# Measuring Progress in Multirobot Research With Rating Methods—The RoboCup Example

Armin Shmilovici, Foaid Ramkddam, Beatriz Lopez, and Josep Lluis de la Rosa

*Abstract*—**Rating the intelligence of artificially made systems is important for measuring progress in scientific and engineering methods. Unfortunately, there is currently no universal agreement about what is considered an intelligent system, and how to measure its intelligence.**

**This research focus on measuring the progress in the robotic technologies deployed for the RoboCup competitions, since one of the original premises of those competitions was to advance the development of intelligent robotic systems.**

**A method used for rating the competence of human chess players is adapted for measuring the advancement in the competence of robotic teams. The results indicate significant yearly improvements in the capabilities of the robotic teams. The same method can be used to indirectly quantify the benefits in specific technology choices.**

*Index Terms*—**Chess rating, multirobot research, RoboCup.**

## I. Introduction

A VAST body of research has been published on the development of so called "intelligent systems," yet there are fundamental disagreements about what are the basic attributes of an intelligent system. Some definitions (e.g., [1], [21]) focus on the building blocks of such systems (e.g., sensors and actuators) and specifically on its *cognitive processes* (e.g., learning, knowledge representation, and behavior generation). Other definitions focus on the *performance* of the system at different *tasks* [2]: "intelligence will be defined as an ability of a system to act appropriately in uncertain environments, where appropriate action is that which increases the success, and success is the achievement of behavior sub-goals that support the systems ultimate goal." This correspondence adopts the second definition.

The ability to measure the capabilities of an intelligent system is important for measuring scientific progress and to reward success. It is difficult to design a metric for intelligence of systems when there is disagreement about the definition of intelligence, or the separation between the capabilities of the "body" of the system (e.g., sensors accuracies) and the "mind" of the system (e.g., intelligent control), or what part of the "performance" of the system (whatever that means) is attributed to its intelligence [3]. Saridis [4] proposed the "principle of increasing precision with decreasing intelligence for intelligent machines," a hierarchical structure that implements it, and a performance measure that is a solution to the minimization of the system's operational costs subject to its reliability constraints [5]. Unfortunately, computing the performance measure requires developing quality mathematical models for systems with complex behaviors, which are often not well understood.

One approach for measuring the performance of intelligent systems is by comparison to human performance for similar tasks (e.g., the famous Turing test [6], [22]). This approach suffers from the limitation of humans in executing some tasks that machines execute very well (e.g., arithmetics). A different approach is to measure indirectly the machine intelligence on carefully constructed benchmark problems. Many benchmarks and challenge problems were defined over the past years. One famous challenge problem came to the end of its life in 1997, when the computer "Deep Blue" won a chess tournament against Gary Kasparov—the chess world champion. There are also some results in empirical research methodology and experimentation [7], [8].

Multirobot systems (MRS) are nowadays, an important area within robotics and artificial intelligence (AI) and a growing number of systems have been recently presented in the literature [9], [10]. Special attention has been given to MRS developed to operate in dynamic environment, where uncertainty and unforeseen changes can happen due to the presence of robots and other agents that are external to the MRS itself. Special metrics such as the "social entropy" of [11] for behavioral diversity, the inter-agent interference of [12], or the "average duration of disagreement" of [13] have been defined in the literature. A significant body of work has originated from motivations that are essentially of engineering nature, where MRS are designed and realized in order to improve the effectiveness, performance, and robustness of a robotic system.

A significant boost to the work on MRS has recently been given by the robotics competitions and the RoboCup challenge problem [14] in particular. RoboCup is an international joint project to promote AI, robotics, and related fields. The declared long-term goal of the RoboCup competition is to "develop a team of fully autonomous humanoid robots, that by 2050, can win a game of soccer, complying with the official rules of the FIFA, against the world champion team in soccer." The short term goal of the RoboCup competition is to advance intelligent robotic technologies such as control, planning, communication, advanced sensors, and information processing by providing a standard problem where wide range of technologies can be integrated and examined. There are different leagues, such as the simulation league, the small size robotic league, the legged robotic league, and the middle size robotic league. Since the small size robotic league (F180) is the only league were the annual changes in the competition regulations were relatively insignificant, it was used here to demonstrate the progress measurement procedure. Following, is a short description of the rules:

Two teams of up to five small sized robots (one of which could be a goalkeeper) play on a green carpeted table-tennis sized field marked with sloping edges. A camera is perched on top of the field to act as the robots' global vision system that tracks the players, opponents and the ball positions. During a game these robots can use wireless communication to talk to each other via a computer off the field, but no human intervention is allowed. The robots have to play by rules similar to FIFA's (e.g., robots can be fouled off the field). There are two 10-min halves in each game with a 20-min break for servicing the robots (e.g., battery replacements). A human referee monitors the match and is the only one allowed to move the robots if necessary. After each goal, the game is reset to an initial "kickoff position."

An implicit assumption of the RoboCup challenge problem, is that the more advanced technologies a team uses, the higher its chance to win in games and eventually win the first prize. Indeed, in recent competitions, leading teams used advanced technologies such as multiagents and role-switching.

The purpose of this correspondence is to suggest a new measure for indirectly gauging the annual progress in the capabilities of the RoboCup teams. The key idea is to use methods developed for rating the capabilities of individuals and teams in competitive sports. That

Manuscript received March 5, 2003; revised August 7, 2003. This paper was recommended by Associate Editor H. Qiao.

A. Shmilovici is with the Department of Information Systems Engineering, Ben-Gurion University, Beer-Sheva 84105, Israel (e-mail: armin@bgumail.bgu.ac.il).

F. Ramkddam is at Reinaerstr 66, Gent 9000, Belgium.

B. Lopez and J. L. de la Rosa are with the Department of Electronics, Information and Control, University of Girona, 17071 Girona, Spain (e-mail: blopez@eia.udg.es; peplluis@eia.udg.es).

Digital Object Identifier 10.1109/TSMCB.2003.818536

is each team would receive a rating point at the end of each tournament—under the assumption that unlike with human capabilities, artificial capabilities cannot decline. The yearly progress in the rating points of each team and the league as a whole will indicate a measure for the technological advances. The next section will introduce the rating method used.

## II. Measuring Progress With Chess Rating Methods

### A. Introduction to Chess Rating

Rating methods were introduced as a convenient method to asses the capabilities of teams (or individuals) based on their previous performance, and before an actual new test of their capabilities. In competitive sports, the implicit assumption is that when a high rated team plays against a low rated team, than the higher rated team has a higher probability of winning the competition, and that the difference in the probability of winning is somehow related to the rating difference between the teams. This makes rating useful for arranging tournaments between similarly rated teams, and for betting purposes. A big part of the success of professional sports is attributed to the introduction of rating systems and compensation methods that give incentive to players to compete and improve their ratings.

The most famous rating method is the Elo method [15] that was designed to rate the strength of chess players. Variations of this method are also used for rating other sports such as Tennis and Table Tennis players. The Elo method is based on several principles/

1) The probability of team $i$ winning a competition against team $j$ depends only on the rating difference $R_i - R_j$ between the teams $i$, $j$, and is captured by the following:

$$\Pr(i \quad \text{deafeats} \quad j) = \frac{1}{1 + 10^{-\frac{(R_i - R_j)}{400}}}. \tag{1}$$

2) A large rating difference in (1) represents a *dominance* of one team over the other team. In order to determine the absolute rating of a player, a *calibration*, or baseline rating is needed. Several calibration methods were proposed over the years. In chess [16], the calibration point is determined such that the rating of chess players is in the range [0,3000].
3) Each player needs to have some kind of a rating, so that it could participate in tournaments with similar strength players. New players that do not have a sufficient playing history against rated players will get a *provisional* rating.
4) The rating of each team (provisional or not) will be updated based on the results in official tournaments, such that good performance (wins) will increase the rating while bad performance will decrease the rating. The update is

$$R_{\text{post}} = R_{\text{pre}} + K(S - S_{\text{exp}}) \tag{2}$$

where $R_{\text{pre}}$, $R_{\text{post}}$ are the pre-tournament and post-tournament ratings, respectively. $S$ is the player's total score in the tournament, $S_{\text{exp}}$ is the expected total score estimated from the player's pre-tournament rating and the opponents pre-tournament ratings and can be calculated by summing the estimated winning expectancies for each game using (1). $K$ is an attenuation factor that determines the weight that should be given to a player's performance relative to the pre-tournament rating (e.g., 32 for the United States Chess Federation for amateur players). For more details and for the provisional rating update formula, see [16].

### B. Implementation for the Robocup Small Size League

It was decided to use the Elo chess rating system for the RoboCup small size robotic league for the following reasons.

1) It uses easy to obtain information—the tournament scores, ignoring the technical heterogeneity of the different robotic teams.
2) It is mathematically and experimentally well understood.
3) It is used for rating individuals (e.g., chess players), teams (e.g., couple's Tennis), and machines (e.g., chess programs playing against humans).
4) It will be also useful for the long-term goal of rating robotic teams playing soccer against human teams.

The first RoboCup competition was held in Nagoya, Japan in 1997. There is one yearly international tournament in the summer, and there are also spring regional tournaments with a smaller number of participants. For this research, the official tournament results as published in [14] were used.

*Assumptions:* Since after each goal, the game is reset to an initial "kickoff position," each game is considered as a tournament of consecutive independent games. The results of each game are used to estimate the team's *probability of winning*: If the number of goals in the game was $n_i : n_j$ between players $i$, $j$ respectively, then the Laplace ratio is used to estimate the probability that $i$ defeats $j$

$$\hat{\Pr}(i \quad \text{defeats} \quad j) = \frac{n_i + 1}{n_i + n_j + 2}. \tag{3}$$

Note that this probability estimate is better suited for small numbers and can deal with zero goals.

From the probability of winning, and inverting (1), *the rating difference* between the teams is estimated as

$$R_i - R_j \approx 400 \log_{10} \left( \frac{n_i + 1}{n_j + 1} \right). \tag{4}$$

The base (*calibration*) rating was arbitrarily selected to be 1000 for the winner of the 1998 competition—since it was the first competition with a reasonable number (10) of competitors.

For determining a *provisional* rating for each team, the rating process is iterated backward (from best team to worst team) using (1) and (3). Using the provisional ratings, the rating process is iterated forward following the games in the competition. Equation (2) is used with $K = 32$, to update the rating of each team until each team gets a *final rating* for that competition. If the final rating is too different[1] from the provisional rating, it is used as provisional rating and the process is repeated.

It is assumed that *ratings can not decrease between consecutive competitions*.[2] Searching all the teams that participated in the sequel competition, find the weakest team which can transfer its ratings as the new base point for the sequel competition, such that none of the other teams which participated in both competitions will have its rating reduced.

*Examples:* The results of the 1998 finals CMU vs. Roboroos were 3:1. From formula (4) the rating difference is estimated to be 120. Since the rating of CMU is defined as 1000, than Roboroos backward rating (its provisional rating) is estimated as 880.

The results of the 1998 semi-finals CMU vs. Cambridge were 3:0, which results in a provisional rating estimate of 759 for Cambridge. The results of the game Cambridge vs. Roboroos were 5:4. From (1)

[1] More than 20 rating points. In practice, 2–3 iterations of the algorithm were sufficient to obtain this accuracy.

[2] Teams must disclose their technology in the proceedings that are published after each competition. Thus we assume the teams are modified only in ways that can improve the performance.

TABLE I
RATINGS FOR TEAMS AND COMPETITIONS

| TEAMS | Nagoya 1997 | Paris 1998 | Stockholm 1999 | Amsterdam 2000 (European Champ.) | Melbourne 2000 | German Open 2001 | Japan Open 2001 | Seattle 2001 | German Open 2002 | Fukuoka 2002 |
|---|---|---|---|---|---|---|---|---|---|---|
| CMU/CMDragons | **495** | 1000 | 1251 | - | - | - | - | 1180 | - | 2122 |
| Roboroos | - | 842 | 1289 | - | 1539 | - | - | 1686 | - | 1739 |
| Cambridge | - | 809 | - | - | - | - | - | - | - | - |
| Paris-8 | - | 707 | - | - | - | - | - | - | - | - |
| UVB | - | 728 | - | 1107 | - | - | - | - | - | - |
| 5 DPO | - | **907** | **907** | 998 | - | 1643 | - | 1664 | 1707 | - |
| J-Star | - | 525 | 982 | - | - | - | - | - | - | - |
| Paris-6 | 185 | 689 | - | - | - | - | - | - | - | - |
| iXs | - | 552 | - | - | - | - | - | - | - | - |
| I-Space | - | 713 | - | - | - | - | - | - | - | - |
| RogiTeam | 196 | - | 949 | 1069 | 1350 | - | - | 1545 | - | 1545 |
| NAIST | 234 | - | - | - | - | - | - | - | - | - |
| AllBotz | - | - | 866 | - | 1174 | - | - | - | - | - |
| BigRed |  | - | 1811 | - | 1811 | - | - | 1884 | - | 2530 |
| RobotIS | - | - | 1558 | - | - | - | - | - | - | - |
| Linked | - | - | **1095** | **-** | **-** | **-** | **1095** | - | - | - |
| TPOTS | - | - | 801 | - | 1015 | - | - | 1259 | - | - |
| TUD | - | - | - | - | 792 | - | - | - | - | - |
| SingPoly | - | - | 1161 | - | - | - | - | - | - | - |
| Crimson | - | - | 1058 | - | 1356 | - | - | - | - | - |
| Owaribitos | - | - | 889 | - | - | - | 1472 | 1580 | - | 1678 |
| FU-Fighter | - | - | **1371** | **1371** | **1609** | **1609** | - | **1851** | **1851** | 2401 |
| LuckyStar | - | - | 1561 | - | 1810 | - | - | 2023 | - | 2417 |
| CFA UPMC | - | - | - | 925 | 1476 | - | - | - | - | - |
| MUCows/Roobots | - | - | - | - | 1103 | - | - | 1426 | - | 2076 |
| 4 Stooges | - | - | - | - | 684 | - | - | 951 | - | - |
| CIIPS Glory | - | - | - | - | 881 | - | - | - | - | - |
| ViperRoos | - | - | - | - | 1029 | - | - | 1328 | - | - |
| Field-Rangers | - | - | - | - | 1517 | - | - | 1793 | - | 2116 |
| Yale Frobocup | - | - | - | - | 953 | - | - | - | - | - |
| Sharif CESR | - | - | - | - | - | - | - | 1293 | - | 1639 |
| Team Canuck | - | - | - | - | - | - | - | 1099 | - | 1590 |
| Robosix | - | - | - | - | - | **1296** |  | **1296** | 1416 | 1657 |
| HWH Cats | - | - | - | - | - | - |  | 1256 | - | - |
| OMNI | - | - | - | - | - | - | 1332 | 1375 | - | 1409 |
| KU-Boxes | - | - | - | - | - | - | 1580 | 1461 | - | 1096 |
| IUT Flash | - | - | - | - | - | - | - | - | 1498 | 1744 |
| IUB Team | - | - | - | - | - | - | - | - | 1164 |  |
| MAXIMUM | 495 | 1000 | 1811 | 1371 | 1811 | 1643 | 1580 | 2023 | 1851 | 2530 |
| MEDIAN | 215 | 720.5 | 1095 | 1069 | 1262 | 1609 | 1402 | 1426 | 1498 | 1739 |
| MINIMUM | 185 | 525 | 801 | 925 | 684 | 1296 | 1095 | 951 | 1164 | 1096 |

we can estimate the expected score for Cambridge out of 9 games as $S_{\mathrm{exp}} = 9^{*}0.3326 = 2.9932$. Applying formula (2) with $S = 5$ results in a revised rating for Cambridge: $759 + 32^{*}(5 - 2.9932) = 823$. For Roboroos $S_{\mathrm{exp}} = 9^{*}0.6674 = 6.0066, S = 5$, and the revised ratings for Roboroos is: $880 + 32^{*}(4 - 6.0066) = 816$.

Four teams that competed in the Paris 1998 finals also competed in the Stockholm 1999 finals. It was found that fixing 5DPO's rating of 907 across competitions did not reduce the rating of any of the other three teams, so this was used as the calibration anchor.

Table I presents the final rating estimates, for each team in each competition. The bold numbers in black boxes indicate rating transfers across competitions. The last three lines in the Table I present the tournaments' minimum,[3] maximum, and median ratings, respectively. Note the distinct yearly progress in all three categories. (The regional competitions included only a small number of participants, thus are less accurate as progress indicators).

Since assumption 4 may introduce *inflation* in the rating, it is preferable to consider the rating differences rather than absolute ratings. An approximate *deinflation* can be obtained by subtracting the yearly minimum rating result from each tournament's ratings. Even with that very conservative computation, we can still notice a significant yearly increase in the median and maximum ratings.

*Technical Comments:*

a) As noted by [16], the Elo method is a kind of Bayesian statistics approximation to the probability of (1), where an *a priori* rating together with the new score information are used to compute the *a postiori* rating with (2) and update the probability (1). A team's playing strength (thus, its rating) is assumed to be *fixed* within a tournament. In practice, the robotic teams are built with experimental rather than industrial-grade technology and partial and full team failures are not uncommon during games. Here we implicitly ignore the effects of hardware failures on the ratings, since the failures are not well documented in the game records, and since reliability can be considered as an important ingredient in a team's strength. Alternatively, it is possible to use the rating procedure proposed in [17]—based on a Bayesian statistics approximation[4]—that considers a Gaussian-like rating distribution function with mean and variance. Reduced team reliability will increase the rating variance of that team.

b) The parameter 400 in (1) is an arbitrary scaling parameter that was taken from the chess system. The parameter $K$ in (2) is an adaptation speed parameter also taken from the chess system. Its size affects the number of iterations before convergence to the final ratings.[5]

[3] There is an effectively bi-annual increase in the minimum qualification requirements for a participating team.

[4] The current rating is effectively an estimate of the mean of the Bayesian rating distribution.

[5] It was felt that the current parameters provide sufficient accuracy for the purpose of this research.

## III. Discussion

To paraphrase William Lord Kelvin, when you can measure something and put some numbers to it, then you know something about it. Though due to the small dataset used, the numbers in Table I are only approximate, by defining the notions of "strong dominance," and "generations" we can still draw some interesting conclusions about the progress of multirobot research. Strong dominance will be defined as above 200 rating difference that practically guarantees a victory. A new generation will be defined as above 500 rating difference that practically guarantees that the weak team will not score even one goal against the strong team.

From the results (of the international events, the above median teams), we can conclude that there were about four distinct technological generations, approximately in the years 1997, 1998, 2000, and 2002.

One of the conditions of entry in the competition is that the teams must disclose their technology in the yearly proceedings[6] that is published after each competition (e.g., [18], [19]). This not only helps to disseminate the knowledge among all the participants (as indicated by the yearly progression in the ratings of the weak teams), but can also be used to measure the contribution of specific robotic technologies: The first two generations of robotic teams can be characterized as pure reactive systems: locate ball, go to ball, kick in direction of goal. The third generation introduced path-planning and ball passing between team members, while the fourth generation introduced game strategy such as player role-switching.

In the old debate about the value of improving the "intelligence" of a system (e.g., better path planning algorithms) compared to improving its "physical capabilities" (e.g., better sensors and actuators) there is strong evidence for the importance of the latter. For example, the strong dominance of Cornell University's "BigRed" team in the 1999 competition can be mostly attributed to an improved mechanical ball manipulation device. The strong dominance of the "All Botz" team over the "4 Stooges" team in the 2000 competition—both teams belong to the same research group in the University of Auckland—can be mostly attributed to different robotic locomotion mechanisms. In a similar way, the impact of other technologies such as completely distributed sensing and decision making (as practiced by the "CIIPS Glory" team), or only distributed decision taking (as practiced by the "RogiTeam") can be measured. Exactly allocating the yearly rating increase between hardware improvements (e.g., faster processor) or "intelligence" improvement (e.g., game strategy) is difficult. The number of identifiable hardware innovations was relatively small, and since the leading teams effectively adopt any successful innovations of their competitors (within a delay of up to two years), than the rest of the yearly increase in the competence of the robotic teams is due to the increase in their AI.

It can be claimed that a better measure of the yearly progress of a team can be obtained by making that team play against the older version of the same team (e.g., older software and hardware). In practice, due to budgetary and human constraints,[7] old teams are not maintained in an operational state. On the other hand, within the framework of the RoboCup simulation league, it is possible to tweek older algorithms to compete against the modified ones. [20] used extensive experimentation between teams for measuring the progress in performance robustness tradeoffs within the RoboCup simulation league. Robustness was defined as graceful degradation for changes in the operating environment, and the goal difference was used as the major metric for dominance measurement. That research does not consider explicitly the different "intelligent methods" used by different teams. Also, using a simulated environment, the experiments depend on some simplified modeling of the physical aspects of the robots (such as kinematics) neglecting other aspects (e.g., wheel slippage).

RoboCup is an international joint project—an attempt to foster AI and intelligent robotics research by providing a standard problem where wide range of technologies can be integrated and examined. The method proposed here for quantifying the yearly progress in RoboCup, managed to demonstrate for the first time that the RoboCup competition is not yet another "sports event," but it definitely fulfills its premise to stimulate substantial scientific progress. The progress measure here is also an indirect indication of the progress in the domain of "intelligent system."

Though the proposed technique seems interesting only in competitive environments, it could be used also in the general framework of MRS: we can let two teams compete, which are identical except for one change. The rating difference between the teams would indicate the value of the change for the team performance. As far as we know, a standard technique does not exist for measuring the performance of a MRS. This technique could be used to quantify the improvement due to a possible hardware change (e.g., improved sensor resolution) or software change (e.g., tighter cooperation).

## References

[1] A. Newell, "The knowledge level," *Artif. Intell.*, vol. 18, no. 1, pp. 87–127, 1982.

[2] J. S. Albus, "Outline for a theory of intelligence," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC–21, pp. 473–509, May/June 1991.

[3] A. Meystel, "Measuring performance of systems with autonomy: metrics for intelligence of constructed systems," in *Proc. Measuring the Performance and Intelligence of Systems*, E. Messina and A. Meystel, Eds., Aug. 14–14, 2000.

[4] G. N. Saridis, "Analytic formulation of the principle of increasing precision with decreasing intelligence for intelligent machines," *Automatica*, vol. 25, no. 3, pp. 461–467, 1989.

[5] U. Lima and G. N. Saridis, *Design of Intelligent Control Systems Based on Hierarchical Stochastic Automata, Series on Intelligent Control and Intelligent Automation*, Singapore: World Scientific, 1996, vol. 2.

[6] A. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, pp. 433–460, 1950.

[7] P. Cohen, *Empirical Methods for Artificial Intelligence*. Cambridge, MA: MIT Press, 1995.

[8] S. Hanks, M. E. Pollack, and P. Cohen, "Benchmarks, test beds, controlled experimentation, and the design of agent architectures," *AI Mag.*, vol. 14, no. 4, pp. 17–42, 1993.

[9] G. Dudek, M. Jenkin, E. Milios, and D. Wilkes, "A taxonomy of multirobot systems," in *Robot Teams: From Diversity to Polymorphism*, T. Balch and L. E. Parker, Eds. Natick, MA: AK Peters, 2002.

[10] L. E. Parker, "Current state of art in multi-robot teams," in *Distributed Autonomous Robotic Systems*. New York: Springer-Verlag, 2000, vol. 4, pp. 3–12.

[11] T. Balch, "Hierarchic social entropy: an information theoretic measure of robot team diversity," *Auton. Robot.*, vol. 8, no. 3, pp. 209–238, 2000.

[12] D. Goldberg and M. J. Mataric, "Interference as a tool for designing and evaluating multi-robot controllers," in *Proc. 14th Natl. Conf. Artificial Intelligence*, New Providence, RI, 1997, pp. 637–642.

[13] G. A. Kaminka and M. Tambe, "Robust multi-agent teams via socially attentive monitoring," *J. Artif. Intell. Res.*, vol. 12, pp. 105–147, 2000.

[14] RoboCup Federation, Tokyo, Japan.. [Online]. Available: www.robocup.org

[15] A. E. Elo, *The Rating of Chess Players Past and Present*. New York: Arco, 1978.

[16] M. E. Glickman and A. C. Jones, "Rating the chess rating system," *Chance*, vol. 12, no. 2, pp. 21–28, 1999.

[17] M. E. Glickman, "Parameter estimation in large dynamic paired comparison experiments," *Appl. Stat.*, vol. 48, pp. 377–394, 1999.

[18] *RoboCup'98: Proc. 2nd Robot World Cup Competition Conferences*. New York: Springer-Verlag, 1998.

[19] *Proc. Intl. RoboCup Symp.*, G. A. Kaminka, P. U. Lima, and R. Rojas, Eds., Fukuoka, Japan, June 24–25, 2002.

[20] A. Kaminka, I. Frank, K. Arai, and T. I. Kumiko, "Performance competitions as research infrastructure: large scale comparative studies of multi-agent teams," *J. Auton. Agents Multi-Agent Syst.*, vol. 7, no. 1–2, pp. 121–144, 2003.

[21] A. Newell and H. Simon, "GPS: a program that simulates human thought," in *Computers and Thought*, H. Feigenbaum and H. Feldman, Eds. New York: McGraw-Hill, 1963.

[22] H. Feigenbaum and H. Feldman, *Computers and Thought*. New York: McGraw-Hill, 1963.

[6] Unfortunately, there are no standardized formal requirements about the sufficient level of details to reveal. Thus, the computation of other—technically oriented—benchmarks is not possible.

[7] Many teams are developed by frequently changing graduate students.