7-10-2007

A methodology for the assessment of the behavior and performance of artificial agents

Jessica Ryan

A Methodology for the Assessment of the Behavior and Performance of Artificial Agents

by

Jessica Morgan Ryan
Physics (B.S.)
New Mexico Institute of Mining and Technology
2004

THESIS

Submitted in Partial Fulfillment of the Requirements
for the Degree of

Master of Science
Computer Engineering

The University of New Mexico
Albuquerque, New Mexico

May 2007

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Caudell, for planting the seed of an idea that turned into this thesis. His constant guidance and tutoring on the finer points of surviving in ‘the real world’ were the incalculable lessons that I took from this time when our paths were in parallel.

I am grateful for the critical reminders from the ECE secretaries, Maryellen Tow and Elmyra Grelle, without whom I probably would’ve missed some important deadline or another.

I also would like to acknowledge Dr. Caudell’s support in the use of his Visualization Lab at the HPC, which was key in the performance of my experiment.

Lastly, I would like to thank my parents and fiance for doing invaluable proofing, and for keeping the support coming.

A Methodology for the Assessment of the Behavior and Performance of Artificial Agents

by

Jessica Morgan Ryan
Physics (B.S.)
New Mexico Institute of Mining and Technology
2004

ABSTRACT OF THESIS

Submitted in Partial Fulfillment of the Requirements
for the Degree of

Master of Science
Computer Engineering

The University of New Mexico
Albuquerque, New Mexico

May 2007

A Methodology for the Assessment of the Behavior and Performance of Artificial Agents

by

Jessica Morgan Ryan
Physics (B.S.)
New Mexico Institute of Mining and Technology
2004

Submitted in Partial Fulfillment of the Requirements
for the Degree of
Master of Science
Computer Engineering
The University of New Mexico
Albuquerque, New Mexico
May 2007

ABSTRACT

Computer games are becoming more popular for both entertainment and educational applications. The growth of this technology and its realm of use creates a new demand for artificial intelligence (AI) systems: as AI becomes more prevalent, it becomes crucial for it to have a natural, human feel to it in order to best support its application. Consequently, the need for a reliable means of testing and comparing the behavioral development of the artificial intelligence used within game applications becomes important.

Turing's test has been the staple in evaluating the "intelligence" of artificial agents in applications ranging from testing chatterbots to stopping web abuse. It is currently being used in evaluating the performance of specific artificial agents in particular games. In the following thesis, a methodology has been developed to provide a new contribution to the field of AI assessment. After bringing the perceptions of the human and AI onto the same level, the Turing test is used to evaluate the "humanness" of diverse agents in generalized environments. Results of a pilot study probing the validity of this methodology are presented.

Table of Contents

List of Figures

List of Tables

1. Introduction

As artificial agents become more prevalent in modern computer applications, the need for their smooth relations with their human users becomes more apparent. For entertainment and serious games alike, a more human-like entity will naturally be easier to interact with from the point of the user. Additionally, since human intelligence is generally accepted to be the baseline with which we compare to others, one can see that one goal of artificial intelligence (AI) development is to approach human intelligence. The definition and assessment of intelligence is a difficult matter, however, so instead this thesis focuses on using a variation of the Turing test to evaluate the behavior and performance - the 'humanness' - of a given AI. Like the original Turing test, the proposed methodology limits the communication between entities, however it does so in a new manner - by bringing the perceptual abilities of the human participant down to the level of the AI with which it is interacting.

1.1 Evolution of AI in Games

As long as there have been people, there have been games; and as long as there have been computers, there have been computer games. Immediately, programmers began to address the computer’s role in a game. Rather than merely offering the game environment and its physics, the computer could provide opponents and allies with varying personalities as well.

Players came to expect more advanced non-player characters as technology progressed [1]. In order to maintain players’ expectations within the limitations of technology, developers have utilized an assortment of methods to give the impression of intelligent or complex behavior [1]. The most basic behavior for computer-controlled agents is simple looping repetition, for instance: the path of a Koopa Troopa in “Super Mario Bros” by Nintendo [2].

This simple behavior was then enhanced using random number generators to deliver a sense of sophistication and surprise to the gamer. An opponent who has a slightly randomized behavior pattern appears to be more of a skillful, non-telegraphing adversary than his fully predictable counterpart. Furthermore, an environment in which random events occur has a higher excitement and replay value than one in which traps and challenges can be avoided with practice and a good memory. Another use of randomness can be seen in non-player character (NPC) difficulty. For instance, a well-written behavioral algorithm for an opponent will yield flawless performance for the NPC, but this is not always appropriate because most game players want a challenging, but not impossible, opponent. In order to adjust an NPC skill or performance level, random imperfections are introduced in its behavior; for example: the guards in Rarewares's "GoldenEye 007 [3]." On the easier player settings, the artificial intelligence (AI) controlled guards will act quite unintelligent and unskillful. They might not react to a shot fired past their ear, or they may fire many rounds in your direction before hitting you. Yet at the most difficult settings, these same guards are deadly in their perceptions and 'decisions'. Randomness, however, has its limitations in the use of opponent behavior: if an enemy's route is randomly altered, it may put itself in the path of an oncoming missile that it had just randomly avoided.

A step above random behavior is the use of state machines. In this case, higher level 'states' will motivate lower level behaviors such as running when wounded or attacking when being provoked. Many modern games such as "Neverwinter Nights” by Atari employ such logic. Several factors such as the environment, health, the presence of allies or enemies, and even player directives (for computer controlled / player guided allies) will affect the lower level behaviors of an agent. In these cases, the agent’s high-level behavior is usually controlled by a carefully designed finite state machine whose transitions are affected by such things as the above mentioned factors [1]. Yet, even with this more complicated performance process, computer controlled game agents invariably have noticeably stilted, mechanical, and particularly, uncreative behavior - especially for the repeat player.

Further limiting the computer’s improvisational skills, developers tend to create the game in such a way as to supply the computer agents with metadata about their environment rather than design the agents to utilize the same sensory-based information that their human counterparts use. This shortcut, like the others, reduces the processor complexity of the computer agent; however it also limits the computer to a world preconceived by its developers. While this may be somewhat acceptable for simple entertainment games, it is unfavorable for serious games, or games intended for simulation and learning. For these applications, it is beneficial for the computer controlled agents to be as natural as possible - that is, they perceive and react to the environment using the same information as human players.

Serious games have an educational role in addition to being entertaining. Such games are meant to develop real world physical or mental skills in their users and are employed in the military and medical fields, among others [4, 5, 6]. Serious games are generally designed for the common personal computer or game console rather than ultra expensive high-end computing equipment, making them widely accessible to educational institutions.

One example is Flight Simulator, written by Bruce Artwick in 1977 and distributed by his company subLOGIC to various computer platforms including the Apple II and Commodore Amiga. This program lives today under the guise of Microsoft Flight Simulator and continues to provide inexpensive, easily accessible simulations for pilots in training. Another area in which serious gaming, or simulators, are widely used is the medical field [4, 6]. One particular example is in training surgeons for such procedures as endoscopic surgeries [4, 6]. It has been found that the skill and practice developed from virtual surgeries saves time and money in the training and preparation of the surgeons for real life cases [4]. A similar example is the University of New Mexico's High Performance Computing Center's Toma module. This application exists in a collaborative virtual environment and is used to teach paramedics valuable skills in the emergency treatment of roadside accident victims [6].

With increasing evidence that serious gaming has a positive effect on human participants [4], it is thought by some researchers that it will have similar benefits for artificial participants [4, 1, 7]. If a computer controlled agent has the ability to assimilate its environment in a natural way and can freely process this information, it, like a human, may be able to learn and advance by practicing in a simulation environment; and, as an added bonus, the computer agent may even find creative solutions that the designers did not anticipate. However, the judgement of the success for an artificial agent ranges in difficulty as much as the agent's application can range in complexity. It is important to define what 'success' means in such an assessment.

Since human intelligence is the baseline by which we compare other intelligences, a reasonable metric for ‘success’ is the human-like qualities exhibited by an AI.

1.2 The Turing Test

Alan Turing (1950) proposed a test that could be used to help answer the question, “Can machines think?” [8] Since the definition of thought is unclear, this question is difficult to answer. Instead, Turing described a situation - a game played between two people and a computer, that would exemplify-by-trial the computer’s cognitive abilities. In the original version of this game, known as “The Imitation Game,” one person is female and she and the computer are behind separate closed doors so as to conceal their identities. The other person acts as an interrogator and can be of either gender. The interrogator’s job is to communicate with the other two players in order to determine which is the female. The communication happens in such a way so that it does not reveal any information about the hidden players; for example, through a textual display. The female’s goal is to aid the interrogator with the identification, and the computer’s goal is to cause the interrogator to choose incorrectly. The question then becomes: “Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a *man* and a woman?" (emphasis added) [8] This new problem removes the ambiguities of the original question while still addressing the fundamental issues regarding the cognitive abilities of a machine - for if the computer can fool the interrogator at least as much as a man, then the machine should, by most reasonable definitions of the word, be considered intelligent.

Over time arguments have been made against Turing's test and its ability to demonstrate a machine to be intelligent [9, 10]. For instance, just because a machine can imitate a woman doesn't necessarily imply that it knows that it is imitating her; and just because a machine can perform one task well doesn't sufficiently prove that it can generalize - that is, perform different but similar tasks well. [10] However, it is the author's opinion that these arguments are missing the essence of Turing's test. Rather than be THE decisive test on intelligence, the imitation game should be viewed more as a single example of a test that could be used to infer intelligence of an entity - short of actually *being* that entity. This view is expressed by

James Moor in his paper ‘An Analysis of the Turing Test’ where he expresses the opinion that the imitation game, while not being the operational definition of intelligence, can be regarded rather as a sample of inductive evidence for the hypothesis that machines can think [9]. Moreover, the author feels that Turing’s test is an exemplary form of any such test for machine intelligence. It is difficult (indeed, it has not been done to date [11]) to pass the test - requiring considerable skill on the part of the computer - and the test is well defined and data is easily collected, making it an ideal model to follow [9].

1.3 Assessing Intelligence in Games

The concept of non-player characters in a computer game naturally lends itself to Turing’s test. A non-player character is, as the name implies, a character in a game that is not controlled by the player. In non-computer games, the NPCs are often controlled by the person who is describing all of the other relevant environmental factors to the players. In Dungeons and Dragons for example, NPCs are controlled by the dungeon master. It is clear to all players who the intelligence is behind these types of NPCs; however, as games were moved to the computer world, it be-came necessary for NPC control to become the computer’s responsibility. In the early stages of computer games, it was again obvious who (or rather, what) was the intelligence behind the character - the computer by two observations: there were no other human players involved, and the NPC’s behavior was extremely unintelligent. Now it is not as clear who or what the intelligence is behind a given character. With more sophisticated artificial intelligence software and the use of networked game playing, the perceived division between non-player and player characters is becoming blurred.

As the use of artificial intelligence in serious games and simulations increases, it can be assumed that the next step is the increased use of artificial intelligence in the real world; but before this can occur, a method of expressly testing the intelligence, as expressed via its behavior and performance, of a computer agent needs to be contrived in order to better understand what AIs are better suited to what worlds or tasks. Based on the idea behind Turing’s test, assessing the intelligence of an artificial agent is essentially assessing how human it behaves - its ‘humanness’. Since human behavior and intelligence is the benchmark by which others are judged, it follows to correlate the humanness of an AI with its skill level or abilities. Thus, a more advanced or better developed AI will seem more human. While there are those [10] that think this approach is deceptive, it is felt by the author that given the current stage of technology this approach will lead to insightful AI developments. Given that an AI is not built specifically to mimic a human teacher (thus having no general understanding of the factors behind the decisions), then the more indistinguishable the AI's performance is from that of a human, the more advanced that AI has become.

1.4 Similar Research

1.4.1 Non-Game Turing Test Applications

Currently, the most frequent uses of Turing's test are found in the Loebner Prize competition and in Human Interactive Proofs (HIP). The Loebner Prize is the first formal instantiation of the Turing test [11]. In 1990, Dr. Hugh Loebner pledged a grand prize of $100,000 for the first computer whose responses were indistinguishable from a human's. This developed into an annual competition that rewards programmers for the most believable chatterbot - a program designed to maintain an intelligent conversation with a human. For the contest, a programmer must develop a chatterbot which will then hold a conversation with a human judge using a predetermined protocol. While, starting with the 2007 contest, the judges will be required to start the conversations, there are no other restrictions on the conversational content. The finalists and winners of the competition are based on their ability to respond intelligently. This is very much in the spirit of Turing's thought experiment. Thus far, no entry has been able to win the grand prize by deceiving the judges in either a text-only test or a full-blown textual/visual/auditory test. [11]

The other common application of the Turing Test is in human interactive proofs, a type of reverse Turing test. The motivation behind human interactive proofs came from the internet and the malicious use of bots. It didn't take long before hackers realized their potential to wreak havoc on internet users and servers; for instance, by consuming bandwidth, harassing chatters, or sending spam from collected free email accounts. This generated a need to create automatic methods that could tell whether the entity requesting to use a web-based service was human or not.

These methods are collectively known as human interactive proofs and are simple tests administered by a computer to a client. Based on the response, the computer then determines if the client is human or machine. [12] A commonly used test is known as CAPTCHA or "Completely Automated Public Turing test to tell Computers and Humans Apart" by Carnegie Mellon University. This test (see figure 1) involves the recognition of letters or numbers that are distorted in ways to prevent optical character recognition, a type of automated translation of characters into machine-readable text. [13]

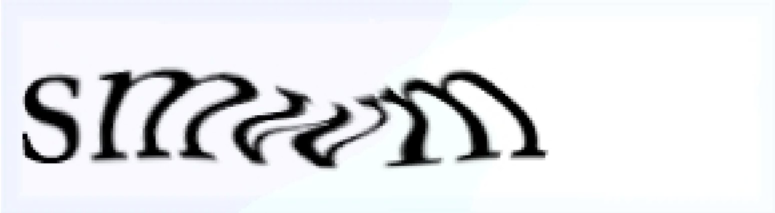


Figure 1: Example CAPTCHA of "smwm" obscured by distortion. Image courtesy of Wikipedia.

Research is continuously being done in the field of human interactive proofs, which is beneficial in two ways. If an HIP algorithm is developed that cannot be broken, it will be valuable in protecting online services from bot attacks, but if an HIP algorithm is defeated, then it signifies that artificial intelligence has become that much more sophisticated. [14] A new method of HIP is being developed under the name of ARTiFACIAL which uses the reverse Turing test and facial features, rather than text recognition, to segregate humans and computers [14]. The algorithm generates a distorted image of a face and the user must identify six particular points on that face to correctly pass the test. This takes advantage of both humans' innate facial recogniting abilities and the simultaneous difficulty that computers have with the same task.

Another recent HIP technique is to apply the same idea to speech recognition. Since speech based services are gaining in popularity, for convenience and accessibility reasons, and building an algorithm to understand and manipulate spoken language is quite manageable, creating an audio HIP will be both useful and necessary [12]. Again, in this method, a reverse Turing test is applied using synthesized speech that is distorted in such a way as to render it likely that automated speech recognition algorithms would fail the test [12].

1.4.2 Game Analysis Methodologies

While the fundamental focus of this thesis is to establish a methodology for assessing various AIs, it is relevant to explore the current research regarding the analysis of the worlds in which they inhabit because in future research, the analysis and parameterization of the game environment will provide important context in the evaluation of the AI.

In “Formal Models of Game Design,” Steven Grünvogel [15] has created a new flexible, mathematical formalism for the analysis of games similar to that of classical game theory analysis. Such formal models can be used in the creation of a language for some facets of the game design. Such a language could then be used to discover relationships between game elements and games as a whole. It is important, however, to have an accurate yet simplified model to represent the game for examination. The precursor to representing the game model in mathematical language is to define and characterize the game in question. To do this, one must perform a critical analysis of the game in order to capture its basic elements.

In their research, “Game Analysis: Developing a methodological toolkit for the qualitative study of games,” Nathan Dutton and Mia Consalvo [16] have outlined a four layered approach to systematically and critically analyze games. The four areas are: a) object inventory, b) interface study, c) interaction map, and d) gameplay log. While the study of a single area is useful, the authors intend for all areas to be studied together to present an overall picture of the game. The areas were chosen such that they represent fundamental game components as well as static/ dynamic, un/changeable game elements. Object inventory analysis helps the researcher address larger issues such as what role objects have within the game, what utility or purpose objects have, and what economic and social structure is apparent from the use of objects. Interface study yields important data regarding the choices and information that is presented to the user. Like a language, the interface guides and shapes the available thoughts of the user. The interface also helps define what is important within a game, for instance score or health, as well as the importance and purpose of information that is withheld. A more difficult aspect of a game to study is the interaction map. In this case, the focus is on the players' choices as they pertain to other players be they human or not. Due to the broad range of possibilities, many analyses are dismissed due to this complexity. Yet, even simply asking if the interactions are limited or if they change over time will still help the researcher understand the freedom allowed to the player. Lastly, the overall world is considered in order to detect such things as emergent behavior or other situations within the game, as well as total feel of the game to the player.

A slightly different, but relevant, tact was taken by John Sterman of MIT [17] in his research on the testing of behavioral models via direct experiment. Though the models to which he refers are database driven decision tools rather than biologically inspired frameworks, his methods of evaluating them are pertinent to this paper's research. Sterman argues that in the case when sufficient data cannot be collected to accurately model a system, one can essentially guess the values of the system's parameters and then evaluate the model's accuracy by direct experimentation. He reasons that since the purpose of a simulation model is to mimic the real situation, it must exhibit decision making behavior "as it is, and not as it might be if the decision makers were omniscient optimizers." This line of thought applies today with the analysis of artificial agents - if it behaves right, even in complex environments, then it must be approaching 'right'.

1.4.3 Turing Test in Games

This subsection reviews similar research showing how the Turing test is used in game applications.

Laird and Duchi [18] composed an artificial player for the computer game, Quake®. This robot, called the Soar Quakebot, is parameterized along four dimensions: decision time, aggressiveness, aiming skill, and tactical knowledge. The humanness of this robot was then tested using a modification of Turing's test. Several humans of varying skill competed against an expert player and recordings were made from their viewpoints. The Quakebot then played against the expert player and its viewpoint was also recorded. Human judges then viewed the recordings in a blind survey (where the judges did not know if the recordings were of a human or computer player) and evaluated the humanness and skill of the behavior they observed using a 1-to-10 scale. In addition, they gave an overall rating of whether it was an artificial or real player. This data was then tallied and averaged over all responses. The methodology for Laird and Duchi's experiment identified trends, particularly relating the bots' humanness to their decision time and aiming skill, and indicated that it could be used to explore their research in more detail.

McGlinchey and Livingstone [19] conducted a similar study regarding the believability of AI players in which they tested the human-like qualities of AI Pong players. The AI was a self-organized map trained on human data and was able to replicate the distinct behaviors of various players [19]. Like the previous study, a number of Pong games were recorded and played back for observers. For each game, the observers were asked which bat (left, right, both, or neither) was controlled by a human player. The observers also had the opportunity to answer why and how they made their decisions. The results of the research showed that while the AI could successfully imitate different playing styles, this imitation was not enough to fool human observers in believing it was a human player. The AI performed well, being identified as human as much as the humans were; however, subtle movement differences, such as jerkiness, distinguished it as not being human.

Gorman, Thurau, et al [20] performed a study that involved Quake II® - very similar to Laird and Duchi's study - only their synthetic robot emphasized imitation learning using a

Bayesian-based approach for the derivation and mimicry of human behavior and motion patterns. Their robot learned the mappings between an expert player's status and his actions, and consequently could adapt to situations that the player did not face. It was specially designed to move like a human in order to successfully deceive observers where an AI like McGlinchey and Livingstone's failed. They classified three distinct metrics that applied to the analysis of imitation based agents: statistical analysis of the accuracy with which the agent reproduces human behavior, believability testing to rate how much the agent is perceived as human, and performance assessment of the agent in competition with other players. The believability testing was done using a modification of Turing's test similar to the previous two studies: subjects viewed isolated video clips of in-game play through the player's eyes. They then rated the player's humanness using a given scale. This information, along with the subjects' game playing experience, was used to create a weighted representation of the degree of humanness of each clip and thus, the robot itself.

Their comprehensive methodology has given reliable results of their imitation agent in comparison to standard artificial agents and human players. The main difference between this method and the others is the manner in which the data were compiled. Gorman et al [20] used similar Turing-like tests to gather the data, but then averaged and weighted the data which yielded more uniform and comparable results than raw data.

1.5 Introduction of a New Methodology

The above research areas each contribute to a particular aspect of the research that follows in this paper. As computer-controlled agents are becoming more widespread, their formal study is becoming important within the research community. While Turing tests still abound in current research, particularly in that of games and the artificial agents within them, no research has been located in the literature that focuses on creating a methodology for the study and comparison of non-specific agents in such a manner that brings the human down to the computer's sensory level. This 'leveling of the playing field' is important and parallels Turing's original thought experiment. Turing used the *thoughts* of the entities as the humanness indicator and thus equalized the human and computer by removing all physical contact and observations from the experiment (with the use of a text-only channel of communication). This new methodology uses the *behavior* of the entities as the humanness indicator and thus equalizes the human and computer by limiting the human’s sensorium to that of the computer agent. Thus, the human cannot detect anything more about the environment than the AI. Additionally, this methodology is not limited to a specific type of AI or environment and thus allows for the study of a broad range of possibilities.

2. Approach

The proposed methodology allows for the ‘plugging’ of different AIs into a non-specific environment in order to assess their performance and behavior via rating their humanness. A key aspect is the equalization of human interrogator and computer opponent. Because technology is not advanced enough to bring the computer opponent up to the level of the human, it is necessary, for a fair application of Turing’s test, to bring the human down to the computer’s level. While eventually the methodology shall include the parameterization of both the AI and its environment, the focus of this thesis will be in showing that the proposed methodology demonstrates trends indicating that it will be useful in future research on the systematic assessment of AI performance.

This methodology will be tested by a pilot study that will establish a baseline showing that the AI and human opponents can indeed be characterized via an assessment of their humanness. The systematic method of AI testing will allow comparisons and progressive development to be accomplished, and thus future research should then be able to indicate how more sophisticated AIs will become less discernible from its human counterparts. Since the environment and AI implementation is not specified, the methodology will be applicable to any comparable simulation or real world situation thus making this research valuable outside of itself.

This thesis describes a two-part experimental design that implements the proposed methodology. In part I, the human player, or rather, the interrogator, will play against an opponent not knowing if the opponent is human or artificial. The interrogator interacts with the opponent in a first person manner and is limited in his perceptions in such a way as to receive only the same environmental information as the computer opponent and no more. The second part of the experimental design serves to provide another perspective in order to test the validity of the methodology. In the second part, the interrogator observes two agents from a third person perspective where one, both, or neither of the agents may be computer-controlled. Following is a description of the pilot study approach.

2.1 Approach Setup

The experiment utilizes three pieces of software developed at the University of New Mexico's High Performance Computer

Center: Flatland for the visualization of the game, Flatworld to simulate the robot and the physics of the 2D world in which it lives, and eLoom to code the neural network implementations of the AI controllers. Please refer to figure 2 for a visual explanation of how these applications work together.

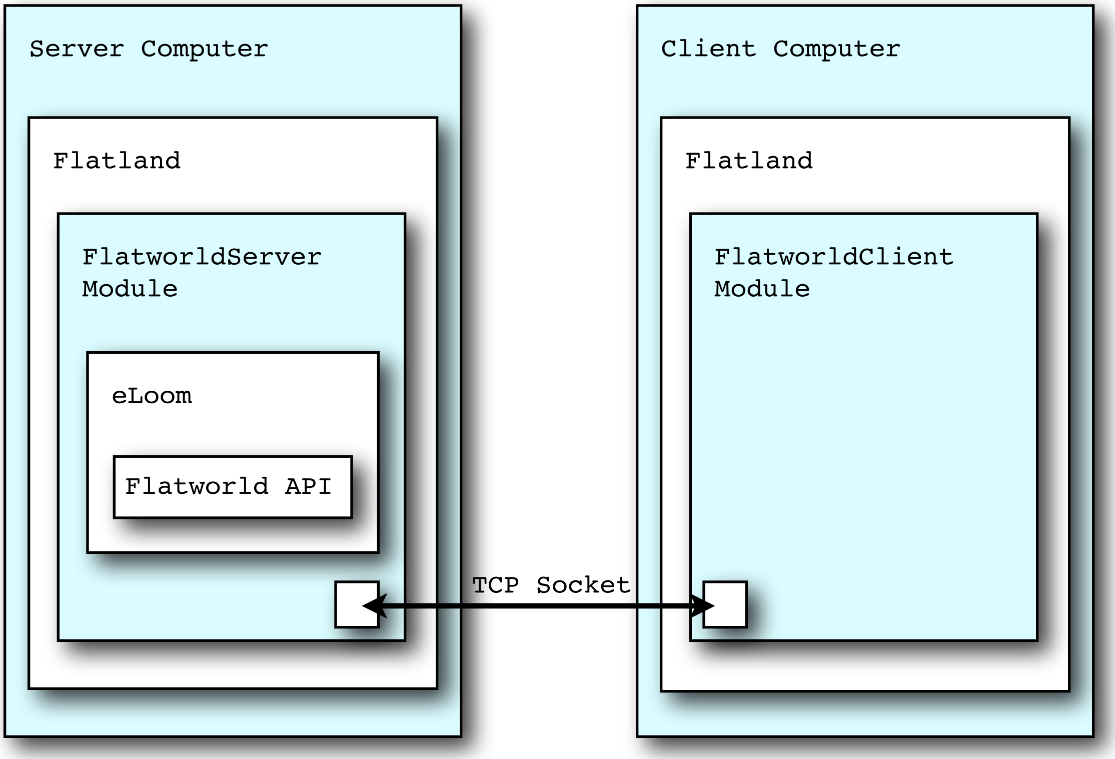


Figure 2: software and hardware set up

Flatland is a virtual environment based on OpenGL that allows for the visualization of and interaction with complex graphical representations of data [21]. The specific version used for this experiment is Flatland D for Macintosh.

Flatland is a multithreaded application that uses dynamically linked and shared libraries to assemble user-created modules that modify the environment to the application-defined needs of the developer. For this experiment, two modules were created and loaded into Flatland on two different computers as shown in figure 2. One module is a client module (called FlatworldClient) that receives and displays data from the other, server, module (called FlatworldServer). The server module is the main module that combines and visualizes the other two components: Flatworld and eLoom.

Flatworld is the environment in which the AI controlled robots and human driven robots exist; its API is a set of function calls that return the current details regarding the environment. In essence, the Flatworld API is the world definition and the robots' ability to sense it (ie: the robots' "body" - eyes, ears, etc). The world contains three types of items: 'good', 'neutral', and 'bad' items. Good items will charge a robot's battery a fraction, bad items will discharge it a fraction, and neutral items will have no effect. Each turn in the game, the battery depletes a small amount, and so the goal of the game is to survive as long as possible by keeping the battery charged.

eLoom is a simulation environment in which to implement the AI controller for the robot and is responsible for processing the robot's perceptions of the world and commanding the actions of the robot's body.

The FlatworldServer module's purpose is to display the dynamic world created by Flatworld in Flatland. This was accomplished by sharing memory and allowing Flatland to access Flatworld's object information data. The module creates a thread whose sole responsibility is to regularly invoke eLoom's core scheduler function, a user defined function. In this case, the scheduler invokes each of the robot's senses and stores the returned values. The robot then processes this data to make a decision about its world observations. Following this, the robot performs its decided action and observes its internal states to learn from its decision. In this preliminary research, a simple neural architecture is being simulated using conventional algorithms (please see Appendix A.2 for a detailed description of the algorithm). It is the server module that will provide the opponent for the interrogator, be that opponent human or artificial.

The FlatworldClient module’s role is to receive the data from the server that pertains to the subject’s robot. It displays this data and sends control requests back to the server. It does not interact directly with Flatworld and eLoom. It is the client module that the interrogator will use to participate in the experiment. Detailed descriptions of all the code involved in this project are provided in Appendix A.

The experimental design is divided into two sections. In part I: a human subject, the interrogator, will control his or her own robot and will compete against another robot in the world. The opponent robot can be controlled either by another human or by the computer-based AI. The goal of the subject will be to survive while simultaneously observing the opponent robot in order to assess its humanness. The subjects are first asked to estimate their experience level in playing computer games using a scale of 1 to 5 (see Appendix B.3 for the actual survey tool used). This allows for the weighted consideration of their observations, as described later, and for the computation of a confidence index which is useful for comparisons among different studies. The subjects will then play a series of games under various world and robot complexity relationships (however, in this pilot study, one world and one robot only are used), each time examining the behavior of the opponent and indicating his humanness using a scale of 1 to 5 (see Appendix B.3 for the actual assessment tool used). Please refer to figure 3 on the following page for snapshots of the first part of the experiment as seen from the subject’s perspective.

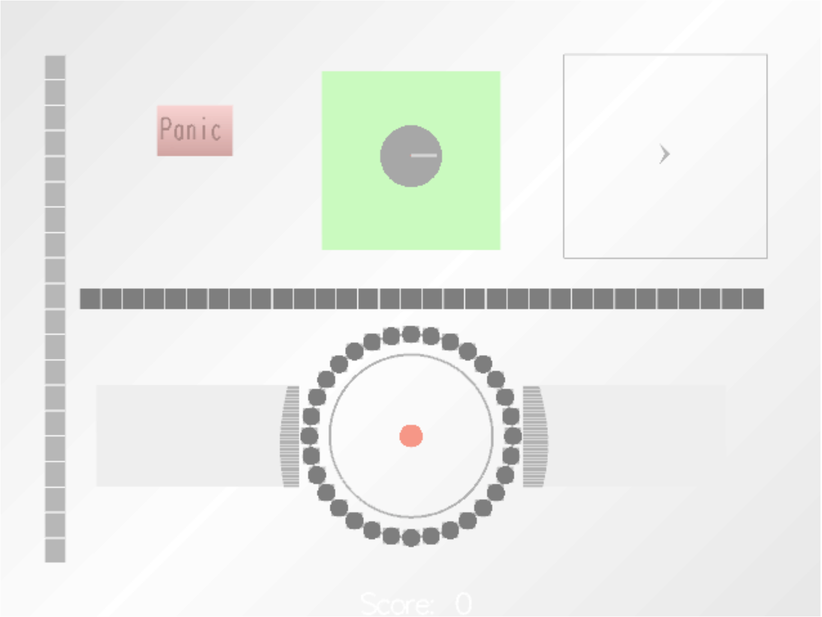
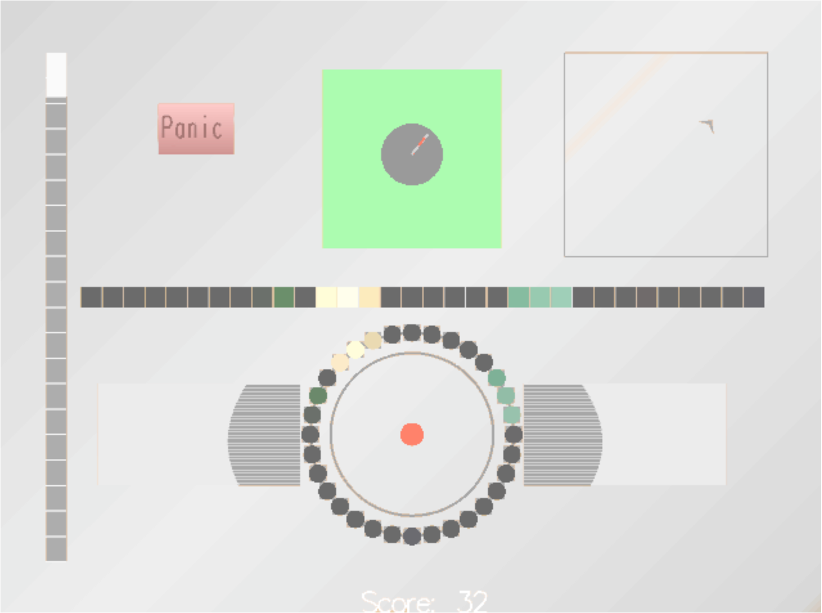


Figure 3: Views as seen during the first part of the experiment. The bottom view shows three objects in the robot's field of view (from left to right, a green object, a tan object, and another green object). For detailed descriptions of the components, see appendix B.2.



In part II of the experimental design, the subject observes pre-recorded games from a third person perspective. After answering the same questions regarding game play experience, the subject then rates each robot in the movie using the same scale as described above (see Appendix B.4 for the actual survey and assessment tool used). Please refer to figure 4 on the following page for snapshots of the second part of the experiment as seen from the subject’s view.

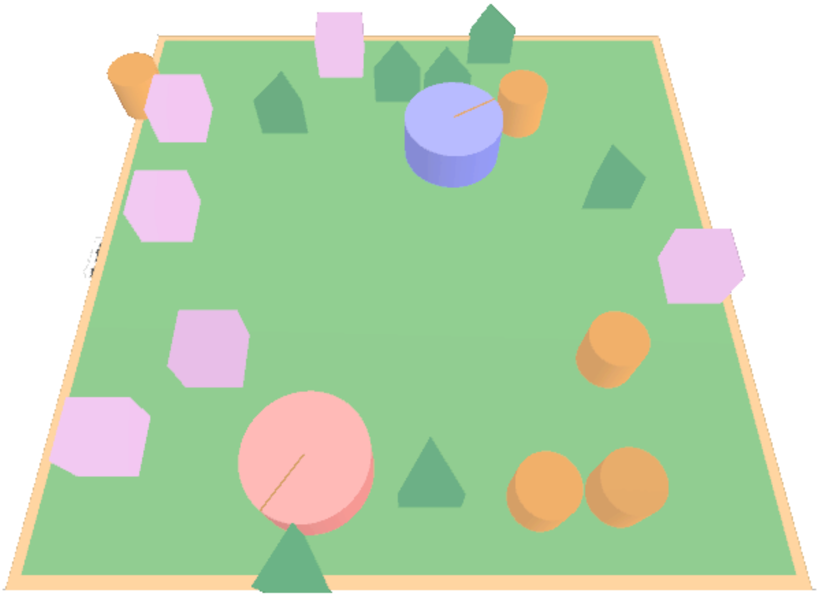
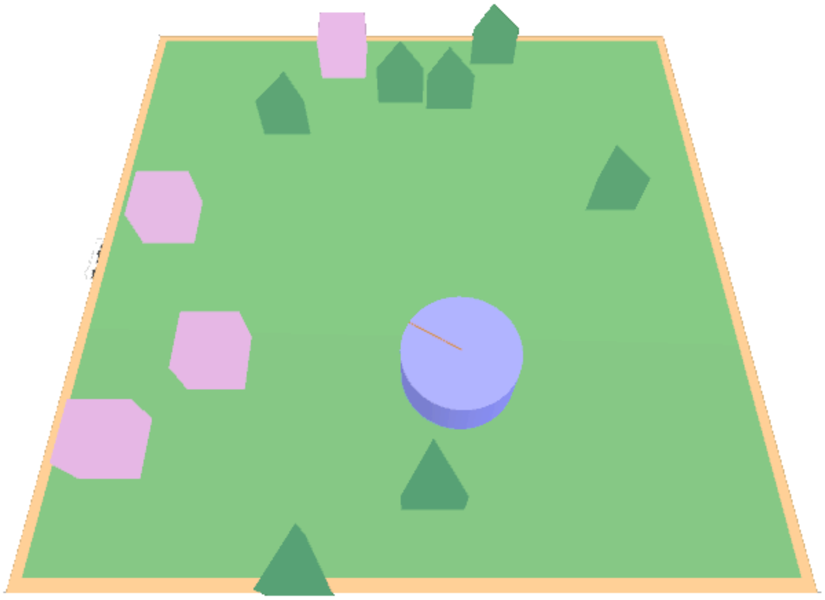


Figure 4: Views as seen in the second part of the experiment. The large red and blue cylinders are the two robots, the small cylinders are 'good' items, the squares are 'neutral' items, and the triangles are 'poison' items. Bottom figure shows the same game several rounds later where all the good items have been consumed and the red robot got lost and wandered off screen.



The subjects consisted of a convenience pool of adults, aged 18 and above, including volunteers from the University of New Mexico's Electrical and Computer Engineering Games class. This sampling of volunteers was chosen to provide experimental subjects who have interest and experience in game playing so as to increase the confidence index. In addition to compiling the free style comments given by the subjects to help assess the methodology and experimental design, the experiment follows the model described by Gorman et al [20] for numerically assessing the believability, or humanness, of artificial agents.

Recall that each opponent is rated on a scale of 1 (definitely human) to 5 (definitely artificial). Since the true value of the opponent player is always either 1 or 5, the degree to which the player persuaded the subject that it was human during a particular encounter, *i*, can be expressed as the normalized difference between that subject's rating and the value corresponding to artificial:

$$h_s(p_{o,i}) = \frac{|r_s(p_{o,i}) - A|}{\max(h)} \tag{a}$$

where h_s is the degree to which subject *s* regarded player *o* of game *i* as human, *rs* is subject *s*'s rating of player *o* of game *i*, *A* is the value on the rating scale which corresponds to 'artificial' (5), and *max(h)* is the maximum possible difference between a player's rating and the value of 'artificial' (4). For example, $h_s(p_{o,i}) = 0$ if the subject identified the player as artificial, and 1 if he identified it as human, and somewhere in between if he chose one of the 'probably' or 'don't know' options.

This humanness degree is then weighted according to the subject's game experience level:

$$w_s(p_{o,i}) = \frac{e_s\, h_s(p_{o,i})}{avg(e)} \tag{b}$$

where e_s is the experience level of subject *s*, and *avg(e)* is the average experience level of all subjects.

Finally, to compute the overall believability of a player *o*, the weighted ratings by each subject are summed over all games of that player and averaged:

$$b_o = \frac{\sum_{s=0}^{n} \sum_{i=0}^{m} w_s(p_{o,i})}{nm} \tag{c}$$

where b_o is the believability index of player *o*, *n* is the number of subjects, and *m* is the number of games played by player *o*.

Following Gorman et al., a confidence index is computed to aid in the comparison of data across different studies:

$$c = \frac{avg(e)}{\max(e)} \tag{d}$$

where *avg(e)* is the subjects' average experience level and *max(e)* is the maximum experience level of all subjects.

2.2 Approach Execution

In order to gain insight into the application of the methodology, this pilot study was separated into two distinct, yet similar, phases that were concerned with slightly different levels of subject involvement.

2.2.1 Phase I

In phase I, subjects were asked to come to the Center for High Performance Computing's (CHPC) Visualization Laboratory at the University of New Mexico. Each subject was given an informed consent (see Appendix B.1) to sign and was oriented as to the nature of the experiment. Each subject was then given a typed briefing (see Appendix B.2) explaining the program that he would be using. It defined the on-screen interface (visual sensors, battery meter, etc) as well as the program controls. It gave a brief explanation of the world and the subject's goal in the game.

The investigator then answered any subject questions without disclosing pertinent information that would bias the experiment. The subject was then given up to ten minutes to work in a practice world in which no opponent was present. This gave the subject a chance to become familiar with the game controls and learn which items in the world were beneficial and which were not. Again, questions were answered by the investigator, and when the subject felt comfortable enough with the game, testing began.

This phase of the experiment involved two computers: a desktop Macintosh G4 computer located in the CHPC Visualization Laboratory and a Powerbook Macintosh G4 laptop computer that was removed from the room so as to minimize artificial influences on the subject's decisions. The FlatworldServer module was run in Flatland on the Powerbook, while the FlatworldClient module was run in Flatland on the desktop - each one able to see the other player, interacting via a wireless TCP/IP socket protocol. Three rounds, each lasting about 3 minutes, were performed. Following each round, the subject via iChat, an instant messaging program, was asked to rate the humanness of the opponent using a simple questionnaire (see Appendix B.3) and then reply when he was ready to begin the next round.

Subjects arrived to participate in the experiment in a random order. Odd numbered subjects played against a human opponent, the investigator, for all three rounds while the even numbered subjects played against the artificial opponent for all three rounds. The subjects were not told that they were playing against the same opponent each time - they were only informed that they may or may not be competing against a human. Following each round/game, the subjects answered one question regarding their judgement on the humanness of their opponents. At the conclusion of the experiment, this questionnaire was collected along with the signed consent. Results are discussed in chapter three.

2.2.2 Phase II

Phase II of the experiment was done in the CHPC Visualization Laboratory and involved the same Macintosh Powerbook G4 as the first phase in tandem with the big screen projector system present in the laboratory. In this phase, subjects participated in groups of up to 3 people at one time. Subjects were asked to observe 10 pre-recorded games, some of which came from phase I, others were created solely for the purpose of recording for the experiment. The recordings were displayed using the FlatworldPlayer module running in Flatland on the Powerbook. In the total of 10 games, there were 20 opponents viewed. Fifty percent of the opponents viewed were computer, the other fifty were human. Thus, each opponent type made the same number of appearances and in various combinations (see table 1 for actual combinations used).

The subjects were told that each game had an independent combination of human and computer players and after viewing the game in a third person overhead view, were asked to rate the humanness of each opponent. Some of the subjects had participated in the first phase, while some had not - and this was indicated on the questionnaire for later review.

Game	**Red Robot**	**Blue Robot**
1	Artificial	Human
2	Artificial	Human
3	Human	Human
4	Human	Human
5	Artificial	Artificial
6	Artificial	Artificial
7	Human	Human
8	Human	Human
9	Artificial	Artificial
10	Artificial	Artificial

Table 1: Details of the actual combinations of artificial and human opponents used in phase II of the experiment.

3. Results

This thesis proposed a methodology, founded on the Turing test, that would enable the assessment of artificial agents' human-like behavior. A two-part experimental design was described and implemented in order to explore the validity of the methodology. In part I, the human interrogator played against an opponent not knowing if the opponent is human or artificial. The interrogator interacted with the opponent in a first person manner and was limited in his perceptions in such a way as to receive only the same environmental information as the computer opponent and no more. The second part of the experimental design served to provide another perspective in order to further test the validity of the methodology. In the second part, the interrogator observed two agents from a third person perspective where one, both, or neither of the agents were be computer-controlled.

Following are the pilot study results - particularly those showing the humanness ratings of the opponents, the demography of the players' experience, the calculated believability and confidence indices, and finally the freeform comments of the subjects regarding their decisions.

3.3 Experimental Results - Phase I

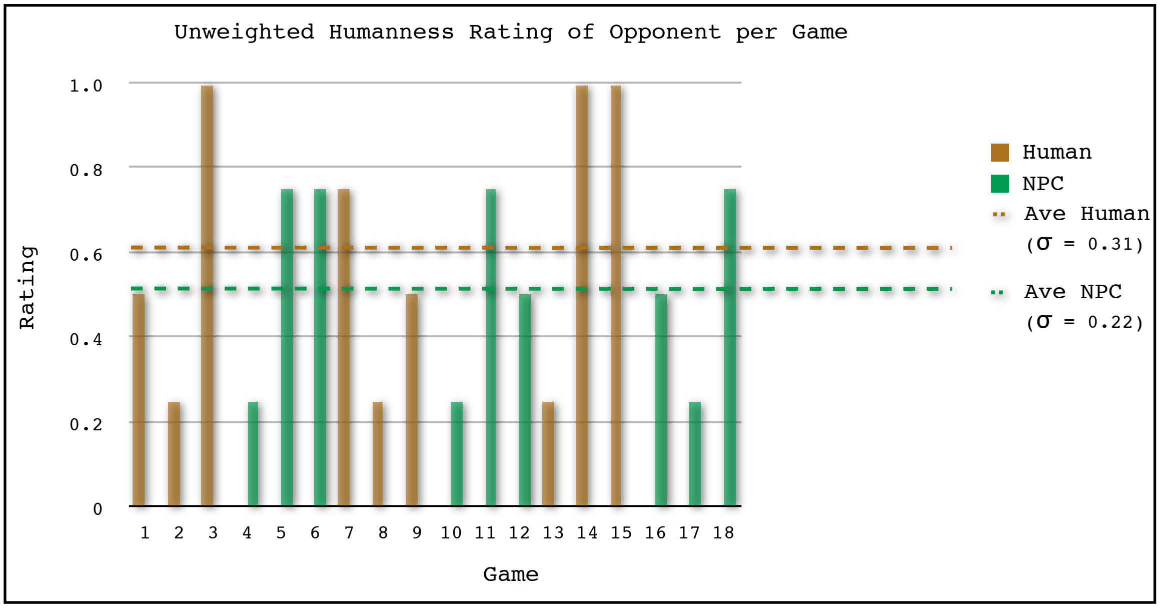


Figure 5: This graph shows the unweighted rating as collected via the questionnaire. There were 6 subjects that played 3 games each, hence 18 games total.

On average, the human opponent had a humanness ranking of 0.61 while the artificial opponent had a humanness ranking of 0.53, as indicated by the black and blue dashed lines respectively.

Figure 5 shows the unweighted humanness ratings of the opponents as calculated using equation (a). Each game played is plotted along the x-axis, and the value of the humanness rating for the opponent of that game is plotted along the y-axis.

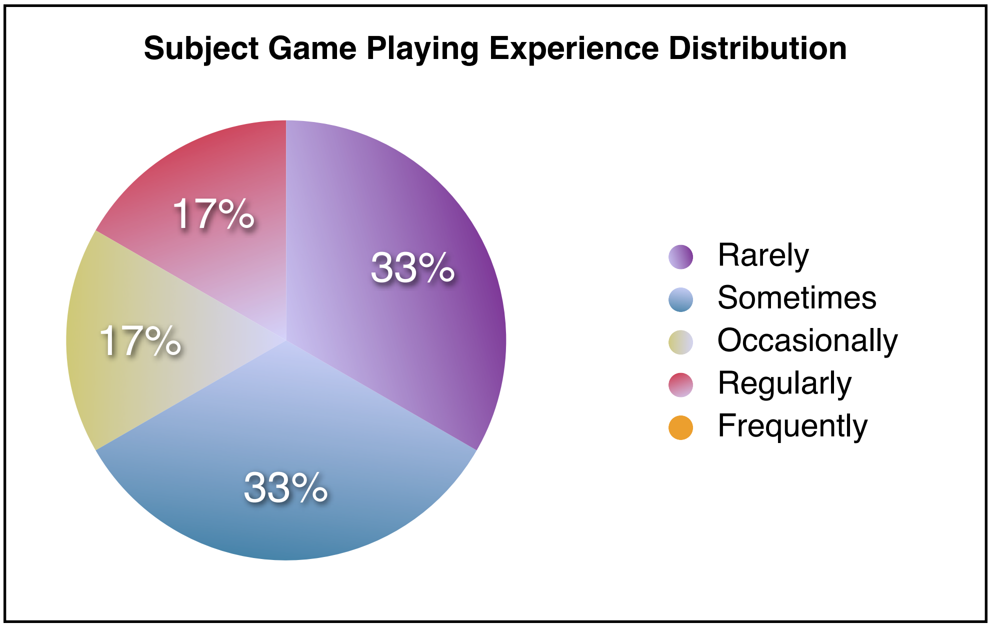


Figure 6: The subjects' game playing experience as collected via survey. Note that no subjects considered themselves frequent game players.

Opponent	Believability	Confidence
Human	0.67	0.54
Artificial	0.47	

Table 2: Believability and Confidence Indices as computed using formulae specified in section 2.2. Thus, the human opponent was correctly identified as human 67% of the time while the artificial opponent was misidentified as a human 47% of the time.

Table 2 shows the believability and confidences indices as calculated using equations (c) and (d) respectively. As an example which may be useful in understanding the meaning of the believability indices, consider a perfect group of subjects. If these subjects did not make any mistakes in the assessment of the opponents, the believability indices of the human and artificial opponents would be 1 and 0 respectively. Thus, in the pilot study, the subjects correctly classified the human opponent as human 67% of the time and incorrectly classified the artificial opponent as human 47% of the time.

Judged Human Because...	**Experience**	**Actual Opponent**
“Limited contact, but not sure”	2	Artificial
“I got a better score”	2	Human
“Better score”	2	Human
“seemed to be fewer ‘good’ items, thus I believe that a human was obtaining them before I was”	3	Artificial
“Grabbed food ahead of me like it knew what it was doing”	1	Human
“The other robot was as lost as myself”	4	Human

Table 3: Subjects’ reasoning for judging the opponent as human; the subjects game playing experience; and the opponent’s actual type

Judged Artificial Because...	**Experience**	**Actual Opponent**
“Seemed uninterested”	2	Artificial
“It seemed slow in the actions but I got very low score”	2	Human
“’good’ items seemed to be readily available”	3	Artificial
“Slower movement out of reach”	1	Human
“I think the other is finding the food quicker”	4	Human

Table 4: Subjects’ reasoning for judging the opponent as artificial; the subjects game playing experience; and the opponent’s actual type

Couldn’t Tell Because...	**Experience**	**Actual Opponent**
“Seemed to move with a purpose, but whether that’s human or not, I don’t know”	2	Artificial
“Little contact with red. Hard to judge”	3	Artificial
“I couldn’t imagine what the opponent was doing”	4	Human

Table 5: Subjects’ reasoning for being unable to judge the opponent; the subjects game playing experience; and the opponent’s actual type

Tables 3 - 5 show the freeform written comments given by subjects justifying their choices as well as the experience of the subject making the comment, and the value of the actual opponent to which the comment refers.

3.4 Experimental Results - Phase II

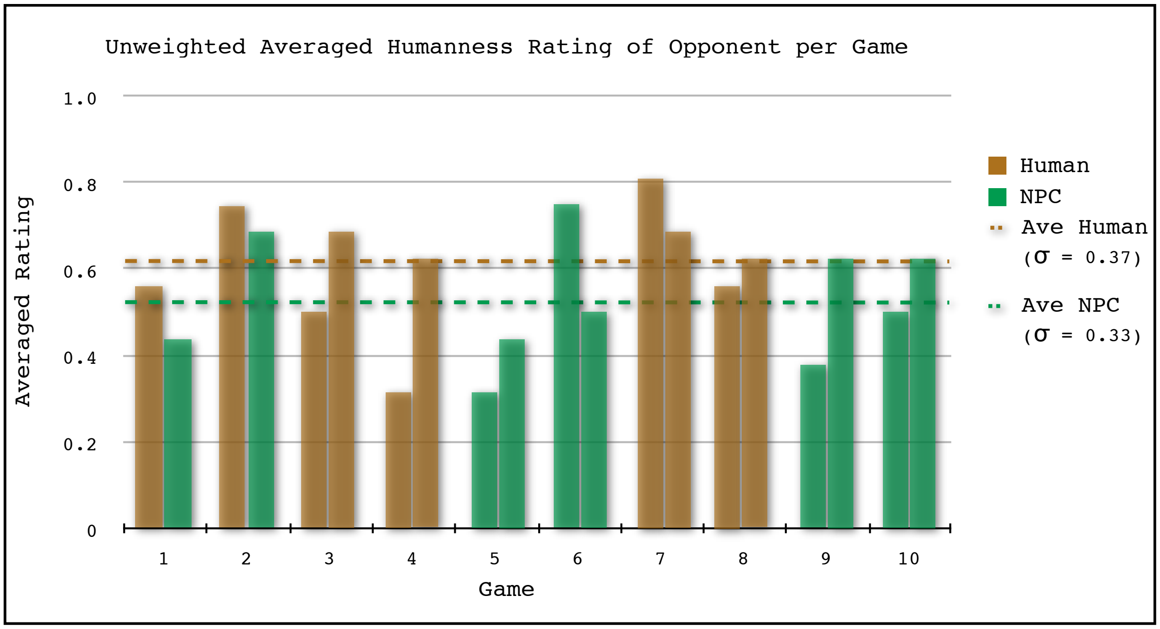


Figure 7: Chart showing the humanness rating of each opponent per movie averaged over the individual respondents as well as the overall average rating of each opponent type for the second phase of experiment.

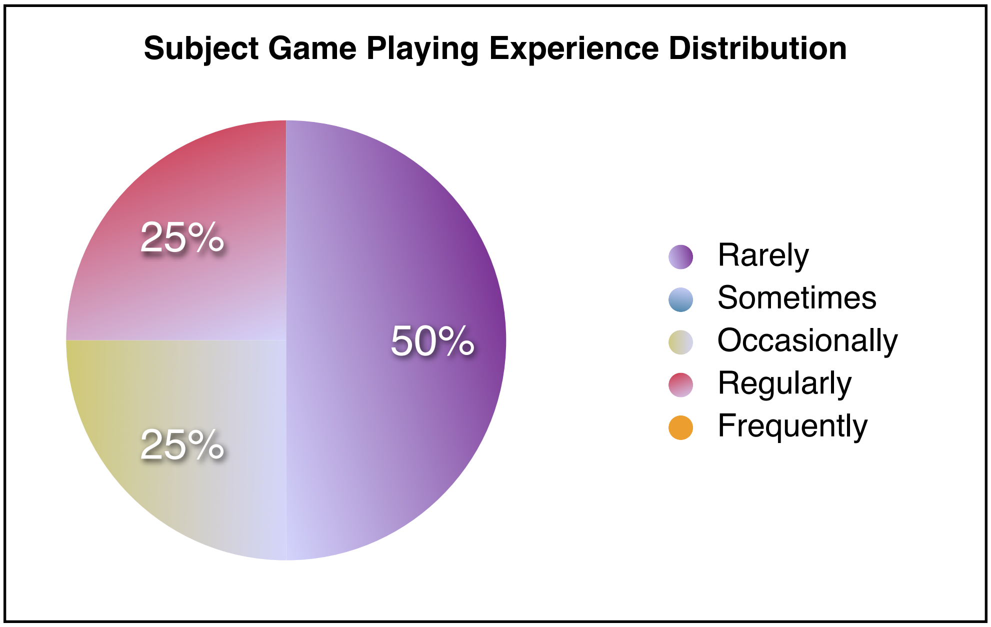


Figure 8: The subjects' game playing experience as collected via survey.

Opponent	Believability	Confidence
Human	0.73	0.56
Artificial	0.57	

Table 6: Believability and Confidence Indices of all subjects participating in second part of experiment.

Opponent	Believability	Confidence
Human	0.85	0.67
Artificial	0.65	

Table 7: Believability and Confidence Indices of subjects who participated in both parts of the experiment.

Opponent	Believability	Confidence
Human	0.37	0.25
Artificial	0.31	

Table 8: Believability and Confidence Indices of subjects who participated in only the second part of the experiment.

Judged Human Because...	Experience	Actual Opponent
“seems to move with more attention to goal of eating ‘good’ items”	3	Human
“seems to become ‘confused’. I assume this to happen because it is trying to only eat the ‘good’ item next to the ‘bad’ one. I think an algorithm would be quicker to choose correctly.	3	Artificial
“blue seems human as it ‘tracked’ red as red passed.”	3	Human
“seemed to get confused and moves off of the map	3	Artificial
“both go off the map as I think a person playing who is confused would”	3	Artificial
“Red slowly moves trying to find food”	3	Human
“seems human”	3	Artificial
“gameplay seems natural, not algorithmic”	3	Artificial
“red seems to eat everything like a human who forgets to turn ‘eat’ off and blue seems confused and eats nothing (like a player not familiar enough with the game)”	3	Humans

Table 9: Subjects’ reasoning for judging the opponent as human in second part of the experiment; the subjects game playing experience; and the opponent’s actual type

Judged Artificial Because...	Experience	Actual Opponent
“seemed to track ‘good’ items relatively well”	3	Human
“tracks ‘good’ items and consumes them”	3	Artificial
“moves in very set move-and-track methods”	3	Human

Table 10: Subjects’ reasoning for judging the opponent as artificial; the subjects game playing experience; and the opponent’s actual type

These results will now be discussed in the context of the methodology in the following chapter.

4. Discussion

With the use of artificial intelligence on the rise, it becomes important to have a dependable means of testing new agents and comparing them to others. This study successfully probed the effectiveness of the proposed methodology.

The results of the preliminary research are compelling and lead one to believe that the methodology is valid. In both parts I and II, the results indicated that a difference could be discerned between the human and artificial opponents. Tables 2 and 5 show that the believability index, or humanness rating, for human opponent is above average while that for the artificial opponent ranges from 16% to 20% lower. This indicates that the methodology was successful in discerning between the two opponents - and as the artificial agent used in the study was considerably skillful at the game (approximately equal in performance to the human opponent), it appears that the methodology will be useful even as the sophistication of artificial agents increases and the gap between human and agent performance narrows.

The use of the believability index developed by Gorman et al. coincided closely with the simple averaged 'humanness' results, lending strength to its use and dependability of the results of the pilot study. In each case, the human opponent was more believably human than the artificial agent - an expected result due to the simplicity of the AI used in the experiment.

While the distribution of subjects' game playing experience was a little lower than hoped (the average subject experience was slightly below average experience), this may only lend more credibility to the methodology because it is believed by Gorman et al and the author that the more experienced the subjects, the better able they are to distinguish human and artificial opponents; and since the less-than-average subject population was still able to appropriately tell the difference between the human and artificial opponents then it is expected that a more skilled subject population will yield even more distinctive results. However, due to low subject population, much more research is necessary to test this conjecture.

Furthermore, in the second part of the experiment, a noticeable difference was found in the judgement of subjects who participated in both parts I and II versus those subjects who participated in only the second part. As noted in table 7, the believability indices between human and artificial of those subjects participating in only part II of the experiment differ only by 6%, leading one to possibly conclude that both opponents were nearly indistinguishable from each other for those subjects. Note also the low value of said indices which also indicates that neither opponent seemed overly human-like in behavior or performance. Since these results also overlap with the low subject experience level, it is possible that it is not so much due from the fact that the subjects did not participate in the first part of the experiment but simply from their low exposure to human and agent game opponents in general. However, at the beginning of the experiment, part II, questions were asked (most noticeably by subjects who hadn't participated in part I) regarding how to tell the difference between human and agent robots. Comments were made that the watcher, having not played the game in part I, does not have much to base his judgement on - that he can rate skill, but not necessarily correlate that into a judgement on humanness. It is felt by the author, however, that this is one aspect of the concept of humanness - how exactly is humanness defined? While there isn't necessarily a clear definition of humanness, it is still a perceivable target.

The subjects were not given any set criteria with which to rate humanness of the opponents, and thus they were encouraged to use their own experience and judgement. From the comments received by subjects explaining their reasoning, one can see that they had various methods of rating humanness: some were accurate, while others yielded a completely opposite rating than the true value of the opponent. Even still, the results were well within expected - and correct - values. The human opponents were correctly identified as human 67% and 73% of time in parts I and II respectively. The artificial opponents were mistakenly identified as human 47% and 57% of the time in parts I and II respectively.

As the artificial agents increase in sophistication, it is hoped to see their believability index approach that of the human opponent thus indicating that the artificial agents approach humans in behavior and performance in the general context of simulation.

5. Summary and Future Work

5.1 Summary

As computer game technology continues to grow, the prevalence of artificial intelligence becomes greater. With computer-controlled agents becoming more popular for both entertainment and educational applications, the need for reliable means of testing and comparing the development of the artificial intelligence used within them becomes imperative.

Since human intelligence is the basis to which we compare all others, this comparison can extend into the artificial realm. However, since intelligence itself is difficult to measure, one must find a similar metric with which to make comparisons. That metric is how human-like an entity's behavior and performance are - its 'humanness'.

Turing's test has been used in many forms since its conception in 1950. This has been the staple in evaluating the behavior and performance of artificial agents in applications ranging from testing chatterbots to stopping web abuse. It is also being used in evaluating the performance of artificial agents in games, not unlike what this research has done. However, instead of using the thoughts of an entity as a humanness indicator (and thus equalizing the communicative abilities of both entities involved), this thesis is uses the *behavior* of the entity as the humanness indicator and thus equalizes sensory abilities of both entities (and due to the current technological limitations, this means limiting the human sensorium to that of the AI). Additionally, this methodology is independent of the type of AI used and the environment in which it is tested.

The decision to split the experiment into two parts was to evaluate which, if either, method of interacting with the agents provided the best insight as to their nature. Part I, in first person, had similar results to part II, the third person view. Most of the participants of part II, however, also participated in part I, giving them insight into the difficulty of playing the game as one of the agents.

5.2 Future Work

More research will be necessary to see if this additional experience affected the results for part II or if the two parts of the experiment are essentially redundant. It would also be enlightening to see how subjects' game playing experience affects their assessment of their opponents for either part of the experiment. Further research in the parameterization of the robot and world complexity needs to be done to see the relationship between the perceived humanness of the robot when it is immersed in environments of varying relative complexity. As technology advances it will also be interesting and beneficial to perform studies using AIs with more sophisticated sensing abilities (and the associated cognitive processing as well).

In conclusion, this research has been interesting and has shown promising results for future work. This methodology should provide other researchers a valuable starting point for the assessment of their artificial agents, and with consistent evaluation will come the development and progress that has been so eagerly awaited.

List of Appendices

Appendix A: Code Details

A.1 Flatland

Flatland D, Macintosh is the version used in the experiment. Modification to code outside that constructed for the purposes of this experiment was the addition of the following lines to the MouseKeyboardLocomotion module, at the end of the animateFunc() function:

```
  flSendMessage( TrackerClientObject, "FlatworldClient", buf);
  flSendMessage( TrackerClientObject, "FlatworldServer", buf);
```

where buf is the buffer containing the position and orientation of the vessel within Flatland. The Flatworld modules use this information to correctly position the HUD (heads up display).

The FlatworldServer is the main module created for the experiment. Its necessary code resides in the Flatworld-Server directory found in the Flatland/usr_modules CVS repository. This module conforms to the standard Flatland module layout. The draw callback function is responsible for generating the graphics while the eLoom scheduling function is called in a separate thread and then stores the relevant information in memory accessible by the graphics thread (that which runs the drawing callback function).

The relevant data is then sent to the client module via sockets in yet another thread dedicated to this task. The eLoom thread is what regularly runs the eLoom scheduling function which is the heart of the eLoom/Flatworld interactions. This function does one ‘round’ of data collection, decision making, and action for each robot in the world. The data collection phase consists of using the Flatworld API functions. The decision making phase accesses the eLoom core neural network manipulation functions if the robot is to be controlled by artificial means, or it merely skips if the robot is controlled by a human. The action phase once again uses the Flatworld API functions to perform the desired actions. When this round is complete, the thread records the specified actions of the robot (as determined in the action phase of the scheduler) for use in the second part of the experiment, sleeps a specified time interval, and the cycle continues until one or more robots have depleted their battery reserves. The socket thread is responsible for regularly communicating with the client module which is running on another machine. The socket thread waits until the eLoom thread indicates (via shared memory) that there is new data to send, then it sends the data and receives any new control directives which it then passes to the eLoom thread (via shared memory) for use in the action phase of the human controlled robot. Meanwhile, the graphics thread continually displays the current state of the world and robot to the user.

The FlatworldClient module is very similar to the Server module only that it has no direct interaction with eLoom or Flatworld. It has a drawing callback which, like the server module, is responsible for displaying the current state of the world and robot. It has a separate socket thread which is responsible for the acquisition of data from the server module, and also sends any control directives back to the server. All of the necessary code for the FlatworldClient module can be found in the usr_modules/ FlatworldClient directory in Vis Lab's CVS repository.

For the second phase of the experiment, one Flatland module is used: FlatworldPlayer. This module works with eLoom and Flatworld in exactly the same manner as FlatworldServer. The only difference is that this module reads in control directives from a file rather than listening for user directives or using an artificial agent. The drawing callback function does not display the HUD but rather displays a third person view of the world and all robots within it.

This allows the users to judge both robots simultaneously and without the distraction of playing. The code necessary to run the FlatworldPlayer module can be found in the usr_modules/FlatworldPlayer directory in the vis lab CVS repository.

A.2 Flatworld

Flatworld version 5 is what was used for this experiment. It is similar to version 4 except that it allows for multiple robots to exist peacefully within the world (no robot-robot interactions are currently allowed). The upgrade from version 4 was done by Dr. Caudell and Jessica Ryan. Additionally, the code was changed slightly to allow for the inclusion of the robots as actual, perceivable objects in Flatworld. This allows for them to be seen by other robots and sets the stage for allowing them to be interactable with other robots. No additional modifications were necessary for this experiment. The exact code used in the experiment is found in the FlatworldServer module directory as it is used directly with that module.

A.3 eLoom

eLoom version 1 was used for this experiment, with some modifications done to the user_execution_scheduling_funcs.c file (contains the schedule responsible for doing one ‘round’). Modifications pertained to the upgrade of Flatworld from v4 to v5 - dealing with more than one robot which could be controlled either by human or computer. Additionally, a function was added that supplemented the movement of the robot object type in Flatworld (allowing for the correct visualization of the robot objects). A function that pertains to the movement control of the computer controlled robot was added as well - simulating an artificial neural architecture using conventional algorithms. This function, taking place of the eLoom neural calls, causes the robot to scan its visual sensors until it finds (the first) one that reports the color pattern representing a ‘good’ object. The robot will then turn to face this object and approach it until it eats it or, in the case of the object having been eaten by the opponent before the robot reached it, it reaches the world boundary. If no ‘good’ objects are found, the robot does a random walk, observing world boundaries, until another ‘good’ item is found. All of the relevant and updated code for eLoom resides in the FlatworldServer module directory as it is used directly by this module.

Appendix B: Experiment Documentation

B.1 Informed Consent

Informed Consent for a Study on Artificial Agent Behavior and Performance

Introduction
You have volunteered to participate in a research study conducted by Master's student Jessica Ryan, from the Electrical and Computer Engineering Department at the University of New Mexico. This study is being conducted for inclusion in a Computer Engineering Master's thesis.

You have been selected for this study because you have volunteered with the understanding that there are no risks or benefits to your person involved. By signing this consent for you acknowledge that you have no medical issues that stand in the way of your use of standard computers or viewing of projector screens.

Purpose of the Study
The purpose of this experiment is to test the capabilities of a proposed methodology for the assessment of the behavior and performance of artificial agents.

Procedure
Your participation in this experiment is strictly voluntary and you receive no compensation. The experiment will last no longer than an hour and will involve either:

□ Up to one hour of participation in a series of simple computer games and the completion of a survey related to this experience. The survey asks the following questions:
- Age and gender
- General gaming experience (on a scale of 1 to 5)
- For each game:
  - a rating of the opponent (on a scale of 1 to 5)
  - a subjective explanation of your decision

□ Up to one hour of participation in the viewing of a series of movies of pre-recorded computer games and the completion of a survey related to this experience. The survey asks the following questions:
- Age and gender
- General gaming experience (on a scale of 1 to 5)
- For each movie:
  - a rating of each game player (on a scale of 1 to 5)
  - a subjective explanation of your decision

Potential Risks and Discomfort

There are no psychological risks associated with this experiment. It possible, but rare, that you may experience typical symptoms of computer use such as pain associated with carpel tunnel syndrome or motion sickness. Individual susceptible to such symptoms should choose not to participate, and at any time during the experiment if you wish to discontinue, you may do so.

Potential Benefits to Participants and Society

The are no individual benefits of this study other than your entertainment. The benefits gained from this research effect mostly society as a whole rather than individual participants. As artificial agents are becoming inherently more common in society, defining a useful and accurate methodology for their behavior and performance assessment is key in the positive and fruitful development and research of said agents.

Confidentiality

All information obtained in connection with this study will not be identifiable with you and thus there is no risk for any breaches in your privacy.

Participation and Withdrawal

You can choose whether or not to participate in this study. If you volunteer to participate, you may withdraw at any time without penalty. You may also refuse to answer any questions that you do not want to answer and still remain in the study. The investigator may withdraw you from this research if circumstances arise which warrant doing so.

Identification of Investigator and Review Board

If you have any questions or concerns about this research, please feel free to contact: Jessica Ryan (jryan@ece.unm.edu) or Dr. Thomas Caudell (tpc@ece.unm.edu). If you have other concerns or complaints, contact the Institutional Review Board at the University of New Mexico, Dr. William Gannon, Chair Human Subjects Institutional Review Board (wgannon@unm.edu, (505) 277-3488) for more information.

I understand the procedures described above. My questions have been answered to my satisfaction, and I agree to participate in this study. I have been provided a copy of this form.

Participate Name (printed)	Participant Signature	Date
Jessica Ryan		
Investigator Name (printed)	Investigator Signature	Date

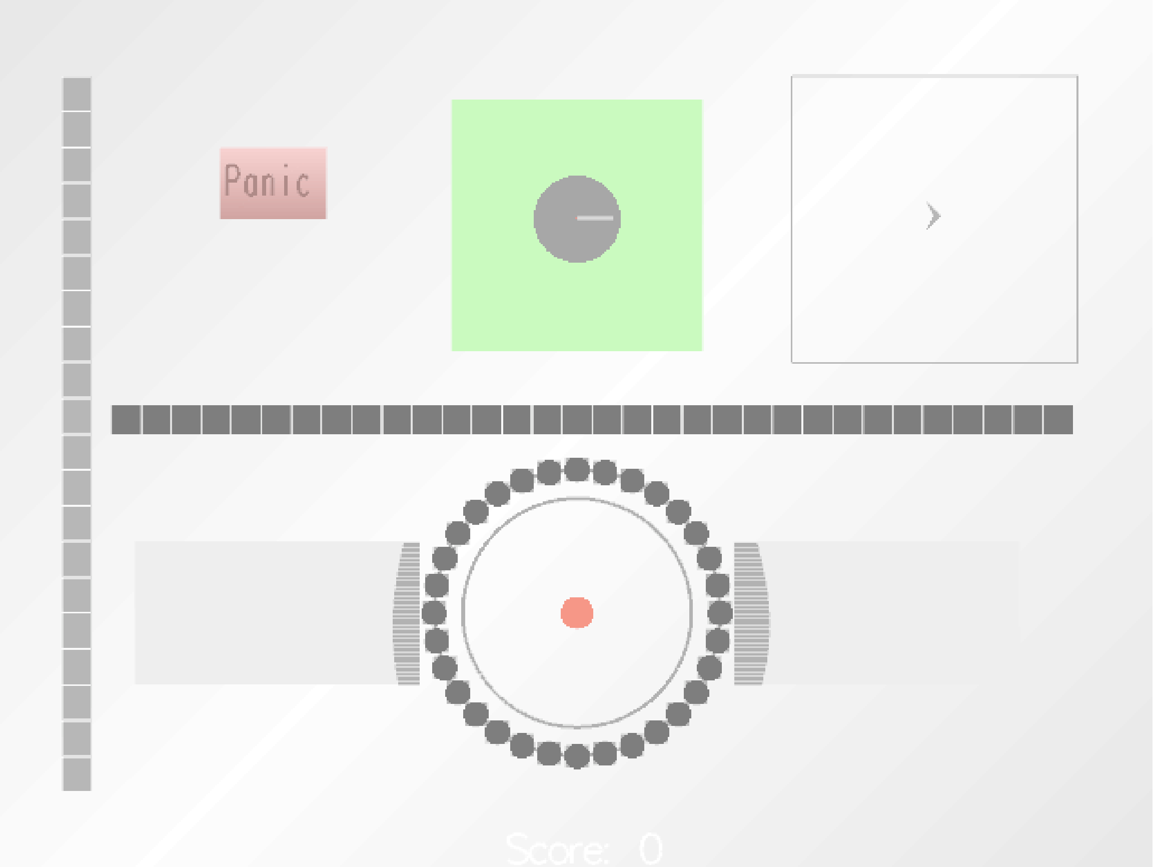
B.2 Subject Briefing

Assessment of the Behavior and Performance of Artificial Agents

Instructions and Explanation

Game Overview:
Your mission, should you chose to accept it, is to play this simple game and evaluate your opponent's 'humanness'. The world in which you will be playing is a very simple world with 4 types of objects: robots, food objects, neutral objects, and poison objects. There will be two robots - you and your opponent. The non-robot objects are stationary and each type has its own distinct shape, color, and audio frequency pattern. The food objects will charge your battery a fraction, the neutral objects will have no effect, and the poison objects will discharge your battery a fraction.

Your perceptions are limited to the same observations that your opponent has. The game display, with explanations, is shown below:



1. Life meter - this indicates the battery level, or life, of your robot. When it reaches zero, your robot dies and the game is over.
2. Panic button - this stops all movements and resets the robot's eat flag.
3. Orientation - this serves as both an indicator and a control for robot orientation. The white dash indicates the front of the robot. When moving forward or backward, a line extends out that indicates the speed and direction of the movement. To turn the robot, you may click anywhere in the green field and the robot should turn towards the point clicked. This is still buggy however and sometimes the robot behave as expected.
4. Location - this is a vague map that gives you a general sense of where your robot is in the world. It is not exact and doesn't give information pertaining to the world size, however. It is merely to aid in your sense of movement and orientation.
5. Visual sensors

    a) this is a flat array of all 32 visual sensors to help you construct a 1-dimensional image of what your robot is sensing. You may click on a sensor to have the robot turn towards and face that direction (useful for targeting a perceived object).

    b) this is a circular array of the same visual sensors as above, but as they are located on the robot's perimeter. This arrangement of the sensors gives you a spatial feel for the location of the perceived objects as they relate to robot position. You may click on a sensor to have the robot turn towards and face that direction (useful for targeting and orientation in general).
6. Audio sensors - these are the robot's ears, one located on each side as indicated in their position on the diagram. Each object gives off a distinct frequency pattern and with practice you can identify objects aurally. The audio sensors are useful for targeting faraway objects, and staying within the location of the objects (ie: not walking into oblivion). Each line represents one frequency, and the larger the lines, the louder the frequency. The robot does correlate what it hears, meaning what you see is a mix of all of the frequencies within hearing range of the robot. It is up to you to discern what this means.
7. Touch sensor - this ring glows blue when the robot is in contact with one or more objects. It is possible to move through objects and be in contact with more than one at a time. Use your visual and audio sensors to identify the object and determine if you are touching more than one at once.
8. Eat indicator - when this indicator is red, your robot will NOT eat any objects with which collides. When it is green it will eat any objects with which it is in collision. Take care not to have this on when unintended.

Controls:
You control your robot using a combination of mouse and keyboard. You may use keyboard only, if you wish. The mouse is used as described above: clicking on the panic button to stop all movement, clicking on the green orientation area to turn the robot (not the most reliable), and clicking on any visual sensor you wish to turn towards and face.

Keyboard controls:

‘a’ - panic: stop all movement and reset eat flag to off
‘s’ - turn left
‘f’ - turn right
‘e’ - move forward
‘d’ - move backward
‘x’ - strafe left
‘c’ - strafe right
‘space’ - toggle the eat flag

A note on movement: The robot will continue in the direction of movement indicated, ie: with one press of the ‘e’ key, the robot will continue to move forward until the action is cancelled by either pressing once on the ‘d’ key or by pressing the panic button/key. The same goes for turning: one press of the ‘s’ key will cause the robot to continuously turn until you cancel the action by turning in the opposite direction, ‘f’, or press panic. You may move in bigger increments by pressing a key multiple times, for instance move forward in bigger ‘steps’ by pressing the ‘e’ key multiple times. Consequently it will take that many times of pressing the ‘d’ to come to a stop. The same goes for turning. It is possible to turn and go forward at the same time. It is also possible to be cruising along and use the mouse to click on a visual sensor to change direction, though if you are traveling too fast, you may miss your intended target.

Objective:
The goal of the game is to survive as long as possible. The goal of my research is for you to play the game long enough to get a feel for the opponent robot and make a judgement on its humanness. Your opponent may or may not be controlled by a human. Based on what you observe in the game (how quickly it wins, its action pattern, your gut instinct, etc) you will rate its humanness using the given questionnaire.

Instructions:
You will be given up to 10 minutes to play the game in practice mode in order to get a feel for the controls and to learn how to discern objects and determine which ones are beneficial and which are not. Following that you will be guided through a series of short games (up to nine) in which your opponents may or may not be controlled by a human. Following each game, please indicate on the questionnaire your judgement of the humanness of the opponent and give a brief explanation of how you came to this conclusion.

THANK YOU!!

B.3 Assessment Tool Phase I

Assessment of Behavior and Performance of Artificial Agents

Part I - First Person Assessment

Please answer the following:

Age:______ Gender:______

General Gaming Experience:

1. Never or rarely play
2. Sometimes play (infrequently)
3. Occasionally play (monthly)
4. Regularly play (weekly)
5. Frequently play (daily)

Please answer the following for each game played:

Game 1:

Humanness of opponent:

1. Human
2. Probably human
3. Don't know
4. Probably artificial
5. Artificial

Reason:

Game 2:

Humanness of opponent:

1. Human
2. Probably human
3. Don't know
4. Probably artificial
5. Artificial

Reason:

Game 3:

Humanness of opponent:

1. Human
2. Probably human
3. Don't know
4. Probably artificial
5. Artificial

Reason:

Game 4:

Humanness of opponent:

1. Human
2. Probably human
3. Don't know
4. Probably artificial
5. Artificial

Reason:

Game 5:

Humanness of opponent:

1. Human
2. Probably human
3. Don't know
4. Probably artificial
5. Artificial

Reason:

Game 6:

Humanness of opponent:

1. Human
2. Probably human
3. Don't know
4. Probably artificial
5. Artificial

Reason:

B.4 Assessment Tool Phase II

Assessment of Behavior and Performance of Artificial Agents

Part II - Third Person Assessment

Please answer the following:

Age:______ Gender:______

If you participated in part I, please check:
□

General Gaming Experience:

1. Never or rarely play
2. Sometimes play (infrequently)
3. Occasionally play (monthly)
4. Regularly play (weekly)
5. Frequently play (daily)

Please answer the following for each movie viewed:

Movie 1:

Humanness of red player:

1. Human
2. Probably human
3. Don't know
4. Probably artificial
5. Artificial

Humanness of blue player:

1. Human
2. Probably human
3. Don't know
4. Probably artificial
5. Artificial

Reasons:

Movie 2:

Humanness of red player:

1. Human
2. Probably human
3. Don't know
4. Probably artificial
5. Artificial

Humanness of blue player:

1. Human
2. Probably human
3. Don't know
4. Probably artificial
5. Artificial

Reasons:

Movie 3:

Humanness of red player:

1. Human
2. Probably human
3. Don't know
4. Probably artificial
5. Artificial

Humanness of blue player:

1. Human
2. Probably human
3. Don't know
4. Probably artificial
5. Artificial

Reasons:

References

[1] Rost, S. (2004) "Evolution of Formidable Player AI in Tactical 3D Games" www.mit.edu. M Cambridge.

[2] Miyamoto, S. (1985) "Super Mario Bros." Nintendo Entertainment System. Nintendo.

[3] Hollis, M. (1997) "GoldenEye 007" Nintendo 64. Rareware.

[4] Stone, R. (December 2005) "Serious Gaming" Defence Management Journal, 31, pgs: 142-144.

[5] Arnseth, H. C. (December 2006) "Learning to Play or Playing to Learn - A Critical Account of the Models of Communication Informing Educational Research on Computer Gameplay" The International Journal of Computer Game Research, Vol. 6, No. 1.

[6] Alverson, D. C., et al. (2005) "Distributed Immersive Virtual Reality Simulation Development for Medical Education" JIAMSE Vol. 15, pgs: 19-30.

[7] Genesereth, M., Love, N., Pell, B. (2005) "General Game Playing: Overview of the AAAI Competition" AI Magazine, Summer 2005, pgs: 62-72.

[8] Turing, A. M. (1950) "Computing Machinery and Intelligence" Mind, 49, 433-460.

[9] Saygen, A. P., Cicekli, I., and Akman, V. (2000) "Turing Test: 50 Years Later" Minds and Machines, 10, pgs: 463-518.

[10] Gunderson, K. (April 1964) "The Imitation Game" Mind, New Series, Vol. 73, No. 290, 234-245.

[11] http://www.loebner.net/Prizef/loebner-prize.html

[12] Kochanski, G., Lopresti, D., Shih, C. (2002) "A Reverse Turing Test Using Speech" Proceedings of the Seventh International Conference on Spoken Language Processing

[13] "CAPTCHA" www.wikipedia.com

[14] Rui, Y., Zicheng, L. (May 2004) "ARTiFACIAL: Automated Reverse Turing test using FACIAL features" ACM Multimedia Systems Journal. Springer.

[15] Grünvogel, S. M. (October 2005) "Formal Models and Game Design" Game Studies: the International Journal of Computer Game Research, Vol. 5, No. 1.

[16] Dutton, N., Consalvo, M. (December 2006) "Game analysis: Developing a methodological toolkit for the qualitative study of games" Game Studies: the International Journal of Computer Game Research, Vol. 6, No. 1.

[17] Sterman, J.D. (February 1986) "Testing Behavioral Simulation Models by Direct Experiment" http://hdl.handle.net/1721.1/2133 Sloan School of Management, Massachusetts Institute of Technology.

[18] Laird, J.E. and Duchi, J. C. (2001) "Creating Human-like Synthetic Characters with Multiple Skill Levels: A Case Study using the Soar Quakebot" American Association for Artificial Intelligence.

[19] Livingstone, D. (January 2006) "Turing's Test and Believable AI in Games" ACM Computers in Entertainment, Vol. 4, No. 1.

[20] Gorman, B., Thurau, C., Bauckhage, C., and Humphrys, M. (2006) "Believability Testing and Bayesian Imitation in Interactive Computer Games" Proceedings of the 9th International Conference on the Simulation of Adaptive Behavior (SAB '06), Vol. LNAI. To appear.

[21] Caudell, T. (September 2003) "Guide to Flatland D" University of New Mexico.