific goal [8]. Game theory has studied predator-prey games as a decision making problem where an equilibrium is the goal for competitors - collaborators [6], [11]. Also, machine learning algorithms such as reinforcement learning [14] or simple neural networks [5] have been adapted or evolved under a prey-hunter scenario to obtain optimality in the behaviours that would be closer or even better than biological efficiency. Another approach is the study and simulation of animal evasion behaviours through mobile robots and simulation, such as group evasion dynamics like swarms and schooling, with a clear translation to the automatic control field in protean controllers.

Protean controllers [5] have been developed as artificial neural networks, but not with the objective of achieving a complex system capable of autonomous navigation of a robot, but to apply genetic encoding to the parameters and study the effect of protean behaviours on evolutionary dynamics. The reported results of these studies reflect the importance of randomness translated into noise for the controllers embedded

in the prey, with the objective of achieving a protean behaviour [16]. However, these results have not been extrapolated to the design of an optimal autonomous controller for a robot that can be used in real life tasks such as search and rescue in hostile environments. Multi-agent systems, reinforcement learning and game theory have been focusing their efforts on designing efficient predators, but not efficient prey that can defeat them successfully.

This first attempt at proposing a more realistic control structure is based on two premises. First, it does not try to model or simulate an accurate dynamic high risk environment, as it is not the purpose to study in the control system. The control structure tries to represent the relevant variables with a realistic approach. Second, the priority is the design and analysis of an effective autonomous navigation controller to succeed in the escaping task, in a way likely to imitate animal behaviour. The goal is not trying to find the shortest path through a search algorithm, but emulating biological reasoning.

Biologically inspired behaviours have been incorporated into widely used learning methods and decision making strategies for the design of the controller. Particular interest is devoted to the prey, its reactions as evasion behaviours, and how to use them to control the navigation of an autonomous mobile robot.

Section II describes briefly the preliminary considerations in the design, including the definitions of concepts used in the paper. A brief outline of the design is presented in Section III. Section IV presents the results obtained in a specific scenario used to evaluate the proposed controller. Finally, Section V includes the conclusions.

II. Preliminary considerations

A. The environment

Formal definitions for all the components in the scenario (obstacles, threats, opponents) are stated as follows: an *obstacle* corresponds to any object or entity that blocks the free navigation of a robot in a spatial location. Hostility or *hostile threats* in this case means the presence of an imminent risk (collapsing of structures, mobile obstacles) that will cause an increase in the level of danger. An *opponent* is a human or machine whose objectives are chasing and/or destroying the robots, equivalent to a predator.

B. Quantification of risk

With these three anteriorly defined entities located in the scenario, a method to evaluate their presence and subsequently their immediate importance is proposed. In previous biological-psychological studies, a way to determine the individual fear has been based on the flight initiation distance [12]. The studies show the relevance of factors such as environmental (refuges), experiential (previous experience with predators) and physical characteristics of the predators, in the assessment of risk and therefore the beginning of the flee [12].

The logical extension of animal studies about risk assessment to robotics is impractical as some specific factors mentioned in [12] become impossible to achieve, or irrelevant to evaluate, such as morphological changes in hardware. After a realistic and practical analysis, the only aspects that will be considered relevant for the studies conducted in this paper are the position of the opponent (equivalent of a predator) and estimation of its capabilities, hostile threats in the environment itself, eliminating any possibility of a refuge, and previous experience, seen as an evaluation of the history. Although this aspect will not be covered in the work of this paper, it is intended for the near future.

In order to measure and evaluate the aspects mentioned before, with the purpose of incorporating this knowledge to the design of a more realistic controller, the concept and a mathematical model of *danger* is proposed. This model considers two parts: first, an evaluation of the probable risks that may cause the damage or destruction of the robot considering the history and the actual conditions. As explained before, history is not incorporated in the design. Second, an evaluation of the opponents, which can be simply measured through the proximity, or with a complex model to predict their behaviour.

Danger is defined as a function of three variables: *(i)* the opponents, *(ii)* static obstacles, and *(iii)* dynamic obstacles. Although some attempts have been made in biology to define and measure the risks encountered by a prey due to a predator, in this case distance has been the only measured factor that determines the degree of danger and the space of possible movements that the robot can make in a precise instant of time after a deliberative process.

III. Design

In a hostile, dynamic environment, some resources can be very limited, such as the time for deliberation. With this specific consideration in mind, the design of the controller has been oriented to the adaptability to fast decision making, from measuring the *danger* present in the current environment surrounding a mobile robot. The following subsections describe the aspects that were modified from a reinforcement learning based controller to achieve a faster deliberation, according to a coherent measure of danger.

A. Structure of the controller

The structure of the controller has been divided in three sections for its design: *(i)* the mapping from the environment to a measure that allows shaping parameters in order to achieve more efficiency of resources while deliberating; *(ii)* the learning method, in this case a reinforcement learning algorithm; and *(iii)* the biological influence in both the selection of the action and the learning process.

B. Reinforcement learning

Reinforcement learning is based on a mapping from states of the environment to actions after a Markov decision making process. An agent has to discover by itself which action is optimal through the reward (or reinforcement signal) that it gets when choosing and performing the action. At the end of

the process, the agent has an optimal policy or set of actions [13].

In reinforcement learning, the goal for the agent i is to maximise the discounted reward of executing a determined action a_i and causing a transition of the state, at each time step k. In order to choose the optimal action in a reinforcement learning mechanism, the agent computes the value of the action or the value of the pair action-state (Q value) measured in the terms of the reward, using Bellman equations [15].

C. Q-learning and Sarsa (λ)

Two different popular reinforcement learning algorithms were used to obtain a comparison in performance for the specific selected task with the incorporation of the biologically inspired module: Q-learning and Sarsa (λ) [13] with radial basis functions neural networks.

In the Q-learning algorithm described in detail in [13], the Q values are updated at a certain learning rate α from a discounted reward R:

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha(R + \gamma \max_{a'} Q(s',a'), \quad (1)$$

where γ is the discount rate, and $Q(s',a')$ is the Q value for the actions of next predicted state.

The Sarsa (λ) algorithm [13] updates a series of parameters $\vec{\theta}$ from a bootstrap of λ steps of past history evaluated with an eligibility trace $\vec{e}$, and an error δ:

$$\vec{\theta} \leftarrow \vec{\theta} + \alpha\delta\vec{e}, \quad (2)$$

$$\delta \leftarrow \delta + \gamma Q(s',a'). \quad (3)$$

A radial basis functions neural network (RBFNN) has been used to approximate the Q values in the Sarsa algorithm as well as the reward function R. A RBFNN [3] with inputs, $\vec{x}$, and outputs, y_k, is defined by a relation between weights and M radial basis functions

$$y_k = \sum_{j=1}^{M} w_{kj}\phi_j(\vec{x}) + w_{k0}, \quad (4)$$

$$\phi_j(\vec{x}) = exp\left(-\frac{\|\vec{x} - \vec{\mu}_j\|^2}{2\sigma_j^2}\right), \quad (5)$$

where $\vec{\mu}_j$ are the centers and σ_j are the parameters for each j gaussian radial basis function.

In the escaping task, several implications in a reinforcement learning scheme can be observed due to the necessity of a faster deliberation. First, the compromise between exploration and exploitation needs to be changed favouring the second to look for the best immediate solution. Second, the learning rate and the chances to run the learning algorithm to find the optimal solution (number of iterations, number of episodes) need to be adjusted to reduce the time. Finally, the bootstrapping has to be bounded to a reasonable measure to facilitate the processing.

D. Immediate decisions

In order to adjust the parameters of the reinforcement learning method, a numerical determination of danger is needed. Danger shortens the time for decision making, and it is defined inversely proportional to the distance r_i in centimeters from the robot to any of the factors of risk in the environment: opponents and obstacles (expressed as a total n)

$$D = \sum_{i=1}^{n} \frac{d_i}{n} \quad (6)$$

$$d_i = \frac{1}{r_i}, r_i > 1. \quad (7)$$

The function D has a maximum value of 1.

E. Biologically inspired module

The presence of an identified threat triggers several kinds of escaping behaviours in animals [4], where a decision has to be made as fast as possible. This necessity of immediate response has been modelled from a measure of the danger, by shaping the parameters of the algorithm to make it faster and more precise versus exploring alternatives. The use of dynamic reinforcement learning parameters such as the learning rate or the discount rate has been proposed in few articles like [1]. However, typically a fixed value is taken as enough, and the number of iterations is not important when trying to look for the best optimal solution in long time. In this case, functions for the parameters have been proposed

$$\alpha = 0.01\exp(3.219D), \quad (8)$$

$$\gamma = 0.6\exp(0.511D), \quad (9)$$

$$Iterations = 300\exp(-3.401D), \quad (10)$$

$$\epsilon = 0.5\exp(-1.609D). \quad (11)$$

The bootstrapping parameter λ in the Sarsa algorithm is fixed.

The biologically inspired module so far is in charge of measuring the danger and choosing a set of parameters to perform the decision making process. Besides, the module creates the map for the rewards in order to suggest not to collide against obstacles and opponents, which can be observed in the results. In the future, protean behaviours are intended to be implemented as part of the autonomous navigation to complete the module.

F. Deliberation process

The controller containing the three sections mentioned before executes the following deliberation process:

1) Reading the environment, which is considered static at time t to obtain the best actions to follow according to the circumstances.
2) Mapping the environment into probabilities of colliding against obstacles and opponents, where a probability of 1 denotes the location of an obstacle or opponent.

3) Selecting a reward function R and the parameters: learning rate α, discount rate γ, number of iterations and probability for exploration-exploitation ϵ.
4) Performing a reinforcement learning method with the parameters obtained before.
5) Obtaining the best policy corresponding to the best set of actions for the static scenario, which is true when the scenario has not changed significantly.
6) If the scenario changes, then a new learning process to obtain a new set of actions is needed, and the process starts again.

IV. Results

A square scenario of 10x10 meters with obstacles and opponents randomly distributed was used as the test bed in the evaluation. An example is shown in Figure 1. The robot is located in an initial position at the centre of the grid (cell $x = 5$, $y = 5$), and it has to navigate to a target point defined for the scenario in Figure 1 at cell $x = 5$, $y = 9$. For a discrete set of tests, the scenario is divided with a grid of 1x1 meters. The robot is allowed to move to 9 different positions: towards, forwards, right, left, towards right, towards left, forwards right, forwards left and stay in the same place.

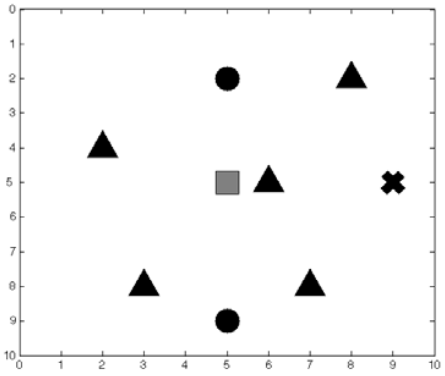

Fig. 1. Example of scenario used for the tests, where triangles correspond to fixed obstacles, circles to opponents, a square to the initial position, and a cross to the target.

From the environment shown in Figure 1, a map of rewards was obtained after analysing the probabilities of colliding against obstacles or opponents when the robot moves through different positions in the grid. The resulting maps are presented in Figure 2.

A run of the Q-learning or Sarsa (λ) corresponds to a set of iterations where the parameter for exploration-exploitation, ϵ, is decreased linearly by $\epsilon/iterations$. 100 Monte Carlo runs of each reinforcement learning algorithm were made with the fixed parameters shown in Table I. With this, the variation of the mean in the average of needed movements to go from the initial position to the target was obtained as a measure for exploitation. A map of the average of visits to each cell was also computed as a measure of exploration to identify the most visited regions in the grid.

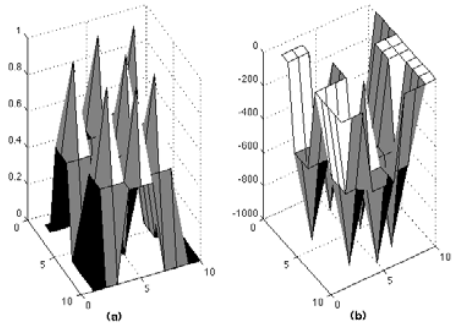

Fig. 2. (a) Map of tabular probabilities obtained after analysing the scenario shown in Figure 1, and (b) its corresponding map of rewards.

TABLE I
Fixed parameters

PARAMETER	FIXED VALUE
Learning rate α	0.1
Discount rate γ	0.7
Iterations	100
Bootstrap λ for Sarsa	0.5
Initial exploration-exploitation probability ϵ	0.5
Radial basis functions for Sarsa	30

Figure 3 shows the exploitation, with a mean of 188.4497 movements to reach the target for the Q-learning and 187.4431 movements for the Sarsa, and a standard deviation of 18.1919 and 27.0804 respectively. Figure 4 shows a map of the average of visits to each cell, where it is possible to observe the similarity in the exploration of both methods.

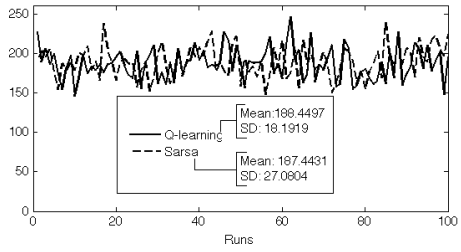

Fig. 3. Average of movements in each run for Q-learning and Sarsa(λ).

Once incorporating the selection of parameters based on the measure of danger, several experiments were conduced as before, with 50 Monte Carlo runs over each set of iterations for both methods Q-learning and Sarsa. This time, one of the four dynamic parameters was adjusted with the equations (8) to (11) and the other three fixed in the sets shown in Table II. The measure of danger was variated from 0 to 1 to observe the effect of the adjustment of each one of the parameters separately from the others in the exploitation or average of movements from an initial position to the target.

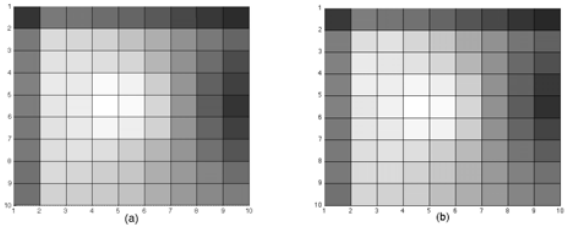

Fig. 4. Map of exploration for *(a)* Q-learning and *(b)*Sarsa(λ), where the darker the cell the lower number of visits.

The expected results were: *(i)* reducing the time of deliberation when increasing danger, *(ii)* a bounded mean in the number of movements needed to go from the initial position to the target similar to the obtained with parameters of Table I, and *(iii)* a standard deviation closer to the shown in Figure 3 when adjusting the parameters for a faster deliberation.

TABLE II
PARAMETERS FOR EVALUATION OF THE EFFECT OF DANGER

PARAMETER	SET 1	SET 2
Learning rate α	0.07	0.19
Discount rate γ	0.7	0.9
Iterations	227	73
Initial exploration-exploitation probability ϵ	0.4	0.2

Figures 5, 6, 7 and 8 show the mean and standard deviation for the variation of the learning rate α, the discount rate γ, the number of iterations, and the exploitation-exploration parameter ϵ, respectively. When variating α, γ and ϵ the mean remained bounded with some oscillations, not the case of the variation in the number of iterations, where a great oscillation is present for danger $= [0.5, 1]$. The standard deviation also remained bounded with small oscillations, excepting when variating the number of iterations again, where its growth is obvious.

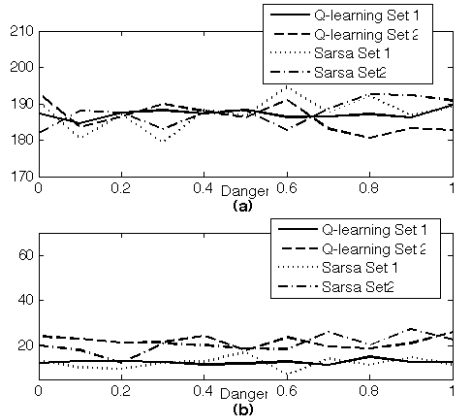

Fig. 5. *(a)* Mean and *(b)* standard deviation of the movements when varying α and combinations of two sets of fixed parameters shown in Table II

Finally, Figure 9 shows the behaviour of the mean and standard deviation when variating the danger and using equations (8) to (11) for all the parameters. According to Figure 9, the time of deliberation is reduced, but the standard deviation increases significantly. Although the results stay bounded, a larger standard deviation introduces an error in the obtained policy, but it is a trade off when a faster deliberation is needed.

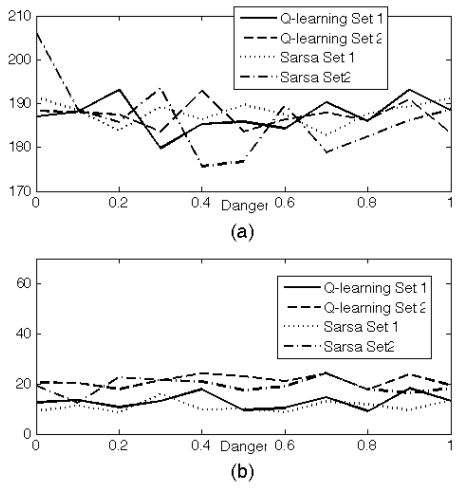

Fig. 6. *(a)* Mean and *(b)* standard deviation of the movements when varying γ and combinations of two sets of fixed parameters shown in Table II

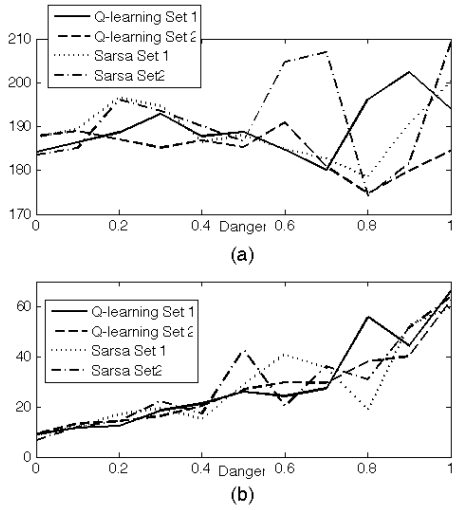

Fig. 7. *(a)* Mean and *(b)* standard deviation of the movements when varying the number of iterations and combinations of two sets of fixed parameters shown in Table II

V. CONCLUSION

A novel controller that incorporates biologically inspired behaviours has been proposed and explained. As mentioned before, the necessity of realistic algorithms to be used in real life demanding tasks such as the one presented here introduces constraints in terms of using as less time as possible

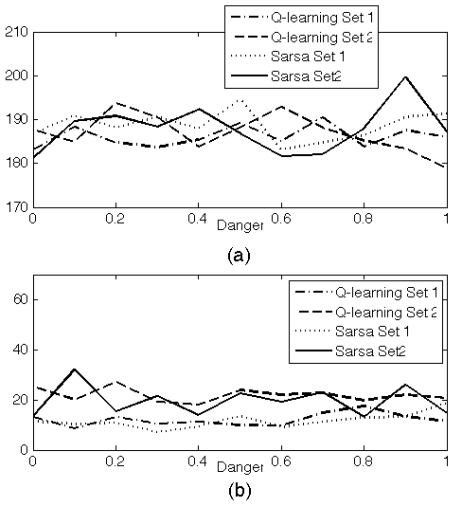

Fig. 8. *(a)* Mean and *(b)* standard deviation of the movements when varying ϵ and combinations of two sets of fixed parameters shown in Table II

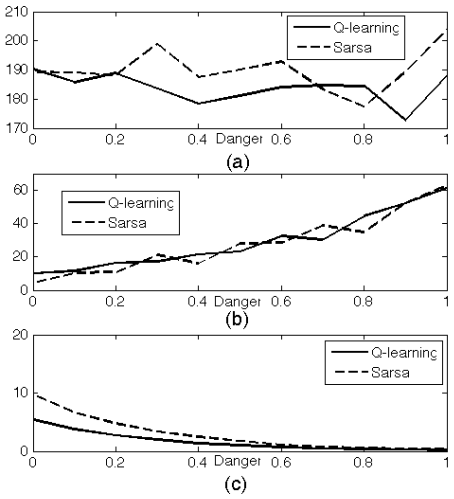

Fig. 9. *(a)* Mean and *(b)* standard deviation of the movements, and *(c)* mean of deliberation time (seconds), when varying all the parameters according to danger.

to give more emphasis to the reaction, without being a reactive controller. A biologically inspired module has been added to commonly used reinforcement learning algorithms to find the best policy, in order to modify the parameters such as the learning rate and find a solution in less time.

The results show that the biological module works successfully as expected, decreasing the total time of processing as a result of an increase in danger but with some associated costs (e.g. an increase in standard deviation, presence of oscillations). Therefore, a tradeoff between the costs has to be reached according to the possibilities allowed by the task.

Further work is planned to analyse the effect of the resolution of the grid in the measure of danger to formulate a balance between the precision of the movements and the complexity of the deliberation process in terms of hardware. Moreover, the use of artificial potential fields in the reward functions is being contemplated for future comparisons between the effect of the rewards obtained through a function of danger and some others based on potentials.

REFERENCES

[1] M.R. Akbarzadeh, H. Rezaei, and M.B. Naghibi. A fuzzy adaptive algorithm for expertness based cooperative learning, application to herdin problem. In *Proceedings of the 22nd International Conference on the North American Fuzzy Information Processing Society*, pages 317–322, 2003.

[2] T. Arai, E. Pagello, and L.E. Parker. Advances in multi-robot systems. In *IEEE Transactions on Robotics and Automation*, volume 18, pages 655–661, 2002.

[3] C.M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, 1995.

[4] S. Edut and D. Eilam. Protean behaviour under barn-owl attack: voles alternate between freezing and fleeing and spiny mice flee in alternating patters. *Behavioural Brain Research*, 155:207–216, 2004.

[5] D. Floreano and S. Nolfi. Adaptive behavior in competing co-evolving species. In *Proceedings of the fourth European Conference on Artificial Life*, pages 378–387. MIT Press, 1997.

[6] J.P. Hespanha, M. Prandini, and S. Sastry. Probabilistic pursuit-evasion games: A one-step nash approach. In *Proceedings of the 39th IEEE Conference on Decision and Control*, pages 2432–2437, 2000.

[7] D.A. Humphries and P.M. Driver. Protean defence by prey animals. *Oecologia*, 5:285–302, 1970.

[8] C. Laugier and R. Chatila, editors. *Autonomous navigation in dynamic environments*. Springer Berlin / Heidelberg, 2007.

[9] S.W. Lee. A bio-inspired group evasion behaviour. Technical report, Department of Computer Science, The University of North Carolina at Chapel Hill, 2008.

[10] G.F. Miller and D. Cliff. Co-evolution of pursuit and evasion I: biological and game-theoretic foundations. Technical Report CSRP311, School of Cognitive and Computing Sciences, University of Sussex, 1994.

[11] B. Scherrer and F. Charpillet. Cooperative co-learning: A model-based approach for solving multi-agent reinforcement problems. In *Proceedings of the 14th International Conference on Tools with Artificial Intelligence*, pages 463–468. IEEE Computer Society, 2002.

[12] T. Stankowich and D.T. Blumstein. Fear in animals: a meta-analysis and review of risk assessment. *Proceedings B*, 272(1581):2627–2634, 2005.

[13] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 1998.

[14] H. Tamakoshi and S. Ishii. Multi-agent reinforcement learning applied to a chase problem in a continuous world. *Artificial Life Robotics*, 5:202–206, 2001.

[15] N. Vlassis. *A concise introduction to multi-agent systems and distributed artificial intelligence*. Morgan & Claypool, 2007.

[16] M. Wahde and M.G. Nordahl. Evolution of protean behavior in pursuit-evasion contests. In *Proceedings of the fifth International Conference on Simulation of Adaptive Behavior on From animals to animats 5*, pages 557–561. MIT Press, 1998.