# Discovery of Player Strategies in a Serious Game

**Hua Li[1], Hector Munoz-Avila[2], Lei Ke[1], Carl W. Symborski[1], Rafael Alonso[1]**
[1]SAIC; 4001 North Fairfax Drive, 4th Floor, Arlington, VA 22203
hua.li@saic.com, | lei.ke@saic.com | carl.w.symborski@saic.com | rafael.alonso@saic.com
[2]Department of Computer Science & Engineering; Lehigh University; Bethlehem, PA 18015
munoz@cse.lehigh.edu

### Abstract

Serious games are popular computer games that frequently simulate real-world events or processes designed for the purpose of solving a problem. Although they are often entertaining, their main purpose is to train or educate users. Not surprisingly, users exhibit different game play behaviors because of their diverse background and game experience. To improve the educational effectiveness of these games, it is important to understand and learn from the interaction between the users and the game engine. This paper presents a study attempting to apply machine learning techniques to the game log to discover: a) strategies that are common to players interacting with serious games and b) variances in the demographics of the player base for these strategies. This is an empirical study with end-user data while playing *Missing*, a serious game developed to help mitigate biases that people may exhibit when analyzing plausible hypothesis for observed events. We found a set of common strategies and interesting variances in player demographics associated with these strategies.

## Introduction

Serious games are popular games genre that are used for a purpose other than just for the sake of entertainment (Michael and Chan, 2005). Frequently, serious games simulate real-world events or processes designed for the purpose of gaining experience in tasks such as problem solving (Susi and Johannesson, 2007). As such, serious games frequently have a pedagogical theme or context and a focused application domain such as military recruiting, education and health awareness tasks (Susi and Johannesson, 2007). There has been a large volume of machine learning research on serious games including learning of effective game play strategies by using feedback loops (Tesauro, 1995), building opponent's models (Davidson et al., 2000), learning patterns of social interaction and dialogue in a restaurant setting (Orkin and Roy, 2010), and even classifying the spectators of a game (Cheung and Huan, 2011). In particular, Tesauro (1995) was focusing on learning to play well. Surprisingly, little attention has been paid to the problem of discovering common strategies that human players follow when playing computer games. In this work we address this problem by investigating the following two questions:

- Do players exhibit common strategies while playing a game?
- Are there any variances in the players' demographics for the different strategies?

We are particularly interested in investigating these questions in the context of Serious Games. There are many applications of these games including using them to teach the player a skill or make the player aware of certain personal biases. It is in this context, that our research takes place. We used the game "*Missing:* The Pursuit of Terry Hughes", which is a serious computer game developed by an SAIC-lead research team with the aim of helping the players to recognize and mitigate common cognitive biases that they may have when investigating an event[1]. In this role-playing game, Terry Hughes is missing and the player is immersed in a 3D environment on a journey to find Terry. The player will examine objects, meet and question people and overcome challenges as we pick up Terry's trail. Along the way the player is exposed to specific bias invoking situations where cognitive bias exhibited by the player are measured during two or three sequential game

[1] The team consists of researchers from SAIC, Carnegie Mellon University (CMU) and Creative Technologies Incorporated (CTI). This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory.

episodes. After each episode is played there is an after-action review that teaches about specific biases, offers feedback on game performance, and reinforces the point with a story.

Players in *Missing* can examine different objects, use smart phones to take pictures and communicate with non-player characters (NPCs), and navigate in the scenario, among other choices. As a result, this is a good setting to investigate if there are common strategies pursued by the players.

- It may reveal flaws in the game and enable the game developers to make corrections. For example, the game may prompt a substantial number of players to find the easy way out instead of exploring aspects of the game that might be more time consuming but more beneficial for the instructional goal.
- Understanding the kinds of strategies employed is particularly important in the context of serious games as it may help devise the training means to achieve the instructional goals.
- Variance in the demographics of the players for the different strategies may reveal the need for additional instruction to specific demographic groups.

One technical challenge is how to identify these strategies and demographic variances. In serious games, a gamer may play for a long session where he/she might take hundreds of gaming actions. There can be dozens or even hundreds of players and hence a manual analysis of the game log for each session is simply not feasible.

In this paper we present the results of a study of *Missing* in which we identify common strategies pursued by game players and discover variances in player demographic information for different strategies. We believe that this is the first work of its kind.

## Related Work

Cheung and Huang (2011) collected hundreds of stories from spectators watching Blizzard Entertainment's Starcraft games (Huhh, 2008). Starcraft is a real-time strategy game aiming to provide entertainment for its players. Cheung and Huang manually annotated the data collected and ran clustering algorithms on the annotated data. From this, they extracted categories of spectators. In our work, we automatically collect actual game traces from a serious game and identify common behavior from the players.

Many works exist that use game traces as input. Bares *et al*. (1994) used traces to make predictions about the outcomes of the game. They established some of the usual practices of the game analysis field: using traces as inputs and making comparisons between traces to make predictions. Cox and Kerkez (2006) assumed that action semantics (preconditions and effects) are given as input (e.g., defined manually by a human engineer). In addition, they assumed that the traces are annotated with the intermediate states observed. This enables their system to perform plan recognition tasks. They use state abstraction techniques to improve the plan recognition. Some works learn from the traces to develop counter-strategies in adversarial games. Sharma *et al*. (2007) concentrated on transfer learning tasks by extracting hierarchies from the traces. Auslander *et al*. (2008) learned counter-strategies by using reinforcement learning techniques. Guillespie *et al*. (2010) learned counter-strategies from observations by mimicking the observed traces. In our work, we use the traces to identify common strategies performed by players. Other important differences between our work and these works are: (1) in the games mentioned before, game traces are extracted from zero-sum games but *Missing*, like many serious games, is not a zero-sum game; it aims to instruct players about their own biases. (2) The games used by previous researchers are adversarial games. *Missing* is not an adversarial game; there is no win/lose scenario within *Missing*. (3) Many of the other games used are classified as recreational games whereas *Missing* is classified as a serious game. Finally, (4) unlike other research, we also explored demographic variances in the data collected from players for the different common strategies discovered.

## Game Traces in *Missing*

*Missing* is a serious game designed with the purpose of teaching players to become aware of their own cognitive biases and learn ways to combat these biases (Evans, 2007; Forster and Liberman, 2007). *Missing* is divided into two or three episodes tailored towards teaching the player about common cognitive biases including confirmation bias, fundamental attribution error, and blind spot bias. Each episode consists of a number of segments or task puzzles (bias vignettes) for the player to complete in sequence. The average length of an episode is 30 minutes.

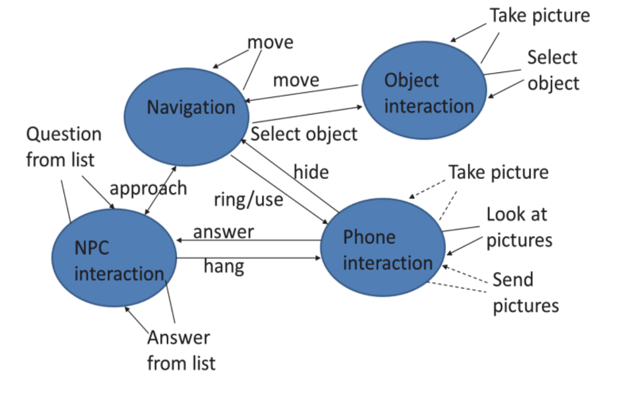


*Figure 1. Player interactions with the game engine*

Figure 1 shows a finite state machine for the interactions that players can perform with the game engine. The circles denote the states in which the player-controlled character, or avatar, can be in. The avatar can find itself in one of four states: navigation, object interaction, NPC interaction, and phone interaction. The arrows indicate the transitions from one state to another. For example, when the avatar is in the *navigation state* it can select an object, in which case the avatar moves to the *object interaction* state. For the sake of clarity some transitions are omitted in the figure.

*Table 1. Example Event in the Game Log*

```
<Event
  type="Message_ScreenActivated"
  realTimeElapsed="109.56"
  gameTimeElapsed="109.56">
  <Param name="m_screen"
    value="Inspect Object Screen(Clone)
          (InspectObjectScreen)"/>
  <Param name="m_stackID" value="Game"/>
</Event>
```

A trace is a sequence of steps or events made by the player. In *Missing*, these traces are recorded as XML events. These events record the actions taken by the avatar, such as going to a location or inspecting an object. We filter the trace and extract only a subset of the events. This can be seen as a generalization because irrelevant events are removed from consideration. Table 1 shows an example of an object inspected event in XML format from the game log.

## Strategy Extraction Architecture

To extract common strategies among players we look for clusters of traces. Traces within the same cluster are similar and, hence, reflect a common strategy. From these clusters we extract features that reflect the distinguishable characteristics within each cluster. We use these features to describe the common strategies followed in each cluster.

The overall architecture for feature extraction is shown in Figure 2. There are three phases:

(1) The feature extractor selects a collection of features π from the game trace. Each π is stored in a feature trace library Π.
(2) The feature trace library Π is partitioned into clusters.
(3) From these clusters, strategies are extracted.

The following discussion addresses each of these three phases in detail.

### Feature Extractor

We apply ideas from case abstraction whereby details are excluded and more general features are collected (e.g., see Bergmann and Wilke (1995) for an example of abstraction from plan traces). Specifically, the Feature Extractor receives a play trace log as input and automatically outputs the relevant features, i.e., low level events (LLEs) for the trace and their categorization into high level events (HLEs). The *Missing* game design supports a fixed set of LLEs, which we manually group into 7 semantic categories, or HLEs. The principle for grouping is as follows. For a particular HLE, the component LLEs are semantically consistent with the nature of the HLE. In other words, each LLE can be considered as an indicator of the corresponding HLE. For example, HLE "Photo" contains two LLEs: Message_PhotoAdded and Message_ScreenActivated-SmartPhonePhotoAlbumScreen. HLE "Movement" consists of four LLEs:

a) *MessageRequestToMoveToPoint*,
b) *MessagePathAttemptResult*;
c) *MessageMoveToPointFinished*, and
d) *Message_ScreenActivated-MapScreen*.

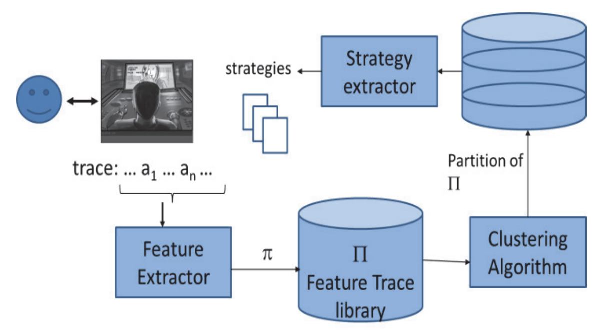


*Figure 2. Strategy extraction architecture*

The seven HLE categories are:

1. **Phone**. LLEs corresponding to using the cell phone are grouped in this category.
2. **Photo**. LLEs corresponding to taking picture, managing the photo album, viewing photos and the like are grouped in this category.
3. **Movement.** LLEs corresponding to moving including using the map for navigation purposes are grouped in this category.
4. **Inspection.** LLEs corresponding to inspecting objects, rotating them, and the like are grouped in this category.
5. **Instant messaging.** LLEs corresponding to sending and receiving messages are grouped in this category.
6. **Communication.** LLEs corresponding to face-to-face communications between the NPCs and the avatar are grouped in this category.

7. **Pester.** LLEs corresponding to communication from an NPC called Chris to the avatar are grouped in this category. The longer it takes for the avatar to complete the episode, the more messages it will get from the NPC Chris.

There are three episodes and we collect HLE data for each episode except for the *Pester* HLE because it only occurs in Episode 1. So the total number of HLE features is 6×3 + 1 = 19.

The component LLEs for a particular HLE are not independent. In some cases, they are sequential actions of the same HLE. In other cases, they may be at a different level of abstraction for the same HLE. As a result, a simple sum of the component LLE counts will overestimate the HLE count. To address this, we use the average count.

The formula for computing the average presumes an HLE consists of m types of component LLEs (cLLEs); we define the average HLE count as:

*Average of all cLLE counts = sum(cLLE) / m*

## Clustering Algorithm

To partition the feature trace library Π into clusters, a clustering algorithm was applied to the HLE events extracted from the game traces and outputs a set of clusters. The clustering algorithm uses the similarity metric $\text{sim}_{\text{trace}}(\bar{\pi}, \pi')$, between a game trace $\bar{\pi}$ and a game trace π', is defined as follows:

$$\text{sim}_{\text{trace}}(\bar{\pi}, \pi') = \sum_{i=1}^{n} w_i * sim_i(\bar{\pi}, \pi')$$

where n is the number of HLE events, $w_i$ is the weight of each HLE event obtained, and $\text{sim}_i(\bar{\pi}, \pi')$ is a local similarity metric for the HLE event *i*, which is computed as the linear interpolation:

$$1 - (|x - y|/(\max - \min))$$

where max and min are the maximum and minimum possible values for the HLE event and x and y are the HLE event $i$'s values for $\bar{\pi}$ and π'.

Since we didn't know the number of clusters to partition the library a priori, we ran the X-means clustering algorithm (Pelleg and Moore, 2000), which extends k-means (MacQueen, 1967) by estimating the number of clusters to partition the observations.

## Strategy Extractor

We annotated each event trace with the cluster ID to which it belonged. Then we ran the correlation-based feature subset selection algorithm (Hall, 1998) on the feature trace library to determine the HLE events that correlate to the cluster membership. We chose this algorithm because it balances between the accuracy of prediction of every HLE event with respect to the clusters and the redundancy of other HLE events thereby finding a small subset of the HLE events that are correlated with each cluster.[2]

# Results

*Missing* was played in three rounds, or cycles, of experiments to investigate the factors that affect its efficacy in reducing cognitive biases. For this paper, only data from the first cycle were used. As part of the experimental protocol, demographic information for each subject was collected. The subjects were recruited in Pittsburgh, PA from Craigslist.com, an online classified advertisements website, and Carnegie Mellon University's Research Participant Pool website. The average age of the subjects was 28.4 years with a standard deviation of 11.2. At least 96.1% had some experience playing video games and the majority used a desktop or laptop computer on a daily basis. A total of 284 subjects participated in one of three experimental conditions: 1) Game, where Missing was played with game trace logged; 2) Video, where a video about cognitive bias was watched; and 3) Control, where a nature documentary on the oceans of the world was watched. Obviously, our analysis can only use the subjects from the Game condition.

We had 166 subjects who successfully played the *Missing* game. Eighty of these subjects played a long version of the game with 3 episodes whereas the rest played a short version with 2 episodes. For this initial study, we focused on the subjects with 3 episodes. Of the 80 game traces, 13 were incomplete or flawed due to either a system problem or a user error. The remaining 67 traces had complete data and were used in our analysis of strategy. Note that despite the sample size being relatively small, the risk of overfitting was low because the number of clusters identified was quite small (i.e. 3 clusters) (Hamerly and Elkan, 2003).

## Discovered Strategies

Out of the 67 participants that completed the three episodes, the clustering algorithm identified three clusters as shown in Table 2.

The first cluster contains 22 traces, the second contains 21 traces and the third contains 24 traces. The feature correlation algorithm identified 4 out of the 19 HLE features as relevant to distinguish between the clusters:

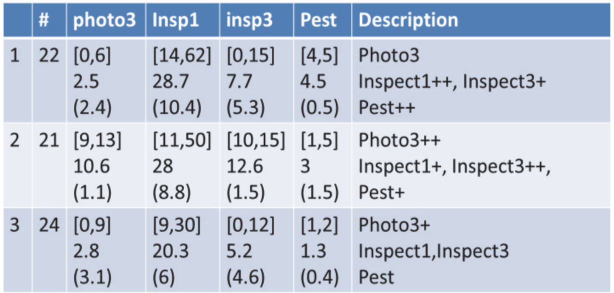
- Photos taken in episode 3
- Inspections done in episode 1

[2] The algorithm does not necessarily find the smallest subset of HLE events.

- Inspections done in episode 3
- Pester

*Table 2. Identified Clusters Representing Strategies*

*(Notations are explained in the body of the paper.)*



| | # | photo3 | Insp1 | insp3 | Pest | Description |
|---|---|---|---|---|---|---|
| 1 | 22 | [0,6]<br>2.5<br>(2.4) | [14,62]<br>28.7<br>(10.4) | [0,15]<br>7.7<br>(5.3) | [4,5]<br>4.5<br>(0.5) | Photo3<br>Inspect1++, Inspect3+<br>Pest++ |
| 2 | 21 | [9,13]<br>10.6<br>(1.1) | [11,50]<br>28<br>(8.8) | [10,15]<br>12.6<br>(1.5) | [1,5]<br>3<br>(1.5) | Photo3++<br>Inspect1+, Inspect3++,<br>Pest+ |
| 3 | 24 | [0,9]<br>2.8<br>(3.1) | [9,30]<br>20.3<br>(6) | [0,12]<br>5.2<br>(4.6) | [1,2]<br>1.3<br>(0.4) | Photo3+<br>Inspect1,Inspect3<br>Pest |

Table 2 shows for each cluster and each of these HLE events, the range (shown as *[lower bound, upper bound]*), the mean (shown as a number) and the standard error (shown in parenthesis *(standard error)*). The "Description" column shows an informal convention of the values: The suffix ++ denotes that the HLE event has highest values on average for that cluster compared to the values of that feature for other two clusters (such as *Photo3++* in Cluster 2). The suffix + denotes an HLE event having intermediate values for that cluster compared to the values of that HLE event for other two clusters (such as *inspect1+* in Cluster 2). No suffix indicates an HLE event having the lowest value for that HLE compared to the values of that HLE event for other two clusters (such as *Pest* in Cluster 3).

The following is a description of the strategies resulting from these analyses:

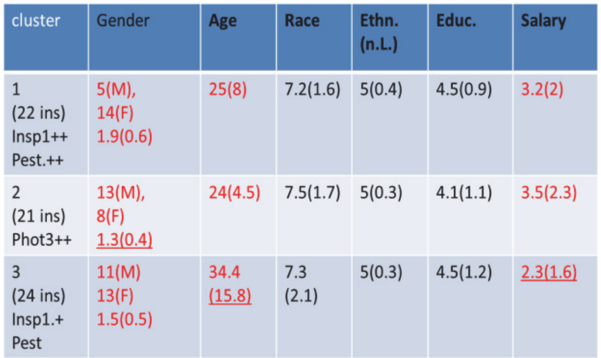
- Players in Cluster 1 tended to be more inquisitive in Episode 1. They inspected the most objects in that episode and took more time to search for clues. This is demonstrated by the fact that they were pestered by the NPC "Chris" more than the other players.
- Players in Cluster 2 also tended to inspect a lot of objects in Episode 1 but didn't take as much time in it as demonstrated by the fact that they were pestered less by Chris than players in Cluster 1. However they did tend to take a lot of pictures in Episode 3.
- Players in Cluster 3 were in a hurry compared to players in the other two clusters; they inspected fewer objects and took fewer pictures. They spent less time in Episode 1 and, as a result, they were pestered the least.

## Demographic Variances

We also investigated if there were any interesting demographic differences between players in the three clusters. The following 5 demographic features were collected from each player:

1. **Gender**: 1 = Male, 2= Female, 3=don't want to disclose
2. **Age**: a natural number
3. **Race**: 3 = Multiracial, 4 = American Indian/Alaska Native, 5 = Asian, 6 = Native Hawaiian or Other Pacific Islander, 7 = Black or African American, 8 =White, 9 = I do not wish to answer this question
4. Ethnicity: 4 = Hispanic or Latino, 5 = Not Hispanic or Latino, 6 = I do not wish to answer this question
5. **Education**: 1 = Grammar School, 2 = High School or equivalent, 3 = Vocational/Technical School (2 year), 4 = Some College, 5 =College Graduate (4 year), 6 = Master's Degree (MS, MA, etc.), 7 = Doctoral Degree (PhD), 8 = Professional Degree (MD, JD, etc.), 9 = Other
6. **Salary**: 1 = Under $25,000, 2 = $25,001 - $49,999, 3 = $50,000 - $74,999, 4 = $50,000 - $74,999, 5 = $100,000 -$149,999, 6 = $150,000 and over, 7 = I do not wish to provide this information

*Table 3. Demographic Variances Associated with Clusters*



| cluster | Gender | Age | Race | Ethn.<br>(n.L.) | Educ. | Salary |
|---|---|---|---|---|---|---|
| 1<br>(22 ins)<br>Insp1++<br>Pest.++ | 5(M),<br>14(F)<br>1.9(0.6) | 25(8) | 7.2(1.6) | 5(0.4) | 4.5(0.9) | 3.2(2) |
| 2<br>(21 ins)<br>Phot3++ | 13(M),<br>8(F)<br>1.3(0.4) | 24(4.5) | 7.5(1.7) | 5(0.3) | 4.1(1.1) | 3.5(2.3) |
| 3<br>(24 ins)<br>Insp1.+<br>Pest | 11(M)<br>13(F)<br>1.5(0.5) | 34.4<br>(15.8) | 7.3<br>(2.1) | 5(0.3) | 4.5(1.2) | 2.3(1.6) |

Table 3 shows how these 5 demographic features map to the three clusters. The average value and standard deviation (in parentheses) for each feature are shown. For gender, the numbers of male ("M") and female ("F") players in the cluster are also indicated. The values for gender, age and salary have the following statistically significant differences among the clusters:

- **Gender**. Cluster 1 tended to have more females than males whereas Cluster 2 tended to have more males than females. Cluster 3 is about even.
- **Age**. Cluster 3 participants were older than the ones in Clusters 1 and 2. Clusters 1 and 2 participants had similar ages.
- **Salary**. Cluster 3 participants had a lower income than those in Clusters 1 and 2. Participants in Clusters 1 and 2 had similar income levels.

These differences are statistically significant (with Student's t-test at the significance level of 0.05, one-tailed distribution, two-sample unequal variance). The underlined number shows the base for the one-tailed distribution.

## Discussion

It is interesting to observe that players in Cluster 3 were older and had less salary while at the same time being the players who inspected fewer objects and took fewer pictures. We speculated that perhaps they were less interested in playing the game compared to the other two groups.

Players in Clusters 1 and 2 were about even in age and income but Cluster 1 players were primarily female whereas Cluster 2 where primarily males. Players in Cluster 1 were more inquisitive in Episode 1 whereas players in Cluster 2 took more pictures in Episode 3. Episode 1 involves investigating clues in an apartment consisting of a few rooms. While playing in Episode 1, players needed to navigate through rooms to search for clues. In contrast, in Episode 3, players stayed in one place (a table in a bar) and they needed to examine another person's belongings while that other person was not there. We speculated that female players might find the first scenario more compelling whereas male players found the third scenario more compelling.

In summary, we have presented a study of how we have automatically extracted common strategies followed by players and found variances in the players' demographics across the different strategies. Understanding the common strategies is particularly important in the context of serious games as it has implications in devising training means and improving the game design to achieve the instructional goals. Variance in the demographics of the players for the different strategies may be translated into the practical need for additional instructions for specific demographic groups. In addition to the two questions addressed in the paper, we are also interested in the following question: do these strategies correlate with bias mitigation performance? We hypothesize that the player's strategy plays a role in the game's effectiveness. In our future work, we will investigate this important question. This is the first study on this subject for a serious game. But similar efforts should be of interest in many other games, including games for leisure and "gamification" applications, where game ideas are used in non-game settings such as GUIs.

## Acknowledgement

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

## References

Auslander, B., Lee-Urban, S., Hogg, C., and Munoz-Avila, H. 2008. Recognizing the Enemy: Combining Reinforcement Learning with Strategy Selection using Case-Based Reasoning. In *Proceedings of the 9th European Conference on Case-Based Reasoning (ECCBR-08)*.

Bares, M., Canamero, D., Delannoy, J. F., and Kodratoff, Y. 1994. XPlans: Case-based reasoning for plan recognition. *Applied Artificial Intelligence* 8: 617-643.

Cheung, G., and Huang., H. 2011. Starcraft from the stands: understanding the game spectator. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 763-772.

Cox, M. T. and Kerkez, B. 2006. Case-based plan recognition with novel input. *Control and Intelligent Systems*. 34(2): 96-10.

Davidson, A., Billings, D., Schaeffer, J., Szafron, D. 2000. Improved opponent modeling in poker. In *International Conference on Artificial Intelligence (IC-AI'2000)*, 1467-1473.

Evans, J. S. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgment*. New York: Psychology Press.

Forster, J., and Liberman, N. 2007. Knowledge activation. In A. W. Kruglanski and T. E. Higgins (Eds.) *Social Psychology: Handbook of basic principles (2nd ed.)*, 201-231. New York: Guilford Press.

Hall, M. A. 1998. *Correlation-based Feature Subset Selection for Machine Learning*. Hamilton, New Zealand.

Hamerly, G., and Elkan C. 2003, Learning the K in K-Means, *Neural Information Processing Systems*, MIT Press.

Huhh, J. 2008. Culture and Business of PC Bangs in Korea. *Games and Culture*, 3(1): 26–37.

MacQueen, J. B. 1967. Some Methods for classification and Analysis of Multivariate Observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. 281–297. University of California Press.

Michael, D.R., and Chen, S.L. 2005. Serious Games: Games that Educate, Train, and Inform. Muska and Lipman/Premier-Trade.

Orkin, J., and Roy, D.K. 2010. Capturing and generating social behavior with the restaurant game. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, 1765-1766.Pelleg, D., and Moore, A.W. 2000. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML '00)*, 727-734.

Sharma, M., Holmes, M., Santamaria, J.C., Irani, A., Isbell, C., and Ram, A. 2007. Transfer learning in real-time strategy games using hybrid CBR/RL. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence* (IJCAI-07)*,* 1041–1046.

Susi, T., and Johannesson, M. 2007. Serious Games: An Overview. IKI Technical Reports, HS-IKI-TR-07-001, University of Skövde, School of Humanities and Informatics.

Tesauro, G. 1995. Temporal Difference Learning and TD-Gammon. *Communications of the ACM*, 38 (3): 58-68.