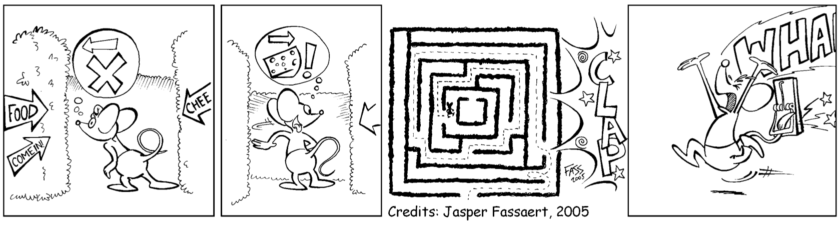
Internal Simulation of Behavior has an Adaptive Advantage

Joost Broekens
Leiden Institute of Advanced Computer Science (*LIACS*), Leiden University, The Netherlands
email: broekens@liacs.nl

Abstract: In this paper we test the hypothesis that internal simulation of behavior has a robust adaptive (learning) advantage. From an evolutionary perspective, it seems plausible that agents that simulate behavior have additional survival perspective compared to those that do not, at least if internal simulation is an advantageous hereditary feature. To test this, we present experimental results with a computational model of learning and decision-making. Our experiments are based on biasing the agent's action-selection by a simulation of future interactions. Using our model, we show that this influence of simulation on learning results in a significant learning advantage. If we assume that this increased individual adaptation is an advantageous hereditary feature, this is a relevant result for the evolutionary plausibility of the simulation hypothesis, specifically since our simulation mechanism mainly uses the agent’s existing processing mechanisms.

Keywords: action-selection; adaptive agent; reinforcement learning; simulation hypothesis; computational model.

Hypothesis: action-selection biased by internal simulation enhances learning



Credits: Jasper Fassaert, 2005

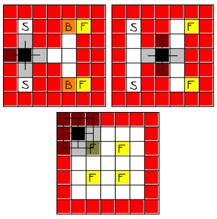
The *simulation hypothesis* (Hesslow, 2002) states that thinking consists of internal simulation of interaction with the environment. An important aspect is evolutionary continuity: no large gap between those agents that simulate and those that do not. Agents that internally simulate behavior

1.) use existing evaluation mechanisms to evaluate imagined future states, and

2.) use existing sensory-motor processes to drive internal simulation.

Is an agent that uses internal simulation to bias its action-selection more adaptive?

Method/Tasks: a simulated agent that learns to gather food in a gridworld



We have studied the influence of simulation on the learning performance of an adaptive agent for a variety of parameters (agent task, learning rate, amount of simulation). Our experiments are based on an agent whose brain is a computational model. Our agent lives in gridworlds in which it must autonomously learn to forage. Simulation consists of:

1.) a small amount of anticipatory simulated interaction with the gridworld, concurrent with the agent’s reactive mode of operation.

2.) a bias of the agent's action-selection induced by this anticipation

Learning and Action: action-selection based on a hierarchical-state prediction of expected benefit—learned through continuous interaction.

The agent's memory structure is modelled by a directed graph. The memory is adapted while the agent interacts with its environment and builds a predictive interaction-based model (cf. Bickhard, 1998):

1.) Select an action *a,* execute *a* and combine with resulting perception *p* into a situation s_1=<*a, p*>. Add *s* to the memory if *s* does not yet exists.

2.) Do another action, resulting in s_2. Add s2 and connect s_1 to s_2 by creating an interactron node I_1.

3.) Recursively apply this process (use *I*s to predict situations).

Continuous interaction results in a predictive stochastic model of the environment that conditionally predicts potential interactions.

Use RL principles to estimate the expected benefit of interactions (Sutton and Barto, 1998).

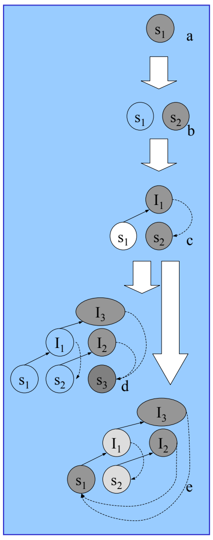
Action-selection is based on these predicted potential benefits.

$$\mathrm{P}(x \mid y) = \upsilon_x \Big/ \sum_{i=1}^{|X_y|} \upsilon_{x_i}$$

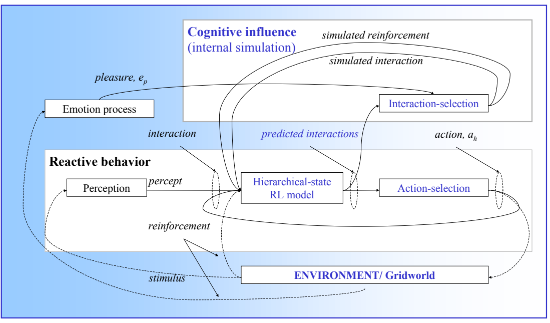
$$\lambda^{t+1}{}_y = \lambda^t{}_y + (r^t - \lambda^t{}_y) \times \rho$$

$$\nu^{t+1}{}_y = \sum_{i=1}^{|X_y|} \mu^t(x_i \mid y) \times \mathrm{P}(x_i \mid y)$$

$$l^t(a_h) = \sum_{i=1}^{k} \sum_{j=1}^{|X_{y_i}|}{}^{*} \mu^t(x^i{}_j \mid y_i) \times \mathrm{P}(x^i{}_j \mid y_i)$$



Simulation of Behavior: select and simulate a subset of predicted interactions, bias expected benefit accordingly, use these biased benefits for action-selection



The process of internal simulation consists of the following 5 steps:

1.) *Interaction-selection*: select a subset of predicted interactions.

2.) *Simulate-and-bias-predicted-benefit*: send the subset of selected interactions to the model as if they were real interactions (including predicted benefit). The memory advances to time *t+1*.

3.) *Reset-memory-state*: to be able to select an appropriate action, reset the memory's state to the previous timestep, i.e., time *t*.

4.) *Action-selection*: select the next action using the action-selection mechanism described above. Thus, the propagated markers of the simulated predicted interactions directly bias action-selection. Our anticipation mechanism is best understood as *state anticipation* (Butz *et al*, 2003).

5.) *Reset-biased-predicted-benefit*: reset μ, λ and ν of the interactions that were changed at step 2 (simulation) to the values of μ, λ and ν of these interactions before step 2.

We have investigated the effect on learning for four selection criteria (simulation strategies / selection thresholds). Step 1 either selects;

a.) nothing, no simulation (NON).

b.) the predicted best interaction (BEST). Step 2 simulates the winning interactions of WTA selection (Step 1). Real interactions are accompanied by a reinforcement signal. Now we simulate, and thus lack such a signal. Instead, this signal is simulated by using the predicted benefit, μ, of the winning interaction as reinforcement. We simulate the predicted interaction and its associated value.

c.) the predicted 50% best interactions (BEST50).

d.) all of the predicted interactions (ALL).

Results: the hypothesis holds for a variety of settings.

In general (BEST/BEST50/ALL), the influence of simulation on learning, by providing an extra action-selection bias, has a significant learning advantage even if the agent simulates just one step ahead.

1.) The ALL simulation strategy has a robust adaptive advantage compared to the other strategies. ALL is either among the best-performance strategies or there is no difference between strategies at all. The effect is produced by the fact that ALL either converges faster or better (or both).

2.) This positive effect occurs in three different learning tasks, and for a variety of learning rates as well as rates of forgetting.

3.) If increased individual adaptation is an hereditary advantageous feature, this is a relevant result for the evolutionary plausibility of the simulation hypothesis.

References

Bickhard. M.H. (1998). *JETAI, 10*, pp. 179-215.
Broekens J. & Verbeek, F.J. (2005). In: *Proc. MNAS Workshop*, in press
Butz, M.V. Sigaud, O. & Gerard, P. (2003). *In: LNAI 2684*, pp.86-109
Hesslow, G. (2002), *TICS 6,* pp. 242-247.
Sutton, R.S. & Barto, A.G. (1998). Reinforcement Learning.