The RoboCup-98 Teamwork Evaluation Session: A Preliminary Report

Gal A. Kaminka[1]

Information Sciences Institute and Computer Science Department
University of Southern California
Marina del Rey, CA 90292
galk@isi.edu

Abstract. Increasingly, agent teams are used in realistic and complex multi-agent environments. In such environments, dynamic and complex changes in the environment require appropriate adaptation of the teamwork (collaboration) among team-members. As RoboCup proposes to provide multi-agent researchers with a standard test-bed for evaluation of methodologies, it is only natural to use it for investigating this essential capability. During the RoboCup-98 workshop and competition a unique event took place: a comparative evaluation of the teamwork adaptation capabilities of 13 of the top competing teams. An evaluation attempt of this scale is a novel undertaking, and presents many novel challenges to researchers in the multi-agent community. This preliminary report describes the data-collection session, the experimental protocol, and some of the preliminary results from analysis of the data. Rather than proposing solutions and well understood results, it seeks to highlight key challenges in evaluation of multi-agent research in general, and of teamwork in particular..

Introduction

Agent teamwork (collaboration) is an important and challenging research area, as teams of agents are increasingly becoming a common, often required, theme in many dynamic, complex, multi-agent environments. Such application environments range from virtual environments for training (Johnson and Rickel 1997), through distributed large-scale simulations (Tambe, Johnson et al. 1995) and robotic soccer (Kitano, Tambe et al. 1997), to future space missions. These application domains present many challenges to managing teamwork (collaboration) among the agent team-members: Agents may have conflicting or incomplete local views, the environment may impose restrictions on their ability to observe each other or otherwise communicate, and individual, localized, failures may lead to global, team-

[1] We thank Kay Schroter and Prof. Hans-Dieter Burkhard for their help during data-collection phases.

wide difficulties. Agent teams deployed in such realistic settings must therefore be able to recognize when such situations occur and adapt individually and team-wise to the changing conditions.

Generally, evaluation of adaptive teamwork capabilities is limited to investigators using the application domain. Each application is therefore generally evaluated on absolute terms, not with respect to other adaptation techniques. Partially to address this problem, RoboCup has been proposed as a standard research domain and test-bed for multi-agent and robotics research (Kitano, Tambe et al. 1997). In particular, the RoboCup simulation environment shares many of the characteristics that make realistic domains so challenging. It involves multiple interacting agents, both in collaborative and adversarial modes, uncertainty in perception and action, random environmental effects (such as weather conditions), etc. Annual competitions, attended by researchers and programmers from across the world present an environment where the adaptive capabilities of teams may be investigated and evaluated comparatively.

Indeed, the IJCAI-97 RoboCup Synthetic Challenge (Kitano, Tambe et al. 1997) calls for rigorous scientific evaluation of teamwork techniques (among other research topics) using the simulation as the standard test-bed environment. Since 1997, over 60 different agent teams, developed independently by different groups, have been built for RoboCup. Unfortunately, most have been only evaluated on the basis of their overall performance in the annual competitions, rather than on the basis of their scientific contributions in particular areas of multi-agent research.

To remedy this situation a unique event took place during the RoboCup-98 workshop and competition: a large-scale comparative evaluation of teamwork adaptation under controlled failure condition of 13 different simulation teams. The evaluation of each team consisted of playing the team against a fixed opponent four times, as up to 3 of its players were disabled. It was not only the first evaluation session of RoboCup teams, but, to the best of our knowledge, the first multi-agent teamwork evaluation of this kind, and on this scale.

The evaluation methodology presents novel challenges to researchers concerned with teamwork. Issues such as statistical significance, evaluation protocols, comparative measures of teams, etc. all raise important questions about our understanding of teamwork in particular, and multi-agent research evaluation in general. As an example of such a difficult issue, despite the availability of an obvious measure of overall *team performance* (the score difference at the end of a game), it isn't clear how *teamwork performance* should be measured.

This short, preliminary, report provides an overview of the evaluation session from the perspective of the organizers. It describes the controlled conditions under which experimentation took place. It provides a description of the data collection protocol and the motivation for its different phases. It discusses a preliminary example how the data may be analyzed, and points at some of the challenges and questions that emerge. Rather than point out specific answers and argue for the correctness of particular methods or the preliminary results, this paper seeks to promote discussion of the underlying challenges involved in this undertaking, and the important questions that arise from them.

This report is organized as follows. Section 2 describes the data-collection controlled conditions and protocol. Preliminary results are provided in section 3. Section 4 provides a discussion of these results and points to emerging questions. Section 5 concludes.

The RoboCup-98 Evaluation Session

During the RoboCup'98 competition and workshop in July 1998 (Paris, France), a special two-day evaluation session was organized, beginning an annual tradition of rigorous scientific evaluation of RoboCup teams. The session consisted of 13 simulation teams each playing against the same fixed opponent four times, as incrementally, up to 3 of their players were disabled. The following describe the experiment protocol and controlled conditions.

Participation

Participation in the evaluation session was open to all teams who wanted to take place, but was strongly encouraged for all teams who have made it through the round-robin round to the double-elimination rounds (representing the top 16 teams in 37). All in all, 13 different teams participated in the evaluation:

- CMUnited-98 (Stone and Veloso 1998)
- AT Humboldt '98 (Burkhard, Wendler et al. 1998) and '97 (Burkhard, Hannebauer et al. 1998)
- Kasuga-Bito II (Maeda, Kohketsu et al. 1998)
- ISIS'98 (Marsella, Adibi et al. 1999)
- PaSo Team (Montesello and Pagello 1998)
- Gemini (Gemini 1998)
- Andhill'98 (Andou 1998)
- AIACS (Lubbers and Spaans 1998)
- CAT Finland (Riekki 1998)
- Darwin United (Andre and Teller 1998)
- Mainz Rolling Brains (Polani, Weber et al. 1998)
- Windmill Wanderers (Corten and Rondema 1998)

Fixed settings

All participating teams competed against a fixed opponent—the previous year's world champion "AT Humboldt'97" (Burkhard, Hannebauer et al. 1998), which was only slightly modified to accommodate changes in the simulation software made between 1997 and 1998. Note that the fixed opponent was also a strong competitor in 1998, and was also evaluated against itself.

Hardware settings (computers, network conditions, etc.) were identical to those of the actual competition: Teams were allowed to use up to 8 Sun machines each for the clients. Two different games ran in parallel, using two different machines to run the servers. Actual competition versions of the players were used. The protocol prohibited using any special versions of the code for the purpose of evaluation. Indeed, teams did not know about the evaluation session until three days before it took place, and did not know the details of the protocol until the beginning of the evaluation session. Except for the disabling of players (which was the controlled variable), the games strictly followed competition rules, with a referee and representatives of each team present during the matches. No tuning of program parameters was allowed between phases.

Evaluation Phases

Each team played four half-games (each lasting 3000 simulation "ticks", about 5 minutes) against the fixed opponent. Each such half-game constitutes an evaluation phase, in which a single change to the number of disabled players was made. These phases are denoted A through D:

- **Phase A.** The control phase. The team played against the fixed opponent under normal competition rules. No players were disabled.
- **Phase B.** The team played against the fixed opponent with a *single player disabled*. The player was randomly selected by the computer—but was not allowed to be the goalie. Thus, a different player was chosen for each team.
- **Phase C**. The team played against the fixed opponent with two players disabled:
  - The same player randomly selected in phase B, and
  - A player selected by the fixed opponent's representative with the intention of disabling the evaluated team's most valuable player. (But not the goalie)
- **Phase D**. The team played against the fixed opponent last final half-game, with three players disabled: The two players disabled in phase C and the evaluated team's goalie.

The motivation for this evaluation protocol in general was to check how well teams are able to adapt to loss of members. Phase A was intended to establish a base-line for the evaluated team's performance under normal conditions. Phases B through D provided the experimental worsening conditions. In all of these phases, the ideal would have been, for comparison's sake, to disable the same player in all teams, to see how their adaptive capabilities face to the same problem. The intention in using a randomly chosen player was to make sure that the teams could not have prepared in advance for particular evaluation settings. However, different teams assign different player numbers to the similar roles—thus randomly selecting a player number and then disabling the same player number is all teams would make little sense.

The next logical alternative is to look at the role of the players: Randomly select a role (e.g., top midfielder) and then disable the player who plays this role in each team. But again teams greatly vary in their team strategies. Thus not all teams have

the same roles. The most common role was that of the goalie (disabled in phase D) but even for the goalie there was at least one team that had players take over the role of the goalie on a regular basis (Andou 1998).

Nevertheless, some random element was required, to make sure that teams did not know in advance which players were going to be disabled. We have therefore decided to randomly select a player for each team that would first be disabled in phase B, but then balance this randomness in phase C by allowing the representative of the fixed opponent to select a player that would potentially damage the evaluated team's soccer-playing ability the most. Our hope was that at the very latest, all teams would face similar difficulties when they reach phase C.

Data Collection: Experiment Execution

For each of the different phases, for each of the evaluated teams, the soccer-server log files were saved and tagged appropriately. These provide complete records of the game, with the exception of communicated messages exchanged between the players. These logs were made available publicly (Repository 1998) to any and all interested parties.

Players were disabled in their initial position and facing direction on the field at the beginning of the game, but were left on the field. The server ignored any commands sent by their respective clients, so disabled players could not communicate, move, nor turn. However, they were visible to other players from their own team and the opponent teams.

In phase B, a list of random numbers in the range 1-11 was generated by the C library's pseudo-random generator, and the numbers were assigned to the different teams in order of participation. The randomly selected player could not be the goalie: if the random number was that of the goalie for the given team, it was skipped and the next different number on the list was used instead. The randomly assigned number was not revealed to the evaluated team until it was actually disabled.

In phase C, when the representative of fixed opponent (Burkhard, Hannebauer et al. 1998) was to choose the next player that would be disabled, such that it would potentially harm the evaluated team the most, there were two potential cases of conflicts in interest: (a) when the fixed opponent was playing itself, and when the fixed opponent (AT Humboldt 97) was used to evaluate AT Humboldt 98, which was developed by the same programming team. In both of these cases, the selection of the representative of the fixed opponent was independently corroborated by a neutral party.

The players disabled for each team in each phase are presented in the Table 1. For each team, the table shows the players disabled in each evaluation phase (i.e., phases B through D).

Preliminary Results and Analysis

Very early on it became clear that actual performance of *teamwork*, rather than the *team*, is difficult to measure. With most teams, qualitative changes in team performance were not observed, and even in cases where qualitative differences were found, they were not sufficient for rigorous comparative evaluation. Quantitative measures are required which can allow us to compare teams and their performance in general, and their collaboration and coordination skills in particular.

Team Name	Disabled Players			Team Name	Disabled Players		
CMUnited 98	5			Rolling Brains	4		
	5	10			4	10	
	5	10	1		4	10	1
Darwin United	5			ATH 98	10		
	5	11			10	3	
	5	11	1		10	3	1
Windmill Wanderers	2			Kasuga-Bito II 98	8		
	2	8			8	2	
	2	8	11		8	2	1
Andhill 98	2			ISIS 98	2		
	2	8			2	9	
	2	8	1		2	9	11
CAT Finland	10			Gemini	4		
	10	9			4	9	
	10	9	1		4	9	1
AIACS	7			Paso Team	5		
	7	10			5	7	
	7	10	1		5	7	11
ATH 97	10						
	10	7					
	10	7	1				

Table 1. Disabled players for each team, in each phase (B-D).

Our expectation is that any quantitative measure used will show a trend of declining performance as more and more players are disabled. A more adaptive team would have a slower decline in performance, while a less capable team would have a sharper decline. Intuitively, a more adaptive team would have less reduction in performance when it loses team-members, as the remaining members would be able to compensate for those disabled. Of course, in practice we cannot expect team members to be successful in compensating for an arbitrary number of disabled team-members (teams, after all, are used most often when tasks are simply too complex and to big for a single agent to undertake). It is also useful to think of a purely theoretical ideal of a zero-slope performance trend, in which a theoretical team so successfully compensates for disabled players that there is no change in performance. This allows measurement of a team's performance not only relative to other teams, but also on an absolutely (0 decline being an ideal best).

At least three quantitative measures are immediately available in the domain of soccer: The number of goals scored by the evaluated team, the number of goals

scored by the opponent, and the score-difference resulting from it. As an example, the score-difference results are shown graphically in figure 1. For each team, for each phase (A-D), the score-difference at the end of the half are shown. The scores are normalized for each team on the basis of the team's performance in phase A. In other words, the results of phases B through D for each team are shown relative to the team's performance in phase A, rather than on an absolute scale. The results are plotted as a function of the number of disabled players.

The results are difficult to interpret. Some teams seem to react with increased performance to loss of players, at times even outdoing their performance in the control phase (A). However, not all teams have responded in this way. Some have shown no effect at times, while others show the expected declining trend in measured performance. Plotting these means, we can see that on average, the evaluated teams show a reduction in performance as we hypothesized (Figure 2).

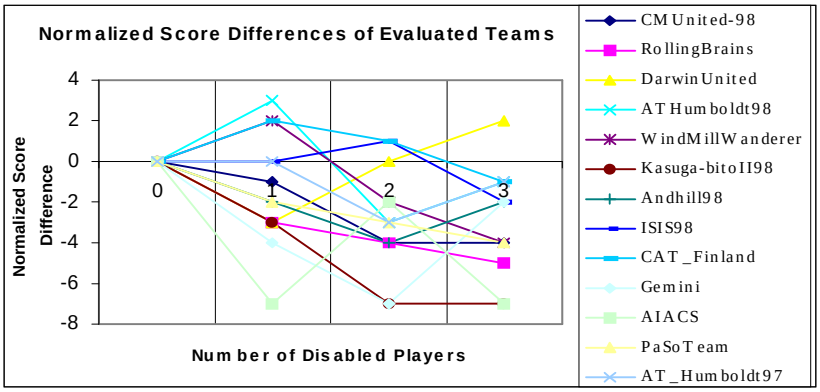

Fig. 1. Normalized Score Differences of Evaluated Teams.

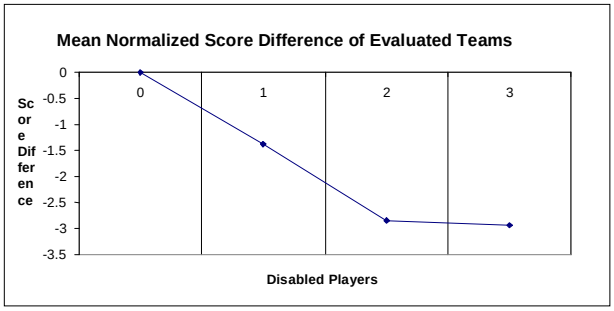

Fig. 2. Mean of Normalized Score Differences.

We use linear regression to draw a line that best represents the performance trend of the evaluated team. Figure 3 shows the computed regression slope values for each of the evaluated teams, as a function of the number of disabled players.

Maintaining performance in face of disabled players is graphically equivalent to a horizontal line, and so the more "horizontal" the performance trends of teams are, the closer they are to this theoretical ideal. Table 2 shows the rankings in our example. It should be emphasized again that these results are based on a *preliminary example analysis*. Indeed, the results are *not statistically significant.* These issues and others are discussed in the next section.

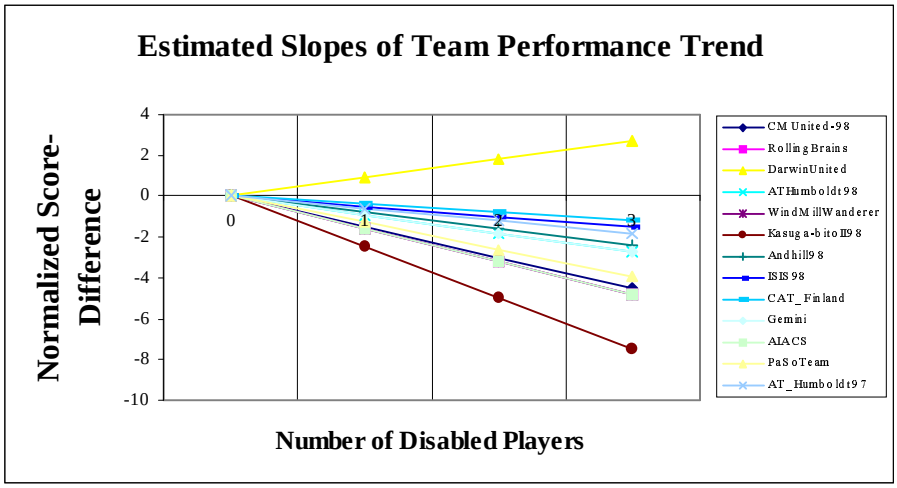

Fig. 3. A Plot of Estimated Team Performance Trends.

Team Name	Estimated Performance Trend
CAT Finland	-0.4
ISIS98	-0.5
AT_Humboldt97	-0.6
Andhill98	-0.8
ATHumboldt98	-0.9
Gemini	-0.9
Darwin United	+0.9
PaSo Team	-1.3
CMUnited-98	-1.5
Windmill Wanderer	-1.6
Rolling Brains	-1.6
AIACS	-1.6
Kasuga-bitoII98	-2.5

Table 2. Estimated ranking of teams' adaptivity (estimates n*ot* statistically significant).

Discussion

Rather than providing definitive answers to questions about actual evaluation results, or about evaluation methodology, the preliminary results and analysis seem

to suggest that there is much that we have yet to understand about teamwork and its evaluation.

Measurement

Our choice of the score-difference variable as the focus of measurement raises several important issues. First, our arbitrary choice of the score-difference variable was arbitrary, and was made for demonstration purposes. The intuition behind the selection of this immediately available task performance measure is that task-performance is associated with teamwork performance. But this is an assumption on our part (see more on this issue below).

Second, even if task performance is indeed correlated with teamwork, we are still left with the problem of selecting a measurement variable. Such a variable may be more or less sensitive to the effects of teamwork. For instance, we could have chosen to measure performance by the number of passes, etc.

Third, confidence in the measurement is a serious concern. For example, the analysis results above are not, for the most part, statistically significant. Although an Analysis-of-Variance (ANOVA) of the normalized score-differences shows that the mean score-difference of each phase are significantly different ($p=0.005$), this is likely due to the normalized mean of 0 (with 0 variance) in phase A. The number of data points (4) is simply not enough to give us a clear picture of each team' s response to the different evaluation phases. Although the trend suggested by the mean normalized goal scores supports our expectations of declining team performance, we need to investigate methods by which this analysis can be stated with greater confidence.

One way to do this would simply to repeat each test a relatively large number of times. For instance, if we repeat each of the phases for an evaluated team 40 times, we stand a greater chance of discovering statistically significant results. However, we should be aware that sometimes the surprising result is that even a great number of games may be insufficient to give us the level of confidence we want. Random effects in the environment and the unpredictable responses of the opponent, issues such as machine and network load, all interfere with our controlled conditions and make evaluation more difficult.

For instance, Tambe et al. (1999) report on a series of over 230 games of the ISIS' 98 team against fixed opponents in which the only independent variable was the use of communications (about half the games were played with communications allowed between the evaluated team' s players, and the rest were played with no communications). Despite the relatively large number of games, no statistically significant result was found. However, when a different measure was used, the average-time-to-agreement measure (ATA), very statistically significant differences were found between the communicating and non-communicating teams (the statistical significance levels exceeded 99%).

Another problem with measurement is that it may not make it clear to what extent changes in the performance (even when significant) are due to adaptation (or lack

thereof) on the part of the evaluated team, rather than other factors, such as adversarial re-planning (for instance, an advanced fixed opponent might recognize that the evaluated team-member is disabled, and adapt its attack to take advantage of this weakness). It may also be that limits of performance are reached -- if an evaluated team is very strong indeed, it may be successful in showing no change in performance despite its lack of adaptations. Here, agents are able to individually compensate for the lack of adaptation on the part of the team.

Perhaps a more productive, and certainly more challenging way of approaching this problem is to change the measures used. This paper demonstrated an initial form of using the score difference to analyze *team performance*. This measure has generally been used by researchers involved in RoboCup to evaluate their teams. However, it may very well be that this measure is simply insufficient for our purposes, and that techniques that measure *teamwork performance* more directly are more useful for our purposes. The ATA measure mentioned above (Tambe et al. 1999), for example, measures the average time it takes team-members to come to agreement on chosen team-plans (tactics). But other measures can be found in the literature. Balch (1998) for instance investigated the use of Social Entropy to measure diversity in soccer, and found a positive correlation between diversity and performance. Goldberg and Mataric (1997) suggest a different measure, based on inter-agent interference.

Analysis

The preliminary results in the previous section raise many issues in our underlying intuitions and knowledge of teamwork and coordination. For instance, we have introduced a theoretical ideal of perfect adaptation as 0-slope performance trend. In other words, we have chosen to look at no change in performance as a characteristics of a perfectly adaptive team. However, this really does not take into account theoretical limits that are likely to exist on the number of agents that are required (as a lower limit) or optimal (as an upper limit) for performing a team task. For instance, perhaps soccer cannot be done with less than 9 agents, in which case disabling more than 2 players will always result in a sloping trend, even with the best possible theoretical team.

Another issue is that we assume in our earlier analysis that degradation is linear. But this may or may not have basis in reality. It could be that, as some of the results may suggest, degradation in performance is non-linear with respect to the number of disabled players. For instance, some teams show an increase in performance when some players are disabled, and a decrease as others are disabled (Figure 1). These changes may be the result of the team moving back and forth from sub-optimal organizational structures to optimal ones as a result of our disabling of clients, but they may also reflect degradation phenomenon due to other factors.

We have mentioned above that our choice (for the example analysis) of a task-performance variable assumed that task performance is correlated with teamwork. But in actuality, such correlation between teamwork and performance is an open

research question. For instance, teams that have done well in the competition did not necessarily rank high in the analysis above (again, this may be due to the insignificance of the results). CMUnited'98 (Stone and Veloso 1998), which was the champion in the RoboCup-98 competition, ranks fairly low in terms of the estimated performance trend slope, while the leading team in those terms is the CAT Finland'98 (Riekki 1998) team, which proved to be a middle-level entry in terms of its placement in the competition. However, CMUnited'98 players, in particular, have been observed to send messages to the screen indicating that they correctly recognize that specific players have been disabled. The question remains why this recognition capability, which no other team was able to demonstrate, did not carry with it the implied compensation in terms of performance.

Rather than argue for the correctness of a particular technique and the implied evaluation results, we seek here to use our earlier analysis to demonstrate the possibility of evaluation on one hand, and the challenges involved in it on the other. The intent here is to stimulate discussion of evaluation methods, not of the obvious lacking of the one we have been using as a demonstration. The analysis provided earlier is a demonstration of a possible approach, and we warn again that its results are preliminary, and are not statistically significant.

Summary

The evaluation session and protocol were organized and executed with the intent of providing the multi-agent research community with a substantial data set that may be used for research in teamwork, adversarial planning, coordination, etc. The data collection and evaluation protocols were developed with this intent in mind, and we hope the research community will find them useful. However, they certainly can be improved upon. In particular, the data may not be statistically significant for certain tests, and the evaluation protocol itself can certainly be improved. We therefore welcome any and all comments and suggestions from investigators on how the evaluation methods and collected raw data may be improved.

The soccer-server log files collected during this evaluation session are to serve as the raw-data upon which the actual evaluation takes place. There are different ways of measuring the teams' performance, and these can be compared using this standard data repository. The logs present the data from the games of all 13 different participating teams, and are intended for use by the scientific community. They are publicly available for all interested parties (Repository, 1998). We hope useful, innovative measures will be developed and presented to the scientific community as a result of this data set. Perhaps most importantly, if RoboCup is truly to make an impact on AI and Robotics research, our investigations should result in evaluation techniques that generalize beyond RoboCup to other multi-agent domains.

Acknowledgements. We thank all participating teams, the RoboCup'98 Simulation-League committee (Chair: Itsuki Noda), Kay Schroter and Hans-Dieter Burkhard for their cooperation and help. Milind Tambe, Stacy Marsella, Nachum Kaminka, and

Shlomit Kaminka deserve many thanks for very useful comments.

References

Andou, T. (1998). Andhill-98: A RoboCup Team which Reinforces Positioning Observability. RoboCup-98, Paris, France.

Andre, D. and A. Teller (1998). Evolving Team Darwin United. RoboCup-98, Paris, France.

Balch, T. (1998). Behavioral Diversity in Learning Robot Teams. Ph.D. thesis, Georgia Institute of Technology.

Burkhard, H.-D., M. Hannebauer, et al. (1998). AT Humboldt - Development, Practice, and Theory. RoboCup-97: Robot Soccer World Cup I. H. Kitano, Springer. **LNAI 1395:** 357-372.

Burkhard, H.-D., J. Wendler, et al. (1998). AT Humboldt in RoboCup-98. RoboCup-98, Paris, France

Corten, E. and E. Rondema (1998). Team Description of the Windmill Wanderers. RoboCup-98, Paris, France.

Gemini, 1998. A Competing entry in RoboCup-98. No team description or other reference was available for this team.

Goldberg, D. and Mataric, M. (1997). Interference as a tool for designing and evaluating multi-robot controllers. In Proceedings of the 14th National Conference on Artificial Intelligence (AAAI-97), pp. 637-642.

Johnson, W. L. and J. Rickel (1997). Steve: An Animated Pedagogical Agent for Procedural Training in Virtual Environments. SIGART Bulletin. **8:** 16-21.

Kitano, H., M. Tambe, et al. (1997). The RoboCup Synthetic Agent Challenge ' 97 the International Joint Conference on Artificial Intelligence (IJCAI-97), Nagoya, Japan.

Lubbers, J. and R. R. Spaans (1998). The Priority/Confidence Model as a Framework for Soccer Agents. the Second RoboCup Workshop (RoboCup-98), Paris, France.

Maeda, K., A. Kohketsu, et al. (1998). Ball-Receiving Skill Dependent on Centering in Soccer Simulation Games. RoboCup-98, Paris, France.

Marsella, S. C., J. Adibi, et al. (1999). On Being a teammate: Experiences acquired in the design of RoboCup teams. the Third International Conference on Autonomous Agents (Agents-99), Seattle, WA, ACM Press.

Montesello, F. and E. Pagello (1998). PaSo-Team' 98: Learning the ' when' RoboCup Competition. RoboCup-98, Paris, France.

Polani, D., S. Weber, et al. (1998). A Direct Approach to Robot Soccer Agents: Description for the Team MAINZ ROLLING BRAINS Simulation League of RoboCup' 98RoboCup-98, Paris, France.

Repository, (1998). The RoboCup Teamwork Evaluation Homepage. **1998**. At: http://www.isi.edu/~galk/Eval/

Riekki, J. (1998). Reactive Task Execution of a Mobile Robot. Infotech Oulu and Department of Electrical Engineering. Oulu, Finland, University of Oulu. (Ph.D. Dissertation)

Stone, P. and M. Veloso (1998). The CMUnited-97 Simulator Team. RoboCup-97: Robot Soccer World Cup 1, LNAI 1395. H. Kitano, Springer**:** 389-397.

Tambe, M., W. L. Johnson, et al. (1995). "Intelligent Agents for Interactive Simulation Environments." AI Magazine **16**(1).

Tambe, M., G. A. Kaminka, et al. (1999). Two Fielded Teams and Two Expert Agents: A RoboCup Challenge Response from the Trenches. International Joint Conference on Artificial Intelligence (IJCAI-99), Stockholm, Sweden.