Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

SEN

Software Engineering

Software ENgineering

COllective INtelligence with Task Assignment

Pieter Jan 't Hoen, Sander M. Bohte

REPORT SEN-E0315 DECEMBER 18, 2003

CWI is the National Research Institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organization for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-369X

COllective INtelligence with Task Assignment

ABSTRACT
In this paper we study the COllective INtelligence (COIN) framework of Wolpert et al. for dispersion games (Grenager, Powers and Shoham, 2002) and variants of the EL Farol Bar problem. These settings constitute difficult MAS problems where fine-grained coordination between the agents is required. We enhance the COIN framework to dramatically improve convergence results for MAS with a large number of agents. The increased convergence properties for the dispersion games are competitive with especially tailored strategies for solving dispersion games. The enhancements to the COIN framework proved to be essential to solve the more complex variants of the El Farol Bar-like problem.

1998 ACM Computing Classification System: I.2.6
Keywords and Phrases: COllective INtelligence, Multi-Agent Systems, Reinforcement Learning

COllective INtelligence with Task Assignment

Coordinating choices in Multi-Agent Systems

Pieter Jan 't Hoen Sander M. Bohte
{hoen,sbohte}@cwi.nl
CWI, Centre for Mathematics and Computer Science
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Abstract

In this paper we study the COllective INtelligence (COIN) framework of Wolpert et al. for dispersion games (Grenager, Powers and Shoham, 2002) and variants of the EL Farol Bar problem. These settings constitute difficult MAS problems where fine-grained coordination between the agents is required.

We enhance the COIN framework to dramatically improve convergence results for MAS with a large number of agents. The increased convergence properties for the dispersion games are competitive with especially tailored strategies for solving dispersion games. The enhancements to the COIN framework proved to be essential to solve the more complex variants of the El Farol Bar-like problem.

1 Introduction

A computational optimization problem can be considered as a resource allocation problem (Wellman, 1996a; Wellman, 1996b). Borrowing from the insights of economics, it is however becoming increasingly clear that few concepts for resource allocation scale well with increasing complexity of the problem domain. In particular, centralized allocation planning can quickly reach a point where the design of satisfying solutions becomes complex and intractable. A conceptually attractive option is to devise a distributed system where different parts of the system each contribute to the solution for the problem. Embodied in a so-called distributed Multi-Agent System (MAS), the aim is thus to elicit "emergent" behavior from a collection of individual agents that each solve a part of the problem.

This emergent behavior relies implicitly on the notion that the usefulness of the system is expected to increase as the individual agents optimize their behavior. A weak point of such systems has however long been the typical bottom-up type of approach: researchers first built an intuitively reasonable system of agents and then used heuristics and tuned system parameters such that – hopefully – the desired type of behavior emerged from running the system. Only recently has there been work on more top-down type of approaches to establish the conditions for MASs such that they are most likely to exhibit good emergent behavior (Barto and Mahadevan, 2003; Lauer and Riedmiller, 2000; Guestrin, Lagoudakis and Parr, 2002).

In typical problem settings, individual agents in the MAS contribute to some part of the collective through their individual actions. The joint actions of all agents derive some reward from the outside world. To enable local learning, this reward has to be divided among the individual agents where each agent aims to increase its received reward by some form of learning. However, unless special care is taken as to how reward is assigned, there is a risk that agents in the collective work at cross-purposes. For example, agents can reach sub-optimal solutions by competing for scarce resources or by inefficient task distribution among the agents as they each only consider their own goals (e.g. a Tragedy of the Commons (Hardin, 1968)).

The COllective INtelligence (COIN) framework by Wolpert et al. suggests how to engineer (or *modify*) the rewards an agents receives for its actions in *private utility functions*. Optimization of each agent's private utility here leads to increasingly effective emergent behavior of the collective, while discouraging agents from working at cross-purposes.

The effectiveness of this top-down approach and their developed utilities are demonstrated by applying the COIN framework to a num-

ber of example problems: network routing (Wolpert, Tumer and Frank, 1998), increasingly difficult versions of the El Farol Bar problem (Wolpert and Tumer, 1999), Braess' paradox (Tumer and Wolpert, 2000), and complex token retrieval tasks ('t Hoen and Bohte, 2003). The COIN approach proved to be very effective for learning these problems in a distributed system. In particular, the systems exhibited excellent scaling properties. Compared to optimal solutions, it is observed that a system like COIN becomes relatively *better* as the problem is scaled up (Wolpert and Tumer, 1999).

In this paper we investigate distributed Reinforcement Learning (RL) for allocation of n agents to k tasks. Agents acting in parallel and using local feedback with no central control must learn to arrive at an optimal distribution over the available tasks. Such problems are typical for a growing class of large-scale distributed applications such as load balancing, niche selection, division of roles within robotics, or application in logistics. These problems are, for example, presented in the literature as dispersion games (Grenager, Powers and Shoham, 2002), minority games (Challet and Zhang, 1997) or variants of the El Farol Bar problem. We investigate the performance of the COIN framework for these general classes of problems relative to general RL approaches and present extensions to the COIN framework for improved convergence results.

This document is structured as follows. In Section 2, we describe the COIN framework and the used RL algorithm. In Section 3 we present dispersion problems that require coordinated joint actions of a MAS. We do not show results for minority games due to lack of space and comparable results as for the presented dispersion games. In Section 3 we also present extensions to the COIN formalism and report on the performance improvements. In Section 4, we present results for the more difficult task of an El Farol Bar-like problem. In Section 5 we discuss future work and conclude.

2 COllective INtelligence

In this Section, we briefly outline the theory of COIN as developed by Wolpert et al., e.g. (Wolpert, Wheeler and Tumer, 1999; Wolpert and Tumer, 1999; Wolpert and Tumer, 2001). Broadly speaking, COIN defines the conditions that an agent's private utility function has to meet to increase the probability that learning to optimize this function leads to increased performance of the collective of agents. Thus, the challenge is to define a suitable private utility function for the individual agents, given the performance of the collective.

In particular, the work by Wolpert et al. explores the conditions sufficient for effective emergent behavior for a collective of independent agents, each employing, for example, Reinforcement Learning (RL) for optimizing their private utility. These conditions relate to (i) the learnability of the problem each agent faces, as obtained through each individual agent's private utility function, (ii) the relative "alignment" of the agents' private utility functions with the utility function of the collective (the *world utility*), and lastly (iii) the learnability of the problem. Whereas the latter factor depends on the considered problem, the first two in COIN are translated into conditions on how to shape the private utility functions of the agents such that the world utility is increased when the agents improve their private utility.

Formally, let ζ be the joint moves of all agents. A function $G(\zeta)$ provides the utility of the collective system, the *world utility*, for a given ζ. The goal is to find a ζ that maximizes $G(\zeta)$. Each individual agent η has a private utility function g_η that relates the reward obtained by the collective to the reward that the individual agent collects. Each agent will act such as to improve its own reward. The challenge of designing the collective system is to find private utility functions such that when individual agents optimize their payoff, this leads to increasing world utility G, while the private function of each agent is at the same time also easily learnable (i.e. has a high *signal-to-noise* ratio, an issue usually not considered in traditional mechanism design). In this paper, ζ represents the choice of which of the k tasks each of the n agent chooses to execute and the challenge is to find a private function for each agent such that optimizing the local payoffs optimizes the total task execution.

Following a mathematical description of this issue, Wolpert et al. propose the **Wonderful Life Utility** (WLU) as a private utility func-

tion that is both *learnable* and *aligned* with G, and that can also be easily calculated.

$$WLU_\eta(\zeta) = G(\zeta) - G(CL_{S_\eta^{eff}}(\zeta)) \qquad (1)$$

The function $CL_{S_\eta^{eff}}(\zeta)$ as classically applied[1] "clamps" or suspends the choice of task by agent η and returns the utility of the system without the effect of agent η on the remaining agents $\hat{\eta}$ with which it possibly interacts. For our problem domain, the clamped effect set are those agents $\hat{\eta}$ that are influenced in their utility by the choice of task of agent η. Hence $WLU_\eta(\zeta)$ for agent η is equal to the value of all the tasks executed by all the agents minus the value of the tasks executed by the other agents $\hat{\eta}$. If agent η picks a task τ, which is not chosen by the other agents, then η receives a reward of $V(\tau)$, where V assigns a value to a task τ. If this task is however also chosen by any of the other agents, then the first term $G(\zeta)$ of Equation 1 is unchanged while the second term drops with the value of $V(\tau)$ as agent η competes for completion of the task. Agent η then receives a penalty $-V(\tau)$ for competing for a task targeted by one of the other agents $\hat{\eta}$. The WLU hence has a built in incentive for agents to find an unfulfilled task and hence for each agent to strive for a high global utility in its search for maximizing its own rewards.

Compared to the WLU function, other payoff functions have been considered in the literature for distributed Multi-Agent Systems: the Team Game utility function (TG), where the world-utility is equally divided over all participating agents, or the Selfish Utility (SU), where each agent only considers the reward that it itself collects through its actions. The TG utility can suffer from poor learnability, as for larger collectives it becomes very difficult for each agent to discern what contribution is made (low signal-to-noise ration), and the SU suffers from – potentially – poor alignment with the world-utility, i.e. agents can work at cross purposes. In Sections 3 and 4, we study the performance of the SU and TG relative to the variants of the WLU.

We use Q-learning (Sutton and Barto, 1998) as RL algorithm for each of the n agents in the MAS. A learner's input space consists of the available k tasks. The policy π is stochastic according to a softmax function; in the policy, a random task k_i is chosen for state s and constant c (set at 50) with normalized chance in $[0, 1]$ of $\frac{c^{Q(s,k_i)}}{\sum_j c^{Q(s,k_j)}}$. As each agent only must choose one task/action, we use a single state per agent. The discount factor γ is set to 0.95. The learning rate α, unless specified otherwise, is set at 1 as this produced best results for all the utility functions considered. The next section presents application of the RL learners for a MAS task assignment problem.

3 Dispersion Games

Dispersion games (Grenager, Powers and Shoham, 2002) are a general class of problems where n agents each have to decide which of the k tasks they are to undertake. In this section we investigate the case when $n = k$ which we call the *full dispersion game*. Full utility is achieved only when all k tasks are chosen by exactly one of the n agents. We first discuss some results from (Grenager, Powers and Shoham, 2002) for this specific problem setting where agents use different strategies[2] for choosing their tasks.

As analyzed in (Grenager, Powers and Shoham, 2002), for $n = k$, the expected time to successful allocation for a naive strategy with random choices by agents is $n^n/n!$. This is exponential in n. Similar long time to convergence results where found for Fictitious play, even with slight modifications to the updates of beliefs to avoid oscillatory behavior within sets of suboptimal outcomes. Better results where found using RL with a Q-learning Algorithm with a Boltzman exploration policy. The agents learned the expected reward for choosing a specific task. The (selfish) reward for each of the agents is a function of the number of agents that use the same action. For this setting, with a well chosen temperature decay trajectory, a polynomial time to convergence was found for convergence to the optimal solution. Similar convergence results are found for the *Freeze* strategy where an action is chosen randomly by an agent until the first time it is alone in choosing an action, at which point

[1] Ongoing work investigates more general clamping functions.

[2] See (Grenager, Powers and Shoham, 2002) for details and references.

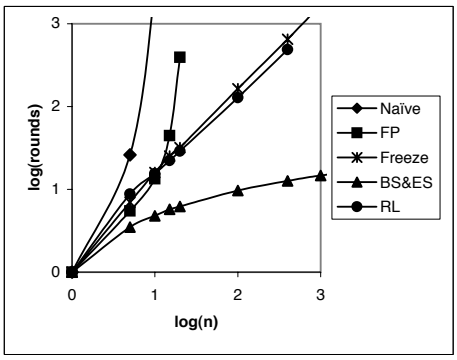

Figure 1: Results reprinted from (Grenager, Powers and Shoham, 2002) with permission.

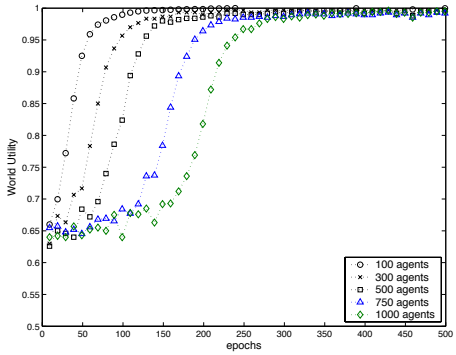

Figure 2: WLU for dispersion games

the agent replays that action indefinitely. Best results where found for the *Basic Simple Strategy* (BS) and the *Extended Simple Strategy* (ES) where agents quickly focus on a task when they are the only candidate and otherwise stochastically choose from the *remaining* tasks that are still under contention. See Figure 1 (reproduced with permission) for an overview of the results.

Figure 2 shows the results for the WLU for increasing number of agents and corresponding number of tasks ($n = k$). The reward for executing a task by an agent is 1. The convergence results improve on the used reinforcement learning algorithm of (Grenager, Powers and Shoham, 2002) and are competitive with the BS en ES strategies, with however a much more local signal as tasks that still need to be resolved are not communicated to the agents and an agent will have to explore for its "own" task. The RL signal for agent η is purely based upon how many agents $\hat{\eta}$ choose the same task. Agents using the SU (not shown) quickly reach a maximum fitness of ≈ 0.8. The agents using the SU however have difficulty in targeting the last 20% of the tasks as they continue to compete for tasks. The TG utility (not shown) performs even worse as a maximum utility of 0.7 is reached for 10 agents and a utility of ≈ 0.65 for a larger number of agents as the signal-to-noise ratio decreases. In contrast, the penalties imposed by the WLU successfully drive agents to efficiently disperse.

As the number of agents increases, the point at which individual agents choose a task is delayed. Agents for longer periods compete for tasks in their early exploratory behavior and the issued penalties can not yet push unsuccessful agents to unfulfilled tasks while this incentive for correct dispersion is necessary for the system to quickly converge. To improve on convergence of the COIN framework, we define two new extensions of the WLU. Both are based on the observation that the WLU as defined in Section 2 is symmetric. If, for example, two agents a_1 and a_2 both choose task k_i, then both agents, according to equation 1, receive a penalty when calculating $WLU_{a_1}(\zeta)$ and $WLU_{a_2}(\zeta)$ respectively. This however can lead to slower convergence as *both* agents then may be forced to target different tasks while only *one* of the agents need choose a different task. This slower convergence becomes more dramatic as more than one agent, say $l > 2$ agents, focuses on the same task and $l-1$ agents need to "switch". This phenomenon partially explains the trend in slower convergence of the WLU for an increasing number of agents in Figure 2.

We break the symmetry in the penalties of the WLU in two ways. First of all, we consider the case where one of the agents η targeting a task k is *r*andomly chosen as the winner and is awarded the positive reward while the other $\hat{\eta}$ agents choosing the same task k are penalized. We name this the $WLUr$ as we consider a *r*andom winner in which of the agents happen to arrive at a specific task. Secondly, as a more refined variant of the WLUr, we consider the case where the positive reward is assigned to the agent that is most likely to choose action k. We reward agent η with the highest Q value for this task. We name this the $WLUm$ from **m**ost likely.

In Figure 3 we show typical results for the

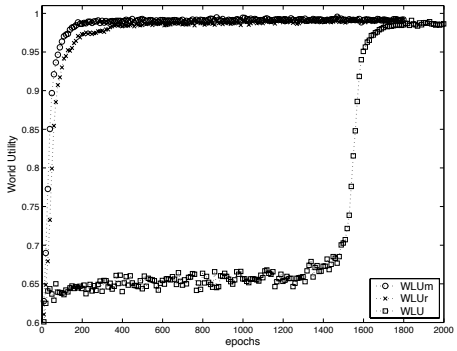

Figure 3: Improved convergence

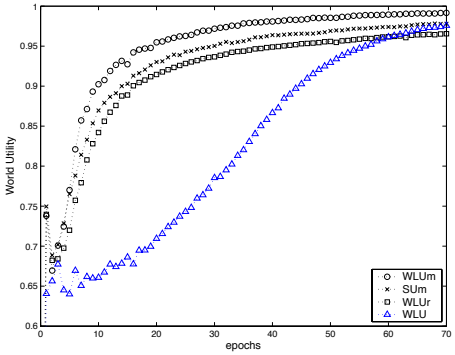

Figure 4: Performance of SUm

new semantics, in this case for 2500 agents.[3] The $WLUr$ and $WLUm$ converge dramatically faster than the classic WLU, even for a large number of agents. The $WLUm$ outperforms the $WLUr$ similarly in all experiments for the range of agents studied in Figure 2. Agents using the WLUm can most quickly converge to a task and drive other agents to choose another task. Note that the adaptions of the WLU's still only involve local use of information per task in the problem domain and no global information is used while the WLUr and WLUm are competitive with the ES and BS strategies of (Grenager, Powers and Shoham, 2002).

We investigate the influence of the adaption for the WLUm to the SU, which we name SUm. Like for the $WLUm$, the agent **m**ost likely to choose a task is given the reward. Penalties to contenders for the same task are however not given. In Figure 3 we show typical results for a 100 agents.[4] The performance of the SUm is inbetween that of the WLUr and WLUm while all learning methods converge in the limit to optimal results. In Section 4 we show that this property does not hold for the SUm in the more difficult task choice problems.

In the above experiments we found best convergence results for all RL algorithms while using a large learning rate α for the individual Q learners. Increasing α from 0.1 to 1 with increments of 0.1 led to continuous increased performance as agents then most quickly choose an individual task to execute. A preference for a large α is in contrast to earlier work ('t Hoen and Bohte, 2003) where agents must coordinate on *sequences* of moves in a MAS token retrieval task. In this setting, a small α led to best results as a too large α led to strong fluctuations in valuations of moves early in a sequence and this easily disrupts the fine-grained coordination required along the *entire* trajectory of an agent.

The next section presents a more difficult dispersion game.

4 El Farol Bar

In our interpretation of the El Farol Bar Problem (Arthur, 1994) that is inspired by the notion of dispersion games, agents have to decide on what day week they will visit one of a given set of bars. Good solutions can be hard to reach in a distributed setting as agents oscillate in their choice of attendance.[5] In this paper we model this problem as n agents that have to choose between 7 tasks that each give a reward of 1 to the first $n/7$ agents that choose the task. Reward for attendance is however only given if at least $n/7$ agents choose a task. In terms of dispersion games, we study $k = 7$ tasks that require 7^{c-1} agents to fulfill for a total of $n = 7^c$ agents, for some constant $c \geq 1$.

Figure 5 shows the results for the various learning algorithms for 49 agents ($c = 2$). Each bar is interpreted as a task that requires 7 agents for a total reward of 7, or 1 per agent helping to accomplish the "task". The SU and TG perform badly as both cannot locally interpret the RL signal to optimize their actions. The variants of the WLU all perform well, as

[3] We did not explore settings with more agents due to memory restrictions with the current implementation.

[4] Similar results held for 10, 1000, 1500, and 2500 agents.

[5] No one goes there nowadays, it's too crowded. (Yogi Berra)

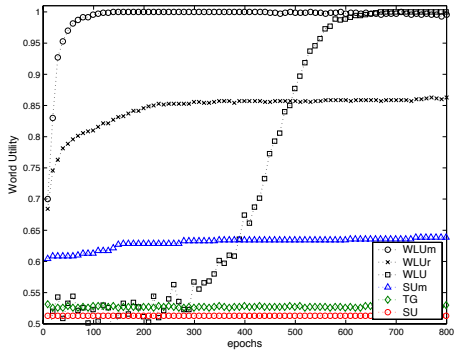

Figure 5: Bar attendance for 49 (7^2) agents

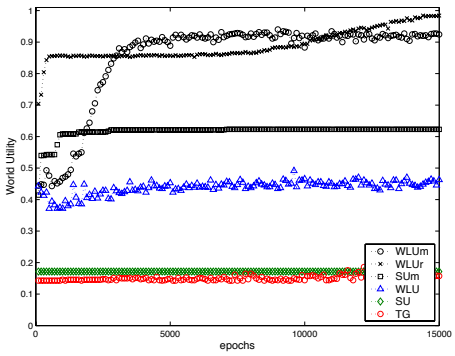

Figure 6: Bar attendance for 343 (7^3) agents

expected. The enhancements to the WLU introduced in Section 3 significantly increase the convergence rate of the system. The SUm, which showed comparable performance for the dispersion tasks of that Section, however does not have a sufficient added benefit for this more difficult task. It converges to a maximum utility of 0.8 after $30,000$ epochs. The issuing of penalties as defined for the WLU's is fundamental for convergence to full utility.

In Figure 6, we show results for 343 (7^3) agents. The WLUr and WLUm as exceptions are both able to achieve good results. We however only achieved these best convergence results for all RL utilities by changing the used learning rate α to an unconventional high level of 10. Similar good results are found by slightly adjusting the reward for the individual bar attendance to 51 instead of 49.[6] Both solutions resulted in stronger convergence by forcing an agent to choose a task. The WLU, even with this enhancement, however did not improve beyond its shown level even after $150,000$ epochs. The WLUr for this problem shows surprising results in performance relative to the WLUm when compared to the results of Section 3.

For this interpretation of the El Farol Bar problem with such a large number of agents we are reaching the limit of the straightforward application of the WLU and even of the proposed enhancements. We had to resort to modifications in the parameters of the learning algorithm or the used reward structure. We are hence reaching a point where we are moving beyond straightforward application of the COIN framework as an engineering approach. This problem hence merits further study to arrive at more fundamental solutions and insights.

[6] $51/49 > 1$ per task reward to help push the softmax function of the Q learners to one task.

5 Discussion and Conclusion

In this paper we studied the COllective INtelligence (COIN) framework of Wolpert et al. for dispersion games (Grenager, Powers and Shoham, 2002) and variants of the EL Farol Bar problem. Essentially, agents have to learn to choose individual tasks to execute. We observed that for complex problems the COIN framework is able to solve difficult MAS problems where fine-grained coordination between the agents is required, in contrast to multi-agent systems that use more common decentralized coordination.

We enhanced the COIN framework to dramatically improve convergence results for difficult problems. The increased convergence properties for the dispersion games are competitive with especially tailored strategies for solving these task assignment problems. The enhancements to the COIN framework proved to be essential to solve the more complex variants of the El Farol Bar-like problem.

The dispersion games of (Grenager, Powers and Shoham, 2002) we believe form an interesting testbed for learning methods applied to Multi-Agent Systems (MASs). The task assignment for n agents to k tasks is straightforward to implement, yet can quickly become difficult for distributed approaches due to parallel, asynchronous learning by the agents and the lack of global information. A fundamental question for learning methods is at what point to they begin to fail as the problems are scaled (increasing n or more difficult tasks). Can this point be de-

layed by increasing communication between the agents and at what cost? MAS learning is a growing research area. Dispersion games can form an interesting benchmark problem to research the limits and possibilities of this new field.

Acknowledgment

This work has been carried out under theme SEN4 "Evolutionary Systems and Applied Algorithmics". This research has been performed within the framework of the project "Distributed Engine for Advanced Logistics (DEAL)" funded by the E.E.T. program in the Netherlands. This research was supported by a partial travel grant by the NASA-Ames Research Center Moffett Field, California.

References

Arthur W. B. Inductive reasoning and bounded rationality. *Am. Econ. Assoc. Papers and Proc*, 84(406), 1994.

A. Barto and S. Mahadevan. Recent advances in hierarchical reinforcement learning. *Discrete-Event Systems journal*, 2003. to appear.

C. D. Chang and Y.-C. Zhang Emergence of cooperation and organization in an evolutionary game. *Physica A*, 246(407), 1997.

T. Grenager, R. Powers, and Y. Shoham. Dispersion games: general definitions and some specific learning results. In *AAAI 2002*, 2002.

C. Guestrin, M. Lagoudakis, and R. Parr. Coordinated reinforcement learning. In *Proceedings of the ICML-2002 The Nineteenth International Conference on Machine Learning*, 2002.

G. Hardin. The tragedy of the commons. *Science*, 162:1243–1248, 1968.

M. Lauer and M. Riedmiller. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *Proc. 17th International Conf. on Machine Learning*, pages 535–542. Morgan Kaufmann, San Francisco, CA, 2000.

R. Sutton and A. Barto. *Reinforcement learning: An introduction*. MIT-press, Cambridge, MA, 1998.

P.J. 't Hoen and S.M. Bohte. COllective INtelligence with sequences of actions. In *14th European Conference on Machine Learning*, Lecture Notes in Artificial Intelligence. Springer, 2003.

K. Tumer and D. Wolpert. COllective INtelligence and Braess' paradox. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pages 104–109, Austin, Aug. 2000.

M. P. Wellman. The economic approach to artificial intelligence. *ACM Computing Surveys*, 28(4es):14–15, 1996.

M. P. Wellman. Market-oriented programming: Some early lessons. In S. Clearwater, editor, *Market-Based Control: A Paradigm for Distributed Resource Allocation*. World Scientific, River Edge, New Jersey, 1996.

D. Wolpert and K. Tumer. An introduction to COllective INtelligence. Technical Report NASA-ARC-IC-99-63, NASA Ames Research Center, 1999. A shorter version of this paper is to appear in: Jeffrey M. Bradshaw, editor, Handbook of Agent Technology, AAAI Press/MIT Press, 1999.

D. Wolpert and K. Tumer. Optimal payoff functions for members of collectives. *Advances in Complex Systems*, 2001. in press.

D. H. Wolpert, K. Tumer, and J. Frank. Using collective intelligence to route internet traffic. In *Advances in Neural Information Processing Systems-11*, pages 952–958, Denver, 1998.

D. H. Wolpert, K. R. Wheeler, and K. Tumer. General principles of learning-based multi-agent systems. In O. Etzioni, J. P. Müller, and J. M. Bradshaw, editors, *Proceedings of the Third Annual Conference on Autonomous Agents (AGENTS-99)*, pages 77–83, New York, May 1–5 1999. ACM Press.