IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 6, 2017, accepted January 14, 2018, date of publication January 31, 2018, date of current version March 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2800016

Performance Evaluation Gaps in a Real-Time Strategy Game Between Human and Artificial Intelligence Players

MAN-JE KIM[1], KYUNG-JOONG KIM[1], SEUNGJUN KIM[2], AND ANIND K. DEY[3]

[1]Department of Computer Engineering, Sejong University, Seoul 143-747, South Korea
[2]Institute of Integrated Technology, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea
[3]Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Corresponding author: Kyung-Joong Kim (kimkj@sejong.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea, Ministry of Science, ICT and Future Planning, under Grant 2017R1A2B4002164, and in part by the GIST Global University Project in 2018.

ABSTRACT Since 2010, annual StarCraft artificial intelligence (AI) competitions have promoted the development of successful AI players for complex real-time strategy games. In these competitions, AI players are ranked based on their win ratio over thousands of head-to-head matches. Although simple and easily implemented, this evaluation scheme may less adequately help develop more human-competitive AI players. In this paper, we recruited 45 human StarCraft players at different expertise levels (expert/medium/novice) and asked them to play against the 18 top AI players selected from the five years of competitions (2011–2015). The results show that the human evaluations of AI players differ substantially from the current standard evaluation and ranking method. In fact, from a human standpoint, there has been little progress in the quality of StarCraft AI players over the years. It is even possible that AI-only tournaments can lead to AIs being created that are unacceptable competitors for humans. This paper is the first to systematically explore the human evaluation of AI players, the evolution of AI players, and the differences between human perception and tournament-based evaluations. The discoveries from this paper can support AI developers in game companies and AI tournament organizers to better incorporate the perspective of human users into their AI systems.

INDEX TERMS Video game, Starcraft, game, artificial intelligence, game AI competition, human factor, human computer interaction.

I. INTRODUCTION

As one of the most successful real-time strategy games in history, StarCraft has had considerable impact in changing gaming culture around the world (Figure 1). Soon after its 1998 release by Blizzard, professional StarCraft teams sponsored by big companies and broadcasting channels have emerged, and they regularly take part in large StarCraft tournaments. Interestingly, pro-gamers have gained celebrity-level fame and attracted young fans to their matches. Cheung *et al.* studied these spectators to understand why they watch the matches, the differences between them and other game-related stakeholders, and their viewing experiences [1]. In addition to watching the show in person, fans share professional and amateur matches recorded in files through online video game communities. In fact, one dedicated StarCraft replay site contains more than 250,000 files (bwreplays.com).

In video games, creating artifacts that can interact with human players to provide a more immersive and engaging gaming experience has become increasingly important. Usually, artificial intelligence (AI) players, also known as NPCs (Non-Player Characters), act as enemies of human players [2]. However, recent studies on game AI have investigated new AI player roles for modern video games. For example, AI players can work alongside human players, replace human players as avatars, and teach novice players by demonstrating various game skills, abilities, and strategies [3], [4]. Because human players are unable to simultaneously take control of every aspect of contents in the latest mobile & wearable games, they widely adopt automated function (e.g., auto combat), which leads to the wide use of AI players during the play.

Since 2010, StarCraft has been available for researchers to develop customized AI players using the Brood War

Application Programming Interface (BWAPI), a specialized programming interface created by hackers. The availability of this program has opened new research opportunities to create human-level AI players for a complex video game that involves real-time reaction, uncertainty, simultaneous control of hundreds of units, and strategic Decision Making. For the past five years, the number of participants in annual StarCraft AI competitions has increased steadily [8]. Currently, there are three annual StarCraft AI competitions that promote the development of improved AI players [5]. In these competitions, an AI player is ranked based on its win-loss ratio of matches against all other AI players. To determine the winner of the competition, the competition usually runs thousands of AI vs. AI matches.

Unsurprisingly, the rules of competition significantly impact AI developers' designs as they investigate how to increase their win ratio to achieve a higher ranking in the competition, and how to design their AIs to beat the best AIs from the previous year. Until now, an evaluation scheme based on win ratio has been successfully used in AI gaming competitions, including StarCraft. However, it is still unclear whether such AI-oriented evaluation can actually improve AI or bot performance from the perspective of a human player. While only the outcome (win/loss) of a match is currently used to guide the evaluation of AIs, humans are likely to evaluate their opponents in different ways.

To address this concern about the discrepancy in how current tournaments and humans would likely evaluate AI players, we collected data from 270 matches between 45 human players (ranging in expertise from novice to expert) and 18 AI players (selected from tournaments over the past five years), and matches from AI only games. From this study, we aim to get 1) an in-depth view of how human players judge current state-of-the art AI players and 2) an understanding of AI player progress over the last 5 years based on competitions driven by the win-loss ratio criterion.

The contribution of our study is summarized as follows:

First, we attempt to understand the differences between AI assessment and human assessment. A key difference between the two assessments is to involve human players in the evaluation of AI players. Until now, there have been a few studies that explored this discrepancy in the two evaluation schemes, but those are quite limited in scope (see the next section for more details). We establish that AI-only competitions are not a good way of measuring AI progress for playing against humans. Outside of these tournaments, AI players are designed to play against human players, so human assessment of these AI players would seem to be very important. However, the current AI-only competitions produce AIs that only have a high win ratio against AI players.

Second, this study explores the progress of AI players over the last 5 years. The effort to create better AIs has continued for several years but it is important to see whether actual progress has been made. We use two different assessment schemes: human (described above) and longitudinal. Traditionally, only AI vs. AI matches have occurred during a single year, with only the resultant win rates considered for assessment. In this context, our study provides the first longitudinal analysis and results for AI vs. AI matches for AIs developed over the past 5 years and also the results of AI vs. human matches.

FIGURE 1. Screenshot of the StarCraft game (the left-bottom mini-map visualizes player/opponents' units/buildings as yellow/red dots and the black/dark areas represent areas that are obscured by the fog-of-war. In the left top, the blue objects are minerals the primary resource necessary to produce units and buildings. The screen shows the Zerg attack units down below attempting to invade the Terran territory located up on the hill.

Finally, this study will help connect game AI developers and game users to design a better gaming experience by introducing the importance of tailoring AI players for users' expertise. It provides justification for AI developers to design multiple AI players fitting their target users instead of building a single strong AI for a specific purpose.

We will now describe the video game AI competitions and how Starcraft is played to provide background for understanding the use of AI players. Following that, we will describe our methods for evaluating AI players and the results of our evaluation. We will conclude with a discussion of the implications of those results.

II. VIDEO GAME AI COMPETITIONS

In game AI communities, several competitions have been organized as special events to promote the development of AI players [6]. For example, there have been Angry Birds, StarCraft, Pac-Man, Super Mario, First-Person Shooting (FPS), and Racing Game AI competitions. Although most of the competitions focus on win ratio, lap time, or scores recorded by AI players, some competitions have introduced special rules to determine winners. For example, the Bot Prize competition uses the game Unreal Tournament (FPS genre) and focuses its evaluation on ''Human-Likeness'', with their tournament designed as a video game Turing test [7]. In the competition, human players play with AI players but the identity (human or bot) of each player is hidden. The more an AI player is believed to be a human, the higher the bot is ranked.

Table 1 summarizes the video game AI competitions and their evaluation methods. While we believe it is essential

TABLE 1. Evaluation methods in video game AI competitions.

Type	Game	Track	Ranking Criterion
Solo play	Super Mario [[9]]	Play	Scores
		Learning	Scores
		Level Generation [[10]]	**Human Evaluation**
	Racing Games [[11]]	Play	Lap Time
	Ms. Pac Man	Play	Scores
	Angry Birds [[12]]	Play	Scores
	General Video Game Playing [[13]]	Play	Scores
Multi play	StarCraft	Play	Win Ratio
		Post Competition Match [[14]]	**Human Evaluation**
		Post Competition Evaluation [[15]]	**Human Evaluation**
	Unreal Tournament [[7]]	Turing Test	**Human Likeness**

to incorporate with human evaluation in the evaluation of AI players, which will eventually interact with human players, so far this has only been done in a very limited manner. As described above, the Bot Prize competition scores bots based on the number of human players who are misled to think that a particular bot is a human player [7]. In Super Mario, human evaluators play two AI-generated levels and choose the one that is more fun to play [10]. In StarCraft, there are two groups that have taken a similar approach to our proposed approach: evaluation based on a match between a human player and an AI. Churchill reported on the AIIDE competition where an AI played 2-3 matches against a human [14]. Weber *et al.* [15] evaluated their EISBot by playing 250 games against human players as part of the International Cyber Cup (ICCup) after the AIIDE 2010 competition. While both of these are promising, they are limited in that they do not help us understand the differences between the traditional and human player-based evaluation, nor do they allow us to see how AI players have progressed, if at all, over time.

III. StarCraft GAMEPLAY

In the StarCraft game, players need to select one of the three races (Protoss, Terran, and Zerg). The three races have different types of units and buildings and use different strategies, but are balanced in terms of overall performance.

After selecting a race, players produce military units and create buildings to win the game. To win, a player must destroy the opponent's buildings, or the opponent must surrender. Because of the ''fog-of-war'' gameplay element which obscures a player's vision, players can only see the area around their own units; in the early stages of the game, each player only has a limited view of the other player's territory. After the game starts, human players usually send out a ''scouting unit'' to find an opponent's position and observe unit production and buildings. Because players try to hide their plan from opponents, they usually attack the ''scouting unit'' as quickly as possible, and either kill the scouting unit or force it to retreat to survive. Because players have a limited view of their opponent's territory, they need to infer the activity of opponent players hidden under the fog.

In the early stages of the game, the player strategy is reflected on a build order (similar to the opening approach in board games) which determines the sequence of building creation and unit production. For experienced players, their build order is highly optimized to produce the right number of units and buildings to fulfill the goal of the player's strategy, although, it needs to adapt if it is not suitable for the changing situation. For example, if you recognize that an opponent is preparing an early stage attack and is investing all of its resources in producing attack units quickly, it is reasonable to change one's build order to hold it back first instead of taking a more traditional action sequence of building balanced unit types.

After the initial stages of the game, players need to expand their territory to new mineral and gas areas. While games usually only take an average of 10 minutes to complete, some games can continue for over an hour; in this situation, it is important that the players maintain enough resources to stay strong/competitive for long stretches of time. In the middle or latter parts of the game, it is important to control unit production and resource management effectively in order to overcome opponents and eventually win the game.

IV. METHODS

In our study, we compare human evaluations of AI players with the traditional evaluation of ranking AI players based on the win-loss ratios of AI vs. AI matches. We used 18 StarCraft AI players from across the last five years and invited 45 human players (15 novice, 15 medium, and 15 expert players). If every human player played every AI player, each human player would play 18 games. With an average playing time of 10 minutes per game, this would take about 6 hours, not including breaks. As this amount of time was not feasible for our human player participants, we arranged matches so that each AI player has 5 matches against human players from each expertise group. In other words, each AI will play against 5 novice, 5 medium, and 5 expert players. In total, this requires 270 matches (18 AIs $\times$ 15 matches). This means that each human player will play a single match with six different AIs (45 human players $\times$ 6 matches), for an average total of 1 hour of gameplay. The exact process for our experiment is described below:

- **Step 1**: The human player's experience level is identified as either novice, medium, or expert based on their official game record, license, or trainee experience (more details on this below). When we were unsure about the player's experience level, we let the player play a single match against an expert player (who holds a semi-professional license) to determine his/her appropriate expertise group.
- **Step 2**: The human player answers our pre-questionnaire about demographics, StarCraft experience, preferred game race, and what he/she thinks the most important skills are for the game.
- **Step 3**: The human player conducts a single match against each of the six AI players assigned to them (six matches total), with a 5-minute break in between games. After each game, the player answers a post-game questionnaire to evaluate the AI just played in terms of several StarCraft skills. The evaluation includes player's numeric scores and their justification for seven different criteria (see below).
- **Step 4**: Finally, the human player provides a relative ranking of all six AIs played and provides some overall comments on StarCraft AIs.

A. EVALUATION CRITERIA

To guide human players in their evaluation of AI StarCraft players, we adopted several different criteria and combined them to provide a final evaluation. The seven evaluation criteria were derived by expert (human) players from the StarCraft community.

- **Human Likeness (HL)**: This measures the similarity of AI and human players. If the AI plays like a human, it gets high scores. The AI does not necessarily need to be a strong player to get high human-likeness scores.
- **Decision Making (DM)**: This measures the quality of the AI's decision making. In StarCraft, players need to make a lot of decisions. However, in this study, we focus only on decisions related to combat to make this evaluation easier for the human players. In combat, decisions are needed about whether to advance or withdraw forces against the enemy's army.
- **Production (PD)**: This measures the quality of unit production for offense. If an AI player is very successful, it can produce attack units at high volume and speed.
- **Operation (OP)**: From the middle of the game, players need to expand their territories to new resource areas and manage lots of different production/combat activities. Usually, it requires highly balanced actions among resource (minerals and gas) management, and the creation, movement, and maintenance of units and buildings over large maps.
- **Build Order (BO)**: In the early stages of the game, time and resources are highly limited. It is important to decide which buildings or units must be prioritized over others. For example, if the player intends to attack very quickly and surprise opponents, it is necessary to ignore long-term plans and instead invest all their resources in producing attack units as quickly as possible. This metric also includes the players' skills in changing the build order as necessary based on scouting information.
- **Micro Control (MC)**: When there is combat between two players, it is important to control units carefully. Even if a player has more units than their opponent, it is possible to lose the combat through poor unit control. Professional players maximize the chance of combat wins by positioning units effectively and attacking enemy units selectively. This metric also includes control of non-attacking units (scouting units or workers).
- **Performance (PM)**: This metric is the overall evaluation of the performance of the AI player. It is not just a simple summation of the individual skill scores but is separately defined.

B. AI BOTS FROM THE StarCraft AI COMPETITION

StarCraft AI Competitions have been organized since 2010 and there are currently three competitions per year for the game: the IEEE CIG StarCraft AI Competition, the AAAI AIIDE StarCraft AI Competition, and the Student StarCraft AI Tournament (SSCAIT). In an archive site on StarCraft AI [16], source code or executable files for AIs for the IEEE CIG and AIIDE competitions can be downloaded. Because AIIDE has more entries than the IEEE CIG events, we adopted the AIs from the AIIDE events held from 2011-2015. The number of entries in the AIIDE StarCraft AI competition increased from 13 (2011) to 22 (2015). In this study, we consider the top three performing AIs for each year's event. The rank of each AI was determined using its win-loss ratio from the full round robin style tournament results of all entries submitted in a particular year. For AIIDE 2015, each AI played against every other AI 90 times. The 22 AIs played about 20,000 games ($(22 \times 21)/2 \times 90$) to determine the final ranks. For the five years we considered, the top ranked AI for each year usually achieved an 80-90% win ratio. In addition to using the three top ranked AI players from each year, from 2015 we also include the three AIs ranked 4th through 6th to help understand the single year variation for 2015. In total, 18 AIs over the five years were selected for our study.

Table 2 summarizes the names and races of the eighteen StarCraft AI players. Human players tend to have a strong preference for using a particular race rather than playing all races. AI players are similar in that developers focus on only one race for their bots. Very rarely, some players and AIs will choose to play with a randomly selected race.

In the early days of the competition (2011-2013), the number of AI participants was small (8-13), and three Protoss-based AIs (UAlbertaBot, Skynet, and Aiur) dominated the competitions. However, since 2014, this situation has changed, with the more diverse winning races including Zerg and Terran. In addition, from 2014, the number of AI entries increased to 18 and 22. A number of these AIs are updated based on the previous year's performances and are

TABLE 2. Name and race of eighteen StarCraft AI bots chosen from 2011 to 2015 (P = Protoss, T = Terran, Z = Zerg, and R = Random race).

	Rank			
Year	1st	2nd	3rd	4th~6th
2015	Tscmoo (Z)	ZZZKbot (Z)	Overkill (Z)	UAlberta (R)
				Aiur (P)
				Ximp (P)
2014	IceBot (T)	Ximp (P)	LetaBot (T)	-
2013	UAlberta (P)	Skynet (P)	Aiur (P)	-
2012	Skynet (P)	Aiur (P)	UAlberta (P)	-
2011	Skynet (P)	UAlberta (P)	Aiur (P)	-

re-entered each year. For example, the Aiur AI has participated in all five of the annual competitions. To distinguish the same AIs with different performances over several years, the bot is named with the year of participation. For example, Skynet12 refers to the Skynet AI from the 2012 competition.

Each AI player has a win ratio record based on the matches against other bots from the same year, but these records do not tell us anything about the relative rankings across years. It is impossible to directly compare two win ratios from different competitions and conclude anything meaningful. In this study, we run a new competition amongst the AIs we selected, following the exact rules of the AIIDE competitions. Each AI player can then be ranked based on the win ratio against the remaining AI players across all years.

C. StarCraft AI COMPETITIONS WITH AIs OVER FIVE YEARS

Currently, the ranking of the StarCraft AI competitions is based on the win ratio of an AI bot against other AIs. Because such an evaluation is performed only with bots submitted in the same year, it is hard to see whether AI performance is improving over the years. To assess this, we wanted to run a StarCraft AI competition using all 18 AI players selected over the past five years. However, we were unable to use three 2011 AI bots (Skynet11, UAlberta11, and Aiur11) because they were based on an old version of BWAPI with the current StarCraft tournament manager does not support. In total, 15 AI bots (2012-2015) participated in the competition. The competition follows the same rule used in AIIDE competition with 90 rounds. A total of 9,450 AI vs. AI matches were conducted using 10 different game maps.

Table 3 summarizes the win ratio of the 15 AI bots from the cross-year tournament. The 2015 bots were ranked from 1st to 5th except for Aiur15. It means that the top 2015 AI bots were an improvement over the previous years' AIs, when the criterion is simply a win-loss ratio. The overall winner was Tscmoo15 (winner of AIIDE 2015). It is interesting that each bot has a different win ratio against all AIs, AIs from the same year, from different years, from past competitions, and from future competitions. For example, IceBot14, which ranked 1st in 2014, had a very good record against other 2014 AI bots (87%) but performed poorly against AIs from other years

TABLE 3. Win ratio (%) of AI bots from the cross-year competition (shading indicates the highest value).

		vs. ALL	*vs.* Same Year	*vs.* Diff Year	*vs.* Past	*vs.* Future
Tscmoo15		**79**	62	**88**	88	-
ZZZKBot15		76	57	86	86	-
Overkill15		69	54	76	76	-
UAlbertaBot15		65	70	62	62	-
Ximp15		59	25	79	79	-
LetaBot14		58	38	62	**94**	29
Ximp14		58	25	64	89	**38**
IceBot14		53	**87**	47	67	27
Aiur15		45	32	52	52	-
UAlbertaBot13		41	60	38	81	24
Skynet13		40	73	34	77	20
Skynet12		37	80	29	-	29
Aiur13		28	17	30	38	28
Aiur12		28	37	27	-	27
UAlbertaBot12		14	33	11	-	11
Average		50	50	52	74	25

as a whole (47%); this was based on medium performance against past AIs (67%), but low performance against future AIs (27%). Averaged across all AIs, the win ratio against past AIs was 74% but the win ratio against future bots was 25%, indicating progress over time in the quality of AI players.

D. PARTICIPANTS IN THE HUMAN PLAYER EVALUATION

In this study, we recruited 45 human players. Gamers aged between 20 and 30 years old often have some StarCraft experiences because the game was so popular when they were teens. To see the impact of player expertise on the evaluation of AI players, we recruited 15 novice, 15 medium, and 15 expert players. To determine the expertise levels of a player, we used the following rules.

- Win Ratio on BattleNet: Usually, StarCraft was played through BattleNet (an online game network from Blizzard). If human players had their BattleNet game records, we used that data to determine the expertise level of players. We used the following expertise decision criteria: win ratio $\geq$ 70% for expert, 50% $<$ win ratio $<$ 70% for medium, and win ratio $\leq$ 50% for novice players.
- License or Professional gamer experience: When StarCraft was very popular, the official E-sports organization related to StarCraft tournaments issued special licenses to talented players to allow them to enter the leagues. If the player has a license, it meant that he was a semi-professional player. Also, some players have experience as trainees on professional teams sponsored by companies. Being a trainee meant that the player was a very skillful player. We group both the licensed players and professional trainees into the expert player category.
- Test Game: If the player has no official record, license, or career experience, a special match was arranged against an expert player (a member of our research team). During the match, the expert player

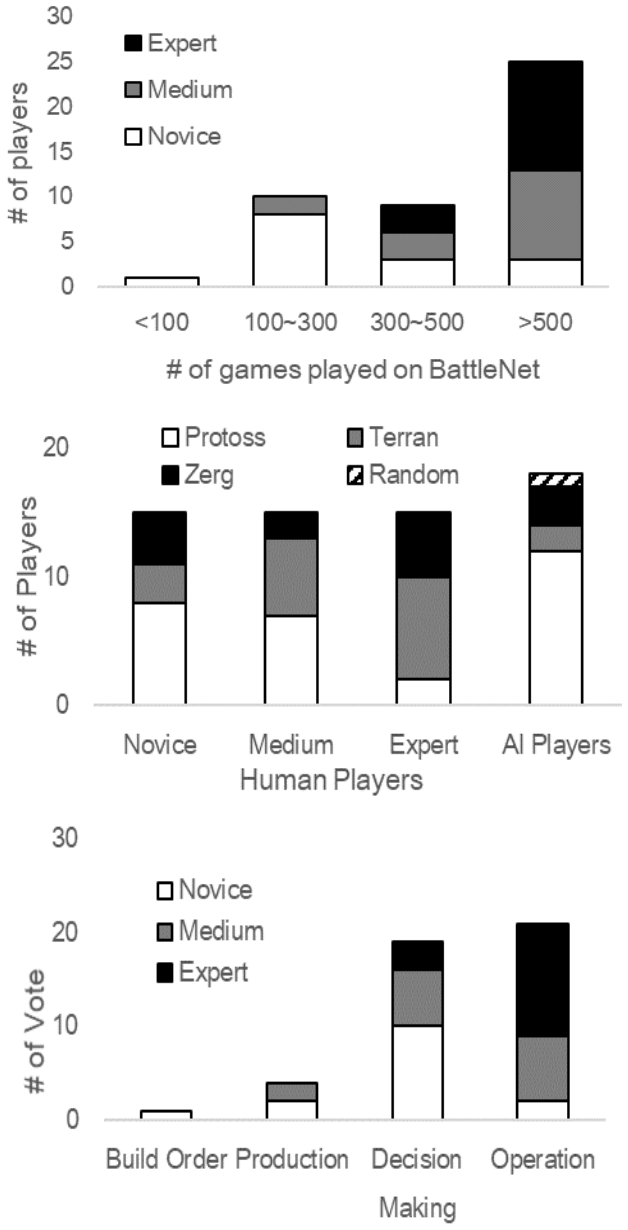


FIGURE 2. Pre questionnaire about # of games played so far, race distribution, and opinion on the most important factor to win StarCraft games.

attempted to assess the player's expertise, purposely fostering special game situations and evaluating the player's response. The expert player told us that a single match was sufficient for assessing expertise into one of our three groups.

Of our 45 participants, 29 had their expertise assessed through a test match against the expert player. 12 players (two novice, 2 medium, and 8 experts) were categorized based on their official win/loss records from BattleNet. 3 players had an official semi-professional license and 1 player was a trainee on an e-sports team. Figure 2 shows the basic statistics of the human players. It shows that most members of the novice group played between 100-300 games, while most members of the medium and expert groups played over 500 games on BattleNet.

Regarding race distribution, Protoss is popular amongst novice and medium players, but Terran and Zerg are dominant in the expert group. AI Players' race distribution is similar to the novices with Protoss as the most popular. Participants were asked to select the most important factor for winning a StarCraft game from the six measures (HL, DM, PD, OP, BO, and MC).

The result shows that participants felt that decision making and Operation are the most important things to consider to win. Novices put more emphasis on decision making, but medium expertise players were divided between decision making and Operation. Interestingly, expert players stress the importance of Operation (∼80%).

V. RESULTS AND ANALYSIS

Here we present the results of our study of human assessment of AI StarCraft players.

A. DIFFERENCE BETWEEN AI AND HUMAN ASSESSMENT OVER FIVE YEARS

After each match, the human player evaluates the AI they just played against using the post-game questionnaire. It asks human players to evaluate the AI's play using a 5-point Likert-scale from "very bad" to "very good" levels with the seven criteria (HL, DM, PD, OP, BO, MC, and PM) and also to write down their reasons behind the scores. Of the 270 matches, human players won 237 games (88%) and lost 33 games (12%). Novice players contributed to the majority of the losses with 25 lost games (28%). Medium players lost 7 games (8%), and expert players only lost 1 of their 90 matches. It is interesting that almost half of the human player losses came from the matches against the 2015 AI bots (45%).

Table 4 summarizes the average of scores by human players for each AI player. Interestingly, our initial hypothesis that humans judge the quality of AI players differently from what a simple win-loss ratio can represent is supported. The evaluation results show that the top AIs judged by human players do not exactly match the highest ranking AIs from our cross-year tournament (Table 2) (Spearman's rank correlation coefficient = 0.018). It is surprising that except for Tscmoo15, AIs with a high win ratio were placed in the middle (ZZZKBot15) and bottom (Overkill15) by human evaluators. In contrast, the Skynet series of bots (Skynet11, Skynet12, and Skynet13) were scored very highly by human players, while performing very poorly in our cross-year AI competition.

The average evaluation scores for the different criteria are in the range of 2.2-2.5 except for Micro Control (3.0) and Build Order (2.8) (Table 4). These results mean that current AI players are recognized by human players as having better higher performances Micro Control and Build Ordering capabilities compared to the other skills. Figure 3 shows the graphical relationship between the human players' evaluation

TABLE 4. Evaluation scores of seven criteria by human players (sorted by Performance).

	Average Scores from Human Players							
Name	Overall Performance (PM)	Human Likeness (HL)	Decision Making (DM)	Production (PD)	Operation (OP)	Build Order (BO)	Micro Control (MC)	Rank in Five-Year Tournament
Skynet13	**3.5**	3.3	**3.3**	3.7	3.3	3.4	**3.4**	11
Skynet12	3.3	**3.5**	2.8	**4.1**	3.3	3.3	2.9	12
Tscmoo15	3.2	3.3	3.0	3.8	3.2	3.3	3.0	**1**
Skynet11	3.1	3.3	2.9	3.5	**3.4**	3.2	2.6	-
IceBot14	2.9	3.1	2.9	2.8	2.8	2.9	2.8	8
UAlbertaBot15	2.7	2.4	2.3	3.8	2.3	3.2	2.3	4
Aiur13	2.7	2.7	2.3	3.5	2.7	2.5	2.3	13
ZZZKBot15	2.6	2.9	2.1	2.9	2.1	**3.6**	2.6	2
UAlbertaBot13	2.3	2.1	1.9	3.5	2.1	2.7	2.1	10
Aiur15	2.2	2.2	2.3	2.5	2.1	3.3	2.0	9
Aiur12	2.1	2.3	1.9	3.1	2.0	2.6	1.6	14
UAlbertaBot11	2.1	2.1	1.9	2.5	2.0	2.1	2.4	-
Ximp15	2.1	2.1	1.9	3.2	1.8	2.7	2.1	5
Aiur11	2.0	2.3	2.2	2.5	2.3	2.7	2.1	-
Ximp14	1.8	2.1	2.0	2.1	1.9	2.1	2.0	7
Overkill15	1.8	1.7	1.5	1.9	2.1	2.0	2.3	3
UAlbertaBot12	1.7	1.8	1.9	3.3	1.7	2.2	1.8	15
LetaBot14	1.7	2.0	1.3	2.0	1.5	2.4	2.3	6
Average	2.4	2.5	2.2	3.0	2.4	2.8	2.4	-

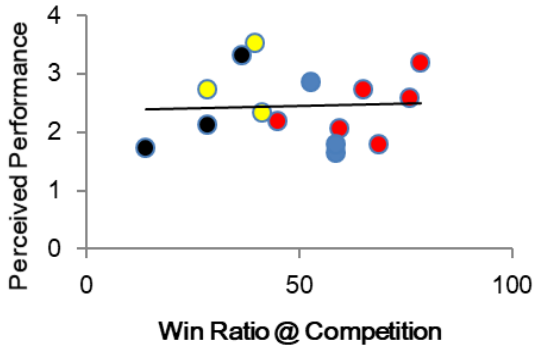


FIGURE 3. Win ratio versus perceived Performance by human players (Red = 2015, Blue = 2014, Yellow = 2013, Black = 2012).

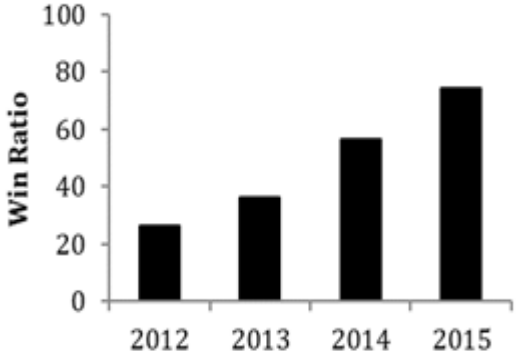


FIGURE 4. Win ratio progress over five years (For 2015, only the top three players' win ratio was averaged).

of overall performance and the actual win ratio in the AI vs. AI competition. It shows that there is no relationship or correlation between these two evaluation measures (Pearson correlation = 0.053). Human evaluation of the AI players seems to be independent of the win-loss ratio performance and thus is not a good proxy for human evaluation.

Table 5 shows a Pearson correlation analysis of the human player's evaluation for the different skills and win ratio from the AI vs. AI competition. It shows that the win ratio (WR) also has very little relationship with the human evaluation scores. The correlations are between −0.29 and 0.33. However, there are very high correlations among human evaluation of AI players in the categories of Human Likeness, Decision Making, Operation, and Performance. These are all scored between 0.89-0.94. This indicates that if human players rate one of the criteria very highly, other measures will also be rated very highly as well. The correlation values for these measures are so high that we could treat the measures as equivalent. Simply put, if the AI bots are rated as good (or bad) in one of the four measures (Decision Making, Operation, Human-Like, or Performance), then they will be evaluated as being similarly good (or bad) in the other 3 measures.

Figure 4 shows the progress of the win ratio over five years. It means that the AI players definitely have improved in their ability to win against other AI players. However,

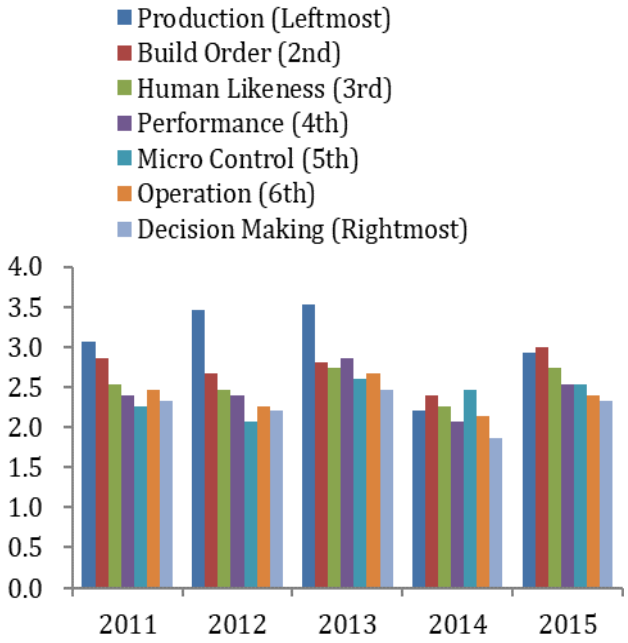


FIGURE 5. The progress of human evaluations over five years.

TABLE 5. Pearson correlation analysis among the seven evaluation criteria and win ratio (WR).

		Perceived						
	WR	HL	DM	PD	OP	BO	MC	PM
HL	**0.13**							
DM	**0.02**	**0.89**						
PD	**0.29**	0.62	0.66					
OP	**0.01**	**0.91**	**0.91**	0.62				
BO	**0.21**	0.77	0.72	0.58	0.60			
MC	0.33	0.80	0.73	0.39	0.76	0.58		
PM	**0.05**	**0.94**	**0.91**	0.75	**0.92**	0.77	0.80	

the human's evaluations on the five years of AI players show no progress on any of the criteria except for Micro Control with a slight increase (Figure 5). A simple linear regression analysis shows that the evaluation score changes by-0.15 for Micro Control, +0.09 for Micro Control, and −0.03 - +0.02 for other criteria each year. The human evaluation seems to be the lowest for the AIs from 2014. For Micro Control, the scores from 2011 to 2013 increase but drop significantly in 2014 before slightly increasing again in 2015. Unlike the win ratio increase, there is no evidence that the AIs become better over the years from a human player's viewpoint. In terms of overall performance, 2013 was the best year (average of 2.9) and 2014 was the worst (average of 2.1).

B. TOWARDS TAILORED AI PLAYERS BASED ON HUMAN PLAYERS' EXPERTISE

In our study, we grouped the human players into three categories based on their expertise. Although the initial analysis was averaged over all human players, it is important to understand evaluation differences in players with different

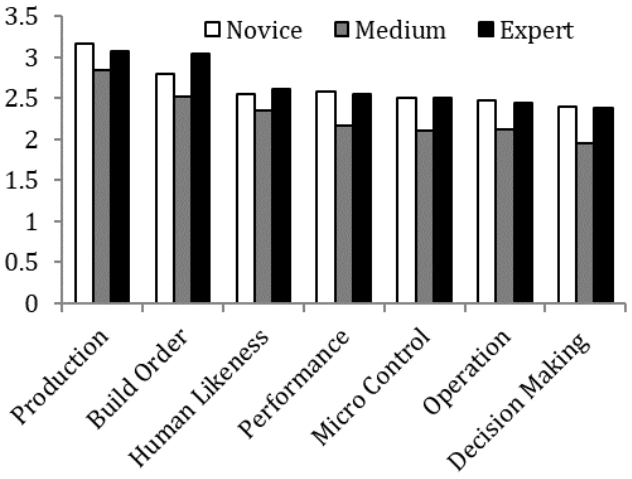


FIGURE 6. Comparison of evaluation scores for novice, medium, and expert players.

expertise levels. The number of games lost by novice, medium, and expert players are 25, 7, and 1, respectively. It means that for novice players, the AIs were not necessarily easy opponents and sometimes able to beat the novices. For medium and expert players, the AIs were not a significant threat.

Figure 6 summarizes the average evaluation scores for the three human player expertise groups. It is interesting that the medium expertise players evaluate the AIs a bit lower than the other two groups (average score over all seven criteria = 2.64, 2.29, and 2.65 for the novice, medium, and expert groups). The fact that this difference in rating exists is very important for AI developers to consider when building their AIs. Concretely, medium expertise players will be less accepting of the StarCraft AI players that we tested compared to the novice and expert players.

The novice players usually had less experiences with StarCraft and they tended to evaluate AI players more favorably if the AI player played better than they did. For expert players, who have much knowledge and experience with the game and rarely lose, they tended to evaluate the weak players without overstatement. However, they also tended to focus on the potential of opponents instead of the results of the single match played against the AI. Today's AIs are obviously weaker than expert players. Expert players were more generous in their evaluations. Medium expertise players usually have knowledge of StarCraft but are less experienced than the expert players, and therefore were not as generous in their evaluations.

It is interesting to see that the expert players evaluate the AI player's Build Order (early stage strategy) as highly as the production skill. So far, the prevailing thoughts in the general StarCraft AI community are that the AIs were only strong in the Micro Control measures (and not Build Order), due mostly to their very high Actions per Minute (APM) metric [18]. It is true that the AIs can maintain a very high

TABLE 6. Pearson correlation analysis on the evaluations from different expertise groups.

	Novice *vs.* Medium	Novice *vs.* Expert	Medium, *vs.* Expert
HL	0.47	0.57	**0.70**
DM	**0.24**	**0.30**	**0.71**
PD	0.58	0.54	0.55
OP	0.50	0.51	**0.75**
BO	**0.19**	0.46	0.38
MC	**0.21**	0.39	0.32
PM	0.48	0.51	**0.67**
Average	0.38	0.46	0.58

APM compared to human players (usually more than 10 times greater). Although it is not always true, this higher APM can be useful for enhancing Micro Control. Therefore, it is not surprising that the AIs received a good evaluation on Micro Control, matching the current thoughts of the StarCraft AI community.

However, the fact that the Build Order is highly evaluated by expert players, in particular, helps us see the progress that AI players have made over time in their early stage strategies. In the early years of the competitions (2011∼2012), the AIs usually only had a single build order without scouting behaviors. However, recent AIs have multiple strategies and adapt theirs based on the scouting information. This new approach is a definite improvement over the earlier AIs, even though the numeric evaluation of progress (Figure 5) does not reveal this impact.

A correlation analysis shows that the medium and expert player evaluations have high similarity with each other (Table 6). It means that although the medium players evaluate the scores as being lower than the other two groups, their evaluation patterns are similar to the expert level players. For HL, DM, and OP, the medium and expert groups show a correlation of ≥ 0.7. Although the medium and expert groups are well correlated, the Build Order (BO) and Micro Control (MC) is less well correlated. It means that the two groups have different viewpoints on these two skills.

After the human players finished their six assigned matches, they were asked to evaluate the current quality of AI players on a scale between 1 (novice player) and 5 (professional player) (Figure 7). Most of the human players (25 players, 55%) thought that the current AIs could only be rated as 2 out of 5. Six novice players rated the AIs as level three and only one expert player rated the AI players as level four. One expert commented:

"It would be very nice if the AI slightly adjusted its strategy. Production seems to be perfect. In general, It's good except for occasional bad unit control and losing focus in the middle of the game.

TABLE 7. Example of categories for text answers.

Evaluation Criteria	Categories for human players' text answers on the reason of their evaluation
HL	1) Similar strategy used by humans 2) Similar attack timing used by humans 3) Similar Decision Making with humans 4) Lower Decision Making than humans 5) Better performance than previous AIs 6) Similar performance to previous AIs 7) Very old strategies not used by humans 8) Different Micro Control skills to humans 9) Different building placement (*e.g.*, grid style) 10) *Etc.*

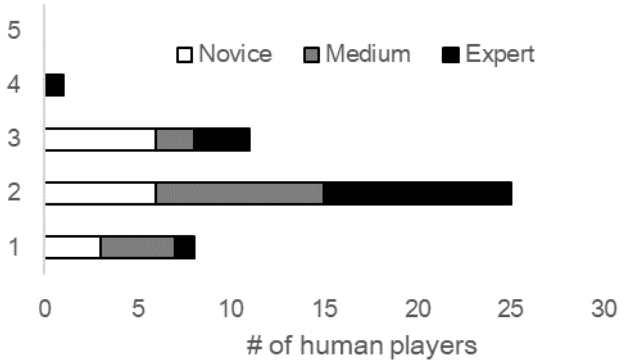


FIGURE 7. Level of current StarCraft AIs perceived by human players. (Level five means professional player).

C. UNDERSTANDING HUMAN PLAYERS NEEDS FROM SHORT TEXT RESPONSES

For each match, human players recorded numeric scores and the reason for their evaluation in free-text. To understand the textual feedback, we recruited three human coders who had official semi-professional player licenses. They reviewed all the comments from the 45 human players (1890 answers = 7 criteria $\times$270 matches). In the coding, each coder reviewed the text response and assigned one or more categories suitable for it. The categories were defined by the expert players after a review of the comments (Table 7).

Each coder reviewed all 1890 text comments and the coding results were aggregated amongst the coders. For example, the coders categorize the user's response on performance evaluation into 10 categories (Figure 8). Six of them are positive reasons while three of them are negative and one is neutral. Human players had 633 comments that were categorized into the positive reasons, but 870 categorized as negative reasons (Figure 7). It means that human players tend to negatively evaluate AI players. The main reason for the negative feedback is a lack of Decision Making capability amongst the AI players. For StarCraft AI developers,

TABLE 8. Summary of the popular comments for each evaluation criteria.

	Popular Comments
HL	1) Worse Decision Making than human players 2) Unit control style is different
DM	1) Wrong combat decisions on location and units 2) Building construction and unit production unsuitable for current situation
PD	1) Producing units without stopping (or break) 2) Constructing an adequate number of buildings for production
OP	1) Bad understanding of situations 2) Using extreme early strategies without thinking about the future (later part of the game)
BO	1) Using unexpected strategies for human players 2) Using customized strategies for opponents
MC	1) Attacking weak opponents (low health) first 2) Attack targeting (who to attack first) is often wrong

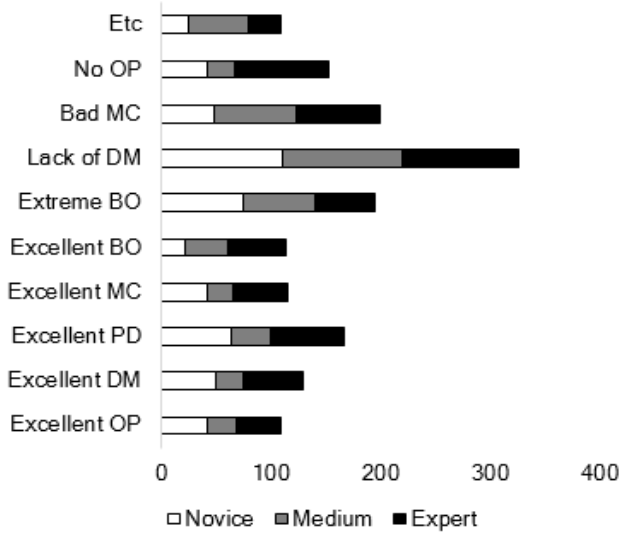


FIGURE 8. Categorization of user responses for Performance (PM).

implementation of Decision Making is one of the most challenging issues because it is based on having extensive experiences. There are some research works attempting to make progress on AI Decision Making [17]. It is interesting that the main complaints from the human players are focused on three issues: Decision Making, Micro Control, and Operation. Table 8 summarizes the top two categories for each criterion.

As you can see in Figure 8, the lack of DM is the biggest problem of the AI, according to the evaluations by humans. To assist solving this problem, we analyzed the DM evaluation data with more depth.

Figure 9 shows the distribution of DM scores given to AI players by human players. We analyzed the comments by human players who gave 1 or 2 points using human coding. Through this analysis, we were able to identify three problems with the lack of DM (Table 9).

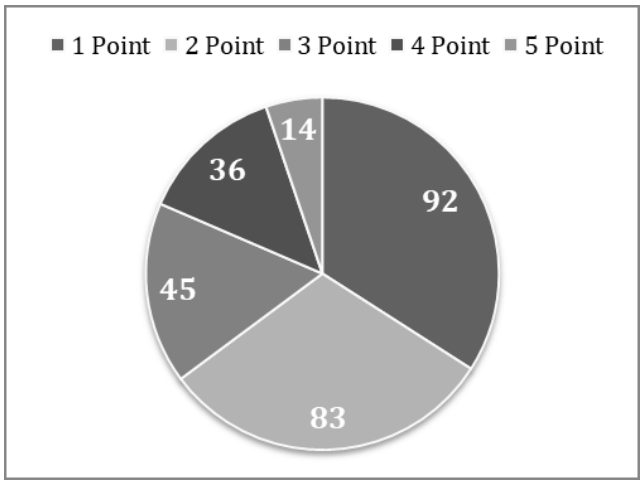


FIGURE 9. Distribution of evaluation scores for user's Decision Making (DM).

TABLE 9. Popular and detailed answers on the lack of DM.

	Most answers regarding the lack of DM
1st	Produce units that are not optimal against to the opponent unit.
2nd	Fight in an unfavorable battlefield.
3rd	Give up other worthy elements to avoid losing the battle.

According to most expert human players, AIs made their units adhere to their original strategy even after conducting reconnaissance of their opponents. Because of the nature of RTS games, it is crucial to be as well-suited to your opponent as you are to your own army and to remain flexible with your strategy. However, the AIs did not produce well-suited units to their opponent units. As a result, AIs did not maintain their advantageous situation. In addition, they ignored favorable terrain for their units and fought on terrain favorable to their opponents. This means that AIs do not consider combat terrain particularly well. Finally, they abandon crucial factors to avoid small damage; for example, if damage is likely to be incurred when occupying a strategic point, most AIs avoid such battles because they prefer to avoid taking damage even though the occupation of this place is of great help to victory. This occurs because AIs judge an engagement only in terms of losses in a single confrontation between units. Most games get closer to victory if you minimize the loss of units. However, in a RTS game, various elements such as resource harvesting, unit production, upgrades, and terrain also affect victory, so sometimes units can be sacrificed for preserving other elements of greater influence. Minimizing loss is not the only approach that should be pursued by AI players in RTS games. In fact, in the popular game Go, players deliberately

use their stones as bait to catch the opponent; they lose a stone, but the result is a bigger reward. Expert human players predicted that if AI could solve these problems and engage in a similar approach as Go players, AI performance would be greatly improved. In addition, this performance improvement will not only improve the performance of the DM, but will also affect other evaluations positively.

VI. CONCLUSIONS

In the game AI community, the objective performance measures (scores, lap time, and win ratio) have been widely used to indicate the quality of an AI. Although an objective measure is clearly defined and easy to implement to allow automated evaluations, it can lead to the development of undesirable AI players for human players.

In this study, we analyzed the responses of 45 human players who played against the 18 top StarCraft AI players from the last five years. The StarCraft AI Competition, which is based on a win ratio against other AIs, has shown clear progress over the years with an increasing win ratio. However, our analysis showed that a higher win ratio is not correlated with better human evaluations. The best AI players judged by human players came from 2011∼2013 period and those that did not have high win ratios, indicate that the current win ratio being used to assess AI players cannot be used as a proxy for human assessment. Human players value a different set of criteria than just whether the AI player plays well enough to win or not. A detailed analysis of the seven evaluation criteria by human players shows that there are neutral or negative relationships between human evaluation and the win ratio of AI players.

Also, while AI players have clearly progressed in terms of ability to beat other AI players from previous years, an assessment by human players does not reveal this change. In fact, the majority of the most recent AI players are ranked quite low by human players. From these two observations on the win ratio assessment and AI progress, we recommend that AI gaming competitions include a human assessment component in their annual rankings of AI players to guide the development of AI players with better means. The AI players that humans prefer should be ranked more highly. We also recommend that AI developers in game companies use human evaluation for improving their AI players outside competition settings.

Additionally, we also identified fairly distinct differences in the assessment of AI players by human players in various expertise levels. Expertise was an important factor in the scoring of the AI players. For example, the medium and expert group players showed similar evaluation patterns on Operations but not on Build Order or Micro Control. The novice players deviated from the medium and expert players in scoring the AI players. Based on these results, we recommend that AI designers consider the expertise of the human players when deciding which AI players to deploy and what characteristics/skills those AI players should have. The text responses from human players also point to a number of key issues that AI designers can use to create more human-preferred AIs, including attack and building strategy choices, Decision Making, and adaptivity to opponent choices.

Solving the problems, they pointed out can be an effective way to improve AI performance. If AI designers can tackle the problem of a lack of DM in particular, we can expect big performance improvements. Although recruiting human players can be difficult, the advice of experienced and expert players can help tackle the lack of expertise of AIs currently available. Moreover, the number of complex cases in the RTS game sector is far too numerous to handle with only a number of conventional cases. We believe that using talented expertise to reduce this efficiently will help develop excellent RTS game AIs for all game genres.

REFERENCES

[1] G. Cheung and J. Huang, "StarCraft from the stands: Understanding the game spectator," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2011, pp. 763–772.

[2] I. Millington and J. Funge, *Artificial Intelligence for Games*. Boca Raton, FL, USA: CRC Press, 2009.

[3] M. O. Riedl and A. Zook, "AI for game production," in *Proc. IEEE Conf. Comput. Intell. Games*, Aug. 2013, pp. 1–8.

[4] M. Treanor *et al.*, "AI-based game design patterns," in *Proc. 10th Int. Conf. Found. Digit. Games*, 2015. [Online]. Available: http://www.fdg2015.org/papers/fdg2015_paper_23.pdf

[5] S. Ontanon, G. Synnaeve, A. Uriarte, F. Richoux, D. Churchill, and M. Preuss, "A survey of real-time strategy game AI research and competition in StarCraft," *IEEE Trans. Comput. Intell. AI in Games*, vol. 5, no. 4, pp. 293–311, Dec. 2013.

[6] J. Togelius, "How to run a successful game-based AI competition," *IEEE Trans. Comput. Intell. AI in Games*, vol. 8, no. 1, pp. 95–100, Mar. 2016.

[7] P. Hingston, "A Turing test for computer game bots," *IEEE Trans. Comput. Intell. AI in Games*, vol. 1, no. 3, pp. 169–186, Sep. 2009.

[8] D. Churchill. *A History of StarCraft AI Competitions*. Accessed: Dec. 6, 2017. [Online]. Available: http://www.cs.mun.ca/~dchurchill/starcraftaicomp/history.shtml

[9] J. Togelius, N. Shaker, S. Karakovskiy, and G. N. Yannakakis, "The Mario AI championship 2009–2012," *AI Mag.*, vol. 34, no. 3, pp. 89–92, 2013.

[10] N. Shaker *et al.*, "The 2010 Mario AI championship: Level generation track," *IEEE Trans. Comput. Intell. AI in Games*, vol. 3, no. 4, pp. 332–347, Dec. 2011.

[11] D. Loiacono *et al.*, "The 2009 simulated car racing championship," *IEEE Trans. Comput. Intell. AI in Games*, vol. 2, no. 2, pp. 131–147, Jun. 2010.

[12] J. Renz, X. Ge, S. Gould, and P. Zhang, "The angry birds AI competition," *AI Mag.*, vol. 36, no. 2, pp. 85–87, 2015.

[13] D. Perez-Liebana, S. Samothrakis, J. Togelius, S. Lucas, and T. Schaul, "General video game AI: Competition, challenges and opportunities," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 4335–4337.

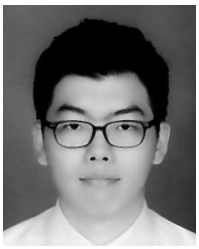
[14] D. Churchill. (2015). *AIIDE StarCraft AI Competition Report*. Accessed: Dec. 6, 2017. [Online]. Available: http://www.cs.mun.ca/~dchurchill/starcraftaicomp/report2015.shtml

[15] B. G. Weber, M. Mateas, and A. Jhala, "Building human-level AI for real-time strategy games," in *Proc. AAAI Fall Symp. Adv. Cognit. Syst.*, 2011, pp. 329–336.

[16] *StarCraft AI Competition—Data Archive*. Accessed: Dec. 6, 2017. [Online]. Available: http://www.cs.mun.ca/~dchurchill/starcraftaicomp/archive.shtml

[17] G. Robertson and I. Watson, "A review of real-time strategy game AI," *AI Mag.*, vol. 35, no. 4, pp. 75–104, 2014.

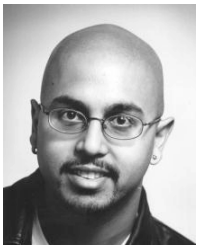
[18] B. Weber, "APM is not everything in StarCraft," *GAMASUTRA Blog*, 2011, accessed: Dec. 6, 2017. [Online]. Available: https://www.gamasutra.com/blogs/BenWeber/20110505/89447/APM_is_not_everything_in_StarCraft.php



MAN-JE KIM received the B.S. degree in computer science from Sejong University, Seoul, South Korea. He is currently pursuing the master's degree with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology. His research interests include artificial intelligence, video game, HCI, human-like AI, reinforcement learning, algorithms for game AI. He is developing artificial intelligence for real-time games every year. He was a co-organizer of the IEEE CIG 2015 Starcraft AI Competition. He received the third place in the CIG Fighting Game AI Competition in 2015 and 2017, respectively.

KYUNG-JOONG KIM received the B.S., M.S., and Ph.D. degrees in computer science from Yonsei University in 2000, 2002, and 2007, respectively. He was a Post-Doctoral Researcher with the Department of Mechanical and Aerospace Engineering, Cornell University, in 2007. He is currently an Associate Professor with the Department of Computer Science and Engineering, Sejong University. His research interests include artificial intelligence, game, and robotics.

SEUNGJUN KIM received the B.S. degree in electrical and electronics engineering from the Korea Advanced Institute of Science and Technology, and the M.S. and Ph.D. degrees in mechatronics from the Gwangju Institute of Science and Technology (GIST), South Korea, in 2006. He is currently an Assistant Professor with the Institute of Integrated Technology, GIST, and an Adjunct Faculty Member with the Human-Computer Interaction Institute, Carnegie Mellon University. He currently leads research and development projects concerning human–vehicle interaction, wearable UI/UX technologies, human–robot interaction, sensory augmentation with haptics and augmented reality, and cyber-learning with a sensor support. His research interests are at the intersection of human–computer interaction (HCI) and sensor data mining to create intelligent systems that improve the quality of HCI experience based on human attention and cognition.



ANIND K. DEY received the B.S. degree in computer engineering from Simon Fraser University, Burnbay, BC, Canada, and the M.S. degree in aerospace engineering and the Ph.D. degree in computer science from Georgia Tech, Atlanta, GA, USA. He is currently a Professor with the Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA. His research interests include feedback/intelligibility and control in ubiquitous computing, context-aware computing, toolkits and end-user programming environments, sensor-rich environments, information overload, ambient displays, privacy, human–computer interaction, and machine learning. His works have been published in a top-tier conference at ACM SIGCHI, and his work was nominated as the best paper for six times.