Evaluation of StarCraft Artificial Intelligence Competition Bots by Experienced Human Players

Man-Je Kim
Cognition and Intelligence Lab
Dept. of Computer Engineering
Sejong University
Seoul, South Korea
jaykim0104@gmail.com

SeungJun Kim
HCI Institute
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA, 15213 USA
sjunikim@cs.cmu.edu

Kyung-Joong Kim
Cognition and Intelligence Lab
Dept. of Computer Engineering
Sejong University
Seoul, South Korea
kimkj@sejong.ac.kr

Anind K. Dey
HCI Institute
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA, 15213 USA
anind@cs.cmu.edu

CHI'16 Extended Abstracts, May 07-12, 2016, San Jose, CA, USA
ACM 978-1-4503-4082-3/16/05.
http://dx.doi.org/10.1145/2851581.2892305

Abstract

StarCraft is one of the most successful real-time strategy (RTS) games and is also actively being researched by artificial intelligence (AI) communities. Since 2010, game AI researchers have hosted annual AI competition events to develop human-level RTS AIs using StarCraft. It ranks the AI bots by their winning ratio from thousands of AI vs. AI matches without human involvement. It is questionable whether successful AI bots are also competitive and preferable to human players. In this study, we invited 20 experienced players with varying expertise to evaluate skill levels, overall performance and human likeness of AI bots. Results show that human's ranking of AI bots are not identical to the current one from AI competitions. It suggests the need for developing new AI competitions that consider human factors ("human-likeness" or "adaptation"). Also, it revealed that the expertise levels of human players have high impact on overall performance and human-likeness evaluations of AI bots. It supports the concept of dynamically adjusting AI bots to satisfy different levels of human players. The outcomes of this study will also be useful to incorporate human factors in other active video AI competitions (e.g., Angry Birds, Fighting Game, and General Game Playing).

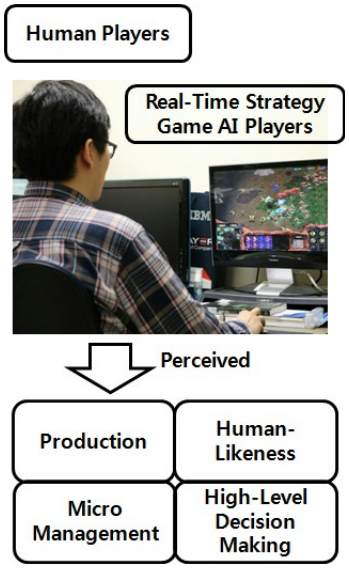

Figure 1: Human players are invited to play matches against successful StarCraft AI players. After games, human players provide reviews on AI players in terms of game skills, performance and human-likeness.

Author Keywords

Real-time Strategy Games; AI bots; AI Competitions; Turing Test.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous;

Introduction

Artificial intelligence (AI) has played an important role in enhancing gaming experiences of human players [1]. It can not only be a supporter in the background but also a trainee, co-creator, role-model and opponent in the foreground. For example, gamers can train AI creatures to perform desirable actions in video games [2]. As a co-creator, it automatically creates game contents from the interaction with players [3].

In real-time strategy (RTS) games, the use of AI is still too under-developed to satisfy human players [4]. In this game genre, players should respond quickly by handling a lot of different things at the same time such as unit control, resource utilization, building/unit production, and strategic decision making. It is a highly challenging task given the current state-of-the art AI techniques because of its complexity and real-time constraints [5].

Since 2010, annual RTS game AI competitions have been launched to promote the development of successful AI players [6]. They use StarCraft (by Blizzard), one of the most successful RTS games and associated AI programming interface (by hackers). Currently, there are three representative StarCraft AI competitions run by IEEE CIG, AIIDE, and SSCAIT[1] with 10~50 entries. In a competition, each entry plays many games against other opponents on different maps. In sum, they require more than several thousand matches to identify the winner [7] and select the best AI purely based on the results of AI vs. AI matches.

Eventually, AI bots need to interact with human players and it is important to consider human factors. In this study, we aim to see whether the traditional evaluation method for RTS AIs produces satisfactory results for human players. Until now, human-based evaluation of RTS AI players has been performed in a limited manner. Weber *et al*. tested only their EIS[2] bot against human players at an online gaming site using win/lose ratio without skill-level analysis or human-likeness testing [8]. In AIIDE competition, the winning AI plays several games against one human expert player [9]. However, it didn't include broad evaluation by many experienced players with different levels of expertise. In this study, we invited 20 experienced human players with different expertise to play matches against 7 successful StarCraft AI bots (see Figure 1). It can help us to understand the potential difference between human players' evaluation and the current AI rankings. In addition, it can show us how varying expertise impacts their viewpoints and evaluations.

Experimental Method

StarCraft was first released in 1998 and was a commercial success, selling 10 million copies (See

[1] CIG (IEEE Conf. on Computational Intelligence and Games), AIIDE (AAAI Conf. on Artificial Intelligence and Interactive Digital Entertainment), and SSCAIT (Student StarCraft AI Tournament)

[2] Expressive Intelligence Studio

Figure 2. A screenshot of StarCraft game for Terran race (the yellow callouts are added to illustrate some features of the game)

Figure 2). It is a good platform to experiment with because it has lots of experienced human players with different expertise and diverse AI players from annual AI competitions. Usually, it is played online and supports multi-player games. Before starting a game, each player needs to select one of three races (Protoss, Terran, and Zerg) for an army. During gameplay, each player commands military units to mine minerals/gas, scout opponent's territories, create buildings, and combat. The goal of the game is to eliminate all the buildings of opponents and it usually takes about 5 min ~ more than 1 hour. Although there are lots of different settings of game matches, in this study, we used single human player versus AI player match. A popular map "Fighting Sprit 1.3" was used.

We invited 20 experienced StarCraft gamers who had a history of 800 or more matches (age *M*=25.7, *SD*=3.7, age range: 18-31, win rate *M*=64.85, *SD*=12.58, win rate range: 41-89 (%), years' experience *M*=7.8, *SD*=2.9, year's experience range: 2-14 (years), gender: male 90% and female 10%). Win rate indicates their performance at the US West BattleNet (a game server for StarCraft). Using their win rates, we classified the players into four categories: A (win rate ≥ 80%, 2 players), B (≥ 70%, 5 players), C (≥ 60%, 7 players), and D (≥ 50% and below, 6 players). Because it is not practical to master all the three races, human players mainly use only one race based on preference. Their chosen races were Protoss (45%), Terran (45%), and Zerg (10%).

We used six AI bots highly ranked (win rate over 50%) from the IEEE CIG 2014 StarCraft AI Competition and the winner of the CIG 2013 competition[3] (win rate M=71.7, SD=12.9, win rate range: 55-91 (%)). Win rate is the result of thousands of matches among AI players submitted in the year's competition (13 submissions in 2014, 8 in 2013). We included the winner of the 2013 competition to see the improvement from 2013 to 2014. The seven AI bots included were five Protoss and two Terran bots, and were developed by students, researchers and freelancers. The bots are ICEBOT (rank = 1st, win rate = 83%), Ximp (2nd, 78%), LetaBot (3rd, 68%), AIUR (4th, 66%), UAlbertaBot (5th, 60%), and MaasCraft (7th, 55%) from IEEE CIG 2014 and SkyNet (1st, 91%) from IEEE CIG 2013.
We instructed each human player to play one match against each bot (the order of the bots was not randomized). In total, 140 games were played (20 players × 7 bots). After the games, each player filled out questionnaires on the perceived performance of the AI players. It includes five questions about the player (win ratio, years of experience, gender, StarCraft race, and play style), six questions about the AI players (indication of the top three players for each criteria and comments) and overall comment. The survey was designed based on two professional (officially licensed) and one semi-professional players' opinion and included five criteria.

- **Production (PD)**: Capability to produce units/buildings massively and efficiently
- **Micro Management (MM)**: Skills in controlling individual units
- **Combat (CB)**: Skills in controlling armies to win combats
- **Decision Making (DM)**: Strategic/Tactical decision making under uncertainty

[3] http://cilab.sejong.ac.kr/sc_competition/

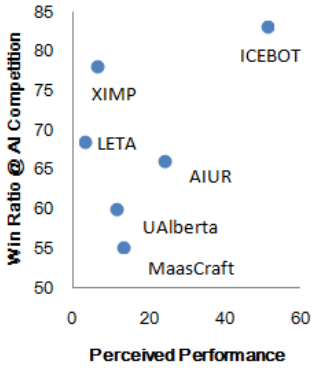

Figure 3. Relationship between "perceived" performance by human players and "objective" win ratio from AI competitions.

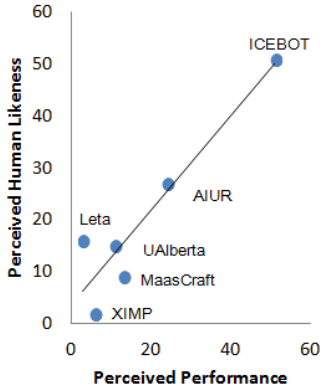

Figure 4. High correlation between perceived performance and human-likeness

- **Performance (PM)**: Overall evaluation

For each criterion, each human player selected three AIs and ranked them (1st = 3 points, 2nd = 2 points, and 3rd = 1 point). Their final ranking was calculated by summing all the scores given by 20 human players. The score for an AI bot thus ranges from 0 to 60.

Evaluation of "human-likeness" is a bit complex. It is important that the human evaluator has no information about the identity of their opponents (AI Bot or Human). In a First-person shooting study, additional human players were used as opponents to confuse human evaluators to critique "human-likeness" [10]. Similarly, we added one human player who plays both of Protoss and Terran races. The 20 human players played an additional eight matches (randomly with 2 human (switching race) and 6 AI bots (SkyNet was not included)). Because they played the games remotely, opponents' identities are hidden from them. They evaluated the 8 opponents by selecting top three AIs for "Human-Likeness" (HL).

Analysis Result 1 – Difference between Human Players' and Traditional Evaluation

Table 1 compares the seven AI bots "objective" performance at the 2014 competitions sorted by win ratio and the subjective scores from human players' evaluations. It shows that the best AI (ICEBOT) also was evaluated as a top player by humans. Except for "production," ICEBOT's score was about two times higher than the runner-up (the bot with the second largest score in each category). Although the ICEBOT was seen as the dominant player by humans, its win ratio (83%) at the AI competition is quite close to the runner-up XIMP (78%). It means that the current AI competitions' outcome does not reflect the performance gap observed by human players exactly.

AI Competition			Perceived by Human					
Bot Name	Rank	Win Ratio (WR) (%)	Production (PD)	Decision Making (DM)	Micro Management (MM)	Combat (CB)	Performance (PM)	Human Likeness (HL)
ICEBot	1	83	27	56	54	47	51	51
Ximp	2	78	3	2	12	20	6	2
Leta	3	68	1	6	7	15	3	16
AIUR	4	66	36	23	8	2	24	27
UAlberta	5	60	16	6	12	4	11	15
MaasCraft	6	55	8	3	11	8	13	9
SkyNet	2013 Winner	-	29	24	16	24	12	-

Table 1: Evaluations from traditional AI-oriented competitions and scores by experienced human players. The shaded box shows the best AI for each criterion.

Figure 3 shows the correlation between human's perceived performance and objective AI's win ratio at the competition. It shows two AI bots (XIMP and Leta) evaluated in the opposite way. Although they're highly ranked (2nd and 3rd place) in AI competitions, their "perceived" performance is lowest in scoring. In fact, XIMP (2nd place) exploits a strategy only successful against AI bots but not humans. It starts a game in a very defensive way but waits until it prepares enough strong attack units. Although this defense is not a tough one, most AI bots fail to pass the wall. In the end, the AI bots lose the game because of XIMP's massively strong attack units.

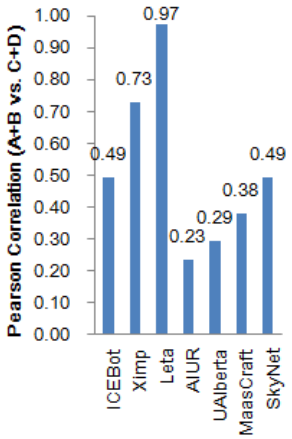

Figure 5. Pearson correlation of evaluations from two groups (A+B vs. C+D).

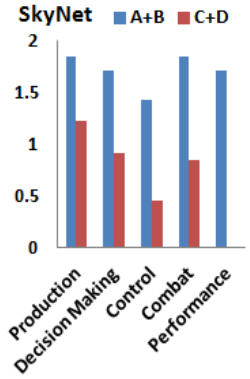

Figure 6. Comparison of evaluations from two different groups (A+B vs. C+D) for SkyNet

Table 2 shows the correlation of "objective" and "perceived" evaluation measures. It shows that the win ratio at the competition is not highly correlated with "perceived" measures except for "Combat" (0.83). It means that players' evaluations cannot predict success in the AI competition. Currently, to win the competition, it is very important for the AI to have good combat skills. It explains why the XIMP was ranked high in the competition although it is ranked near the bottom in all "perceived" measures except combat skills. Among the four major skills (PD, MM, DM and CB), the micromanagement and decision making skills are highly related to the final overall perceived performance (PM) by humans. It means that human players usually give high weight on these two skills. This is a big difference between AI competition (combat is important) and human players (micromanagement and decision making are important).

	AI	**Perceived by Human**				
	WR	PD	DM	MM	CB	PM
PD	0.12					
DM	0.62	0.69				
MM	0.66	0.34	0.88			
CB	0.83	0.14	0.76	0.88		
PM	0.51	0.61	0.93	0.89	0.65	
HL	0.49	0.71	0.97	0.83	0.65	0.93

Table 2: Pearson correlation among seven measures (one from AI competition and six from human players) and the shaded box shows correlation > 0.8.

In the human-like testing, we included one human player with other AI bots to confuse human evaluators. However, it's revealed that no human player is fooled by the human opponent. All evaluators can identify humans from AI bots successfully. It reveals the significant difference between human players and AI bots in terms of "human-likeness". Human players noted that the AI players are usually showing strange behaviors (not seen in human matches) or mechanical placements of units/buildings. As a result, the perceived "performance" is highly related to "human-likeness" with correlation (0.93) (see Figure 4).

Based on the results, we have found that the "combat" skills are more important to win more games in AI competitions than other skills. However, for human players, "micro management" and "decision making" are highly important to evaluate AI players. For RTS AI game developers, it's important to focus on the MM and DM to satisfy experienced human players. For AI competition organizers, it's important to devise a way to make AI bots with strong MM and DM can win the competition.

Analysis Result 2 – Human Players' Evaluation with Different Expertise

To see viewpoints of human players with different expertise, we grouped A+B (7 players, win ratio ≥70%) and C+D (13 players, win ratio < 70%). Figure 5 shows the Pearson correlation between two groups' evaluations (six perceived measures). If the correlation is high, it means that the two groups evaluated the AI bots similarly. If not, the two groups show differences in their evaluations.

To the more expert group (A+B), it was more important to have balance between several basic skills rather than having strength or completeness in each skill. However, the C+D group gave higher scores to bots with fully developed skills instead of overall balance or harmony amongst them. XIMP (2nd) and

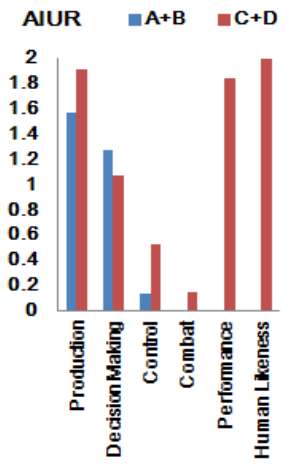

Figure 7. Comparison of evaluations from two different groups (A+B vs. C+D) for AIUR

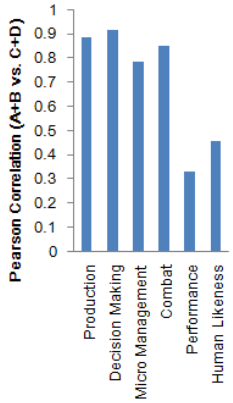

Figure 8. Pearson Correlation of each measure between two groups (A+B vs. C+D)

Leta (3rd), were evaluated similarly by the two groups (Pearson > 0.7). However, the evaluation results differed for the other bots (see Figure 5). For SkyNet, the two groups show similar patterns of evaluations but the A+B group more highly evaluated than the C+D group (see Figure 6). The SkyNet has skills that are well balanced although each skill is not fully developed.

The more expert players give high weight to the overall balance and harmony instead of the strength of individual skills. However, the less expert players (C+D) were more likely to be biased based on the strength of individual skills although they are not well balanced. For example, the C+D group included the AIUR (Pearson = 0.23) in the top three best players for the performance and human-likeness measures (see Figure 7). However, A+B group did not.

This analysis shows that the two groups have different viewpoints when evaluating overall performance. Figure 8 shows the correlation of the two groups for the six perceived measures. It's interesting that the four basic skills (Production, Decision Making, Micro Management, and Combat) has high correlation between the two groups (>0.78). However, their correlation is low in "performance" and "human-likeness" evaluations (Pearson = 0.33 and 0.45, respectively). The design implication of these results is that game AI developers need to incorporate the dynamic skill weighting for different levels of AI players. For example, if the game player has low expertise, it is desirable to increase the strength of each basic skill, rather than spending resources for integration or balancing. However, for the highly expert group, it is moreimportant to give high weight on the balancing of basic skills to satisfy them. This information will help to design AI bots dynamically based on the expertise of human game players for game AI developers.

Limitations and Future Works

In this study, we showed that there are interesting mismatches between human evaluations and rankings by the current AI competitions in RTS games. Also, human players have different viewpoints of AI bots depending on their expertise in the game. It opens a new research discussion on the best way to evaluate the AI bots (e.g., ranking AI bots with a "Turing-Test", special competition tracks targeting human factors in game AIs and officially including "Machine vs. Human matches"). Also, for AI designers, it is important to design dynamically adjustable AI players in consideration of human opponent's expertise.

One limitation of our work is that a small number of games may not allow human players to see the full capability of AI players. Recently, some AI bots have included an "adaptation" capability to change their strategy across games. To see this ability, it is necessary to allow human players to play multiple games against such a bot. Although this study focuses on the user's evaluation over whole game, it may also be desirable to see correlation between human players' feedback and specific events in a game [14]. Finally, the findings of this study need to be further validated by using other video game AI competitions [11] such as for Angry Birds [12], Fighting Game [15], Geometry Friends [16], and General Game Playing [13].

Acknowledgements

This work was supported by the Korean government (Ministry of Science, ICT & Future Planning) (2013 R1A2A2A01016589, R22151510080001002).

References

1. M. Treanor *et al*. 2015. AI-based game design patterns. 10th International Conference on the Foundations of Digital Games.
2. I. V. Karpov, V. K. Valsalam, and R. Miikkulainen. 2011. Human-assisted neuroevolution through shaping, advice and examples. Proc. Of the 13th Annual Genetic and Evolutionary Computation Conference.
3. N. Shaker, J. Togelius, and M. J. Nelson. 2015. Procedural Content Generation in Games: A Textbook and an Overview of Current Research. Springer.
4. S. Ontanon, G. Synnaeve, A. Uriarte, F. Richoux, D. Churchill, and M. Preuss. 2013. A survey of real-time strategy game AI research and competition in StarCraft. IEEE Trans. on Computational Intelligence and AI in Games. 5, 4, 293-311.
5. M. Buro. 2003. Real-time strategy games: A new AI research challenge. Proc. Of the Int. Joint Conference on AI. 1534-1535.
6. M. Buro, and D. Churchill. 2012. Real-time strategy game competition. AI Magazine. 33, 3, 106-108.
7. S. Farooq, I.-S. Oh, M.-J. Kim, and K.-J. Kim. 2016. StarCraft AI Competition: A Step Toward Human-Level AI for Real-Time Strategy Games. AI Magazine.
8. B. G. Weber, M. Mateas, and A. Jhala. 2011. Building human-level AI for real-time strategy games. AAAI Fall Symposium: Advances in Cognitive Systems.
9. D. Churchill. 2015. AIIDE StarCraft AI Competition Report. https://webdocs.cs.ualberta.ca/~cdavid/starcraftaicomp/report2015.shtml
10. P. Hingston. 2009. A Turing test for computer game bots. IEEE Trans. on CI and AI in Games. 1, 3, 169-186.
11. K.-J. Kim and S.-B. Cho. 2013. Game AI competitions: An open platform for computational intelligence education. IEEE Computational Intelligence Magazine. August issue.
12. D.-M. Yoon and K.-J. Kim. 2015. Challenges and opportunities in game artificial intelligence education using Angry Birds. IEEE Access. June.
13. M. Swiechowski, H.-S. Park, J. Mandziuk and K.-J. Kim. 2015. Recent advances in general game playing. The Scientific World Journal. August.
14. G. Robertson and I. Watson. 2014. An improved dataset and extraction process for StarCraft AI. Proc. Of 27th Int. Florida AI Research Society Conference.
15. F. Lu, K. Yamamoto, L. H. Nomura, S. Mizuno, Y.M. Lee, and R. Thawonmas. 2013. Fighting game artificial intelligence competition platform. IEEE Global Conference on Consumer Electronics. 320-323.
16. J. Quiterio, R. Prada, F. S. Melo. 2015. A reinforcement learning approach for the circle agent of geometry friends. IEEE Conf. on Computational Intelligence and Games. 423-430.