# On the Advantages of Non-Cooperative Behavior in Agent Populations

Alexander Pudmenzky ⋆

*School of Land and Food Sciences, The University of Queensland, Brisbane 4072, Australia*

**Abstract**

We investigate the amount of cooperation between agents in a population during reward collection that is required to minimize the overall collection time. In our computer simulation agents have the option to broadcast the position of a reward to neighboring agents with a normally distributed certainty. We modify the standard deviation of this certainty to investigate its optimum setting for a varying number of agents and rewards. Results reveal that an optimum exists and that (a) the collection time and the number of agents and (b) the collection time and the number of rewards, follow a power law relationship under optimum conditions. We suggest that the standard deviation can be self-tuned via a feedback loop and list some examples from nature were we believe this self-tuning to take place.

*Key words:* agent population, co-operation, reward collection, armed bandit search, optimum standard deviation, exploitation and exploration

## 1 Introduction

We investigate the behavior of a collective systems of agents that are able to communicate locally. The aim is to analyze how communication between agents can be optimized to fulfill a larger common goal such as the minimization of time taken to search for and collect randomly distributed rewards. We begin with a definition of the terminology we will be using in section 2 which enables us to formulate a generic description of the problem in section 3. Section 4 then introduces the parameter values we investigate followed by presentation of the results in section 5. The discussion of the results contained in

⋆ Tel: +61-7-33469465; fax: +61-7-33651177.
*Email address:* a.pudmenzky@uq.edu.au (Alexander Pudmenzky).

section 6 is followed by section 7 were we suggest the existence of naturally occurring examples. Finally, we conclude by offering some closing remarks in section 8.

## 2 Definitions

We use the generic term *agent* for an artificial or biological entity playing a part in the behavior of a population. An agent can be a gene or an animal such as an insect or human being, or an artificial entity such as a software-agent, a router in a communications network, a central processor in a multi-CPU cluster or a mechanical robot, to just name a few examples. We also use the word *population* as a generic term for a collection of agents in a defined environment. A population can be represented by terms such as "genome", "group", "swarm", "ant-colony", "collective" or similar. Agents will generally try to collect *rewards* located at *targets* in a certain problem domain (context) which we will call their *world*. Those rewards can consist of food or completed tasks. Total reward collection time is to be always minimized and reverse-proportional to the fitness of the population. Using this terminology we will now attempt a more general formulation of the problem under investigation.

## 3 Problem Description

Located in a $d$-dimensional world of size $A_{world}$, at each trial are $K$ targets with a total of $R$ equally distributed rewards so that each individual target consists of $R/K$ rewards. The size of the targets, $A_{target}$, is chosen so that an arbitrary ratio $a = A_{world}/A_{target}$ is achieved. A population of $N$ agents of zero extent (i.e. point-size agents) are uniform-randomly placed into this world at each iteration (cf. Fig. 1 for configuration). If an agent happens to be placed in a target area, the agent removes $c$ (carrying capacity) rewards at this iteration. The agent will remain at this position and continue to remove $c$ rewards at subsequent iterations until all rewards at this target position have been taken. The agent(s) may be joined by other agents that discover the target location at a later iteration. All agents participating in reward collection will again be participating in target location once the reward is exhausted. If all rewards are exhausted the number of iterations is recorded as $T$ and the trial is complete. In addition to discovering a target by direct placement into a target by chance, the discovery of targets can be achieved via communication between agents. An agent located in a target will broadcast its position to all surrounding agents. If agents capture the message, they will join the broadcasting agent in the target at the next iteration. However, the

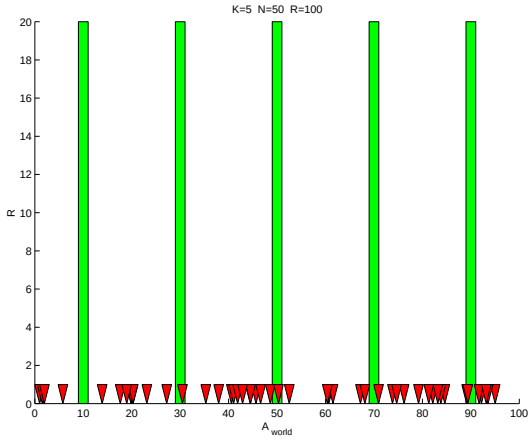


Fig. 1. Example configuration of $K = 5$ targets (bars) containing a total of $R = 100$ rewards and $N = 50$ agents (triangles) in a $d = 1$ -dimensional world of size $A_{world} = 100$. Targets are occupying 10 percent ($a = 0.1$) of the world size.

probability that surrounding agents will capture this broadcast depends on their distance from the broadcasting agent according to a normal distribution with standard deviation $\sigma$ (once an agent is committed to a target, it will not accept broadcasts from other agents at different targets anymore). The choice of $\sigma$ will therefore determine how many agents are on average being recruited for reward collection at each iteration. These recruited agents will therefore be unavailable for further target search, which will have some effect on the total reward collection time. We expect to find an optimum value that minimizes this time. This optimum value for $\sigma$ will depend on the number of agents $N$, targets $K$, rewards $R$, the ratio $a$, the carrying capacity $c$ and the dimensionality of the world $d$.

We call the problem we designed here the *bandit-search* problem, since it is reminiscent of the so-called *$K$-armed bandit* problem in which an optimal strategy has to be found for selecting from a number ($K$) of one-armed slot machines (bandits) with unknown reward probability to maximize the total payoff [10, p.2,3]. Both problems share a common theme; the essence of the bandit problem is the systematic search for the balance between exploration and exploitation necessary for effective optimization. It is known that each bandit is distributing rewards, what is not known is how much because the reward amount is stochastic. In comparison, the *bandit-search* problem consists of finding the bandit in the first place; once found, the reward is not stochastic but predictable and continuous until depleted. The optimization thereby shifts from a strategy of "bandit-selection" to a strategy of "bandit handle-pulls versus bandit search". Both problem models deal with the determination of the optimal tradeoff between balancing the benefits of gathering information (exploration) versus that of maximizing short-term payoff (exploitation).

## 4 Simulations

Since our *bandit-search* problem can not be satisfactorily analyzed anymore using a simple mathematical approach, we investigate the effects and influences of a variation in parameter values via computer modelling as follows. We select a varying number of targets (bandits) $K = 1, 5, 20, 50$ and agents $N = 1, 50, 100, 500$ with a fixed ratio $r = 2$ for the number of rewards per agents $r = R/N$ to exclude the likely influence of total number of rewards on the results (we did not separately investigate the influence of a varying number of total rewards $R$). We choose a 1-dimensional world of size $A = 100$ and a fixed ratio $a = 0.1$ for the target to world size $a = A_{world}/\sum A_{target}$ and a carrying capacity of $c = 1$. For each of the 16 resulting combinations of agents and targets we investigated a minimum of 26 standard deviations ranging from $\sigma = 0, 1, 2, 3, 4, 5, 10, 15, 20, \ldots, 100$ and $\infty$, and reported the mean of 1000 trials for each instance ($16 \cdot 26 \cdot 1000 = 416,000$ trials in total). Some more points were calculated where deemed necessary.

## 5 Results

The results of these runs are represented in graphic form in Fig. 2 for $1, 5, 20$ and 50 agents. The significant points of those graphs, namely mean total collection times $T$ for $\sigma = 0, \sigma = \sigma_{opt}$ and $\sigma = \infty$, are summarized in a single graph, Fig. 3, for better comparison.

From these results we draw the following conclusions.

(1) The optimum rate of assistance $\sigma_{opt}$ (or successful information transfer) that minimizes the total collection time $T$, changes with the number of agents $N$ and number of targets $K$.
(2) With an increase in number of targets $K$, the advantages of providing reliable assistance are shifted in favor of providing no assistance. Reliable assistance is closer to the optimum for smaller number of targets and optimal for a single target.
(3) No assistance is always worse than optimum assistance. However, with a large number of targets present, no assistance approaches optimum.
(4) Reliable assistance is either worse than optimum assistance or at least the same.
(5) No assistance can be either worse or better than reliable assistance, dependent on the number of targets $K$. The more targets are present, the more advantageous it becomes if no assistance is provided.
(6) The time taken for reward collection is decreased with an increase in agents and this relationship follows a power-law as Fig. 4 shows. For an

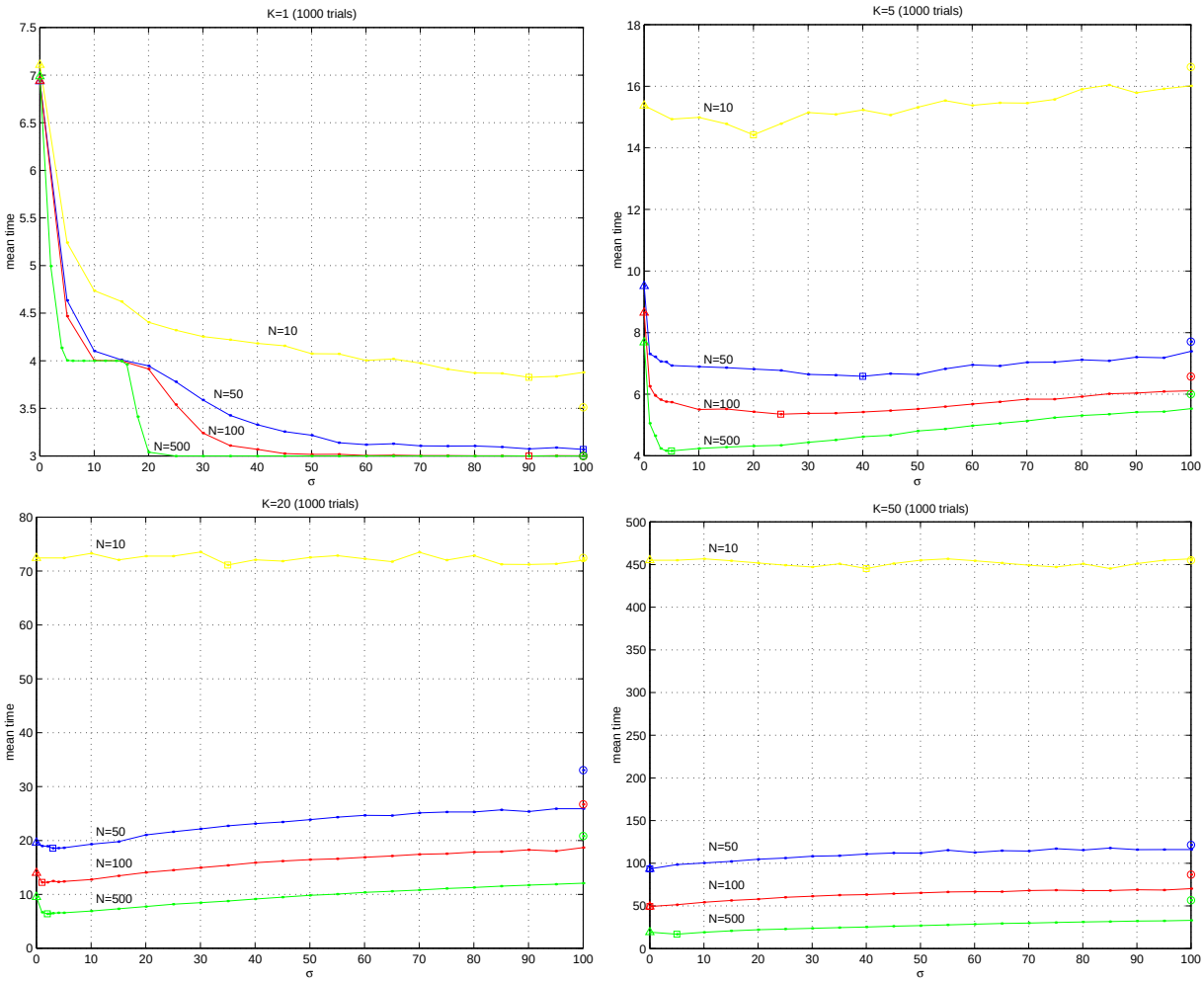


Fig. 2. Reward collection times $T$ with varying number of targets (separate graphs for $K = 1$,$K = 5$,$K = 20$,$K = 50$) for a population of $N = 10, 50, 100$ and 500 agents with $a = 0.1, r = 2, c = 1$ and $d = 1$. Each point represents the mean of 1000 trials. Points marked $\triangle$ at $\sigma = 0$ (left) represent times resulting from sole collection of agent(s) finding the target while all other agents were free to explore the world for additional targets, i.e. no information transfer occurred. Points marked $\odot$ at $\sigma = 100$ (right) are not to be confused with the points calculated from $\sigma = 100$, but instead represent times resulting from collection by all agents immediately after a target had been located by a single agent (or more than one agent simultaneously). Points marked $\square$ represent the optimum value for $\sigma$ for target location broadcasting, resulting in minimum reward collection time.

optimum value of the standard deviation (marked by the symbol $\square$), we notice an exponential decay in efficiency for an increased number of agents. The larger the number of distributed targets $K$ is, the more advantages will be gained from an increase in agents $N$ in terms of reducing the overall collection time $T$. This effect is reduced with increasing number of agents.

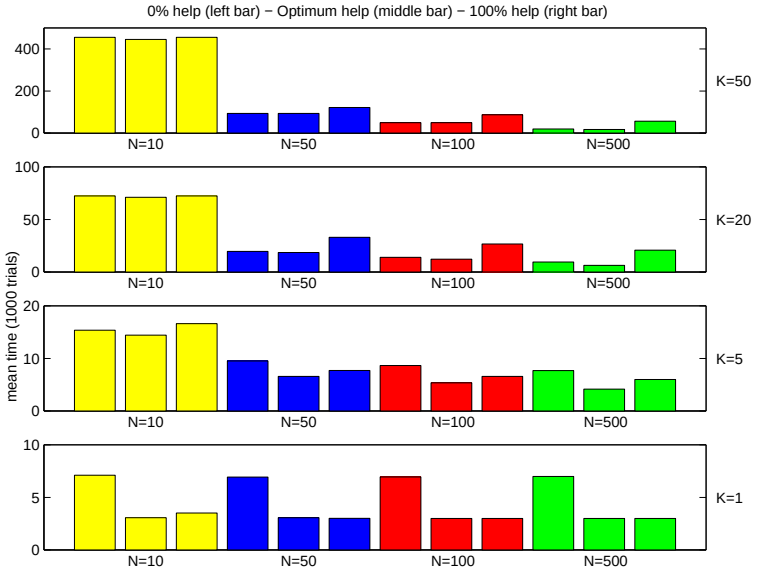


Fig. 3. Comparison of reward collection times $T$ for a population of $N = 10, 50, 100, 500$ agents and $K = 1, 5, 20, 50$ targets (summary of previous Fig. 2 now representing points marked $\triangle, \odot, \square$ there as bars). Each bar represents the mean of 1000 trials (please note the different scales for the mean time). All parameters as in Fig. 2. Three bars are shown for each combination of $N$ and $R$. The bar on the left represents the time taken to collect all rewards without communicating the location of targets to any other agent (no assistance). The bar on the right represents the time taken if all agents are collecting the rewards immediately after their discovery by at least one agent (reliable assistance). The bar in the center, which is always either lower or at least the same height as the other two bars, indicates the time taken when only the *optimum* number of agents participated in the collection of rewards while all other agents continued the search for new targets (optimum assistance). This optimum number was determined from simulation runs using $\sigma = 0, 1, 2, 3, 4, 5, 10, 15, 20, \ldots, 100$ and marked with $\square$ in previous graphs.

## 6 Discussion

The power law relationship of the collection time with both variables, number of agents $N$ and number of rewards $K$ displayed in Fig. 4, is reminiscent of the Danish theoretical physicist Per Bak's *self-organized criticality* [3]. "Self-organized" is often associated with the word "emergent" [6, p.99]. It seems that if a population of agents would be able to regulate its own standard deviation via a feedback loop to tune it optimally, it could lead to the emergent ability of the system to minimize collection time which in turn shows a power law relationship with process variables poising the system to reside at the edge of chaos. Kauffman [9] remarked that adaptive agents tune internal redundancy and couplings with one another to achieve a self organized critical state. The optimum amount of collaboration between agents minimizes the reward collection time which is equivalent to a maximization of agent-population

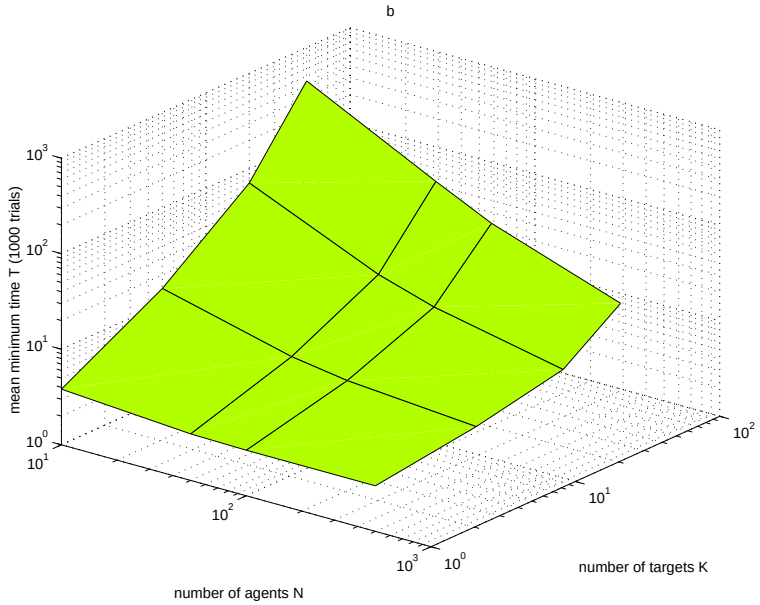


Fig. 4. Mean reward collection times $T$ under optimum conditions ($\sigma = \sigma_{opt}$) for varying number of agents $N$ and targets $K$ (mean of 1000 trials). All parameters as in Fig. 2. The logarithmic plot shows the power law relationship between both, the number of agents and the mean minimum collection time and also the number of targets and the collection time. Under optimal conditions an increase in agents will have a positive effect on the overall collection time $T$. For a large number of targets this effect is dramatic but weakens with an increase in agent numbers.

fitness. Furthermore, using this power law relationship with a known quantity of agents and targets, we can predict the minimal collection time under optimal conditions. Our results are an extension of finding a balance between exploration and exploitation into the temporal dimension.

It has not escaped our attention that the functionality of the standard deviation as used in our context resembles that of genetic epistasis. From a biological perspective, epistasis entails the interaction (enhancement or suppression) of alleles of different genes. Kauffman's [8] epistasis value $K$ determines how many other genes, from a total of $N$ genes, a single gene interacts with. Similarly the standard deviation $\sigma$ employed in our investigations determines how many other agents a single agent interacts with. Both values, $K$ and $\sigma$, can possess optimum values that maximize fitness as Kauffman and we have shown. The standard deviation $\sigma$ can therefore be understood as the stochastic counterpart of the static epistasis value $K$ given by Kauffman.

Below we list some examples where we suggest that feedback loops exist that enable optimal self-tuning of behavior in this way.

## 7 Examples

Deneubourg [4] noticed an "error" (as he called it) during communication among ants of the location of a food source. Some ants would not follow the instructions given to them by other ants to help exploit a known source, and wander away instead. Those ants were however free to discover new sources and Deneubourg showed, via computer simulations, that this behavior maximizes the total intake of scattered food for the colony over time. The standard deviation of this communication error can evolve to an optimum value that matches the degree of distribution of food normally encountered by particular species of ants.

Heck and Ghosh [7] computer-simulated the behavior of an ant colony in an attempt to study "synthetic creativity". The authors differentiated between "normal" ants that find food, collect food and follow artificial pheromone trails to food, and "creative" ants that only find food and establish a trail to it. The results of their simulation shows a U-shaped collection time for an increasing number of "creative" ants [7, p.62]. The authors however omitted to explain the importance of tuned standard deviations in this context.

Seeley *et al* [13] observed a similarly tuned probabilistic behavior in bees. Bees report the location of a patch of flowers via a well known wiggle dance. The standard deviation from the true direction to the location of a newly discovered patch allows other bees to explore the limits of this patch.

Allen and McGlade [1] report on different hunting strategies observed in fishermen, one of the few remaining examples of ancestral hunting activities in humans. Hunting contains elements of discovery and exploitation as opposed to agriculture, which only contains the latter. Their research is concerned with an improved model for fishing behavior in a fisheries management plan by the Government. The authors argue that fishery models should be based on the Lotka-Volterra equation rather than the logistic equation and base their conclusions on a case study of the groundfish fisheries of Nova Scotia, a Province in SE Canada. The logistic equation only considers fish populations but the Lotka-Volterra equation includes the behavior of fishermen as well. The inclusion of their behavior into the model reveals the advantage that a combination of two kind of fishing strategies have for the whole of the population, as observed in the case study. The first strategy is to search randomly with the risk of finding nothing, the second is to go to an area of known best return regardless of how low it is. If only the second strategy is chosen the result will be disastrous, all fishing activities will virtually shut down. The fishing fleet and individual catches will be small since efforts will only be concentrated on a single location. The first strategy however will result in the maintenance of the whole fishing area with a larger fishing industry and larger catches. A

reduction in information of areas of best return thereby allows a random response by the boats which will in turn explore less visited parts of the system. This involves more risk for skippers but results in avoiding certain disaster when discovery is totally abandoned in the opposite case. Allen and McGlade refer to the types of searches *exploration* and *discovery*. They liken discovery to invention and creation and stress the importance of the fact that they can be achieved through non-rational behavior, as in his examples of fishing strategies. The complementarity of the two behaviors results in a mixed strategy that optimizes the desired outcome. Here a deviation of behavior is spread over a number of individuals in a population. Two types of fishermen are identified, high risk-taking "hunters" and low risk-taking "followers". The authors generalize that a balance of both strategies maximizes the efficiency of the whole population and propose that a society should encompass both, in the form of freedom for discovery and preservation of traditional strategies.

In another paper [2], the same authors publish details of their model that shows that evolution is able to act on the rate of change. They hypothesize that fidelity of reproduction is a hereditary characteristic which could itself vary. Their simulations show that if a population contains a certain amount of random variability, selection is able to operate on it to regulate the variability necessary for hill climbing or countering the evolution of other species. We [12] have also demonstrated that this is indeed possible using a simple computer simulation of a population of agents containing a feedback loop with their environment. The processes not only include the selection of fit individuals but also the creation of new types. This makes variability itself part of a species' strategy and produces populations with not optimal behavior but the ability to learn. Adaptation and change now become a permanent feature of evolutionary strategy. Evolution is driven by the noise to which it leads and consists not only of the selection of optimal behavior but selection of species that can produce change thereby enabling an ability to cope with change.

## 8 Conclusion

Using a simple computer simulation of an interacting population of reward collecting agents, we have shown that there exists an optimum amount of cooperation between agents that minimizes the population's reward collection time. We investigated some parameters for the number of agents and the number of rewards and found that the optimum is sensitive to those parameters and located somewhere between agents providing no assistance and full assistance. Our simulations have also highlighted the power law relationships between the optimized collection time and the number of agents in one case, and the number of rewards in the other. This hints towards a self organized critical *edge of chaos* state that is thought to be the preferred operational state

of dissipative systems [11].

We suggest that a feedback loop can positively affect the discovery of the optimum setting and present some examples from nature where such a process can be observed. In all these cases, either artificial or natural, the quantitative aspects of cooperation between agents and the inclusion of feedback loops and algorithms to tune cooperation to an optimum value should receive increased attention.

## 9 Acknowledgements

The author would like to thank Piero Giorgi for discussions and comments on the manuscript.

## References

[1] Allen, P. M., McGlade, J. M., *Dynamics of discovery and exploitation: the case of the scotian shelf groundfish fisheries*, Can. J. Fish. Aquat. Sci., Vol. 43, 1187–1200, 1986.

[2] Allen, P. M., McGlade, J. M., *Evolutionary Drive: The Effect of Microscopic Diversity, Error Making and Noise*, Foundations of Physics, Vol. 17, No. 7, 723–738, 1987.

[3] Bak, P., Tang, C., Wiesenfeld, K., *Self-organized criticality*, Phys. Rev. A, Vol. 38, 364, 1988.

[4] Deneubourg, J. L., *Probabilistic Behaviour in Ants: A Strategy of Errors?*, J. theor. Biol., 105, 259–271, 1983.

[5] Gazzaniga, M., *Mind Matters: How Mind and Brain Interact to Create our Conscious Lives*, Bradford Books, 1988.

[6] Gell-Mann, M., *The Quark and the Jaguar*, New York: Freeman and Company, 1994.

[7] Heck, P. S., Ghosh, S., *A study of synthetic creativity: Behavior modeling and simulation of an ant colony*, IEEE Intelligent Systems and their Applications, Vol. 15, No. 6, 58–66, 2000.

[8] Kauffman, S. A., *The Origins of Order: Self-Organization and Selection in Evolution*, Oxford University Press, 1993.

[9] Kauffman, S. A., *At Home in the Universe: The Search for Laws of Self-Organization and Complexity*, Oxford University Press, 1995.

[10] Macready, W. G., Wolpert, D. H., *Bandit Problems and the Exploration / Exploitation Tradeoff*, IEEE Transactions on Evolutionary Computation, Vol. 2, No. 1, 2–22, April 1998.

[11] Prigogine, I., *From Being To Becoming: Time and Complexity in the Physical Sciences*, W. H. Freeman and Company, 1980.

[12] Pudmenzky, A., *Teleonomic Entropy: Measuring the Phase-Space of end-directed Systems*, submitted to: Applied Mathematics and Computation, Elsevier, December 2003.

[13] Seeley, T. D., Camazine, S., Sneyd, J., *Collective decision-making in honey bees: how colonies choose among nectar sources*, Behavioral Ecology and Sociobiology, Springer Verlag, 28, 277–290, 1991.