# Reward Functions For Agent-Based Ecosystem Modeling

Claes Strannegård[1,2] and Niklas Engsner[2]

[1]Dept. of Applied Information Technology, University of Gothenburg, Sweden
[2]Dept. of Molecular Medicine and Surgery, Karolinska Institutet, Sweden
claes.strannegard@gu.se

*Agent-based models* (Macal and North, 2010) are used for studying agents that interact in a shared environment. They are widely used in science (Rollins et al., 2014), for example in ecosystem modeling, where animals of different species interact with other organisms and the physical environment (DeAngelis and Grimm, 2014). A common practice in agent-based ecosystem modeling is to hand-code models of animal behavior (Jadhav et al., 2024; Widyastuti et al., 2022). However, this is challenging because the mechanisms of animal behavior are highly complex and not fully understood (Tohyama et al., 2025). An alternative approach is to use *reinforcement learning* (RL) (Sutton and Barto, 2018), in particular *multi-agent reinforcement learning* (MARL), which focuses precisely on multiple agents interacting in a shared environment (Zhang et al., 2021). MARL has been used to model animal behavior in predator–prey dynamics (Yang et al., 2017; Sunehag et al., 2019) and in ecosystem modeling (Strannegård et al., 2025). Reinforcement learning mechanisms have been observed throughout the animal kingdom (Neftci and Averbeck, 2019; Barron et al., 2010). Typically, a reward signal is computed by the brain's reward system, which integrates several internal and external cues into signals controlling the release of dopamine and other neurotransmitters associated with learning (Arias-Carrión et al., 2010). A natural strategy in RL—forming the basis of *homeostatic* RL—is therefore to model natural reward signals by defining functions that convert homeostatic signals into scalar rewards (Yoshida et al., 2024a). While animals use RL to learn during their lives, many artificial agents only learn before deployment. Here, we study how the choice of reward function influences the performance of MARL-based ecosystem models from three perspectives: animal lifespan, population size, and resilience, defined as the ability of a set of species to coexist under environmental change.

## Ecosystem environment

We conducted experiments with different reward functions using an environment defined by the agent-based ecosystem model in (Strannegård et al., 2025), representing an Alpine landscape in Italy with wolves, chamois (mountain goats), and three species of grass. The geographical area is represented by a grid of cells, where each cell has its own populations of organisms and its own land-cover class. The animal agents can move, eat, drink, and reproduce under certain conditions. They are all mortal and can die from starvation, thirst, predation, or old age (after 500 time steps). The behavior of each animal species is controlled by a neural network (policy network), which determines what each individual agent will do at each time step. The observations and actions of each chamois or wolf agent are as follows:

**Observations:** The agent observes its own internal levels of energy and hydration, both in the range $[0, 1]$, and its own species (chamois or wolf). It also observes the properties of the cells in the $3\times3$ neighborhood surrounding it. Thus, it is aware of the presence of grass, water, chamois, and wolves in its nearby environment. It also senses aggregated smells of these objects in the same cells, informing it about distant objects.

**Actions:** The agent eats and drinks automatically when in a cell where food or water is present. It moves as specified by three output values of the policy network: a direction given by two coordinates, and a speed.

The policy networks were trained across a range of environments to enable the agents to survive in multiple environment that they have never encountered before.

## Reward functions

The following reward functions were used during training, cf. Konidaris and Barto (2006) and Yoshida et al. (2024a,b):

***Classic*:** Reward for eating or drinking: chamois gain reward for eating grass or drinking water; wolves for eating a chamois or drinking water.

***Survival*:** Reward +1 at each time step as long as the agent remains alive.

***Euclidean*:** Negative Euclidean distance between the present homeostatic state and an ideal setpoint $(1, 1)$.

***Delta Euclidean*:** Change in Euclidean reward from one time step to the next.

***Homeostatic*:** Sum of the levels of energy and hydration.

***Delta Homeostatic*:** Change in homeostatic reward from one time step to the next.

***Wellness*:** Capped quadratic function of energy and hydration that provides no additional reward when both exceed 0.8 (satiation).

## Method

We used the RL algorithm PPO from Stable-Baselines3 (Raffin et al., 2021) and the seven reward functions described above to train a total of 49 policy networks, each with 144 inputs, two hidden layers, and three outputs. Since randomness affects the training process, we trained seven policy networks per reward function. All policy networks were trained on a large number of randomly generated Perlin worlds of varying difficulty. To evaluate the policy networks, we used a test set consisting of 190 environments. By varying the amount of grass in the original model, we produced ten environments at each of 19 levels of difficulty, ranging from simple to impossibly hard. All environments were populated with 100 chamois and 25 wolves randomly distributed on the map. Each of the 49 policy networks was run once on each environment, for a total of $49 \times 190 = 9310$ simulations. Each simulation lasted until either the chamois or the wolves died out, or for a maximum of 4000 steps. The Python code for the project is available at https://gitlab.com/ecotwin/alife-2025-workshop.

## Results

To test whether differences between reward functions were statistically significant ($p < 0.05$), we applied the non-parametric Friedman test (Friedman, 1937), which ranks functions by average performance (1 = best, 7 = worst). Results are shown in critical difference diagrams (e.g., Fig. 1), where a red bar between two reward functions indicates no significant difference, and its absence indicates a significant one.

**Average lifespan** The average lifespan for wolves ranged from 140 to 162 time steps across the seven rewards, and for chamois from 64 to 70. For chamois, Wellness and Delta Euclidean produced the longest lifespans, while for wolves it was Survival, Wellness, and Euclidean (Fig. 1).

**Population size** The average number of wolves was 13–15 across reward functions, while chamois populations ranged from 170 to 200. Classic yielded the largest chamois populations, whereas Homeostatic and Wellness favored larger wolf populations (Fig. 2).

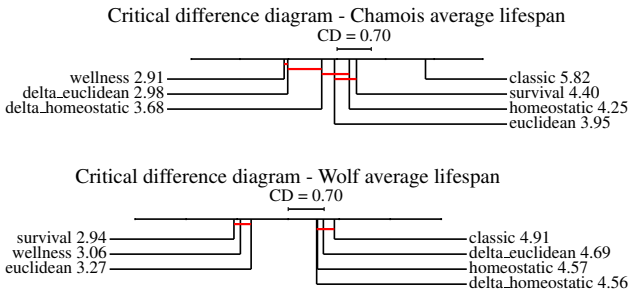


Figure 1: Average lifespans of chamois and wolves.

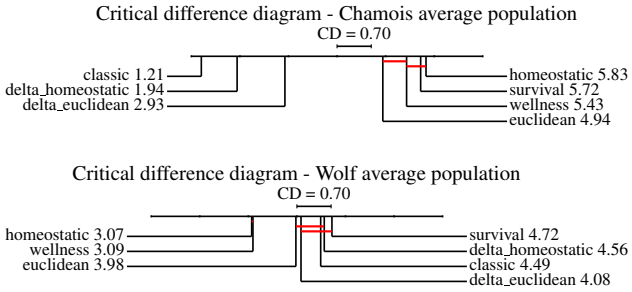


Figure 2: Population sizes of chamois and wolves.

**Resilience** Resilience (average episode length) ranged from 1526 to 2151 time steps. Homeostatic, Wellness, and Euclidean led to the most resilient animal populations (Fig. 3).

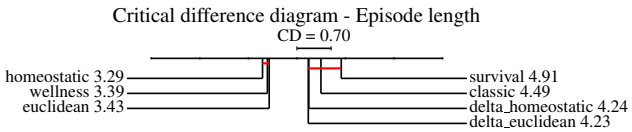


Figure 3: Resilience (ability of the species to coexist).

## Conclusion

Several statistically significant differences were detected between the reward functions, but none emerged as a clear overall winner or loser. All seven reward functions, except *Survival*, could potentially benefit from reward shaping—for example, by adjusting the relative weights assigned to energy and hydration. In contrast, *Survival* can be applied without modification to construct behavioral models of diverse animal species. Agents optimized for survival are, *a fortiori*, also optimized for locating food and water, evading predators, conserving energy, and navigating terrain.

## Acknowledgements

This research was supported by the Sten A Olsson Foundation for Research and Culture and Stiftelsen Sävstaholm. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725. ChatGPT was used for language improvement.

## References

Arias-Carrión, O., Stamelou, M., Murillo-Rodríguez, E., Menéndez-González, M., and Pöppel, E. (2010). Dopaminergic reward system: a short integrative review. *International archives of medicine*, 3(1):24.

Barron, A. B., Søvik, E., and Cornish, J. L. (2010). The roles of dopamine and related compounds in reward-seeking behavior across animal phyla. *Frontiers in behavioral neuroscience*, 4:163.

DeAngelis, D. L. and Grimm, V. (2014). Individual-based models in ecology after four decades. *F1000Prime reports*, 6.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701.

Jadhav, V., Pasqua, R., Zanon, C., Roy, M., Tredan, G., Bon, R., Guttal, V., and Theraulaz, G. (2024). Collective responses of flocking sheep (ovis aries) to a herding dog (border collie). *Communications Biology*, 7(1):1543.

Konidaris, G. and Barto, A. (2006). An adaptive robot motivational system. In *International conference on simulation of adaptive behavior*, pages 346–356. Springer.

Macal, C. M. and North, M. J. (2010). Tutorial on agent-based modelling and simulation. *Journal of Simulation*, 4(3):151–162.

Neftci, E. O. and Averbeck, B. B. (2019). Reinforcement learning in artificial and biological systems. *Nature Machine Intelligence*, 1(3):133–143.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. (2021). Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8.

Rollins, N. D., Barton, C. M., Bergin, S., Janssen, M. A., and Lee, A. (2014). A computational model library for publishing model documentation and code. *Environmental Modelling & Software*, 61:59–64.

Strannegård, C., Palak, M., Engsner, N., Stocco, A., Antonelli, A., and Silvestro, D. (2025). Predicting ecosystem resilience using multi-agent reinforcement learning. *bioRxiv*.

Sunehag, P., Lever, G., Liu, S., Merel, J., Heess, N., Leibo, J. Z., Hughes, E., Eccles, T., and Graepel, T. (2019). Reinforcement learning agents acquire flocking and symbiotic behaviour in simulated ecosystems. In *Artificial life conference proceedings*, pages 103–110. MIT Press.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tohyama, T., Nagashima, U., Higashino, I., Arima, F., Hiyoshi, K., Nagase, M., Yada, Y., Honda, N., and Watabe, A. M. (2025). Aversive experiences induce valence plasticity of instructive signals to change future learning rules in mice. *Communications Biology*, 8:1002.

Widyastuti, K., Reuillon, R., Chapron, P., Abdussalam, W., Nasir, D., Harrison, M. E., Morrogh-Bernard, H., Imron, M. A., and Berger, U. (2022). Assessing the impact of forest structure disturbances on the arboreal movement and energetics of orangutans—an agent-based modeling approach. *Frontiers in Ecology and Evolution*, 10:983337.

Yang, Y., Yu, L., Bai, Y., Wang, J., Zhang, W., Wen, Y., and Yu, Y. (2017). A study of AI population dynamics with million-agent reinforcement learning. *arXiv preprint arXiv:1709.04511*.

Yoshida, N., Arikawa, E., Kanazawa, H., and Kuniyoshi, Y. (2024a). Modeling long-term nutritional behaviors using deep homeostatic reinforcement learning. *PNAS nexus*, 3(12):pgae540.

Yoshida, N., Daikoku, T., Nagai, Y., and Kuniyoshi, Y. (2024b). Emergence of integrated behaviors through direct optimization for homeostasis. *Neural Networks*, 177:106379.

Zhang, K., Yang, Z., and Başar, T. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. In Chowdhury, S. R., Jiang, X., and Yang, Z., editors, *Handbook of Reinforcement Learning and Control*, volume 310 of *Studies in Systems, Decision and Control*, pages 321–384. Springer, Cham.