# Performance Competitions as Research Infrastructure: Large Scale Comparative Studies of Multi-Agent Teams

Gal A. Kaminka (galk@cs.cmu.edu)
*Computer Science Department, Carnegie-Mellon University, Pittsburgh, PA 15213, U.S.A*

Ian Frank (ianf@fun.ac.jp)
*Future University-Hakodate, 116-2 Kamedanakano, Hakodate-shi, Hokkaido, 041-8655 Japan*

Katsuto Arai (karai@cogsci.l.chiba-u.ac.jp)
*Department of Cognitive and Information Sciences, Faculty of Letters, Chiba University, Japan*

Kumiko Tanaka-Ishii (kumiko@ipl.t.u-tokyo.ac.jp)
*Interfaculty Initiative in Information Studies, Graduate School of Interdisciplinary Information Studies, University of Tokyo, Tokyo, Japan*

**Abstract.** Performance competitions (events that pit many different programs against each other on a standardized task) provide a way for a research community to promote research progress towards challenging goals. In this paper, we argue that for maximum research benefit, any such competition must involve *comparative studies* under closely controlled, varying conditions. We demonstrate the critical role of comparative studies in the context of one well-known and growing performance competition: the annual Robotic Soccer World Cup (RoboCup) Championship. Specifically, over the past three years, we have carried out annual large-scale comparative evaluations—distinct from the competition itself—of the multi-agent teams taking part in the largest RoboCup league. Our study, which involved 30 different teams of agents produced by dozens of different research groups, focused on *robustness*. We show that (i) multi-agent teams exhibit a clear performance–robustness tradeoff; (ii) teams tend to over-specialize, so that they cannot handle beneficial changes we make to their operating environment; and (iii) teams improve in performance more than in robustness from one year to the next, despite the emphasis by RoboCup organizers on robustness as a key challenge. These results demonstrate the potential of large-scale comparative studies for producing important results otherwise difficult to discover, and are significant both in the lessons they raise for designers of multi-agent teams, and in understanding the place of performance competitions within the multi-agent research infrastructure.

## 1. Introduction

Recent years have seen an explosion in multi-agent research, spanning both theory [17, 12, 23] and practice [39, 33, 41, 32, 15, 24]. But to keep pace with this progress, evaluation techniques must also evolve. Ideally, new techniques should be evaluated empirically not only in absolute, functional terms (e.g., via a lesion study), but also in comparison to existing techniques and the state of the art, via *comparative* studies. In particular, comparative studies are important for the identification of the relative strengths and weaknesses of competing claims, since they can evaluate systems across a wide variety of controlled conditions

However, comparative studies can be difficult to conduct. For example, researchers wanting to compare their own techniques to those published in the literature often find they need to reconstruct the state-of-the-art from the published descriptions. One way to transfer the burden of conducting empirical evaluation from the individual to the community is through standard benchmarks and performance competitions [14, 21, 28], where researchers are invited to submit programs to carry out a well-defined task. But the designers and maintainers of such contests face the following twin questions: "What should the performance task be?" and "How should performance be measured?"

In this paper, we argue that for maximum research benefit the answer to these questions must involve comparative studies under closely controlled, varying conditions. To demonstrate this, we report our experience with a new and growing performance competition: the simulation league of the RoboCup Robotic Soccer World Cup Initiative (henceforth, RoboCup). Independently of the competition itself, we have set up and carried out a large-scale comparative evaluation effort, in which we invited all the RoboCup teams to participate. To the best of our knowledge, the scale of the resulting evaluations is unprecedented. Over three years, we rigorously tested 30 different teams of agents—the research products of dozens of different research groups.

In particular, we focussed on the broad concept of *robustness*. This is a general property of particular interest to multi-agent systems researchers, covering notions of 'fault-tolerance', 'graceful degradation', 'adaptability', and the ability to maintain performance in the presence of *unanticipated* domain changes [4, 44]. In multi-agent research, robustness has been a key motivation for much of the work on collaboration [10, 18, 38, 20, 22] and coordination [16, 37]. Robustness has also been emphasized as a key challenge area for RoboCup multi-agent teams [21].

This paper reports the controlled experiments we conducted to test RoboCup teams for robustness in the face of unanticipated domain changes such as the disablement of players and changes in the uncertainty of perception information and in the reliability of actuators. We analyze the collected data using an advanced measurement tool, the *Statistics Proxy Server* [43], which provides measurements of close to 40 different team behavior features. We use a novel application of linear regression to quantitatively estimate the robustness of a team along any one of these behavioral features. This analysis reveals the following conclusions:

- **There is a performance-robustness tradeoff**. The teams that perform best in playing soccer under normal conditions tend to degrade less gracefully in the face of changes in the environment. That is, performance is negatively correlated with robustness.
- **Over-specialization is a problem**. Most teams over-specialize, i.e., their performance degrades even in the face of beneficial changes we make to their operating environment.
- **The pace of improvement in robustness is slow**. Despite the explicit emphasis on robustness by RoboCup organizers [21], teams clearly improve in soccer performance from one year to the next, but show little improvement in robustness.

These conclusions would have been difficult to reach using either lesion studies or analysis of the competition results. Instead, the evaluation sessions take advantage of the scale offered by an existing performance competition, and incorporate an extra element of comparative design that carefully defines and uses control and test phases. Thus, our results produce lessons both for the designers of multi-agent teams and also for those who promote research evaluation via performance competitions.

Specifically, in terms of the question "What should the performance task be?" our results suggest that performance competitions are most effective when they involve comparative studies, and that care should be taken to avoid biasing the research effort of participants towards too narrow a goal. Also, in terms of "How should performance be measured?" we believe the in-depth analysis techniques we present here are a significant first step. However, as a secondary contribution, we further make all our evaluation data publicly available [19]. We hope that this repository will encourage researchers to experiment with new measures and analysis methods that address this challenge.

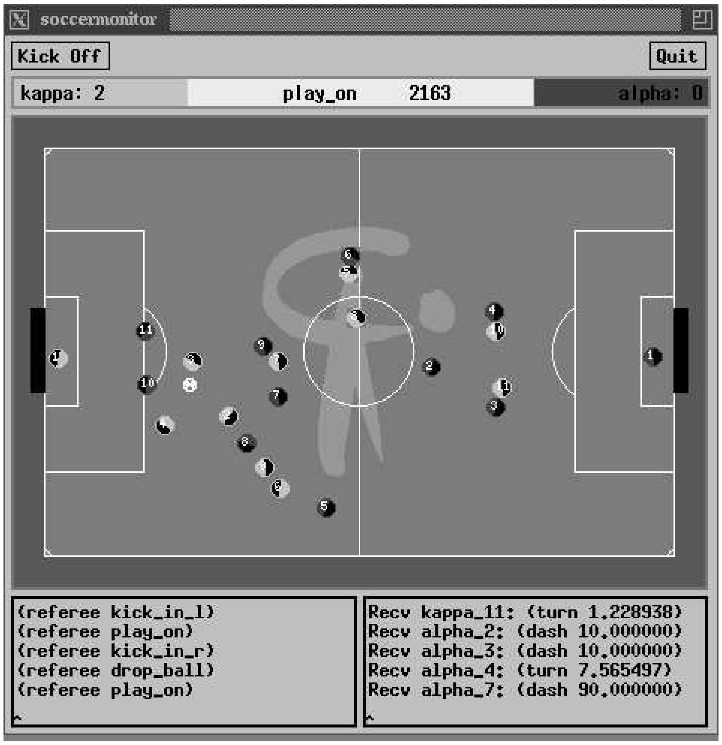
The remainder of this paper is organized as follows. In Section 2 we give some background on RoboCup, and then describe our evaluation sessions in detail in Section 3. In Section 4, we explain our measurement techniques and introduce the Statistics Proxy Server. In Section 5 we present our results. Finally, we discuss related work and emerging challenges in Section 6, and then conclude in Section 7.

## 2. RoboCup: A Short Overview

A robotic soccer team capable of beating the human world champions may seem like a pipe dream. But this is the challenge goal at the heart of the Robot Soccer World Cup (RoboCup) Games and Conferences. Since the early 1990's, researchers have been working towards establishing soccer as an effective multi-disciplinary test-bed for evaluating competing theories, algorithms, and architectures for robotics, learning and multi-agent systems [21]. Most prominently, this effort has led to the establishment of an annual Robot Soccer World Cup tournament.

The Robot Soccer World Cups began in 1997, when RoboCup-97 was staged in conjunction with IJCAI-97 (the International Joint Conference on Artificial Intelligence), in Nagoya, Japan. In this inaugural event, 39 teams participated in three leagues (two robotic leagues and one simulation league), and a parallel workshop featured over 20 paper presentations. One year later, RoboCup-98 was held in Paris at the same time as ICMAS-98 (the International Conference on Multi-agent Systems) and the 1998 FIFA Soccer World Cup. Next, RoboCup-99 was staged in conjunction with IJCAI-99 in Stockholm, Sweden, and the 2000 contest was held in conjunction with PRICAI (The Pacific Rim Conference on Artificial Intelligence) in Melbourne. By 2000, the tournament had grown to include 90 teams, and also included a new "legged robot league" based on Sony's AIBO pet dog robots.

Here, we restrict ourselves to giving an overview of the league that forms the basis for our evaluation sessions: the simulation league. The basis of this league is the *Soccer Server* platform [30, 29], which models a virtual soccer field populated by two teams of 11 players (see Figure1). In a game, the actions of each player are controlled by a separate client program that connect independently to the Soccer Server. All in all, 22 client programs, controlling 11 players on each team, are connected the soccer server during a game. The server carries out the commands issued by the clients and acts as a referee. It also sends the clients relevant sensor information such as stamina, vision, and auditory information. Vision and communication are limited by the physical constraints of the game (e.g., players cannot see behind themselves, and cannot hear their teammates' shouts at long distances). The server simulates uncertainty in the environment (e.g., wind), in sensing (e.g., in the ability to identify the ball's position), and in acting (e.g., in kicking the ball at a specified angle).



```
(referee kick_in_l)
(referee play_on)
(referee kick_in_r)
(referee drop_ball)
(referee play_on)
```

```
Recv kappa_11: (turn 1.228938)
Recv alpha_2: (dash 10.000000)
Recv alpha_3: (dash 10.000000)
Recv alpha_4: (turn 7.565497)
Recv alpha_7: (dash 90.000000)
```

*Figure 1.* Screen capture of a game played on the RoboCup Soccer Server.

The simulation league allows researchers to develop soccer-playing teams free from concerns over issues such as the interpretation of real-world scenes and robot motor control. In practice, every year has seen approximately 40 different research groups focus very effectively on questions such as controlling dynamic multi-agent systems (e.g., [40]), and multi-agent learning (e.g., [34]). Another significant benefit of working with the simulation league is that the logistical problems of physically setting up games between teams—and ensuring that all players are functioning as intended—is far easier than with their robotic counterparts. In particular, the Soccer Server is easily parameterized to allow for changes to the operating conditions of teams.

## 3. The RoboCup Evaluation Sessions

The winners of the RoboCup competitions are decided based solely on the performance of teams in tournament matches. Unfortunately, this does not easily permit rigorous evaluation. For example, the overall task performance (determining the winners, and a ranking of the participants) takes precedence over a careful study of the relative abilities of the participating teams across the spectrum of tasks required to perform well at soccer. Also, different teams meet different opponents, and each team frequently undergoes significant modification over the course of the RoboCup event. Therefore, luck plays a role in the team's success. Indeed, even for real soccer teams, Wagenaar [45] has shown that luck plays a large part in determining the team that emerges as the tournament winner. Thus overall, the competition results may mask interesting lessons simply because of bad luck, or by focussing attention on teams that win rather than on the techniques that allow teams to excel at sub-tasks such as individual control, communications, collaboration, or planning.

To address this challenge, we organized the annual RoboCup simulation league evaluation sessions, which have taken place in each of the last three years. Rather than focusing on the single goal of "winning" games or a tournament, these evaluation sessions provide large-scale, rigorous comparative evaluations of participating teams, under carefully controlled conditions. To the best of our knowledge, the scale of these evaluations is unprecedented: close to 40 different teams of agents, the research results of dozens of research groups, have been evaluated in these three years. This undertaking presents multi-agent researchers with novel opportunities, lessons and challenges, which we highlight in the remainder of this paper.

### 3.1. Administering the Evaluation Sessions

Robustness is a key issue in multi-agent research [18, 38, 44, 20]. Indeed, the IJCAI 1997 RoboCup Synthetic Agent Challenge identifies robustness as one of the key teamwork challenge areas [21]. Therefore, we designed our experiments to test this property. The key here was to evaluate teams based on their reactions to unanticipated changes to the environment—changes that go beyond the task of playing soccer, and beyond the normal operating conditions of the simulated environment. Our evaluation sessions conducted controlled experiments that tested simulated soccer teams for robustness in three specific areas: (i) the ability to handle disabled team-members; (ii) the ability to handle varying levels of uncertainty in perception; and (iii) the ability to handle varying levels of uncertainty in the effects of actions.

The evaluation sessions themselves were carried out during the actual competitions, with the full support of RoboCup organizers. This enabled us to use the existing equipment set up for the competitions. Participation in the sessions was open to all, but was mandatory for all teams reaching the last 16 of the competition (the top to medium teams, in terms of soccer performance levels). Each year, the specific nature of the tests was kept secret from the research groups until the evaluation sessions began, and researchers were then forbidden from modifying their software in any way to accommodate the test. However, each test was repeated in the following year, to allow researchers to respond to the challenges of the evaluation tests with special versions of their software. Thus, each year's evaluation session consisted of two tests: one secret, and one known in advance (repeating the test from the previous year).

Every participating team is evaluated against the same fixed opponent, under identical conditions. Each test has several phases. The first phase of each test is always played under normal competition rules and conditions, to establish a performance base-line, i.e., control data. In the remaining phases, changes are introduced incrementally to the simulated world, and the team again plays against the same fixed opponent. The number of phases depends on the particulars of the test, and is also limited by pragmatic constraints (each team can take up to 10–15 minutes for each phase). However, the intention is always to provide sufficient data to explore the extremes of the phenomena being tested.

Crucially, the changes introduced in each test phase affect only the evaluated team. This prevents contamination of the results by changes in the behavior of the fixed opponent. For example, if a test phase increases the perception uncertainty (see Section 3.3), only the players on the team being evaluated experience this change; the players on the team of the fixed opponent maintain their normal perception capabilities. Below, we describe the design of each evaluation test.

### 3.2. The Disabled Players Test (1998 & 1999)

The motivation for this evaluation protocol was to test a team's robustness to the disabling of some of its members. Disabled players were left on the field in their initial position at the beginning of the game, but unable to move, kick the ball, or communicate. However, they were still visible to their teammates and opponents. All participating teams competed against a fixed opponent (the 1997 World Champion AT Humboldt97 [8]). Each team played four half-games against the fixed opponent. Each such half-game constituted an evaluation phase, in which a single change to the number of disabled players was made:

- **Phase A.** The control phase. No players were disabled.

- **Phase B.** The team played with a single player disabled (randomly chosen).

- **Phase C.** The team played with two players disabled: the same player selected in phase B, and a player selected for its importance to the evaluated team (most valuable player).

- **Phase D.** The team played with three players disabled: the two players disabled in phase C and the evaluated teams goalie.

Ideally, we would like to have disabled precisely the same players for each team in phases B, C and D, to ensure evaluation under the exact same conditions. However, since different teams use players differently, it is actually impossible to find a player whose role is common to all teams. Even the goalie, whose role was most popular, was in fact being replaced on a regular basis by at least one team [1]. To limit biasing in the test, we thus sought to include both a random element (phase B), as well as elements that touch on deeper understanding of the evaluated team's usage of roles or coordination (phases C and D).

### 3.3. The Sensory Uncertainty Test (1999 & 2000)

Motivated by robotics literature, which often mentions the uncertainty associated with sensor readings, this test examined team behavior under varied levels of perception uncertainty. The RoboCup simulation already includes perceptual uncertainty, but at a given level. We varied this uncertainty level by a multiplicative factor, causing individual players (only on the evaluated team) to perceive the world around them with greater or smaller uncertainty. The fixed opponent in this test was the CMUnited-98 team [36]. For this test, there were three phases:

- **Phase A.** The control phase.

- **Phase B.** The team played with *no perception uncertainty*, i.e., with improved conditions compared to normal operating conditions.

- **Phase C.** The team played with much increased perception uncertainty levels (12 times the normal level).

Including Phase B in this test allowed us to investigate not only whether teams were able to handle degradation in their operating environment, but also whether they were over-specialized, so that they could not even handle beneficial changes in the environment.

### 3.4. The Action Uncertainty Test (2000)

To complement the perception uncertainty test, in 2000 we introduced an actuator uncertainty test to vary the levels of uncertainty in the *actions* of each player on the evaluated team. Already, the Soccer Server adds Gaussian noise to the parameters of the commands issued by each team's players. In this evaluation test, we used a simple multiplicative factor to vary this noise level. The fixed opponent was the CMUnited-99 team [35], and the three phases were:

- **Phase A.** The control phase.
- **Phase B.** The team played with twice the normal uncertainty level.
- **Phase C.** The team played with five times the normal uncertainty level.

In contrast to the perception uncertainty test in the previous section, where we looked at the two extremes of "no uncertainty" and "much uncertainty" to examine over-specialization, here we were more interested in exploring the degradation curve as the operating conditions of the environment slowly become worse.

## 4. Measurement and Analysis Methodology

Analyzing the logs of the evaluation sessions presents two challenges: (i) turning the raw player movement data into quantitative measures that highlight significant team behavior features; and (ii) formalizing a method for assessing robustness based on these measures. Here, we describe our solution to each of these problems.

### 4.1. The Statistics Proxy Server

There are three obvious quantitative measures that 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 these two. These measures are directly associated with a team's overall performance on the task, but convey no information on a team's style, tactics and strategies, or learning and adaptation capabilities.

To provide more detailed information on team performance, we have developed a sophisticated analysis tool, the Statistics Proxy Server [43], which evaluates team behavior along dozens of measures. This Statistics Proxy is based on the MIKE commentator system [42, 26], which tracks and interprets a game of soccer using six Soccer Analyzers. The complete repertoire of over 50 statistics produced by the Statistics Proxy includes many that are directly related to team play (summarized in Table I). We indicate which statistics we use in our analysis (the ones most appropriate for drawing lessons about robustness—see discussion below), as well as others that may be useful to researchers interested in conducting their own analysis of the data [19].

Most of these statistics should be fairly self-explanatory. For instance, the *XAvr* statistic measures the average position of all the players on a team, expressed as a distance (positive or negative) past the centerline. However, two notions that might not be familiar are *winning pass-work patterns* and *compactness*. The first of these is defined as any chain of three players from the same team $A$, $B$, $C$, such that $A$ has passed to $B$ once or more, $B$ has passed to $C$ once or more, and $C$ scores at least one goal. The second is designed to measure the way that (at least in real soccer) the entire formation of a team follows the movement of the ball across the field. We define the *compactness* of a team as the X-distance between its front-most player and its rear-most player (excluding the goalie). Teams with lower compactness tend to be more dynamic, and benefit from having more players closer to the ball.

### 4.2. A Measure of Robustness

Our abstract definition of robustness is the ability of a team to maintain its performance in the face of unanticipated changes to its operating conditions. For instance, when a team suddenly has players disabled, there is a direct effect on the team's passing ability. In order to maintain its overall performance (e.g., the score difference at the end of the game), a robust team may adapt by modifying its passing patterns to exclude any disabled players, or by having players dribble the ball more than before.

We therefore choose to measure robustness by plotting a performance degradation curve showing how an evaluated team's performance changes as its operating conditions change, i.e., as players are disabled or as sensory uncertainty is changed. Then, we use linear regression to approximate the general slope of the curve. More robust teams will have graphs with smaller (absolute values) of slopes, since this would signify smaller changes to the team's performance despite the changes to its operating conditions. The following example illustrates this process.

In Figure 2, we have plotted the results of two RoboCup teams (CMUnited-98 [36], and ISIS-98 [40]) on the 1998 disabled players test.

Table I. Team performance measures computed by the Statistics Proxy Server, and whether they are used for computing robustness in the Disabled Players (DP) Test, and the Sensory Uncertainty (SU) Test. Note that the log files for the most recent Action Uncertainty Test are still being verified, so we have not yet tested them, and there is no AU column in this table.

| Command | Explanation | DP test | SU test |
|---|---|---|---|
| *S*core | score | ● | ● |
| *X*Avr | average X-location of all players (m) | | |
| *Y*Avr | average Y-location of all players (m) | | |
| *X*Var | deviation all players' X-locations (m) | | |
| *Y*Var | deviation of all players' Y-locations (m) | | |
| *P*osession | possession rate of a team | ● | ● |
| *B*allAtEachSide | territorial advantage (ball in opposition half) | ● | ● |
| *P*assLenAvr | pass length average(m) | | |
| *D*ribbleLenAvr | dribble length average(m) | | ● |
| *P*assLenVar | pass length deviation(m) | | |
| *P*assChainNum | number of pass chains | | |
| *P*assBackNum | number of backwards passes | | |
| *P*assChainPlayerNumAvr | average number of players in one pass chain | | |
| *P*assChainLenAvr | pass chain average length(m) | | |
| *P*assNum | total number of passes | | |
| *P*assLongNum | long passes (> 12m) | | ● |
| *D*ribbleNum | total number of dribbles | | |
| *S*hootNum | total number of shots on goal | ● | ● |
| *S*tealNum | total number of steals | | ● |
| *C*ompactnessAvr | average compactness (m) | | |
| *C*ompactnessVar | variance of compactness ($m^2$) | | |
| *B*allPlayerDisAvr | average distance of ball and players (m) | | |
| *B*allPlayerDisVar | deviation of distance of ball and players (m) | | |
| *W*inningPassPatternNum | number of winning passwork patterns | | ● |
| *I*nactivePlayerNum | number of players making zero passes | | |
| *P*assSuccessRate | average pass success rate | | ● |
| *O*ffSideNum | total number of off sides | ● | ● |
| *F*reeKickNum | total number of free kicks | ● | ● |
| *K*ickInNum | total number of kick ins | ● | ● |
| *C*ornerKickNum | total number of corner kicks | ● | ● |
| *G*oalKickNum | total number of goal kicks | | ● |
| *D*istanceAvr | average distance covered by one player (m) | | |
| *S*hootSuccessRate | average shoot success rate | ● | ● |

Here, the x-axis shows the number of disabled players in the different evaluation phases, and the y-axis shows a very simple measure of team performance: the score-difference at the end of each evaluation game. In addition to the plots of each team's performance (thin lines), we have also shown the line fitted to the data points by linear regression (thick lines). In terms of soccer *performance*, the CMUnited-98 team is superior to the ISIS-98 team, as it scores higher against the same fixed opponent. However, if we consider instead the slope of the linear regression lines, the superiority is reversed. The slope of the line for CMUnited-98 is larger, reflecting the team's inability to maintain its level of performance as well as ISIS-98.

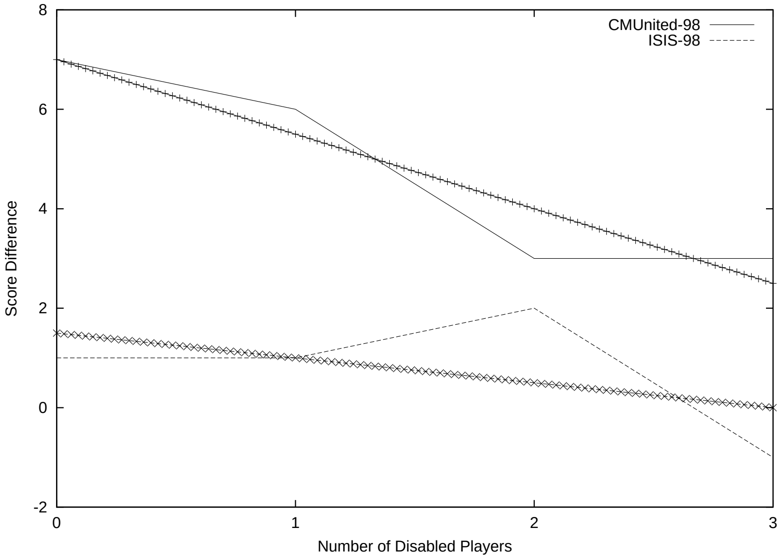


*Figure 2.* Score-difference trends for two 1998 teams

For each of the measures produced by the Statistics Proxy (plus some other more straightforward measures such as score-difference), we plot the values of the measure in question against the controlled evaluation conditions, e.g., Possession percentage vs. number of disabled players, or score-difference vs. uncertainty factor. We then use linear regression to produce a line that best represents the performance trend of the evaluated team for that measure. The slope of the fitted line for each team signifies its robustness: the closer its value to zero, the more robust the team. For a perfectly robust team, we would expect to see a horizontal line: a zero-slope performance trend.

Note that a team can respond in various ways to changes in its operating conditions. For instance, as previously discussed, a team may respond to a disablement of a player by having the remaining player dribble more. On the other hand, we would not expect a team to deal with disabled players by striving to decrease the number of shots it makes on the opponents' goal. Thus, when we examine the slopes generated for different measures by the linear-regression process described above, we must be careful. For certain measures, a large slope may indicate a change in a team's behavior that signifies lack of robustness (for instance, a change in the result of the game). But for other measures, a large slope may indicate a change that is a result of a team adapting itself to the change in its environment.

We therefore restrict our attention to measures that we expect to be stable in an ideal (perfectly robust) team. These depend on the test. For each of the tests we conducted, we chose the appropriate measures guided by discussions with RoboCup researchers. The selected measures are shown in Table I. For example, the goalie is one of the players that is disabled in the Disabled Players test, so we cannot use the *GoalKickNum* measure, since this would be expected to qualitatively change. In the Sensory Uncertainty test, we can include more measures related to pass-work, since the changes to the domain have less impact on the roles of individual players. Restricting the measures we consider in this way and using our linear regression technique to quantify robustness allows us to carry out a wide and detailed analysis.

## 5. Results: Robustness and Performance

This section presents the results of measuring robustness and performance in the tests described in the previous sections. We draw conclusions regarding the relationship between robustness and performance, and the design choices made by RoboCup researchers. Our intent here is to demonstrate that the evaluation sessions provide a basis for a concrete exploration of such questions.

### 5.1. Robustness-Performance Tradeoffs

We first examine the relationship between robustness and task performance. As described above, we use the Statistics Proxy to generate measurements, and then apply linear regression to generate a slope describing each team's change in behavior for all of the appropriate measures. We do this for each of the three different evaluation tests (Disabled Agents 1998, 1999 & Sensory Uncertainty 1999; the 2000 data is still being analyzed).

As an example, Table II presents the raw results and the computed slopes for the 1998 champion team, CMUnited-98 [36], in the disabled-players test conducted in 1998. The first column shows names of the measures output by the Statistics Proxy. We show all 33 measures, plus the number of disabled players and the score difference. The second column shows the value computed for each measure in the control phase (no disabled player). This value is used to rank the team's performance. The next three columns then show the values of each measure as progressively more players are disabled. Thus along each of the measures (rows), one can plot a curve that shows how the measured behavior of the team changes with the disabling of players. The computed slope of this curve is shown for each measure in the last column.

We can examine the overall relationship between robustness and performance by computing the correlation (over all teams) between the control value of a measure (which represents the team's performance under normal conditions) and the slope that represents robustness. Table III shows the correlation computed in this way for each of the appropriate measures in the disabled players test of 1998 (13 teams), the disabled players test in 1999 (17 teams) and the sensory uncertainty test in 1999 (16 teams). Each correlation value is between -1.0 and 1.0. Here, a negative value indicates that robustness improves with performance (since the linear regression slope decreases with increasing robustness). A positive value means that robustness decreases with performance.

These results, which cover all medium- and top-level RoboCup teams, raise an important lesson: *teams trade robustness for performance.* While not all measures show this trend, for many of the key measures, the correlation figures are large and positive. Indeed, a graphical plot of these measures often shows a clear trend: as performance improves, robustness worsens. However, to inspect the results graphically, we must first apply a normalizing factor to compare performance and robustness on a uniform scale. To do this, we rank the teams twice (and separately for each measure) based on performance and on robustness. That is, for any given measure, the team with the best (largest) performance result is ranked 1 in performance. The second best team is ranked 2, and so on, until the worst team is assigned the largest rank. Similarly, the team with the best robustness (smallest slope) is ranked 1 in robustness, the second smallest slope is ranked 2, and so on.

Figures 3–5 plot the team performance and robustness rankings produced in this way, for a number of different measures. The vertical axis in these figures shows the ranking, and the horizontal axis denotes the teams, which are sorted by the performance ranking from left to right. As we move from left to right, the monotonically increasing curve is the performance ranking curve, which indicates worsening performance from left to right. The other curve (often decreasing from left to right) shows the robustness ranking for the same teams.

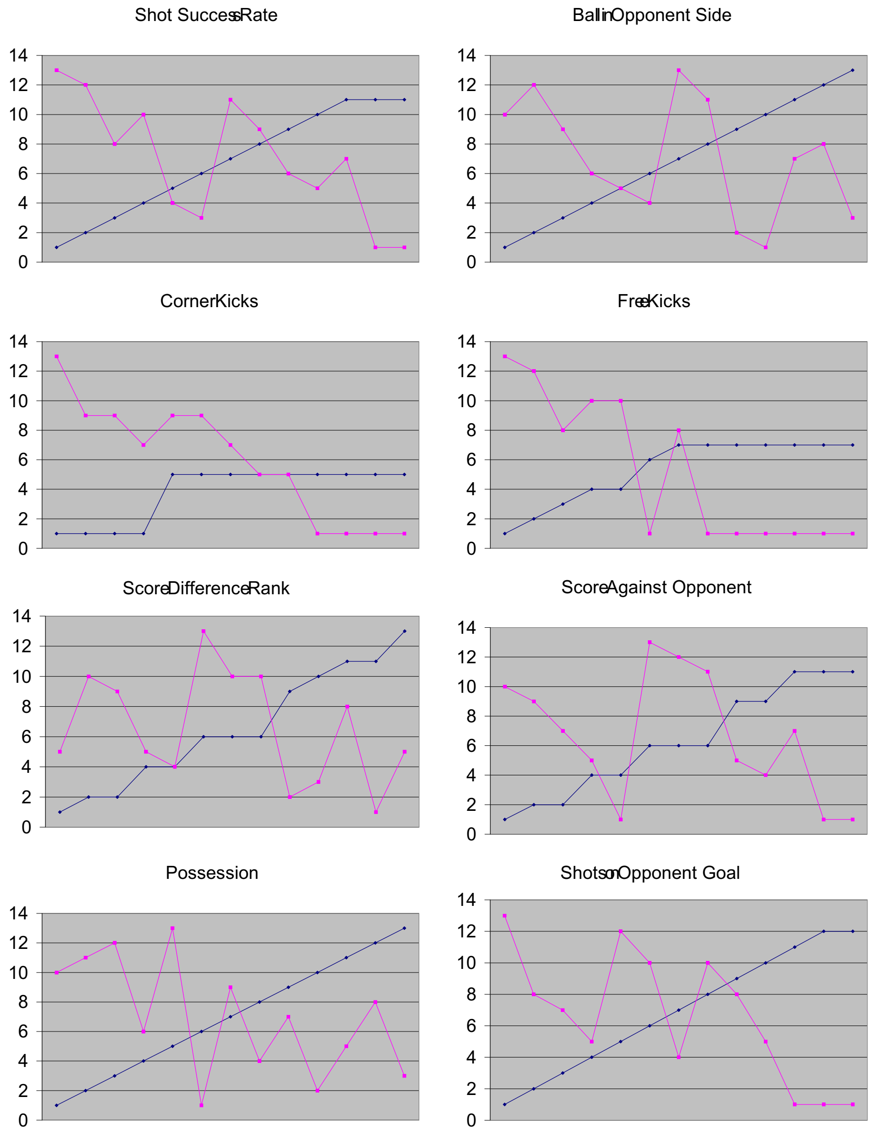
Table II. Analysis results for CMUnited-98, disabled-players test, 1998.

| Measure | Control (A) | Phase B | Phase C | Phase D | Slope |
|---|---|---|---|---|---|
| Num Disabled Players | 0 | 1 | 2 | 3 | |
| Score Difference | 7 | 6 | 3 | 3 | -1.5 |
| XAvr | 1.436 | -2.796 | 8.633 | 8.009 | 3.1149 |
| YAvr | 2.024 | -2.335 | 3.471 | 3.233 | 0.9433 |
| XVar | 23.64 | 25.813 | 22.496 | 22.664 | -0.6248 |
| YVar | 13.79 | 16.76 | 16.055 | 14.681 | 0.1978 |
| Posession | 63.59 | 72.71 | 70.418 | 65.836 | 0.4431 |
| PassNum | 30 | 26 | 28 | 13 | -4.9 |
| PassLenAvr | 11.87 | 12.885 | 15.464 | 18 | 2.098 |
| PassLenVar | 5.30 | 6.733 | 9.2909 | 11.92 | 2.2399 |
| PassLongNum | 7 | 6 | 11 | 3 | -0.7 |
| PassBackNum | 6 | 6 | 8 | 5 | -0.1 |
| PassSuccessRate | 80.70 | 79.365 | 69.118 | 73.81 | -3.0924 |
| ShootNum | 14 | 8 | 4 | 4 | -3.4 |
| DribbleNum | 9 | 9 | 8 | 8 | -0.4 |
| StealNum | 21 | 19 | 19 | 11 | -3 |
| Score | 7 | 6 | 3 | 3 | -1.5 |
| ShootSuccessRate | 50 | 62.5 | 75 | 75 | 23.75 |
| DistanceAvr | 690.34 | 559.97 | 545.23 | 382.16 | -93.929 |
| PassChainNum | 6 | 5 | 8 | 5 | 0 |
| PassChainPlayerNumAvr | 3.83 | 4 | 3.875 | 3.4 | -0.1425 |
| PassChainLenAv | 37 | 53 | 54.875 | 37.2 | 0.2475 |
| DribbleLenAvr | 14.33 | 16.556 | 20 | 13 | -0.0556 |
| InactivePlayerNum | 1 | 2 | 2 | 5 | 1.2 |
| CompactnessAvr | 46.72 | 62.3 | 62.092 | 58.69 | 3.5688 |
| CompactnessVar | 8.24 | 10.98 | 9.9237 | 12.077 | 1.04514 |
| BallPlayerDisAvr | 29.67 | 31.15 | 30.031 | 35.05 | 1.5 |
| BallPlayerDisVar | 19.31 | 20.59 | 18.097 | 22.008 | 0.5595 |
| WinningPassPatternNum | 2 | 3 | 1 | 0 | -0.8 |
| BallAtEachSide | 67.03 | 63.1 | 39.263 | 71.31 | -1.1007 |
| CornerKickNum | 0 | 1 | 0 | 0 | -0.1 |
| GoalKickNum | 0 | 0 | 0 | 0 | 0 |
| FreeKickNum | 1 | 1 | 4 | 0 | 0 |
| KickInNum | 0 | 1 | 0 | 3 | 0.8 |
| OffSideNum | 1 | 2 | 1 | 0 | -0.4 |

Table III. Performance/Robustness correlation (over all teams).

| Measure Name | Disabled 1998 | Disabled 1999 | Sensory 1999 |
|---|---|---|---|
| Score Diff | 0.4354 | 0.0579 | 0.7437 |
| ShootNum | 0.5859 | 0.8475 | 0.8898 |
| Score | 0.3913 | 0.4298 | 0.9852 |
| ShootSuccessRate | 0.7145 | 0.5499 | 0.9717 |
| CornerKickNum | 0.6465 | 0.9203 | 0.4236 |
| FreeKickNum | 0.9854 | 0.9393 | 0.8502 |
| KickInNum | 0.0855 | 0.9383 | 0.4036 |
| OffSideNum | 0.8023 | 0.5960 | 0.7078 |
| Possession | 0.4400 | 0.0375 | -0.0476 |
| BallAtEachSide | 0.3398 | 0.1111 | 0.8191 |
| GoalKickNum | - | - | 0.6068 |
| DribbleLenAvr | - | - | 0.2697 |
| StealNum | - | - | 0.3187 |
| WinningPassPatternNum | - | - | 0.9351 |
| PassLongNum | - | - | 0.5003 |
| PassNum | - | - | 0.4267 |
| PassSuccessRate | - | - | 0.3046 |
| PassChainLenAvr | - | - | -0.2145 |
| PassChainPlayerNumAvr | - | - | -0.2147 |
| DribbleNum | - | - | 0.2486 |
| PassChainNum | - | - | -0.0937 |
| DistanceAvr | - | - | 0.1041 |
| PassLenAvr | - | - | 0.3185 |
| Average | 0.5099 | 0.5352 | 0.4207 |

These figures demonstrate the trade-off between improved performance and improved robustness. That is, the best soccer-playing teams are typically not robust, and the most robust teams often don't play as well as others. For instance, Figure 5 shows the results for the sensory uncertainty test. In particular, the score-difference rankings (third row, left column) show very clearly that as the performance rankings grow to the right (i.e., we look at teams with lower performance), robustness improves. While the number of games is an issue here (each team played each phase of the evaluation test once), we note that that the same trends are repeated for the majority of the measures, despite the measures not being directly dependent on each other. And while some of the performance measures (such as kick-ins) may be noisy, others


Shot Success Rate

Ball in Opponent Side

Corner Kicks

Free Kicks

Score Difference Rank

Score Against Opponent

Possession

Shots on Opponent Goal

*Figure 3.* Plots of robustness vs. performance in the disabled players test, 1998

(such as average dribble length or possession) aggregate dozens of data points collected through the game.

In our discussions of this robustness result with other researchers, we sometimes hear the comment that a negative correlation between performance and robustness is not surprising[1]. The typical argument

[1] We find that reaction is fairly evenly split: either the result completely fails to surprise, or it surprises completely.

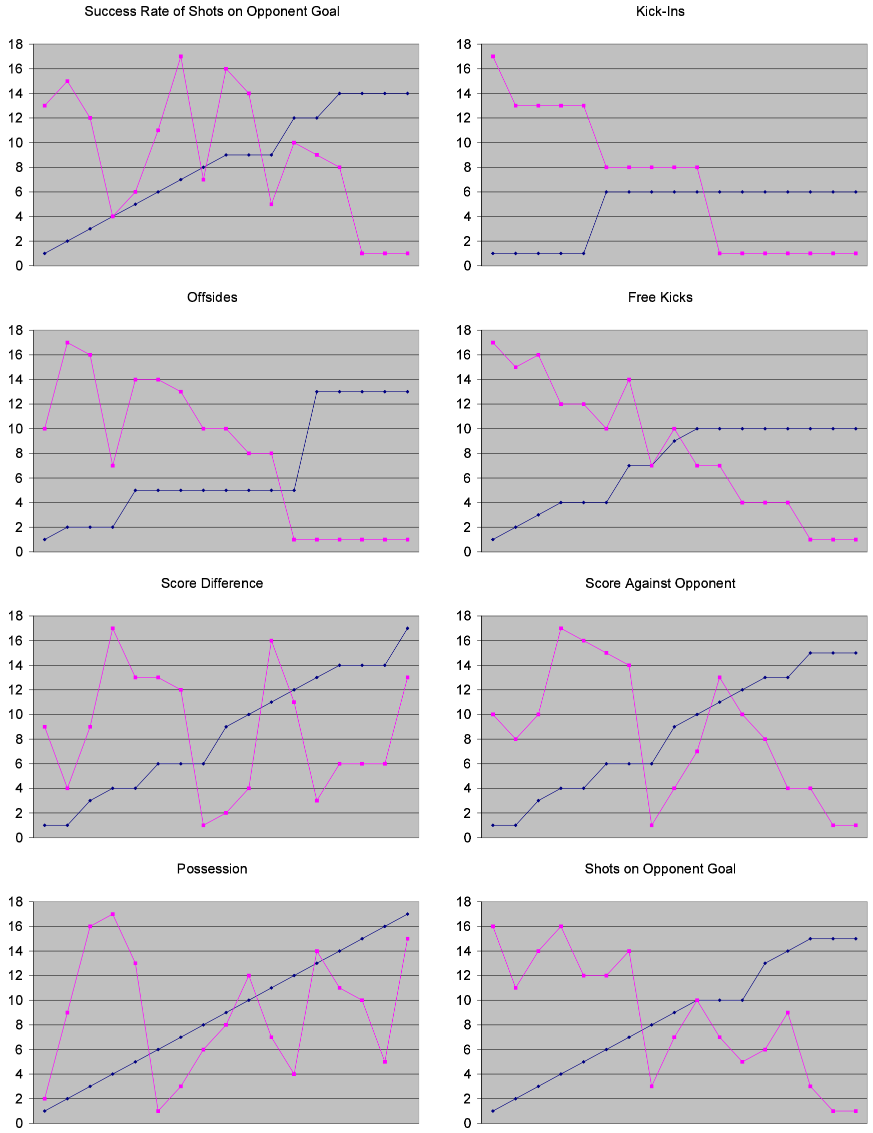


*Figure 4.* Plots of robustness vs. performance in the disabled players test, 1999

is that since better teams have more to lose, bad teams appear more robust. In the extreme case, a team that moves randomly may appear to be very robust.

However, the robustness measure we propose is orthogonal to the performance measure, since a performance measure such as score-

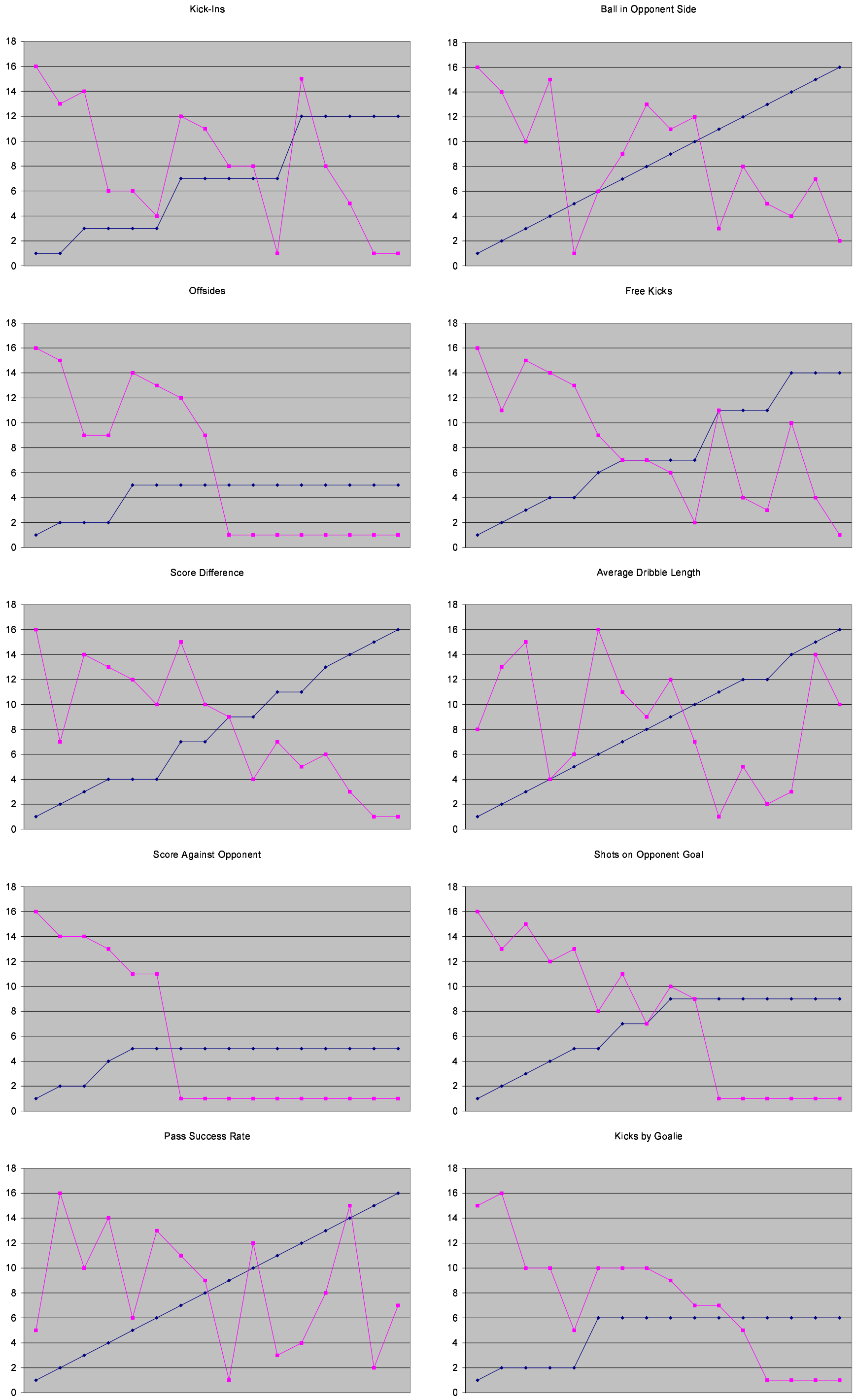


*Figure 5.* Plots of robustness vs. performance in the sensory uncertainty test, 1999

difference is essentially unbounded both above and below[2]: teams can win or lose by margins far larger than those present in the data. Thus a top-level team has no more to lose, in principle, than a poorer team—their respective ability to perform can degrade at the same rate. As an example, in the disabled-players test of 1998, three teams had a fairly high robustness of 0.9: AT Humboldt'98 [13], Gemini [31], and Darwin United [2]. Yet Gemini and AT Humboldt ranked respectively first and third (out of 13) in terms of score-difference performance, while Darwin United ranked last (13). AT Humboldt and Gemini are therefore examples of teams that are both robust and successful (in performance). What our findings indicate is that these teams are the exception, rather than the rule.

### 5.2. Over-specialization in Competition

A second important lesson emerges from the results when we look at the sensory uncertainty test. Unlike the disabled players test, this test has no gradual degradation in conditions. Instead, teams are tested on extreme degradation (twelve times the normal uncertainty level), but also under conditions of extreme *beneficial* change in their environment—no uncertainty in sensing. To our surprise, the evaluation data reveals that the performance of many teams suffers when they operate in a perfect-sensing environment. In other words, not only do specialized teams fail to degrade gracefully, but in fact they are over-specialized so that their performance degrades even under beneficial changes (at least with respect to managing this kind of sensory uncertainty).

Figure 6 shows the score difference results for each team in the sensory uncertainty test. For each of the 16 evaluated teams, the left bar shows the score difference under normal uncertainty conditions (note that score differences of zero produce bars of height zero), while the right bar shows the score difference for each team under conditions of perfect sensors—no uncertainty. In 10 out of the 16 cases, the right bar shows a lower value than the left bar, signifying that the team's performance has degraded. The team are presented in order of decreasing performance, from left to right. This allows us to see that over-specialization occurs both among the top-ranking teams, as well as poor-performance teams. These results were determined to be significant at the 94% significance level using a t-test ($p = 0.06$). The same trend can be observed at differing significance levels with the majority of the other performance measures used in the sensory uncertainty test.

[2] Ceiling and floor effects typically occur in the ±20–30 range, far from the score difference results in the data we analyzed.

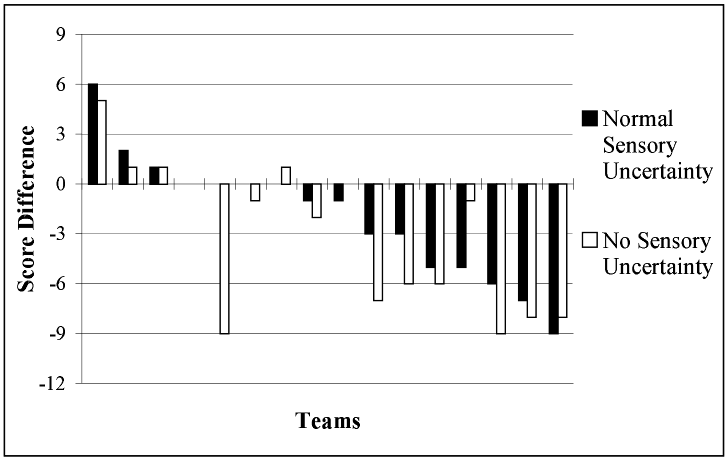


*Figure 6.* Score difference with and without sensory uncertainty

The conclusion that teams may be over-specializing for the expected competition environment is supported to some extent by another finding. In comparing the results of the disabled players tests of 1998 and 1999, we find that in general, teams improve in performance more than in robustness. For instance, we find that the average score difference has nearly doubled from 1998 to 1999. However, the average robustness value for the score-difference measure has improved by less than 20%. Using 1-tailed t-tests to assess the change between 1998 and 1999 in each of the performance and robustness measures of the disabled players test, we find that other measures also improved in performance, at the significance level ($p < 0.05$) or close to it ($p < 0.15$). In contrast, in all measures but one, we find no statistically significant increase in robustness from one year to the next (though slight improvements can be seen in many measures). Thus, despite explicit emphasis by RoboCup organizers on robustness as a key challenge [21], teams improve in performance more than in robustness.

This result can be a source of concern to organizers of research-oriented competitions, such as RoboCup, who seek to use the competitions in order to promote research into qualities such as robustness, and may want to avoid the over-specialization side-effect. Indeed, we are currently collaborating with RoboCup organizers to address this issue in future competitions (see also next section).

## 6. Related Work and Emerging Challenges

This section will focus on challenges to the AI and Multi-Agent research community raised by the evaluation sessions. We highlight these challenges in the context of related work.

### 6.1. Related Work on Research Infrastructure

We can group the existing infrastructure related to our evaluation sessions into two broad categories. The first and most obvious of these is other research-oriented competitions. The second is research test-beds, benchmarks, and repositories. We introduce the characteristics of each below, and then discuss how they relate to the question of research methodology in general, and to our evaluation sessions in particular.

*Competitions* Other research competitions related to our work include the AAAI Robot Competition [3], the AIPS Planning System Competition [28], and also the RoboCup contests themselves [21]. These contests differ in the way that they interpret the notion of "competition". For example, at a very simple level the basic organization of each contest is different. In RoboCup, for instance, the number of participants prohibits exhaustively playing all competitors against fixed opponents. But in both the AAAI and AIPS competitions, all competing systems tackle the same tasks. Defining a competition also presents decisions at a more fundamental level, as illustrated by this extract from McDermott's discussion of the 1998 AIPS Planning Systems competition [28, p. 45]:

> ... there was an intricate negotiation involving the committee and the community of potential contestants. The committee wanted to encourage the research community to try new things; the community wanted the committee to focus on the areas their planners did well in.

We believe that this is a general tendency for any research-based contest. As it becomes established, it is pressured towards the second category we distinguished above: research test-beds, benchmarks, and repositories.

*Test-beds, Benchmarks, and Repositories* Whereas contests are typically focussed on task-specific performance, repositories are more concerned with promoting comparative evaluations of techniques. Repositories exist in many different fields and include examples such as the Irvine Machine Learning repository [7], the plan-recognition repository [25], and the Planning Systems Competition problem set [27]. Certainly, at a fundamental level there can be many similarities between test-beds, benchmarks, repositories, and competitions. But more than anything else, what turns them into different kinds of infrastructure is how they are used; the methodology they are designed to support.

We have been much influenced by two seminal works on empirical research methodology. Cohen [9] provides not only a guide to conducting empirical research, but also discusses the motivation for conducting such research. Hanks, Pollack, and Cohen [14] discuss experiment-oriented research methodologies, and provide a survey of benchmark problems and research test-beds that are of use to planning systems researchers. Both of these works emphasize the importance of controlled experimentation in investigating complex systems, and in particular raise the problem of generating significant lessons from experiments. As both works point out, it can be difficult to learn general lessons from controlled experiments in a particular domain: Cohen devotes a closing chapter to this issue, while Hanks et al. devote much of their discussion to this problem.

*Discussion and Comparison to Our Work* We view our evaluation sessions as a bridge spanning each of the types of infrastructure (competitions, test-beds, benchmarks, and repositories). The design of this bridge is intended to allow the RoboCup competitions (or any other contest run in the same way) to make simultaneous use of multiple helpful methodologies.

For instance, competitions have a tendency to focus on overall task performance. Thus, issues such as robustness are evaluated only qualitatively, often with the implicit assumption that a better-performing program is probably more robust. The evaluation sessions bring to this scenario the aspect of controlled experimentation. Indeed, one of our main goals in organizing the evaluation sessions was to address the key methodology issue mentioned above: the generality of results. Certainly, empirical results established in the RoboCup domain do not necessarily carry over to other domains of interest. But the evaluation sessions introduce the opportunity to compare the independently-developed research results of dozens of research groups, all tested under the exact same physical conditions, down to machines, network, and location. Care is taken to ensure that the evaluation sessions maximize the comparability of different teams: the sessions are held annually in a centralized location, with all evaluated teams using the exact same conditions. Thus, the evaluation sessions provide us with a greater degree of confidence in establishing lessons in a particular environment, without being affected by the interaction between the environment and a particular agent embodying a technique.

Since we also make the log files of the evaluation sessions available, the evaluation sessions also have some of the nature of repositories. A repository is intended for long-term use, allowing researchers to carry out off-line, post-fact analysis and to produce temporally- and physically-distributed comparisons. This makes truly general comparisons more difficult, but allows for greater flexibility. One of the reasons for making the data from the evaluation sessions publicly available was to recover some of this flexibility. However, the evaluation sessions are different from repositories in two significant ways. First, repositories make results public for comparative purposes, and thus rely on standardized measurement for comparison. However, in the evaluation sessions, measurement is only loosely defined, using different rankings for each measure. The work we present in this paper takes a specific approach to measurement of performance and robustness, but other approaches are also possible, and we certainly hope that one of the results of our work will be that others consider this issue of measurement more closely (see the following subsection). Second, we were careful to keep the evaluation sessions dynamic, by introducing a new secret test annually. This prevents over-specialization by researchers towards particular benchmarks or areas of interest, a concern with benchmarks that are known in advance.

This final point on the need to conceal some of the testing details brings us back to the main message of this paper: that the evaluation sessions allowed us to uncover important results (e.g., the relationship between performance and robustness) that would not have come to light within the framework of the contest itself. Since teams over-specialized despite the RoboCup organizers emphasizing the need for robustness, the contest itself was in some sense unsuccessful in one of its goals (the promotion of robustness). This echoes the tension between organizers of research-oriented competitions and the research community described by McDermott above. We have dealt with these dual interests in RoboCup so far by conducting the evaluation session with no competitive aspects (that is, evaluated teams were not ranked based on their evaluation results). Also, we worked with RoboCup organizers to make participation in the evaluation sessions a requirement of all medium-level through top-level teams. However, the RoboCup organizers are now working to integrate aspects of the evaluation sessions into the competition itself. Thus, teams will have to demonstrate a breadth of capabilities, in addition to good performance in the main competition task. In order to prevent teams again over-specializing on the specific evaluation tests given in any year, it will be probably be important to continue our current practice of creating new tests each year — and keeping the nature of some of these tests secret.

### 6.2. Measurement and Analysis

Measurement is a central challenge for multi-agent researchers. The multi-agent community thus far lacks domain- and task- independent tools for quantifiable analysis of key concepts such as coordination, teamwork, and robustness. Furthermore, even task-specific tools such as the Statistics Proxy face difficulties in (i) analyzing the behavior of individuals in isolation from the effects of the team in which they are members, and in (ii) analyzing team behavior in isolation from the effects of other teams on it. In particular, two specific problems are highlighted by our experiences. First, key properties of interest to the research community (such as coordination, teamwork, cooperation, resource allocation, and load balancing) often defy quantitative measurement. Second, existing quantitative measures do not facilitate comparison of teams across tasks and domains. We address each of these issues below.

*Measuring Behavioral Features.* While task performance is certainly a key factor in analyzing the behavior of a multi-agent system, it is by no means the only one. Other factors such as robustness, teamwork, and adversarial modeling capabilities are all issues that may be of extreme importance. Yet, when we come to analyze a team's behavior, we often find that it is difficult to measure the relative strengths and weaknesses in these areas directly. We often rely on lesion studies using overall performance to guide our decisions. For instance, many RoboCup teams evaluate techniques by demonstrating that their own team, when using a particular technique, out-performs their own team without the technique in question. While certainly such lesion studies should be encouraged, they fundamentally avoid direct measurement of key areas of interest, because they evaluate techniques only in the context of the team's infrastructure, and in the context of overall task performance. A hypothetical team that coordinates well, but cannot kick accurately, may face difficulties in showing the strength of its coordination mechanisms through such tests.

The evaluation session data is publicly available to be analyzed [19], and it is our hope that it will encourage researchers to come up with new and innovative behavioral measures, which more directly evaluate a team in a particular area of interest. The Statistics Proxy takes a first step towards this goal, by measuring team behavior along behavioral features that correspond to qualitatively different styles of soccer teamwork. For instance, it analyzes dribbling and passing behaviors, thus contrasting teams of loosely-banded individual players (dribble more), with teams that emphasize coordination and cooperative achievement of goals (pass more).

*Cross-Task, Cross-Environment Comparisons.* A measure used in one task is not necessarily going to be useful for comparison with a team working on a different task. For instance, even knowing the robustness of a soccer team, and also the robustness of a different team in a different domain (under a similar test of, e.g., disabled agents), it is difficult to compare the robustness of the two. Part of the difficulty here is the lack of direct measures of key behavioral features (as discussed above). However, a second key difficulty is that the values of the measures (e.g., the robustness slopes) depend also on the task. For instance, we cannot compare the robustness slopes of teams when disabling three agents if a soccer team (which has 11 agents) is compared to a volleyball team (which has 6). Moreover, we cannot compare the robustness slope along score-difference in the disabled-players test, with the slope along average dribble length in the sensory uncertainty test. These two slopes use different units (e.g., score-difference change per disabled player, vs. change in average dribble length per uncertainty multiplier).

Recently, some researchers have begun to address these two measurement issues within the domain of soccer and outside of it. For example, Balch [6] investigates the use of Social Entropy [5] to measure behavioral *diversity*, establishing that positive correlation exists between diversity and performance in soccer. Goldberg and Mataric [11] suggest a different measure, based on inter-agent *interference* (which is measured as the time agents spend avoiding each other). Kaminka and Tambe [20] propose a measure of teamwork quality that is based on the average duration of disagreements in a team (ATA), and show that teams that have better perception of their environment are less dependent on communications. All of these measures show promise for directly measuring features of interest to multi-agent researchers (i.e., diversity, coordination, teamwork), and all also show at least some potential for cross-task, cross-environment comparability. However, they are only in the preliminary stages of investigation. None have been applied to the evaluation session data, as all three currently suffer from a limitation of relying on accurate execution traces of the internal decisions made by the agents. Such execution traces are not available (and cannot be expected to be available) with arbitrary teams, as is the case in the evaluation sessions.

## 7. Summary and Future Work

We have presented the RoboCup evaluation sessions that we have carried out over the last three years, and analyzed the results of these sessions in terms of robustness—a key general feature often sought and investigated by multi-agent researchers. Most notably, we showed that for the RoboCup teams in the evaluation sessions (i) performance is negatively correlated with robustness, (ii) over-specialization is a serious concern, and (iii) there is improvement in performance, but little corresponding improvement in robustness.

We draw lessons from our results both in terms of the design of performance competitions ("What should the performance task be?") and in terms of the general challenges of the evaluation of multi-agent systems ("How should performance be measured?"). In terms of the design, our ability to draw these conclusions demonstrates the potential of evaluation sessions in facilitating comparative studies. Without the sessions, the question of the existence of a general robustness-performance tradeoff would have been difficult to explore empirically. In terms of measurement, we believe the in-depth analysis techniques we have presented are a first step in addressing evaluation challenges in the context of robustness. However, as a secondary contribution, we also make all our evaluation data publicly available [19] to encourage researchers to experiment with new measures and analysis methods. This data presents a key challenge to the community, to develop the necessary tools and techniques to quantitatively evaluate multi-agent behavior along key areas of interest to the community, such as coordination, teamwork, and robustness.

We hope that our results will also encourage organizers of research-oriented competitions to increase and emphasize the research infrastructure that competitions provide for participants. Such infrastructure includes recording (and making public) the execution traces of the systems being evaluated for later analysis, for instance by video, or by providing standard logging facilities. It also includes dedicating sufficient time for conducting controlled experiments that evaluate systems beyond their ability to carry out the competition tasks, and ideally to highlight the techniques that allow them to do so.

Indeed, we are planning to continue working on the challenges raised by the evaluation sessions ourselves. We are currently collaborating with RoboCup organizers (who have been supportive of this work in the last three years) to integrate the evaluation sessions more firmly into the RoboCup tournament itself. Further, we are investigating new measurement techniques that we hope will facilitate and encourage yet stronger empirical studies of multi-agent systems in the future.

## Acknowledgements

We thank all participating teams in the evaluation sessions and the RoboCup simulation league organizing committees (1998 Chair: Itsuki Noda, 1999 Chair: Peter Stone, 2000 Chair: Paul Scerri). We thank Kay Schroter, Hans-Dieter Burkhard, Daniel Polani, Hiroaki Kitano, Mikhail Prokopenko, Marc Butler, Tom Howard, and Patrick F. Riley for their help and cooperation in making the evaluation sessions a successful annual event. In addition, we thank Milind Tambe, Stacy Marsella, Orna Raz, Nachum Kaminka, Shlomit Kaminka, and the anonymous reviewers for very useful comments and many useful discussions. As always, K. Ushi deserves special thanks.

## References

1. Ando, T.: 1999, 'Andhill-98: A RoboCup team which reinforces positioning observability'. In: M. Asada and H. Kitano (eds.): *RoboCup-98: Robot soccer world cup II.* Springer-verlag, pp. 373–388.
2. Andre, D. and A. Teller: 1999, 'Evolving team Darwin United'. In: M. Asada and H. Kitano (eds.): *RoboCup-98: Robot soccer world cup II.* Springer-verlag, pp. 346–352.
3. Arkin, R. C.: 1998, 'The 1997 AAAI Robot Competition and Exhibition'. *AI Magazine* **19**(3), 13–17.
4. Atkins, E. M., E. H. Durfee, and K. G. Shin: 1997, 'Detecting and Reacting to Unplanned-for World States'. In: *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97).* Providence, RI, pp. 571–576.
5. Bailey, K. D.: 1990, *Social Entropy Theory.* State University of New York Press.
6. Balch, T.: 1998, 'Behavioral Diversity in Learning Robot Teams'. Ph.D. thesis, Georgia Institute of Technology.
7. Blake, C. and C. Merz: 1998, 'UCI Repository of machine learning databases'. http://www.ics.uci.edu/~mlearn/MLRepository.html.
8. Burkhard, H.-D., M. Hannebauer, and J. Wendler: 1998, 'AT Humboldt—development, practice, and theory'. In: H. Kitano (ed.): *RoboCup-97: Robot soccer world cup I*, Vol. LNAI 1395. Springer-verlag, pp. 357–372.
9. Cohen, P.: 1995, *Empirical methods for Artificial Intelligence.* Cambridge, MA: MIT Press.
10. Cohen, P. R. and H. J. Levesque: 1991, 'Teamwork'. *Nous* **35**.
11. Goldberg, D. and M. J. Mataric: 1997, 'Interference as a tool for designing and evaluating multi-robot controllers'. In: *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97).* Providence, RI, pp. 637–642.
12. Grosz, B. J. and S. Kraus: 1996, 'Collaborative Plans for Complex Group Actions'. *Artificial Intelligence* **86**, 269–358.
13. Gugenberger, P., J. Wendler, K. Schroter, and H.-D. Burkhard: 1999, 'AT Humboldt in RoboCup-98'. In: M. Asada and H. Kitano (eds.): *RoboCup-98: Robot soccer world cup II.* Springer-verlag, pp. 358–363.

14. Hanks, S., M. E. Pollack, and P. Cohen: 1993, 'Benchmarks, Test Beds, Controlled Experimentation, and the Design of Agent Architectures'. *AI Magazine* **14**(4), 17–42.
15. Horling, B., B. Benyo, and V. Lesser: 2001, 'Using Self-Diagnosis to Adapt Organizational Structures'. In: *Proceedings of the Fifth International Conference on Autonomous Agents (Agents-01).* pp. 529–536.
16. Horling, B., V. R. Lesser, R. Vincent, A. Bazzan, and P. Xuan: 1999, 'Diagnosis as an Integral Part of Multi-Agent Adaptability'. Technical Report CMPSCI Technical Report 1999-03, University of Massachusetts/Amherst.
17. Jennings, N. R.: 1993, 'Commitments and conventions: the foundations of coordination in multi-agent systems'. *Knowledge Engineering Review* **8**(3), 223–250.
18. Jennings, N. R.: 1995, 'Controlling cooperative problem solving in industrial multi-agent systems using joint intentions'. *Artificial Intelligence* **75**(2), 195–240.
19. Kaminka, G. A.: 1998, 'The Multi-Agent Systems Evaluation Repository'. http://www.cs.cmu.edu/∼galk/Eval/.
20. Kaminka, G. A. and M. Tambe: 2000, 'Robust Multi-Agent Teams via Socially-Attentive Monitoring'. *Journal of Artificial Intelligence Research* **12**, 105–147.
21. Kitano, H., M. Tambe, P. Stone, M. Veloso, S. Coradeschi, E. Osawa, H. Matsubara, I. Noda, and M. Asada: 1997, 'The RoboCup Synthetic Agent Challenge '97'. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-97).* Nagoya, Japan.
22. Kumar, S. and P. R. Cohen: 2000, 'Towards a Fault-Tolerant Multi-Agent System Architecture'. In: *Proceedings of the Fourth International Conference on Autonomous Agents (Agents-00).* Barcelona, Spain, pp. 459–466.
23. Kumar, S., P. R. Cohen, and H. J. Levesque: 2000, 'The Adaptive Agent Architecture: Achieving Fault-Tolerance Using Persistent Broker Teams'. In: *Proceedings of the Fourth International Conference on Multiagent Systems (ICMAS-00).* Boston, MA, pp. 159–166.
24. Lenser, S., J. Bruce, and M. Veloso: 2001, 'CMPack: A Complete Software System for Autonomous Legged Soccer Robots'. In: *Proceedings of the Fifth International Conference on Autonomous Agents (Agents-01).* pp. 204–211.
25. Lesh, N.: 1995, 'The Plan-Recognition Data Repository'. http://www.merl.com/people/lesh/prdata.html.
26. Matsubara, H., I. Frank, k. Tanaka-Ishii, I. Noda, H. Nakashima, and K. Hasida: 1998, 'Automatic Soccer Commentary and RoboCup'. In: M. Asada (ed.): *the Second RoboCup Workshop (RoboCup-98).* Paris, France, pp. 7–22.
27. McDermott, D.: 1998, 'The AI Planning System Competition Problem-Set Repository'. ftp://ftp.cs.yale.edu/pub/mcdermott/domains/.
28. McDermott, D.: 2000, 'The 1998 AI Planning Systems Competition'. *AI Magazine* **21**(2), 35–55.
29. Noda, I. and I. Frank: 1998, 'Investigating the Complex with Virtual Soccer'. In: J.-C. Heudin (ed.): *Virtual Worlds.* Springer Verlag, pp. 241–253.
30. Noda, I., H. Matsubara, K. Hiraki, and I. Frank: 1998, 'Soccer Server: A Tool for Research on Multiagent Systems'. *Applied Artificial Intelligence* **12**(2–3), 233–250.
31. Ohta, M. and T. Ando: 1998, 'Cooperative reward in reinforcement learning'. In: *Proceedings of the 3rd JSAI RoboMech Symposia.* pp. 7–11.

32. Pechoucek, M., V. Marik, and O. Stepankova: 2001, 'Towards Reducing Communication Traffic In Multi-Agent Systems'. *Journal of Applied System Studies* (Special Issue on Virtual Organizations and E-Commerce Applications).
33. Rickel, J. and W. L. Johnson: 1999, 'Animated Agents for Procedural Training in Virtual Reality: Perception, Cognition, and Motor Control'. *Applied Artificial Intelligence* **13**, 343–382.
34. Stone, P.: 1998, 'Layered Learning and Flexible Teamwork in Multi-Agent Systems'. Ph.d., Carnegie-Mellon University.
35. Stone, P., P. F. Riley, and M. Veloso: 2000, 'The CMUnited-99 Champion Simulator Team'. In: *RoboCup-98: Robot soccer world cup III.* Springer-verlag.
36. Stone, P., M. Veloso, and P. F. Riley: 1999, 'The CMUnited-98 Champion Simulator Team'. In: *RoboCup-98: Robot soccer world cup II.* Springer-verlag, pp. 61–76.
37. Sugawara, T. and V. R. Lesser: 1998, 'Learning to Improve Coordinated Actions in Cooperative Distributed Problem-Solving Environments'. *Machine Learning* **33**(2/3), 129–153.
38. Tambe, M.: 1997, 'Towards Flexible Teamwork'. *Journal of Artificial Intelligence Research* **7**, 83–124.
39. Tambe, M., W. L. Johnson, R. Jones, F. Koss, J. E. Laird, P. S. Rosenbloom, and K. Schwamb: 1995, 'Intelligent agents for interactive simulation environments'. *AI Magazine* **16**(1).
40. Tambe, M., G. A. Kaminka, S. C. Marsella, I. Muslea, and T. Raines: 1999, 'Two fielded teams and two experts: A RoboCup Challenge Response from the Trenches'. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-99)*, Vol. 1. pp. 276–281.
41. Tambe, M., D. V. Pynadath, N. Chauvat, A. Das, and G. A. Kaminka: 2000, 'Adaptive Agent Integration Architectures for Heterogeneous Team Members'. In: *Proceedings of the Fourth International Conference on Multiagent Systems (ICMAS-00).* Boston, MA, pp. 301–308.
42. Tanaka-Ishii, K., I. Noda, and I. F. et.al.: 1998, 'MIKE: An Automatic Commentary System for Soccer — System Design and Control —'. In: *Proceedings of International Conference on Multi-Agent Systems '98.* Paris, France, pp. 285–292.
43. Tanaka-Ishii, K., I. Noda, I. Frank, and H. Matsubara: 1999, 'A Statistical Perspective on the RoboCup Simulator League: Progress and Prospects'. In: *The 3rd Proceedings of RoboCup Workshop.*
44. Toyama, K. and G. D. Hager: 1997, 'If at First You Don't Succeed...'. In: *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97).* Providence, RI, pp. 3–9.
45. Wagenaar, W.: 1988, *Paradoxes of gambling Behaviour.* Lawrence Erlbaum Associates Ltd. ISBN 0-86377-080-0.